ENERGY 2012

The Second International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies

ISBN: 978-1-61208-189-2

March 25-20, 2012

St. Maarten, Netherlands Antilles

ENERGY 2012 Editors

Pascal Lorenz, University of Haute Alsace, France

Kendall Nygard, North Dakota State University, USA

ENERGY 2012

Foreword

The Second International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY 2012), held between March 25-30, 2012 - St. Maarten, Netherlands Antilles, continued the inaugural event considering Green approaches for Smart Grids and IT-aware technologies. It addressed fundamentals, technologies, hardware and software needed support, and applications and challenges

There is a perceived need for a fundamental transformation in IP communications, energy-aware technologies and the way all energy sources are integrated. This is accelerated by the complexity of smart devices, the need for special interfaces for an easy and remote access, and the new achievements in energy production. Smart Grid technologies promote ways to enhance efficiency and reliability of the electric grid, while addressing increasing demand and incorporating more renewable and distributed electricity generation. The adoption of data centers, penetration of new energy resources, large dissemination of smart sensing and control devices, including smart home, and new vehicular energy approaches demand a new position for distributed communications, energy storage, and integration of various sources of energy.

We welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard forums or in industry consortia, survey papers addressing the key problems and solutions on any of the above topics short papers on work in progress, and panel proposals.

We take here the opportunity to warmly thank all the members of the ENERGY 2012 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to ENERGY 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We hope that ENERGY 2012 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in energy-aware technologies.

We are certain that the participants found the event useful and communications very open. The beautiful places of St. Maarten surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

ENERGY 2012 Chairs
Pascal Lorenz, University of Haute Alsace, France
Petre Dini, Concordia University - Montreal, Canada / China Space Agency Center - Beijing, China

ENERGY 2012

Committee

ENERGY 2012 Technical Program Committee

Amir Abtahi, Florida Atlantic University - Boca Raton, US
Nizar Al-Holou, University of Detroit Mercy, USA
Cosimo Anglano, Universitá del Piemonte Orientale - Alessandria, Italy
Luca Ardito, Politecnico di Torino, Italy
Mehdi Bahrami, Islamic Azad University, Iran
Fabio Luigi Bellifemine, Telecomitalia, Italy
Frede Blaabjerg, Aalborg University, Denmark
Gerben Broenink, TNO, The Netherlands
Antonio Caló, NorTech Oulu -Thule Institute / University of Oulu, Finland
Davide Careglio, Universitat Politècnica de Catalunya - Barcelona, Spain
Mari Carmen Domingo, Barcelona Tech University, Spain
Dave Cavalcanti, Philips Research North America - Briarcliff Manor, USA
Mohamed Cheriet, ETS/GreenStar - Montreal, Canada
Howard Choe, Combat Systems - Plano, USA
Delia Ciullo, Politecnico di Torino, Italy
Peter Corcoran, College of Engineering & Informatics, NUI Galway, Ireland
Margot Deruyck, Ghent University/IBBT, Belgium
Marco Di Girolamo, Hewlett-Packard Company, Italy
Yong Ding, Karlsruhe Institute of Technology (KIT), Germany
Venizelos Efthymiou, Electricity Authority of Cyprus, Cyprus
Eugene A. Feinberg, Stony Brook University - New York, USA
Alexandre Peixoto Ferreira, IBM Austin Research Laboratory, USA
Steffen Fries, Siemens Corporate Technology - Munich, Germany
Brian P. Gaucher, IBM Research Division - Yorktown Heights, USA
Erol Gelenbe, Imperial College London, UK
Hamid Gharavi, National Institute of Standards and Technology, USA
Georgios B. Giannakis, University of Minnesota - Minneapolis, USA
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Manimaran Govindarasu, Iowa State University, USA
Mesut Günes, Freie Universität Berlin, Germany
Helmut Hlavacs, University of Vienna, Austria
Chun-Hsi Huang, University of Connecticut - Storrs, USA
Canturk Isci, IBM TJ Watson Research Center, - New York, USA
Philip Johnson, University of Hawaii - Honolulu, USA
Young Sun Kim, Korea Electrotechnology Research Institute (KERI), Korea
Dejan Kostic, EPFL, Switzerland
Paul J. Kühn, University of Stuttgart, Germany
DongJin Lee, University of Auckland, New Zealand
Eugene Litvinov, ISO New England, USA
Jaime Lloret Mauri, Universidad Politécnica de Valencia, Spain

Thair Shakir Mahmoud, Edith Cowan University - Western Australia, Australia
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Jean-Marc Menaud, Ecole des Mines de Nantes, France
Avi Mendelson, Microsoft R&D, Israel
George Michailidis, University of Michigan, USA
Nicolas Montavont, Telecom Bretagne, France
Daniel Mossé, University of Pittsburgh, USA
Gero Mühl, Universität Rostock, Germany
Giorgio Nunzi, NEC Laboratories Europe, Germany
Dragan Obradovic, Siemens AG - München, Deutschland
Jacob Østergaard, Technical University of Denmark, Denmark
Cathryn Peoples, University of Ulster, UK
Mirek Piechowski, Meinhardt (Vic) Pty Ltd. - Melbourne, Australia
Jean-Marc Pierson, Université Paul Sabatier - Toulouse, France
Philip W. T. Pong, The University of Hong Kong, Hong Kong
Darren Robinson, University of Nottingham, UK
Sebastian Rohjans, OFFIS - Institute for Information Technology - Oldenburg, Germany
Sebnem Rusitschka, Siemens AG - München, Germany
Eliot Salant, IBM Haifa Research Labs / Haifa University, Israel
Dave Saraansh, University of Bristol / Toshiba Research Europe Limited, UK
Dirk Uwe Sauer, ISEA / RWTH - Aachen University, Germany
Harald Schrom, Technische Universitaet Braunschweig , Germany
Sandra Sendra, Universidad Politécnica de Valencia, Spain
Gerard Smit, University Of Twente - Enschede, The Netherlands
Pavel Somavat, CCS Haryana Agricultural University, India
Grzegorz Swirszcz, IBM Watson Laboratory, USA
Zhibin Tan, Wayne State University - Detroit, USA
Mahmoodi Toktam, King's College London, UK
Bernard Tourancheau , University Joseph Fourier of Grenoble, France
Jean-Philippe Vasseur, Cisco Systems, Inc., France
Stéphane Vialle, SUPELEC, France
Le Yi Wang, Wayne State University, USA
Guanghua Yang, The University of Hong Kong, Hong Kong
Francis Zavoda, Hydro-Quebec, Canada
Albert Zomaya, University of Sydney, Australia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Analysis of Performance of CCHP Systems for Large Hospitals

Francesco Patania, Antonio Gagliano, Francesco Nocera, Aldo Galesi
Department of Industrial Engineering and Mechanics
University of Catania, Catania, Italy
{fpatania, agagliano, fnocera, agalesi}@diim.unict.it

Abstract—**Hospitals are large public buildings that have a significant impact on the environment because they use large amounts of energy and water and produce large amounts of waste. For these reasons, hospitals are natural candidates for sustainable design. The design of heating ventilation and air conditioning (HVAC) plants for large hospitals must aim to some top priority objectives: improve their energy efficiency, forecast the use of clean innovative technologies of self production of energy, reduce both the operating costs of HVAC plants and polluting impacts on environment, guarantee the continuity of energy supply from every case of critical states and black-out of electrical energy or natural gas. In the context of cooling/heating efficient energy, we illustrate the benefits obtained by means the use of a system of "Combined Cooling Heating and Power" for the San Marco Catania's Hospital. This paper shows in what way is guaranteed the continuity of energy supply for any possible critical state, the economic gain through the auto-production of energy that reduces of about 1.000 M€ the operating cost of HVAC plants, the environmental benefits due to the reduction of about 25 Mtons of carbon dioxide equivalent.**

Keywords-HVAC systems; saving energy; CCHP; greenhouse emissions.

I. INTRODUCTION

Hospitals are large public buildings that have a significant impact on the environment and economy of the surrounding community. They use large amounts of energy and water and produce large amounts of wastes so they are natural candidates for sustainable design. In today's expanding energy hungry world, sustainability is no longer an option, it has become the design standard for design professionals.

The Combined Cooling Heat and Power (CCHP) plant provide cooling alongside heat and power from the same energy source into a 'tri-generation' scheme.

CCHP plants have a higher efficiency than systems that producing only heating or power because CCHP system uses waste heat from electricity generation to produce steam for heating and cooling. Fig. 1 shows a typical scheme of CCHP plant.

The performance of CCHP systems has been studied before [1][2][3]. The CCHP plants are particularly useful for buildings, like hospitals, that have large amounts of air conditioning needs.

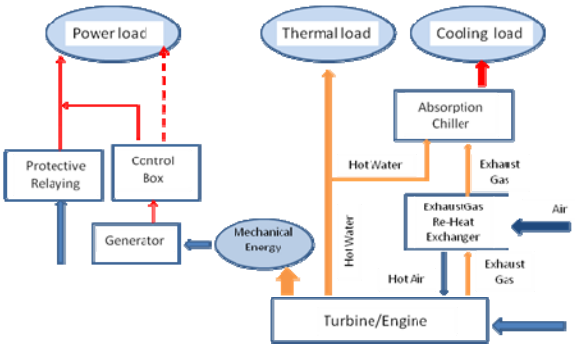


Figure 1. Typical CCHP system

Because CCHP uses waste heat to produce thermal energy for heating and cooling, hospitals equipped with CCHP systems are more energetically efficient.

Hospitals are ideal candidates for CCHP systems because they work 365 days for year and require round the clock energy. In the world, the number of hospitals using CCHP systems has grown steadily in recent years. Combined systems enable hospitals to reduce energy costs, increase reliability on energy availability and improve environmental performance.

With less fuel consumed, greenhouse gases (GHG) and air pollutants (nitrogen oxides and sulfur dioxides) are reduced [4]. Resources saved could be redirected to improve patient care [5]. Moreover, hospitals must perform critical, life-saving functions even when a widespread disaster interrupts their supply of natural gas and electricity from the utility grid. CCHP systems can be designed to maintain critical life-support systems, operate independently of the grid during emergencies and be capable of black start (the ability to come online without relying on external energy sources). Blackout of electrical national grid or further possible negative happenings (terrorist outrages, oil crisis caused by war events in oil producing countries, and so on) directs people efforts to have more sources of energy supplying and reach the maximum possible of energetic independence of the hospital. Because of they are already up and running, CCHP systems can offer a seamless, reliable power alternative than traditional emergency generators.

The CCHP systems, however, are not automatically configured for backup capability; hospitals must ensure that

 ISBN: 978-1-61208-189-2

the systems have automatic transfer capability and that the energy output can be matched with energy demand.

The CCHP systems are not always advantageous for all typologies of hospitals: their cost-effectiveness needs to be evaluated on a case by case basis. In determining a CCHP system's viability, there are several important considerations: interconnection agreements, site issues, and permits all should be discussed with the partnering utility. Local and national incentives including direct financial grants, tax incentives, and low interest loans should be determined. The type of driving unit and cooling device distinguish different kinds of CCHP systems. A CCHP driving unit module can be a steam turbine, a gas turbine, a reciprocating engine, or a fuel cell. A turbo or absorption chiller generally produces the cooling energy from a CCHP system, the choice depends on the required output power and operating regime [6].

Designers must choose equipment that best fits the hospital's thermal and electrical loads and power quality requirements. The CCHP systems often can be installed for less cost upfront than renewable energy options such as photovoltaic systems of a similar scale. When matched to suitable loads, some CCHP systems can provide a simple payback in the five to ten year range, depending on system size and energy costs [7].

This work provides a quantitatively analysis of the significant benefits that can be brought by the CCHP systems.

The paper has the following structure: Section II describes the main characteristics of the building complex; Section III reports the main energy flow and the financial advantages obtainable by the CCHP system. We conclude in Section IV.

II. THE SAN MARCO HOSPITAL

The "San Marco Hospital and Orthopedically Institute of Excellence" is located in the metropolitan area of Catania city, close to many facilities as like as airport, harbor and link roads with Sicilian motorway as Figure 2. It takes up an area of about 230.500 m^2, the buildings of whole complex cover a surface about 28.500 m^2 and the total buildings cubature is about 405.000 m^3.

Figure 2. San Marco Hospital and Orthopaedical Institute of Excellence with Building Area in evidence

Figures 3 and 4 show some views of hospital's buildings. San Marco Hospital is an ideal candidate for a CCHP system because it operates 365 days for year and requires round the clock energy. A combined system could enable the hospital to reduce energy costs, improve environmental performance, and increase energy reliability. During the Italy blackout of electrical national network of 28 September 2003, many hospitals had failures in their backup power generators. Blackout of electrical national grid and further possible negative happenings (terrorist outrages, oil crisis caused by war time events in oil producing countries, anomalies in national electrical distribution system, and so on) directs people efforts towards design of HVAC plants in such a way both to render energy supplying not dependent by an unique source and to reach the maximum possible as regard to energetic independence of the hospital.

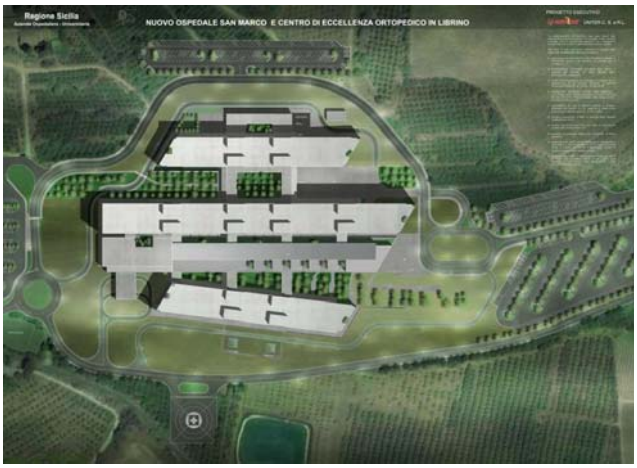


Figure 3. San Marco Hospital and Orthopaedical Institute of Excellence

Figure 4. San Marco Hospital and Orthopaedical Institute of Excellence

A. *Hospital's energy needs*

The Hospital's thermal needs (cooling, space heating, domestic hot water, steam, and ventilation) have been calculated using the MC4 software [8], which is a multi-use software program that calculates load and energy efficiency

 ISBN: 978-1-61208-189-2

for any type of building. The electrical needs (lighting, large medical equipment, distributed medical equipment & plug loads) have been calculated in function of the electric power of all the utilities operating simultaneously.

The calculated power peak loads are shown in Table I.

TABLE I. DATA OF PEAK DEMAND

Thermal power (Winter)	**kW**
Domestic Hot Water	1.425
Space Heating and Ventilation	5.607
Hot Water Separated Circuits	1.415
Various Utilities	1.650
General Amount	*10.097*
Steam needs (Summer)	**kg/h**
Domestic Hot Water	2.052
Absorption Chillers	18.400
Ventilation	1.346
Various Utilities	2.376
General Amount	*24.174*
Electrical power (Summer)	**kW**
Ventilation	815
Indoor Lighting	350
Outdoor Lighting	60
Electro-medical Equipments	1.200
Chillers	670
Operating Theatres	560
Various Utilities	936
General Amount	*3.366*
Cooling power (Summer)	**kW**
Space Cooling, Ventilation	10.392
Cold Water Separated Circuits	679
General Amount	*11.071*

B. *Architecture of the energy system*

Hospitals have to be able to generate their own base load power, in conjunction with or as a backup for the main electricity in the event of an unexpected loss of electricity and/or emergency. System energy supply has been designed to operate independently of the grid during emergencies, and to be able to come online without relying on external energy sources maintaining the critical life support systems. For these reasons, designers have foreseen to utilize three distinct energetic sources:

- Natural gas, to feed the gas turbine during normal service and, in case of emergency or for request of peak loads , three steam generators
- Electric energy coming from national network during periods of maintenance or out of order of turbine.
- Gas oil to feed in emergency (lack of natural gas) three steam generators.

The architecture of the energy system feeding to the various machines and equipments of HVAC plant is shown in Figure 5.

A turbine using natural gas produces electric power. Gas turbine produces high quality exhaust heat that can be used in CCHP configurations to reach overall system efficiencies (electricity and useful thermal energy) of 70 to 80 percent [6][9].

Another essential directive has been design and built a power plant that as far as possible would be compatible with the environment, by drawing on the best available techniques for the production processes and the process machinery.

Based on the calculations to satisfy the energy demand of hospital, the designers have forecast the machines and equipments as following summarized:

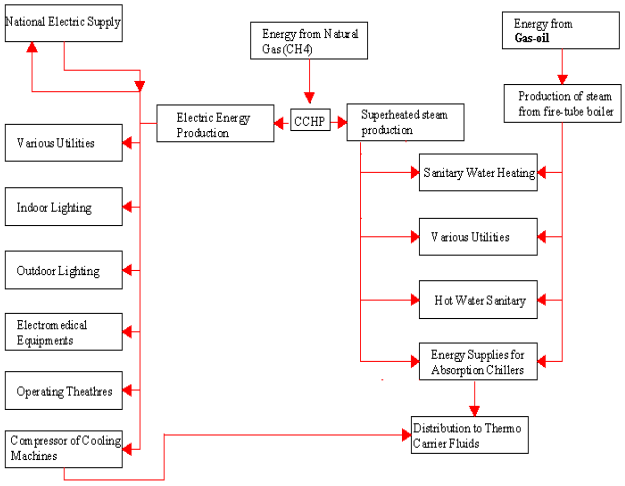


Figure 5. Architecture of System

- gas turbine, able to produce 5.250 kW of electric power
- unfired heat recovery steam generator that produces 12.750 kg/h of steam at 10 bar used into a thermal loop for hospital space heating during winter months or to feed single effect absorption chillers that provide cooling power during the summer
- electric power generators, able to produce 600 kW of electric power
- three steam generators, each one able to produce about 5.000 kg/h of steam, that is about 2.780 kWt
- Two electric chillers, each one able to produce 4.079 kW of cooling power
- Two absorption chillers, each one able to produce 4.037 kW of cooling power

For standard operating conditions, the energetic demands will be satisfy as shown in Table II.

TABLE II. POTENTIAL OF SUPPLYING

Peak Demand	**Machines**	**Supply**	**%**
Thermal (kW)	Heat recover steam generator	8.850	87
10.097	Steam generator	2.730	23
	Total	*11.580 kWt*	100
Steam (kg/h)	Heat recover steam generator	12.750	53

 ISBN: 978-1-61208-189-2

24.174	Steam generators (n.3)	15.000	47
	Total	*27.750 kg/h*	
Cooling(kW)	Absorption chillers (n.2)	8.174	74
11.071	Electric Chiller (n.1)	4.079	26
	Total	*12.253 kWf*	100
Electric (kW)			
4.591	Turbo-alternator	5.125 kWe	100

The plant has supplemented by other machines and equipments like: Air Treatment Units (ATU), cooling towers, compressors, pumps, tanks, vessels, heat exchangers and so on. The Gas Turbine is able to produce about 5,25 MWe at steady state condition, and, in meantime, to discharge a flow of about 87.840 kg/h of exhaust gas at 500°C.

The thermal energy contained in exhaust gas is utilized to feed the heat recovery steam generator able to generate about 12.750 kg/h of superheated steam in is turn utilized to feed subsequent thermal exchanges :

- heat exchangers to produce hot water at 80°C for the heating utilities;
- heat exchangers to produce hot sanitary water at 48.5 °C, whilst the temperature forecasted in hot water storage tank is about 62°C
- the two absorption chillers
- post heating exchangers of air treatment units (ATU)
- various utilities ad equipments of hospital

C. Critical States of Energy Supplying

There are three possibilities to have critical state for the supplying of energy to systems:

- Critical state n.1: blackout in national electric network.
- Critical state n.2: ordinary or extraordinary maintenance of the gas turbine
- Critical state n.3: blackout for the grid distribution of natural gas

For each one of previous critical states, thanks both to three possibilities of energy supply and to all the machines and equipments foreseen, the hospital system will be always able to offer health services to the users. The strategies to solve the problem of energy supply, caused by previous critical states, are shown in Table III.

TABLE III. ENERGY SUPPLY IN CRITICAL EVENTS

Critical State 1	**Peak Demand MW**	**Machines and Equipments avalaible**	**Available Power MW**	**% of backup**
Blackout in National Electric Network	Thermal load 10,10	Heat recovery steam generator n.1 Steam generator	8,85 2,73 **11,58**	100
	Cooling Load 11,07	n.2 Absorption chillers n.1 Chiller	8,17 4,08 **12,25**	100
	Electric Load 11,07	Gas Turbine	**5,12**	100
	Steam load 24,17 t/h	Heat recover steam generator n.3 Steam generators	12,75 15,00 **27.750 t/h**	100
Critical State 2	**Peak Demand MW**	**Machines and Equipment**	**Available Power MW**	**% of backup**
Gas Turbine Maintenance or fault	10,10	n.3 Steam generators	**8.340 kWt**	82
	11,07	n.2 Absorption chillers	**8.158 kWf**	74
	11,07	National electric network	**4.591 kWe**	100
	14,47 t/h	n.3Steam generators	**15,0 t/h**	100
Critical State 3	**Peak Demand MW**	**Machines and Equipment**	**Available Power MW**	**% of backup**
Blackout for Gas Supply (only Gas-oil storage 60.000 l)	10,10	n.3 Steam generators (Gas-oil feeding)	**8,34**	82%
	11,07	n.2 Chillers	**8.158**	74%
	4,59	National electric network	**4,591**	100
	17,66 t/h	n.3 Steam generators (Gas-oil feeding)	**15,0 t/h**	85%

The storage tank of design containing about 60.000 l of Gas oil will be able to satisfy energetic demand of critical states reported in Table III for 60 hours; that is one period of time quite enough to eliminate the cause of critical state.

III. ENERGY FLOW AND FINANCIAL ADVANTAGES

CCHP often can offer financial advantages over power purchased from a local utility or produced by other energy systems. Moreover, CCHP system creates an additional revenue stream by allowing hospitals to sell surplus electricity back to their utilities. A hospital's ability to do this depends on the net metering and rate policies of its utility. Typically, "selling back" during off-peak hours is not profitable for a hospital, but, given the right circumstances, it can be a revenue generator during peak hours.

The baselines of Italian rules starts up the so called "Green Certificates Market" that allows to producers of electric energy by alternative sources to sell their Green

Certificates to producers that do not make use renewable sources.

In accordance with Italian legislation [10][11][12], on the efficient use of fuels, the CHP plants can obtain incentives, like exemption from having to buy Green Certificates and priority dispatching of the electricity it produce, if the Energy Recovery Index (ERI) and the Thermal Limit (TL) are, respectively, at least 10% and 15%.

The Energy Recovery Index quantifies the saving of primary energy achieved by a section of cogeneration compared to the separate production of the same quantities of electrical and thermal energy.

The Thermal Limit quantifies the amount of useful thermal energy produced annually compared to the total production of electricity and heat.

Another verification is request for the Energy Index (Ien) and it fixes the smallest values (0,51) to identify CCHP as renewable source;

The Energy Recovery Index defined by Decree No. 79/99, corresponding to the PES (Primary Energy Saving) introduced with the European directive 2004/8/CE that fixes the "high efficiency of plants" if it is verified a PES more than 10%. More punctual technical definition of previous indexes are reported in Italian Regulation previously referred.

The design of plants has been developed with the aim even to produce yearly excess of electric energy in relation to that one strictly necessary to satisfy the demand of the hospital. In this way it is possible to market the energy surplus and to have yearly economic benefits.

Table IV shows the values of main energy indexes of designed CCHP compared with that one required by technical manager of the Hospital. The value obtained by designers confirms that the CCHP system may be considered, in force of Italian regulation, as a renewable source. Table IV shows the value of main energy indexes of designed CCHP in relation to that one required by terms of contract by technical manager of Hospital Enterprise. The value obtained by designers confirms that the plants may be considered, in force of Italian regulation, as renewable sources.

TABLE IV. ENERGY INDEXES

	by Contract	by Design
Ien	0,51	0,63
TL	0,50	0,65
ERI/ PES	0,25	0,28

On the basis of these indexes values, the works council of Hospital will be able to accede in Green Certificate Market to sell that amount of electric energy yearly produced but none self-consumed.

People have calculated the monthly financial fluxes shown in Table V. The reported data have calculated considering the energy fluxes, reported in Fig.6, and the average prices of free market both for natural gas (source A.S.E.C. S.p.a: €/m3 0,48) and electric energy (source Enel €/kWh 0,09 average over 24 hours).

The CCHP systems give also considerable results in terms of respect of environment. In fact these systems have higher whole system efficiency than systems that split heating and power generation. With less fuel consumed, GHG and air pollutants (like nitrogen oxides and sulphur dioxides) are reduced more than 25,000 tons per year.

TABLE V. *MONTHLY FINANCIAL FLUXES*

Month	Costs for natural gas supply (1) €	Costs savings for Electric Energy (2) €	Revenues for Sales of Electric Energy (3) €	Financial fluxes (4=2+3-1) €
January	272.237	246.297	161.001	135.061
February	246.340	222.461	130.848	106.969
March	189.064	158.910	32.350	2.196
April	263.937	238.352	191.290	165.705
May	272.734	246.297	182.958	156.521
June	264.204	238.352	15.566	-10.286
July	454.848	351.800	123.263	20.215
August	454.848	351.800	189.134	86.086
September	104.928	238.352	81.437	214.861
October	142.642	175.937	50.716	84.011
November	175.887	175.937	61.694	61.744
Total Yearly	3.114.404	2.820.432	1.270.973	977.001

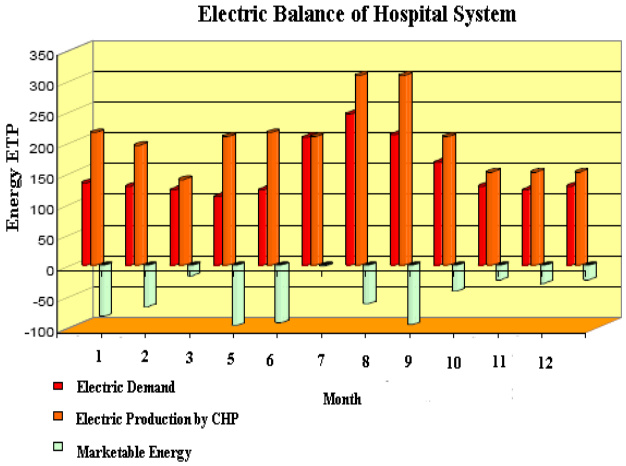


Figure 6. Monthly Trend of Electric Energy

IV. CONCLUSION

The CCHP technique can be utilized in large public buildings with considerable results in terms both of energy saving and financial advantage.

The architecture of the proposed CCHP system allows guaranteeing 100 percent redundant power source for

inpatient areas and will provide the continuity of energy supply during every critical state. The following percentage of supplies have been calculated in function of "peak demands": in the worse case (blackout in natural gas feeding), the system will be able to guarantee the 82% of the power thermal demand, the 74% of the power cooling demand and the 100% of power electric demand. The study predicts a large potential for financial savings, approximately €1,00 M€ year.

By enabling hospitals to supply their own power, CCHP systems also provide a hedge against the rising cost of electricity.

The technique gives also considerable results in terms of respect of environment: CCHP systems have higher whole system efficiency than systems that split heating and power generation. With less fuel consumed, greenhouse gases and air pollutants (nitrogen oxides and sulphur dioxides) are reduced more than 25,000 tons per year of carbon dioxide equivalent.

REFERENCES

[1] A. Moran, P.J. Pago, and L.M. Chamra, "Thermo-economic modelling of micro-CHP (micro-cooling, heating and power) for small commercial applications", International Journal of Energy Research, vol. 32, pp. 808–823, July 2008.

[2] A.S. Szklo, J.B. Soares, and M.T. Tolmasquim "Energy consumption indicators and CHP technical potential in the Brazilian hospital sector" Energy Conversion Management, vol. 45, pp. 2075-2091, 2004

[3] M. Vio, Impianti di Cogenerazione, ed. Delfino, 2009.

[4] Practice Greenhealth's energy impact calculator website at http://practicegreenhealth.org/tools-resources/energy-impact-calculator

[5] Electronic Publication: Report on National Center for Energy Management and Building Technologies Task 15, 2005

[6] S. Okamoto "Saving energy in a hospital utilizing CCHP technology" International Journal of Energy and Environmental Engineering, vol. 2, pp. 45-55, 2011

[7] Electronic Publication: US Department of Energy, "Hospitals Discover advantages to using CHP systems", Energy Efficiency & Renewable Energy, July 2011

[8] MC4 website, at www.mc4software.com [Last access March 16, 2012]

[9] Midwest Clean Energy Application Center website, at http://www.midwestcleanenergy.org [Last access March 16, 2012]

[10] Parliament of Italy Legislation Recourses website, at http://www.parlamento.it/parlam/leggi/deleghe/99079dl.htm [Last access March 16, 2012]

[11] Rete Ambiente Legislation Resources website, at http://www.amministrativo.it/ambiente/osservatorio.php?num=88 [Last access March 16, 2012]

[12] Parliament of Italy Legislation Recourses website, at www.camera.it/parlam/leggi/deleghe/testi/07026dl.htm [Last access March 16, 2012]

 ISBN: 978-1-61208-189-2

Short-Term Energy Pattern Detection of Manufacturing Machines with In-Memory Databases - A Case Study

Christian Schwarz
Hasso Plattner Institute, University of Potsdam
Potsdam, Germany
Email: firstname.lastname@hpi.uni-potsdam.de

Felix Leupold, Tobias Schubotz
SAP Labs Inc.
Palo Alto, United States
Email: firstname.lastname@sap.com

***Abstract*—Todays energy companies mainly use generalized demand sets to predict the required amount of energy of their customers on a high aggregation level. This is sufficient in an energy consumption oriented power grid, having enough resources to produce and transmit the requested and produced amount of energy. With the increasing amount of renewable energy sources, the power grid evolves from a purely consumption controlled supply network to a production controlled grid. In that environment the need for detailed short term energy demand predictions increases. A first step to predict the demand of energy is to find generalizable patterns within the energy consumption data that can later on be used for early predictions on real time data. To study the possibility to predict energy patterns nearly real-time, we created an environment, where metering data is collected every second and used for real-time pattern matching. We developed and implemented pattern recognition algorithms that use the abilities of in-memory databases with the collected metering data in order to detect energy consumption patterns.**

***Keywords*-energy pattern recognition; machine learning; in-memory database;**

I. Introduction

In industries energy expenses can make up to 43% of all operational expenses. Many companies monitor their energy usage on a detailed level and reduced their energy consumption, for example by 58% in the aluminum industry since 1975 [1].

A study by the US department for energy has shown that 71% of private households changed their energy usage behavior after they got the possibility to monitor their energy consumption based on in-home displays, even if the expected savings are only between 5 to 15% [2], [3].

Today's energy companies use generalized demand sets to forecast the required amount of energy for a given period of time. The amount of data recorded within an Advanced Metering Infrastructure (AMI), that is expected to collect metering data every 15 minutes, would enable more detailed energy demand predictions [4]. In contrast, short-term demand prediction based on real-time smart metering data is not used to reduce the gap between demand and supply of energy. That gap comes with high costs and can lead to bankruptcies of energy companies in extreme cases [5]. One of the reasons is the amount of data produced within an AMI, where each smart meter produces more than 35,000 meter readings per year.

We evaluated how a new trend in the database market helps to handle the amount of data produced by an AMI: in-memory databases. In-memory databases offer the possibility to run analytical workloads on transactional data within seconds, the used column-oriented data layout is not optimal by default for write-intensive workloads like the smart grid [6]. To enable a high transactional throughput, like it occurs within an AMI, in-memory databases use techniques like write-optimized buffers and bulk loading [7].

In this paper, we will focus on short-term pattern recognition of manufacturing machines. Short-term pattern recognition in that case means, that we are able to detect usage patterns within real-time enabling the energy provider to adjust their energy production for the remaining period of the pattern. A completed consumption pattern describes an energy usage pattern that occurs when a machine creates a product. Therefore, we have installed a metering infrastructure that enables us to collect energy usage data from several devices within a one second interval. In our scenario we use the energy consumption data created by a fully automated coffee machine, as well as several other devices. We detect the consumption pattern of that machine and forecast its future short-term energy demand based on historically recorded data within the in-memory database. This coffee machine is able to produce different types of coffees creating a more or less unique energy footprint during its production period. The goal is to detect the produced coffee as soon as possible and to predict the amount of energy required within the remaining production time.

The rest of the paper is organized as follows. Section V presents related work in the area of energy data pattern recognition The used pattern recognition and prediction methods are presented in Section II, while the experimental setup is described in detail in Section III. The evaluation of our pattern recognition algorithms is presented in Section IV followed by the conclusion and outlook in Section VI.

II. Pattern Recognition

In general, the field of pattern recognition is associated with the automatic discovery of similarities in data, which is

 ISBN: 978-1-61208-189-2

achieved through machine learning. Pattern matching can be used for regression analysis and classification [8]. Regression analysis models a function from a set of data points in order to interpolate and extrapolate between this data set. Classification decides to which of n classes a given data set belongs.

We use *supervised learning methods*, a technique that takes a training set of patterns for learning and undertakes to generalize from this training set to also identify untrained patterns [9]. Given an *input vector* X the algorithm calculates an *output vector* Y. The actual class of X is called the *target vector* T. In supervised learning the algorithm goes through a *learning* phase, where it is given a set of training vectors $X = \{X_1, X_2, ..., X_i\}$ with the corresponding target vectors $T = \{T_1, T_2, ..., T_i\}$. The algorithm then tries to find a function $y_k = Y(X_k)$ so that the deviation of the output vector for each x is minimal. During our project, we implemented multiple pattern matching algorithms from which we chose three to present within this paper: *inter-quartile range coverage (IQRC)*, a *multi class support vector machine (MCSVM)* and a *k-nearest neighbor (knn)* algorithm.

A. Inter-Quartile Range Coverage

We developed the IQRC pattern matching algorithm for our scenario to classify recorded patterns. Given a set of training vectors we calculate the upper and lower quartile for each dimension of all X_i that have an identical target vector T_k, which means they lie in the same class. The range between the upper and lower quartile is called *inter-quartile range (IQR)*. Given an input vector X, we sum up the dimensions of X that lie in the IQR of the same dimension in the training patterns. This is done for each classifier and compared against a threshold. If the threshold of classifier i is exceeded the algorithm identifies the input series as product i.

For classes with a high deviation amongst vectors the IQR will be larger than for classes with a small deviation. To take that into account, the value of a dimension lying in the IQR of a class is $\frac{1}{1+\|IQR\|}$, rating the values that lie in smaller IQR higher than values that lie in a greater IQR.

This step is done for each of the products. Our algorithm will output the product that exceeds its threshold the most. The algorithm can be formalized as follows: Let $X = (x_1, x_2, ..., x_n)$ be the input vector and j_i^k be a vector with all values from the training patterns of class k in dimension i. $\delta(k)$ denotes the threshold of class k.

$$y(x) = argmax_{k \in Classes} \left((\sum_{i=0}^{n} w(x_i, j_i^k)) - \delta(k) \right) \quad (1)$$

$$\text{where } w(x_i, j_i^k) = \begin{cases} \frac{1}{Q_{.75}(j_i^k) - Q_{.25}(j_i^k) + 1}, & \text{if } x_i \in IQR(j_i^k), \\ 0, & \text{else.} \end{cases} \quad (2)$$

$$\text{and } IQR(j_i^k) = \{w | w \in j_i^k \ \wedge \ Q_{.25}(j_i^k) < w < Q_{.75}(j_i^k)\} \quad (3)$$

If more than one beverage has an IQRC above a certain threshold, we chose maximum overstepping. Due to the relatively high warp amongst patterns, it is difficult to find a threshold that is exceeded by all positive but by none of the negative examples.

To optimize the thresholds we use a modified hill climbing algorithm [10]. The threshold for all products are initially chosen so high that none of the training patterns are recognized at all. We then order the beverages descending by the number of their occurrences in the training set and start to decrement their threshold until the overall matching performance reaches a maximum. This threshold is then fixed for the product and we continue with the next product. After processing all products, we start again with the first one. Contrary to the the first pass, the threshold of all other products are now at a fairly good value. We do this until all thresholds stay constant for one round that is, the optimum does not change anymore. In our scenario that happens after four iterations.

B. Multi-Class Supported Vector Machine

We also implemented a Support Vector Machine (SVM) algorithm inside the database management system [11]. Normally a SVM only separates between two classes, but there are approaches that extend the original SVM to solve multi-class classification problems as well. The most common approach is called *one-versus-all* [12]. Given there are n classes $c_1, c_2, ..., c_i, ..., c_n$ we will create a binary SVM that is trained with all patterns from c_1 for its first target and with the rest of the patterns for the other target. This is repeated with all of the n classes. In matching an incoming pattern is passed to each SVM. Ideally only one machine detects a positive result. If there is more than one SVM classifying the input as c_i, the one with the largest result vector is used. If there is no SVM classifying the input as c_i the one with the smallest negative result vector is chosen. Assuming n classes this approach needs n iterations.

C. K-Nearest Neigbor

We have seen that patterns we collected have a considerable variance for the same target vector. The clusters of each classifier are therefore rather big and overlapping. Nonetheless, since irregularities seem to be the rule, we might find the corresponding class by looking for the pattern that is closest to the input pattern. This way even in a subspace with many patterns of class A, a pattern of class B varying from the others might still be found. This is what the knn algorithm does. In our case k is set to 1. Given an input vector X, knn returns the classification y from the trained vector $\bar{X}$ for which the squared distances $d(X, \bar{X})$ is minimal [13].

An advantage of the nearest neighbor algorithm is that it requires less computational power than the other algorithms, because the calculation of squared distances is not expensive. The training phase only creates value series and does not involve learning.

 ISBN: 978-1-61208-189-2

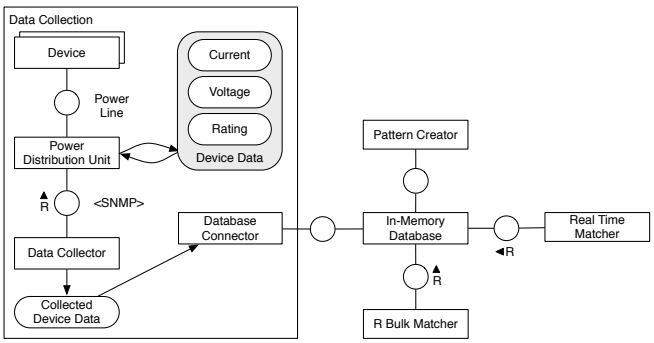


Figure 1. Experimental setup as FMC block diagram

III. Experimental Setup

To be able to perform short-time energy demand prediction detailed data like voltage, current and power consumption is necessary. These information need to be collected and stored in a database where further processing takes place. This section gives a detailed overview about the experimental setup, the data collection and energy demand prediction.

A. Data Collection

In smart grids, the (AMI) is expected to be applied for collecting necessary data. It is a multi-level architecture, where smart meter readings are transmitted from smart meters over concentrators to a central system, meaning a database. In order to detect the short energy usage patterns of our coffee machine, we increased the measuring granularity. In our experiment, we assume that every installed device can be measured independently from each other. This might not yet be the case in private households, but companies typically install multiple smart meters in order to monitor different consumers like air conditioning, lightning, office and manufacturing devices independently.

Because it is nearly impossible to get self-measuring devices, we used an Emerson Network Power Rack Power Distribution Unit (PDU) with a Liebert MPX control module which is capable of measuring the voltage, current and rating of every single power plug [14]. Depicted in Figure 1, to the PDU, we connected several devices, the PDU in turn is connected to a local area network and can be queried via the Simple Network Management Protocol (SNMP). There are in total 18 devices connected to the PDU. This setup results in an approximately 16,000 times higher data volume than the AMI proposes.

The data collector queries the PDU as often as possible to collect the current data for each device. This data is then transferred into the in-memory database. Table I depicts the used table schema for `device_readings`. Detected patterns in the stream of readings are written into the table `pattern_recognition`. The resulting transmission interval from PDU to database is between 0.5 and 2 seconds depending on the current network load.

Table I
Schema of the tables used for prediction

DEVICE_READINGS		
DEVICE_ID	:	INTEGER
DATETIME	:	INTEGER
VALUE	:	DOUBLE

PATTERN_RECOGNITION		
DATETIME	:	INTEGER
VALUE	:	DOUBLE
PRODUCT	:	VARCHAR

B. Data Processing

1) Hardware Setup: We use an instance of SAP's in-memory database with a column-oriented data layout that is best suited for analytical workloads [6]. Our implementation uses the GNU R interface of the databases development version [15]. The database is installed on a HP ProLiant DL580 G7 series server that is equipped with four Intel Nehalem X7560 CPUs and 256 GB main memory at 1066 MHz. The server runs a 64-bit version of openSUSE 11.2 (kernel 2.6.31.14).

2) Data for Training and Testing: In order to match patterns amongst the monitored energy consumption we use a set of features of that data to perform our matching algorithms on. Beside the raw data for electronic power consumption in watt hours [Wh], we also use the following more abstract features of the gathered data:

- Number of peaks
- Greatest delta
- Total amount of energy used
- Pattern duration
- Gradients values
- Moving average, and
- Histogram

3) Precision Benchmarking: We divide the pattern set into two parts, a training set with input and target vectors and a set with input and target vectors used for testing the performance. We use a cross validation technique to achieve reliable results. The pattern set of each product is split into five parts. In each iteration of the algorithm, we use four chunks for training and one for testing purposes. The overall performance is calculated by taking the average of all iterations. This technique is called *leave-some-out cross validation* [16].

4) Performance Benchmarking: For the daemon, we calculated the average time for one cycle in the algorithm over multiple hours of operation. When we repeated the measurement for the different algorithms, we measured the same time slots on different days. We define one cycle as querying the database for new data plus the time used for matching given there is a pattern detected.

 ISBN: 978-1-61208-189-2

Algorithm	Shortest duration	Longest duration	Average Duration
knn	3ms	806ms	9ms
SVM	3ms	1116ms	10ms
IQRC	3ms	1547ms	13ms

Table II
ITERATION TIMES AND THEIR DEVIATION DURING REAL-TIME DETECTION

IV. EVALUATION

A. Real-Time Performance

For interactive applications the critical limit for interaction response is two seconds [17]. Therefore, queries that are triggered by the real-time matcher have to be answered within this time interval. Real-time matching is implemented as a daemon. Although there is no critical time limit for this daemon, having as many matching cycles as possible allows discovering the patterns really close to their occurrences.

When the machine is idle, one cycle takes about three milliseconds. If the coffee machine is currently producing, it still takes less than a second. Table II shows the cycle times for the different algorithms. We can see that nearest neighbor matching performs best. Matching with Support Vector Machine takes about 30% percent longer. IQRC matching needs about twice the time compared to nearest neighbor. The reason is that we have to optimize the thresholds for all products individually in each cycle as they depend on the pattern length. Still, all algorithms have satisfying performance, as they are below the critical limit of two seconds.

B. Bulk Pattern Recognition

In addition, we analyzed the performance for bulk pattern recognition. Figure 2 shows the execution times of the MCSVM algorithm, implemented as a GNU R function, depending on the number of readings in the `device_readings` table for different numbers of used cores using nearest neighbor matching. The values represent the averages of ten measurements with a standard deviation of 11%. The execution times for support vector lie within the standard deviation of the execution times for knn. Therefore, we conclude that both algorithms perform equally well.

The comparison shows that the bulk matching scales linearly for more than 1000 values. This is not surprising, since we are iterating over device readings, which gives the algorithm a complexity of $\mathcal{O}(n)$.

The *rminer* package, which is used in the GNU R implementation does not parallelize its computation, therefore we parallelized the execution by distributing the available input values among the number of processors. Each of the n cores could then independently work on $\frac{1}{n}$th of the total values.

We also laid focus on parallel execution, utilizing all cores of our target system. As we can see, the usage of multiple CPUs decreases the execution time. However, this speedup is not linear as we might expect, because of the massive parallelization. According to Ahmdal's law the speedup is determined by the serial fraction of the algorithm [18]. In our case, this fraction is determined by the initialization

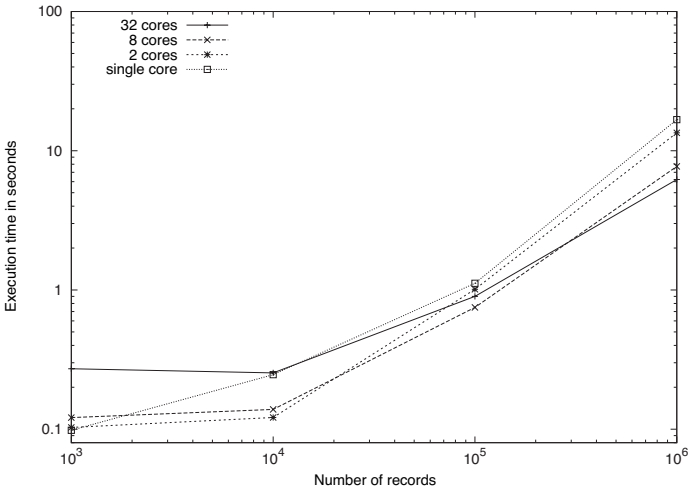


Figure 2. Comparison of execution times for bulk pattern recognition depending on the number of reading for different number of CPU cores

of the matcher and the merging of different parts of the `device_readings table`. We found out that the merging of one part of the list for a total of one million datasets took 10 to 20 ms. Although we tried to parallelize the merge, the jump from eight to 32 processes, increases the execution time for less than 200,000 value. The overhead in the merge is not outrun by the smaller number of device readings, each process has to analyze. However, 32 cores outperform eight cores for 200,000 readings. With an increasing number of readings, the gain from executing the computing expensive operations in parallel increases.

C. Precision Results

Fig 3 shows the hit rates for all features using the presented inter-quartile range coverage, k-nearest neighbor and Support Vector Machine algorithms. For the IQRC we were not able to perform the benchmark with the histogram feature, because all patterns of one product result in one histogram. They do not have any deviations or quartiles. The lower boundary for the matching performance with eight different products is 12.5% which would be the accuracy of chance. The PDU, which was used for measuring the consumption data, has an accuracy of $\pm 10\%$ for its measurements [14].

As we can see in Figure 3, the multidimensional features consumption, gradient and moving average perform equally. They are also the best performing features in total. The more dimensions are available, the more information can be used for classification, which lead to higher hit ratios.

One dimensional features perform significantly worse than multidimensional features.

The histogram feature could only be implemented for knn and SVM. Though it performs slightly better than the presented one dimensional feature, it is still noticeably worse than the multidimensional features. Reason is that the measured values hardly differ in size the histogram has a lot of different values with low frequencies.

 ISBN: 978-1-61208-189-2

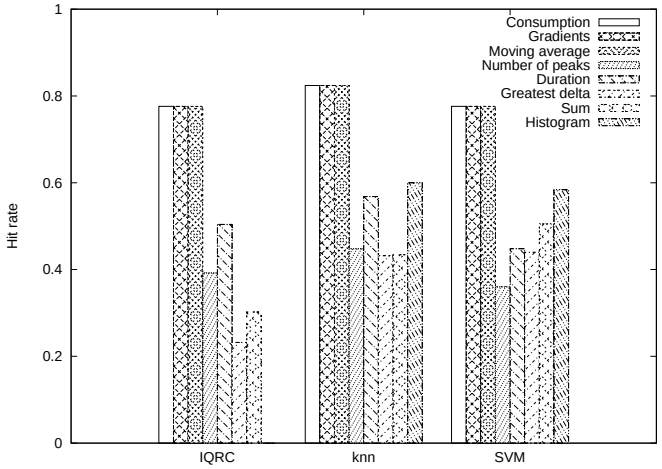


Figure 3. The comparison of three algorithms with different data features

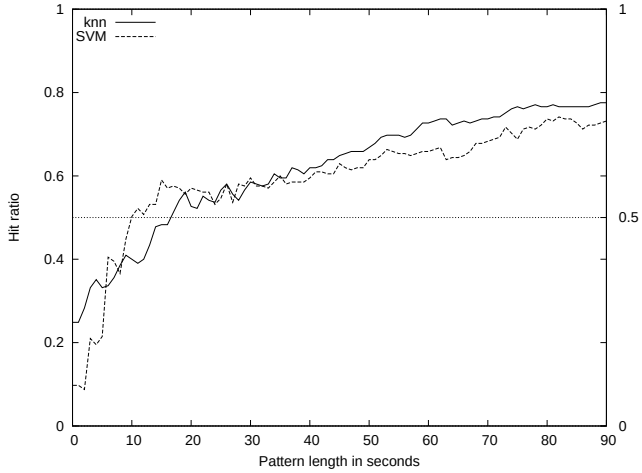


Figure 4. Hit rate depending on the length of the input vector

If we compare the algorithms with each other, we can see that the knn performs slightly better than the other two. It performs about 5 - 10 % better than the other in all features except for the greatest delta and sum feature. For multidimensional features the IQRC algorithm has the same hit rate as SVM. Nonetheless, IQRC outperforms SVM slightly considering the number of peaks and duration features. SVM on the other hand has the strongest results for the greatest delta and sum criteria. The implementation of multi-class SVM using *one-versus-all* is susceptible to misclassification if all machines calculate a negative result [19]. Due to the high deviation amongst patterns in our scenario, this case occurs more often. Therefore the overall accuracy of SVM is not as high as expected.

D. Short-Term Energy Consumption Prediction

If a pattern is detected, its subsequent values can be used for predicting the future consumption of a device. The earlier we recognize the product that is currently produced, the result gets more useful because the predicted remaining interval gets longer. Early matching has to be performed on incomplete consumption data and is therefore not as accurate as matching after the complete consumption. Fig 4 shows the accuracy of the knn and SVM algorithm depending on the length of the patterns. If we pass a pattern with length n we cut all training patterns down to that length and apply the algorithms.

We consider a hit rate of 0.5 to be sufficient in order to speak of successful pattern recognition. There are eight possible beverages, a hit rate bigger than 50% would be four times better than chance. As we can see, we break the 0.5 accuracy line at approximately 20 seconds. This means approximately one third of the pattern are sufficient for pattern recognition. If we transfer that finding to industrial manufacturing processes that take multiple hours, this forecasting range is valuable for utility companies, as it is sufficiently long for trading e.g. at the EEX spot market [20].

After twenty seconds of production we are able to identify the product. This means we can predict the succeeding ten to seventy seconds using the information from our trained patterns. The easiest way to predict values for the short term is to take the nearest neighbor and predict its course for the next values. We decided to use the knn approach for short term prediction also because its accuracy outruns the SVM's in the long run. An SVM approach could be implemented using the values of the closest support vector of the corresponding class as a prediction.

When we predict the subsequent consumption of a pattern after 20 seconds, we have an average deviation of 25%. After 40 seconds the deviation falls below 20%. This accuracy might not be convincing at first sight. Considering that the consumption values of the coffee machine even under load lie between 0.1 and 0.8 watt seconds, a predicted value that only differs by .01 watt seconds may lead to a deviation of 10%. Therefore we would have to predict three decimal places correctly to fall below that number. If we subtract measuring errors of the PDU [14], we end up with a deviation around 10%. For high performance industrial machines, the consumption is higher than for the coffee machine. Therefore the precision in industrial use cases is expected to be higher.

V. Related Work

Smart grids are considered as the continuation of the classical power grid in the information age [21]. In order to avoid different conflicting standards amongst its participant countries, the European Union has instantiated the Smart Meter Coordination Group (SMCG) [22]. OPENmeter is a project that is financed by the European Union and has proposed the AMI to be used for the smart grid [23]. On top of that manufacturers found a platform called Open Metering Systems that collaborates with SMCG and also assumes the AMI [22], [24]. Regardless of its changes, our experimental setup is comparable to this AMI.

Collected smart meter data can be used to to optimize consumer contracts [25]. More detailed monitoring of power data implies, that the energy consumption of current machines used in a data center highly depend on the computed task [26].

 ISBN: 978-1-61208-189-2

Predicting the energy consumption for medium and shorter terms has been done using artificial neural-networks [27], [28]. Related work has been focussing on comparing different algorithms for efficient pattern matching over event streams [29].

The Support Vector Machine is one of the most popular algorithms used in machine learning. It was introduced by Vapnik in 1992 and has proved to provide significantly better classification performance than other algorithms in most use cases [11], [8]. SVM can solve binary classification problems by finding an n-dimensional function that is able to distinct all data points of one class from another. A discussion on the various multi-class implementations of SVMs is done by Duan and Keerthi [12].

VI. Conclusion and Future Work

In our experiment we have shown, that it is possible to determine the type of coffee, that is produced by a coffee machine only based on a series of smart meter readings. We used those readings from a small dataset to train our different pattern recognition algorithms. The used in-memory database has shown that it is indeed able to process the collected amount of data, which was approximately 471 million tuples within the `device_readings` table and match it with the pre-calculated patterns in real time, meaning running a complete detection cycle below two seconds.

We have shown, that bulk pattern recognition can be scaled to multiple CPUs to enable an energy providing company with the possibility to cluster and aggregate the consumption data of multiple customers at a time. In the future unsupervised learning algorithms need to be applied to automatically create patterns from the energy consumption data. Thus, clustering customers in multiple groups helps companies to optimize their energy rate offers.

If a manufactured good can be determined by its energy footprint, those pattern recognition techniques might also be used to detect an early machine break down in the area of predictive maintenance. This can reduce the costs caused by unscheduled machine downtime within a productive environment.

Future work will include revalidating the results with different types of manufacturing machines, consuming a higher amount of energy and producing different energy usage footprints.

References

[1] T. N. Project, "Energy Consumption," *Intermediate Energy Infobook*, pp. 1–5, 2011.

[2] L. A. Butler, "In-Home Display Pilot," *US Department of Energy - Energy Efficiency and Renewable Energy*, pp. 1–9, Jul. 2011.

[3] S. Darby, "The Effectiveness of Feedback on Energy Consumption," *Environmental Change Institute, University of Oxford*, pp. 1–24, Apr. 2006.

[4] The OPENmeter Consortium, "Report on the identification and specification of functional, technical, economical and general requirements of advanced multi-metering infrastructure, including security requirements," *Deliverables*, June 2009.

[5] R. Weron, *Modeling and forecasting electricity loads and prices*, ser. a statistical approach. Wiley, Dec. 2006.

[6] H. Plattner, "A common database approach for OLTP and OLAP using an in-memory column database," *Proceedings of the 35th SIGMOD International Conference on Management of Data*, Jun. 2009.

[7] J. Krueger, M. Grund, C. Tinnefeld, and H. Plattner, "Optimizing write performance for read optimized databases," *Database Systems for Advanced Applications*, 2010.

[8] S. Marsland, *Machine Learning An Algorithmic Perspective*. Chapmann&Hall/CRC, 2009.

[9] C. M. Bishop, *Pattern Recognition And Machine Learning*. Springer, 2007.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition)*, ser. Prentice Hall series in artificial intelligence. Prentice Hall, 2002.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, pp. 273–297, 1995.

[12] K. Duan and S. S. Keerthi, "Which is the best multiclass svm method ? an empirical study," in *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, 2005, pp. 278–285.

[13] *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing)*. The MIT Press, 2006.

[14] E. E. Co., "User manual - network interface card for the liebert rack pdu family of power distribution products," Monitoring For Business-Critical Continuity, Tech. Rep., 2009.

[15] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[17] Galitz and W. O., *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*. Wiley & Sons, 2007.

[18] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," *Proc. AFIPS 1967 Spring Joint Computer Conf. 30 (April), Atlantic City, N.J.*, pp. 483–485, 1967.

[19] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines." *IEEE Transactions on Neural Networks*, vol. 13, pp. 415–425, 2002.

[20] "Transparency in energy markets," European Energy Exchange AG, Tech. Rep., 2011.

[21] M.-P. Schapranow, R. Kühne, A. Zeier, and H. Plattner, "Enabling Real-Time Charging for Smart Grid Infrastructures using In-Memory Databases." IEEE, pp. 1040–1045.

[22] Arbeitsgemeinschaft Für Sparsame und Umweltfreundlichen Energieverbrauch E.V., "Smart Meter - Intelligente Zähler."

[23] "Requirements of AMI," OPENmeter, Tech. Rep., 2009.

[24] H. Baden and P. .Gabriel, "Open metering system specification," Open Metering System Group, Tech. Rep., 2011.

[25] R.-I. C. Chi-Cheng Chuang, Jimi Y. C. Wen, "Consumer Energy Management System: Contract Optimization using Forecasted Demand."

[26] M. M. G. P. Antonio Vetro', Luca Ardito, "Monitoring IT Power Consumption in a Research Center: Seven Facts."

[27] S. A. and Kalogirou, "Applications of artificial neural-networks for energy systems," *Applied Energy*, vol. 67, no. 1-2, pp. 17 – 35, 2000.

[28] P. A. Gonzalez and J. M. Zamarreto, "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network," *Energy and Buildings*, vol. 37, pp. 595 – 601, 2005.

[29] J. Agrawal, Y. Diao, D. Gyllstrom, and N. Immerman, "Efficient pattern matching over event streams," in *Proceedings of the 2008 SIGMOD International Conference on Management of Data*, 2008, pp. 147–160.

 ISBN: 978-1-61208-189-2

A Parallel Rolling Horizon Scheme for Large Scale Security Constrained Unit Commitment Problems with Wind Power Generation

Eugene Feinberg
Department of Applied Math and Statistics
Stony Brook University
Stony Brook, New York 11794-3600
email: Eugene.Feinberg@sunysb.edu

Jiaqiao Hu
Department of Applied Math and Statistics
Stony Brook University
Stony Brook, New York 11794-3600
email: jqhu@ams.sunysb.edu

Eting Yuan
Department of Applied Math and Statistics
Stony Brook University
Stony Brook, New York 11794-3600
email: eyuan@ams.sunysb.edu

***Abstract*—The Unit Commitment Problem (UCP) is an important category of power planning problems. The purpose of UCP is to determine when to start up and shut down the generator units and how to dispatch the committed units to meet the electricity demands, ancillary services requirements and security constraints. In this paper, we improve the traditional Lagrange Relaxation (LR) approach and analyze the effectiveness of using parallel computing in solving large unit commitment problems with wind penetration and investigate the potential of combining parallel computing with a rolling horizon scheme to improve the solution quality when a large amount of wind power is present. In particular, we first formulate a security constrained unit commitment problem by taking into account power generation costs, ancillary costs, wind power and a variety of security constraints employed in real New York State day-ahead power market. We then propose a parallelized version of the LR method to solve the problem in a single step, analyze the scalability issue of parallel computing, and investigate the impact of increased wind energy penetration. Finally, when a large amount of wind power is present, we further propose an approach that combines parallel computing with a rolling horizon technique to solve the UCP online.**

***Index Terms*—unit commitment; ancillary service; wind power; parallel computing; rolling horizon.**

I. INTRODUCTION

The goal of unit commitment problems is to find the optimal production schedule for the power generation units and the production level of each unit over a short term period in order to minimize the operational cost of the power grid [1]. To maintain the security of the electric grid, a variety of security constrained, for example, reserve constraints and transmission constraints, are enforced, and the resulting problem is usually called Security Constrained Unit Commitment (SCUC) problem. In New York State, UCP is solved by New York Independent System Operator (NYISO) in the day-ahead power market based on the generation and ancillary service bids, which give generation and ancillary service cost of each power generator, from Independent Power Producers (IPP), Loads bid from Load Service Entities (LSE) and Security Constraints set by NYISO and other power regulation authorities. Because of the importance of UCPs, broad and intensive study has been carried out in this field, and many methods have been proposed in literature and used in practice [2].

Depending on the system configuration of a power grid, different optimization objectives and security constraints are considered. For the basic UCP formulation, the objective is simply to minimize the power generation cost subject to the electricity demand. However, as the liberalization of the electricity markets and advancement of optimization techniques, more and more elements are introduced. In [3], a security constrained unit commitment problem (SCUC) with transmission constraints was tackled using a lagrangian relaxation approach, where the transmission and reserve constraints were relaxed to form a dual problem and subsequently solved using subgradient methods. The test result showed that the proposed direct method can reduce the generation cost over the indirect method that does not consider transmission constraint in the dual optimization process. The algorithm was improved in [4] to address the feasibility issue, and a unit de-commitment step was added to achieve a better solution. The AC constraints were considered in [5], and Bender's decomposition technique was used to solve the problem. Furthermore, ancillary services has been gradually introduced into the unit commitment process. In [6], Z. Li and M. Shahidehpour used Lagrangian Relaxation technique to solve the security constrained UCP with the ancillary service constraints and costs; moreover, they also calculated the market clearing price of both generation and ancillary costs. Their work is important because there is a conflict between generation service and ancillary service when a generator is turned on. Additionally, some environmental elements, such as carbon tax, were introduced into the UCP in the past two years [7]. Because of its complexity, it is unusual

The research of Eugene Feinberg was partially supported by NSF grants CMMI-0900206 and CMMI-0928490. The work of Jiaqiao Hu was supported in part by the Air Force Office of Scientific Research under Grant FA95501010340, and by the National Science Foundation under Grants CMMI-0900332 and CMMI-1130761.

 ISBN: 978-1-61208-189-2

for ISOs to solve the problem in a single step. Instead, a multi-step approach is often adopted. For example, in the New York State, different constraints are added at different steps to decrease the complexity of the problem [8]; however, this will decrease the solution quality. Therefore, how to solve a SCUC problem in a single step with a certain time limit becomes a challenging problem.

The presence of renewable energy sources such as wind power can further increase the complexity of the unit commitment problem, and a common method to handle this is to use the scenario tree technique [9] to simulate the uncertainties and dynamics of wind power. However, to make the problem computationally tractable, only a very limited number of scenarios can be used. Many research projects proposed to use a rolling horizon optimization scheme rather than the traditional day-ahead scheme; see for example, the Wilmar project [10][11][12]. An alternative method is to use a probability measure to set up a probability level to limit the probability of power outage within the prescribed threshold [13]. To meet these probability requirements, one needs to set the operation reserve based on the variability of wind power. A rolling horizon approach can also be used to dynamically locate the operation reserve when new wind forecast information is available. Note that the rolling horizon approach is more computationally demanding as compared to traditional UCPs in day-ahead scheduling, as decisions need to be made at every time step in an online manner and every decision requires solving a nonlinear optimization problem involving both continuous and discrete decision variables.

In this work, an improved Lagrangian Relaxation (LR) method, which is adapted for parallel computing, is proposed to solve a large scale SCUC problem. Because linear generation cost functions are used, a greedy algorithm is proposed to optimize the generation and ancillary service when a generator is on. By using the proposed algorithm, we expect to solve the large-scale SCUC in a single step and dramatically reduce the computational time. A system based on the features of power system of New York Control Area (NYCA) is simulated to test the significance of our algorithm.

To address the variability of wind power, we follow the idea of [13] and use a probabilistic reserve constraint to describe the uncertainty of wind power. Since the wind power forecast is more accurate over shorter time periods [14][11], the probabilistic reserve constraints method is combined with a rolling horizon scheme to dynamically update the reserve constraints when more accurate wind forecast becomes available. the computational time issue is addressed by implementing our proposed solver on a parallel computing facility, and the research results show that parallel computing has the potential to satisfy this computational speed requirement of Rolling Horizon Scheme.

The organization of this paper is as follows. A security constrained unit commit model is formulated in Section II, and a solution algorithm is given in Section III. Section IV gives the probabilistic formulation of reserve constraints and the method used to handle the constraints. In Section V, we provide case study to illustrate the performance of our algorithms. The nomenclature is given in the Appendix at the end of this paper.

II. Formulation of SCUC Model

In this work, a SCUC model is formulated based on the realistic problem solved in New York State. Both generation service and ancillary services, including reserve services and regulations services, are optimized to minimize the total operational costs. Realistic security constraints, including balance constraints, ancillary service requirement, load capacity constraint, transmission constraints, etc, are considered in this research. Unlike some other research, which is based on benchmark problems with up to 100 generators and limited security constraints, this work is trying to solve large-scale problem with more than 600 power generators and realistic security constraints enforced in daily power planning process in New York States. The formulation is given in the subsections below.

A. Objective Function

In this work, the total operational cost, including both the power supply cost and ancillary services cost are optimized. The power supply cost includes power generation cost and start-up cost. The ancillary service cost includes reserve service cost and regulation service cost. Moreover, reserve service is divided into spinning service and non-synchronous service. The formulation is given in (1).

$$\begin{aligned} cost \quad = \quad & \sum_{m=1}^{M}\sum_{i_m=1}^{I_m}\sum_{t=1}^{T}(F_{m,i_m,t}(p_{m,i_m,t}) + S_{m,i_m,t}(z_{m,i_m,t})) \\ & + \sum_{m=1}^{M}\sum_{i_m=1}^{I_m}\sum_{t=1}^{T}(R_{m,i_m,t}(r_{m,i_m,t})) \\ & + \sum_{m=1}^{M}\sum_{i_m=1}^{I_m}\sum_{t=1}^{T}(CReg_{m,i_m,t}(reg_{m,i_m,t})), \end{aligned} \tag{1}$$

where

$r_{m,i_m,t} = (r10s_{m,i_m,t},\ r10ns_{m,i_m,t},\ r30s_{m,i_m,t},\ r30ns_{m,i_m,t})$.

The first line in equation (1) includes power generation function, which is formulated as a piecewise linear function, and a startup function, which is formulated as a stepwise linear function. The second line includes the reserve cost function, which is a linear function with respect to the 10-minute and 30-minute spinning reserves and non-synchronous reserves. The third line is the linear regulation cost function. It should be noted that a generator can provide spinning reserve or regulation services only when it is turned on, and provide non-synchronous reserve service only when it is turned off.

B. Load Balance

In this work, we assume that wind power can be integrated at no additional cost, and that there is no wind curtailment. Thus, wind power will always be delivered to customers. Then, wind power can be considered as negative load in this research.

 ISBN: 978-1-61208-189-2

Here we define the term "net load" as the difference between the electricity demand and the predicted wind power, i.e., the load that needs to be supplied by the traditional generators, including steam generator, gas turbine and hydro power. In day-ahead planning, the hydro and thermal plants should meet the sum of net loads of certain control areas. The mathematical formulation is given in equation (2).

$$\sum_{m=1}^{M}\sum_{i_{m=1}}^{I_m} p_{i,i_m,t} = \sum_{m=1}^{M}(d_{m,t} - w_{m,t}),\ \forall t \tag{2}$$

C. Ancillary service requirements

For the entire control area, there are three kinds of reserve requirements: 10-minute spinning reserve, 10-minute total reserve, and 30-minute total reserve requirements. The mathematical formulations are given in (3), (4), and (5), respectively.

$$\sum_{m=1}^{M}\sum_{i_m=1}^{I_M} r10s_{m,i_{m,}t} \geq Res_{10spin},\ \forall t \tag{3}$$

$$\sum_{m=1}^{M}\sum_{i_m=1}^{I_M} (r10s_{m,i_{m,}t} + r10ns_{m,i_m,t}) \geq Res_{10t},\ \forall t \tag{4}$$

$$\sum_{m=1}^{M}\sum_{i_m=1}^{I_M} (r30s_{m,i_{m,}t} + r30ns_{m,i_m,t}) \geq Res_{30t},\ \forall t \tag{5}$$

In real power market, a control area is often divided into several individual areas, which are usually called "zones". And for certain collection of zones, there are several location based reserve constraints. Letting Λ_j be the jth collection of zones, the location based reserve constraints are given by (6), (7), and (8).

$$\sum_{m\in\Lambda_j}\sum_{i_m=1}^{I_M} r10s_{m,i_{m,}t} \geq ResLB_{j,10spin},\ \forall t \tag{6}$$

$$\sum_{m\in\Lambda_j}^{M}\sum_{i_m=1}^{I_M} (r10s_{m,i_{m,}t} + r10ns_{m,i_m,t}) \geq ResLB_{j,10t},\ \forall t \tag{7}$$

$$\sum_{m\in\Lambda_j}^{M}\sum_{i_m=1}^{I_M} (r30s_{m,i_{m,}t} + r30ns_{m,i_m,t}) \geq ResLB_{j,30t},\ \forall t \tag{8}$$

In a control area operated by an ISO, those collection of zones are called "super-zones". Equations (6), (7), and (8) should be satisfied for all those super-zones.

Additionally, due the fluctuation of power demand and wind power, ISOs has certain requirements for regulation services, which are expressed in (9).

$$\sum_{m=1}^{M}\sum_{i_m=1}^{I_M} r10s_{m,i_{m,}t} \geq Reg_t,\ \forall t \tag{9}$$

D. Transmission constraints

Transmission constraints between different zones are also considered. The modeling of transmission constraints follows the method used in [4], and is given in (10).

$$\sum_{m-1}^{M}\Gamma_{l,m}(\sum_{i_m=1}^{I_m} p_{m,i_m,t} + w_{m,t} - d_{m,t}) \leq Tram_{l,max},\ \forall l,t \tag{10}$$

where $\Gamma_{l,m}$ is the line flow distribution factor for the transmission line l due to the net power injection of zone m.

E. Single generator capacity constraints

A generator that has been turned on might provide generation and reserve simultaneously. In practice, the sum of these services should be within the maximum power output limit. These requirements are given in (11), (12), (13), and (14). When a generator is off-line, it might be used to provide non-synchronous reserve services, the sum of which should also be within the maximum output limit. The formulations are given in (16), (17), and (18). In addition, regulation service should also be within a certain limit, which is given in equation (15).

$$p_{m,i_m,t}+r10s_{m,i_m,t}+r30s_{m,i_m,t} \leq p_{m,i_m,max}z_{m,i_m,t},\ \forall m,i_m,t \tag{11}$$

$$p_{m,i_m,min}z_{m,i_m,t} \leq p_{m,i_m,t} \leq p_{m,i_m,max}z_{m,i_mt},\ \forall m,i_m,t \tag{12}$$

$$r10s_{m,i_m,t} \leq r10s_{m,i_m,max}z_{m,i_m,t},\ \forall m,i_m,t \tag{13}$$

$$r30s_{m,i_m,t} \leq r30s_{m,i_m,max}z_{m,i_m,t},\ \forall m,i_m,t \tag{14}$$

$$regs_{m,i_m,t} \leq regs_{m,i_m,max}z_{m,i_m,t},\ \forall m,i_m,t \tag{15}$$

$$r10ns_{m,i_m,t}+r30ns_{m,i_m,t} \leq p_{m,i_m,max}(1-z_{m,i_m,t}),\ \forall m,i_m,t \tag{16}$$

$$r10s_{m,i_m,t} \leq r10s_{m,i_m,max}(1-z_{m,i_m,t}),\ \forall m,i_m,t \tag{17}$$

$$r30s_{m,i_m,t} \leq r30s_{m,i_m,max}(1-z_{m,i_m,t}),\ \forall m,i_m,t \tag{18}$$

F. Other constraints

Some other constraints such as minimum up time and down time constraints are also considered in our model. The detailed formulation can be found in, e.g., [15] and [16]. In addition, the maximum number of stops will be considered in our model, which means that a generator can only be turned off for a limited number of times on a single day. This constraint has never be considered in previous studies.

 ISBN: 978-1-61208-189-2

III. Solution Algorithm

As proposed in [4], a direct method is used to solve the SCUC problem. Compared with the work in [3], we introduced the ancillary service costs in the objective function, and single generator capacity constraints are added to the model. Moreover, a parallel computing scheme is developed to enhance the computational speed.

A. Lagrangian Relaxation Algorithm

To solve this problem, Lagrangian Relaxation method is used to relax the demand, reserve, transmission, and regulations constraints. A dual problem is thus obtained, and its objective function is given in equation (19).

$$
\begin{aligned}
Dual\,Cost = Cost + \sum_{t=1}^{T}\{ & \\
& \lambda_{d,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} p_{m,i_m,t} - \sum_{m=1}^{M}(d_{m,t} - w_{m,t})] \\
& + \lambda_{10s,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} r10s_{m,i_m,t} - Res_{10spin}] \\
& + \lambda_{10t,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m}(r10s_{m,i_m,t} + r10ns_{m,i_m,t}) \\
& \qquad - Res_{10t}] \\
& + \lambda_{30t,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m}(r30s_{m,i_m,t} + r30ns_{m,i_m,t}) \\
& \qquad - Res_{30t}] \\
& + \sum_{j=1}^{J}\{\lambda_{j,10s,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} r10s_{m,i_m,t}\mathcal{I}(m \in \Lambda_j) \\
& \qquad - Res_{j,10spin}] \\
& + \lambda_{j,10t,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m}(r10s_{m,i_m,t} + r10ns_{m,i_m,t})\mathcal{I}(m \in \Lambda_j) \\
& \qquad - Res_{j,10t}] \\
& + \lambda_{j,30t,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m}(r30s_{m,i_m,t} + r30ns_{m,i_m,t})\mathcal{I}(m \in \Lambda_j) \\
& \qquad - Res_{t,10t}]\} \\
& + \lambda_{reg,t}[\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} regs - Reg_{10spin}] \\
& + \sum_{t=1}^{T}\sum_{l=1}^{L}\{\lambda_{tran,l,t}[Tran_{l,max} \\
& - \sum_{m=1}^{M}\Gamma_{l,m}(\sum_{i_m=1}^{I_m} p_{m,i_m,t} + w_{m,t} - d_{m,t})]\}, \quad (19)
\end{aligned}
$$

where "$Cost$" equals the cost function in equation (1). The second line of the equation (19) is due to the relaxation of demand constraints; lines $3-7$ are due to the relaxation of reserve constraints of the whole control area, while lines $8-13$ are due to the relaxation of location reserve constraints, line 14 is due the relaxation or regulation constraints, and lines $15-16$ are the relaxation of transmission constraints. $\lambda_{d,t}$, $\lambda_{10s,t}$, $\lambda_{10t,t}$, $\lambda_{30t,t}$,$\lambda_{j,10s,t}$, $\lambda_{j,10t,t}$, $\lambda_{j,30t,t}$, $\lambda_{reg,t}$, and $\lambda_{tran,l,t}$ are the corresponding lagrangian multipliers. For simplicity, we define $\lambda_t = \{\lambda_{d,t}, \lambda_{10s,t}, \lambda_{10t,t}, \lambda_{30t,t}, \lambda_{j,10s,t}, \lambda_{j,10t,t}, \lambda_{j,30t,t}, \lambda_{reg,t}\}$. The dual problem is to find $\max_{\lambda_t}[\min(Dual\,Cost)]$. A single generator problem is defined in equation (20).

$$
\begin{aligned}
DualCost_{m,i_m} = \sum_{t=1}^{T}\{ & F_{i,i_m,t}(P_{m,i_m,t}) + S_{m,i_m,t}(z_{m,i_m,t}) \\
& + R_{m,i_m,t}(r_{m,i_m,t}) \\
& + CReg_{m,i_m,t}(reg_{m,i_m,t}) \\
& - p_{m,i_m,t}(\lambda_{d,t} + \sum_{l=1}^{L}\lambda_{tran,l,t}\Gamma_{l,m}) \\
& - r10s_{m,i_m,t}[\lambda_{10s,t} + \lambda_{10t,t} + \sum_{j=1}^{J}\mathcal{I}(m \in \Lambda_j)(\lambda_{j,10s,t} + \lambda_{j,10t,t}) \\
& - r10ns_{m,i_m,t}[\lambda_{10t,t} + \sum_{j=1}^{J}\mathcal{I}(m \in \Lambda_j)\lambda_{j,10s,t}] \\
& - r30s_{m,i_m,t}[\lambda_{30t,t} + \sum_{j=1}^{J}\mathcal{I}(m \in \Lambda_j)\lambda_{j,30t,t}] \\
& - r30ns_{m,i_m,t}[\lambda_{30t,t} + \sum_{j=1}^{J}\mathcal{I}(m \in \Lambda_j)\lambda_{j,30t,t}] \\
& - reg_{m,i_m,t}\lambda_{reg,t} \quad (20)
\end{aligned}
$$

Then the dual cost can be express as follows.

$$
Dual\,Cost = \sum_{m=1}^{M}\sum_{i_m=1}^{I_m} DualCost_{m,i_m} + Extra \quad (21)
$$

where $Extra$ is the difference between equations (19) and (21). It could be seen that the term $Extra$ does not depend on the status of power generators, and is a constant if the values of multipliers are given.

A subgradient method is used to solve the dual problem. The value of λ_t is initialized first, and its value can be used to minimize dual cost function. Because the term $Extra$ is a constant, we just need to minimize the term $DualCost_{m,i_m}$ individually. Dynamic problem is used to solve the single generator problem. To calculate the one-step reward, we need to optimize the allocation of generation services and ancillary services when a generator is on or off. When a generator is on, it could provide generation services, spinning reserve services, and regulation services. The one-step cost optimization problem (22) should be solved subject to constraints (11), (13), (14), and (15). On the other hand, when a generator is off, it could provide non-synchronous reserve services. Another one-step cost optimization problem (23) should be solved subject to constraints (17) and (18).

 ISBN: 978-1-61208-189-2

$$\begin{aligned}\min\{&F(p_{m,i_m,t}) + R_{m,i_m,t}(r_{m,i_m,t}) \\ &+ CReg_{m,i_m,t}(reg_{m,i_m,t}) \\ &- p_{m,i_m,t}(\lambda_{d,t} + \sum_{l=1}^{L}\lambda_{tran,l,t}\Gamma_{l,m}) \\ &- r10s_{m,i_m,t}[\lambda_{10s,t} + \lambda_{10t,t} \\ &+ \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)(\lambda_{j,10s,t}+\lambda_{j,10t,t})] \\ &- r30ns_{m,i_m,t}[\lambda_{30t,t} + \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,30t,t}] \\ &- reg_{m,i_m,t}\lambda_{reg,t} \end{aligned} \quad (22)$$

$$\begin{aligned}\min\{&R_{m,i_m,t}(r_{m,i_m,t}) \\ &- r10ns_{m,i_m,t}[\lambda_{10t,t} + \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,10s,t}] \\ &- r30ns_{m,i_m,t}[\lambda_{30t,t} + \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,30t,t}] \end{aligned} \quad (23)$$

General linear programming techniques can be used to solve these two problems; however, it is time consuming. Since piecewise generation cost functions and linear ancillary service cost functions are used in this work, we propose to use a greedy algorithm, which can significantly reduce the computational time. Let

$$F(p) = b_0 + \begin{cases} b_1 p & a_0 \le p < a_1 \\ (b_1-b_2)a_1 + b_2 p & a_1 \le p < a_2 \\ ... & \\ \sum_{k=2}^{K}(b_{k-1}-b_K)a_{k-1} + b_k p & a_{K-1} \le p \le a_K, \end{cases}$$

where we have $b_1 < b_2 < ... < b_k$ and $0 = a_0 < a_1 < ... < a_K$. We can transform it into another formula

$$F(p) = G(x_1, x_2, ..., x_K) = b_0 + \sum_{k=k}^{K} b_k x_k,$$

where $p = \sum_{k=1}^{K} x_k$, $0 \le x_1 \le a_1 - a_0$ and $0 \le x_k \le (a_k - a_{k-1})\cdot\mathcal{I}(x_{k-1} = a_{k-1} - a_{k-2})$, $k \ge 2$. The ancillary service costs are formulated as below:

$$\begin{aligned} R(r) \quad = \quad & rc10s\cdot r10s + rc10ns\cdot r10ns \\ & + rc30s\cdot r30s + rc30ns\cdot r30ns \end{aligned}$$

$$CReg(reg) = creg\cdot reg,$$

where $rc10s$, $rc10ns$, $rc30s$, $rc30ns$, and $creg$ are constant cost coefficients. Then equation (22) can be equivalently written as follows.

$$\min F'_{m,i_m,t}(p_{m,i_m,t}) + R'_{m,i_m,t}(r_{m,i_m,t}) + CReg'_{m,i_m,t}(reg_{m,i_m,t}) \quad (24)$$

where

$$\begin{aligned} F'_{m,i_m,t}(p_{m,i_m,t}) &= G'(x_1, x_2, ..., x_K) \\ &= b_0 + \sum_{k=k}^{K}(b_k - \lambda_{d,t} - \sum_{l=1}^{L}\lambda_{tran,l,t}\Gamma_{l,m})x_k \end{aligned}$$

$$\begin{aligned} R'_{m,i_m,t}(r_{m,i_m,t}) &= [rc10s_{m,i_m,t} - \lambda_{10s,t} - \lambda_{10t,t} \\ &- \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)(\lambda_{j,10s,t}+\lambda_{j,10t,t})]\cdot r10s_{m,i_m,t} \\ &+ [rc30s_{m,i_m,t} - \lambda_{30t,t} - \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,30t,t})]\cdot r30s_{m,i_m,t} \end{aligned}$$

$$CReg'_{m,i_m,t}(reg_{m,i_m,t}) = (creg_{m,i_m,t} - \lambda_{reg,t})\cdot reg_{m,i_m,t}$$

and equation (23) is equivalent to (25):

$$\min R'_{m,i_m,t}(r_{m,i_m,t}), \quad (25)$$

where

$$\begin{aligned} &R'_{m,i_m,t}(r_{m,i_m,t}) \\ &= [rc10ns_{m,i_m,t} - \lambda_{10t,t} - \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,10t,t}]\cdot r10ns_{m,i_m,t} \\ &+ [rc30ns_{m,i_m,t} - \lambda_{30t,t} - \sum_{j=1}^{J}\mathcal{I}(m\in\Lambda_j)\lambda_{j,30t,t})]\cdot r30ns_{m,i_m,t} \end{aligned}$$

When a generator is "on", the greedy algorithm works as follows:

1) Initialize the power generation and ancillary service levels to 0.
2) Sort the linear cost coefficients, including those in generation cost function F' and ancillary cost functions R' and $CReg'$, in equation (24), to form a non-decreasing list $\{c_h\}$, where $h \in \{1, 2, ..., K+3\}$.
3) Let $p_{m,i_m,t} = p_{m,i_m,min}$.
4) From $h = 1$ to $h = K+3$, consider the following cases:
   a) c_h is a generation cost coefficient; if $c_h > 0$, stop; else, c_h should be the generation cost coefficient of the kth segment and
   $$a_k + r10s_{m,i_m,t} + r30s_{m,i_m,t} \le p_{m,i_m,max},$$
   then $p_{m,i_m,t} = \max(a_k, p_{m,i_m,min})$.
   b) c_h is a 10-minute spinning reserve cost coefficient; if $c_h > 0$, stop; else,
   $$\begin{aligned} r10s_{m,i_m,t} = min(&r10s_{m,i_m,max}, \\ &p_{m,i_m,max} - p_{m,i_m,t} - r30s_{m,i_m,t}). \end{aligned}$$
   c) c_h is a 30-minute spinning reserve cost coefficient;

if $c_h > 0$, stop; else

$$r30s_{m,i_m,t} = min(r30s_{m,i_m,max}, p_{m,i_m,max} - p_{m,i_m,t} - r10s_{m,i_m,t}).$$

d) c_h is a regulation cost coefficient;
if $c_h > 0$, stop; else, $r30s_{m,i_m,t} = reg_{m,i_m,max}$.

When a generator is "off", a similar algorithm can be applied:

1) Initialize ancillary service levels to 0.
2) Sort the linear cost coefficients of ancillary cost function R' to form a non-decreasing list $\{c_1, c_2\}$.
3) For $h = 1\, or\, 2$, consider the following cases:
   a) c_h is a 10-minute non-synchronous reserve cost coefficient;
   if $c_h > 0$, stop;
   else,

$$r10s_{m,i_m,t} = min(r10ns_{m,i_m,max}, p_{m,i_m,max} - r30ns_{m,i_m,t}).$$

   b) c_h is a 30-minute non-synchronous reserve cost coefficient;
   if $c_h > 0$, stop;
   else,

$$r30s_{m,i_m,t} = min(r30ns_{m,i_m,max}, p_{m,i_m,max} - r10ns_{m,i_m,t}).$$

B. A Parallel computing scheme

In the LR algorithm, the dual problem is decomposed into identical single generator problems. Therefore, it is natural to assign those problems to individual CPU's and solve them simultaneously. In every iteration, the root CPU will "broadcast" the values of lagrangian multipliers to the branch CPU's, where the single generator sub-problems are solved, and the solutions are "collected" to the root CPU to update the value of the multipliers. The computing scheme is illustrated in figure III-B. Suppose there are N power generators, then the computational time required to solve the dual problem can be estimated by equation (26) if the computational load is equally distributed to each branch CPU.

$$t_{dual} \approx (\frac{N}{\#\, of\, CPUs} \times t_{single} + t_{update} + t_{comm}) \times n_{iteration} \quad (26)$$

where t_{dual} is the computational time needed to solve the dual problem, t_{single} is the computational time in solving a single generator problem, t_{comm} is the communication time between the root CPU and branch CPU's, and $n_{iteration}$ is the number of iterations in the subgradient search process.

IV. Probabilistic Unit Commitment Model and Rolling Horizon (RH) Scheme

For simplicity, we only consider the probabilistic reserve constraints. Without loss of generality, we can replace the reserve constraint in Section II-C with the following constraint (27).

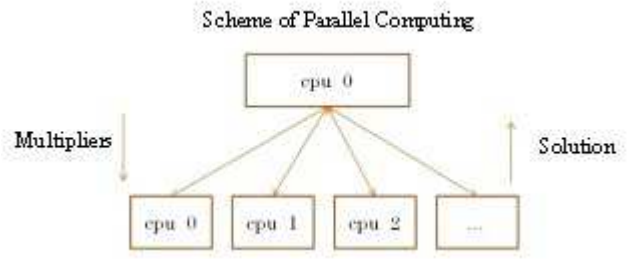


Figure 1. Parallel Computing Scheme to Solve UCP

$$\mathcal{P}\{\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} p_{m,i_m,t} + \sum_{m=1}^{M}\sum_{i_m=1}^{I_m} res_{m,i_m,t} \geq \sum_{m=1}^{M} d_{m,t} - \sum_{m=1}^{M} w_m, \forall t\} \geq 1-\alpha, \quad (27)$$

where $res_{m,i_m,t}$ is the general reserve service level. Using Bonferroni's inequality, we can transform equation (27) to another equation (28).

$$\mathcal{P}\{\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} p_{m,i_m,t} + \sum_{m=1}^{M}\sum_{i_m=1}^{I_m} res_{m,i_m,t} \geq \sum_{m=1}^{M} d_m - \sum_{m=1}^{M} w_{m,t},\} > 1-\frac{\alpha}{T} \quad (28)$$

Here, we assume that $w_{m,t}$ follows a normal distribution $N(\mu^w_{m,t}, (\sigma^w_{m,t})^2)$. Thus, the above equation can be written as (29).

$$\sum_{m=1}^{M}\sum_{i_m=1}^{I_m} p_{m,i_m,t} + \sum_{m=1}^{M}\sum_{i_m=1}^{I_m} res_{m,i_m,t} \geq \sum_{m=1}^{M} d_m - \sum_{m=1}^{M} \mu^w_{m,t} + z_{1-\frac{\alpha}{T}} \sum_{m=1}^{M} \sigma^w_{m,t} \quad (29)$$

Equation (29) may not accurately describe the reserve requirement, because the wind power in different zones are in general correlated. Nevertheless, for simplicity we assume that they are independent, and the general correlated case will be considered in our future research. This reserve constraints is very similar to the constraints in Section II-C, and a similar LR algorithm can be applied to solve the problem.

If we assume that wind power forecasts are updated every hour, we can update equations (2) and (29) and solve the corresponding UCP on a hourly basis. The RH scheme considers the updated wind power information in both unit commitment and economic dispatch processes, while in traditional day-ahead scheme, the updated wind power information is only considered in economic dispatch process. Thus, by involving more accurate information in the optimization process, we expect to get better solutions with decreased operational costs.

 ISBN: 978-1-61208-189-2

V. CASE STUDIES

A. New York Control Area

To study the effectiveness of our approach on large scale problems, we simulated a SCUC problem based on the characteristics of New York State control Area. NYCA is divided into 11 sub-zones with transmission interface between adjacent sub-zones. The detailed zone map is given in figure V-A. One feature of power grid in New York State is that most of the electricity demand comes from the southeast area of New York state, i.e., Long Island and New York City, while a large portion of the power resources are located in the west and north parts of the state. Additionally, in the near future, most of the power farms will be located in zones $A-E$ [17], and will bring more burden to transmission lines. This uneven distribution of power generation sources and power demand makes transmission constrained unit commitment an important problem in NYCA. Moreover, locational reserve requirements are enforced to maintain the safety operation of the power grid.

We follow the practice of NYISO and divide NYCA into two super-zones, where west super-zones include load zones $A-E$ and east super-zones include zones $F-G$. Additional reserve requirements are enforced for east super-zones. In addition, similar reserve requirements are also enforced on zone K, which is Long Island. The reserve requirements can be formulated in the forms of equations (6), (7), and (8). The transmission constraints are formulated in the form of equation (10).

In accordance with the day-ahead power market in New York State [18], piecewise linear generation cost functions and stepwise startup cost functions are used. Each generation cost function can have up to 12 pieces. A total of 641 power generators, including nuclear plants, hydro plants, steam plants, and gas turbines, are simulated in this work. The net load for each zone is calculated by subtracting the forecasted wind power from the forecasted electricity demands. Four different wind penetration level cases are used: $1275MW$, $4250MW$, $6000MW$, and $8000MW$. According to the penetration level, the regulation requirement is adjusted as proposed in [17]. A single day (24 hour period) in August is used for the study. Because currently, the report [17] by NYISO indicates that the current reserve level is enough for the $8000MW$ penetration of wind power, so we will not consider the probabilistic reserve level management in this case.

The LR algorithm was coded in C++ and implemented on New York Blue Gene, a distributed-memory supercomputing cluster. Up to 50 nodes were used, while each node has two $700MHz$ PowerPC processors and $1G$ DDR memory. Figure 3 shows the computational time required to execute the algorithm v.s. the number of CPUs used. The minimum computational time is around 180 seconds, which is much less than the 1600 seconds computational time when 1 CPU is used. The total operation costs, which is the sum of generation costs and ancillary costs, are given in table I. It is interesting to note that the costs are all negative. This is reasonable because

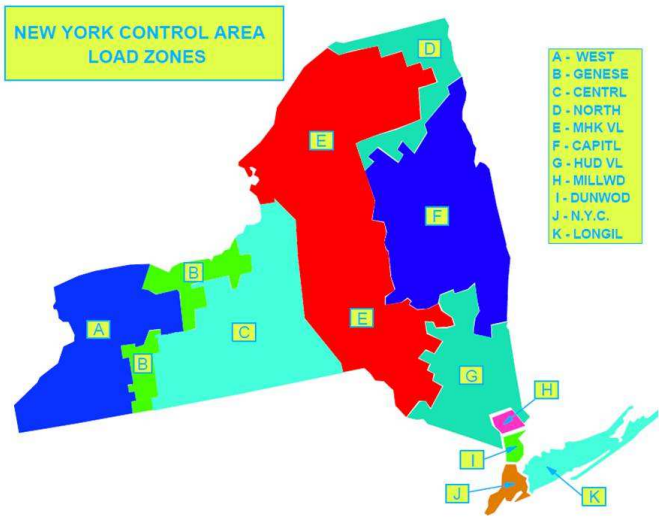


Figure 2. New York State Control Area

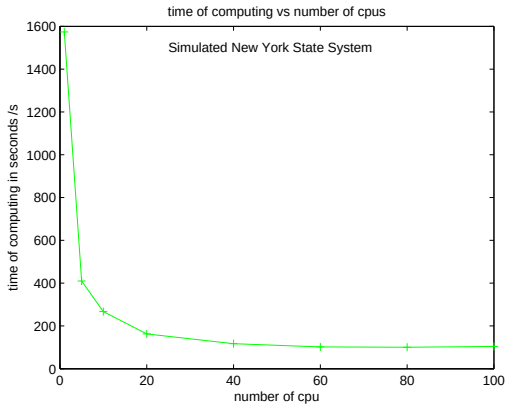


Figure 3. Computational time in seconds vs number of CPU's for NYCA case

for some IPP's, they want to assure that some plants will be selected for generation, for example, nuclear plants and some coal steam plants. Because whatever they bid for generation, they will be paid by the positive market clearing price. Thus, they have the incentive too keep those cheap power sources online. From the table, it is obvious that high penetration of wind power will save money for the New York control area. The plots for marginal regulation costs are given in figure 4. Because the increment in regulation requirements, the marginal costs for regulation generally increase. Because of the location-based reserve services, different sub-zones might have different marginal reserve cost services. Besides, we note that the Lagrangian multiplier for transmission constraints increases as the penetration level of wind energy grows. This is reasonable for New York state because most of the wind power resources in NYCA are located in north parts while most of electricity consumption are located in southeast regions. Thus, the increased power penetration will bring more pressure on transmission lines in New York State.

 ISBN: 978-1-61208-189-2

Penetration Level (MW)	Operation Costs
1275	-1.12×10^7
4250	-1.23×10^7
6000	-1.27×10^7
8000	-1.30×10^7

Table I
TOTAL OPERATION COSTS FOR DIFFERENT PENETRATION LEVEL OF WIND POWER

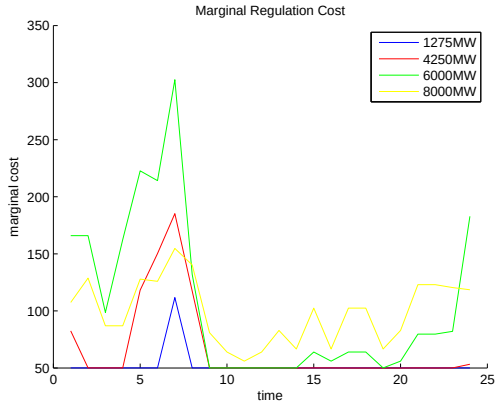


Figure 4. Marginal Regulation Costs for different wind penetration level

B. Rolling horizon study

The proposed method has been applied to solving large sized problems based on the ten-unit system of [19], which has been repeated 100 times so that the problem comprises 1000 units. The generator parameters are slightly perturbed because it is unrealistic to have so many identical generators. The load profile is based on the System D in [20], which has been multiplied by 100 accordingly. We assume 25% of wind penetration level, which is close to the $8000MW$ case in NYCA. Figure 5 shows that the computational time decreases dramatically when the number of processors increases from 1 to 20, and the minimum computational time is about 20% of the sequential computational time. This result is not as good as that for the NYCA case, which involves much more constraints. For NYCA case, the computational time decreases from 1600 seconds (around 0.5 hours) to 3 minutes, which make it reasonable to restart the UCP solver every hour when new information is available.

The result of rolling horizon approach was compared with the traditional day-ahead planning method. In current market, the stochastic problem was solved once every day, and only dispatch problem was solved when the real data was available. The operation costs of the next 24 hours of both approaches were compared. The result is shown in Figure 6. A significant reduction of cost is observed when applying rolling horizon approach. For stochastic problem with wind energy, the cost reduction (approximately 3%) is more significant.

VI. CONCLUSION

In this paper, we have formulated a security constrained unit commitment problem by incorporating complex ancillary services, security, and local reserve constraints, and applied this model to the New York Control Area. We investigated the impact of the increasing penetration of wind power on the New York state day-ahead power market. Additionally, the test results show that parallel computing can significantly reduce the computational time, which makes it possible for rolling horizon implementation of the algorithm. Our testing results on a standard test system show that the rolling horizon approach may lead to significant cost reduction over the traditional day-ahead approach.

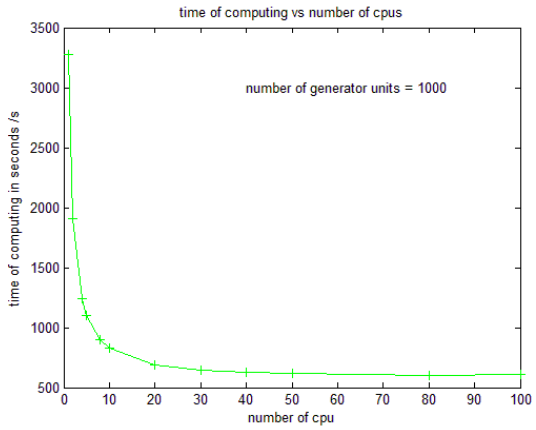


Figure 5. Relationship between computational time (in seconds) and number of CPU's

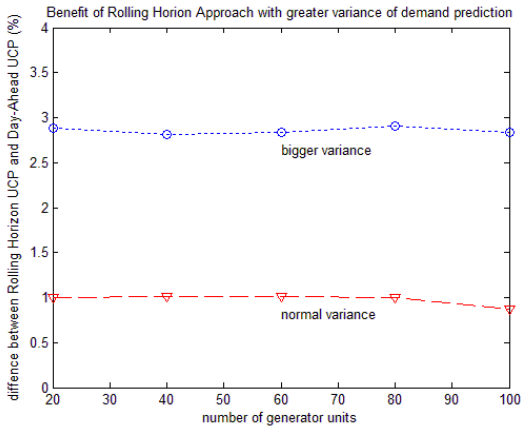


Figure 6. Comparison between Rolling Horizon Approach and Day-ahead Approach

REFERENCES

[1] A. Cohen and V. Sherkat, "Optimization-based methods for operations scheduling," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1574 – 1591, Dec. 1987.

[2] N. Padhy, "Unit commitment-a bibliographical survey," *Power Systems, IEEE Transactions on*, vol. 19, no. 2, pp. 1196 – 1205, May 2004.

[3] J. Shaw, "A direct method for security-constrained unit commitment," *Power Systems, IEEE Transactions on*, vol. 10, no. 3, pp. 1329 –1342, Aug. 1995.

[4] C.-L. Tseng, S. S. Oren, C. S. Cheng, C. Li, A. J. Svoboda, and R. B. Johnson, "A transmission-constrained unit commitment method in power system scheduling," *Decision Support Systems*, vol. 24, no. 3-4, pp. 297 – 310, 1999.

 ISBN: 978-1-61208-189-2

[5] H. Ma and S. Shahidehpour, "Unit commitment with transmission security and voltage constraints," *Power Systems, IEEE Transactions on*, vol. 14, no. 2, pp. 757 –764, May 1999.
[6] Z. Li and M. Shahidehpour, "Security-constrained unit commitment for simultaneous clearing of energy and ancillary services markets," *Power Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 1079 – 1088, May 2005.
[7] R. Barth, H. Brand, P. Meibom, and C. Weber, "A stochastic unit-commitment model for the evaluation of the impacts of integration of large amounts of intermittent wind power," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*, June 2006, pp. 1 –8.
[8] NYISO, *Day-Ahead Scheduling Manual*, June 2001.
[9] S. Takriti, J. R. Birge, and E. Long, "A stochastic model for the unit commitment problem," *IEEE Transaction on Power System*, vol. 11, no. 3, pp. 1497–1508, 1996.
[10] "Wind power integration in liberalised electricity markets (Wilmar) project." *Available at http://www.wilmar.risoe.dk <retrieved: Jan., 2012>*.
[11] A. Tuohy, E. Denny, and M. O'Malley, "Rolling unit commitment for systems with significant installed wind capacity," in *Power Tech, 2007 IEEE Lausanne*, 1-5 2007, pp. 1380 –1385.
[12] A. Tuohy, P. Meibom, E. Denny, and M. O'Malley, "Unit commitment for systems with significant wind penetration," *Power Systems, IEEE Transactions on*, vol. 24, no. 2, pp. 592 –601, 2009.
[13] U. Ozturk, M. Mazumdar, and B. Norman, "A solution to the stochastic unit commitment problem using chance constrained programming," *Power Systems, IEEE Transactions on*, vol. 19, no. 3, pp. 1589 – 1598, 2004.
[14] B. C. Ummels, M. Gibescu, E. Pelgrum, W. L. Kling, and A. J. Brand, "Impacts of wind power on thermal generation unit commitment and dipatch," *IEEE Transaction on Energy Conversion*, vol. 22, no. 1, pp. 44–51, 2007.
[15] J. Birge and F. Louveaux, *Introduction to stochastic programming*. New York, Berlin, Heidelberg: Springer, 2000, vol. Vol.II.
[16] M. Carrion and J. Arroyo, "A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 21, no. 3, pp. 1371 –1378, Aug. 2006.
[17] NYISO, "Final report of the NYISO 2010 wind generation study," institution, Tech. Rep., 2010.
[18] ——, *Market Participants User's Guide*, May 2011.
[19] S. Kazarlis, A. Bakirtzis, and V. Petridis, "A genetic algorithm solution to the unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 11, no. 1, pp. 83 –92, Feb 1996.
[20] U. A. Öztürk, "The stochastic unit commitment problem: A chance constrained programming approach considering extreme multivariate tail probabilities," Ph.D. dissertation, University of Pittsburgh, 2003.

APPENDIX

NOMENCLATURE

$p_{m,i_m,t}$ electricity output level of generator i_m in zone m at time period t.
$z_{m,i_m,t}$ binary variable that is 1 if generator $i_m$in zone m is on during time period t; 0 otherwise.
$F_{m,i_m,t}$ fuel cost function of generator i_m in zone m at time period t.
$S_{m,i_m,t}$ startup cost function of generator i_m in zone m at time period t.
$R_{m,i_m,t}$ reserve service cost function of generator i_m in zone m at time period t.
$r10s_{m,i_m,t}$ 10-minute spinning reserve level of generator i in zone m at time period t.
$r10ns_{m,i_m,t}$ 10-minute non-synchronous reserve service level of generator i_m in zone m at time period t.
$r30s_{m,i_m,t}$ 30-minute spinning reserve level of generator i in zone m at time period t.
$r30ns_{m,i_m,t}$ 30-minute non-synchronous reserve service level of generator i in zone m at time period t.
$r_{m,i_m,t}$ reserve service vector defined as $(r10s_{m,i_m,t}, r10ns_{m,i_m,t}, r30s_{m,i_m,t}, r30ns_{m,i_m,t})$.
$CReg_{m,i_m,t}$ regulation cost function of generator i_m in zone m at time period t.
$reg_{m,i_m,t}$ regulation service level of generator i_m in zone m at time period t.
$d_{t,m}$ prediction of electricity demand of time period t in zone m.
$w_{t,m}$ prediction of wind power of time period t in zone m.
Res_{10s} 10-minute spinning reserve requirement for the whole ISO control area.
Res_{10t} 10-minute total reserve requirement for the whole ISO control area .
Λ_j super-zone j or the jth collection of zones.
$ResLB_{j,10s,t}$ 10-minute spinning reserve requirement for sub control area j at time t.
$ResLB_{j,10t}$ 10-minute total reserve requirement for sub control area j at time t.
$p_{m,i_m,max}$ maximum output when generator i_m in zone m is on.
$p_{m,i_m,min}$ minimum output when generator i_m in zone m is on.
$\Gamma_{l,m}$ line flow distribution factor for the transmission line l due to the net power injection of zone m
$Tran_{i,max}$ maximum transmission capacity of transmission line l in designate direction.
$Ms_{i_m,m}$ maximum number of times that generator $i_m$in zone m is allowed to be shut down.

A Survey on Smart Grid Technologies in Europe

Luca Ardito, Giuseppe Procaccianti, Giuseppe Menga, Maurizio Morisio
Dipartimento di Automatica ed Informatica
Politecnico di Torino
Torino, Italy
e-mail: name.surname@polito.it

***Abstract*—The old electricity network infrastructure has proven to be inadequate, with respect to modern challenges such as alternative energy sources, security, electricity demand and energy saving policies. Moreover, ICT technologies development seems to have reached an adequate level of reliability and flexibility in order to support an entirely new concept of electricity network - the Smart Grid. In this work, we try to give a definition of what a Smart Grid is. Moreover, we will analyse the state-of-the-art of Smart Grids, not only in their technical details, but also in their management and optimization.**

Keywords-smart grid; renewable energy; grid intelligence; energy efficiency; energy storage;

I. INTRODUCTION

Over the past 50 years, electricity networks evolved from the ”local grid” networks of the beginning of the century, to interconnected electric grids, based on generating stations of notable scale (1000-3000 MW) distributing power to major load centres, which divided energy to a large number of individual consumers. The generating stations, or power plants, were built in order to provide massive amounts of energy, due to the nature of power generation technologies in use (hydroelectric, coal, oil, and gas). By the end of the 20th century, however, this model proved to be unreliable and inadequate. First of all, the demand forecast techniques and the data processing technologies could not efficiently provide the desired energy at the desired time; thus, power distribution was based upon rough average classifications. Moreover, the emerging environmental issues and the geopolitical interdependence of power sources limited the development of economies of scale. The main challenges that a modern electricity network has to face are: [1]

- Privacy issues between energy suppliers and customers
- Security threats from cyber attack
- National goals to employ alternative power generation sources
- Significantly more complexity in maintaining stable power with intermittent supply
- Conservation goals that seek to lessen peak demand surges during the day so that less energy is wasted in order to ensure adequate reserves
- High demand for an electricity supply that is uninterrupted
- Digitally controlled devices that can alter the nature of the electrical load and result in electricity demand that is incompatible with a power system that was built to serve an analog economy.

These challenges require the development of an intelligent, self-balancing, integrated electric network that makes use of the modern ICT techniques to manipulate and share data. The Smart Grid technology tries to answer these needs.

In this survey, we propose an overview of the main aspects of Smart Grids development and implementation. In Section II, we give some definitions of the smart grid concepts, from different points of view. In Section III, we will analyse the management process of the Smart Grid. In Section IV, we will review its technical aspects. In Section V, we will see how a Smart Grid can be optimized. In Section VI, some conclusions are given.

II. WHAT IS A SMART GRID?

The Smart Grid is a complex system, and can be described in various ways. Here, we report two different definitions. The first one sums up the ”European” view of the Smart Grid:

”A Smart Grid is an electricity network that can intelligently integrate the actions of all users connected to it - generators, consumers and those that do both - in order to efficiently deliver sustainable, economic and secure electricity supplies. A Smart Grid employs innovative products and services together with intelligent monitoring, control, communication, and self-healing technologies. Smart Grids development must include not only technology, market and commercial considerations, environmental impact, regulatory framework, standardization usage, ICT and migration strategy, but also societal requirements and governmental edicts.” [2]

The second one, written in the Statement of Policy on the Modernization of Electricity Grid of the United States Government [3], characterizes the Smart Grids by means of a list of achievements. The most relevant are: the use of digital information to improve reliability, security and efficiency; integration of distributed resources and generation; ‘smart’ technologies for metering, communication and automation;

 ISBN: 978-1-61208-189-2

deployment of energy storage technologies (i.e., electric vehicles).

According to [4] it is clear that both definitions combine two dimensions: kWh and bytes. It is not argued the key role of ICT in developing a smart grid, and both viewpoints recognize the growing role of renewable technologies, distributed generation and energy storage. In addition, both visions are focused on efficiency, but with different objectives: the U.S. definition is more energy intensive (and therefore less efficient) than the European one, which sets mandatory targets mainly in energy consumption reduction. The main differences consist in current and future business models: on the European side, unbundling and retail competition will be predominant; on the contrary, the U.S. side will be predominantly characterized by vertically-integrated monopolies. As a result, the European vision will lead to a greater complexity in the definition of business cases, and in a redesign of the industry value chain.

III. Business Management

The development of a Smart Grid does not involve replacing the existing electricity network. Such a process would be impossible for technical and economical reasons. Instead, the Smart Grid development is an enhancement of the existing network, by means of implementing new services and features, while maintaining, as much as possible, the old physical infrastructure. We have to define what functions a Smart Grid must provide. According to the United States Department of Energy's Modern Grid Initiative report, [3] these functions are:

- Self-healing
- Consumer Participation
- High Quality Power
- Support for different types of storage and generation
- Higher efficiency

The Smart Grid Technology also changes radically the Energy Market scenario: new actors may arise, such as Energy Retailers and Traders, Distributed Generation operators, and so on [5] (see Figure 1).

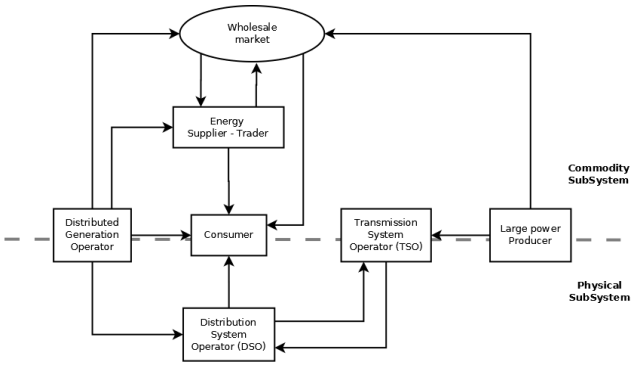


Figure 1. Overview of transactions within the electricity market. [5]

In [6], authors describe three elements as the "pillars" of the Smart Grid. Those elements are:

- *Smart Customer*: the set of technologies that enable consumers to observe and control their consumption
- *Smart Utility*: the utility that implements monitoring, control and pricing, and demand response
- *Smart Market*: an economically efficient market structure to integrate technology, decision making and information

Authors also identify Real-Time Pricing (RTP) as a fundamental tool to realize the Smart Market, because it provides consumers with a transparent way to control their energy bill, and utilities with a rate flexibility that allows them to increase their competitivity and implement Demand-Side Management.

In this new scenario, the roles of producer and consumer get closer. The consumer is now able to produce energy, through distributed renewable energy sources. This new emerging entity is called the *prosumer*, which is discussed in [7]. Authors define it as an economically motivated entity that:

1) Consumes and produces power
2) Operates a small or large power grid, thus transports electricity
3) Optimizes the economic decisions regarding its energy utilization

The prosumer may not be strictly a physical entity, but rather a combination of components: energy sources, loads, an electric grid, controls to operate his system, and a market, or other economic decision making system.

This new market must be supported through a management system that takes into account these new figures. An example of a strategic approach for a complete energy management system is BEMI (Bidirectional Energy Management System). [8] BEMI is an energy management system designed for installation at Low Voltage grid connection points. Its main task is to optimize the so-called Controllable Distributed Electrical (CDE) units, which means locally connected loads or generators. This optimization is done accordingly to consumption and generation tariffs, set by an energy service provider through a Pool-BEMI system. BEMI supervisions the CDE unit switching and operation, and also provides grid costumers complete information about the variable tariffs, energy cost and device schedules. The BEMI System is shown in Figure 2.

Another project worth to be mentioned is SmartGen [9], an Italian project driven by several industries and two different research institutes (University of Bologna and Genova). This project aims at finding and implementing industrial solutions for Smart Grid management. The authors propose the definition of a DMS (Distribution Management System) for each portion of the grid, able to control and optimize power flows, distributed generation and load balancing. The

 ISBN: 978-1-61208-189-2

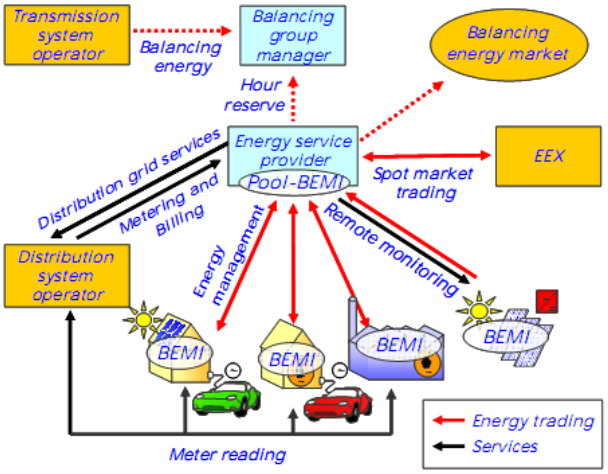


Figure 2. BEMI System in the liberalized energy market. [8]

base function of a DMS can be divided into:

- Supervisory Control and Data Acquisition (SCADA);
- Control Stations

The SCADA system provides specific monitoring and real-time control operations, in an automated way, while the Control Stations allow human operators to interact with the system.

IV. Technology

In this section, we will review some technical aspects of the implementation of a Smart Grid and its features.

A. Distributed Generation

Distributed generation (DG) is a driving factor for the Smart Grid implementation. Its integration in the energy network proved to bring many benefits [10] for customers, energy efficiency, and network operation itself. This integration is enabled through a number of different technologies [11], some of which are discussed in this survey:

- Advanced Metering Infrastructure (see Section IV-B)
- Energy Storage Systems (see Section V)
- Advanced Distributed Management Systems (see Section III)

In [11], authors provide a methodology for integration of DG in a Smart Grid Network. It is based upon the connection of a distributed generator with a feeder, combined with an Automatic Voltage Control (AVC) system and a Dynamic Line Rating (DLR) function. In their study, they also provide an economic feasibility study, as a series of steps, which may be extended to a general project involving DG technologies. The defined steps are:

- Define the installation, operation and maintenance costs of the project
- Define additional financial parameters like electricity rates, discount rates, inflation rate, etc.
- Quantify additional benefits brought to the network, in terms of a premium to the electricity rate per output unit
- Evaluate externalities, such as Greenhouses gases (GHG) reduction, to add them as a benefit
- Calculate the economic parameters internal rate of return (IRR) and net present value (NPV) to evaluate the feasibility of the project.

B. Metering

In order to efficiently implement a Smart Grid, a smart metering infrastructure is essential. Traditional metering devices, provided by energy distribution companies for their customers, typically measure energy consumption only in terms of total energy consumed, thus not giving any information about when and how it is consumed. This information can lead to a more intelligent energy provision, finely tuned to suit specific customer needs - and optimizing energy distribution over all the network. In this context, it is impossible not to cite AMI (Advanced Metering Infrastructure). AMI features include [12] [13]:

- Two way communication to the electric meter to enable information interchange
- Self registration of metering points
- Auto-configuration after a failure in communications
- AMI system interconnection to utility billing, outage management systems, and other applications

In [12], the authors name the integration between Smart Grid and AMI as AGI - standing for Advanced Grid Infrastructure. The AGI has the following enhancements [12]:

- **Outage**: *Improved Customer Service*
  Utilizing the AMI infrastructure, a utility can know when an outage occurs. The AMI system can notify the trouble call system automatically, facilitating rapid crew deployment and reduced outage times.
- **Loss Detection**: *Improved Network Operation*
  By connecting information nodes at key points of the medium voltage distribution lines and distribution transformers, it is possible to directly calculate the system technical and non-technical losses. This enables better tracking and efficiency on the distribution network.
- **State Estimation**: *Integration of Renewable Sources*
  By utilizing information from the customer site, medium voltage lines, and transformers, accurate load models can be computed allowing accurate load estimation on the distribution grid. This information is critical to understanding the impact and benefit of connecting renewable energy sources to the distribution grid.

Also in [14], a more detailed view of the AMI infrastructure is given.

An open issue regarding metering is determining power consumption of ICT equipment. According to [15], ICT technologies have a relevant impact over power consumption. Thus, the problem of finding precise models and figures

 ISBN: 978-1-61208-189-2

to calculate and estimate this consumption has become a priority for the professionals of the sector. In [16], authors expose a case of study regarding the analysis of power consumption data from a data center, in a time period of about a year. The analysis was mostly focused on servers, which represent the main ICT power consuming device in these structures (if we exclude cooling and illumination). This work has shown that, undoubtedly, software has a relevant impact over servers' power consumption (up to 10%). This means that, in order to correctly forecast ICT power consumption, many factors have to be considered, such as the usage profiles of the equipment.

C. Forecasting

Forecasting is a key functionality of a Smart Grid system. Through forecasting, the Grid is able to balance loads, optimize power distribution and handle failures. The main problem of forecasting, in a modern electricity network, is given by the Renewable Energy Sources (RES). The energy produced by RES can vary, and its variation depends on several parameters (climate conditions, source plant location, etc.) In this sense, their contribute in terms of energy can be difficult to predict, because the variables to observe are too many. In this sense, it is worthy to cite the work of Bertani et al. [17], where is presented a central dispatcher with the following functions: short-term forecast of the power produced by renewable energy sources (RES), short-term load forecast and day-ahead load profile prediction, distribution system state estimation, day-ahead economic dispatching and on-line scheduling of the optimal distributed resources' operating conditions. The forecast algorithm was based on a neural network. The results can be seen in Figure 3.

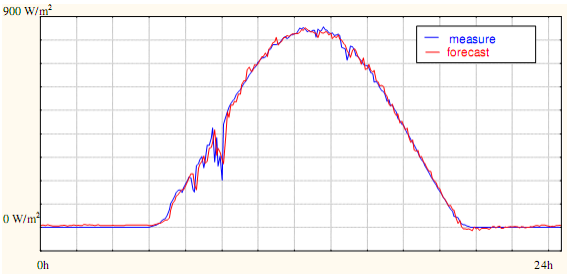


Figure 3. Example of forecasted vs. measured global radiation. [17]

D. Communication

A key issue for successfully realizing the future-oriented energy system is integrating information and communication on an Internet-based infrastructure,able to provide access to energy information in a simple, quick and economic way. This is because energy providers, either centralized or decentralized, need a constant flow of real-time updates regarding the energy demands, in order to provide the precise amount of energy needed. Optimization of energy consumption is based on integrated and near-real-time electronic communication between producers and loads on all levels of the grid.

This infrastructure will also be profitable for consumers. In fact, by developing intelligent end devices, customers will be able to monitor their actual power consumption in real time, and consequently they can optimize the usage of their devices in order to reduce costs. [18]

An example of integration between the Smart Grid and the ICT is a solution proposed in [19]. In modern energy distribution systems, generation and demand need to be always matched in real-time. This means that the modern grid is a real-time distributed system, thus it needs a precise synchronization between its devices. Modern grid infrastructures realize several functions, such as protection testing, fault detection, load balancing and scheduling through synchronization. The authors propose a solution based on the implementation of the Network Time Protocol (NTP) over 802.11 networks along with an optimisation technique to reduce the energy usage of a common Wireless Sensor Network (WSN) synchronisation protocol. [19]

The Service-Oriented Architecture (SOA) provides concepts particularly suitable for an energy distribution network. In fact, it decouples functionalities from implementation, integrating them through message exchange protocols in a dedicated Service Bus. Moreover, it is not needed to develop interfaces between every application: each application only needs to be interfaced to the integration platform.

The only issue of the SOA is finding the correct semantics for data. Without open interface definitions and a standard semantic structure for message exchange, it is not possible to realize an efficient energy network.

Web Services are especially useful in the context of Smart Houses. An Energy-Aware Smart House is a residential building equipped with a Smart Metering system (see Section IV-B) able to measure and control in real-time the power consumption of every electrical device installed. In [20], a Web-Oriented Application Framework for embedded devices is presented. The framework is based on a RESTful architecture.The embedded devices represent sensor nodes, which may provide all sort of information (power consumption, for instance). This solution has shown a response time for querying each device lower than 60 ms, even with high workload.

The suitability of Web Service architectures for the Smart Grid/Smart Houses is also stressed in [5].

V. Optimization

In this section, we will present how a Smart Grid network can be optimized through new technologies and approaches.

A. Agents

A Smart Grid is, by itself, a decentralized network, where intelligence is distributed across several devices. These de-

 ISBN: 978-1-61208-189-2

vices may have to take autonomous decisions, in order to react quickly and efficiently to changes in energy demands, faults, and such events.

Thus, the Software Agents paradigm may provide a way to implement a system like that. In fact, in this paradigm, it is possible to design a distributed system with specific functionalities through the cooperation of autonomous, intelligent components.

In [21], authors present a Multi-Agent System (MAS) simulating a Smart City. The simulated entities were:

- Houses
- Appliances (Single devices, of different classes, installed into a house)
- Vehicles (Electric Vehicles able to store energy into batteries)
- Cities
- Power Stations

The system was implemented using JADE (Java Agents Development Environment). Each entity was represented by a software agent. Then, an energy controller agent is able to act in order to balance power demand and power generation (for example, turning off some devices when power consumption is too high). In Figure 4 are shown the results of the balancing activity. The proposed scenario involved 300 houses evenly divided into three cities, and a total of 3840 appliances.

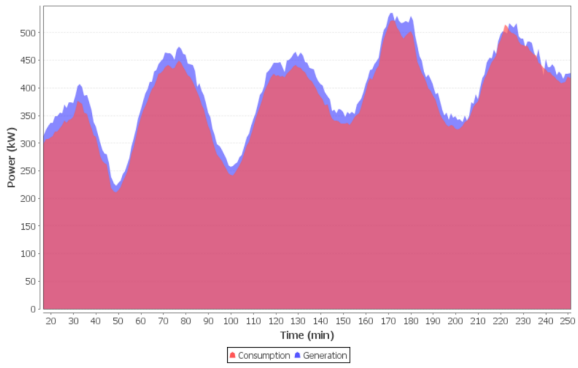


Figure 4. Consumption and generation chart [21]

MAS are often associated to *electronic markets*, computing frameworks for distributed decision making based on microeconomics and Game Theories. By applying this paradigm to the energy distribution networks, we can make use of the already developed techniques and methodologies to realize the so-called *Market-Based Control*.

Several works following this idea have been proposed. In [22], for example, a complex multi-agent architecture is presented, composed of different components: : the Problem Formulator and Attributes Evaluator (PROFATE), the Scenarios Builder, the Electricity Market Multi-Agent System (EMMAS), the Decision Making Assistant (DMA).

Of these components, the most interesting is without any doubt the EMMAS. In order to forecast market prices, both at medium and long-term, accurate simulation models are needed, able to react to structural changes. The EMMAS realizes these models, by means of a complex taxonomy of software agents that represent every actor in the transaction process

Another example of a MAS designed for controlling energy networks is *PowerMatcher*.

"The Power Matcher is a general-purpose coordination mechanism for balancing demand and supply in clusters of Distributed Energy Resources. These 'clusters' might be electricity networks with a high share of distributed generation or commercial trading portfolios with high levels of renewable electricity sources, to name a few.
Within a PowerMatcher cluster, the agents are organized into a logical tree. The leafs of this tree are a number of local device agents and, optionally, a unique objective agent. The root of the tree is formed by the auctioneer agent, a unique agent that handles the price forming, i.e., the search for the equilibrium price. In order to obtain scalability, concentrator agents can be added to the structure as tree nodes." [23]

B. Energy Storage

Another aspect that can substantially improve the efficiency of a Smart Grid is the energy storage. Basically, it is the problem of keeping energy available directly on the grid, in storage components efficient enough to avoid energy losses. In cases of energy production peaks, when there is an overproduction of energy, having a distributed storage system can increase the overall efficiency.

An agent-based solution is exposed in [24]. Basically, they propose a game-theoretic framework, that analyses the Nash equilibrium of an electricity network, and develops learning strategies for agents that dynamically adapt to the energy market. As regards storage devices, they embrace the so-called Vehicle-to-Grid (V2G) view, where the unused energy is stored in the batteries of electric vehicles (EVs) or Plug-in Hybrid Electric Vehicles (PHEVs). Since this solution can raise problems of peaks in energy demand, a Multi-Agent System is adopted in order to optimise usage and storage of electricity. In particular, the proposed system models a situation where each device is represented as an intelligent software agent, and every agent can try to "buy" the needed amount of energy at every time, meanwhile learning which is the most profitable amount of energy to buy, according to the specific usage. The authors claim that, implementing their solution, a single consumer may save up to 13% on his electricity bill. [24]

C. Unit Commitment Problem

One of the key objectives of a Smart Grid architecture is dispatching energy from all the available sources in order to

 ISBN: 978-1-61208-189-2

meet the electric load. In other terms, there is a problem of coordination between energy demand and generation. This problem has been formalized under the name of Unit Commitment. Unit commitment (UC), also known as pre-dispatch, is the problem of scheduling the production of energy by generation units of a power system. The objective is to minimize total production costs, while observing several operating constraints. Thus, UC is a complex mathematical problem, based on both integer and continuous variables. In order to solve this problem, an optimized algorithm is needed, because complete enumeration of all the possible solutions would require excessive computation time. [25] For this survey's purposes, we analysed two possible solutions, which involve different approaches for solving the Unit Commitment Problem. In [26], authors propose a solution based on Adaptive Dynamic Programming (ADP).

The solution presented by the authors focuses on a specific family of ADP: the Heuristics Dynamic Programming (HDP).

*"The implementation is divided into **action network**, **critic network** and **model network**. The function of **action network** is to determine the feasibility region of operation of the power systems and to detect the emergency state with corresponding violations under different contingencies. The function of **critic network** is for post-optimization process, evaluation and assessment of control options during contingencies. And the function of **model network** is to read power system parameters and obtain distribution function for state estimation of measurement errors inherent in data, ascertain and improve accuracy of data. The aim of all these kinds of methods is to approximate the cost-to-go function which is relative to the output of critic network."* [26]

Another approach for the UC problem is presented in [27], where the authors introduce a solution using Genetic Algorithms (GAs).

The application of the GAs to the UC problem included encoding each solution with a simple binary alphabet. At first, a number of initial binary-coded solutions (*genotypes*) are produced randomly to form the initial population. Then, a fitness value is given to each solution, calculated as a sum of penalties for violating certain problem constraints. Afterwards, a new offspring genotype (new solution) is produced by means of the two basic genetic operators: *crossover* (combining different solutions by mixing their binary codes) and *mutation* (modifying randomly chosen bits of the offspring genotypes. The above procedure is repeated until a new set of genotypes is produced, which is considered as the new generation of solutions. The new generation totally replaces the parents. By also implementing some adjustments to the fitness calculation, the GA technique has proven to converge in the order of hundreds of generations.

VI. CONCLUSIONS

In this work, we surveyed the Smart Grid project from different points of view, analysing the efforts that the scientific community is making to implement this infrastructure. We presented complete management solutions, communication systems, and different kinds of optimization techniques.

One of the facts that this survey has shown is that, from a technological point of view, there are plenty of solutions already available. Several management systems have been tested and are ready for deployment. However, another fact is evident: although many different standards exist, especially for data communication and protocols, few of them have been widely accepted for application in energy distribution networks. To give an idea of the problem, below some of the standard communication protocols are listed:

- Application Level:
  - IEC TS62351 (data and communication security)
  - IEC 62443 (safety)
  - IEC 61968 (integration of applications into electricity supply facilities)
- Transport Level:
  - IEC 61850 (Station automation)
  - IEC 62055 (Electric Meter)
  - ISO/IEC 14543-3(KNX)
- Communication Media Level:
  - IEC 60255 (Protection Installation)
  - IEC 81334 (PLC)

This can be an issue, because one of the keys to an efficient energy network is interoperability between different energy providers. A partial solution can be using Web Services and system integration techniques, but there has to be a standard definition for data structures and models in order to enlarge the scope of the network. The biggest obstacle to standardization, and in general to Smart Grid implementation in Europe, from our point of view, is given by the complex situation of the European energy market, where regulated and liberalized regimes still coexist. In regulated markets, the main grid operator establishes a monopoly business, that does not allow consumers to choose among different technologies, as regards, for example, metering services, forecasting, and so on. Also, energy retailers, although present on the territory, are not able to assume their innovative role in the Future Energy Market depicted in Section III, in terms of demand response, consumer services and network operation.

As far as it concerns the research activity, what should be done is embracing a common view of the problem, focusing on interoperability and supporting the creation and affirmation of technology standards. In this way, the development of solutions and optimization techniques can be immediately followed by field testing and deployment, speeding up the overall infrastructure realization process.

 ISBN: 978-1-61208-189-2

REFERENCES

[1] Smart Grid Working Group - Energy Future Coalition, "Challenge and opportunity: Charting a new energy future, appendix a: Working group reports," 2002.

[2] Smart Grids European Technology Platform, "Smartgrids - strategic deployment document for european electricity networks of the future," Apr. 2010.

[3] US Government, Approved by US Congress in December 2007, "Energy independence and security act - sec. 1301 - 1308," 2007.

[4] R. Bigliani, "Why smart grids are different in europe and the u.s." 2009. [Online]. Available: http://http://idc-insights-community.com/posts/f2c76f6bec

[5] C. Warmer, K. Kok, S. Karnouskos, A. Weidlich, D. Nestle, P. Selzam *et al.*, "Web services for integration of smart houses in the smart grid," in *Grid-Interop - The road to an interoperable grid, Denver, Colorado, USA*, Nov. 2009.

[6] R. Tabors, G. Parker, and M. Caramanis, "Development of the smart grid: Missing elements in the policy process," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, Jan. 2010.

[7] S. Grijalva and M. Tariq, "Prosumer-based smart grid architecture enables a flat, sustainable electricity industry," in *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES*, Jan. 2011.

[8] J. Ringelstein and D. Nestle, "Application of bidirectional energy management interfaces for distribution grid services," in *CIRED*, 2009.

[9] A. Borghetti, C. Nucci, M. Paolone, A. Morini, F. Silvestro, and S. Grillo, "Generazione diffusa, sistemi di controllo e accumulo in reti elettriche," *AEIT*, no. 11/12, 2010.

[10] P. Daly and J. Morrison, "Understanding the potential benefits of distributed generation on power delivery systems," in *Rural Electric Power Conference, 2001*, 2001.

[11] R. Hidalgo, C. Abbey, and G. Joos, "Integrating distributed generation with smart grid enabling technologies," in *Innovative Smart Grid Technologies (ISGT Latin America), 2011 IEEE PES Conference on*, Oct. 2011.

[12] D. Hart, "Using AMI to realize the smart grid," in *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, July 2008.

[13] A. Weidlich and S. Karnouskos, "Integrating smart houses with the smart grid through web services for increasing energy efficiency," in *10th IAEE European Conference, Energy, Policies and Technologies for Sustainable Economies, Vienna, Austria*, Sep. 2009.

[14] S. Karnouskos, P. G. da Silva, and D. Ilic, "Assessment of high-performance smart metering for the web service enabled smart grid era," in *ICPE*, S. Kounev, V. Cortellessa, R. Mirandola, and D. J. Lilja, Eds. ACM, 2011.

[15] The Climate Group, "Smart 2020: Enabling the low carbon economy in the information age," GeSi, Tech. Rep., 2008.

[16] A. Vetro', L. Ardito, M. Morisio, and G. Procaccianti, "Monitoring it power consumption in a research center: Seven facts," in *Proceedings of ENERGY 2011, The First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*. Mestre, Italy, 2011.

[17] A. Bertani, A. Borghetti, C. Bossi, O. Lamquet, S. Massucco, A. Morini *et al.*, "Management of low voltage grids with high penetration of distributed generation: concepts, implementations and experiments," in *Proceedings of CIGRE*, 2006, cIGRE, Paris, 2006.

[18] BDI, "Internet of Energy - ICT for Energy Markets of the Future," 2010.

[19] J. Shannon, H. Melvin, R. O'Hogartaigh, and A. Ruzzelli, "Synchronisation challenges within future smart grid infrastructure," in *Proceedings of ENERGY 2011, The First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*. Mestre, Italy, 2011.

[20] A. Kamilaris, A. Pitsillides, and V. Trifa, "The smart home meets the web of things," *IJAHUC*, vol. 7, no. 3, 2011.

[21] S. Karnouskos and T. N. de Holanda, "Simulation of a smart grid city with software agents," in *Proceedings of the 2009 Third UKSim European Symposium on Computer Modeling and Simulation*, ser. EMS '09. Washington, DC, USA: IEEE Computer Society, 2009.

[22] E. Gnansounou, S. Pierre, A. Quintero, J. Dong, and A. Lahlou, "A multi-agent approach for planning activities in decentralized electricity markets," *Knowl.-Based Syst*, vol. 20, no. 4, 2007.

[23] J. K. Kok, M. J. J. Scheepers, and I. G. Kamphuis, "Intelligence in electricity networks for embedding renewables and distributed generation," in *Intelligent Infrastructures*, ser. Intelligent Systems, Control And Automation: Science and Engineering. Springer Netherlands, 2010.

[24] P. Vytelingum, T. D. Voice, S. D. Ramchurn, A. Rogers, and N. R. Jennings, "Agent-based micro-storage management for the smart grid," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1 - Volume 1*, ser. AAMAS '10, 2010.

[25] D. Zhang, L. G. Papageorgiou, N. J. Samsatli, and N. Shah, "Optimal scheduling of smart homes energy consumption with microgrid," in *Proceedings of ENERGY 2011, The First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*. Mestre, Italy, 2011.

[26] J. Momoh and Y. Zhang, "Unit commitment using adaptive dynamic programming," in *Intelligent Systems Application to Power Systems, 2005. Proceedings of the 13th International Conference on*, Nov. 2005.

[27] S. Kazarlis, A. Bakirtzis, and V. Petridis, "A genetic algorithm solution to the unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 11, no. 1, Feb 1996.

 ISBN: 978-1-61208-189-2

Error Rate Performance of QPSK-Transmitted Signal for Power Line Communication under Nakagami-like Background Noise

Youngsun Kim, Hui-Myoung Oh, and Sungsoo Choi
Power Telecommunication Research Center
Korea Electrotechnology Research Institute (KERI)
Ansan-city, Gyeonggi-do, Republic of Korea
Email: yskim, hmoh, sschoi@keri.re.kr,

***Abstract*—Power line communication is a candidate for communication technology in electric vehicles and home area networks of smart grid. This paper derives the error rate performance of a QPSK-transmitted power line communication system. The Nakagami-m distribution is used as the closed-form background noise model, which was presented by a previous work. We have evaluated the symbol error rate of QPSK systems with both analytical values and computer simulations, and verified its validity through computer simulations.**

***Keywords*-Power Line Communication; Background Noise; Symbol Error Rate; Nakagami-m distribution.**

I. INTRODUCTION

Power Line Communication (PLC) has evolved so significantly that it can support various services, ranging from low-speed command and control to high-speed multimedia transmission. There has been commercial deployment of PLC technology, e.g., for Automatic Meter Reading (AMR), in recent years [1]. The performance of AMR using PLC has been analyzed by various researchers (e.g., [2], [3]). PLC is also of great interest due to its possibility as a communication technology for smart grids. The requirements for PLC are presented as an energy management and facility automation systems in smart grids [4]. Furthermore, the possibility of vehicle communications using PLC has been studied. The channel characteristics of an in-vehicle power line were measured and analyzed for PLC [5].

Despite the versatility of PLC, its channel is time-variant and corrupted by unpredictable impulsive noises. Thus extensive research activities on effective channel modeling of background noise, and impulsive noise have been conducted [6], [7]. It was proposed that the amplitude of background noise in power-line channel follows the Nakagami-m distribution [8] [9]. In our previous work, a closed-form expression for the real part of background noise was derived [10]. Using this closed-form expression, we derived the Bit Error Rate (BER) performance of (Binary Phase Shift Keying) BPSK transmission over power line channels and verified its validity by simulations [11].

This paper considers the Symbol Error Rate (SER) of a quadrature-phase shift-keying (QPSK) system. The system model for the QPSK system is presented in Section II. As an imaginary portion of the background noise is added to the QPSK system, we derive the Probability Density Function (PDF) of the imaginary portion of the background noise in Section III. In Section IV, we analyze the SER of the QPSK-transmitted signal for PLC with Nakagami-like background noise. We verified the accuracy of the analyzed performance by simulations in Section V. The conclusions are presented in Section VI.

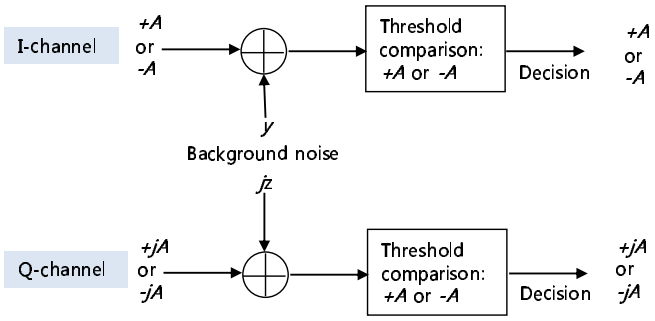


Figure 1. System model for QPSK transmission system

II. SYSTEM MODEL

Fig. 1 shows a conventional baseband system model for QPSK transmission. There are in-phase and quadrature channels that are used for the data transmission from the component set, $\{\pm A\}$. As seen in Fig. 1, the QPSK system can be observed as two BPSK systems. Thus, the data recovery is performed in both BPSK systems in a parallel manner. A simple comparison shows the transmitted data has a threshold value 0. Between the transmission and the decision, both channels have background noises, y and z. We consider the background noise only for the ease of analysis. Applying the channel model added by impulsive noise remains as a subject for future work.

As introduced in the previous section, the amplitude spectrum of the background noise of power line can be modeled as a Nakagami-m distribution. Nakagami-m distribution is conventionally used to model a fading channel of wireless communication due to its versatility [12]. From the random

 ISBN: 978-1-61208-189-2

$$f(z) = \frac{1}{\sqrt{\pi}\Gamma(m)}\sqrt{\frac{m}{\Omega}}e^{-\frac{mz^2}{\Omega}}\left\{\frac{\Gamma(\frac{1}{2}-m)}{\Gamma(1-m)}\left(\frac{mz^2}{\Omega}\right)^{m-\frac{1}{2}}{}_1F_1\left(\frac{1}{2},\frac{1}{2}+m,\frac{mz^2}{\Omega}\right)+\frac{\Gamma(m-\frac{1}{2})}{\sqrt{\pi}}{}_1F_1\left(1-m,\frac{3}{2}-m,\frac{mz^2}{\Omega}\right)\right\} \quad (11)$$

variable for noise amplitude, α, which follows a Nakagami-m PDF,

$$f(\alpha) = \frac{2m^m\alpha^{2m-1}}{\Gamma(m)\Omega^m}e^{-(m/\Omega)\alpha^2}, \quad \alpha \geq 0 \quad (1)$$

where $\Gamma(\cdot)$ is the Gamma function, m and Ω are parameters defined as $\Omega = E[\alpha^2] = \overline{\alpha^2}$, $E[\beta] = \overline{\beta}$ denoting the expected value of α and $m = \frac{\left(\overline{\alpha^2}\right)^2}{(\overline{\alpha^2}-\overline{\alpha^2})^2} > 0$. Thus Ω is viewed as the power of the amplitude, α. Eq.(1) is close to white Gaussian ($m \approx 1$) at high frequencies (about 25MHz), whereas it becomes a one-sided Gaussian ($m < 1$) at low frequencies (about 5MHz).

The noise amplitude can be divided into real and imaginary parts, y and z, respectively. Then, we apply two components by θ, which is a uniformly distributed random variable from $-\pi$ to π, and we get two random variables as

$$y = \alpha\cos\theta, \quad (2)$$

$$z = \alpha\sin\theta. \quad (3)$$

In a previous work [10], we derived the PDF of the real part of noise, y,

$$f(y) = \frac{1}{\sqrt{\pi}\Gamma(m)}\sqrt{\frac{m}{\Omega}}e^{-\frac{my^2}{\Omega}}\left\{\frac{\Gamma(\frac{1}{2}-m)}{\Gamma(1-m)}\left(\frac{my^2}{\Omega}\right)^{m-\frac{1}{2}} \times{}_1F_1\left(\frac{1}{2},\frac{1}{2}+m,\frac{my^2}{\Omega}\right)+\frac{\Gamma(m-\frac{1}{2})}{\sqrt{\pi}} \times{}_1F_1\left(1-m,\frac{3}{2}-m,\frac{my^2}{\Omega}\right)\right\} \quad (4)$$

for $0 < m < 1$ and $m \neq \frac{1}{2}$. The confluent hypergeometric function of the first kind, ${}_1F_1$, is defined as [13, Chap. 9.2] :

$${}_1F_1(a,b,z) = 1+\frac{a}{b}\frac{z}{1!}+\frac{a(a+1)}{b(b+1)}\frac{z^2}{2!}+\frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!}\cdots \quad (5)$$

We derive the PDF for the imaginary portion of noise, z, in a later section.

III. PDF FOR IMAGINARY PORTION OF NOISE

In this section, we derive the PDF for the imaginary portion of noise, z, in Eq.(3). From Eqs.(1) and (3),

$$\frac{dz}{d\alpha} = \sin\theta, \quad (6)$$

then, a conditional PDF of z subjected to θ, $f(z)|_\theta$, can be expressed as

$$\begin{aligned} f(z)|_\theta &= \frac{f(\alpha)}{dz/d\alpha} \\ &= \frac{2m^m\alpha^{2m-1}}{\Gamma(m)\Omega^m}e^{-(m/\Omega)\alpha^2}\frac{1}{dz/d\alpha} \\ &= \frac{2m^m z^{2m-1}}{\Gamma(m)\Omega^m\sin^{2m-1}\theta}e^{-(m/\Omega)\frac{z^2}{\sin^2\theta}}\frac{1}{\sin\theta} \\ &= \frac{2z^{2m-1}}{\Gamma(m)\sin^{2m}\theta}\left(\frac{m}{\Omega}\right)^m e^{\frac{mz^2}{\Omega\sin^2\theta}} \end{aligned} \quad (7)$$

using Eqs.(1), (3) and (6). From the fact that the θ is uniformly distributed over $[-\pi,\pi]$, the joint PDF for z and θ, $f(z,\theta)$, is represented as

$$\begin{aligned} f(z,\theta) &= \frac{2z^{2m-1}}{\Gamma(m)\sin^{2m}\theta}\left(\frac{m}{\Omega}\right)^m e^{-\frac{mz^2}{\Omega\sin^2\theta}}\frac{1}{2\pi} \\ &= \frac{z^{2m-1}}{\pi\Gamma(m)\sin^{2m}\theta}\left(\frac{m}{\Omega}\right)^m e^{-\frac{mz^2}{\Omega\sin^2\theta}}. \end{aligned} \quad (8)$$

The PDF of the imaginary part of noise, $f(z)$, can be obtained

$$\begin{aligned} f(z) &= \int_{-\pi}^{\pi} f(z,\theta)d\theta \\ &= \int_{-\pi}^{\pi}\frac{z^{2m-1}}{\pi\Gamma(m)\sin^{2m}\theta}\left(\frac{m}{\Omega}\right)^m e^{-\frac{mz^2}{\Omega\sin^2\theta}}d\theta \\ &= 4\int_0^{\pi/2}\frac{z^{2m-1}}{\pi\Gamma(m)\sin^{2m}\theta}\left(\frac{m}{\Omega}\right)^m e^{-\frac{mz^2}{\Omega\sin^2\theta}}d\theta. \end{aligned} \quad (9)$$

Letting $\sin^2\theta = t$ gives $d\theta = \frac{dt}{2\sqrt{t}\sqrt{1-t}}$, then, Eq. (9) is

$$\begin{aligned} f(z) &= 4\int_0^1\frac{z^{2m-1}}{\pi\Gamma(m)t^m}\left(\frac{m}{\Omega}\right)^m e^{-\frac{mz^2}{\Omega t^2}}\frac{dt}{2\sqrt{t}\sqrt{1-t}} \\ &= \frac{2z^{2m-1}}{\pi\Gamma(m)}\left(\frac{m}{\Omega}\right)^m\int_0^1 t^{-(m+\frac{1}{2})}(1-t)^{-\frac{1}{2}}e^{-\frac{mz^2}{\Omega t^2}}dt, \end{aligned} \quad (10)$$

which has same integration as that of the real part of noise, $f(y)$, in [10]; thus, we get a closed-form PDF as in Eq.(11) for $0 < m < 1$ and $m \neq \frac{1}{2}$. Then, we can conclude that the amplitude noise of the power line shows the same statistical behavior in both real and imaginary parts of noise.

IV. DERIVATION OF THE ERROR RATE PERFORMANCE

In this section, we derive the error rate performance of the QPSK-modulated signal with noise as described in section II. Fig. 2 shows the signal constellation for the QPSK signal transmission. A transmitter sends the complex symbol, $\pm A\pm$

 ISBN: 978-1-61208-189-2

$$P_1 = \frac{\Gamma(\frac{1}{2}-m)}{2m\sqrt{\pi}\Gamma(m)\Gamma(1-m)}\left[x^{2m}\,{}_2F_2\left(m,m;\frac{1}{2}+m,m+1;-x^2\right)\right]_{x=\sqrt{\frac{m}{\Omega}}A}^{\infty} + \frac{\Gamma(m-\frac{1}{2})}{\sqrt{\pi}\Gamma(m)}\left[x\,{}_2F_2\left(\frac{1}{2},\frac{1}{2};\frac{3}{2}-m,\frac{3}{2};-x^2\right)\right]_{x=\sqrt{\frac{m}{\Omega}}A}^{\infty} \quad (17)$$

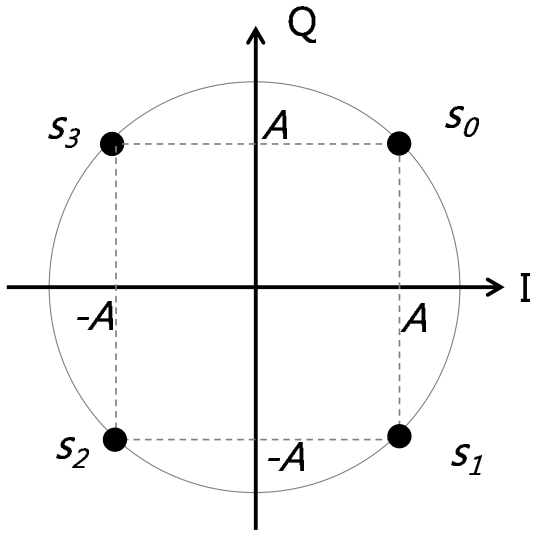


Figure 2. Signal constellation for QPSK

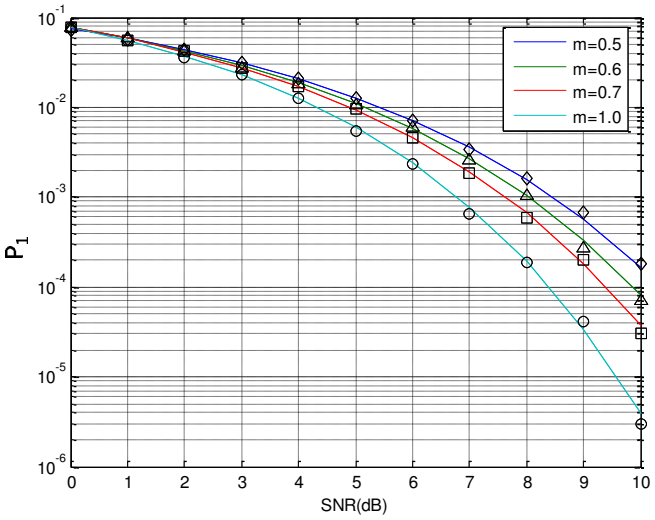


Figure 3. Simulated and analyzed evaluations of P_1 with various m values (reprinted from [10])

jA , according to the symbol set, $\{s_0, s_1, s_2, s_3\}$ then, the complex decision metric at the receiver, r, is defined as

$$r = \pm A \pm jA + y + jz, \quad (12)$$

where I and Q channel data, $\pm A \pm jA$, are corrupted by a complex Nakagami background noise, $y + jz$. We assume that four signals are equally probable; thus, the optimal threshold for decision is x and y-axes. For example, if signal s_0 is transmitted and a noise-corrupted received signal falls in the first quadrant, then the decision is correctly made as s_0; otherwise, it fails.

When a symbol s_0 is sent, the probability of symbol s_0 is correctly decided, $p(C|s_0)$, and is represented by

$$p(C|s_0) = p(y > 0|s_0)\; p(z > 0|s_0), \quad (13)$$

which means a noise-corrupted symbol must fall on the first quadrant. We get

$$\begin{aligned} p(y > 0|s_0) &= 1 - \int_{-\infty}^{0} p(y|s_0)dy \\ &= 1 - \int_{A}^{\infty} f(y|s_0)dy \\ &= p(z > 0|s_0). \end{aligned} \quad (14)$$

The probability of s_0 being correctly decided is

$$\begin{aligned} p(C|s_0) &= \left(1 - \int_A^{\infty} f(y|s_0)dy\right)^2 \\ &= 1 - 2\int_A^{\infty} f(y|s_0)dy + \left(\int_A^{\infty} f(y|s_0)dy\right)^2. \end{aligned} \quad (15)$$

Finally, the symbol error probability, P_e, is given as

$$\begin{aligned} P_e &= 1 - p(C|s_0) \\ &= 2\int_A^{\infty} f(y|s_0)dy - \left(\int_A^{\infty} f(y|s_0)dy\right)^2 \\ &= 2P_1 - P_1^2, \end{aligned} \quad (16)$$

where we can evaluate the integral part, P_1, using the result from [11] as Eq.(17). The simulated and the analyzed evaluations of P_1 are depicted in Fig. 3. The X-axis is a dB-scaled Signal-to-Noise Ratio (SNR) value, which is defined as A^2/Ω. For higher values of SNR, P_1^2 becomes very small; then, Eq. (16) can be approximated as

$$P_e \approx 2P_1, \quad (17)$$

which means that the symbol error rate is approximately twice the BER of BPSK. The result is that the SER of QPSK is poorer than that of BPSK in symbol error sense.

 ISBN: 978-1-61208-189-2

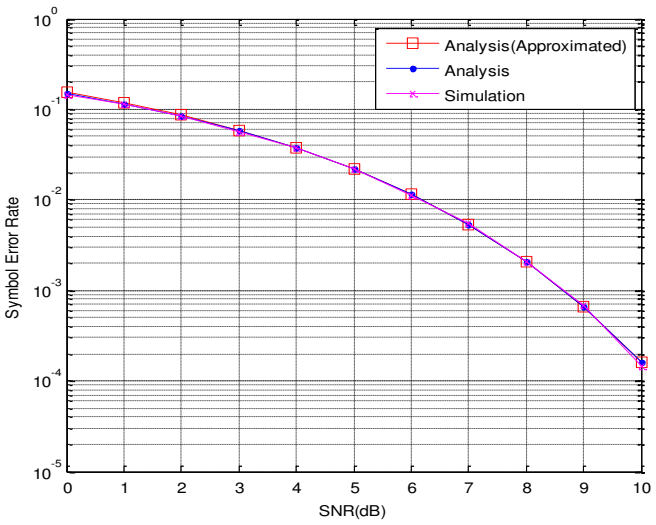


Figure 4. Analyzed and simulated SER performances under Nakagami-like background noise with $m = 0.6$

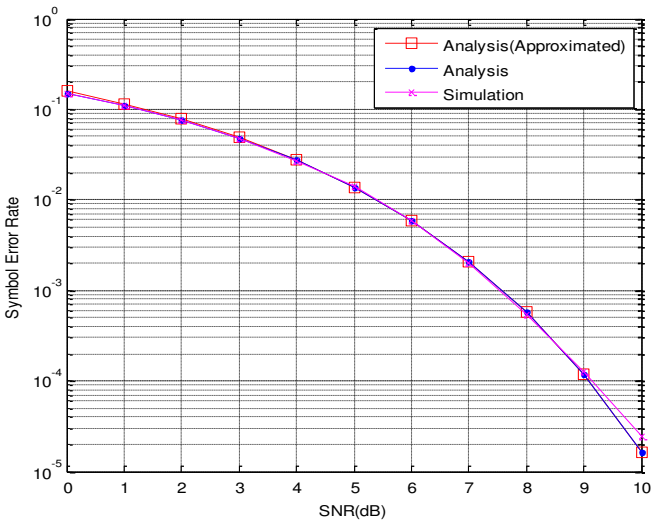


Figure 5. Analyzed and simulated SER performances under Nakagami-like background noise with $m = 0.9$

V. Simulation Results

The SER performance of QPSK is simulated with the Monte Carlo method. 10^6 symbols are randomly generated at the transmitter. The generated symbols are corrupted by complex Nakagami-like background noise; then, the demodulated symbols are decided by a threshold comparison as presented in Fig. 1.

Fig. 4 shows the analyzed and the simulated SER performances for $m = 0.6$. The approximated analysis curves obtained from Eq.(17). The simulation results match well with those of the analysis. The accuracy of the approximated analysis is credible for all SNR values. Fig. 5 shows the analyzed and the simulated SER performances for $m = 0.9$. We can observe that a similar accuracy exists between the simulated and the analyzed performances.

VI. Conclusion

This paper analyzed and simulated the SER performance of the QPSK-transmitted signal under Nakagami-like background noise. We derived the imaginary portion of the Nakagami-like background noise. The derived result shows that the statistical behavior of the real and the imaginary parts of noise are the same. Next, we derived the SER performance of the QPSK-transmitted signal. For higher values of SNR, the SER performance is approximated to twice the error rate performance of a binary transmitted signal. Simulation was done for randomly generated 10^6 symbols through the Monte Carlo method. The analyzed SER performances agree well with the simulations using various values of m. The BER performance of the QPSK-transmitted signal would be the focus of the future work.

References

[1] Y. Zhang and S. Cheng, “Power Line Communications,” IEEE Potentials, vol. 23, no. 4, pp. 4-8, Oct.-Nov. 2004.

[2] A. Zaballos, A. Vallejo, M. Majoral and J. M. Selga, “Survey and performance comparison of AMR over PLC standards,” IEEE Trans. Power Del., vol. 24, no. 2, pp. 604-613, April 2009.

[3] B. Sivaneasan, E. Gunawan and P. L. So, “Modeling and performance analysis of automatic meter-reading systems using PLC under impulsive noise interference,” IEEE Trans. Power Del., vol. 25, no. 3, pp. 1465-1475, July 2010.

[4] G. Mumiller, L. Lampe, and H. Hrasnica, “Power line communication networks for large-scale control and automation systems ,” IEEE Communications Magazines, vol. 48, Issue 4, pp. 106-113, 2010.

[5] M. Lienard, M. O. Carrion, V. Degardin and P. Degauque, “Modeling and analysis of in-vehicle power line communication channels,” IEEE Trans. on Vehicular Technology, vol. 57, Issue 2, pp. 670-679, 2008.

[6] M. Zimmermann and K. Dostert, “A multipath model for the powerline channel,” IEEE Trans. Commun., vol. 50, no. 4, pp. 553-559, April 2002.

[7] M. Zimmermann and K. Dostert, “Analysis and modeling of impulsive noise in broadband powerline communications,” IEEE Trans. Electromagn. Compat., vol. 44, no. 1, pp. 249-258, Feb. 2002.

[8] M. Nakagami, “The m-distribution - A general formula of intensity distribution of rapid fading,” in Statistical Methods in Radio Wave Propagation, pp. 3-36, Pergamon Press, Oxford, U.K, 1960.

[9] H. Meng, Y. L. Guan and S. Chen, “Modeling and analysis for noise effects on broadband power-line communications,” IEEE Trans. Power Del., vol. 20, no. 2, pp. 630-637, April 2005.

 ISBN: 978-1-61208-189-2

[10] Y. Kim, H-M. Oh and S. Choi, "Closed-form expression of Nakagami-like background noise in power-line channel," IEEE Trans. Power Del., vol. 23, no. 3, pp. 1410-1412, Oct. 2008.

[11] Y. Kim, Y-H. Kim, H-M. Oh and S. Choi, "BER performance of binary transmitted signal for power line communication under Nakagami-like background noise," Proc. Energy 2011 : The First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, pp. 126-129, May. 2011.

[12] M. K. Simon and M.-S. Alouini, Digital Communications Over Generalized Fading Channels : A Unified Approach to Performance Analysis, New York: Wiley, 2000.

[13] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, 6th ed., San Diego, CA: Academic, 2000.

 ISBN: 978-1-61208-189-2

Degrees of Freedom in Sharing Control of Smart Grid Connected Devices

A framework for comparison of cross-organizational control sharing mechanisms for balancing supply & demand

Kristian Helmholt,
Department: Business Information Services
TNO
Groningen, The Netherlands
Kristian.Helmholt@tno.nl

Gerben Broenink,
Department: Information Security
TNO
Groningen, The Netherlands
Gerben.Broenink@tno.nl

***Abstract*—Electricity networks require a balance between supply & demand of power in order to maintain stability and to provide a good power quality. The growth of renewable energy sources makes obtaining balance more difficult, because of their intermittent power profiles. Financial incentives for producing 'green electricity' locally also increase complexity due to the larger geographical distribution of electricity generation. Not surprisingly, more sophisticated (distributed) control mechanisms for balance in (smart) electricity grids are being proposed. Some of these proposals attempt to solve the problem of balance by managing demand, and thus introduce the concept of sharing control of devices connected to the grid. However, sharing control could introduce imbalances in 'societal' power between governments, companies and consumers. We propose that all parties involved should consciously decide on what amount of control they want to share. We provide a framework for comparison of control sharing mechanisms.**

Keywords-Smart grid; control sharing; privacy.

I. INTRODUCTION: WHERE DOES IT SAY SHARING IN 'GRID CONTROL'?

The concept of 'control of an electricity grid' can have different meanings. In this paper, we mean control with respect to obtaining balance between supply and demand of power in electricity grids. In an electricity grid, it is quintessential that the total consumption of power is continuously equal to the total production of power. If this is not the case, the quality of the provided power will degrade. In classic grids (as opposed to future 'smart' grids), control mechanisms are already put into place in order to deal with the variation in demand by power consumers. When consumers demand more power from the grid, power producing parties connected to the grid have to provide more power as a whole (group). In the future, more will be demanded from control mechanisms [11]. They have to be able to deal with the increase of more distributed and renewable power sources with variable output (wind, solar, wave, etc.).A solar cell or wind turbine cannot be powered down without wasting valuable energy. Also, wind turbines can not be immediately shut down by turning them away from the wind. Another problem is the fact that the power flow is changing from one way to two way. In the classic grid, there are a few 'centralized' large power plants and many distributed users. In future grids, there might be many distributed small power plants: home-owners with a wind mill, solar panels, geothermal installations, etc. that have a surplus in electricity production. This does not only reduce the accuracy of prediction the production of power – since it is now closely related to the weather–, but also the accuracy of the prediction of power transported across the grid, since locally generated electricity is consumed 'first' before more power is demanded from the grid. Another reason why more intelligence in the energy grid is needed, is that the rise of the usage of Plug-in (Hybrid) Electrical Vehicles (P(H)EV) seems to become a real challenge [12]. It is not unlikely that PHEVs will be plugged into the grid at almost the same time (when people come home from work). This will create a huge demand for power in a relatively short time, possibly resulting in a grid overload. The grid was not dimensioned with all this in mind. With the current grid it seems likely new control mechanisms have to be put in place.

Currently, 'Demand Response' (DR) of devices connected to the grid is being used in several research projects as a new means of control [13]. Depending on the amount of power that is consumed by devices, it can make sense to switch devices on and off in order to attain balance in the grid. Since DR almost always requires somebody or something else than the owner of the device to (automatically) switch on or switch off the device, device owners are no longer fully in control. For example, when DR is applied at charging PHEVs, the charging process may have to wait for a signal 'from the electricity network' that tells the car to start charging.

As a society, we should decide how much we want others to be in control of the grid-connected devices we own. For example, do we want to control our own devices in our own home, as we do now, or do we want having our devices controlled by some 'entity' in the electricity grid in order to have balance in the electricity grid? To make this decision we need a framework for comparison, which we provide in the remaining sections of this paper.

In the next section, related research on this topic will be given. We will see that much research is done, however almost no research is done in comparing different solutions

 ISBN: 978-1-61208-189-2

with each other. After that, our problem description is given, followed by our contribution and methodology. Section V describes our framework, which can be used to compare different demand response mechanisms. After that, the consequences of choices in the framework are explained in Section VI. As an example of the application of our framework to a real situation, Section VII compares two real systems with each other, and mentions their differences. In Section VIII we will draw our conclusions, and finally, in Section IX, the future work will be described.

II. RELATED RESEARCH

The main goal of our research was to be able to compare control mechanisms for the management of supply & demand in smart grids. The comparison should be useful for different stakeholders in society. We want them to be able to decide on the application of these control mechanisms, based on the consequences for their societal position. We did not find research (yet) on that specific subject.

With respect to recent research on control mechanisms themselves we did find different approaches. First of all there is research with a focus on the control of the grid itself. In their description of a High Assurance Smart Grid (HASG) model Overman & Sackman put emphasis on the issue of admission control [1]. They describe a Smart Grid with '*a control system architecture characterized by a distributed architecture that is designed to mitigate against widespread failures when control system components themselves are compromised*'. More on this can be found a later paper from Overman et al., where 'a Three-Part Model for Smart Grid Control Systems' is described [3]. They note that while "*energy flow is now more interconnected and less hierarchical, the energy control system architecture is still largely hierarchical*."**.** Furthermore, they elaborate on a distributed control signaling architecture "*such that some level of device collaboration can be done even when there are losses of control capability from the still dominant hierarchical control system architecture. This is a key feature required for a self-healing grid.*". We suspect that this will be an important aspect of future intelligent networks: distributed control, where no single entity has total control over the entire network. Not only because of the ability to deal with attacks on the network (which is an important aspect in the 'three-part model'), but also because of the fact that one or a few central entities cannot handle all the dynamics of supply & demand with energy sources with an intermittent profile.

Another example of what we think is innovative thinking in control of energy grids themselves, is described by Belkacemi et al. [6]. They use the concept of the Human Immune System (HIS) in order to "*perform self-healing and control of the grid by automatic fault location and isolation, reconfiguration and restoration*." They see the HIS as a Multi-Agent System (MAS), which consists of many different agents that carry out separate tasks with a certain level of autonomy, in order to achieve a goal at a higher level. There is no single control entity which is directly carrying out all control tasks, but control tasks are distributed across nodes. In this paper, however, we want to focus on sharing control between stakeholders. Grid stability and optimization is largely within the realm of a network operator. We also want to be able to take into account parties that the grid for 'energy logistics' and which (may) have to share control. There is also research carried out in this area. For example, an architecture for distributed control of power consuming and producing devices which are attached to the grid, is described by Tariq et al. [4]. They state that "*the advent of renewable generation technologies has resulted in increased complexity, requiring more powerful EMS applications. Regulatory changes in market structures frequently require modifications to these applications*". EMS meaning 'Energy Management Systems'. Next to stating this requirement, they "*describe the elements required for implementation of a "Prosumer" based distributed control architecture for smart grid*". In their description the authors describe four layers of control, that have no knowledge about the workings of the other layers and which only interact on the basis of interfaces between the layers:

- **Device Layer,** concerned with the physical connectivity of electric components.
- **Local Control Layer,** concerned with the control mechanisms of the devices. Examples named are the LTC control of a transformer and the battery charger of an electric vehicle.
- **System Control Layer.** According to the authors Energy Management Systems (EMS) and Distribution Management Systems (DMS) applications are examples of systems control layer for Independent System Operator (ISO) and electric utilities. Also the authors see 'corresponding' system control layers at the level of microgrids, buildings, homes, etc.
- **Market Layer**. Decision control processes at the level of available resources, where economic objectives are taken into account. This layer generates control actions for the system control layer or price signal for the external world, based on information from the system control layer, where market strategies are taken into account.

We state that the concept of 'layering' allows for a necessary separation of concerns, which enables us to deal with the complexity of future smart grids. Another model for control with separation of concerns is described by Molderink et al. [7]. They present a "*three-step control methodology….focused on domestic energy streams*". They refer to an important issue of sharing control from domestic environments: the comfort of residents. Different stakeholders in a smart grid have different goals and/or desires. While network operators might target at system

 ISBN: 978-1-61208-189-2

stability, domestic users may 'just' want to have their devices consume energy in order to provide comfort. At the same time a government might want to target at reducing CO2 emissions by increasing energy (use) efficiency. This requires a combination of local and (more) global optimization. In their paper the authors provide three steps: local prediction, global planning and local scheduling. Together, these steps form one iteration. In this way, the authors think it is possible to combine different goals at different levels of control.

From a quantitative point of view one might state that controlling devices in a domestic environment will not be a real issue for the future since what amount of power can actually be shifted in time in homes? While this is an issue of debate (it also depends on the amount of electrification of heating and cooling equipment in homes), there is one development which certainly cannot be marginalized [10]. This is described by Erol-Kantarci et al. They state that the charging load of Plug-in Hybrid Electric Vehicles (PHEV) can cause several problems if left unmanaged. To that, they discuss an admission control system [8]. If everybody with a PHEV plugs in their PHEV after work, the grid has to transport a lot of power at the same time. Current grids have not been designed with this in mind. A (probably costly) solution would be increasing the capacity of the grid, another solution is to carry out some kind of 'congestion management' where PHEVs are charged 'on-a-turn-basis'. This means sharing control, since the person wanting the PHEV getting charged is probably not the only one deciding on the time of charging if a congestion management mechanism is put into place. More on the important role of PHEVs or 'gridable vehicles' can be found in a paper from Venayagamoorthy, who talks about the complexity of Cyber-Physical Energy System (CPES) [5]. He does not only see 'gridable' vehicles as a consumer of power, but also as a possible producer. This adds an extra dimension of control to gridable vehicles, since this means two-way flow of power, making the control problem more complex.

III. PROBLEM DESCRIPTION

As stated at the beginning of Section II the main goal of our research was to be able to compare (distributed) control mechanisms for supply & demand management, based on consequences for 'societal' power. Our problem then became answering the question 'what is an efficient and useful means of comparison – to be used by different stakeholders - with respect to sharing control of devices connected to the grid, when focusing on consequences for societal power?'

Answering this question required a structured overview of what we call the 'solution space': what kind of variations in Demand Response Management can be distinguished with respect to consequences for sharing control of devices connected to the grid and thus for societal power. To that we needed a overview of what the consequences would actually be when choosing for a specific model of sharing control.

IV. OUR CONTRIBUTION & METHODOLOGY

At TNO we are involved at different research projects, ranging from technical pilots, simulation studies economic evaluation of multi-stakeholder analysis, legislative view. We carry out this research on behalf of different customers. What we describe in this paper is derived from the experience we have had in these projects. The demonstration case "PowerMatcher" is a technology which we use in several other projects.

Our contribution in this paper is a framework for comparison, which contains a structured overview of the degrees of freedom for sharing control of devices connected to the grid. Also it contains a list of consequences caused by choices made with respect a certain degree of freedom. We arrived at this model by analyzing different (partial) Smart Grid designs from the viewpoint of sharing control, while focusing on consequences of design decisions. This resulted in the framework title 'Degrees Of Freedom In Sharing Control' (DOFISC) for Smart Grids (4SG). This approach resembles the approach we took in defining the DOFIS-4SG model, which was focused on 'information sharing' [1].

Like the DOFIS-4SG, the basic structure of the model is a set of axes. Each axis is a 'degree of freedom' and represents an aspect *controlling a device or a group of devices* of another owner. This means not all aspects of smart grid control are included. When control is within one domain of one owner (e.g. network operator), there is very little sharing going on, so these aspects are not taking into account. The aspects of sharing control we did include are related to differences in owner or user of a device connected to the grid. Just as with the DOFIS-4SG model, the basis for the included aspects was found in literature on Smart Grids and our experience with control architectures in other domains (i.e. telecommunications). We distilled the greatest common denominators and make additions where they were necessary in order to provide for making comparisons. And, once again, just as with DOFIS-4SG, this meant that we did not mathematically derive these aspects, but carried out a selection process based on the criteria 1) 'relevant to control devices attached to the grid' and 2) the axes being 'orthogonal'. For more information on the concept of orthogonal axis see [1]. We arrived at a list of possible consequences in a similar fashion. Currently, we suspect that the axes, their subdivisions and the list of consequences can be used to assess and compare aspects sharing control of devices connected to the grid. Providing proof for this should be included in further research, just as it was the case with DOFIS-4SG, which is being evaluated at the moment.

V. FRAMEWORK DESCRIPTION

In this section, we present our framework, which can be used to classify and compare demand response management systems in smart grids. Our framework consists of 3 axes which are orthogonal to each other. An design choice on one axis does not influence an design choice on another axis. Before presenting the axes, we want to make a distinction with respect to different types of control of a device. We base this distinction on the three 'modes of control' of

 ISBN: 978-1-61208-189-2

devices by Overman et al. [3]. With respect to devices *attached* (i.e. not *in*) to the grid we see three types of control:

1. **manual** control, A light that is switched on or off by home owners operating a switch. Another example is a electricity generator running on diesel. This type of control is difficult to share between parties which are not located at the location of the device.
2. **automatic** control, A light that is switched on by movement (e.g. infra-red sensor). Another example are solar panels that produce electricity once the sun is shining. This type of control is difficult to share between the owner of and others, because the control depending on the solar intensity.
3. **remote** control. A washing machine that is switched on by a control device outside the washing machine. This type of control can be shared, especially because of the fact that it is 'remote'.

Our focus in the framework is on remote control, which can be carried out from any location. Also, we understand sharing control to be the sharing of control between different persons and/or organizations. A network operator that uses distributed control mechanisms using multi-agent systems for 'network stability' does not automatically 'share' control with consumers. Only when the network operator has some (indirect) influence with respect to the control of power consuming devices, we consider this to be sharing control.

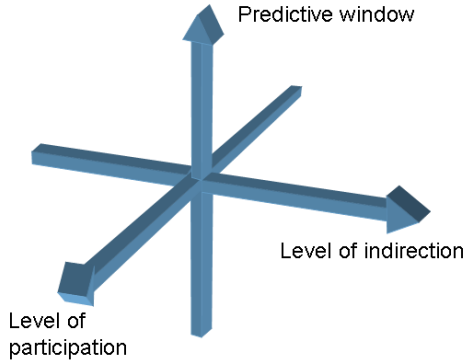


Figure 1. The three degrees of freedom

A. *Predictive window*

The predictive window is a time horizon of a controlling party. A controlling party has to make decisions and communicate them with others parties involved. It matters how far in advance the controlling party has to make its decisions. Some supply & demand designs prescribe a predictive window of 15 minutes, while others might have predictive windows of a day or even more.

An important question with respect to the predictive window is how far in advance does the controlling entity need to plan? For example, in Figure 2, if the controlling entity has a predictive window of 1 hour, it does not know that turning the washing machine on in 15 minutes, will cause a heavy load in 2 hours, when the electric car starts loading. However, if a controlling entity has a predictive window of three hours, it can foresee that starting the washing machine in 15 minutes, will cause a heavy load in 2 hours. And therefore, the controlling entity could make another decision.

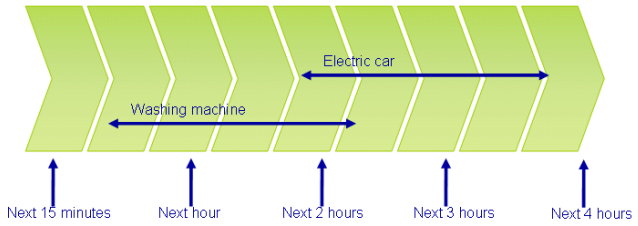


Figure 2. Distance in time

B. *Level of indirection*

The level of indirection determines the amount of freedom in control left after a control decision has been made by a controlling party. An extreme example is 'direct and total control', where the controlling party directly controls a device connected to grid.

There is a fundamental difference between direct control and indirect control. In case of direct control, the user is subjected to the control of the controlling party, and in case of indirect control, the controlling party gives directions with respect to power consumption. This can be in terms of constraints, within which is some freedom left to the consuming party to control devices connected to grid. Also a set point can be given as a direction, where the consuming party has to consume the specified amount of power. The controlling party does not specify which devices have to be switched on or off. An example of direct control is: a washing machine (2000 Watt) is turned on at 16:00, and the electric car (3000 Watt) will be loaded at 18:00. While an indirect example could be: the user will not be able to consume more than 3500 Watts. In both ways, the peak load is avoided, however in the indirect version, the user can chose to the order of the washing machine and the electric car. While in the direct version, the consumer has no choice. In any design of a smart grid, a decision has to be made about which stakeholders is in charge of making which (in)direct control decisions. Note that different stakeholders can provide different constraints to each other. For example, a supplier of power can set a maximum for the amount of power and a network operator can set a maximum for the amount of power which can be transported. In Figure 3, this concept is shown graphically. Three examples of possible solutions are shown:

- Red line: a supplier provides no constraints. The network operator is giving only high level constraints. And the direct control is given at device level.

 ISBN: 978-1-61208-189-2

- Blue line: a supplier is providing no constraints, the network operator is providing some constraints, the home environment provides more restrictive constraints, and finally, the direct control decisions are made at device level.
- Black line: the supplier is providing some constraints, and the network operator takes direct control decisions.

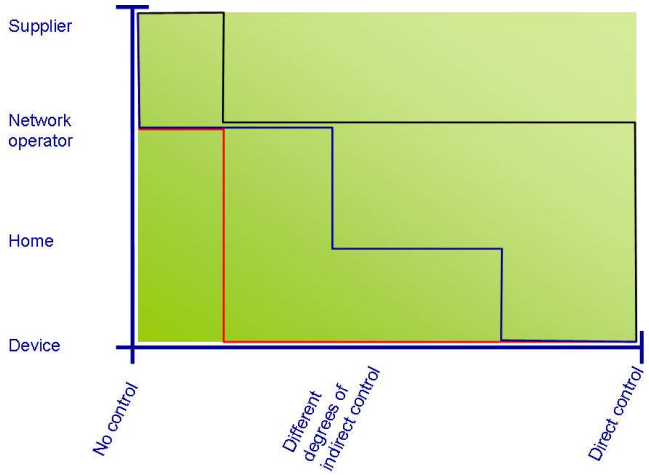


Figure 3. Level of indirection

C. *Level of participation in control decisions*

The level of participation is an important part of the control space. In different smart grid architectures, there are different levels of participation of owners of devices connected to the grid in control decisions. It is theoretically possible that the owner has no say in the control decision at all. For example, in the current energy network, in the Netherlands, no one is allowed to consume a peak load of more then 16 amps on one group of devices (each house can have several groups of devices). This control decision is based on the infrastructure (the infrastructure supports no more than 16 amps), and a consumer has no say in this decision (unless he is willing to pay for a special connection to the energy network). Another extreme is a smart grid, in which all consumers publish there preferences, and a distributed algorithm makes a control decision, satisfying as much as possible participants.

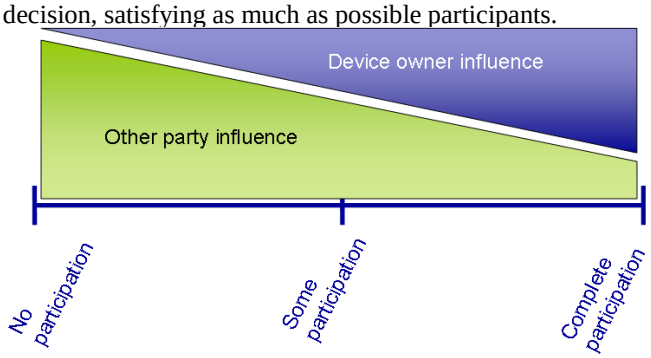


Figure 4. Level of participation

Note that a high level of participation differs from a high level of indirection. The level of indirection is about the decision itself (it is a property of the decision) While the level of participation is about the process of decision making.

In Figure 4, a graphical overview of the possible degrees of participation is provided. One extreme is no influence from owners/users of devices connected to grid, and the other extreme is almost no influence from another controlling party with whom control could be shared.

D. *Applying the framework*

To apply the framework, an inventory of all control decisions - made in a particular smart grid design - has to be made. Important for this phase is to recognize that there can be many different kind of decisions with respect to the control of power in one and the same design, so the framework may have to be applied many times in order to compare different designs. For example, one can imagine a design where parties make (indirect) control decisions: a power supplier who makes a control decision about the minimal load which is to be delivered, a network operator who makes control decisions about the maximal load, and a consumer who makes the direct decision to turn on device like a washing machine. In this case, at least three types of control decisions are made here.

To make things worse, a smart grid can have different ways to treat different devices. For example, to load an electric vehicle demand response management can be used, while for switching a lamp on and off, no demand response management is used. As a result, the framework may need to be applied multiple times to one architecture.

VI. CHOICES HAVE CONSEQUENCES

Choosing a position on the axis in the framework has consequences. In this section, we discuss consequences that relate to societal aspects of Smart Grids. We do not claim that this list is exhaustive.

A. *Consequences for balance of societal power*

The choices, which are made on one of the axes of our framework, have consequences. One of those consequences is the impact on the balance of what we call ‘societal’ power. With that we mean the power to determine the behaviour of other people and/or organizations. In the current energy grid, that kind of power is distributed. Each consumer has the right to turn his own devices on and of, and the energy suppliers take care of the energy balance on the net. Demand response management will affect this balance. As soon as control of the end used devices is shared with the network operators, the network operates will have more societal power in the energy grid, and the end user will have less. We do not put any direct qualification to a shift in the balance of societal power. We do want to state that a distributed balance of societal power is a natural barrier to misuse of power. In Table I, an overview is made of the impact the three axes of

 ISBN: 978-1-61208-189-2

the frame work have on the balance of control. In this Table, it is shown how changing the position on one of the axes will have consequences.

TABLE I. CONSEQUENCES FOR THE BALANCE OF SOCIETAL POWER

Degree of freedom	Impact on balance of control
Predictive window	The size of the predictive window has a small impact on the balance of control. It determines how long in advance the controlling party has to make its decisions.
Level of indirection	The level of indirection has more impact on the balance of power. The more energy consuming and producing devices are directly controlled by one party, the more direct control this party has on the entire smart grid, and thus the people connected to it.
Level of participation	With a low level of participation, the controlling party has a high level of control, while a high level of participation will result in a lower level of control.

B. Consequences for network stability and optimization

In general increasing the degrees of freedom may seem to result in less network stability, or at least more difficulty in obtaining it. When more demands and wishes have to be taken into account, more sophisticated decisions have to be made. Whether or not this will be the case depends on the interaction of the actual control mechanism and the power consuming and producing behaviour of the different stakeholders. In this paper, we cannot provide the reader with a general 'rule-of-thumb' in this area. We can state that if network operators have no control at all and energy suppliers and consumers do share control with respect to balance in supply & demand, a situation could occur where there is balance in supply & demand from an energy point of view, but which cannot be implemented physically, due to network constraints. Also, by sharing control with a network operator, it could carry out more network usage optimisation and thus minimize the transmission costs. In any case, a decision on whether or not to share control with a network operator influences the possible usage of Demand Response to optimize network usage.

TABLE II. CONSEQUENCES FOR THE NETWORK STABILITY AND OPTIMIZATION

Degree of freedom	Impact on stability and optimization.
Predictive window	In a larger predictive window there are more possibilities for an optimization algorithm to find a optimal solution. For example, with a distance in time of 1 hour, it will not be possible to find a solution which involves a decision about another devices which has to be turned on in 3 hours. So, depending on the used algorithm for control, a large predictive window could result in a more stable and more optimized network.
Level of indirection	The impact of the level of indirection on the grid depends highly on the optimization and stabilization algorithms which are used. Some algorithms need a higher level of direction for a more stable and optimal network.
Level of participation	The higher the level op participation, the less influence of the optimization algorithms is used. Therefore, a higher level of participation will probably result in a less optimized and less stable network.

C. Consequences for privacy

The privacy discussion often focuses on the information which is gathered about the people. However, not only the information which is gathered about them influences the privacy. Also the amount of self-control influences privacy. In an extreme example: when all electricity is cut off after 22:00, most people will be forced to go to bed early. As a result, people have less control about their own live, and are forced to apply to the rules given by the smart grid. The different axes in the framework have different impacts on privacy. In Table III those impacts are explained.

TABLE III. CONSEQUENCES FOR PRIVACY

Degree of freedom	Impact on privacy
Predictive window	A large predictive window forces a consumer to make decisions about his energy consumption early. For example, when the decision to use no energy after 22:00 is made at 18:00, a consumer cannot change his mind at 21:00.
Level of indirection	The higher the level of indirection, the more choice the consumer has, so the lower the impact on privacy. For example, when the only control is that the consumer may not consume more than 3000 Watts, the consumer can decide for himself how he uses the 3000 Watts. However, when the grid decides that the consumer cannot watch TV, because he will be using the washing machine, there will be a huge impact on privacy.
Level of participation	When the control decisions are made by an external party, the owner of the device connected to the gird loses a lot of privacy. However, when the consumer is participating in the control decisions, the impact on privacy will be lower. The more participation there is in the decision making process, the more privacy *with respect to self-control* is left for the consumer.

D. Consequences for green 'intermittent' energy

Last but not least, we come to the consequences of sharing control for a driver behind 'smart grids': creating space for the integration of energy from renewable sources with an intermittent profile and limited prediction (wind, solar, etc.). When there is no sharing of control in the balance of supply & demand, attaining balance must be achieved by creating and or introducing other suppliers (e.g. more gas turbine based electricity plants) that can

 ISBN: 978-1-61208-189-2

compensate for the variation on the 'intermittent' renewable supply side. Demand Response management (i.e. sharing control) reduces the need for extra flexible suppliers.

TABLE IV. CONSEQUENCES FOR 'GREEN' ENERGY

Degree of freedom	Impact on 'green' energy
Predictive window	A larger predictive window is difficult to obtain for power sources like wind and solar power plants.
Level of indirection	More or less indirection with respect to demand response management does not directly impact 'green energy'. However, when consumers are allowed to tell which energy resources have to be used ('green power'), direct control over the feeding of power to the grid does impact the use of green energy.
Level of participation	Sharing control for supply & demand management can be used to stimulate energy usage when renewable sources are 'at peak level'. However, a higher level of participation (be it consumers or produces) does not directly impact 'green energy'.

VII. FRAMEWORK APPLICATION

As a demonstration of how one could apply this framework, we now apply it in comparing the classical grid and *a* possible future smart grid with respect to sharing control of balancing supply & demand. For this grid of tomorrow we use the 'PowerMatcher' distributed control mechanism [9]. In a Powermatcher world there are so called consumer and producer nodes. They are represented by a 'software agents'. The agents exchange 'bids' on electricity. They express to what degree an agent is willing to pay (consumer) or receive (producer) for which amount of electricity. This is done through a mechanism based on micro-economic markets in which bids are aggregated and the market clearing price is determined as the equilibrium where supply meets demand. The 'market clearing price' is returned as a response to a bid. Agents then have to follow the allocated energy profile. Note that in the classical grid (in the Netherlands) a kind of auctioneering mechanism also takes place (e.g. the APX) at the level of Program Responsible parties (the electricity providers and network operators) The huge difference of the PowerMatcher lies within the fact that in a 'PowerMatcher world' devices attached to the grid partake into the bidding through their agents. We can now compare these two situations using the framework and focus on the consequences.

Predictive window. In the classical grid, consumers of power do not tell producers of their need for energy directly.. Only very large consumers have specific contracts with energy suppliers and network operators. For the mass of domestic and SME users statistics are used, where previous behavior is used to determine future behavior. In a PowerMatcher grid consumers do communicate their energy need – be it indirectly – by bidding prices for amounts of energy. The distance in time are market rounds. The distance in time is related to the time it takes for a market round to complete and if it is allowed to bid for market rounds in the future.

Level of indirection. In the classical grid, the consumer has a certain level of indirect control of balancing supply & demand. The entire group of consumers of controls the entire group of producers by demanding energy through the grid, which the producers have to supply, as long as the consumers pay. Since the price of energy for consumers hardly varies (in time), the amount of control of balancing supply & demand the producer has is very little. In a 'PowerMatcher world' there is more direct control from the producers: by demanding a higher price in the next market round, they can 'push' the agents of consuming devices into not consuming power. Or by lowering the price, they can 'push' agents towards consuming, depending on the actual need for energy of course.

Level of participation. In the classical grid there is little participation in control with respect to balancing supply & demand by both consumers and producers. The group of producers responds to a total predicted demand arriving at centralized markets (national, international). The group of consumers responds to prices on the market by energy providers. In a 'PowerMatcher world' the level of participation is increased significantly. Each consuming device influences the market (indirectly) by bidding and each producer influences the market by demanding a certain price for their electricity production. However, the amount of participation in a 'PowerMatcher world' is closely linked with the type of mechanism is used by the specific price determining agents used for determining the 'market clearing price'. For example, when real money is used in an auctioneering type of mechanism, the consumer willing to spend the highest amount of money available has more influence than parties who have less to spend. Also, producers demanding the least amount of money for their electricity will probably have more influence.

VIII. CONCLUSION

In this paper, we presented a framework for comparing (distributed) control mechanisms for balance of supply & demand, along three axes. The framework is focused on sharing control of 'end-user devices' in a smart grid. We saw that changes – along these axes - in the design of a control architecture for a smart will have consequences. We discussed these consequences with respect to issues of network stability, balance of 'societal' power, privacy and green energy. These issues are intertwined. For example, in an architecture, with a low level of indirection where energy suppliers can directly influence the usage of energy at consumers, a possibly unwanted consequence is the impact on privacy with respect to self-control. Increasing the level of indirection by having the energy suppliers influence the total amount of power instead of devices, the impact on privacy is less, but due to the shared influence of both the supplier and the consumer, there is more impact on network

 ISBN: 978-1-61208-189-2

stability and optimization. Due to the possible large impact of design choices, we also argue that these design decisions should not be taken lightly. Future work

Although we already use the framework for comparison of control mechanisms for the management of balance of supply & demand in smart grids, we do think there is room for improvement. As the reader might have noticed the framework is not as fine-grained as it might need to be in order to make efficient and useful comparisons. Future work will have to determine if this is the case. To that, we want to integrate framework on 'sharing control' with our earlier framework on 'sharing information'. We will be doing so in 'Reference Model for Supply & Demand Management on Smart Grids' (working title), which TNO is currently working on.

ACKNOWLEDGMENT

We would like to thank the European research project 'Web2Energy' in the FP7 call: "*ENERGY.2009.7.3.5. Novel ICT solutions for smart electricity distribution networks* for their funding to this research.

REFERENCES

[1] G. Broenink, and K. Helmholt, "Degrees of Freedom in Information Sharing on a Greener and Smarter Grid" Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY), 2011 The First International Conference on , pp. 141-147, 22 May 2011

[2] T.M. Overman, and R.W. Sackman, "High Assurance Smart Grid: Smart Grid Control Systems Communications Architecture", Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, pp.19-24, 4-6 Oct. 2010

[3] T.M. Overman, R.W. Sackman, T.L. Davis and B.S. Cohen, "High-Assurance Smart Grid: A Three-Part Model for Smart Grid Control Systems," Proceedings of the IEEE , vol.99, no.6, pp.1046-1062, June 2011

[4] M.U. Tariq, S. Grijalva and M. Wolf, "Towards a Distributed, Service-Oriented Control Infrastructure for Smart Grid," Cyber-Physical Systems (ICCPS), 2011 IEEE/ACM International Conference on, pp.35-44, 12-14 April 2011

[5] Venayagamoorthy and Ganesh Kumar, "Innovative smart grid control technologies", Power and Energy Society General Meeting, 2011 IEEE, pp.1-5, 24-29 July 2011,

[6] R. Belkacemi, A. Feliachi, M.A. Choudhry and J.E. Saymansky, "Multi-Agent systems hardware development and deployment for smart grid control applications," Power and Energy Society General Meeting, 2011 IEEE, pp.1-8, 24-29 July 2011

[7] A. Molderink, V. Bakker, M.G.C. Bosman, J.L. Hurink and G.J.M. Smit, "Management and Control of Domestic Smart Grid Technology," Smart Grid, IEEE Transactions on , vol.1, no.2, pp.109-119, Sept. 2010

[8] M. Erol-Kantarci, J.H. Sarker and H.T. Mouftah, "Analysis of Plug-in Hybrid Electrical Vehicle admission control in the smart grid," Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2011 IEEE 16th International Workshop on, pp.56-60, 10-11 June 2011

[9] M.P.F. Hommelberg, C.J. Warmer, I.G. Kamphuis, J.K. Kok and G.J. Schaeffer, "Distributed Control Concepts using Multi-Agent technology and Automatic Markets: An indispensable feature of smart power grids," Power Engineering Society General Meeting, 2007. IEEE, pp.1-7, 24-28 June 2007

[10] D.J.C. MacKay, "Sustainable Energy – without the hot air." UIT Cambridge, 2008. ISBN 978-0-9544529-3-3.

[11] G. Strbac, A. Shakoor, M. Black, D. Pudjianto and T. Bopp, "Impact of wind generation on the operation and development of the UK electricity systems", Electric Power Systems Research, vol. 77, issue 9, July 2007 pp. 1214-1227

[12] C. Framer, P. Hines, J. Dowds and S. Blumsack, "Modeling the impact of increasing PHEV loads on the distribution infrastruction" System Sciences (HICSS), 43rd Hawaii International conference on, 2010, pp 1-10

[13] M.H. Albadi and E.F. El-Saadany, "Demand Response in Electricity Markets: An Overview", Power Engineering Society General Meeting, 2007, Tampa, FL, pp 1-7.

 ISBN: 978-1-61208-189-2

MIMO-OFDM based Broadband Power Line Communication using Antenna and Fading Diversity

Jeonghwa Yoo
Wireless Communication Laboratory
The Catholic University of Korea
Bucheon, Republic of Korea
mundade@hanmail.net

Sangho Choe
Wireless Communication Laboratory
The Catholic University of Korea
Bucheon, Republic of Korea
schoe@catholic.ac.kr

Abstract—We present multiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) based broadband power line communication (BPLC) using antenna and fading diversity. We evaluate the proposed MIMO-OFDM system over multi-conductor power line channels, with or without cross-talk between the antenna paths. The proposed scheme employs maximum ratio combining (MRC) that effectively combines both the multiple antenna diversity gain and the multipath fading diversity gain over 3-phase (2×2 MIMO, outdoor) power line channels. Simulation results prove that the proposed scheme improves the bit error rate performance over the existing schemes, irrespective of whether cross-talk exists.

Keywords- MIMO; OFDM; broadband power line communication (BPLC); maximum ratio combining (MRC)

I. INTRODUCTION

Smart grid (SG) is a future power grid network based on regenerative green energy; it requires high-rate data transmission technology for bidirectional information exchange among electric power providers, electricity industries, and consumers. Broadband power line communication (BPLC) that allows a reliable high-rate transmission over power cables is available at a low cost because it does not require any additional infrastructure; further, it is ubiquitous because it is available anywhere where there is electricity and is easy-to-access with a plug-in power cable. Hence, a BPLC network is a promising medium for SG and its associated services such as advance metering infrastructure. Moreover, since an international BPLC standard, IEEE 1901 [1], was adopted in 2010, there has been a growing interest in various other BPLC applications including home networks, high-speed Internet, and emergency backup networks.

A BPLC signal transmitted via an electric power cable experiences severe channel distortions such as multipath fading and impulse noise. In this study, we use the Middleton class A model [2] for handling the impulse noise and the Zimmermann frequency model [3] for handling the power line multipath fading.

In the case of a 3-phase 4-wire power cable, we implement a 2×2 multiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) BPLC with maximum ratio combining (MRC). A recent literature search revealed several MIMO-OFDM BPLC schemes [4-6] that use antenna MRC (AMRC) for obtaining the spatial diversity gain. In particular, the authors in [4] first introduced a space frequency (SF)-coded MIMO-OFDM over power line channels. The authors in [5] implemented a space time (ST)-coded MIMO-OFDM for outdoor multi-conductor power cables and performed an experiment showing the capacity loss caused by antenna coupling. The authors in [6] compared the simulation results of SF-, ST-, and space time frequency (STF)-coded MIMO-OFDM over indoor power line channels and demonstrated the superiority of STF coding.

Contributions. The contribution of this paper is a novel SF-coded MIMO-OFDM scheme employing antenna and fading MRC (AFMRC) that effectively combines both multiple antenna and multipath fading diversity. Through a computer simulation, in this study, we prove that a coupling effect is not negligible with respect to the performance of a MIMO system. Simulation results also verify that the proposed scheme is superior to conventional schemes, irrespective whether cross-talk between antenna channels exists. The proposed scheme improves the bit error rate (BER) performance not only in the 2×2 MIMO system via a 3-phase 4-wire power line but also in the single-input single-output (SISO) system via a single-phase 2-wire power line; note that the SISO system uses fading MRC (FMRC) instead of AMRC. We also evaluate the system design parameters by comparing the BER when the impulse noise index A is varied.

Organization of the paper. The rest of this paper is organized as follows: Section II explains the power line channel characteristics including impulse noise and multipath fading and presents the proposed MIMO-OFDM BPLC system model over multi-conductor power line channels. Section III presents the simulation results of the proposed scheme, which are compared with those of the existing schemes. Finally, concluding remarks are given in Section IV.

 ISBN: 978-1-61208-189-2

II. SYSTEM MODEL

A. Impulse Noise and Fading Channel in BPLC

A BPLC channel can be characterized by both impulse noise and multipath fading owing to the multiple signal reflections caused by a power line impedance mismatch. First, for handling the impulse noise, we use the Middleton class A model [2], whose *pdf* (probability density function) is defined as

$$p_X(x) = \sum_{m=0}^{\infty} e^{-A} \frac{A^m}{m!} \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{x^2}{2\sigma_m^2}} \quad (1)$$

$$\sigma_m^2 = \sigma^2 \frac{m/A + \tau}{1+\tau}$$

where $\sigma^2 = \sigma_G^2 + \sigma_I^2$ (σ_G^2 is the Gaussian noise variance and σ_I^2 is the pure impulse noise variance), $\tau = \sigma_G^2 / \sigma_I^2$, and A is the impulse index.

Second, for handling the multipath fading, we employ the Zimmermann frequency PLC channel model [3], whose transfer function at the *j*th antenna path is expressed as

$$H_j(f) = \sum_{l=1}^{L} H_{j,l}(f)$$
$$H_{j,l}(f) = g_{j,l} \cdot e^{-(\alpha_0 + \alpha_1 \cdot f^u) d_{j,l}} \cdot e^{-j2\pi f (d_{j,l}/v_p)} \quad (2)$$

where L is the number of fading paths. α_0, α_1, and u are the power line cable parameters, and $|g_{j,l}| \le 1$ is the weighting factor of the *j*th antenna and the *l*th fading path [3]. $d_{j,l}/v_p$ is equivalent to the corresponding path delay $\tau_{j,l}$ (where $d_{j,l}$ represents the path length) as follows:

$$\tau_{j,l} = \frac{d_{j,l} \cdot \sqrt{\varepsilon_r}}{c_0} = \frac{d_{j,l}}{v_p}$$

where ε_r is the non-insulation dielectric constant of the cable and c_0 is the speed of light. Typically, each OFDM subcarrier has flat (constant) frequency channel characteristics because of its narrow bandwidth; hence, the frequency selective fading transfer function of (2) can be translated (digitized) and approximated as follows:

$$H_j(f)|_{f=f_c+k\Delta f} \cong H_j(k) = \sum_{l=1}^{L} H_{j,l}(k) \quad (3)$$

where f_c is the carrier frequency (which is herein assumed to indicate the lower limit of the OFDM bandwidth (BW)), Δf is the subcarrier spacing, and the frequency index $k = 0, 1, \ldots, N - 1$.

B. MIMO-OFDM BPLC System

In this study, we design a MIMO-OFDM system that contains I (transmit antennas) × J (receive antennas). In the OFDM transmitter, the kth subcarrier modulation signal, $S(k)$, experiences the following inverse fast Fourier transform (IFFT)

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) e^{j2\pi nk/N} \quad (4)$$

where $s(n)$ is the nth (= 0, 1, …, N-1) time sample and N is the number of subcarriers.

In a MIMO BPLC system, a pair of electrical wires is converted into a single antenna channel; hence, the number of transmitting and receiving antennas is typically limited to two for a 3-phase 4-wire power line and one for a single-phase 2-wire power line. Therefore, MIMO-OFDM is used either indoors or outdoors with a 3-phase 4-wire power line, whereas SISO-OFDM is mostly used indoors with a single-phase 2-wire power line. Fig. 1(a) shows the block diagram of a 2×2 MIMO-OFDM BPLC system that uses a 3-phase 4-wire power line (Fig. 1(b) shows its cross-cut interior structure). This 2×2 MIMO system has two antenna paths that consist of a single antenna path formed by C1 and C2 and another single antenna path formed by C3 and C4; C0 assumes the role of a ground connection [4]. An SF encoder is used for reducing the error probability caused by the interference in the MIMO channel. The following two SF encoder vectors $\mathbf{S_1}$ and $\mathbf{S_2}$ are formed by arranging the same subcarrier signal samples in an appropriate order (i.e., vector $\mathbf{S}_2$ is the circular-shifted version of $\mathbf{S}_1$ [4]) for this SF encoder.

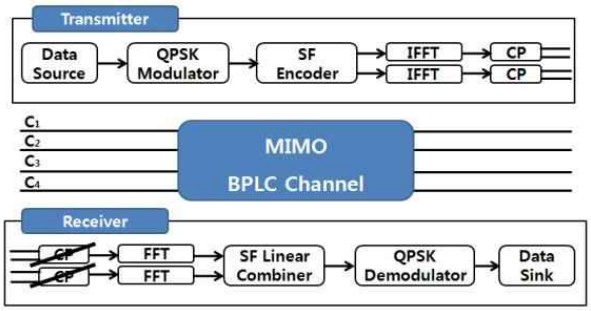
$$\mathbf{S}_1 = [S_1(0), \ldots, S_1(\tfrac{N}{2}-1), S_1(\tfrac{N}{2}), \ldots, S_1(N-1)]^T$$
$$\mathbf{S}_2 = [S_2(\tfrac{N}{2}), \ldots, S_2(N-1), S_2(0), \ldots, S_2(\tfrac{N}{2}-1)]^T \quad (5)$$

where $S_1[k] = S_2[k]$, (k = 0, 1, …, N-1) and $(\cdot)^T$ refers to the transpose of $(\cdot)$. $\mathbf{S_1}$ and $\mathbf{S_2}$ are respectively converted to the corresponding time sample vectors, $\mathbf{s_1}=IFFT\{\mathbf{S_1}\}$ and $\mathbf{s_2}=IFFT\{\mathbf{S_2}\}$, through the IFFT process (see (4)) and then transmitted to the receiver via each antenna path. In the system shown in Fig. 1(a), this transmission process occurs at the signal encoder and modulator, and the corresponding receiving process is processed at the linear combiner and detector. A cyclic prefix (CP) is added to the OFDM modulated sample vectors ($\mathbf{s_1}$, $\mathbf{s_2}$) before their transmission to prevent inter-symbol interference (ISI) caused by the multipath delay. The signal received via the fading channel undergoes the SF decoding process, i.e., fast Fourier transform (FFT), an inverse circular-shift operation, and then the MRC process, to recover its data stream after the removal of the added CP.

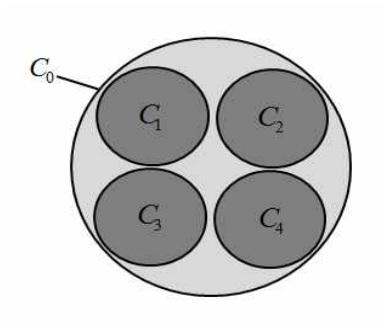
 ISBN: 978-1-61208-189-2

1) MRC without Cross-Talk

For simplifying the simulation, we assume that there is no coupling between the two antenna paths (this assumption is practically reasonable for the carrier frequency f_c < 25 MHz [6].).



(a)



(b)

Figure 1. (a) 2×2 MIMO-OFDM PLC system block diagram; (b) 3-phase 4-wire power line interior structure.

The diversity gain of a conventional system that uses AMRC is obtained by multiplying its optimum weight to the different spatial antenna paths. The system proposed in this paper employs AFMRC, a combined technique of AMRC and FMRC. The proposed system has one receiver per fading path to achieve the FMRC gain. Assuming the same transmit signal via the jth antenna L fading paths, i.e., $S_j(k) = S_{j,1}(k) = S_{j,2}(k) = \ldots = S_{j,L}(k)$, we can write the jth (j = 1,2) antenna received signal $Y_j(k)$ at the kth subcarrier as

$$Y_j(k) = \sum_{l=1}^{L} Y_{j,l}(k) = \sum_{l=1}^{L} \sqrt{\frac{E_s}{2}} H_{j,l}(k) S_j(k) + X_{j,l}(k) \qquad (6)$$

where E_s represents the average energy of the transmit signal. $X_{j,l}(k)$ represents the jth antenna path and the kth subcarrier noise component, that is the result of the FFT operation of the time axis impulse plus Gaussian noise signal $x_{j,l}(n)$ with the variance σ^2 (see (1)). In this study, we assume ideal fading channel estimation to simplify the simulation [7].

In the application of the proposed MRC (AFMRC) to the single-phase 2-wire SISO BPLC (just using FMRC) and the 3-phase 4-wire 2×2 MIMO BPLC, we express the output $\hat{S}$ (detected signal) of the maximum likelihood (ML) receiver as

$$\hat{S}(k) = \begin{cases} \arg\min\limits_{S \in \mathcal{S}} \left| \sum\limits_{l=1}^{L} Y_l(k) H_l^*(k) - S \right|^2 & \text{for SISO} \\ \arg\min\limits_{S \in \mathcal{S}} \left| \sum\limits_{j=1}^{J} \sum\limits_{l=1}^{L} Y_{j,l}(k) H_{j,l}^*(k) - S \right|^2 & \text{for MIMO} \end{cases} \qquad (7)$$

where $(\cdot)^*$ refers to the conjugate of $(\cdot)$ and $\mathcal{S}$ indicates the signal constellation set. AFMRC improves the system performance as compared to the conventional MRC (AMRC), where $\hat{S}(k) = \arg\min\limits_{S \in \mathcal{S}} \left| \sum\limits_{j=1}^{J} Y_j(k) H_j^*(k) - S \right|^2$. However, as shown in (7), the receiver complexity of the AFMRC-based SISO/MIMO system increases L-fold due to the addition of FMRC. Even in the case of the indoor single-phase 2-wire SISO channel, the proposed scheme has the FMRC diversity gain.

2) MRC with Cross-Talk

The proposed MIMO system, shown in Fig. 1(a), might have cross-talk (its capacity loss might not be negligible (but less than 16%), especially in the case of $f_c \geq 25$ MHz [6].) between two parallel antenna channels; this cross-talk degrades the system performance. The 2×2 MIMO channel matrix $\mathbf{H}$ with non-zero cross-talk terms, indicating the ith transmit and jth receive antenna path gain $H_j^i(k) \neq 0$ (where $i \neq j$), can be expressed as follows:

$$\mathbf{H} = \begin{bmatrix} H_1^1(k) & H_2^1(k) \\ H_1^2(k) & H_2^2(k) \end{bmatrix} \qquad (8)$$

Let the channel capacity with and without the cross-talk be denoted as C_{ct} and C_{nct}, respectively. The capacity-loss ratio (CR) estimated by using the cross-talk can be defined as [5]

$$CR = \frac{C_{nct} - C_{ct}}{C_{nct}} \times 100\% \qquad (9)$$

where $C = \mathrm{BW} \log_2 \det(\mathbf{I}_I + \frac{\mathrm{SNR}}{I} \mathbf{H}\mathbf{H}^H)$. I is the number of transmit antennas; SNR, the signal-to-noise ratio; and $\mathbf{I}_I$, an identity matrix of size I. The output $\hat{S}$ of the proposed MIMO MRC receiver can be expressed as

$$\hat{S}(k) = \arg\min_{S \in \mathcal{S}} \left| \sum_{j=1}^{J} \sum_{l=1}^{L} Y_{j,l}(k) H_{j,l}^{j*}(k) - S \right|^2, \qquad (10)$$

 ISBN: 978-1-61208-189-2

where $Y_{j,l}(k) = \sum_{i=1}^{I} \frac{\sqrt{E_s}}{2} H^i_{j,l}(k) S_i(k) + X_{j,l}(k)$.

III. SIMULATION RESULTS

We simulate the proposed system model with the QPSK constellation under the power line channel conditions, and compare its uncoded BER results to that of a conventional system [6]. We assume a multipath fading PLC channel with L = 6, whose simulation parameters are the same as those given in Table I of [6]. For the sake of simplicity, we also assume the same fading channel parameters for the two antenna paths[1)]. We set N = 1024, CP size = 120 (unit: sample), f_c = 25 MHz, Δf (frequency spacing) = 1 KHz, and BW = 1.024 MHz for the simulation; hence, the maximum data rate is approximately 1.83 Mbps.

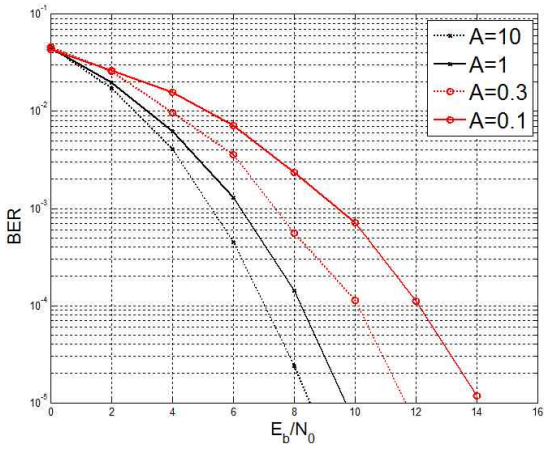


Figure 2. BER Comparison of 2×2 MIMO PLC for different values of A.

Fig. 2 presents a comparison of the BER of the 2×2 MIMO BPLC system when the impulse noise index A is varied. For this experiment, we set τ = 0.1. While for a large value of A (≥ 1), the noise channel characteristics approach those of Gaussian noise, for a small value of A (<1), they are similar to those of impulse noise. Hence, the BER decreases when the value of A increases, as shown in Fig. 2. For example, in the case of A = 10, we can obtain approximately 1-dB gain at BER = 10^{-5} as compared to the case of A = 1, 3-dB gain as compared to the case of A = 0.3, and 5.5-dB gain as compared to the case of A = 0.1. Practically, in the BPLC channel environment, A has a value within the range of 0.0001 to 0.35; hence, we choose A = 0.3 for the next experiment [6].

Fig. 3 shows a comparison of the BER performance between the conventional AMRC and the proposed AFMRC in the SISO/MIMO-OFDM systems. First, in the case of the MIMO system, AFMRC results in a performance gain of approximately 1 dB at BER = 10^{-5} as compared to AMRC. Fig. 3 also shows a comparison of the conventional method (with no MRC) and the proposed method (with FMRC) in the case of the SISO-OFDM system; it shows a 1.5-dB improvement in the case of BER = 10^{-5} for the proposed scheme. Therefore, the simulation verifies that the proposed SISO/MIMO-OFDM system is more effective, in the case of both an indoor single-phase power line and an outdoor 3-phase power line, than the conventional SISO/MIMO-OFDM system.

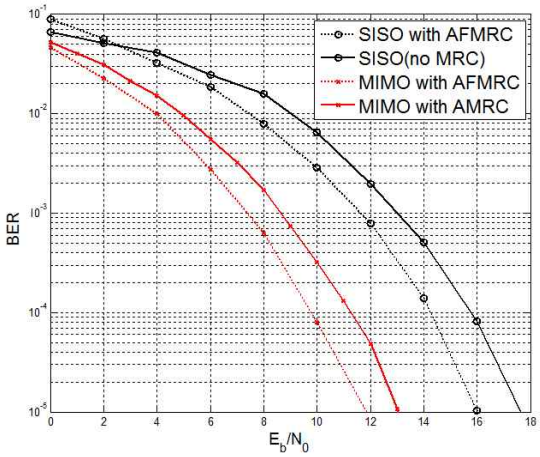


Figure 3. Performance comparison of SISO/MIMO-OFDM with different MRC schemes (assuming A = 0.3).

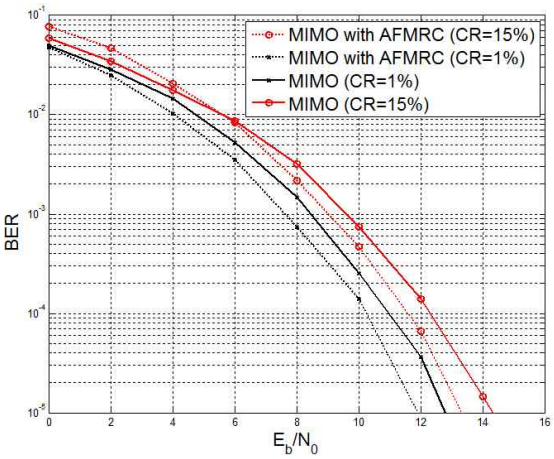


Figure 4. Performance of MIMO-OFDM with cross-talk (assuming A = 0.3).

Fig. 4 shows the performance of an SF-based MIMO BPLC system with cross-talk. The proposed scheme has a gain of approximately 0.7–0.8 dB over the conventional scheme at the BER = 10^{-5}. When the CR value increases from

[1)] In the case of PLC channels using 3-phase 4-wire power line, the cha nging of channel parameters between the antenna paths is almost neglig ible in practice [8].

 ISBN: 978-1-61208-189-2

1% to 15%, the BER performance of both schemes degrades by approximately 1 dB. As a result, it is observed that the effect of crosstalk on the performance of an MIMO-OFDM system is not negligible.

IV. CONCLUSIONS

We proposed an MIMO-OFDM-based BPLC system with antenna and fading MRC (AFMRC) that effectively combines multiple antenna and multipath fading diversity. We evaluated the proposed MIMO-OFDM system over multi-conductor power line channels with or without cross-talk between the antenna paths. The computer simulation verified that the proposed scheme was more efficient in terms of system performance in the case of both the indoor single-phase (SISO) and the outdoor 3-phase (MIMO) BPLC applications than the conventional schemes.

ACKNOWLEDGMENT

This work (Grant No. 00047470-1) was supported by Business for Cooperation between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration 2011.

REFERENCES

[1] IEEE 1901, "IEEE standard for broadband over power line networks: Medium access control and physical layer specifications," 2010.

[2] N. Andreadou and F.-N. Pavlidou, "PLC channel: impulse noise modeling and its performance evaluation under difference array coding schemes," IEEE Trans. on Power Delivery, vol. 24, no. 2, pp. 585-595, April 2009.

[3] M. Zimmerman and K. Dostert, "A multipath model for the powerline channel," IEEE Trans. Commun., vol. 50, no. 4, pp. 553-559, April 2002.

[4] C. L. Giovaneli, B. Honary, and P. G. Farrell, "Space-frequency coded OFDM system for multi-wire power line communications," Proc. IEEE ISPLC 2005, pp. 191-195, April 2005.

[5] L. Hao, and J. Guo, "A MIMO-OFDM scheme over coupled multi-conductor power-line communication channel," Proc. IEEE ISPLC 2007, pp. 198-203, March 2007.

[6] B. Adebisi, S. Ali, and B. Honary, "Space-frequency and space-time-frequency M3FSK for indoor multiwire communications," IEEE Trans. on Power Delivery, vol. 24, no. 4, pp. 2361-2367, October 2009.

[7] Y. H. Kim, "Multipath parameter estimation for PLC channels using the GEESE algorithm," IEEE Trans. on Power Delivery, vol. 25, no. 4, pp. 2339-2345, October 2010.

[8] R. Hashmat, P. Pagani, and T. Chonavel, "MIMO communications for inhome PLC networks: Measurements and results up to 100 MHz," Proc. IEEE ISPLC 2010, pp. 120-124, March 2010.

 ISBN: 978-1-61208-189-2

GREASE Framework
Generic Reconfigurable Evaluation and Aggregation of Sensor Data

Matthias Vodel, Rene Bergelt, and Wolfram Hardt
Dept. of Computer Science
Chemnitz University of Technology
Chemnitz, GERMANY
Email: { vodel | berre | hardt }@cs.tu-chemnitz.de

***Abstract*—The proposed research work represents a generic, energy-efficient concept for the synchronised logging, processing and visualisation of any kind of sensor data. The concept enables a chronological coordination and correlation of information from different, distributed sensor networks as well as from any other self-sufficient measurement systems. Based on the achieved relation between the several sensor sources, the information quality can be increased significantly. Therefore, the system-wide description of the monitoring scenario and each data set is realised in XML. Accordingly, the aggregated, heterogeneous sensor information are convertible into multiple output formats. Dependent on the application specific requirements for the visualisation, we are able to consider additional meta-information from the test environment to optimise the data representation. The definition of advanced data fusion techniques and pre-processing mechanisms allows a selective data filtering to shrink the network load. To evaluate of basic usability requirements and the efficiency of the proposed concept, an automotive sensor network represents a capable test system for the proposed framework. Within the demonstrator, the available on-board measurement systems were extended by high-precision sensor nodes, which establish wireless sensor network topology. Afterwards, the correlated measurement information were converted and visualised for several professional data analysis tools, e.g., jBEAM, Google Earth, and FlexPro.**

Keywords-Data Aggregation; Data Fusion; Data Synchronisation; Heterogeneous Wireless Sensor Network; Sensor Actuator Systems.

I. INTRODUCTION

Actual research projects in the field of wireless sensor networks operate on different, proprietary hardware platforms and contain multifaceted types of sensors. Currently, each measurement scenario consists of several application-specific and independent operating processes for the data-collection, -storage and -analysis.

It does not exist any uniform synchronisation techniques between the autonomous sensor systems. Accordingly, a detailed and target-oriented post-processing of the data sets within a shared knowledge base is not feasible. In consequence, we are not able to create unique relations between the different measurement information. Due to these missing relations, it is very hard to create a common primary index for the given, heterogeneous sensor platforms.

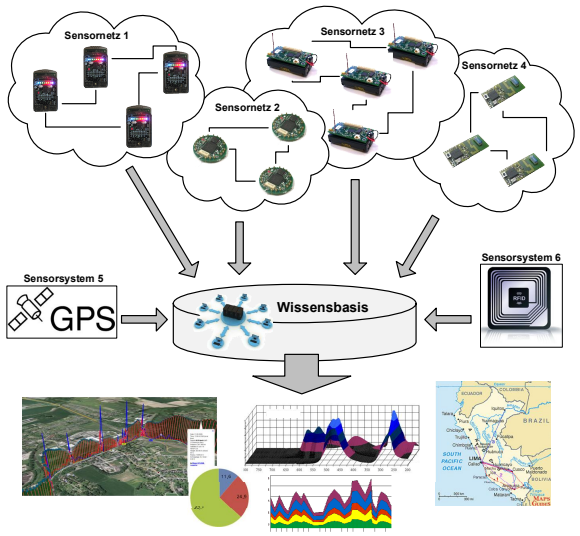


Figure 1. GREASE - A synchronised data-processing framework allows an efficient integration of different, autonomous sensor networks.

To solve this problem, we present *GREASE* - a Generic Recon-figurable Framework for the Evaluation and Aggregation of heterogeneous Sensor Data (see *Figure* 1). In order to introduce this integrated data processing concept, this paper is structured as follows: After this introduction, section *II* provides an overview about heterogeneous, distributed sensor environments, the data processing flow and respective challenges. The proposed GREASE framework is introduced in Section *III*, including conceptual fundamentals, basic requirements, system parameters, and the top level structure (Section *IV*). Accordingly, Section *V* provides implementation details of the GREASE software architecture as well as the overall application flow within the framework. Section *VI* specifies the application scenarios with all integrated components and the environmental conditions. The respective data analysis is described and discussed in Section *VII*. Finally, the paper concludes with a summary and an outlook for future work in this research project.

 ISBN: 978-1-61208-189-2

II. Related Work

During the last two decades, a couple of commercial tools for the measurement data recording and monitoring were developed. Unfortunately, most of them have functional or conceptual restrictions. Some of the vendors offer exclusive, hardware-specific analysis tools, which require special devices, predefined product series or vendors. Other sensor systems do not have any special software tools for extracting the measured data sets. There is no support for further post-processing steps.

In consequence, *LabView* from National Instruments [1], *jBEAM* from AMS [2] or FlexPro [3] offer multiple features to enhance the restricted vendor tools, which only provides a small set of general data recording an handling functions. These applications allow the interpretation of offline data from data bases or files as well as the live analysis of a given data source. Both jBeam and LabView operates platform-independent and all of these related tools support a lot of established data formats and communication interfaces. Especially jBeam, which integrates the *ASAM* standard *(Association for Standardisation of Automation and Measuring Systems)* [4], enables an easy and modular extension with user-defined components. Furthermore, FlexPro includes a lot of additional visual plugins and represents a complete visualisation framework for the given measurement data.

The very high system requirements of all the related software applications represent a critical disadvantage. Accordingly, these tools are not capable for resource-limited data recording environments. Thus, such frameworks have to be used in a second data processing step on dedicated workstations with sufficient hardware components. Hence, small and energy saving hardware system, which are used exclusively for collecting and storing multiple data from different sensor sources, are not able to use the data aggregation [5] and visualisation features of these software frameworks. Due to these circumstances, most of the ongoing sensor system projects use proprietary software solutions to organise and synchronise the collected measurement data [6]. In fact, there are many critical compatibility problems between such software tools. In consequence, modifications of the measurement scenario or the system configuration take a lot time and bind important resources. In conclusion, the user wants a universal software tool for collecting and analysing the entire pool of sensor information in an application-specific and resource-efficient way. Automated or semi-automated data visualisation techniques represent further essential requirements.

III. Concept

We are now looking for generic utilities and standards to route information from different sensor systems into a common data processing unit in a synchronised way, considering scenario-specific configuration schemes and sensor parameters. Hereby, synchronised time stamps for the heterogeneous sensor data sets are very important to allow correct correlations in the common knowledge base. Furthermore, such utilities allow us to define user-specific data analysis procedures during the measurement runtime and advanced data fusion techniques [7][8][9] to shrink the data volume directly within the sensor nodes.

Figure 2. GREASE - Integrated data processing flow for heterogeneous measurement topologies.

To provide such features, GREASE represents a software framework based on a capable and lightweight data management concept, which is able to bypass the already mentioned disadvantages. It combines advanced sensor network configuration features with resource-efficient operating parameters. GREASE integrates the entire data processing flow in an efficient way, including all stages like the data measurement, the data processing, data storage and finally data analysis tasks. *Figure* 2 illustrates this flow.

The concept focuses on resource-limited systems and has to be feasible for a wide variety of application scenarios. Thus, the primary objective is a dynamic and flexible processing environment, which is adaptable to modifications in the configuration or in the analysis requirements. Furthermore, the data processing core has to be separable into two spatial, chronological and platform-specific operating modes. All components for the data measurement are working within the first mode. All relevant modules for the data analysis as well as possible visualisation plugins operate independent within the second mode. Based on this requirement, we are able to map different data processing functions to predefined configuration scenarios. In contrast, other related software tools do not separate the data handling process into different phases in an efficient way.

GREASE deals with a standardised data transport and definable synchronisation parameters for collecting information from several distributed sensor components. Accordingly, changes in the data analysis process have no effects on the components of the data recording. This feature provides significant benefits, especially for complex sensor systems or inaccessible measurement environments.

In addition, all GUI *(Graphical User Interface)* actions, which are accessible by the user, also have to be executable and controllable in an automated or semi-automated way. This feature represents another important difference to other related software tools, which not provide any script-based operating mode without GUI. But especially for continuous maintenance-free sensor measurement scenarios, the scripting of user-defined activities is essential.

 ISBN: 978-1-61208-189-2

All central requirements for an synchronised data logging, processing and visualisation framework, specially in the field of heterogeneous sensor network systems can be summarised as follows:

- Synchronisation of different, autonomous sensor systems
- Modular extension with plugins and an easy modification / adaptation
- Using *XML ((Extensible Markup Language)* as common data exchange format instead of proprietary data types
- Offline and live data analysis from files, network file systems or databases
- Graphical User Interface for configuration and maintenance
- Automated or semi-automated data analysis and data representation mechanisms

Based on the proposed concept, the developed framework act as coordinator between application-specific components. The framework itself operates as a generic coordinating unit and includes no application-specific logic.

IV. Structure

As already mentioned, the structure of the proposed concept is divided into two operating modes. The first one encapsulates the data recording, synchronisation and correlation. A second mode processes the data analysis and generates a user-defined representation of the information sets. Due to the modular operating concept, the sensor net framework is completely independent from the given sensor configuration. Therefore, the environment uses an end-to-end communication design, called the *hourglass architecture*, which enables maximum interoperability between the several components. It means, that multiple sensor units and corresponding data processing units are connected by a dedicated *SensorController*. The sensor controller ensures a specific, universal mapping of the sensor information into a predefined format (see *Figure* 3). For the data output, the *SensorLogReader* component provides different modules for the information representation. The result is a very high diversity on both data input and data output components. In contrast, the binding middle part shows a strict uniformity.

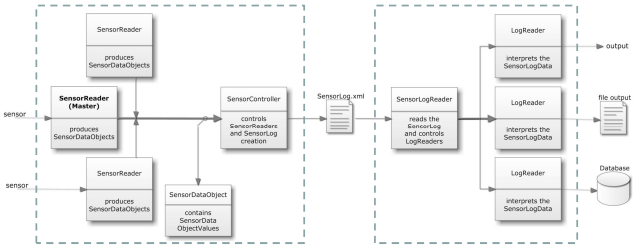


Figure 3. Sensornet framework structure. The separation into two operation modes for the data collecting (left) and data analysis (right) is described.

This structure fulfils the central requirement of a separated data processing core for gathering and analysing the sensor measurements. Hereby, the common XML representation of the entire scenario is essential and allows us to transfer the information for any kind of application. All internal and external parameters of the environment as well as special meta-information regarding the measurement scheme are correlated together with the data sets. Accordingly, researchers are able to reconstruct the whole test scenario with synchronised data, time stamps and a detailed system configuration. The reuse factor, for instance in the field of automotive testing scenarios, increases substantially.

Furthermore, we have the ability to modify existing sensor configuration in a time efficient way. The structure allows us to add or remove single components without changing the data flows within the overall monitoring system. In principle, it is also possible to create a direct interconnection between data collecting and visualisation. This increases the runtime significantly and reduces the used resources. But without the common XML representation, is not possible to capture the entire measurement scenario in a reusable way.

V. Implementation & Application Flow

To enable a platform-independent operation, the proposed concept was implemented in *Java*. Thereby, the resulting framework is also able to include external components, which are written in another programming language. A precondition is the common interface specification within the SensorController. To provide such a feature, the proposed framework implementation contains a *central core library*, which encapsulates the entire data processing logic. This library can also be used for the development of further modules. The communication between the SensorController and the SensorLogReaders with its set of modules uses predefined protocols. All of these protocols are designed as generic as possible to allow a universal usage. This flexibility also simplifies the integration of third party modules and ensures the compatibility during further developments [10]. Due to the fact, that the sensor configurations and all kinds of scenario parameters are also transmitted within the XML representation, possible enhancements for customer-specific applications can be done in an easy way.

Regarding to the application flow, the initialisation of the SensorController starts with loading a application-specific configuration file. This file contains all information for the actual project as well as the structure of all corresponding *SensorReader* components. Accordingly, the SensorController loads and activates all necessary sensor components and starts the data recording. Each SensorReader module operates simultaneous as a dedicated thread. When a SensorReader receives a data set, one *SensorDataObject* will be generated, which is predefined by the framework configuration scheme. Additional meta-information, for example physical measurement units or special indicators, are

 ISBN: 978-1-61208-189-2

also included. Afterwards, the SensorController receives the respective signal for the completed object. The object will be transmitted. During the following data processing, the controller analyses all correlations to information from other SensorReaders. Finally, the correlated and classified data set will be stored as an XML data structure.

If the *SensorLogReader* is reading and analysing an XML log file in a reverse process, each data set will be converted into a predefined, framework-conform object and accordingly provided to the data analysis components. The data sets can be reused for multiple representation or visualisation output formats. An essential advantage of this framework system is the fact that the data source is not restricted to local log files. Also data base systems or network file systems can act as data input for further analysis and post-processing steps.

In respect of the communication tasks in the sensor environment, we also have to discuss security features. Due to the fact, that GREASE focuses on research & development environments, we actually do not consider further security aspects for the distributed handling and storage of the sensor data. Within the different development stages of a given system, engineers design and implement complex test environments for getting valid and high-quality measurement results. Accordingly, the risks, which result from general communication threats are negligible. Anyway, we actually cooperate with related German car manufacturers in regards to this weak point. Several research projects focus on the development of advanced, energy-efficient security features for embedded, resource-limited sensor network topologies. In this context, the main challenge is the maintenance of a lightweight software architecture, which provides stable and flexible modules for diversified application scenarios. Here, advanced security mechanisms have a direct impact on the data throughput and the resource consumption. Accordingly, our goal is to find a good trade-off between runtime performance and security capabilities within GREASE.

VI. Application Scenarios

The proposed concept was developed to manage several sensor net scenarios at our computer engineering department. To clarify functional aspect of the implemented framework, we describe the data processing flow by a real-world automotive measurement system [11]. For this monitoring scenario, the existing sensor components of a given research vehicle were upgraded with high-definition sensor nodes. These nodes are placed at predefined positions to monitor the entire environment and provide independent measurement data about the current temperature, light intensity as well as the acceleration in two axes and the magnetic field strength. Thus, the established wireless sensor network provides meta-information about the measurement environment and external parameters. *Figure* 4 illustrates the measurement scenario.

The wireless sensor communication infrastructure is based on the *IEEE 802.15.4* and *ZigBee* [12][13]. Additionally, mobile sensor nodes are worn by the passengers. In order to realise localisation features for these nodes, they are equipped with *nanoPAN* ultra-low power network interfaces [14], which provide *RSSI*-based *(Received Signal Strength Indication)* distance information. Both communication technologies are using the 2.4 GHz frequency spectrum for the data transmission. A multi-interface, multi-standard data sink is able to handle both communication standards simultaneously. Robust communication stacks with adapted layer 2 and layer 3 protocols minimise interference-based influences on the communication behaviour.

In addition, we integrated a high-resolution *GPS (Global Positioning System)* sensor, which enables the correlation between absolute positioning information, speed, altitude and the available on-board vehicle data.

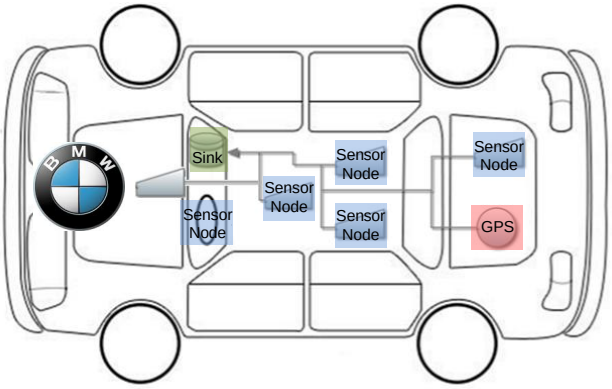


Figure 4. Measurement system - All data from the sensor nodes and the GPS module are transmitted to the data sink in the vehicle, represented by the proposed framework.

By providing a synchronised knowledge base of all sensor information, a detailed analysis of specific driving situations and the driver behaviour is possible. Thereby, the GPS data allows a verification of these situation based on available track information. Accordingly, we are able to calculated and predict driver profiles. The results are used to adjust and to optimise the characteristics of the entire vehicle, for instance, the engine management system or the suspension dynamics. Furthermore, an analysis of the wear measuring quantity provides interesting statements about the vehicle lifetime.

For this measurement environment, special SensorReader modules for the data sink communication interfaces were implemented. Incoming data from the sensor topologies are classified and converted into abstract data objects, which are transmitted to the controller. Another SensorReader module implements features for the GPS data input. Thereby, the *NMEA 0183 (National Marine Electronics Association 0183)* protocol for the positioning data is required. Accordingly, all kinds of GPS hardware, which supports the NMEA

 ISBN: 978-1-61208-189-2

protocol and the serial port as communication interface, are supported.

For the synchronisation of the several sensor data, one specific SensorReader has to be predefined during the initialisation of the measurement scenario.

In our exemplary case, the system provides a high-resolution GPS unit. Besides positioning information, this sensor also provides an accurate time signal. In consequence, the given time stamps from the GPS sensor represent the global *synchronisation master*. Thereby, the framework is not restricted for using a time stamp as master index. Especially for the integration of multiple, autonomous sensor systems and a missing central scheduling entity, a user-defined choice for the synchronisation master provides important benefits.

VII. Data Analysis

For post-processing the collected data, two data analysis components were implemented. The first one is an export module, which prepares the sensor data sets for the storage in a given database system and accordingly transmits the chosen information. A second module is responsible for converting the sensor data with dedicated visualisation plugins, e.g., for Google Earth. Hence, the data output of this module is a KML representation of all correlated sensor information. *Figure* 5 describes the data flow.

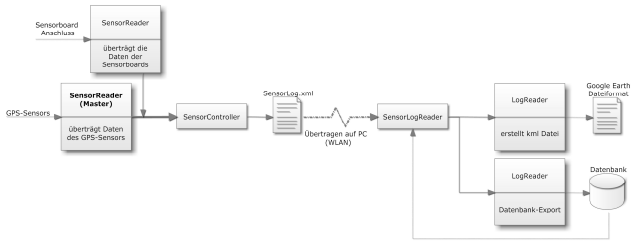


Figure 5. Application scenario of the proposed framework.

Within Google Earth, an additional 3D altitude track extends the visualised measurement curves (represented in *Figure* 6). The entire data processing flow integrates all proposed features for the data recording, handling and visual representation. All developed components are as generic as possible. This also includes a high compatibility level for both hardware and software environments [15][16].

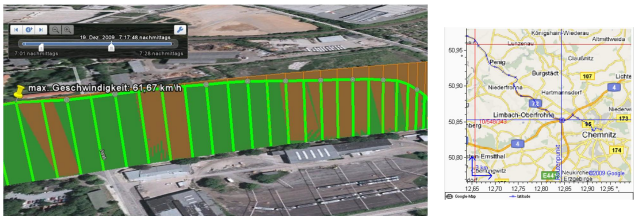


Figure 6. Data visualisation in Google Earth.

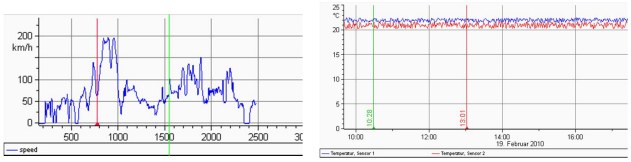


Figure 7. Data visualisation in FlexPro.

Other commercial software tools, e.g., FlexPro, or jBeam (see *Figure* 7, are able to import these information for the database for advanced, sectoral data post-processing tasks. Therefore, a dedicated *CSV (Comma Separated Values)* export module ensures a universal exchange interface.

VIII. Conclusion and Future Work

The proposed research work described the implementation of a comprehensive data processing environment for heterogeneous sensor systems. The basic concept provides generic structures for many further research projects in the field of novel data aggregation and data fusion techniques. For an easy data collecting and data analysis process, we are now able to synchronise and correlate the single data sets also on resource limited and embedded computer systems. The result is a common and extensive knowledge base, which integrates all information sources into complex data sets.

In comparison to other related software tools, the proposed framework fulfils essential requirements for a flexible usage, a resource-efficient runtime behaviour as well as a automated or semi-automated operating mode. The presented framework is used for several wireless sensor and actuator network projects at the Chemnitz University of Technology. We developed a standardised process for monitoring and archiving data from a heterogeneous sensor network topology in a synchronised way. Besides storing basic information from the sensor data sets, the system also integrates meta-information from the environment to increase the reuse factor of the measurement scenario. The universal XML data representation and a modular plugin system ensure a generic usage for all kind of sensor scenario. Multiple data input and output interfaces provides a high level of compatibility to other software tools and data formats.

Regarding the presented automotive application scenario, the proposed framework enables correlations between the measured sensor data from the test track and specific driver profiles. Accordingly, these information allow dynamic adaptations of the driving parameters within the vehicle. This offers novel and interesting possibilities to optimise a vehicle for the specific characteristics of its driver.

 ISBN: 978-1-61208-189-2

REFERENCES

[1] National Instruments. LabView. http://www.ni.com/labview/, 2010. [Online, retrieved: January, 2012].

[2] AMS GmbH. jBEAM. http://www.jbeam.de/german/produkte/jbeam.html, 2010. [Online, retrieved: January, 2012].

[3] Weisang. FlexPro. http://www.weisang.com/, 2010. [Online, retrieved: January, 2012].

[4] ASAM Consortium. Association for Standardisation of Automation and Measuring Systems. http://www.asam.net/, 2010. [Online, retrieved: January, 2012].

[5] L. Krishnamachari, D. Estrin, and S. Wicker. The impact of data aggregation in wireless sensor networks. In *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops*, pages 575–578. IEEE Computer Society, November 2002.

[6] M. Vodel, M. Lippmann, M. Caspar, and W. Hardt. Distributed high-level scheduling concept for synchronised, wireless sensor and actuator networks. *Journal of Communication and Computer*, 11(7):27–35, November 2010.

[7] V. Gupta and R. Pandey. Data Fusion and Topology Control in Wireless Sensor Networks. *WSEAS Trans. Sig. Proc.*, 4(4):150–172, 2008.

[8] H. Qi, S. S. Iyengar, and K. Chakrabarty. Distributed Sensor Fusion - A Review Of Recent Research. *Journal of the Franklin Institute*, 338(1):655–668, 2001.

[9] H. Qi, X. Wang, S. S. Iyengar, and K. Chakrabarty. Multisensor Data Fusion In Distributed Sensor Networks Using Mobile Agents. In *Proceedings of International Conference on Information Fusion*, pages 11–16, August 2001.

[10] A. Brown. *Component-Based Software Engineering*. Wiley-IEEE Computer Society Press, 1996.

[11] M. Vodel, M. Lippmann, M. Caspar, and W. Hardt. A Capable, High-Level Scheduling Concept for Application-Specific Wireless Sensor Networks. In *Proceedings of the World Engineering, Science and Technology Congress*, pages 914–919. IEEE Computer Society, June 2010.

[12] IEEE Computer Society. Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs). http://standards.ieee.org/getieee802/download/802.15.4-2006.pdf, 2007. [Online, retrieved: January, 2012].

[13] Zigbee Alliance. Zigbee specification. http://www.zigbee.org/en/spec_download/zigbee_downloads.asp, 2007. [Online, retrieved: January, 2012].

[14] Nanotron Technologies. Nanotron's transceiver enables iso compliant real time locating systems. In *The International Organization for Standardization and the International Electrotechnical Commission*, volume 24730-5:2010. New standard for Real Time Locating Systems (RTLS), April 2010.

[15] M. Vodel, W. Hardt, R. Bergelt, and M. Glockner. Modulares Framework fr die synchronisierte Erfassung, Verarbeitung und Aufbereitung heterogener Sensornetzdaten. In *Proceedings of the Dresdner Arbeitstagung Schaltungs- und Systementwurf*, pages 67–72. Fraunhofer Institute for Integrated Circuits, May 2010.

[16] M. Vodel, R. Bergelt, M. Glockner, and W. Hardt. Synchronised data logging, processing and visualisation in heterogeneous sensor networks. In *Proceedings of the International Conference on Data Engineering and Internet Technology*. Springer, March 2011.

 ISBN: 978-1-61208-189-2

Distributed Optimization of Energy Costs in Manufacturing using Multi-Agent System Technology

Tobias Küster, Marco Lützenberger, Daniel Freund, and Sahin Albayrak
DAI-Labor, Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Telephone: +49 (0)30 - 314 74000, Fax: +49 (0)30 - 314 74003
{*tobias.kuester|marco.luetzenberger|daniel.freund|sahin.albayrak*}*@dai-labor.de*

***Abstract*—While widely endorsed, the increased provision of electricity from renewable sources comes with the concern that energy supply will not be as reliable in the future as it is today, due to variations in the availability of wind and solar power. However, fluctuations in energy supply also give rise to volatility of the price for short-term energy procurement, and therefore bear the opportunity to save costs through shifting energy consumption to periods of low market prices. In a previous work, we presented an evolution-strategy-based optimization of production schedules with respect to day-ahead energy price predictions, yielding good results, but – being a stochastic optimization – not always arriving at the best solution. In this paper, we extend our framework by agent-based mechanisms for distribution and parallelization of the optimization, to increase scalability and reliability of the approach.**

***Keywords*-multi-agent systems; production planning; energy efficiency**

I. Introduction

In recent years, environmental-friendly production of energy and goods has gained more and more importance, with both customers and governments demanding for "green" production and increased provision of renewable energies. However, critics claim that by relying more on variable and volatile energy sources such as wind and solar power, the future energy supply will not be as stable as today, with variations in available energy over time [1].

However, these fluctuations in energy supply also give rise to the volatility of the price for short-term energy procurement, e.g., via the *European Energy Exchange* (EEX) [2], taking into account the predicted feed-in of wind energy, the current oil price and previous EEX results. This bears the opportunity for manufacturing industries to decrease production costs by shifting energy-intense production steps to periods of high wind and solar energy availability – and thus low energy prices – at the same time also fostering the use of environmental-friendly renewable energy sources. Today, demand-response mechanisms like this are only beginning to be implemented in the industry, but are expected to gain currency in some countries, as fluctuation of short term energy price becomes more distinctive [3].

In previous work [4], we presented an approach for optimizing a production schedule with respect to day-ahead energy price predictions, using evolution-strategy. The optimization yields good results most of the time; but being a stochastic algorithm, it does not always arrive at the best solution, getting stuck in local optima instead. Thus, to find the global optimum, the optimization needs to be run more than once on a specific process graph. However, depending on the complexity and granularity of the process graph, the optimization can be a time consuming process, and with the restricted time frame available between receiving energy price forecasts and the end of the bidding period, it becomes necessary to distribute and to parallelize the optimization procedure.

In this paper, the optimization framework is extended by agent-based mechanisms. Using a simple interaction protocol, the optimization can transparently be distributed to multiple servers to increase scalability and reliability, as individual optimization runs are executed in parallel. As we evaluated the results, we found out that even very few runs, or "populations", are sufficient to reliably arrive at a near-optimal solution, without increasing the time to find these results significantly.

We start this paper by outlining the principles of our so far work (Section II) and proceed by describing on how the multi-agent system paradigm can be used for the purpose of manufacturing process optimization with respect to dynamic energy procurement (Section III). We proceed with the evaluation of the optimization framework (Section IV) and subsequently compare our approach to related work (Section V). Finally, we wrap up with a conclusion (Section VI).

II. Previous Work

In the following, we will provide a short recapitulation of our work so far [4]. We start by presenting the production process meta model and proceed with details on the simulation and optimization algorithms. Subsequently, we explain how the system has been implemented.

A. Production Process Meta-Model

In our approach, any production process is modeled as a bipartite graph of *activities* and *resources*, similar to a

 ISBN: 978-1-61208-189-2

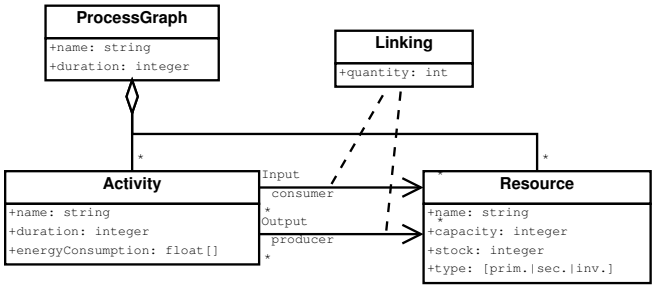


Figure 1. Process meta-model [4]

Petri net [5]. Activities have input and output resources, a duration of multiple atomic time steps, and a variable energy consumption over this duration, which can also be negative for energy sources and storage devices. Resources have a minimum and maximum capacity and a current stock. Resources are subdivided into primary resources (e.g., raw materials, intermediate products), secondary resources (e.g., pressurized air, gas, waste heat) and inventory resources (e.g., machines). The meta-model is shown in Figure 1.

Since electrical energy is the main concern of the optimization, it is not modeled as a resource but treated separately. Unlike other resources, electrical energy is available in (for all practical purposes) unlimited quantity and at a variable price, based on the energy market.

When an activity is executed, its input resources are consumed and its output resources are produced, and it will add to the overall energy consumption of the production process. Primary and inventory resources are consumed/allocated in the first step and produced/de-allocated in the last step of the activity's execution; secondary resources are produced or consumed in every step of the respective activity.

Using this simple meta-model, a wide range of production processes can be modeled. At the same time it facilitates the simulation and optimization of energy consuming activities in other domains, such as e.g., the utilization and charging schedules of electric vehicles.

B. Simulation and Optimization

The purpose of the optimization is to find the best possible *production schedule* for a given process model. In the implementation at hand, cost optimization is conducted mainly on the basis of day-ahead price forecasts, e.g., for the EEX electricity spot market. The optimization consists of three major steps: (1) The simulation of a given production schedule, (2) measuring the quality of that simulated schedule, and (3) finding the schedule with the highest quality.

1) Simulation: The simulation of a production schedule keeps track of the resource stocks and the energy consumption in each step of the simulation for the duration of the process, checking which activities are to be started, which activities are still running, and which activities are to be ended in the current step, producing and consuming resources and energy accordingly.

Concerning energy consumption and cost, two parameters of the simulation can be adjusted to reflect different determining factors: First, an *energy price curve* can be provided, for instance from the day-ahead energy market. Second, a *base energy* level can be specified, being the amount of energy the facility acquires via a flat fee. Energy consumption up to this level has already been paid for, so the *energy price curve* does not apply for it.

2) Quality Measurement: The *quality* of a production schedule p is determined by the inverse of its *defect*, which is the weighted sum of the total energy costs ($p_e \cdot w_e$) and the total over- and undershootings of the several resources' capacities ($p_{r,d} \cdot w_{r,d}$) over all steps of the simulation.

$$defect(p) = p_e \cdot w_e + \sum_{r \in \{p,s,i\}} \sum_{d \in \{l,h\}} p_{r,d} \cdot w_{r,d}$$

Different weights ($w_{r,d}$) can (and should) be used for resource stocks being too low and those being too high ($d \in \{l, h\}$) and for the different kinds of resources ($r \in \{p, s, i\}$).

Production schedules, which exceed the maximum or minimum capacities of a resource are not discarded, but are merely given a lower quality rating. For many optimization algorithms this is necessary in order to overcome local optima. For example, a schedule might be highly improved by swapping two activities. During this swap, there may be a phase in which the activities will both occupy a shared resource, but the benefits of the new schedule may be big enough to compensate for this temporary defect.

3) Optimization: For finding an energy- and cost-efficient production schedule, making best use of a given energy price curve, we use Evolution Strategy [6], which is similar to genetic algorithms.

As the name implies, Evolution Strategy is inspired by natural evolution: Using a $(\mu/\rho + \lambda)$ strategy, an initial "population" of μ individuals is generated. In the system at hand, each individual represents one production schedule. Based on these μ "parents", λ "offspring" are generated by recombining a random selection of ρ parents and slightly "mutating" the result. Finally, the quality of each of the parents and offspring is determined and the μ best individuals are selected to be the parents of the next generation. This process is repeated until a satisfactory production schedule is found.

The initial population is created by a very simple scheduler, aligning production activities as long as and as early as the primary resources permit, or until a desired quantity of products has been produced. In order to mutate an individual, either a random activity is inserted into or removed from the schedule, or one or more activities are moved to another position in the schedule, thus being executed earlier or later.

 ISBN: 978-1-61208-189-2

C. Optimization Framework and Tools

The approach is currently being evaluated using a prototypical implementation, which can be used to design the manufacturing process to be simulated, to configure and to run the actual optimization, and to visualize the results.

The process meta-model and a simple graphical editor for creating and configuring process models have been implemented as an extension to the Eclipse development environment. Following the usual notation for Petri nets, activities are represented by rectangles and resources by circles (see the example in Section IV).

Regarding the optimization, a generic optimization framework has been created, which can be used for optimizing different domains using different optimization algorithms. The actual Evolution Strategy algorithm as well as the process model domain have been implemented as plug-ins for this framework, targeting the optimization of comparably complex and heterogeneous industrial manufacturing processes. With respect to other domains, such as charge optimization of a large numbers of electric vehicles, other algorithms may be expedient.

For the manufacturing domain, the system features a large domain-specific area, providing controls for configuring the simulation and optimization (e.g., the energy price curve to use) and for showing the best production schedule found so far in a Gantt-like diagram. Once the optimization has come to an end, additional charts are available, showing the energy consumption and stocks of individual resources over the course of the simulation, as well as the development of these charts over the course of the entire optimization as a three-dimensional plot. Finally, the optimized process plan can be saved to file.

III. Agent-based Optimization of Manufacturing Schedules

While Evolution Strategy yields good results most of the time, it is also possible, as with other stochastic local search algorithms, that the optimization gets stuck in local optima. To increase the chances of arriving at a solution close to the global optimum, the optimization should be applied on more than one "population", and since the individual populations are independent of each other, they can easily be parallelized.

As described in the introduction, we combine the optimization framework with agent-oriented technologies to distribute and parallelize the optimization, and find global optima within reasonable time.

In the following we describe a simple interaction protocol, which makes a number of optimization servers (i.e., agents conducting the optimization) available to more than one optimization client. Further, we explain how the protocol was implemented using the JIAC V agent framework.

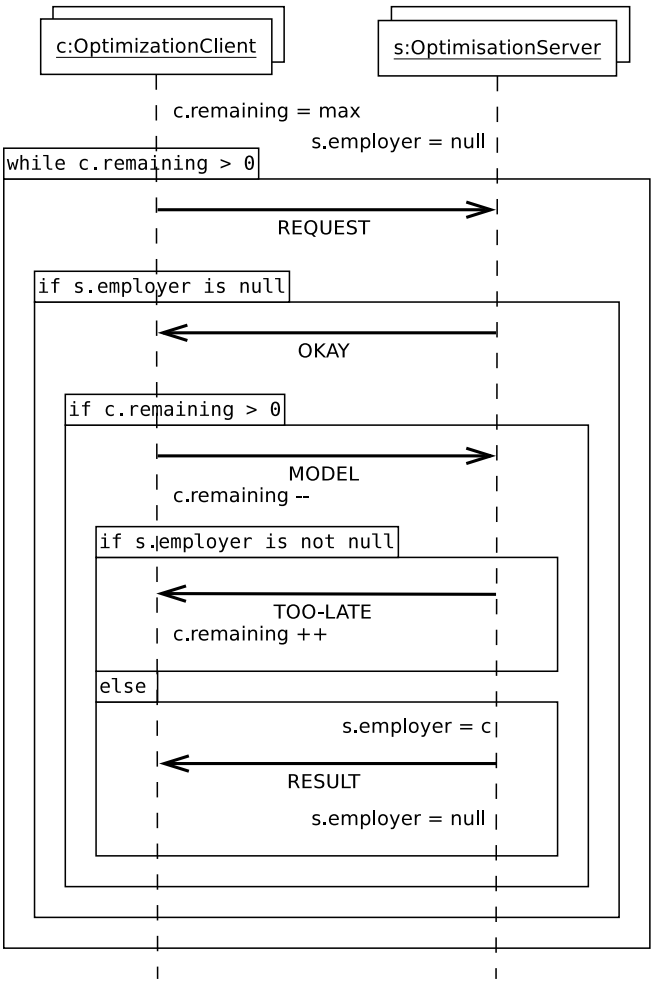


Figure 2. Interaction protocol used in the distributed optimization.

A. Interaction Protocol

Using the simple interaction protocol outlined in this section, each of the populations of a $(\mu/\rho+\lambda)$ optimization can be distributed to another agent. Since each run of the optimization, or each population respectively, is independent from the others, this does not introduce any noteworthy communication overhead. An interaction diagram of the protocol is shown in Figure 2.

The roles in the protocol are:

- *optimization client*, requesting an optimization
- *optimization server*, conducting the optimization

Obviously, there should be more than one optimization server agent for the distribution to provide any benefit at all, and there may be multiple clients, as well, sharing those servers. In the following we will describe the several interactions comprised in the protocol.

1) The protocol starts with a client broadcasting a REQUEST message to all the servers.
2) Each server receiving the message checks whether it

already has an "employer", i.e., whether it is currently running an optimization. If not, it replies with an OKAY message.

3) The client received the OKAY message, and if it still requires the server (i.e., if there have not been enough replies from other servers yet), it replies by sending the actual MODEL to be optimized to that server. The number of remaining optimization runs is reduced.
4) On receiving the MODEL message, the server will check again whether it already has an employer, as in the case of multiple clients, it might have sent OKAY messages to other clients, which may already have sent their MODEL messages.
   - If so, the server replies with a TOO LATE message. The client received this messages and corrects the number of remaining optimizations.
   - Otherwise, the server accepts the client as its new employer and starts the optimization run, and finally sends a message holding the RESULT back to the client.
   - At any time, the client can send an ABORT message, stopping the optimization.
5) The client continues sending out REQUEST messages until the desired number of optimizations has been conducted.

B. The JIAC V Multi-Agent Framework

JIAC V (Java Intelligent Agent Componentware, Version 5) is a Java-based multi-agent development framework and runtime environment [7]. Among others, JIAC features communication, tuple-space based memory, transparent distribution of agents and services, as well as support for dynamic reconfiguration in distributed environments, such as component exchange at runtime. Individual JIAC agents are situated within Agent Nodes, i.e., runtime containers, which also provide support for strong migration. The agents' behaviors and capabilities are defined in a number of so-called *Agent Beans*, which are controlled by the agent's life cycle.

C. Implementation

The protocol has been implemented by means of two JIAC Agent Beans, namely the *Optimization Client Bean* and *Optimization Server Bean*. Just like the optimization framework introduced in Section II, the Agent Beans were kept generic so that they – and thus the protocol – can just as well be used with domain-models other than the one presented in this work, and even with different optimization algorithms.

Using asynchronous messaging, the implementation with JIAC (or a similar multi-agent framework) has some advantages over traditional approaches using remote procedure calls or web services:

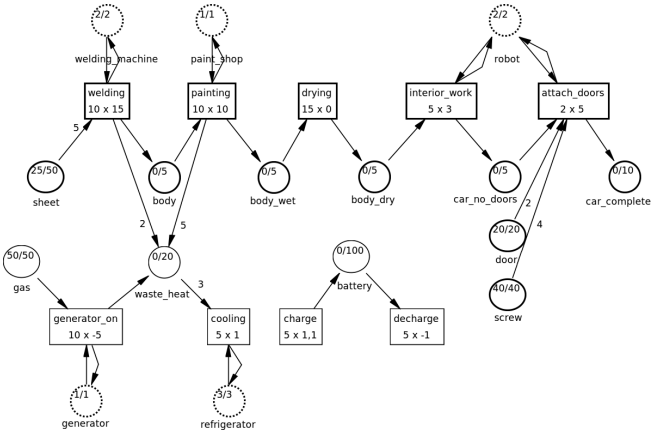


Figure 3. Example process: automobile construction (simplified)

- Both the Client Nodes and the Server Nodes can be distributed to any computer in the local network, with no need to configure IP addresses or ports.
- With each JIAC agent running in a separate thread, a node with multiple agents can be deployed to a multi-core server computer, and will automatically make ideal use of the several CPUs.
- Optimization procedures can be aborted ahead of time by sending the appropriate message. Similarly, the servers can send back intermediate results, to provide a trend for long-running optimizations.

Besides advantages over alternative means of distributed systems, the agent-based approach performs as expected with respect to previous local optimization. It yields good results in reasonable time and the variability of results decreases with an increased number of populations.

IV. EVALUATION

We evaluated our optimization algorithm and the benefits of distribution and parallelization using a fictional process, a production goal of five completed cars, a hill-shaped energy price curve and an evolution strategy with $\mu = 3$ and $\lambda = 8$.

A. Example Process

For the evaluation, a simple, fictional example process inspired by automotive industry was used (Figure 3). The process starts with two energy-intensive activities, which induce lots of waste-heat besides their primary production purpose: welding and painting the car chassis. Once the paint has dried, some interior works are performed, and finally the doors are attached to the chassis. For each of the intermediate products, a specific primary resource is created. The resulting production process graph is supplemented with utility activities and resources such as cooling, on-site electricity storage and a gas-powered cogeneration unit. The latter two elements can be used to temporarily decrease the grid energy consumption, but costs for the

 ISBN: 978-1-61208-189-2

corresponding increase in gas consumption will in turn add to the production schedule's penalty.

While the process surely is very simplified, it comprises most of the aspects that can be realized in the process model, for example

- the modeling of the basic production chain,
- one kind of machinery being used for two activities,
- the use of resources associated with a cost, or
- cooling facilities and other supporting processes.

B. Optimization Results

To asses the benefit of distribution and parallelization, the example process was optimized several times with different numbers of populations. The size of populations ranged from one to thirteen, and ten runs of the optimization were performed in each case. The results are shown in the logarithmic plot in Figure 4.

As can be seen, using only one population, the quality of the optimized process plan varies greatly. While there are some results with near-optimal quality, many populations apparently get stuck in local optima, and obtain a low overall-quality. For up to four populations, results start to look better, but are still noticeably scattered. For five and more populations, the results become reliable, with almost each optimization run resulting in near-optimal quality.

It may be noticed that the maximum quality reached – around 0.05 – is still far from the theoretically possible 1.0. The reason for this is that energy costs, no matter whether they could be improved any further, still add to the defect of the process. Thus, with minimum energy costs of around 20 (in no specific currency), the quality can not be much greater than 0.05.

Also to be noted is the gap in quality between around 0.015 and 0.045. This gap separates results, which still have resource conflicts, and those merely suffering from less-than-optimal energy costs. In the evaluation, the weight of resource conflicts was set to add greatly to the overall result's defect.

Further, we noted that there is little to none correlation between the time an individual optimization run takes, and the resulting quality (see Figure 4), i.e., a quick optimization run can yield a very good result, while a long-running optimization does not guarantee to bring a good result. Thus, one possibility to improve the performance could be to start a large number of optimization in parallel, and to abort the remaining optimization runs once the first few results to choose from have arrived.

V. Related Work

Industry has long since discovered, that process optimization is able to increase revenues significantly. As a result, there are many sophisticated applications available today. In this section we outline the latest optimization tools. Due to the broad range of existing approaches we focus our survey on works, which influenced us the most. We conclude this section by discussing significance of our work against the backdrop of contemporary applications.

Highly interesting for our work is the approach of *Santos et al.* [8], as it puts focus on energy related criteria. Yet, as opposed to our objective, the aim of *Santos et al.* is to reduce energy consumption in general, while we try to adapt our manufacturing schedules to a given objective function. *Bernik et al.* [9] developed a similar approach, although they do not account for energy criteria. The approach is capable to propose manufacturing schedules, which are able to fulfill a given production target. In addition to the manufacturing schedule, resource requirements are calculated and assigned to the production depots. *Schreiber et al.* [10] describe a similar application, which optimizes manufacturing schedules with respect to a specified given production target. As opposed to the approach of *Bernik et al.*, the application is able to calculate so called lot-sizes, which are defined as the number of pieces, which are processed at the same time at one workplace with one-off (time) and at the same costs investment for its set up [10].

In addition to the above mentioned academic works, there are many commercial software packages available.

The *Siemens Plant Simulation Software* [11], for instance, is a commercially available software, which facilitates the optimization of production systems and controlling strategies. Business- and logistic- processes may be supported as well. Processes are captured in compliance with an object oriented domain model. The *SIMUL8* framework [12], *Arena* [13] and *GPSS/H* [14] provide similar features and are able to simulate entire production processes, from warehouse capacities, to equipment utilization, to logistics-, transportation-, military- and mining applications. Beyond that, *SIMUL8* additionally accounts for real life requirements, such as maintenance intervals and shift patterns. Other types of software packages as for instance *Simio* [15] and *ShowFlow* [16] do not explicitly focus on the optimization of production processes, but on their visualization. For this purpose, most of the mentioned applications apply sophisticated 3D engines.

Thus far, the mentioned works are focused on the optimization of production processes. Yet, over the last years, the idea of general purpose frameworks emerged. Instead of focusing on a particular domain or problem, general purpose frameworks are able to optimize processes in general. Foundation to these frameworks is a generic meta-model, which is able to capture process structures.

PACE [17] and *AnyLogic* [18] for instance feature an arbitrary level of detail for process design. While *PACE* uses hierarchically arranged *High-Level-Petri-Nets* for this purpose, *AnyLogic* applies an object-oriented meta-model to capture process structures. *SLX* [19] takes a layered approach to process modeling. Most commonplace processes are handled in *SLX*'s upper layers, while unique

 ISBN: 978-1-61208-189-2

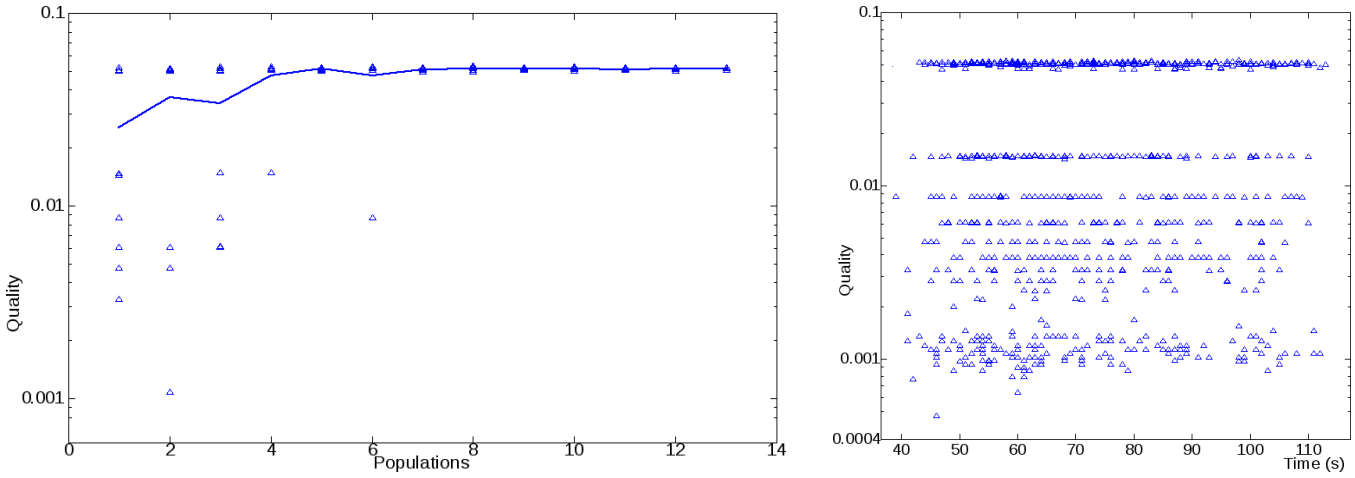


Figure 4. Left: Correlation of number of populations to expected result quality. The graph indicates average quality values. Right: Correlation of time of optimization run to result quality.

and more complex problems can be captured with *SLX*'s lower layers. The *Microsaint* [20] package totally avoids hierarchically structures and facilitates readability as well as easy comprehensibility. The framework entirely relies on flow charts as meta process language. In most analyzed frameworks, process design is usually supported by visual editing tools. The *ADONIS* framework [21] for instance provides an impressive graphical editor for the design and manipulation of the examined process system.

Finally, we analyzed tools which have been developed for similar optimization problems, but for domains different from manufacturing. Business processes for instance have a striking resemblance to manufacturing processes and as there are optimization frameworks for business processes, we want to mention the most prominent members of this realm as well.

To start with, *ProcessModel* [22] is a business process optimization software, which supports optimization from problem analysis to efficiency evaluation. The tool is able to visualize many aspects such as money savings or the efficiency of analyzed processes to serve customers. A similar application is *SIMPROCESS* [23]. In addition to the capabilities of *ProcessModel*, *SIMPROCESS* is able to handle hierarchical process structures and comes along with a set of sophisticated tools for the process design. Both applications apply means of simulation in order to verify optimized processes and to estimate their overall quality.

In this section, we gave a comprehensive overview on state of the art concepts and applications. To sum up; our idea of optimizing production with respect to dynamic energy tariffs is adopted by none of the examined applications. Further, we can state that energy related criteria are currently not comprehensively covered by state-of-the-art solution, as only the approach of *Santos et al.* facilitates such factors. We learned that evolutionary algorithms can be used to increase the performance of optimization algorithms and thus applied such principle [6]. Finally, the *AnyLogic* framework convinced us to apply mechanisms of distributed computing, namely the agent paradigm.

VI. Conclusion and Future Work

This paper proposed the use of a multi-agent system for the distributed process optimization with respect to energy consumption. A set of software agents has been designed and deployed to a physically distributed client-server-architecture, implementing an interaction protocol for the dynamic coordination of optimization processes and result aggregation. Feasibility and performance of the system were verified by using an exemplary manufacturing process. Results show that an increased number of parallel populations significantly decreases the variability of simulation outcomes and the probability of receiving a suboptimal result. Due to parallelization, the duration of the optimization did not vary noticeably with an increased number of populations.

In this study, only the distribution of many equivalent optimization jobs to several agents is evaluated. However, the agent-based optimization system is designed to accomplish variations between the individual jobs, e.g., using different settings for the optimization. Furthermore, diverse optimization strategies besides Evolution Strategy can be introduced as plug-ins to the system. As an example, there may be distinct independent sub-problems in the manufacturing site's overall process graph, such as the optimization of manufacturing processes on the one hand and the charging of forklift trucks on the other hand, where distinct optimization algorithms perform better or worse.

Future work will be dedicated to the evaluation of quality and performance gains through the aforementioned extensions to the system; namely diversification of population parameters (e.g., number of parents and offspring), diversifi-

 ISBN: 978-1-61208-189-2

cation of optimizing algorithms, and breaking down process graphs into sub-problems to distribute them among different agents.

Further, it is our intention to exploit the agent-paradigm stronger. In this work, we focused on the aspect of distribution and neglected other important characteristics of software agency, as for instance autonomy, pro- or reactiveness. The reason for this decision is simple, as we see the contribution of this particular paper in the distributed structure of our formerly centralized solution. For the future, we want to use this distributed structure as a basis for further extensions. Having this objective in mind, we aspire an autonomous energy procurement of additional energy and also an autonomous brokering of energy surpluses, based on predicted energy demands. In addition, we want to enhance our distributed optimization by load balancing capabilities. Optimization clients will be aware of the local load and be able to migrate to machines with free capacity.

VII. Acknowledgments

This work is funded by the *Federal Ministry of Economics and Technology* under the funding reference number 0327843B.

References

[1] H. Holttinen, P. Meibom, A. Orths, F. van Hulle, B. Lange, M. O'Malley, J. Pierik, B. Ummels, J. O. Tande, A. Estanqueiro, M. Matos, E. Gomez, L. Sder, G. Strbac, A. Shakoor, J. Ricardo, J. C. Smith, M. Milligan, and E. Ela, "Impacts of large amounts of wind power on design and operation of power systems, results of iea collaboration," *Wind Energy*, vol. 14, no. 2, pp. 179–192, 2011.

[2] European Energy Exchange AG, "EEX homepage," 2012. [Online]. Available: http://www.eex.com

[3] J. Torriti, M. G. Hassan, and M. Leach, "Demand response experience in europe: policies, programmes and implementation," *Energy*, vol. 35, no. 4, pp. 1575–1583, 2010. [Online]. Available: http://centaur.reading.ac.uk/23573/

[4] T. Küster, M. Lützenberger, and D. Freund, "An evolutionary optimisation for electric price responsive manufacturing," in *Proceedings of the 9th Industrial Simulation Conference, Venice, Italy*, S. Balsamo and A. Marin, Eds. EUROSIS-ITI, June 2011, pp. 97–104.

[5] T. Murata, "Petri nets: Properties, analysis and applications." in *Proceedings of the IEEE*, april 1989, pp. 541–580.

[6] I. Rechenberg, *Evolutionsstrategie : Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, ser. Problemata. Stuttgart-Bad Cannstatt: Frommann-Holzboog, 1973.

[7] B. Hirsch, T. Konnerth, and A. Heßler, "Merging agents and services – the JIAC agent platform," in *Multi-Agent Programming: Languages, Tools and Applications*, R. H. Bordini, M. Dastani, J. Dix, and A. El Fallah Seghrouchni, Eds. Springer, 2009, pp. 159–185.

[8] A. Santos and A. Dourado, "Global optimization of energy and production in process industries: a genetic algorithm application," *Control Engineering Practice*, vol. 7, no. 4, pp. 549–554, 1999.

[9] I. Bernik and M. Bernik, "Multi-criteria scheduling optimization with genetic algorithms," in *Proceedings of the 8th WSEAS International Conference on Evolutionary Computing*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2007, pp. 253–258.

[10] P. Schreiber, P. Vazan, P. Tanuska, and O. Moravcik, "Production optimization by using of genetic algorithms and simulation model," in *DAAM International Scientific Book 2009*, B. Katalinic, Ed. DAAM International, 2009.

[11] Siemens, "Plant simulation — plant, line and process simulation and optimization," Project Broschure, 2010, siemens Product Lifecycle Management Software Inc. [Online]. Available: http://www.plm.automation.siemens.com/en_us/Images/7541_tcm1023-4957.pdf

[12] K. Concannon, M. Elder, K. Hunter, J. Tremble, and S. Tse, *Simulation Modeling with SIMUL8*, 4th ed. Visual Thinking International Ltd., 2003.

[13] M. D. Rossetti, *Simulation Modeling and Arena*, 1st ed. Wiley, 2009.

[14] R. C. Crain, "Simulation using GPSS/H," in *Proceedings of the 29th Winter Simulation Conference*, December 1997, pp. 567–573.

[15] J. A. Joines and S. D. Roberts, *Simulation Modeling with SIMIO: A Workbook*. Simio LLC, 2010.

[16] I. S. Solutions, "The ShowFlow website," 2011. [Online]. Available: http://www.showflow.com/

[17] B. Eichenauer, "Optimizing business processes using attributed petri nets," in *Proceedings of the 9th Symposium about Simulation as Commercial Decision Help*, March 2004, pp. 323–338.

[18] P.-O. Siebers, U. Aickelin, H. Celia, and C. W. Clegg, "Understanding retail productivity by simulating management practices," in *Proceedings of the Eurosim 2007, Ljubljana, Slovenia*, 2007, pp. 1–12.

[19] J. O. Henriksen, "SLX: The x is for extensibility," in *Proceedings of the 32nd Winter Simulation Conference*, 2000, pp. 183–190.

[20] D. W. Schunk, W. K. Bloechle, and K. R. L. Jr., "Micro saint: Micro saint modeling and the human element," in *Proceedings of the 32nd Winter Simulation Conference*, 2002, pp. 187–191.

[21] S. Junginger, H. Kühn, R. Strobl, and D. Karagiannis, "Ein Geschäftsprozessmanagement-Werkzeug der nächsten Generation – ADONIS: Konzeption und Anwendungen," *Wirtschaftsinformatik*, vol. 42, no. 5, pp. 392–401, 2000.

[22] ProcessModel, Inc., "The ProcessModel website," 2011. [Online]. Available: http://www.processmodel.com/

[23] CACI, "The SIMPROCESS website," 2011. [Online]. Available: http://simprocess.com/

 ISBN: 978-1-61208-189-2

A Privacy Preserving and Secure Authentication Protocol for the Advanced Metering Infrastructure with Non-Repudiation Service

Chakib Bekara, Thomas Luckenbach
RESCON Department
Fraunhofer FOKUS Institute
Berlin, Germany
{*chakib.bekara, thomas.luckenbach*}*@fokus.fraunhofer.de*

Kheira Bekara
LOR Department
Institut TELECOM Sud-Paris
Evry, FRANCE
kheira.bekara@it-sudparis.eu

***Abstract*—In the smart grid, smart meters play an important role in keeping a real-time balance between energy production and energy consumption. The advanced metering infrastructure is responsible of collecting, storing, analyzing and providing metering data from smart meters to the authorized parties, and also carrying commands, requests, messages and software updates from the authorized parties to the smart meters. As such, advanced metering infrastructure is one of the important components of the smart grid, and securing it is a prerequisite for guaranteeing a large acceptance and deployment of the smart grid. One step towards securing it is to provide authentication, integrity and confidentiality services. Another important step is to preserve the privacy of the end-customer equipped with a smart meter, where all its related-data (including metering data) are kept secret, such as energy consumption, billing and which smart appliances are used in its premises. In this paper, we propose an ID-based authentication protocol for the advanced metering infrastructure, which provides source authentication, data integrity and non-repudiation services, while preserving the end-customer's privacy.**

Keywords-Identity-based Cryptography*; *Key Establishment*; *Data Source Authentication*; *Smart Grid*; *Advanced Metering Infrastructure*; *Smart Meter*; *User's privacy

I. INTRODUCTION

The term SM (Smart Meter) designs an advanced digital utility meter (electricity, gas, water, heat, etc.) equipped with a two-way communication interface. In the context of the SG (Smart power Grid), a SM is installed at an end-customer premises, and is able to communicate with the utility (energy utility), by sending metering data (e.g., energy consumed/locally produced, grid status, meter status, etc.), and also by responding to messages from the utility (e.g., software update, realtime pricing, load shedding, energy cut-off, etc.).

The SM is a key element in the AMI (Advanced Metering Infrastructure), as shown in Figure 1. The AMI [1] is responsible for collecting, analyzing, storing and providing the metering data sent by the SMs to the appropriate authorized parties (e.g., energy provider, utility, SG operator, etc.). The AMI is also responsible for transmitting requests, commands, pricing-information and software updates from the authorized parties to the SMs. Figure1 presents a simplified view of the AMI as a part of the SG, where we differentiate the following components[2]:

- End-customer's HAN: this includes the SM, which is the main component, in addition to SAs (smart appliances), e-car (electric car) and local renewable energy sources in a customer's premises.
- GW (Gateway): the GW acts as an interface between a set of SMs and the AMI head-end. The GW plays the role of a concentrator, collecting data from several SMs using a local short-distance communication infrastructure (e.g., ZigBee, Bluetooth, WiFi, PLC, etc.), then sending them using a long-distance communication infrastructure (e.g., Internet, GSM, WiMax, GPRS) to the AMI head-end. It could also play the role of a firewall, by protecting the end-customers HAN from outside attackers. There may be several levels of GWs: a GW per a residential building block (a set of HANs), an upper level GW per a set of building blocks, etc.
- AMI Communication Infrastructure: in its simplest form, this network provides a communication path from the SM/GW to the AMI head-end, and reciprocally.
- AMI head-end: is responsible for two-way communication with SMs/GWs. Thus it serves as a gateway between the end-customer's HAN (mainly the SM) and the AMI back-end.
- AMI back-end: the components in the AMI back-end manage SMs, use data collected from SMs (for billing, grid status estimation, consumption/production forecasting, outage management, etc.) and send data and requests (software/firmeware update, real-time pricing, load shedding, etc.) to SMs. The MDMS (Meter Data Management System) [2] is responsible for storing all the exchanged data with the SMs, then dispatching them to the authorized receivers.

The well-functioning of the SG is based on the trustworthy of the overall data flow(assuming authentic origin/source and data integrity), including the data exchanged in the the AMI [3]. Attacking the AMI will impact the utility and the end-customer, and, as consequence, threaten the

 ISBN: 978-1-61208-189-2

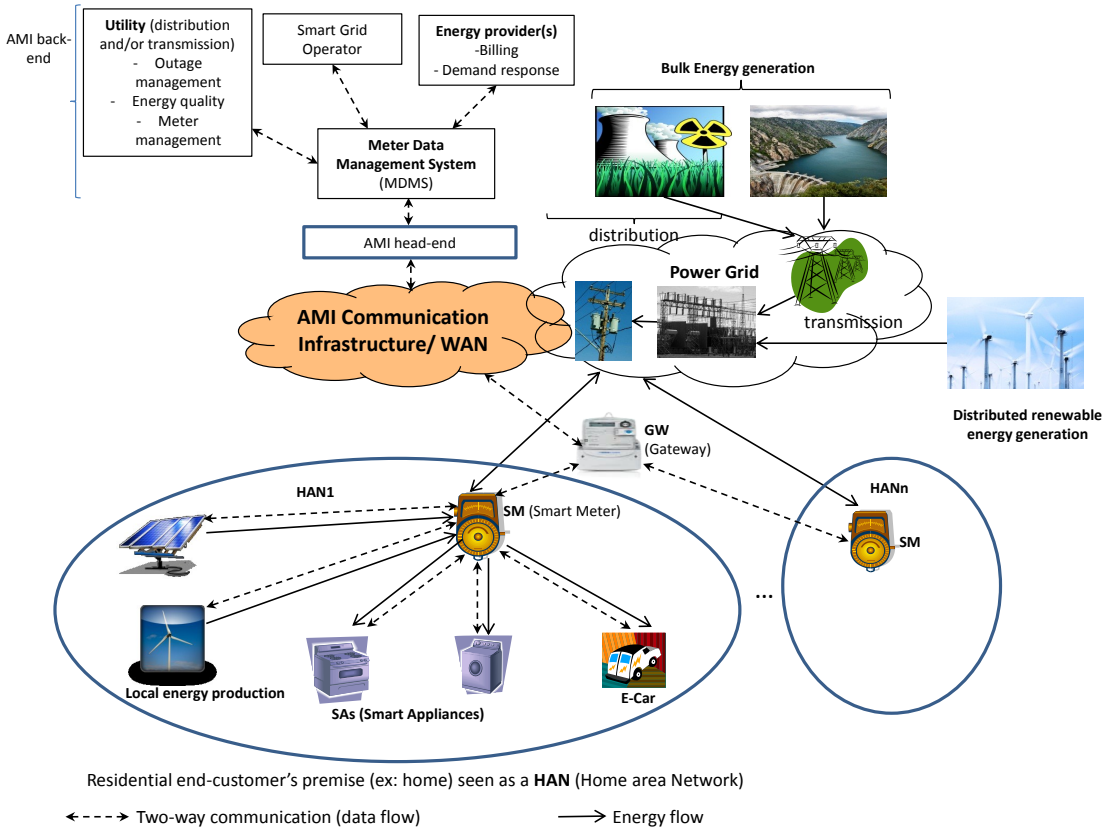


Figure 1. A general view of the AMI as part of the SG

security of the SG. An attacker could impersonate the SMs and send false meter readings to the utility. This may results on financial loss for the utility, unbalanced energy production/consumption and may lead to power outages. Moreover, an attacker could target the end-customers HAN, by masquerading as the utility and sending fake commands (e.g., disconnection or load shedding), fake pricing messages (low price during peak period and high price during off-peak period), resulting on financial loss for the end-customer. The attacker could also impersonate the GW to which is connected the SM and thus drop all messages sent from/to the SM, impacting both the utility and the end-customer. Finally, the attacker could also impersonate the SM to the local SAs inside the end-customers HAN, making them to operate during peak period .

Securing the flow of data in the AMI requires establishing secure (end-to-end) communication channels between communicating parties in the AMI, providing the basic security services: data source authentication, non-repudiation and confidentiality [3], where:

1. Data source authentication allows the verification of the identity that one party claim to have (origin authentication), and that the data received from that party was not altered en-route (data integrity). This service protects the AMI from the following attacks: identity impersonation, MITM (Man-In-The-Middle) [4] attack and data injection/modification.
2. Non-repudiation prevents a sender from denying sending a message. This service is required to avoid that a SM (end-customer) deny sending some metering data (energy consumption), or the energy provider deny sending some real-time pricing. This service is useful when responsibilities in case of dispute need to be clearly identified.
3. Confidentiality protects data from being legible to unauthorized parties.

In our paper, we mainly investigate how to efficiently provide the two first security services. In addition, we consider a third security service, which is related to the end-customer's HAN privacy. Information about the SAs inside the end-customer's HAN (e.g., identity, type, etc.) should not be divulged, even when a SM and a SA need to mutually authenticate.

In this paper, we present an ID(Identity)-based authen-

 ISBN: 978-1-61208-189-2

tication protocol for the AMI, providing data source authentication and non-repudiation services, and preserving the privacy of the end-customer. In Section II, we review some related works about authentication in AMI. In Section III, we describe our motivation towards our proposal, give the assumptions we made and summarize the notations we use. Section IV presents our ID-based authentication protocol for the AMI, and Section V discusses its security and gives a brief comparison with the related works. Section VI gives some perspective works and concludes the paper.

II. Related Works

Several authentication protocols (APs) were proposed for the SG and the AMI, all of them rely on the use of one or more of the following cryptographic keying materials/schemes:

1. Using traditional PKC (public-key cryptography); coupled with a PKI [5] (public key infrastructure), as in [6].
2. Using IBC (ID-based Cryptography) [7]; coupled with a single trusted PKG (Private Key Generator), as in [8].
3. Using solely symmetric-key cryptography; coupled with a single trusted KDC (Key Distribution Center), as in [9] [10].

The proposed authentication protocols achieve authentication between two parties A and B, in one of the following two ways:

- Prove the possession of a shared secret key (pre-shared or established), that only both parties know/share. This is typically done by generating/verifying a *MAC* (message authentication code) using the shared key. However, MACs do not provide non-repudiation service, since the used key is known to the two parties.
- Prove the possession of a private information (called private key) without revealing it, that the other party could verify using an authentic public information related to the prover (called public key). This is typically done by generating digital signatures using the signer's private key, while the receiver uses the signer's public key to check the signatures. Digital signatures provide non-repudiation service, *if and only if* the signer's private key is known to the signer only.

In [6], Fouda et al. investigate the authentication between two parties A and B of the SG: mainly a SM and a GW. Each entity possess a pair of self-generated private/public keys, while the public key is certified by one of the CAs (Certification Authorities) of the PKI. After verifying the certificate of each others (is valid and not revoked), A and B use the authenticated DH key establishment mechanism [4], in order to securely establish a secret shared key K_{A_B}, that both use to provide data source authentication. If non-repudiation is required, each party can use its private key to generate signatures.

In [8], So et al. propose an ID-based encryption and signature protocol for the AMI, based on IBC. The authors

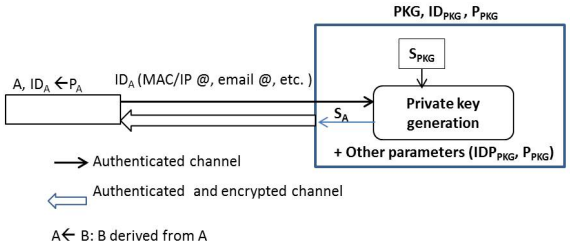


Figure 2. Private/Public Keys generation in IBC System

assume the existence of a widely trusted PKG that possesses a master private (S_{PKG})/public (P_{PKG}) keys. The PKG generates the ID-based private key of each entity A in the AMI using ID_A and S_{PKG}, while A's ID-based public key could be easily derived from A's identity ID_A (see Figure 2). Again, to provide data source authentication, A uses its private key to generate signatures over the messages it sends, while a receiver B uses ID_A to derive A's public key, then checks the authenticity of the messages received from A. Since the protocol is based on IBC, certificates are not used. The authenticity of A is implicitly checked if its signature is successfully verified, which means that A owns the right private key issued by the PKG.

In [10], Ayday et al. investigate authentication in end-customer's HAN, where each HAN contains: a GW, a SM, and a set of SAs. The authors assume that the utility plays the role of a widely trusted KDC in the SG, and that each GW possesses a unique IP address (serving as its ID) issued by an ISP (Internet Service Provider). The KDC shares a long-term unique secret key LTK_A with each entity A. Two entities A and B, with corresponding unique identifiers ID_A and ID_B, trying to establish a secure communication for the first time, send first an authenticated key-establishment request carrying both ID_A and ID_B, to the KDC. The KDC serves the requests as follows:

- The KDC checks that the messages are authentic using LTK_A and LTK_B.
- If the requests originate from a SM and a GW and are authentic, the KDC first uses the service of a third party providing localization information, to make sure that SM and GW are collocated (belong to the same HAN). Mainly, the KDC knows the location Loc_{SM} of SM due to the billing address of SM, then sends ID_{GW} and Loc_{SM} to the ISP (internet service provider) of GW, which could determine whether or not Loc_{GW}=Loc_{SM} . In this way, the KDC avoids a wrong/malicious key-establishment between a GW and SM of different HANs. If SM and GW are collocated, the KDC generates a secret key K_{SM_GW} and securely sends it encrypted using LTK_{SM} to SM and using LTK_{GW} to GW.
- If the key-establishment is between GW/SM and a SA,

 ISBN: 978-1-61208-189-2

the KDC first verifies that the GW and the SM related to the SA have already established a secret key. Finally, the KDC issues a secret key K_{SA_GW} (resp. K_{SA_SM}) and securely sends it to SA and GW (resp. SA and SM).

In [9], Yan et al. present an authentication and encryption protocol for the AMI, where SMs are interconnected through a multi-hop wireless network, and form a logical linear communication path to reach a remote collector node (GW). The authors assume the existence of trusted KDC (utility), which shares a secret key with each SM and securely issues a secret key to each pair of SMs in the communication path. A SM encrypts its reading using the key shared with the GW, and authenticates the data it sends (including data received from previous SMs in the path) using the key shared with the next SM in the path. In this way, each SM could verify the authenticity of the data it receives, and also guarantee the confidentiality of the metering data that it generates.

A. discussion

Authentication protocols based on symmetric-key cryptography coupled with KDC [9] [10], and those based on IBC coupled with a PKG [8], are known for their relatively low induced overheads (computation, storage, transmission), and lightweight management requirements (no need for PKI, since certificates are not used). In the other side, authentication protocols based on PKC are known for their relatively expensive overheads, especially for resources-constrained devices, and require the costly deployment of a PKI to issue and manage a large number of digital certificate. However, the fine performance of [8] [9] [10], comes at the expense of some security issues and some hard to fulfil assumptions:

- All the three protocols assume the existence of a *single trusted* entity (usually the utility) in the whole SG, playing the role of the PKG in [8] and the role of the KDC in [9] [10]. Assuming the existence of a single trusted entity (PKG or KDC) in the AMI is a hard to fulfil assumption, and do not scale to a so large network (millions of SMs, SAs, etc.). It is not trivial that the large number of manufacturers of SMs, SAs, GWs, e-cars, etc., would accept to trust the same single entity for key management. Moreover, would this single entity be able to efficiently manage the security of a large number of entities and systems involved in the AMI?
- In [8] based on IBC, the PKG issues the private keys of all entities of the SG. As a consequence, there is a key-escrow problem, and thus, non-repudiation service could not be guaranteed.
- In [9] [10], the KDC issues a secret shared key for any pair of communicating entities A and B in the AMI (SM, GW, SA, etc.). As a consequence, the privacy of communications in the AMI is not completely preserved, since the KDC can easily eavesdrop on the encrypted messages, or even impersonate any entity without being detected. Moreover, the end-customer's privacy in [10] is not preserved, since information related to the SAs in its HAN could be divulged during the key-establishment phase between a GW/SM and SAs.
- [10] assumes the existence of a GW per HAN/SM, whereas in practice there is one GW to serve a set of SMs. Requiring a GW per HAN, either means integrating a GW (with all its advanced features) with each SM, or deploying a separate GW per SM/HAN. Both solutions are financially expensive due to the large number of deployed SMs. Assuming that one GW serves a set of HANs, the use of ISP to check whether the SM and the GW are collocated will fail, since the GW and the SM are now in two different locations. As a consequence, preventing wrong/malicious key-establishment could not be guaranteed.
- Assuming that the utility plays the role of the KDC or the PKG is problematic, since in many countries (e.g., USA) several independent electricity utilities may operate in the same region. In this case, we will end up with several KDCs/PKGs, which make all the previous authentication protocols not directly applicable.

Finally, all the proposed protocols do not consider the case where SAs of HAN_i could successfully mutually authenticate with the SM of a neighboring HAN_j. This situation could make the SAs to be controlled by the SM of HAN_j instead of the SM of HAN_i. Meanwhile, the same problem occurs, except for [8], when establishing secure communications between a SM and the *right* GW: e.g., a SM communicate with the GW of a neighboring building and not the GW of the the local building.

III. Motivation, Assumptions and Notations

A. Motivation and Assumptions

To provide an efficient authentication for the AMI, we mainly make use of symmetric-key cryptography and IBC, and rarely consider using classical PKC (RSA, DSA, and even ECC) [4], since it requires the use of certificates, where certificate distribution/fetching and certificate validation will add extra overheads that some resource-constrained devices in the AMI (SMs, SAs, etc.) could not easily offer. The only exception is for PKGs, which still use classical ECC [11].

Moreover, in order to provide non-repudiation service and protect the end-customer's HAN privacy, we do not rely on a KDC for authentication and key management. Instead, we use a variant of IBC called certificate-less IBC [12], where each entity's private key is partially issued by the PKG, the other part of the key being securely issued by the entity itself. Using certificate-less IBC, we can now provide non-repudiation service for the AMI, which was not possible using the basic IBC as in [8].

Unlike [8] [9] [10], we assume the existence of several trusted key management authorities (PKG), where entities

 ISBN: 978-1-61208-189-2

Table I
USED NOTATIONS

SM	Smart meter
SA	Smart appliance
GW	Gateway
EP	Energy Provider
ID_A	unique identity of entity A
N_A	a nonce value generated by A
P_A	A's public-key
S_A	ID-based private key of A, or simply A's private key
PKG_i	the i^{th} Private Key Generator
P_{PKG_i}	Public key of PKG_i
S_{PKG_i}	master-secret Private key of PKG_i
MAC	Message Authentication Code
K_{A_B}	an l-bit secret shared key between entities A and B
MAC_K(M)	a MAC generated over message M using key K
Enc_K(M)	message M encrypted using key K
$\sigma_{S_A}(M)$	a signature generated over message M using S_A
p	a large prime of length $\mid p \mid$>160 bits
F_p	a finite field, F_p=[0, p-1]
$E(F_p)$	an elliptic curve defined over F_p
aP	scalar to point multiplication: addition of P a times
$x \parallel y$	concatenation of x and y

served by different PKGs could authenticate and establish secure communications. In this way, the scalability is improved, no single point of failure exists, and the end-customer's privacy is enhanced.

Finally, we assume that the SMs and the GWs deployed on the AMI network are owned and managed by the utility, which owns also the electricity distribution and/or transmission power network. In the other side, SAs inside the end-customer's HAN are owned and managed by the end-customer. In our paper, we make a distinction, which is not made by the other works, between the utility and the energy provider:

- The utility owns the physical electric infrastructure until the end-customer's electricity point of delivery (SM), over which electricity is delivered.
- The energy provider; with which the end-customer signs a contact, supplies the end-customer by energy (by buying it from energy produces), and bills it for the consumed energy. The utility is informed about each signed contract between any end-customer (SM) and any energy provider.

B. Notations

Table I summarizes the notations used through the remaining of the paper.

IV. OUR PROPOSED ID-BASED AUTHENTICATION PROTOCOL FOR THE AMI

In this section, we propose and describe an ID-based authentication protocol for the AMI, which provides source authentication, data integrity, non-repudiation and preserves the end-customer's HAN privacy.

Our protocol involves three phases: System setup phase, Node initialization/Private key generation phase and Data Source Authentication phase.

A. Phase I-System Setup

We assume the existence of t trusted entities in the SG, PKG_i $i = 1, ..., t$, playing the role of private key generation authorities. It is evident that t is infinitely negligible compared to the number of entities involved in the SG. The t PKGs agree on the use of the same elliptic curve $E(F_p)$, with the simplified domain parameters $Param_E = (a, b, p, n, P)$, where:

- p is the order of the finite field F_p over which is defined $E(F_p)$.
- $a, b \in F_p$ are the coefficients of $E(F_p)$.
- $P \in E(F_p)$ is a generator point of a cyclic subgroup of $E(F_p)$, and n is a big-prime and is the order of P.

The PKGs also agree on the use of two n-degree cyclic groups: $G_1 \subset E(F_p)$ (additive group, e.g., the subgroup of $E(F_p)$ defined by P) and G_T (multiplicative group, e.g., an extension field of F_n), on a symmetric pairing function e [13] and on two hash functions:
$H_1 : \{0,1\}^* \times G_1 \rightarrow G_1$ and $H_2 : G_T \times G_T \rightarrow \{0,1\}^l$.

e has the following properties [13]:

- *Bilinearity*: $e(aR, bS)$=$e(abR, S)$=$e(R, abS)$=$e(R, S)^{ab}$, $\forall$ a,$b \in F_n^*$, $\forall$ $R, S \in G_1$
- *Non-degeneracy*: $e(P, P) \neq 1_{G_T}$, where 1_{G_T} is the identity element of G_T, P is the generator of G_1.
- *Computability*: there exists an efficient algorithm implementing and computing e.

Then, each PKG_i performs the following:

- Picks a random master-secret private key $S_{PKG_i} \in {F_n}^*$ and computes its public-key P_{PKG_i}=$S_{PKG_i}P$, then sets its IBC system parameters: $Param_i$=$(n, G_1, G_T, e, P, P_{PKG_i}, H_1, H_2) \cup Param_E$.
- Securely gets the identity ID_{PKG_j} of each remaining PKG_j along with an authentic copy of P_{PKG_j}. Then, PKG_i issues a cross-domain certificate to each PKG_j, where T_{exp} is the certificate's expiration date:

$$Cert_{i \rightarrow j} = \underbrace{ID_{PKG_i}, ID_{PKG_j}, P_{PKG_j}, T_{exp}}_{\pi} \sigma_{S_{PKG_i}}(\pi)$$

  PKG_i makes $Cert_{i \rightarrow j}$ available by publishing it on a public repository referenced to it by $@Dir_i$.

For signature generation/verification over the certificates, we use the ECDSA signature scheme [11]. Figure 3, summarizes Phase I

B. Phase II-Node Initialization/Private Key Generation

Each entity involved in the AMI (SM, GW, SA, e-car, etc.) is issued a *partial* ID-based private key by one of the t PKGs at the manufacturing phase. Each manufacturer signs a contract with one or more PKGs, in order to securely provide each device it produces with its partial ID-based private key, along with the necessary IBC system parameters. We assume that the initialization step is performed over an

 ISBN: 978-1-61208-189-2

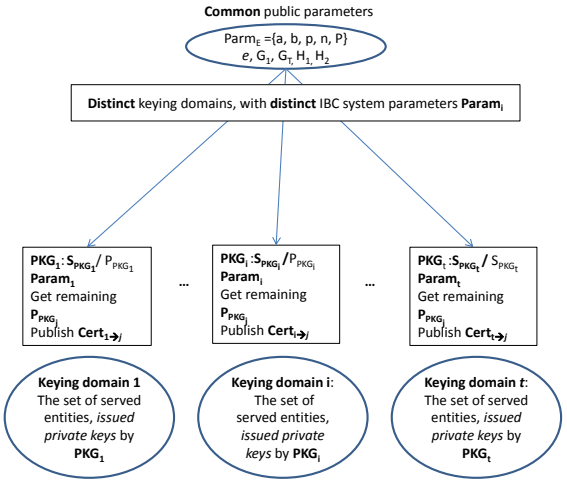


Figure 3. System Setup Phase

already established *secure channel* between PKG_i and the manufacturer.

Thus, a device A with a unique built-in identity ID_A (e.g., serial number, MAC address, IPv6 address, etc.), served by PKG_i will be *securely* initialized as follows:

- A is loaded with $Param_i$, $@Dir_i$
- Then, A picks a random $k_A \in F_n^*$ and sends to PKG_i: ID_A, k_AP.
- PKG_i sends to A its partial private key: D_A=$S_{PKG_i}H_1(ID_A, k_AP)$
- A computes its total ID-based private key S_A=k_AD_A, and sets P_A=$< k_AP$, $k_AP_{PKG_i} >$. A securely stores k_A and S_A, since it needs them for signature generation, public-key decryption and key-establishment.

At the end of the initialization, A is the *only* entity knowing S_A. Even PKG_i does not know it since it could not know k_A. A is said to belong to the keying domain i defined by PKG_i (see Figure 3). In the same way, each EP will securely get its partial private key, then computes its total private key.

Finally, each entity A=$\{SM, GW\}$ purshased by the utility, goes through a second step of initialization before its deployment. Mainly, the utility pre-loads the SM/GW with its identity $ID_{Utility}$ and a unique long term shared secret key LTK_A.

C. Phase III-Data Source Authentication

In this phase, two communicating parties A and B of the AMI (EP, GW, SM, SA, etc.), first mutually authenticate (the first time they met), then securely exchange messages. Two cases can be distinguished:

- **Case I**: Intra-domain communications: both A and B belong to the same keying domain i defined by PKG_i.
- **Case II**: Inter-domains communications: A and B belong to two different keying domains i and j, defined by PKG_i and PKG_j respectively.

We only consider **Case II**, since Case I is trivial. We consider three scenarios of communication, where all the communicating entities belong to two different keying domains:

- between EP and SM.
- between SM and GW.
- between SM and SA

For simplicity, we assume that during a communication the initiator always belongs to domain i, whereas the responder belongs to domain j. Also, we give the details of the ID-based signature generation/verification we use at the end of the section, and not for each scenario.

1) Communication between EP and SM: When the end-customer signs a new contract with an EP to be its energy supplier, the EP initiates a communication with the corresponding SM, by by setting a request $Req1$ (association with a new EP) and generating an ID-based signature over $M1$, then sends the following message to the SM:

$$\underbrace{Req1, ID_{EP}, P_{EP}, ID_{PKG_i}, N_{EP}, ID_{SM}}_{M1} \sigma_{S_{EP}}(M1) \quad (1)$$

Upon reception of (1), the SM checks the message freshness (not replayed) from the received nonce value N_{EP}. If expired or not already held, the SM checks from $@Dir_j$ whether or not there exists a certificate $Cert_{j\to i}$ issued from PKG_j to PKG_i. Assuming it exists, the SM checks $Cert_{j\to i}$'s validity using P_{PKG_j} following the ECDSA signature scheme [11]. If the certificate is valid (certificates signature is valid), the SM trusts P_{PKG_i}, and as consequence assumes that P_{EP} is a valid public-key. Finally, SM checks the signature on (1). If valid, the SM sends a request $Req2$ to the utility to check whether ID_{EP} is a valid energy provider, and whether or not it must accept $Req1$:

$$\underbrace{Req2, ID_{SM}, ID_{EP}, N_{SM}}_{M1a}, MAC_{LTK_{SM}}(M1a) \quad (2)$$

Upon reception of (2), the utility checks that the message is fresh and authentic, then checks if there is a new association between ID_{SM} and ID_{EP} (a new energy supply contract has been signed). If it is the case, the utility sends the following approval confirmation message (otherwise sends a deny message):

$$\underbrace{OK, ID_{EP}, ID_{SM}}_{M1b}, MAC_{LTK_{SM}}(M1b \parallel N_{SM}) \quad (3)$$

Upon reception of (3), the SM checks if it is fresh and authentic (MAC is valid) and that the response of is 'OK'.

 ISBN: 978-1-61208-189-2

If all verifications are positive, the SM considers that it has successfully authenticated the EP. The SM responds to the EP to both authenticate itself and also to confirm the acceptance of $Req1$:

$$\underbrace{OK, ID_{SM}, P_{SM}, ID_{PKG_j}, N_{SM}, ID_{EP}}_{M2}, \sigma_{S_{SM}}(M2 \parallel N_{EP}) \quad (4)$$

Upon reception of (4), the EP verifies that it is fresh. If expired or not already held, the EP checks from $@Dir_i$ whether or not there exists a certificate $Cert_{i \to j}$. Assuming it exists and is valid, the EP trusts P_{PKG_j}, and consequently assumes that P_{SM} is valid. The EP checks the ID-based signature on 4. If the signature is valid and the response is 'OK', it considers that it has successfully authenticated the SM and sends the following message to conclude the mutual authentication phase:

$$Finish, \sigma_{S_{EP}}(Finish, ID_{EP}, ID_{SM}, N_{SM}) \quad (5)$$

Now, the EP and the SM can successfully generate and verify signatures over their exchanged messages, thus ensuring data source authentication and non-repudiation. They could also establish a shared secret key, as described in Section IV-C2, to secure their communications if non-repudiation is not mandatory.

2) Communication between SM and GW: A newly deployed SM in a residential building needs first to authenticate the GW of the building, before sending its metering data to the AMI head-end via this GW. The deployment of the SM is done by an authorized employee of the utility. To avoid associate the SM with a wrong GW (neighboring GW or a fake GW), all what is needed is that the employee indicates to the SM the ID_{GW} to which the SM needs to communicate. Assuming the employee securely initializes the SM with ID_{GW}, the following steps are performed between the SM and the GW:

The SM sets an *attach-req* request to ask to be attached to the GW, generates an ID-based signature over $M3$, then sends the following message to ID_{GW}

$$\underbrace{attach-req, ID_{SM}, P_{SM}, ID_{PKG_i}, N_{SM}}_{M3}, \sigma_{S_{SM}}(M3) \quad (6)$$

Upon reception of (6), the GW checks its freshness. If expired or not already held, the GW checks from $@Dir_j$ whether or not there exists a certificate $Cert_{j \to i}$. Assuming it exists and is valid, the GW checks the validity of the signature on (6) using P_{SM}. If valid, using secret values k_{GW} and S_{GW}, and public-key P_{SM}, the GW computes a secret key K=K_{GW_SM}=

$$H_2(e(S_{GW}, k_{SM}P)e(k_{GW}H_1(ID_{SM}, k_{SM}P)k_{SM}P_{PKG_i}))$$

then sends the following message to the SM, to notify the acceptance of the request:

$$\underbrace{attach-ok, ID_{GW}, ID_{PKG_j}, N_{GW}}_{M4}, MAC_K(M4 \parallel N_{SM}) \quad (7)$$

Upon reception of (7), the SM checks its freshness and makes sure that it originates from the pre-configured ID_{GW}. If expired or not already held, the SM checks from $@Dir_i$ the existence of a certificate $Cert_{i \to j}$. Assuming it exists and is valid, the SM computes using secret values k_{SM} and S_{SM}, and public-key P_{GW}, the secret key K=K_{SM_GW}=

$$H_2(e(S_{SM}, k_{GW}P)e(k_{SM}H_1(ID_{GW}, k_{GW}P), k_{GW}P_{PKG_j}))$$

then checks the received MAC. If the verification succeeds, then SM concludes that the GW is authentic and that K_{SM_GW}=K_{GW_SM} (otherwise detects that the GW is misbehaving and stops communication with it). Then, the SM sends the following message to authenticate itself to the GW:

$$finish, MAC_K(finish, ID_{SM}, ID_{GW}, N_{GW}) \quad (8)$$

The GW verifies the authenticity of (8). If the MAC is valid, the GW concludes that the SM is authentic and that K_{GW_SM}=K_{SM_GW}. Henceforth, the SM and the GW use K_{SM_GW} to provide data source authentication. However, if they need to provide non-repudiation, they can still use their private keys. Now let prove that K_{SM_GW}=K_{GW_SM}, for simplicity we omit H_2 in the proof.

We have K_{GW_SM}=$H_2(\alpha\ \beta)$, where

$$\begin{aligned} \alpha &= e(S_{GW}, k_{SM}P)) \\ &= e(k_{GW}S_{PKG_j}H_1(ID_{GW}, k_{GW}P), k_{SM}P) \\ &= e(k_{SM}H_1(ID_{GW}, k_{GW}P), k_{GW}S_{PKG_j}P) \\ &= e(k_{SM}H_1(ID_{GW}, k_{GW}P), k_{GW}P_{PKG_j}) \end{aligned}$$

$$\begin{aligned} \beta &= e(k_{GW}H_1(ID_{SM}, k_{SM}P), k_{SM}P_{PKG_i}) \\ &= e(k_{GW}H_1(ID_{SM}, k_{SM}P), k_{SM}S_{PKG_i}P) \\ &= e(k_{SM}S_{PKG_i}H_1(ID_{SM}, k_{SM}P), k_{GW}P) \\ &= e(S_{SM}, k_{GW}P) \end{aligned}$$

From α and β, we can easily deduce that K_{GW_SM}=$H_2(\alpha\beta)$=$H_2(\beta\alpha)$=K_{SM_GW}

3) Communication between SM and SA: When a new SA is deployed in the end-customer's HAN, it first needs to authenticate itself to (associate itself with) the SM of the HAN, and also requires the authentication of the SM. To avoid associating a SA with a wrong SM (e.g., the SM of a neighboring HAN), the end-customer explicitly indicates to the SA ID_SM to which the SA needs to associate. Assume that the SA is provided with a data input interface (e.g., small keyboard), or could be connected to a PC and then be accessible through a software interface. In this case, the end-customer (SA's owner) could initialize the SA with the

 ISBN: 978-1-61208-189-2

appropriate ID_{SM} (the SM of its HAN). As a consequence, the SA and the SM could mutually authenticate and establish a secret key K_{SA_SM} for data source authentication, as described in Section IV-C2, without involving any online third party (KDC or PKG).

4) ID-based Signature Generation/Verification: We based our ID-based signature on the Hess ID-based signature scheme [14], while providing some modification to reflect the use of Certificate-less IBC. Let A belonging to domain i be the signer, B belonging to domain j be the verifier, and M the signed message. In addition, assume that B already trusts PKG_i. Moreover, S_A=$k_A S_{PKG_i} H_1(ID_A, k_A P)$ is A's private key and P_A=$< k_A P, k_A P_{PKG_i} >$ is A's public key.

A generates the signature $< R, v >$ as follows:

- Picks $k \in {F_n}^*$, and $P_1 \in {G_1}^*$, then computes

$$r = e(kP_1, P) \tag{9}$$

- Computes

$$v = H(M \parallel r) \tag{10}$$

and sets

$$R = vS_A + kP_1 \tag{11}$$

where H is a one-way hash function (e.g., SHA1)

- Outputs $\sigma_{S_A}(M)$=$< R, v >$

B verifies the signature $< R, v >$ using ID_A, P_A and P_{PKG_i} as follows:

- Computes r', where:

$$r' = e(R, P)e(-vH_1(ID_A, k_A P), k_A P_{PKG_i}) \tag{12}$$

- Accepts the signature *only and only if*

$$v = H(M \parallel r'). \tag{13}$$

Now, let prove that if the signature is verified (equation 13 held), then A really generated the signature over M using its private key S_A corresponding to ID_A and P_A, where S_A is partially generated PKG_i.

From (9), (10) and (13), we deduce that

$$r' = e(kP_1, P) \tag{14}$$

Finally, from (12) and (14), we have:

$$\begin{aligned} e(kP_1, P) &= e(R, P)e(-vH_1(ID_A, k_A P), k_A P_{PKG_i}) \\ &= e(R, P)e(-vH_1(ID_A, k_A P), k_A S_{PKG_i} P) \\ &= e(R, P)e(-vk_A S_{PKG_i}(ID_A, k_A P), P) \\ &= e(R, P)e(-vS_A, P) \\ &= e(R - vS_A, P) \end{aligned}$$

Now, we get that

$$e(kP_1, P) = e(R - vS_A, P)$$

By equality, term to term of the both pairing functions, we have

$$kP_1 = R - vS_A \Rightarrow R = vS_A + kP_1$$

Thus, we find here the initial signature as generated by A in 11.

V. Security Analysis and Comparison

Our ID-based authentication protocol for the AMI achieves secure authentication and non-repudiation. It also improves the privacy for the end-customer's. Any pair of communicating nodes of the AMI could mutually authenticate each other, then if authentic, securely exchange data either by signing them (if non-repudiation is required), or by establishing a shared secret key and generating MACs over the data (if non-repudiation is not required). Moreover, a SM and a SA could mutually authenticate without involving an online third party as the KDC in [9][10], thus, preserving the end-customer's privacy. Finally, the inclusion of a nonce value in each exchanged message protects the receiver from replay attacks. Indeed, a receiver accepts an authentic message from A as a *fresh*, *only and only if* the new nonce N_A carried in the message is grater than the last nonce received from A.

In our proposed solution, two nodes A and B are able to mutually authenticate, *if and only if*:

- They belong to the same keying domain i: In this case, both are issued a partial private key from the trusted PKG PKG_i.
- They belong to two different keying domains: In this case, they could not directly trust each other's public-key, since they do not trust the public-key of the PKG of the other domain. Each node needs to get a cross-domain certificate, issued by the PKG of its local domain to certify the public key of the PKG of the other domain. If such cross-certificate exists, then both entities could mutually authenticate, and also establish a shared key to protect their communications. If such certificate does not exist, then they will not be able to communicate.

A node X cheating on its ID or the ID of its PKG, is not able to pass the authentication phase in the three scenarios, since it could not have the appropriate private key corresponding to its identity and to its public-key. Consequently, X is not able to generate a valid signature that the other side will successfully verify, and cannot generate the same shared key as generated by the other party. As a consequence, X will be detected as misbehaving/cheating, and the communication with him will be stopped.

Our ID-based authentication protocol covers communications between the SM and the EP, a feature that is not described in the related works presented in Section II. Authentication in our protocol is very useful, especially since the end-customer is able to freely move from one EP

 ISBN: 978-1-61208-189-2

to another, in the context of SG. Also, in the new SG, the tendency is to make a separation between the utility (infrastructure provider) and the EP (electricity provider), since they are two separate entities. However, the intervention of the utility will remain, as seen in Section IV-C1. Indeed, the utility protects the SM from malicious/illegal association to any new EP.

The use of certificate-less IBC removes the key-escrow problem in [8], and thus guarantees non-repudiation. Indeed, an entity A is the only entity knowing its private key S_A. Thus, A could not repute a signature that it has generated and that could be verified using P_A, by claiming that another party (e.g., PKG_i) generated it, since PKG_i does not know S_A.

Finally, for authentication between SM/GW, and SM/SA, an authorized human intervention is performed to avoid a wrong/malicious association. Indeed, for SM/GW, an authorized field personal from the utility will indicate to the SM the identity of the associated GW, in order to send its meter readings and receive messages from utility and its EP. For SM/SA, the owner of the SA will indicate the identity of the SM to which the SA need to communicate with.

VI. Conclusion and Perspectives

Authentication is an important requirement to protect the AMI from several attacks, such as impersonation and data modification. In this paper, we present an ID-based authentication protocol for the AMI that induces low overheads, provides non-repudiation and authentication services, allows efficient key-establishment and preserves the end-customers privacy. Moreover, our solution is scalable since it considers several key management authorities. As a future work, we will evaluate the performances of the protocol, through simulation and implementation. Then, we will extend it to provide also confidentiality for communication in AMI, and enforce end-customers privacy through anonymization techniques, so that the generated metering data could not be linked to a particular SM/end-customer.

References

[1] D. P. Chassin, "What can the smart grid do for you? and what can you do for the smart grid," *Electricity Journal, Elsevier*, vol. 23, no. 5, pp. 57–63, June 2010.

[2] ASAP-SG, "The advanced security acceleration project," Open SG User Group, Online, http://osgug.ucaiug.org/utilisec/amisec/, (retrieved: January 2012), Technical Guidline, June 2010.

[3] F. M. Cleveland, "Cyber security issues for advanced metering infrasttructure," in *IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the* 21^{st} *Century*. Pittsburgh, PA: IEEE, July 2008, pp. 1–5.

[4] A. J. Menzes, P. C. V. Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. CRC Press, 1997.

[5] C. Farell and S. Adams, "Internet x.509 public key infrastructure certificate management ptotocols. rfc 2510."

[6] M. M. Fouda, Z. M. Dadlulah, N. Kato, R. Lu, and X. Shen, "Towards a light-weight message authentication mechanism tailored for smart grid communications," in *Int. Workshop on Security in Computers, Networking and Commmunications*. Shanghai, China: IEEE, April 2011, pp. 1035–1040.

[7] L. Chen, "Identity-based cryptography," International School on Foundations of Security Analyses and Design, http://www.sti.uniurb.it/events/fosad06/ (retrieved: January 2012), 2006.

[8] H. K. H. So, S. H. M. Kwok, E. Y. Lam, and L. King-Shan, "Zero-configuration identity-based signcryption scheme for smart grid," in *IEEE Int. Conf. on Smart Grid Communications*, Gaithersburg, USA, October 2010, pp. 321–326.

[9] Y. Yan, Y. Qian, and H. Sharif, "A secure and reliable in-network collaborative communication scheme for advanced metering infrastructure in smart grid," in *Wireless Communications and Networking Conference*. Caucun, Mexique: IEEE, March 2011, pp. 909–914.

[10] E. Ayday and S. Rajagopal, "Secure, intuitive and low-cost device authentication for smart grid networks," in *IEEE Consumer Communications Networking Conference*, Las Vegas, USA, January 2011, pp. 1161–1165.

[11] D. Hankerson, A. J. Menzes, and S. Vanstone, *Guide to Elliptic Curve Cryptography*. New-York: Springer-Verlag, 2004.

[12] G. Omahen, A. Souvent, and B. Luskovec, "Advanced metering infrastructure for slovenia," in *IEEE Conf. on Electricity Distribution*, Prague, Slovenia, June 2009, pp. 673–676.

[13] A. Menzes, "Introduction to pairing-based cryptography," Online, http://www.math.uwaterloo.ca/ ajmeneze/publications/pairings.pdf (retrieved: January 2012).

[14] F. Hess, "Efficient identity based signature schemes based on pairings," in *SAC02 Revised Papers from the* 9^{th} *Annual International Workshop on Selected Areas in Cryptography*. London UK: Springer-Verlag, 2003, pp. 310–324.

 ISBN: 978-1-61208-189-2

Decision Support Independence in a Smart Grid

Kendall E. Nygard, Steve Bou Ghosn, Md. Minhaz Chowdhury, Ryan McCulloch, Davin Loegering, Anand Pandey, Md. M. Khan
Department of Computer Science
North Dakota State University
Fargo, ND, USA
{Kendall.Nygard, Steve.Boughosn, Md.Chowdhury, Ryan.Mcculloch, Davin.Loegering, Anand.Pandey, Mahbuburrahman.Khan}@ndsu.edu

Prakash Ranganathan
Department of Electrical Engineering
University of North Dakota
Grand Forks, ND
prakashranganathan@mail.und.nodak.edu

***Abstract*—We describe a novel framework for designing and implementing agent based simulations of the smart electrical grid. The framework is based on two primary concepts. First, the eletrical grid system is separated into semi-autonomous microgrids, each with their own set of hierarchically organized agents. Second, models for automating decision-making in the grid during crisis situations are independently supported. Advantages of this framework are scalablity, modularity, coordinated local and global decision making, and the ability to easily implement and test a large variety of decision models. We believe that simulators based on this kind of framework will be valuable for evaluating the effectiveness and reliability of alternative methodologies for configuring automated self-healing in the grid with little human intervetion. The primary achievement of work is the software design for directly supporting decision model independence.**

Keywords-Multi-agent Multi-agent System; SmartGrid; Distributed Computing; Intelligent Systems; Self-healing.

I. INTRODUCTION

A primary objective of smart grid software architectures is to provide intelligence and communications technology to support powerful and efficient automation. A power system is exposed to faults created by natural calamity, terrorism and equipment or operator failures. Once a fault occurs in a power system, it is necessary to quickly isolate the malfunctioning components from the rest of the network to minimize outages. A power failure can range in magnitude and impact from a relatively modest curtailment to a catastrophic regional blackout. Because most power failures cannot be prevented [6], it is desirable for the Smart Grid to have self-healing capabilities that respond appropriately to disruptions when they occur, restore the power system to a healthy state, minimize consumer outages, and involve little or no manual intervention.

Due to the large scale and complexity of the Smart Grid, anticipating all possible scenarios that lead to performance lapses is difficult [7]. There is a high degree of uncertainty in accurately estimating the impact of disruptions on the reliability, availability and efficiency of the power delivery system. These uncertainties result in hesitation on the part of decision makers in committing to smart systems for grid management. We report here on research that is focused on the use of simulation models to promote trust in Smart Grid solutions in safe and cost effective ways.

Recently developed Smart Grid simulators and analysis tools include GridLAB-D and the Graphical Contingency Analysis (GCA). Both are projects developed at the U.S. Department of Energy's (DOE) Pacific Northwest National Laboratory (PNNL). GridLAB-D is a sophisticated simulator that provides detailed information of the power grid's state, including power flow, end-use loads, and market functions and interactions. The GCA is a visual analytic software tool that aids power grid operators in making complex decisions. By using human friendly visualizations and classifications of critical areas and by allowing the operators to simulate possible actions and their consequences, the tool helps human operators to analyze large amounts of data and make decisions in a reasonable amount of time. Due to the large amounts of data representing the grid status at any given time, even when aided by simulation and analysis tools, there are still limitations on how quickly human operators can make efficient decisions in near real time.

Because of the limitations of human operators in comparison with automated control, there is considerable research being done on how to fully automate control of the electrical grid by using software agents. A software agent is an encapsulated software system situated in an environment where it can conduct flexible and autonomous actions to meet its design objectives [2]. A Multi Agent System (MAS) is composed of multiple interacting intelligent agents that can sense, act, communicate and collaborate with each other. In our previous work [1], we presented guidelines for an agent-oriented smart grid simulation design. The agents in our MAS exhibit autonomy or partial autonomy, are decentralized, and have local views and knowledge. This design is related to other agent-based simulators that have been developed [4][5][16][17].

A fully automated grid relying on a multi-agent system will also presents some challenges and disadvantages along with its many advantages however. For example, developing agents able to function on par with human experts for the various scenarios that can happen in the smart grid, this will require a significant amount of research and experimentation. Relying on autonomous agents will also introduce a number of security issues. An agent could be hacked and controlled by an attacker who could manipulate the decisions and communications of the agent to perform malicious behavior. The trustworthiness of any particular agent or even the system as a whole could be called into question because of both the security risks and the general difficulty in replicating human expertise. Much must be done in order to overcome these inherent disadvantages of autonomous agent based systems.

 ISBN: 978-1-61208-189-2

Supporting knowledge bases and decision-making capabilities within individual agents is a key challenge in Smart Grid designs. In order to successfully evaluate intelligent systems that allow agents to employ appropriate decision models for self-healing scenarios that can occur in the Smart Grid, intensive testing should be done using simulators. We are carrying out extensive experimentation with our own simulator based on the discussed framework. The following types of questions are being answered through experimentation within the simulator.

- What decision model is most appropriate for handling self-healing in a particular situation in the smart grid?
- What decision model can guarantee that reliability and efficiency is maximized for a given power system?
- How can decisions be made quickly enough to avoid potential cascading failures, yet be deliberate enough to maintain high efficiency in the overall system?

Unfortunately, previous smart grid simulators, despite their valuable contributions, do not provide suitable simulation frameworks for answering these kinds of questions.

In this work, we present a simulator design aimed at directly supporting the research and experimentation required for full automation of distributed decision making in the grid. Our design assumes that the Smart Grid is naturally modularized into microgrids. A microgrid is a regional grouping of electrical generation, storage and consumption units that can be isolated from the centralized grid. microgrids typically have some ability to function autonomously if necessary. The new framework is fundamentally based on Distributed Control and Decision Model Independence. This paper is divided as follows: Section II discusses the previous research works in more details. Section III discusses the Distributed Control aspects of the simulation framework, including its advantages challenges. Section IV discusses Decision Model Independence. In section V we present our conclusions and describe future work.

II. LITERATURE REVIEW

There have been several attempts to create simulation systems for a smart grid using multi agent systems. In one of the earliest works [3], the authors created an accurate hardware simulation of a simple microgrid using MATLAB and Simulink to model the functionality of low level electrical circuits. Their agent implementation was very simple, with voltage monitoring to activate a circuit breaker and secure critical loads, with no complex decision making. Their focus was on showing that a microgrid can be managed as part of the global grid and is still able to work autonomously in an islanded mode. Their simulation of agent interaction and collaboration was not thoroughly tested, evaluated or analyzed. Their contribution was valuable in showing that the smart grid could be modularized into smaller independent units with their own agents and that those modules could work autonomously and as part of the whole.

In [10], a centralized multi-agent framework for power system restoration was designed. In [11] and [12], improvements were done to the framework while still maintaining a centralized design. Multi-Agent systems that utilize centralized control are typically not able to respond quickly enough to perform global decisions and actions in near real time. Thus, such systems fall short of being able to address critical situations like a cascading failure that could have catastrophic consequences if not dealt with promptly. In [13] and [14], new multi-agent frameworks for the power grid based on decentralized multi-agent systems are presented. The frameworks presented in [13] and [14] are decentralized multi-agent systems, but have the disadvantage of only allowing nodes to communicate with their neighbors. When nodes only acquire information from their neighbors, it greatly limits the quality of the decisions that the nodes can make, due to insufficient data. In [15], a hybrid multi agent framework that combined centralized and decentralized architectures was proposed. All of these approaches are topology dependent, with the exception of [14] and [15], which used a topology independent framework. By allowing the framework to work irrespective of the physical structure of the grid, a high level of flexibility and scalability can be obtained. Other agent-oriented Smart Grid designs are described in [4] and [5].

In [8] and [1], we describe our MAS simulator, including innovations in supporting power grid topology, dynamic agent generation, and scalability. The simulator is based on a topology independent framework that places the physical aspects of the power grid and the actual agents in separate independent layers.

A major goal of our simulator is to support the appropriate decision models for the various self-healing scenarios that can occur. This addresses the previously unmet need to allow researchers to easily configure the type of decision model used by the agents, to compare how agents perform in the same case scenario when using different reasoning processes. More importantly, we also address the need for researchers to develop their own models and easily integrate them in the simulator for testing.

The framework we present in this research work addresses many of the shortcomings of traditional centralized and decentralized schemes by utilizing a hierarchical distributed control scheme. It also supports direct comparisons among different decisions models, by implementing them in a separate independent layer from the agents. We see considerable potential for the simulator to help in the building of smarter electrical grid architecture.

III. DISTRIBUTED CONTROL

To address the large size and complexity of the smart grid, we break the problems that were handled by a centralized controller into smaller problems handled by multiple distributed controllers. It is critical to utilize a granularity that allows the units to work independently and autonomously as well as to integrate and coordinate with each other in order to guarantee that the system works efficiently as a whole.

Our design is inspired by the concept of an Intelligent Autonomous Distributed Power System (IDAPS) that was proposed by the Advanced Research Institute of Virginia Tech. [9]. An IDAPS is essentially a microgrid that contains sufficient intelligence and resources to be fully autonomous, yet function within the global grid. We specifically design microgrids so that they are capable of disconnecting themselves from the rest of the grid under certain situations and work autonomously in islanded mode [3].The intelligence in the microgrid handled by the multi-agent system associated with the microgrid and its quality depends directly on the multi-agent design employed.

A. Multi-Agent Design

In our design, we establish hierarchical relationships among the agents, where the agents on higher levels supervise

 ISBN: 978-1-61208-189-2

those in the levels below. With this arrangement, an agent can either be entirely autonomous (acts on its own) or semi-autonomous (works under direction from a supervisor) [18]. In our framework, we implement a three level hierarchical system of agents, in which the agents in the second level are supervised by the agents in the first level and agents in the third level are supervised by those in the second. In general, this means that the agents in the first level exhibit full autonomy in that they act of their own accord without direct instruction, while the agents on the second and third layer act semi-autonomously in that they receive instructions form their supervisory agents.

In our modeling of autonomous micro-grids, we use a three layered design, the two upper layers of which consist of the two hierarchical agent levels as described above. The bottom layer is a simple hardware simulation called the physical layer. The purpose of this bottom layer is to mimic the behavior of the electrical components themselves. This layer simulates devices such as relays, transformers, capacitors, power lines, consumers and generators that run on their own, with no intervention or added intelligence. This separation between the intelligent agent layers and the physical simulation allows researchers to run two basic scenarios in the simulator, one, at the base hardware level without autonomous actions like the grid would normally operate, and another, using intelligent autonomous agent support.

The top layer, called the management layer, hosts the management agents that make the high level decisions. These agents are continuously sending data that represents the status of the microgrid at a given point in time. A management agent organizes, analyzes, and parameterizes models with the data, in order to detect situations in the grid that require healing. If such a situation is detected, it creates a strategy to handle the disruption and heal the system. This strategy is expressed as a set of different roles to be performed by the middle layer agents, known as corrective behaviors. The management agent is the main decision maker in our simulation framework, but the agents in the middle layer do have a certain amount of autonomy in how they carry out the high level decisions generated by the management agent.

We refer to agents in the middle layer as distributed energy resource (DER) agents, user agents, device agents, and control agents. These agents collaborate with each other as well as report to and follow the instructions of the management agents. We describe each in turn.

1) User agents: act on behalf of consumers to ensure that businesses, organizations, homes, or other electricity consumers have the power they need.

2) DER agents: act on behalf of the DERs within the microgrid. DERs are generation sources that are independent from the main power distribution circuit. These generators are typically small companies or special consumers that also participate in the power market, such as a wind power unit or geothermal generator. DER agents act on behalf of their generators by engaging in power supply negotiations with users.

3) Device Agents: act on behalf of the individual electrical grid components (such as switches and transformers). The device agent reports the device's sensor and meter readings to the control agents. It also has the ability to perform actions such as using a relay to reroute power, or closing a circuit breaker.

4) Control agents: are in charge of monitoring a section of the grid by collecting data from all of the agents in that area. These agents carry out data fusion and are sent on to the management agent as the representation of the current status of the system.

In order for the agents in the middle layer to communicate with simulated physical components in the first layer, it is fundamental to have middleware to facilitate the communication. This is achieved by encapsulating the bottom layer within an environment agent. An environment agent contains all information about the grid topology and all device status information at any given time. Middle layer agents can query the environment agents to gain information about the physical grid. They also communicate with the environment agent when they seek to alter the behavior of the physical grid. This simulates, at an abstract level, how the agents can be integrated with smart meters and other sensors in a real grid.

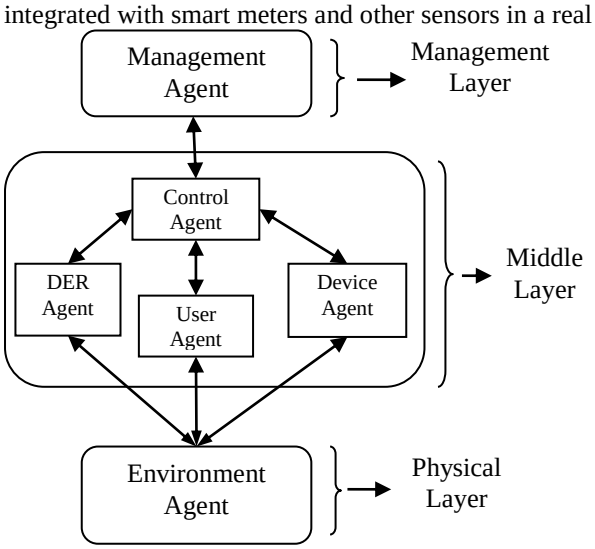


Figure 2. Graphical representation of communication between layers

Figure 1 shows a graphic representation of the communication flow between the layers, illustrating the kinds of intelligence that is integrated into each layer. High level decisions are carried out by a management agent, and the physical components with no intelligence reside at the bottom. The middle layer agents have limited intelligence.

IV. THE PRIMARY ADVANTAGES OF THIS FRAMEWORK USING DISTRIBUTED CONTROL ARE MODULARITY, SCALABILITY AND EFFECTIVE LOCAL AND GLOBAL DECISION MAKING

A. Modularity

The complexity of large systems such as the smart grid can be managed with a divide and conquer approach. As much as possible, each autonomous unit is responsible for managing and solving local problems that pertain to the unit itself. When a critical situation occurs in the unit it is simple to isolate that unit from the global grid to prevent the crisis from propagating to other sections of the grid, which happens easily for cascading failures. Modularity also allows the different agents within the microgrid to tune themselves and adapt to use the decision models and strategies that optimize resource usage and maximize efficiency for that particular microgrid, with its own topology, organization and operation.

 ISBN: 978-1-61208-189-2

B. Scalability

Separation into well-defined independent autonomous units simplifies the process of testing at the unit level. Integration testing, referring to testing how the modules coordinate and integrate together, is also facilitated. The design supports the scalability of the system, because every microgrid enjoys autonomy in that all the subordinate agents in the microgrid report to their management agents and communication is done exclusively between elected management agents in each microgrid. This means that if we verify that a number of units work properly when tested independently and also work properly together when tested through integration testing, that is good evidence that adding more units would also scale well. The main purpose of our design is to allow for easy scaling of the system. With our approach we can create a larger macrogrid by connecting several microgrids together. It is convenient for a user to create a simulation of a large grid by developing several different microgrids in the simulator, followed by connecting them together.

Once a user has generated several microgrids and established connections between them, the simulator should give him/her some flexibility on how to run those microgrids. Many microgrids could be set to run on the same computer or one microgrid per computer. Therefore, the only limits to the scalability of the system are the resources available to the user. The simulator itself shouldn't have specific limits, since different instances of it can be run to support each microgrid, and the microgrids if connected can communicate and interact with each other.

C. Local and Global Decision Making

In designing a distributed control system, it is often very difficult to coordinate effective local and global decision making. For example, in distributed systems, if agents are only able to communicate with neighboring agents, there is a severe restriction on the quality and quantity of the data that can be collected. For local decisions this restriction of data is acceptable for the decisions that are correspondingly limited in scope, but for global decisions this data restriction can easily result in ineffective and potentially harmful decisions. However, centralizing decision making and forcing the nodes to report to and be controlled by a single entity is also problematic due to excessive communication requirements. Our design combines the two extreme approaches by allowing local microgrid-based decisions to be handled locally, while still supporting certain global interactions.

Separating the extremely large and complex electrical grid system into small independent microgrids facilitates effective local decision making. These decisions are handled by the management agents contained in every microgrid. These agents constantly receive data reflecting the status of the local microgrid in near real time, and therefore are enabled to make decisions that optimize the local performance of the microgrid. However, there are circumstances under which it is desirable for microgrids to adjust their level of autonomy. This tuning of autonomy levels depends greatly on the specific decision model in effect in the management agents. We describe this in more detail in the next section.

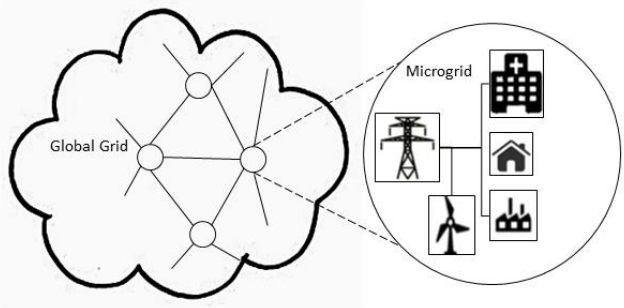


Figure 3. Local and Global Control Interactions

In the electrical grid, it is highly desirable to distribute decision-making. Our design allows each microgrid to essentially act as a single node in the larger grid. For example, If an individual microgrid requires electrical power from outside, it can communicate and negotiate with its connected neighbors establish a contract for that power. Under outage conditions, a microgrid node can island itself from its neighbors to avoid propagation of the disruption. Islanding to avoid cascading failures is described in [19]. A key advantage of a Smart Grid is the ability to access an open market for power with speed and agility [20]. This means that a microgrid node that requires additional power can access the power market, select its preferred provider based on attributes such as price or location, and then negotiate an automated contract with the provider in near real time. This ability can make significant advances in removing inefficiencies that are pervasive in the standard grid.

V. DECISION MODEL INDEPENDENCE

We envision that the true promise of the Smart Grid lies in the development of multiple types of decision models that carry out their calculations automatically and trigger actions that are appropriate to the situation with little or no human intervention. The following types of decisions are candidates for automation:

- Power rerouting. When devices or power lines fail, models that are equipped with details of the network topology and distribution costs and parameters can be charged with rerouting power along alternative pathways. The decisions must avoid exceeding the capacities of the available lines and devices, honor reliability requirements, and head off possibilities for cascading failures.
- Resource allocation. When it is critical to rapidly access new or reserve power supply sources, the decisions must consider many available combinations and prioritize them in terms of their advantages and disadvantages. Cost, transmission distance and routing options, risks and reliability, and contract terms are all factors that must be included in parameterizing these models.
- Dynamic pricing. When it is advantageous to shift power sources or limit power consumption at prescribed times to achieve cost savings, dynamic pricing models can negotiate and establish new power supply schedules at reduced cost.

An important new innovation of our simulation framework is that of supporting the decision models independently within the design. By placing the decision model agents on an independent layer, the monitoring and action-oriented agents can carry out their functions with no encumbrance from extensive special interactions and communications. This is

 ISBN: 978-1-61208-189-2

accomplished through middleware that adheres to communication standards between the decision support agents and the others.

A. Decision_Process

The management agents play a central role in the decision-making process. Management agents receive extensive near-real time data that characterizes the state of the system. The management agents are equipped with meta-level reasoning capabilities that determine if the current state of the grid is acceptable or in need of healing. In the latter case, it generates a strategy that is evaluated by calling upon decision model agents that carry out and return results aimed at corrective actions.

Multiple decision models have been designed and prototyped and are at some stage of maturity. These include the following:

- Integer linear programming. These models are capable of identifying optimal rerouting options and resource allocations. They are designed for decomposition so that local management agents can invoke only the portions of the global model that pertains to their microgrid.
- Fuzzy logic. These models are based on fuzzy set membership functions that capture degrees of fit with key resource allocation parameters, such as cost, distance, risk and reliability. The fuzzy sets drive a rule-based expert system that produces the suggested allocations. Although heuristic in nature, this type of decision model can function very quickly and easily in near-real time decision making.
- Bayesian Belief Networks. These models are based on probabilities associated with system states. These models are capable of polling the grid for additional information that forms the basis for producing posterior probabilities with enhanced accuracy.
- Market-driven pricing models. These models work within a maket economy in which energy resources are traded. This type of model includes dynamic pricing based on smart building and smart meter infrastructure, and provides an area of great promise in improving grid performance and efficiency.

In general, state variable data is made available via middleware in an Application Programming Interface (API). A key advantage of this approach is that researchers are free to readily test the models above as well as any other developed decision model. The model builder must convert data from the API to match the data types of the decision model. This method is analogous to the presentation layer in the OSI networking model.

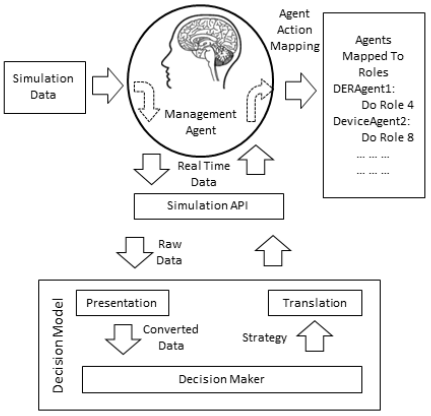


Figure 4. Decision Model Design & Decision Process in our Framework

After a decision model is invoked, it has defined a set of actions and corrective behaviors that can ultimately be carried out middle layer agents.

B. Adjustable_Agent_Autonomy

In some cases it is desirable for certain agents to be only semi-autonomous in that they make decisions only in a context controlled by a management agent. For example, a semi-autonomous user agent may have preferences and configurations that can be compromised by a decision model for the betterment of the microgrid. Fully autonomous agents must adhere to their settings regardless of decision model recommendations. . Semi-autonomous agents have a "wait" state in which they take actions to deal with a problem only upon a directive from a management agent.

Management agents can strategically choose to elevate the autonomy level of a semi-autonomous agent, based on the current decision model and system state. For example, a component of a management agent strategy to heal the system could be to promote some semi-autonomous agents to fully autonomous status either temporarily or permanently. The inherent flexibility of adjustable autonomy is a powerful capability that allows an agent-oriented system to respond to events that cannot be foreseen. It is critical that an agent-based system for a large complex system such as the Smart Grid support only agents with a high level of trust, to alleviate suspicions expressed by people that agent decisions could go awry. Adjustable, situated autonomy increases the level of trust. Thus, we believe that adjustable autonomy is an important element of decision making in the Smart Grid.

The key advantage of Decision Model Independence is the ease in experimenting with and evaluating alternative decision models. Any single scenario can be evaluated with multiple models for side-by-side comparisons.

VI. CONCLUSIONS

Distributed multi-agent control accomplishes modularity, scalability, and a balance between locally and globally effective decisions. We have discussed the advantages of an agent-based framework as a methodology for fully automating the electrical grid. By supporting decision models separate from the monitoring and action agents, alternative models can be easily evaluated. Adjustable and situated agent autonomy adds further depth and power to the design.

VII. REFERENCES

[1] Boughosn, S., P. Ranganathan, S. Salem, J. Tang, D. Loegering, and K. Nygard, Agent-Oriented Designs for a Self Healing Smart Grid, First IEEE International Conference on Smart Grid Communications, pp. 461-466, 2010.

[2] Jennings, N. and Wooldridge, M., Agent-Oriented Software Engineering, Artificial Intelligence, vol. 117, pp. 277-296, 2000.

[3] Pipattanasomporn, M. H.. Feroze, and S. Rahman, Multi-Agent Systems in a Distributed Smart Grid: Design and Implementation, IEEE Power Systems Conference and Exposition (PSCE'09), pp. 1-8, 2009.

[4] Noorian, S. , H. Hosseini, and M. Ulieru, "An Autonomous Agent-based Framework for Self-Healing Power Grid," IEEE Conference on Systems, Man and Cybernetics, pp. 1983-1988, 2009.

[5] Karnouskos, S. and T. Nass de Holanda, Simulation of a Smart Grid City with Software Agents, Third European Symposium on Computer Modeling and Simualtion, pp. 424-429, 2009.

[6] Solanki, J., S. Khushalani, and N. N. Schulz, "A Multi-agent Solution to Distribution Systems Restoration," IEEE Transactions on Power Systems, vol. 22, no. 3, pp. 1026–1034, 2007.

 ISBN: 978-1-61208-189-2

[7] Lemmon, M., G. Venkataramanan, G., and P. Chapman, Using Microgrids as a Path towards Smart Grids. Position Paper, ICEE Workshop, 2009.

[8] Nygard, K., S. Bou Ghosn, D. Loegering, M. Chowdhury, M. Khan, R. McCulloch, A. Pandey, and P. Rangnathan, Implementing a Flexible Simulation of a Self-Healing Smart Grid, International Conference on Modeling, Simulation and Visualization Methods, pp. 106-111, 2011.

[9] Rahman, S. M., Pipattanasomporn, and Y. Teklu, Intelligent DistributedAutonomous Power Systems (IDAPS), IEEE PES Annual General Meeting, Tampa, Florida, pp. 1-8, 2007.

[10] Nagata, T. and H. Sasaki, "A Multi-agent Approach to Power System Restoration," IEEE Trans. on Power Systems, vol. 17, pp. 457–462, 2002.

[11] Nagata, T., H. Fujita, and H. Sasaki, "Decentralized approach to normal operations for power system network", Proc. 13th International Conf. on Intelligent Systems Application on Power Systems, Cambridge, pp. 407–412, 2005.

[12] Momoh, J. and O. Diouf, "Optimal Reconfiguration of the Navy Ship Power System Using Agents," Proc. IEEE PES Transmission and Distribution Conf. and Exhibition, Dresden, pp. 562–567, 2006,.

[13] Nordman, M. and M. Lehtonen, "Distributed Agent-based State Estimation for Electrical Distribution Networks," IEEE Trans. on Power Systems, vol. 20, pp. 52–658, 2005.

[14] Solanki, J., S. Khushalani, and N. Schulz, "A multi-agent solution to distribution systems restoration," IEEE Trans. on Power Systems, vol. 22, pp. 1026–1034, 2007.

[15] Sutanto, D, Y. Ye, and M. Zhang, "Design of an Intelligent Self-Healing Smart Grid using a Hybrid Multi-Agent Framework", Journal of Electronic Science and Technology, vol. 9, pp. 17-22, March 2011.

[16] Hossack, J., S. Mcanhur, J. Mcdonald, J. Stokoe, and T. Cumming, A Multi-agent Approach to Power System Disturbance Diagnosis, In Proc. International Conference on Power System Management and Control,, Vol. 488, pp. 317-322, 2002.

[17] McArthur, S., E. Davidson, J. Hossack, and R. McDonald, Automating Power System Fault Diagnosis through Multi-agent System Technology, Hawaii International Conference on System Sciences, pp. 947-954, 2004.

[18] Camponogara, E. and S.N. Talukdar, Agent Cooperation: Distributed Control Applications, International Conference on Intelligent System Application to Power Systems, pp. 1-6, April 1999.

[19] Pahwa, S., A. Hodges, C. Scoglio, and S.Wood, Topological Analysis of the Power Grid and Mitigation Strategies Against Cascading Failures, IEEE International Systems Conference, 2010.

[20] Ketter, W., J. Collins, J., and C. Block, "Smart Grid Economics: Policy Guidance Through Competitive Simulation," ERIM Reprot Series, 2010.

 ISBN: 978-1-61208-189-2

Optimization of Energy and Emissions in High-Performance Grid Computing Data Centres

Mikko Majanen, Olli Mämmelä
Seamless Networking Team
VTT Technical Research Centre of Finland
Kaitoväylä 1, Oulu, Finland
Email: mikko.majanen@vtt.fi, olli.mammela@vtt.fi

***Abstract*—At an early stage of information and communications technology and high-performance computing, performance and reliability were two important factors in research and development. It was highly essential that the hardware was able to function adequately, performing strategic operations. Energy consumption was not considered as a serious topic, since the technical characteristics of hardware and software were limited and the amount of computing nodes in a computing cluster, i.e., a data centre was small. Gradually the situation has evolved a lot: nowadays there are multiple data centres located in geographically diverse locations and the software has become more complex. Modern data centres are equipped with a large amount of computing nodes having vast computing power. However, all this progress has not come without a price, since more computing power equals more total power consumption. Consequently, energy consumption has become a major topic nowadays. This work presents two algorithms for optimizing energy and emissions in high-performance grid computing, in which multiple data centres are interconnected to each other. The algorithms are validated in a simulation environment by comparing them to standard round-robin algorithm. Our simulation experiments show that the solution is able to reduce energy consumption and emissions drastically without increase in job turnaround or wait time.**

Keywords-HPC; grid computing; energy; emissions.

I. Introduction

Energy consumption is an increasingly important consideration in computing. Data centres consume substantial amounts of energy, at an increasing financial and environmental cost. In 2006, U.S. servers and data centres consumed around 61 billion kilowatt hours (kWh) at a cost of about 4.5 billion U.S. Dollars [1]. This is equal to about 1.5% of the total U.S. electricity consumption or the output of about 15 typical power plants. High energy consumption naturally causes huge environment pollution. It has been estimated that ICT, as a whole, covers 2% of world's CO_2 emissions [2].

In High-performance Computing (HPC), the ever-growing demand for higher performance seems to increase the total power consumption, even though more flops per watt are achieved. In order to provide even greater computing capabilities, HPC data centres can be interconnected to each other to form larger, federated or HPC grid data centres. The connection is implemented by using special grid software (e.g., UNICORE [3]) that manages the job submissions to all data centres belonging to the grid.

The energy consumption between the data centres may vary radically due to the different characteristics of the centres. For example, the server hardware in each centre may be different and consume different amount of energy. The centres may also locate geographically far from each other and the surrounding climate can cause large differences in the needed cooling, i.e., the Power Usage Effectiveness (PUE) [4] values between different centres may vary due to the surrounding climate. Also, since the energy sources can differ between the centres, the CO_2 emissions of the data centres may vary radically depending on the available energy sources. The differences between the data centres naturally enable optimizations regarding energy consumption and CO_2 emissions. In this paper we introduce two algorithms for selecting the data centre inside the grid in energy- and CO_2-aware manner. The performance of the algorithms is studied by simulations and the results show significant savings in energy consumption and CO_2 emissions.

The rest of the paper is organized as follows: Section II describes the related work. Section III introduces the cluster selection algorithms. Sections IV and V present the simulation model and scenario, respectively. Simulation results are presented in Section VI. Conclusion and future work are presented in Section VII.

II. Related work

As described in [5], several methods for saving energy in single HPC data centres have been studied. The methods include mainly the use of energy-efficient or energy proportional hardware, Dynamic Voltage and Frequency Scaling (DVFS) techniques, shutting down idle hardware components at low system utilizations, power capping, and thermal management. In our prior work [5], we used an energy-aware job scheduler to schedule

 ISBN: 978-1-61208-189-2

the jobs inside single data centres and shut down idle computing nodes whenever possible. We also noted that merely the choice of a different scheduling algorithm can affect the energy consumption of a data centre. In this paper we extend our scope from single HPC data centres to HPC grid data centres, and introduce two algorithms for selecting the data centre inside the grid in energy- and CO_2-aware manner.

To the best of our knowledge, there has not been much previous research that addresses the energy efficiency or CO_2 emissions of the grids from the whole grid perspective; mainly only optimizations inside a single data centre have been studied. Perhaps the most similar approach to our approach is Heterogeneity Aware Meta-scheduling Algorithm (HAMA) [6]. HAMA first selects the most energy-efficient cluster for the job based on the power consumption of the servers and the efficiency of the cooling system. Additionally, when running the job, DVFS is used to reduce the power consumption of the CPU. The simulation results show that HAMA can reduce up to 23% energy consumption in the worst case and up to 50% in the best case as compared to other algorithms (EDF-FQ, which prioritizes jobs based on a deadline and submits jobs to resource sites in earliest start time (FQ) manner with the smallest waiting time). Without DVFS, HAMA can still result in power savings of up to 21%.

Lynar *et al.* [7] have explored the effect on energy consumption by using different resource allocation mechanisms, both in a cluster and in a grid. The results show that different resource allocation methods can result in a significantly different energy usage while computing a stream of tasks. The Pre-processed Batch Auction (PPBA) and batch auctions almost always result in significantly lower energy use than a random resource allocation. By using a simple batch auction allocation method, energy consumption can be reduced up to 37.5%, and possibly even more by using the PPBA method.

Patel *et al.* [8] have presented an energy-aware policy for distributing computational workload in the Grid resource management architecture. They introduce a data centre energy coefficient that is taken into account as a policy when making allocation decisions for compute workloads. This coefficient is determined by the thermal properties of each data centre's cooling infrastructure including regional and seasonal variations. The estimated energy savings in case of three data centres located in two different time zones were large enough to give sufficient reason for the economic viability of the approach.

Shah and Krishnan [9] also analyze the climatic conditions as a means to reducing cooling energy costs. They show that dynamic optimization of the thermal workloads based on local weather patterns can reduce the environmental burden by up to 30% in their case study. Additionally, the data centre operational costs can be potentially reduced by nearly 35%. Due to the variability of fuel mixes encountered in a global grid, they also found that the use of pure energy consumption as a metric for environmental sustainability — a common practice in the ICT literature — can be erroneous.

The GREEN-NET framework [10] consists of an ON/OFF model, which includes prediction heuristics and green advice for the users and takes the decision to switch on or off the nodes, and an adapted energy efficient Resource Management System (RMS) at the grid level.

III. Optimization in the HPC Grid

The optimization algorithm in the HPC grid focuses on optimizing the scheduling process in the UNICORE middleware [3]. The scheduling process is triggered by submitting a job from the UNICORE Commandline Client, or from the UNICORE Rich Client to the UNICORE Workflow Engine. The UNICORE Workflow Engine queries a UNICORE Service Orchestrator (USO), on which cluster the job should be submitted. As a default, the USO uses round-robin algorithm for choosing the cluster. After cluster decision, the job is submitted to the RMS of the chosen cluster. The RMS takes care of executing the job according to the used scheduling algorithm, e.g., FIFO or backfilling.

In this work, we focus on reducing the energy consumption and the CO_2 emissions. The CO_2/energy related optimizations should not affect the current Service Level Agreement (SLA) or QoS agreements, or alternatively, a new green SLA [11] could be used. In HPC, there are no clear SLAs between users and data centres, but a reasonable turnaround time can be seen as sort of a QoS agreement. A possible green SLA for HPC data centres could mean that the users allow certain delay for the execution of their job. As a bonus, they will get some extra computing time for free.

For decreasing CO_2 emissions and/or energy consumption in federated HPC data centres, the optimization algorithm will be used for performing the cluster selection in CO_2/energy-aware manner. In addition to cluster selection algorithms, also energy-aware single site scheduling algorithms will be used. As depicted in Figure 1, the USO in UNICORE receives job requests coming from the users. The jobs include the requirements for the needed resources (e.g., number of nodes/cores, RAM, etc.). If the user wants to use the green SLA, it is also included in the job requirements. The grid optimization algorithm is used to select the most suitable cluster for the job and the job is subsequently submitted to the RMS of the selected cluster. The RMS uses energy-aware job scheduling algorithms to schedule the job and power off idle servers. The energy-aware job scheduling algorithms

 ISBN: 978-1-61208-189-2

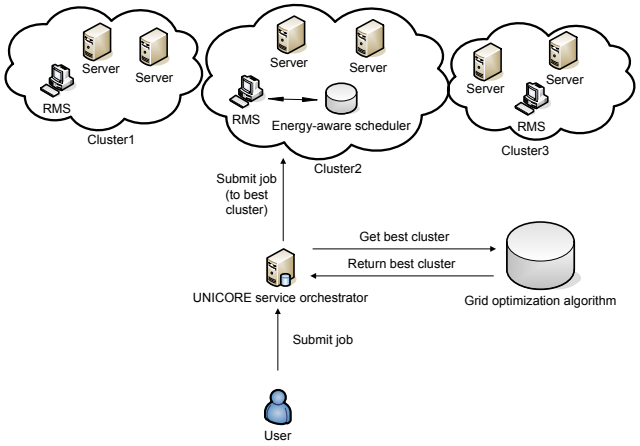


Figure 1. Job submission in a federated HPC data centre

for single site data centres were defined in our previous work [5].

A. Carbon Usage Effectiveness

Carbon Usage Effectiveness (CUE) is a sustainability metric developed by the Green Grid organization [12]. The main purpose of the metric is to address carbon emissions associated with data centres. The CUE can be calculated as follows:

$$CUE = \frac{SiteEmissions}{ICTEnergy}, \quad (1)$$

where $ICTEnergy$ is the energy consumpted by the ICT equipment in the data centre. An alternative approach for calculating the CUE is to multiply the Energy Source Coefficient (ESC) by the data centre's PUE:

$$CUE = ESC * PUE, \quad (2)$$

where PUE is a metric for defining how efficiently the power in the data centre is used, i.e., how much power is actually used by the ICT equipment and how much power is used for cooling and other equipment. ESC is defined as follows:

$$ESC = \sum ESP * EEC, \quad (3)$$

where Energy Source Percent (ESP) indicates the percentage of the energy generation source, and Energy Emission Coefficient (EEC) indicates how many kilograms of CO_2 are emitted per 1 kWh of energy. Example values of the EEC can be found in Table I [13]. By using the formulas described earlier and the values in Table I, we are able to estimate how much emissions are caused by data centres with different energy sources:

$$SiteEmissions = CUE * ICTEnergy \quad (4)$$
$$= PUE * ESC * ICTEnergy. \quad (5)$$

Table I
ENERGY EMISSION COEFFICIENT FACTORS

Generation type	Conversion factor ($kgCO_2$ per kWh)
Closed cycle gas turbine	0.360
Coal	0.910
Electricity, France interconnector	0.083
Electricity, Ireland interconnector	0.699
Non pumped storage hydro	0.0
Nuclear	0.0161
Open cycle gas turbine	0.479
Oil	0.610
Pump storage	0.0
Other	0.610

B. Algorithm/policy description

This subsection describes the functionalities of the default round-robin cluster selection algorithm, as well as the two developed algorithms for optimizations: Fastest possible (FB) that tries to minimize the waiting time, and CO_2-aware (CUE) that tries to minimize the CO_2 emissions.

1) Round-robin: Round-robin (RR) algorithm is generally used in USO for selecting the cluster. Round-robin algorithm balances the number of jobs between different clusters by always choosing the next cluster compared to the previous selection. After the last cluster, the selection is started again from the first cluster.

2) Fastest possible: Fastest possible (FB) cluster selection algorithm tries to select the cluster that could possibly execute the job with minimal waiting time. For this, the algorithm first checks if there are enough idle nodes/cores in some cluster for executing the job. If yes and the cluster's queue is also empty, the job is submitted to that cluster. If not, an estimated waiting time for the job in each cluster is calculated by using the current status of each cluster: number of nodes and cores, status of running jobs, number of jobs in the queue, and walltimes of each queued job. The cluster with the shortest estimated wait time is then selected.

The algorithm relies on the dynamic cluster properties (status of nodes and queues), which can be obtained by a single site monitoring system. Otherwise, this dynamic information is not available for the USO, so the normal cluster selection algorithms can exploit only static cluster information for the decision making.

It should be noted that the wait time can only be estimated. The walltimes of the jobs are given by the users and, in general, they are inaccurate [14], [15]. Also, the used scheduling algorithm affects in which order the jobs are executed (especially backfilling). Thus, it is possible to calculate only the maximum wait times for the jobs, not the exact wait times.

3) CO_2-aware: This algorithm tries to find the cluster with the smallest amount of estimated CO_2 emissions.

 ISBN: 978-1-61208-189-2

The CO_2 emissions of the job are CUE * ICTEnergyOfTheJob. The simplest way is to select the cluster with the smallest CUE value. This works if the clusters have significant differences in their CUE values (CUE = ESC * PUE). If there are only small differences in the CUE values, then additional estimations should be done, since the job may consume different amount of ICT energy in different clusters due to the different computing node properties (CPU, RAM, etc.), and this difference may become a greater factor than CUE for the CO_2 emissions. The ICT energy of the job can be estimated by using the job requirements (number of nodes/cores, walltime) and cluster's computing node properties (CPU, RAM, etc.) as inputs for power consumption models such as those described in [5] and [16].

However, selecting always the cluster with the least amount of estimated CO_2 emissions would cause huge load and queue on the cluster with the least CO_2 emissions. This would mean large delay for the users. Thus, some form of load balancing is needed for this algorithm. In the conducted simulations (described in the next sections), we used a queue size limit: If the queue exceeded its size limit, the job was submitted to the cluster with the second least CO_2 emissions, and so on. In the case of green SLA, the users set a certain deadline for the completion of their job. This limit can be used for load balancing: The estimated completion time for the job can be calculated as a sum of the estimated wait time and walltime of the job. If this is in the limits, the cluster can be chosen. If not, the same calculations should be made to the cluster with the second least CO_2 emissions, and so on. If the user sets too strict a time limit for the job that none of the clusters can fulfill, the job should be either denied or the cluster should be chosen by the Fastest possible algorithm.

If CO_2 emission related information is not available for the cluster, a similar kind of algorithm can be used for selecting the cluster with minimal energy consumption by replacing CUE by PUE.

IV. HPC Grid Simulation model

The simulation model has been developed with the OMNeT++ discrete event network simulator [17] and the INET Framework [18]. The design of the model is similar as in [5], except that the model is extended from a single site scenario to a federated site scenario.

Figure 2 illustrates the network topology used in the simulations. It consists of three backbone routers, three gateway routers, three data centre modules, five clients and a USO module. In this scenario, the clients send HPC job requests to the USO, which is responsible for choosing an appropriate data centre, i.e., an HPC cluster, for executing the job. The USO has been adapted for the simulation so that it is capable of using the developed optimization algorithms and making decisions based on the dynamic properties of the cluster. Normally, only static information of the cluster is available for the USO.

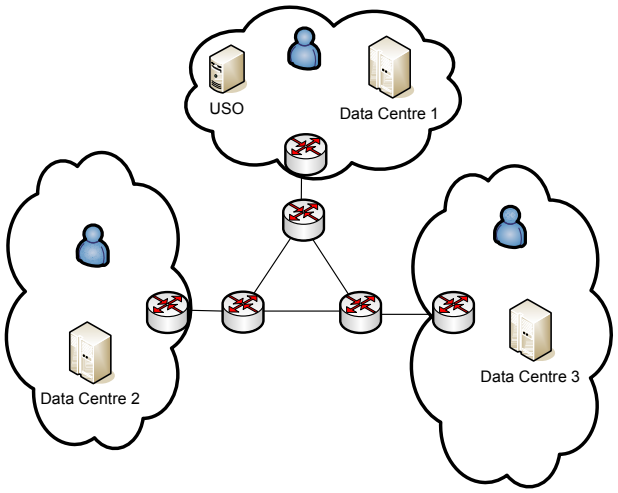


Figure 2. Network topology

For the decision making, the USO can query the status and properties of each cluster from the corresponding RMS. Once the cluster is chosen, the USO forwards the job request to the RMS of the chosen cluster. The RMS uses the policies and scheduling algorithms of the cluster to choose suitable servers for job execution. When the job execution finishes, the RMS informs the USO, which again forwards the information to the client that submitted the job for execution.

The data centre module can be seen in Figure 3, which is similar as in the single site scenario. It contains a RMS, a fixed number of servers and a router between them. The RMS handles all incoming job requests arriving to the data centre and allocates the jobs to the servers for execution according to the selected policies and algorithms. Thus, the RMS also functions as a scheduler in the simulation. The RMS supports 6 different scheduling algorithms: standard FIFO, Backfill First Fit and Backfill Best Fit algorithms and their energy-aware counterparts developed previously (see [5] for more information on these).

The RMS module includes parameters for the PUE and the CUE. By using these two values, the USO is able to select a cluster that is the most energy-efficient or produces the least amount of CO_2 emissions.

V. Simulation scenario

In this section, we describe the simulation scenario and parameters. For evaluation we consider a scenario that includes three data centres and 75 clients that are sending job requests to the USO. The simulation is stopped once 1500 jobs have been completed. During the simulation

 ISBN: 978-1-61208-189-2

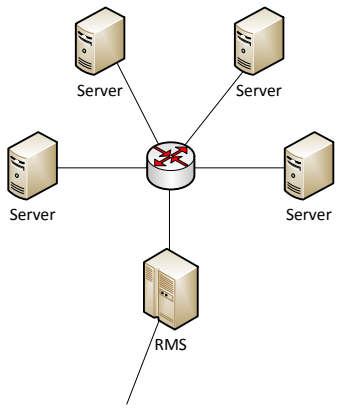


Figure 3. Data centre module

Table II
Simulation parameters

Parameter	Value
Simulation runs	10
Number of jobs	1500
Number of data centres	3
Number of clients	75
Number of gateway routers	3
Number of backbone routers	3
USO cluster selection algorithm	RR, FB, CO_2-aware
RMS scheduling algorithm	FIFO, BFF, BBF
Server memory	4 * 2 GB = 8 GB
Server cores per CPU	2
Server CPUs	2
Server CPU idle power	15 W
Server core voltage	1.2 V
Client job cores	1, 2, 4
Client job load	Uniform(30,99)
Client job nodes	Uniform(1,20)
Client job memory	Uniform(100MB, 2GB)
Client job run time	Uniform(600s, 86400s)

we measure the energy consumed by each data centre and present the obtained results in the next section. General simulation parameters are presented in Table II. Uniform(a,b) means randomly selected value according to a uniform distribution between a and b.

In Table III, we can see the parameters for the three clusters in the considered federated HPC data centre. The clusters have different characteristics, such as, the number of servers, PUE, and ESC. The energy sources (O = Oil, C = Coal, H = Hydro, N = Nuclear) for the clusters were selected so that both extreme ends in terms of ESC were represented in the simulations, while the third one represents something in the middle of them. Also, servers have different operating systems (OS) and processor architectures.

VI. Results

In all of the following figures, the algorithms are shortened as follows:

Table III
Data centre parameters

Parameter	Cluster 1	Cluster 2	Cluster 3
Servers	30	40	50
Energy source	C 50% H 20% N 30%	C 80% O 20%	O 20% H 40% N 40%
PUE	1.5	1.8	1.3
ESC	0.45983	0.85	0.12844
CUE	0.689745	1.53	0.166792
OS	Linux	Windows	Linux
CPU arch.	AMD	Intel	Intel

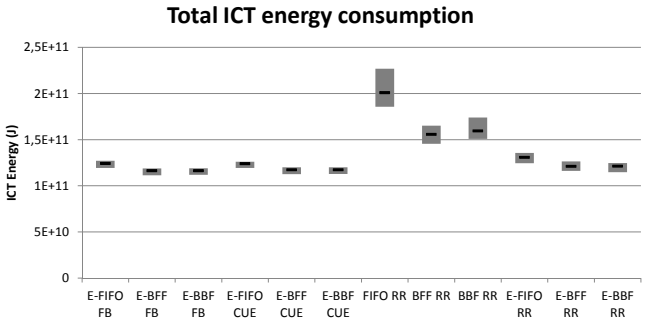


Figure 4. Total ICT energy consumption. Black lines represent the average value and the floating bars show the range of values from minimum to maximum

- FB = Fastest possible USO cluster selection algorithm
- CUE = CO_2-aware USO cluster selection algorithm
- RR = Round-robin USO cluster selection algorithm
- FIFO = First In, First Out job scheduling algorithm
- BFF = Backfilling first fit job scheduling algorithm
- BBF = Backfilling best fit job scheduling algorithm
- E-FIFO, E-BFF, E-BBF = energy-aware counterparts for the job scheduling algorithms (idle nodes are powered off whenever possible)

Figure 4 presents the total ICT energy consumption of the three clusters for different USO cluster selection and job scheduling algorithms. As can be seen, RR with normal job scheduling algorithms consumes the most amount of energy. RR with normal job scheduling algorithms represents a generally used, un-optimized algorithm combination in federated HPC data centres. Thus, it serves as a comparison point when calculating the energy savings and CO_2 emission reductions.

Figure 5 presents the energy savings achieved by using Fastest possible and CO_2-aware USO cluster selection algorithms instead of the default RR algorithm, and by using energy-aware job schedulers on each cluster. The energy-aware job schedulers are compared to their normal counterparts; for example, the first bar (E-FIFO FB) means the savings compared to FIFO RR. The last three bars present the savings when using RR but with energy-aware job scheduling. It can be seen that by using

 ISBN: 978-1-61208-189-2

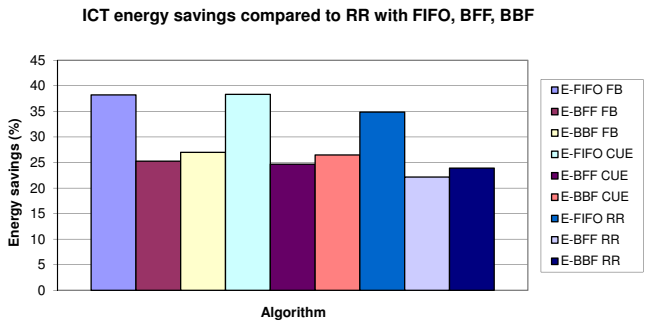


Figure 5. ICT energy savings compared to un-optimized, generally used RR with FIFO, BFF, and BBF

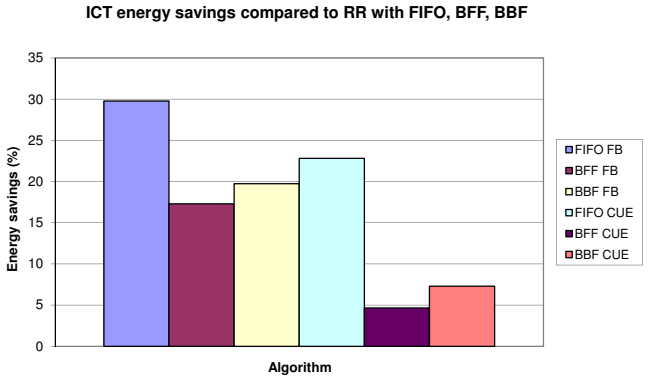


Figure 6. ICT energy savings compared to RR with normal job scheduling

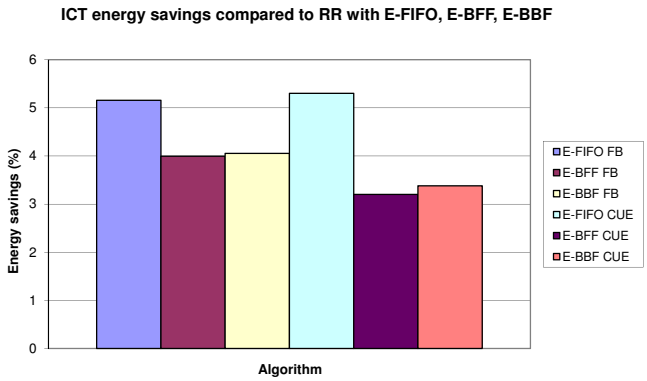


Figure 7. ICT energy savings compared to RR with energy-aware job scheduling

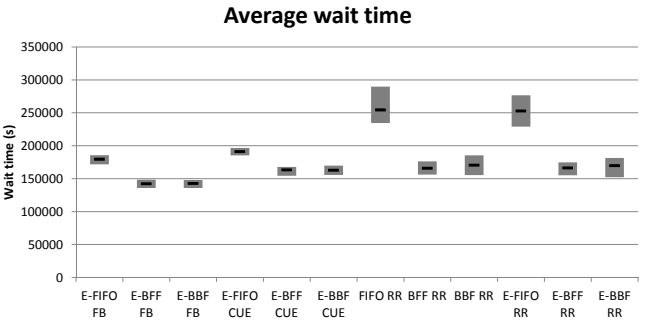


Figure 8. Average job wait times

energy-aware job scheduling, 22% to 35% energy savings can be achieved. Together with FB and CUE cluster selection, the savings are about 25% to 38%. However, if we only change the cluster selection algorithm, and keep the normal job scheduling algorithms, we can see from the Figure 6 that with FB we can save 17% to 30 %. Since the cluster selection is performed before job scheduling, we can say that about 8 % of the total savings are due to the energy-aware job scheduling, while the rest is due to the FB cluster selection. When comparing to RR with energy-aware job scheduling (as depicted in Figure 7), we can see that FB and CUE cluster selection algorithms can save additionally about 3% to 5%. For the explanation, we have to take a look at the jobs' average wait and turnaround times and the simulation duration.

Figure 8 presents the average wait times of the jobs in case of different USO cluster selection and job scheduling algorithms. As can be seen, the average wait time is clearly shorter with the FB USO algorithm. The CUE USO algorithm with backfilling has about the same average queuing time as RR, even though RR with FIFO clearly has the longest waiting time. Also, there are basically no differences between RR with energy-aware and normal job scheduling. This is true also in general, as reported in [5]: energy-aware job scheduling does not cause significant increase in wait time.

Figure 9 depicts the simulation duration, i.e., how long a time it took to execute all the 1500 submitted jobs. The graph shows the same as Figure 8: because the wait times are longer with RR USO cluster selection, also the simulation duration is longer.

Figure 10 presents the average job turnaround times in case of different scheduling algorithms. The story is the same as in previous figures: RR is slower due to the longer queuing time.

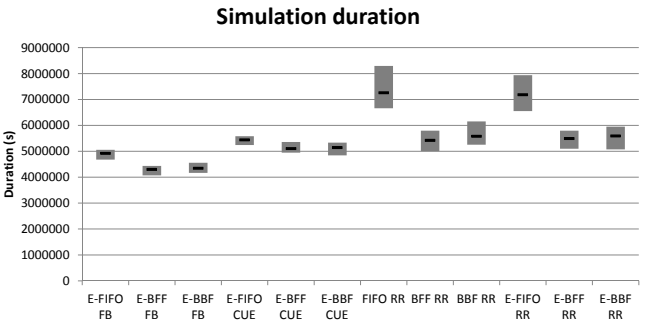


Figure 9. Average simulation duration

 ISBN: 978-1-61208-189-2

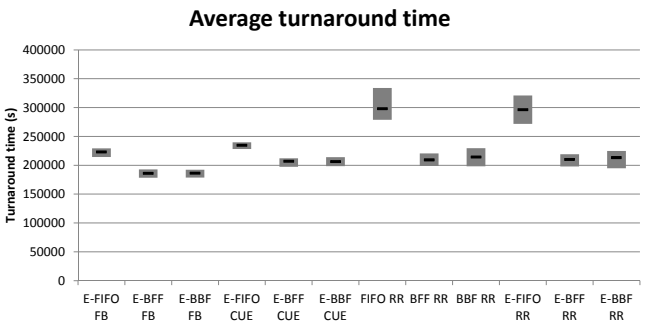


Figure 10. Average job turnaround time

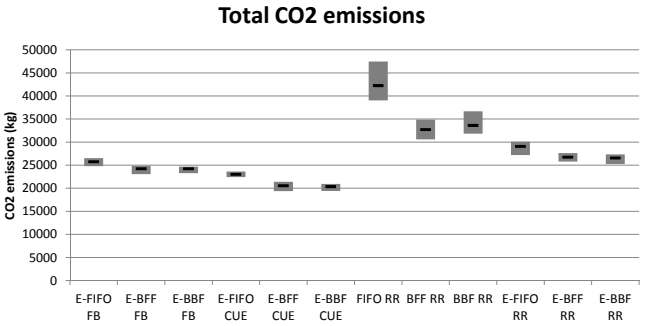


Figure 11. Total CO_2 emissions

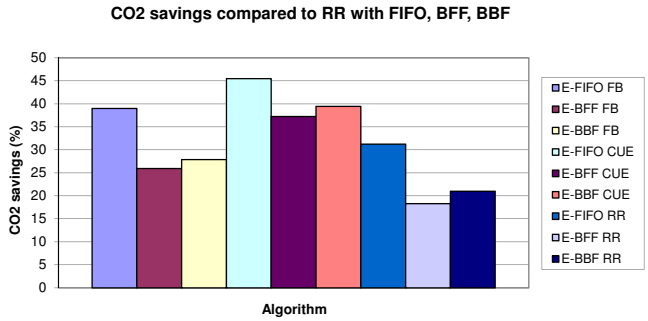


Figure 12. CO_2 savings compared to RR with FIFO, BFF, and BBF

Based on the results above, we can conclude that RR cluster selection with normal job scheduling algorithms can be very inefficient in terms of energy. This is because RR only balances the number of jobs among the clusters. It does not take into account the differences in the clusters (e.g., number of nodes/cores) or the differences in the submitted job characteristics (e.g., number of nodes/cores, walltime estimate). This can lead to a situation where one cluster is over utilized with many jobs waiting in the queue, while the other clusters can be under utilized at the same time, with nodes running idle. The energy-aware job schedulers (E-FIFO, E-BFF, E-BBF) power off the idle nodes whenever possible, and this is why a substantial amount of energy can be saved. On the other hand, the FB cluster selection algorithm inherently takes into account the differences in the clusters and submitted jobs: it always selects the cluster with the estimated minimal wait time, and thus balances the utilization between the clusters. Then fewer nodes are running idle and energy is saved.

Figure 11 presents the total CO_2 emissions of the federated HPC data centre. As can be seen, RR with normal job scheduling causes the largest CO_2 emissions. Using energy-aware job scheduling reduces the emissions due to the reduced energy consumption. Using FB cluster selection reduces the energy consumption still a bit more due to the better load balancing among clusters, and thus the CO_2 emissions are also smaller. CUE cluster selection algorithm favours the cluster with the best CUE value, i.e., least amount of CO_2 emissions, and hence achieves the greatest savings in CO_2 emissions, about 37% to 45% compared to RR with normal job scheduling. The CO_2 savings are depicted in Figure 12 as percentages.

VII. Conclusion and Future Work

The results show that the generally used round-robin cluster selection algorithm can lead to unbalanced utilizations among clusters. This can be very inefficient in terms of energy consumption and CO_2 emissions. Using energy-aware job scheduling to power off idle computing nodes whenever possible greatly enhances the energy-efficiency. Load can also be balanced by replacing round-robin cluster selection by the Fastest possible selection algorithm. This leads to energy savings due to the better utilization of clusters and shorter wait times. Using both energy-aware job scheduling and FB cluster selection simultaneously leads to greater energy savings than using only one of them. The greatest CO_2 emission savings can be achieved by using CUE cluster selection algorithm to favour the cluster with least CO_2 emissions. The actual savings in each case depends on the cluster and job characteristics. In these simulations, for example, the energy sources were chosen so that one cluster had rather small CUE, another one rather big CUE, while the third one was something between them. With smaller differences in CUE, also the possible savings in CO_2 emissions would be smaller.

Based on the simulation results presented above, we propose to use FB cluster selection algorithm for the jobs without green SLA, since it leads to energy and CO_2 emission savings due to the better utilization of the clusters, and to better QoS due to the shorter wait time. For the jobs with green SLA, we propose to use the CUE cluster selection algorithm, since it can lead to even greater CO_2 emission savings than FB, while keeping the QoS (in terms of time) at the user specified level. It can be used also without green SLA, if some

 ISBN: 978-1-61208-189-2

other parameter (e.g., queue size limit) is used for load balancing to prevent excessive load on the "greenest" cluster.

Previous research in the energy-efficiency of HPC grid computing has mainly focused on performing optimizations inside a single data centre. This work presented a global view by taking into account the whole grid: the characteristics of the data centres, compute nodes and the computing hardware. The most comparable approach to our work is HAMA, described in [6]. The results of HAMA are similar to our approach: energy savings are between 23% and 50%.

Our future topics include building a high-performance computing test bed and testing our proposed solution in a laboratory environment. In the simulation studies presented in this paper, the PUE was assumed to be constant. However, PUE is not static in the long term. Instead, it changes over time as a function of outside temperature, for example. Future work would be to investigate the impact of dynamic PUE on the energy consumption and explore how the proposed solution is able to cope with it.

Acknowledgment

This work was supported by EU FP7 project FIT4Green [19]. The authors would like to thank all the colleagues working in the project. Special thanks to André Giesler from Jülich Supercomputing Centre for his comments.

References

[1] Y. Liu and H. Zhu, "A survey of the research on power management techniques for high-performance systems," *Softw. Pract. Exper.*, vol. 40, pp. 943–964, October 2010.

[2] Gartner. (2012, Jan.) Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO_2 Emissions. [Online]. Available: http://www.gartner.com/it/page.jsp?id=503867

[3] D. Erwin and D. Snelling, "UNICORE: A Grid Computing Environment," in *Euro-Par 2001 Parallel Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2001, vol. 2150, pp. 825–834.

[4] C. Belady, "Green Grid Data Center Power Efficiency Metrics: PUE and DCIE," Green Grid, Tech. Rep., 2008.

[5] O. Mämmelä, M. Majanen, R. Basmadjian, H. De Meer, A. Giesler, and W. Homberg, "Energy-aware job scheduler for high-performance computing," *Computer Science - Research and Development*, pp. 1–11, 2011. [Online]. Available: http://dx.doi.org/10.1007/s00450-011-0189-6

[6] S. K. Garg and R. Buya, "Exploiting Heterogeneity in Grid Computing for Energy-Efficient Resource Allocation," *Seventeenth Annual International Conference on Advanced Computing and Communications (ADCOM 2009)*, 2009.

[7] T. Lynar, R. Herbert, Simon, and W. Chivers, "Reducing grid energy consumption through choice of resource allocation method," *2010 International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*, May 2010.

[8] C. Patel, R. Sharma, C. Bash, and S. Graupner, "Energy aware grid: Global workload placement based on energy efficiency," Hewlett Packard, HP Laboratiories Palo Alto, Tech. Rep., November 2002.

[9] A. J. Shah and N. Krishnan, "Optimization of global data center thermal management workload for minimal environmental and economic burden," *IEEE Transactions on Components and Packaging Technologies*, vol. 31, no. 1, pp. 39–45, March 2011.

[10] G. Da Costa, J.-P. Gelas, Y. Georgiou, L. Lefevre, A.-C. Orgerie, J.-M. Pierson, O. Richard, and K. Sharma, "The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems," *HPPAC 2009 : High Performance Power Aware Computing Workshop in conjunction with IPDPS 2009*, May 2009.

[11] S. Klingert, T. Schulze, and C. Bunse, "GreenSLAs for the eco-efficient management of data centres," in *2nd International Conference on Energy-Efficient Computing and Networking 2011 (E-energy 2011)*, New York, NY, USA, 2011.

[12] C. Belady, D. Azevedo, M. Patterson, J. Pouchet, and R. Tipley, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric," Green Grid, Tech. Rep., December 2010.

[13] RealtimeCarbon.org. (2012, Jan.) CO_2 conversion factors. [Online]. Available: http://www.realtimecarbon.org/resources/RealtimeCarbonMethodology.pdf

[14] W. Cirne and F. Berman, "A comprehensive model of the supercomputer workload," in *IEEE International Workshop on Workload Characterization, WWC-4. 2001*, dec. 2001, pp. 140–148.

[15] C. Bailey Lee, Y. Schwartzman, J. Hardy, and A. Snavely, "Are user runtime estimates inherently inaccurate?" in *Job Scheduling Strategies for Parallel Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3277, pp. 253–263.

[16] R. Basmadjian, N. Ali, F. Niedermeier, H. De Meer, and G. Giuliani, "A methodology to predict the power consumption of servers in data centres," in *Proc. of the ACM SIGCOMM 2nd Int'l Conf. on Energy-Efficient Computing and Networking (e-Energy 2011)*. ACM, 2011.

[17] OMNeT++. (2012, Jan.). [Online]. Available: http://www.omnetpp.org

[18] INET Framework. (2012, Jan.). [Online]. Available: http://inet.omnetpp.org/

[19] FIT4Green project. (2012, Jan.) FIT4Green: Energy aware ICT optimization policies. [Online]. Available: http://www.fit4green.eu

 ISBN: 978-1-61208-189-2

Reducing Power Consumption Using the Border Gateway Protocol

Shankar Raman*, Balaji Venkat‡ and Gaurav Raina‡
India-UK Advanced Technology Centre of Excellence in Next Generation Networks
**Department of Computer Science and Engineering, ‡ Department of Electrical Engineering*
Indian Institute of Technology Madras, Chennai 600 036, India
Email: mjsraman@cse.iitm.ac.in, balajivenkat@tenet.res.in, gaurav@ee.iitm.ac.in

***Abstract*—In this paper, we propose a framework to reduce the aggregate power consumption of the Internet using a collaborative approach between Autonomous Systems (AS). We identify the low-power paths between the AS and then use Traffic Engineering techniques to route packets along the paths. Such low-power paths can be identified by using the available power-to-bandwidth (PWR) ratio as an additional constraint in the Constrained Shortest Path First algorithm. For re-routing the data traffic through these low-power paths, the inter-AS Traffic Engineered Label Switched Path that spans multiple AS can be used. Extensions to the Border Gateway Protocol can be used to disseminate the PWR ratio metric among the AS thereby creating a collaborative approach to reduce the power consumption. Since calculating the low-power paths can be computationally intensive, a graph-labelling heuristic is also proposed. This heuristic reduces the computational complexity but may provide a sub-optimal low-power path. The feasibility of our approaches is illustrated by applying our algorithm to a subset of the Internet. The techniques proposed in this paper for the inter-AS power reduction require minimal modifications to the existing features of the Internet.**

Keywords-Border Gateway Protocol; Autonomous systems; Traffic engineering.

I. INTRODUCTION

Estimates of power consumption for the Internet predict a 300% increase, as access speeds move from 10 Mbps to 100 Mbps [2]. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering approaches [3].

Low-power router and switch design aim at reducing the power consumed by hardware components such as transmission link, lookup tables and memory. In [7], it is shown that the link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence, the authors suggest having more line cards and fully utilize them. Operating at full throughput will lead to less power per bit. Therefore, larger packet lengths will consume lower power. The two important components that have received attention for high power consumption are static and dynamic RAM-based buffers (SRAM, DRAM) and Ternary Content Addressable Memories (TCAM). A 40 Gb/s line card would require more than 300 SRAM chips and consume 2.5 kW [1]. Some variants of TCAMs have been proposed for high speed lines with reduced power consumption [10]. But these schemes cannot scale forever.

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [9], a subset of IP router interfaces are put to sleep, using an Energy Aware Routing (EAR) after calculating shortest path trees of the network from each router. Such a technique is useful in setting up paths within an Autonomous System (AS). In [5], the authors provide a way to introduce hardware standby primitives and apply traffic engineering methods to coordinate and reduce power consumption under given network operational constraints. Power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 W. Hence, instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption. A detailed review on energy efficiency of the Internet is given in [6].

Multilayer traffic engineering based methods make use of parameters such as resource usage, bandwidth, throughput and Quality of Service (QoS) measures, for power reduction. In [13], an approach for reducing intra-AS power consumption for optical networks using Djikstra's shortest path algorithm is proposed. The input assumes the existence of a network topology for constructing an auxiliary graph. This topology is easy to obtain for intra-AS scenario. Traffic is then re-routed through the low-power optimized links.

We propose a collaborative approach that uses inter-AS power reduction. Multi-Protocol Label Switching (MPLS) label switched paths that traverse multiple AS carry traffic from a head-end to a tail-end. AS use the Border Gateway Protocol (BGP) for exchanging routing and topology related information. One of the attributes of BGP namely, AS-PATH-INFO is used to derive the topology of the Internet at the AS level. The Constrained Shortest Path First algorithm (CSPF) uses the AS level topology with available power-to-bandwidth (PWR) ratio as a constraint, to determine the low-power path from the head-end to the tail-end. The PWR ratio can be exchanged among the collaborating AS using BGP. Explicit routing can be achieved between the head and tail-ends through the low-power paths connecting the AS using inter-AS Traffic Engineered Label Switched

 ISBN: 978-1-61208-189-2

Path (TE-LSP) that span multiple AS. Since calculation of such low-power paths can be computationally intensive certain heuristics may be needed to reduce the computation time. A graph-labelling heuristic is proposed to reduce the computation time, which may lead to sub-optimal low-power paths. We illustrate our approaches by applying it to a subset of the Internet topology.

The uniqueness of our approach is that it can be used for inter-AS power reduction and requires cooperative effort from Internet Service Providers (ISP). Further, we use BGP the existing protocol in the Internet not only for detecting the topology but also to exchange power information and then construct low-power path.

The rest of the paper is organized as follows: In Section II, we discuss in detail the pre-requisites for the algorithm. Section III introduces the proposed technique for calculating the low-power path. We also show that by using a graph-labelling technique, we can reduce the computational complexity of the low-power path algorithm, but may obtain a sub-optimal low-power path. In Section IV, we discuss the implementation issues. We present our conclusion and future work in Section V.

II. Pre-requisites for the proposed method

In this section, we discuss the pre-requisites for the implementation of the proposed scheme.

A. Constructing network topology using BGP strands

The inter-AS topology can be modelled as a directed graph $G = (V, E, f)$ where the vertices (V) are mapped to AS and the edges (E) map the link that connect the neighbouring AS. The direction (f) on the edge, represents the data flow from the head-end to the tail-end AS. To obtain the inter-AS topology, the approach proposed in [12] is used. In this approach, it is shown that a sub-graph of the Internet topology, can be obtained by collecting several prefix updates in BGP. This is illustrated in Figure 1 which shows the different graph strands of an AS recorded from the BGP packets. Each vertex in this graph is assigned a weight according to the available power-to-bandwidth (PWR) ratio of the AS, as seen by an Autonomous System Border Router (ASBR) that acts as an entry point. Figure 2 shows the merged strands forming the topology sub-graph where the weight of the vertices are mapped to the ingress edges. A reference AS level topology derived from 100 strands of AS-PATH-INFO received by an AS in the Internet is presented in Figure 3. For a detailed discussion on completeness of Internet topology information using BGP refer to [8], [11]. Any other algorithm that gives a complete AS topology could also be used.

B. PWR ratio calculation

In the topology sub-graph, each AS shares its PWR ratio. To calculate this ratio we need the available power and maximum bandwidth with an ASBR.

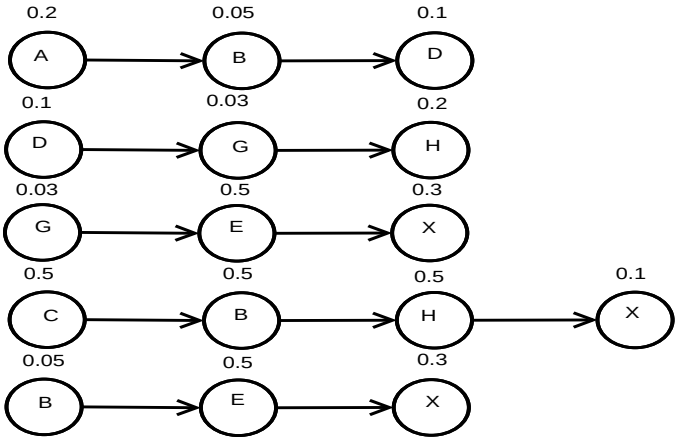


Figure 1. Strands obtained from BGP updates, vertices A,B,C,D and G are the head-end AS; D,H and X are the tail-end AS. The vertex weights represent the PWR ratio of an AS, and the link direction shows the next AS hop.

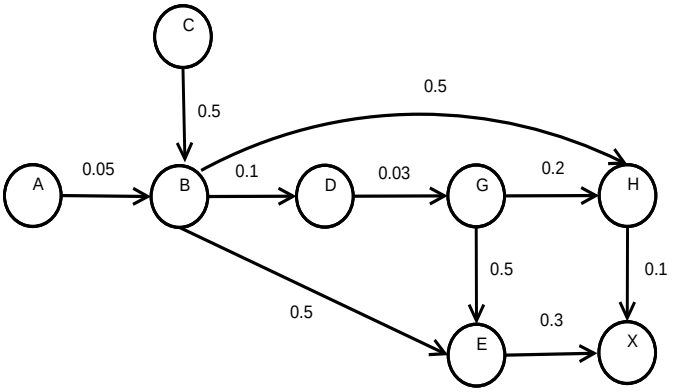


Figure 2. Strands combined to get the Internet topology. The PWR ratio is mapped to the ingress link of the ASBR.

The entry point to the AS is through ASBRs that advertise the prefixes reachable through the AS. Hence, the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We obtain the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. These can be obtained by using any of the intra-AS power calculation techniques. The average available power is obtained by subtracting the consumed power from the maximum power rating, summing the values for all the routers and then dividing the result by the number of routers. Other alternatives include using a weighted average depending on the category of the router advertising the consumed power, or to take the average or sum of the maximum power rating of all the routers within an AS. The average available power is divided by the maximum bandwidth available at each of the ASBR's egress link. This step is necessary as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the bandwidth available in the ASBR's egress

 ISBN: 978-1-61208-189-2

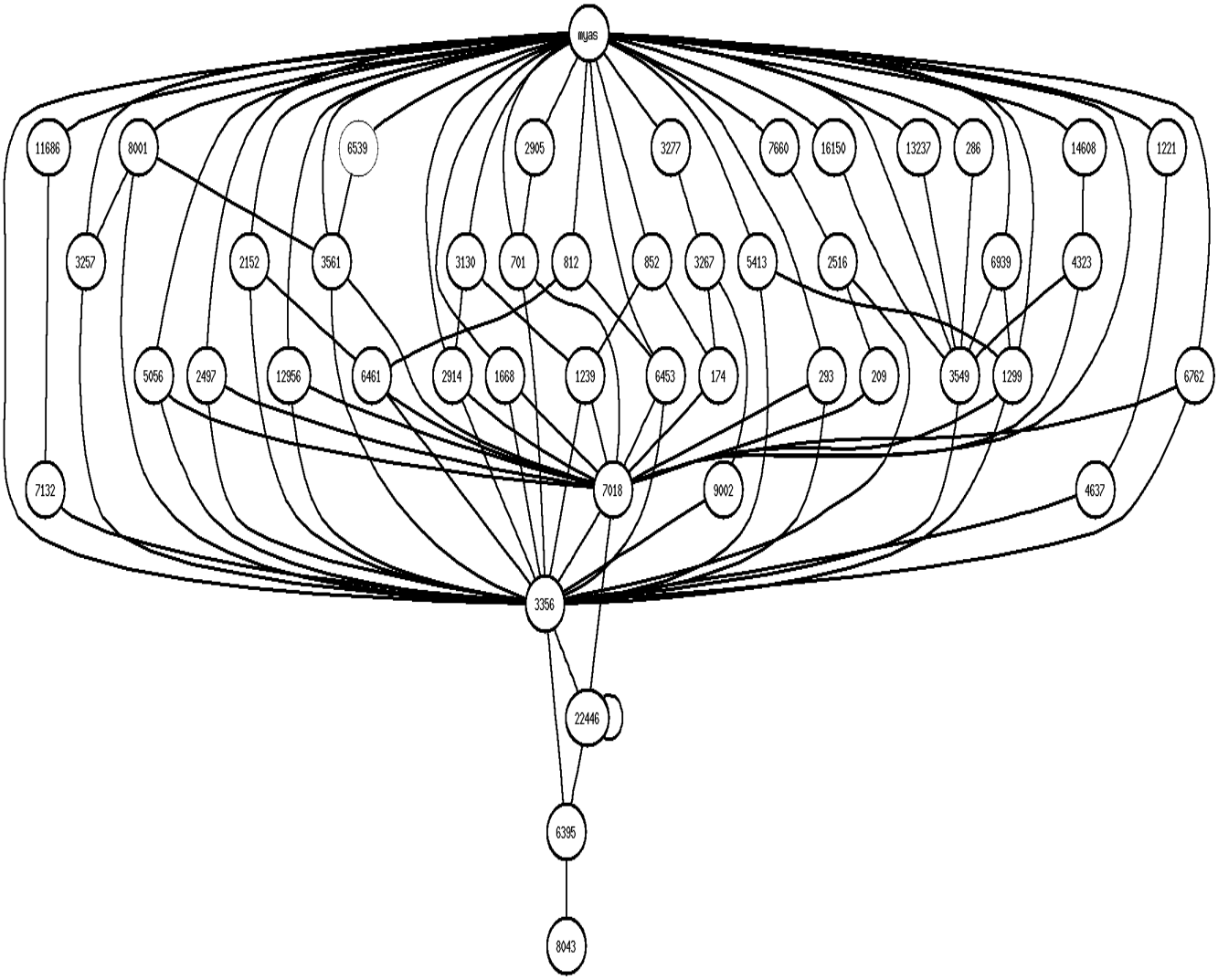


Figure 3. Internet topology graph derived from 100 strands of AS-PATH-INFO attribute by an AS through an ASBR. The top-most node (myas) represents the head-end and the bottom-most node (8043) represents the tail-end AS.

links. Simple Network Management Protocol can be used to extract this power information [4].

The highest available bandwidth amongst the egress links of the ASBR is used as the denominator in the PWR ratio computation. This PWR ratio must be computed and disbursed much ahead of time before the inter-AS TE-LSP explicit path is computed using the CSPF algorithm. The correctness of this ratio is of importance to compute the inter-AS TE-LSP route through the low-power AS. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, it is possible that an head-end AS might see different PWR ratios for an intermediate AS.

As an illustration, consider an AS X which is one of the AS in the vicinity of another AS Y. Let this ASBR of X have 3 egress links denoted as $E(1)$, $E(2)$ and $E(3)$, and 2 ingress links labelled $I(1)$ and $I(2)$. We now calculate the PWR ratio for $I(1)$ and $I(2)$. Assume that the routers in X have average available power of 200 kW/hour. From Figure 4 we can calculate the PWR ratio for $I(1)$ and $I(2)$ as 200 kW/$(60 * 60 * 1.5$ Gb$) = 3.7037 * 10^{-8}$. We could scale this to 0.37037. This ratio is a mapping function defined for each of the ingress link of the ASBR of an AS. Note the absence of ingress link for the head-end AS.

The PWR ratio can then be advertised to the other neighbouring AS through the control plane using BGP extensions. BGP ensures that the information is percolated to other AS. On receipt of this PWR ratios by the AS at the far-end of the Internet, the overall AS level topology can be constructed. Note that view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in

 ISBN: 978-1-61208-189-2

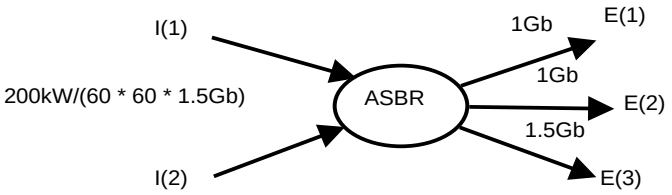


Figure 4. Calculation of PWR ratio by an ASBR of an AS. The I's are ingress links and E's are egress links. 200 kW/hour is the average available power in the AS. 1.5 Gb is the maximum available ASBR egress link bandwidth.

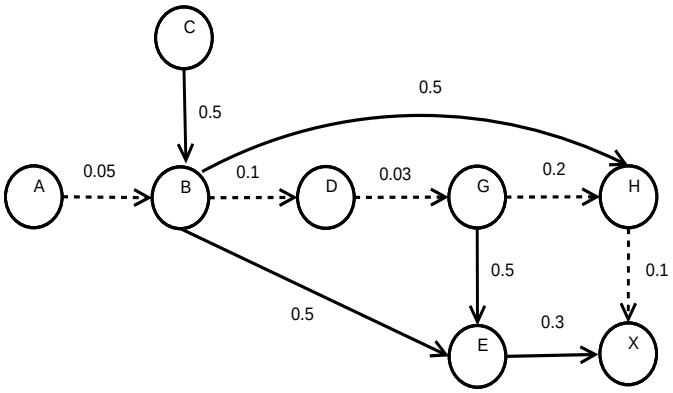


Figure 5. Dotted lines represent low-power path but has a longer number of hops than the shortest path.

Figure 2 are obtained by using such a mapping function on the ingress links.

C. Explicit routing using TE-LSPs

The head and tail-ends may reside in different AS and the path could span multiple intervening AS. To generate this path we can use Traffic Engineered Label Switched Paths (TE-LSPs). TE-LSPs can influence the exact path (at the AS level) for the traffic and this path can be realized by providing a set of low-power consuming AS to a protocol like Resource Reservation Protocol (RSVP). RSVP-TE then creates TE-LSPs or tunnels, using its label assigning procedure. The routers use these low-power paths created by the explicit routing method rather than using the conventional shortest path algorithm. This influences the exclusion of a number of high power AS on the path from the head-end to the tail-end AS. For example, the dotted line in Figure 5 represents the explicit route that is chosen by making use of such TE-LSPs from head-end AS "A" to the tail-end AS "X". Note that if the metric used is the number of hops, then the route chosen could be different.

III. Low-power paths

In this section, we present the low-power path calculation algorithm. The algorithm consists of two sub-algorithms: the first algorithm is executed by all the ASBRs in the network and the other by all the Path Computation Elements (PCEs) in their respective AS. PCEs have been proposed by the Internet Engineering Task Force (IETF) for path computation activities. We can use the existing PCE architecture for our algorithm. The algorithms for the ASBRs and PCEs are given as Algorithm 1, 2 and 3.

Algorithm 1 ASBR low-power path algorithm

Require: Weighted Topology Graph T=(AS, E, f)

```
1: Begin
2: if ROUTER == ASBR then
3:    /* As part of IGP-TE */
4:    Trigger exchange of available bandwidth on bandwidth change, to the AS internal neighbours;
5:    BEGIN PARALLEL PROCESS 1
6:    while PWR ratio changes do
7:       Assign the PWR ratio to the Ingress links;
8:       Exchange the PWR ratio with its external neighbours;
9:       Exchange the PWR ratio with AS's (internal) ASBRs;
10:   end while
11:   END PARALLEL PROCESS 1
12:   BEGIN PARALLEL PROCESS 2
13:   while RSVP packets arrive do
14:      Send and Receive TE-LSP reservations in the explicit path;
15:      Update routing table with labels for TE-LSP;
16:   end while
17:   END PARALLEL PROCESS 2
18: end if
19: End
```

A. Illustration

We illustrate the technique with a simple example. Consider the AS level topology sub-graph shown in Figure 5 constructed using the strands shown in Figure 1. The PWR ratio calculated at an ASBR is assigned to the ingress link. AS "H" has two edges coming into it: one from "B" and the other from "G". Note that the power metrics for the two strands are different. "G" to "H" is lower than that of "B" to "H". This means that the lower power metric into "H" is better if the path from "G" to "H" is chosen rather than "B" to "H". The dotted lines in Figure 5 represent low-power path.

To construct a path with "A" as the head-end and "X" as the tail-end in the AS level topology the paths "A", "B", "H", "X" and "A", "B", "E", "X" have the same number of hops. However by using CSPF with the PWR ratio as the constraint, the path "A", "B", "D", "G", "H", "X" is power efficient. The routing choice will depend on the reservation of the bandwidth on this path. If available bandwidth exists to setup a TE-LSP, then the explicit path

 ISBN: 978-1-61208-189-2

Algorithm 2 PCE low-power path algorithm

Require: Weighted Topology Graph T=(AS, E, f)
Require: Source and Destination for inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: if ROUTER == PCE then
3:    Calculate the shortest paths from the head-end to the
      tail-end using CSPF with PWR ratio as the metric;
4:    if no path available then
5:       Signal error;
6:    end if
7:    if path exists then
8:       Send explicit path to head-end to construct path;
9:    end if
10:   Continue passively listening to BGP updates to update
      T=(AS, E, f);
11: end if
12: End
```

"A", "B", "D", "G", "H", "X" is chosen. The Resource Reservation Protocol (RSVP) adheres to its usual operation and tries to setup a path. If bandwidth is not available in the low-power path thus calculated, then we may fall back to other shortest paths, provided there is available bandwidth. The low-power path algorithm given as Algorithm 2 is executed by the PCE. Algorithm 1 prepares the topology and feeds it as input to the PCE as a weighted topology graph.

Using the CSPF algorithm to calculate the route from source to destination could be time consuming for large networks. But the topology is dynamically updated and hence the computation of the shortest path can be triggered based on need. We now give a heuristic method based on graph-labelling that reduces the computation time but could trade-off the low-power path.

B. Equivalence class with total ordering

The heuristic is based on avoiding high PWR ratios by partitioning the weighted links into equivalence classes based on a range of PWR values. For each partition a label is applied such that each link in the partition has the same label. A total ordering relationship is then defined on the equivalence class. The heuristic starts including partitions with minimum label value iteratively until we get a connected component, which includes the head-end and tail-end AS. We apply the CSPF algorithm with the weights as label values on this sub-graph to obtain the low-power path. The modified algorithm which uses this scheme is given as Algorithm 3. It should be noted that this algorithm could provide sub-optimal power paths as the intermediate steps carry incomplete Internet topology information.

Algorithm 3 PCE low-power path algorithm with graph labelling

Require: Weighted Topology Graph T=(AS, E, f)
Require: Source and Destination for inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: if ROUTER == PCE then
3:    Group the links into "N" partitions with a label for
      each partition depending on the PWR ratio
4:    Sort the labels in ascending order.
5:    repeat
6:       Include the links that have the least label value;
7:       Remove the partition with this label;
8:    until there is a path from the head-end to tail-end AS
9:    Calculate the low-power path using labels from the
      head-end to the tail-end using CSPF ;
10:   if no path available then
11:      Signal error;
12:   end if
13:   if path exists then
14:      Send explicit path to head-end to construct path;
15:   end if
16:   Continue passively listening to BGP updates to update
      T=(AS, E);
17: end if
18: End
```

C. Illustration of graph labelling

We briefly illustrate the graph-labelling algorithm in Figure 6. In this figure, the links are categorized into three partitions based on the PWR ratio. PWR ratio less than 0.1 are labelled as "G", between 0.1 to 0.3 are labelled as "Y" and the rest "R". The total ordering is defined as "G" $<$ "Y" $<$ "R", where the "G" links have low PWR ratios than the "Y" links. The path could be established through the AS that has "G" as the ingress link; the path being 1245, 1339, 34234, 23411 and 16578.

IV. Implementation notes and Discussion

In this section, we present some notes on feasibility of implementation of our scheme in a live network.

First, the requested bandwidth should be available on the low-power path, but the CSPF algorithm is run with multiple constraints, one of which is the bandwidth requirement for the flows to be transported through the TE-LSP. The PWR ratio can then be applied to the available paths thus computing the low-power paths. Second, as we are using traffic engineering with link state routing protocols, there is a reliable flooding process that gets triggered when updates about the change in characteristic arise. We propose addition of some attributes with no change to the protocol implementation. There may be a time lag when the far ends

 ISBN: 978-1-61208-189-2

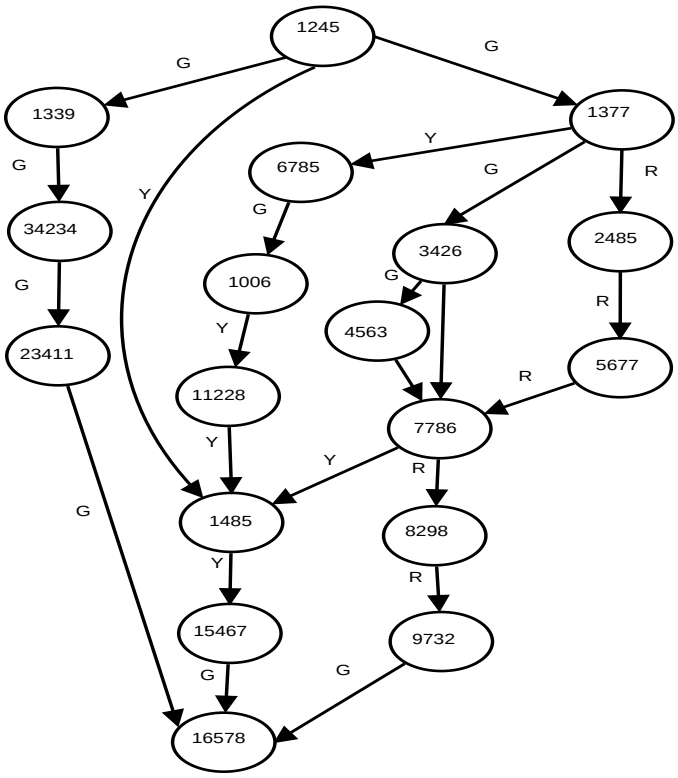


Figure 6. Application of the graph-labelling heuristic. We consider 3 labels "G" < "Y" < "R". Using algorithm 3 the "G" path from the head-end 1245 to the tail-end AS 16578 is chosen in the first iteration.

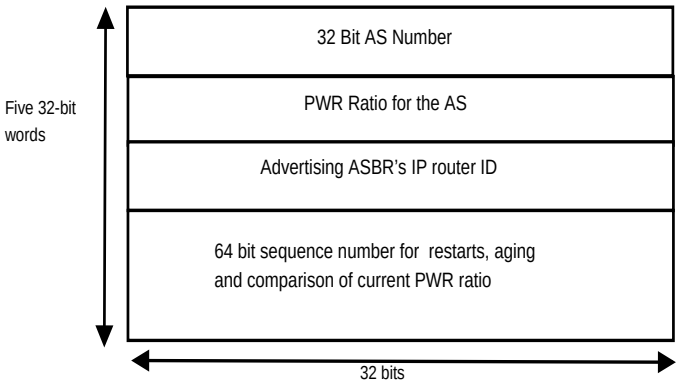


Figure 7. Proposed PDU format with a new attribute for AS-PATH-POWER-METRIC.

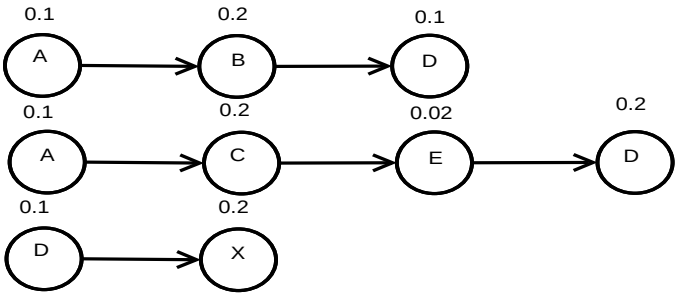


Figure 8. Example of strands where more than one PWR ratio is advertised by "D".

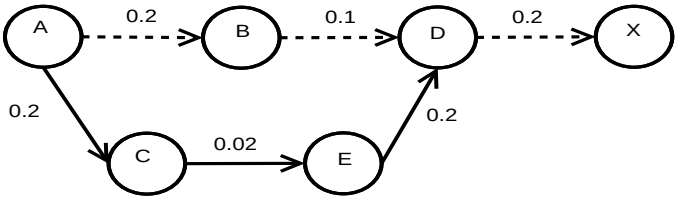


Figure 9. Low-power path derived using the algorithm that uses low value ingress link but through the same AS.

of the Internet receive the attribute and the time it originated. This however cannot be avoided as with other attributes and metrics.

In MPLS-TE, when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR ratio as well. The proposed PWR ratio is advertised to the neighbouring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute can be implemented as shown in Figure 7. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised as the PWR ratio for the AS. This AS-PATH-POWER-METRIC is filled into an appropriate transitive non-discretionary attribute and inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or be the transit prefixes. The filled vector is sent to the ASBR of the neighbouring AS, and later propagated to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is to be identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR ratio has changed in the originating AS.

The power consumed by each router may fluctuate over short time intervals. In order to dampen these fluctuations, which can cause unnecessary updates, power can be measured when falling within intervals of suitable size (say a range of values). This is as opposed to measuring power as a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change.

Multiple ASBRs advertising differing PWR ratio can lead to AS that have low PWR ratio through an ingress link and not through other. Consider the case of multiple ASBRs that belong to the same AS, advertising differing PWR ratios. This could lead to power values that belong to different classes with intervening classes in between. These advertised

PWR ratios could lead to one ASBR being preferred over the other thus taking a different path from head-end to tail-end. This also entails that there may be multiple paths to the AS through these different ASBRs. As an example, consider Figure 8 which shows a set of strands that derive a topology as in Figure 9. Here, "D" is reachable via two paths but the PWR ratios differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

V. Conclusion and Future Work

In this paper, we have proposed a scheme for reducing the power consumption of the Internet using collaborative effort between AS. The topology of the Internet is depicted as a graph using the strands obtained from the AS-PATH attribute of the BGP updates. CSPF algorithm is run on this topology by using the PWR ratio as a constraint. The PWR ratio is advertised through the ingress links of the ASBRs associated with AS using BGP updates. The CSPF algorithm determines the low-power consuming path between AS and routes data packets from head-end to tail-end. Explicit routing is handled through the use of TE-LSPs. Since using CSPF can be time consuming a heuristic algorithm to derive the low-power paths using graph-labelling is proposed. Our work complements the current schemes for reducing power consumption within a router such as switching off or bringing to power-idle-state certain select components within the forwarding and lookup mechanisms.

The scheme proposed in this paper assumes that the PWR ratio information is reliable. It is possible that ISPs could fake the PWR ratio information. However, ISPs usually have service level agreements (SLAs) for carrying traffic. One method is to link up each ISP with a power application level gateway to ensure that proper ratios are advertised. This could be mandated at least amongst the cooperating ISPs. Further the proposed algorithms might lead to increased latency as the number of hops increase, which could be critical for time sensitive applications. Since the PWR ratio could vary dynamically with traffic, the impact of traffic on the algorithm would also be of interest. Our future work will quantify and analyse these issues.

Acknowledgments

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC.

References

[1] G. Appenzeller, "Sizing router buffers", Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.

[2] J. Baliga, K. Hinton and R. S. Tucker, "Energy consumption of the Internet", Proc. of joint International Conference on Optical Internet, June 2007, pp. 1–3, doi: 10.1109/COINA-COFT.2007.4519173.

[3] A. Bianzino, C. Chaudet, D. Rossi and J. Rougier, "A survey of green networking research", IEEE Communications and Surveys Tutorials, preprint, pp. 1–18, doi: 10.1109/SURV.2011.113010.00106.

[4] F. Blanquicet and K. Christensen, "Managing energy use in a network with a new SNMP power state MIB", IEEE Conference on Local Computer Networks, October 2008, pp. 509–511, doi: 10.1109/LCN.2008.4664214.

[5] R. Bolla, R. Bruschi, A. Cianfrani and M. Listani, "Enabling backbone networks to sleep", IEEE Network, vol. 25, no. 2, March/April 2011, pp. 26–31, doi: 10.1109/MNET.2011.5730525.

[6] R. Bolla, R. Bruschi, F. Davoli and F. Cucchietti, "Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures", IEEE Communications Surveys and Tutorials, vol. 13, no. 2, second quarter 2011, pp. 223–244, doi: 10.1109/SURV.2011.071410.00073.

[7] J. Chabarek, J. Sommers, P. Bardford, C. Estan, D. Tsiang and S. Wright, "Power awareness in network design and routing", Proceedings of the IEEE INFOCOM 2008, April 2008, pp. 457–465, doi: 10.1109/INFOCOM.2008.93.

[8] H. Chang, R. Govindan, S. Jamin, S. J. Shenker and W. Willinger, "Towards capturing representative AS-level Internet topologies", Computer Networks, vol. 44, April 2004, pp. 737–755, doi: 10.1016/j.comnet.2003.03.001.

[9] A. Cianfrani, V. Eramo, M. Listanti and M. Polverini, "An OSPF enhancement for energy saving in IP networks", Computer Communications Workshops, INFOCOM 2011, April 2011, pp. 325–330, doi: 10.1109/INFCOMW.2011.5928832.

[10] W. Lu and S. Sahni, "Low-power TCAMs for very large forwarding tables", IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948–959, doi: 10.1109/TNET.2009.2034143.

[11] R. Oliveira, D. Pei, W. Willinger, B. Zhang and L. Zhang, "The (in)completeness of the observed Internet AS-level structure", IEEE/ACM transactions on Networks, vol. 18, no. 1, February 2010, pp. 109-122, doi: 10.1109/TNET.2009.2020798.

[12] B. Venkat et.al, "Constructing disjoint and partially disjoint InterAS TE-LSPs", USPTO Patent 7751318, Cisco Systems, 2010.

[13] M. Xia, M. Tornatore, Y. Zhang, P. Chowdhury, C. Martel and B. Mukherjee, "Greening the optical backbone network: A traffic engineering approach", IEEE ICC Proceedings, May 2010, pp. 1–5, doi: 10.1109/ICC.2010.5502228.

 ISBN: 978-1-61208-189-2

Energy Efficiency of Server Virtualization

Jukka Kommeri
Helsinki Institute of Physics, Technology program, CERN, CH-1211 Geneva 23, Switzerland
jukka.kommeri@cern.ch

Tapio Niemi
Helsinki Institute of Physics, Technology program, CERN, CH-1211Geneva 23, Switzerland
tapio.niemi@cern.ch

Olli Helin
Helsinki Institute of Physics, Technology program, CERN, CH-1211 Geneva 23, Switzerland
olli.helin@cern.ch

***Abstract*—The need for computing power keeps on growing. The rising energy expenses of data centers have made server consolidation and virtualization important research areas. Virtualization and its performance have received a lot of attention and several studies can be found on performance. So far researches have not studied the overall performance and energy efficiency of server consolidation. In this paper we study the effect of server consolidation on energy efficiency with an emphasis on quality of service. We have studied this with several synthetic benchmarks and with realistic server load by performing a large set of measurements. We found out that energy-efficiency depends on the load of the virtualized service and the number of virtualized servers. Idle virtual servers do not increase the consumption much, but on heavy load it makes more sense to share hardware resources among fewer virtual machines.**

***Keywords*-virtualization; energy-efficiency; server consolidation; xen; kvm; invenio; cmssw**

I. Introduction

The need for computing power in data centers is heavily growing due to cloud computing. Large data centers host cloud applications on thousands of servers [1], [2]. Traditionally, servers have been purchased to host one service each (e.g., a web-server, a DNS server). According to many studies the average utilization rate of a server is around 15% of maximum but depends much on the service and it can be even as low as 5% [3], [4].

This level of utilization is very low compared to any field in industry. A common explanation for the low utilization is that data centers are build to manage peak loads. However, this is not a new data center specific issue, since high peak loads are common in many other fields. Even with this low level of utilization the computers are usually operational and consuming around 60% of their peak power [5]. Low utilization level is inefficient through the increased impact on infrastructure, maintenance and hardware costs. For example, low utilization reduces the efficiency of power supplies [6] causing over 10% losses in power distribution. Thus, computers should run in near full power when they do value adding work, because then they operate most efficiently when comparing consumed energy per executed task [3].

A solution for increasing utilization is server consolidation by using virtualization technologies. This enables us to combine several services into one physical server. In this way, these technologies make it possible to take better advantage of hardware resources.

In this study, we focus on energy efficiency of different virtualization technologies. Our aim is to help the system administrator to decide how services should be consolidated to minimize energy consumption without violating quality of service agreements.

We studied energy consumption of virtualized servers with two open source virtualization solutions; KVM and Xen. They were tested both under load and idle. Several synthetic tests were used to measure the overhead of virtualization on different server components. Also a couple of realistic test cases were used; an Invenio database service and CMSSW, The CMS Software Framework, physics analysis software. The results were compared with the results of the same tests on hardware without virtualization. We also studied how overhead of virtualization develops by sharing resources of physical machines equally among different number of virtual machines and running the same test set on each virtual machine set.

II. Related work

Virtualization and its performance is a well-studied area but these studies mainly focus on performance, isolation, and scheduling. Even though energy efficiency is one of the main reasons for server consolidation and virtualization, it has not received much attention. Many of these studies evaluate overhead differences between different virtualization solutions.

Regola et al. studied the use of virtualization in high performance computing (HPC) [7]. They believed that virtualization and the ability to run heterogeneous environments on the same hardware would make HPC more accessible to bigger scientific community. They concluded that the I/O performance of full virtualization or para-virtualization is not yet good enough for low latency and high throughput applications such as MPI applications.

Nussbaum et al. [8] made another study on the suitability of virtualization on HPC. They evaluated both KVM and Xen in a cluster of 32 servers with HPC Challenge benchmarks. These studies did not find a clear winner.

 ISBN: 978-1-61208-189-2

They concluded the performance of full virtulization is far behind that of paravirtulization. Also running workload among different number of virtual machines did not seem to have an effect. Verma et al. [9] also studied the effect of having the same workload on different number of virtual machines. They found out that virtualization and division load between several virtual machines does not impose significant overhead.

Padala et al. [10] carried out a performance study of virtualization. They studied the effect of server load and virtual machine count on multi tier application performance. They found OS virtualization to perform much better than paravirtualization. The overhead of paravirtualization explained by L2 cache misses, which in the case of paravirtualization increased more rapidly when load increased.

Another study from Deshane et al. [11] compare scalability and isolation of a paravirtualized Xen and a full virtualized KVM server. Results said that Xen performs significantly better than KVM both in isolation and CPU performance. Surprisingly, a non-paravirtual system outperforms Xen in I/O performance test.

As we can see from previous work, the energy-efficiency has not received much attention. Our study focuses mainly on the energy-efficiency of virtualization.

III. Tests and Test Environment

Our tests aimed at measuring the energy consumption and overhead of virtualization with a diverse test set. We used both synthetic and real applications in our tests and measured how performance is affected by virtualization. We compared the results of the measurements, that were done on virtual machines, with the results of the same tests on physical hardware. First we measured the idle consumption of virtualized machines using different number of virtual machines. Then we compared different virtualization technologies and operating systems.

A. Test Hardware

The tests were conducted in our dedicated test environment. The test computer was a Dell PowerEdge R410 server with two Intel Xeon E5520 processors and 16 GB of memory. Hyper-threading was disabled for the processors and clock speed was the default 2.26 GHz for all cores. The Turbo Boost option of Intel was enabled. The system had a single 250 GB hard disk drive. As a client computer we used a server with dual Xeon processors running at 2.80 GHz. Network was routed through D-Link DGS-1224T gigabit routers. Power usage data was collected with a Watts up? PRO meter via a USB cable. Power usage values were recorded every second. For physics analysis tests (CMSSW) and some idle tests, we used a dual processor Opteron 2427 server with 32GB of memory and 1 TB hard disk.

B. Used Virtualization Technologies

The operating system used in all machines, virtual or real, was a standard installation of 64-bit Ubuntu Server 10.04.3 LTS. The same virtual machine image was used with both KVM and Xen guests. The image was stored in a raw format, i.e., a plain binary image of the disc image. We used the Linux 3.0.0 kernel. It was chosen as it had the full Xen hypervisor support. With this kernel we were able to compare Xen with KVM without a possible effect of different kernels on performance.

For CMSSW tests, a virtual machine with Scientific Linux 5 was installed with CMSSW version 4.2.4. For these tests real data files produced by the CMS experiment was used. These data files were stored on a Dell PowerEdge T710 server and shared to the virtual machines with a network file system, NFSv4.

C. Test Applications

Our synthetic test collection consisted of Linpack [12], BurnInSSE [1], Bonnie++ [13] and Iperf [14]. Processor performance was measured with an optimized 64-bit Linpack test. This benchmark was run in sets of thirty consecutive runs and power usage was measured for whole sets. In addition, processor power consumption measurements were conducted with ten minute burn-in runs with 64-bit BurnInSSE collection using one, two and four threads. Disk input and output performance was measured using Bonnie++ 1.96. The number of files for a small file creation test was 400. For a large file test the file size was 4 GB. For Bonnie++ tests, the amount of host operating system memory was limited to 2.5 GB with a kernel parameter and the amount of guest operating system memory was limited to 2 GB. For hardware tests, a kernel limit of 2 GB was used. The tests were carried out ten times. Network performance was measured using Iperf 2.0.5. Three kinds of tests were run: one where the test computer acted as a server, another where it was the client and a third where the computer did a loopback test with itself. Testing was done using four threads and a ten minute timespan. All three types of tests were carried out five times.

As real world applications, we used two different systems. The first one was based on the Invenio document repository [15]. We used an existing Invenio installation, which had been modified for the CERN library database. The Invenio document repository software suite was v0.99.2. The document repository was run on an Apache 2.2.3 web server and a MySQL 5.0.77 database management system. All this software were run on Scientific Linux CERN 5.6 inside a chroot environment. Another server was used to send HTTP requests to our test server. The requests were based on an anonymous version of a real-life log data from similar document repository in use at CERN. The requests were sent

[1] http://www.roylongbottom.org.uk

 ISBN: 978-1-61208-189-2

using the Httperf web server performance test application [16].

The second real application was a physics data analysis that used the CMSSW framework [17]. This analysis task is a very typical one in high-energy physics. We used real data created at CERN. The data was stored in a ROOT image[18] files, which our case were of size 4GB. Normally, a data analysis with this data takes days to perform, thus we limited the number of events of one analysis task to 300. With this limitation the analysis takes 10 minutes on the Opteron hardware. The data was located on network file system, NFS, and reading it caused very little network traffic, 2kB per task.

IV. Results

A. Idle consumption

First we studied idle energy consumption with different virtualization solutions and with different number of virtual machines. We also tested the effect of having different operating systems in virtual machines.

Figure 1 shows the power consumption of two different virtualization solutions. We had three virtual machines running idle in both cases. The figure shows how energy consumption of two different virtualization solutions behave when the servers are idle. It shows how overhead of virtualization depends on the virtualization solution and kernel version. The difference between KVM and hardware is less than 3%, which is already a big improvement compared to three separate physical machines running idle. This test was run with the Dell R410 server.

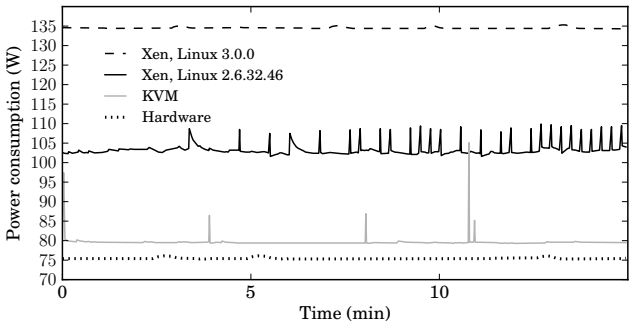


Figure 1. Typical idle power consumption

The second idle measurement was run on the dual processor Opteron server. Test measured the energy comsumption of the physical hardware for 20 minutes. Figure 2 shows how the operating system affects the idle consumption. The same test was repeated with different number of virtual machines on the same physical hardware. The energy consumption cumulates with the virtual machine count when we have Scientific Linux 5 (SLC5) but with Ubuntu it remains almost the same as hardware.

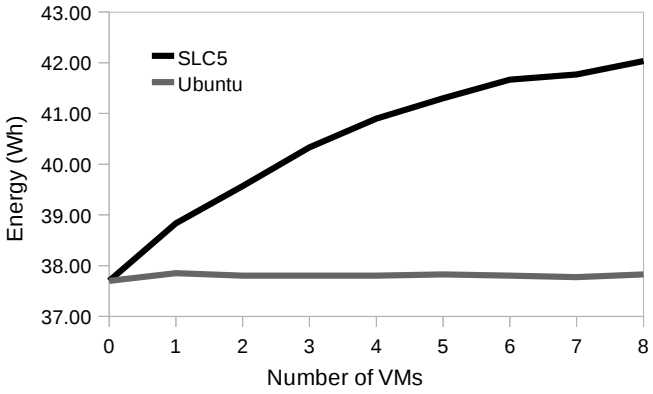


Figure 2. Energy consumption of idle virtual machines

B. Synthetic tests

We used synthetic tests to stress different server components; CPU, I/O and network. With these tests we studied in which situations virtualization causes the most overhead.

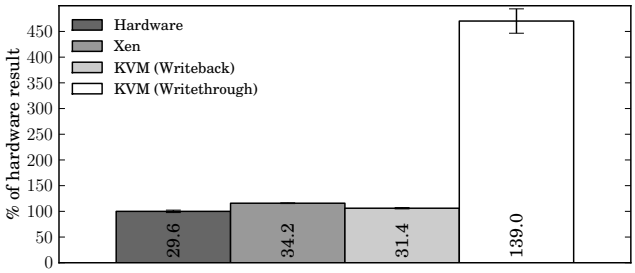


Figure 3. The energy consumption of Bonnie++ (Wh)

As seen in Figure 3, when running a set of synthetic disk operations Xen uses slightly more energy compared to hardware. With KVM the situation is different. When using the default writethrough cache, KVM uses around 350% more energy than hardware. About 90% of the test time is spent doing the small file test part of Bonnie++. Switching to writeback cache, results of KVM are actually slightly better than hardware results. Writeback cache writes only to a cache and stores data to the disk only just before the cache is replaced. This is a cache mode that is not safe for production use and is available mainly for testing purposes.

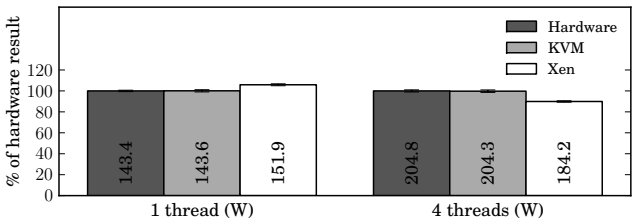


Figure 4. Power consumption under high CPU load with BurnInSSE

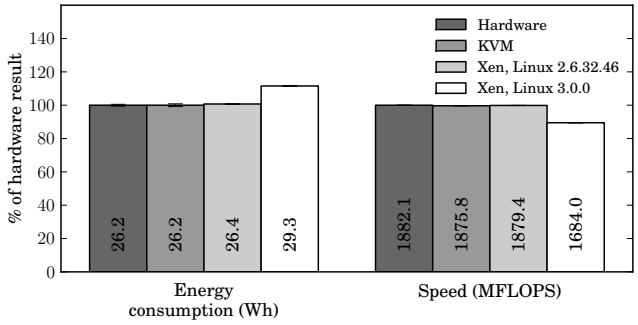


Figure 5. Energy consumption under high CPU load with Linpack

In Figure 4, power consumption is tested under full CPU load between 1 and 4 threads running BurnInSSE64. With 1 thread, KVM and hardware use the same amount of power, but Xen uses around 10% more. With 4 threads the situation is the other way around: Xen uses less power than KVM and hardware. The explanation to this can be seen in Figure 5. Even though Xen uses more power in the single-threaded LINPACK test, it is slower: the CPU is not running at its full turbo boosted speed, but Xen has a systematic overhead in power consumption compared to the others. With 4 threads, Xen's CPUs are not running at full speed so the power usage is not so great as with hardware or KVM, and the effect of overhead in power consumption is overshadowed by the power usage of 4 computing threads.

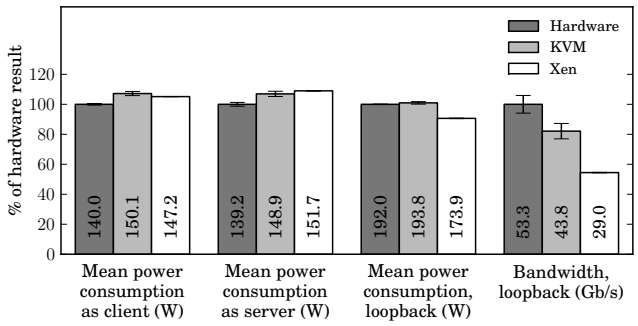


Figure 6. Power consumption under Iperf network traffic test

Iperf test results in Figure 6 show a similar trend: KVM uses slightly more power than hardware while Xen consequently uses slightly more power than KVM. Interestingly, when a Xen virtual machine was running as server it used slightly more power than when running as client. With KVM and hardware it was the other way around. In the loopback mode, one can find similar results with Xen as in the LINPACK test in Figure 5: for some reason, Xen's performance is capped and consequently bandwidth in the loopback mode is much worse than with KVM or hardware, and on the other hand mean power consumption is lower.

C. Realistic load

Realistic tests were designed such that we would get better understanding of energy usage in two different real world situations: web services and physics analysis.

The first realistic test was a CERN document server repository case. In this test, we sent HTTP requests, which were based on CERN library log data, to a virtualized web server. We measured both performance and power consumption. We ran the same test with and without virtualization. We compared two virtualization solutions and hardware to measure the overhead of virtualization. In all Invenio tests, the Invenio installation was in a chroot environment with a complete SLC5 installation. To assure that chroot between the operating system and the Invenio web application did not have any negative effects on test results, a comparative test was performed between the base system and another chroot environment using a copy of the base system as the new root.

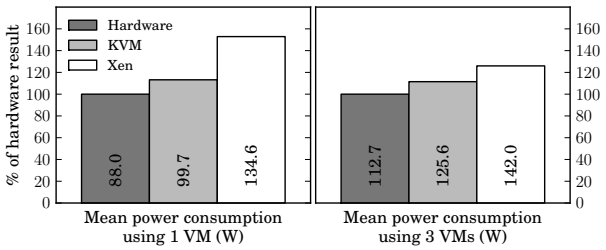


Figure 7. Power consumption of different virtualization solutions with different number of virtual machines in the repository test

In Figure 7, we have the results for Httperf rates 5 and 15. On the left side we have one virtual machine with rate 5 workload and on the right side 3 virtual machines with load 5. It shows the power consumption levels when we have more virtual machines and load.

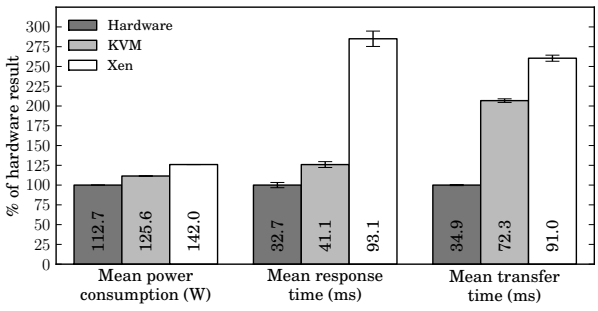


Figure 8. Power consumption and Httperf results of different virtualization solutions

Figure 8 shows the results of comparison of hardware, Xen and KVM with different number of virtual machines.

We tested overhead of virtualization by increasing the number of virtual machines with similar workload. These

 ISBN: 978-1-61208-189-2

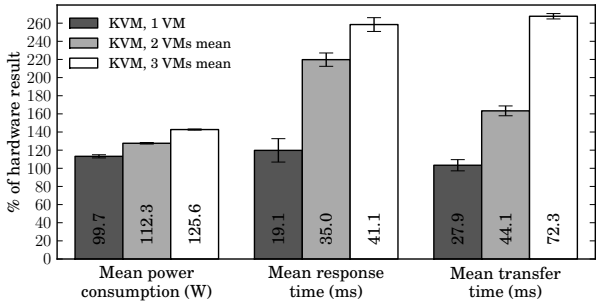


Figure 9. The effect of workload on virtual machine performance with KVM using different amounts of virtual machines

results are illustrated in Figure 9

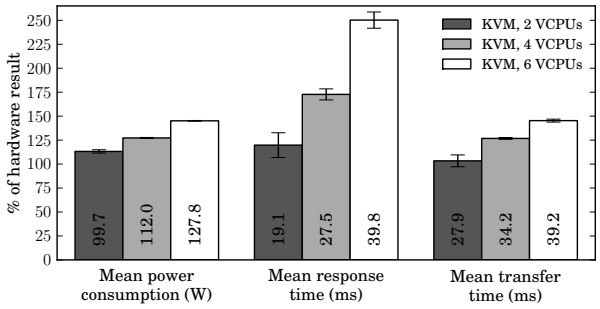


Figure 10. The effect of workload on virtual machine performance with KVM using different virtual machine resources

In Figure 10, we have the results of a test where, instead of growing the number of virtual machines, we increased the load and resources of one virtual machine. Table IV-C shows the rates and resources given to virtual machines in these tests. The MaxClients setting refers to the maximum clients setting in Apache web server configuration.

Table I
SETTINGS FOR CHANGING LOAD AND RESOURCES OF A SINGLE VIRTUAL MACHINE

VCPUs	Memory (GB)	MaxClients	Request rate
2	5	8	5
4	8	15	10
6	15	24	15

Figures 11 and 12 illustrate the effect of virtualization on quality of service. They show a cumulative distribution of response times the Httperf test application reported for the HTTP requests.

In Figure 11, we have the results of running workload of 15 queries per second on 3 virtual machines and on hardware for comparison. Corresponding performance results are shown in Figure 8. These figures show how the response times increase when the same workload is divided into 3 virtual machines.

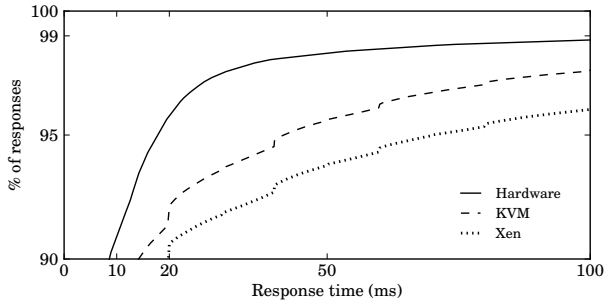


Figure 11. The impact of virtualization solution on quality of service

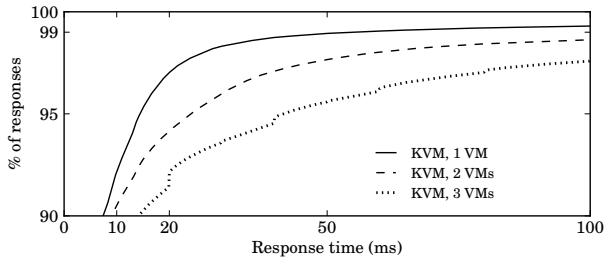


Figure 12. The impact of different loads on quality of service

To better show the virtualization overhead effect on different workload and number of virtual machines we compared KVM with rates 5, 10, and 15. The results of this experiment can be seen in Figure 12. Corresponding performance measurements are shown in Figure 9.

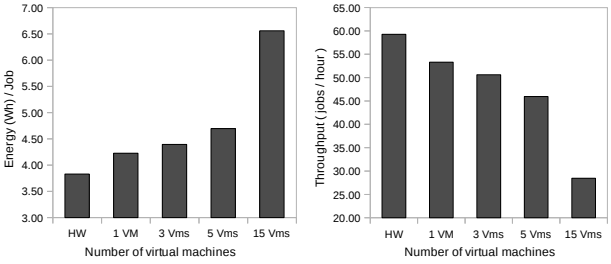


Figure 13. Running 15 jobs in different number of virtual machines

As our second realistic load, we had a physics analysis application. Here we consider one run of the test application as a job. In Figure 13, we have the results of running 15 jobs in 5 different virtual machine sets. The figure illustrates how the energy efficiency degrades as the number of virtual machines increases. 15 virtual machines running one job is 6.8 times less energy efficient than running 15 jobs on one virtual machine.

Figure 14 shows the effect of workload on energy-efficiency. We tested different workloads on 5 identical virtual machines sharing the the same physical server.

 ISBN: 978-1-61208-189-2

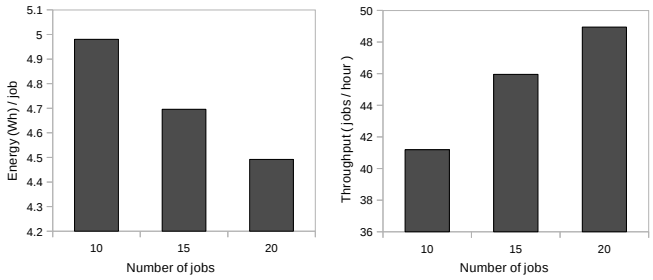


Figure 14. Running different workload on 5 virtual machines

V. CONCLUSIONS

The overhead of virtualization is a well-known fact reported in many publications. Although the technologies have been improving a lot during the past five years, the performance of a virtualised system is still far from the hardware level. However, this does not mean that virtualization could not be useful in improving energy-efficiency in large data centers but it means that one should know how to apply this technology to achieve savings in energy consumption.

We studied the energy-efficiency virtualization technologies and how different load affects it. Our research indicates that idle power consumption of virtualized server is close to zero. However, this depends a lot on the operating system running on the virtual machine, but it is always a small number compared to idle energy consumption of a physical server. Our study also indicates that virtualization overhead has great impact on energy-efficiency. This means that it would make more sense to share the physical resources among few virtual machines with heavy load instead of larger set of light-loaded ones.

ACKNOWLEDGMENT

Many thanks to Jochen Ott from CMS@CERN experiment for providing help in installing the CMSSW and providing with a typical analysis job. Also we would like to thank Salvatore Mele, Tibor Simko, Jean-Yves LeMeur of CERN library and Invenio developers, for providing realistic data and a test case for our analysis.

REFERENCES

[1] B. Schäppi, F. Bellosa, B. Przywara, T. Bogner, S. Weeren, and A. Anglade, "Energy efficient servers in europe," Austrian Energy Agency, Tech. Rep. October, 2007.

[2] E. STAR, "Report to congress on server and data center energy efficiency," U.S. Environmental Protection Agency ENERGY STAR Program, Tech. Rep., 2007.

[3] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, pp. 33–37, 2007.

[4] W. Vogels, "Beyond server consolidation," *Queue*, vol. 6, pp. 20–26, January 2008.

[5] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," in *Proceeding of the 14th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS '09. Washington, DC, USA: ACM, 2009, pp. 205–216.

[6] U. Hölzle and B. Weihl, "High-efficiency power supplies for home computers and servers," Google, Tech. Rep., 2006.

[7] N. Regola and J.-C. Ducom, "Recommendations for virtualization technologies in high performance computing," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 30 2010-dec. 3 2010, pp. 409–416.

[8] L. Nussbaum, F. Anhalt, O. Mornard, and J.-P. Gelas, "Linux-based virtualization for hpc clusters," *Network*, pp. 221–234, 2009.

[9] A. Verma, P. Ahuja, and A. Neogi, "Power-aware dynamic placement of hpc applications," in *Proceedings of the 22nd annual international conference on Supercomputing*, ser. ICS '08. New York, NY, USA: ACM, 2008, pp. 175–184.

[10] P. Padala, X. Zhu, Z. Wang, S. Singhal, and G. Shin, K., "Performance evaluation of virtualization technologies for server consolidation," *Work*, no. HPL-2007-59, p. 15, 2007.

[11] T. Deshane, Z. Shepherd, J. Matthews, M. Ben-Yehuda, A. Shah, and B. Rao, "Quantitative comparison of xen and kvm," in *Xen summit*. USENIX association, June 2008.

[12] J. Dongarra, P. Luszczek, and A. Petitet, "The linpack benchmark: past, present and future," *Concurrency and Computation Practice and Experience*, vol. 15, no. 9, pp. 803–820, 2003. [Online]. Available: http://doi.wiley.com/10.1002/cpe.728

[13] B. Martin, "Using bonnie++ for filesystem performance benchmarking," *Linuxcom*, vol. Online edi, 2008.

[14] M. Egli and D. Gugelmann, "Iperf - network stress tool," *Source*, pp. 1–2, 2007.

[15] J. Caffaro and S. Kaplun, "Invenio: A modern digital library for grey literature," in *Twelfth International Conference on Grey Literature*, Prague, Czech Republic, Dec 2010.

[16] D. Mosberger and T. Jin, "httperf - a tool for measuring web server performance," *SIGMETRICS Perform. Eval. Rev.*, vol. 26, pp. 31–37, Dec 1998.

[17] F. Fabozzi, C. Jones, B. Hegner, and L. Lista, "Physics analysis tools for the cms experiment at lhc," *Nuclear Science, IEEE Transactions on*, vol. 55, pp. 3539–3543, 2008.

[18] I. Antcheva and et al., "Root a c++ framework for petabyte data storage, statistical analysis and visualization," *Computer Physics Communications*, vol. 180, no. 12, pp. 2499 – 2512, 2009.

 ISBN: 978-1-61208-189-2