ICDT 2015

The Tenth International Conference on Digital Telecommunications

ISBN: 978-1-61208-396-4

April 19 - 24, 2015

Barcelona, Spain

ICDT 2015 Editors

Tapio Saarelainen, National Defence University, Finland

ICDT 2015

Foreword

The Tenth International Conference on Digital Telecommunications (ICDT 2015), held between April 19th-24th, 2015 in Barcelona, Spain, continued a series of special events focusing on telecommunications aspects in multimedia environments. The scope of the conference was to focus on the lower layers of systems interaction and identify the technical challenges and the most recent achievements.

High quality software is not an accident; it is constructed via a systematic plan that demands familiarity with analytical techniques, architectural design methodologies, implementation polices, and testing techniques. Software architecture plays an important role in the development of today's complex software systems. Furthermore, our ability to model and reason about the architectural properties of a system built from existing components is of great concern to modern system developers.

Performance, scalability and suitability to specific domains raise the challenging efforts for gathering special requirements, capture temporal constraints, and implement service-oriented requirements. The complexity of the systems requires an early stage adoption of advanced paradigms for adaptive and self-adaptive features.

Online monitoring applications, in which continuous queries operate in near real-time over rapid and unbounded "streams" of data such as telephone call records, sensor readings, web usage logs, network packet traces, are fundamentally different from traditional data management. The difference is induced by the fact that in applications such as network monitoring, telecommunications data management, manufacturing, sensor networks, and others, data takes the form of continuous data streams rather than finite stored data sets. As a result, clients require long-running continuous queries as opposed to one-time queries. These requirements lead to reconsider data management and processing of complex and numerous continuous queries over data streams, as current database systems and data processing methods are not suitable.

We take here the opportunity to warmly thank all the members of the ICDT 2015 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICDT 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICDT 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICDT 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of digital communications.

We are convinced that the participants found the event useful and communications very open. We hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

ICDT 2015 Advisory Chairs:

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Abdulrahman Yarali, Murray State University, USA
Michael Grottke, University of Erlangen-Nuremberg, Germany
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain
Saied Abedi, Fujitsu Laboratories of Europe Ltd. (FLE), UK
Gerard Damm, Alcatel-Lucent, USA
Dan Romascanu, Avaya, Israel
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD - Darmstadt, Germany

ICDT 2015

Committee

ICDT Advisory Chairs

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Abdulrahman Yarali, Murray State University, USA
Michael Grottke, University of Erlangen-Nuremberg, Germany
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain
Saied Abedi, Fujitsu Laboratories of Europe Ltd. (FLE), UK
Gerard Damm, Alcatel-Lucent, USA
Dan Romascanu, Avaya, Israel
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD - Darmstadt, Germany

ICDT 2015 Technical Program Committee

Antonio Marcos Alberti, INATEL - Instituto Nacional de Telecomunicações, Brazil
Abdullah M. Alnajim, Qassim University, Saudi Arabia
Maria Teresa Andrade, FEUP / INESC Porto, Portugal
Iosif Androulidakis, MPS Jozef Stefan, Slovenia
Regina B. Araujo, Federal University of São Carlos, Brazil
Khaled Assaleh, American University of Sharjah, United Arab Emirates
Anteneh Ayanso, Brock University, Canada
Francisco Barcelo-Arroyo, Technical University of Catalonia, Spain
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Abdelouahab Bentrcia, King Saud University – Riyadh, Kingdom of Saudi Arabia
Andrzej Borys, Gdynia Maritime University - Gdynia, Poland
Damian Bulira, Wroclaw University of Technology, Poland
Ladislav Burita, University of Defence in Brno / Tomas Bata University in Zlin, Czech Republic
Andi Buzo, University Politehnica of Bucharest, Romania
Lee-Ming Cheng, City University of Hong Kong, Hong Kong
Alberto Coen-Porisini, Università degli Studi dell'Insubria – Varese, Italy
Doru Constantin, University of Pitesti, Romania
Gerard Damm, Alcatel-Lucent, USA
Klaus Drechsler, Fraunhofer-Institut für Graphische Datenverarbeitung IGD - Darmstadt, Germany
Roger Pierre Fabris Hoeffel, Federal University of Rio Grande do Sul, Brazil
Peter Farkas, FEI STU – Bratislava, Slovakia
Christophe Feltus, Luxembourg Institute of Science and Technology (LIST), Luxembourg
Gerardo Fernández-Escribano, University of Castilla-La Mancha, Spain
Mário Ferreira, University of Aveiro, Portugal
Pierfrancesco Foglia, University of Pisa, Italy

Copyright Information

Table of Contents

Power Consumption Analysis of FTTH Networks 1
Kun Wang, Maria Kihl, Anders Gavler, Manxing Du, and Christina Lagerstedt

A Dynamic Scalable Cryptosystem Based-on Reduced Key Size ECC 7
Jia Wang and Lee-Ming Cheng

Analysis of Prefetching Schemes for TV-on-Demand Service 12
Manxing Du, Maria Kihl, Ake Arvidsson, Christina Lagerstedt, and Anders Gavler

Unmanned Aerial Vehicles as Assisting Tools in Dismounted Company Attack 19
Tapio Saarelainen

Indoor Positioning and Navigation System for Interior Design Augmented Reality 28
Fady Adel, Mina Makary, Mohamed El-Nahas, and Mustafa ElNainay

Power Consumption Analysis of Energy-aware FTTH Networks

Kun Wang, Anders Gavler, Manxing Du, Christina Lagerstedt
Dept. Network and Transmission Lab
Acreo Swedish ICT
Kista, Sweden
Email: kunw@kth.se

Maria Kihl
Dept. Electrical and Information Technology
Lund University
Lund, Sweden
Email: maria.kihl@eit.lth.se

***Abstract*—With increasing usage of the Internet, energy consumption of network equipment has become a crucial challenge from both an economic and an environmental point of view. This paper combines users' behavior of accessing the network with energy saving algorithms for energy-aware network equipment, and investigates potential energy savings in the access network. The study is based on a set of traffic data that collected from a real residential fiber-to-the-home (FTTH) network during three continuous months in 2013. The results show that on average every household link in the access network can potentially save at least 18% energy consumption with sleep-mode enabled equipment.**

Keywords- energy efficiency; user behaviour; traffic measurement; FTTH; access netowork.

I. INTRODUCTION

With increasing usage of Internet and its related services, the volume of global IP traffic is growing enormously [1]. Recent reports states that the world's Information-Communication-Technologies (ICT) ecosystem approaches 10% of the world's electricity generation [2]. Other papers estimate that about 37% of the ICT electricity is generated in the telecommunication networks, whereas the rest is mainly generated in data centers and user devices [3]. Therefore, energy consumption has become a particularly important economic and environmental interest when the global IP traffic grows at such dramatically fast pace.

In the recent years, a number of studies have been published focused on energy consumption in telecommunication networks. For example in [4], the power consumption of nodes and links in a backbone network was investigated. Further, according to [5][6], the access networks account for about 70% of the networks' total electricity generation. Also, according to [7], in a Fiber-To-The-Home (FTTH) network, a major part of the electricity consumption is due to the optical network unit (ONU) at the end-user's home and the access line utilization is quite low, which means that the majority of the spent power could be saved with power-saving solutions. In [8], a benchmark for current network devices was presented, and the results showed that the power consumption of the Ethernet switch depends on the number of active ports, which means that switching off unused ports can reduce power consumption.

Further, in [5][9] some existing energy saving approaches used in fixed network and optical access networks were summarized. Two common approaches are idle mode (also called sleep mode) and adaptive link rate (ALR) [10][11].

The idle mode concept is adopted by the Energy Efficient Ethernet (EEE, also known as IEEE 802.3az) [12][13], which was published in 2010. EEE allows network components to sleep, i.e. standby at a low power idle state to save the energy consumption, when there is no data packets transmitted in the link. The ALR approach dynamically associates the energy consumption of the network equipment with the actual data rate or workload transmitted in the network, so that the energy dissipation can be allocated according to the traffic load pattern. In [12][13], the potentials of EEE for reducing energy consumption was evaluated, however, not including data from real operational networks.

In order to develop efficient ways of reducing energy consumption in the networks, it is essential to also understand the users' behavior when accessing the networks [14][15]. However, there are only a few papers that use real data traces to investigate energy consumption and energy saving solutions for telecommunication networks. In [4], a real aggregated traffic profile was used in order to study the power consumption of nodes and links in a backbone network. In [16], a daily Internet usage pattern from 2008 was used to calculate potential energy saving from an ALR approach. However, both these papers used one generic example of the traffic pattern for their analysis. In reality, different households can have very varying traffic patterns depending on each household's Internet usage behavior.

This study uses traffic measurement data from real residential users in a FTTH access network, and estimates the potential amount of energy savings when sleep-mode enabled energy-aware equipment were applied. The main novelties are: firstly our investigation covers 2627 households with ultra-fast internet access service in a FTTH network during the 3 months period, so that our results give a practical evaluation of the performance of energy saving approaches in connection to a real network user case; secondly this paper proposes a method of mapping users' traffic behavior to link states, which then can be mapped to

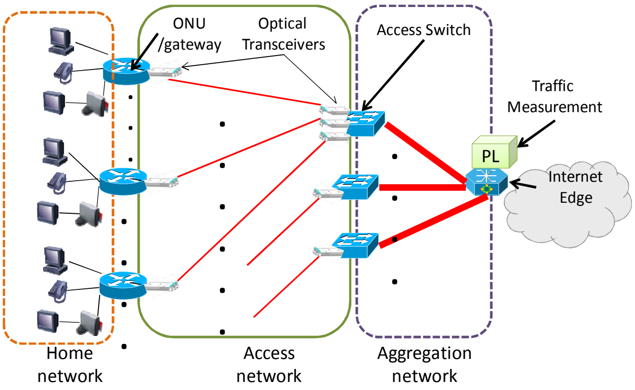


Fig.1. Network architecture

corresponding power consumption in the network devices; thirdly, two ONU energy models are proposed and compared according to the characteristics of traffic measurement data, the first model includes sleep-mode feature, and the second model includes both sleep-mode and off-mode feature.

The paper is organized as follows. First a description of the network scenario, energy consumption models, and the scope of the work are presented in Section II. In Section III, we propose a mapping between Internet user behavior and energy models, and then analyze how much energy can be saved in different energy consumption models when Internet user behavior pattern applied in Section IV. A conclusion is drawn in Section V.

II. METHODOLOGY

A. Network and measurements

The analysis in this paper is based on measurements in a real operational FTTH residential network in Sweden. The FTTH network is an active optical network (AON), which is currently the most deployed fiber access solution in Europe representing 78% of overall deployment at the mid of 2012 [17]. As Fig. 1 shows, every household has a 100 Mbps Internet service subscription and connects to the network via an Optical Network Unit (ONU) acting as a home gateway. Further, the ONU is connected to the network operator's access switch via a dedicated fiber link.

The access Ethernet switch, also called optical line termination (OLT) for Point-to-Point (PtP) FTTH, is located at the border between the aggregation network and access network. Every downlink port of the access Ethernet switch is connected to one household. The uplink interface towards aggregation network is shared among all households connected to the switch.

Traffic measurements were performed by the network operator, and post processed, anonymous data were made available to the authors for analysis. The traffic measurement probe was PacketLogic (PL) [18], a commercial traffic management device, which can track and identify several hundred thousand simultaneous connections. The PL was placed at the Internet edge of the network, see Fig. 1, and the volumes of traffic for each household, in both inbound and outbound directions, were recorded at 5 minutes intervals. The collected data comprise 2627 households and the study is based on data from 2013-March-01 to 2013-May-31.

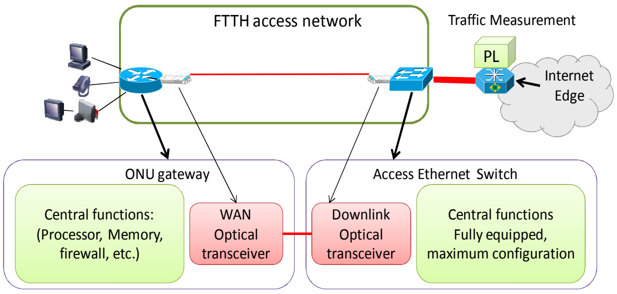


Fig.2. Network scenario

B. Investigated scenario

In this paper, we focus on the access (first mile) part of the network, more specifically the ONU (optical network unit) with home gateway functions, the access Ethernet switch and optical transceivers as described in Fig. 1. The ONU terminates the fiber optical signal from the FTTH network and converts it to electrical signal which is then communicated with other home network devices, e.g., TV, PC, IP telephony, etc.

The power consumption of the ONU can be assumed to come from two parts. The first part is the home gateway central functions, for example, the processor, memory for routing, firewall, OAM and user interfaces [19]. The second part is the wide area network (WAN) interface, i.e., the optical transceiver, which directly connects to the access switch.

Further, in this paper, we consider the power consumption of the downlink Ethernet switch center functions and optical transceiver that connects towards users. However, the uplink interface was excluded, since we assume that the possibility of this shared interface to be idle is low.

A more detailed illustration of the investigated network scenario is shown in Fig. 2.

III. ENERGY CONSUMPTION

In this section, we describe the proposed energy models used in this paper. The objective of our analysis has been to investigate the potential energy-savings that can be obtained with energy-aware equipment, based on data from a real operational network.

The European commission code of conduct for 2013 [19] defines three power states for energy-aware equipment using idle or sleep mode approach. The first power state is the "Active state", which means that the link is actively used and that the device is processing user traffic at its best performance. The second power state is the "Idle state", which indicates that the device is not processing or transmitting any user traffic, but that the link is established and ready to detect activity. The third power state is the "Off state", which corresponds to that the device is totally

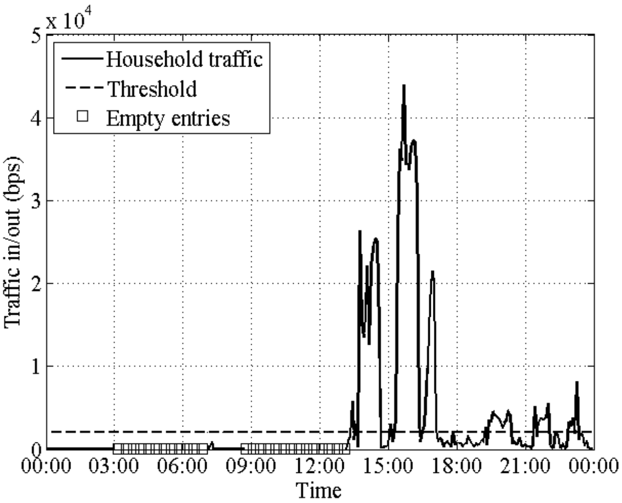


Fig. 3. An example of measured household traffic volumes during one day. The measured traffic rate is shown by the solid line. Empty entries are intervals where no traffic was detected. The threshold value is used to separate the Active and Idle power states.

powered off and not providing any functionality, which means that the link cannot send any traffic.

A. Mapping from traffic volumes to link states

To be able to analyze the effects of energy-aware equipment, we need to map the measured traffic volumes to link states, which then can be mapped to corresponding power consumption in the devices. In this subsection we describe the mapping from the measured traffic volumes to link states.

The measurements registered the average traffic rate (in bits per second) during five minutes intervals for each household, as illustrated in Fig. 3. For each interval, the total incoming and outgoing traffic rate was measured. If substantial traffic was observed by the PL, the household was assumed to be actively using the Internet, and therefore, the link was determined to be in the Active state. If a low rate of traffic was observed by the PL the network devices are powered on, but the household was assumed to not actively use the Internet, and therefore, the link can be determined to be in the Idle state. The small amount of traffic was assumed to be control and management (C&M) communications between the ISP and the ONU gateway.

If no traffic was observed by the PL during an interval, this was registered as an empty entry in the database. This can be due to two main reasons. Either the household's traffic was too small to be recorded, or the ONU gateway was powered off by the household. When there was an empty entry in the database, the link was considered to be in an Off state.

To separate the Active and Idle link states, we need a threshold value. In this paper, we use two cases for the threshold value. In the first case we use a threshold value of 100 bps. In this case, we assume that the C&M communications constitute maximum 100 bps for each household. If the registered data rate is smaller or equal to 100 bps we assume that the gateway is not processing and transmitting any user requested data, therefore we determine that the link is in the Idle state. In the second case, we use a threshold value of 0 bps, which leads to an extreme case where the Idle state can be activated only when there is no traffic observed at all in the ONU gateway.

TABLE I AVERAGE POWER OF ONU AND ACCESS SWITCH. VALUES WERE EXTRACTED FROM [19]

Link	ONU gateway	Access switch
Active	G_{active}=5 watt	E_{active} = 4 watt / port
Idle	G_{idle}=2.9 watt	E_{idle} = 2.7 watt / port*
Off	G_{off}=0 watt	E_{idle}

*This value was not available in [19]. We assumed 2/3 of active power

B. Power values

The power values for the ONU gateways and the Ethernet access switches used in this paper were extracted from [19]. TABLE I summarizes the proposed link states and corresponding power of the network equipment. The power values for the ONU gateway are denoted as G_{active}, G_{idle}, and G_{off} for the respective link states. The power values for the Ethernet access switch are denoted as E_{active} and E_{idle} for the respective power states. The Ethernet access switch is assumed to not have an energy-aware Off state. Therefore, the switch can use the same power as in the Idle state also when there is no traffic.

C. Energy models

In this subsection, we present our proposed energy models that we use to investigate the potential energy-savings with energy-aware equipment.

For one household *k*, the total power consumption, P_k, is modeled as (1). $P_g(k)$ is the total power consumption of the household's ONU gateway, and $P_e(k)$ is the total power consumption related to the household's port in the access Ethernet switch.

$$P_k = P_g(k) + P_e(k) \tag{1}$$

To estimate the power consumption for each device, we need to know the amount of time that the device works in its respective power state. We estimate the power states by using our traffic measurements. Each data point in the measurement (corresponding to a five minutes interval) is mapped to a corresponding power state, according to the procedure described in the previous subsection. The total time that the devices for household *k* is considered to be in Active power state is denoted $T_{active}(k)$, the total time for Idle power state is denoted $T_{idle}(k)$, and the total time for the Off power state is denoted $T_{off}(k)$.The values are calculated as accumulated five minutes intervals.

Thereafter, the total power consumption for the Ethernet access port switch corresponding to household *k* can be derived as

$$P_e(k) = E_{idle} \cdot (T_{idle}(k) + T_{off}(k)) + E_{active} \cdot T_{active}(k) \tag{2}$$

TABLE II AVERAGE ENERGY CONSUMPTION PER HOUSEHOLD DURING THREE MONTHS. STD = STANDARD DEVIATION. 95% CI = 95% CONFIDENCE INTERVAL

	ONU	**Access switch / port**	**ONU + switch**	**STD**	**95% CI**
No Energy-awareness	11.04 kwh	8.83 kwh	19.87 kwh	0	0
ONU Model 1 TH 100bps	7.28 kwh (-34.1%)	6.50 kwh (-26.4%)	13.78 kwh (-30.7%)	1.39	0.05
ONU Model 2 TH100bps	4.59 kwh (-58.4%)		11.09 kwh (-44.2%)	3.06	0.12
ONU Model 1 TH 0bps	8.76 kwh (-20.7%)	7.42 kwh (-16%)	16.18 kwh (-18.6%)	2.36	0.09
ONU Model 2 TH0bps	6.07 kwh (-45.0%)		13.49 kwh (-32.1%)	4.34	0.17

where E_{idle} and E_{active} are the power values for the Ethernet access switch found in TABLE I. Since the access switch is assumed to not have an energy-aware Off state, both the household's mapped Idle and Off states correspond to the Idle power value.

In this paper, we propose and evaluate two energy models for the ONU gateway, in order to show the effects when an energy-aware Off state is implemented in the ONU. Therefore, the first ONU gateway model (in the following called ONU1) only has "Active" and "Idle" power states, whereas the second ONU gateway model (in the following called ONU2), instead has all power states, "Active", "Idle" and "Off.

For model ONU1, the total power consumption for the ONU gateway belonging to household k is given by

$$P_g(k) = G_{idle}\cdot(T_{idle}(k)+T_{off}(k)) + G_{active}\cdot T_{active}(k) \qquad (3)$$

where G_{idle} and G_{active} are the respective power values for the ONU gateway found in TABLE I. Both the link's Idle and Off states correspond to the Idle power value.

For model ONU2, the total power consumption for the ONU gateway belonging to household k is instead given by

$$P_g(k) = G_{off}\cdot T_{off}(k) + G_{idle}\cdot T_{idle}(k) + G_{active}\cdot T_{active}(k) \qquad (4)$$

IV. RESULTS AND DISCUSSION

The main objective of our investigations was to evaluate the amount of energy savings that can be achieved with energy-aware equipment based on the households' Internet usage. Each household's energy consumption during the three months measurement period was therefore estimated using both energy models and the two threshold values for the links.

A. *Average energy savings*

TABLE II shows the estimated average energy consumption per household when compared to a benchmark case without energy-aware equipment. The potential energy savings are shown in the parentheses of TABLE II). As can be seen in the table, each ONU gateway consumes in average 11 kWh whereas each connected port on access switch consumes in average 8.8 kWh without energy-aware equipment. However, with energy-aware equipment, large energy savings can be achieved. The most striking result is that 58% of the ONU energy consumption can be saved in the case where the ONU uses an Off mode (ONU Model 2) and the link state threshold 100bps. Even when ONU Model 1 is used, without Off mode, and the link state threshold value of 0 bps is applied (which is very strict), there is still a potential of 21% average energy savings for the ONUs. Energy-aware access switches can also have significant potential energy savings, about 26% and 16% on average for the two cases respectively.

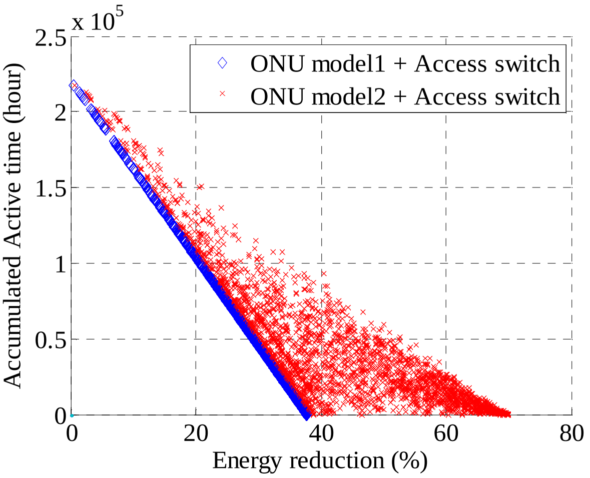


Fig. 4. Relationship between accumulated active time per household and the potential energy savings. Each data point corresponds to one household.

B. *Households' Internet usage and energy-savings*

Fig. 4 shows the relationship between the accumulated active time (T_{active}) for each household and the potential energy savings calculated from the energy models with threshold of 100bps. Every data point in the figure corresponds to one household. The results clearly show that the most active households only have small benefits from using energy-aware equipment. However, for the households who have less active time, significant potential energy savings can be achieved by using energy-aware equipment. Using ONU model 1, without Off mode, the households who are the least active can save about 36% of the energy consumption, while for ONU model 2, with Off mode, the savings can reach as high as 67% compared to the case without energy-aware equipment. When energy-aware equipment is not deployed in the access networks, these "light" Internet users will consume the same amount of energy as the "heavy" Internet users.

Further, the graph indicates the potential benefits of introducing an Off state in the ONU gateway. The shorter active Internet time a household has, the more energy-savings from the Off state will the household have.

C. *Cumulative distribution functions*

Fig. 5 shows the cumulative distribution function (CDF) for the households in relation to the potential ratio of energy savings that the households can achieve, when ONU Model

ISBN: 978-1-61208-396-4

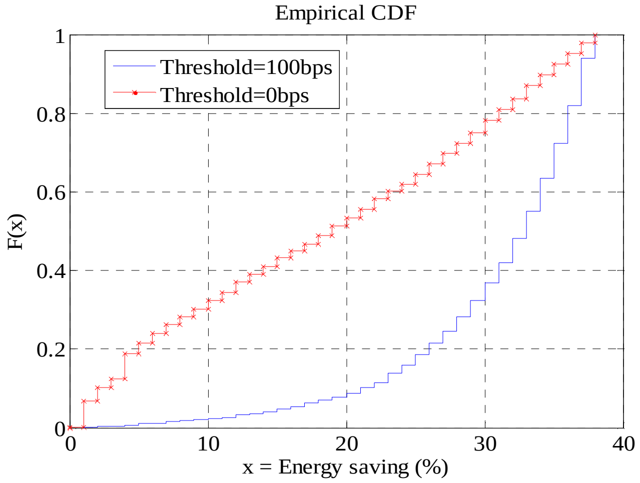


Fig. 5 Distribution of households in relation to the amount of energy savings using ONU model 1. The blue line corresponds to the link state threshold of 100bps, and the red crossed marker corresponds to the link state threshold of 0bps.

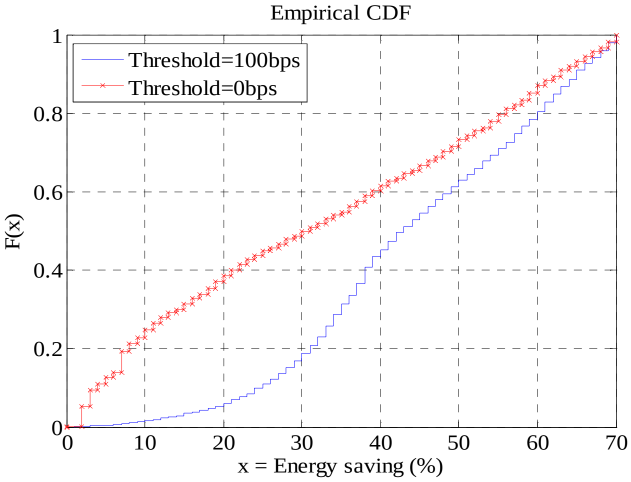


Fig. 6 Distribution of households in relation to the amount of energy savings using ONU model 2. The blue line corresponds to the link state threshold of 100bps, and the red crossed marker corresponds to the link state threshold of 0bps.

1, without Off state, is applied. Both link state thresholds are shown in the figure. As can be seen in the figure, for the 100 bps threshold case, the CDF curve increases exponentially. About 30% of the households in the network can save only up to about 30% of the energy using energy-aware equipment, whereas the rest of households can save from 30% to 36% of the energy with energy-aware equipment. There is a significant difference in the 0bps threshold case, where the CDF shows a clear linear trend. In both cases, the maximum energy saving is about 36%. Since it is the households with the lowest active time that have the maximum energy savings this result indicates that the households with the most energy saving have too little accumulated active time to be influenced by the threshold values.

Fig. 6 shows the cumulative distribution function using ONU model 2, with Off state, and both link state thresholds. In this case, the maximum potential energy-saving reaches as high as 67%. With the 100 bps threshold, about 83% of the households can save more than 30% of the energy, whereas even in the 0 bps threshold case there are still 51% of households that can achieve more than 30% energy savings.

V. Conclusion

The study uses real traffic data traces from a residential FTTH access network to investigate energy consumption and energy saving solutions of energy-aware equipment. The paper has proposed a method of mapping users' traffic behavior to link states that then can be mapped to corresponding power consumption in the network devices. The results show that on average every household link in the access network can potentially save at least 18% energy consumption with sleep-mode enabled equipment. With an additional off-mode feature on ONU, the amount of energy can be further reduced about 14% when compared to sleep-mode only ONU. The link state threshold value has significant impact on the amount of energy savings.

Currently, 70% energy of the whole telecommunication network is consumed in the access network, when we put our experiment results of access network into an end-to-end network scope, we can see that the total potential energy savings of a whole network can reach up to 31% with energy-aware equipment. Even if in a strict case, when ONU Model 1 is used, without Off mode, and the link state threshold value of 0 bps is applied, the estimated end-to-end network energy savings can achieve 13%

Acknowledgment

This work has partly been financed by the Swedish Governmental Agency for Innovation Systems (VINNOVA) in the EFRAIM project, and Celtic-plus NOTTS project. Maria Kihl is a member of the Lund Center for Control of Complex Engineering Systems (LCCC) and the Excellence Center Linköping – Lund in Information Technology (eLLIIT).

References

[1] Index, Cisco Visual Networking. "Forecast and Methodology, 2012–2017." White Paper , 2013.

[2] M. P. Mills, “The cloud begins with coal: Big data, big networks, big infrastructure, and big power”, Technical Report, Digital Power Group, 2013.

[3] J. Malmodin, Å. Moberg, D. Lundén, G. Finnveden, and N. Lövehagen, "Greenhouse gas emissions and operational electricity use in the ICT and entertainment & media sectors." Journal of Industrial Ecology 14.5, 2010, pp. 770-790.

[4] L. Chiaraviglio, M. Mellia, and F. Neri. "Energy-aware backbone networks: a case study." Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on. IEEE, 2009.

[5] 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."Communications Surveys & Tutorials, IEEE 13.2 , 2011, pp. 223-244.

[6] L. Christoph, D. Kosiankowski, R. Weidmann, and A. Gladisch. "Energy consumption of telecommunication networks and related improvement options." Selected Topics in Quantum Electronics, IEEE Journal of 17.2 , 2011, pp. 285-295.

[7] A. Otaka, "Power saving ad-hoc report.", Technical presentation, IEEE 802 LAN/MAN standards committee, 2008.

[8] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan. "A power benchmarking framework for network devices." NETWORKING 2009. Springer Berlin Heidelberg, 2009. 795-808.

[9] J.-I. Kani, S. Shimazu, N. Yoshimoto, and H. Hadama, "Energy efficient optical access network technologies," Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference , 2011, vol., no., pp.1,3, 6-10.

[10] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall; "Reducing Network Energy Consumption via Sleeping and Rate-Adaptation." NSDI. Vol. 8. 2008.

[11] C. Gunaratne, K. Christensen, B, Nordman, and S. Suen. "Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR)," Computers, IEEE Transactions on , vol.57, no.4, pp.448,461, April 2008.

[12] K. Christensen, P. Reviriego, P., B. Nordman, M. Bennett, M. Mostowfi, and J. A. Maestro, "IEEE 802.3 az: the road to energy efficient ethernet." Communications Magazine, IEEE 48.11 ,2010, pp. 50-56.

[13] P. Reviriego, J.A. Hernández, D. Larrabeiti, and J.A. Maestro. "Performance evaluation of energy efficient ethernet." Communications Letters, IEEE 13.9 ,2009, pp. 697-699.

[14] G. Hasslinger, J. Mende, R. Geib, T. Beckhaus, and F. Hartleb, “Measurement and characteristics of aggregated traffic in broadband access networks,” in Managing Traffic Performance in Converged Networks, vol. 4516 of Lecture Notes in Computer Science, pp. 998–1010, Springer-Verlag Berlin Heidelberg, 2007.

[15] M. Kihl, C. Lagerstedt, Å. Aurelius and P. Ödling. "Traffic analysis and characterization of Internet user behavior." Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2010 International Congress on. IEEE, 2010.

[16] C. Lange, and A. Gladisch. "Energy efficiency limits of load adaptive networks." Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC). IEEE, 2010.

[17] Valerie Chaillou. (2012). Inventory of FTTH/B in Europe. IDATE DigiWorld Institute. Available: http://www.idate.org/en/News/Inventory-of-FTTH-B-in-Europe_765.html, retrieved: March, 2015

[18] “Procera networks.” http://www.proceranetworks.com, retrieved: March, 2015

[19] European Commission. "Code of Conduct on Energy Consumption of Broadband Equipment ver. 4.1." (2013).

A Dynamic Scalable Cryptosystem Based-on Reduced Key Size ECC

Jia Wang, Lee-Ming Cheng
Department of Electronic Engineering
City University of Hong Kong
Kowloon, Hong Kong
E-mail: jwang244-c@my.cityu.edu.hk

***Abstract*—This paper proposed a dynamic scalable cryptosystem based on reduced key size Elliptic Curve Cryptography (ECC) used for low-cost mobile devices or Radio Frequency Identification (RFID)-like devices, which are extremely constrained in memory, computing power and battery life. Supplementary algorithms including dynamic parameter assigning, random base point selection and session base point synchronization are designed in order to enhance the security level of the system. Synchronization is also performed for the consistency of the curve used between the server and the client. The proposed approach will provide a composite *ρ-sec* of (110+22) = 132 >128 and requires less storage and computing power, which provides good security alternative for the wireless transformation between devices with simpler and smaller processors.**

Keywords-ECC; scalable; cryptosystem; RFID security.

I. INTRODUCTION

Recent years have seen the great convenience brought to all walks of life by various kinds of applications built around the overly general concept called the Internet of Tings (IoT). A growing number of distributed diverse embedded devices get connected to the IoT platforms, sharing the facility as well as the security threatening due to the availability of tracking devices on communicating via RF. Security assurance, e.g., access control and eavesdropping prevention, must be guaranteed on the IoT before sending out of ID; mechanisms should be established to address the privacy concerns [1][2] on IoT. Deploying of a IoT in new context [3] which is not specifically designed may also imposing threats to system.

Traditional cryptographic algorithms used in protocols proposed for transformation security, such as the widely adopted Diffie-Hellman (DH) or Rivest, Shamir and Adleman (RSA) algorithms, often involve intensive computations and create a big challenge for the embedded devices in the IoTs which are usually designed with small and simple processors for low-cost purpose. Recently, the breakthrough in analysis discrete logarithm (DL) problem by Barbulescu et al [4], Adj et al [5] and Granger et al [6] has leaded into the belief that DL based algorithms like DH and RSA will soon to break and will be phased out much earlier than its' anticipation. The new challenge facing DH and RSA over time in term of being crypto-analysis and the proliferation of smaller and simpler devices hold them back to be used in the area of IoT cryptography.

Meanwhile, Elliptic Curve Cryptographic (ECC) algorithms, which developed by Neil Kobliz and Victor Miller in 1985, requires smaller key size and less power consumption while providing equivalent security compared with other crypto-algorithms, illustrated in Table I. This feature leads to significant performance advantages especially for IoT platforms where computing power, memory and battery life of devices are extremely constrained.

However, ECC belongs to the class of stream cipher and their hardly scalable and complex in operations are the major obstacles of making them to be popular when compared with RSA. In addition, the traditional way of implementing secure elliptic curves will not be suitable to provide a solution for unified platform for a diversified devices and servers with processors ranging from 8 to 256 bits as it requires at least 200 bits key size capable scalable to 521 bits or more. In order to address the problems mentioned above, in this paper we proposed a dynamic scalable cryptosystem based on reduced key size elliptic curves with supplementary algorithms to enhance its security. The supplementary algorithms include Dynamic Parameter Assigning (DPA), Random Initial Base Point Selection (RIBPS) and Session Base Point Synchronization (SBPS). A synchronization protocol is also designed for the consistency of the curve used between the server and the client.

TABLE I. SECURITY COMPARISON OF DIFFERENT CRYPTO-ALGORITHMS

Security Equivalent	**Bit/modular bit size**					
Symmetric Algorithm	56	80	112	128	192	256
RSA&DH	512	1024	2048	3072	7680	15360
ECC	112	160	224	256	384	521

The rest of the paper is organized as follows. Section II gives an overview of ECC. The architecture of the proposed dynamic scalable cryptosystem based on reduced key size ECC is given in Section III. In Section IV we introduce the synchronization protocol designed for the consistency of the transformation between the server and the clients. Simulation results are presented in Section V. The security and complexity are analyzed in Section VI. Finally we summarize the paper and discuss our future work in Section VII.

II. Overview of ECC

In this section, we give a brief introduction of ECC. Good reference for ECC could be found in [7].

The notations used are defined as follows.

- q is the order of the underlying finite field.
- F_q is the underlying finite field of order q.
- E is an elliptic curve defined over F_q.
- $E(F_q)$ is the set of points on E both of whose coordinates are in F_q, together with the point at infinity.
- P is a point in $E(F_q)$.
- Q is another point in $E(F_q)$.
- a and b are elements of finite field F_q, c is a integer.
- n is the order of the point P.
- k is a random integer selected in the interval $[2,n\text{-}2]$.

Elliptic curves used in cryptography are plane curves defined over finite field which consists of points satisfying the equation $y^2 = x^3 + ax^2 + \mathrm{b}$, along with the infinity point O. $E(F_q)$ together with the group operation of elliptic curves construct an Abelian group. Rules of elliptic curve arithmetic such as point addition, multiplication could be found in [7]. The order of point P is defined as the smallest positive integer n such that $nP = O$, which are usually large prime numbers in practical applications. ECC uses points in $E(F_q)$ as the elements for encryption. When a specific curve is chosen, P is randomly selected from all the points on the curve to generate the public key Q by field multiplication $Q = kP$, where k is called the private key. Q is used for encryption/signature verification and k is used for decryption/signature generation. The curve $E(F_q)$ and the base point Q are public in typical ECCs.

An example is given as simple application of ECC. Suppose *Alice* wants to sends *Bob* messages secretly. For *Alice*, she will randomly generate an integer k_a and compute $Q_a = k_aQ$ and then make Q_a public. *Bob* will randomly generate an integer k_B and compute $Q_b = k_bQ$ and make Q_b public.

To send message M_a to *Bob*, *Alice* will do:

1) Calculate the cipher $C_a = M_a + k_aQ_b$.
2) Send C_a to *Bob*.

To decrypt massage M_A from *Alice*, *Bob* will do:

1) Calculate $D_a = k_bQ_{a.}$
2) Decrypt $M_a = C_a – D_a$.

The security of ECC relies on the difficulty of solving the Elliptic Curve Discrete Logarithm Problem (ECDLP), that is given E defined in F_q and two points P, $Q \in E(F_q)$, find a integer k such that field multiplication $kP = Q$, provided such an integer exists. This problem is considered computationally infeasible to solve.

Not all elliptic curves are secure. Super singular curves could be cracked much faster over none singular one and a great deal of effort have been expended at finding curves that suitable for cryptography. Aranha et al [8] proposed the use of Curve22519 and Curve1174 which are Montgomery curve and Edward curve in the form of $y^2 = x^3 + ax^2 + x$ and $x^2 + y^2 = 1 + ax^2y^2$ respectively with prime $2^b - c$.

According to [8], more secure curves can be designed using different '*a*'s and '*b*'s. A set of known secured Montgomery curves with *(a, b, c)* are given in Table II. These parameters can be utilized to building a scalable structure by presetting the parameters *(a, b, c)*, in term of different key sizes.

TABLE II. Scalable ECC using various curve parameters

	ρ-sec	*a*	*b*	*c*	No. (*l*)
Montgomery Curve	110	117050	221	3	0
	128	486662	255	19	1
		61370	256	189	2
		240222	256	765	3
		55790	254	$-127\text{x}2^{240}-1$	4
	192	2065150	383	187	5
	256	530438	511	187	6

TABLE III. Architecture of the Curve List

No. (*l*)	Curve Parameters				Initial Point (*P*)
0	a_0	b_0	c_0	n_0	P_0
1	a_1	b_1	c_1	n_1	P_1
2	a_2	b_2	c_2	n_2	P_2
……	…	…	…		……
L-1	a_{L-1}	b_{L-1}	c_{L-1}	n_{L-1}	P_{L-1}

III. Architecture of the Proposed Dynamic Scalable Cryptosystem

The dynamic parameter assigning is dedicated to selecting a preset group of shared parameters '*a*'s, '*b*'s and '*n*'s, which constructed a reduced secure curve, details are demonstrated in part *A*. Since a new set of curves requires computing a group of new base points, the traditional way of point generation will not be appropriate. A random initial base point selection approach [9] needs to be established. A cache will be designed to house these initial points. The Parameter Synchronization method similar to [10] is needed to provide a session base point selection.

A. Dynamic Parameter Assigning

This approach is very similar to the minimal list approach for use in RFID [11]. RFID tags store a list of random parameters or pseudonyms for authorization [11]. Each time the tag is requested, it will send out the next pseudonym in the list, going back to the beginning when the pseudonym is exhausted. The proposed DPA works similar to pseudonym flow where the DPA list holds a set of parameters and each time it is requested, the next set of parameters will be sent out and cycling back to its beginning when it hits the end of the list.

B. Random Initial Base Point Selection

Input dynamic parameters *a*, *b*, *c* and *n* (*n*>0) as integers, the algorithm will select an effective random point *P* on the desired curve and form the base point by calculating *kP*. The scalar multiplication value is used to verify the correctness of base point selection on an elliptic curve.

The base point choice algorithm of ECC on Montgomery Curve is given as an Example below.

Input: *a, b, c, n, k*.

Output: Effective base point *G*.

Steps:

1) Randomly choose *x* (0<*x*<*n*).
2) Calculate $y^2 = (x^3 + ax^2 + x) \bmod (2^b - c)$.
3) Check if v belongs to quadratic residue of $\bmod (2^b - c)$, if so y is found, select $P = (x, y)$ go to 4), if not, go to 1).
4) Compute $G = kP$, then check whether *G* meets $y^2 = (x^3 + ax^2 + x) \bmod (2^b - c)$ and *G* is not infinite point. If so, *G* is set to be the base point, and go 5), if not, go to 1) .
5) Return *G*.

As shown in Table III, each curve in the list has been allocated some space to store the initial point *P* and the value of *P* should be overwritten as *G* each time after the curve having been used for security enhancement.

C. Session Base Point Synchronization

De-synchronization will raise concern in real applications although it can provide protection on denial-of-service attacks. One possible approach to solve de-synchronization problem is to maintain a list not just of current *parameters*, but also of values from several future time-steps. This approach is similar to that of [10] involving tag resynchronization by checking plus and minus one step parameters when de-synchronization occurs. The advantage of our proposal is that it would permit a certain amount of synchronization, but would still not leak any sequence values.

IV. Synchronization mechanism of the server and the clients

A. Symbol notations

- S server
- C client
- m Pseudo-Random Number Generated by the server's PRNG
- r Pseudo-Random Number Generated by the client's PRNG
- *f(x)* hash function of x
- L total number of curves in the list
- l the randomly selected curve, 0<=*l*<=*L-1*

B. Protocol Flow

As shown in Figure 1, the protocol sequences are as follows:

1) *S* generates a random number *m* of 16 bits by its PRNG and sends it to the client *C*.
2) *C* generates a random number *r* of 16 bits and sends it to *S*.
3) *S* forms the message $M_1 = r\ Xor\ m$, $M_2 = (M_1\ mod\ L)\ Xor\ l$, and $f(M_1)$, then sends *M*2 and $f(M_1)$ to *C*.
4) *C* computes $f_1 = f(r\,Xor\ m)$ and compares it with $f(M_1)$ if $f_1 = f(M_1)$ then computes $l = ((r\ Xor\ m)\ mod\ L)\ Xor\ M_2$ as the selected curve.

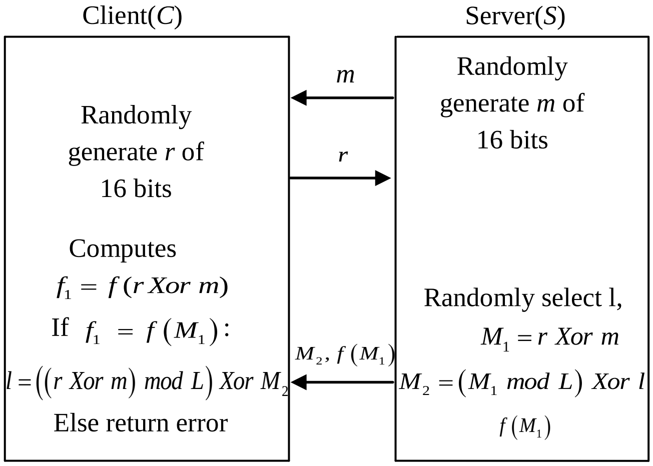

Client(C)
Server(S)
m
r
Randomly generate r of 16 bits
Computes $f_1 = f(r\,Xor\ m)$
If $f_1 = f(M_1)$:
$l = ((r\ Xor\ m)\ mod\ L)\ Xor\ M_2$
Else return error
Randomly generate m of 16 bits
Randomly select l,
$M_1 = r\ Xor\ m$
$M_2 = (M_1\ mod\ L)\ Xor\ l$
$f(M_1)$
$M_2, f(M_1)$

Figure 1. The proposed synchronization protocol.

An example is given here.

Suppose that there are 32 curves in the list which are numbered from 0 to 31, let $l = 11$. Assuming that $m = 0111010101010101,\ r = 1010101111001011$

First of all, the server sends *m* to the client and the client returns r back to the server. The server then computes:

M1 = r Xor m

$$= 1010101111001011 \; Xor \; 0111010101010101$$
$$= 1101111010011110$$
$$M2 = 11110 \; Xor \; 01011$$
$$= 10101$$

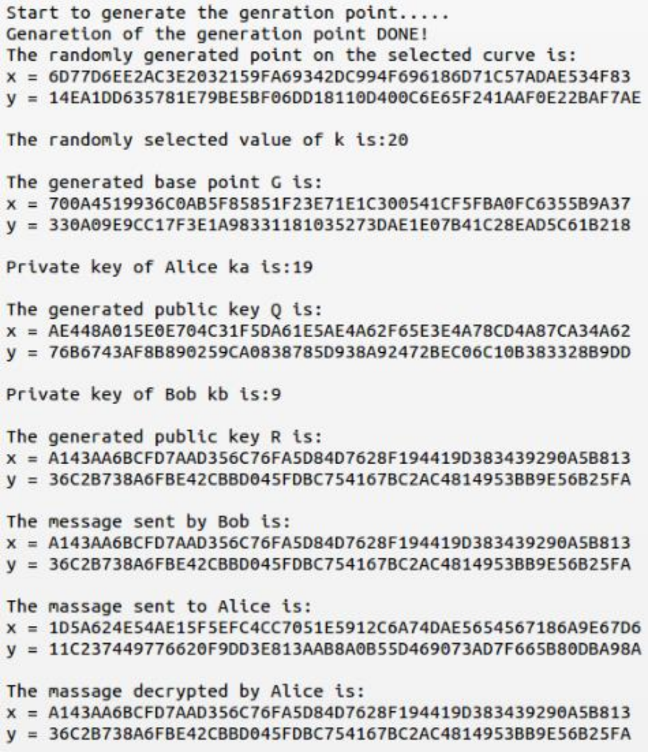
Suppose the transformation of *r* and *m* is performed in secure channel and f_1 equals to $f(M_1)$, then the client calculates:

$$(r \; Xor \; m) \bmod L = 11110,$$
$$l = ((r \; Xor \; m) \bmod L) \; Xor \; M_2$$
$$= 11110 \; Xor \; 10101$$
$$= 01011$$

The result indicates that the curve numbering 11 is chosen to be used by the server in the cryptography.

V. Simulation results

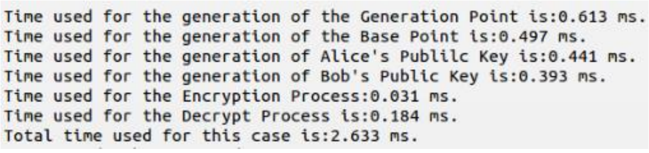
The proposed scheme implemented on MIRACL crypto library with GCC in C programming language on Linux platforms with 2.8GHz Intel processor i7 and memory 4.0GB.



```
Start to generate the genration point.....
Genaretion of the generation point DONE!
The randomly generated point on the selected curve is:
x = 6D77D6EE2AC3E2032159FA69342DC994F696186D71C57ADAE534F83
y = 14EA1DD635781E79BE5BF06DD18110D400C6E65F241AAF0E22BAF7AE

The randomly selected value of k is:20

The generated base point G is:
x = 700A4519936C0AB5F85851F23E71E1C300541CF5FBA0FC6355B9A37
y = 330A09E9CC17F3E1A98331181035273DAE1E07B41C28EAD5C61B218

Private key of Alice ka is:19

The generated public key Q is:
x = AE448A015E0E704C31F5DA61E5AE4A62F65E3E4A78CD4A87CA34A62
y = 76B6743AF8B890259CA0838785D938A92472BEC06C10B383328B9DD

Private key of Bob kb is:9

The generated public key R is:
x = A143AA6BCFD7AAD356C76FA5D84D7628F194419D383439290A5B813
y = 36C2B738A6FBE42CBBD045FDBC754167BC2AC4814953BB9E56B25FA

The message sent by Bob is:
x = A143AA6BCFD7AAD356C76FA5D84D7628F194419D383439290A5B813
y = 36C2B738A6FBE42CBBD045FDBC754167BC2AC4814953BB9E56B25FA

The massage sent to Alice is:
x = 1D5A624E54AE15F5EFC4CC7051E5912C6A74DAE5654567186A9E67D6
y = 11C237449776620F9DD3E813AAB8A0B55D469073AD7F665B80DBA98A

The massage decrypted by Alice is:
x = A143AA6BCFD7AAD356C76FA5D84D7628F194419D383439290A5B813
y = 36C2B738A6FBE42CBBD045FDBC754167BC2AC4814953BB9E56B25FA
```

Figure 2. Simulation results of the proposed scheme

In the implementation, we generated a curve table using a set of nonsingular ellicptic curves with shorter key-sizes and then labelled them with curve numbers. By using schronizaiton mechanics proposed in Section IV, *Alice* chose a curve and shared with *Bob*. In this experiment case, the curve with *No. 0* is selected. Without loss of generality, the private keys of *Alice* and *Bob* are randomly generated as small integers for simplicity. The simulation result is shown in Figure 2. Details of time consumption of this case are inllustrated in Figure 3.



```
Time used for the generation of the Generation Point is:0.613 ms.
Time used for the generation of the Base Point is:0.497 ms.
Time used for the generation of Alice's Publilc Key is:0.441 ms.
Time used for the generation of Bob's Public Key is:0.393 ms.
Time used for the Encryption Process:0.031 ms.
Time used for the Decrypt Process is:0.184 ms.
Total time used for this case is:2.633 ms.
```

Figure 3. Time consumption of the experiment case

VI. Security and complexity analysis

In this section, we studied the proposed scheme according to its security level, storage requirement and power consumption, which are demonstrated in part *A*, *B*, *C* respectively.

A. Security level

Given the security of ECC denoted by ρ-sec in term of bit size [8], from Table II it is *c*lear that in order to achieve *ρ-sec* of 128, the ECC module bit size should be around 256. By randomly selecting lower ρ-sec curves generated with different *(a, b, c)* of the same *ρ-sec*, the same high level can be achieved because the random prime choice can bridge the extra security required. Table II gives examples of various configurations of Montgomery Curves with *ρ-sec*=128 [12].

In our simulation, *ρ-sec* curves of 110 is used, the prime bit size will be 220. Table III shows the bit size required for symmetric and asymmetric crypto-algorithms. According to RSA/DH security, a total of 2^{22} primes can be found. The randomly selected prime will provide addition *ρ-sec* of 22. This approach will provide a composite *ρ-sec* of (110+22) = 132 >128.

B. Storage requirements

As mentioned earlier, due to the use of randomly selected curve, the traditional way to generate the base point is not feasible as all the points would be stored. We store one initial point instead of storing all the points on the curve by using RIBPS, which extremely reduces the required memory space.

C. Efficiency and power consumption

Reduced key size elliptic curves are used in the proposed dynamic scalable cryptosystem in order to provide flexible security mechanism for the diverse devices and servers with processors ranging from 8 to 256 bits in the IoT platforms. As illustrated in Table IV, our scheme used reduced key-size curves to gain even higher security. Shorter key-size leads to great reduction of the computation load as well as running time.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-396-4

TABLE IV. EFFICIENCY COMPARISON

	ρ-sec	*Key-size (bit)*
Our scheme	> 132	221
Typical ECCs	128	256

VII. CONLUSION AND FUTURE WORK

A dynamic scalable cryptosystem based on reduced key size elliptic curves is proposed for the IoT platforms. Less storage and lower power consumption but higher level of security can be achieved with supplementary algorithms include dynamic parameter assigning, random initial base point selection and session base point synchronization. It could be used to address the security problems targeting IoTs, Cloud platforms and Cyber-Physical Systems, which usually connected with highly distributed low-cost smart devices, based point of view of ECC. It's expected to strengthen Cloud security and support Cloud revolution that enables unprecedented interconnection of networked processes operated in a blurred real and virtual world boundary.

The future work covers the development of novel verification method, as well as the building of the benchmark for evaluating the performance and the threat/risk from cyber-physical attacks.

ACKNOWLEDGMENT

This work supported by City University of Hong Kong Strategic Grant Number 7004228 is acknowledged.

REFERENCES

[1] J. A. Stankovic, "Research Directions for the Internet of Things,"Internet of Things Journal, IEEE, vol.1, pp.3-9, 2014, doi: 10.1109/JIOT.2014.2312291.

[2] B. Pranggono, Y. Yang, K. McLaughlin and S. Sezer, "Intrusion Detection System for Critical Infrastructure," The State of the Art in Intrusion Prevention and Detection, A-S. K. Pathan, Ed., ed:London, UK: CRC Press, pp.150-170, 2014.

[3] W. Aman and E. Snekkenes, "An Empirical Research on InfoSec Risk Management in IoT-based eHealth" MOBILITY 2013 : The Third International Conference on Mobile Services, Resources, and Users, 2013, pp.99-107, ISSN**:** 2308-3468, ISBN: 978-1-61208-313-1.

[4] R. Barbulescu, P. Gaudry, A. Joux, and E. Thom´e, "A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic," Advances in Cryptology— EUROCRYPT 2014, volume 8441 of LNCS, Springer, 2014, pp.1–16, doi: 10.1007/978-3-642-55220-5_1.

[5] G. Adj, A. Menezes, T. Oliveira and F. Rodr´ıguez-Henr´ıquez, "Weakness of $F_{36\cdot 509}$ for discrete logarithm cryptography," Pairing-based Cryptography—Pairing 2013, volume 8365 of Lecture Notes in Computer Science, Springer, pp. 20-44, 2013.

[6] R. Granger, T. Kleinjung, and Jens Zumbrägel, "Breaking '128-bit secure' super singular binary curves," CRYPTO (2) 2014, pp. 126-145, doi:10.1007/978-3-662-44381-1_8.

[7] Hankerson D, Menezes A, and Vanstone S, Guide to elliptic curve cryptography, Springer, Berlin, 2004.

[8] D.F.Aranha, P.S.L.M. Barreto, G.C.C.F. Pereira, and J.E. Ricardini, "A note on high-security general-purpose elliptic curves," Cryptology ePrint Archive, Report 2013/647 (2013), Available from: http://eprint.iacr.org/.

[9] M. Roy1, N. Deb, and A. J. Kumar, 2014. "Point Generation And Base Point Selection In ECC," An Overview, International Journal of dvanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6711-6713, May 2014.

[10] L.M. Cheng, CW So, and L.L. Cheng. An improved forward secrecy protocol for next generation EPCGlogal tag, Development and Implementation of RFID Technology, I-Tech Education and Publishing KG, Editor Cristina Turcu , Jan. 2009, pp. 317-332, ISBN 978-3-902613-54-7.

[11] A. Juels, "Minimalist cryptography for low-cost RFID tag," Conference on Security in Communication Networks – SCN'04, LNCS, Amalfi, Italia, September (2004), Springer-Verlag, pp. 149-164, 2004.

[12] J. W. Bos, C. Costello, P. Longa, and M. Naehrig, Selecting elliptic curves for cryptography : An efficiency and security analysis, IACR Cryptology ePrint Archive, 2014:130, 2014. Available from: http://eprint.iacr.org/.

Analysis of Prefetching Schemes for TV-on-Demand Service

Manxing Du*†, Maria Kihl†, Åke Arvidsson‡§, Christina Lagerstedt* and Anders Gavler*
*Acreo Swedish ICT, Sweden, Email: firstname.lastname@acreo.se
†Dept. of Electr. and Inform. Technology, Lund University, Sweden, Email: firstname.lastname@eit.lth.se
‡Business Unit Support Solutions, Ericsson, Sweden, Email: firstname.lastname@ericsson.com
§Department of Computer Science, Kristianstad University, Sweden, Email: firstname.lastname@hkr.se

***Abstract*—TV-on-Demand service has become one of the most popular Internet applications that continuously attracts higher user interests. With rapidly increasing user demand, the existing network conditions may not be able to ensure low start-up delay of video playback. Prefetching has been broadly investigated to cope with the start-up latency problem which is also known as user perceived latency. In this paper, we analyse request patterns for TV programs from a popular Swedish TV service provider over 11 weeks. According to the analysis, we propose a prefetching scheme at the user end to preload videos before user requests. Our prefetching scheme significantly improves the cache hit ratio compared to terminal caching and we note that there is a potential to further improve prefetching performance by customizing prefetching schemes for different video categories. We further present a cost model to determine the optimal number of videos to prefetch. Finally, we discuss available time for prefetching and suggest that when to make prefetching decisions depends on the user demand patterns of different video categories.**

Keywords–TV-on-Demand services; user perceived latency; prefetching;

I. INTRODUCTION

Nowadays, other than air broadcasting, cable networks and physical media like Video Home System (VHS) or Digital Video Disc (DVD), Internet has become a popular medium for distributing multimedia content like TV shows, movies, and user generated videos. The massive amount of multimedia traffic has imposed a significant burden on the Internet. Consequently, users sometimes have to endure long access delay for filling up the playout buffer before the content is displayed. In [1], the result shows that user's tolerance of waiting time for downloading web pages is about 2 seconds. Results in [2] suggest that the more familiar users are with a web site, the more sensitive they are to delays. Although web caching is widely used as a solution to lessen the web traffic congestion and improve the network performance, the benefit of caches is limited. To further reduce the user perceived latency (UPL), prefetching has become a popular technique. The objective of the prefetching system is to proactively preload certain content to the cache even before users request it.

Thorough summaries of web caching and prefetching approaches and performance measures can be found in [3] [4] [5]. Domènech *et al.* in [6] compared different prefetching architectures and found that the maximum latency reduction of 67.7% can be obtained if the predictor is placed at a proxy while the collaborative prediction between proxy and server can reduce the latency more than 97.2%. However, the results are obtained based on the most ideal scenario that the prediction can be 100% correct, thus the results can be seen as the upper limits of latency reductions.

Most of the existing prefetching approaches are access-history based which predict user future requests depending on the observed content access patterns. Márquez *et al.* applied a Double Dependency Graph (DDG) prediction algorithm to a mobile web and observed that the performance of prefetching approaches rely on the underlying networking technologies [7]. Another history based model is the Markov model, which is an effective scheme to predict what users intend to request based on the sequence of the historical access [8]–[11]. The prefetching schemes proposed in [12]–[15] use the data mining technique to discover users' access patterns.

Popularity-based prefetching approches are also widely used, especially for prefetching multimedia content. In [16], a trace driven simulation was performed to investigate prefetching schemes for YouTube videos. Their prefetching scheme is to prefetch the top 25 videos from each video's related video list to a proxy server. Combining both prefetching and conventional proxy caching, 80.88% hit ratio can be obtained and it only introduces 2% increase in the network load.

However, this recommendation-based prefetching scheme requires an effective recommendation system which can be a big challenge. Krishnappa *et al.* in [17] applied a prefetching top-100 videos scheme on a trace of Hulu traffic in a campus network and compared the performances of prefetching and conventional proxy caching which proves prefetching is very effective for online TV service.

From these papers, we note that the research focus is mainly on proxy prefetching whereas the performance of terminal prefetching is less well known. To the best of our knowledge, our analysis is the first that focuses on using terminal storage to do prefetching for the TV-on-demand service. To prefetch content to the terminal can eliminate the delay between proxy and user and further reduce the playback start-up latency preceived by users.

In this study, we have used data from one of the most popular Swedish TV channels which provides all their broadcasting content online for subscribed users. Each TV-on-Demand program consists of a series of episodes which have high consistency regarding content, thus the index of each episode is a good indicator of user's future access. We propose to use the intrinsic structural information of episodes belonging to a TV series to make prefetching decisions. The criterion for prefetching is based on the index of episodes within each TV program. First, we analyse the potential of prefetching in our dataset, followed by investigating the optimal choice of prefetching. We show that high terminal gain can be obtained by prefetching two adjacent episodes in a series for each viewed episode with minimum cost. This study shows the potential of implementing terminal prefetching for TV-on-

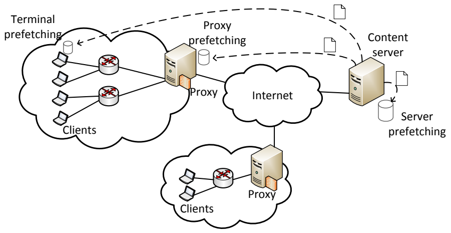


Figure 1. Network architecture with prefetching system

Demand service both effectively and economically. In addition, we investigate whether there is enough time to perform prefetching.

The remainder of the paper is organized as follows. Section II describes the infrastructure of a prefetching system and the evaluation metrics. Our dataset and prefetching methods are presented in Section III. Section IV shows the results and evaluations of our prefetching scheme. In Section V, we further discuss the available time for prefetching and we conclude in Section VI.

II. Video prefetching system

By implementing a prefetching system, the multimedia streams can be retrieved from the content provider and saved in a cache. In this way, users can quickly get access to their preferred content. In this section, we first introduce the components of a prefetching system, and then we present the evaluation metrics used for measuring the performance of prefetching schemes.

A. The infrastructure of a prefetching system

The prefetching system consists of two elements: prediction engine and prefetching engine. The prediction engine is responsible for forseeing what users would watch next before they request the content. The prefetching engine proactively stores the video prefix to the local cache. In this case, both the first-time views and repeats of the prefetched videos can be served from the local cache with short start-up delay. These two engines can be placed at any part of the web architecture which is shown in Figure 1: terminals [16], proxies [13] [17], servers [18] [19] or between these elements [6] [7].

In order to bring content closer to end users, so as to the most extent reduce user's perceived latency, we assume the prefetching engine is implemented at the user end which is called terminal prefetching. This means that the prefetched content will be stored in a terminal cache at the user end. Anything from a short part of the video to the entire video can be prefetched. Typically, to reduce the start-up latency, there is no need to prefetch the entire video, which requires more network resources, thus it may introduce longer delay. A solution is to prefetch only parts of video streams. When a user requests a video, it can be played directly from the cache which gives more time for setting up the transport and filling up the playout buffer for the remaining parts of the video. Considering that users may not all favour the beginning of each video, in [20] [21], the performance of caching different parts of the videos are discussed. In this paper, we ignore the portion of prefetched video and treat each video as a video unit, thus the cache size is not considered. In this scenario, every video request from a user is first sent to a prefetching engine which checks whether the video has already been retrieved earlier or not. If so, the content will be served from the local cache instead of the remote server.

B. Evaluation metrics

In order to evaluate the performance of a prefetching scheme, three metrics are used in this study. The first metric is precision (P) which is defined as the number of videos that are prefetched and requested by users (h_v) over the total number of prefetched videos (p_v).

$$P = \frac{h_v}{p_v} \tag{1}$$

A high precision value suggests that more prefetched content is requested by end users, thus the prediction engine works efficiently.

The second metric is recall (R) which is the share of prefetched and requested videos (h_v) over total number of requested videos (r_v).

$$R = \frac{h_v}{r_v} \tag{2}$$

Higher recall indicates that a larger share of the content requested by users can be correctly predicted by the prediction engine.

Precision and recall are constrained by each other. In principle, higher recall could be obtained by prefetching as many videos as possible in order to cover more content. Thus it would be more likely that the prefetched content contains the videos which will be requested by users. However, in this case, p_v will increase. Consequently, the prediction engine's precision will decrease. To prefetch a great amount of data which will not be used by users will also deteriorate the network congestion. These two metrics need to be balanced when designing a prefetching system. F_1 score (balanced F-score) [22] is a weighted average of the precision and recall which is used in this study to show the effectiveness of a prediction. The closer F_1 score is to 1, the more effective the prediction is.

$$F_1 = 2\frac{PR}{P+R} \tag{3}$$

To evaluate whether the prefetched content are highly demanded by users, we introduce the last metric which is the cache hit ratio (H). It is defined as the number of requests to videos (h_r) which are retrieved from the prefetching cache over the total number of requests (t_r).

$$H = \frac{h_r}{t_r} \tag{4}$$

Cache hit ratio shows the share of repeated requests for videos which can be served directly from the cache. A high hit ratio suggests that more requests can be served with reduced start-up delay and a high utilization of the prefetched content.

TABLE I. EXAMPLE OF USER REQUEST

Title	The bridge episode 13
User Id	1044197
Start time	2012-12-31 09:00:08
End time	2012-12-31 09:21:48
Viewing minute	22
avgBitrateMbps	2.599
Program	The bridge
Category	TV series

III. EXPERIMENT

In this section, we will describe the experiments conducted using the prefetching method in this paper. The objective of our analysis has been to investigate which episodes to prefetch for each viewed video to obtain the best performance of the prefetching system. The analysis is carried out by measuring the performance of prefetching and terminal caching in terms of effectiveness and hit ratio. To optimize the number of videos to prefetch, a cost model is proposed to find an appropriate trade-off between the cost of prefetching and the potential of response time improvement.

A. Dataset

Our study is conducted using the video requests from one of the most popular Swedish TV providers. The data was collected by Conviva which provides online video analytic solutions to media content providers around the world. The data is based on recorded TV requests for a subset of users in a city in Sweden. The users are so called subcribed users who can get access to all the TV content online provided by that TV channel by paying a monthly fee. Thus the users who do not have subscriptions are not included in this study either. All the users are anonymized and no data can be traced back to any specific user.

There are nine video categories defined by the service provider as follows: *Children, Documentary, TV series, Home and leisure time, Entertainment, Default, Mixed, Sports and News.* There is too little data in the *Default* and *Mixed* category and usually the *Sports* and *News* content are distributed by live streaming which cannot be prefetched like TV-on-demand content. Therefore, in the following sections, only the TV-on-demand content which is categorized as *Children, Documentary, TV series, Home and leisure time* and *Entertainment* are included and all the TV programs in each category consist of a series of episodes. To ensure unique identifiers for each episode, our dataset only includes requests of programs with only one season available.

Table I shows an example of the available information which is contained in each data entry. There are 7933 subscribed users who generated 104845 requests to 2427 videos which belong to 253 series over 11 weeks from December 31, 2012 to March 17, 2013. Figure 2 depicts the Cumulative Distribution Function (CDF) of viewing time per request. Around 20% of the total requests result in viewing times smaller than 2 minutes. We filtered out these short viewing sessions to eliminate the impact of user's random clicks which are not suitable to serve as predictors. We should also note that users may request the prefetched videos after the end of time period in our dataset. Thus, the result of hit ratio is underestimated due to the finite time period.

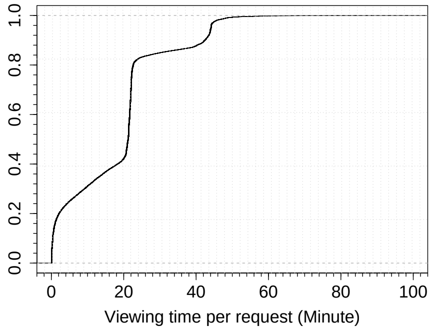


Figure 2. CDF of viewing time per request

B. Video prefetching selection methods

In order to implement prefetching, the prediction engine needs to determine what content should be preloaded based on user's viewing histroy of the same series. In the following, we propose and evaluate a scheme for the prediction engine to make prefetching decisions.

In this study, we consider two extreme cases as baselines for evaluating the performance of the proposed prefetching scheme. One is to prefetch all the available episodes in a series to the end user as long as a user watched an episode in that series. In reality, the content provider has the information of the number of episodes in each TV series, both released and unreleased. In our dataset, the total number of episodes in each series is unknown, so we scan users' viewing histories and collect unique video sets of each series. We assume the videos in each set are all the available videos in each series. This coarse scenario may waste significant amounts of bandwidth since some of the content would never be accessed by users.

The other extreme scenario is conventional terminal caching which passively caches all of user demand and nothing will be preloaded prior to user requests. Comparing with prefetching, terminal caching is also called passive caching. When terminal caching is enabled, the repeated requests for each video can be served directly from the local cache. Terminal caching can help to reduce the initial delay only if the cached video is requested again.

Our approach is based on the intrinsic structure of TV content. Since each TV program contains a series of episodes which have high consistency and similarity in content, we propose to prefetch N adjacent episodes for each viewed episode. Different from videos on the traditional VoD websites like YouTube, a TV series consists of a series of episodes which will be released regularly. User's request patterns of episodes in the series will be examined so that the prediction engine could decide which episodes to prefetch.

C. Cost model definition

Any prefetching scheme makes inaccurate predictions which inevitably downloads more content than a system without prefetching, consequently, the traffic overhead caused by prefetching may impose more burden on a bandwidth sensitive network. The congested network may lead to packet loss, longer transmission delay and poor quality of experience (QoE). Nervertheless, prefetching more content increases the probablity of meeting user's demand and potentially reduces the user's perceived latency. Sometimes spare network capacity is available during off-peak hours, e.g, during nights. It can be profitable to prefetch as much content as possible during that time to achieve high hit ratio. Hence, we propose a cost model to quantify the cost of prefetching in order to choose optimal number of videos to prefetch.

We assume the cost of real time video delivery equals 1 monetary unit per video and the cost of off-peak prefetching is x monetary unit per video which is smaller than 1, since prefetching can be done during the off-peak hours which costs less than the real time downloading. Besides, the possible cost for poorer QoE can also be seen as a reason that real-time downloading is more expensive than prefetching.

We define the number of videos which are requested by each user but not prefetched as video miss (M). The number of videos which are prefetched by each user is P. The cost of prefetching for n users is:

$$C = 1 \cdot \sum_{i=1}^{n} M_i + x \cdot \sum_{i=1}^{n} P_i, \quad 0 < x < 1 \tag{5}$$

IV. Results

In this section, we present how to find the optimal number (N) of episodes to prefetch in order to achieve a high hit ratio which represents the potential of improving the response time.

We first show the potential of the prefetching scheme, followed by comparing the prefetching performances when different values of N are selected. The cost metric can be seen as a metric of measuring performance of each prefetching scheme regarding transport cost, QoE cost and so forth. It helps to make optimal prefetching decision when network condition changes.

A. The potential and benchmarks of prefetching

Before we get into depth of prefetching schemes, it is essential to evaluate the potential of prefetching. We assume a clairvoyant scheme that once an episode of each series has been watched by a user, the following episodes in that series which he will watch later can be 100% predicted and prefetched by the prefetching system. In this case, only the first requested video in each series cannot be predicted and preloaded. The precision of this prediction is 100%. In this case, the optimal recall equals 73% which means in principle, 73% of the clicks to new videos are predictable. This result suggests that, if all the episodes which are watched after the first one can be correctly predicted and stored in the local terminal cache, 73% of the requests to new videos will be served without delay. The corresponding F_1 score equals 0.84 which can be seen as the upper limit of the prediction effectiveness that our study can obtain based on this dataset.

In order to evaluate the performance of a prefetching scheme, the two extreme cases decribed in section III-B serve as benchmarks for comparison. First, we present the non-prefetching system which only has enabled terminal caching. The terminal cache yields a hit ratio of 13.77% which means even with an infinite terminal cache, only 13.77% requests can be served from the local cache. From this result, it shows a great potential of prefetching since terminal caching leaves over 85% of the requests unattended.

The other extreme case is to prefetch all the episodes in a series when an episode is watched, the cache hit ratio can reach up to 77%. This result is the upper limit of the hit ratio that the prefetching system can achieve in this study. When all the episodes are prefetched, the maximum recall value 73% is obtained. However, the precision is only 17% since the prefetching engine prefetched too much redundant data which users are not interested in. As a result, the effectiveness of this prefetching scheme is only 0.28. Clearly we need to intelligently select the content to prefetch and store.

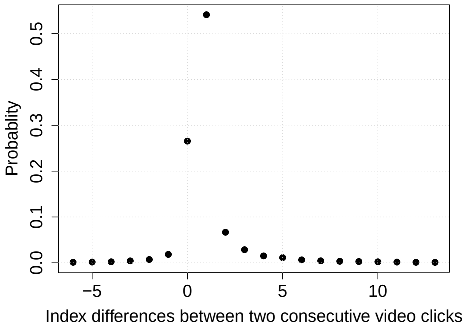


Figure 3. Transition probability of user requests within the same series

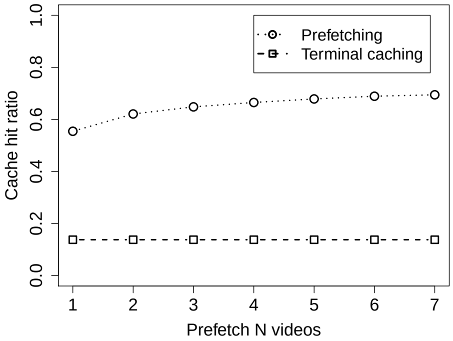


Figure 4. Hit ratios of terminal cache and prefetch

B. Prefetch N episodes

In order to avoid prefetching too much data and deteriorating network congestion, we propose to limit the number of videos to be prefetched by using a prefetching scheme, which only prefetches N videos in each series for each user.

Figure 3 shows the probability that a request for an episode n will be followed by a request for episode $n+k$ as a function of k. We find that for a user that has watched episode n of a series, the probablity of that user watches episode $n+1$ next is over 50%. Over our measurement period, there are about 26% of users that will not watch any episode after watching episode n. According to Figure 3, if we prefetch the videos with index of: $n+1$, $n+2$, $n+3$, $n-1$, $n+4$, $n+5$, and $n-2$, this will account for 95% of the requests for a next video.

Now we need to decide the value for N and which videos to prefetch. In Figure 4, we notice that when episode $n+1$ is prefetched, the hit ratio can reach up to 55% which is a big improvement comparing to the 13.7% hit ratio with terminal caching. To prefetch more videos besides $n+1$ and $n+2$ gives very little increase in hit ratio.

We also repeated the above analysis for different video categories. The request pattern of episodes in each video category

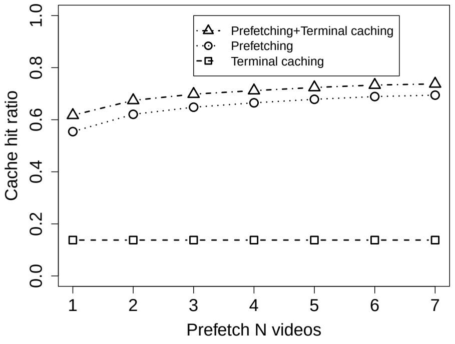


Figure 5. Hit ratio of combining prefetching and terminal caching

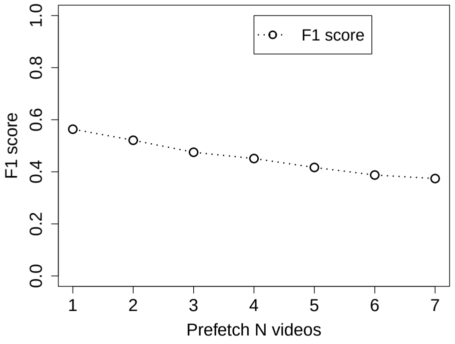


Figure 6. F1 score of prefetching system

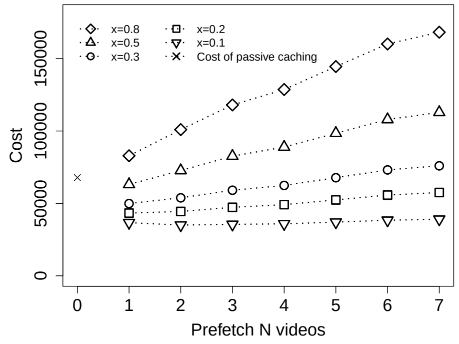


Figure 7. Total cost vs. Number of prefetched videos

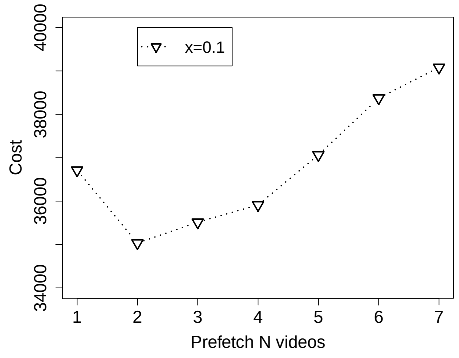


Figure 8. Total cost vs.Number of prefetched videos when cost factor = 0.1

were examined separately and we found that TV series and entertainment programs exhibit predictible consecutive request patterns while children's programs, documentaries and home and leisure time programs exhibit more random request pattern. When we prefetch videos according to the request pattern for each video category, the terminal hit ratio increase is about 1 percentage point comparing to the results in Figure 4. Even if the improvement is not significant, which may due to the limited number of videos in each category, the impact of customizing prefetched videos according to video category should be further investigated by using larger datasets.

Considering that prefetched videos and cached videos are two different sets, combining prefetching and terminal caching will increase the cache hit ratio, thus more content can be served from the cache with short start-up delay. Figure 5 shows the hit ratio improvement when prefetching and terminal caching are both adopted in a system. When a user requests a video which is not prefetched, the video will be downloaded from the server and at the same time cached at the user's device. By adding the passive terminal cache, hit ratio increases about 6 percentage points which represents the gain for not deleting the videos which are requested by users but not prefetched. In our study, we treat all requests of same video as cache hits. However, users may watch a part of a video and after some time continue to watch the rest of it. If we only treat the views after a user finishes watching the whole video as hits, the contribution from caching will be lower and the benefit brought by conventional passive terminal caching will be more limited.

Now, we present the effectivness of prefetching schemes with different values of N. Figure 6 shows the F_1 scores of the prefetching system when different N values are applied. In general, the more videos prefetched, the less accurate and less effective the prefetching system can be. However, the hit ratio increases when more videos are prefetched, as shown in Figure 4 and Figure 5. The decrease of effectiveness is caused by the amount of prefetched videos which are not requested by users. If the network condition is suitable to cope with these extra traffic, then a small hit ratio improvement with relatively large decrease of prefetching effectiveness is still profitable. Next, we apply the proposed cost model in the next section in order to quantify the cost of prefetching and to find the optimal number of videos to prefetch.

C. *Cost model*

In this section, we present the measurement of cost of prefetching as a performance metric to find the optimal number of videos to prefetch.

Each point on the curves in Figure 7 shows the prefetching cost value (see section III-C) versus the number of videos prefetched. Three curves represent the prefetching cost x in equation 5 equals 0.8, 0.5, 0.3, 0.2, and 0.1, respectively. The cross mark to the left shows the total cost of passive terminal caching when nothing is prefetched. Even though the top four curves show that the total cost of prefetching has a linear increase as more videos are prefetched, when off-peak downloading costs less than 30% of real time downloading, to prefetch up to 7 videos is still cheaper than passive caching. When off-peak downloading costs half of the real time downloading, to prefetch more than one video costs more than passive caching. But in this case, to prefetch only one video still outperforms passive caching. The curve of $x = 0.1$ is shown separately in Figure 8. An interesting phenomenon in Figure 8 is that the total cost of prefetching declines when N increases from 1 to 2. It suggests that when prefetching two videos, the decrease of cost generated by video miss (M) is faster than cost increase generated by prefetching more videos. However, when N is larger than 2, the cost of prefetching starts to raise again. It indicates that there are more additional prefetched videos which are not used by users. When the cost of off-peak downloading is 10% of the cost of real-time downloading, to prefetch two videos yields the least cost.

V. Time to prefetch

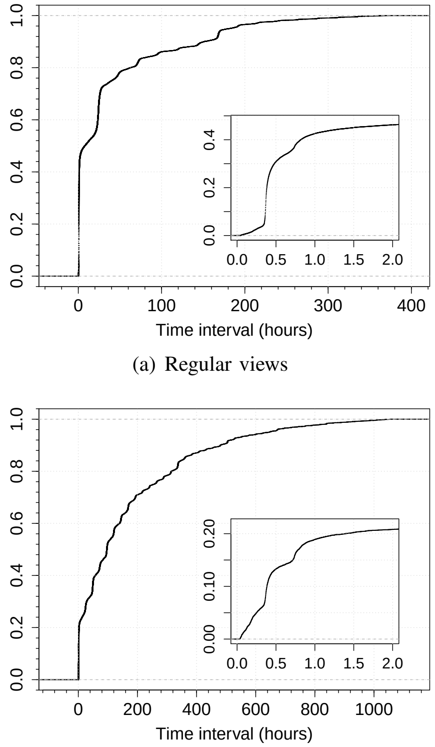
In this section, we estimate the available time for prefetching by measuring the time interval between two consecutive viewing sessions. The time between the start of a viewing session and the start of the next one is defined as the upper bound of the time to make prefetching decisions. The time between the end of a viewing session and the start of the next session serves as the lower bound of prefetching time. We differentiate the behaviour of watching the $n + 1$ video and watching any video not in sequential order as regular view and irregular view correspondingly.

As is shown in Figure 9(a), two steep increases occur at 24h and 1 week respectively. Some programs are released on a daily or weekly basis and the increases suggest that a number of users follow these programs at the same pace. The inner graph shows the same CDF but zooms into two hours scale. It shows that 30% of regular views arrive within 20 minutes. Comparing with the results in Figure 9(b) for irregular views, within 20 minutes, only less than 15% of the requests are generated. Figure 10 shows the CDF of the lower bound time. In general, it shows similar trend as in Figure 9. The difference is shown in the inner figures, which suggests that about 40% of regular views are generated within 1 minute after the end of a viewing session and only 15% of the irregular views are generated within the same time period. It indicates that if we choose to prefetch videos at the end of the current session, we have rather limited time. As is shown in Figure 3, the risk of prefetching for the regular view is less. Thus, it is more reasonable to prefetch the next video in order during the current session. For irregular views, the request pattern shows that people watch episodes in the same program on daily basis, which leaves us longer time for prefetching and it can be delayed until off-peak time.

Finally, we perform the same analysis for different video categories and we focus particularly on the lower bound time in this part. We observe that for TV series, 40% of the requests of the $n+1$ video are within 1 minute and 30% of the requests within 24 hours. It suggests that the next episode can be either pre-downloaded during the current episode's playtime or be downloaded during off-peak time. The entertainment programs show a very similar pattern to the TV series. For children's programs, 60% to 70% of the requests are generated within 3 minutes no matter which episode is watched next. Similar patterns are observed for the videos of home and leisure time programs and documentaries. It suggests that in this case to prefetch during the video playback is more critical since user has a high probablity to immediately request another video after he watches the current one.

VI. Conclusions

In this paper, in order to explore the potential of reducing the start-up latency of streaming media serices, we have proposed a prefetching scheme and performed an analysis to evaluate its performance based on data from a Swedish TV-on-demand service.



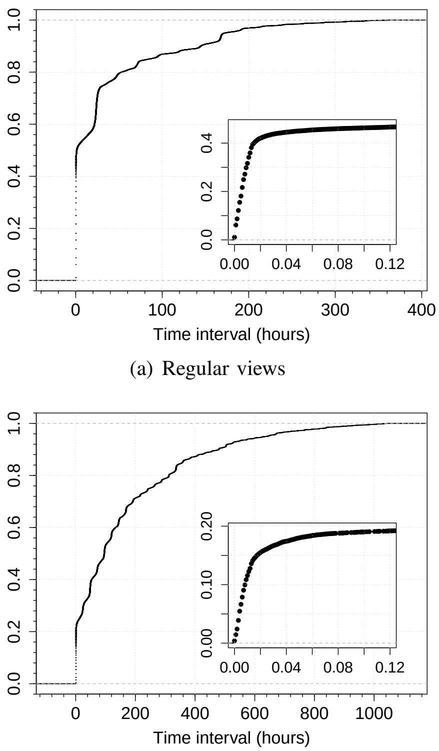
(a) Regular views

(b) Irregular views

Figure 9. CDF of time interval between two view events (Upper bound)

First, the paper has demostrated that in the ideal scenario with 100% prediction accuracy, 73% of the requests are predictable. It suggests a great potential for prefetching. We proposed to use the intrinsic structure of TV series in our data set and prefetch N adjacent videos to terminal devices. We found that the more videos are prefetched the higher hit ratio can be obtained which implicates that more requests can be served directly from the local cache with short delay. A cost model was proposed to quantify the cost of prefetching and to provide an optimal solution for prefetching. The result shows that prefetching two adjacent videos yields 62% hit ratio which is more than four times as terminal caching can obtain. We also demostrate that with a simple prefetching scheme as we proposed, 69% of all the requested videos can be correctly predicted which is very close to the ideal value 73% in this study. When prefetching costs 10% of the cost of real time downloading, to prefetch two videos costs the least which is the optimal choice of prefetching in this study. Moreover, prefetching combined with terminal caching can further improve the hit ratio. We also found that, the time for prefetching depends on user request patterns. For TV series, it is more reasonable to prefetch the next episode before the end of the current viewing session. For irregular requests, videos can be prefetched during off-peak hours. For programs which have a more random request pattern, like children's programs, it is better to make prefetching decisions during the current video playback time, even for irregular views.

In this work, only viewing sessions longer than 2 minutes can trigger prefetch. However, users may be less tolerant



(a) Regular views

(b) Irregular views

Figure 10. CDF of time interval between two view events (Lower bound)

to delays in short sessions than in long sessions. Thus, to prefetch for short session is worth to be further investigated. Another prediction limit of our study is the number of first seen episodes in each series. In future work, we plan to extend our research into cluster-based prefetching mechanisms to find user clusters and make prefetching decisions based on similar users' behaviour to even predict the first seen episode in each series.

Acknowledgment

This work has partly been financed by the Swedish Governmental Agency for Innovation Systems(VINNOVA) in the EFRAIM project and the NOTTS project. Maria Kihl is a member of the Lund Center for Control of Complex Engineering Systems(LCCC) and the Excellence Center Linkping - Lund in Information Technology(eLLIIT)

References

[1] F. F.-H. Nah, "A study on tolerable waiting time: how long are web users willing to wait?" Behaviour & Information Technology, vol. 23, no. 3, 2004, pp. 153–163.

[2] D. F. Galletta, R. Henry, S. McCoy, and P. Polak, "Web site delays: How tolerant are users?" Journal of the Association for Information Systems, vol. 5, no. 1, 2004, pp. 1–28.

[3] W. Ali, S. M. Shamsuddin, and A. S. Ismail, "A survey of web caching and prefetching," Int. J. Advance. Soft Comput. Appl, vol. 3, no. 1, 2011, pp. 18–44.

[4] J. Xu, J. Liu, B. Li, and X. Jia, "Caching and prefetching for web content distribution," Computing in Science Engineering, vol. 6, no. 4, July 2004, pp. 54–59.

[5] Y. Jiang, M.-Y. Wu, and W. Shu, "Web prefetching: Costs, benefits and performance," in Proceedings of the 7th international workshop on web content caching and distribution (WCW2002). Boulder, Colorado. Citeseer, 2002.

[6] J. Domenech, J. Sahuquillo, J. A. Gil, and A. Pont, "The impact of the web prefetching architecture on the limits of reducing user's perceived latency," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, ser. WI '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 740–744. [Online]. Available: http://dx.doi.org/10.1109/WI.2006.166

[7] J. Marquez, J. Domenech, J. Gil, and A. Pont, "Exploring the benefits of caching and prefetching in the mobile web," in Second IFIP Symposium on Wireless Communications and Information Technology for Developing Countries, 2008.

[8] M. Deshpande and G. Karypis, "Selective markov models for predicting web page accesses," ACM Trans. Internet Technol., vol. 4, no. 2, May 2004, pp. 163–184. [Online]. Available: http://doi.acm.org/10.1145/990301.990304

[9] X. Chen and X. Zhang, "Popularity-based ppm: an effective web prefetching technique for high accuracy and low storage," in Parallel Processing, 2002. Proceedings. International Conference on, 2002, pp. 296–304.

[10] D. Joseph and D. Grunwald, "Prefetching using markov predictors," SIGARCH Comput. Archit. News, vol. 25, no. 2, May 1997, pp. 252–263. [Online]. Available: http://doi.acm.org/10.1145/384286.264207

[11] Z. Ban, Z. Gu, and Y. Jin, "An online ppm prediction model for web prefetching," in Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management, ser. WIDM '07. New York, NY, USA: ACM, 2007, pp. 89–96. [Online]. Available: http://doi.acm.org/10.1145/1316902.1316917

[12] G. Pallis, A. Vakali, and J. Pokorny, "A clustering-based prefetching scheme on a web cache environment," Computers & Electrical Engineering, vol. 34, no. 4, 2008, pp. 309–323.

[13] W.-G. Teng, C.-Y. Chang, and M.-S. Chen, "Integrating web caching and web prefetching in client-side proxies," Parallel and Distributed Systems, IEEE Transactions on, vol. 16, no. 5, May 2005, pp. 444–455.

[14] Z. Su, Q. Yang, and H.-J. Zhang, "A prediction system for multimedia pre-fetching in internet," in Proceedings of the Eighth ACM International Conference on Multimedia, ser. MULTIMEDIA '00. New York, NY, USA: ACM, 2000, pp. 3–11. [Online]. Available: http://doi.acm.org/10.1145/354384.354394

[15] C. Bouras, A. Konidaris, and D. Kostoulas, "Predictive prefetching on the web and its potential impact in the wide area," World Wide Web, vol. 7, no. 2, 2004, pp. 143–179.

[16] S. Khemmarat, R. Zhou, D. K. Krishnappa, L. Gao, and M. Zink, "Watching user generated videos with prefetching," Image Commun., vol. 27, no. 4, Apr. 2012, pp. 343–359. [Online]. Available: http://dx.doi.org/10.1016/j.image.2011.10.008

[17] D. K. Krishnappa, S. Khemmarat, L. Gao, and M. Zink, "On the feasibility of prefetching and caching for online tv services: A measurement study on hulu," in Proceedings of the 12th International Conference on Passive and Active Measurement, ser. PAM'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 72–80. [Online]. Available: http://dl.acm.org/citation.cfm?id=1987510.1987518

[18] Z. Zeng and B. Veeravalli, "Hk/t: A novel server-side web caching strategy for multimedia applications," in Communications, 2008. ICC '08. IEEE International Conference on, May 2008, pp. 1782–1786.

[19] Z. Zeng, B. Veeravalli, and K. Li, "A novel server-side proxy caching strategy for large-scale multimedia applications," J. Parallel Distrib. Comput., vol. 71, no. 4, Apr. 2011, pp. 525–536. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2010.06.008

[20] S. Seny, J. Rexfordz, and D. Towsley, "Proxy prefix caching for multimedia streams," in INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 3, Mar 1999, pp. 1310–1319 vol.3.

[21] W. Liu, C. T. Chou, Z. Yang, and X. Du, "Popularity-wise proxy caching for interactive streaming media," in Local Computer Networks, 2004. 29th Annual IEEE International Conference on, Nov 2004, pp. 250–257.

[22] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in Proceedings of the 27th European Conference on Information Retrieval, 2005, pp. 345–359.

Unmanned Aerial Vehicles as Assisting Tools in Dismounted Company Attack

Major Tapio Saarelainen, PhD, IARIA Fellow
Army Academy
Army Research and Development Division
Lappeenranta, Finland
tapio.saarelainen@mil.fi

***Abstract*—This paper presents an idea-phase introduction of how to utilize swarms of Unmanned Aerial Vehicles (UAVs) in performing a dismounted company attack. UAVs are used in the process of data gathering and as tools to increase overall Situational Awareness (SA). A company represents a basic military unit performing time-critical tasks whose completion is mandatory for operational survival. The discussed idea-stage solution relies on using UAVs as tools of military commanders to act as data collectors, hub-stations and tools to gather data from the areas of interest. Since the tempo of operations at tactical level (battalion and below) has increased, the amount and type of data gathered are crucial in terms of operational success. The need for timely and accurate data from a designated area is necessary for improved decision making process, which is dependable on Situational Awareness and Common Operational Picture (COP). Once the critical data have been gathered and analyzed, UAVs act as versatile assisting tools in military operations in the roles of collecting and forwarding data to support processes of control and command. The main contribution of this paper is to identify the possibilities and the process of how to improve the overall performance of military troops by utilizing UAVs as assisting tools in gathering real-time data required for decision making.**

Keywords-Unmanned Aerial Vehicles (UAVs), dismounted company attack, real-time data, military decision making process.

I. INTRODUCTION

Militaries all over the world continue developing methods for saving the lives of own troops. The reason for this is the downsizing process of armies in western countries. Battlespace is the environment where military operations are executed. The number of soldiers in combat units decreases, and simultaneously the number of personnel to take care of the logistics and maintenance issues increases as the maintenance of machinery utilized in the battlespace asks for ever increasing resources. The overall aim of militaries continues to be the sustained capability to improve operational performance despite the downsizing and minimized number of soldiers in combat. The key for this is in attempting to increase SA. This asks for the use of UAVs to produce the required data for analyzing purposes.

Creating a simulation to model a UAV assisted operation is at present unfortunately not feasible on the grounds that neither funding nor facilities are available. To measure the utilization of UAVs in a dismounted company attack would require real functioning swarms of UAVs. No funding is available for this at the moment.

A company attack, typically a dismounted company attack, is a demanding military operation, which requires constant real-time data to allow executing all the phases of an operation to achieve the set goal. It comprises several phases of action, the first of which is reconnaissance. An attacking unit must find out the composition and location of the opposite entity before the attack can be executed. Once the reconnaissance data have been gathered, the planning sequence of the operation begins by implementing the process of Military Decision Making Process (MDMP). In this process, it is possible to benefit from automated assisting tools. When the MDMP has been carried out, it results in different types of Courses of Action (COAs). These COAs represent different alternatives to military commanders on how to organize the attack. Once the optimal COA has been chosen, the maneuvre named dismounted company attack can be ignited. A Dismounted Company Attack is composed of the phases outlined in the following Figure 1: Assembly area, dismount line, line of departure, engagement, combat and the objective.

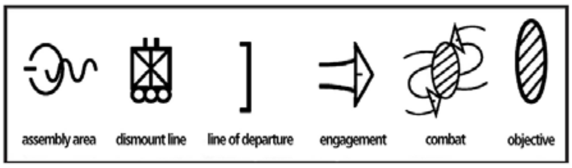


Figure 1. Dismounted Company Attack as a process [1.]

As an example of examining COAs, Figure 2 depicts two different types of COAs in the battlespace. In the first COA, the objective is to stop an armored enemy by deploying a flanking movement, whereas in the second COA, the objective is to destroy an enemy command post by direct engagement.

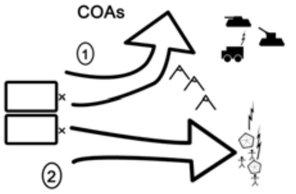


Figure 2. Different types of Courses of Actions.

The utilization of UAVs can be seen as having a central role when real-time data are required for rapid decision making. UAVs can be seen as a resource for military commanders in that they offer the advantage of surprise in an attack by producing the real-time data needed for decision making. If a commander fails in surprising the enemy, he or she loses the possibility to take the initiative in the operation. The ability to take and maintain the initiative is usually a must in a successful military operation, especially in a dismounted company attack. When real-time data are available, an improved decision making process and its outcome become possible.

The utilization of Unmanned Aerial Vehicles is necessary to get the correct information from enemy space as rapidly as possible in the hectic tempo of operations, save soldiers' lives and support the morale of troops performing the maneuvres in an operational area. Furthermore, the use of unmanned machines can be seen as part of optimizing the existing resources of military units as machines continue performing fearlessly.

When militarily utilized, UAVs can be used for varying tasks: performing terrain reconnaissance in the area of objective, performing reconnaissance of the use of Nuclear, Biological and Chemical substances, maintaining reconnaissance engagement by providing data on enemy movements, and monitoring the actions of enemy troops (direction and speed of movement, location, size, formation, action). UAVs can also be utilized in the targeting process in that UAVs can be sent airborne to the designated areas for the reconnaissance and targeting of the potential enemy targets. UAVs can act as relay stations for sustaining the constant capability to communicate and control the troops and machines in the hostile area.

Collected data from the designated areas is utilizable in the MDMP. This process is linked to gaining SA and choosing between different COAs. These data can be collected more safely, if an appropriate number of UAVs of the necessary type are available for these operations. When UAVs are used aside the soldiers, to complement the data gathering, it is optimal to combine the capabilities of soldiers and machines. Machines can be sent out to the most dangerous locations while soldiers remain in charge of the less lethal task, if ever possible in a battlespace.

The use of machines gives an advantage for commanders. First, UAVs can be sent to the areas of interest, days ahead, if required. UAVs can be shut down or reactivated via electrical signals when necessary. This gives commanders the advantage to transport the UAVs as reconnaissance resources to the designated areas well in advance and at the chosen moment. When UAVs have been flown to the area and shut down, they will not consume any energy, the saved energy can be used at a chosen moment and in the wanted operation, which supports the own battle plan and own objectives. Commanders can plan these actions well in advance and have UAVs transported to the areas when appropriate and safe. UAVs can be ready and standing by well in advance, in a chosen area, with the chosen sensors embedded in the UAV platform.

To summarize: commanders can benefit from the use of push and pull factors. Commanders can push the UAVs to the optimal area of battlespace at a chosen moment. The optimized location of UAVs guarantee the improved performance in own oncoming military maneuvers. The pull factor means pulling the data from the designated areas at a chosen moment. These data are pulled to increase the performance of own operations in analyzing the collected data. The collected data are raw material for defining SA and COP. The accrued data are utilized in the MDMP and in the process of choosing between different COAs.

Communication is critical in the execution process of command and control. The lack of communication results in an ineffective military operation. When the possibilities offered by the Wireless Polling Sensor Network (WPSN) are taken into active use together with using One Time Pads (OTPs), two communication goals become achievable: the covert network and security in messaging. This is discussed more in Section V.

When machines, such as UAVs, are utilized, Service Oriented Architecture (SOA) becomes applicable. In digitized battlespace and in digitized operational planning, SOA is utilizable in the allocation processes of own existing resources and optimizing the use of troops in correct time and in the correct operational area. SOA is also useful in offering assistance in the overall planning process. SOA can be used in optimizing the timing of the different actions, also while the dismounted company attack is in progress, and in automating the MDMP. By automating the MDMP and using UAVs, commanders can save time, resources and lives as well as achieve the set objective.

The rest of this paper is organized as follows. Section II concentrates on the related work, Section III discusses Unmanned Aerial Vehicles, and Section IV concentrates on the essence of communication. Section V deals with Wireless Polling Sensor Networks and the use of One Time Pads, Section VI deals with Military Decision Making Process, and Section VII focuses on the Situational Awareness and Common Operational Picture. Section VIII discusses the significance of sensors, and Section IX looks at SOA in relation to MDMP and reorganizing the chain of command in troops. Section X comprises discussion, Section XI concludes with the results, and Section XII addresses the requirements for further studies.

II. Related Work

Several researchers have been studying the use of UAVs in accruing data to support SA, COP and MDMP by increasing the performance of different types of networks. Moreover, the studies listed here have concentrated on increasing the speed, safety and capability to communicate in an improved manner.

As demonstrated in [2], ad-hoc networks can create a UAV access net ensuring communication among mobile or stationary users. These ad-hoc networks support Blue Force Tracking, as indicated in [3]. When UAVs are equipped with Free Space Optics (FSO) communication links, operations can be executed by avoiding to become sensed by means of electrical reconnaissance detection, as concluded in [4]. FSO represents an optical communication technology that uses light propagating in free space to transmit data from point-to-point (and multipoint) by using low-powered infrared lasers, which can also be used for localization purposes, if range and orientation information is available FSO-technology offers high-speed, up to 10 Gb, reliable and cost-effective connectivity for heterogeneous wireless services provision in both urban and rural deployments when Dense Wavelength Division Multiplexing (DWMD) is utilized in Radio-on-FSO (RoFSO) system.

In vision-based tracking, pan-tilt gimbaled cameras using Commercial off-the shelf (COTS) components can be used as well as calculation algorithms and advanced controlling systems for integrated control of a UAV and an onboard gimbaled camera, see [5]. Along with the availability of both low-cost and highly capable COTS-based UAVs and Unmanned Ground Vehicles (UGVs) and communications equipment, it is reasonable to apply quick and inexpensive means for surveillance, tracking and location purposes, as discussed in [6]. UAVs of varying types and sizes can be used in aerial surveillance and ground target tracking, see [7]. To boost the performance of a single UAV, swarms of small UAVs can rely on airborne MANETs, as indicated in [2]. Transmit antennas are significant in the process of operating UAVs, as indicated in [8]. When swarms of UAVs are utilized for navigation, localization and target tracking, information synchronization is important, as discussed in [9]. In present battlespace miniature UAVs are becoming increasingly significant among surveillance applications, as shown in [10]. Remotely controlled UAVs can act as an assisting tool in tracking and monitoring, as discussed in [5]. Remotely controlled UAVs can enhance SA, Blue Force Tracking (BFT), thereby enforcing the probability of success in missions, even when operating beyond line-of-sight, see [11]. The means for exploiting UAVs and UGVs in the processes of data collection and the distribution of near real-time COP to be implemented in Shared SA are discussed in [12]. Battle Management Language (BML) can be seen as a common language enabler between machines and interfaces along with almost ubiquitous swarms of UAVs [9.]. For example, networks utilizing COTS components mounted on of UAVs add survivability and remove the need for a line-of-sight connection, as described in [6.].

This present paper examines the topic from a different angle by focusing on how to facilitate a dismounted company attack with the use of UAVs. This means aiming at optimizing the use of existing resources and automating the attack to the extent feasible. The objective is to contribute to the overall goal of increasing safety in military operations by means of improved use of resources resulting in decreasing numbers of casualties as well as increasing the tempo of own military operations.

III. Unmanned AerialVehicles

Unmanned Aerial Vehicles utilized in a dismounted company attack can be autonomous or guided platforms built with COTS material ensuring the relatively inexpensive price tags on the UAVs. The main function of UAVs is to produce real-time data for commanders for decision making purposes. The use of swarms of UAVs ensures the gathering of data behind the visual horizon. Distances between command link and the swarms of UAVs are typically few kilometers. Typically, if a UAV has been identified by the actions of an adversary, the particular UAV ends up becoming annihilated. Therefore UAVs have to be built to be disposable elements. Once the UAVs are used as swarms, the combat survivability of the system can be increased.

This paper examines only UAVs because of their versatility compared to the other Unmanned Vehicles (UVs), such as Unmanned Ground Vehicles (UGVs). When a small tactical level military unit, such as a company, is performing a complicated maneuver, a dismounted company attack, the UAVs represent the only reasonable type of UVs to be utilized. UAVs are capable of monitoring the designated target areas and transmitting real-time data to the base-station simultaneously when monitoring the area. See Figure 3.

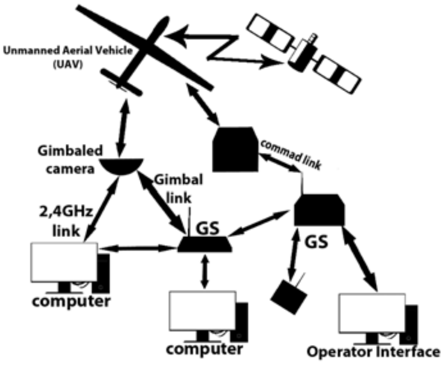


Figure 3. UAVs and the data transmission.

Typically, when flight times are short, less than an hour, engines and sensors embedded into UAVs can be powered by liquid fuel batteries to ensure adequate level of energy. Liquid Polymer (LiPo) batteries are utilizable for their capacity. Electrical surveillance components, guidance systems, and command systems are depend on adequate electricity level.

Typically, the distance between a communication link and a swarm of UAVs is few kilometers, and therefore 2,4 GHz Ultra-Wideband Network system between the communication link and the base station is applicable for these distances. The typical speed of swarms of UAVs is tens of kilometers per hour. This is a chosen speed to balance the energy consumption and the range of transmission power and movement.

UAVs are most versatile with their capability for quick deployment. UAVs tend to be miniature-sized airplanes, drones and helicopters, weighing few kilograms. The range of these vehicles can vary from few hundred meters to few kilometers as can their mass and size. The same applies to the payload. The payload can be measured from tens of grams to few hundred grams depending on the use and measurements of UAVs.

The typical payload of UAVs can comprise varying sensors, such as: acoustic-, seismic-, magnetic-, visible image-, shortwave infrared (SWIR)-, thermal-, infrared-, low-light television (LLTV)-, and sensors for laser tracking and spotting, and for facial recognition. These sensor packages can be also deployed into the area of interest at a desired moment. If the area of interest is known well in advance, a UAV and the sensor package can be flown into the perimeter in advance and been dropped in the chosen area. In addition, a UAV can be guided close to the area of interest. The UAV can then be parked, for example, on rooftops and cliffs to wait for the command to start the reconnoitering mission. This saves time and positions the UAV in nearby perimeters of the desired area. The UAVs remain hidden and hard to detect, and when detected, it is too late anyway.

IV. COMMUNICATION

Communication between different troops and inside the company remains vital from the perspective of a successful military operation. The sustained capability to communicate between troops (soldiers) and machines (UAVs) must be maintained throughout a military operation. Without communication there is neither command nor control between the entities. Communication can be described as comprising three layers. These layers are sensor-layer, C^4I^2SR –layer and shooters –layer. The layers are connected with the existing communication networks. Different layers are depicted in Figure 4.

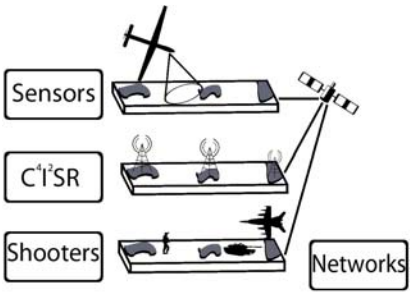


Figure 4. Communication network from sensor to shooter.

The depicted layers communicate and forward data via UAV –radios, which can utilize the output of GPS/GLONASS receivers for automatic position reporting. When Software Defined Radios (SDRs) are embedded onboard UAVs, typical SDRs' features can be included into transmission protocols. These include: Multiband 30 – 512 MHz, multimode, multi-mission, software programmable architecture, Low Probability of Detection (LPD) and Low Probability of Identification (LPI), simultaneous voice and data, near-real time data transfer from sensor to decision maker and onwards to the shooter applications. Lightweight UAV –implemented radios offer low consumption transceivers operating in the frequency of 2.4 GHz Industrial, Scientific and Medical (ISM) band. The use of this frequency offers very low Radio Frequency (RF) signature modulation scheme based on spread spectrum technology (DSSS) and provides a robust, reliable and low probability of detection time division multiple access (TDMA) waveform. Radios typically feature advanced encryption standard (AES) encryption providing a very high security level on the transferred audio and data systems, while security keys can be downloadable. UAV -radios can provide several tactical communication services: full-duplex voice conferencing, GPS reporting, e-mail, chat, file transfer, and real-time video streaming.

When properly adopted into active use, UAVs can be seen as flying hubs or flying relay-stations, tools of communication. When UAVs are used to secure communication, as depicted in Figure 4, the throughput of communication can be maximized. Furthermore, the swarms of UAVs end up creating an own data communication system, as depicted in Figure 5. This ensures that the data transmission distances between UAVs remain short and become operationally secure. This aids in meeting the requirements of Low-Probability of Detection (LPD) and Low-Probability of Identification (LPI). The described delicate system introduced is a new one and based on ideas that can be executed by utilizing existing COTS- technology [13].

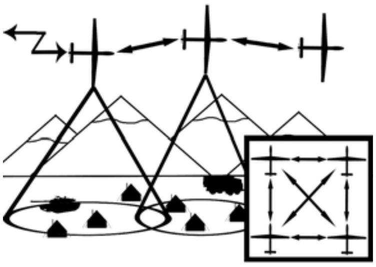


Figure 5. A data exchange process inside the swarm of UAVs.

Figure 5 describes the idea of using UAVs as swarms. The number of UAVs used in each scenario varies depending on the commanded mission and its speed and other set requirements.

V. WIRELESS POLLING SENSOR NETWORKS

A battlespace tends to be embedded with different types of sensors found from the soil, airborne or attached into various types of manned and unmanned vehicles. The utilization of Wireless Polling Sensor Network (WPSN) together with OTPs can be seen as one possible solution for communication system between sensors and UAVs, as indicated in [14]. WPSN can be viewed one possible solution when gathering data from different sensors and sensor networks. When a swarm of UAVs are utilized in forming an ad-hoc network and polling a large number of fixed sensor nodes, a secure network system can be created. The WPSN system is more robust in the military environment than traditional Wireless Sensor Networks (WSNs). Although WSNs have been used for a long time, they demonstrate particular disadvantages. These include the fact that multi-hop transmission fails when nodes are destroyed in military environments, battery lifetime creates limitations, and security challenges remain unsolved. A WPSN has advantages in all of these areas compared to other proposed solutions. WPSN comprises a small mobile ad-hoc network of UAVs and a high number of fixed ground-based sensors, which are periodically polled by the UAVs.

The advantages concerning WPSN and OTP include that in the WPSN solution the fixed sensor nodes remain concealed, yet active, because the sensor nodes of WPSN do not communicate with each other but only respond to polling by the mobile nodes. The WPSN node communicates with a UAV through encrypted messages. Thereby, WPSN responds only after a UAV has submitted a polling request with a specific code. The routes of UVs can be fed into the systems early enough to gain the needed information from the designated areas [14].

When speaking about a battlespace and actions taking place in this hostile environment, it is mandatory that some of the UAVs be shot down or destroyed by other means of contemporary warfare. This possibility must be recognized prior to engagement. UAVs must be designed so that once malfunctioning, they will get automatically destroyed (software and hardware) to become instantly useless for the adversary. Yet, this destruction of one UAV does not jeopardize the concept of sustained secure communication, for the network composed by remaining UAVs will reroute itself automatically.

VI. Military Decision Making Process

In military operations performed at tactical level, i.e., battalion and below, the significance of tempo and timing becomes critical. In the MDMP all the raw data collected by UAVs have to be analyzed and taken into account as they are directly connected with targeting systems, weapon selection processes, COP, SA and Control, Command, Computers, Communication, Information, Intelligence, Surveillance and Reconnaissance (C^4I^2SR).

Here, automation and mathematics can be seen as assisting tools in the process of making rapid and reliable decisions. When mathematical methodology is implemented, measuring the additive value model is the simplest and most commonly used mathematical model in multiple objective decision analysis. As described in [15], the additive value model is given by the equation:

$$v(x_j) = \sum_{i=1}^{n} w_i v_i(x_{ij}) \tag{1}$$

where

$v(x_j)$ is the total value of alternative *j*,

$i = 1$ to n are the value measures specified in the qualitative value model,

x_{ij} is alternative *j*'s score (raw data) on value measure *i*,

$v_i(x_{ij})$ is the single-dimensional value of alternative *j* on value measure *i*,

and w_i is the swing weight of value measure *i*.

Equation (1) is the simplest and most commonly used mathematical model in multiple objective decision analysis.

Obviously, mathematics alone cannot solve the dilemma of making the correct operational decision quickly in a chaotic combat setting. When a decision is made between different COAs, mathematics and probability prognosis can only be seen as assisting tools. The human commander is the only one who is responsible for sensible and applicable decision which can be converted into commands to be issued and executed in an operation.

VII. Situational Awareness and Common Operational Picture

The term Situational Awareness has been given an apt definition in the Army Field Manual 1-02. SA can be understood as knowledge and understanding of the current situation, which promotes timely, relevant and accurate assessment of friendly, competitive and other operations within the battlespace in order to facilitate decision making. SA, furthermore, equals an informational perspective and skill that fosters an ability to determine quickly the context and relevance of events that is unfolding. The term SA comprises three levels: 1) perception, 2) comprehension and 3) projection. [16]. SA, or, the lack of it, remains critical in performing military operations successfully. The means to increase SA can and must be fostered and developed, since the loss or deterioration of SA results in inaccuracies, human errors, and eventual casualties and fratricide. The military operation in progress usually fails because of poor level of SA.

Situational Awareness has a strong relation to COP. COP represents an overall understanding of the prevailing situation in the battlespace. COP can be displayed on the screen of a computer or a digital device, and by using markers and traditional maps. COP features elements, such as individuals of friendly forces, neutral entities and the adversary, presented by symbols of various types.

To complete the list of phenomena affecting the MDMP, C^4I^2SR needs to be taken into account. UAVs are utilized to assists the MDMP performed in C^4I^2SR environment. When combined together as swarms, UAVs form tools for accruing data, forwarding and analyzing these data into the form of information to create COP and increase SA.

To sum up, all these listed elements are linked to the MDMP. The decisions made as part of the MDMP can also be seen as tools in targeting and weapon selection processes. Figure 6 explains the relations and functions inside MDMP when the use case is related to targeting and weapon selection systems. In the MDMP the end-come is the optimal use of weapon systems to avoid collateral damage and fratricide.

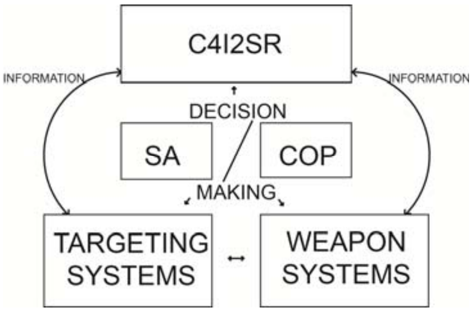


Figure 6. Decision making system in targeting and weapon systems.

VIII. Sensors

In order to achieve the set objective in a given military setting, it makes sense to utilize maximally the data produced by various types of sensors when accruing data from hostile environments. In some cases, especially when the weather conditions are challenging, for example, the wind speed exceeds 10 metres per second, the UAVs cannot be used or they are too slow and there has to be an alternative possibility to deploy sensors for accruing real-time data. Some of these sensors can be deployed to the area of interest with the assistance of artillery fire produced by mortars or cannons. Rapidly deployable airborne sensors represent relatively inexpensive and versatile tools for low-level battalion and company operations. As indicated in [17], light sensor munition can be deployed behind enemy lines. An example of sensors' deployment, when UAVs are not applicable, is Sensor Element Munitions (SEMs). SEMs can be manufactured of composites surfaced with materials capable of absorbing radar beams, making the SEM less visible in enemy counter-artillery radars. In any military operation, airborne sensors are important for missions, such as force protection, perimeter control and intelligence utilization, as discussed in [18]. Transmitting the accrued data to prevent cases of fratricide and to ensure success in operations presupposes optimal communications. WiMAX transmission offers applicable possibilities in forwarding collected data. The distances in the transmission process are relatively short, ranging from 1 kilometre to few kilometres in conditions of clear Line-Of-Sight.

The sensor package inside SEMs (Sensor Elements, SE) is made of COTS-products comprising sensors capable of sensing most of the phenomena occurring in the electromagnetic spectrum. Overall, COTS-products are relatively inexpensive and reliable in terms of function, as explained in [19]. Sensor Elements can contain the same sensors as UAVs. The command post has the capability for the data fusion of all the accrued sensor information.

Once an SE is airborne, it immediately starts to transmit the gathered data to friendly troops either directly or, if the transmission distance exceeds the capability of the transmission unit, the SE transmits the data to another airborne device, which acts as a relay station in relation to own troops. The SE communicates with the receiver station and other sensor element packages over a 2,4 GHz Ultra-Wideband Network system. The accrued data are encrypted for security reasons. The composition of SEMs is depicted in Figure 7.

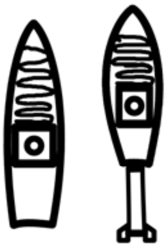


Figure 7. Structures of Sensor Element Munitions (SEMs): An artillery SEM (left), a mortar SEM (right) [17].

SEMs can be deployed to the target area with manned artillery weapons or unmanned remotely controlled pieces of artillery or by using mortars, as mentioned earlier. The process of deploying SEM to the area of interest is depicted in Figure 8. SEM ejects the Sensor Element (SE) which in turn reports the gathered data to the base station [17].

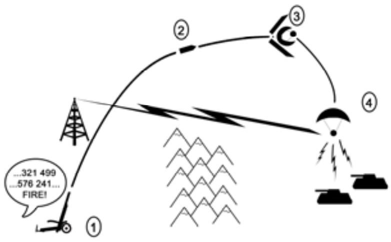


Figure 8. Process on how to deploy an SE to an enemy territory [17].

Figure 8 presents a typical use case, in which a company is executing a military operation supported with an artillery or mortar unit. The reconnoitering range tends to vary from one to few kilometers. When a dismounted company attack is supported with units of UAVs tailored for Close-Air Support (CAS), the data exchange transmission process for the target data is depicted in Figure 9.

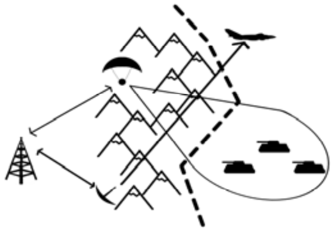


Figure 9. The process of detecting target to the shooter [17].

The UAVs of CAS units optimize the speed and destruction power used in proximity to destroy the designated targets. When a small unit operates, it needs to achieve results in short time in order to maintain the initiative and reach the set objective. A company is a small military unit, which has to maximize the momentum offered by the performance produced by CAS units. UAVs must be utilized as tools to evaluate the outcome of the executed CAS fire-mission. If the result of CAS fire-mission is reported not to fulfil the requirements set, the new round of CAS fire-mission must be performed to destroy the chosen target.

IX. SERVICE ORIENTED ARCHITECTURE

SOA offers a variety of possibilities to improve the performance in military operations. SOA can be exploited when needing to reorganize the military organization after casualties affect the chain of command in a dismounted company. This process is described in [20]. Military operations nowadays usually demonstrate features of Network Centric Warfare (NCW) in which one key aspect is to be able to offer a valid and accurate COP for the operating troops in the battlespace. A basic requirement for a military commander is the ability to command the troops and sustain optimal SA and COP. An important aspect in distributing information in the battlespace is the amount and quality of information shared at different levels. SOA can be seen as a useful tool in distributing data in a preprogrammed manner. The amount of information allocated must be set to a level where the decision maker can perform timely and draw accurate conclusions. In the battlespace units suffer from casualties and the chain of command never remains intact.

A constructive idea in SOA is in its process ideology. In a dismounted company, the composition of the unit and its performance are critical in executing the operations. Military units suffer from casualties and their mathematical performance value tends to change in an unpredictable manner. The performance of a military organization, such as a dismounted infantry company, can be mathematically calculated as explained in [20].

Behind these mathematical values is a Psycho Physical Factor, described in [20]. The process of creating this factor is described in [21] and the formula may be useful in calculating the performance of a military unit.

The implementation of SOA is described in [20]. A key aspect in the presented architecture [20] is the dynamically changing architecture. Benefitting from the possibilities offered to orchestrate data and services with the assistance of SOA allows for improved performance in the execution of operations. SOA can be utilized in the process of choosing between different COAs. Eventually, the chosen COA will be fine-tuned into commands and maneuvres of a dismounted company attack.

In applying SOA paradigms, loose coupling, dynamic binding and independency of development technologies, platforms and organizations, as well as locations, all these become advantages in that the use of SOA typically encourages reusing services. The identified assets belong to military units, but military units offer their responsibilities through services, and capability deployment requires invoking and integrating a number of services. Figure 10 depicts the relations of services, assets and capabilities of a military unit.

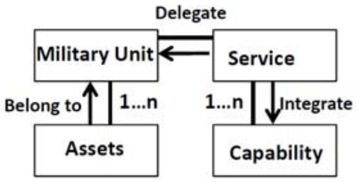


Figure 10. Conceptual model of C^4I^2SR capabilities based on SOA [22].

In a dismounted company attack, all operating military platoons use the same services and assets of a dismounted company. In this case, SOA itself offers a flexible approach to identify C^4I^2SR capabilities when several platoons benefit from the services and assets of a dismounted company. SOA enables a company to respond more quickly to the changing battlespace situations and requirements to execute the given missions in the given time and with the allocated resources.

X. DISCUSSION

When the UAVs are successfully utilized in the different phases in a dismounted combat attack, the results can be optimally utilized. These gained results can be identified and evaluated in relation to the different stages of the process of a dismounted company attack. When the need of a requested service is identified, automated systems assist to fulfil the need of any type (need of data, resupply, firepower, evacuation).

As noted in the Related Work –Section, the use of UAVs has been identified as critical and effective for a successful military operation. UAVs can be used for collecting near real-time data, as a flying hub-station and in assessing the impact of artillery fire. Using swarms of UAVs enables quick, reliable and effective data collecting from a specified area. Furthermore, when UAVs function as the communication link, the chain of command and control remains secure as regards communication.

By exploiting the data accessed by means of using UAVs it is possible to enhance a dismounted military operation: readjusting the direction and action of combat units and increasing their speed. The communication between UAVs and ground base-stations is encrypted. This ensures that the data collected and communication transmitted remain intact and coherent. UAVs may fly via automated waypoints or serve as fighter-operated systems. UAVs can be designed to be disposable, self-destroyable, once their task has been completed, or in case of malfunction, or if encountered by enemy. The use of SEMs becomes applicable in cases when the weather conditions are challenging, for example, the wind speed exceeds ten metres per second, or if data concerning a target must be rapidly accessed.

Compared to traditional WSN-systems, WPSN allows for improved security protocol in the communication between UAVs and sensors. Data collecting systems gather raw data on battle space phenomena, for example troop movements and action. These raw data feed the MDMP and facilitate speeding up the decision making. Using mathematical models and –programs produces improved SA and COP, compared to non-automatized human decision making performance. The improved SA and COP allow for significant increase in efficiency as regards planning and implementing the tactical use of destructive fire power.

As the raw data collected by UAVs are already in electronic form, SOA can be utilized in planning, distributing and optimizing resources: evacuation, supplies, use of artillery fire. When the described systems for data collecting, analyzing, and communicating function as planned, it becomes possible to carry out an automized, computer-assisted attack as described in Figure 11. Utilizing FSO communication links fosters reliable, secure and coherent communication in command and control processes.

If and when all the accrued data can be properly processed and analyzed in MDMP with the assistance of SOA, the performance of troops can result in an automated dismounted company attack as depicted in Figure 11.

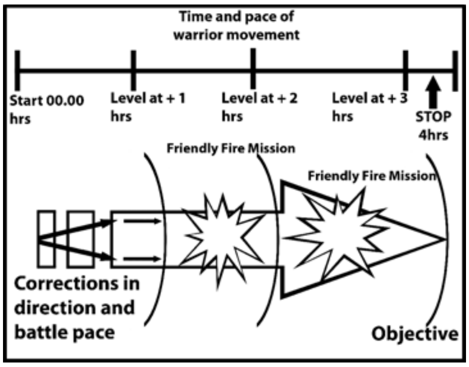


Figure 11. An automated attack operation. [1]

Figure 11 aims to visualize the goal of commanding military troops with the assistance of a computerized Artificial Intelligence (AI). The final commands for the military units to move and execute are given by a human commander, not by a machine. UAVs can be seen as tools in monitoring and assisting in an operation when re-adjusting its pace: If the pace of the units or an individual soldier is too slow, the data transmitted by UAVs is utilized to fine-tune the speed and direction of the operating troops. The SA data acquired by means of UAVs must be used in taking the iniative and translating it into success in battle and eventually meeting the set objectives of the given military operation.

XI. Conclusions

This paper has focused on observing how to utilize the real-time data collecting ability provided by UAVs in order to improve the performance of a dismounted company attack. The approach adopted equals idea-stage examination and as such aims to examine grounds for further planning of how to execute military attack missions with the assistance of UAVs. This paper introduces a concept of benefitting from the use of UAVs as part of a dismounted company attack. When doing so, it also points out the necessity of rapid data collection to support the fast MDMP.

The utilization and performance of UAVs can be seen as data collectors in the battlespace. Sensors embedded into the UAVs used can produce different types of data from the designated area. UAVs can be deployed to the designated areas in most weather conditions and immediately when required, as there are no latency times. The data collected by UAVs are then transmitted to the command posts. The collected real-time data remain critical for the MDMP. The data accrued must be in a pre-defined digitized form, which is applicable in digitized decision making systems exploited in the battlespace. The level and quality of SA continues to be critical at the soldier level, whilst the level of COP plays an important role in command posts, where operations are planned, commanded and controlled.

The data for the MDMP are collected by using various types of sensors embedded into UAVs and SEMs. The key issue is the speed of deploying the sensor package to the area. The prevailing combat situation in the battlespace determines the selecting of the type of UAV and sensor package embedded. The data accrued must be in a digitized form applicable in the software environment used. SOA can be used in re-organizing troops and allocating resources.

The ultimate goal of a dismounted company attack is to execute the mission with the resources allocated and to obtain the set objective. This asks for sustaining timely performance with a minimal number of casualties, no instances of fratricide and with the least possible amount of collateral damage. The objective is difficult to obtain, when using soldiers prone to making human errors. However, operational performance can be improved if the data collected via UAVs is reliable and can be adopted in active use in near-real time. This may result in increased individual and collective performance. With improved levels of SA and COP, the safety of operations may be sustained.

With computers acting as assisting tools in the MDMP offering suggestions as commands to the commander of attacking troops, the role of the commander is to either approve or reject the suggested commands. Thus a human decision maker remains critical in the chain of command.

XII. Future Work

Because a dismounted company attack represents a time-critical maneuvre in the category of tactical military actions performed at a company level, any efforts to improve the company's capability are worthwhile. This asks for developing a ruggedized system based on the idea-phase description outlined in this paper. Attention must be paid to planning the utilization of UAVs together with accounting for the operational security issues concerning using software and hardware in the battlespace.

The usability of UAVs to create a functional communication network requires field testing in combat exercise settings prior to any operational use. UAVs have to be remotely destroyable both physically and digitally. The capability of UAVs to self-destroy when malfunctioning or having been shot down must be tested. Other identified challenges are related to maintaining an adequate level of constant energy flow and protecting against violations caused by electronic warfare.

The use of SOA in assisting the MDMP has to be studied in combat exercise settings as well in order to gain realistic and relevant data on human commanders evaluating COAs when planning a dismounted company attack.

REFERENCES

[1] T. Saarelainen and J. Jormakka, “Interfacing collaboration and command tools for crises management military command and control systems”, International Journal of Electronic security and Digital Forensics, vol 3, No. 3, 2010, pp. 249 – 264.

[2] S. Chaumette, R. Laplace, C. Mazel, and A. Godin, “Secure cooperative ad hoc applications within UAV fleets”, IEEE Conference on Military Communications Conference (MILCOM 2009), 18-21 Oct. 2009, Boston, MA, pp. 1 – 7, doi 10.1109/MILCOM.2009.5379819.

[3] E. Loren., L. Riblett, and J. Wiseman, “TACKNET: Mobile ad hoc secure communications network”, in Proceedings of 41st Annual IEEE International Carnahal Conference on Security Technology, 8-11 October 2007, Ottawa, Ontario, Canada, pp. 156 – 162.

[4] C. Chlestil, et al., “Reliable optical wireless links within UAV swarms”, in Proceedings of Transparent Optical Networks, 18 – 22 June, 2006, Nottingham, Great-Britain, pp. 39 – 42, doi 10.1109/ICTON.2006.248491.

[5] V.N. Dobrokhodov, I.I. Kaminer, K.D. Jones, and R. Ghabcheloo, “Vision-based tracking and motion estimation for moving targets using small UAVs”, Proceedings of the 2006 American Control Conference Minneapolis, June 14 - 16, Minnesota, USA, pp. 1428 – 1433, doi 10.1109/ACC.2006.16564418.

[6] D. Hague, H.T. Kung, and B. Suter, “Field experimentation of cots-based UAV networking”, in Proceedings of IEEE Conference on Military Communications (MILCOM2006), 23-25 Oct. 2006, pp. 1 – 7, doi 10.1109/MILCOM.2006.302070.

[7] A. Ruangwiset, “Path generation for ground target tracking of airplane-typed UAV”, Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics (ROBIO), Bangkok, Thailand, February 21 – 26, 2009, pp. 1354 – 1358, doi 10.1109/robio.2009.4913197.

[8] C-M. Cheng, P-H. Hsiao, H.T. Kung, and D. Vlah, “Transmit antenna selection based on link-layer channel probing”, in Proceedings of IEEE Conference on World of Wireless, Mobile and Multimedia Networks, 18-21 June 2007, Cambridge, MA, U.S.A. pp. 1 – 6, doi 10.1109/WOWMOM.2007.4351703.

[9] Y. Qu, Y. Zhang, and Q. Zhou, “Cooperative localization of UAV based on information synchronization”, in Proceedings of the 2010 IEEE International Conference on Mechatronics and Automation, August 4 – 7, 2010, Xi’an, China, pp. 225 – 230, doi 10.1109/ICMA.2010.5589081.

[10] M. Pachter, N. Ceccarelli, and P.R. Chandler, “Vision-based target geo-location using camera equipped MAVs”, in Proceedings of 46th Conference on Decision and Control (CDC2007), 2007, pp. 2333 – 2338, doi 10.1109/CDC.2007.4434038.

[11] B. Cummings, T. Zimmerman, B. Robinson, and M. Snyder, “Voice over blue force tracking”, Proceedings of IEEE Conference on Military Communications Conference (MILCOM 2006), 23-25 Oct. 2006, Washington, DC, pp. 1 – 5, doi 10.1109/MILCOM.2006.302173.

[12] J. Harrald and T. Jefferson, “Shared situational awareness in emergency management mitigation and response”, in Proceedings of 40th Annual Hawaii International Conference on System Sciences (HICSS 2007), Jan. 2007, Waikoloa, HI, pp. 23 – 23, doi 10.1109/HICSS.2007.481.

[13] M. C. Zari, et al., “Personnel identification system utilizing low probability-of-intercept techniques: prototype development and testing”, in Proceedings on The Institute of Electrical and Electronics Engineers 31st Annual 1997 International Carnahan Conference on Security Technology, 15-17 Oct 1997, Canberra, ACT, pp. 224 – 230, doi 10.1109/CCST.1997.626274.

[14] J. Jormakka and T. Saarelainen, “UAV-based sensor networks for future force warriors”, International Journal On Advances in Telecommunications, vol 4, numbers 1 and 2, 2011, ISSN:1942-2601, pp. 58 – 71.

[15] R. Dees, S. Nestler, R.Kewley, and K. Ward, “WholeSoldier performance: A value-focused model of soldier quality”, 77th MORS Symposium, WG20- Manpower and Personnell, 35 pages, 7 Dec 07, 21.6.2010, accessed on 5.2.2015.

[16] Field Manual FM 1-02, www.armypubs.army.mil/doctrine/Active_FM.html FM 1-02, accessed on 8.12.2012.

[17] T. Saarelainen, “Targeting situational awareness beyond the event horizon by means of sensor element munition,” ICDT 2012, The Seventh International Conference on Digital Telecommunications, pp. 8 – 14.

[18] P. Buxbaum, “Denying access”, Special Operations Technology, July 2010, Vol 8, Issue 5, pp. 26 – 27.

[19] R. Kozma, et al., “Multi-modal sensor system integrating COTS technology for surveillance and tracking”, Radar Conference, 10-14 May 2010, pp. 1030 – 1035, doi 10.1109/RADAR.2010.5494467.

[20] T. Saarelainen and J. Timonen, “Tactical management in near real-time systems,” IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, (CogSIMA2011), Miami Beach, Florida, 22.–24 Feb 2011, U.S.A., pp. 240 – 247, 10.1109/COGSIMA.2011.5753452.

[21] M. Phillips, “Air-to-Ground and ground-to-air communications”, Military Technology, Vol XXXVII, Issue 6/2013, pp. 66 – 67.

[22] Z. Ying, W. Zhixue, L. Xiaoming, and C. Li, “C4ISR capability analysis based on service-oriented architecture”, The Fifhth IEEE International Symposium on Service Oriented System Engineering, 2010, pp. 179 – 180, DOI 10.1109/SOSE.2010.41.

Indoor Positioning and Navigation System for Interior Design Augmented Reality

Fady Adel, Mina Makary, and Mohamed El-Nahas
Computer and Communications Dept.
Alexandria University
Alexandria, Egypt
e-mail: fady.adel@mena.vt.edu
mina.makary@mena.vt.edu
mohamed.elnahas@mena.vt.edu

Mustafa ElNainay
Computer and Systems Engineering Dept
Alexandria University
Alexandria, Egypt
e-mail: ymustafa@alexu.edu.eg

***Abstract*—The problem of indoor localization and navigation has become increasingly important for a range of different applications. One of these applications is interior design augmented reality. Due to the inability of the global positioning system (GPS) to navigate indoors, and the lack of any global indoor localization scheme, research has turned to develop several approaches and solutions to indoor localization. Accuracy of these techniques ranges from few centimeters to few meters depending on the technology used. High accuracy is achieved through special hardware equipment. For interior design applications, users are interested in getting high accuracy in certain positions depending on the design layout. It is also desirable to use market tablets and mobile phones rather than special hardware for such applications. In this paper, a hybrid WiFi fingerprinting - sensor fusion scheme is proposed to achieve accurate indoor positioning and navigation using Android tablets. Moreover, an algorithm has been developed to determine the WiFi Access Points (APs) positions given the apartment layout. Results show high accuracy for both the positioning and navigation parts.**

Keywords–Indoor positioning; Indoor navigation; WiFi fingerprinting; Sensor fusion; Interior design augmented reality.

I. Introduction

Indoor localization systems work to locate and track objects within a closed environment such as a building or an office. Sample applications in such systems include positioning patients in hospitals, objects within a warehouse, employees in an office, customers in malls or locating people within a burning building. Indoor localization systems can also be used to enhance the clients experience in applications like Interior Design Augmented Reality. Indoor localization systems typically require a carefully planned environment involving Access Points (APs), sensors or other stationary or mobile equipment. Furthermore, there has been an increased demand on these systems to deliver high-accuracy with minimal cost and initial set-up.

Previous researches concerning this topic have achieved good results using WiFi fingerprinting, dead reckoning, camera-based techniques. Furthermore, integrated approaches have been developed to combine two or more techniques together. However, most of the proposed techniques do not target specific application scenario and hence, the resulted accuracy was not enough to be used in various applications that require very high accuracy, e.g., augmented reality applications. As we are biased towards low cost and high accuracy systems, we have been investigating the use of a hybrid system that combines the WiFi fingerprinting technique with the step detection technique to achieve better accuracy for interior design augmented reality applications.

This paper provides a number of contributions focusing on Interior Design Augmented Reality applications. Firstly, an algorithm is developed to determine the adequate number and locations for the APs used for the WiFi-Fingerprinting in order to differentiate easily between different Points of Interests (positions users are likely to stop at to view the interior design model). Secondly, to tackle the problem of different signal strengths in different cardinal directions, a map for each direction was used. Moreover, the orientation detection module was enhanced using an algorithm that estimates the step angle using the last five angles while excluding outlier values. Finally, a hybrid WiFi fingerprinting - sensor fusion scheme is developed to achieve accurate indoor positioning and navigation using Android tablets.

The remainder of this paper is organized as follows: In Section II, related work of research of high-accuracy indoor localization systems is surveyed. The problem assumptions adopted in this work are listed in Section III. The proposed solution for the automatic selection of the APs positions is presented in Section IV. In Section V, the proposed indoor positioning and navigation hybrid scheme for the interior design augmented reality application is presented. Performance evaluation details and results are discussed in Section VI. The paper is concluded in Section VII.

II. Related Work

Related work in indoor localization systems have achieved good results using WiFi fingerprinting, dead reckoning, camera-based techniques. Furthermore, integrated approaches have been developed to combine two or more techniques together. Various methods rely on multi-lateration of radio waves and location fingerprinting [1][2][3]. However, methods relying solely on multi-lateration usually provide relatively low accuracy (within 1-3 meter) as well as being dependent on existing infrastructure and often involve calibration. Methods relying on location fingerprinting usually provide slightly better accuracy than multi-lateration methods while also providing better opportunity for integration with other localization methods (e.g., dead reckoning) [4].

Shih et al. [5] attempt to deal with problems for fingerprint differences due to environment changes (weather changes, doors opening and closing, etc.). The main idea in the work is the clustering of reference points (RPs) based on the similarity of the path-loss exponent values and then uses the path loss propagation model to calculate RSS fingerprints dynamically in these RP clusters or regions. This calculation is achieved via sensors that re-measure the RSS value for each cluster. The work also contributes an algorithm to calculate accurate placement of these sensors. The work maintains a higher accuracy of 0-2 meters but incurs additional computational cost for the initial k-means clustering which also adds cost to the calibration phase.

The problem of noisy data (from sudden movements, band interference, etc.) has been tackled by Zhao et al. [6]. The system introduces three signal strength filters: the max filter, the limit filter and the move average filter. It was found that two of these filters (max filter and limit filter) provide higher accuracy to the k-NN algorithm employed for location estimation; increasing accuracy in 2.5 m range of error to 96% and reaching 98% in an error range of 3 m. The work also introduces a path tracking assistance that prunes the search space for the forecasted location to a predetermined subspace. This requires a stage whereby these subspaces are defined before real-time location determination but significantly lessens the localization execution time. Although the work by Zhao et al. [6] provided interesting insights, it failed to tackle heightened accuracy (less than 1 m) and also did not incorporate any smartphone sensor data into its localization algorithm.

So et al. [7] discuss a novel method to compare the run-time fingerprint to the patterns in the fingerprint data set instead of the Euclidean and probabilistic approaches. The authors propose two schemes to replace the Euclidean distance approach. The first scheme filters out large Received Signal Strength Indicator (RSSI) differences and hence when a difference increases beyond a certain threshold, the distance is no longer increased. The second scheme reduces the distance coefficient to 1 (instead of 2 as in Euclidean distance). This results in reducing the effect of large RSSI differences and give higher value to the exact match. The proposed schemes achieve a much better accuracy than the Euclidean distance metric. However, they also introduce additional problems of determining the proper threshold in the first scheme. Another problem is that it increases the sensitivity to small RSS differences (which is likely occurring due to fading).

Another category depends on device sensors, called dead reckoning, has shown promise in several approaches. Alzantot and Youssef [8] recognize and solve some of the problems of estimating motion based on accelerometer. The presented approach uses Support Vector Machine (SVM) to detect gait (walk, jog, etc.) and calculates displacement based on step count and step length estimations (derived from the gait). One of the work novel contributions is that its step detection (and hence count) is done via a Finite State Machine (FSM) that models the various states in an acceleration signal during a step. This removes problems of the smartphone changing the orientation as well as removes any need of preprocessing the accelerometer signal. Accuracy of navigation is about 4 m and 97% accuracy on the SVM classifier. Yim [9] presents a system designed to handle users walking around in large exhibition spaces. Its main contributions are improvements on the FSM proposed by Alzantot and Youssef in [8] as well as periodically assessing if the user is standing to watch a particular exhibit (via assessing the accelerometer y-axis magnitudes standard deviation). The system however, depends on the observation that users in exhibition spaces moved very slowly; an assumption that is not safe to make for other environments.

Integrated methods often utilize sensor fusion to combine inputs from multiple sensors in the Inertial Navigation System (INS) and RSSI for room-level navigation. The approach proposed by Holčík [10] uses sensor fusion to combine heading information from gyroscope and accelerometer sensors and utilizes a state model to model transition from one state to another based on three factors: the RSSI probability map, step length and stride length. This achieves a relatively bad accuracy of 2.3 meters. Chai et al. [11] combine inputs from barometer, WiFi fingerprinting and accelerometer. The proposed approach uses an Adaptive Kalman Filter (AKF) to incorporate measurement with the proposed state model. Although it achieves a slightly better accuracy at 1.65 meters, it also has the disadvantage of using individual sensors rather than smartphone ones. Le [12] uses a probabilistic model to combine RSSI and INS information. The model relies on a combination of likelihoods, i.e., the likelihood that location is accurate given RSS information and Previous location from INS, is equivalent to the probability of WiFi positioning based on RSS * probability of current location given the previous location, calculated according to a likelihood function. This is one of the best accuracies recorded for integrated systems and achieves an accuracy of 0.7 m.

Aside from the popular approaches discussed so far, other techniques have also been suggested in the literature for the localization problem that do not involve location fingerprinting or an integration of location fingerprinting and dead reckoning. For example, Filonenko et al. [13] provide 0.1 meter accuracy but changes the regular hardware used for localization to use four ultrasound microphones to provide for ultrasound (and inaudible) multi-lateration.

Previous work either uses special hardware to achieve high accuracy or cannot achieve the required accuracy for intrior design augmeneted reality applications. Our focus is to achieve high accuracy using commercially available Android tablets depending on the assumption that users of such applications will likely stop at certain positions to view the design and navigate between these positions. The next section presents our system assumptions.

III. System Assumptions

In our work, it is assumed that the location/apartment map with dimensions and the positions where the user will like to stop and view the interior design is known; we call these positions the Points of Interests (PoIs). The positioning and navigation module shall utilize this information to choose the number and positions of the WiFi APs and then builds a Radio Map of the signal strengths at each of the PoIs in an offline phase. Using the WiFi fingerprinting techniques, the positioning and navigation will map the user position to the nearest PoI when the module detects that the user has stopped, otherwise the step detection and counting module along with the orientation/heading direction are used to update the user position while moving from one PoI towards another one. It is also assumed that the user's direction is the same as the tablet's heading direction which is compatible with interior design applications needs. Finally, a subset of the corners will be chosen as the positions of the APs based on the algorithm described in Algorithm1 in the next section. Positioning the APs in corners limits the search space to a finite number of combinations. The following sections describe the system architecture and its main building blocks in more details.

IV. Determination of number of APs and their locations

The overall algorithm steps are described in Algorithm 1. At the beginning, PoIs are manually placed on the location map. The coordinates of these PoIs and the coordinates of the corners of the location are the inputs of the algorithm which determines the minimum number of APs and the best locations for them. This algorithm will differentiate between PoIs based on readings of the signals strength of the APs.

Algorithm 1 Determination of number of APs and their locations

```
Data: n(initial number of APs) = area / coverage area
Result: optimal number of APs and their locations
while true do
    c = get all combination of 'n' corners;
    for each combination do
        for each PoI do
            estimate the RSS from each AP;
        end
        if no more than 2 PoIs have the same RSSs range then
            return this combination;
        end
    end
    n++;
end
```

For a significant reduction in the processing time and resources, a mapping between signals strength and distance is used. It takes into consideration, other factors that affect the signal strength such as walls, doors, ceilings and floors. Thus, what are processed through the algorithm are Euclidean distances between each PoI and every corner, added to them the effects of the previously mentioned factors. Then, a distinct byte value is given for each range of distances. These values are then used to fill a 2D array where first dimension is number of PoIs and the second dimension is the number of corners that can be nominated to be the APs locations. Each step is explained in details, in the following sections.

A. Inputs and Combinations of corners

The inputs are the coordinates of the PoIs and the coordinates of the corners of the location as mentioned earlier. After this step, the corners will undergo enumeration process and get stored in an array. An initial number of APs needed, is calculated by dividing area of the location over the average coverage distance of the APs used. Each AP can be located in one of the corners. Thus, a combination module is developed to output all the possible combinations of corners (possible locations of the APs) to be used in the following step.

B. Calculating distances between PoIs and indexed Corners

Using the coordinates of the corners and the coordinates of the PoIs, the distance between each PoI and each corner is calculated and mapped to a byte value depending on its range from each AP location using the following ranges with gaps ([0m, 1m], [1.5m, 3.5m], [4.5m, 7.5m], > 8.5m) that have been determined through evaluation of different types and different brands of access points.

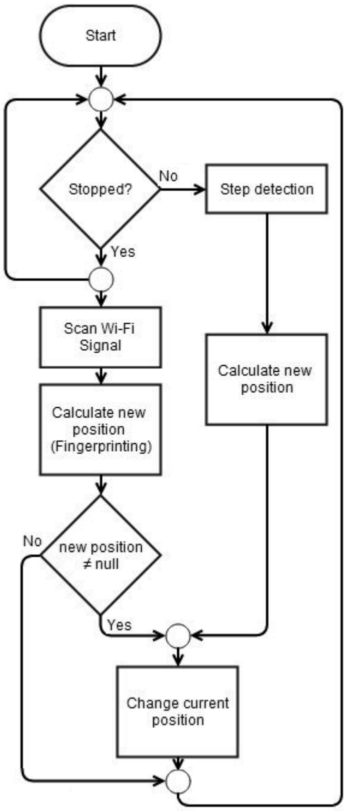


Figure 1. Overall System modules and interactions among them

C. Bitwise pattern and the Output

The 2D array is then filled where the elements in each row are the Byte-values resulting from the calculated distances that have been mapped to Bytes. Thus, each row is the bitwise pattern for the respective PoI. Each bitwise pattern of a PoI is XORed with the rest of the bitwise patterns of the rest of the PoIs. If any XOR operation resulted in a value equals zero (There are two PoIs that couldn't be differentiated from each other), this means that another combination needs to be tested. If all combinations failed for the current number of APs, the number of APs will be incremented and the new combinations that include the additional AP will be tested. Otherwise, if all the XOR operations for a combination resulted in a non-zero value this means that the corners in this combinations can be selected as the locations for the APs and can be successfully used to differentiate between all PoIs. Finally, the outputs are the number of APs needed and their locations.

V. Proposed Indoor Positioning and Navigation Scheme

The proposed hybrid scheme is depicted in Figure 1. The scheme consists of two main modules: Step detection and counting using Mobile Inertial Sensors and Localization using WiFi fingerprinting. The WiFi fingerprinting Localization is responsible for determining the user's location according to the nearest PoI ***where the PoI are predefined points provided by the Interior Designer (based on the developed design)***.

The WiFi fingerprinting module consists of two phases:

- Offline Phase: where the Radio Map is being built.
- Online Phase: where the user location is estimated.

On the other hand, the Step Detection and Counting module using the Inertial Sensors is responsible for determining the user's location while moving between PoIs. It contains two sub modules:

- The Step Detection Sub Module which determines the step boundaries based on the raw sensor values.
- The Orientation Detection Sub Module that estimates the user orientation/heading direction.

The two modules, the WiFi fingerprinting and Step detection and counting, will then be integrated to construct our Localization System.

A. Localization using the Inertial Sensors

This module determines the user's location while moving between PoIs. It contains two sub modules. The Step Detection Module which determines the step boundaries based on the raw sensor values and the Orientation Detection Module which estimates the user orientation/heading direction.

1) Step Detection Module: This module has three main components: the step sensitivity algorithm that is developed to estimate the user's step shape, the stop detection algorithm that is used to detect the stopping event of the user, and the step detection algorithm that is used to detect the steps of the user. The following subsections describe each component in details.

a) Step Sensitivity Algorithm: The basic idea behind this algorithm is to get estimation for the user's step shape (range of positive and negative accelerometer peaks). The user starts walking 20 steps during which the accelerometer values are being stored. Then, the positive and negative peaks are calculated with the average accelerometer values and are used to calculate a certain threshold that is used during the user's movement (in Section V-A1c) to estimate if the current accelerometer values represent a step or not.

b) Detect Stop Algorithm: In order to use the step detection algorithm in the hybrid system (in Section V-B2), the user's stopping event has to be detected. This algorithm depends on a simple check that verifies if the difference between the current time and the last step time is higher than a certain threshold (2 sec for example), which indicates that the user must have stopped.

c) Step Detection Algorithm: This algorithm is used to detect each step the user performs while he is moving. It simply searches in the current accelerometer values for a positive peak that is higher than the positive peak threshold specified previously during the Step Sensitivity process (in Section V-A1a), then the algorithm searches for a matching negative peak that is less than the negative peak threshold specified previously (in Section V-A1a). It then checks if the difference between those two peaks is less than a certain threshold and that the difference between the current time and the last step time is higher than 0.5 s, then it is considered as a step.

2) Orientation Detection Module: The aim of this module is to get the most accurate direction of the user during each step performed in order to determine the distance in the X direction and the Y direction after each step. First, the direction relative to the north is calculated using the magnetometer and accelerometer values. Each time the magnetometer and accelerometer values are renewed, a new direction is calculated and added to the angle list. When a step is detected, the last five angles in the angle list are retrieved and an algorithm is used to exclude two outlier angles (out of 5 angles) and get the average of the remaining 3 angles representing the step's angle.

3) Step Length: In our work, there different ways to estimate the user's step length have been investigated that range from the simplest to the most accurate.

- Put an average step length for all users (around 0.7m).
- Let the user enter his height as an input and use the Height-Step Length equation to get the corresponding step length
$$step\,length = 0.42 \times height$$
- Let the user walk a specified number of steps, then divide the distance travelled by that number to get the average step length for that specific user.

Finally, in order to determine the current user's location, first, we calculate Δx and Δy, which represent the change in location, and then add these values to the previous location. Δx = Lsinθ and Δy = Lcosθ where L represents the average step length and (θ) is the orientation angle, which is defined from the Orientation Detection Module (in Section V-A2).

B. WiFi Localization

There are generally two phases for location fingerprinting. First one is the offline phase in which AP' Received Signal Strength (RSS) samples are collected at RPs to build the Radio Map. Second one is the online phase in which the user location is estimated based on the RSS from each AP and the Radio Map prepared in the offline phase.

1) Offline Phase- Building the Radio Map: There are two steps to build the Radio Map. First, identify the RPs, then collect the RSS samples at each RP. In our system, we treat the PoI as the RPs, at which we collect the RSS samples. To build the Radio Map, fingerprints are collected at each RP. For each RP the user has to wait around 25 seconds for the system to collect 5 RSS samples. These values are then averaged to get one fingerprint for each RP and save it in the Radio Map. The user has to repeat this step for all the specified RPs. At first, all samples were collected in one direction. This results in low accuracy if the user stops at any PoI with a direction different than the one used in collecting RSS samples. To solve this problem, the radio map is built with samples in 4 main directions (North, South, East, and West). Four maps are generated from the radio map builder and used to locate the user while stopping at one of the PoIs. This approach increases the time needed to build the radio map(s) in the offline stage but increases the accuracy of the localization system in the online stage as well. ***The current orientation of the user when stopping at one of the PoIs is mapped to the nearest map.***

2) *Online Phase- Estimating the user location:* Fingerprinting algorithm works as follows; when the user stops, three APs' RSS readings are taken. After the first reading, a short list of the nearest PoIs (estimated by the Inertial Sensors Localization module in Section V-A) is constructed. Euclidean distances between the RSS and each fingerprint of this short list (stored in the Radio Map with the corresponding direction, mentioned in Section V-B1) are calculated. If the Euclidean distance of the Nearest Neighbor (NN) is less than a specific threshold, the current estimated position is set to the NN's location. After the third reading, a short list of the PoIs locating in specific covered range (calculated from the number of estimated steps and the average error per step) is constructed. Euclidean distances between RSS of each reading and each fingerprint of this short list are calculated to determine the NN from each reading. Then, a voting method is used to obtain the best candidate. If the Euclidean distance between the winning candidate and the RSS is less than a specific threshold, the current estimated position is modified by the candidate's location. The last PoI detected is not included in the short lists if the user takes a few steps away from the PoI and doesn't return back to prevent the fingerprinting from resetting the current position to the last PoI. Therefore a condition is added to check if the difference between the numbers of steps taken in opposite directions is smaller than certain value before adding the last PoI in the short list.

VI. Performance Evaluation

The testing environment is a location of $91m^2$ in area. It consists of 2 rooms and a hall. The first room is $33.5m^2$, the second is $22.5m^2$, while the hall is $35m^2$ in area. The location has 12 corners in total. Six points in this location are chosen to be the PoIs marked by symbol "'O"' in Figure 2. The coordinates of these corners and these PoIs are stored in a file to be used in the following scenarios.

A. Determination of number of APs and their locations

The file is then used as the input to the algorithm that was explained in Section IV. After testing the corner combinations, 2 APs can differentiate between all PoIs, and their locations are the corners marked by symbol "'X"' in Figure 2.

B. WiFi Localization

The WiFi fingerprinting map is then created using the 2 APs determined in the previous step and the wifi online phase is tested by standing at a each PoI and verifying the correctness of the detction. This experiment is repeated several times for all the PoIs to get an estimated average of the WiFi fingerprinting module accuracy. These experiments resulted in an average accuracy of 93%.

C. Step detection module

To evaluate the performance of the step detection module. A certain number of steps are performed, then the actual number of steps is compared with the number of steps detected by the developed navigation module. This experiment is repeated several times to get an estimated average of the step detection module accuracy. It can be stated that 90% of the steps are correctly detected by the step detection algorithm.

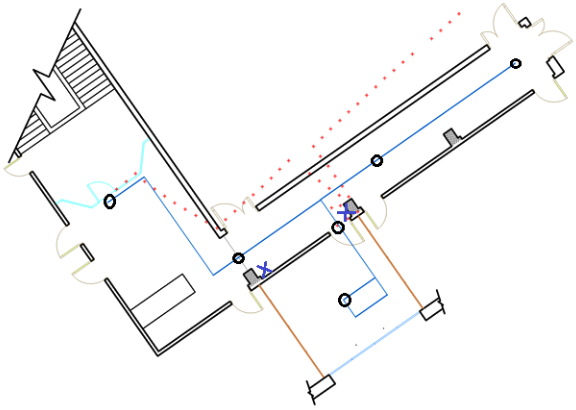


Figure 2. Test results of Localization using the Inertial Sensors

D. Orientation detection module

Testing scenario: Angles from a real compass were compared with angles provided by the sensors fusion. This experiment was repeated several times in different cardinal directions in order to calculate an approximate estimate of the Orientation detection module accuracy. The maximum angle deviation was in the range of -7 to +7 degrees with an average around 2 degrees.

E. Localization using the Inertial Sensors

The localization system using the inertial sensors is then tested without the integration with the WiFi fingerprinting. A certain number of steps is performed to get from a starting position to ending position and the difference between the actual ending position and the position given by the application is calculated. This experiment is repeated several times to get an estimated average of the step and orientation modules accuracy. The red dots in Figure 2 represent the actual positions of the user while the blue dots represent the estimated positions. The accumulated step detection error was up to 2.6m, which is a relatively low accuracy, as it can be seen in Figure 2.

F. Overall Hybrid Localization System

The previous testing scenario is repeated for the integrated hybrid WiFi fingerprinting and sensor fusion system. The red dots in Figure 3 represent the actual positions of the user while the blue dots represent the estimated positions. Using the integrated system, the step detection error was corrected constantly each time the user stopped (using the Wifi fingerprinting module), which led to a less localization error ranging from 0.5 m to 0.7 m.

The overall integrated system accuracy was acceptable to the interior design augmented reality applications where the user is likely need to view the design at the PoIs with high accuracy and can accept slightly less accurate view while moving among PoIs.

VII. Conclusion and Future Work

In this paper, a hybrid WiFi fingerprinting - sensor fusion scheme was presented to achieve accurate indoor positioning and navigation using Android tablets. The target application for our system is the interior design augmented reality applications where users are interested in getting high accuracy in certain positions depending on the design layout and can accept less accuracy while moving among these positions. An algorithm

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-396-4

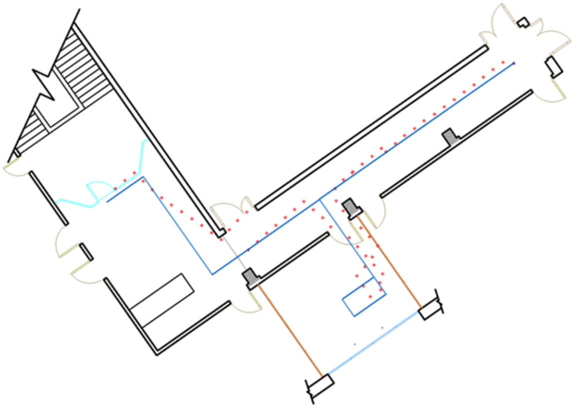


Figure 3. Test results of Localization using the Hybrid System

has been developed to determine the WiFi Access Points (APs) positions given the apartment layout. The hybrid localization system uses WiFi fingerprinting to cancel the sensor fusion-based navigation error at certain posistions. Results show high accuracy of 90% for less than 0.7m for the hybrid integrated system.

Future extensions of this work include the comparison with other related techniques in the same environment and with the same system assumptions. The replacement of the WiFi access points with beacons is currently investigated as promising cheaper alternative. Better integration with the augmented reality part through the utilization of the tablet's multiple cores is also of interest.

ACKNOWLEDGMENT

This work has been supported by a grant from the Information Technology Academia Collaboration (ITAC) program of the Information Technology Industry Development Agency (ITIDA).

REFERENCES

[1] M. Youssef and A. Agrawala, "The horus wlan location determination system," in Proceedings of the 3rd International Conference on Mobile Systems (MobiSys '05), Applications, and Services, Seattle, Washington. ACM, 2005, pp. 205–218.

[2] L. F. M. de Moraes and B. A. A. Nunes, "Calibration-free wlan location system based on dynamic mapping of signal strength," in Proceedings of the 4th ACM International Workshop on Mobility Management and Wireless Access (MobiWac '06), Terromolinos, Spain. ACM, 2 Oct. 2006, pp. 92–99.

[3] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg, "Compass: A probabilistic indoor positioning system based on 802.11 and digital compasses," in Proceedings of the 1st International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH '06), Los Angeles, CA, USA. ACM, 2006, pp. 34–40.

[4] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, no. 6, Nov. 2007, pp. 1067–1080.

[5] C.-Y. Shih, L.-H. Chen, G.-H. Chen, E.-K. Wu, and M.-H. Jin, "Intelligent radio map management for future wlan indoor location fingerprinting," in 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 1-4 Apr. 2012, pp. 2769–2773.

[6] Y. Zhao, Q. Shen, and L. Zhang, "A novel high accuracy indoor positioning system based on wireless lans," Progress in Electromagnetics Research C, vol. 24, 2011, pp. 25–42.

[7] J. So, J.-Y. Lee, C.-H. Yoon, and H. Park, "An improved location estimation method for wi-fi fingerprint-based indoor localization," International Journal of Software Engineering and Its Applications, vol. 7, no. 3, 2013, pp. 77–86.

[8] M. Alzantot and M. Youssef, "Uptime: Ubiquitous pedestrian tracking using mobile phones," in 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 1-4 Apr. 2012, pp. 3204–3209.

[9] J. Yim, "A smartphone indoor positioning method," International Journal of Smart Home, vol. 7, no. 5, 2013, pp. 9–18.

[10] M. Holčík, "Indoor navigation for android," Master's thesis, Faculty of Informatics, Masaryk University, Brno, Spring 2012.

[11] W. Chai, C. Chen, E. Edwan, J. Zhang, and O. Loffeld, "2d/3d indoor navigation based on multi-sensor assisted pedestrian navigation in wi-fi environments," in Ubiquitous Positioning, Indoor Navigation, and Location Based Service (UPINLBS), Helsinki, Finland, 3-4 Oct. 2012, pp. 1–7.

[12] M. H. V. Le, "Indoor navigation system for handheld devices," Ph.D. dissertation, Worcester Polytechnic Institute, 2009.

[13] V. Filonenko, C. Cullen, and J. D. Carswell, "Indoor positioning for smartphones using asynchronous ultrasound trilateration," ISPRS International Journal of Geo-Information, vol. 2, no. 3, 2013, pp. 598–620.