IARIA

AFIN 2013

The Fifth International Conference on Advances in Future Internet

ISBN: 978-1-61208-300-1

August 25-31, 2013

Barcelona, Spain

AFIN 2013 Editors

Renzo Davoli, University of Bologna, Italy

Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania

AFIN 2013

Foreword

The Fifth International Conference on Advances in Future Internet [AFIN 2013], held between August 25-31, 2013 in Barcelona, Spain, continued a series of events dealing with advances on future Internet mechanisms and services.

We are in the early stage of a revolution on what we call Internet now. Most of the design principles and deployments, as well as originally intended services, reached some technical limits and we can see a tremendous effort to correct this. Routing must be more intelligent, with quality of service consideration and 'on-demand' flavor, while the access control schemes should allow multiple technologies yet guarantying the privacy and integrity of the data. In a heavily distributed network resources, handling asset and resource for distributing computing (autonomic, cloud, on-demand) and addressing management in the next IPv6/IPv4 mixed networks require special effort for designers, equipment vendors, developers, and service providers.

The diversity of the Internet-based offered services requires a fair handling of transactions for financial applications, scalability for smart homes and ehealth/telemedicine, openness for web-based services, and protection of the private life. Different services have been developed and are going to grow based on future Internet mechanisms. Identifying the key issues and major challenges, as well as the potential solutions and the current results paves the way for future research.

We take here the opportunity to warmly thank all the members of the AFIN 2013 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 AFIN 2013. 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 AFIN 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AFIN 2013 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 future internet.

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.

AFIN 2013 Chairs:

Jun Bi, Tsinghua University, China

Eugen Borcoci, University Politehnica of Bucharest, Romania

Petre Dini, Concordia University - Montreal, Canada / China Space Agency Center - Beijing,China

AFIN 2013

Committee

AFIN Advisory Chairs

Petre Dini, Concordia University - Montreal, Canada / China Space Agency Center - Beijing, China
Eugen Borcoci, University Politehnica of Bucharest, Romania
Jun Bi, Tsinghua University, China

AFIN 2013 Technical Program Committee

Rocío Abascal Mena, Universidad Autónoma Metropolitana - Cuajimalpa, México
Marie-Hélène Abel, University of Technology of Compiègne, France
Alessandro Aldini, University of Urbino "Carlo Bo", Italy
Javier A. Barria, Imperial College London, UK
Jun Bi, Tsinghua University, China
Alessandro Bogliolo, University of Urbino, Italy
Eugen Borcoci, University Politehnica of Bucharest, Romania
Christos Bouras, University of Patras and Research Academic Computer Technology Institute, Greece
Tharrenos Bratitsis, University of Western Macedonia, Greece
Chin-Chen Chang, Feng Chia University, Taiwan
Maurizio D'Arienzo, Seconda Università di Napoli, Italy
Guglielmo De Angelis, CNR - ISTI, Italy
Sagarmay Deb, Central Queensland University, Australia
Gayo Diallo, University of Bordeaux Segalen, France
Daniel Díaz-Sánchez, University Carlos III - Madrid, Spain
Sudhir Dixit, HP Labs India - Bangalore, India
Jonas Etzold, Fulda University of Applied Sciences, Germany
Florian Fankhauser, TU-Wien, Austria
Wu-Chang Feng, Portland State University, USA
Liers Florian, TU Ilmenau, Germany
Alex Galis, University College London, UK
Ivan Ganchev, University of Limerick, Ireland
Rosario G. Garroppo, Università di Pisa, Italy
Christos K. Georgiadis, University of Macedonia, Greece
Apostolos Gkamas, Higher Eccelesiastic Academy Vellas of Ioannina, Greece
George Gkotsis, University of Warwick, U.K.
William I. Grosky, University of Michigan-Dearborn, USA
Vic Grout, Glyndwr University, U.K.
Adam Grzech, Wrocław University of Technology, Poland
Puneet Gupta, Infosys Labs, India
Dongsoo Han, Korea Advanced Institute of Science and Technology(KAIST), Korea
Sung-Kook Han, Won Kwang University, Republic of Korea
Gerhard Hancke, Royal Holloway, University of London, UK

Copyright Information

Table of Contents

Inter-Operator Traffic Differentiation based on Multiscale Analysis

Nelson Coelho, Paulo Salvador, António Nogueira
DETI, University of Aveiro/Instituto de Telecommunicações
Aveiro, Portugal
email: {nelsonmiguel, salvador, nogueira}@ua.pt

***Abstract*—Web 2.0 changed the interaction paradigm of Internet users, placing them in a more active role as both producers and consumers of digital contents. This concept has also triggered the appearance of social networks and cloud computing services, which have an increasing contribution to the total traffic amount. The increasing Internet complexity brings new challenges to network operators and managers, which need to understand new applications and know the exact properties of the generated traffic. The ability to accurately map traffic patterns to their corresponding application can be used to build efficient traffic and user profiles, which can be extremely helpful in several critical tasks like network resources optimization, service differentiation and personalization, network management and security. This paper proposes a classification approach that is able to accurately differentiate traffic flows in a core network and associate them with their underlying applications, allowing the construction of accurate traffic and user profiles. By performing a wavelet decomposition and analyzing the obtained scalograms, the captured traffic can be fully characterized in terms of its time and frequency components. As the different frequency components of the traffic are inferred, an appropriate communication profile characteristic of each application type can be defined. This way, it is possible to identify the distinct applications that are being used by the different connected clients and build useful user profiles.**

Keywords - traffic identification; profiling; multi-scale analysis; wavelet transform.

I. INTRODUCTION

The ability to accurately build efficient traffic and user profiles has a crucial importance in many network operation and management tasks; it can be used to infer the most appropriate bandwidth and delay requirements for each user or group of users, allowing and optimized distribution of the network resources and improving the values of the Quality of Service (QoS) parameters; it will allow network managers to easily create groups of users requesting similar contents, easing the delivery of appropriate and related contents and services; security standards can be improved because it will allow the detection of users presenting illicit profiles or profiles including unknown applications, triggering alarms and providing counter-actions while allowing the remaining connected clients to experience better QoS levels.

This paper proposes a methodology for the creation of traffic profiles based on the classification of collected traffic flows, that is, based on their mapping to the corresponding generating applications. The proposed classification approach performs a wavelet decomposition at several scales of analysis; it is known that lower scales comprise low frequency events, which are typically created by user clicks and applications synchronization events; mid-range frequency components are related to the creation of Internet sessions; higher scales of analysis capture higher-frequency events, such as packet arrivals and packet bursts. So, by decomposing the traffic generated by different clients running diverse applications and analyzing it at various time scales, the methodology will be able to build a *multi-scale application profile* depicting the different frequency components that are characteristic of the mostly used applications.

Figure 1 represents the generic architecture of a traffic classification system with QoS support. Aggregated traffic is monitored by several network probes, which collect the necessary traffic flows at representative time periods. Using wavelet transforms, the most relevant components of the traffic flows are extracted, resulting in scalograms that are used to build accurate profiles for each one of the Internet applications whose traffic belongs to the aggregate. The profiles of the different applications are stored in a Profiles Database, which at bootstrap, only contains known profiles created in controlled environments or classified using deep-packet inspection (these are known as *training traces*). While capturing and classifying traffic, the different traffic profiles can be updated with the newly inferred profiles, after a validation process that may include payload inspection or human validation. The classifier associates the captured traffic to different service classes, characterized by several QoS parameters. All inferred/calculated data will also feed the User Profiling module, which is responsible for updating the different user profiles.

The efficiency of the proposed traffic classification methodology will be evaluated by applying it to aggregated inter-operator traffic, captured in the backbone network of a tier 1 Internet Service Provider (ISP). The results obtained show that the approach is efficient, being able to accurately differentiate Internet applications and, thus, having the potential to be the key component of a traffic and user profiling architecture.

The rest of the paper is organized as follows: Section II presents some of the most relevant related work on traffic classification and profiling; Section III provides some background on multi-scale analysis; Section IV describes the proposed classification methodology; Section V presents the traffic traces that are used to test the proposed methodology; Section VI presents and discusses the main results obtained and, finally, Section VII presents the main conclusions.

II. BACKGROUND ON TRAFFIC CLASSIFICATION AND PROFILING

Traffic classification efforts started by simple port-based identification approaches, where ports used by the different traffic

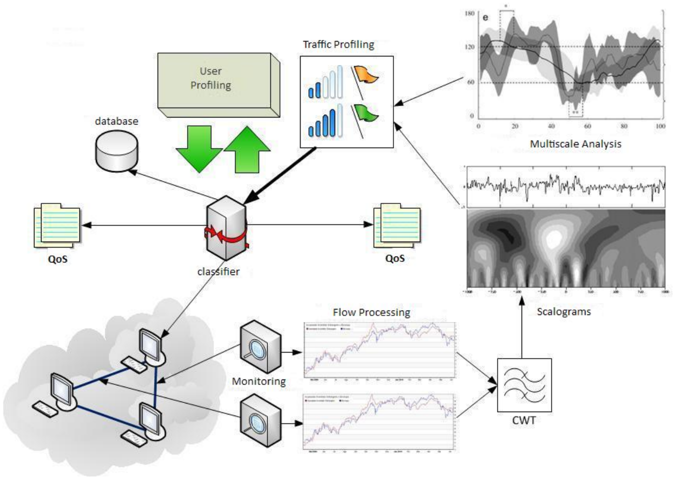


Fig. 1. Architecture of a traffic classification and user profiling system with QoS support.

flows were exclusively used to identify the applications that generated them. Since many protocols started to use random port numbers or ports generally associated to other protocols for bypassing *firewalls* and proxies, port-based approaches could no longer provide an accurate identification of Internet traffic [1].

Payload-inspection appeared as an evolutionary solution, inspecting the payload of captured packets in order to search for application level signatures of known applications. This approach relies on the use of extensive databases, containing known signatures and patterns of many Internet protocols, which are used as a comparison term whenever any new captured traffic has to be classified. This methodology was able to achieve good classification results, being used by several of the currently available commercial products [2] [3]. However, the databases associated to the classification approach need to be constantly updated in order to comply with new and emerging protocols. Besides, legal restrictions prevent Internet Service Providers from analyzing the contents of the users packets [4], while technical issues such as scalability on high-speed links also prevent researchers and Internet Service Providers from using payload inspection approaches.

Statistical analysis of the traffic flows appeared as the solution that could overcome these restrictions [5]. Moore *et al.* [6] proposed several flow discriminators and machine learning techniques to select the best discriminators for classifying flows. Hu *et al.* [7] built behavioral profiles describing dominant patterns of the studied applications and the classification results obtained showed that the approach was quite promising. In Huang *et al.* [8], authors attempted to describe negotiation behaviors by capturing traffic discriminators available at early negotiation stages of network flows and several machine learning algorithms were deployed to assess the classification accuracy. This way, they were able to conclude that the approach was suitable for *real-time* application identification. In a recent work [9], multi-dimensional probabilistic approaches were used to model the multi-scale traffic patterns generated by several Internet applications and to match the analyzed traffic to its generating application(s). However, these techniques can not efficiently differentiate between similar web-applications in scenarios where there is no access to layer 3 (and above) information and payloads. Hybrid classification approaches have also been used: in Tavallaee *et al.* [10], for example, a two-level hybrid approach in which payload analysis is combined with machine-learning algorithms was used to classify unknown traffic based on its statistical features.

There are several definitions of user profile [11], but a common definition can state that an user profile consists of a description of the user interests, behaviors and preferences. Therefore, the process of creating an user profile can be seen as the process of gathering the appropriate information until all these characteristics are obtained. In Claffy *et al.* [12], a parameterizable methodology for profiling Internet traffic flows at different granularities was proposed. Flows were defined based on traffic satisfying various temporal and spatial locality conditions, as observed at internal points of the network instead of only end-point definitions. In Xu *et al.* [13], a real-time behavior profiling system for high-speed Internet links was proposed, using flow-level information from continuous packet or flow monitoring systems and relying on data mining and information-theoretic techniques to automatically discover significant events based on the communication patterns of end-hosts. Reverse Domain Name System lookups, which are used to determine the domain name associated with an Internet Protocol (IP) address, have also been used to provide a simple association between a domain and the services it is known to run. A similar work was carried out in Trestian *et al.* [14], where authors stated that all information needed to profile any Internet endpoint is available in the Internet itself: therefore, accurate profiles were built by simply querying the most used search engine (Google) and dividing the querying results into several tags describing the requested services. However, inspection techniques can not be applied in scenarios where layer 3 and layer 4 information is not available, such as networks where authentication and encryption mechanisms are deployed.

III. Multi-Scale Analysis

The use of a wavelet decomposition through the Continuous Wavelet Transform (CWT) allows the analysis of any process in both time and frequency domains, being widely used in many different fields such as image analysis, data compression and traffic analysis. The CWT of a process $x(t)$ can be defined as [15]:

$$\Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{+\infty}^{-\infty} x(t)\psi^*(\frac{t-\tau}{s})dt \qquad (1)$$

where $*$ denotes the complex conjugation, $\frac{1}{\sqrt{|s|}}$ is used as an energy preservation factor, $\psi(t)$ is the *mother wavelet*, while τ and s are the translation and scale parameters, respectively. By varying these parameters, a multi-scale analysis of the entire captured process can be performed, providing a description of the different frequency components present in the decomposed process together with the time-intervals where each one of those components is located.

The wavelet scalogram can be defined as the normalized energy $\hat{E}_x(\tau, s)$ over all possible translations (set $\mathbf{T}$) in all analyzed scales (set $\mathbf{S}$), and is computed as:

$$\hat{E}_x(\tau, s) = 100 \frac{\left|\Psi_x^{\psi}(\tau, s)\right|^2}{\sum_{\tau' \in \mathbf{T}} \sum_{s' \in \mathbf{S}} \left|\Psi_x^{\psi}(\tau', s')\right|^2} \quad (2)$$

The volume bounded by the surface of the scalogram is the mean square value of the process. The analysis of these scalograms enables the discovery of the different frequency components, for each scale (frequency) of analysis. Assuming that the process $x(t)$ is stationary over time, several statistical metrics can be obtained, such as the standard deviation:

$$\sigma_{x,s} = \sqrt{\frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} (\hat{E}_x(\tau, s) - \mu_{x,s})}, \forall s \in \mathbf{S} \quad (3)$$

where $\mu_{x,s} = \frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} \hat{E}_x(\tau, s)$, and $|\mathbf{T}|$ denotes the cardinality of set $\mathbf{T}$.

IV. Classification Methodology

By defining characteristic regions of the scalogram statistics, for the different applications, in different frequency subsets, it is possible to identify profiles presenting components characteristic to each one of the applications. Such regions are inferred from the scalograms obtained from the decomposition of the *training traces* of each web-application.

Let us consider the (positive) region R_a^+ as the region defined as a function of a frequencies (positive) sub-set $\mathbf{s}_a^+$ and energy variation (positive) sub-set $\mathbf{\Sigma}_a^+$ for which we always have the characteristic statistical values of application a. Moreover, let us define the (negative) region R_a^- as a function of a frequencies (negative) sub-set $\mathbf{s}_a^-$ and energy variation (negative) sub-set $\mathbf{\Sigma}_a^-$ for which we never have characteristic statistical values of application a.

$$R_a^+ = f(\mathbf{s}_a^+, \mathbf{\Sigma}_a^+) \quad (4)$$

$$R_a^- = f(\mathbf{s}_a^-, \mathbf{\Sigma}_a^+) \quad (5)$$

A traffic trace process $x(t)$ is classified as belonging to application a if for all scales belonging to sub-set $\mathbf{s}_a^+$ the energy standard deviation $\sigma_{x,s}$ belongs to region R_a^+ and, simultaneously, for all scales belonging to sub-set $\mathbf{s}_a^-$ the energy standard deviation $\sigma_{x,s}$ does not belong to region R_a^-:

$$C(x) = a \Leftarrow \forall s \in \mathbf{s}_a^+, \sigma_{x,s} \in R_a^+ \wedge \forall s \in \mathbf{s}_a^-, \sigma_{x,s} \notin R_a^- \quad (6)$$

The classification decision can be made as soon as all conditions are met. Note that, even if time $\mathbf{T}$ grows and allows more classification precision, decisions can nevertheless be made with small $\mathbf{T}$ sub-sets (short-time analysis and decision).

The inference of regions R_a^+ and R_a^- (defined by $\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-$) can be performed by solving the following optimization problem:

$$\max_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \in \mathbf{I}_a} C(i) == a \right) \wedge \min_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \notin \mathbf{I}_a} C(i) == a \right), \forall a \quad (7)$$

where $==$ represents a comparison function witch outputs 1 if both terms are equal and 0 if terms are different. $\mathbf{I}_a$ represents the subset of processes (known as) belonging to web-application a. This optimization problem was solved (not for the optimal solution) using exhaustive search. However, more advanced algorithms can be applied to find (sub)optimal solutions.

Several regions can be created, in the various frequency subsets, for each studied application a. The higher the number of regions of an application, the higher the ability to analyze the different frequency components and consequently, a more accurate traffic mapping can be achieved. An algorithm was used to automatically define such regions (obviously satisfying the above conditions) using known simple geometrical equations, such as ellipses.

V. Traffic Traces

The traffic traces used in the evaluation tests were obtained from the Cooperative Association for Internet Data Analysis (CAIDA) [16]. Traces of this repository have one hour duration and were collected every month in each one of the passive monitors managed by the organization [17]. Specifically, the datasets used in this paper were collected on July 21, 2011, after 12:59 PM, at the Equinix data center, located in Chicago, Illinois, which is connected to the backbone line of a level 1 ISP between Chicago and Seattle, Washington. The connection is bidirectional, so traces contain packets circulating in both directions. After data collection, CAIDA proceeded to data anonymization due to legal and privacy requirements [16]. Captures are divided into several ".cap" files whose duration varies between fifty and fifty nine seconds. In our evaluation studies, we have considered the traffic traces corresponding to the first five minutes, in both directions. Traffic flows are identified by the traditional 5-tuple definition (source IP address, destination IP address, source port, destination port and protocol). Note that both cities are located on different time zones, which helps explaining the differences observed on the traffic flows in both directions. All tests were made on a desktop computer equipped with an Intel Core i5 CPU 650 @ 3.20GHz x 4 processor and running the *Ubuntu* 12.04 LTS operating system. *Tshark* and *wireshark* [18] were used to calculate the relevant statistics of the traffic traces.

The following applications, all contributing to relevant percentages of current Internet traffic, were considered in this study: Hypertext Transfer Protocol (HTTP), Simple Mail Transfer Protocol (SMTP), Real Time Streaming Protocol (RTSP), Mobile Status Notification Protocol (MSNP) and XBOX Live. For each service, four contexts were analyzed: downstream and upstream traffic on the client side, downstream and upstream traffic on the server side. However, due to space restrictions, we will only present here results corresponding to the client side, although a similar analysis also applies to the other scenarios.

VI. Evaluation Results

As an illustrative example (similar plots were also made for all the other applications), Figure 2 shows the traffic volume over

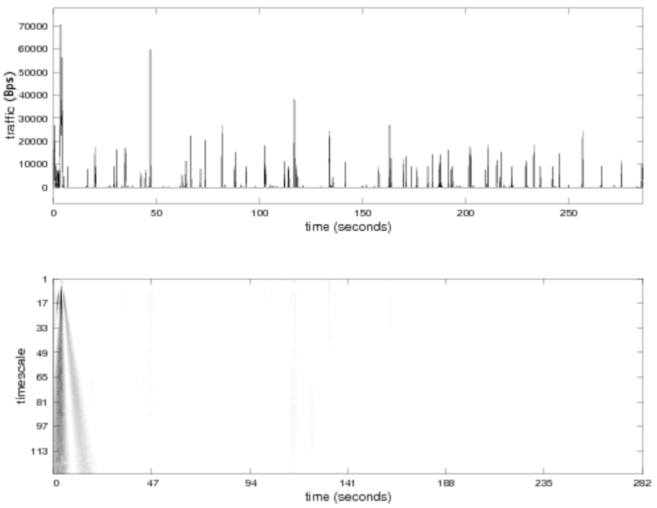


Fig. 2. HTTP downstream (from the client side) traffic: (up) traffic volume over time; (bottom) scalogram.

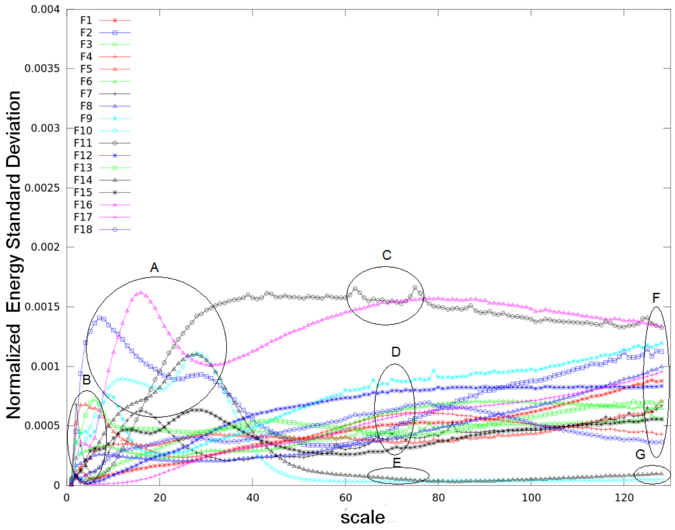


Fig. 3. Energy standard deviation of different HTTP downstream traffic flows.

time and the scalogram corresponding to HTTP downstream traffic. In this case, traffic peaks have low frequency, low amplitude and are not periodic, which indicates that this user is performing typical browsing. The corresponding scalogram exhibits visible frequency components at the beginning of the temporal scales range. Note, however, that HTTP traffic flows corresponding mainly to video downloads or file sharing have significantly different behaviors, so HTTP traffic presents a high diversity of patterns due to the wide spectrum of services that run on top of this protocol. This diversity is visible when analyzing flows from all the different HTTP usage profiles.

Figure 3 represents the energy standard deviation of different HTTP downstream traffic flows. Several regions can be identified, corresponding to different types of human and/or network events. Region A includes events having low frequency and moderate energy variation, usually generated by user clicks when accessing to online news, browsing photos and using social networks. Region B includes low frequency events with small energy variation, typical of the visualization of online video sites. Region C contains two flows (8 and 11) including medium frequency events with high energy variation, related to the creation of a high number of Transport Control Protocol (TCP) and HTTP sessions. Region D involves medium frequency events with a slower energy variation. Events included in region E have a very small energy variation, that is, a low number of TCP and HTTP sessions, typical of social networks applications (flows 10 and 14). Region F includes events with small and moderate energy variations, so flows located in this region have a moderate number of high frequency events, corresponding to the arrival of a reasonable number of packets. Finally, region G exhibits high frequency components with a very reduced energy variation, so flows 10 and 14 (which are located in this region) are generated by applications that are responsible for a reduced number of packets, like for example photo sharing or email clients.

In Figure 4, a single region is able to include all low frequency flows. The energy variation of these flows is small, indicating that they were generated by infrequent user clicks. In the medium frequencies range, we can find two flows (10 and 12), responsible for events with some energy variation (region B); region C includes the remaining flows, characterized by a small number of created sessions, which is responsible for their reduced energy variation. These characteristics are usually related to social networks applications and visualization of photos and videos. In the high frequency segment, two distinct regions can be identified: region D contains a flow (flow 12) characterized by a considerable percentage of high frequency events, which is a clear sign that this flow is responsible for significant upstream traffic originated at this client; region E incorporates the remaining flows and is characterized by events with high frequency components but with a reduced energy variation (much smaller when compared to region D), which corresponds to a reduced upstream packet rate characteristic of scenarios where users are not browsing in a very active way (which can correspond to video visualization scenarios, reading specific emails or reading news feeds on social networks' sites).

Figure 5 represents the energy standard deviation of different SMTP downstream traffic flows. In this case, most of the traffic flows have a similar behavior, being included in three different frequency regions: B, D and F. These flows have a small percentage of frequency components for all scales, which is associated to a low number of user clicks, few TCP sessions that were opened and a reduced number of packets that were exchanged. These results are according to the expected behavior of the client download traffic associated to SMTP. However, there are some flows diverging from this pattern. In region A (flows 2, 7, 11, 15 and 18) it is possible to find low frequency events with moderate energy variation, which implies more user clicks when comparing to flows located in region B. Region C (flows 7, 11, 15 and 18) incorporates medium frequency events with moderate energy variation; so, these flows have more HTTP and TCP interactions when compared to flows from region D. Finally, region E contains

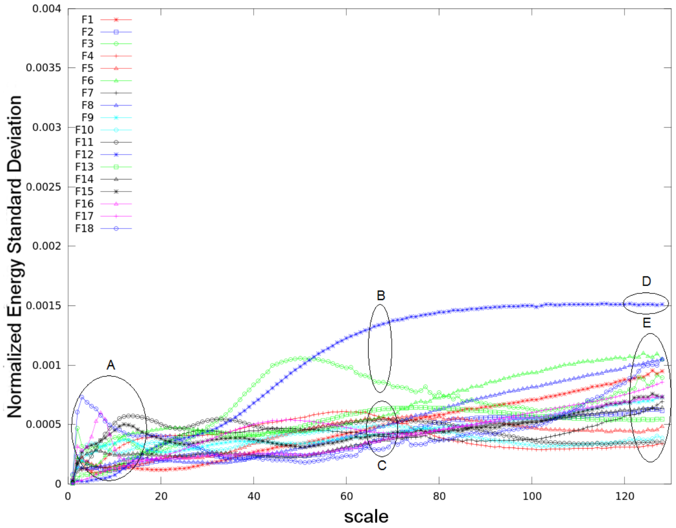


Fig. 4. Energy standard deviation of different HTTP upstream traffic flows.

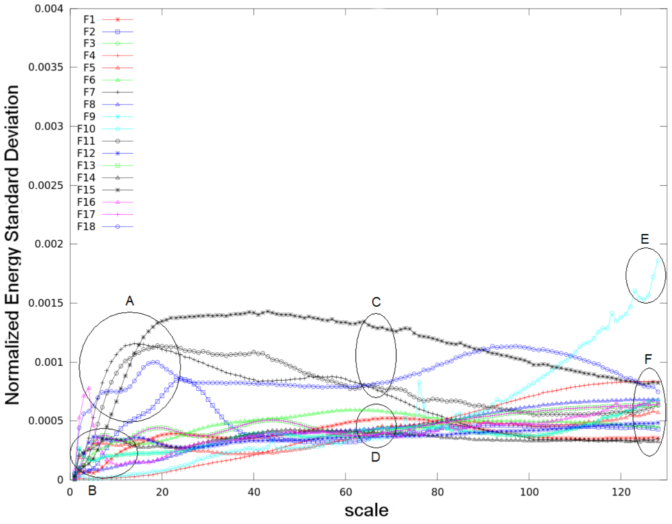


Fig. 5. Energy standard deviation of different SMTP downstream traffic flows.

a single flow (flow 10) with a very high energy variation, indicating that this flow corresponds to sending one or two large volume emails due to the high number of packets that were detected.

Figure 6, which represents the energy standard deviation of different SMTP upstream traffic flows, shows two regions in the low frequencies segment: region A includes very low frequency events, generated by the initial download of the email application interface and by the subsequent automatic synchronizations of the client email box. Region B corresponds to situations where, besides these events, the client performs other operations in his email interface, like sending emails to other contacts. Region C includes all traffic flows in the medium frequencies segment, having small to moderate energy variation, which correspond to the establishment of some TCP sessions during the analyzed time period. Finally, in the high frequencies segment two regions can be identified: region D, with a significant percentage of high frequency

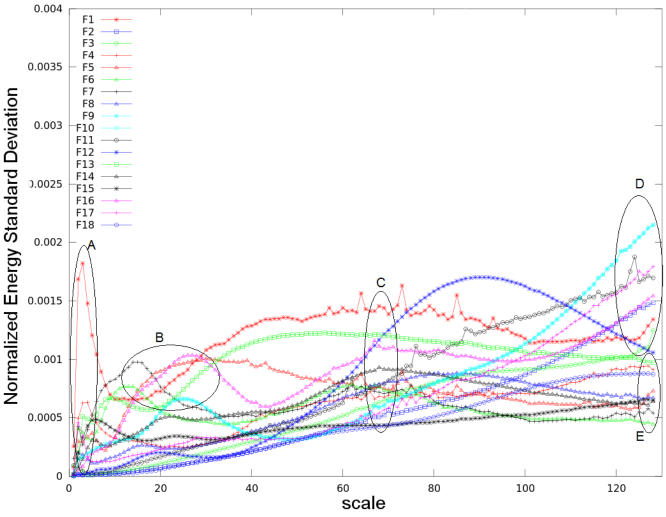


Fig. 6. Energy standard deviation of different SMTP upstream traffic flows.

components, and region E with a reduced percentage of high frequency components. Thus, region D is associated to events generated by sending large quantities of packets from the client to the server (specially, large emails), while region E is associated to situations where the client activity is more reduced and the packet exchange is only limited to control and synchronization data between the interface of the email application and the SMTP server.

Figure 7 represents the energy standard deviation of different RTSP downstream traffic flows. It is possible to verify the existence of several traffic flows corresponding to very low frequency events with high energy variation, which are generated by events such as automatic and periodic data synchronizations from the server and responses to clicks that were made by the client itself. Region B involves low frequency events with considerable energy variation, usually associated to requests for new contents; in the specific case of this protocol, this can correspond to the choice of new streams to visualize. In the medium frequencies segment, traffic flows are divided in two regions: region C includes flows with a considerable energy variation, which are obviously related to TCP and RTSP interactions; in fact, energy variations having this kind of amplitude should be related to client requests for establishing TCP and RTSP sessions dedicated to visualize streams. Region F includes high frequency events with a reduced percentage of high frequency components: the existence of few events in this region can be attributed to the reduced downstream packet transmission rate, which is quite unusual if streaming transmissions work as expected. So, we can assume that in this case the stream content was transmitted for a small period of time or some problems have occurred in the transmission while trying to visualize specific contents. Regarding region E, detected events have a considerable energy variation, which proves that the downstream packet transmission for these flows is higher than the one corresponding to flows of region F. Note that flows 8 and 9 have irregular peaks on the standard deviation of the normalized energy, specially in some specific intervals: region A, medium frequencies segment an very high

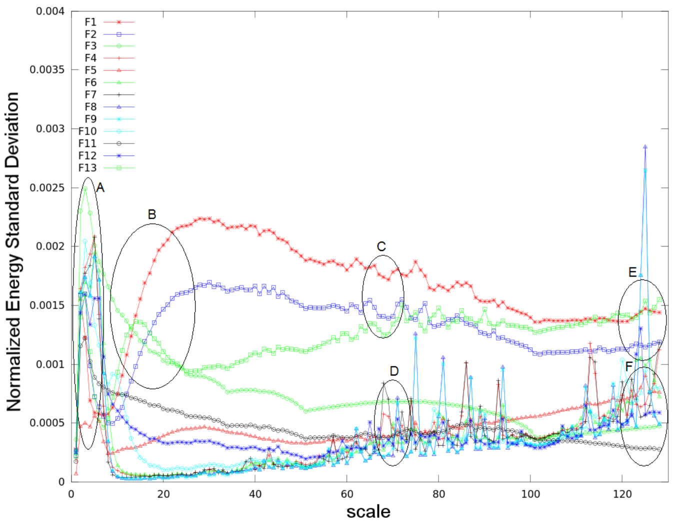


Fig. 7. Energy standard deviation of different RTSP downstream traffic flows.

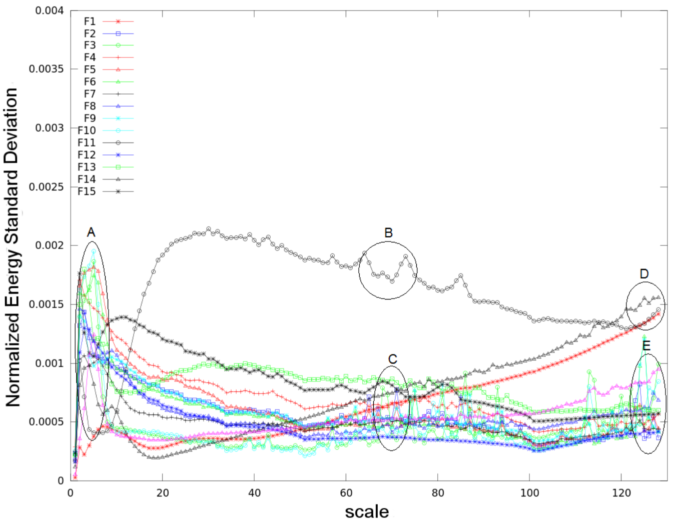


Fig. 8. Energy standard deviation of different RTSP upstream traffic flows.

frequencies segment. This irregularity in the energy pattern can be explained by the fluctuations on the packet transmission rate of video streaming applications: there are periods with high packet transmission rates (video is immediately processed and displayed), followed by periods where transmission rates are lower, resulting in some video freezing occurrences.

Figure 8 represents the energy standard deviation of different RTSP upstream traffic flows. Region A corresponds to very low frequency events, generated by rare occurrences: this should correspond to a very low number of client solicitations that are sent to the server or to the exchange of very small size packets. The medium frequencies segment contains two regions (B and C), although region B only contains one flow. This flow consists of events presenting energy variations higher than those corresponding to events from region C, so for this specific traffic flow there are more TCP and RTSP interactions and more TCP sessions established upon client request. We can assume that the client is trying to access various streams (in order to select the one that presents the highest quality) or is forced to update his browser because the quality of the connection/visualization is not as good as it should be. Regarding the high frequencies segment, we can observe that most of the analyzed traffic flows are located in region E and have small energy variation. The three flows located in region D correspond to events having higher energy variation than the one corresponding to events associated to region E, which means that there is a higher upstream packet transmission rate. Note that flows 3 and 9 have an irregular shape in the medium and very high frequencies segments, probably due to the instability of the connection between client and server.

Figure 9 represents the energy standard deviation of different MSNP downstream traffic flows. In the low frequencies segment, there is only one region, region A, including all traffic flows. Events corresponding to this region have a very low frequency, typically presenting a periodic pattern that makes sense if the time intervals between writing and reading messages is similar. In the medium frequencies segment, all flows (except flow 4) belong to region C and present a small energy variation because sending text messages to other online users creates a small number of sessions and few User Datagram Protocol (UDP) and MSNP interactions. Flow 4 stands out from the others because its events show a considerable energy variation. This can be explained by the fact that this MSNP user is talking at the same time as other online users, resulting in a higher number of MSNP and UDP interactions. In the high frequencies segment, we can see that most of the analyzed flows belong to region E and present a reduced packet arrival rate, which is expected if we take into account that text messages that are usually sent by this application require small packet sizes. Flows 4 and 7 do not follow this rule, being located at region D, where the energy variation of the generated events is slightly higher than the one corresponding to flows from region E because the packet arrival rate itself is higher. This again can be explained by the fact that this MSN client has established new conversations with other users.

Figure 10 represents the energy standard deviation of different MSNP upstream traffic flows. Flows located in region A present the same behavior of the corresponding flows represented in region A of Figure 9, although in that case packets were originated at the server while now they are originated at the client. The medium frequencies segment is divided in two regions: region C includes traffic flows with small energy variation, while region B includes flows with moderate energy variation, corresponding to more MSNP and UDP interactions, that is, in these flows the client interacts with more users of the same application. In the high frequencies segment, we can see that all flows (except flow 10) are located in region E. Here, packet transmission rates are usually small, suggesting that the client interacts with few users of the same application. Regarding region D, it only contains flow 10: events associated to this flow generate a quite reasonable packet transmission rate from the client, meaning that this particular client interacts with others in a very active way.

Figure 11 represents the energy standard deviation of dif-

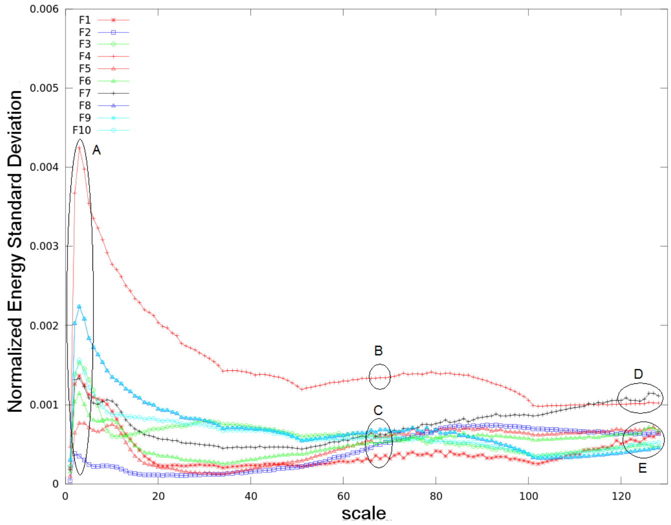


Fig. 9. Energy standard deviation of different MSNP downstream traffic flows.

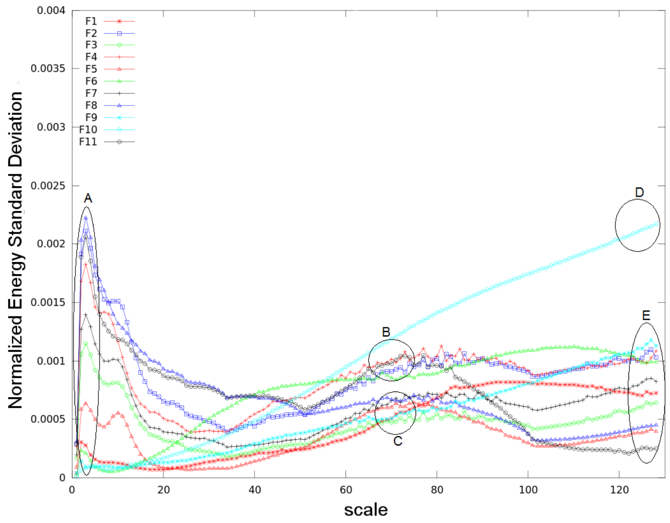


Fig. 10. Energy standard deviation of different MSNP upstream traffic flows.

ferent XBOX downstream traffic flows. There are two regions in the low frequencies segment: regions A and B. The first one includes very low frequency events, which rarely happen, such as the user activity in the XBOX Live service interface or automatic synchronizations between the remote server and the client terminal in order to assure an appropriate quality level for the connection. Region B includes low frequency events with small energy variation amplitude. Region C is located in the medium frequencies segment and includes all traffic flows because for this application energy variation does not differ significantly from one flow to another. The low value of energy variation indicates that few UDP sessions are created while the client is playing. In the high frequencies segment, three regions can be identified: region F, where the packet rate is very low; region E, where the packet rate is higher and region D where the packet reception rate is quite high (flows 4, 8, 9 and 18). It is possible to assume that flows of region D are associated to games where the user has a more active role, like

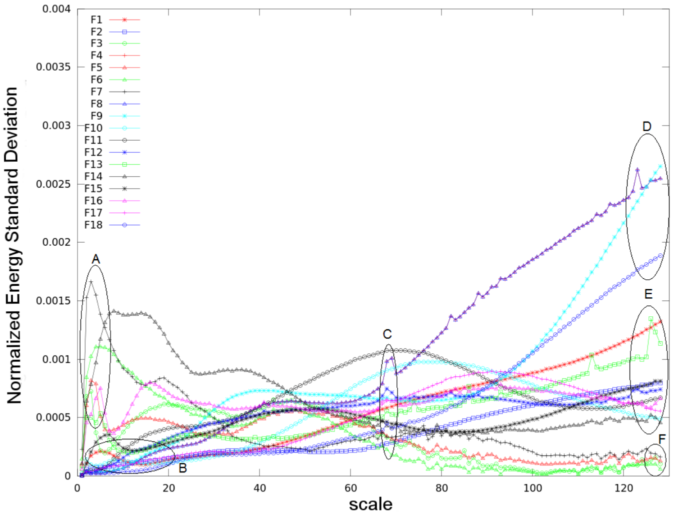


Fig. 11. Energy standard deviation of different XBOX downstream traffic flows.

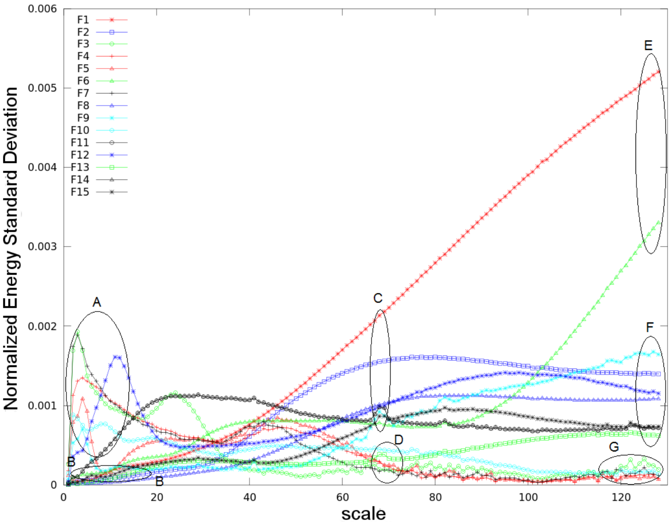


Fig. 12. Energy standard deviation of different XBOX upstream traffic flows.

personalized action games, sports or adventure games. Flows of region E can be associated to question/answer or to strategy games.

Finally, Figure 12 represents the energy standard deviation of different XBOX upstream traffic flows. Region E includes two flows (1 and 6) with high energy variation and high packet transmission rate between client and server, possibly corresponding to scenarios where the XBOX user is making a lot of clicks or playing a game that requires a high activity level. Region F, corresponding to a lower transmission rate, can be associated to games where the user is not so active and there are some inactivity periods between the different user actions (like, for example, games of the question/answer type).

From this discussion, we can conclude that a multi-scale analysis based on the wavelet transform is able to efficiently highlight the most important distinguishing features/characteristics of each Internet application. By performing a wavelet decomposition of the traffic flows, the different

TABLE I
PERCENTAGE OF CORRECTLY CLASSIFIED FLOWS.

Application	Download	Upload
HTTP	90.2%	93.1%
SMTP	77.8%	74.9%
RTSP	84.5%	82.3%
MSNP	91.1%	89.7%
XBOX	84.2%	88.6%

time and frequency components can be identified from the scalogram of the traffic metrics, allowing an efficient mapping of the captured traffic to the corresponding application.

Table I shows the classification results obtained by applying the previously discussed methodology. First of all, a set of *training flows* is used to define the elliptic areas that are used to differentiate the most relevant characteristics of each application; then, a set of *testing flows* is classified based on those areas in order to evaluate the efficiency of the classification approach. As can be seen, the classification results are quite good, with an identification accuracy higher than 70% for all cases. These classification results correspond to 50 runs and 95% confidence intervals were also built: since their widths are very small, they were omitted in the table.

VII. CONCLUSIONS

The complexity of current Internet forces network operators and managers to understand the underlying mechanisms of applications/services and the exact properties of the generated traffic. The ability to accurately map traffic patterns to their corresponding application can be used to build efficient traffic and user profiles that can be very useful in several operational and management tasks. This paper proposed a classification approach that is able to accurately differentiate traffic flows in the core network of a tier 1 ISP and associate them with their underlying applications. By performing a wavelet decomposition and analyzing the obtained scalograms, the captured traffic can be fully characterized in terms of its time and frequency components. This way, appropriate application profiles can be built, allowing the identification of all distinct applications that are being used by the different connected clients and the definition of useful user profiles.

REFERENCES

[1] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," Lecture Notes in Computer Science, vol. 3431, 2005, pp. 41–54.

[2] [retrieved: June, 2013] Snort home page. [Online]. Available: http://www.snort.org/

[3] [retrieved: June, 2013] Cisco IOS Intrusion Prevention System (IPS) - Products and Services. [Online]. Available: http://www.cisco.com/en/US/products/ps6634/index.html

[4] F. McSherry and R. Mahajan, "Differentially-private network trace analysis," in Proceedings of ACM SIGCOMM, 2010, pp. 123–134.

[5] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," IEEE Communications Surveys Tutorials, vol. 10, 2008, no. 4, pp. 56–76.

[6] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in Proceedings of ACM SIGMETRICS, 2005, pp. 50–60.

[7] Y. Hu, D.-M. Chiu, and J. Lui, "Application identification based on network behavioral profiles," 16th International Workshop on Quality of Service, 2008, pp. 219–228.

[8] N.-F. Huang, G.-Y. Jai, and H.-C. Chao, "Early identifying application traffic with application characteristics," in IEEE International Conference on Communications, May 2008, pp. 5788 –5792.

[9] E. Rocha, P. Salvador, and A. Nogueira, "Detection of illicit network activities based on multivariate gaussian fitting of multi-scale traffic characteristics," in IEEE International Conference on Communications, Jun. 2011.

[10] M. Tavallaee, W. Lu, and A. Ghorbani, "Online classification of network flows," in Seventh Annual Communication Networks and Services Research Conference, May 2009, pp. 78 –85.

[11] D. Godoy and A. Amandi, "User profiling in personal information agents: a survey," Knowledge Engineering Review, vol. 20, Dec. 2005, no. 4, pp. 329–361.

[12] K. Claffy, H. Braun, and G. Polyzos, "A parameterizable methodology for internet traffic flow profiling," IEEE Journal of Selected Areas in Communications, vol. 13, Oct. 1995, no. 8, pp. 1481–1494.

[13] K. Xu, F. Wang, S. Bhattacharyya, and Z.-L. Zhang, "A real-time network traffic profiling system," in 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Jun. 2007, pp. 595 –605.

[14] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Googling the internet: profiling internet endpoints via the world wide web," IEEE/ACM Transactions on Networking, vol. 18, Apr. 2010, no. 2, pp. 666 –679.

[15] J. Slavic, I. Simonovski, and M. Boltezar, "Damping identification using a continuous wavelet transform: application to real data," Journal of Sound and Vibration, vol. 262, 2003, no. 2, pp. 291 – 307.

[16] [retrieved: June, 2013] Caida - The Cooperative Association for Internet Data Analysis". [Online]. Available: http://www.caida.org/home/

[17] M. Fomenkov and K. Claffy, "Internet measurement data management challenges," in Workshop on Research Data Lifecycle Management, Princeton, NJ, Jul, 2011.

[18] [retrieved: June, 2013] Wireshark: go deep. [Online]. Available: http://www.wireshark.org/

An Inter-channel and Intra-channel Dynamic Wavelength/Bandwidth Allocation Algorithm for Integrated Hybrid PON with Wireless Technologies for Next Generation Broadband Access Networks

N. Moradpoor
School of Engineering, Computing and Applied Mathematics
University of Abertay Dundee
Dundee, UK
e-mail: n.moradpoor@abertay.ac.uk

G. Parr, S. McClean, and B. Scotney
School of Computing and Information Engineering
University of Ulster
Coleraine, UK
e-mails: gp.parr@ulster.ac.uk
si.mcclean@ulster.ac.uk
bw.scotney@ulster.ac.uk

***Abstract*—Optical and wireless technology integration schemes merge the high-speed and high-capacity of the optical networks with the low-cost, wide-coverage and mobility features of wireless counterparts for Subscriber Stations (SSs). It is also financially viable for the telecommunication service providers particularly in rural areas. In order to successfully integrate the two technologies, there are some technical concerns in terms of Architectural aspects, Physical Layer features and Media Access Control (MAC) related issues. This paper is mainly focused on the analysis of the key topics in MAC-related issues over the converged scenario and proposes an Inter-channel and Intra-channel Dynamic Wavelength/Bandwidth Allocation (IIDWBA) algorithm where the hybrid Passive Optical Network (PON) acts as a backhaul technology for the wireless counterpart. Performance of the proposed algorithm is evaluated through conducted simulation scenarios in terms of different Quality of Service (QoS) metrics where the IIDWBA algorithm shows a better performance when it is compared with the scenario in which it has not been employed.**

Keywords: optical and wireless technology integration, Media Access Control (MAC), resource allocations

I. INTRODUCTION

PON [11], [12], and [27] is the most promising candidate among optical access solutions in terms of maintainability and robustness. There have been various efforts on PON multiplexing techniques such as Time Division Multiplexing (TDM), Wavelength Division Multiplexing (WDM) and Code Division Multiplexing (CDM). TDM-PON reduces the cost per subscriber and has inexpensive network components as it requires only one transmitter in the Optical Line Terminal (OLT), as well as only one type of transmitter in Optical Network Units (ONUs) [23]. However, TDM-PON sacrifices the maximum available bandwidth per subscriber and limits the number of supported subscribers up to 32 [21]. On the other hand, WDM-PON provides multiple wavelength channels with a good security and protocol transparency which offers higher bandwidth and supports more subscribers [6]. In WDM-PON, a WDM transmitter, particularly in a subscriber side, is the most critical component where the associated transmitter should be precisely aligned with the allocated channel [21]. Unlike TDM-PON, in WDM-PON, the OLT needs to have an array of transmitters with one transmitter for each ONU. Every ONU also needs to have a wavelength-specific laser. Generally speaking, although PON has been viewed as an attractive solution for the “first/last mile” bandwidth bottleneck problem, extending fibre-based infrastructures of PON to the rural area is either too costly or inaccessible. Moreover, PON is unable to provide wireless access services.

World Wide Interoperability for Microwave Access (WiMAX, IEEE 802.16 [10]) standard comes into play as a wireless matching part for PON technology. WiMAX aims to reduce the equipment, operation and maintenance costs and capable of providing the low-cost, wide coverage, fixed and mobile broadband access connections with the QoS provisioning scheme [5]. It also provides wireless access services for rural areas where the development of copper-based technologies or fibre-based broadband is too expensive or inaccessible. However, WiMAX copper-based backhaul technology is still a controversial issue. This is the point where optical and wireless technology integrations come into play. PON can be used as a scalable, cost-aware and potential solution for WiMAX backhaul problem whereas WiMAX can extend PON infrastructure to rural areas with relatively lower cost. In order to successfully integrate the optical and wireless technologies, there are some challenging issues that need to be addressed efficiently and effectively in order to provide the smooth End-To-End (ETE) technology integrations.

To the best of our knowledge, the traditional single channel TDM-PON has been addressed in most of the existing work related to the optical and wireless integration scenarios. Therefore, we are motivated to combine the wavelength routing and high-capacity of WDM-PON, the power-splitting and lower-cost of TDM-PON with high coverage and mobility features of the wireless counterpart for the integrated scenario. In order to provide the full dynamic wavelength/bandwidth allocations across hybrid PON integration with wireless technology, an IIDWBA algorithm is proposed. The remainder of this paper is prepared as follow. In Section II, the existing work related to the optical and wireless technology integrations is briefly discussed. The proposed IIDWBA algorithm is discussed fully in Section III. Sections IV and V include the simulation model and the captured results, respectively followed by the conclusions in Sections VI and references.

ISBN: 978-1-61208-300-1

II. RELATED WORK

To date, a wide range of research has been carried out on the successful integration of the optical and wireless technologies. The integrated scenario has been considered in three categories: *Architectural*, *Physical layer* and *MAC layer* issues. The *Architectural* aspects [1], [3], [7], [13], [16], [17], [24], and [25] include the way two technologies connect to each other. In *Physical layer* issues, most of the works were focused on providing the cost-effective and reliable Radio over Fibre (RoF) systems [19]. As the research in this paper is related to the MAC-related issues of the converged scenario, the previous works related to this aspect are selected and discussed as follow. The MAC-related issues for the integrated scenario were discussed for the first time in [7]. The authors raised several issues for the bandwidth allocations, packet scheduling and QoS support. Ou et. al [29] investigated the scheduling techniques aimed at improving the performance and guaranteeing the QoS for different class of services. Ou et. al [3] proposed a slotted Dynamic Bandwidth Allocation (S-DBA) algorithm which aimed to increase the bandwidth utilizations by reducing the signaling overhead caused by the cascading bandwidth requests and grants. Ou et. al [22] proposed an intra ONU-Base Station (BS) scheduling algorithm termed Hybrid Priority Weighted Fair Scheduling (HPWFS) to progress the QoS performance without bandwidth starvations for the lower priority class of services. Ou et. al [13] proposed a DBA algorithm for the suggested Optical-Optical-Wireless (OOW) architecture which was executed in three levels. Ou et. al [20] investigated the possible challenging issues for the integrated structures of the TDM-PON and WDM-PON with WiMAX and Wireless Fidelity (Wi-Fi) networks. Performance evaluations of the existing scheduling techniques for three popular service classes were studied which showed the strong impact of using an efficient up-link scheduler in converged scenario. Please refer to [4], [8], [15], [18], and [25] for more work related to the MAC-layer aspects of the converged scenario. To the best of our knowledge, TDM-PON has been addressed in most of the existing work related to the integration scenario. Using the traditional single channel TDM-PON, where a group of ONUs (typically 16) sharing a single channel as a backhaul for 802.16 BS, provides each BS with ~ 62.5 Mb/s capacity which is almost matched the 802.16 channel capacity (~ 70 Mb/s over a 20 MHz channel). However, 62.5 Mb/s does not seem to be enough when a given ONU is employed as a backhaul for more than a single BS. This is the point where WDM-PON comes into play where multiple wavelengths will be available over a same fibre channel. Thus higher bandwidth can be provided by the OLT for a given ONU, more BSs and, finally, more number of SSs can be supported. In terms of MAC-related issues, an IIDWBA algorithm is proposed in this paper to provide the full distributed and dynamic wavelength/bandwidth allocations across OLT - ONUs as well as ONU - BSs. While a given DBA algorithm deals with bandwidth allocations inside a given channel, the IIDWBA algorithm works on top of a given DBA algorithm inside the OLT and ONU in association with multiple channels.

III. IIDWBA ALGORITHM

In order to save space, a given Server Station is termed SST, which can be the OLT or ONU, and a given Client Station is termed CS, which can be the ONU or BS. An SST is an element that provides resources for a given CS periodically and a given CS is an element that asks for resources from a given SST regularly. A given SST can be a CS and vice versa at any time. In this paper, a given ONU is a CS when it asks for resources from the OLT and is a SST when it provides resources for a given BS. Moreover, a given BS is a CS when it asks for resources from the associated ONU and an SST when it provides resources for the associated SSs. However, the OLT and the SSs are the ultimate SST and CSs, respectively. The IIDWBA algorithm works in three phases as follow.

A. *Phase one: Initialisation phase*

Phase one includes the auto-discovery and registration processes during which the CSs join the converged network and a wavelength will be assigned to each of them. In this phase, the IIDWBA algorithm first identifies the total number of the supported channels as well as the total number of the CSs associated per SST. The average number of the CSs per channel will be calculated next. Then it starts randomly allocating the channel identifiers (IDs) to all the CSs in such way that the number of the CSs per channel will be the same. Finally, the allocated channel IDs will be acknowledged to all the CSs associated with a given SST. In phase one, the objectives are as follow:

1) Identifying the average number of the CSs per channel per SST.
2) Assigning a default channel ID to all the CSs associated with a given SST.
3) Finishing the registration process for each CS and receiving the first associated queue status.

B. *Phase two: Intra-channel bandwidth allocations phase*

Phase two will be executed immediately after phase one and will be accomplished once per service cycle inside all the channels per SST. This phase is responsible for allocating bandwidth inside a given channel by considering the actual bandwidth requests and minimum guaranteed bandwidth per CS per service cycle. At the end of a given bandwidth allocation cycle, the local information such as the total number of the heavily loaded CSs and associated MAC addresses, the total excess bandwidth, the total excess requested bandwidth and the generated service cycle will be delivered to the phase three. In order to distinguish how the bandwidth is granted from a given SST to the associated CSs during a given service cycle, the *Limited Bandwidth Allocation Scheme* [14] and [50], is discussed first as follows. In Limited Bandwidth Allocation Scheme, if the requested bandwidth from a given CS is less than the minimum guaranteed bandwidth, the requested bandwidth will be granted; otherwise the minimum guaranteed bandwidth will be granted. This approach provides excess bandwidth remaining from the CSs which requested bandwidth less than the minimum guaranteed bandwidth (lightly loaded CSs).

One solution to employ the excess bandwidth is to distribute it fairly among those CSs, which requested bandwidth more than the minimum guaranteed one (heavily loaded CSs), as proposed in [8]. However, the difference between the work in this paper and the work in [8] is that, in this paper, the excess amount of bandwidth, which is remaining from a given channel, will be used globally if it is not employed by the local CSs associated with it. In phase two, the objectives are as follow:

1) Allocating bandwidth to local CSs on a given channel.
2) Scheduling the local service cycle.
3) Capturing the total number of the heavily loaded CSs and their MAC addresses, total local excess bandwidth and total local excess requested bandwidth.
4) Sending objective two and three to the Global scheduler.

C. Phase three: Inter-channel bandwidth allocations phase

Phase three will be executed immediately after the phase two and includes three real-time stages of *Collect*, *Schedule* and *Distribute*.

During the *Collect* stage, the IIDWBA algorithm collects the local information from each channel including the number or the local heavily loaded CSs with associated MAC addresses, total excess bandwidth, total excess requested bandwidth and the latest scheduled service cycle. During the *Schedule* stage, based on the number of the total heavily loaded CSs on all the channels (globally heavily loaded CSs), global excess requested bandwidth, and the global excess available bandwidth, the IIDWBA algorithm schedules the global excess bandwidth among all the globally heavily loaded CSs. During the *Distribute* stage, the globally scheduled excess bandwidth will be distributed among the globally heavily loaded CSs inside the associated service cycle and be immediately broadcast to all the channels. In phase three, the objectives are as follow:

1) Receiving the total number of the heavily loaded CSs and their MAC addresses, total local excess bandwidth, and total local excess requested bandwidth from all the channels on a given SST.
2) Receiving the latest scheduled service cycle from all the channels.
3) Identifying the lightly/heavily loaded channels.
4) Calculating the total number of the heavily loaded CSs across all the heavily loaded channels.
5) Calculating the total excess bandwidth across all the lightly loaded channels.
6) Calculating the average granted excess bandwidth from each lightly loaded channel to a given globally heavily loaded CS.
7) Allocating the global excess bandwidth to the globally heavily loaded CSs according to the actual need.
8) Scheduling and embedding the global allocated excess bandwidth inside associated service cycle.
9) Sending all the service cycles for all the channels to broadcaster.

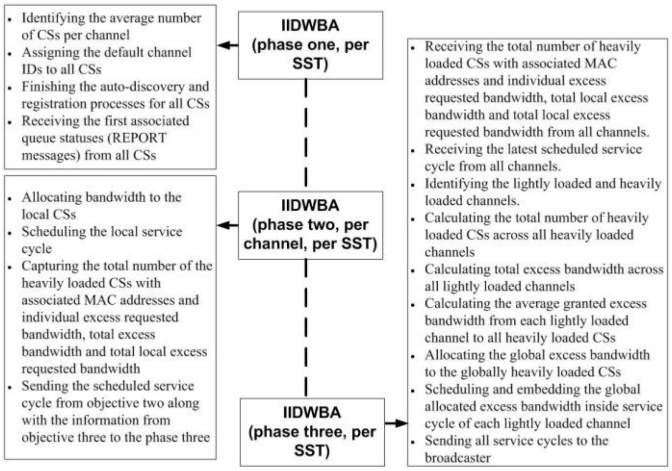

IIDWBA (phase one, per SST)
• Identifying the average number of CSs per channel
• Assigning the default channel IDs to all CSs
• Finishing the auto-discovery and registration processes for all CSs
• Receiving the first associated queue statuses (REPORT messages) from all CSs
IIDWBA (phase two, per channel, per SST)
• Allocating bandwidth to the local CSs
• Scheduling the local service cycle
• Capturing the total number of the heavily loaded CSs with associated MAC addresses and individual excess requested bandwidth, total excess bandwidth and total local excess requested bandwidth
• Sending the scheduled service cycle from objective two along with the information from objective three to the phase three
IIDWBA (phase three, per SST)
• Receiving the total number of heavily loaded CSs with associated MAC addresses and individual excess requested bandwidth, total local excess bandwidth and total local excess requested bandwidth from all channels.
• Receiving the latest scheduled service cycle from all channels.
• Identifying the lightly loaded and heavily loaded channels.
• Calculating the total number of heavily loaded CSs across all heavily loaded channels
• Calculating total excess bandwidth across all lightly loaded channels
• Calculating the average granted excess bandwidth from each lightly loaded channel to all heavily loaded CSs
• Allocating the global excess bandwidth to the globally heavily loaded CSs
• Scheduling and embedding the global allocated excess bandwidth inside service cycle of each lightly loaded channel
• Sending all service cycles to the broadcaster

Figure. 1 Specific tasks for each phase of the IIDWBA algorithm

The specific tasks for each phase of the IIDWBA are specified in Fig. 1 per channel per SST.

The next Section evaluates the performance of the IIDWBA algorithm through simulated scenarios.

IV. SIMULATION MODEL

The performance of the proposed IIDWBA algorithm is evaluated by conducting a simulated scenario using OPNET Modeler [9]. The simulation scenario, Fig. 2, includes: a single OLT in Central Office (CO), which supports four channels (w_1,…,w_4) and is related to 16 ONUs, a 1:4 sized Arrayed Waveguide Grating (AWG) with a co-located amplifier and four 1:16 sized TDM Splitters. AWG and TDM Splitters are seated between the OLT and ONUs, respectively. A given ONU is also assigned to 16 BSs and supports the same four channels (w_1,…,w_4). A 1:4 sized AWG with a co-located amplifier and four 1:16 sized TDM Splitters are also located between a given ONU and BSs, respectively. Each BS is also associated with 10 wireless SSs which will be increased to 100 SSs in stage by 10 to evaluate the performance of the IIDWBA algorithm under the different load values. A given BS also supports the same four channels (w_1,…,w_4). Simulation parameters, see Table I, are employed for the experiments. Moreover, the simulation scenarios ran for three seed values of 128, 166 and 90. However, we depicted the average plots from three runs in this paper. Based on widely used configurations for the converged scenario and traditional TDM-based PON [1] and [14], the simulated scenario is carried out as follow. The buffer sizes inside the ONUs and BSs are set to finite 10 Mbytes and the maximum cycle time is considered as 2 ms. A fixed 192µs is considered as the Round Trip Time (RTT) delay for each CS in every service cycle. Moreover, 100 Mb/s and 1Gb/s are the upstream data rates between a given BS and the associated ONU as well as a given ONU and the OLT. A fixed guard time of 5µs is considered for the light sources on ONUs and BSs. Moreover, the MPCP Extension protocol [2] is employed to support all the communications among the components. An uneven

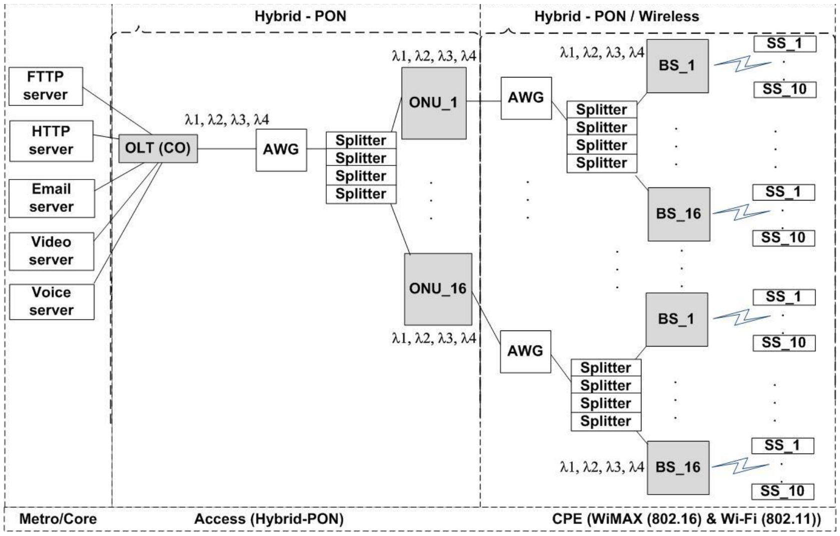


Figure. 2 Network topology for evaluating the proposed IIDWBA algorithm

traffic pattern is considered across all the channels on a given SST where lightly loaded and heavily loaded CSs are always available during the simulation run time. Two traffic patterns termed first traffic pattern and second traffic pattern are considered for the experiments. In the first traffic pattern, the load on w_1 is gradually increased from 10 SSs to 100 SSs per BS (total of 160 to 1600 SSs per ONU) while the number of the SSs on other channels are fixed, up to 10 SSs (total of 160 SSs per ONU), in order to distinguish how CSs on w_1 benefit from the available free bandwidth on the other channels when the traffic builds up. In the second traffic pattern, the load on all channels is increased by gradually raising the number of the SSs per BS from 10 to 100 (total of 160 to 1600 SSs per ONU) to distinguish the performance of the CSs associated with w_1 from the results captured in the first traffic pattern.

TABLE I. SIMULATION PARAMETERS

Traffic pattern	Burst (uneven across SSs)
ON and OFF state time (sec)	20% and 80% of simulation time
Traffic start time	even across all SSs
Traffic stop time, Packet size	Never, 500 bytes (constant)
Number of SSs per BS	10 to 100 BY 10
Traffic class	Best Effort (BE)
Simulation time, Seed	30 sec, 99, 128, 166
value per static, update interval	1600, 300000

V. CAPTURED RESULTS

After employing the first traffic pattern, as Fig. 3 reveals the proposed IIDWBA algorithm is successful in decreasing the average queuing delay to almost 14% for the ONUs associated with w_1, when it is compared with the scenario without applying the IIDWBA algorithm [8]. It is because, when the load starts increasing on channel one, w_1, the IIDWBA algorithm starts looking for the excess bandwidth on neighbouring channels (w_2, w_3,w_4) which will be collected and distributed among the ONUs associated with w_1. The average queuing delay for the ONUs associated with w_2, w_3,w_4 is also captured with and without applying the IIDWBA algorithm, Fig. 4, in which the first traffic pattern is employed. As it reveals, the IIDWBA algorithm has almost zero negative affect on the performance of the ONUs associated with w_2, w_3,w_4. It is because the allocated bandwidth from w_2, w_3,w_4 channels is the unutilised local excess bandwidth. The average extra requested bandwidth from the ONUs associated with w_1 is presented in Fig. 5, with and without employing the IIDWBA algorithm in which the first traffic pattern is utilised. As it reveals, the IIDWBA algorithm is capable of keeping the average extra requested bits from the ONUs associated with w_1 under the minimum guaranteed bandwidth by allocating the available excess bandwidth from the other channels (w_2, w_3,w_4) to them during each service cycle. In the second traffic pattern, the load on all channels is gradually increased by raising the total number of the SSs connected per BS from 10 to 100 to distinguish how the increased load on three channel (w_2, w_3,w_4) will affect the performance of the ONUs over w_1.

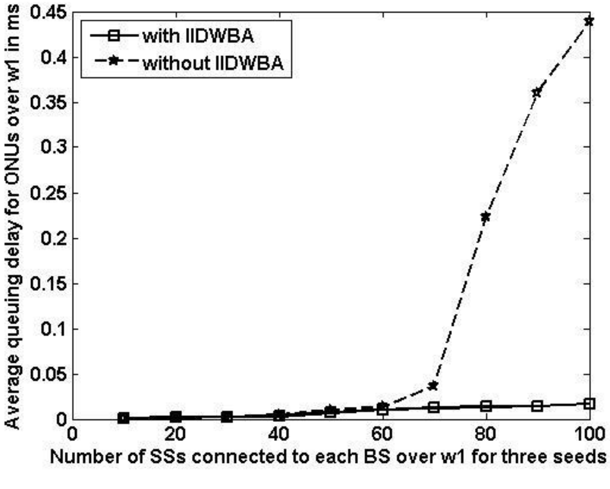


Figure. 3 Average queuing delay for the ONUs on channel one, employing first traffic pattern

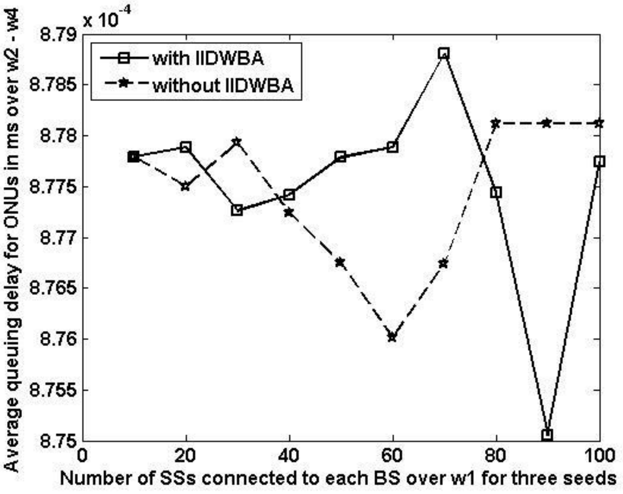


Figure. 4 Average queuing delay for the ONUs associated with channel two to four, employing first traffic pattern

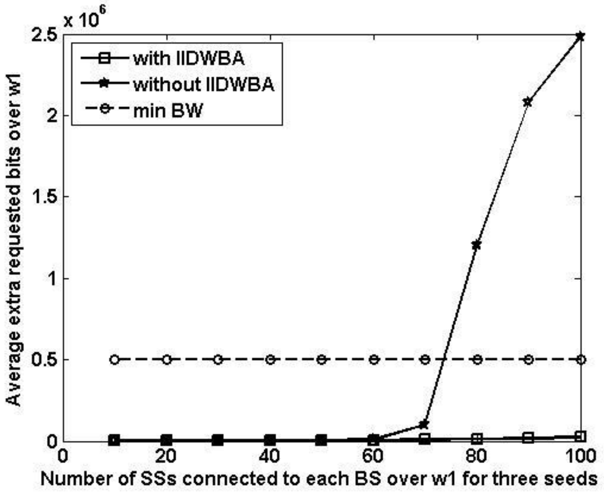


Figure. 5 Average extra requested bandwidth from the ONUs associated with channel one, employing first traffic pattern

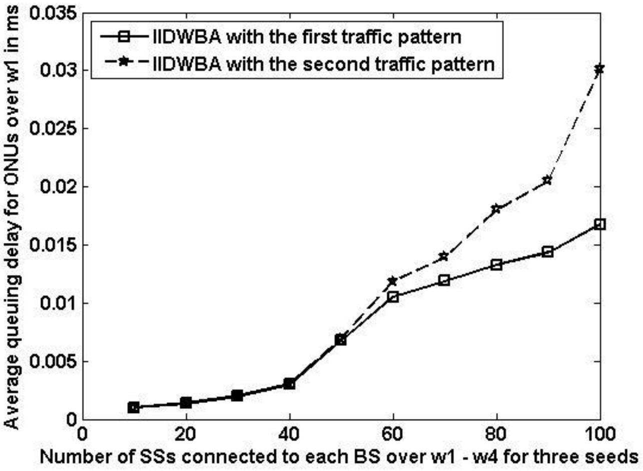


Figure. 6 Average queuing delay for the ONUs associated with channel one employing the first and second traffic patterns

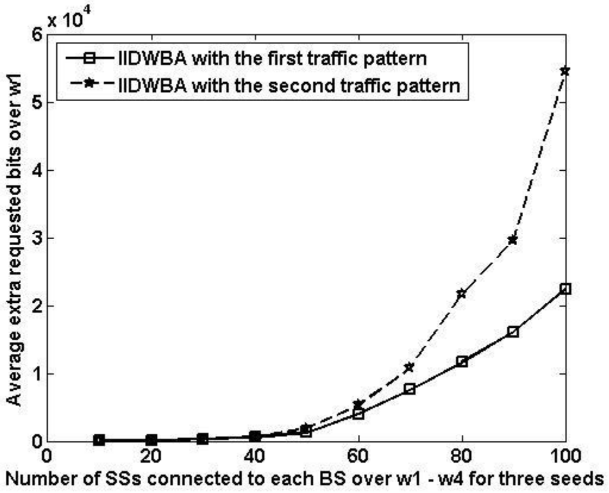


Figure. 7 Average extra requested bandwidth from the ONUs associated with channel one employing the first and second traffic pattern

When the number of the SSs connected per BS on all the channels reaches to 60, Fig. 6, the average queuing delay inside the ONUs associated with w_1 starts increasing constantly and reaches to almost 0.03 sec when it gets to 100 SSs per BS. The reason behind this degradation is that when the load on three channels (w_2, w_3,w_4) is increased gradually, the IIDWBA algorithm cannot find as much excess bandwidth for the ONUs associated with w_1 when it is compared to the first traffic pattern, where the loads on three channels is almost fixed. However, as it reveals in Fig. 6, the queuing delay for the ONUs over w_1, under the second traffic pattern, is still much lower than the scenario when the IIDWBA is not employed.

The average extra requested bandwidth (bits) over w_1 is also displayed in Fig. 7, in which the second traffic pattern is employed, and then compared with the captured results from the first traffic pattern. As it reveals, when the number of the SSs connected per BS is gradually increased from 50 to 100 SSs on all channels, the average extra requested bits over w_1 starts increasing to almost twice more than the first traffic pattern scenario. It is because, when the second traffic pattern is employed, the load on all the channels starts building up gradually. Therefore the IIDWBA algorithm cannot find as much as excess bandwidth for the ONUs associated with w_1 over w_2, w_3,w_4, when compared to the first traffic pattern. This behaviour results in accumulating more packets in the queues associated with the ONUs over w_1. Thus longer queuing delay and larger average extra requested bits will be produced.

VI. CONCLUSION

In this paper, the IIDWBA is proposed for the multi-channel PON integration with wireless technologies where the extra bandwidth from all the available channels associated with a given SST will be identified, collected, scheduled and allocated to the heavily loaded CSs, which may be scattered over the different channels in the same service cycle. Through conducted simulation experiments, it is demonstrated that the proposed algorithm is capable of collecting the excess bandwidth from the lightly loaded channels and spreading them across the heavily loaded CSs, which may be scattered over different channels. The proposed algorithm shows a better performance in terms of average queuing delays, utilisations and throughput when compared with the same simulation scenario not employing the IIDWBA algorithm.

VII. Acknowledgment

The authors would wish to acknowledge the support of the University of Ulster and University of Abertay Dundee for funding this work.

References

[1] B. Jung, J.Y. Choi, Y. T. Han, M. G. Kim, and M. Kang, “Centralized Scheduling Mechanism for Enhanced End-to-End Delay and QoS Support in Integrated Architecture of EPON and WiMAX,” Journal of Lightwave Technology, vol. 28, no. 16, August. 2010, pp. 2277-2288.

[2] M. P. McGarry, M. Reisslein, and M. Maier, “WDM Ethernet Passive Optical Networks,” IEEE Optical Communications, vo1.44, no. 2, February. 2006, pp. 15-22.

[3] S. Ou, K. Yang, and H. H. Chen, “Integrated Dynamic Bandwidth Allocation in Converged Passive Optical Networks and IEEE 802.16 Networks,” IEEE Systems Journal, vol. 4, no. 4, December. 2010, pp. 467- 476.

[4] I. S. Hwang, J. Y. Lee, C. W. Huang, and Z. D. S, “Advanced Dynamic Bandwidth Allocation and Scheduling Scheme for the Integrated Architecture of EPON and WiMAX,” Tenth International Conference on Mobile Data Management Systems Services and Middleware, 2009, pp. 655-660.

[5] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

[6] A. Banerjee, Y. Park, F. Clarke, H. Song, S. Yang, G. Kramer, K. Kim, and B. Mukherjee, “Wavelength-division multiplexed passive optical network (WDM-PON) technologies for broadband access—A review [Invited],” OSA Journal of Optical Networking - Special Issue Optical Access Networks, vol. 4, no. 11, November. 2005, pp. 737–758.

[7] G. Shen, R. Tucker, and C. J. Chae, “Fixed mobile convergence architectures for broadband access: Integration of EPON and WiMAX,” IEEE Communications Magazine, vol. 45, no. 8, 2007, pp. 44–50.

[8] A. R. Dhaini and P. H. Ho, “MC-FiWiBAN an Emergency-Aware Mission-Critical Fiber-Wireless Broadband Access Network,” IEEE Communications Magazine, January. 2011, pp. 134-142.

[9] OPNET Modeler 16.0, available at: www.opnet.com [retrieved: June, 2013]

[10] C. Eklund, R. B. Marks, K. L. Stanwood, and S. Wang, “IEEE Standard 802.16: A Technical Overview of the Wireless MAN Air Interface for Broadband Wireless Access,” IEEE Communications Magazine, vol. 40, no.6, 2002, pp. 98-107.

[11] ITU-T G.9S4.x (GPON), at: http://www.itu.int/rec/T-REC-G/e [retrieved: June, 2013]

[12] ITU-T G.983.1 (BPON), at: http://www.itu.int/rec/T-REC-G.983.1-200501-I/en [retrieved: June, 2013]

[13] A. Ahmed, and A. Shami, “A New Bandwidth Allocation Algorithm for EPON-WiMAX Hybrid Access Networks,” IEEE Global Communication Conference (Globecom), December. 2010, pp. 1-6.

[14] G. Kramer, B. Mukherjee, S. Dixit, Y. Ye, and R. Hirth, “On Supporting Differentiated Classes of Service in Ethernet Passive Optical Networks,” Journal of Optical Networking, vol. 1, no. 8, 2002, pp. 280–298.

[15] T. Tang, G. Shou, Y. Hu, and Zh. Guo “Performance Analysis of Bandwidth Allocation of Convergence of WiMAX and EPON,” International Conference on Networks Security, Wireless Communications and Trusted Computing, April. 2009, pp. 662-665.

[16] Y. Luo, T. Wang, S. Weinstein, M. Cvijetic, and S. Nakamura, “Integrating Optical and Wireless Services in the Access Network,” Optical Fibre Communication (OFC) Conference, 2006.

[17] S. Sarkar, S. Dixit, and B. Mukherjee, “Hybrid Wireless-Optical Broadband-Access Network (WOBAN) A Review of Relevant Challenges,” Journal of Lightwave Technology, vol. 25, no. 11, November. 2007, pp. 3329-3340.

[18] K. Yang, S. Ou, K. Guild, and H. H. Chen, “Convergence of Ethernet PON and IEEE 802.16 Broadband Access Networks and its QoS-aware Dynamic Bandwidth Allocation Scheme,” IEEE Journal on Selected Areas in Communications (JSAC), vol. 27, no. 2, February. 2009, pp. 101–116.

[19] Z. Jia, J. Yu, A. Chowdhury, G. Ellinas, and G K. Chang, “Simultaneous Generation of Independent Wired and Wireless Services Using a Single Modulator in Millimetre-Wave-Band Radio-Over-Fibre Systems,” IEEE Photonics Technology Letters, vol. 19, no. 20, October. 2007, pp. 1691-1693.

[20] N. Moradpoor, G. Parr, S. McClean, B. Scotney, and G. Owusu, “Hybrid Optical and Wireless Technology Integrations for Next Generation Broadband Access Networks,” 6th IFIP/IEEE International Workshop on Broadband Convergence Networks, BCN 2011.

[21] D. J. Shin, D. K. Jung, H. S. Shin, J. W. Kwon, S. Hwang, Y. Oh, and C. Shim, “Hybrid WDM/TDM-PON with Wavelength-Selection-Free Transmitters,” Journal of Lightwave Technology, vol. 23, no. 1, January. 2005, pp. 187-194.

[22] C. Ranaweera, E. Wong, C. Lim, and A. Nirmalathas, “Quality of Service Assurance in EPON-WiMAX Converged Network,” International Topical Meeting on Microwave Photonics Conference, 2011, pp. 369-372.

[23] G. Kramer, B. Mukherjee, and G. Perawnto, “Ethernet PON (ePON): Design and Analysis of an Optical Access Network,” Photonic Network Communications, vol. 3, no. 3, July. 2001, pp. 307–19.

[24] W. T. Shaw, S. W. Wong, N. Cheng, K. Balasubramanian, X. Zhu, M. Maier, and L. G. Kazovsky, “Hybrid Architecture and Integrated Routing in a Scalable Optical–Wireless Access Network,” Journal of Lightwave Technology, vol. 25, no. 11, November. 2007, pp. 3443-3451.

[25] K. Yang, S. Ou, K. Guild, and H. H. Chen, “Convergence of Ethernet PON and IEEE 802.16 Broadband Access Networks and its QoS-aware Dynamic Bandwidth Allocation Scheme,” IEEE Journal on Selected Areas in Communications (JSAC), vol. 27, no. 2, February. 2009, pp. 101–116.

[26] G. Kramer, “Ethernet Passive Optical Networks,” New York, McGraw- Hill, 2005.

[27] IEEE 802.3ah (EPON)at: http://www.ieee802.org/3/ah/index.html [retrieved: June, 2013]

IPv6 Hash-Based Addresses for Simple Network Deployment

Renzo Davoli
Computer Science and Engineering Department
University of Bologna
Bologna, Italy
Email: renzo@cs.unibo.it

***Abstract*—The configuration of an IPv6 network is a rather daunting and error-prone procedure for system administrators. Each node must be provided with its own (128 bit long) IPv6 address and with a domain name manageable by human beings. Autoconfiguration methods can give addresses to interfaces but do not provide any means of configuring the DNS. This paper presents some methods based on hash functions which highly simplify the network configuration process. System administrators just need to define the fully qualified domain names of all the networking nodes (servers and clients) and the networking prefixes for each LAN or subnet. Each node will acquire its own IPv6 address and the DNS server will be automatically configured to support name resolution for all the nodes. The whole process does not require system administrators to type any IPv6 addresses, and it is fully compatible with existing protocols for autoconfiguration and name resolution.**

***Index Terms*—IP networks; TCPIP; Domain Name System; Next generation networking;**

I. Introduction

IPv6 standard provides two different ways for auto-configuration - stateless and stateful. In stateless auto-configuration [1] each networking node broadcasts a router solicitation request to obtain a list of address prefixes active on that local area network. The node then self-assigns an IPv6 address by combining each prefix it receives with the EUI-64 interface hardware address. Stateless auto-configuration provides the node with valid addresses and routing information but it does not configure the DNS (Domain Name Service) server address, nor does it configure its own DNS entries.

As described in [2] and [3], this kind of auto-configuration can cause privacy problems, since the hardware address, usually the Ethernet MAC (Media Access Control) address, is part of the final IPv6 address. So, it is possible to track the movements of personal computers, laptops, tablets, smart-phones etc.. The more these devices are personal *digital extensions*, the more this permits the tracking of people, their physical locations and habits. Stateless auto-configuration changes the IPv6 address of a node in cases of network adapter substitution. On the other hand, stateful configuration [4] is based on DHCPv6 (Dynamic Host Configuration Protocol). In stateful configuration, a node broadcasts a DHCP request using its own link-local IPv6 address. The server then replies, providing the node with its own global IPv6 address (or addresses).

Stateful auto-configuration is not self-configuration. In fact, whilst no configuration effort is required from the client, the mapping between hardware and IP addresses must be configured at the server, by hand.

Another problem related to address configuration is DNS mapping. Forward and reverse DNS mapping is compulsory for servers, but it is useful to give symbolic names to clients, too. Numeric addresses, especially 128 bit IPv6 addresses, are hard to use: symbolic names facilitate the management, e.g., the tracking of networking problems.

Both DHCP and DNS configuration involve the typing of several IPv6 addresses: sequences of 32 hexadecimal numbers. Naturally this is highly prone to error.

The results presented in this paper introduce a set of methods that can help the system and network administrators to set up an IPv6 network in a simple and effective way. These methods provide the networking nodes of a local area network with their addresses, given an IPv6 prefix, a domain name of the LAN (Local Area Network) and the list of host names to be configured. Each node just needs to know its own Fully Qualified Domain Name (FQDN). The only hand typed IPv6 address is the prefix of the LAN. Those methods also provide forward name resolution and, if required, reverse name resolution.

This paper is organized as follows. The next section introduces the idea of hash-based address by presenting some implementation scenarios. Section III discusses the limits of the proposed approach and section IV analyzes the cases of address collision. The final part of the paper include sections about a proof-of-concept implementation, a comparison with the related work available in the literature, and final remarks, also discussing the future developments of this project.

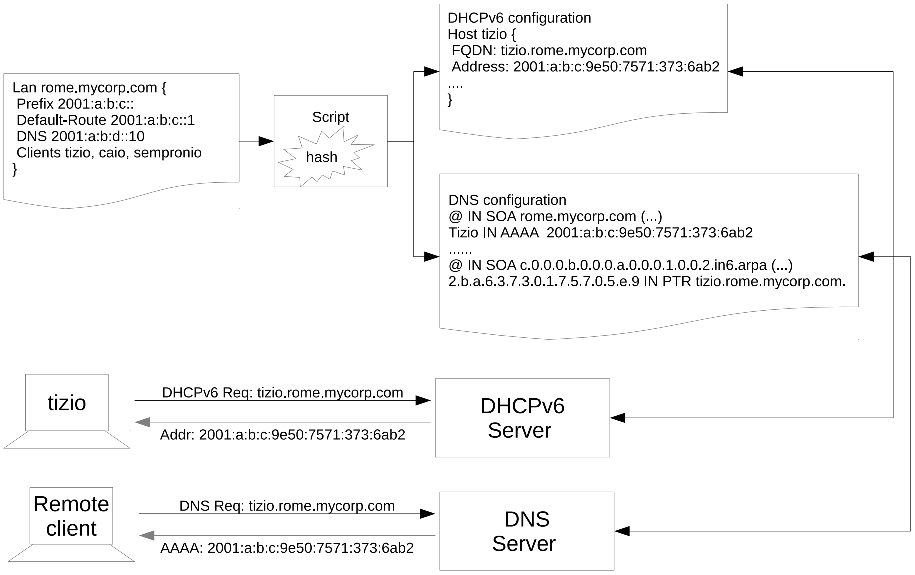
II. Hash-based Addresses

The core idea is to compute a 64 bit encoding of the FQDN of each node and to use it as its host address, following the 64 bit prefix fixed for each specific LAN. This idea can be implemented in many ways. Each one has pros and cons.

A. Assisted DHCP and DNS management.

Given the list of nodes, a script computes both DHCP and DNS tables by generating a hash-based address for each FQDN (see Fig. 1). The router must be configured for stateful

ISBN: 978-1-61208-300-1


```
Lan rome.mycorp.com {
 Prefix 2001:a:b:c::
 Default-Route 2001:a:b:c::1
 DNS 2001:a:b:d::10
 Clients tizio, caio, sempronio
}
```

```
DHCPv6 configuration
Host tizio {
 FQDN: tizio.rome.mycorp.com
 Address: 2001:a:b:c:9e50:7571:373:6ab2
....
}
```

```
DNS configuration
@ IN SOA rome.mycorp.com (...)
Tizio IN AAAA  2001:a:b:c:9e50:7571:373:6ab2
......
@ IN SOA c.0.0.0.b.0.0.0.a.0.0.0.1.0.0.2.in6.arpa (...)
2.b.a.6.3.7.3.0.1.7.5.7.0.5.e.9 IN PTR tizio.rome.mycorp.com.
```

Script
hash
tizio
DHCPv6 Req: tizio.rome.mycorp.com
Addr: 2001:a:b:c:9e50:7571:373:6ab2
DHCPv6 Server
Remote client
DNS Req: tizio.rome.mycorp.com
AAAA: 2001:a:b:c:9e50:7571:373:6ab2
DNS Server

Fig. 1. Assisted DHCP and DNS management.

auto-configuration, both DHCP servers and clients must use the fully qualified host names as client identifiers. When a node starts, it learns from a router advertisement packet that stateful auto-configuration is required. Then it sends a DHCPv6 request using its FQDN as its identifier. The DHCP server assigns to each node its specific address. This method uses standard DHCP and DNS servers and clients. Node addresses are not related to hardware addresses, so they do not change in case of network adapter substitution. DHCP and DNS tables must be recomputed if a name gets modified or a new node is added. Given an IPv6 prefix, a LAN domain name and the list of host names to configure, this method provides the networking nodes of a local area network with their addresses. The use of hash-based addresses should be preferred to other assignment rules, for instance, by a script which enumerates the hosts, as there is no need for system administrators to store the mapping between each host name and its address. The address can be retrieved by running the hash function when needed. The management of the accountancy for unused and reassigned host addresses is also unnecessary.

B. Hash address self-assignment and automatic DNS forward resolution.

This method does not even require a list of nodes. Each node autonomously compute its own address by combining the prefix learned from the router advertisement and the hash of its FQDN (see Fig. 2). A custom DNS server for the subnet provides a forward resolution for any name within the domain of the local area network. This method simplifies the deployment of local area networks composed of client computers. Provided no nodes share the same FQDN and there are no hash collisions, the IPv6 address assignment and the DNS forward resolution require no configuration. Name collisions are very rare events, as shown in section IV. This solution requires a specific DNS server and a program to self-assign the address.

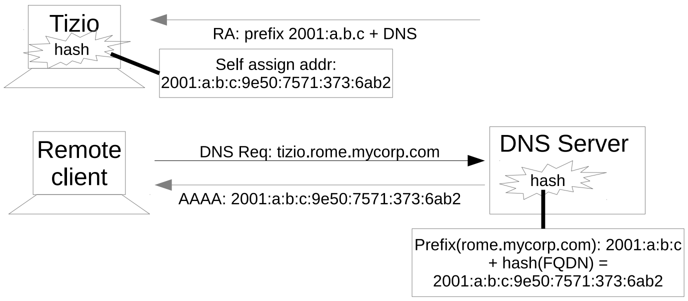

Tizio
hash
RA: prefix 2001:a.b.c + DNS
Self assign addr:
2001:a:b:c:9e50:7571:373:6ab2
Remote client
DNS Req: tizio.rome.mycorp.com
AAAA: 2001:a:b:c:9e50:7571:373:6ab2
DNS Server
hash
Prefix(rome.mycorp.com): 2001:a:b:c
+ hash(FQDN) =
2001:a:b:c:9e50:7571:373:6ab2

Fig. 2. Hash address self-assignment.

C. Hash address self assignment plus DNS forward and reverse resolutions

By itself, the DNS can obtain the address of any host in the subnet by computing the hash of the FQDN, but it is not possible to perform the reverse resolution automatically. The hash function is not injective. In any case, it is possible for the DNS to cache the mapping between names and their addresses so as to be able to give correct answers to reverse DNS resolution, requests. The DNS should not store any DNS resolutions as it may refer to non-existent nodes: an attack based on DNS requests for large numbers of random host names could cause table overflows and delays. The simplest solution is to store the reverse mapping of a host only if the request comes from the address which is being resolved. This means that a host can request its own reverse resolution by introducing itself to the DNS server: it sends a DNS forward resolution of its own FQDN. The DNS server gets the request,

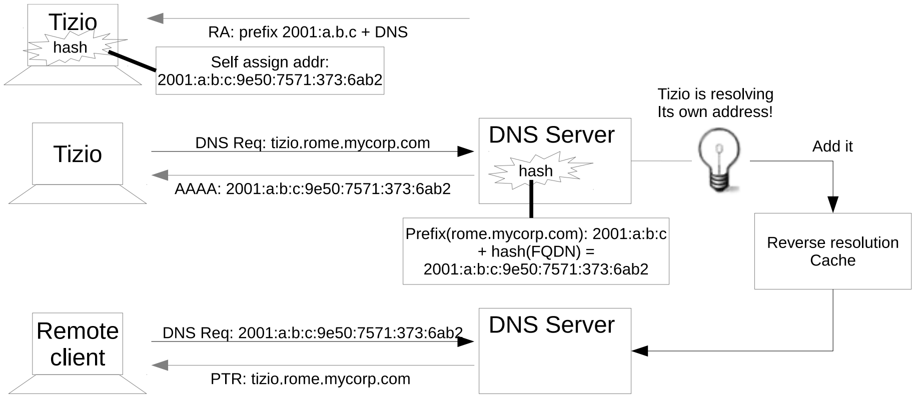

Tizio
hash
RA: prefix 2001:a.b.c + DNS
Self assign addr:
2001:a:b:c:9e50:7571:373:6ab2
Tizio
DNS Req: tizio.rome.mycorp.com
AAAA: 2001:a:b:c:9e50:7571:373:6ab2
DNS Server
hash
Tizio is resolving
Its own address!
Add it
Prefix(rome.mycorp.com): 2001:a:b:c
+ hash(FQDN) =
2001:a:b:c:9e50:7571:373:6ab2
Reverse resolution
Cache
Remote
client
DNS Req: 2001:a:b:c:9e50:7571:373:6ab2
PTR: tizio.rome.mycorp.com
DNS Server

Fig. 3. Hash address self-assignment supporting reverse DNS mapping

uses the hash mapping and recognizes that the request comes from the same address, so it saves the mapping for the reverse resolution (Fig. 3). In this way, it is not possible for external DNS clients to add fake reverse mapping items. However, some assumptions are needed to grant this security property: the internal nodes and the routing must be trusted. Otherwise, misbehaving internal nodes, or nodes able to forge source addresses of the local network, can attack the DNS and add fake entries.

D. Further methods

The methods described in II-B and II-C have been designed for an environment where each LAN is assigned its specific prefix and all the nodes on the LAN belong to the same domain. Although this is quite a common situation, hash-based addresses can also support different and more general scenarios.

- Several Domains, one prefix. The hosts of several domains use the same 64 bit prefix for their IP addresses. This scenario is already supported using the methods described in II-B and II-C: the hash-based address is computed on the FQDN, which is different, so the addresses will not interfere with each other (except for the address collision problem, see IV).
- One or Several Domains, Several Prefixes. Hosts of several domains are working on the same LAN but each domain has its own prefix, routing rules, etc. While the previous methods are still valid for the node resolution by remote nodes, it is not possible for nodes to self-assign their addresses. In this situation each node receives advertisements from the router/routers for all the prefixes and cannot autonomously choose the right one for its domain. In this case, the hash-based assignment can be implemented in the DHCPv6 server (see Fig.4). Each host is able to acquire its address using a standard stateful configuration interaction.

III. Limits

The hierarchical structures of fully qualified names and IP addresses are not necessarily related. Symbolic names are useful for humans to reach servers and services, whilst IP addresses are machine oriented representations used by routers to dispatch packets towards the right networking nodes. The methods introduced in this paper require direct mapping between a DNS domain and a prefix. In real applications, this is quite usual for small business firms, client nodes in computer labs, or office LANs. A company may have several networks, and then it would need to use several (sub) domains. e.g., the company `famouscorp.com` may have several IPv6 sub-nets. Hash configuration requires the subnets to be named as sub-domains, e.g., `lab.famouscorp.com`, `adm.famouscorp.com`, `londonoffice.famouscorp.com`, etc. In general, there is no problem assigning FQDN to clients, as client names are for internal use, for system and network administrators, so the more specific they are, the clearer they are. On the other hand, institutional servers (e.g., `www.famouscorp.com`) should not refer to the internal sub-netting structure. It is always possible to assign names to externally visible servers using DNS CNAME entries, e.g., define `www.famouscorp.com` as a CNAME of `www.londonoffice.famouscorp.com`. CNAME entries can provide network administrators with a means of defining all the symbolic names needed. Using the hash address assignment, the administrator just writes symbolic names. Thus, no IPv6 address typing is required, reducing the probability of configuration errors.

IV. Addresses' Birthdays

It is possible that multiple FQDNs generate the same hash code, so that they collide, pretending to use the same IPv6 address. The discussion is focussed on one network where all the hosts share the same networking prefix, as no collision can take place if the prefixes are different. This section shows that there is a very low probability of this event, so it can safely be ignored. Should it happen, it is possible to avoid it by changing one of the colliding names. If we consider that the hash keys follow a uniform distribution, the probability of collision becomes an instance of the Birthday Problem (also known as Birthday Paradox, see page 507 in [5]). The probability of

ISBN: 978-1-61208-300-1

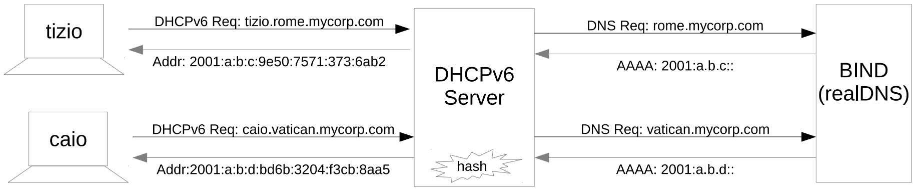

tizio
DHCPv6 Req: tizio.rome.mycorp.com
Addr: 2001:a:b:c:9e50:7571:373:6ab2
caio
DHCPv6 Req: caio.vatican.mycorp.com
Addr:2001:a:b:d:bd6b:3204:f3cb:8aa5
DHCPv6 Server
hash
DNS Req: rome.mycorp.com
AAAA: 2001:a.b.c::
DNS Req: vatican.mycorp.com
AAAA: 2001:a.b.d::
BIND (realDNS)

Fig. 4. Hash address assignment using a modified DHCPv6 server for multiple-prefixes LAN

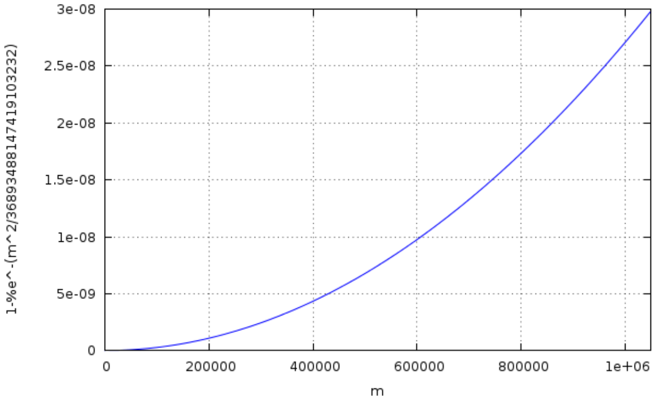


Fig. 5. Probability of address collision in a 64 bits hash

m nodes choosing the same random key, amongst n possible keys, is

$$Pr[(n,m)] = 1 - \frac{n!\binom{m}{n}}{m^n} \quad (1)$$

Using a Taylor series approximation (as $m/n << 1$) the probability can be estimated as follows:

$$Pr[(n,m)] \approx 1 - e^{-\frac{m^2}{2n}} \quad (2)$$

Figure 5 shows the probability function (2) plotted for up to 1Mi (i.e., 2^{20}) computers. The probability of two addresses colliding is less than one in 30 million for a LAN connecting more than 1 million hosts. In more realistic cases, network connecting less than one thousand nodes, the probability has the order of magnitude of one in $3 \cdot 10^{14}$.

Anyway, it is possible to implement methods in order to warn system administrators if such a collision should occur. Using the technique described in II-A, the address assignment script can take care of the collision detection. For II-C the updating function of the address cache for reverse resolution can reveal the collision problem and start some warning procedure.

V. Proof-of-concept implementation

Some software tools have been implemented to test the effectiveness of the approaches proposed in this paper. A first tool is a command to compute the address of a host. This command, named `hashaddr`, uses the MD5 algorithm [6] to compute the hash of the FQDN.

`hashaddr` takes two arguments: the first is a prefix (or a base address, as explained in the following) P and the second is the FQDN.

The final address is computed as follows:

$$hash = md5hash(FQDN) \quad (3)$$

$$address = P \oplus hi64bits(hash) \oplus lo64bits(hash) \quad (4)$$

where the function $hi64bits$ returns the 64 most significant bits, and $lo64bits$ returns the 64 least significant bits.

For example, a test running of `hashaddr` together with its output follows:

```
$ ./hashaddr 2001:a:b:c::  tizio.rome.mycorp.com
2001:a:b:c:9e50:7571:373:6ab2
```

The MD5 hash of `tizio.rome.mycorp.com` is `a2ea0c2518c2756b3cba79541bb11fd9`. The final address is computed as follows (the IPv6 encoding of 128 bit numbers is used for clarity):

$$\begin{aligned} addr &= 2001:a:b:c:: \oplus :: a2ea:0c25:18c2:756b \\ &\quad \oplus :: 3cba:7954:1bb1:1fd9 \end{aligned}$$

`hashaddr` can be used in the scripts to compute the tables for the DHCPv6 and DNS servers as presented in II-A. These scripts must be run each time there is a change in the set of managed hosts to update the tables. No modifications are needed for DHCPv6 and DNS servers, because they simply use automatically generated tables instead of manually inserted data. Clearly the performance of the stateful auto-configuration and name resolution operations is not affected by the usage of hash computed addresses.

`hashaddr` uses `getaddrinfo` [7] to get the prefix (or base address), so it can be expressed in a symbolic way (using DNS resolution). For example, if the base address `2001:a:b:c::` for rome.mycorp.com has been manually added to the DNS table, this is the result:

```
$ ./hashaddr rome.mycorp.com  tizio.rome.mycorp.com
2001:a:b:c:9e50:7571:373:6ab2
```

In the examples, we have used a prefix as the second argument, i.e., the 64 least significant bits of the address were zeroes. When a general address is used instead, the 64 low order bits and the hash of the FQDN are combined by a XOR operation. In this case we name this *base address*, as it allows us to provide a specific mapping.

TABLE I
PROXY IMPLEMENTATION PERFORMANCE

Test #1: 100 DNS lookups

	average	standard deviation
DNS(no proxy)	35.3	5.22
Proxy	75.6	11.27

Test #2: `scp` of 1k file

	average	standard deviation
DNS(no proxy)	156,6	4.72
Proxy	158.9	5.89

This idea can be used to implement an extension of the method to preserve some address privacy. IPv6 architecture and philosophy does not include the idea of masqueraded addresses for clients (unique local addresses [8] are site-local and not routed to the Internet). If the hash-based address resolution of local clients and DNS resolution for base addresses are allowed only for local clients, it is harder for a remote attacker to guess the IP addresses. He cannot just compute the hash of some known FQDN. The knowledge of the entire base address is needed to succeed.

The proof-of-concept implementation for the method described in II-B, hash address self-assignment and automatic DNS forward resolution, uses a specific DNS proxy program: `hashdns` (see Fig. 6). This program takes three or more arguments. The first argument is the IP address the proxy DNS server will use; the second is the address of a real existing DNS, where the requests will be forwarded; the third and following arguments are a list of domains managed by the proxy.

```
./hashdns 2001:a:b:c::2 2001:a:b:c::3 rome.mycorp.com
```

All the DNS requests received by the proxy (at the address `2001:a:b:c::2`) get forwarded to the DNS server specified by the second argument (`2001:a:b:c::3`) except those asking for an AAAA record of a host of one of the managed domains. In this case (e.g., `tizio.rome.mycorp.com` in the example above) the proxy forwards the request for the domain (`rome.mycorp.com`) by stripping away the preceding part of the FQDN. When the real DNS server returns the address corresponding to the domain, `hashdns` performs the same process described for `hashaddr` to compute the hash-based address, using the address received by the real DNS server as its base address. The result is returned by the proxy as its reply to the original request.

Table I shows some of the proxy performance figures. The test environment is consists of a dual core 2Ghz CoreDuo2 processor, 3GiB RAM. The proxy is running on the same machine using the user mode stack LWIPv6 [9]. The first test evaluates the time needed for name resolution using a real DNS server and the proxy supporting hash-based addresses. The time in the table is in milliseconds for 100 calls of `getaddrinfo`. Although the overheads are high, the proxy takes about twice the time of the DNS, the second test shows that it is not appreciable for a normal networking usage. In fact, the second test measures the time needed to transfer one 1KB file by `scp` from the same machine: the address returned by the DNS or proxy resolution is one of the addresses of the same computer. If this is not the worst case for an `scp` transfer, it is an unfavorable situation: the transfer of a small file on a very fast line. The cost of name resolution would impact less on large files or slow lines. The computed overhead is less than 2%.

VI. RELATED WORK

Hash-based addresses are very useful when applied to the Internet of Threads. This new idea, presented in [10], changes the concept of communication endpoints in the Internet: each process can be provided with its own IPv6 address and can be a node of the network.

This change of perpective permits the development of many new services and a higher level of flexibility but, at the same time, increases the number of IP addresses to manage.

Cryptographically Generated Address (CGA) protocol, defined in [11], uses a one-way hash function to define the least-significant 64 bits of the IPv6 address. The meaning and the purpose of the operation is completely different from those defined in this paper: CGA computes a hash function on the public key of the sender to enforce a secure communication; hash-based address is computed on the FQDN of the host to ease the network management.

IPv6 privacy extension [2] generates the least-significant 64 bits of the IPv6 address in a random manner to preserve the privacy of the hardware controller MAC address. A new random address is generated if an address collision is detected. This method is for clients and does not provide methods to update a DNS server map.

VII. CONCLUSION AND FUTURE WORK

This paper has explained how to use hash-based IPv6 addresses to simplify the network deployment by system administrators. The proposed methods open a new perspective on IPv6 network configuration. The use of the MD5 algorithm as well as the proxy based implementation are just examples to support the effectiveness of these methods. It is possible to envision several further developments for this project. A non-exhaustive list of ideas by the author at the time of writing this paper follows:

- Per site defined hash function: each company, institution, etc., can design its own hash function. All the methods work provided that the same hash function is shared between the node, for address self assignment, and the DNS resolution.
- Native support of hash-based addresses in DNS servers. This enhancement would provide better performance to the name resolution process.
- DHCPv6 proxy for multiple domain support. An implementation of the method proposed in section II-D is still missing.
- It is possible to support the co-existence of static IPv6 addresses and hash-based ones. For example, the DNS proxy can forward the request for the entire FQDN to

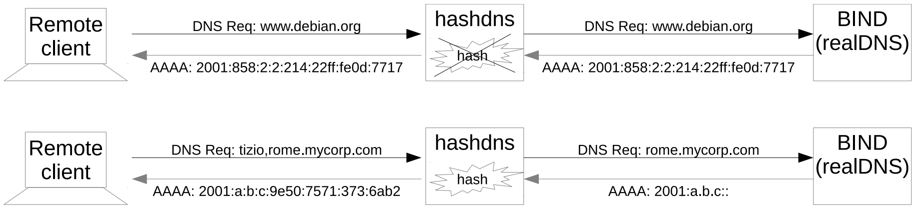


Fig. 6. hashaddr: DNS proxy proof-of-concept implementation

the real DNS server and perform the hash-based name resolution only when the FQDN resolution fails.

- It is common for companies to have different DNS servers to provide different views of their networking structure for local and external users. For example, customers should be able to access the company's web server, but it is useless (and potentially dangerous) to provide the name resolution of the accountancy office personal computer. An extension for the DNS proxy can provide filters to decide the visibility boundary of the hash-based address resolution for each domain, i.e., which source IP addresses are allowed to know the resolution for the hosts in each domain.

The source code to test the experiments presented in this paper can be downloaded from `svn://svn.code.sf.net/p/view-os/code/branches/hashaddrtest` and has been released under the GNU General Public License (GPL) v. 2 or newer. The programs are intended as just a proof-of-concept to show the effectiveness of the ideas introduced here.

Acknowledgment

I would like to thank the anonymous reviewers for their valuable suggestions and their comments.

References

[1] S. Thomson, T. Narten, and T. Jinmei, "IPv6 Stateless Address Autoconfiguration," RFC 4862 (Draft Standard), Internet Engineering Task Force, Sep. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4862.txt

[2] T. Narten, R. Draves, and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6," RFC 4941 (Draft Standard), Internet Engineering Task Force, Sep. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4941.txt (Retrieved: June 15, 2013)

[3] M. Tortonesi and R. Davoli, "User untraceability in next-generation internet: a proposal," in *Proceeding of Communication and Computer Networks 2002 (CCN 2002)*, IASTED, Ed., November 2002, pp. 177 – 182.

[4] R. Droms, J. Bound, B. Volz, T. Lemon, C. Perkins, and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)," RFC 3315 (Proposed Standard), Internet Engineering Task Force, Jul. 2003, updated by RFCs 4361, 5494, 6221, 6422, 6644. [Online]. Available: http://www.ietf.org/rfc/rfc3315.txt (Retrieved: June 15, 2013)

[5] D. E. Knuth, *The art of computer programming, volume 3: sorting and searching*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1973.

[6] R. Rivest, "The MD5 Message-Digest Algorithm," RFC 1321 (Informational), Internet Engineering Task Force, Apr. 1992, updated by RFC 6151. [Online]. Available: http://www.ietf.org/rfc/rfc1321.txt (Retrieved: June 15, 2013)

[7] R. Gilligan, S. Thomson, J. Bound, J. McCann, and W. Stevens, "Basic Socket Interface Extensions for IPv6," RFC 3493 (Informational), Internet Engineering Task Force, Feb. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3493.txt (Retrieved: June 15, 2013)

[8] R. Hinden and B. Haberman, "Unique Local IPv6 Unicast Addresses," RFC 4193 (Proposed Standard), Internet Engineering Task Force, Oct. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4193.txt (Retrieved: June 15, 2013)

[9] R. Davoli and M.Goldweber, "Virtual square: Users, programmers and developers guide," 2011. [Online]. Available: http://www.cs.unibo.it/ renzo/virtualsquare/ (Retrieved: June 15, 2013)

[10] R. Davoli, "Internet of threads," in *Proc. of the The Eighth International Conference on Internet and Web Applications and Services, ICIW 2013. To appear*, 2013.

[11] T. Aura, "Cryptographically Generated Addresses (CGA)," RFC 3972 (Proposed Standard), Internet Engineering Task Force, Mar. 2005, updated by RFCs 4581, 4982. [Online]. Available: http://www.ietf.org/rfc/rfc3972.txt (Retrieved: June 15, 2013)

Evaluation of Time Dimensional Traffic Engineering with Storage Aware Routing

Shigeyuki Yamashita and Miki Yamamoto
Faculty of Engineering Science
Kansai University
3-3-35 Yamate-cho, Suita-shi, Osaka, Japan
Email: k859805, yama-m@kansai-u.ac.jp

Tomohiko Yagyu
Cloud System Research Labo, NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki,
Kanagawa, Japan
Email: yagyu@cp.jp.nec.com

***Abstract*—Recently, the amount of traffic on the Internet continues to grow by increment of rich content such as video every year. Not only the overall traffic increase but also the time variation becomes large and the difference in the amount of traffic during the time of peak and off peak is very large. Therefore, it is difficult to use the link bandwidth efficiently. In this paper, we propose a new system of content distribution, named Storage Aware Routing (SAR). With SAR, routers can exploit the links with low utilization by using their large storages. Our performance evaluation of SAR by binary integer programming formulation shows that SAR is able to smooth the link utilization.**

Keywords-Storage; Content Distribution; New Generation Network; Traffic Peak Shift.

I. INTRODUCTION

Traffic volume transferred in the Internet has been steadily increasing with continuous growth of video traffic. According to Cisco white paper [1], it increases eight-fold in 5 years and will be approaching 1.3 zeta byte by the end of 2016. Sophisticated network management is indispensable for effective accommodation of increasing traffic demand. Content Delivery Network (CDN) [2] and traffic engineering [3] are promising examples for effective network management. These techniques aim to distribute traffic in space dimension by adequate routing. This means traffic can be managed only in space dimension in these conventional techniques.

Another impact of current Internet traffic than traffic volume is the difference between peak and minimum traffic volume in day-time scale. According to Japanese government report [4], difference between peak and minimum traffic in day-time scale significantly increases recently. In the conventional traffic management technique, bandwidth of the most congested link should be designed to accommodate peak traffic. Average link utilization should be small due to bursty nature of traffic, which means network resources cannot be effectively utilized only with spatial traffic management techniques.

Large storage can be used for time dimensional traffic management. While the network is congested, traffic stored in a network, i.e., at routers, can be transmitted during off-peak time period. Delay Tolerant Network (DTN) [5] is one example of usage of storage for traffic shift. It enables asynchronous forwarding, but its main purpose is providing reliable communications between a wide range of networks having poor and disparate performance, which means DTN is not a technology for traffic management. Some methods that control the timing of transmission with nodes' storage are proposed [6] [7]. Because these methods do not consider deadline of delivered data, they cannot guarantee the arrival of priority data in time. Another method of asynchronous forwarding is also proposed for inter-datacenter bulk transfer, called NetStitcher [8]. In NetStitcher, bulk data transfer is scheduled to adapt to resource fluctuation in diurnal pattern. NetStitcher is applied for world wide datacenter systems and bulk transfer is scheduled so that off-peak in diurnal fluctuation is to be used. Purpose of NetStitcher is very similar to ours, traffic management in time dimension, but it can only be applied to worldwide scale datacenter systems because it adapts to resource fluctuation due to local time difference.

In near future, new usage scenario of video services can be deadline-aware, such as a user reserves her preferable video services with her convenient time. One promising scenario for this deadline-aware video service is that in the morning a user on a commuter train reserves video service with evening time on which she can enjoy it at her home. In current video on demand, e.g., YouTube, peak traffic period is exactly the peak demand time period and may be prime time in the evening. In contrast, in deadline-aware video service, there is time leeway for receiver to receive content (video) file. Another example of deadline-aware delivery is backup of critical data between distant data centers for Business Continuity Planning (BCP). When a router has large storage space and can storage content traffic in the case of the following link is in its peak traffic period, on-peak traffic can be shifted to off-peak period.

In this paper, we propose a novel traffic engineering technique which manages traffic in not only space but also time dimension, SAR. SAR is one of traffic engineering technique and aims at adapting to link utilization fluctuation. In SAR, a router stores content traffic at its storage and time-shifts on-peak traffic to off-peak time period. Storage at a router is assumed to be large enough to enable time-shift of on-peak traffic, i.e., plentifully large storage than current router queue. In this paper, we formulate SAR traffic management as time-space dimension routing problem. With numerical examples binary integer programming problem for time-space dimension model, we comparatively evaluate SAR. Our numerical examples show that SAR can effectively adapt to fluctuation of link utilization. And we investigate condition that SAR can work well and show that SAR generally works well. In this paper, we investigate the feasibility and the fundamental benefits of deadline-aware content delivery with SAR.

This paper is structured as follows. First, in Section 2, we explain our proposed new traffic engineering technique, named SAR, in detail. In Section 3, we evaluate performance of SAR by using binary integer programming. Finally, we conclude a paper in Section 4.

II. STORAGE AWARE ROUTING

A. Design Concept of SAR

In the current Internet, sharp increase of content traffic volume is one of the most important technical issues. Especially, difference between peak and off-peak traffic volume is significantly increasing. When some content traffic on peak period can be shifted to off-peak time period, smoothed content traffic volume will enable effective accommodation of increasing content traffic. So, we propose Storage Aware Routing which enables time-dimensional traffic shift for deadline-aware content distribution.

An end-user is interested only in content itself. She does not care about content retrieval time and content can be retrieved anytime before her preferable time to watch it. SAR tactically utilizes time difference between reserved timing and content watching timing. According to the above mentioned content retrieval nature, content can be retrieved between these two timings. In SAR, each session (end-to-end session for one content retrieval) schedules its hop-by-hop transmission timing so that fluctuation of link utilization along transmission path is minimized. With this smoothed link utilization, network can reserve its resources to sessions generated later.

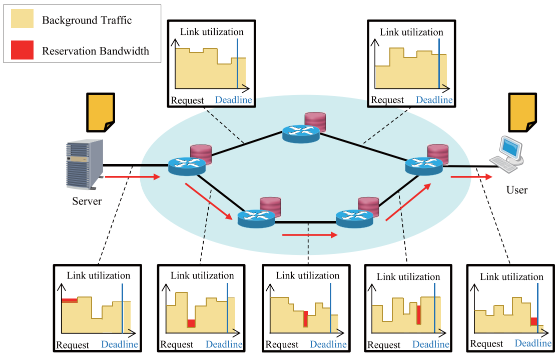


Figure 1. Storage Aware Routing

B. Storage Aware Routing

Fig. 1 shows overview of SAR. Framework of SAR is composed of the following 3 elements.

1) Reservation by end user
2) Scheduling of transmission at the server
3) Routing and scheduling in the network

SAR is time dimensional traffic engineering technique by making use of time difference between reservation and content watching. At an end-user side, reservation for content is a trigger to SAR (reservation by end user). A server and networks schedule transmission timings of a corresponding content file so that fluctuation of link utilization is minimized. A server decides its transmission timing (scheduling of transmission at the server). A network decides a path for content file transmission and schedules hop-by-hop transmission timing (routing and scheduling in the network).

In Fig. 1, SAR chooses route and transmission timing which enables low link utilization. A graph for each link depicts link utilization variance in time. SAR first chooses a route enabling low link utilization. Then SAR schedules hop-by-hop transmission timing. On the first link in Fig. 1, content file is not transferred in the lowest utilization time period. This is because of law of causality for low utilization periods on the following links. When content file is transferred in the lowest utilization time period on the first link, content must be transferred in high utilization time period of the following link so as to reach before deadline. SAR designs not only routing but also transmission timing at each hop on the route, i.e., routing and scheduling.

III. NUMERICAL EXAMPLES

One of the most advanced aspects of SAR is time dimensional scheduling enabled by large storage at a router. In this paper, basic performance of SAR in this aspect is evaluated by numerical examples with binary integer programming formulation.

A. Problem Formulation of SAR

SAR design can be formulated as a routing problem in space-time diagram. Fig. 3 shows space time diagram for the tandem topology model shown in Fig. 2. Horizontal axis shows time dimension and vertical axis shows space dimension. Space dimension corresponds to tandem network shown in Fig. 2. The origin ($[S][T]=$ 00) of this time space diagram shows the content server at time 0. Time is normalized with file transmission time on a link (we simply assume that a content file is transmitted on a link in 1 time unit). In a sample path in Fig. 3, the server transmit content file to router 1 at time 0. Router 1 does not store it and forwards it to router 2 immediately after it receives a whole file at time 1. At routers 2 and 3, content file is stored during one unit time, which is time dimensional scheduling. In this example, deadline of content retrieval is time 8 and a content file is retrieved just at this deadline.

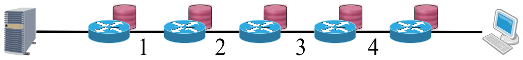


Figure 2. Tandem topology model

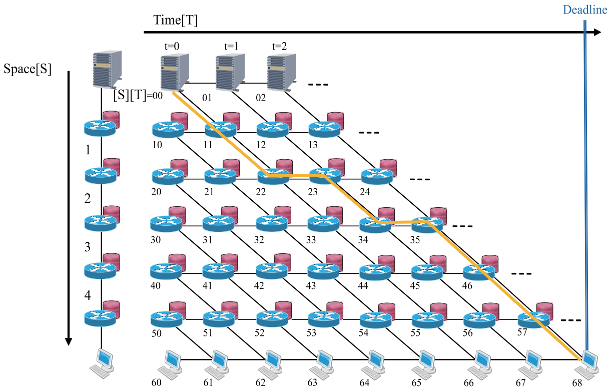


Figure 3. Space-time diagram

SAR design can be formalized as the following minimization of total cost for a "path" in space time diagram. In our design, we set the link cost to the square of utilization to smooth the traffic in the networks. We set the cost of storage to 0 supposing that storage capacity is very large. The impact on other cost functions will be studied in future.

$$minimize \sum_{(i,j)\in E,t\in T} COST_{i,j,t}*X_{i,j,t}+\sum_{i\in V,t\in T} COST_{i,i,t}*X_{i,i,t} \quad (1)$$

$subject \quad to$

$$BT_{i,i,t} + d * X_{i,i,t} \le C_{i,i} \quad (2)$$

$$BT_{i,j,t} + d * X_{i,j,t} \le C_{i,j} \qquad (i \neq j) \quad (3)$$

$$\sum_{j:(i,j)\in E,t\in T} X_{i,j,t} + \sum_{i\in V,t\in T} X_{i,i,t} - \sum_{j:(j,i)\in E,t\in T} X_{j,i,t-\Delta t} - \sum_{i\in V,t\in T} X_{i,i,t-\Delta t} = 0 \qquad (i_t \neq s, r) \quad (4)$$

$$\sum_{j:(s,j)\in E,R\in T} X_{s,j,R} + \sum_{tr\in T} X_{s,s,R} - \sum_{j:(j,s)\in E,R\in T} X_{j,s,R-\Delta t} - \sum_{tr\in T} X_{s,s,R-\Delta t} = 1 \qquad (i_R = s) \quad (5)$$

$$\sum_{j:(r,j)\in E,D\in T} X_{j,r,D-\Delta t} + \sum_{D\in T} X_{r,r,D-\Delta t} - \sum_{j:(j,r)\in E,D\in T} X_{r,j,D} - \sum_{D\in T} X_{r,r,D} = 1 \qquad (i_D = r) \quad (6)$$

$$COST_{i,j,t} = ((BT_{i,j,t} + d)/C_{i,j})^2 \qquad (i \neq j) \quad (7)$$

$$COST_{i,i,t} = 0 \quad (8)$$

TABLE I
NOTATION USED IN THE PAPER

Notation	Meaning
$X_{i,j,t}$	- Binary variable denoting reservation of link (i, j) $(for \quad i \neq j)$ - Binary variable denoting reservation of storage at router i $(for \quad i = j)$ for time period between t and $t + \Delta t$
$COST_{i,j,t}$	- Transmission cost on link (i, j) $(for \quad i \neq j)$ - Storage cost at node i $(for \quad i = j)$ for time period between t and $t + \Delta t$
$BT_{i,j,t}$	- Background traffic volume of link (i, j) $(for \quad i \neq j)$ - Stored background traffic volume at router i $(for \quad i = j)$ for time period between t and $t + \Delta t$
$C_{i,j}$	- Link capacity of link (i, j) $(for \quad i \neq j)$ - Storage capacity of router i $(for \quad i = j)$
d	Traffic demand for each session
E	Set of links
V	Set of nodes
T	Set of discrete time periods (denoting each time unit)
s	Currently requesting sender
r	The receiver at the corresponding deadline in time-space diagram
i_t	Router i at time t in time-space diagram
D	Corresponding deadline
R	Generation time of request for the currently focused traffic

Table I shows the notations used in this paper. Equation (1) shows the objective function of SAR. The first term is summation of cost for file transfer on link (i, j) along the space dimensional path (set of diagonal lines on time space diagram path). The second term is summation of storage cost at routers (set of horizontal lines on time space diagram path). So, the objective function of SAR aims for minimizing accumulation of link cost and storage cost. Link cost should have close relationship with link utilization, i.e., link cost should be large for high link utilization. Equations (2) and (3) give capacity constraint for storage and link, respectively. The first term of each equation gives accumulated link and storage usage for background traffic. The second term shows link and storage usage offered by currently reserved session. So, when (2) and (3) are satisfied, newly reserved traffic can be accommodated without overutilization. Equation (4) gives session conservation at each transient node. It means that the traffic coming into a node between $t-\Delta t$ and t must be equal to the traffic going out between t and $t + \Delta t$ for any node except requesting node and the receiver node. Equations (5) and (6) show session conservation at the requesting time of the sender and the deadline for the corresponding session, respectively. They mean that the traffic must go out from the currently requesting sender node at the requesting time and come into the receiver node at the corresponding deadline. The cost of link and storage are defined as (7) and (8), respectively.

Because we investigate fundamental effectiveness of traffic smoothing with SAR, we assume that storage capacity is enough large and much cheaper than link cost. Therefore, we set storage cost as 0 in this evaluation. It is possible to consider storages capacity limitation with defining cost function for the storage as link cost. We use binary integer

TABLE II
NOTATION USED IN ALGORITHM

Notation	Meaning
scr	Sender node of content
dst	Receiver node of content
dl	Deadline time of content distribution
G	Graph of network topology
G′	Time-space diagram for *G*
E′	Set of links in *G′*
unit_size	Chunk size that can be sent in unit time
src(*rqtime*)	*Src* at request time
dst(*dl*)	*Dst* at deadline time
P	Shortest path from *src*(*rqtime*) to *dst*(*dl*) in *G′*

programming for fundamental evaluation. It is also possible to use Dijkstra's algorithm to calculate shortest path in space-time diagram. Route calculation algorithm of SAR is shown in Algorithm 1. Notations used in Algorithm 1 are shown in Table II. This system assumes that link capacity in future can be reserved. We suppose to use centralized controller such as OpenFlow [9]. Centralized controller has a potential to manage link bandwidths in the network and to dictate scheduled transmission timing to routers.

Algorithm 1 Route calculation algorithm in SAR

```
$Calc\_SAR\_Route\ (G, src, dst, dl, size)$
    $G' = ExtractGraphToTimeDimension\ (G, src, dst, dl)$
    $SetCost\ (l\ in\ E')$
    $m = content\_size\ /\ unit\_size$
    for i = 1 to $m$ do
        $P = CalcShortestPath\ (G', src(rqtime), dst(dl))$
        $UpdateCost\ (l\ in\ P)$
    end for
```

B. Basic Performance

In this section, we evaluate SAR with simple tandem model with 5 routers (see Fig. 2). In this model, background traffic is artificially given as normal distribution with mean of 40 and standard deviation of 10. Capacity of each link is 100. Link usage of each session is simply given by fixed value 3.

SAR is comparatively evaluated with the following two design policies.

1. SAR-edge

   In SAR-edge, the sender, i.e., the content server, can schedule its transmission time so as to minimize total cost of link and storage on the path. In SAR, all routers on the path can store content file but in SAR-edge, only edge router can store content file.

2. Fastest Reservation (FR)

   At request generation timing, transmission of content file is scheduled as fast as possible on condition that link usage of scheduled traffic is up to link capacity.

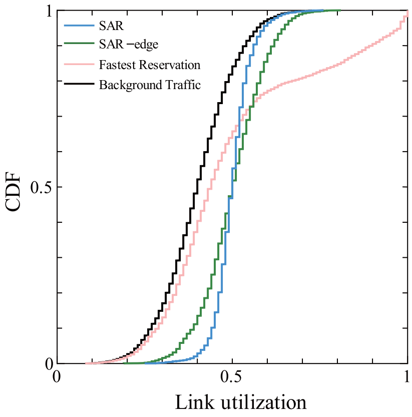


Figure 4. CDF of unit-time link utilization

Fig. 4 shows CDF of unit-time link utilization of SAR, SAR-edge and FR. CDF of background traffic is also shown in this figure.

In FR, large portion of content file transmission is scheduled at high link utilization time period. SAR-edge has smoothed link utilization distribution than FR, which means only with the sender scheduling smoothing of link utilization can be realized on some level. With scheduling of the router storage, SAR has more smoothed link utilization and has the best characteristics among three. As shown in this result, time dimensional scheduling enabled by large storage at a router has great effect on smoothing link utilization.

C. Accumulated Session Model

1) Link utilization smoothing effect of SAR: In this section, we evaluate SAR in more sophisticated model, accumulated session model. Fig. 5 shows mesh model with 5 senders, 5 receivers and 5 tandem routers. Sessions are set up for 5 server-receiver pairs of (a, j) (b, f), (c, g), (d, h) and (e, i). The latter 4 pairs generate the following traffic before the first pair so that generated traffic of these 4 pairs is background traffic for the first pair. At first, 1200 sessions (300 sessions for each pair) with short allowable reservation time (rather short time interval between request generation time and deadline) are generated for the latter 4 pairs. Their request generation is normally distributed with mean value denoted in Table III and standard deviation of 7.5. In pattern *a*, server *e* has the earliest generation mean time of its request and the latest one in pattern *b*. Mean deadline value

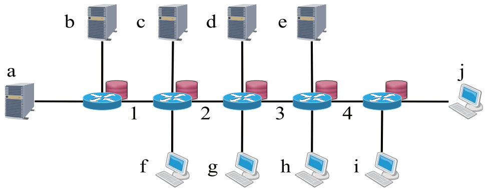


Figure 5. Mesh model

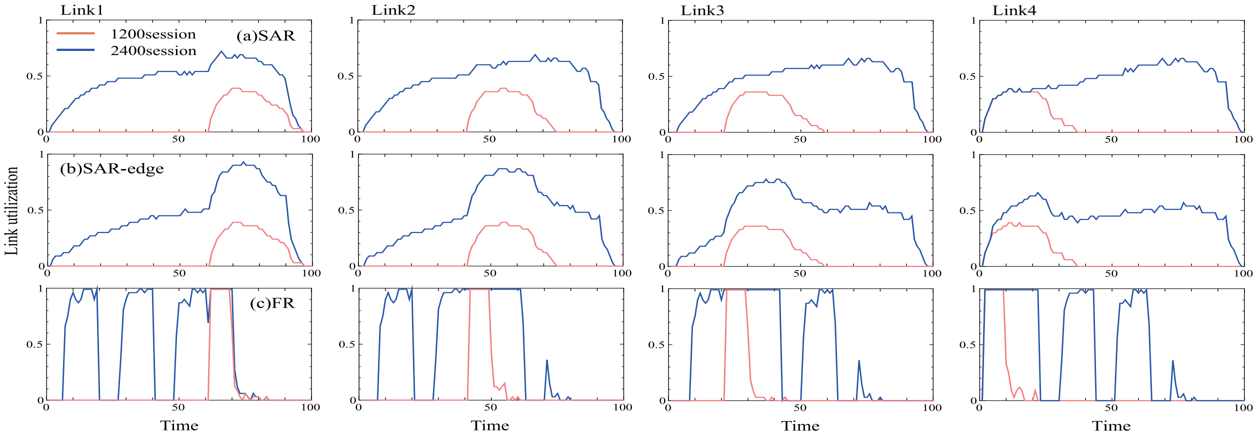


Figure 6. Link utilization - pattern *a*

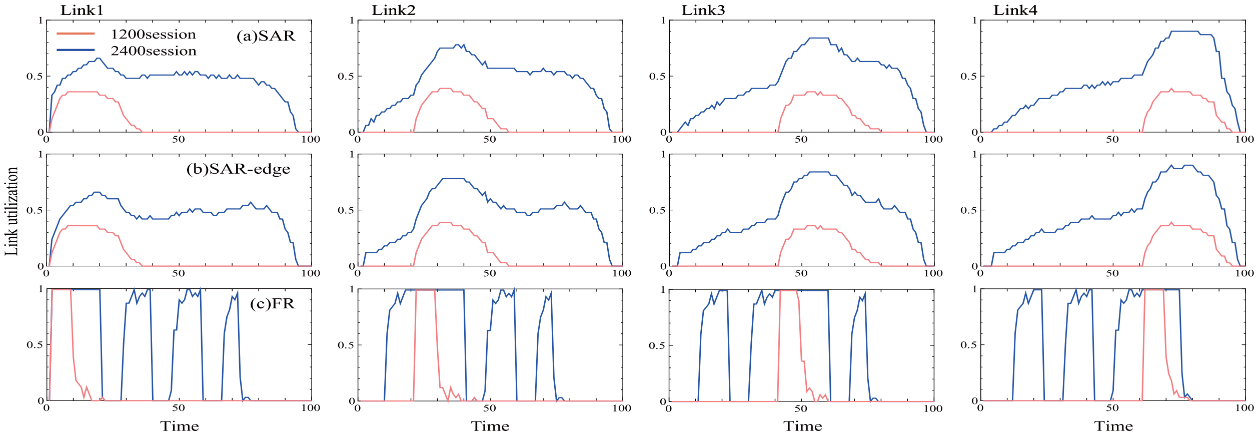


Figure 7. Link utilization - pattern *b*

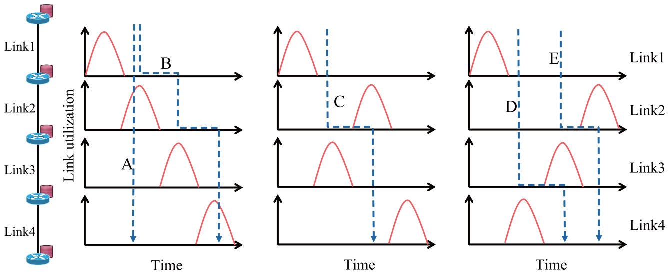
is exponentially distributed with 20 unit time later than mean generation time, e.g., server *e* has mean deadline of 40 in pattern *a*. With these assumptions, background traffic has average 20 unit time allowance for its reservation.

TABLE III
MEAN VALUE OF NORMAL DISTRIBUTION

server-receiver pair	pattern *a* mean value	pattern *b* mean value
b – *f*	80	20
c – *g*	60	40
d – *h*	40	60
e – *i*	20	80

After these generations of background traffic, Node *j* starts reservation to Server *a* for its generated 1200 sessions. Deadline of these 1200 sessions is normally distributed with mean 80 and standard deviation of 7.5. Request generation time is exponentially distributed between 0 and deadline assigned by the above described normal distribution. Figs. 6 and 7 show transition of link utilization for background traffic (1200 sessions at first generated: red line) and total 2400 sessions (blue line). As shown in Fig. 6, for background traffic in pattern *a*, SAR can schedule latter generated 1200 sessions so as to make link utilization smoothly flattened. In SAR-edge, high link utilization time interval generated by background traffic for each link is still high link utilization time interval after latter 1200 sessions are accommodated (see Fig. 6(b)). This is because in SAR-edge a router has no ability of time-shift of traffic peak and traffic peak generated at different timing at each link cannot be avoided only with time-dimensional scheduling at the sender side. FR starts its transmission as soon as possible, which leads to several time intervals with extremely high link utilization.

As shown in Fig. 7, SAR-edge and FR has poor performance from the viewpoint of smoothing link utilization. In pattern *b*, SAR also happen to have poor performance. This is because even with time-dimensional scheduling inside a network, high utilization time interval cannot be avoided. Even when a session can be scheduled to avoid peak traffic interval at a certain router, at any router beyond high utilization interval cannot be avoided. Fig. 8 simply explains this situation. As shown in Fig. 8(a), when peak traffic intervals line up in series, like pattern *b*, a path in time-space diagram cannot be found so as to avoid peak traffic intervals at all links. However, this happens only when peak time intervals line up clearly in series. When this sequence of peak time is broken even at one pair of links (see Fig. 8(b)), we can find a path avoiding peak link utilization. This means SAR can generally find a good path in time-space diagram. Even when there is a path(s) unfortunately having peak traffic interval in sequence, space dimensional routing might find another good path. In our future work, we will investigate and evaluate this space dimensional design of SAR.



(a)Case1(Pattern b) (b)Case2 (c)Case3

Figure 8. Path of session in case of every peak traffic interval

2) Number of rejected sessions: In this section, we evaluate SAR from viewpoint of number of rejected sessions. Further 1200 sessions are injected after accommodating 2400 sessions described in the previous section. These 1200 sessions are generated by the same distribution of session generation time and deadline as in the previous section. Table IV shows number of rejected sessions for session generation and deadline distribution of pattern *a* and *b*. In both cases, SAR has the lowest number of rejected sessions. So, SAR can design content file transmission timing so as to leave more room for upcoming sessions. It means that SAR can allow networks to accommodate traffic generated by users who want to watch video immediately.

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new traffic engineering technique for content distribution named SAR. When there is time difference between request generation time and content watching time, a user just cares about content is completely retrieved before her watching time and do not care about when it is retrieved. A router can schedule its transmission time in time-dimension when it has large storage. SAR makes use of this time-dimensional traffic control of scheduling of content file transmission time at each router in addition to the conventional space-dimension traffic control, i.e., routing. Numerical examples with binary integer programming formulation of SAR show that SAR can accommodate traffic so as to smooth link utilization at all links on the path. Our numerical example shows that there is limited condition where SAR cannot avoid high utilization interval. This limited condition of high link utilization interval set up in sequence along the path hardly happens, which means SAR generally works well. Even though this situation happens in a certain path, spatial routing can find another good path. Our next step is design of this spatial dimension routing of SAR. It is expected that the combination of spatial dimension routing and time dimension scheduling has more effect. Moreover, our future works are congestion avoidance on the storages and multicast using storages.

TABLE IV
NUMBER OF REJECTED SESSIONS

	pattern *a*	pattern *b*
SAR	4	224
SAR-edge	255	243
FR	657	669

ACKNOWLEDGMENT

This research is partly supported by the National Institute of Information and Communications Technology, Japan.

REFERENCE

[1] Cisco Visual Networking Index: Forecast and Methodology, 2012-2017, May. 2013. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html. [retrieved: June, 2013]

[2] I. Lazar and W. Terill, "Exploring Content Delivery Network," IEEE IT Professional, vol. 3, no. 4, pp. 47-49, 2001.

[3] Y. Wang and Z. Wang, "Explicit Routing Algorithms for internet Traffic Engineering," in Proc. IEEE ICCCN, Oct. 1999, pp. 582-588.

[4] Grasp of total traffic amount on the Internet in Japan, Ministry of Internal Affairs and Communications, http://www.soumu.go.jp/main_content/000211328.pdf, Mar 2013. [retrieved: June, 2013]

[5] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," in Proc. ACM SIGCOMM, Aug. 2003, pp. 27-34.

[6] S. Jain, S. Gopinath, and D. Raychaudhuri, "STAR: Storage aware routing protocol for generalized Delay Tolerant Networks," in Proc. IEEE WoWMoM, Aug. 2011, pp. 1-4.

[7] N. Somani, A. Chanda, S. C. Nelson, and D. Raychaudhuri, "Storage aware routing protocol for robust and efficient services in the future mobile Internet," in Proc. IEEE ICC, Jun. 2012, pp. 5849-5853.

[8] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-Datacenter Bulk Transfers with NetStitcher," in Proc. ACM SIGCOMM, Aug. 2011, pp. 74-85.

[9] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," in ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, pp. 69-74, 2008.

Towards Decentralized Networlds

Current Developments in the P2Life Project

Hauke Coltzau
Department of Communication Networks
FernUniversität in Hagen
Hagen, Germany
hauke.coltzau@fernuni-hagen.de

***Abstract*—Decentralized networlds can serve as platforms both for interaction between users and for content and service delivery. They may well be the next evolutionary step of the WWW. To gain acceptance of users and businesses, content and service providers must be able to keep their independence as in today's WWW and still maintain the impression of a coherent unlimited virtual world. Existing approaches emerged from the area of networked virtual environments have so far failed to fulfill these basic requirements. In this article, we present an architecture to build unlimited networlds, in which interaction between user and content provider is handled directly. This allows among others for real money trading as in the WWW. Additionally, we discuss our approach to build and maintain dynamic maps of the virtual world without need for centralized instances or any other form of global knowledge.**

Keywords-networlds; networked virtual environments.

I. Introduction

In the late 1990s, the question came to focus, how users would experience the upcoming WWW in their everyday life. Among others, the dutch architect Rem Koolhaas observed hopes, expectations and social changes through the integration of the WWW as follows:

> What about the widely heralded hope that cyberspace will be the new street, the piazza in our sprawling cities of connected isolation; that *cyberflâneurs* will promenade around the Net like Baudelaire taking a stroll around Paris? (see [1])

In the context of this article, it is not so much the question whether these changes actually have come into effect today, but instead, it is the term *cyberflâneur* that shall be of interest. According to Koolhaas (and others unmentioned here), the vision existed that users would be strolling the WWW from site to site, without the need for a specific goal to go to. The early WWW supported this kind of browsing paradigm, since no centralized (and, therefore, dominant) instances as, e.g., search engines existed. The users were ”forced” to stroll, and they could do so anonymously.

Fifteen years later, the publicist Evgeny Morozov takes on the idea of cyberflânerie – and declares it dead [2]. He argues that almost all browsing activity is channelled through a few central nodes, especially search engines. Moreover, in his opinion, the anonymity of the early WWW got lost due to the omnipresence of social networks, in which everyone shares everything with everybody else.

Koolhaas and Morozov do not provide scientific foundation for their arguments, but rather illustrate personal experiences and implications. Nevertheless, with the idea of cyberflânerie, they give a sketch of an interesting and enriching way of using and experiencing contents as well as services in the WWW. In other words: A coherent 3D networld consisting of virtual objects representing content and services, in which users can interact with each other but also perform trading with content and service providers as in today's WWW, could be the next evolutionary step of what we call 'the net' today.

As approaches like *SecondLife* [3], *Kaneva* [4] and *3D-City* [5] show, it has already been tried to put the idea of a coherent, browseable and intuitively understandable virtual content- and service environment into practice. They all have in common to provide a platform for a virtual environment, into which users can add interactive objects to display content and provide services. Although a remarkable interest from the user side has been brought towards such interaction and trading platforms (e.g., *SecondLife* has $\approx$ 60.000.000 users), none of the existing approaches has come so far yet to be recognized as a replacement for even just a few current WWW contents or services.

Even though virtual worlds provide new ways of displaying products and of interacting with potential customers, businesses are remarkably cautious about investing resources into virtual worlds and use them merely as an additional platform for advertising, but not for actually trading their products and services. This is mainly due to three reasons:

1) Businesses cannot control the availability (visibility) of their objects, but instead have to rely completely on the availability of the platform provider.
2) An unjustifiable amount of unidirectional trust into the platform provider is necessary, since all virtual objects, which may need to contain internal business information, must be hosted on the providers machines.
3) Usually, trade is done on base of virtual currencies (e.g., the *Lindendollar* in SecondLife). This leads to unforeseeable exchange rate risks for businesses, rendering safe trading almost impossible.

These problems can be overcome by distributed architectures, as they have emerged from Massively Multiplayer

Online Games (MMOGs). Decentralized virtual environments, usually referred to as Networked Virtual Environments (NVEs, [6]), have in common that the software components needed for managing a coherent and consistent virtual world are linked together on the base of local algorithms and without any global knowledge or control.

NVEs have remarkable advantages over centralized approaches for users and businesses:

- **Reduced dependency and necessity of trust:** Availability and visibility of contents and services do no longer have to depend on single platform providers. There is no need to upload business critical data to a third party platform provider.
- **Scalability:** Decentralized approaches can scale with both the number of users as well as the size of the virtual world.

But, none of the existing approaches for NVEs can be used as base for a next evolutionary step of the WWW, because they all expect deeper cooperation between the participating peers, whereas in a decentralized interaction and *trading* platform, the autonomy of each single participant (i.e., WWW server) is of highest importance.

A structure that provides both a coherent virtual world without making service and content providers dependent on centralized instances or other participants of the virtual world, shall be referred to as *networld* for the remaining part of this article.

In Section II, a brief overview is given on concepts and projects related to this topic. Section III displays a summary of requirements on networlds, which we have already developed in previous works (see e.g., [7] and [8]). In Section IV, the architecture of *P2Life*, our approach to build and maintain virtual decentralized networlds is presented. Section V shows, how this structure can be used to build maps of the whole virtual world efficiently and still with high redundancy without using centralized instances. Section VI concludes the article and gives an outlook on future work.

II. Related Work

Since networked virtual environments focus on interaction (and not, as networlds, on flânerie), the most important aspect in terms of scalability certainly is, how event multicasting and filtering is performed.

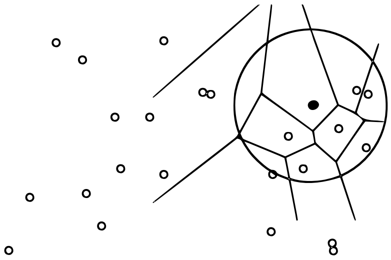


Fig. 1. Nodes in a Voronoi-based Overlay Network (VON).

Makbily et. al. (see [9] and [10]) suggest, that participants use their current position to calculate *Update-Free-Regions (UFR)* pairwise for every other participant. These are regions of the virtual world, in which participants cannot propagate events to each other. They could result, e.g., from obstacles, which make participants invisible for each other. But since the UFRs have to be calculated pairwise amongst the participants, the calculation does not scale, because the number of messages needed is in $O(N^2)$. Steed and Angus also filter events based on mutual visibility [11] [12]. In their approach, the virtual world is partitioned into cells. For each cell, a rough estimation is possible, which other cells are visible from there. This information can be used to generate so-called *Frontier Sets*, out of which users can *not* see each other. Bharambe et. al. propose *Donnybrook* [13], in which the number of event messages is reduced by estimating, on which other participants a user currently focuses (*Interest Set*). A similar approach has been introduced by Najaran and Krasic [14].

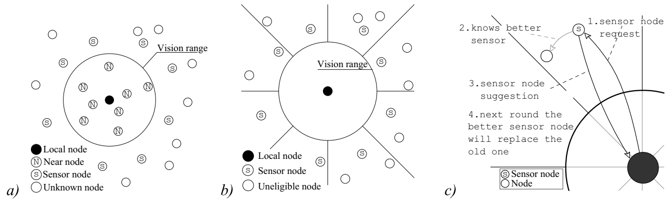


Fig. 2. Network from the perspective of a single node in *pSense*. a) Point of view and roles of the nodes, b) segmentation to select *sensor-nodes*, c) Finding the optimal *sensor-node* of a segment (Source: [24])

Approaches that are not only used for event management but also for object management actually create and maintain virtual worlds. They can be classified into *static* and *dynamic segmentation* approaches, the first partitioning the virtual world into segments of fixed size and location, while the latter dynamically decides on size, shape, number and location of segments.

A. Static Segmentation

Knutsson et. al., who authored *SimMud* [15], assume that the structure and content of the virtual world is static and hence is available offline on every client. For avatar management, the world is partitioned into fixed segments, in which events are forwarded to all avatars located in it. Participants can register for events of several regions. Events starting from single participants are forwarded in a Pastry network [16] [17] using Scribe multicasts [18] [19].

In the *Zoned Federation of Game Servers* proposed by Iimura et. al. [20], each partition (zone) is assigned to a zone server managing all objects and events within. The zone servers are linked together based on a dynamic hashtable, which allows users to find out, which zone server is responsible for any given zone. Similar approaches for static segmentation with only minor variations have amongst others been proposed by Lee and Sun [21], Yamamoto et. al. [22], and de Oliveira et. al. [23].

B. Dynamic Segmentation

Although a lot more approaches for dynamic segmentation exist, only two shall be mentioned here. Schmieg et. al. [24] assign each object to a node (peer). Objects (including avatars) are not bound to a position, but can freely move through the virtual environment. Each node manages a circle-shaped field of view around it. All other nodes within this circle (*near-nodes*) receive events from the node in the center. Additionally, each node links itself to eight other nodes outside of its field of view (*sensor-nodes*), one – namely the closest – in each eighth of the compass rose (see Fig. 2).

The most advanced project based on dynamic segmentation was proposed by Hu and Chen [25] [26] and is based on a Voronoi-based Overlay Network (VON) [27]. Each node is linked to its closest neighbours (in terms of distance in the VE, see Fig. 1). This structure is used to provide *spatial publish-subscribe*, i.e., dynamically forming regions of arbitrary size, in which events are forwarded and received. The task of filtering and forwarding is no longer assigned to single peers, but instead solved cooperatively amongst all peers in a region. The approach is prone to consistency problems (see [28]), which can nevertheless be handled locally.

All dynamic segmentation approaches have in common that they do not have to limit the size of the virtual world.

III. Requirements on Decentralized Networlds

A decentralized networld, which shall be used as both an interaction and a *trading* platform must allow trading based on real money to happen in the same way as in today's WWW. Businesses and customers shall be able to use well known and well established payment methods as, e.g., credit cards or online payment services for transactions. Existing approaches for NVEs, of which a lot can undoubtly be used as interaction platforms, are however not designed to support trading.

A system that aims to merge the advantages of today's WWW with the advantages of NVEs, must fulfill specific requirements, which mostly result from real world trading and, therefore, go far beyond requirements on NVEs. In fact, when real money comes into consideration, a change of paradigm for creating a decentralized coherent virtual environment is necessary, in which cooperation between the participating peers is reduced to a minimum. Each provider must be able to maintain its autonomy and thus be in control of the data and system security.

Requirement 1 – Object ownership: In most legal systems, trading is understood as a bilateral legal act, in which the participants have to be uniquely identifiable. Since trading in the WWW as well as in the proposed networld is performed through virtual objects, these objects need to be assigned to a legal body, i.e., the contracting party of the trade. In contrast, current approaches for NVEs dynamically assign objects to peers based on structural circumstances, i.e., often an explicit owner for each object is not given.

Requirement 2 – Independent management layer: For an object to be visible in the virtual world, a management layer is needed that allows to lookup, which objects are located on a given position or within a given region. When the management layer is based solely on the structure of the virtual world, the peers managing the *neighbourhood* of an object could limit its visibility. Therefore, the management layer must be independent from the structure of the virtual world.

Requirement 3 – Unlimited virtual world: The size of the virtual world, i.e., the number of participants, is not limited. No artificial shortage of the resource *space* is allowed.

When looking at the existing NVE approaches, none of them fulfills all three requirements, which leads to the conclusion that a new approach is needed to create virtual networlds as interaction- and trading platforms.

IV. Architecture

The main part of the proposed approach, which we have named *P2Life*, contains an independent decentralized management layer, in which content providers can register their contents at specific coordinates and users (browsers) can lookup all objects located at a given coordinate or region.

A. Structure of the P2Life Decentralized Networld

The unlimited P2Life networld is divided into square-shaped parcels of equal size. Other shapes as, e.g., hexagons could be used, but nevertheless we stay with square parcels for simplicity reasons. Each parcel is either assigned to exactly one provider, or it is empty. The provider assigned to a parcel 'owns' it, i.e., he is the one to be contacted for any content located on it.

Interaction between users and providers happens in the same manner as in today's WWW. When a user enters a parcel, the browser contacts the provider of the parcel to get a description of the objects existing on it, e.g., in form of a markup language like 3DMLW. It is not necessary to decide for *one* system-wide object description language. It is well possible to use application specific languages, as long as the browser is able to understand them. Having received the object descriptions, the browser is now able to render the objects for the user and handle interactions between user and objects, possibly leading to subsequent data exchange between user and provider (analogue to input fields in HTML).

The illusion of a coherent virtual world is created by the user's browser. Whenever a parcel is visited, the browser automatically contacts the providers of the neighbouring parcels and gathers information about their displayable objects. This information can be used to display a coherent environment, in which parcel borders need not to be visible at all. It solely depends on the user's *Area of Interest* and the available data transfer rate, which parcels are displayed.

B. Registration and Lookup Architecture

To be assigned to a previously empty parcel, a provider has to register its (IP-)address for the parcel. Users can then generate a lookup for the according coordinate and receive the address of the provider to contact for more information. It is worth mentioning that using human readable names for identifying a host as in the *domain name service* still remains possible. Identifying a provider using its coordinate is to be understood as an additional way of addressing, specifically useful in virtual environments.

To achieve this, a registration and lookup infrastructure is needed, which is the core architectural component of the networld described here. According to the autonomy requirement, this infrastructure must be organized in a decentralized manner. Additionally, it must not rely on the structure of the virtual world itself.

The management layer is based on a *Content-Addressable-Network* (CAN), which is a natural choice due to its planary structure. Providers, who have occupied at least one parcel in the virtual world, are automatically assigned to maintain a region in the CAN. Other DHT-like architectures could also be used, yet, the most important advantage of CANs in this context is their robustness against *man-in-the-middle* attacks, since they provide a large amount of disjunct paths between any two peers.

Coordinates are mapped to 2-dimensional keys in $[0..1] \times [0..1]$ using a system wide hash function with an adaptive number of digits. This way, the unlimited character of the coordinate space is mapped into the arbitrary granularity of the key space. The hashing guarantees that the structure of the virtual world is not reflected in the management layer's neighbourhood. The CAN-Peer managing the key for any given coordinate holds a reference (address) to the provider assigned to the respective parcel. If no provider address is assigned to that key, the parcel is assumed to be empty.

When a provider wants to register itself for any given coordinate, it generates a registration request containing the hash value of the requested coordinate and the provider's IP-address. The request is signed using the provider's credentials and then forwarded to an arbitrary CAN-peer. The peer can either directly handle the request, because it manages the required hash value, or it can forward it in the CAN until the managing CAN-peer is found. This peer can now directly contact the provider and ask for the coordinate of the parcel, for which the provider wants to register.

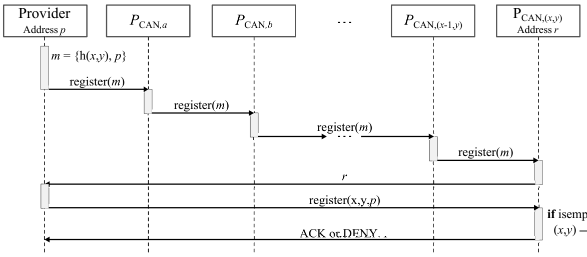


Fig. 3. Registration of a provider

If the required coordinate is empty, the CAN-peer can acknowledge the request and store the provider's IP-address in its local database. Otherwise, the request is declined.

The lookup for finding out, which provider is registered for any given coordinate, is performed in a similar way. The requesting user (browser) generates the hash value for the requested coordinate and forwards a lookup request containing the hash value and the user's IP-address to any CAN-peer. The message is forwarded in the CAN until the managing CAN-peer is found. This CAN-peer can contact the browser directly and ask for the plaintext coordinate. If a provider address is stored for this coordinate, it is returned to the client. Otherwise, the requested parcel is assumed to be empty.

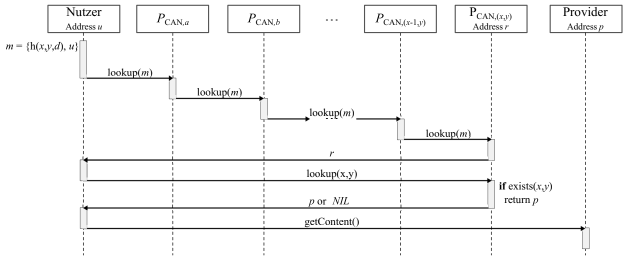


Fig. 4. Lookup scheme

C. Hashing

The average number of messages to reach an arbitrary peer in the CAN is in $O(\sqrt{N})$. While this is efficient enough for registering a provider (because it can be assumed that this happens only once per provider) as well as for lookups, browsing would become a problem. As described above, the illusion of a coherent virtual world is created, because the browser automatically looks up and contacts the providers in the neighbourhood of a requested parcel. But since neighboured parcels are usually not managed by neighboured CAN-Peers, each of these lookups would also be in $O(\sqrt{N})$ time complexity, which would disturb a smooth browsing experience.

For this reason, each CAN peer maintains for each coordinate it manages, direct links to the CAN peers that manage the four neighboured parcels of that coordinate. The resulting structure allows to route requests in keyspace, thus independent from the structure of the virtual world, but additionally provides constant time lookups for neighboured coordinates.

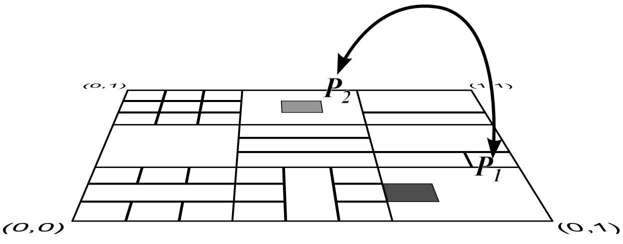


Fig. 5. Direct link between to distant CAN-peers managing keys for neighboured coordinates

But since these links have to be maintained for every key (even for those representing empty parcels) and since each key represents an unlimited number of coordinates (due to the fact that the key has a limited number of digits), a scalability problem arises: Each peer would have to maintain links to (almost) all other CAN peers. A solution for this problem is to use a hashfunction with a characteristic, which could best be described as *collision-preserving*:

Let (x_{c1}, y_{c1}) and (x_{c2}, y_{c2}) be two arbitrary coordinates mapping on the same d-digit hashvalue, i.e., $h_i(x_{c1}, y_{c1}, d) = h_i(x_{c2}, y_{c2}, d)$. Furthermore, let $(x_{c1}+1, y_{c1})$ and $(x_{c2}+1, y_{c2})$ be the right neighbours of (x_{c1}, y_{c1}) and (x_{c2}, y_{c2}). Then these two values also have to be mapped on an identical hashvalue, i.e., $h_i(x_{c1}+1, y_{c1}, d) = h_i(x_{c2}+1, y_{c2}, d)$. This also needs to be true for the other three neighbouring coordinates, so that for each key only four links to other CAN peers need to be maintained.

This way, the Content-Addressable-Network evolves to a structure, which reflects both the lattice structure of the virtual world *and* the independent keyspace, in which the neighbourhood of providers is deliberately broken up. The advantages of both structures can be used without being forced to accept their disadvantages.

D. Summary

The proposed architecture allows to manage an unlimited coherent virtual networld built by independent content and service providers. Each of the providers can act the same way as today's WWW servers, which includes direct real money trade with their customers.

Due to the nature of the underlying Content-Addressable-Network, neighbourhood related attacks as well as man-in-the-middle attacks are more difficult than in previous decentralized approaches for virtual environments. Still, with the imprinted lattice structure, neighbourhood based routing and, as a result, flânerie is supported in an efficient manner.

V. Decentralized Maps

In the current version of P2Life, no restrictions exist on which parcels a provider can register for. Especially, no relation exists between the position of a provider and its content. Hence, with a growing amount of participating providers, navigation and orientation become a challenge for the user.

A likely solution for this problem is to let the system provide a coherent map of the virtual environment, through which the user can navigate in different levels of detail. In the highest level of detail, a map fraction representing the top view on a single parcel is displayed. On the lowest level of detail, all nonempty parcels are shown in a single map for a general overview. The intermediate levels let the user display map fractions containing more (lower level of detail) or less (higher level of detail) parcels.

The algorithms used to create, maintain and provide these maps of course have to be decentralized to avoid generating a dependency on centralized services. Additionally, the algorithms have to comply with the general requirements on decentralized networlds (see Section III):

- **Authorative Owners:** Each Provider manages the top views for each of their parcels on their own.
- **Verifyability:** Except for a single top view, each request for a map fraction can be answered by several independent peers so that malicious behaviour of single peers can be revealed.
- **Redundancy:** With growing level of a map fraction (i.e., with increasing number of parcels, a map fraction represents), the number of peers maintaining this map fraction also increases. Therefore, local disturbances only have local influence.

Fulfilling these requirements leads to an architecture, in which malicious behaviour of single peers can neither disturb the consistency nor the availability of the map as a whole or of its fractions. The approach described in this section covers two aspects:

1) A feasible **map structure** to manage and link the map fractions
2) A strategy for an efficient **redundant assignment** of data sets to peers maintaining them.

A. Map Structure

The map data structure is organized in such a way that different zooming levels can be provided. Therefore, the virtual world is divided into square disjunct map fractions for each level e. The division depends on a system constant b, which controls the size of the map fractions in combination with the zooming level as follows: In the lowest zooming level $e_0 = 0$, each map fraction contains a top view on a single provider, on the highest level $e_{max} \approx \log N$, the only existing map fraction represents the whole nonempty part of the unlimited virtual world. Between those two extremes, additional levels $0 < e < e_{max}, e \in \mathbb{N}$ exist, in which each map fraction represents b^{2e} parcels (see Fig. 6b).

The centre coordinates (x, y) of the map fractions are chosen in such a way that each map fraction in level $e > 0$ fully contains exactly b^2 map fractions of level $e-1$:

$$x = c_x \cdot b_m^e + \frac{b_m^e - 1}{2}, y = c_y \cdot b_m^e + \frac{b_m^e - 1}{2}, c_x, c_y \in \mathbb{Z} \qquad (1)$$

It is obvious that an appropriate data structure to manage the map fractions on different levels must have the nature of a tree, in which each node represents a part of the map. Nodes close to the root represent map fractions of higher levels (i.e., of lower detail), while leaf nodes represent top views on single providers.

The map is generated bottom-up, i.e., the providers manage the information of the leaf nodes. Each provider maintains for each of its parcels all information necessary to display them in the map. The map fractions of the higher levels are generated using the information available from their child map fractions of lower levels. Changes in the map are also generated on level 0, e.g., by adding or removing providers. For each map fraction a dataset exists that contains the following elements:

1) one or more displayable *representations* of the map fraction.
2) either 0 or 2 to b_m^2 links to *child maps*, i.e., map fractions that represent a part of the virtual world that is located within the same area as the current map fraction. Hence, child maps always have a lower level than the map they are a child of.

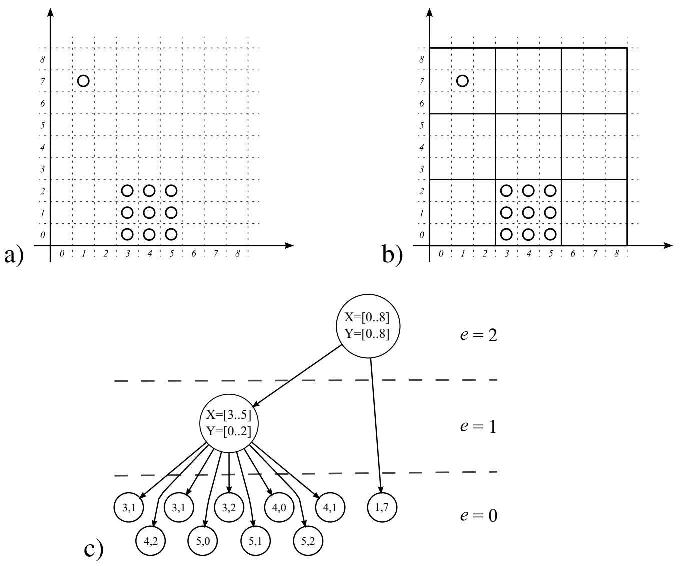


Fig. 6. a) Providers in an example virtual world, b) map partitioning, and c) resulting tree of map fractions.

3) 0 or 1 link to a *parent map*, i.e., the map fraction that this map is a child map of. If no such map exists, the current map is the root map and represents the whole nonempty virtual world.

The tree contains only relevant map fractions. There is no need to maintain a map fraction that contains no data, because no providers are located within the respective area of the virtual world. In the same way, a map fraction that only has one child map is obsolete, since it is fully represented by that single child map. The child maps are chosen in such a way that starting from any of the b_m^2 (max.) child maps, each map fraction of level 0 (i.e., each single provider) that is represented by the respective child map, can be reached. Additionally, the child maps are chosen to have the lowest level possible. This reduces the number of reorgs necessary to handle changes in the map datastructure.

In Fig. 6.a, an example of a virtual world with 10 providers (circles) is shown. There is a area with higher provider density around parcel $(4,1)$ and a single outlying provider on parcel $(1,7)$. Fig. 6.b shows the resulting partitioning on levels 0, 1 and 2. The base for the partitioning has a value of $b = 3$. The resulting tree of map fractions in Fig. 6.c starts with a root map of level 2, i.e., with an edge length of $3^2 = 9$ and its center at $(4,4)$. The area with higher provider density around $(4,1)$ is represented by an intermediate map of level 1, which contains 9 child maps of level 0. The outlying provider at $(1,7)$ is instead directly linked to the root map.

This way, the number of map fractions necessary to represent any given area of the virtual world roughly scales with the number of *providers* within it instead of its *extent*. Additionally, the time and message complexity to reach an arbitrary map fraction of level 0 starting from the root map remains in $O(\log N)$.

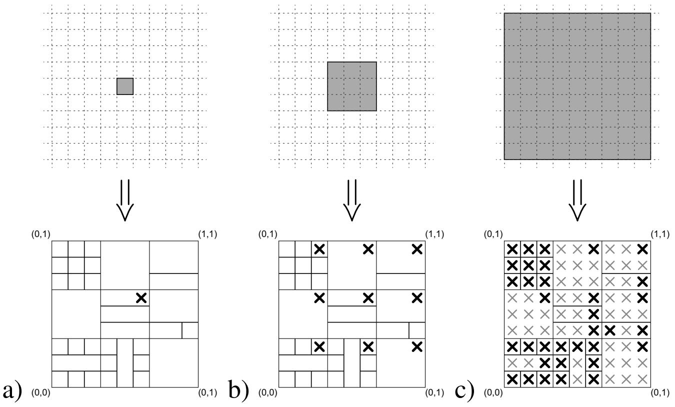


Fig. 7. Carrier assignment example for a) level $e = 0$, b) level 1 and c) level 2 in a CAN with base $b = 3$.

B. Assignment of map fractions to peers

The maintenance of the map fractions is integrated into the already existing management structure (see Section IV-B). A CAN-Peer, who manages a given map fraction, is called a *carrier* of it. Only the map fractions of level 0 have a single carrier. All map fractions of higher levels have multiple carriers. The assignment of a map fraction of level to its carriers is done by building the 2-dimensional hash value (k_x, k_y) of the map's center coordinate. through the key (see Section IV-C). A CAN-Peer is carrier for a map fraction of level e, who's center coordinate is mapped to $k = (k_x, k_y)$ in hashspace, when the CAN region of the peer covers a key matching k with the leading e digits masked out. For a single provider, i.e., for $e = 0$, the CAN-Peer managing k is the only carrier. For the root map, i.e., $e = e_{max}$, (almost) all CAN-Peers are carriers.

In Fig. 7, the assignment of map fractions to carriers is demonstrated with three example maps having the same center coordinate but different levels. In subfigure c), some over-determined carrier assignments can be seen, i.e., a CAN-Peer manages several keys matching the masked hashvalue of the according map's center coordinate. In such a case, the carrier will of course only maintain a single copy of the map fraction.

Events, which can influence the appearance of the map fractions, are always initiated on level 0, i.e., by providers.

- **Registration:** A provider registers for a formerly empty parcel. The according carrier is contacted during the registration process and receives all information necessary to manage a map fraction of level 0 from the provider.
- **Update:** The information for a single parcel has changed. The provider generates an update message for the according carrier.
- **Unregistration or provider fail:** A provider has unregistered or is assumed to be ultimatively offline for other reasons.

In all three cases, the according carrier on level 0 is contacted by the provider. The carrier performs the necessary changes on the map fraction data and then contacts the b^2 carriers of the next higher level (i.e., the carriers of the parent map), which can be done with local knowledge. Each of these carriers updates the map fraction data for the parent map and forwards the event to their parent map carriers, and so on.

To maintain locality and to reduce network load, each carrier of level e forwards events only to carriers within the same CAN-region (see Fig. 8), i.e., each carrier regardless of its level only has to generate a maximum of b^2 parallel update messages to perform the forwarding.

Attack- or failure scenarios derived from misbehaviour of single peers can be handled, if

1) each carrier of level e sends an event to *few* arbitrary carriers of level $e+1$ *outside* of the current carrier's CAN-region. Additionally, the carrier forwards the event to 1 or 2 arbitrary carriers of level e *within* its CAN-region.
2) carriers of identical map fractions randomly cross-check their data from time to time.

This way, updates do reach a carrier on different ways without influencing the overall performance. Malicious behaviour of single peers can be revealed in short time.

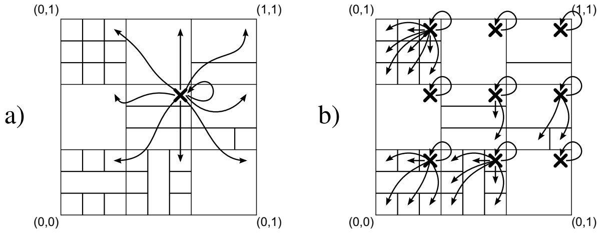


Fig. 8. Example event forwarding from a) level 0 to level 1, b) level 1 to 2.

C. Performance

The most important aspect in practical matters is, whether the proposed architecture allows users to navigate through the map in such a way, that they are able to contact any peer in the CAN and reach the carrier for any map fraction efficiently.

It is obvious that each node in the tree of map fractions can be reached in $O(\log N)$ time complexity, when the CAN distance between two carriers is ignored. Hence, if each carrier maintains a small cache with direct links to carriers of parent- and child maps, the number of hops to reach an arbitrary map fraction carrier is also in $O(\log N)$. Without such a cache, the number of hops is in $O(\sqrt{N})$.

The second aspect of the performance evaluation covers the handling of map events. To avoid that each event has to be forwarded up all the way to the root node, updates are only forwarded when they perform a noticeable change in the parent map. Otherwise, the update event is stored locally and merged with further events, until an effect on the parent map actually occurs. This way, the *average* number of forwards in the tree for each event is constant, i.e., in $O(1)$.

Due to limited space, a proof is omitted here and only an example simulation showing the average number of hops over different network sizes is given in Fig. 9a.

As mentioned above, for each forwarding of an event to the next higher level, each carrier has to send b^2 physical messages through the CAN. Each of this messages has an average path length of $\sqrt{N/b^{2e}}$. Hence, the total number of physical messages to forward an event from level e to level $e+1$ is

$$b^{2(e+1)} \cdot \sqrt{N/b^{2e}} = b^{e+2} \cdot \sqrt{N} \tag{2}$$

If again a threshold Δ is used to limit forwarding to events that actually have an effect on the next higher level, the total number of messages sums up to

$$\frac{\sqrt{N}}{\Delta} \sum_{e=0}^{e_{max}-1} \frac{1}{b^e} < \frac{\sqrt{N}}{\Delta} \frac{1}{b-1} \in O(\sqrt{N}) \tag{3}$$

without using caches (see Fig 9b).

When caches are used, i.e., when instead of using the CAN for message delivery, each carrier can contact the b^2 carriers of the next higher level directly, the average number of physical messages is reduced to

$$\frac{1}{\Delta} \sum_{e=0}^{e_{max}-1} \frac{b^{2(e+1)}}{b^{2(e+1)}} = \frac{e_{max}-1}{\Delta} \in O(\log N) \tag{4}$$

D. Summary

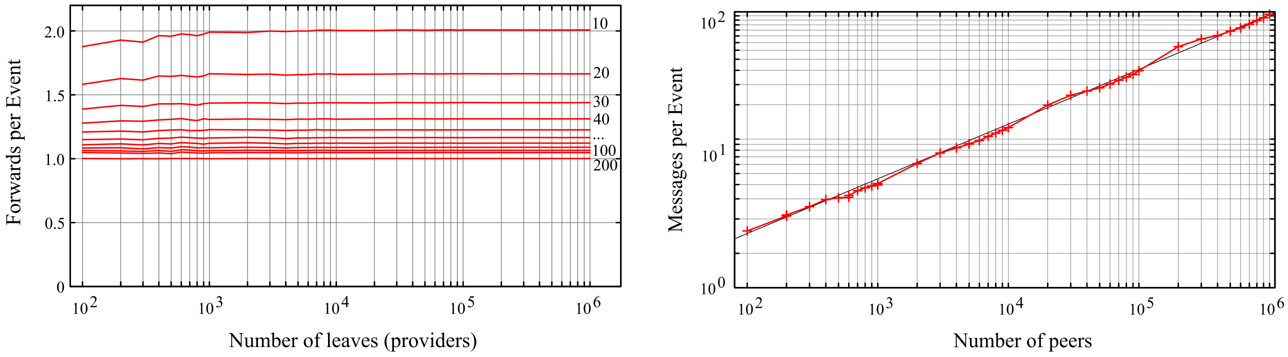
As long as no semantic navigation is possible in virtual worlds, navigable maps are an essential tool for orientation. Since the main idea behind this article is to provide a *decentralized* networld, creation and maintenance of such maps may not be conceded solely to centralized instances. The proposed approach utilizes the already existing architecture to store a tree of map fractions in a massively redundant way. By using simple threshold mechanisms, both maintenance and usage of the resulting maps can be done efficiently.

VI. Conclusion

Virtual networlds as platforms for interaction, content and service delivery do have the potential to be the next evolutionary step of the WWW. To do so, they have to allow content and service providers to keep their independence as in today's WWW and still maintain the impression of a coherent unlimited virtual world. Existing approaches in the area of networked virtual environments fail to fulfill these basic requirements. Therefore, we have presented an architecture to build unlimited networlds, through which users can browse as cyberflâneurs without the need for a specific target or to give up their anonymity. Interaction between user and provider is handled directly, which allows among others for real money trading as in the WWW.

Navigation and Orientation is provided through dynamic maps of the nonempty parts of the networld. The maps are stored in a decentralized way utilizing the already existing architecture to maintain the networld, with a high level of redundancy and yet with low maintenance costs in terms of hops and physical messages necessary.

Yet, a single reality virtual world will inevitably reach its limits, when the number of content and service providers increases. It is therefore necessary to allow a kind of semantic navigation, i.e., to build individual views onto the available content based on a user's current context.



(a) Average number of forwards in the tree of map fractions to handle an event originating from level 0 using thresholds from 10 to 200.

(b) Average number of *physical* messages in the CAN to handle an event, when a threshold Δ exists and no caches are used.

Fig. 9. Simulation results for handling events in the tree of map fractions in a decentralized manner.

References

[1] N. Gardels, “The changing global order: world leaders reflect”. Wiley-Blackwell, 1997.

[2] E. Morozov, “The death of the cyberflâneur,” http://www.nytimes.com/2012/02/05/opinion/sunday/the-death-of-the-cyberflaneur.html, retrieved June 2013

[3] SecondLife website, http://secondlife.com/, retrieved June 2013

[4] Kaneva project website, http://www.kaneva.com, retrieved June 2013

[5] 3D-City project website, http://www.3dcity.de/, retrieved June 2013

[6] S. Singhal and M. Zyda, “Networked virtual environments: design and implementation”, ser. SIGGRAPH series. Addison-Wesley, 1999.

[7] H. Coltzau, “P2life: An infrastructure for networked virtual marketplace environments,” *IJIIP*, vol. 1, no. 2, 2010, pp. 1–13.

[8] H. Coltzau and B. Ulke, “Navigation in the p2life networked virtual marketplace environment,” in Autonomous Systems: Developments and Trends, ser. Studies in Computational Intelligence, H. Unger, K. Kyamakya, and J. Kacprzyk, Eds. Springer Berlin / Heidelberg, vol. 391, 2012, pp. 213–227.

[9] Y. Makbily, C. Gotsman, and R. Bar-Yehuda, “Geometric algorithms for message filtering in decentralized virtual environments,” in Proceedings of the 1999 symposium on Interactive 3D graphics, ser. I3D ’99. New York, NY, USA: ACM, 1999, pp. 39–46.

[10] A. Goldin and C. Gotsman, “Geometric message-filtering protocols for distributed multiagent environments,” *Presence: Teleoper. Virtual Environ.*, vol. 13, no. 3, Jul. 2004, pp. 279–295.

[11] A. Steed and C. Angus, “Enabling scalability by partitioning virtual environments using frontier sets,” in Presence: Teleoper. Virtual Environ., vol. 15, no. 1, Feb. 2006, pp. 77–92.

[12] S. Avni and J. Stewart, “Frontier sets in large terrains,” in Proceedings of Graphics Interface 2010, ser. GI ’10. Toronto, Ont., Canada, Canada: Canadian Information Processing Society, 2010, pp. 169–176.

[13] A. Bharambe, J. R. Douceur, J. R. Lorch, T. Moscibroda, J. Pang, S. Seshan, and X. Zhuang, “Donnybrook: enabling large-scale, high-speed, peer-to-peer games,” SIGCOMM Comput. Commun. Rev., vol. 38, no. 4, 2008, pp. 389–400.

[14] M. T. Najaran and C. Krasic, “Scaling online games with adaptive interest management in the cloud,” in Proceedings of the 9th Annual Workshop on Network and Systems Support for Games, ser. NetGames ’10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 9:1–9:6.

[15] B. Knutsson, H. Lu, J. C. Mogul, and B. Hopkins, “Architecture and performance of server-directed transcoding,” in ACM Trans. Internet Techn., vol. 3, no. 4, 2003, pp. 392–424.

[16] A. Rowstron and P. Druschel, “Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems,” in IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), Nov. 2001, pp. 329–350.

[17] M. Castro, M. B. Jones, A.-M. Kermarrec, A. Rowstron, M. Theimer, H. Wang, and A. Wolman, “An evaluation of scalable application-level multicast built using peer-to-peer overlays,” in Proceedings of Infocom’03, Apr. 2003, pp. 1510 – 1520.

[18] A. Rowstron, A.-M. Kermarrec, M. Castro, and P. Druschel, “Scribe: The design of a large-scale event notification infrastructure,” in Networked Group Communication, Third International COST264 Workshop (NGC’2001), ser. Lecture Notes in Computer Science, J. Crowcroft and M. Hofmann, Eds., vol. 2233, Nov. 2001, pp. 30–43.

[19] M. Castro, P. Druschel, A.-M. Kermarrec, and A. Rowstron, “Scribe: A large-scale and decentralized application-level multicast infrastructure,” IEEE Journal on Selected Areas in Communication (JSAC), vol. 20, no. 8, 2002, pp. 100–110.

[20] T. Iimura, H. Hazeyama, and Y. Kadobayashi, “Zoned federation of game servers: a peer-to-peer approach to scalable multi-player online games,” in Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, ser. NetGames ’04. New York, NY, USA: ACM, 2004, pp. 116–120.

[21] H.-H. Lee and C.-H. Sun, “Load-balancing for peer-to-peer networked virtual environment,” in Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games, ser. NetGames ’06. New York, NY, USA: ACM, 2006, p. Art.14.

[22] S. Yamamoto, Y. Murata, K. Yasumoto, and M. Ito, “A distributed event delivery method with load balancing for mmorpg,” in NETGAMES, 2005, pp. 1–8.

[23] J. C. de Oliveira, D. T. Ahmed, and S. Shirmohammadi, “Performance enhancement in mmogs using entity types,” in Proceedings of the 11th IEEE International Symposium on Distributed Simulation and Real-Time Applications, ser. DS-RT ’07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 25–30.

[24] A. Schmieg, M. Stieler, S. Jeckel, B. Kabus, Patric Kemme, and A. Buchmann, “psense - maintaining a dynamic localized peer-to-peer structure for position based multicast in games,” in Proceedings of the 8th International Conference on Peer-to-Peer Computing 2008 (P2P 2008), pp. 247–256.

[25] S.-Y. Hu, C. Wu, E. Buyukkaya, C.-H. Chien, T.-H. Lin, M. Abdallah, J.-R. Jiang, and K.-T. Chen, “A spatial publish subscribe overlay for massively multiuser virtual environments,” in International Conference on Electronics and Information Engineering (ICEIE), Aug. 2010, pp. 314–318.

[26] S.-Y. Hu and K.-T. Chen, “Vso: Self-organizing spatial publish subscribe,” in Proceedings of IEEE SASO, Oct 2011, pp. 21–30.

[27] S.-Y. Hu, J.-F. Chen, and T.-H. Chen, “Von: a scalable peer-to-peer network for virtual environments,” IEEE Network, vol. 20, no. 4, 2006 pp. 22–31.

[28] H. Backhaus and S. Krause, “Voronoi-based adaptive scalable transfer revisited: gain and loss of a voronoi-based peer-to-peer approach for mmog,” in Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games, ser. NetGames ’07. New York, NY, USA: ACM, 2007, pp. 49–54.

Synergic Effects among Plural Extensions of Breadcrumbs for Contents Oriented Networks

Tomohiko Yagyu
Cloud System Research Laboratories
NEC Corporation
Kawasaki-shi Kanagawa, Japan
Email: yagyu@cp.jp.nec.com

***Abstract*—Data traffic in mobile networks are rapidly growing because of the diffusion of smartphones and rich applications. It is effective to cache contents data in networks to reduce traffic load. Breadcrumbs was proposed to discover cached contents efficiently. Each router records trace information of content download, which is called Breadcrumbs, and redirects request messages along the Breadcrumbs trail. Furthermore, several extensions of Breadcrumbs, e.g., Active BC, Hop-aware BC, MSCR, BC Scoping, have been proposed. In this paper, we evaluate mutual effects among those extensions when they are working in the networks at the same time. Evalution results revealed that combination of ABC and MSCR has the synergic effects regarding improvement of cache hit ratio.**

***Keywords*-Contents Oriented Network; In-network Guide; Breadcrumb; Cache Discovery; Synergic effect**

I. Introduction

Because of the diffusion of broadband access and smartphones, it becomes ordinary to consume rich contents such as video through the networks. According to the Cisco's report [1], a smartphone generates 24 times more traffic than a non-smartphone. It also reports that traffic in mobile networks will grow 6.3EB per month by 2015. It estimates that video traffic will account for 66% of total traffic. Following the expansion of contents distribution and M2M (machine to machine) communication, the number and sort of contents in the networks will be enormous. In order to avoid congestion of network due to contents flood, it is effective to cache popular contents in the network [2]. Paul et al. proposed Cache and Forward architecture [3], which the contents are cached by routers in the networks. Furthermore, the concept of Data Centric Networking are proposed [4][5]. It is convenient for users to be able to get desired contents without knowing their locations but with using content IDs. For the cache networks, because of the complication to manage all the locations of cached contents, it is desirable to route request messages to the nearest caching node with content IDs through the networks. Breadcrumbs [6] is proposed to efficiently discover cached contents in the cache networks. With Breadcrumbs, each router guides request messages to the cached contents with in-network guide information called Breadcrumbs (BC). In order to solve various drawbacks of Breadcrumbs, several extensions are proposed, e.g., Active BC (ABC) [7], Hop-aware BC (HBC) [8], BC Scoping [9] and Mapping Server with Cache-location Resolution (MSCR) [10]. Because these extensions are not exclusive each other, they can be used together. When using some of them together, a certain synergy or conflict will be supposed. In this paper, mutual effects among several extentions are evaluted with the prototype system implementing Breadcrumbs and its extensions. Evalution results revealed that combination of ABC and MSCR has good synergic effect regarding the cache hit ratio. It is also observed that some extensions can compensate loss of performance each other. The remainder of this paper is organized as follows. Section II briefly explains the basic scheme of Breadcrumbs. Section III explains several extensions of Breadcrumbs. Section IV shows environments and assumptions for the evaluation. In Section V, evaluation results are discussed regarding performance metrics and network traffic. Finally, Section VI concludes this paper.

II. Cache Discovery with Breadcrumbs

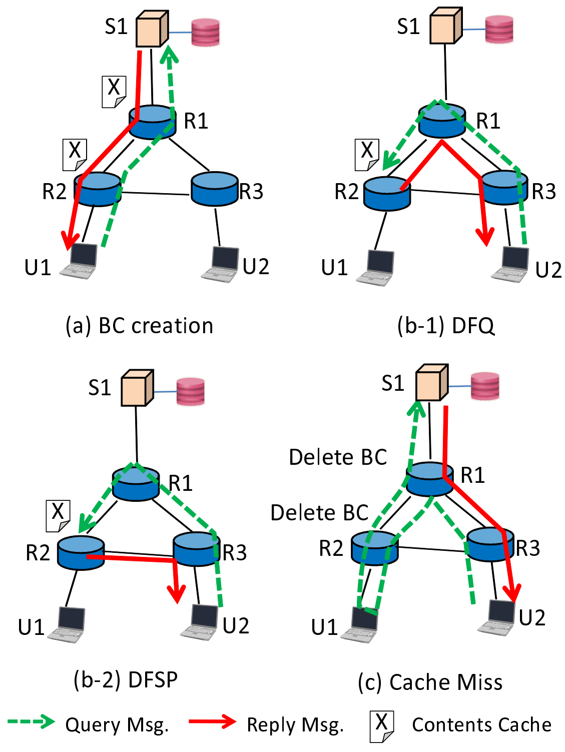
Breadcrumbs is used in the cache networks comprising cache-capable routers. Routers can cache the forwarded contents and record the trail information called Breadcrumbs along the download path. And they can guide messages to request contents (query) along the BC trail to the cached contents. Cached contents are replaced based on the Least Recently Used (LRU) policy when cache storage in a router is exhausted. A BC entry includes following 5 information. It is assumed that each content has global unique ID.

- Content ID
- ID of node from which the content arrived (Upstream Node)
- ID of node to which the content was forwarded (Downstream Node)
- Most recent time the content passed through the node (Download Time)
- Most recent time the content was requested at the node (Request Time)

Fig.1 shows an example of the cache network. The network consists of a contents server S1, three routers R1˜R3 and two user nodes U1˜U2. The basic bahavior of

Breadcrums is explained with Fig.1. First, user U1 requests content X. The query message sent by U1 reaches S1 via R2 and R1 (dotted arrow in Fig.1 (a)). The contents server S1 replies the data of content X to U1 via R1 and R2 (solid arrow in Fig.1 (a)). While forwarding the content X, router R1 and R2 cache the data and record the direction of the download in BC. Upstream and Downstream nodes of BC entries in R1 and R2 are (S1, R2), (R1, U1) respectively. A router also records Download time and Request time in BC entry.

Suppose that user U2 requests same content after R1 deleted it from cache storage. When the query sent by U2 reaches R1, R1 redirects the query to R2 because R1 has no cache but the valid BC entry for the content X (dotted arrow in Fig.1 (b-1) and (b-2)). BC entry becomes valid only if the Download time is within T_f or the Request time is within T_q. If the recorded time exceed such thresholds, BC entry expires. When the query reaches R2, the cache of content X in R2 is hit. Then R2 starts sending the data to U2. There are two options for download path to the requesting node. One is DFS (Download Follows Query), which forwards the contents along the reverse path of the query(solid arrow in Fig.1 (b-1)). Another one is DFSP (Download Follows Shortest Path), which forwards the contents through the shortest path from the contents holder to the requesting node(solid arrow in Fig.1 (b-2)). When the content is downloaded from R2 to U2, R1 and R2 overwrite their BC entries and R3 creates new BC entry. Upstream and Downstream node of BC entries in R2, R1 and R3 become (R1,R1), (R2,R3), (R1,U2) in DFQ case, and (R1,R3), (S1,R2), (R2,U2) in DFSP case respectively. If no caches exist along the BC trail, the query is forwarded to S1 (dotted arrow in Fig.1 (c)). When R1 receives the query from R2, R1 can recognize that no cache exists in the direction of R2 anymore. Therefore R1 deletes the BC entry for content X. While the content is downloaded from S1 to U2 (solid arrow in Fig.1 (c)), R1 creates BC entry (S1,R3) and R3 creates BC entry (R1,U2). R1 and R3 also create cache of the content X.



(a) BC creation

(b-1) DFQ

(b-2) DFSP

(c) Cache Miss

Figure 1. Basic behavior of Breadcrumbs

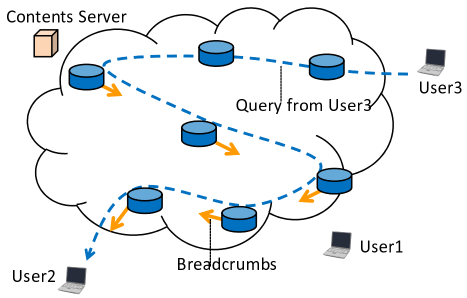


Figure 2. Lengthened BC Trail

III. Extensions of Breadcrumbs

A. Active Breadcrumbs

Active Breadcrumbs (ABC) is the extension of Breadcrumbs proposed in [7]. With Breadcrumbs, routers always overwrite BC entries to the newest direction of download. Therefore if a content is downloaded from a node having the cache, BC entries in its adjacent routers are overwritten to the direction of download users. For example in Fig.2, after User1 downloads a content from Server and User2 downloads it from User1, BC trail is formed like short arrows in Fig.2. Therefore the query sent by User3 travels a long trail along the dotted line. Even if the query passes close to the server and User1 which have desired content, it is redirected toward the further node User2. The objective of ABC is to avoid such ineffective redirection of queries. In addition, ABC can draw queries travling near the nodes which have the contents cache even if no adjacent routers have BC entries. As a result, ABC can improve cache hit ratio and reduce the load of contents server. In Fig.2, suppose that User1 distributes ABC to the adjacent router, the query from User3 is guided to User1. So User3 can download contents from User1. If User1 removed the contents from its cache storage, it also invalidates ABC for the removed contents. The range of distribution of ABC can be adaptively determined in terms of contents popularity, load of the node and so on.

ISBN: 978-1-61208-300-1

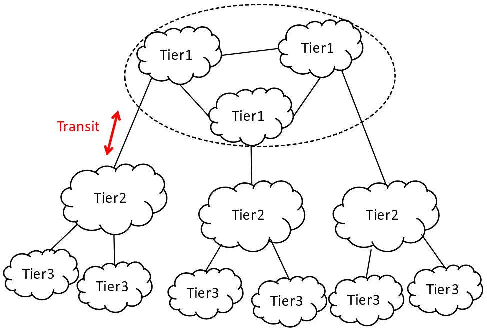


Figure 3. Hierarchy of ASes

B. Hop-aware Breadcrumbs

Hop-aware Breadcrumbs (HBC) is proposed in [8]. As same objective as ABC, HBC avoids redirection of queries near the server to further node. HBC aims to decrease delay of contents discovery. HBC extends BC entry to have the download node in addition to the downstream node of BC trail. A router compares hop count to the contents server and that to download node. If contents server is closer than the download node, the router ignores BC information and forwards the query to the server direction.

C. BC Scoping

BC Scoping is proposed in [9]. BC Scoping assumes a hierarchical structure of the network as shown in Fig.3. The Internet consists of a large number of Autonomous Systems (AS). ASes have vertical relations called Tier. Lower Tier ASes have transit links to connect upper Tier ASes. When lower ASes send or receive traffic through the transit links, they need to pay to upper ASes according to the amount of traffic. Transit cost per Mbyte in US was predicted $2.34 in 2012 [11]. Therefore it will be possible that lower ASes reduce their transit cost with utilizing contents cache in their own networks. In addition, when users can download desired contents within the belonging AS, users have benefits such as faster download and so on. Since original BC doesn't take hierarchy of the network into consideration, queries will be guided toward the different ASes via transit links. BC Scoping permits only routers which belong to the same ASes as the download user to create BC entries. This scheme restricts inter-AS redirection of queries and increases cache hit ratio within the same AS.

D. MSCR

Mapping Server with Cache-location Resolution (MSCR) is proposed in [10]. With Breadcrumbs, it is necessary for requesting users to know the location of contents server prior to sending a query. MSCR assumes Mapping Server (MS) which can provide server locations correspond to the content IDs. Users who want some contents inquire contents server information to MS. With MSCR, MS also keeps Prospective Cache Location (PCL) in addition to the server locations. PCL includes location of users who recently inquired the server location, in other words, locations where desired contents are potentially cached. When other user inquires server location of the contents to MS, MS replies PCL which is the closest to the user, too. The user puts the PCL information to the query message. When a router receives the query with PCL, the router forwards the query not toward the contents server but toward the location where the PCL indicates. If no caches exist on the way to the location, PCL information is removed from the query message and the query is forwarded toward the contents server. As a result, since the users can get the contents from closer location, network traffic can be reduced. For the prototype used in the evaluation, MS selects best PCL as the following priority order.

1) PCL having the same domain which the requesting user belongs to.
2) PCL having the location address which is longest-matched with requesting user's address.
3) PCL registered most recently.

E. Priority among extensions

Our prototype implemented above mentioned four extensions based on Breadcrumbs. Because the four extensions can be used simultaneously, routers may have several options for query forwarding. Following is the order of priority for routers to forward queries.

1) ***ABC***: If there exists ABC, the query is forwarded by ABC.
2) ***HBC***: If valid HBC exists AND contents server is closer than download node in HBC, query is forwarded toward the server. If server is not closer, the query is forwarded to the downstream node of HBC.
3) ***MSCR***: If PCL exists in the query, the query is forwarded toward the location indicated by PCL.
4) Otherwise the query is forwarded toward the contents server.

IV. Evaluation Environment

Breadcrumb-based contents distribution system was developed on Linux with C language. The system implements basic Breadcrumbs and four extensions, i.e., ABC, HBC, BC Scoping and MSCR. The structure of the system is shown in Fig.4. Every router, user node and contents server runs the contents distribution module. The system also has a mapping server which provides location (i.e., IP address) of cotents servers to users. All modules are implemented as user level processes. All messages among processes are defined in XML.

In this evaluation, the performance metrics and network traffic were measured when plural extensions work at the same time. 14 patterns in which more than one extensions are used were compared with basic BC case. Measured

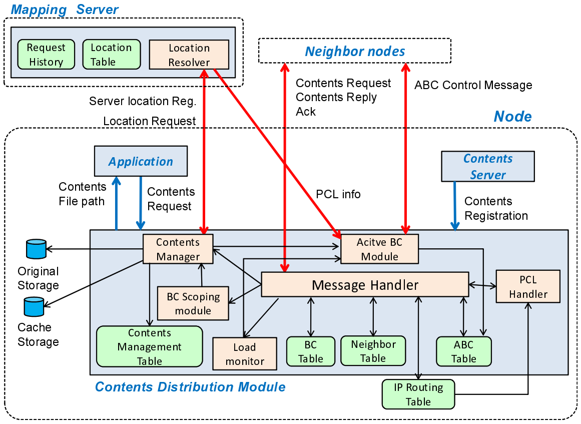


Figure 4. Structure of breadcrumbs system

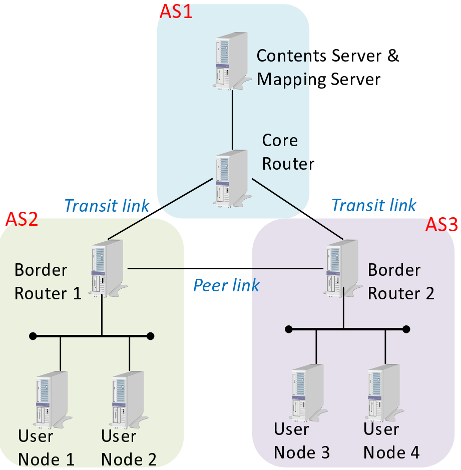
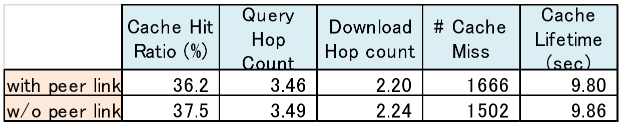


Figure 5. Evaluation Topology

Table I
PARAMETERS

Param	Value
Router Cache Size	10
User Cache Size	10
Download policy	DFSP
Cache Replace Policy	LRU
T_f	1800sec
T_q	600sec
Total Eval. time	1000sec

Table II
RESULTS WITH BASIC BC



	Cache Hit Ratio (%)	Query Hop Count	Download Hop count	# Cache Miss	Cache Lifetime (sec)
with peer link	36.2	3.46	2.20	1666	9.80
w/o peer link	37.5	3.49	2.24	1502	9.86

metrics are cache hit ratio, query hop count, download hop count, number of cache miss and average cache lifetime.

A topology used for the evaluation is shown in Fig.5. The network consists of one contents server (CS), which has the role of both mapping server and contents server, one core router (CR), two border routers (BR1˜2) and four user nodes (UN1˜4). There are three Autonomous Systems (AS) in the network. It is assumed that AS1 is upper AS and AS2 and AS3 are lower ASes. Two topologies with and without a peer link between AS2 and AS3 were evaluated. All network interfaces equiped in the nodes are FastEther (100Mbps).

Only routers can forward queries with BC. User nodes do not create any BC entries. However user nodes can distribute ABC when contents in its cache are hit. The range of ABC distribution depends on the hit count of cached contents. Because hit count is higher, the content is supposed to be more popular. The pair of (Cache Hit Count, ABC TTL) are set (1,1), (5,2), (10,3) in this evaluation. Contents data are downloaded along the shortest path. Every router and user node have cache storage, which size is as 10 contents. Total number of contents is 1000 and size of a content is 100Kbytes. The popularity of k^{th} content $p(k)$ is calculated in eq.(1) as followed in Zipf distribution. In this experiment, $\alpha = 1.0$ was used.

$$p(k) = \frac{1/k^{\alpha}}{\sum_{n=1}^{1000} 1/n^{\alpha}} \qquad (1)$$

All user nodes send one query per second. Requested contents are chosen with probability in proportional to $p(k)$. Becasue one trial lasts 1000 seconds, the number of total reqeusts from all user nodes is 4000. Three trials were performed and results are averaged. Parameters configured for the evaluation are shown in Table I.

V. EVALUATION RESULTS

A. Performance of Contents Distribution

The results in the case using only basic BC are shown in Table II. The results between topologies with and without peer link have no big differences. Cache hit ratio is slightly better than 36%. Query hop count is about 3.5, a little longer than the hop count from user node to content server (3 hops). However, download hop count is around 2.2, shorter than 3. It means that users can get the contents from closer cache. Cache miss events happen more than 1500 times out of 4000 requests. Each content cache can survive less than 10 secounds on average.

Compared with this basic BC results, Table III and IV show increase or decrease percentage of the results for combinations of extensions. The results in the topology with peer link are shown in Table III, and those in the topology without peer link are shown in Table IV. Shaded cells mean degradation from basic BC results. Bold figures mean the best results among the combinations. Same results regarding each metric are also shown in Fig.6 ˜10. Circled bars in the figures mean the best results.

Table III
INCREASE/DECREASE % IN TOPOLOGY WITH PEER LINK

	Cache Hit Ratio	Query Hop Count	Download Hop count	# Cache Miss	Cache Lifetime
ABC	4.93	−2.24	−2.05	−4.90	3.37
BC Scope	−14.07	−22.92	−0.34	−42.37	2.14
MSCR	2.79	−3.11	−1.25	0.68	1.33
HBC	−14.46	−22.56	−0.11	−41.69	1.43
ABC +BC Scp	−9.33	−23.43	−1.14	−43.91	2.24
ABC +MSCR	**6.34**	−5.28	**−2.61**	−5.42	**4.49**
ABC +HBC	−4.37	**−24.15**	−2.05	**−45.35**	3.88
BC Scp +MSCR	−14.08	−14.90	0.23	−24.66	1.12
BC Scp +HBC	−13.84	−22.56	−0.45	−40.41	1.33
MSCR +HBC	−11.97	−14.68	−0.45	−23.96	1.53
ABC +BC Scp +MSCR	−5.23	−16.70	−1.70	−29.15	2.65
ABC +BC Scp +HBC	−9.07	−23.43	−0.91	−44.51	2.14
ABC +MSCR +HBC	−3.45	−15.33	−1.36	−27.47	2.86
BC Scp +MSCR +HBC	−11.04	−15.69	−1.14	−24.97	1.94
All	−4.72	−16.92	−1.93	−28.75	3.37

Table IV
INCREASE/DECREASE % IN TOPOLOGY W/O PEER LINK

	Cache Hit ratio	Query Hop count	Download Hop count	# Cache Miss	Cache Lifetime
ABC	2.29	−0.29	0.00	−2.04	1.72
BC Scope	−13.96	−24.30	−3.57	−35.41	1.42
MSCR	−1.55	1.79	0.67	1.04	−1.01
HBC	−17.56	−23.37	−2.23	−35.21	1.12
ABC +BC Scp	−11.17	−24.44	−2.90	−39.20	2.43
ABC +MSCR	**2.84**	0.07	−1.00	−2.15	2.33
ABC +HBC	−12.27	**−24.59**	−2.57	−41.31	1.83
BC Scp +MSCR	−15.27	−12.26	−1.90	−42.49	0.81
BC Scp +HBC	−18.76	−22.80	−1.45	−33.70	0.51
MSCR +HBC	−18.04	−16.20	−1.79	−35.52	0.81
ABC +BC Scp +MSCR	−9.85	−13.12	−2.34	**−45.64**	1.83
ABC +BC Scp +HBC	−10.52	−24.01	−2.90	−37.36	2.33
ABC +MSCR +HBC	−10.92	−17.49	−2.79	−40.02	1.93
BC Scp +MSCR +HBC	−13.28	−18.06	**−3.79**	−36.85	2.94
All	−8.04	−17.99	**−3.79**	−39.56	**3.45**

1) cache hit ratio: The result of cache hit ratio in both topologies are shown in Fig.6. As shown in Fig. 6, all combinations of extensions do not significantly improve cache hit ratio. This is because the topolgy is almost tree structure, the core router can surely redirect queries only with basic BC. Since MSCR can guide queries to the region in which ABC is locally distributed, the combination of ABC and MSCR can improve cache hit ratio best. It can be said that MSCR can enhance the benefit of ABC. Since the network is samll in this evaluation, synergy of ABC and MSCR is limited. The benefit of synergy between ABC and MSCR will be more in larger networks. On the other hand, BC Scoping and HBC worsen cache hit ratio compared with basic BC. This is because both extensions will restrict query redirection at the core router, so more queries tend to go to the contents server. When ABC is used with HBC in the topology with peer link, degradation of cache hit ratio can be minimized with keeping shorter query and download hop count. This is because ABC can redirect queries between BRs directly via peer link only if cache in another BR is available.

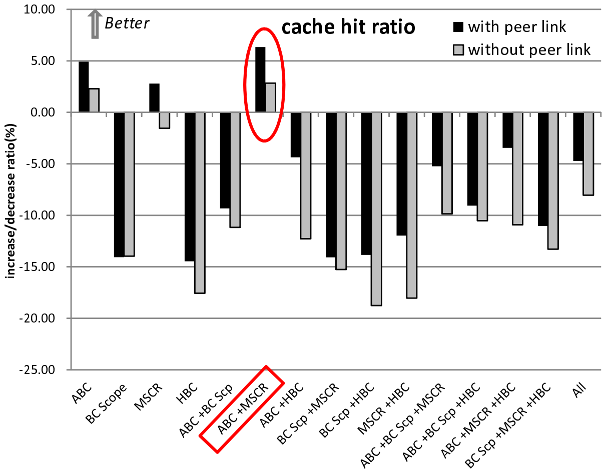


Figure 6. Increase/decrease of Cache Hit Ratio

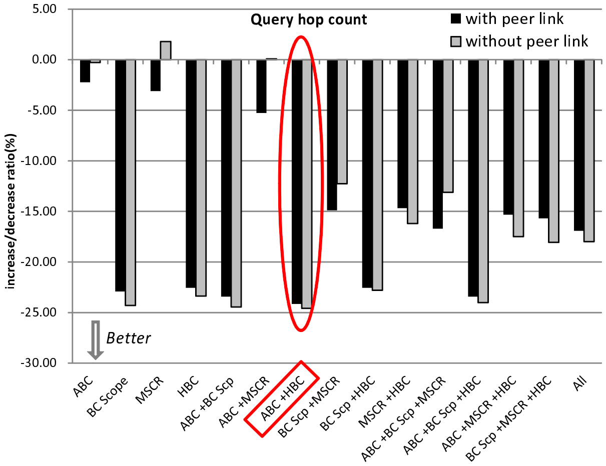


Figure 7. Increase/decrease of Query Hop Count

2) query hop count: Query hop count is related to the time to find contents. Shorter the query hop is, faster the user can find contents. According to Fig.7, thanks to the restriction of redirection at CR, HBC and BC Scoping can shorten query hop count by more than 20%. With using only ABC or MSCR individually, reduction of query hop count is limited. This is because the server is as close as cache locations in BRs and UNs in the evaluation topology. However, combination of ABC and HBC/BC Scoping can enhance query hop reduction. This is because queries for only popular contents are redirected between BRs by ABC, but queries for less popular contents are not redirected from CR to BRs due to BC Scoping or HBC. Therefore combination of ABC and HBC/BC Scoping is good choice to reduce query hop count with less degradation of cache hit ratio.

3) download hop count: Download hop count is related to both time to download contents and traffic load in the

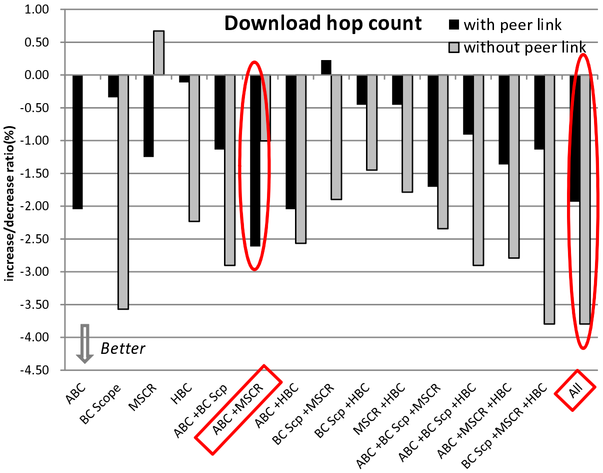


Figure 8. Increase/decrease of Download Hop Count

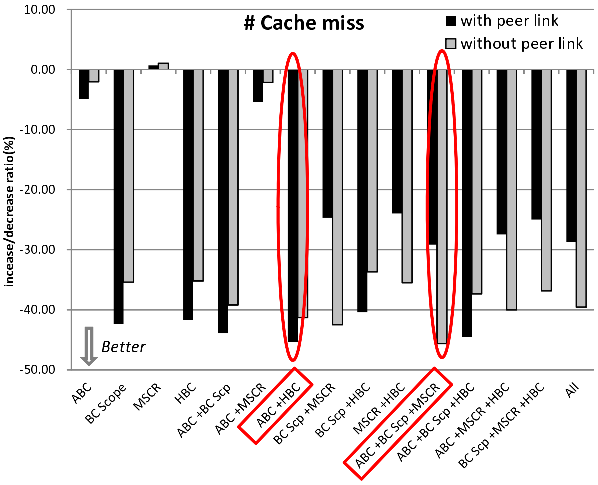


Figure 9. Increase/decrease of number of cache miss

network. Shorter the download hop is, faster the user can download contents, also lighter the traffic load in the network becomes. The results of increase or decrease of download hops are shown in Fig.8. In case of the topology with peer link, the combination of ABC and MSCR is best. ABC only and ABC+HBC also marked good scores. It means that ABC contributes downloading contents cache from BRs via peer link. Furthermore, MSCR and ABC encourage download between UNs in the same AS. Suppose that UN1 has a content cache and distribute ABC to BR1. Even if BR1 has no BC pointing to UN1, ABC in BR1 can guide queries from UN2 to the cache-holding UN1. MSCR can also guide queries directly to the neighbor UN in the same AS based on the PCL given by MS. Hop count between UN1 and UN2 in the same AS is only 1 hop. Therefore combination of ABC and MSCR can reduce more download hop count than other combinations in this topology. Unlike the case with peer link, using all extensions can reduce download hop most in the topology without peer link. When a UN downloads a content from other UN in the different AS, hop count will be 4. Since hop count from server to UN is 3, download between UN in different ASes lengthen the download hop count. HBC and BC Scoping can avoid such longer download path by restricting queries to go to other AS. On the one hand, ABC and MSCR help download between UNs in the same AS. This synergy causes the best result in terms of download hop reduction.

4) cache miss frequency: According to Table II, cache miss events happen more than 1500 times for total 4000 queries. It means that cache storage size in routers and user nodes are so small while BCs survive much longer than cache lifetime. According to Fig.9, HBC and BC Scoping can reduce cache miss event by around 40% because they restrict query redirection from CR to BRs with BC. Owing to the decrease of cache miss, useless forwarding of queries are suppressed. As the result, total messages processed in the network can be reduced by 7%. ABC reduces cache miss because ABC can guide queries to surely existing cache. When the cache in the node is removed, ABC distributed to the adjacent nodes are also purged. Therefore ABC does not worsen cache miss situation. Regarding MSCR, since PCL in the MS is not related to the removal of cached contents, MSCR is more prone to cache miss than ABC. In the topology with peer link, ABC+HBC is the best combination to minimize cache miss. This is because ABC can surely find cached contents via peer link, and HBC prevents unsure redirection of query at CR. On the other hand, in the topology without peer link, ABC+BC Scoping+MSCR is the most effective. With MSCR, MS possibly gives PCL in same AS as the requesting user. Therefore if neighbor UN in the same AS doesn't have the cached contents, a query merely goes to BR. Since BC Scoping prevents CR from creating BC, CR forwards queries to another AS with ABC only if BR or UNs in another AS surely have cached contents.

5) cache lifetime: Because average cache lifetime shown in Fig.10 is around 10 seconds, BC lifetime (T_f) should be similar value as the cache lifetime. If BC lifetime is too long, cache miss will happen more frequently. Conversely BC lifetime is shorter than cache lifetime, cache hit ratio must be degraded. However we cannot know the average cache lifetime in advance and it will be changeable due to many factors such as popularity of the contents, user distribution, network topology, router capacity and so on. Therefore how to decide optimal BC lifetime is still open issue. Since ABC is distributed when the cache is hit in the node, ABC will guide queries for popular contents and cache of the popular contents can be used more frequently. As the result, cache lifetime become longer when using ABC. In the topology with peer link, the combination of ABC and MSCR can prolong the cache lifetime most. While in the

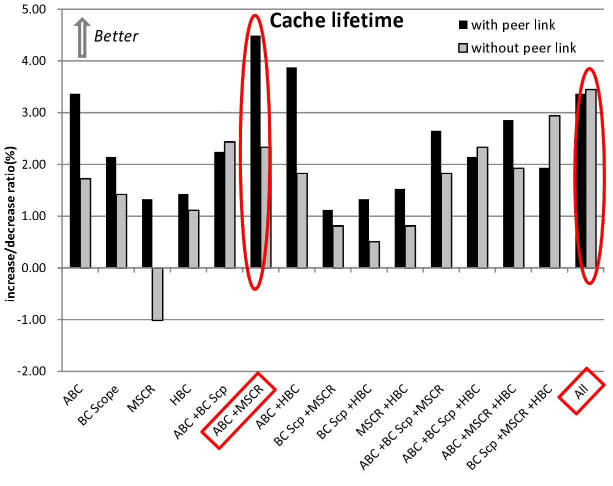


Figure 10. Increase/decrease of average cache lifetime

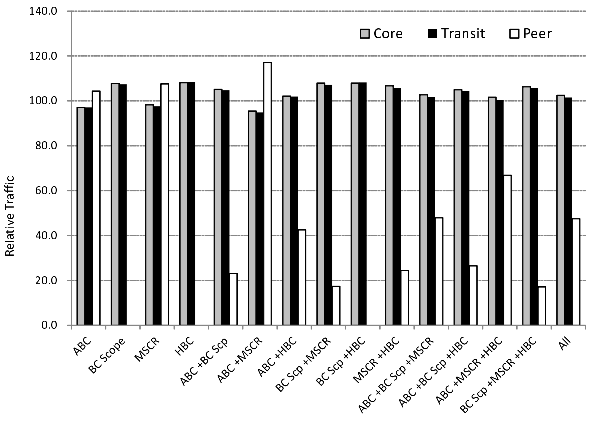


Figure 11. Relative traffic in topology with a peer link

Table V
NETWORK TRAFFIC WITH BASIC BC (MBYTES)

	Core	Transit	Peer	Intra	Total
with peer link	395.9	410.9	38.9	519.7	1365.4
w/o peer link	386.8	485.6	N/A	517.0	1389.5

topology without peer link, the lifetime is the longest when using all extensions. Although cache hit ratio is degraded in most combinations, cache lifetime becomes longer. Main reason is that the cache in the BRs can survive longer. In the topology without peer link, the cache lifetime in UNs tend to be shorter. While in the topology with peer link, that in the CR becomes shorter. However in both topologies, the cache lifetime in BRs becomes longer by 6˜10%. If cache lifetime becomes longer, it is expected that load to replace cache in routers can be reduced.

B. Network traffic

Total amount of network traffic (Mbytes) are shown in Table V when using basic BC individually. $Core$ means traffic on the link between CR and contents server. $Transit$ means traffic on the transit links between CR and BRs. $Peer$ means traffic on the peer link between BRs. $Intra$ means traffic within AS2 and AS3. $Total$ is the traffic throughout the whole network. Total traffic increases in case of no peer link. This is becasue download across ASes travels two transit links instead of the peer link.

For all combinations, relative amount of traffic to basic BC results as 100 are shown in Fig.11 and Fig.12.

1) The topology with peer link: From the viewpoint of operators in lower Tier ASes, it is desirable that users (their customers) can download contents within the AS or can download from other ASes connected with a peer link. According to Fig.11, it reveals that HBC can not utilize the peer link at all. This result depends on the topology. Since the contents server is closer than UNs from CR, HBC always selects to forward queries toward the server in CR. Since BC Scoping also prevents inter-AS download, it doesn't utilize peer link at all, too. In terms of the operation cost, ABC+MSCR is the best combination becasue it can reduce the transit traffic most. This combination can also reduce total traffic most becasue download hop count is minimized by exploiting the peer link.

2) The topology without peer link: As shown in Fig.12, when using HBC or BC Scoping, more transit traffic can be reduced. This is because inter-AS download is restricted. Instead, core traffic increases becasue more users download the contents from the server. BC Scoping+MSCR+HBC is the best combination for the operators in lower ASes because it can reduce tansit traffic most. This combination is also the best from global point of view becasue it can minimize total traffic. From viewpoint of core network operator, ABC+MSCR is the best to reduce the most core network traffic. According to this observation, there is contradiction between the benefits of core network operator and lower AS operators. Core network operators have no incentive to use HBC and BC Scoping, because they want to minimize traffic in core network and maximize transit cost. Lower AS operators want core network to use BC Scoping and HBC to minimize transit cost. If each operator can configure which extensions are used in thier routers independently, the benefit will be reduced. How to compromise the benefit of operators in different ASes and find best mix of extensions for all operators is the essential but difficult issue.

VI. CONCLUSION

In this paper, synergic effects among multiple extensions of Breadcrumbs, which is the distributed mechanism to discover cached contents in the networks were evaluated.

According to the evaluation results, ABC and MSCR can improve cache hit ratio most regardless the existence of peer links. In the topology with peer links, ABC and MSCR

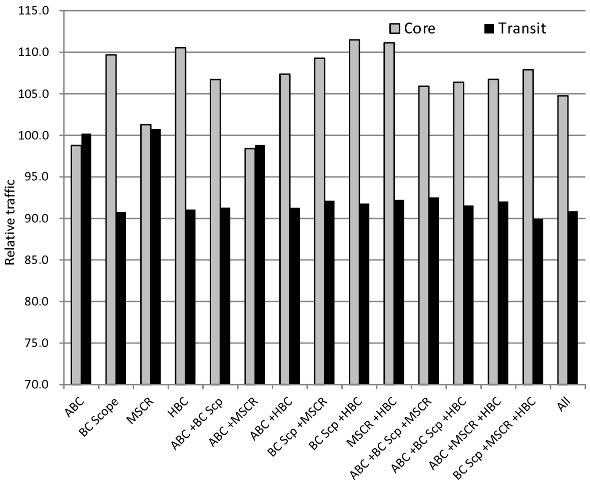


Figure 12. Relative traffic in topology without peer link

is also the best combination for both performance metrics and network traffic. If it is important to reduce control messages, ABC and HBC is the best combination. In the topology without peer links, using all extensions is the best choice to improve performance metrics. From the viewpoint of network traffic, BC Scoping, HBC and MSCR is good combination for lower AS operators.

Because these results fairly depend on the topologies used in the evaluation, evaluation with larger networks will be performed near future. It is also necessary to consider real network scenario for applications of Breadcrumbs system. Automatic BC lifetime adaptation scheme will be another challenge to optimize performance of cache discovery with BC.

Acknowledgment

This work was supported by National institute of Information and Communications Technology (NICT) Japan.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015, Feb. 2011.

[2] G.Tyson, S.Kaune, S.Miles, Y.El-khatib, A.Mauthe, and A.Taweel, ”A Trace-Driven Analysis of Caching in Content-Centric Networks”, Proc. ICCCN’12, July 2012, pp.1-7.

[3] S.Paul, R.Yates, D.Raychaudhuri, and J.Kurose, ”The cache-and-forward network architecture for efficient mobile content delivery service in the future internet,” First ITU-T Kaleidoscope Academic Conference, May 2008, pp.367-374.

[4] T.Koponen, M.Chawla, B-G.Chun, A.Ermolinskiy, K.H.Kim, S.Schenker, and I.Stoica, ”A Data-Oriented (and Beyond) Network Architecture,” in Proc. ACM SIGCOMM 2007, Aug. 2007, pp.181-192.

[5] V.Jacobson, D.Smetters, J.Thorton, M.Plass, N.Briggs, and R.Braynard, ”Networking Named Content,” in Proc. ACM CoNEXT 2009, Dec. 2009, pp.1-12.

[6] E.J.Rosenweig and J.Kurose, ”Breadcrumbs: efficient, best-effort content location in cache networks,” in Proc. IEEE INFOCOM 2009, Apr. 2009, pp.2631-2635.

[7] M.Kakida, Y.Tanigawa, and H.Tode, ” Active Breadcrumbs: Aggressive Distribution Method of In-network Guidance Information for Content-Oriented Networks, ” in Proc. LCN2012, Oct. 2012, pp. 184-187.

[8] K.Hashimoto, Y.Takaki, C.Ohta, and H.Tamaki, “In-network Hop-aware Query Induction Scheme for Implicit Coordinated Content Caching,” in Proc. AFIN2011, Aug. 2011, pp. 69-73.

[9] M.Kakida, Y.Tanigawa, and H.Tode, ” Distribution Method of In-network Guidance Information for Inter-AS Content-Oriented Network Topology, “ World Telecomm. Congress (WTC) 2012, Mar. 2012, Poster session.

[10] H.Kawabata, K.Hashimoto, T.Inamoto, Y.Takaki, C.Ohta, and H.Tamaki, “Content/Location Mapping with Cache-Location Resolution for In-network Guidance”, Proc. AFIN2012, Aug. 2012, pp.1-6.

[11] W.B.Norton, ”Internet transit prices - historical and projected,” tech. rep., Dr Peering International, 2010.

Competitive Algorithms for Online Data Placement on Uncapacitated Uniform Network

Maciej Drwal
Institute of Computer Science
Wroclaw University of Technology
Wroclaw, Poland
e-mail: maciej.drwal@pwr.wroc.pl

***Abstract*—In this paper, we study the iterated problem of placing copies of data objects in a network of storage servers in order to serve request demands with minimal delay. We show how to compute the optimal sequence of decisions for two variants: in general, using dynamic programming algorithm, which requires exponential time, and for uncapacitated uniform network, which requires only polynomial time. For the latter case, we study online algorithms, which return new placement immediately after a new element of input sequence becomes available. We prove that the comeptitive ratio of the problem is bounded by 2. The paper is summarized with computational study, which compares the 2-competitive online algorithm with dynamic programming. Online distributed storage management becomes increasingly important issue in large-scale Internet applications due to the widespread of Content Devlivery Networks, as well as cloud computing and content-aware paradigms.**

***Keywords*—online algorithms; network algorithms; wide-area networks.**

I. Introduction

Optimization of data placement is employed by Content Delivery Networks (CDNs) in order to improve the efficiency of media content distribution [6]. Placing replicated objects in multiple storage servers on behalf of a content owners allows to reduce network congestion and balance processing load on servers. The operator of Content Delivery Network needs to apply appropriate placement algorithms and client redirection methods in order to provide data access services with high performance.

In many practical applications of optimization and control, decisions have to be made immediately after a new piece of data becomes available, without knowledge of the future data. In such circumstances, each single decision from the sequence influences the overall quality of the solution assessed within a longer time horizon. This is especially common in the real time systems where a sequence of actions is uninterruptible and additionally execution time constraints are imposed. Optimization and decision problems with such characteristics are customarily called *online problems*, and the solution methodology falls under the area of *online algorithms* [1], [9].

In this paper, the data placement problem is considered from the perspective of continuous system operation under the stream of requests. Let us consider discrete time intervals, starting at time instants $t_1, t_2, \dots, t_T$. At each of those time instants it is possible to change the placement of objects. It is assumed that time is perfectly synchronized at all network nodes.

The considered system consists of N nodes, where each node is associated with local area network and a storage server. Each ith local area network ($i = 1, \dots, N$) is characterized by request demands w_{ip} for each pth data object, $p = 1, \dots, M$. It is assumed that each node can access any object from a server which holds a copy of it. In particular, if object is stored locally, then no external transmission is required. However, storing object at any jth server requires a fixed cost b_{jp}, which represents a delay needed to fetch the object p from its publisher and install it in the storage.

The clients' demands for each of the considered T time periods are given as a sequence: $\{\mathbf{w}(t)\}_{t=1}^{T}$, where $\mathbf{w}(t) = [\mathbf{w}_1(t) \ \dots \ \mathbf{w}_M(t)]^T$, $\mathbf{w}_p(t) = [w_{1p}(t) \ \dots \ w_{Np}(t)]^T$, and $w_{ip}(t)$ is the mean number of requests for pth object sent from ith client in time interval $[t, t+1)$. All other system's parameters are fixed. The model consists of the following parameters: d_{ij} is the transmission delay of unit of data between nodes i and j ($d_{ii} = 0$ for each i); h_j is the time required to process a single unit of data on server j; s_p is the size of data object p (in the assumed units); S_j is the maximal number of simultaneous connections that server j can handle; R_j is the storage capacity of server j (in the assumed units).

It is assumed that there are single copies of each data object in the network (for convenience, no origin server is considered; instead, any server j may play the role of publisher, for example having fixed zero demand). In this paper, placement decision $\mathbf{z}$ is a three-dimensional matrix, indexed by time, which is interpreted as follows for $t = 1, \dots, T$:

$$z_{ijp}(t) = \begin{cases} 1 & \text{if } p\text{th object is copied from } i\text{th node to } j\text{th} \\ & \text{node at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{N} = \{1, \dots, N\}$. The value of $\mathbf{z}(0)$ is given, as it defines the initial placement, as: $\forall_{i \in \mathcal{N}} \ z_{ijp}(0) = 1$ if object p is initially placed at server j, and zero otherwise. Similarly:

$$x_{ijp}(t) = \begin{cases} 1 & \text{if } i\text{th node is assigned to } j\text{th node for} \\ & \text{accessing object } p \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The optimization variables are grouped in two sequences $\{\mathbf{x}(t)\}_{t=1}^{T}$ and $\{\mathbf{z}(t)\}_{t=1}^{T}$. The objective is to determine the sequence of placement and assignment decisions, so as to minimize the following sum of costs:

$$Q_T(\mathbf{x},\mathbf{z}) = \sum_{t=1}^{T}\sum_{p=1}^{M}\left(\sum_{i=1}^{N} x_{ijp}(t)w_k(t)d_{ij}s_p + \right.$$
$$+ \sum_{j=1}^{N} h_j x_{ijp}(t)\left(\sum_{k=1}^{N}\sum_{q=1}^{M} x_{ijp}(t)w_k(t)s_q\right) + \tag{1}$$
$$\left. + \beta_p \sum_{j=1}^{N} z_{jip}(t)d_{ji}s_p\right),$$

subject to:

$$\forall_t\ \forall_{i\in\mathcal{N}}\ \forall_{p\in\mathcal{M}} \quad \sum_{j=1}^{N} x_{ijp}(t) = 1, \tag{2}$$

$$\forall_t\ \forall_{i,k\in\mathcal{N}}\ \forall_{j\in\mathcal{N}}\ \forall_{p\in\mathcal{M}} \quad x_{ijp}(t) \le z_{kjp}(t), \tag{3}$$

$$\forall_t\ \forall_{i\in\mathcal{N}}\ \forall_{j\in\mathcal{N}} \quad \sum_{i=1}^{N}\sum_{p=1}^{M} x_{ijp}w_{ip}(t) \le S_j, \tag{4}$$

$$\forall_t\ \forall_{i\in\mathcal{N}}\ \forall_{j\in\mathcal{N}} \quad \sum_{p=1}^{M} z_{ijp}s_p \le R_j. \tag{5}$$

The set of data object is denoted by $\mathcal{M} = \{1,\ldots,M\}$. The parameter $\beta_p > 0$ is a replication cost factor, which allows for differentiating the transmission cost of client's data access and the cost of data replication into a new server. It represents an additional overhead on a cache server associated with first-time installation of object transmitted from a publisher.

A problem for which the whole input data sequence $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_T)$ is given before the decision is made is called *offline* problem (a solution algorithm has access to the whole input sequence) [9]. Typically, optimization and decision problems are examined in this fashion. For example, formulation (1)–(5), where $\sigma_t = \mathbf{w}(t)$, can be considered an offline problem, if it is assumed that all its parameters are known in advance. The problem for which we require an immediate decision for each element of the input sequence is called *online* problem, i.e., solution algorithm has access only to the first k elements, for some $k \le T$. Usually, such algorithm is executed T times, after each element σ_k of input data sequence becomes available.

We can observe that in these settings the information about the problem's input itself is the most scarce resource. The lack of this information causes fundamentally different difficulties than the computational complexity of the problem. Even if we allow unlimited computational resources (in terms of time and memory) for solving the problem for each σ_k, it is not guaranteed that we get the optimal solution for the whole sequence σ.

The paper is organized as follows. Section II lists the related works. Section III provides an exact solution algorithm for the formulated problem, while Section IV studies a special case of uniform network. Main results concerning online algorithms and their analysis are provided in Section V. An experimental study is presented in Section VI, and finally, Section VII concludes the paper.

II. Related work

Data replication has been studied extensively in the last decade, motivated by the widespread of large scale Internet applications. This research includes aspects of caching, routing, requests redirection and storage management [3], [7], [10], [12], [14], [16]. Underlying mathematical models, based on the *facility location problems* [11], have been identified to be hard to solve exactly and also hard to approximate efficiently [6]. Due to the dynamic nature of users' activity, the online approach has been proposed. Earliest works on distributed memory management can be found in [8], which initiated online analysis in the area of computer systems performance evaluation. Subsequent results have been obtained on distributed paging [4], file allocation in network [2], data replication and migration [5], [15] and distributed database management [17].

The online algorithms analysis allowed to characterize the performance of distributed systems in terms of competitive ratio, which measures the efficiency of algorithm operating on partial input data. For the introductory material on this subject, we refer to [1], [9]. In this paper, a similar analysis is provided for a special formulation of data placement problem which includes time-varying users' demands.

III. Exact algorithm for general case

As it can be easily seen, computing optimal solution of the problem for any single t, can result in bad initial configuration for $t+1$, which eventually leads to suboptimal solution of (1). Thus, it is necessary to take into consideration the whole input data sequence, and connect the partial solutions appropriately. An obvious approach is to use dynamic programming algorithm.

Let the pair of decision matrices $\hat{\mathbf{x}}(t) = [\mathbf{x}(t), \mathbf{z}(t)]^T$ for each t be called *configuration*. Let:

$$g(\hat{\mathbf{x}}(t)) = \sum_{p=1}^{M}\left(\sum_{i=1}^{N} x_{ijp}(t)w_k(t)d_{ij}s_p + \right.$$
$$+ \sum_{j=1}^{N} h_j x_{ijp}(t)\left(\sum_{k=1}^{N}\sum_{q=1}^{M} x_{ijp}(t)w_k(t)s_q\right)$$
$$\left. + \beta_p \sum_{j=1}^{N} z_{jip}(t)d_{ji}s_p\right). \tag{6}$$

Let $V_{T-t}(\hat{\mathbf{x}}(t))$ be the optimal cost after $T-t$ iterations, assuming that the tth configuration is $\hat{\mathbf{x}}(t)$. We define the following Bellman equation for $t = 0, 1, \ldots, T-1$:

$$V_{T-t}(\hat{\mathbf{x}}(t)) = \min_{[\hat{\mathbf{x}}(t),\hat{\mathbf{x}}(t+1),\ldots,\hat{\mathbf{x}}(T)]} \sum_{s=t}^{T} g(\hat{\mathbf{x}}(s)) =$$
$$= \min\left\{g(\hat{\mathbf{x}}(t)) + V_{T-t-1}(\hat{\mathbf{x}}(t-1))\right\}, \tag{7}$$

with initial condition $V_0(\hat{\mathbf{x}}(T)) = 0$ for any final configuration $\hat{\mathbf{x}}(T)$.

From the above equation we get the recursive formula for computing the optimum of (1) as $Q_T(\mathbf{x}^*, \mathbf{z}^*) = V_T(\hat{\mathbf{x}}(0))$, where $\hat{\mathbf{x}}(0)$ is a given initial configuration. This gives:

$$V_T(\hat{\mathbf{x}}(0)) = \min\{g(\hat{\mathbf{x}}(0)) + V_{T-1}(\hat{\mathbf{x}}(1))\}, \tag{8}$$

which states, that in order to compute the total optimum we can decompose the problem into solving for fixed configuration (i.e., starting from the initial state) and solving analogous problem with one step shorter time horizon. Subsequently we get:

$$V_{T-1}(\hat{\mathbf{x}}(1)) = \min\{g(\hat{\mathbf{x}}(1)) + V_{T-2}(\hat{\mathbf{x}}(2))\}, \tag{9}$$

and continue this until we reach the time horizon $t = T$.

In order to solve this, we start the backward induction by computing:

$$\begin{aligned} V_{T-(T-1)}(\hat{\mathbf{x}}(T-1)) &= \min\{g(\hat{\mathbf{x}}(T-1)) + V_0(\hat{\mathbf{x}}(T))\} \\ &= \min g(\hat{\mathbf{x}}(T-1)), \end{aligned} \tag{10}$$

that is, solving the static case problem under assumption that the current placement configuration is $\hat{\mathbf{x}}(T-1)$. In result, we obtain the final configuration $\hat{\mathbf{x}}(T)$. But, in order to determine the previous configuration $\hat{\mathbf{x}}(T-1)$, we again need to solve the static case problem, this time assuming that the current placement configuration is $\hat{\mathbf{x}}(T-2)$. Continuing this reasoning, we reach initial state $\hat{\mathbf{x}}(0)$.

Unfortunately, there is no simple method to find an optimal solution, as at each stage we need to solve an NP-hard problem of minimizing $g(\hat{\mathbf{x}}(t))$, which can be seen to be at least as hard as solving facility location problem [11]. Assuming that the complexity of this problem is $O(f(N, M))$ the overall complexity of the dynamic algorithm is $O(f(N, M)^T)$. For example, assuming that $h_j = 0$ for all j, the exhaustive search over all placements requires $O(2^{NT})$ time. Nevertheless, the dynamic programming method gives a hypothetical algorithm allowing to compute the optimal solution, upper-bounding the time complexity exponentially in the length of time horizon. We make use of this algorithm for the special case of uncapacitated online data placement in uniform network in Section V (see pseudocode of Algorithm 1), in order to compare it with fast approximation algorithm.

IV. POLYNOMIAL TIME ALGORITHM FOR A CASE OF UNIFORM NETWORK

In Section III, we concluded that the fully general case of iterated data placement problem is hard to solve. In this section, we examine a simplified variant, which still bears practical importance, but for which solution can be computed quickly (both in terms of network size and time horizon length).

A. Formulation of special case problem

Let us define the following case of iterated data placement problem. There are no capacity constraints (4)–(5) imposed on servers (thus we consider the placement of a single object), and the distances in network are uniform, i.e., $d_{ij} = 1$ for all $i \neq j$. Many practical instances of wide-area networks can be considered uniform, since all their edge routers have identical (usually very high) capacity. Moreover, we assume that servers have very high processing speeds, thus $h_j = 0$ for all j.

Given are fully connected graph consisting of N vertices, an initial placement of object $\mathbf{z}(0)$, and a sequence of requests $\sigma = \{\mathbf{w}(t)\}_{t=1}^T$, where $\mathbf{w}(t) = [w_1(t), w_2(t), \ldots, w_N(t)]^T$, and $w_i(t) \in \mathbb{R}_{\geq 0}$ is the demand of ith client LAN in time instant $t \in \{1, \ldots, T\}$. The problem is to decide how to replicate (copy) objects across the graph nodes in order to minimize the total replication cost and service cost. The replication cost is assumed to be equal to b_j per each copy of any object at node j. Any copy of object can be also deleted from the network at no cost (except the last copy). The service cost is equal to the transmission demands, resulting from the magnitudes of demands. If the object is replicated at node i in a given iteration t, then demand $w_i(t)$ is fulfilled at no cost. Otherwise, the cost of servicing ith node equals exactly $w_i(t)$ (since the network is uniform).

In the presented model, it is assumed that request demands change at discrete time instants. Operations of replication can be performed between any change of clients' demands, i.e., at any $t \in \{1, \ldots, T-1\}$.

The following notation is introduced. Let vector $\mathbf{z}(t)$ be defined as:

$$z_j(t) = \begin{cases} 1 & \text{object is available in node } j \text{ before} \\ & \text{the request } t \text{ is processed,} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Using the indicator notation $[P] = 1$ if predicate P is true and $[P] = 0$ if predicate P is false, we define replication and service costs, respectively:

$$d_1(\mathbf{z}, \mathbf{z}') = \sum_{j=1}^{N} b_j [z_j < z'_j], \tag{12}$$

$$d_2(W_t) = \sum_{j=1}^{N} w_j^{(t)} [z_j = 0]. \tag{13}$$

The problem is to minimize the sum of replication and service costs within the time horizon T, with respect to the placement decision variable $\mathbf{z}$:

$$\text{minimize} \quad \sum_{t=1}^{T} [d_1(\mathbf{z}(t-1), \mathbf{z}(t)) + d_2(\mathbf{w}_t)] \tag{14}$$

subject to:

$$\forall_t \quad \sum_{i=1}^{N} z_i(t) \geq 1, \tag{15}$$

$$\forall_{i,t} \quad z_i(t) \in \{0, 1\}. \tag{16}$$

The presented problem is very similar to the *distributed paging* problem introduced in [5], also called the *constrained file allocation problem*. If for all t we have demands of the form $\mathbf{w}_t = [0 \ 0 \ \ldots \ 1 \ \ldots \ 0]^T$ (i.e., there is only one request at one node in each iteration), then the problem is a simple replication problem [8], which is in turn a special case of the *file allocation problem* [2]. This generalized model allows for two different types of request: *read* and *write* (in the formulation above all requests are *read*). *Write* requests require all object replicas to be updated and thus its service cost is proportional to length of minimum Steiner tree (or, in alternative formulations, to minimum spanning tree) instead of shortest path. This model is traditionally used to model memory management in distributed systems, caching in networks or database object management.

It is known that (offline) replication problem is NP-hard in general networks [15]. However, for uniform networks the problem can be solved in polynomial time.

B. Exact off-line algorithm

As it was shown by Lund et al. [15] the (uncapacitated) replication problem of one object in uniform network (as well as more general file allocation problem) can be solved in polynomial time. The proof uses the reduction to the min-cost maximum 1-commodity flow problem on acyclic network. Here, we show a similar reduction for the case of uniform data placement problem (14)–(16).

Theorem 1. *The optimal solution of dynamic data placement problem* (14)–(16) *in uncapacitated uniform network without processing costs can be computed in time polynomial in network size N and time horizon length T.*

Proof: Let $\sigma = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T)$ be a sequence of demand vectors. The following flow network is constructed. The network has $(|\sigma|+1)$ layers of nodes (indexed from 0 to T) and $(2N-1)$ nodes in each layer. Additionally, there is a single source node s and a single terminal node d. Nodes in each tth layer are divided into two subsets $V_t = \{v_1^{(t)}, v_2^{(t)}, \ldots, v_N^{(t)}\}$ and $U_t = \{u_1^{(t)}, u_2^{(t)}, \ldots, u_{N-1}^{(t)}\}$. Nodes in set V_t correspond to the actual nodes in underlying data network, while nodes in set U_t are artificial and are used only to denote the absence of object in the data network. Each layer $t \geq 1$ corresponds to the state of network just before demands $\mathbf{w}_t$ are about to be served. Layer $t = 0$ corresponds to the initial placement $\mathbf{z}(0)$ of object.

There is an arc between each pair of nodes in layers t and $t+1$, for $t = 1, \ldots, T-1$, i.e., between any pair of nodes from $(V^{(t)} \cup U^{(t)}) \times (V^{(t+1)} \cup U^{(t+1)})$. There is an arc between s and node $v_i^{(0)}$ such that $z_i(0) = 1$, and also between s and k first nodes in U_0, where $k = N - |\{j : z_j^{(0)} = 1\}|$. There is an arc between every node in the last layer $t = T$ and the terminal node d.

All arcs in the flow network have unit capacity. Observe that since the network is acyclic and there are exactly N nodes leaving the source node s, and there exists a path from each node in each layer to the terminal node d, the maximum amount of flow that can be transported through this network is exactly N.

Costs of arcs are defined as follows. All arcs leaving source nodes s, as well as all arcs entering terminal node d have costs 0. For all $t \in \{0, 1, \ldots, T-1\}$, costs of arcs entering any node in $U^{(t+1)}$, leaving any node in $V^{(t)} \cup U^{(t)}$, have costs 0. Arcs leaving node $v_i^{(t)}$ and entering node $v_i^{(t+1)}$ have costs $-w_i(t+1)$. Arcs leaving node $v_i^{(t)}$ and entering node $v_j^{(t+1)}$, where $j \neq i$, have costs $b_j - w_j(t+1)$. Finally, arcs leaving node $u_i^{(t)}$ and entering node $v_j^{(t+1)}$, for any j, have costs $b_j - w_j(t+1)$.

Now, observe that if there is a flow entering a node $v_i^{(t+1)}$ from any node in $U^{(t)}$ or node in $V^{(t)}$ different than $v_i^{(t)}$, then we replicate the object at node i in the underlying content provider network, paying the placement cost b_i. If that flow leaves node $v_i^{(t)}$, then the object is already at node i, thus there is no placement cost. In both cases the demand $w_i(t+1)$ is served at zero cost. If there is no flow to some node $v_j(t+1)$, then there is no object at j in the underlying content provider network. This means that the demand $w_j(t+1)$ has to be served from different node. Due to the uniform network distances it does not matter from which node this demand is served, thus the cost is always $w_j(t+1)$.

Let $f_{vv'}^{(t)} = 1$ if there is a flow between node $v \in V^{(t)} \cup U^{(t)}$ and node $v' \in V^{(t+1)}$, and $f_{vv'}^{(t)} = 0$ otherwise. The cost of flow is:

$$F(f) = \sum_{t=0}^{T-1} \left[\sum_{v \in V^{(t)} \cup U^{(t)}} \left(\sum_{\substack{v' \in V^{(t+1)} \\ v \neq v'}} f_{vv'}^{(t)} (b_{v'} - w_{v'}(t+1)) - f_{vv}^{(t)} w_v(t+1) \right) \right]. \quad (17)$$

In order to compute the optimal solution value it is enough to solve the min-cost max-flow problem defined above. We charge all demands of clients from all T iterations in advance, and then add the placement and subtract service costs resulting from the flow assignment. Let f^* be the optimal flow. Then the optimal solution of (14)–(16) has value:

$$Q(f^*) = F(f^*) + \sum_{t=1}^{T} (w_1(t) + w_2(t) + \ldots + w_N(t)). \quad (18)$$

The number of nodes in the constructed flow network is polynomial in N and T. Thus, the claim of the theorem follows from the polynomial time solvability of min-cost max-flow problem.

■

V. Competitive analysis

To examine the performance of online algorithm, denote by $ALG(\sigma)$ the value of solution obtained by algorithm ALG on input sequence σ. By $OPT(\sigma)$ we denote the optimal value of solution obtained by exact offline algorithm on the same input sequence σ. Without the loss of generality let us assume that values of each feasible solution are always positive. The value of $ALG(\sigma)/OPT(\sigma)$ can be regarded as a natural comparison grade [1]. Formally, the *competitive ratio* of online algorithm ALG is defined as $\sup_{\sigma \in I} ALG(\sigma)/OPT(\sigma)$, where I is the set of all allowed input sequences. Equivalently, we say that algorithm ALG is c-competitive, if for all sequences $\sigma \in I$ there exists a constant b, that $ALG(\sigma) \leq c \cdot OPT(\sigma) + b$.

The competitive ratio of 1 (or 1-competitive algorithm) correspond to the best possible online algorithm, however it is rarely the case that such an algorithm exists for a given problem. The value of competitive ratio tells us how much worse can be the online solution, compared to the one computed assuming full knowledge of input data sequence. Optimal solutions $OPT(\sigma)$ can be often computed using dynamic programming algorithms, such as the one presented in Section III for data placement problem.

Considering all algorithms solving online the given problem, the notion of competitiveness can be extended to the problem itself [13], if we consider the performance of hypothetically best algorithm on the worst possible input sequence. Let $\mathcal{A}$ be the set of all algorithms solving given problem. The competitive ratio of this problem is defined as:

$$\inf_{ALG \in \mathcal{A}} \sup_{\sigma \in I} \frac{ALG(\sigma)}{OPT(\sigma)}. \tag{19}$$

Let us return to the uncapacitated variant of the problem on uniform network without processing costs, defined as (14)–(16). In the online settings we need to solve the problem subsequently for each vector of demands $\mathbf{w}_t$. Without the loss of generality we assume that initially there is an object only at node 1. Consider the following four natural algorithms:

Algorithm A. When $t = 1$, place an object in each node. Then do nothing.

Algorithm B. Let $V(t)$ denote the set of nodes storing a copy of object in iteration t. Without the loss of generality, let initially $V(0) = \{1\}$. In each iteration t, let $j = \arg\max\{w_i(t) : i \in V(t)\}$. If $b_j \leq w_j(t)$ then place an object at node j, and let:

$$V(t) \leftarrow V(t-1) \cup \{j\}. \tag{20}$$

Algorithm C. The same as Algorithm B, except that an object is unconditionally placed at node j, corresponding to maximum $w_i(t)$, in each iteration.

Algorithm D. Keep counter $c_j(t)$ on each node $j \in \mathcal{N}$. Initially each $c_j(0) = 0$. For each input vector increase counters: $c_j(t) \leftarrow c_j(t-1) + w_j(t)$. If for any $c_j(t) \geq b_j$ then replicate an object at j.

Observe that Algorithm A always yields a cost equal to $B = \sum_{j=2}^{N} b_j$, regardless of the input sequence. This is optimal only if $\sum_{t=1}^{T} \sum_{j=2}^{N} w_j(t) \geq B$. However, the competitive ratio of this algorithm is unbounded, since the worst-case input sequences for this algorithm would be of the form: $\sigma_t = [0, w_2(t), 0, \ldots, 0]^T$, for $w_2(t) \to 0$ (i.e., zero demands for all nodes except e.g., node $j = 2$, which has very small demand $w_2(t)$). This results in $OPT(\sigma) = \sum_{t=1}^{\infty} w_2(t) < B$. It is always possible to construct such input sequence that $OPT(\sigma)$ will be arbitrarily small, e.g., for any $\epsilon > 0$, $w_2(t) = \frac{\epsilon}{2^t}$, which gives $OPT(\sigma) = \epsilon$, and competitive ratio is $B/\epsilon \to \infty$.

Algorithm B places an object at the node of highest demand, provided that its demand is no less than the placement cost. Although it may seem more reasonable placement method, this also has unbounded competitive ratio, thus may be considered as the least robust for different input sequences. To see this, consider input sequence consisting entirely of $w_j(t) < b_j$ for all j. For this sequence, the algorithm will never replicate an object, but as $T \to \infty$ the sum of demands served can be arbitrarily large. Notice, however, that it performs optimally on the input sequences that give the worst result in case of Algorithm A.

Algorithm C fills all nodes with objects after $N-1$ iterations. Without the loss of generality assume that $w_1 \geq w_i$ for all $i \in \{2, \ldots, N\}$. This algorithm yields a bounded cost:

$$\sum_{j=2}^{N} b_j + \sum_{t=1}^{T} \sum_{i \notin V(t)} w_i \leq$$

$$\leq B + \sum_{i=1}^{N} (N-i) w_i \leq B + \sum_{i=2}^{N} w_1 - \sum_{i=2}^{N} i w_1 \leq$$

$$\leq B + w_1 \left(\frac{1}{2} N^2 - \frac{3}{2} N - 1 \right). \tag{21}$$

It improves the Algorithm B, having a bounded competitive ratio, which however depends on the size of network N.

Algorithm D again improves Algorithm C, as it would defer replicating an object as long as the total demand requested from a node is less than replication cost b_j. It also has the best competitive ratio, bounded by a constant, regardless of network size and even values of parameters. For any j, let t_j denote such iteration t in which counter $c_j(t_j)$ exceeds b_j for the first time. We consider the following cases of input sequences σ:

1) For all nodes j, $\sum_{t=1}^{T} w_j(t) \geq b_j$. Then the cost paid by algorithm on this sequence σ is:

$$\sum_{j=2}^{N} \sum_{t=1}^{t_j - 1} w_j(t) + \sum_{j=2}^{N} b_j \leq \sum_{j=2}^{N} b_j + \sum_{j=2}^{N} b_j = 2B. \tag{22}$$

Since in such case it is optimal to replicate everywhere (i.e., just apply Algorithm A), the optimal solution has cost B. Thus, the competitive ratio of Algorithm D for these sequences is exactly 2.

2) For all nodes j, $\sum_{t=1}^{T} w_j(t) < b_j$. It is optimal to never replicate any object. This is exactly what Algorithm D does for such input sequences.

Any given input sequence σ can be seen as a mixture of cases 1) and 2). For some nodes j we may have the total demand exceeding b_j. From this we conclude that Algorithm D is 2-competitive, and consequently:

Corollary 1. *Competitive ratio of online data placement problem on uniform network without processing costs is upper-bounded by* 2.

VI. Computational experiments

The efficiency of Algorithm D from the previous section has been confirmed with an experimental study. This algorithm allows to compute very good approximate solutions within fractions of seconds even for very large problem instances. Table I contains the summary of these results. Not only these solutions are never worse than 2 times the optimal (as implied by Corollary 1), but for uniformly generated random instances they were usually very close to optimal. In order to compare these results with optimal solutions, dynamic programming

Algorithm 1 Dynamic programming algorithm for data placement on uniform network.

```
Require: Input sequence $\sigma = (\mathbf{w}(1), \mathbf{w}(2), \dots, \mathbf{w}(T))$, placement costs
    $\mathbf{b} = [b_1, b_2, \dots, b_N]^T$.
Ensure: Sequence of optimal placement decisions $\mathbf{z}^*(1), \mathbf{z}^*(2), \dots, \mathbf{z}^*(T)$.
function SOLVE$(t, \mathbf{z})$
    if $t = T$ then
        $\mathbf{z}^*(T) \leftarrow \mathbf{z}(T-1)$
        for $i = 1, \dots, N$ do
            if $z_i(T-1) = 0$ and $w_i(T) \geq b_i$ then
                $\mathbf{z}_i^*(T) \leftarrow 1$
            end if
        end for
        return $\mathbf{z}^*$
    end if
    $v^* \leftarrow \infty$
    $\mathbf{z}^* \leftarrow \mathbf{z}$
    for each binary sequence $\mathbf{s}$ of length $N$ do
        if vector $\mathbf{s}' = \mathbf{s} - \mathbf{z}(t-1)$ does not contain entries $-1$ then
            $\mathbf{z}(t) \leftarrow \mathbf{s}$
            $\mathbf{z}_{\text{next}} \leftarrow$ SOLVE$(t+1, \mathbf{z})$
            $v \leftarrow Q(\mathbf{z}_{\text{next}})$
            if $v < v^*$ then
                $v^* \leftarrow v$
                $\mathbf{z}^* \leftarrow \mathbf{z}_{\text{next}}$
            end if
        end if
    end for
    return $\mathbf{z}^*$
end function
```

TABLE I. EXAMPLE SOLUTIONS COMPUTED BY ONLINE ALGORITHM D. LAST TWO COLUMNS LIST VALUES OF OPTIMAL SOLUTIONS OBTAINED VIA DYNAMIC PROGRAMMING FOR SMALLER INSTANCES, ALONG WITH THE RUNNING TIME OF COMPUTATIONS (IN SECONDS).

N	T	solution value	running time	optimum	running time
5	30	102.01	0.2	74.74	1074
6	10	38.74	0.2	29.34	30
6	15	56.42	0.2	42.81	449
6	20	77.08	0.2	58.62	2675
6	30	112.81	0.2	85.9	38067
6	40	150.94	0.2	115.47	252880
8	10	49.42	0.2	38.3	6100
8	15	80.78	0.3	56.63	15767
10	5	30.16	0.1	25.24	200
10	8	50.0	0.2	37.5	66043
12	3	20.23	0.1	19.77	30
12	5	37.71	0.1	31.87	53750
15	3	25.15	0.2	23.05	1905
15	5	39.35	0.2	N/A	N/A
50	50	1216.37	0.3	N/A	N/A
100	100	5324.7	0.5	N/A	N/A
250	250	32667.61	1.1	N/A	N/A
500	500	128442.61	2.15	N/A	N/A
1000	1000	507175.79	4.56	N/A	N/A

algorithm has been also implemented (see Section III). Due to the prohibitive running time $O(2^{NT})$, optimal solutions only for small problem instances were obtained. This procedure is described in detail as Algorithm 1. Running it for $t = 1$ and initially zero matrix $\mathbf{z} = \mathbf{0}_{N \times T}$ allows to compute optimal placement matrix $\mathbf{z}^*$. Instances were generated by using random demands $w_i \in (0,1)$ and placement costs $b_i \in (1, T)$ from uniform distributions.

VII. CONCLUSION

The general online data placement problem is hard to solve efficiently. Exact dynamic programming procedure requires time exponential in both network size and time horizon length. In this paper, a simplified variant of this problem has been studied, in which storage capacities of servers are neglected, and all transmission delays are treated as (approximately) equal. For such a case, two results were obtained: 1) given access to the full input data, this problem can be solved in time polynomial in both network size and time horizon length; 2) in online settings, when a decision has to be computed after each element of input data sequence in provided, the competitive ratio of the problem is 2, i.e., there exists an online algorithm, which results in overall performance no worse than twice the optimal off-line algorithm.

ACKNOWLEDGMENT

This research is co-financed by the European Union as part of the European Social Fund.

REFERENCES

[1] S. Albers and S. Leonardi. On-line algorithms. *ACM Computing Surveys*, 31(3es):4, 1999.

[2] B. Awerbuch, Y. Bartal, and A. Fiat. Competitive distributed file allocation. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 164–173. ACM, 1993.

[3] I. Baev, R. Rajaraman, and C. Swamy. Approximation algorithms for data placement problems. *SIAM Journal on Computing*, 38(4):1411–1429, 2008.

[4] Y. Bartal. Distributed paging. In Amos Fiat and Gerhard Woeginger, editors, *Online Algorithms*, volume 1442 of *Lecture Notes in Computer Science*, pages 97–117. Springer Berlin / Heidelberg, 1998.

[5] Y. Bartal, A. Fiat, and Y. Rabani. Competitive algorithms for distributed data management. *Journal of Computer and System Sciences*, 51(3):341–358, 1995.

[6] M.H. Bateni and M.T. Hajiaghayi. Assignment problem in content distribution networks: unsplittable hard-capacitated facility location. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 805–814. Society for Industrial and Applied Mathematics, 2009.

[7] T. Bektas, J.F. Cordeau, E. Erkut, and G. Laporte. Exact algorithms for the joint object placement and request routing problem in content distribution networks. *Computers & Operations Research*, 35(12):3860–3884, 2008.

[8] D.L. Black and D.D. Sleator. Competitive algorithms for replication and migration problems. Technical report, Technical Report CMU-CS-89-201, Department of Computer Science, Carnegie-Mellon University, 1989.

[9] A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.

[10] M. Drwal and J. Jozefczyk. Decentralized approximation algorithm for data placement problem in content delivery networks. In *Proc. of 3rd Conference on Computing, Electrical and Industrial Systems*, 2012.

[11] M.T. Hajiaghayi, M. Mahdian, and V.S. Mirrokni. The facility location problem with general cost functions. *Networks*, 42(1):42–47, 2003.

[12] M.R. Korupolu, C.G. Plaxton, and R. Rajaraman. Placement algorithms for hierarchical cooperative caching. In *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 586–595. Society for Industrial and Applied Mathematics, 1999.

[13] E. Koutsoupias and C. Papadimitriou. Beyond competitive analysis. *SIAM Journal on Computing*, 30(1):300–317, 2000.

[14] T. Leighton. Improving Performance on the Internet. *Communications of the ACM*, 52(2):44–51, 2009.

[15] C. Lund, N. Reingold, J. Westbrook, and D. Yan. Competitive online algorithms for distributed data management. *SIAM J. Comput.*, 28(3):1086–1111, 1999.

[16] S. Sivasubramanian, M. Szymaniak, G. Pierre, and M. Steen. Replication for web hosting systems. *ACM Computing Surveys*, 36(3):291–334, 2004.

[17] O. Wolfson, S. Jajodia, and Y. Huang. An adaptive data replication algorithm. *ACM Transactions on Database Systems*, 22(2):255–314, 1997.

Semantic Web Technologies in Business and IT Alignment

Multi-Model Algorithm of Ontology Matching

Ulf Seigerroth, Julia Kaidalova
School of Engineering
Jönköping University
Jönköping, Sweden
Ulf.Seigerroth@jth.hj.se, Julia.Kaidalova@jth.hj.se

Nikolay Shilov
Computer Aided Integrated Systems Laboratory
SPIIRAS
St.Petersburg, Russia
nick@iias.spb.su

Tomasz Kaczmarek
Department of Information Systems
Poznan University of Economic
Poznan, Poland
t.kaczmarek@kie.ue.poznan.pl

***Abstract*— The importance of using appropriate and effective IT means to facilitate business functions have been acknowledged and discussed by practitioners and scientists over the past two decades thus giving a rise to the Business and IT alignment (BITA) problem. In BITA, the issue of sharing and processing large amounts of information from distributed and heterogeneous sources is of significant importance. Ontologies have shown their usability for this type of tasks. However, they also bring new challenges. The paper proposes utilising of the Semantic Web technologies to assist in solving them. Namely, the paper describes application and adaptation of the ontology matching algorithm for the BITA problem. The algorithm integrates different matching models. Its operation is shown via an illustrative example.**

Keywords-ontology; matching; business and IT alignment

I. Introduction

A key issue in today's enterprise activity is information technology (IT) that supports business needs, processes, and strategies [1]. The importance of using appropriate and effective IT means to facilitate business functions have been acknowledged and discussed by practitioners and scientists over the past two decades [2]. The problem of Business and IT alignment (BITA) is even more complex due to the dynamic and evolving nature of both sides – business and IT [3]. Today, BITA is a crucial issue for enterprise success and it is often addressed as a top concern of IT and business practitioners [4, 5]. It is possibly caused by the recognition of organizational benefits that BITA can bring to the table.

Researchers and practitioners discuss and tackle BITA in various ways. Initially, from more general point of view, the problem was studied as linking business plan and IT plan or business strategy with IT strategy. Talking more particularly, one possible way to perceive it is the following: in order to align business and IT perspectives it is required to enable alignment of their representations [6]. Here, models come into play, since in this respect models are often used as a supportive means that are able to capture and represent different aspects and constructs of an enterprise. Subsequently, models are often used as a support in a transition process to take a business from one state to another and to create BITA.

In such tasks, the issue of sharing and processing large amounts of information from distributed and heterogeneous sources (information management systems, experts, electronic documents, real-time sensors, etc.) is of significant importance. Due to such factors, as different data formats, interaction protocols, etc. this leads to a problem of semantic interoperability.

Hence, the information sharing in a business network is highly important for BITA and should be achieved at both technical and semantic levels. The interoperability at the technical level is addressed in a number of research efforts. It is usually represented by such approaches as, e.g., Service-Oriented Architecture or SOA [7] and on the appropriate standards like WSDL and SOAP [8]. The semantic level of interoperability in the flexible supply network is also paid significant attention. As an example (probably the most widely known), the Semantic Web initiative is worth mentioning [10]. The main idea is to use ontologies for knowledge and terminology description.

Ontologies have shown their usability for this type of tasks (e.g., [11], [12], [13]). These are content theories about the sorts of objects, properties of objects and relations between objects that are possible in a specified knowledge domain. Ontologies provide potential terms for describing the knowledge about the domain [14]. An ontological model is used to solve the problem of heterogeneity of descriptions of different enterprise elements. This model makes it possible to enable interoperability between heterogeneous information sources due to provision of their common semantics [9].

However, in open or evolving systems, as in the case of BITA, different parties would, in general, adopt different ontologies. Thus, just using ontologies, like just using XML, does not reduce heterogeneity; it raises heterogeneity problems at a higher level.

Ontology matching is a promising solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of the ontologies. These correspondences can be used for various tasks, such as ontology merging, query answering, data translation, or for navigation on the Semantic Web. Thus, matching ontologies enables the knowledge and data expressed in the matched ontologies to interoperate [15].

The goal of ontology matching is basically solving the two major problems, namely:

- Identify ontology entities, which have different names but have the same meaning.
- Identify ontology entities, which have the same (similar) name but have different meaning.

In 2010, a multi-model approach for on-the-fly ontology matching to be used in smart spaces has been developed [16]. In this paper, the above mentioned concept is further developed and adapted for the BITA problem.

The paper is structured as follows. Section II introduces the problem. The background in the area of ontology matching is presented in Section III. Section IV describes the ontology matching algorithm and its components. The algorithm's adaptation is presented in Section V. The case study and its discussion is given in Section VI. Major results are summarized in the last section.

II. Problem Statement

An example of a BITA problem could be the following situation: there are certain IT systems deployed in an enterprise, which already support its operations. However, the support might be only partial (full would mean an enterprise requiring little human intervention). Second, the environment, in which the enterprise is placed, evolves (customers, suppliers, legal conditions, and general macroeconomic conditions - all change). IT systems have to be modified to reflect these changes. Third, enterprise managers or the owner might want to change the way the business is carried out (even if no particular external change enforces it) – again IT systems have to be modified. BITA therefore is the problem of changing the IT infrastructure and systems to the imagined future state of these systems.

Why is this hard? From a more general point of view considered previously [6], the problem lies in identifying what to change and in what way (how). Now, the modelling comes into the play. Having an "as-is" model and a "to-be" model of an enterprise, let us assume that we want to change a part of the "as-is" model that describes enterprise operations, to reflect the desired "to-be" situation of enterprise operations. The part of the model that describes IT systems is misaligned and has to be changed too.

A. Assumptions

The notations and terminology between models describing the "as is" and "to be" situations are the same. However, the interpretation of the same objects may differ depending on its connections with other objects.

At the moment, the RDF [17]-like formalism (consisting of triples "subject-predicate-object") is considered as the most generic (consisting basically of objects and relationships between them). In this case the models can be considered as ontologies.

B. Approach ideas

The main aim of the approach is to find a way to describe relationships between different models so that these relationships could be transferred between "as is" and "to be" sets of models. Model versioning techniques could be used if the "to be" set is built upon the "to be" set.

Ontology matching originating from Semantic Web could assist in solving the "as is" and "to be" alignment problem.

III. Ontology Matching

The existing ontology matching techniques has been analyzed by Smirnov et al. [16]. All the similarity metrics in the performed state-of-the-art review are based on the two information retrieval metrics, namely metrics of precision and recall. The balance between these is achieved via choosing the right threshold value. The possibility of choosing the right threshold value has to be taken into account in the development of the matching models.

The above mentioned approaches to ontology matching apply the following techniques in different combinations. The techniques include:

- Identification of synonyms
- Similarity metrics (name similarity, linguistic similarity)
- Heuristics (for example two nodes are likely to match if nodes in their neighborhood also match)
- Compare sets of instances of classes instead compare classes
- Rules: for example, if class A1 related to class B1 (relation R1), A2 related to class B2 (relation R2) and B1 similar to B2, R1 similar to R2 therefore A1 similar to A2.

As a result of matching, the following types of elements mapping proximity can be identified:

- One-to-one mapping between the elements (Associate-Professor to Senior-Lecturer)
- Between different types of elements (the relation AdvisedBy(Student, Professor) maps to the attribute advisor of the concept Student)
- Complex type (Name maps to the concatenation of First Name and Last Name)

All methods can be separated into the following four groups.

Linguistic Methods. These methods are focused on determining similarity between entities based on linguistic comparison of these entities (count of the same symbols estimation, estimation of the longest similar parts of words, etc.).

Statistical Methods (Instance-Based). These methods compare instances of the ontology entities and based on this estimation entities can be compared.

Contextual Methods. The aim of the contextual similarity is to calculate a measure of similarity between entities based on their contexts. For example, if parents and children of the ontology classes are the same consequently the classes also the same.

Combined Methods. These methods combine specifics of two or three of the above methods.

In the considered problem domain, the differentiation between instances is not an easy task. Because of this reason, the techniques and methods relying on instances were not considered for further development. Hence, the developed models presented below integrate all of the above techniques (except those dealing with instances) and propose a set of combined methods having features of the linguistic and contextual methods.

IV. MULTI-MODEL APPROACH FOR ON-THE-FLY ONTOLOGY MATCHING

The below proposed approach allows matching of ontologies for the interoperability purposes and is based on the ontology matching model illustrated by Figure 1. The approach takes into account that matched ontologies are responsible for concrete and well-described tasks, which means that they generally should be small–to–medium size and describe only limited domains. A detailed description of the approach can be found in [16].

Ontology is represented as RDF triples, consisting of the following ontology elements: subject, predicate, object. Degree of similarity between two ontology elements is in the range [0, 1]. The approach consists of the following steps:

- Compare all elements of two ontologies and fill the matrix M'. Matrix M' is of size *m* to *n*, where *m* is the number of elements in the first ontology and *n* is the number of elements in the second ontology. Each element of this matrix represents the degree of similarity between two ontology elements.
  - The degree of similarity between equal elements or synonyms is set to 1 (maximum value of the degree of similarity).
  - WordNet or Wiktionary are used to calculate semantic distances based on the synonymy relationship or other relationships (with a lower degree of similarity).
  - The degree of similarity between other string terms of ontology elements is calculated using the fuzzy string comparison method.
- Update values in matrix M, where each new value of elements of M is the maximum value of (M, M')
- Improve distance values in the matrix M using the graph-based distance improvement model.

As a result, the matrix M contains degrees of similarity between elements of two ontologies. This allows determining correspondences between elements by selecting those for which the degrees of similarities are above the pre-selected threshold value.

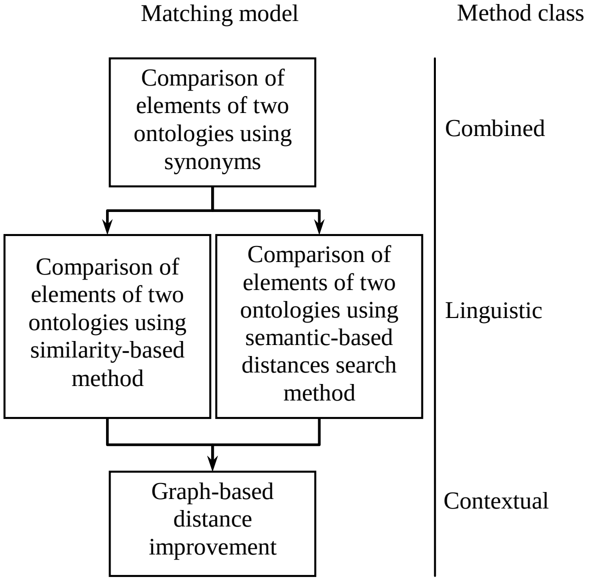

Matching model
Method class
Comparison of elements of two ontologies using synonyms
Combined
Comparison of elements of two ontologies using similarity-based method
Comparison of elements of two ontologies using semantic-based distances search method
Linguistic
Graph-based distance improvement
Contextual

Figure 1. Multi-model approach to on-the-fly ontology matching.

V. ADAPTATION OF THE ALGORITHM AND CASE STUDY

The analysis of the possible application of the above ontology matching approach to the BITA domain has shown that the following elements of the approach that could be applied:

1. Comparison of elements of two models taking into account synonyms (e.g., a synonym can be considered as a description of the same aspect)
2. Comparison of elements of two ontologies using fuzzy string comparison.
3. Graph-based distance improvement (e.g., if an object A in one model is a super class of three other objects, and object B in the other model is a super class of the same three other objects, then it is likely that A and B are the same)

The analysis of the possibility to apply Wiktionary for the experimentation purposes has shown that in the business domain it does not have enough synonyms to produce any significant results in the ontology matching. The situation with the WordNet is better, but the amount of synonyms in the business domain is still not sufficient. This issue is still under research and will be addressed in the future work.

Below, the detailed description of the used techniques can be found.

A. Fuzzy string comparison

The basis of the fuzzy string comparison algorithm is the well-known conventional algorithm that calculates occurrence of substrings from one string in the other string.

1. Perform the comparison based on the above algorithm twice: $FC_1 = FuzzyCompare(Element_1, Element_2)$ and $FC_2 = FuzzyCompare(Element_2, Element_1)$.
2. Calculate the result as an aggregation of the above results in accordance with the following formula: $Re' = n*FC_1+(1-n)*FC_2$, where

n is a weight, $n \in [0;1]$; $n = 0.5$ sets the same weight to the both strings, $n = 0$ searches only Element$_2$ within Element$_1$, and $n = 1$ searches only Element$_1$ within Element$_2$. It is proposed to set $n = 0.5$.

B. *Graph-based distance improvement*

The graph-based improvement model for propagation similarities from one ontology elements to another is presented in Figure 2. The aim of this model is to propagate the degree of similarity between closely matching ontology elements to ontology elements related to them through RDF triples.

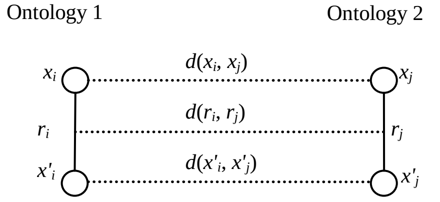


Figure 2. Graph-based distance improvement.

Set $X = (x_1, x_2, ..., x_n)$ is the set of subjects and objects in the ontologies of two knowledge processors, $D_x = (d(x_i, x_j), ...)$ is a degree of similarity between x_i and x_j. Set $R = (r_1, r_2, ..., r_n)$ is the set of predicates in the ontologies of two knowledge processors. Constant Tr is a threshold value which determines if two ontology elements mapped to each other or not.

The following algorithm allows propagating similarity distance to RDF subjects and objects:

```
$d(x_i, x_j)$ = maximum($D_x$)
while ($d(x_i, x_j) > Tr$) do
    for each $d(x'_i, x'_i)$ as $x_i\, r_m\, x'_i$ and $x_i\, r_m\, x'_j$ do
        $d(x'_i, x'_j) = \sqrt[2]{d(x_i, x_j)\, d(x'_i, x_j)}$
    endfor
    Exclude $d(x_i, x_j)$ from $D_x$
    $d(x_i, x_j)$ = maximum($D_x$)
endwhile
```

Today, modern research in the considered and adjacent areas [18] - [20] suppose finding threshold value by experts based on the statistical data. For every task family the threshold will be different but usually it is in range [0.5, 0.95].

VI. Case Study

The ontologies that are used as examples of “as is” and “to be” are two ontologies taken from a real project. In this project two types of ontologies were used: base ontology and domain ontology. The examples in this article are domain ontologies. The domain ontologies were used to give a structure for conceiving organizational work performed by humans and artifacts on behalf of an enterprise. It was an important driver for adopting a multi-layered thinking and for creating alignment between and within different abstraction layers in an enterprise (strategy, practice, process, service, IS/IT infrastructure). This domain ontology served as an essential and necessary enabler for directing attention to what to conceive on different levels of abstraction (focal areas), how to interrelate different levels, and how to gradually increase the degree of formalism on more detailed levels in the framework (lower levels of abstraction).

A. *“As is” ontology*

In order to match the “as is” and “to be” ontologies the “as is” ontology had to be built out of two source domain ontologies. These ontologies are presented in Figure 3. To find the correspondences between classes of the domain ontologies the same ontology matching algorithm was applied. It produced the following results:

<u>Similar objects:</u>

Prerequisite / basis - Basis = 0.55 (the result of fuzzy string comparison)

Product - Product = 1

<u>Matched objects:</u>

Product - Product = 1

It can be seen that 2 pairs of similar classes were found and no similarity propagation has been made. With the threshold being set to 0.75 the pair “Prerequisite / basis” – “Basis” was dropped.

After an analysis of the ontologies, it was concluded that in this particular case, the classes “Prerequisite / basis” and “Basis” are the same. As a result, it was decided to explicitly mark these as synonyms. The second try taking into account the introduced synonymy relationship produced the following results:

<u>Similar objects:</u>

Prerequisite / basis - Basis = 1

Product - Product = 1

<u>Matched objects:</u>

Prerequisite / basis - Basis = 1

Product - Product = 1

It can be seen that 2 matching pairs of classes were found. All other classes are different. As a result the merged “as is” ontology have been built as a union of these two ontologies, where classes “Prerequisite / basis” - “Basis” and “Product” - “Product” have been merged into “Prerequisite / basis” and “Product” correspondingly. The resulting ontology is shown in Figure 4.

B. *“As is” – “to be” ontology matching*

In this subsection, the matching of the above built “as is” ontology and “to be” ontology shown in Figure 5 is performed. The “to be” ontology is outlined in such a way so that its differences from the “as is” ontology would be obvious.

The matching results produced by the multi-model algorithm were as follows:

<u>Similar objects:</u>

Result - Result = 1

Prerequisite / basis - Prerequisite / basis = 1

Activity - Activity = 1

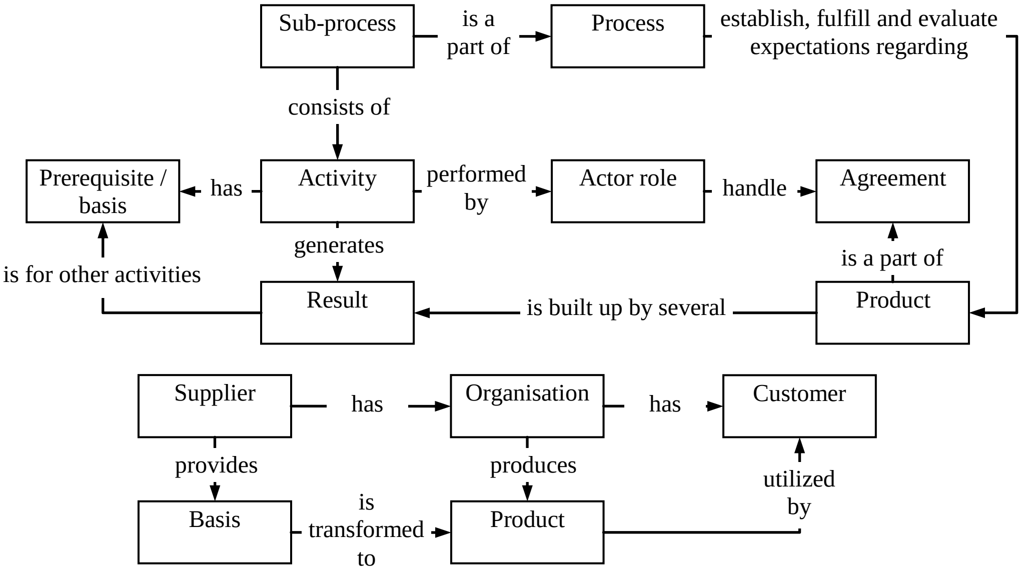


Figure 3. “As is” ontologies.

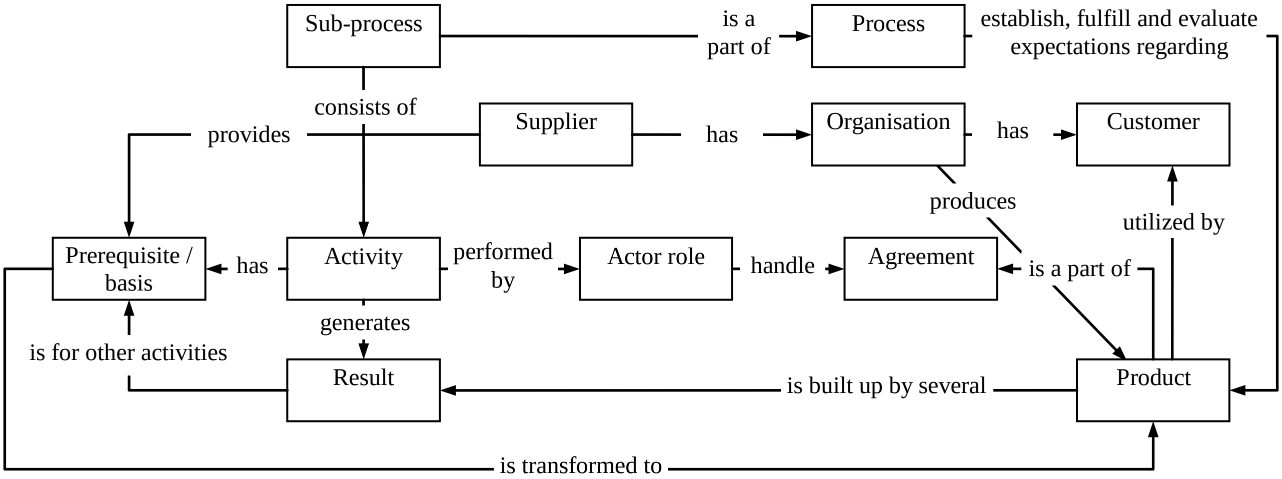


Figure 4. Integrated “as is” ontology.

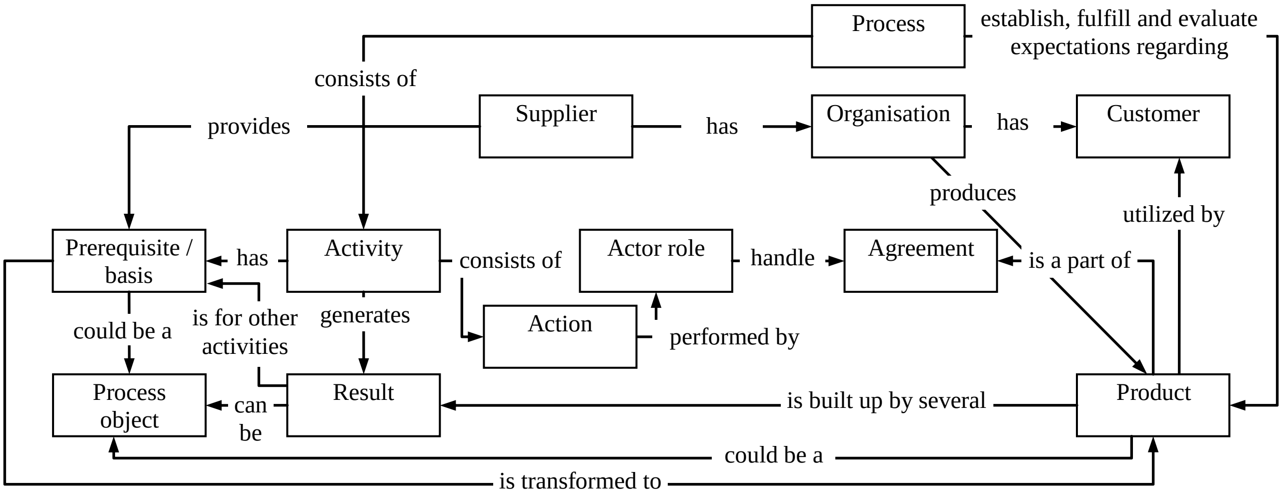


Figure 5. “To be” ontology.

ISBN: 978-1-61208-300-1

Activity - Action = 0.40 (the result of fuzzy string comparison)

Sub process - Process = 0.72 (the result of fuzzy string comparison)

Process - Process = 1

Process - Process object = 0.65 (the result of fuzzy string comparison)

Actor role - Actor role = 1

Agreement - Agreement = 1

Product - Product = 1

Supplier - Supplier = 1

Organisation - Organisation = 1

Customer - Customer = 1

Similarity propagation:

Activity - Activity = 1

Sub process - Process: 0.72 -> 0.85;

The “Sub process” class of the “as is” ontology and the “Process” class of the “to be” ontology matched with the similarity of 0.72. Since they are both related to the classes “Activity” matched with the similarity 1, the graph-based distance improvement algorithm has propagated this fact via increasing the similarity between the “Sub process” and “Process” classes from 0.72 to 0.85 ($\sqrt[2]{1*0.72}$).

Actor role - Actor role = 1

Activity - Action: 0.40 -> 0.64;

The “Activity” class of the “as is” ontology and the “Action” class of the “to be” ontology matched with the similarity of 0.40. Since they are both related to the classes “Actor role” matched with the similarity 1, the graph-based distance improvement algorithm has propagated this fact via increasing the similarity between the “Activity” and “Action” classes from 0.40 to 0.64.

Product - Product = 1

Process - Process object: 0.65 -> 0.80;

The “Process” class of the “as is” ontology and the “Process object” class of the “to be” ontology matched with the similarity of 0.65. Since they are both related to the classes “Product” matched with the similarity 1, the graph-based distance improvement algorithm has propagated this fact via increasing the similarity between the “Process” and “Process object” classes from 0.65 to 0.80.

The final result is as follows:

Matched objects:

Result - Result = 1

Prerequisite / basis - Prerequisite / basis = 1

Activity - Activity = 1 or Activity - Action = 0.64

Sub process - Process = 0.85

Process - Process = 1 or Process - Process object = 0.80

Actor role - Actor role = 1

Agreement - Agreement = 1

Product - Product = 1

Supplier - Supplier = 1

Organisation - Organisation = 1

Customer - Customer = 1

The result is quite predictable. However, there are three interesting facts, which deserve some discussion.

The class “Process” of the “as is” ontology matched to the classes “Process” (with similarity 1) and “Process object” (with similarity 0.80) of the “to be” ontology. This could be interpreted as a sign to the expert to check if the new added class “Process object” is the same with the existing class “Process”, which is not the case in our example.

The class “Activity” of the “as is” ontology also matched to two classes of the “to be” ontology: “Activity” (with similarity 1) and “Action” (with similarity 0.64). This means that the algorithm actually discovered that in the “to be” ontology the class “Activity” was split into two classes “Activity” and “Action”.

Similarly, via finding the match between both the “Sub process” and the “Process” classes of the “as is” ontology with the class “Process” of the “to be” ontology. The algorithm informed the experts that the class “Sub process” was probably merged with the class “Process”, what is the truth in the presented example.

VII. Conclusion and Future Work

The paper presented the adaptation of the ontology matching algorithm originating from research works in the area of Semantic Web to BITA. By far, there are no approaches aimed at comparison of ontologies representing “as is” and “to be” situations. The proposed algorithm is aimed at assisting experts in finding changes between the mentioned ontologies. The algorithm is based on the idea when the least computationally expensive operations are followed by more computationally expensive. As a result, the more computationally expensive operations work with less data resulting in a higher efficiency of the algorithm. The illustrative example presented shows some of the advantages of using this approach. It has been discussed how this algorithm can reveal and inform the experts about situations where classes are split or merged. For the presented example the algorithm works nearly instantly.

At the moment, the algorithm does not implement the matching based on the synonymy relationship. The Wiktionary does not have enough synonyms to produce any significant results in the ontology matching for BITA. The situation with the WordNet is better, but the amount of synonyms in the business domain is still not sufficient. This issue is still under research and will be addressed in the future work.

Acknowledgment

This paper is due to the project COBIT sponsored by the Swedish Foundation for International Cooperation in Research and Higher Education. Some parts of the work have been sponsored by grant # 12-07-00298-a of the Russian Foundation for Basic Research.

References

[1] A. J. G. Silvius, “Business and IT Alignment: What We Know and What We Don't Know,” Proc. International Conference on Information Management and Engineering, IEEE, 2009, pp. 558-563.

[2] J. O. Vargas, “A Framework of Practices Influencing IS/Business Alignment and IT Governance,” Thesis, School of Information Systems, Computing and Mathematics in Brunel University, 2011.

[3] N. J. Luftman, “Measure Your Business-IT Alignment,“ Optimize: Business execution for CIOs Magazine, iss 26, 2003.

[4] Y. E. Chan and B. H. Reich, “IT alignment: what have we learned?” Journal of Information Technology, vol. 22, 2007, pp. 297-315.

[5] J. Luftman and E. R. McLean, “Key issues for IT executives,” MIS Quarterly Executive, vol. 3, no. 2, 2004, pp. 89-104.

[6] T. Kaczmarek, U. Seigerroth, and N. Shilov, “Multi-layered enterprise modeling and its challenges in business and IT alignment,” Proc. International Conference on Enterprise Information Systems, 2012, pp. 257-260.

[7] SOA: Service-oriented architecture definition, 2007. URL: http://www.service-architecture.com/web-services/articles/service-oriented_architecture_soa_definition.html [retrieved June, 2013].

[8] Web Services explained, 2007. URL: http://www.service-architecture.com/web-services/articles/web_services_explained.html [retrieved June, 2013].

[9] M. Uschold and M. Grüninger, “Ontologies: Principles, methods and applications,” Knowledge Engineering Review, vol. 11, no. 2, 1996, pp. 93-155.

[10] Semantic Web, 2006. URL: http://www.semanticweb.org [retrieved June, 2013].

[11] D. J. Bradfield, J. X. Gao, and H. Soltan, “A Metaknowledge Approach to Facilitate Knowledge Sharing in the Global Product Development Process, Computer-Aided Design & Applications, vol. 4, no. 1-4, 2007, pp. 519-528.

[12] E. C. K. Chan, and K. M. Yu, “A framework of ontology-enabled product knowledge management,” International Journal of Product Development, Inderscience Publishers, vol. 4, no. 3-4, 2007, pp. 241-254.

[13] L. Patil, D. Dutta, and R. Sriram, “Ontology-based exchange of product data semantics,” IEEE Transactions on Automation Science and Engineering, vol. 2, no. 3, 2005, pp. 213-225.

[14] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, “What are ontologies and why do we need them?” IEEE Intelligent Systems, vol 14, no. 1, Jan/Feb, 1999, pp. 20-26.

[15] Ontology Matching, 2013. URL: http://ontologymatching.org [retrieved June, 2013].

[16] A. Smirnov et al., “On-the-Fly Ontology Matching in Smart Spaces: A Multi-Model Approach,” Smart Spaces and Next Generation Wired/Wireless Networking, Proc. Third Conference on Smart Spaces, ruSMART 2010, and 10th International Conference NEW2AN 2010, S. Balandin, R. Dunaytsev, and Y. Koucheryavy (Eds.), Springer, LNCS 6294, pp.72-83.

[17] Resource Description Framework (RDF), 2004. URL: http://www.w3.org/RDF/ [retrieved July, 2013].

[18] A. Doan, J. Madhavan, P. Domingos and A. Halevy, “Learning to map between ontologies on the semantic web,” Proc. of the 11th international conference on World Wide Web, 2002, pp. 662-673.

[19] D. AnHai, M. Jayant, D. Pedro, and H. Alon, “Ontology Matching: A Machine Learning Approach,” Handbook on Ontologies in Information Systems (Staab, S., Studer, R., eds.), Springer-Velag, 2003, pp. 397-416.

[20] J. Euzenat, and P. Shvaiko, “Ontology matching,” Springer-Verlag, 2007.

Efficient Content Sharing over Content Centric Networking

Sho Harada
Waseda University
Tokyo, Japan
shoharada1990@akane.waseda.jp

Yong-Jin Park
Waseda University
Tokyo, Japan
yjp@ieee.org

***Abstract*—Content Centric Networking (CCN) is a clean slate network architecture optimized for today's and expected future's demands for the Internet. It is a promising paradigm for the Future Internet architecture among new clean slate network architectures. It realizes the reduction of the load on servers, efficient mobility support, fast content distribution and retrieval, and high security. However, there is room for improvement on CCN when we share contents in a community or we distribute contents as a service provider. In this paper, we propose a method that enables us to share contents more efficiently in a community. Our content sharing model will reduce an extra processing and the network load by 20% on average.**

Keywords—Future Internet; Content Centric Networking; Named Data Networking;

I. INTRODUCTION

In recent years, the number of Internet users is explosively increasing. In the past, we used Internet for exchanging messages. However, the purpose changed and the majority of today's users use the Internet for retrieving contents. To meet today's demands, researchers have been extending the functionalities of Internet protocols. However, in this scenario, we need to continue extending the functionality if the scale of the Internet keeps increasing. One promising solution is to make a clean slate architecture that is designed for large-scale Internet and optimized for today's demands.

In the past five years, Information Centric Networking (ICN) [1] [2] is becoming popular. In ICN, we do not depend on location information such as IP address to request or retrieve contents, but Unified Resource Identifiers (URIs) that indicate resources. It enables us to request contents directly and we do not need to be aware of their exact locations. As ICN architectures, there are many promising architectures. PURSUIT [3], which is the succeeding project of PSIRP [4], has been researched mainly in Europe. It realizes the optimization of network by utilizing centralized management.

Content Centric Networking (CCN) [5] [6] is the most promising one in ICN architectures. CCN enables us to reduce the burden on servers, efficient mobility support, and fast content retrieval by dispersion of network load. However, there is room for improvement. Our proposal realizes cutting out the extra communication and reducing network load when we share contents in a dynamic community.

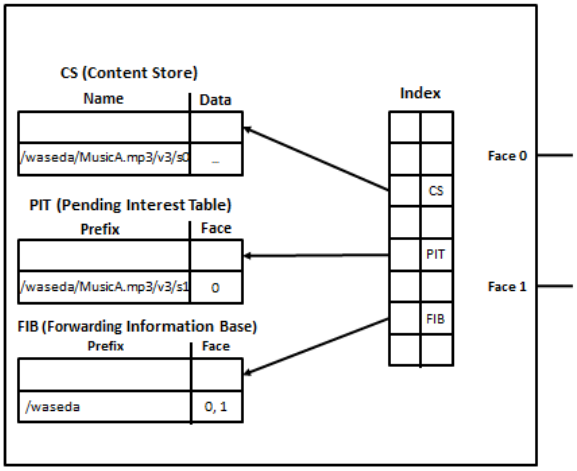


Figure 1. Structure of CCN Router

The remainder of this paper is organized as follows. Section Ⅱ describes the basic CCN architecture and Voice over Content Centric Networking [7] as a related work. In Section Ⅲ, we explain the problem of the basic CCN, our proposed community model, the key distribution method, and the content sharing model. In Section Ⅳ, we show a performance evaluation. Finally, we conclude our paper in Section Ⅴ.

II. CONTENT CENTRIC NETWORKING

A. Basic CCN Architecture

In CCN, we use two kinds of packets. Interest Packets are used to request content by name. Data Packets are used to deliver requested content to the requester in response to an Interest Packet. By using these packets, we can retrieve contents without knowing where they are. In CCN, we use an exclusive router called CCN Router. Fig. 1 shows the structure of CCN Router. CCN Router has three kinds of tables. Content Store (CS) is used to store contents, enables CCN Router to cache contents. When a CCN Router receives an Interest Packet and has the requested content in its CS, it will send the cached content to the requester instead of the content producer. Therefore, users can retrieve contents fast while the load on content producer is reduced. In addition, it supports mobility. When a user moves and fails to retrieve a Data Packet, the user just simply requests the content again.

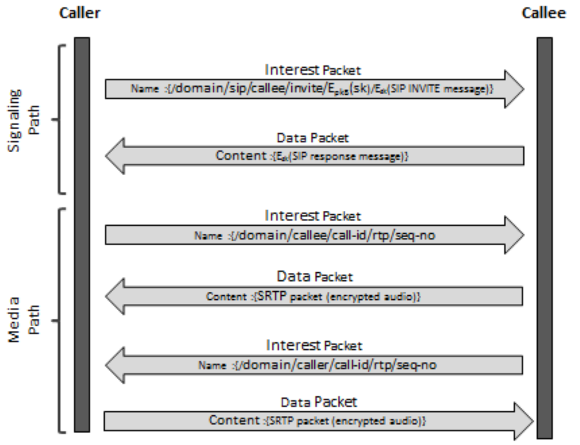

Caller
Callee
Signaling Path
Interest Packet
Name :{/domain/sip/callee/invite/E_{pke}(sk)/E_{sk}(SIP INVITE message)}
Data Packet
Content :{E_{sk}(SIP response message)}
Media Path
Interest Packet
Name :{/domain/callee/call-id/rtp/seq-no
Data Packet
Content :{SRTP packet (encrypted audio)}
Interest Packet
Name :{/domain/caller/call-id/rtp/seq-no
Data Packet
Content :{SRTP packet (encrypted audio)}

Figure 2. Communication Model in VoCCN

In this case, the user can retrieve the content from the adjacent CCN Router, which has cached the content in its CS. Of course, we can use traditional technologies that are used in IP network for mobile support. Pending Interest Table (PIT) saves the face where an Interest Packet carried on and is used to forward Data Packet to the original requester. Forwarding Information Base (FIB) is similar to a routing table and used to forward Interest Packet to the content producer. The detail of FIB and the routing is written in [8].

In CCN, users can retrieve contents by name and do not care about the content locations. In addition, adequate security measure is taken in contents. Therefore, it is difficult for malicious users to attack a specific server nor to tamper contents.

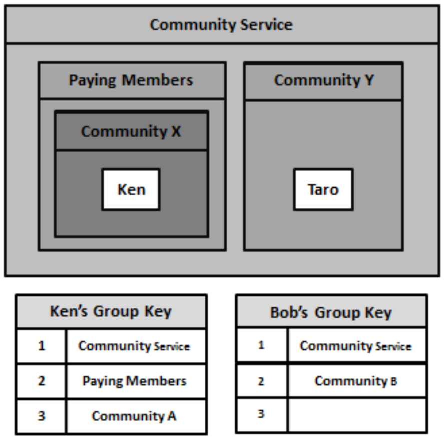
B. *Related Work*

Voice over Content Centric Networking (VoCCN) is a real-time, conversational, telephony application over CCN and simpler, more secure and more scalable than VoIP [9]. In VoCCN, data flows directly from producer to consumer. In addition, CCN architecture enables multipoint routing.

Fig. 2 shows the processes to start a call over CCN. At first, a caller sends an Interest Packet, which includes *"invite message"* and the symmetric key encrypted by the callee's public key combined with its content name. Then, the callee who received the caller's Interest Packet will send a response message encrypted by the caller's symmetric key as a Data Packet. By this method, the caller and the callee share a symmetric key and can communicate each other securely. In our proposal, secure key transmission is very important because our proposal is designed on the premise of it.

III. EFFICIENT CONTENT SHARING MODEL

In this Section, we explain the insufficient of basic CCN and our proposed content sharing model. Our content sharing model will reduce the load on servers. In addition, the average load on network will be cut by 20%.



Ken's Group Key	
1	Community Service
2	Paying Members
3	Community A

Bob's Group Key	
1	Community Service
2	Community B
3	

Figure 3. Proposed Community Model

A. *The Problem of Basic CCN*

In basic CCN, when a producer sends out content, the producer will encrypt the content with a key. Then, the requester will decrypt the content with a key that the requester received from the producer in advance. In this time, the encrypted content will be cached in every CCN Routers. Therefore, users can retrieve the content from these CCN Routers and decrypt it if they have the key. However, when the scope of disclosure is updated, all cached contents have to be deleted and the producer needs to re-encrypt content when they are requested again. It causes extra network load and processing since in a dynamic community or monthly subscription service, the scope of disclosure changes frequently. Therefore, we propose a method that enables CCN Routers to keep cached contents even when a community member list is changed and the scope of disclosure is updated. It will cut the extra network load and processing.

B. *Community Model*

We introduce a community model to realize our efficient content sharing model. Fig. 3 shows the structure of a community. A user belongs to a community that is a part of a big community. In a community, group members share a Group Key that is managed by the group leader or the service provider. When a user distributes contents, the user will encrypt the contents with a Group Key. Based on the scope of disclosure, the user can choose a Group Key. When a list of group members is updated, its Group Key will also be updated.

For example, when a producer wants to send out content only for the members of *"Community X"*, the producer will encrypt the content with the Group Key of *"Community X"*. However, when the producer sends out the content for all members of *"Community Service"*, the producer will encrypt the content with the Group Key of *"Community Service"*. Group Key is managed by the group leader or the service provider.

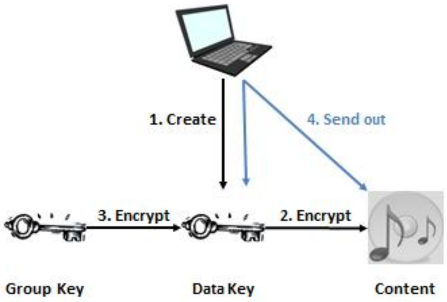


Figure 4. Content Encryption

In this case, the Group Keys of *"Community Service"* and *"Paying Members"* will be managed by the service provider. The Group Keys of *"Community X"* and *"Community Y"* will be managed by the group leaders.

This method enables us to control the scope of content disclosure by the community. In addition, we can control the scope of disclosure only by encrypting contents with key. This model is suitable for CCN architecture because authentication by server is not needed to retrieve contents.

C. Key Distribution

In our model, it is very important to exchange or distribute keys smoothly and securely. Therefore, we propose a key distribution method that applies, simplifies, and optimizes the key transmission of VoCCN for our proposed model.

To distribute a key, client sends Interest Packet by using the Public Key of Group Leader (PKG) and the Symmetric Key of the Client (SKC). The content name of Interest Packet will be like *"Community Service/Paying Members/ Community X/PKG(SKC)/SKC(authentication message and key request message"*. PKG(SKC) means that SKC is encrypted with PKG. In response to this Interest Packet, the group leader will send its group key encrypted with SKC as a Data Packet. Once the group leader needs any information about the client, the group leader will send Interest Packet encrypted by SKC to authenticate the client before sending Group Key.

D. Efficient Content Sharing Model

Fig. 4 shows our content encryption model. To distribute content, the producer creates a Data Key. The content name of a Data Key will be defined like *"Waseda/Network Community/MusicA.mp3/DataKey"*. After creating a Data Key, the producer will encrypt the content with this key.

The Data Key, which was used to encrypt the content, will also be encrypted with a Group Key. The producer will keep several Group Keys. Therefore, the producer will choose one of them to encrypt the content.

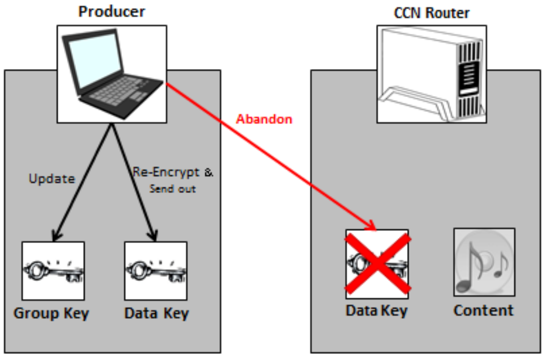


Figure 5. Updating the Scope

In our model, the content name of a Group Key will be defined like *"Waseda/NetworkCommunity/GroupKey/v0"*. Different from the name of Data Key, Group Key has the number of the version in its name. When a scope of disclosure is updated, the Group Key has to be updated. All users need to know it and distinguish the new Group Key. However, the old Data Key must be abandoned and replaced by the new Data Key. Therefore, the name of Data Key must be unique. When this content producer distributes the content, the content and the encrypted Data Key will be sent out at the same time.

In basic CCN, if the list of a community member is updated, the manager of the Group Key needs to update the Group Key and abandon the encrypted content. In our model, the manager will update the Group Key and abandon the encrypted Data Key instead of the content itself (Fig. 5).

Therefore, we do not have to update encrypted content and can continue to use cached contents. It will reduce the re-encryption cost and network load between the producer and the CCN router that caches the content because in many cases, the size of Data Key is much smaller than that of content itself.

IV. EVALUATION

Our proposal will reduce the network load. We defined *Efficiency* as the network load in our proposal divided by the network load in basic CCN. The network load between consumers and routers that cache contents is almost same. However, the load between a producer and the routers will be greatly reduced. The average *Efficiency* of updating the scope will be the following calculation formula.

$$\text{Efficiency} = \frac{0.8\text{Dc} + 2.6\text{I} + 1.8\text{Dk}}{1.0\text{Dc} + 2.0\text{I} + 1.0\text{Dk}} \qquad (1)$$

I in this formula is the data size of an Interest Packet. *Dc* is the size of an encrypted content. *Dk* is the size of a key. The load by *Dk* and *I* increases. However, we can reduce the load by content from 1.0*Dc* to 0.8*Dc*. In many cases, *Dc* is far bigger.

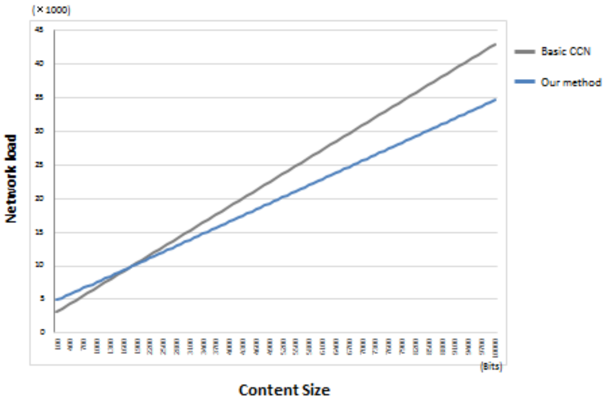


Figure 6. Comparison of Network Load

We compared the network load between our model and basic CCN, as shown in Fig. 6. When the content size is more than 2000 bits, the efficiency of our method is better than basic CCN. The bigger the content size is, the more efficient our proposal will be. However, when the content size is small, the traditional method is better. Therefore, the producer needs to choose which method to use. When we use the evaluation result of [5] as an argument, the calculated efficiency is 80%. In other words, our proposed model can reduce the network load by 20% on average.

In our model, the caching policy is based on on-path caching, which is a basic caching policy in CCN. However, many caching policies are already proposed [10,11,12]. If we combine our model with a appropriate caching policy, the efficiency will improve dramatically. In addition, our proposal is independent of security standards. Therefore, users can use any traditional security methods.

V. CONCLUSION AND FUTURE WORK

Content Centric Networking is a promising architecture in Future Internet research. It enables us to retrieve contents efficiently and reduces the network load. In addition, CCN realizes efficient mobility supports, high security, and fast content retrieval. CCN Routers have storage to cache contents, while acting as intermediate nodes. We can share contents efficiently by using caching function. In basic CCN, we encrypt contents to distribute them to particular group. However, when the scope of disclosure is updated, we have to abandon the cached contents. In this case, we have to re-encrypt the contents and re-send them. In other words, we cannot use caching function effectively. In this paper, we proposed a community model, a key distribution method, and content sharing model. By using the community model and the key distribution model, our proposed content sharing model enables us to share contents efficiently in a dynamic community. In our content sharing model, content producers create Data Key, which is abandoned instead of the content itself. It reduces the network load and extra processions. Fig.6 shows the efficiency of our proposed method. We can reduce the network load by 20% on average because the size of Data Key is much smaller than that of the content itself and we are able to take advantage of CCN caching function.

REFERENCES

[1] B. Ahlgren, C. Dannewitz, C. Imbrenda, and D. Kutscher, “A Survey of Information-Centric Networking”, IEEE Communications Magazine, Vol. 50, Issue. 7, 2012, pp. 26-36.

[2] M. F. Bari, R. Boutaba, and B. Mathieu, "A Survey of Naming and Routing in Information-Centric Networks", IEEE Communications Magazine, Vol. 50, Issue. 12, 2012, pp. 44-53.

[3] N. Fotiou, G.C. Polyzos, and D. Trossen, "Illustrating a publish-subscribe Internet architecture", Journal on Telecommunication Systems, Springer, 2011, pp. 233-245.

[4] N. Fotiou, P. Nikander, D. Trossen, and G.C. Polyzos, "Developing Information Networking Further: From PSIRP to PURSUIT", BROADNETS’10, 2010, pp. 52-58.

[5] V. Jacobson, et al., “Networking Named Content”, CoNEXT ’09, 2009, pp. 1-12.

[6] L. Zhang, et al., "Named Data Networking (NDN) Project", NDN Technical Report NDN-0001, 2010..

[7] V. Jacobson, et al., “VoCCN: Voice-over Content-Centric Nwtworks”, ACM ReArch’09, 2009, pp. 1-6.

[8] L. Wang, et al., "OSPFN: An OSPF Based Routing Protocol for Named Data Networking", NDN Technical Report NDN-0003, 2012.

[9] B. Goode, "Voice over Internet protocol (VoIP)", Proceedings of the IEEE, Vol. 90, Issue. 9, 2002, pp. 1495-1517.

[10] Z. Ming, M. Xu, and D. Wang, “Age-based cooperative caching in Information-Centric Networks”, INFOCOM WKSHPS’12, 2012, pp. 268-283.

[11] I. Psaras, W.K. Chai, G. Pavlou, “Probabilistic in-network caching for information-centric networks”, ICN’12, 2012, pp. 55-60.

[12] K. Cho, et at., “WAVE: Popularity-based and collaborative in-network caching for content-oriented networks”, INFOCOM WKSHPS’12, 2012, pp. 316-321.

Scalable OpenFlow Controller Redundancy Tackling Local and Global Recoveries

Keisuke Kuroki, Nobutaka Matsumoto and Michiaki Hayashi
Integrated Core Network Control And Management Laboratory
KDDI R&D Laboratories, Inc.
Saitama, Japan
e-mail:{ke-kuroki, nb-matsumoto, mc-hayashi}@kddilabs.jp

***Abstract*—OpenFlow is expected to be an enabler that solves the problems of today's network. Thanks to the centralized management with OpenFlow, agile network operation can be achieved with flexible programmability; however, the centralized management implies a significant impact of any outages of the OpenFlow controller. Hence, a high availability technology is indispensable for building the OpenFlow controller, and the high availability system should consider extraordinary events (e.g., power outage) affecting the entire data center as well as anticipated server failures within a local system. In this paper, the high-availability of the OpenFlow controller is investigated, and a redundant method considering both local and global (i.e., inter data-center) recoveries is proposed by using the multiple-controllers functionality that is defined in OpenFlow switch specification version 1.2 and later. The proposed redundant scheme eliminates frontend server causing limitation of performance scalability, while it achieves competitive role change and failover times.**

Keywords-OpenFlow; controller; redundancy.

I. INTRODUCTION

Towards future telecom services, programmability of the network is expected to shorten the service delivery time and to enhance flexibility of service deployment meeting diversified and complex user requirements on various applications (e.g., real-time and non real-time applications). OpenFlow [1] is an enabler of the centralized management solution meeting the aforementioned expectations, and many researches have addressed the scalability issue of the OpenFlow-based solution.

In [1], several OpenFlow controllers are evaluated from the viewpoint of scalability in centralized management and control. Message processing performances of two operation modes (i.e., proactive and reactive) of the OpenFlow controller are evaluated using several existent implementations (e.g., Floodlight, NOX, Trema). In [2], the scalability of the OpenFlow solution for a data center environment is analyzed to show an implementation guideline. The paper concludes that, to achieve lossless and low delay performance in the data center application, the number of OpenFlow switches managed by one controller should be limited to eight. To leverage an advantage of the centralized management, the OpenFlow controller should not be a simple flow switching policy server. In [3], OpenQoS architecture delivers end-to-end quality of service (QoS) with OpenFlow-based traffic control. The OpenFlow controller with OpenQoS has the role of collecting the network state to perform dynamic QoS routing, i.e., the controller has the route calculation function just like the Path Computation Element (PCE). Indeed, in IETF, PCE architecture is growing as a stateful operation supporting the enforcement of path provisioning in addition to its original path computation role. Hence, the importance of the OpenFlow controller is growing with the broader concept of Software Defined Networking (SDN), and thus the high availability of the controller system must be discussed. However, there is little research on the high availability of the OpenFlow controller that must play the important role on SDN.

In this paper, the high availability of the OpenFlow controller is investigated, and a high availability method applicable to multiple OpenFlow controllers is proposed. In the proposed redundant method, "global" repair (i.e., inter data-center redundancy) as well as local repair (i.e., redundancy within a local network) are considered. The proposal achieves a competitive failover time compared with existent redundant schemes (e.g., server clustering), while the proposal does not require any frontend server limiting performance scalability of the OpenFlow controller.

The organization of this paper is as follows: In Section II, we explain the function of multiple controllers defined in OpenFlow switch specification 1.2 [5] and also explain its applicability to achieving redundancy of the OpenFlow controller. We investigate existent approaches of redundant schemes as well. In Sections III-A and B, we evaluate the performance of the redundant method for multiple controllers placed on a single domain. In Sections III-C and D, we propose the redundant method for multiple controllers placed on multiple domains and evaluate the performance. Finally, concluding remarks are given in Section IV.

II. BACKGROUND

Typical implementation of OpenFlow allocates a controller separating the control plane from the data plane, and an OpenFlow switch playing the role of data plane communicates with an OpenFlow controller using the OpenFlow protocol over a Transport Layer Security (TLS) [12] or a TCP connection [13] defined as a "secure channel". The switch tries to forward a packet by looking up flow entries populated in-advance by the controller. If the packet does not match the current flow entries, the switch sends a packet-in message over the secure channel to the controller in order to retrieve a direction on how to treat the packet.

One method handling data plane failure is implementing a monitoring function on OpenFlow switch [11]; however, only the monitoring function in a data plane is not sufficient to achieve high availability of the control plane. In contrust, achieving controller redundancy also contributes to protection of the data plane.In the case of the controller outages, the secure channel connection is lost accordingly, and then the packet-in message cannot be successfully processed by the controller. Hence, new packets that are not matched with the flow entry are simply dropped or allowed to fall in a default operation (e.g., forwarding to a neighbor anyway) that never provides desirable services until the ultimate recovery of the controller.

OpenFlow specification 1.2 introduced the capability of multiple controllers by defining three states (i.e., MASTER, SLAVE, and EQUAL) of a controller. A controller has its own role by using the function of multiple controllers, and the state itself is owned by the switch. In the three states, MASTER and EQUAL have full access to the switch and can receive all asynchronous messages (e.g., packet-in) from the switch. A switch can make secure channel connections to multiple EQUAL controllers, but the switch is allowed to access only one MASTER controller. In the SLAVE state, a controller has read-only access to switches and cannot receive asynchronous messages apart from a port-status message from the switches. A controller can change its own state by sending an OFPT_ROLE_REQUEST message to switches. On receipt of the message, the switch sends back an OFPT_ROLE_REPLY message to the controller. If the switch receives a message indicating the controller's intent to change its state to MASTER, all the other controllers' states owned by the switch are changed to SLAVE. This function enables a switch to have multiple secure channels, and thus the switch is not required to re-establish new secure channels in the event of controller outages. In the multiple-controllers capability, the role-change mechanism is entirely driven by the controllers, while the switches act passively only to retain the role. Therefore, investigating the implementation of the controller side is important to achieve the redundancy; however, that has not been proposed yet. We use the capability of multiple controllers to achieve high availability of the control plane. In the following section, we propose how to use it and explain the effect.

III. PROPOSAL AND DEMONSTRATION

In this section, a proposed architecture for local and global recoveries is described, and recovery operation in the two scenarios (i.e., local and global) is demonstrated. To avoid the secure channel re-establishment that is inevitable in conventional virtual IP-based redundancy, the proposal commonly applies the multiple-controllers functionality [10] to both local and global scenarios. Through the demonstration for the two scenarios, we implemented the controller prototype based on NOX-C++ for OpenFlow 1.2 available in [10].

A. *Proposed Design of Local Recovery*

First, we explain the redundant method in a single domain, which is typically a data-center hosting OpenFlow controllers.

Figure 1 shows a reference model to describe and demonstrate the proposed scheme designed for the local recovery. An OpenFlow Switch (OFS) is connected to two controllers through two secure channels. In a normal operation, the role of the OpenFlow Controller (OFC) 01 is set to MASTER and that of OFC02 is set to SLAVE. OFC01 and 02 have the same flow entry information mirrored between the two OFCs. OFSs are operated under the reactive mode, and send a packet-in message to the controller when it receives a new packet undefined in the flow entry. To evaluate the performance influence in the data plane, a traffic generator continuously generates data packets with 100 packets per second (pps) where every packet has unique flow identifiers for stressing the reactive operation of the controller.

Figure 2 shows an operational sequence of the proposed redundant scheme utilizing the multiple-controller capability. In the proposed scheme, controllers send keep-alive messages (e.g., ICMP echo) to each other every 50

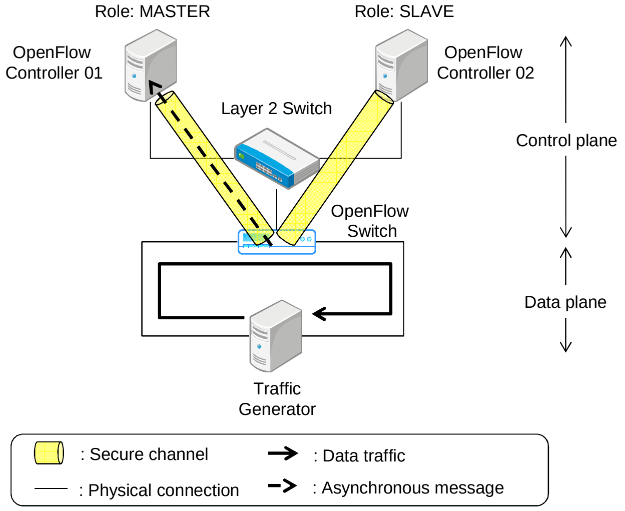


Figure 1. Testbed configuration for the local recovery.

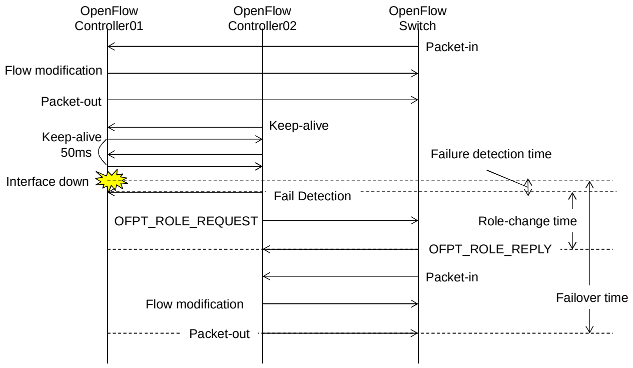


Figure 2. Design of a control procedure for the local recovery.

milliseconds. In a normal operation, OFS sends an asynchronous message such as packet-in to OFC01, since the switch recognizes the role of OFC01 as MASTER and that of OFC02 as SLAVE. OFC01 sends a flow-modification message and packet-out message to respond to the packet-in message from the switch. If the keep-alive message is lost, a controller (i.e., OFC01) is assumed to have failed. Due to the failure of OFC01, OFS cannot send any packet-in messages, and then the data plane cannot continue successful packet forwarding for any new incoming flows. Upon detecting the failure of OFC01, OFC02 sends an OFPT_ROLE_REQUEST message to OFS for changing its own role to MASTER. Then, OFS replies the OFPT_ROLE_REPLY message, and starts sending asynchronous messages to OFC02 after the completion of the role-change process. To respond to the asynchronous messages, OFC02 starts sending flow-modification and packet-out messages, and finally, the packet forwarding in the data plane is restored. As represented in Figure 2, failover time is defined as the duration time from the failure event of OFC01 to the first packet-out message sent by OFC02. Failover time is measured using a traffic generator to obtain the data plane outage time. A role-change time is defined as the duration time from the detection of OFC01 failure to the receipt of OFPT_ROLE_REPLY by OFC02. Role-change time is measured by retrieving the event log of each controller to observe the control message process.

B. Demonstration of Local Recovery

The failover time and role-change time are evaluated with increasing flow entries in order to investigate the influence of the entry size. Figure 3 shows the failover time and role-change time averaged with 10 times measurements. Failover time is around 60-90 milliseconds and role-change time is about 15 milliseconds. Since the failure detection included in the failover time has a timing offset within the keep-alive interval, observed failover time has some fluctuation range. Although the role-change time of the proposal is comparable with that of the virtual address-based redundancy, the failover time of the proposal shows a significant advantage thanks to the seamless handover between multiple secure channels. Figure 3 also shows that entry size on OFCs does not affect the local recovery operation both for role-change time and failover time.

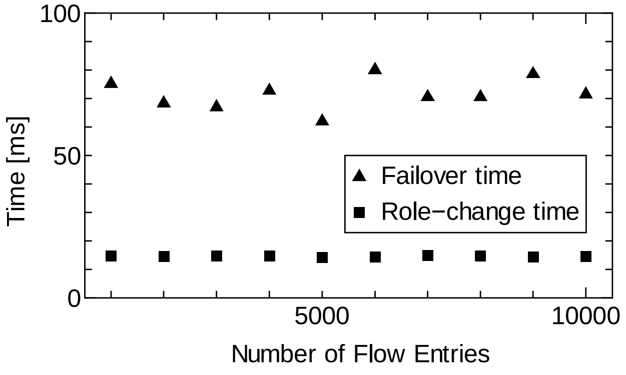


Figure 3. Result of failover and role-change time in a single domain.

C. Proposed Design of Global Recovery

In this section, we explain the redundant method of multiple domains. Figure 4 shows a reference model of the controller redundancy for the global recovery scenario. The global repair should consider tackling extraordinary events affecting, for example, the entire data center. We assume that a controller is installed in each domain to retain its scalability and performance. The controller manages OFSs belonging to the same domain as the MASTER, and the controller manages the other OFSs in the other domains as the SLAVE. The respective roles of the controllers are depicted in the upper side of Figure 4. For example, OFS-A (i.e., some switches belonging to domain-A) recognize the role of OFC-A (i.e., the controller belonging to domain-A) is MASTER and the role of the other controllers is SLAVE. Similarly, OFS-B and OFS-C also recognize the role of the controller that belongs to its same domain is MASTER and the roles of the other controllers are SLAVE. The controller has flow entry information for only OFSs recognizing the controller as MASTER. Thus the controller does not need to have an excessive configuration or receive an excessive message. Additionally, one characteristic of our proposal is the existence of a Role Management Server (RMS). RMS monitors all controllers to manage their role, and RMS has some data such as CPU utilization, role information, configuration of all controllers and domain information of all switches. RMS determines which controller should take over the role of MASTER and relevant configuration data, if a controller has failed. In this regard, we have to be careful to prevent second failures. If OFC-B takes over the role of MASTER for broken OFC-A and places OFS-A under management besides OFS-B, there is the possibility of CPU utilization overload of OFC-B and then OFC-B may fail consequently. Thus we should consider that one failure will induce subsequent failures. That is why RMS monitors CPU utilization and judges multiple controllers should take over the role of MASTER from one controller, if RMS judges

	Role of controller		
OFS A	Master	Slave	Slave
OFS B	Slave	Master	Slave
OFS C	Slave	Slave	Master

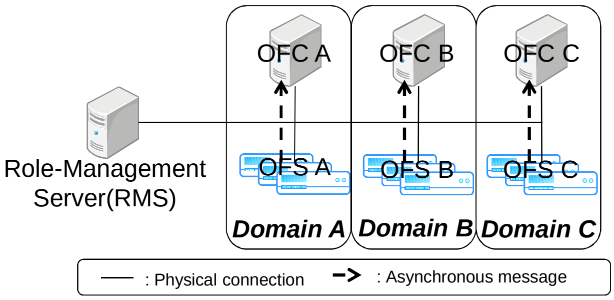
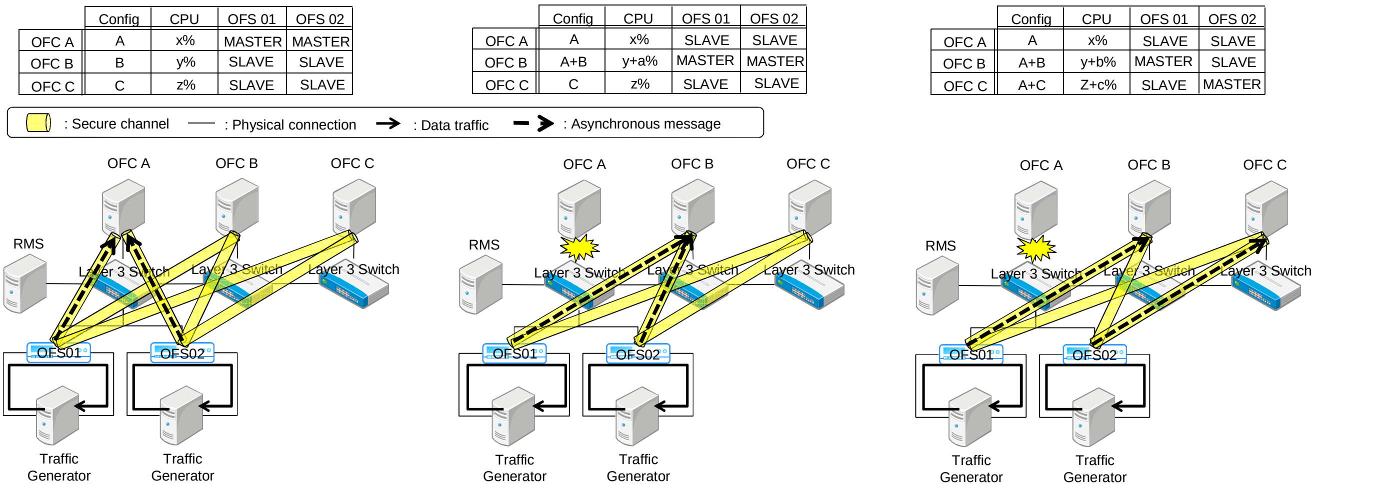


Figure 4. A network model for global recovery.



	Config	CPU	OFS 01	OFS 02
OFC A	A	x%	MASTER	MASTER
OFC B	B	y%	SLAVE	SLAVE
OFC C	C	z%	SLAVE	SLAVE

	Config	CPU	OFS 01	OFS 02
OFC A	A	x%	SLAVE	SLAVE
OFC B	A+B	y+a%	MASTER	MASTER
OFC C	C	z%	SLAVE	SLAVE

	Config	CPU	OFS 01	OFS 02
OFC A	A	x%	SLAVE	SLAVE
OFC B	A+B	y+b%	MASTER	SLAVE
OFC C	A+C	Z+c%	SLAVE	MASTER

(a) Normal state.

(b) Two switches are migrated to a single controller.

(c) Two switches are migrated to two controllers.

Figure 5. Role-change transition in the global controller recovery.

that taking over with single controller raises overload of CPU utilization.

Figure 5 shows the role-change transition for the global controller recovery. Figure 5 (a) shows the initial state, and two switches are connected to three controllers through three secure channels. In the normal operation, both switches recognize that the role of OFC-A is MASTER and the other controllers are SLAVE. So only OFC-A receives some asynchronous messages such as packet-in messages. In this case, the three controllers have different configurations respectively and the information is reflected in the database of RMS. Also RMS has CPU utilization, role information of each controller and the cognition haven by switch regarding the role of the controller in its database. The traffic generator connects OFS01 and OFS02 respectively and the data transfer rate is 100 pps. The two switches receive a new packet and send a packet-in message to the controller at all times as well as the measurement of a single domain.

If OFC-A fails and RMS judges there is no problem to take over the MASTER role by a single controller, the initial state (i.e., Fig. 5 (a)) is changed to Figure 5 (b) where only OFC-B takes over the role of MASTER. The database of RMS is updated accordingly, and both switches start sending asynchronous messages to OFC-B.

In contrast, if OFC-A is failed and RMS judges that a single controller cannot take over the Master role but two controllers can, the initial state is changed to Figure 5 (c) where two controllers take over the role of MASTER. The database of RMS is updated accordingly, and then OFS01 starts sending asynchronous messages to OFC-B. OFS02 sends asynchronous messages to OFC-C.

Figure 6 shows a global recovery scheme in the case of Figure 5 (b). RMS monitors the CPU utilization of all controllers every 50 milliseconds. Since Figure 5 (b) has three controllers, each controller is monitored every 150 milliseconds. The proposed recovery process consists of a judge-phase and a takeover-phase. If RMS is unable to retrieve the information about CPU utilization from OFC-A, RMS does not immediately assume that OFC-A has failed to avoid false positive. To ensure the failure detection, RMS requests that the ICMP echo be sent from the other controllers (OFC-B and OFC-C) to OFC-A. If more than half of the results indicate the failure of OFC-A, RMS determines that OFC-A has failed and starts calculating a new MASTER controller migrating OFC-A’s configuration and OFSs under OFC-A. The process from failure detection to the determination of a failed controller is defined as the judge-phase as indicated in Figure 6. After the judge-phase, RMS moves to the takeover-phase. In the takeover-phase, RMS firstly calculates whether it is no problem for a single controller to take over all switches connected to OFC-A by considering CPU utilization of OFC-A as well as OFC-B and C. If two or more controllers are required to take over all switches of OFC-A, RMS separates the switches based on the ratio of the available CPU resources of new MASTER

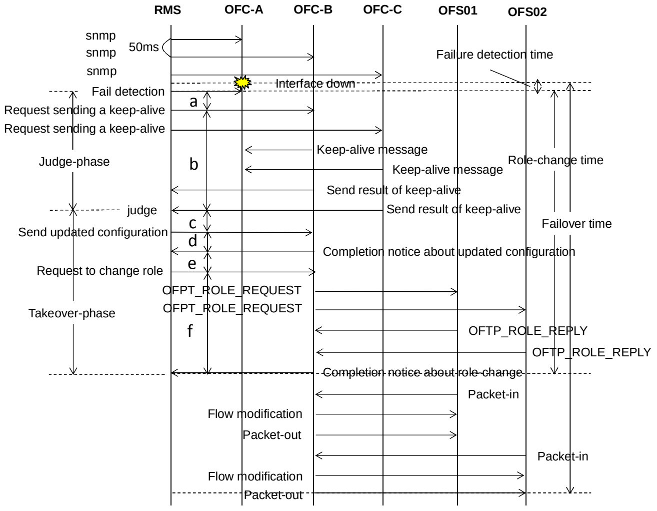


Figure 6. Proposed operational sequence for Figure 5 (b) scenario.

controllers. If RMS decides that OFC-B is sufficiently adequate to become a new single MASTER as shown in Figure 5 (b), RMS integrates OFC-A's configuration into OFC-B's and registers the integrated configuration into OFC-B. Upon receiving the integrated configuration, OFC-B updates its own configuration and then reports the completion of the integration process. Then, RMS requests OFC-B to send OFPT_ROLE_REQUEST to the switches for updating the role of OFC-A to SLAVE and OFC-B as MASTER. The switches send OFPT_ROLE_REPLY after updating the role change process. Then, OFC-B reports the completion of the role-change process to RMS. The process from completion of the judge-phase to completion of the role-change is defined as the takeover-phase. After the take–over phase, the switches OFC01 and 02 start sending asynchronous messages to OFC-B.

D. *Demonstration of Global Recovery*

Figure 7 shows the role-change time and failover time averaged with 10 times measurements in both cases of Figure 5 (b) and (c). Role-change time and failover time increase with the growth of flow entry size. This result shows the difference in behavior compared with the result of a local recovery shown in Figure 3. The major reason for this increase of failover time is that RMS needs integration of multiple configurations of failed OFC and registration of the configuration during the takeover-phase. As different scenarios of the global recovery, RMS selects multiple controllers as the new MASTER as shown in Figure 5 (c), and the scenario takes longer role-change time and failover time as shown in Figure 7. This reason is analyzed using the result of Figure 8 that shows a breakdown of role-change time under 1000 entries in both cases (i.e., Figure 5 (b) and (c)). The characters ("*a*" to "*f*") placed on the x-axis of Figure 8 correspond to the marker shown in Figure 6. As shown in Figure 8, the major performance difference comes from *c* that is the time to integrate configuration in RMS and register it to OFC. Current implementation suffers from the serial processing of the registration of integrated data. This means introducing parallel processing of the registration resolves the delay of role-change for the scenario shown in Figure 5 (c).

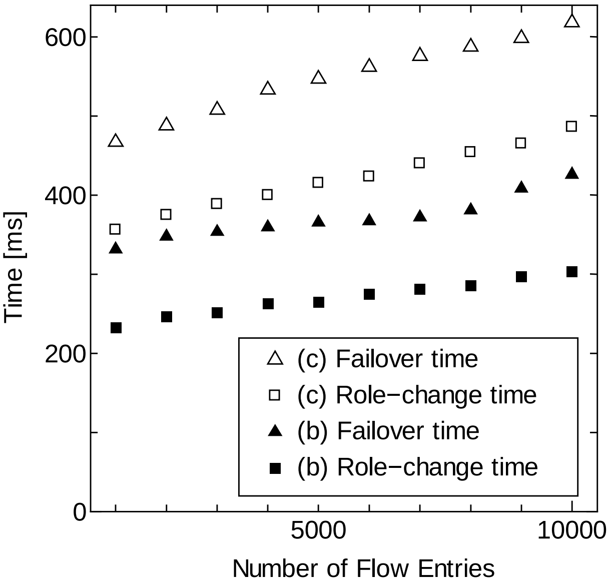


Figure 7. Result of failover time and role-change time in global recovery.

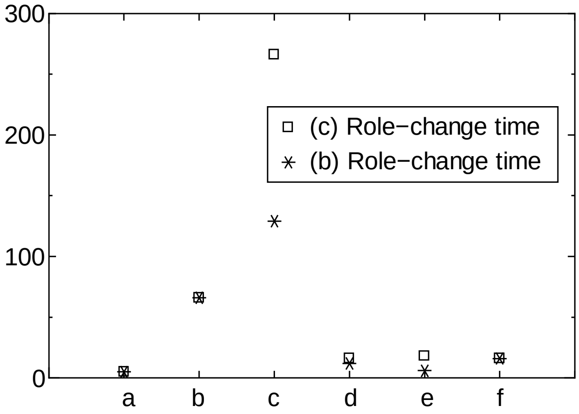


Figure 8. Breakdown of role-change time observed for scenario Figure 5 (b) and (c).

According to Figure 7, role-change time is about 300 milliseconds and failover time is 420 milliseconds in 10000 flow entries, in the case of the scenario in Figure 5 (b). In the case of the Figure 5 (c) scenario, the role-change time is about 500 milliseconds and failover time is about 620 milliseconds. These results indicate that, for both scenarios, our proposal achieves competitive role-change time and faster failover time compared with existent redundant mechanisms [8, 9]. We consider the proposed implementation of multiple controllers achieves high availability controllers for both intra and inter data-center recoveries.

In this paper, we did not explicitly show the redundancy of RMS itself. Although conventional server redundancy mechanisms accompanying relatively longer failover time may be applied to RMS redundancy, single failure of the RMS itself does not directly affect packet forwarding.

IV. RELATED WORK

In [6], the HyperFlow approach improves the performance of the OpenFlow control plane and achieves redundancy of the controllers. HyperFlow introduces a distributed inter-controller synchronization protocol forming a distributed file system. HyperFlow is implemented as a NOX-C++ application and synchronizes all events between controllers by messaging advertisements. In the case of controller failures, HyperFlow requires overwriting of the controller registry in all relevant switches or simply forming hot-standby using servers in the vicinity of the failed controller. Thus, this approach assumes re-establishment of the secure channel, and does not assume the multiple-controllers capability defined in OpenFlow 1.2. Therefore, time duration of the failover operation may increase with the growth of the number of switches managed by the failed controller. Since the failover process of HyperFlow does not consider any server resource, overload of CPU utilization is a potential risk in the event of migrating switches to a new controller especially in the global recovery scenario.

There are several methods of general server redundancy, and such methods may also be effective for OpenFlow controllers. For example, one possible server redundancy can use one virtual IP address aggregating hot-standby or several servers. In [7], failover time is evaluated using the virtual address-based implementation with Common Address Redundancy Protocol (CARP), which is like Virtual Router Redundancy Protocol (VRRP) [8]. According to the analysis, the average time of changing the role between master and backup is 15.7 milliseconds. However there is a concern that the virtual IP-based approach takes a longer fail-over time than our approach, since the virtual IP-based approach fundamentally involves the re-establishment process of the secure channels. Although the virtual IP-based scheme is straightforward if it is applied within single LAN, it cannot simply be applied to multiple locations (e.g., data centers) managed under different addressing schemes. This means that the virtual IP-based scheme alone is not sufficient to tackle global repair. In [9], a server clustering with a mechanism of seamless handover of TCP connection between backend servers was proposed. While each TCP connection is visible to only one back-end server in a normal clustering scheme, the proposal [9] makes the connection visible to at least two back-ends using proprietary backup TCP (BTCP) protocol within a backend network. The connection migrates to a backup, and then the backup is able to resume the connection transparently before the client TCP connection is lost. Using this scheme, the connections are recovered by the backup server within 0.9 seconds including a failure detecting time of 0.5 seconds. This approach is expected to be applicable also for global repair involving multiple locations. However, from the viewpoint of performance scalability of the OpenFlow controller as analyzed in [1, 2], a common frontend server required in the clustering system can be a serious bottleneck of message processing in the control plane (e.g., if the frontend server is broken, all TCP connections are lost). The high availability scheme should avoid such single frontend server to ensure the performance scalability of OpenFlow controllers. In our proposed solution, RMS cannot be a serious bottleneck of processing asynchronous messages because RMS failure itself does not affect any secure channel sessions and thus the data plane is not affected, accordingly. In addition, to tackle global repairs, server utilization should also be considered in the process of migrating many switches. However, conventional approaches do not consider utilization of the server resources (e.g., CPU).

V. CONCLUSION AND FUTURE WORK

In OpenFlow architecture, the controller is an important element to achieve reliable SDN. In this paper, we proposed a redundant scheme to tackle both a single domain (“local”) and multiple domain (“global”) recovery scenarios, which cannot be resolved with conventional redundant schemes. To avoid performance scale-limit due to conventional clustering schemes, our scheme eliminates any frontend server from the redundant system. The demonstration shows that the proposal performs competitive role change and failover times compared with conventional schemes. The role change time observed in a local recovery scenario is about 15 milliseconds regardless of entry size, and that in a global scenario ranges from 200 to 400 milliseconds. CPU resource-aware migration of managed OpenFlow switches in the failover process is successfully achieved by our scheme. The proposal is expected to be an effective high availability scheme necessary for deploying reliable and scalable SDN.

In future work, we will establish CPU-based controller resource modeling to accurately handover many OpenFlow switches in the event of, especially, global recovery where massive nodes may need to be protected.

REFERENCES

[1] N. McKeown et al., “OpenFlow: Enabling innovation in campus networks,” ACM SIGCOMM Computer Communication Review, vol. 38, i 2, April 2008, pp. 69-74.

[2] M. P. Fernandez, “Evaluating OpenFlow controller paradigms,” Proc. International Conference on Networks (ICN2013), January 2013, pp. 151-157.

[3] R. Pries, M. Jarschel, and S. Goll, “On the usability of OpenFlow in data center environments,” Proc. IEEE International Conference on Communications (ICC2012), June 2012, pp. 5533-5537.

[4] H. E. Egilmez, S. T. Dane, K. T. Bagci, and A. M. Tekalp, “OpenQoS: an OpenFlow controller design for multimedia delivery with end-to-end Quality of Service over Software-Defined Networks,” Proc. Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC 2012), Dec. 2012, pp. 1-8.

[5] “OpenFlow switch specification version 1.2,” Open Networking Foundation, Dec. 2011.

[6] A. Tootoonchian and Y. Ganjali, “HyperFlow: a distributed control plane for OpenFlow,” Proc. the 2010 internet network management conference on research on enterprise networking (INM/WREN’10), 2010.

[7] F. Koch and K. T. Hansen, “Redundancy performance of virtual network solutions,” Proc. IEEE Conference on Emerging Technologies and Factory Automation (ETFA’06), Sept. 2006, pp. 328-332.

[8] “Virtual Router Redundancy Protocol (VRRP),” IETF RFC3768, April 2004.

[9] R. Zhang, T. F. Abdelzaher, and J. A. Stankovic, “Efficient TCP connection failover in web server clusters,” Proc. IEEE International Conference on Computer Communications (INFOCOM’04), vol. 2, March 2004, pp. 1219-1228.

[10] OpenFlow 1.2 Tutorial, https://github.com/CPqD/OpenFlow-1.2-Tutorial [retreived: April, 2013].

[11] J. Kempf, E. Bellagamba, A. Kern, D. Jocha, A. Takacs, and P. Skoldstrom, “Scalable fault management for OpenFlow,” Proc. IEEE International Conference on Communications (ICC2012), June 2012, pp. 6606-6610.

[12] “The Transport Layer Security (TLS) Protocol Version 1.2,” IETF RFC5246, August 2008.

[13] “TRANSMISSION CONTROL PROTOCOL,” IETF RFC793, September 1981.

Comparative Analysis of Three Systems with Imperfect Coverage and Standby Switching Failures

Tseng-Chang Yen
Department of Applied Mathematics
National Chung-Hsiung University
Taichung, Taiwan, R.O.C
+886 4 22860133x623
tycen@amath.nchu.edu.tw

Kuo-Hsiung Wang
Department of Computer Science
and Information management
Providence University
Taichung, Taiwan, R.O.C.
+886 4 26328001x18122
khwang@pu.edu.tw

Wu-Lin Chen
Department of Computer Science and
Information Management
Providence University
Taichung, Taiwan, R.O.C.
+886 4 26328001x18109
wlchen@pu.edu.tw

Abstract—**The cloud computing is an emerging new computing paradigm which provides high reliability, high availability, and QoS-guaranteed computing services. The reliability and stability of power supply is one of the most important factors in successful cloud computing. In this paper, we compare three different configurations with imperfect coverage and standby switching failures based on the reliability and availability. The time-to-repair and the time-to-failure for each of the primary and warm standby components are assumed to be exponentially distributed. We derive the explicit expressions for the mean time-to-failure, *MTTF*, and steady-state availability, for three configurations and perform a comparative analysis. Three configurations are ranked based on *MTTF*, steady-state availability, and cost/benefit where benefit is either *MTTF* or steady-state availability.**

Keywords-Reliability; Availability; Imperfect coverage; Standby switching failures

I. INTRODUCTION

Uncertainty is one of the important issues in management decisions. Maintaining a high or required level of reliability and/or availability is especially essential in information industry, communication systems, power plants, etc. With the increasing demand for computing resources, the computing paradigm has evolved from stand-alone computing, distributed computing, grid computing, to cloud computing. Cloud computing hosts and delivers services over the Internet, i.e., information is processed on servers located in cloud data center and cached temporarily on clients via the Internet. A data center usually consists of thousands of servers that are organized in racks and interconnected through gigabit ethernet or other fabrics. Data center consumes a lot of electricity to maintain its normal operation. The power consumption breakdown of a data center includes servers and storage systems, power conditioning equipment, cooling and humidification systems, and networking equipment. In this paper, we discuss the optimal configuration of power electricity for data centers in terms of reliability and availability. Cao [4] first introduced reliability concept into a queueing system with a repairable service station which has exponentially distributed lifetime and generally distributed repair time. The concept of the standby switching failures in the reliability with standby system was first proposed by Lewis [6]. The concept of coverage and its effect on the reliability and/or availability model of a repairable system has been introduced by several authors such as Amari, et al. [1], Arnold [3], Dugan [5], Trivedi [7], and etc. Moreover, the status and trends of imperfect coverage models and its associated reliability analysis techniques were introduced in Amari, et al. [2]. Wang and Chiu [9] investigated the cost benefit analysis of availability systems with warm standby units and imperfect coverage. Wang and Chen [8] performed comparative analysis of availability between three system with general repair times, reboot delay and switching failures. Wang et al. [11] studied the cost benefit analysis of series systems with warm standby components and general repair times. Recently, Wang et al. [10] performed comparisons of reliability and the availability between four systems with warm standby components, reboot delay and standby switching failures.

The problem considered in this paper is more general than the works of Wang et al. [11] and Wang et al. [12]. We first systematically develop the explicit expressions for the $MTTF_i$ and $A_{T_i}(\infty)$ to three configurations with imperfect coverage and standby switching failures. Next, efficient Maple computer programs are utilized to perform a parametric investigation. We provide extensive numerical results to study the effects of various values of system parameters to the cost/benefit ratios. Finally, we rank the configurations for the *MTTF*, the $A_T(\infty)$, and the cost/benefit, based on specific values of distribution parameters, as well as of the costs of the components.

II. PROBLEM STATEMENT

For the sake of discussion, we consider a data center require a 30MW power electricity, and assume that the electricity generation capacity of generators is available in units of 30MW, 15MW, and 10MW. To provide reliable and stable power supply, there are standby generators, and all active and standby generators are continuously monitored by a fault detecting device to identify if they fail or not. We also assume that standby generators are allowed to fail while inactive before they are put into full operation. Each of the active components fails independently of the state of the others and has an exponential time-to-failure distribution with parameter λ. Whenever an active component (or warm standby component) fails, it may be immediately detected and located with a coverage probability c, and the failed component is instantly replaced by a warm standby component with switchover time β_1 if any standby is available. We now assume that each of the available standby component fails independently of the state of all the others and has an exponential time-to-failure distribution with parameter α ($0 < \alpha < \lambda$). Moreover, we define the *unsafe failure* state of the system as any one of the breakdowns is *not covered*. We further assume that active-component failure (or standby-component failure) in the *unsafe failure* state is cleared by a reboot. Reboot delay is assumed to be exponentially distributed with parameter β_2 for an active component (or standby component). The system fails when the remaining electricity generation capacity is less than 30MW. We define such situation as the state of *safe failure*. We assume that there is always the possibility of failures during the switching from standby state to active state. Let us assume that the switching component has a failure probability q . Active components and standby components are considered to be repairable. Whenever a primary component or a standby component fails, it is immediately repaired based on a first-come, first-served (FCFS) discipline. The time-to-repair for each of the primary and warm standby components are assumed to be exponentially distributed with parameter μ . Once a component is repaired, it is as good as new. Further, failure times and repair times are independently distributed random variables.

We consider three configurations as follows: the first configuration consists of one 30 MW active component and one 30 MW warm standby component. The second configuration is composed of two 15 MW active components and one 15 MW warm standby component. We assume the standby component can replace either one of the initially working components in case of failure. The third configuration includes of three 10 MW active components and two 10 MW warm standby component.

III. PROBLEM SOLUTIONS

Let $P_{n,m}(t)$ be the probability that exactly n primary components and m standby components are working at time $t (t \geq 0)$, and let $P_{uf_i}(t)$ be the probability that the system is in unsafe failure states, where $i = 1, 2, 3, 4$.

A. Calculations for configuration 1

A.1. MTTF

Using Trivedi's concept (see Trivedi [7]) and Wang et al.' concept (see Wang et al. [10]), the state-transition-rate diagram of configuration 1 is shown in Figure 1. The probability vector $\mathbf{P}(t)$ of configuration 1 is defined as:

$$\mathbf{P}(t) = [P_{1,1}(t), P_{1,0}(t), P_{0,1}(t), P_{uf_1}(t), P_{uf_2}(t), P_{0,0}(t)] .$$

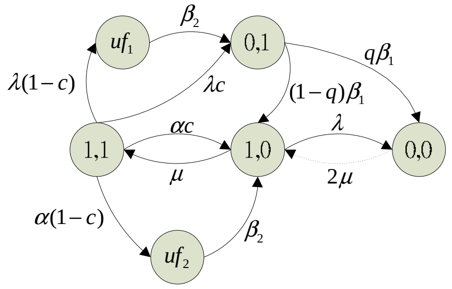


Figure 1. The state-transition-rate diagram of configuration 1

Relating the state of the system at time t and $t + dt$, the steady-state equations for configuration 1 can be expressed as follows:

$$d\mathbf{P}(t) / dt = B_1\mathbf{P}(t) , \qquad (1)$$

where

$$B_1 = \begin{pmatrix} -\lambda - \alpha & \mu & 0 & 0 & 0 & 0 \\ \alpha c & -\lambda - \mu & (1-q)\beta_1 & 0 & \beta_2 & 0 \\ \lambda c & 0 & -\beta_1 & \beta_2 & 0 & 0 \\ \lambda(1-c) & 0 & 0 & -\beta_2 & 0 & 0 \\ \alpha(1-c) & 0 & 0 & 0 & -\beta_2 & 0 \\ 0 & \lambda & q\beta_1 & 0 & 0 & 0 \end{pmatrix} .$$

To evaluate the *MTTF*, we take the transpose matrix of B_1 and delete the rows and columns for the absorbing state(s). The new matrix is called A_1. The expected times to reach an absorbing states is obtained from

$$E[T_{P(0) \to P(\text{absorbing})}] = \mathbf{P}(0)(-A_1^{-1})[1, 1, 1, 1, 1]^T , \qquad (2)$$

where the initial conditions are given by

$$\mathbf{P}(0) = [P_{1,1}(0), P_{1,0}(0), P_{0,1}(0), P_{uf_1}(0), P_{uf_2}(0)] = [1, 0, 0, 0, 0] ,$$

and

$$A_1 = \begin{pmatrix} -\lambda - \alpha & \alpha c & \lambda c & \lambda(1-c) & \alpha(1-c) \\ \mu & -\lambda - \mu & 0 & 0 & 0 \\ 0 & (1-q)\beta_1 & -\beta_1 & 0 & 0 \\ 0 & 0 & \beta_2 & -\beta_2 & 0 \\ 0 & \beta_2 & 0 & 0 & -\beta_2 \end{pmatrix} .$$

For configuration 1, the explicit expression for the $MTTF_1$ is given by

$$E[T_{P(0) \to P(\text{absorbing})}] = MTTF_1 .$$

This implies that

$$MTTF_1 = \frac{\Lambda_1}{\lambda\Delta} - \frac{-\alpha - \lambda + \lambda q}{\lambda\Delta} + \frac{\Lambda_1}{\beta_1\Delta} - \frac{\Lambda_1\Lambda_2}{\beta_2\Delta} - \frac{\alpha\Lambda_1\Lambda_2}{\lambda\beta_2\Delta} , \qquad (3)$$

where $\Lambda_1 = \lambda + \mu$, $\Lambda_2 = -1 + c$, and $\Delta = \mu q + \lambda + \alpha$.

A.2. Availability

To discuss the availability case of configuration 1, we use the following procedure to obtain the steady-state availability. In steady-state, the derivatives of the state probabilities become zero. Thus we have

$$\begin{pmatrix} -\lambda-\alpha & \mu & 0 & 0 & 0 & 0 \\ \alpha c & -\lambda-\mu & (1-q)\beta_1 & 0 & \beta_2 & 2\mu \\ \lambda c & 0 & -\beta_1 & \beta_2 & 0 & 0 \\ \lambda(1-c) & 0 & 0 & -\beta_2 & 0 & 0 \\ \alpha(1-c) & 0 & 0 & 0 & -\beta_2 & 0 \\ 0 & \lambda & q\beta_1 & 0 & 0 & -2\mu \end{pmatrix} \begin{pmatrix} P_{1,1}(\infty) \\ P_{1,0}(\infty) \\ P_{0,1}(\infty) \\ P_{uf_1}(\infty) \\ P_{uf_2}(\infty) \\ P_{0,0}(\infty) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{4}$$

Solving (4) and using the following normalizing condition

$$P_{1,1}(\infty)+P_{1,0}(\infty)+P_{0,1}(\infty)+P_{uf_1}(\infty)+P_{uf_2}(\infty)+P_{0,0}(\infty)=1,$$

we then obtain $P_{uf_1}(\infty)$, $P_{uf_2}(\infty)$, and $P_{0,0}(\infty)$.

Let T_1 represent the time-to-failure of the system for configuration 1. The explicit expression for the $A_{T_1}(\infty)=1-P_{uf_1}(\infty)-P_{uf_2}(\infty)-P_{0,0}(\infty)$ is given by

$$A_{T_1}(\infty)=\frac{2\beta_2\mu\left(\mu\beta_1+\beta_1\lambda+\beta_1\alpha+\mu\lambda\right)}{\beta_1\beta_2\lambda^2+2\mu\beta_1\beta_2\alpha+2\mu^2\Delta_1+\lambda\Delta_2} \tag{5}$$

where $\Delta_1=\beta_1\beta_2+\beta_1\alpha-\beta_1\alpha c+\beta_1\lambda-\beta_1\lambda c+\beta_2\lambda$ and $\Delta_2=\beta_1\beta_2\left(2\mu+\alpha+\mu q\right)$.

B. Calculations for configuration 2

B.1. MTTF

Using Trivedi’s concept (see Trivedi [7]) and Wang et al.’ concept (see Wang et al. [10]), the state-transition-rate diagram of configuration 2 is shown in Figure 2. The $\mathbf{P}(t)$ of configuration 2 is defined as:

$$\mathbf{P}(t)=[P_{2,1}(t),P_{2,0}(t),P_{1,1}(t),P_{uf_1}(t),P_{uf_2}(t),P_{1,0}(t)].$$

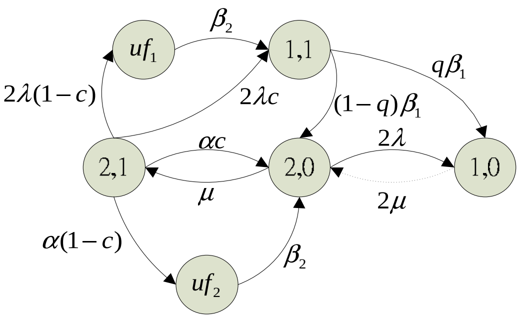


Figure 2. The state-transition-rate diagram of configuration 2.

Relating the state of the system at time t and $t+dt$, the steady-state equations for configuration 2 can be expressed as follows:

$$d\mathbf{P}(t)/dt=B_2\mathbf{P}(t), \tag{6}$$

where

$$B_2=\begin{pmatrix} -2\lambda-\alpha & \mu & 0 & 0 & 0 & 0 \\ \alpha c & -2\lambda-\mu & (1-q)\beta_1 & 0 & \beta_2 & 0 \\ 2\lambda c & 0 & -\beta_1 & \beta_2 & 0 & 0 \\ 2\lambda(1-c) & 0 & 0 & -\beta_2 & 0 & 0 \\ \alpha(1-c) & 0 & 0 & 0 & -\beta_2 & 0 \\ 0 & 2\lambda & q\beta_1 & 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

To evaluate the *MTTF*, we take the transpose matrix of B_2 and delete the rows and columns for the absorbing state(s). The new matrix is called A_2. The expected times to reach an absorbing states is obtained from

$$E[T_{P(0)\to P(\text{absorbing})}]=\mathbf{P}(0)(-A_2^{-1})[1,1,1,1,1]^T, \tag{8}$$

where the initial conditions are given by

$$\mathbf{P}(0)=[P_{2,1}(0),P_{2,0}(0),P_{1,1}(0),P_{uf_1}(0),P_{uf_2}(0)]=[1,0,0,0,0].$$

For configuration 2, the explicit expression for the $MTTF_2$ is given by

$$E[T_{P(0)\to P(\text{absorbing})}]=MTTF_2.$$

This implies that

$$MTTF_2=\frac{1}{2}\left(\frac{\Lambda_1}{\lambda\Delta_1}-\frac{-\alpha-2\lambda+2\lambda q}{\lambda\Delta_1}+\frac{2\Lambda_1}{\beta_1\Delta_1}-\frac{2\Lambda_1\Lambda_2}{\beta_2\Delta_1}-\frac{\alpha\Lambda_1\Lambda_2}{\lambda\beta_2\Delta_1}\right), \tag{9}$$

where $\Lambda_1=2\lambda+\mu$, $\Lambda_2=-1+c$, and $\Delta_1=\mu q+2\lambda+\alpha$.

B.2. Availability

For the availability case of configuration 2, we use the same procedure in 3.1.2 to obtain the steady-state availability. In steady-state, the derivatives of the state probabilities become zero. Thus we have

$$\begin{pmatrix} -2\lambda-\alpha & \mu & 0 & 0 & 0 & 0 \\ \alpha c & -2\lambda-\mu & (1-q)\beta_1 & 0 & \beta_2 & 2\mu \\ 2\lambda c & 0 & -\beta_1 & \beta_2 & 0 & 0 \\ 2\lambda(1-c) & 0 & 0 & -\beta_2 & 0 & 0 \\ \alpha(1-c) & 0 & 0 & 0 & -\beta_2 & 0 \\ 0 & 2\lambda & q\beta_1 & 0 & 0 & -2\mu \end{pmatrix} \begin{pmatrix} P_{2,1}(\infty) \\ P_{2,0}(\infty) \\ P_{1,1}(\infty) \\ P_{uf_1}(\infty) \\ P_{uf_2}(\infty) \\ P_{1,0}(\infty) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{10}$$

Solving (10) and using the following normalizing condition

$$P_{2,1}(\infty)+P_{2,0}(\infty)+P_{1,1}(\infty)+P_{uf_1}(\infty)+P_{uf_2}(\infty)+P_{1,0}(\infty)=1,$$

we then obtain $P_{uf_1}(\infty)$, $P_{uf_2}(\infty)$, and $P_{1,0}(\infty)$.

Let T_2 represent the time-to-failure of the system for configuration 2. The explicit expression for the $A_{T_2}(\infty)=1-P_{uf_1}(\infty)-P_{uf_2}(\infty)-P_{1,0}(\infty)$ is given by

$$A_{T_2}(\infty)=\frac{\beta_2\mu\left(\mu\beta_1+2\beta_1\lambda+\beta_1\alpha+2\mu\lambda\right)}{\beta_1\beta_2\Delta_1+\beta_1\mu^2\Delta_2+2\beta_2\mu^2\lambda} \tag{11}$$

where $\Delta_1 = \mu^2 + 2\mu\lambda + \mu\alpha + 2\lambda^2 + \lambda\alpha + \mu q\lambda$ and $\Delta_2 = \alpha - \alpha c + 2\lambda - 2\lambda c$.

C. Calculations for configuration 3

C.1. MTTF

Using Trivedi's concept (see Trivedi [7]) and Wang et al.' concept (see Wang et al. [10]), the state-transition-rate diagram of configuration 3 is shown in Figure 3. The $\mathbf{P}(t)$ of configuration 3 is defined as:

$$\mathbf{P}(0) = [P_{3,2}(t), P_{3,1}(t), P_{2,2}(t), P_{3,0}(t), P_{2,1}(t),$$

$$P_{uf_1}(t), P_{uf_2}(t), P_{uf_3}(t), P_{uf_4}(t), P_{2,0}(t)].$$

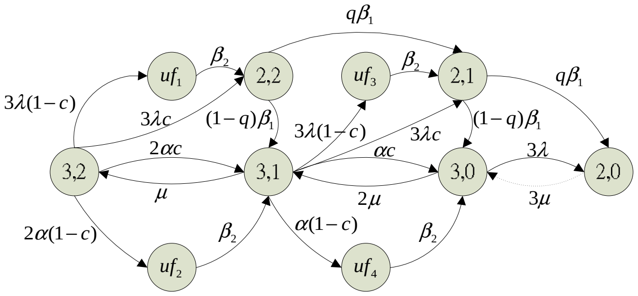


Figure 3. The state-transition-rate diagram of configuration 3.

For the reliability case, the initial conditions are:

$$\mathbf{P}(0) = [P_{3,2}(0), P_{3,1}(0), P_{2,2}(0), P_{3,0}(0), P_{2,1}(0),$$

$$P_{uf_1}(0), P_{uf_2}(0), P_{uf_3}(0), P_{uf_4}(0)]$$

$$= [1, 0, 0, 0, 0, 0, 0, 0, 0].$$

The following differential equations written in matrix form can be obtained:

$$d\mathbf{P}(t)/dt = B_3\mathbf{P}(t). \qquad (12)$$

Hence the matrix B_3 is an (10×10) square matrix whose last column is zero. The matrix B_3 is too spacious to be shown here. For the *MTTF*, we take the transpose matrix of B_3 and delete the rows and columns for the absorbing state(s). The new matrix shall be called A_3. The expected times to reach an absorbing states can now be calculated from

$$E[T_{P(0)\to P(\text{absorbing})}] = \mathbf{P}(0)(-A_3^{-1})[1,1,1,1,1,1,1,1]^T. \quad (13)$$

For configuration 3, the explicit expression for the $MTTF_3$ is given by

$$MTTF_3 = E[T_{P(0)\to P(\text{absorbing})}].$$

The mean time to system failure for configuration 3 $MTTF_3$ is too ample to be shown here.

C.2. Availability

For the availability case of configuration 3, the initial conditions are

$$\mathbf{P}(0) = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0].$$

The following differential equations written in matrix form can be obtained from

$$d\mathbf{P}(t)/dt = B_3\mathbf{P}(t), \qquad (14)$$

where the matrix B_3 can be formulated in a way similar to (12). It is an (10×10) square matrix whose last column, rather than being zero as in (12), is appropriately modified. The resulting matrix is too spacious to be shown here. In steady-state, the derivatives of the state probabilities become zero. That allows us to calculate the steady-state probabilities $P_{uf_1}(\infty)$, $P_{uf_2}(\infty)$, $P_{uf_3}(\infty)$, $P_{uf_4}(\infty)$, and $P_{2,0}(\infty)$ with the following normalizing condition

$$P_{3,2}(\infty) + P_{3,1}(\infty) + P_{2,2}(\infty) + P_{3,0}(\infty) + P_{2,1}(\infty)$$

$$+ \sum_{i=1}^{4} P_{uf_i}(\infty) + P_{2,0}(\infty) = 1.$$

Let T_3 represent the time-to-failure of the system for configuration 3. Again, the explicit expression for the $A_{T_3}(\infty) = 1 - P_{uf_1}(\infty) - P_{uf_2}(\infty) - P_{uf_3}(\infty) - P_{uf_4}(\infty) - P_{2,0}(\infty)$ is too spacious to be shown here.

IV. COMPARATIVE ANALYSIS

A. Comparison for the MTTF

The main purpose of this section is to present specific numerical comparisons for the $MTTF$. Using an efficient Maple program, three configurations will be compared in terms of their $MTTF_i$ $(i = 1, 2, 3)$ with the following values:

$$1/\lambda = 50 \text{ days},\ 1/\alpha = 200 \text{ days},\ \text{and } 1/\mu = 2 \text{ days},$$

or $\lambda = 0.02$, $\alpha = 0.005$, and $\mu = 0.5$.

We consider the following two cases to perform a comparison for the $MTTF_i$ of the configurations 1, 2, and 3.

Case 1: We fix α=0.005, μ=0.5, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary λ from 0.02 to 0.1.

Case 2: We fix λ=0.01, α=0.005, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary μ from 0.001 to 0.5.

The numerical results of $MTTF_i$ for each configuration i $(i = 1, 2, 3)$ are shown in Table 1 for cases 1 and 2.

Table 1. Comparison of the configurations 1, 2, 3 for $MTTF_i$

	Result
Range of λ	
$0.02 < \lambda < 0.04391$	$MTTF_3 > MTTF_1 > MTTF_2$
$0.04391 < \lambda < 0.1$	$MTTF_1 > MTTF_3 > MTTF_2$
Range of μ	
$0.001 < \mu < 0.12652$	$MTTF_1 > MTTF_3 > MTTF_2$
$0.12652 < \mu < 0.5$	$MTTF_3 > MTTF_1 > MTTF_2$

B. *Comparison for the* $A_{T_i}(\infty)$

In this section, we consider the following two cases to compare the $A_T(\infty)$ of the configurations 1, 2, and 3.

Case 1: We fix α=0.0005, μ=0.1, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary λ from 0.001 to 0.1.

Case 2: We fix λ=0.01, α=0.0005, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary μ from 0.01 to 0.5.

The numerical results of $A_{T_i}(\infty)$ for each configuration i $(i = 1, 2, 3)$ are shown in Table 2 for cases 1 and 2.

Table 2. Comparison of the configurations 1, 2, 3 for $A_{T_i}(\infty)$

	Result
Range of λ	
$0.0 < \lambda < 0.00005$	$A_{T_1}(\infty) > A_{T_3}(\infty) > A_{T_2}(\infty)$
$0.00005 < \lambda < 0.01656$	$A_{T_3}(\infty) > A_{T_1}(\infty) > A_{T_2}(\infty)$
$0.01656 < \lambda < 0.1$	$A_{T_1}(\infty) > A_{T_3}(\infty) > A_{T_2}(\infty)$
Range of μ	
$0.01 < \mu < 0.058489$	$A_{T_1}(\infty) > A_{T_3}(\infty) > A_{T_2}(\infty)$
$0.058489 < \mu < 0.5$	$A_{T_3}(\infty) > A_{T_1}(\infty) > A_{T_2}(\infty)$

C. *Comparison of all configurations based on their cost/benefit ratios*

The cost (C_i) of the configuration i $(i = 1, 2, 3)$ are listed in the following:

$$C_1 = \$48 \times 10^6,\ C_2 = \$39 \times 10^6,\ C_3 = \$42 \times 10^6$$

Consider the following two cases¸ we perform a comparison for the cost/benefit ratios, namely, $C_i / MTTF_i$ and $C_i / A_{T_i}(\infty)$ for each configuration i $(i = 1, 2, 3)$. The results are depicted in Figures 4-7, respectively.

Case 1: We fix α=0.0005, μ=0.1, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary λ from 0.001 to 0.1.

Case 2: We fix λ=0.01, α=0.0005, $q = 0.1$, $c = 0.9$, $\beta_1 = 3.0$, $\beta_2 = 2.4$ and vary μ from 0.01 to 0.5.

Figure 4 and Figure 5 show that the $C_i / MTTF_i$ and $C_i / A_{T_i}(\infty)$ increase as λ increases for any configuration. We observe from Figure 4 that the optimal configuration using the $C_i / MTTF_i$ value depends on the value of λ. When $\lambda < 0.0574$, the optimal configuration is configuration 3, but when $\lambda > 0.0574$, the optimal configuration is configuration 1. One observes from Figure 5 that the optimal configuration using the $C_i / A_{T_i}(\infty)$ value depends on the value of λ. When $\lambda < 0.0418$, the optimal configuration is configuration 2, when $0.0418 < \lambda < 0.0757$, the optimal configuration is configuration 3, and when $\lambda > 0.0757$, the optimal configuration is configuration 1.

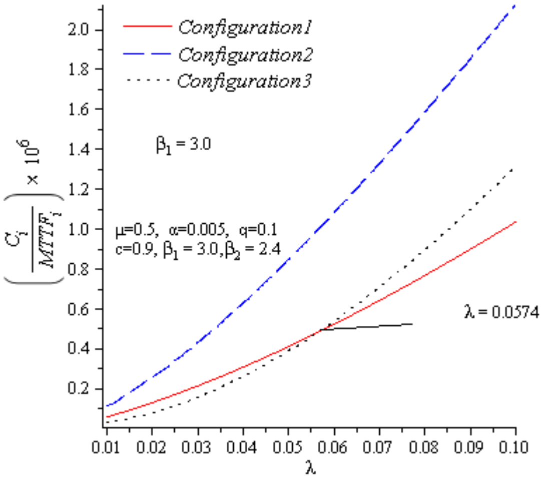


Figure 4. $C_i / MTTF_i$ versus λ.

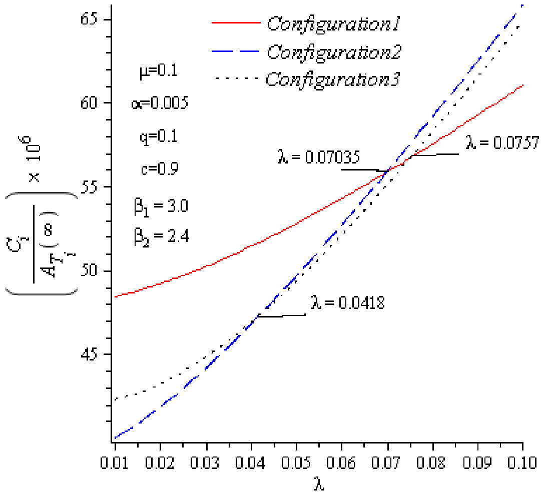


Figure 5. $C_i / A_{T_i}(\infty)$ versus λ.

We can easily see from Figure 6 and Figure 7 that the $C_i / MTTF_i$ and $C_i / A_{T_i}(\infty)$ decrease as μ increases for any configuration. Figure 6 reveals that the optimal configuration using the $C_i / MTTF_i$ value depends on the value of μ. When $\mu < 0.0979$, the optimal configuration is configuration 1, but when $\mu > 0.0979$, the optimal configuration is configuration 3. We observe from Figure 7 that the optimal configuration using the $C_i / A_{T_i}(\infty)$ value depends on the value of μ as well. When $\mu > 0.0241$, the optimal configuration is configuration 2.

ISBN: 978-1-61208-300-1

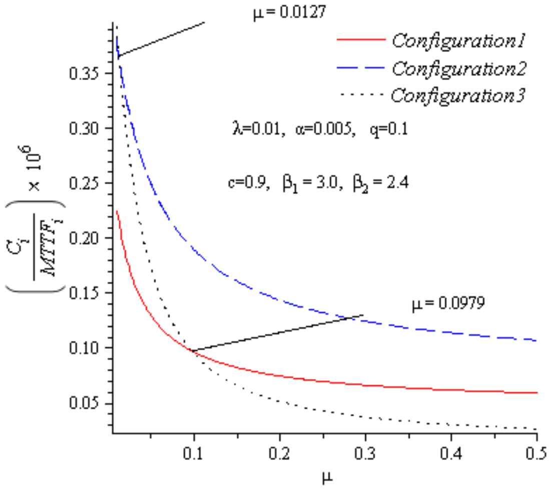


Figure 6. C_i / $MTTF_i$ versus μ .

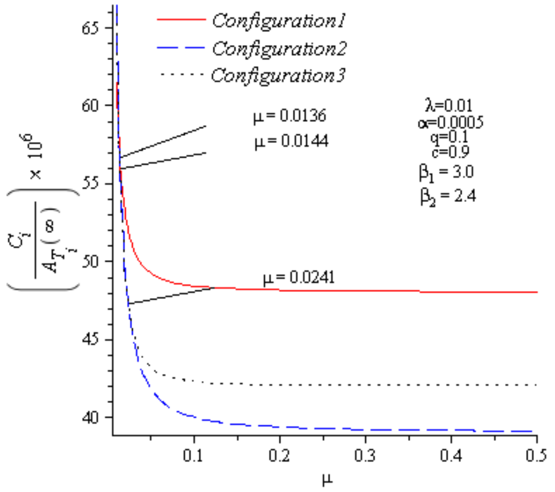


Figure 7. C_i / $A_{T_i}(\infty)$ versus μ .

V. CONCLUSIONS

In this paper, we analyzed three different configurations with imperfect coverage and standby switching failures to study the cost/benefit analysis of three configurations under uncertainty. For each configuration, we present the explicit expressions for the $A_{T_i}(\infty)$ and the *MTTF*. We rank three configurations based on the $A_{T_i}(\infty)$, the *MTTF*, and the cost/benefit where benefit is either steady-state availability or *MTTF*

REFERENCES

[1] S. V. Amari, J. B. Dugan and R. B.Misre,“A separable method for incorporating imperfect fault-coverage into combinatorial models,” IEEE Trans. Reliab., vol. 48, pp. 267-274, 1999.

[2] S. V. Amari, A. F. Myers, A. Rauzy and K. S. Trivedi, Imperfect coverage models: status and trends, in Nisra, K.B. Handbook of Performability Engineering, Springer, Berlin, pp. 321-348, 2008.

[3] T. F. Arnold, “The concept of coverage and its effect on the reliability model of a repairable system, ”IEEE Trans. Comput., vol. C-22, pp. 251–254, 1973.

[4] J. Cao and K. Cheng, “Analysis of M/G/1 queueing system with repairable service station,”Acta Math. Applicate Sinica, vol. 5, pp. 113–27, 1982.

[5] J. B. Dugan and K. S. Trivedi,“Coverage modeling for dependability analysis of fault-tolerant systems,”IEEE Trans. Comput., vol. 38, pp. 775-787, 1989.

[6] E. E. Lewis, Introduction to reliability engineering. 2nd ed., Wiley, New York, 1996.

[7] K. S. Trivedi, Probability and Statistics with Reliability, Queueing and Computer Science Applications, 2nd ed., John Wiley and Sons, New York, 2002.

[8] K.-H. Wang and Y.-J. Chen, “Comparative analysis of availability between three system with general repair times, reboot delay and switching failures,” Appl. Math. and Comput., vol. 215, pp. 384–394, 2009.

[9] K.-H. Wang and L.-W. Chiu, “Cost benefit analysis of availability systems with warm standby units and imperfect coverage,” Appl. Math. and Comput., vol. 172, pp. 1239-1256, 2006.

[10] K.-H. Wang W.-L. Don, and J.-B. Ke, “Comparison of reliability and the availability between four systems with warm standby components and standby switching failures,”Appl. Math. and Comput., vol. 183, pp. 1310-1322, 2006.

[11] K.-H. Wang, Y.-C. Liou and W. L. Pearn,“Cost benefit analysis of series systems with warm standby components and general repair times,”Math. Methods Oper. Res., vol. 61, pp. 329–343, 2005.

[12] K.-H. Wang and W. L. Pearn,“Cost benefit analysis of series systems with warm standby components,”Math. Methods Oper. Res., vol. 58, pp. 247-258, 2003.

ISBN: 978-1-61208-300-1

Method to Minimize Redundancy of Intra-Mode

Kibaek Kim, Dongjin Jung, and Jechang Jeong
Department of Electronics and Computer Engineering
Hanyang University
Seoul, South Korea
k2b0002@hanyang.ac.kr, dling@naver.com,
jjeong@hanyang.ac.kr

Gwanggil Jeon
Department of Electronics and Computer Engineering
University of Incheon
Incheon, South Korea
ggjeon@gmail.com

***Abstract*— In this paper, we present a method to minimize redundancy of intra-prediction mode. To minimize spatial correlation, H.264/AVC standard utilizes intra-prediction approach, which has nine modes for 4x4, 8x8 blocks, and this mode information should be signaled and four bits are needed to represent nine modes in binary. To minimize the average length of mode information, H.264/AVC estimates Most Probable Mode (MPM) and if the MPM is the same as the best intra-predicted mode, only one bit needs to be signaled. In this paper, we propose merging MPM approach to reduce the bits for signaling. By using adaptive scheme of intra-mode signaling, we could achieve 0.801% bit reduction while giving similar performance. In particular, 1.901% bit reduction was achieved in low bit rate condition.**

Keywords-H.264/AVC; Intra-prediction; Coding efficiency; most probable mode.

I. INTRODUCTION

To reduce the coded information of an image within a video sequence, Intra-prediction is an efficient tool. Intra-prediction in the spatial domain was proposed in several proposals and was involved into its current form of H.264/AVC [1-3]. Intra-prediction is to create a predictor block by extrapolation of neighboring coded block's pixels. The predictor block is subtracted from the target block and the residual components are coded by using transform, quantization and variable length coding. Since there are several directions to extrapolate for the target block, we need to determine the direction of prediction from neighboring coded block's pixels. H.264/AVC specifies a DC and 8 directional modes for 4x4 and 8x8 luminance blocks. Only DC, horizontal, vertical and planar modes are available for 16x16 luminance blocks and chrominance blocks. To save mode bit to be sent the decoder, H.264/AVC specifies MPM that estimate prediction mode by selecting direction having small mode number among modes between left and upper block. If the best mode is not equal to MPM, we send best mode to be selected among remaining modes except MPM. Then we need three bits to represent the best mode because one is selected among eight candidate modes (except MPM out of nine modes).

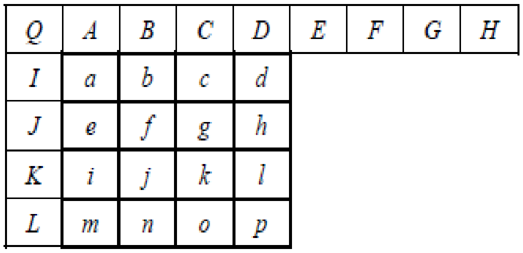
H.264/AVC provides several profiles to support various video services. Among the profiles, H.264/AVC supports baseline profile for low bit rate condition's devices. For low bit rate condition's devices, we can use high quantization parameter (QP) for coding video sequences. In low bit rate conditions, the importance of intra-prediction becomes higher, so we need to reduce the quantity of mode information. To reduce the quantity of mode information, Kim et al. [4] proposed an intra-mode skip method based on adaptive single-multiple prediction and Zhu et al. [5] proposed a clustering approach for reducing the number of intra-modes. By analysis on neighboring block's characteristics, it can reduce the intra-mode bits efficiently. Although some tool can be less complex in the encoder, analysis can be burden in the decoder complexity.

In the paper, we focus on the accuracy rate of MPM and we try to reduce mode bits efficiently. In the decoder, there is no additional analysis, so there is no complex increase. The reminder of this paper is organized as follows. In Section 2, we present the proposed algorithm. Simulation results for a variety of video sequences are provided in Section 3. Finally, conclusions are given in Section 4.

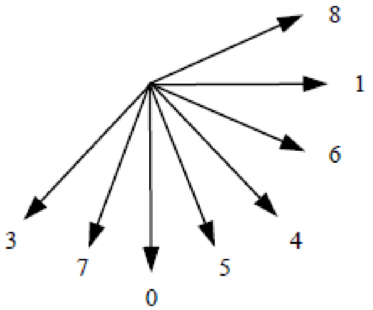
II. PROPOSED ALGORITHM

A. *Intra-prediction in H.264/AVC*

Intra-prediction is conducted in the transform domain, by referring to neighboring samples of previously coded blocks, which are to the left and above the block to be predicted. Macroblock (MB) is the basic coding unit of H.264 and its size is 16x16. According to image characteristic, various block size is applied for prediction. For luminance samples, intra-prediction may be formed for each 4x4 block or for each 8x8 block or a 16x16 block. There are a total of 9 optional prediction modes for each 4x4 and 8x8 luminance block; 4 modes a 16x16 luminance block. Similarly for chrominance 8x8 block, 4 modes are supported.



(a)



(b)

Figure 1. Labelling and direction of intra-prediction, (a) Neighboring pixels and pixels of current block (b) Eight directional prediction modes

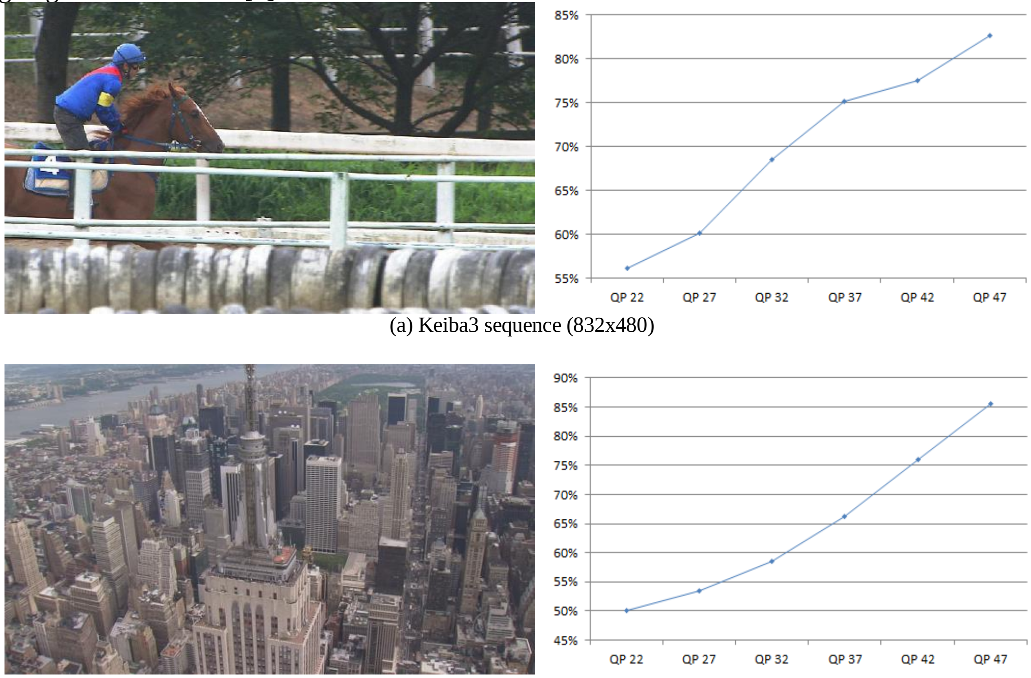
Figure 1 shows labeling and direction of intra-prediction. In Fig. 1(a), the small letter pixels (a~p) are the current 4x4 block and the capital letters (A~L, Q) are the prediction pixels, which are already decoded. Eight of the nine directional predictions are already shown in Fig. 1(b), where DC prediction (mode 2) that uses the average value of the eight boundary pixels (A~D and I~L) as the predictor is not shown. Each block is independently coded as one of the nine prediction modes. One of these nine modes is selected through the Rate-Distortion Optimization (RDO) process [6,7]. To efficiently compress the prediction mode in H.264/AVC, the prediction mode of the current block is estimated from the smaller directional prediction number between the above and left adjacent blocks of the current block. The estimated prediction mode is called the MPM. Just one bit flag is sent to the decoder if the MPM is equal to the best mode of the current block that is determined by computing a Lagrangian cost function [6] in the H.264/AVC encoder. Otherwise, an one bit flag indicating that the MPM does not match the best mode of the current block and an additional three bits are sent to the decoder to represent one of the eight directional modes.

B. *Background about statistic of intra-mode*

Figure 2 shows the sample of Keiba3 and City sequence and its accuracy rate of MPM, respectively. When QP becomes higher, the accuracy rate of MPM also becomes high. As shown in Fig. 3, bit rate is more important than distortion in low bit rate condition, which loss of residual data incurred by big quantization value. RDO is used to decide which mode is the most appropriate for each MB or block by minimizing the following equation:

$$J(s,c,Mode) = SSD(s,c,Mode) + \lambda \times R(s,c,Mode)$$
$$SSD = \sum_{x=i}^{m}\sum_{y=j}^{n}(s(x,y)-c(x,y))^2 \qquad (1)$$

where λ is the Lagrange multiplier for the mode decision and Mode indicates the mode chosen from the possible prediction mode candidates. SSD is the sum of squared differences between the original 4x4 block luminance signal (s) and its reconstruction signal (c) and R represents the number of bits associated with the chosen Mode. s(x,y) and c(x,y) denote the original luminance and reconstructed pixel values, respectively. When QP increases, λ also increases. Mode with short R (that is, MPM) is advantageous in low bit rate condition. High QP means an increase of quantization step size.



(a) Keiba3 sequence (832x480)

(b) City sequence (1280x720)

Figure 2. The sample of Keiba3 and City sequence and its accuracy rate of MPM

ISBN: 978-1-61208-300-1

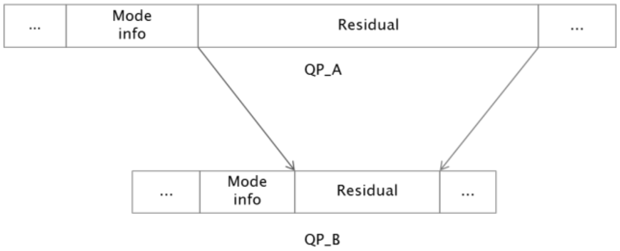
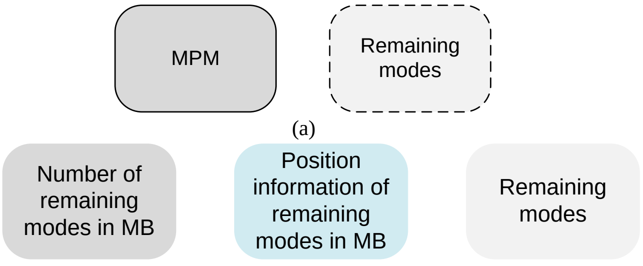


Figure 3. Change of bitstream according to QP increase (QP_A < QP_B)

Therefore, intra-prediction mode should be identical to MPM to reduce mode bits unless there is big difference in terms of distortion. The cases that block's intra-mode is equal to MPM increase according to the increase of QP as shown in Fig. 2. When QP is equal to 47, the accuracy rate of MPM in Keiba3 and City sequence reach 82% and 86%, respectively. It is obvious that the most sequence's accuracy rate of MPM increases according to the increase of QP. That the accuracy rate of MPM becomes higher, which means that entropy becomes lower. In other words, we can use this statistic to reduce the mode information.

C. *Proposed method*

Figure 4(a) shows the composition of intra-mode in H.264/AVC If MPM is selected as best mode of block, it requires one bit. Otherwise, we need four bits to represent best mode by using MPM and remaining modes. Since the accuracy rate of MPM increases according to the increase of QP as shown in Fig. 2, it is wasteful to send MPM per every block.


MPM
Remaining modes
Number of remaining modes in MB
Position information of remaining modes in MB
Remaining modes

(a)

(b)

Figure 4. Bit format of the proposed algorithm, (a) conventional method, (b) proposed method

We place the MPM into other flags to represent the mode information for MBs with a high accuracy rate of MPM. The proposed flag contains the number of MPM in MB along with the position information of the remaining modes. For MBs with a low accuracy rate of MPM, we use the conventional method in H.264/AVC. For distinguishing the boundary between a high and low accuracy rate of MPM, we set a specific threshold as the boundary. If the accuracy rate of MPM in an MB is higher than the threshold, the proposed method is chosen. Otherwise, the conventional method is chosen.

We changed the bit format as shown in Fig. 4. The conventional method to be applied to MB with low accuracy rate of MPM is described in Fig. 4(a), while the proposed method to be applied to MB with high accuracy rate of MPM is described in Fig. 4(b).

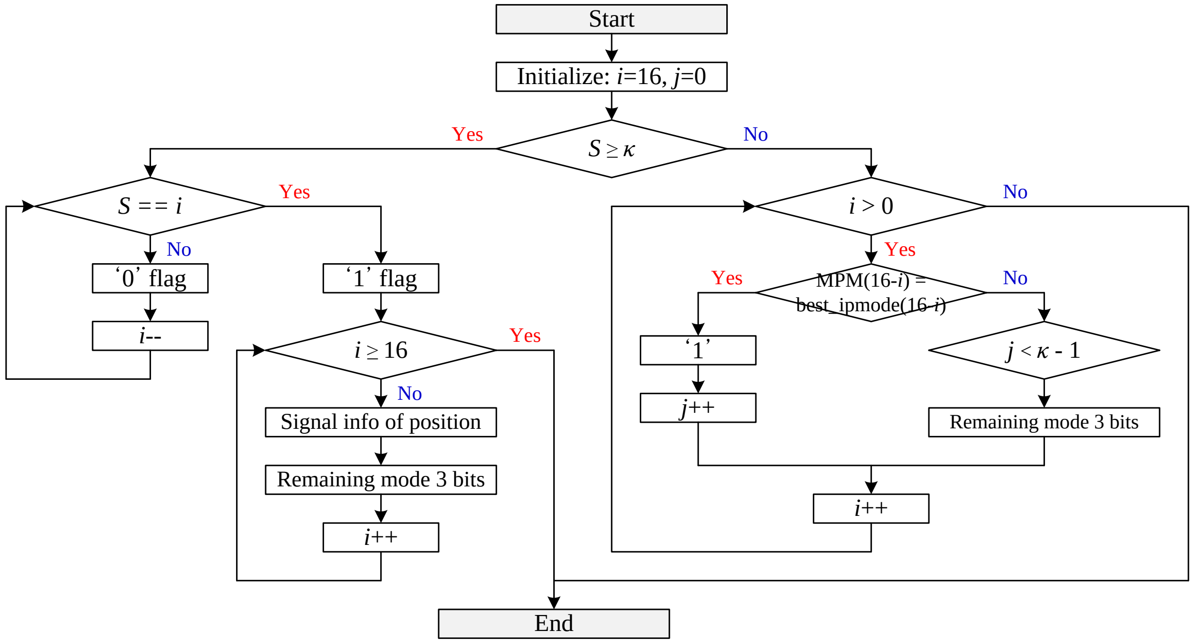


Figure 5. Flowchart of the proposed algorithm

The proposed method replaces MPM in additional flags. If the accuracy rate of MPM in an MB is higher than the threshold, then we send the number of blocks. This is because the best intra-mode in the block is different from MPM, and we denote this as MPM_{NB}, which means the best mode is not equal to MPM. If the best intra-mode is equal to MPM, we denote this as MPM_{BM}.

In the proposed method, blocks with MPM_{NB} are the target, so we should signal the position information of blocks with MPM_{NB}. Then, one mode among the remaining modes except MPM is coded. The proposed method is not effective if there are many blocks with MPM_{NB}. The threshold is determined empirically, and it should be set as a high number. If we set the threshold to thirteen, number of block with MPM_{NB} is equal to three. For number of remaining modes in MB, it requires three bits by using truncated unary binarization [8]. In order to represent the position information of block with MPM_{NB}, each block needs four bits since a MB has sixteen 4x4 blocks. There are three blocks with MPM_{NB}. We need twelve bits. And it requires additional nine bits to represent remaining mode of three blocks. If we apply this in the proposed method, it requires twenty-four bits. In conventional method, we need thirteen bits in blocks with MPM_{BM} and twelve bits in blocks with MPM_{NB}, respectively. Total twenty-five bits are required in conventional method. Through the comparison between two methods, we can know that proposed method can save one bit compared to the conventional method. However, when threshold is lower thirteen, there is no compression effect in the proposed method. This is the reason why threshold should be set as high number. We can know the range of threshold 13 to 16.

The encoder signals one bit per MB to distinguish whether or not the accuracy rate of MPM in an MB is over the threshold. Figure 5 shows the flowchart of the proposed algorithm, where S is the number of blocks to be MPM as the best mode, and κ is the threshold mentioned above. If the number of blocks with MPM_{BM} is less than κ, then conventional mode signaling is used (this is the same as the H.264/AVC case). The proposed algorithm is an adaptive scheme that is composed of the proposed intra-mode signaling combined with conventional intra-mode signaling. The method is determined by the number of blocks with MPM_{BM}. In Fig. 5, the left part shows the proposed method and the right part shows the conventional method.

III. EXPERIMENTAL RESULTS

The proposed method based on the accuracy rate of MPM was simulated in JM 16.0 reference software [9] in order to evaluate its performance. Various types of test sequences were used and a group of experiments were carried out on different QP ranges for evaluating coding efficiency in the low bit rate condition. Experimental conditions were: (a) one set of QP values were 22, 27, 32 and 37 and the other set of QP values were 32, 37, 42, 47, (b) Baseline profile was used, (c) number of frames was 100, and (d) entropy coder was CAVLC.

In this paper, we used κ=13 as the threshold value, which was determined empirically. All frames were coded as Intra. In order to evaluate the performance of the proposed method, it was compared with H.264/AVC. To calculate the efficiency, the proposed method was used to calculate the average BD-PSNR and BD-bitrate [10,11]. The RD performance comparisons are shown in Table 1.

TABLE I. CODING EFFICIENCY RESULTS WITH DIFFERENT QP RANGE

Sequence	Resolution	QP range 1		QP range 2	
		Bit-rate (%)	PSNR (dB)	Bit-rate (%)	PSNR (dB)
Foreman	CIF (352x288)	-0.099	0.006	-1.581	0.093
Container		-0.365	0.026	-1.551	0.092
Paris		-0.260	0.025	-1.219	0.087
Nuts5	WQVGA (416x240)	-1.125	0.062	-1.838	0.111
Nuts3		-1.355	0.080	-2.067	0.136
Keiba3		-0.496	0.035	-2.584	0.158
Flower4	WVGA (832x480)	-0.871	0.062	-1.331	0.074
BQmall		-0.561	0.037	-1.612	0.089
Keiba3		-0.627	0.035	-2.904	0.164
Bigships	HD (1280x720)	-0.223	0.012	-2.154	0.084
City		-0.181	0.013	-1.588	0.072
Crew		-1.486	0.060	-2.117	0.100
Jets		-2.185	0.088	-2.050	0.112
Vidyo3		-1.376	0.091	-2.023	0.138
Average		-0.801	0.045	-1.901	0.109

The average bit saving is 0.801% and the average PSNR gain is 0.045 dB for various test sequences in the Baseline profile. For additional information on the low bit rate condition, the proposed algorithm was also tested at QP range 2 (32, 37, 42, and 47). In QP range 2, the average bit saving is 1.901% and the average PSNR gain is 0.109 dB. The average bit saving increases according to the increase of QP. Especially, it is effective in some sequences. In the case of Foreman and City sequence, the bit saving of QP range 1 is 0.099% and 0.181%, respectively. It is relatively small compared to other sequences. However, the accuracy rate of MPM increases rapidly according to the increase of QP. In low bit rate condition, the bit saving of QP range 2 is 1.581% and 2.904%, respectively. Thus, we can verify that the proposed method is effective in the low bit rate condition. This means that the attempt to reduce redundancy of MPM is successful. As QP increases, the accuracy rate of MPM increases and the number of blocks with MPM_{BM} increase. Hence, mode signaling of the proposed method occurs often in the low bit rate condition.

In H.264/AVC, Intra-4x4 has an advantage regarding the PSNR because of the accuracy of its predictions. However, in the low bit rate condition, it also has a disadvantage in its bit rate because the mode information of 16 blocks is signaled. So, Intra-16x16 is selected as the best partition in the low bit rate condition. However, in the low bit rate condition in the proposed algorithm, Intra-4x4 is more often selected as the best partition compared to H.264/AVC because it can reduce the burden of bit rate. In other words, the number of bits used to represent the intra-mode information is reduced in the proposed algorithm. Table 2 shows the increase of Intra-4x4 partition compared to H.264/AVC. In the table, △I4x4 is used to calculate the number of Intra-4x4 partitions in the proposed algorithm compared to H.264/AVC. △I4x4 is calculated like following:

$$\Delta I_{4\times4} = \frac{I_{4\times4}^{prop} - I_{4\times4}^{H.264}}{I_{4\times4}^{H.264}} \times 100 \qquad (2)$$

TABLEⅡ. INCREASE OF INTRA-4X4 PARTITION COMPARED TO H.264/AVC

Sequence	Resolution	QP = 37	QP = 42
		ΔI_{4x4} (%)	ΔI_{4x4} (%)
Foreman	CIF (352x288)	6.680	11.425
Container		9.697	8.783
Paris		5.594	5.286
Nuts5	WQVGA (416x240)	17.53	36.922
Nuts3		27.910	49.278
Keiba3		6.681	16.130
Flower4	WVGA (832x480)	29.250	31.941
BQmall		9.050	17.125
Keiba3		10.621	25.798
Bigships	HD (1280x720)	17.999	38.404
City		5.796	23.488
Crew		36.205	57.326
Jets		25.880	27.542
Vidyo3		11.381	16.577
Average		15.734	26.145

When QP is equal to 37, the average of △I4x4 is 15.734%. It means that Intra-4x4 partition increases 15.734% compared to Intra-4x4 in H.264/AVC. In particular, when QP is equal to 42, Intra-4x4 partition increases 26.145% as shown in Table Ⅱ.

IV. CONCLUSION

In this paper, we proposed mode signaling to use the accuracy rate of MPM. An adaptive scheme is applied to the blocks according to number of block with MPM_{BM}. Two types of bit format are supported to code intra-mode information and it is selected based on the accuracy rate of MPM. Especially, the proposed algorithm can improve the coding efficiency in the low bit rate condition. According to the experimental results, the average bit reduction is 0.801% and the average PSNR gain is 0.045 dB. In particular, in the low bit rate condition, the average bit rate reduction is 1.901% and the average PSNR gain is 0.109 dB. This paper is enhancing the compression efficiency at low bit rate application such as video-conferencing and video telephony.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government(MOE) (NRF-2011-0011312)

REFERENCES

[1] G. Bjøntegaard, "Coding improvement by using 4x4 blocks for motion vectors and transform," ITU-T/Study Group 16/Video Coding Experts Group (Question 15), Eibsee, Germany, document Q15-C-23, December. 1997.

[2] G. Bjøntegaard, "Response to Call for Proposals for H.26L," ITU-T/Study Group 16/Video Coding Experts Group (Question 15), Seoul, Korea, document Q15-F-11, November. 1998.

[3] G. Conklin, "More Results on New Intra Prediction Modes," Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), Geneva, Switzerland, document JVT-B080, January. 2002.

[4] D. Kim, K. Han, and Y. Lee, "Adaptive Single-Multiple Prediction for H.264/AVC Intra Coding," IEEE Trans. on CSVT, vol. 20, no.4, April. 2010, pp. 610-615.

[5] W. Zhu, W. Ding, Y. Shi, Y. Sun, and B. Yin, “Adaptive intra modes reduction by clustering for H.264/AVC,” ICIP 2011, September. 2011, pp. 1665-1668.

[6] K. Takagi, Y. Takishima, and Y. Nakajima, “A study on rate-distortion optimization scheme for JVT coder,” in Proc. Int. Soc. Opt. Eng., vol. 5150, July. 2003, pp. 914-923.

[7] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” IEEE Signal Process. Mag., vol. 15, no. 6, November. 1998, pp. 74–90.

[8] D. Marpe, H. Schwarz, and T. Wiegand, “Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard,” IEEE Trans. on CSVT, vol. 13, no. 7, July. 2003, pp. 620-636.

[9] Available:http://iphome.hhi.de/suehring/download/old _jm.

[10] G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” ITU-T SG16 Q.6 Doc., VCEG-M33, Austin, US, April. 2001.

[11] S. Pateux and J. Jung, “Improvements of Excel macro for BD-gain computation,” ITU-T/Study Group 16/Video Coding Experts Group (Question 6), Geneva, Switzerland, document SG16-C358, October. 2009.

Reduction of Electricity Consumption and Electricity Demand Peak in Home Environments

Ana Rosselló-Busquet, José Soler and Lars Dittmann
Networks Technology & Service Platforms group, DTU Fotonik,
Technical University of Denmark, Kgs. Lyngby, Denmark
{aros, joss, ladit}@fotonik.dtu.dk

***Abstract*—There is an increase need to become more energy efficient and to reduce peak electricity demands to move towards the so called Smart Grid. In order to succeed, Home Energy Management Systems need to be developed to seamlessly provide the users with the tools to control the household electricity consumption. The Home Energy Management System developed includes a Home Gateway that provides rules to reduce the household consumption and a scheduling algorithm to reduce the peak demands. In addition, web services are used to provide remote access, communicate with the Smart Meter and the Energy Rules Server. Users will be able to donwload/update rules to reduce the electricity consumption and configure their Home Energy Management System remotely by using the JAVA based pilot-application developed. In this article the Home Energy Management System and its components are presented. The results show that the energy management system based on rules reduces the total househapld consumption. The scheduling algorithm results show that it is possible to distribute the consumption with some delays on the appliances tasks.**

***Index Terms*—Smart Grid, Home Energy Management, Home Gateway, Demand Response (DR), Demand Peaks**

I. Introduction

A Home Energy Management System (HEMS) is a system from which the user can control the devices in the home network through an Graphical User Interface (GUI) and apply energy management strategies to reduce and optimize their consumption. This article presents a HEMS, which helps reduce the household electricity consumption, and which includes a scheduling algorithm that spreads electricity consumption over time reducing demand peaks. Demand peaks are high points on power consumption caused by customers using electricity concurrently during the same period of time. From the utilities and electricity production perspective, demand peaks should be reduced or avoided as it is more efficient to have the power consumption distributed over time. Utility companies are forced to develop costly methods to generate enough power to meet the demand peaks. If the utilities cannot meet this demand, this could result in electricity shortages or even blackouts in certain areas.

The appliances found in users premises are usually manufactured by different producers and may use different communication technologies which can lead to interoperability issues between devices. Therefore, the main challenge in home networks is the variety of technologies, providing different communication methods, as well as the diversity of producers, providing different types of devices and services. The herein proposed Home Gateway is technology and device type independent, in order to offer a common pluggable platform to different devices in the home network, which makes them interoperable at the service level.

The HEMS main elements is the Home Gateway. The system can be accessed remotely and communicate with other components using web services:

- Remote access: users can connect to their Home Gateway, change the home appliances status and the energy management system settings among others.
- Communicate with other components of the HEMS: the HEMS can include other components beside the Home Gateway, such as the Smart Meter and bridges to communicate with other devices. Web services are used to handle this communication.
- Energy Rules Server: the Energy Rules Server will contain rules for reducing the electricity consumption. The user can donwload/update the energy rules in their HEMS through web services.

The Home Gateway presented in [1], [2], [3] has been extended to include a scheduling algorithm and the performance on terms of energy savings and energy reduction has been tested in this article. The scheduling algorithm which will distribute the electricity consumption over time. Different appliances are scheduled to consume based on their priority type, which can be changed by using the remote access. The overall goal of this method is to guarantee that a defined electricity consumption limit, provided by the utility and approved by the user, will not be exceeded. This technique could result in a more distributed consumption and lower demand peaks, which can lead to a reduction of greenhouse gasses. Additionally, the system presented can also ease the task of forecasting consumption as the customers will guarantee that they will not consume more than a determined amount of power.

In the past years, research has centered in home gateways for home automation and home energy management. An example of a home gateway using OSGi and ontologies is Domotic OSGi Gateway (DOG) [4] by Politecnico di Torino. The main difference between DOG and the home gateway herein presented is the fact that DOG is focused on domotics, while the home gateway herein is mainly concerned with energy management. The user can use it to define their own energy management system by creating, modifying and deleting rules, which may reduce the total electricity consumption. In [5], a HEMS has been implemented to reduce stand-by consumption by setting a power line network. Similarly, in [6], a HEMS implementation using ZigBee and infrared communication to reduce stand-by consumption of power outlets and lights is presented. The HEMS proposed in article paper has two main advantages over [5] and [6]: 1-the energy management strategy can help reduce the consumption and not only stand-by consumption; 2-the HEMS may communicate using different technologies and not only power line communication or ZigBee. SESAME-S [7] has developed a Home Gateway very similar to the one presented here. They have also used OSGi and have developed their own ontology for energy management. However, SESAME-S does not produce any results on the home gateway related to energy management. Their focus is in the user's acceptance of their system.

The remainder of this article is organized as follows: in Section II, an overview of developed HEMS is provided. Section III describes the rules used to reduce the household consumption and the electricity savings results are provided. The scheduling system and the results obtained are presented in Section IV. Finally, the conclusions are found in Section V.

II. Home Energy Management System

The aim of the HEMS developed is to reduce and schedule electricity consumption of the household. The Home Gateway developed offers the user a variety of basic functionalities:

- Monitor: The user can obtain the current status of any device connected to the home network through the user interface, locally or remotely.
- Control: The user can send basic commands, such as On/Off and parametric commands, such as *Start Program 3* to any home appliance, locally or remotely.
- Data Validation: The systems checks that the commands to be send to a device are actually valid for that device. For instance, if *Start Program 3* is send to a lamp the Home Gateway will detect the error and notify the user.

To build a HEMS and not only a home automation system more advanced and energy related functionalities are provided:

- Power Consumption History: in order to reduce electricity consumption it is important to know how much electricity each appliance consumes. The HEMS herein presented gathers this information.
- Energy Management System: the developed Home Gateway uses energy rules to help reduce the electricity consumption. The rules can be edited, deleted or created at any time and will be effective immediately without having to restart the system. The rules can be introduced into the system by the user. However, this requires that the user has advanced knowledge about the system, which is, in many case, unrealistic. Therefore, the Home Gateway communicates with a Energy Rules Server through web services. The user can connect to this server to browse the rules and an obtain a brief description of each rule so the most suitable rules, according to user's preferences, can be downloaded.
- Scheduling Algorithm: besides reducing the consumption, in order to be more energy efficient the electricity consumption should be distributed over time instead of creating demand peaks. This algorithm guarantees that the total electricity consumption in the household does not exceed a certain limit.

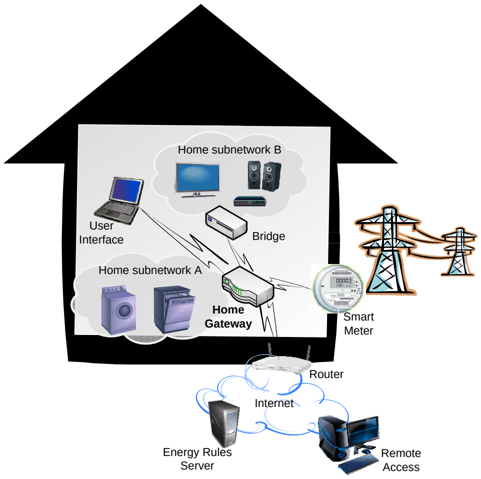


Fig. 1. System Architecture

The following subsections provide more details about the above functionalities and an overview of the system architecture and its implementation.

A. System Architecture

In the HEMS herein designed, the central element is the Home Gateway which can directly communicate with all the home devices or can use a bridge to interconnect with them. Besides being able to communicate with the home devices the Home Gateway can also communicate with the Smart Meter and will enable remote access through the internet. All the elements of this HEMS are depicted in Fig. 1.

B. Implementation

A Home Gateway that manages the home appliances to fulfill the reduction of overall consumption and reduction of peak demands has been developed using OSGi Equinox Framework [8]. The OSGi Framework is an open service platform for the delivery and control of different JAVA-based applications, called bundles. Each bundle has a specific functionality and can interact with other bundles in the same component of with other bundles in another component through web services. Web services are incorporated into the Home Gateway developed to offer modularity as some of the HEMS functionalities are external to the Home Gateway, such as Smart Meter or Energy Rules Server. To incorporate web services into the Home Gateway developed, Apache CXF Distributed OSGi [9] is used. This distribution enables an easy integration of web services into OSGi platform. Furthermore, CXF-DOSGi will auto-generate the Web Services Description Language (WSDL) from the java interface, at the deployment time. Further information of this implementation can be found in [3].

The Home Gateway accesses a knowledge base data repository from where the capabilities of the devices can be obtained. This knowledge base data repository is implemented by an ontology, where the home devices are classified according to their functionalities and capabilities. Using this knowledge

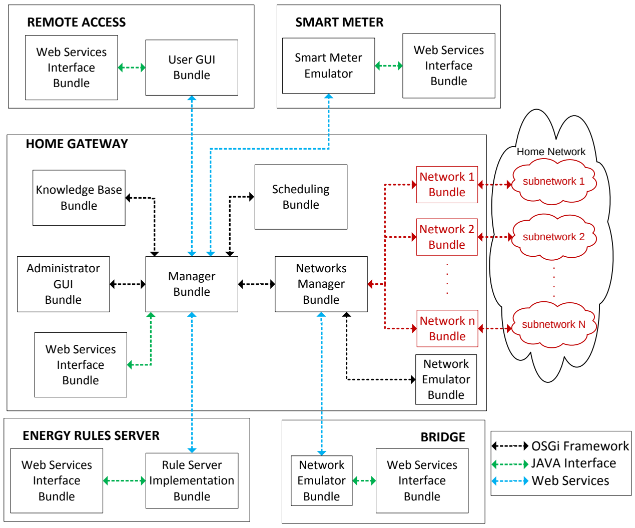


Fig. 2. OSGi Implementation

base data repository, energy management strategies can be performed by applying a set of rules, which are based on the energy consumption of the home devices, information from the electrical grid and users' preferences. The ontology has been included into the JAVA implementation by using Protégé-OWL API 3.4.4 [10]. The rules in this system have been written in SWRL which can be used to reason about the home devices in ontology terms. In order to run this rules from the JAVA platform Jess Rule Engine are used.

The components and its bundles are depicted in Fig. 2. In order to test the Home Gateway, some of the components, such as the Smart Meter, Energy Rules Server, bridges and home appliances have been emulated to test the Home Gateway capabilities.

A brief description of the Home Gateway bundles is provided below:

- Scheduling Bundle This bundle implements the scheduling algorithm and ensures that the maximum consumption is not pass a certain limit set by the user.
- Knowledge Base Bundle This bundle handles the interactions with the knowledge base data repository and rule engine. To implement the Home Gateway knowledge base data repository DogOnt [11] ontology is used. In addition, this bundle contains the means to apply energy management strategies by using rules.
- User Interface Bundle The Home Gateway developed provides a GUI which is contained in this bundle. This interface is used to communicate to the devices, obtain information about them and manage the energy management rules and the scheduling algorithm.
- Network n Bundle The home network found in the users' premises can contain devices using different communication technology, for example power line or wireless. Each of these *Network n* bundles will handle the communication with the devices in the subnetwork n in the home network which use different communication technologies. These bundles will send messages and forward notification messages to/from the devices contained in the n subnetwork.
- Network Emulator Bundle The focus of this article is not on the enabling technologies in the physical layer and their interoperation, but on the software mechanisms that allow use of the different elements, regardless of the connectivity mechanisms towards the Home Gateway. The home network is therefore emulated and an interface is provided to emulate changes in the devices status.
- Networks Manager Bundle Due to the fact that various Network n Bundles may exist, inside the Home Gateway and also in bridges, this bundle is created to handle the communication with these Network n bundles.
- Manager Bundle This is the central bundle which handles the interaction between the different bundles and contains the web services implementation and therefore acts as the server. It uses web services to communicate with bundles found outside the Home Gateway, such as the Smart Meter or the Energy Rules Server and OSGi framework to communicate within the Home Gateway.
- Web Services Interface Bundle It provides the JAVA interface needed to implement the web services.

In addition to the Home Gateway, other components have been developed to test the energy management system and the scheduling algorithm: a *Remote Access*, a *Bridge*, the *Energy Rules Server* and the *Smart Meter*. In order to emulate the remote access, a User Interface has been deployed in another computer, which communicates with the Home Gateway using web services.

The Home Gateway may not provide all the communication technologies found in the home network and the possibility of using bridges to communicate to some of the home appliances is a probable scenario. For this reason and to test the developed Home Gateway, a bridge has been implemented.

Energy Rules Server is used as a rules provider for the energy management system of the implemented Home Gateway. This bundle emulates a rules server that can be found in the internet.

A Smart Meter has been implemented to emulate the possible communication with the Home Gateway. This communication may involve electricity information, such as kWh price and a request to limit consumption, and user information, such us current electricity consumption and acceptance/rejection of requests.

III. Reduction of Electricity Consumption

To test the electricity savings the HEMS can provide, a set of basic rules, which aim to have no effect on the users' comfort, are implemented into the Home Gateway. All rules will be evaluated every time the knowledge base data repository is changed. For instance, if the Home Gateway receives a statues update, the status of the devices will be change in the knowledge base data repository, and then Jess rule engine will evaluate and trigger the necessary rules. When a rule is triggered and a device in a home network has to change status a message is send to the target device by suing the Manager bundle service and the corresponding Network bundle.

ISBN: 978-1-61208-300-1

The energy management strategy for this test is based on a few basic rules in terms of occupancy and irradiance. Therefore, it has been assumed that the home network contains a *presence system*, which by using sensors can indicate the presence of users in the home premises. Light sensors are also used to detect the solar irradiance. The rules used are summarized below:

- Lights and irradiance threshold: If the irradiance detected by the light sensors in a room is higher than the threshold, the lights in that room are turned off. The threshold can be modified according to user preferences through the user interface.
- Lights and no presence: If the sensors detect that there is no one present, all the lights are turned off.
- Standby: Standby power consumption is one of the major energy savings areas as the appliances are consuming without performing any task. Therefore, this rules makes sure that appliances are either turned off or on, but never in standby mode.
- Appliance and no presence: some of the appliances at home can be turned off while the user is away, for instance the printer and Wifi router. A rule for each appliance that the user wants to turn off while away from home has been implemented. This rules can be extended to more or less appliances according to users' preferences.
- Appliance and presence: in the similar way some appliances only have to be turned on when the user is away from the premises, for instance the answering machine or alarm system.

More rules can be added to the HEMS, for instance rules regarding heating, ventilation, and air conditioning (HVAC) systems. However, in other to implement these rules and emulate their consumption, detailed information about the HVAC systems, home architecture and home isolation is necessary. As this information is specific for each home environment and depends on a few factors it has not been included in the simulations presented in the next section.

A. Evaluation

In order to test the HEMS herein described, information about how users interact with the appliances is needed. Most models simulating power consumption in home environments provide time-correlated power consumption for the entire dwelling. However, in order to test the HEMS detail information about, when, for how long, and the instantaneous consumption of each device found in the home environment is needed. The model presented in [12] by Richardson et al. is used. This power consumption model is based on occupant time-use data, where occupant activity is mapped to appliance use. In the same way, detail information about light devices and its usage is also needed. To calculate the light usage, Richardson et al. take into consideration the solar irradiance. This model under-represents the demand during night, as it does not consider users leaving the lights on by mistake. In addition, this models provides time-correlated occupancy data, referred as active occupancy. Active occupancy is defined as the number of people who are at home and awake, this data input is used to model the presence system, which will have status present when active occupancy equals one or more.

TABLE I
HEMS INPUT PARAMETERS AND RESULTS

Cons. 28 days	Cons. w/ HEMS 28 days	Savings 28 days	Estimated Yearly Savings	Savings Percentage
216 kWh	178 kWh	38 kWh	500 kWh	17,72%

Richardson et al. provide an Excel Workbook [13] containing a high-resolution model of domestic whole house electricity demand. This implementation is used to calculate the use of home devices within a single UK dwelling over a 24-hour period at a one-minute time resolution. The simulator incorporates models to calculate the active occupancy and lighting usage. This implementation offers the option to configure the day and month of the year, the total number of users that live in the dwelling and whether a week day or a weekend day is simulated.

To test the work herein described an occupancy of 4 people has been considered. 28 days have been simulated to obtain samples to reproduce the consumption. This 28 days are divided into four weeks (5 week days and 2 weekend days) of each season (Summer, Autumn, Winter and Spring). These data has been used as input into the developed HEMS. A summary of the input parameters used and the results obtained can be found in Table I.

The herein developed HEMS using the rules presented in the previous section can successfully reduce the electricity consumption. Extrapolating the consumption of the 28 days simulated into a yearly electricity consumption, the electricity is reduced from 28.195 kWh to 23.198 kWh. This energy savings represent a decrease of 17,7%. The users could further reduce the electricity consumption by incorporating more rules to the HEMS. The new rules incorporated to the HEMS depend on the users' preferences and the home devices.

IV. SCHEDULING ALGORITHM

The aim of the implemented scheduling algorithm is to spread the electricity consumption and to keep the consumption under certain limit to reduced or even avoid electricity demand peaks. The scheduling bundle contains the scheduling algorithm and communicates with the Manager Bundle to handle the interaction with the rest of the bundles. The main concept behind this approach is the aggregation of home appliances into priority classes and the definition of a maximum power consumption limit, which is not allowed to be exceeded. If the user sets a maximum consumption, every time an appliance is turned on, a requests to consume is send to this bundle. This bundle will reply accepting or declining the request. This bundle is also capable of sending *pause* and *resume* commands to the household appliances. The *pause* command is used to force the appliance to go to into stand-by mode. The *resume* command is used to switch the appliance from stand-by mode to on, where the appliance will then continue its task. The Scheduling Bundle will decide which appliances can be paused and when they should continue their task by following the event driven scheduling algorithm illustrated in Fig. 3 and explained in detail in the next subsection.

A. *Event Driven Scheduling Algorithm*

The event driven scheduling algorithm showed in Fig. 3 is used to keep the total consumption of the household under the determined limit. The algorithm is triggered by two events: request to consume and end of consumption. An end of consumption event is send from Manager bundle when an appliance has been turned off.

The Manager bundle sends a request to consume to the Scheduling bundle when the users switch the appliance on. The Scheduling Bundle then calculates if there is enough power to switch on that appliance without exceeding the maximum consumption. If the consumption of the appliances already switched on, plus the consumption of the new appliance does not exceed the maximum consumption, then the appliance requesting to consume electricity is switched on. On the other hand, if switching on the new appliance would exceed the maximum consumption, the algorithm proceeds to examine the priorities of the appliances already switched on. The bundle tries to find a subset of the appliance(s) switched on, which have lower priority than the appliance requesting to consume. This subset of appliances should free enough electricity consumption so the new appliance can be switched on without exceeding the limit. If a subset fulfilling this condition is found, the appliances in the subset is paused and added to the paused appliances list. The bundle grants access to the new appliance without exceeding the maximum power consumption. In the case, a subset of appliances could not be found, the new appliance is added to the paused appliances list.

When an appliance is switched off, an end of consumption event is send to this bundle. The bundle then checks the paused appliances list. If one or more paused appliances can be switched on without exceeding the maximum power consumption, a *resume* command is send to the paused appliance(s).

In order to get users to accept that the users' low priority appliances will be paused and therefore take longer for them to finish their task, utilities could offer them a reduction in the electricity bill. This can be used as a commercial strategy by utilities to face a more homogeneous consumption by using attractive pricing schemes. It has to be taken into consideration that electricity bills are increasing along with the number of electrical appliances and users are interested in reducing their electricity bill. In particular, during the winter of 2007/08, 20% of Americans could not pay on time their electricity bill and 8.7 million American consumers were disconnected from their electricity utility services [14].

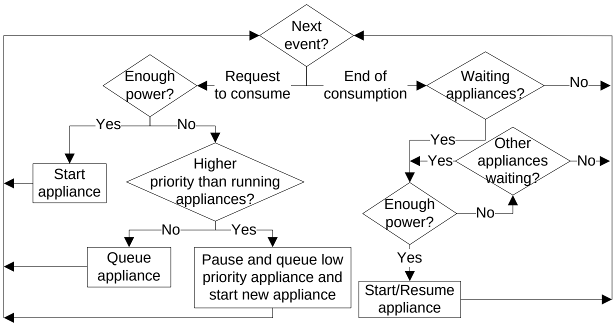


Fig. 3. Event Driven Scheduling Algorithm

TABLE III
RESULTS

	Scenario A	Scenario B
Max Consumption	1000 W	750 W
Mean waiting time	19,3 min	69,6 min

B. *Evaluation*

The delay suffered by the low priority appliances due to the fact that there is a limit on the maximum consumption and therefore, in some occasions, these appliances will have to wait before they can consume power is evaluated in this section. This delay is referred as *waiting time*.

As the priorities of appliances can considerably vary from user to user a simplified scenario that contains a television set, a computer, a washing machine, a dryer and a dishwasher has been considered. In addition, the appliances have been septated into these two priorities: high priority appliances (television set, computer) and low priority appliances (washing machine, dryer and dishwasher). The high priority appliances are considered to need electricity as soon as they are turned on, they cannot be denied power and cannot be paused or delayed in any way. On the other hand, the low priority appliances can be paused and/or delayed as it has been considered that their task duration could be prolonged without affecting the users' comfort. This scenario has been evaluated separately from the energy management system as it does not affect the total consumption, it only limits the instant power consumption.

The household appliances considered in this scenario are presented in Table II. This table also includes the consumption of each appliance, the average usage of the appliance and the duration of this usage. The average usage of the appliance has been taken from [15]. This data attempts to model the appliances usage of a typical day of a family (4-6 persons) between 17:00-00:00, when most of electrical consumption takes place.

100 test cases of one day have been done for each two different maximum consumption 750W and 1000W. The mean waiting time of the low priority appliance for both scenarios is summarized in Table III. The waiting time is the time the low priority appliances have been in *pause* mode, which can happen before the appliance even starts its task or during the task. The waiting can be seen as some extra time the low priority appliances will take to finish their task. For instance, for Scenario A, the washing machine will take in average 149,3 min minutes instead of 130 minutes, which means an increase of 14,8%. For Scenario B, the waiting time is considerably higher, 69,6 minutes.

V. CONCLUSION

The main motivation was to create a simple Home Gateway which would be easily scalable and that had the necessary capabilities to create a HEMS. The HEMS developed will not only offer control of devices through a GUI and run rules to reduce energy consumption, but it will also use web services to communicate with their components and also to

TABLE II
HOME APPLIANCES CHARACTERISTICS

Appliances	Model	Power Consumption	Average Usage [15]	Usage Duration
Television Set	Television: 42LE4900 LG [16] DVD player: DVX550 LG [17] Home theater: S-HS111US Pioneer [18]	239 Watt	4,3 hours/day	120 min
Computer	PC: HP Pavilion Slimline s5670t [19] Monitor: BX2340 Samsung [20]	242 Watt	3,5 hours/day	100 min
Washing machine	WM12S32XEE Siemens [21]	733 Watt	3,1 times/week	131 min
Dryer machine	WTW8658XEE Bosch [22]	609 Watt	4,4 times/week	134 min
Dish washer	SMS69T25EU Bosch [23]	720 Watt	4,1 times/week	100 min

communicated with other components and offer remote access. The HEMS also includes an event driven scheduling algorithm for regulating electricity demand peak. The developed Home Gateway uses web services to offer communication with external devices, such as the Smart Meter and Energy Rules Server, in addition to remote access to the system. Using this remote access the user can change the settings of the energy management system and the scheduling algorithm. The user can modify, delete, or download new rules from the Energy Rules Server to reduce the electricity consumption using the GUI of this HEMS. Basic rules obtained from the Energy Rules Server are downloaded to the HEMS to test if the household electricity consumption is reduced. In the considered scenario, the electricity consumption for a 4 people household consuming 28.195 kWh/year is reduced a 17,7%.

The scheduling algorithm to reduce demand peaks is also tested. In this case, two scenarios are considered, one with a maximum power consumption of 750 W and the other with 1000 W. This scenario will happen when the Home Gateway receives a request to limit the household consumption from the Smart Meter through web services. Using the scheduling algorithm will ensure that the maximum consumption is not exceed, and demand peaks intensity can be reduced. This, however, comes with a cost for the customer, which will have to accept that the low priority appliances take longer to finish. This waiting time depends on the allowed maximum power consumption. The lower the maximum consumption is the longer the low appliances will take to finish their task. For the first scenario, where the limit was set to 750 W, the waiting time of low priority appliances is of 69,6 minutes, which is considerably high, an increase of 53,5%. However, when the limit is set to 1000 W, this waiting time is reduced considerably, 19,3 minutes. This represents an increase of 14,8%, which is a more feasible scenario.

This article has proved that the HEMS developed successfully carry out its two main objectives: reduction the electricity consumption and reduce electricity demand peak in home environments without significatively disturbing the users' comfort.

REFERENCES

[1] A. Rosselló-Busquet, J. Soler, and L. Dittmann, "A Novel Home Energy Management System Architecture," in *13th International Conference on Computer Modelling and Simulation (UKSim)*, April 2011.

[2] A. Rosselló-Busquet, L. J. Brewka, J. Soler, and L. Dittmann, "OWL Ontologies and SWRL Rules Applied to Energy Management," in *13th International Conference on Computer Modelling and Simulation (UKSim)*, April 2011.

[3] Ana Rossello Busquet and Jos Soler, "A Novel Web Service Based Home Energy Management System," in *Proceedings of the Third International Conference on Advances in Future Internet AFIN*, 2011.

[4] D. Bonino, E. Castellina, and F. Corno, "The DOG gateway: enabling ontology-based intelligent domotic environments," *Consumer Electronics, IEEE Transactions on*, November 2008.

[5] S. H. Ju, Y. H. Lim, M. S. Choi, J.-M. Baek, and S.-Y. Lee, "An efficient home energy management system based on automatic meter reading," in *Power Line Communications and Its Applications (ISPLC), 2011 IEEE International Symposium on*, April 2011.

[6] J. Han, C.-S. Choi, and I. Lee, "More efficient home energy management system based on zigbee communication and infrared remote controls," *Consumer Electronics, IEEE Transactions on*, February 2011.

[7] A. Fensel, S. Tomic, V. Kumar, M. Stefanovic, S. V. Aleshin, and D. O. Novikov, "Sesame-s: Semantic smart home system for energy efficiency," *Informatik-Spektrum*, vol. 36, pp. 46–57, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00287-012-0665-9

[8] OSGi Alliance, "OSGi Service Platform Core Specification Release 4," Accessed Dec. 2010. [Online]. Available: http://www.osgi.org

[9] "Apache CXF Distributed OSGi," Accessed April 2011. [Online]. Available: http://www.w3.org/TR/ws-gloss/

[10] H. Knublauch, "Protege-OWL API Programmer's Guide," Accessed Dec. 2010. [Online]. Available: http://protegewiki.stanford.edu/wiki/ProtegeOWL_API_Programmers_Guide

[11] D. Bonino and F. Corno, "DogOnt - Ontology Modeling for Intelligent Domotic Environments," *The Semantic Web - ISWC 2008*, 2008.

[12] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: a high-resolution energy demand model," Accessed April 2013. [Online]. Available: https://dspace.lboro.ac.uk/2134/5786

[13] I. Richardson and M. Thomson, "Domestic electricity demand model - simulation example," Accessed April 2013. [Online]. Available: https://dspace.lboro.ac.uk/2134/5786

[14] "The Smart Grid: An Introduction," U.S Department of Energy (DOE), Tech. Rep., 2008.

[15] "Residential Energy Consumption Survey Data," U.S. Energy Information Administration, Independent Statistics Analysis, Tech. Rep., 2008.

[16] "Model 32LE4900, Specifications, ECO," Accessed: 20/03/2012. [Online]. Available: http://www.lg.com/uk/tv-audio-video/televisions/LG-led-tv-42LE4900.jsp

[17] "Model DVX550, Specifications, POWER," Accessed: 20/03/2012. [Online]. Available: http://www.lg.com/uk/tv-audio-video/video/LG-dvd-player-DVX550.jsp

[18] "Model S-HS111US, Specifications," Accessed: 20/03/2012. [Online]. Available: http://www.pioneerelectronics.com/PUSA/Home/Home-Theater-Systems/S-HS111US

[19] "Model: Pavilion Slimline s5670t series, Specifications," Accessed: 20/03/2012. [Online]. Available: http://www.shopping.hp.com/webapp/shopping/store_access.do?template_type=series_detail&category=desktops&series_name=s5670t_series&jumpid=in_R329_prodexp/hhoslp/psg/desktops/promo_tile/3/dt_promo_tile3_s5670t_113

[20] "Model BX2340, Specifications," Accessed: 20/03/2012. [Online]. Available: http://www.samsung.com/uk/consumer/pcperipherals/monitors/professio-nal/LS23CBUMBV/EN/index.idx?pagetype=prd_detail&tab=specification

[21] "Model WM12S32XEE, Ficha tecnica," Accessed July 2012. [Online]. Available: http://www.siemens-home.es/WM12S32XEE.html

[22] "Model WTW8658XEE, Ficha tecnica," Accessed: 20/03/2012. [Online]. Available: http://www.bosch-home.es/WTW8658XEE.html

[23] "Model SMS69T25EU, Ficha tecnica," Accessed: 20/03/2012. [Online]. Available: http://www.bosch-home.es/SMS69T25EU.html

Collaborative Applications Platform Based on Secure P2P Networks

Chun-Hsin Wang, Chia-Chun Lien and Shao-Hua Lin
Department of Computer Science and Information Engineering
Chung Hua University, Hsinchu, Taiwan, R.O.C.
E-mail: chwang@chu.edu.tw; m10102025@chu.edu.tw; m10102026@chu.edu.tw

***Abstract*—The feature of collaborative applications is to utilize various resources distributed over Internet (or intranet). It is a good choice to develop collaborative applications based on Peer-to-Peer (P2P) networks, which can be applied to integrate resources over peers. Most of popular P2P networks focus on files or content sharing and security problems are not seriously considered, so they are not good enough for developing collaborative applications. In this paper, secure P2P networks are designed by authentication of joining peers and encrypted data communication. To prohibit misuse of resources, peers are classified into three levels with different priorities. Based on secure P2P networks, a scalable and flexible collaborative application platform composed of core services and user defined services is built. Various resources provided by peers can be easily used by service execution. Two collaborative applications are designed to demonstrate the use of services over peers. It can be expected that more creative collaborative applications will be designed based on the proposed platform.**

Keywords-Collaborative applications platform; Secure P2P Networks.

I. Introduction

Network applications have great variation from conventional client-server model to P2P networks. In client-server model, direct communications are not allowed between any two clients, but they may occur between any two peers in P2P networks. The P2P technology has been widely applied to the integration and sharing of network resources. The shared resources can be provided by any peers joining the system instead of few dedicated servers. More creative network applications need resources distributed over peers to work well. For example, some collaborative approaches [1-4] defend Distributed Denial of Service (DDoS) attacks and worm containment from Internet. In [5], collaborative application is applied for data fusion, in which data are from different radars, sensors, and processing nodes. P2P-assisted cloud [6-7] or cloud-assisted P2P [8-10] collaborative models are trying to integrate the services provided by P2P network systems and cloud computing systems. Theses examples reveal that network applications tend to prefer collaboration of joining nodes and require multi-types resources support such as files, content, storage, or computing power.

It is a good choice to develop collaborative applications based on P2P networks. Most of popular P2P network systems focus on files, content sharing, or computation power. For example, BitTorrent-like P2P systems [11] are implemented for file sharing, while PPLive [12] and PPStream [13] are designed for streaming content sharing. The well-known SETI@home [14] project tries to find intelligent life outside Earth by stealing computation power over peers. Security problems in P2P networks are not seriously considered. Peers can join in these P2P networks without any authentication. Malicious nodes can easily launch attacks, such as sybil attacks [15], which generate a large number of shadow identifies that control system operations. Due to the weak of security problems and limited types of resources sharing, current P2P networks are not good enough for developing collaborative applications.

For development of collaborative applications, it is important for peers to easily provide services. The services provided by peers can be defined as executable software modules which can be executed to satisfy requests from other peers. This feature is to fit the collaborative applications which can request some selected peers to provide different types of services for approaching purposes of applications. As a scenario may occur, peers have resources, but can not provide requested services. Some peers in the system should have the ability to publish new services to the requested peers. The available services must be scalable and allowed to be defined by users. The open source vuze P2P system [16] has the similar concept, which is a BitTorrent-like file sharing system implemented in Java. Users can develop and plug their software modules into the system, but current plug-in software modules still focus on enhancement of files, content sharing, or improvement of user friendly interfaces.

In this paper, P2P network systems are enhanced by authentication of joining peers with three levels of priorities and encrypted data communication. Based on secure P2P networks, a scalable and flexible collaborative application platform composed of core services and user defined services is built. Various resources provided by peers can be easily used for service execution. The system model is shown in Fig. 1. Graphic user interface of peers is implemented in shell-like commands to request other peers to execute services. Two collaborative applications are designed to demonstrate the use of services distributed over peers. More creative collaborative applications based on the proposed platform can be easily designed, such as distributed computing, location-aware applications, etc.

The rest of the paper is organized as follows. The proposed secure P2P networks are described in Section II. The collaborative applications platform and its services are described in Section III. Some possible collaborative applications based on the proposed platform are presented in Section IV. Web-based online management system is introduced in Section V. Finally, some concluding remarks and future work are given in Section VI.

ISBN: 978-1-61208-300-1

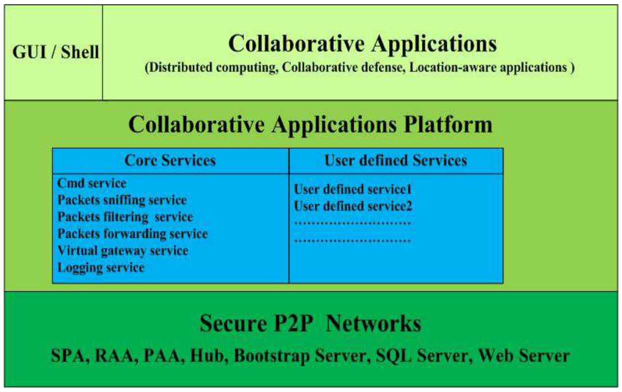

GUI / Shell
Collaborative Applications
(Distributed computing, Collaborative defense, Location-aware applications)
Collaborative Applications Platform
Core Services
User defined Services
Cmd service
Packets sniffing service
Packets filtering service
Packets forwarding service
Virtual gateway service
Logging service
User defined service1
User defined service2
Secure P2P Networks
SPA, RAA, PAA, Hub, Bootstrap Server, SQL Server, Web Server

Figure 1. System model

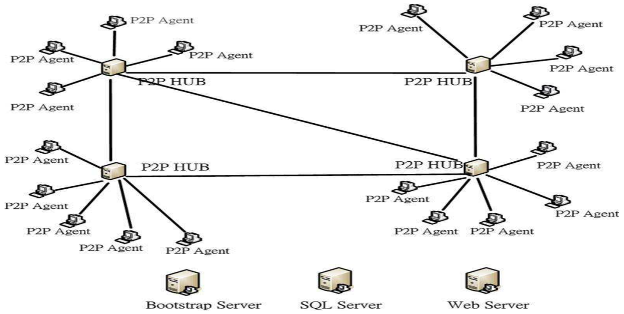


Figure 2. Overlay networks of secure P2P networks

II. Secure P2P Networks

In this section, we first introduce the model of the proposed secure P2P networks. Then, we describe the functions of components in it, system operation and implementation in detail. The proposed secure P2P networks (Fig. 2) are similar to KaZaA [17]. The main members of system are bootstrap server, P2P Hub nodes, and P2P agent nodes. To improve security of P2P networks, anonymous peers are not allowed. A database server (microsoft SQL) and web server are set up to maintain and manage related information of peers.

Peers are classified into three categories, namely Service Passive Agent (SPA), Request Active Agent (RAA), and Publish Active Agent (PAA). SPAs can provide services for RAAs and PAAs which they can issue requests to other peers. Besides the functions of SPA and RAA, PAA can publish new services to extend core services. The role of P2P Hub node is similar to cluster leader in KaZaA, which maintain the IP addresses of its members and associated resources-sharing. Hub nodes are selected from agents instead of hardware appliances. It means that some of agents have also to be roles of Hub nodes. Based on the open source library Lidgren [18], agents, Hub nodes, and bootstrap server are implemented in .net framework 4.0.

A. Bootstrap Server

The software architecture of bootstrap server and relationship of agents are shown in Fig. 3. Two components, ”bootstrap controller” and ”website controller”, can startup bootstrap service by graphic user interface in bootstrap server and by a online browser respectively. The Hub module will be started in a agent only when it is selected to be a Hub node. To maintain information of the owned agents(members) and how much data traffic is exchanged, each Hub node will set up a file-based SQLite [19] database. The maintained information in Hub nodes will be updated to the database in bootstrap server periodically. In our system, bootstrap server maintains the authorities and locations of joining peers, the list of existing Hub nodes, log information of peers, private/public keys for secure communication, and how much data traffic is exchanged in each Hub node. Note agents and Hub nodes connect to bootstrap server only for authentication or reporting information.

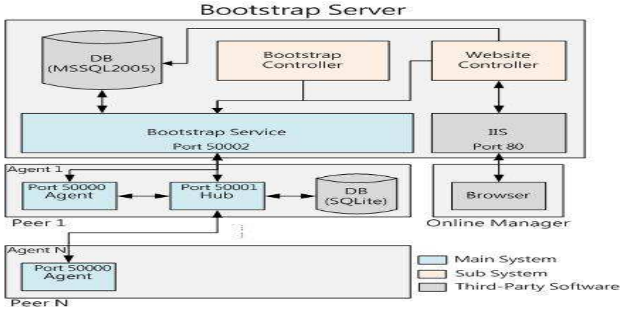


Figure 3. Software architecture of bootstrap server

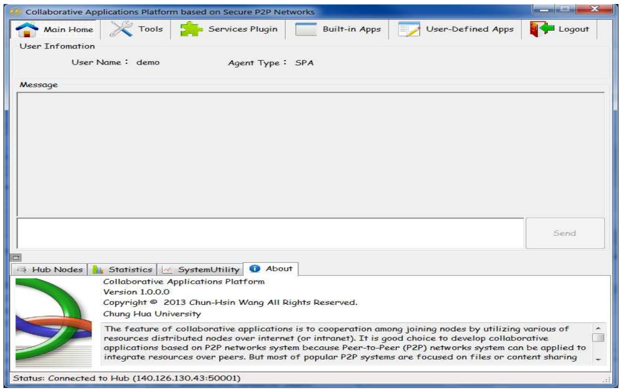
B. System Operation

The bootstrap server is initialized when secure P2P networks are starting. One peer can request it to acquire an account to join the system and becomes a SPA by default. Bootstrap server will give a list of Hub nodes according SPA’s location where its country and city are. To approach this purpose, the database of mapping IP address to geographic location [20] is adopted. If none of Hub nodes exists in the system, the SPA will become the first Hub node. Otherwise, SPA will measure Round Trip Times (RTTs) between it and Hub nodes in the given list. Then SPA will select the Hub node with minimum measured RTT to connect. When the selected Hub node is too far from the SPA (ex. RTT $\leq$ 50ms) or the members of it are too many (ex. over than 50), the SPA will become a new Hub node under performance and load balance consideration. Once a new Hub node is created, it will request bootstrap server to get five numbers of Hub nodes as their neighbors including three local neighbors the same country as it and two random Hub nodes in foreign countries if they exist. The new Hub node will connect their neighbors to join the system. When Hub node leaves system, members of it will automatically rejoin the system after random time. SPA can request bootstrap server to be a RAA or PAA and get related keys for secure communication.

C. Secure Communication

To construct secure P2P networks, message and data exchange among members of P2P system are encrypted. The communication between bootstrap server and the other agents is protected by AES [21]. The RSA [22] is implemented to protect the communication between two Hub nodes, between Hub node and its member, and between two agents. All kinds of agents and Hub nodes have to register in bootstrap server and then get the corresponding keys from it as follows.

- The private keys of AES for communication between bootstrap server and the others (agents and Hub nodes)



User Infomation

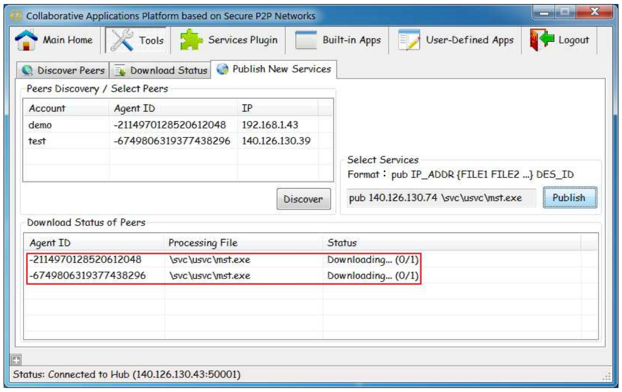
User Name： demo Agent Type： SPA

Collaborative Applications Platform
Version 1.0.0.0
Copyright © 2013 Chun-Hsin Wang All Rights Reserved.
Chung Hua University

The feature of collaborative applications is to cooperation among joining nodes by utilizing various of resources distributed nodes over internet (or intranet). It is good choice to develop collaborative applications based on P2P networks system because Peer-to-Peer (P2P) networks system can be applied to integrate resources over peers. But most of popular P2P systems are focused on files or content sharing

Status: Connected to Hub (140.126.130.43:50001)

Figure 4. GUI of agent



Peers Discovery / Select Peers

Account	Agent ID	IP
demo	-2114970128520612048	192.168.1.43
test	-6749806319377438296	140.126.130.39

Select Services
Format： pub IP_ADDR {FILE1 FILE2 ...} DES_ID

pub 140.126.130.74 \svc\usvc\mst.exe

Download Status of Peers

Agent ID	Processing File	Status
-2114970128520612048	\svc\usvc\mst.exe	Downloading... (0/1)
-6749806319377438296	\svc\usvc\mst.exe	Downloading... (0/1)

Status: Connected to Hub (140.126.130.43:50001)

Figure 5. Publish new service to selected peers

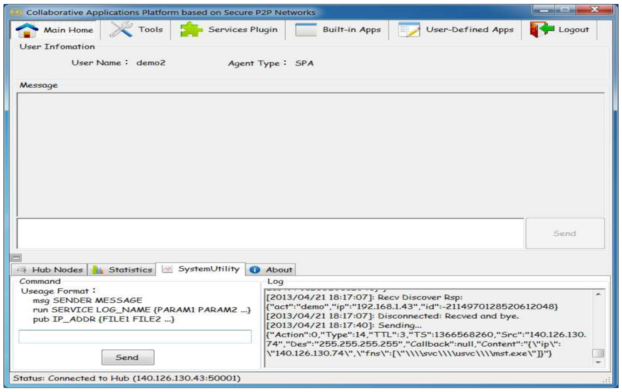
are generated by bootstrap server and given to agents when they log in the system.

- When agents or Hub nodes register to bootstrap server, their associated public and private keys will be generated and saved in database by bootstrap server for preparation of RSA communication.

- When agent want to connect a Hub node, it will query bootstrap sever to get public key of the Hub node first and then use it to encrypt a connection request. The Hub node can decrypt the connection request by its private key and send a request to bootstrap server for verifying the agent. If the agent exists, bootstrap server will response its public key. Then, the Hub node will accept the connection request and response message encrypted by public key of the agent. Otherwise, the connection request will be rejected.

- The connection between two Hub nodes are protected by RSA communication. It occurs when a new Hub node wants to connect its neighboring Hub nodes. In a similar way as mentioned above, the public keys of Hub nodes and verification of Hub nodes are also provided by bootstrap server.

- PAA or RAA agent can request other agents over the system to provide their services. To approach it, the first thing is to discover the associated agents by overlay networks. The public key of requesting agent can be sent with the request of discovering agents. Then, the found agents will response with their public keys back. The communication between two agents can be protected by RSA encryption.



User Infomation

User Name： demo2 Agent Type： SPA

```
Useage Format：
   msg SENDER MESSAGE
   run SERVICE LOG_NAME {PARAM1 PARAM2 ...}
   pub IP_ADDR {FILE1 FILE2 ...}
```

```
[2013/04/21 18:17:07]: Recv Discover Rsp:
{"act":"demo","ip":"192.168.1.43","id":-2114970128520612048}
[2013/04/21 18:17:07]: Disconnected: Recved and bye.
[2013/04/21 18:17:40]: Sending...
{"Action":0,"Type":14,"TTL":3,"TS":1366568260,"Src":"140.126.130.
74","Des":"255.255.255.255","Callback":null,"Content":"{\"ip\":
\"140.126.130.74\",\"fns\":[\"\\\\svc\\\\usvc\\\\mst.exe\"]}"}
```

Status: Connected to Hub (140.126.130.43:50001)

Figure 6. System utilities of agent

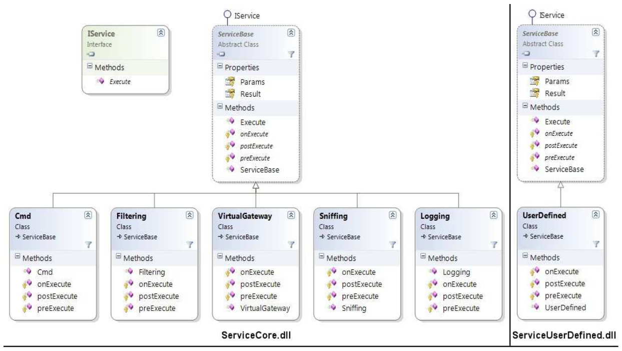


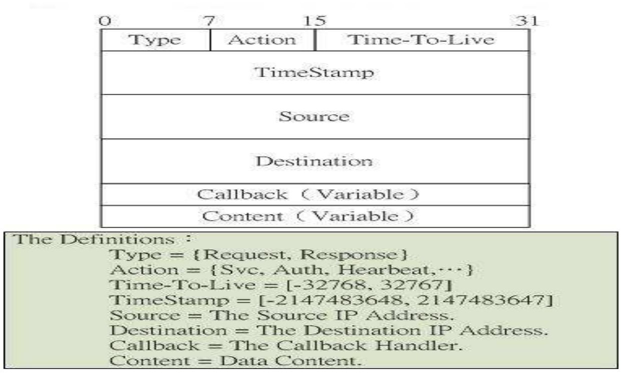
Figure 7. Class diagrams of core and user defined services

D. Implementation of Agents and Hub nodes

Although we have three different types of agents and Hub nodes, a universal Graphic User Interface (GUI) in Fig. 4 is designed for simplification of operations. User name, type of agent, and the connected Hub node are displayed when user logs in the system. A simple chatting function to send messages to all of peers is implemented. Unavailable functions can be seen but can not be used for agents without the priority. For example, the SPA agents can see the function of ”publish new services” but they are not granted to use it.

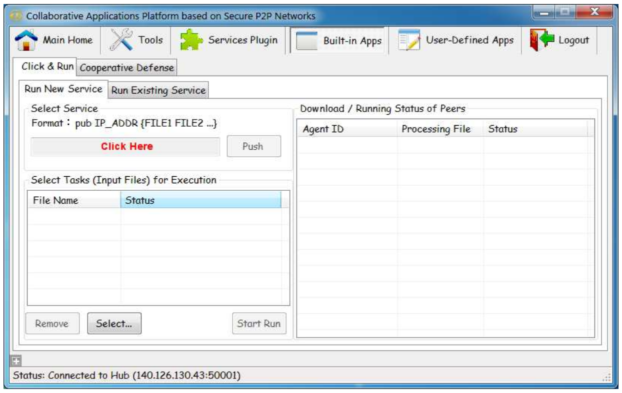
The menu bar on the top consists of ”Tools”, ”Services Plugin”, ”Built-in APPS”, ”User-Defined APPS”, and ”Logout”. Three frequently used tools are designed. The first tool is discovery of peers in a Time-to-Live range. The second tool is used to observe the current download status of services (or files) from other peers. The third tool is used to publish new services which only PAA agents have right to do it. Fig. 5 shows the tool of publishing new services defined by users to the selected peers. The function of ”Services Plugin” is used to check what core services and user defined services are plugged in. ”Built-in APPS” defines bulit-in collaborative applications we have designed. We will discuss it in next two section. ”User-Defined APPs” is reserved for embedding user defined collaborative applications in the future.

The sign of plus (”+”) down the middle of GUI can be clicked to show the optional utilities as in Fig. 4. We



The Definitions：
Type = {Request, Response}
Action = {Svc, Auth, Hearbeat,···}
Time-To-Live = [-32768, 32767]
TimeStamp = [-2147483648, 2147483647]
Source = The Source IP Address.
Destination = The Destination IP Address.
Callback = The Callback Handler.
Content = Data Content.

Figure 8. Application layer protocol between two peers



Select Service
Format：pub IP_ADDR {FILE1 FILE2 ...}

Select Tasks (Input Files) for Execution

File Name	Status

Download / Running Status of Peers

Agent ID	Processing File	Status

Status: Connected to Hub (140.126.130.43:50001)

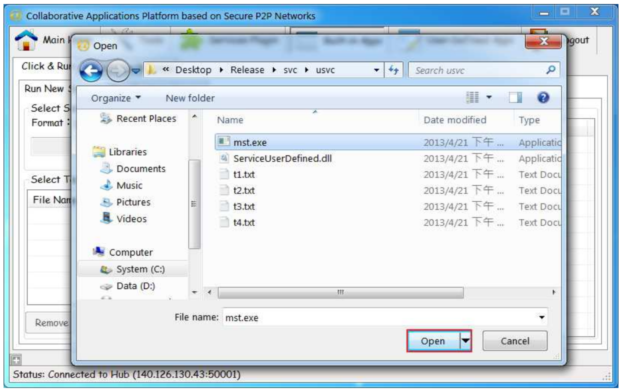
Figure 9. Click and open file browser to select program

can see the information of the proposed collaborative applications platform such as software version, introduction, etc. In addition, information of the connected Hub node can be seen such as its IP address, port number, etc. If the agent is also a Hub node, optional functions can be enabled to observe the statistical information how number of data bytes are received and transmitted. ”System Utility” (Fig. 6) can enable some manual commands to request other peers and show log information of history. Due to the page limitation, only few figures are shown in this paper.

III. Collaborative Applications Platform

Collaborative applications platform is constructed by services distributed over agents in secure P2P networks. It consists of two classes (Fig. 1), core services and user defined services which can be published to be plugged in peers. The core services are composed of basic and useful services we have implemented and will be keeping extension. For example, simple *cmd* service can ask peers to execute services (programs) in background by command shell. *Logging* service can save result of running services and send it back to requesting agents (RAA or PAA).

Some useful services for defending network security problems are implemented such as packet sniffing, filtering, and virtual gateway services. Peers can be asked to monitor and drop harmful packets according to signatures of packets by packet sniffing and filtering services respectively. Since peers can only verify packets they have received, the area of defending network security is limited. It motivates us to design the virtual gateway service which can redirect packets passing through neighboring nodes to the peer running the service. A peer running virtual gateway service will generate ARP reply packets [23] to cheat their neighboring nodes that the MAC address of the real gateway is the MAC address of it. Without the assistance of network switches, all packets from neighboring nodes will be forwarded to the virtual gateway. These core services can be applied to develop collaborative applications to defend network security problems.



« Desktop ▸ Release ▸ svc ▸ usvc

Search usvc

Name	Date modified	Type
mst.exe	2013/4/21 下午 ...	Applicatio
ServiceUserDefined.dll	2013/4/21 下午 ...	Applicatio
t1.txt	2013/4/21 下午 ...	Text Docu
t2.txt	2013/4/21 下午 ...	Text Docu
t3.txt	2013/4/21 下午 ...	Text Docu
t4.txt	2013/4/21 下午 ...	Text Docu

File name: mst.exe

Status: Connected to Hub (140.126.130.43:50001)

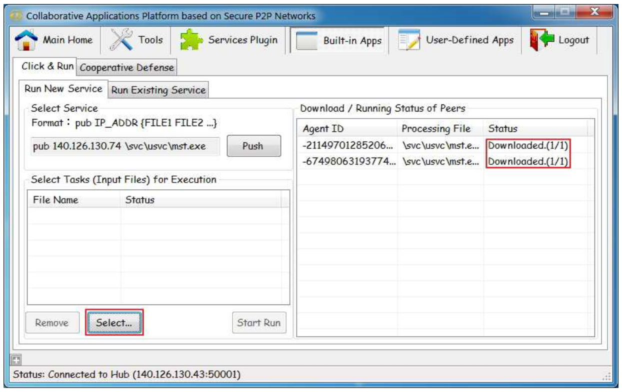
Figure 10. Select a program to be published

Besides core services we have designed, user defined services are allowed and easily plugged in peers. The services are packaged into two dynamic linking libraries. The class diagrams of core services and user defined services are shown in Fig. 7. Core services are provided by ServiceCore.dll and user defined services can be designed and implemented into ServiceUserdefined.dll. The library ServiceSDK.dll including two main components (Iservice and ServiceBase) is provided to let users define their services in ServiceUserdefined.dll and develop collaborative applications. PAA with highest priority can publish new services embedded in ServiceUserdefined.dll to the associated peers. In this way, the collaborative applications platform can be easily scalable. It can be expected that more creative collaborative applications based on our proposed platform can be developed.

After services in platform have been designed, application layer protocol is required to communicate between two peers. The format of protocol and related definition are shown in Fig. 8. Two types, request and response, are defined basically. The field of Time-to-Live is used to limit broadcast area of messages in overlay networks, which is defined by how number of Hub nodes is passing through. To avoid loop of message transmission, the time stamp is adopted to recognize whether the received messages are repeated or not. The filed of ”Action” is trying to define possible requests or responses of peers such as running services (*SVC*), authentication (*Auth*), Hello packets between two Hub nodes (*Heartbeat*), etc. The associated parameters and data with ”Action” can be set in the field of ”Content”. In addition, the filed of ”Callback” is used to assign a predefined program to handle the response of a corresponding request. New requests or responses of peers can be easily and flexibly added in our application layer protocol.

IV. Examples of Collaborative Applications

Based on the proposed collaborative applications platform, collaborative applications can be designed by use of resources



Format：pub IP_ADDR {FILE1 FILE2 ...}

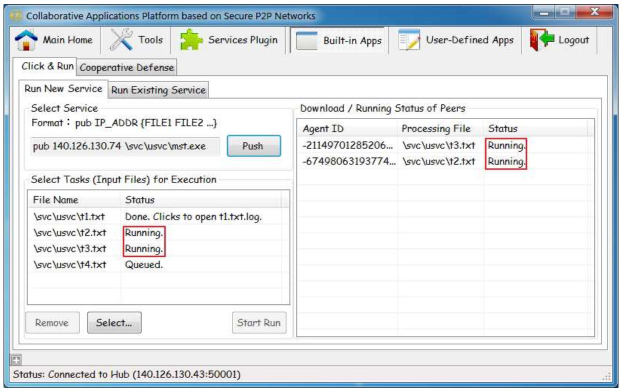
pub 140.126.130.74 \svc\usvc\mst.exe

File Name	Status

Agent ID	Processing File	Status
-21149701285206...	\svc\usvc\mst.e...	Downloaded.(1/1)
-67498063193774...	\svc\usvc\mst.e...	Downloaded.(1/1)

Status: Connected to Hub (140.126.130.43:50001)

Figure 11. Protocol format is automatically generated.



Format：pub IP_ADDR {FILE1 FILE2 ...}

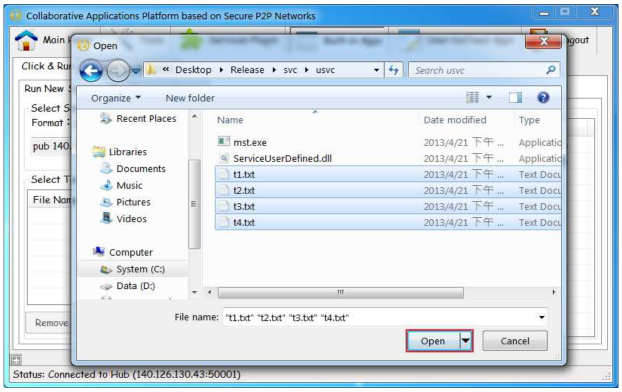
pub 140.126.130.74 \svc\usvc\mst.exe

File Name	Status
\svc\usvc\t1.txt	Done. Clicks to open t1.txt.log.
\svc\usvc\t2.txt	Running.
\svc\usvc\t3.txt	Running.
\svc\usvc\t4.txt	Queued.

Agent ID	Processing File	Status
-21149701285206...	\svc\usvc\t3.txt	Running.
-67498063193774...	\svc\usvc\t2.txt	Running.

Status: Connected to Hub (140.126.130.43:50001)

Figure 13. Execution status of each task

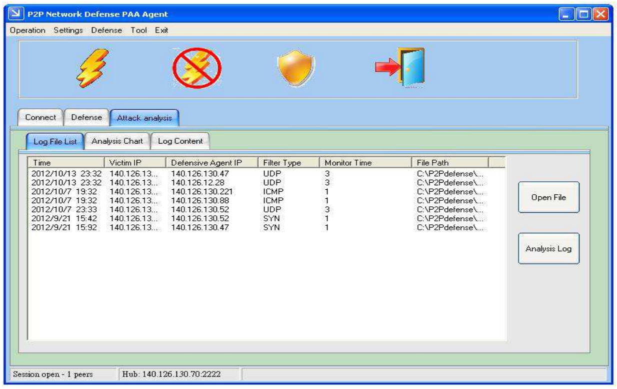


Name	Date modified	Type
mst.exe	2013/4/21 下午 ...	Applicatio
ServiceUserDefined.dll	2013/4/21 下午 ...	Applicatio
t1.txt	2013/4/21 下午 ...	Text Docu
t2.txt	2013/4/21 下午 ...	Text Docu
t3.txt	2013/4/21 下午 ...	Text Docu
t4.txt	2013/4/21 下午 ...	Text Docu

File name: "t1.txt" "t2.txt" "t3.txt" "t4.txt"

Status: Connected to Hub (140.126.130.43:50001)

Figure 12. Select input files by a file browser



Time	Victim IP	Defensive Agent IP	Filter Type	Monitor Time	File Path
2012/10/13 23:32	140.126.13...	140.126.130.47	UDP	3	C:\P2Pdefense\...
2012/10/13 23:32	140.126.13...	140.126.12.28	UDP	3	C:\P2Pdefense\...
2012/10/7 19:32	140.126.13...	140.126.130.221	ICMP	1	C:\P2Pdefense\...
2012/10/7 19:32	140.126.13...	140.126.130.88	ICMP	1	C:\P2Pdefense\...
2012/10/7 23:33	140.126.13...	140.126.130.52	UDP	3	C:\P2Pdefense\...
2012/9/21 15:42	140.126.13...	140.126.130.52	SYN	1	C:\P2Pdefense\...
2012/9/21 15:92	140.126.13...	140.126.130.47	SYN	1	C:\P2Pdefense\...

Session open - 1 peers

Hub: 140.126.130.70:2222

Figure 14. Attack logs from peers

distributed over peers in secure P2P networks. Applications on P2P networks are not limited to file sharing anymore. For demonstration of our platform, two built-in collaborative applications will be introduced and some possible useful collaborative applications will be further studied in the future.

The first buit-in collaborative application is a simple distributed computing application named click&Run. It can request all of peers to execute a new service or existing service. A new service can be a simple execution program. Users can focus on developing their execution programs and request other peers to execute them easily without knowledge of our platform. In our example, we want to request peers to compute minimum spanning trees of random graphs. The execution program, ”mst.exe”, can be viewed as a new service because the program is not available at other peers. We can simply use this application to select it from a file browser and then send to peers. The operations are shown in Fig. 9 and Fig. 10. The protocol format will be automatically generated instead of typing manually. After that, the ”push” button is used to send the program to all of peers. We can observe which peers have already downloaded the program as shown in Fig. 11. In our experiment, there are two peers in the system. Next thing we have to do is to select input files of the published program by a file browser (Fig. 12). When the ”start Run” button is clicked, each input file will be automatically assigned to one of peers for execution. Once peer has finished computing, executing result will be sent back and peer will be given the next input file till no more input files exist. The execution status of each task (input file) can be monitored and executing result can be opened by double clicks as shown in Fig. 13. Running existing services of peers can work in a similar way. This simple built-in collaborative application shows computing resource sharing.

The second buit-in collaborative application is trying to integrate various resources over Internet (or intranet) to defend network security. We have designed defensible services in our core services including the virtual gateway service, packet sniffing, filtering, and logging services. Without the assistance of network switches, peers can be the virtual gateway to investigate the packets from other nodes and then filter malicious packets. The investigated packets will be forwarded to the real gateway if they are valid. This application is migrated from our pervious work [24]. The integration of this collaborative application with our platform is still under working. We only show the idea here. Fig. 14 shows attack logs from virtual gateways which they sniff and send log files back to PAA by core services. The content of log files contains time stamp, IP addresses of victim and collaborative agents, traffic type, duration time of monitoring, and file names of logs. Another tool can display statistical volume of attack traffic according to source IP address as shown in Fig. 15. We can observe what IP addresses are suspects of attack origins. In the future, we will finish integrating this collaborative application into our platform.

V. WEB-BASED MANAGEMENT SYSTEM

For convenient management of our platform, a web server is set up. From the web server, users are allowed to register accounts and download software of agent and development libraries. A general user can get information of logging time,

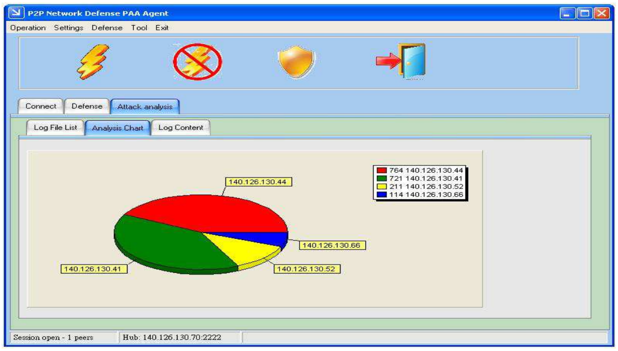
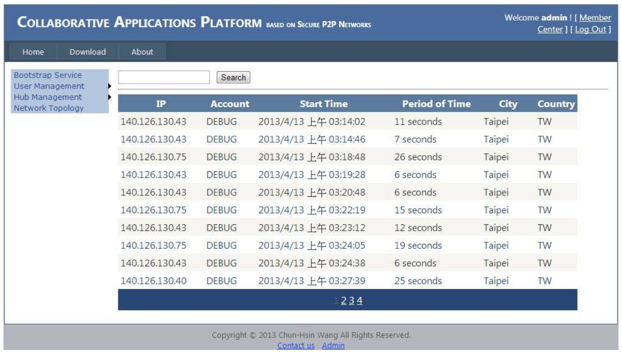


Figure 15. Analysis of attack logs


Collaborative Applications Platform based on Secure P2P Networks
Welcome admin ! [Member Center] [Log Out]
Home
Download
About
Bootstrap Service
User Management
Hub Management
Network Topology
Search

IP	Account	Start Time	Period of Time	City	Country
140.126.130.43	DEBUG	2013/4/13 上午 03:14:02	11 seconds	Taipei	TW
140.126.130.43	DEBUG	2013/4/13 上午 03:14:46	7 seconds	Taipei	TW
140.126.130.75	DEBUG	2013/4/13 上午 03:18:48	26 seconds	Taipei	TW
140.126.130.43	DEBUG	2013/4/13 上午 03:19:28	6 seconds	Taipei	TW
140.126.130.43	DEBUG	2013/4/13 上午 03:20:48	6 seconds	Taipei	TW
140.126.130.75	DEBUG	2013/4/13 上午 03:22:19	15 seconds	Taipei	TW
140.126.130.43	DEBUG	2013/4/13 上午 03:23:12	12 seconds	Taipei	TW
140.126.130.75	DEBUG	2013/4/13 上午 03:24:05	19 seconds	Taipei	TW
140.126.130.43	DEBUG	2013/4/13 上午 03:24:38	6 seconds	Taipei	TW
140.126.130.40	DEBUG	2013/4/13 上午 03:27:39	25 seconds	Taipei	TW

Figure 16. Web-based administration

volume of data traffic exchanged, and personal information. Administrator can startup bootstrap service, manage accounts of users, Hub nodes, and observe where users are from (Fig. 16). Statical volume of data traffic in each Hub node will be considered to design how the proper number of members is and how to select agents to be Hub nodes in the future.

VI. Conclusions and Future Work

In this paper, secure P2P networks are set up by authentication of joining peers and encrypted data communication. Although limitation of users is added, it is necessary for collaborative applications sensible of network security and resource sharing. Based on secure P2P networks, a scalable and flexible collaborative applications platform composed of core services and user defined services is built. Two collaborative applications based on the proposed platform are designed to demonstrate the use of services over peers. In the future, more core services will be added. The services may be combined with cloud services and location where peers are. It can be expected that more creative collaborative applications will be present.

Acknowledgment

This study was supported by the National Science Council of Taiwan, under the Grant No. NSC 101-2221-E-216-038.

References

[1] M. Cai, K. Hwang, Y. K. Kwok, S. Song, and Y. Chen, ”Collaborative Internet Worm Containment,”, *IEEE Security and Privacy*, May/June, 2005, pp. 25-33.

[2] M. E. Locasto, J. J. Parekh, A. D. Keromytis, and S. J. Stolfo, ”Toward Collaborative Security and P2P Intrusion Detection,” IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY, 2005, pp. 333-339.

[3] Y. Chen, K, Hwang, and W. S. Ku, ”Collaborative Detection of DDoS Attacks over Multiple Network Domains,” IEEE Transactions on Parallel and Distributed Systems, Vol. 18, No. 12, December 2007, pp. 1649-1661.

[4] S. Radwane, N. A. Farid, and S. Ahmed, ”A collaborative peer-to-peer architecture to defend against DDoS attacks,” the 33rd IEEE Conference on Local Computer Networks, Oct. 2008, pp. 427-434.

[5] P. Lee, A. P. Jayasumana, H. D. Bandara, S. Lim, and V. Chandrasekar, ”A peer-to-peer collaboration framework for multi-sensor data fusion,” Journal of Network and Computer Applications, May 2012, pp. 1052-1066.

[6] H. M. Xu, Y. J. Shi, Y. L. Liu, F. B. Gao, and T. Wan, ”Integration of Cloud Computing and P2P: A Future Storage Infrastructure,” International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), June 2012, pp. 1489-1492.

[7] J. Xu, J. Yan, L. He, P. Su, and D. Feng, ”CloudSEC: A Cloud Architecture for Composing Collaborative Security Services,” in 2nd IEEE International Conference on Cloud Computing Technology and Science, Dec. 2010, pp. 703-711.

[8] A. H. Payberah, H. Kavalionak, V. Kumaresan, A. Montresor, and S. Haridi, ”CLive: Cloud-Assisted P2P Live Streaming,” IEEE 12th International Conference on peer-to-peer Computing, Step. 2012, pp. 79-90.

[9] K. Graffi et al., ”Towards a P2P Cloud: Reliable Resource Reservations in Unreliable P2P Systems,” in 16th International Conference on Parallel and Distributed Systems, Dec. 2010, pp. 27-34.

[10] J. Dharanipragada and H. Haridas, ”Stabilizing peer-to-peer systems using public cloud: A case study of peer-to-peer search,” in 11th International Symposium on Parallel and Distributed Computing, June 2012, pp. 135-142.

[11] D. Harrison, BitTorrent homepage, http://www.bittorrent.org/, retrieved: July., 2013.

[12] PPLive Inc., PPLive homepage, http://www.pplive.com/, retrieved: July., 2013.

[13] PPStream Inc., PPStream homepage, http://www.ppstream.com/, retrieved: July., 2013.

[14] SETI, http://setiathome.berkeley.edu/index.php, retrieved: July, 2013.

[15] J. R. Douceur, The sybil attack, in First International workshop on peerto-peer systems, 2002, pp. 251-260.

[16] Vuze, http://wiki.vuze.com/, retrieved: July, 2013.

[17] J. Shi, J. Liang, and J. You, ”Measurements and Understanding of the KaZaA P2P Network,”, Current Trends in High Performance Computing and Its Applications, 2005, pp. 425-429.

[18] Lidgren networking library, http://code.google.com/p/lidgren-network-gen3/, retrieved: July, 2013.

[19] SQLite, SQL database engine, http://www.sqlite.org/, retrieved: July, 2013.

[20] MaxMind, Inc., http://dev.maxmind.com/geoip/geolite, retrieved: July, 2013.

[21] J. Daemen and V. Rijmen, ”The Design of Rijndael: AES V The Advanced Encryption Standard,” Springer, 2002.

[22] A. J. Menezes, P. C. V. Oorschot, and S. A. Vanstone, ”Handbook of Applied Cryptography,” CRC Press, the 5th printing in August 2001.

[23] Network Working Group ARP, RFC 826, 1982.

[24] C. H. Wang and C. W. Huang, ”A Collaborative Network Security Platform in P2P Networks,” International Conference on New Trends in Information and Service Science, June/July, 2009, pp. 1251-1256.