International Journal on Advances in Internet Technology

2011 vol. 4 nr. 1&2

The *International Journal on Advances in Internet Technology* is published by IARIA.
ISSN: 1942-2652
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Internet Technology
Volume 4, Number 1 & 2, 2011

Communication Theory, QoS and Reliability

- Adrian Andronache, University of Luxembourg, Luxembourg
- Shingo Ata, Osaka City University, Japan
- Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
- Michel Diaz, LAAS, France
- Michael Menth, University of Wuerzburg, Germany
- Michal Pioro, University of Warsaw, Poland
- Joel Rodriques, University of Beira Interior, Portugal
- Zary Segall, University of Maryland, USA

Ubiquitous Systems and Technologies

- Sergey Balandin, Nokia, Finland
- Matthias Bohmer, Munster University of Applied Sciences, Germany
- David Esteban Ines, Nara Institute of Science and Technology, Japan
- Dominic Greenwood, Whitestein Technologies AG, Switzerland
- Arthur Herzog, Technische Universitat Darmstadt, Germany
- Malohat Ibrohimova, Delft University of Technology, The Netherlands
- Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
- Joseph A. Meloche, University of Wollongong, Australia
- Ali Miri, University of Ottawa, Canada
- Vladimir Stantchev, Berlin Institute of Technology, Germany
- Said Tazi, LAAS-CNRS, Universite Toulouse 1, France

Systems and Network Communications

- Eugen Borcoci, University 'Politechncia' Bucharest, Romania
- Anne-Marie Bosneag, Ericsson Ireland Research Centre, Ireland
- Jan de Meer, smartspace®lab.eu GmbH, Germany
- Michel Diaz, LAAS, France
- Tarek El-Bawab, Jackson State University, USA
- Mario Freire, University of Beria Interior, Portugal / IEEE Portugal Chapter
- Sorin Georgescu, Ericsson Research - Montreal, Canada
- Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore
- Jong-Hyouk Lee, INRIA, France
- Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway
- Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
- Sjouke Mauw, University of Luxembourg, Luxembourg
- Reijo Savola, VTT, Finland

Future Internet

- Thomas Michal Bohnert, SAP Research, Switzerland
- Fernando Boronat, Integrated Management Coastal Research Institute, Spain
- Chin-Chen Chang, Feng Chia University - Chiayi, Taiwan
- Bill Grosky, University of Michigan-Dearborn, USA

- Sethuraman (Panch) Panchanathan, Arizona State University - Tempe, USA
- Wei Qu, Siemens Medical Solutions - Hoffman Estates, USA
- Thomas C. Schmidt, University of Applied Sciences – Hamburg, Germany

Challenges in Internet

- Olivier Audouin, Alcatel-Lucent Bell Labs - Nozay, France
- Eugen Borcoci, University "Politehnica" Bucharest, Romania
- Evangelos Kranakis, Carleton University, Canada
- Shawn McKee, University of Michigan, USA
- Yong Man Ro, Information and Communication University - Daejon, South Korea
- Francis Rousseaux, IRCAM, France
- Zhichen Xu, Yahoo! Inc., USA

Advanced P2P Systems

- Nikos Antonopoulos, University of Surrey, UK
- Filip De Turck, Ghent University – IBBT, Belgium
- Anders Fongen, Norwegian Defence Research Establishment, Norway
- Stephen Jarvis, University of Warwick, UK
- Yevgeni Koucheryavy, Tampere University of Technology, Finland
- Maozhen Li, Brunel University, UK
- Jorge Sa Silva, University of Coimbra, Portugal
- Lisandro Zambenedetti Granville, Federal University of Rio Grande do Sul, Brazil

International Journal on Advances in Internet Technology
Volume 4, Numbers 1 & 2, 2011

CONTENTS

Yuki Minami, Graduate School of Information Science and Technology Osaka University, Japan
Yuki Koizumi, Graduate School of Information Science and Technology Osaka University, Japan
Shin’ichi Arakawa, Graduate School of Information Science and Technology Osaka University, Japan
Takashi Miyamura, NTT Network Service Systems Laboratories NTT Corporation, Japan
Kohei Shiomoto, NTT Network Service Systems Laboratories NTT Corporation, Japan
Masayuki Murata, Graduate School of Information Science and Technology Osaka University, Japan

A Framework for Distributing Scalable Content over Peer-to-Peer Networks

Michael Eberhard*, Amit Kumar†, Silvano Mignanti‡, Riccardo Petrocco§, Mikko Uitto¶
*Klagenfurt University, Klagenfurt, Austria, michael.eberhard@itec.uni-klu.ac.at
†STMicroelectronics, Milan, Italy, amit-agr.kumar@st.com
‡University of Rome Sapienza, Rome, Italy, silvano.mignanti@dis.uniroma1.it
§Technische Universiteit Delft, Delft, The Netherlands, r.petrocco@gmail.com
¶VTT Technical Research Centre, Oulu, Finland, mikko.uitto@vtt.fi

***Abstract*—Peer-to-Peer systems are nowadays a very popular solution for multimedia distribution, as they provide significant cost benefits compared with traditional server-client distribution. Additionally, the distribution of scalable content enables the consumption of the content in a quality suited for the available bandwidth and the capabilities of the end-user devices. Thus, the distribution of scalable content over Peer-to-Peer network is a very actual research topic. This paper presents a framework for the distribution of scalable content in a fully distributed Peer-to-Peer network. The architectural description includes how the scalable layers of the content are mapped to the pieces distributed in the Peer-to-Peer system and detailed descriptions of the producer- and consumer-site architecture of the system. Additionally, an evaluation of the system's performance in different scenarios is provided. The test series in the evaluation section assess the performance of our layered piece-picking core and provide a comparison of the performance of our system's multi layer and single layer implementations. The presented system is to our knowledge the first open-source Peer-to-Peer network with full Scalable Video Coding support.**

***Keywords*-Peer-to-Peer; Scalable Video Coding; Packetizing; Error Concealment; Performance Evaluation**

I. INTRODUCTION

The streaming of content over Peer-to-Peer (P2P) networks becomes more important as the popularity of Internet multimedia services is increasing and the corresponding server costs are rising. One of the major challenges of distributing multimedia content is that different users often require the content in different quality. On the one hand, this is due to the differences in the user's network connections, which can differ depending on the user's location during the content consumption. On the other hand, the users consume the content on various terminals like TV sets or mobile devices, which have different capabilities in terms of resolution, processing power, or power supply.

These problems are addressed by layered streaming systems that provide the content in different qualities within a single bitstream. The architecture of our P2P streaming system supporting scalable content has been originally published in [7]. This paper extends the previous system description and provides an additional evaluation of the system's performance. In this paper we are going to describe our entire framework for the distribution of scalable content in a fully distributed P2P network. The P2P system targeted for the integration is the NextShare system, which is developed within the P2P-Next project [3]. P2P-Next is a research project partially founded by the European Commission in the context of the Framework Program 7, within the ICT (Information and Communication Technology) theme. The main goal of P2P-Next is the development of an open-source next generation P2P content delivery platform, the NextShare system.

The NextShare system has been developed based on the Bittorrent protocol [1] and thus provides an implementation of a fully distributed P2P system. To support Video on Demand (VoD), live streaming and the distribution of scalable content in the NextShare system, a number of modifications to the original Bittorrent protocol have been performed [11], as the original Bittorrent protocol does not support streaming. The scalable codecs used within NextShare are based on the Scalable Video Coding (SVC) extension of the Advanced Video Coding (AVC) standard [14].

One of the main reasons for implementing SVC support for the NextShare system is that there is to our knowledge today no open-source P2P system supporting SVC available that can be downloaded and tested by interested users. The advantages of distributing scalable content compared to simulcast approaches have been evaluated in a number of surveys (see, e.g., [15]). Additionally, we provide a comparison to our implementation for single layer content in Section VI-B to illustrate the advantages of using scalable content.

The remainder of this paper is organized as follows: Section II provides an overview of the related work. In Section III, the approach for the integration of the scalable content into the NextShare system is described. In the following two sections, the producer- and consumer-site of this architecture are described in detail. Section VI provides an evaluation of our implementation in terms of piece download efficiency as well as a comparison to the traditional single layer approach. Finally, future work is addressed in Section VII and Section VIII concludes the paper.

II. Related Work

The distribution of multimedia content over P2P networks has been a popular research topic in recent years. Due to the increasing popularity of streaming high-quality multimedia content over the Internet, P2P provides a cost-efficient alternative to reduce server costs.

The distribution of layered content over P2P systems has also been addressed in the literature before. LayerP2P [10] provides a well defined solution for distribution SVC content over P2P, but does not utilize real SVC codecs for the prototype implementation and relies on the usage of H.264/AVC-compatible codecs that can only be used to test one of SVC's scalability dimensions, the temporal scalability. Thus, one of the goals of the NextShare implementation was to design, implement, and distribute an open-source system with full SVC support. Other systems supporting the distribution of SVC content over P2P are described in [12] and [8]. PALS [12] provides a receiver-driven solution for receiving layered content over P2P. [8] describes how SVC can be integrated into a tree-based P2P system. However, both approaches to not allow an easy integration into existing P2P systems, as the implementations have been based on proprietary systems and protocols. Regarding compatibility an advantage of our implementation is that it has been based on the wide spread Bittorrent protocol and all architectural choices have been performed while ensuring backwards compatibility to existing Bittorrent clients. This allows an easy integration of the new scalable video technology into existing P2P communities. Furthermore, backwards-compatibility of the base layer for existing Bittorrent clients is provided.

III. NextShare Integration

To fully integrate scalable content into the NextShare system, a number of problems had to be addressed. Two main problems, the selection of suitable scalability layers and the mapping of the layers to Bittorrent pieces, are described in detail within this section. While the selection of the scalability layers tries to consider all popular qualities and to support a number of different network connections, the mapping to the scalability layers to the Bittorrent pieces tries to ensure that the best trade-off between flexibility for possible quality switches and overhead in terms of piece management is found.

It should be noted that even though we are using SVC within our NextShare system, all design decisions have been made with the intention to make the architecture codec-agnostic. Thus, if another scalable video codec is utilized within the NextShare system, only the coding and packaging tools need to be replaced, while the integration into the NextShare core will remain suitable for every other layered codec.

Table I
Scalability Layers

Bit Rate	Resolution	Quality	frame/sec
512 kbps	320x240	low	25
1024 kbps	320x240	high	25
1536 kbps	640x480	low	25
3072 kbps	640x480	high	25

A. Scalability Layers

The first step for the integration of scalable content into the NextShare system was the selection of the desired scalability layers. The selected layers are described in Table I.

As illustrated in Table I, four scalable layers were selected for the integration. The main reasons for selecting this layer structure were to maintain a good coding efficiency and to provide all popular qualities. The possibility to add further layers to support HD content is also fully supported by our framework, but has been omitted for the current version due to constraints in the upload bandwidth of our system's users. From the coding-efficiency point of view, the difference between the layers in terms of bit rate should be not too low, as the coding efficiency decreases drastically in such cases [14], while the selected bit rates represent the most popular qualities that are provided nowadays by multimedia portals. Furthermore, it should be noted that the audio bitstream is provided together with the video bitstream of the base layer. Thus, the 512 Kbps for the base layer includes the bit rate for the 128 Kbps audio bitstream. This is necessary to ensure that the audio is always received in time for playback, which can start as soon as the base layer is received.

To ingest the different layers into the P2P system the layers need to be provided as separated files. The base layer is multiplexed with the audio content and provided in a proper container format. The enhancement layers are provided as separate optional files. By using this file structure, Bittorrent clients without SVC support can still download the H.264/AVC-compatible base layer and decide not to download the optional enhancement layers without wasting any bandwidth.

B. Mapping to Bittorrent Pieces

The second step of the integration process is the mapping of the scalability layers to Bittorrent pieces. Firstly, the unit shall represent a synchronization point for dynamic switches between different quality layers. To achieve this goal each unit starts with an Instantaneous Decoding Refresh (IDR) reference frame. Secondly, it should be noted that we do not perform a direct mapping to pieces but to a unit. This unit represents a fixed number of frames for a specific layer and can be mapped to a fixed number of pieces. The reason for this approach is that the piece size might be changed in the P2P system for various reasons, and by basing the mapping

Table II
UNIT MAPPING

Layer	Kb/time slot	KB/time slot	pieces/time slot
BL	$512Kbps * 2.56 \approx 1.310$	$/8 \approx 164KByte$	3 pieces @ 55 KByte/time slot
EL1	$1024Kbps * 2.56 \approx 2.621$	$/8 \approx 328KByte$	6 pieces @ 55 KByte/time slot (3 pieces in previous layers, 3 new pieces)
EL2	$1536Kbps * 2.56 \approx 3.932$	$/8 \approx 492KByte$	9 pieces @ 55 KByte/time slot (6 pieces in previous layers, 3 new pieces)
EL3	$3072Kbps * 2.56 \approx 7.864$	$/8 \approx 983KByte$	18 pieces @ 55 KByte/time slot (9 pieces in previous layers, 9 new pieces)

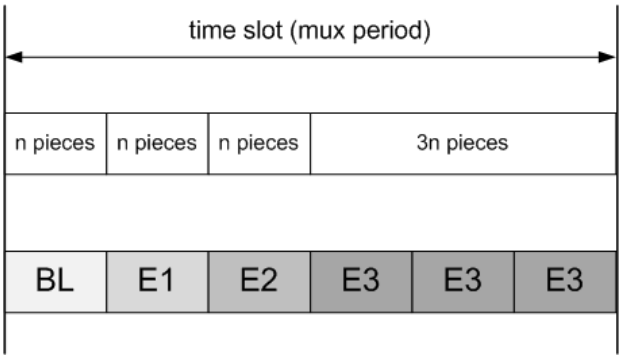

Figure 1. Piece Mapping

on units rather than on pieces only the unit/piece-mapping needs to be updated when the piece size is modified.

The mapping to the units has been performed based on several criteria. First, the units need to be selected large enough to allow for a good coding efficiency. As it should be possible to decode each unit independently (when all lower layer units for the same time stamp are also available) the number of frames within one unit should be high enough to allow for good coding efficiency. Additionally, the number of frames within one unit should be low enough to provide the flexibility to conveniently switch between qualities when the network conditions change.

Based on these considerations, a mapping of 64 frames, which represent 2.56 seconds of content at a frame rate of 25 frames/sec, has been selected. Such a unit is subsequently mapped to three pieces; however, as noted previously, the piece mapping can always be changed based on the requirements from the P2P system. The piece mapping is illustrated in Table II.

The mapping to the 55 KByte pieces results in a small overhead of available bits per piece. However, this overhead is utilized to compensate the small drifts of the constant bit rate (CBR) algorithm utilized during the SVC encoding process (see Section IV-A).

Based on the calculations in Table II, a mapping of the layers to Bittorrent pieces could be performed as illustrated in Figure 1.

The figure shows that the unit for each layer can be mapped to a specific number of actual pieces.

IV. PRODUCER-SITE ARCHITECTURE

The producer-site architecture describes all steps from encoding the SVC bitstream to the ingestion into the core of the P2P system. The topics addressed in this section include the encoding process, the splitting of the bitstream, creating metadata based on the bitstream's supplemental enhancement information (SEI), packetizing the bitstream, and ingesting the bitstream into the core of the P2P system. An illustration of this architecture is provided in Figure 2, more details on each of the processing steps are provided in the following sections.

A. Bitstream Preparation

As the first step of the bitstream preparation process, the raw video (i.e., the YUV video frames) is encoded by an optimized JSVM 9.15 [4] encoder, which uses a CBR algorithm to ensure that the pieces created from the video content have a constant size. The CBR algorithm works at GOP (Group of pictures) level and maintains the bit rate at GOP level throughout the encoded bitstream. However, the CBR algorithm still produces a small offset compared to the desired bit rate. As a constant piece size has to be maintained, a positive offset could results in frame dropping while a negative offset can be easily addressed by using padding bits during the splitting process. To ensure that no frames are dropped in case the small drifts of the CBR algorithm result in a positive offset, the target bit rate for the CBR algorithm is chosen slightly lower (approx. 1-2 % below the target bit rate). Thus, the CBR algorithm produces only negative offset compared to the real target bit rate, which can be easily handled.

The encoded SVC bitstream is subsequently split into the H.264/AVC-compatible base layer (BL) and the enhancement layers (EL) by the Network Abstraction Layer Unit (NALU) demuxer. The demuxer analyzes the NALU headers and splits the access units into separate bitstreams for each layer. Each of these layer bitstreams consists of several pieces of constant size. If within one bitstream the GOP size exceeds the piece size, subsequent NALUs (frames) would be dropped. However, as mentioned in the previous paragraph, such a situation is avoided by setting a slightly lower target bit rate for the CBR algorithm. If the GOP size is less than the piece size, the remaining size bits are filled with padding bits. Additionally, the SEI information at the beginning of the bitstream (i.e., the scalability info message) and the Sequence Parameter Set (SPS) and Picture Parameter Set (PPS) are provided to the metadata creator (see Section IV-C).

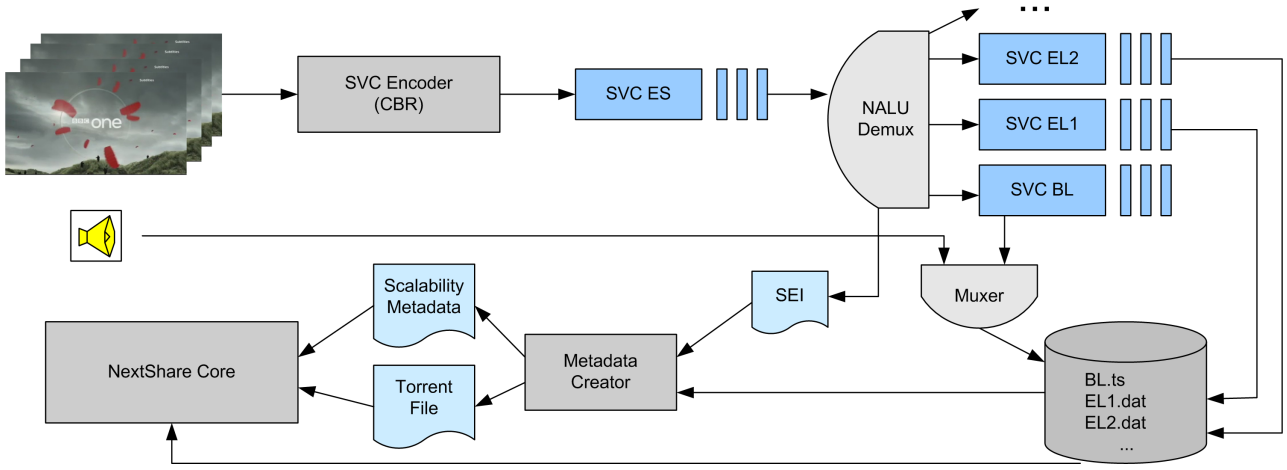

Figure 2. Producer-Side Architecture

The audio data can be provided already encoded, e.g., as an MP3 or AAC audio file. If a raw PCM audio file is provided, the audio content is encoded to the desired audio coding format.

B. Bitstream Packetizing

In the bitstream packetizing step, the base layer of the SVC bitstream is muxed with the audio into a proper container format. The main reason for this step is that the base layer should be provided in a backwards-compatible way, so that also end user P2P clients or terminals that only support H.264/AVC can successfully process the base layer. For such a purpose two different container formats were investigated for our system, the MPEG Transport Stream (MPEG-TS) [5] and the MPEG-4 file format (MP4) [6].

MPEG-TS is a standard able to encapsulate audio and video Packetized Elementary Streams (PESs) and other data and is supported by a majority of systems and applications. The main disadvantage in using MPEG-TS is that it usually has a rather high overhead in terms of bit rate(10-20% in average). An alternative muxing scheme is provided by the MP4 format which provides functionalities similar to the ones of MPEG-TS while having a clearly lower overhead ($\sim$1%). Thus, MP4 is the preferred container format used in our system, while MPEG-TS support is provided for compatibility to older systems.

The overall architecture is codec-independent: the system is able to recognize the container format and apply the corresponding processing. A general problem during the muxing phase is that the output should have a certain fixed size to ensure that a full GOP of video content and the corresponding audio content can be mapped to one unit. Considering that muxing schemes can have variable overheads, it is in principle not possible to a priori know if the output of the muxer for a certain audio and video input will respect the size limits. In case the output size is smaller than expected it will be possible to add padding bits and solve the issue (muxing codecs usually provide routines for that). The real problem is when the muxing output size is higher than the allowed one: in such a case the muxer tries to change its parameters to lower the overhead to the minimum. However, if adjusting the muxer's parameters is not sufficient, it would usually not be possible to meet the size constraints. Thus, as previously mentioned in Section IV-A, the target bit rate is set lower than desired to ensure that only the first case (lower output size) occurs. To avoid possibly wasting too much bit rate on padding bits, the architecture optionally provides support for a feedback mechanisms between the muxer and the encoders to solve this. Thus, in case the output size would be higher than the target size, the muxer asks the SVC and the audio encoders to re-encode both audio and video using a lower target bit rate. For the enhancement layers, the padding mechanism described in the previous section is applied.

C. Scalability Metadata Support

Although the pieces of the video stream are transmitted over the network in a layered way, the de-packetizer at the consumer-site needs to know the properties of the layers for the decoding process and the decoder needs access to the parameters from the beginning of the bitstream (the SPS and PPS elements). Thus, the properties of the layers, which are usually provided by the Supplemental Enhancement Information (SEI) at the beginning of the bitstream, and the parameter sets need to be forwarded to the consumer-site.

To store these metadata and transmit them to the de-packetizer when needed, the SEI message and the parameters are forwarded from the NALU demuxer to the metadata

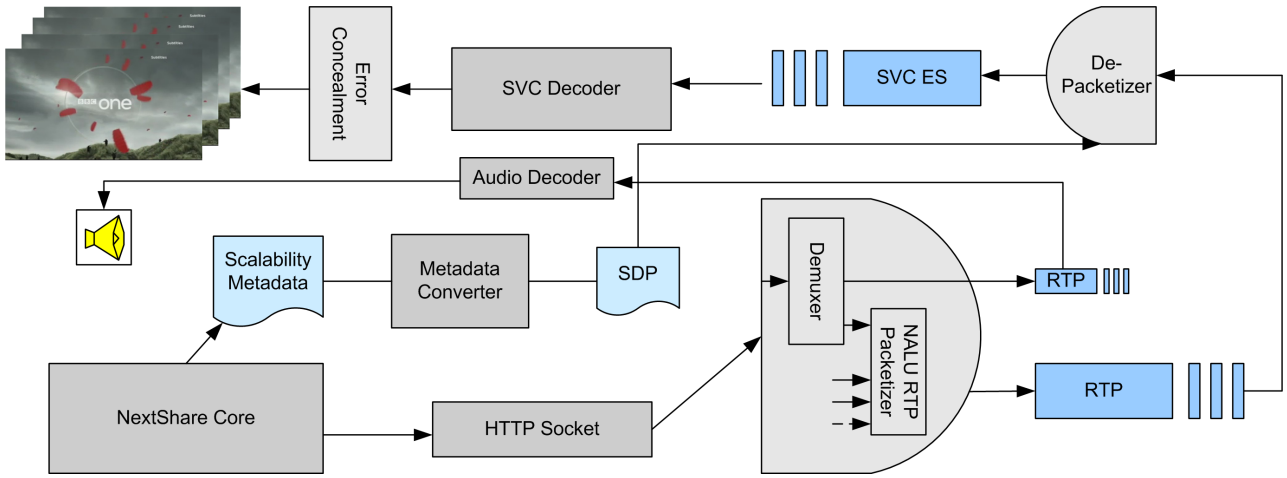

Figure 3. Consumer-Side Architecture

creator. The metadata creator subsequently parses the SEI data and stores the properties of the layers in an XML metadata document. Additionally, the SPS and PPS elements are encoded in base64 to allow their storage in XML and are added to the metadata document. The resulting metadata document contains all the layer information and parameter sets required by the de-packetizer and decoder modules (see Section V for details).

D. Ingest into the Core

The NextShare core represents the P2P engine responsible for creating and injecting the content into the network. The main metadata file required for the ingestion of the content into the P2P system is the torrent file. The torrent file provides the information required for the download of the previously encoded base and enhancement layers, as well as metadata related to the content including the previously created scalability metadata. The created torrent file is compatible with the Bittorrent protocol [1] and can therefore be processed by every peer running a Bittorrent-compatible client. This backwards-compatibility increases the reliability of the torrent swarm and the scalability of the distribution costs, as the torrent file can not only be processed by NextShare-compatible clients, but by all Bittorrent-compatible clients.

The fact that the H.264/AVC-compatible base layer and the audio stream are provided commonly packetized into a proper standardized muxing format enables also clients without SVC support to join the swarm as they are able to consume the stream in base layer quality. Therefore, every peer has an incentive to download at least the base layer, which increases its availability in the swarm. As the base layer is the most important layer (it is sufficient to start the playback and is always required) this really helps to ensure real-time playback for all clients in the swarm.

After the creation of the torrent file, the content is ingested into the NextShare core (i.e., the torrent file is distributed to other peers and the files containing the base- and enhancement layers are seeded). During the ingestion process, the base and enhancement layer files are split into pieces as illustrated in Figure 1. During the mapping process some problems have to be taken into consideration. Firstly, the integrated CBR algorithm does not provide an exactly constant bit rate, but allows for minor drifts. Thus, the piece mapping is always performed with a small overhead. Additionally, the 188 byte size utilized for the MPEG-TS packets cannot be mapped to a power of two, while the pieces ingested into the NextShare core should have a multiple of two as their size. Thus, the possibility of choosing a piece size that differs from the standard power of two has been successfully investigated.

V. Consumer-Site Architecture

The consumer-site architecture describes all steps from receiving the bitstream through the P2P system to decoding the bitstream for displaying it in the media player. The steps described include the local streaming of the received content, the de-packetizing of the content, the signaling of the layer properties using suitable metadata, and the merging and decoding of the received layers. An overview of this architecture is provided in Figure 3 and described in detail in the subsequent sections.

A. Provision of the Received Content

When the content is accessed by the user the layers from the NextShare swarm are downloaded in an intelligent way in order to maximize the Quality of Experience (QoE) for the current available download bandwidth. A great advantage of changing quality by displaying more or less enhancement layers regards the fall-back scenario: in a P2P system peers

are considered to have an unreliable and selfish nature, leaving the swarm and decreasing the total available bandwidth as soon as they have received the desired content.

In such a case the fall-back scenario will occur, where the user will experience a slow decrease of the QoE, as the download engine avoids downloading higher layers for upcoming time slots, if enough pieces of the previous layers are not already available. Retrieving the content from the network is based on a modified approach of the Give-to-Get (G2G) algorithm [11]. The main reason for using G2G is that Bittorrent's Tit-for-Tat algorithm is not suitable for streaming multimedia content.

The original G2G algorithm divides the part of the piece buffer close to the current playback position into high-, mid-, and low-priority sets with regard to the current playback position. The high-priority set is the part following immediately after the playback position. The G2G algorithm selects the pieces in the high-priority set based on their deadline, i.e., the piece with the nearest deadline in the high-priority set is downloaded first to ensure a continuous playback of the content. In the mid- and low-priority set, the pieces are selected using Bittorrent's rarest first strategy. Using this piece-picking policy the G2G algorithm tries to ensure that the pieces are downloaded in time for playback while still ensuring that piece's that are desired by neighbour peers are downloaded as well.

For the layered application of the G2G algorithm, the priority sets are applied to all the active available layers in a proportional way. Thus, for every layer a high-priority set is created where the pieces are selected according to their deadline while the dependency between the layers is considered.

As discussed in the previous section, by providing backwards compatibility with other clients, the availability of the base layer at all peers in the same swarm can be assumed. Therefore, if a download bandwidth of at least the base layer's bit rate is provided by the peer's connection, it can also be assumed that the playback will never stall (if the neighbour peer seeds the content or is ahead in its playback position). Note that once a peer finishes watching the content, the download engine will start retrieving the remaining pieces of all the layers for two major reasons: firstly, to increase the layer's availability in the swarm, and secondly, to enable watching the content at the highest quality again once all the layers have been downloaded. However, if this behaviour is not desired due to bandwidth restrictions it can be disabled in the configuration.

After enough content has been downloaded to guarantee a continuous playback of at least the base layer, the download engine of the NextShare core will initialize the demuxer module. The multimedia data is forwarded to the demuxer utilizing an HTTP socket, which was selected to ensure interoperability between the NextShare core and third-party de-packetizing/decoding solutions. A persistent connection will be established between the demuxer and the NextShare core, allowing the demuxer module to sequentially ask for the available content for the following time slot, depending on the requests it receives from the decoder module. It is important to notice that the available pieces of the following time slot will be sent to the demuxer as late as possible, allowing the NextShare core to manage the major buffer, to increase the quality until the last moment and to try to increase the quality if the user pauses the playback.

B. Bitstream Demuxing and Packetizing

As described before, the demuxer works online: it receives from the swarm at least the stream of the base layer, which is processed by a suitable demuxer. The demuxer firstly removes any possible padding from the production phase and splits the container format stream into the audio content and the H.264/AVC-compatible base layer.

The elementary audio stream is directly encapsulated into a Real-Time Transport Protocol (RTP) packet stream (e.g., according to [9] for an MP3 audio stream) and is sent to the successive module. The elementary SVC base layer video stream is forwarded to the NALU RTP packetizer, which puts the base layer and the received enhancements layers into an RTP packet stream, reordering the packets and forwarding the RTP stream containing all layers to the next module. Furthermore, the demuxer establishes a RTCP channel with the player. This is useful in order to maintain essential synchronization information among the video and the audio layers and also to support playback control commands.

Please note that all the modules represented in Figure 3 are typically running on the same host, i.e., the peer that receives the content. However, the main reason for selecting the RTP protocol to convey the audio and video data was to provide flexibility and to enable a possible integration of the P2P network with a more traditional server-based network. In such a server-based network the NextShare consumer-peer could act additionally as a server, receiving the content from the P2P network and redistributing the scalable content within an RTP streaming network.

C. Scalability Metadata Support

To decode the layers received from the NextShare system, the de-packetizer needs to be aware of the scalability properties of these layers. To provide a generic signaling mechanism for these properties, which could also be used by third-party solutions, we have decided to provide this information as a Session Description Protocol (SDP) document to the de-packetizer. The SDP document is formatted according to [13], which provides the capabilities to signal the properties and dependencies of the scalable layers.

As the metadata is provided by the NextShare core in a NextShare-conforming XML format, the scalability metadata document is firstly converted to the targeted SDP format. Subsequently, the SDP message containing the layer

A) Correctly received EL3 picture: PSNR 42,4 dB

B) Upscaled picture from the base layer: PSNR 31,7 dB

Figure 4. Upscaling Result

properties and the parameters sets is forwarded to the de-packetizer.

D. Bitstream Consumption

After the demuxing and RTP packetizing of the audio and video content, the RTP streams are forwarded. While the audio stream is forwarded to a standard de-packetizer and decoder, the SVC RTP stream is processed by our customized tools.

Firstly, our SVC de-packetizer parses the RTP packets and provides the payload and the time stamps to the SVC decoder. To perform this extraction process, the de-packetizer needs to have SVC layers properties along with the audio and video RTP port information in advance. This information is provided by the SDP file. Additionally, the de-packetizer parses the parameter sets from the SDP and provides them to the decoder, as they are required for the decoding process. The SDP file contains the SVC layer information for each layer and the de-packetizer extracts the suitable information for the desired layer playback. Finally, our highly optimized version of the JSVM 9.15 reference decoder performs the real-time decoding of the SVC content utilizing the error concealment algorithm embedded in the decoder (described in the next section).

E. Error Handling

Error robustness in the video decoder is important since transmission errors are very common in current (especially wireless) video streaming and transmission systems [16]. The transmission errors can lead to very poor quality of experience and in worst case scenario, they can lead to decoder crashes. Usually the error concealment is performed by monitoring the order of the NAL slices and their macroblocks to see if all the NALUs are received. If NALUs are missing suitable concealment operations for the missing macroblocks are performed, e.g., by using a frame copy from the previously correctly received frame [17].

The difference in the error concealment implementation in the NextShare system to such traditional error concealment approaches is that random frames or burst of frames cannot occur, as each piece contains a full GOP for a specific layer. Thus, a GOP can either be received or not received, but single frames cannot be lost during the transmission. As the whole GOP between the IDR pictures can either be present or missing, this can lead to varying resolution in the player if spatial scalability layers exist (as suggested in Table I). However, a major advantage of the layered content provisioning in NextShare is the awareness of the layer dependency. The higher enhancement GOPs are not send to the decoder if not all the pieces from the lower layer GOPs have been previously sent for the current time slot. Additionally, the base layer is always retrieved, which makes the error concealment easier and more effective.

The error concealment is integrated into the optimized JSVM 9.15, which was integrated into the VLC plug-in. To cope with the missing spatial enhancement layers, an upscaling functionality was integrated into the decoder based on normative integer-based 4-tap filters. The upscaling algorithm is provided by the JSVM reference software; please refer to [4] for more details.

The target resolution for the sequence is defined in the SPS NAL unit, which is compared with the received resolution (the resolution of the IDR picture) when starting the

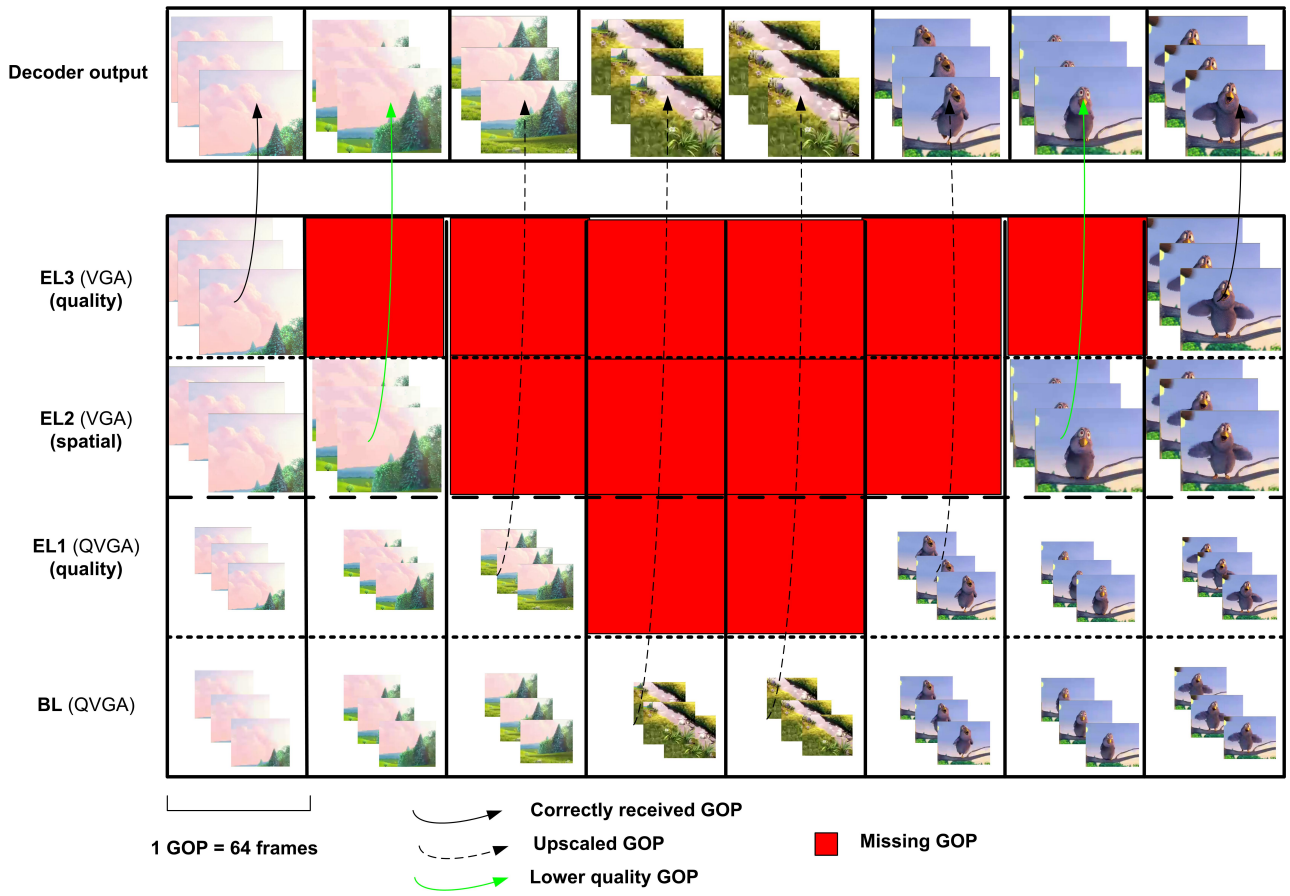

Figure 5. Upscaling Principle

decoding for a new time slot. As mentioned earlier, the SPS information reaches the decoder within an SDP description from the RTP de-packetizer. If the resolutions do not match, frames of the new time slot are up-scaled from the lower resolution to the target resolution to maintain the preferred resolution. Since the layer with lower spatial resolution usually provides lower quality, the upscaled picture is blurred, as illustrated in Figure 4. However, compared to changing the window size during video playback, the blurring is usually the better solution.

The selected structure of layers presented in Table I supports two different resolutions for the video. In some cases the layers for high resolution video cannot be received within the defined time slot, even though the consumer prefers to watch the video in higher resolution. As an example, this could happen while streaming on a heavily congested access point. Figure 5 shows an example situation with a four layer SVC sequence, where the second enhancement layer provides the spatial enhancements whereas the first and third enhancement layers provide quality enhancements. In Figure 5, the number of received layers decreases during the streaming from best quality to the base layer level. In this case, an upscaling of the base layer quality to the higher resolution is performed, to avoid changing the playback window size, which is very disturbing for the consumer. The upscaling algorithm is performed when only the base layer or the base and first enhancement layer are received. As soon as the second enhancement layer is received, no upscaling is necessary as the desired resolution is already provided.

VI. EVALUATION

To investigate the performances of our solution we performed a series of experiments with the implementation of our architecture. The experiments were performed in a lab-only environment composed of several peers connected with each other using heterogeneous connections. We monitored the performance of a peer acting as leecher in the swarm, retrieving the content from at least one seeder. Each of the seeders provides a limited upload bandwidth capacity to the leeching peer, which limits the download bit rate the leecher can achieve.

For our tests we encoded a two minute video sequence using four quality layers. The properties of these layers are

Table III
EVALUATION LAYER STRUCTURE

Bit Rate	Resolution	Quality	frame/sec
512 kbps	640x480	basic	25
1024 kbps	640x480	low	25
1536 kbps	640x480	medium	25
2048 kbps	640x480	high	25

described in Table III. The properties of the sequence used for the evaluation differ slightly from our reference layer structure described in Table I. The main reason for using a different layer structure for the evaluation were to keep a simple uniform unit size for all layers. Additionally, only quality layers were used for this evaluation to enable an easy comparison of the received quality layers in terms of peak signal-to-noise ratio (PSNR).

The evaluation consists of two test series. First, the behaviour of the layered piece download algorithm is illustrated in different scenarios to demonstrate the efficiency of our layered piece-picking implementation (Section VI-A). Second, the received quality of the layered implementation is compared to the single layer implementation of our P2P system to show how improvements in terms of received quality can be achieved (Section VI-B).

For the evaluation process five scenarios for the two minute test video sequence were defined. Four of those scenarios were investigated for both test series, while the first scenario is only interesting for the layered piece-picking series. In the first three scenarios all peers remain in the swarm for the whole time. In the other two scenarios, seeders leave the swarm at specific time points to test the robustness of the system against churn.

- Scenario 1: The leecher peer connects to a single seeder, which provides sufficient bandwidth to download only the base layer. This scenario is only investigated for the first test series.
- Scenario 2: The leecher peer connects to three seeding peers. Together the seeders provide enough bandwidth to download all layers for the test video.
- Scenario 3: The leecher peer connects to two seeders, which provide more than sufficient bandwidth for the download of three layers, but not sufficient to constantly download all layers.
- Scenario 4: At the beginning of this scenario two seeders provide sufficient bandwidth to download all layers. After 40 seconds one of the seeders leaves the swarm and the available bandwidth is decreased.
- Scenario 5: At the beginning of this scenario three seeding peers are in the swarm and sufficient bandwidth for all layers is provided. After 30 seconds, one seeder leaves the swarm and decreases the available bandwidth. After 70 seconds, a new seeder joins the swarm and increases the bandwidth.

As the player of our system can switch between qualities without flickering (see Section V-E) for this evaluation the playback quality is increased as soon as possible, even if only for one time slot. This particular behaviour can be changed to a more conservative approach that might be needed in case the decoder/player module is replaced or a constant quality playback is preferred. To influence the download strategy, the piece-picking algorithm can be configured to only perform switches to higher layers if a specific number of higher layer pieces have been downloaded (this number is one for the presented evaluation, i.e., immediate quality switches are performed).

A. Layered Piece-Picking Test Series

In the following graphs, representing the first test series, all five scenarios are presented to analyze the behaviour of the piece selection when layered content is streamed. In every graph the available upload capacity, the download rate, and the received bit rate are presented. The upload capacity illustrates the download bandwidth that is provided from the seeders to the leecher peer. The download bit rate describes at which rate the leecher peer downloads pieces from the seeders. Please note that the download speed is never calculated instantly, instead it is based on the piece arrivals and therefore averaged over a small period of time. This explains the smoothness of the download curve and avoids spikes in the results. Finally, the received bit rate represents the number of pieces that were received in time for playback. It should be noted that the received bit rate differs slightly from the actual playback bit rate of the video. The reason for this deviation is that the CBR algorithm of our system's SVC encoder does not provide an exactly constant bit rate but allows for small drifts. Thus, padding is used to achieve the constant piece size, as described in Section IV-A.

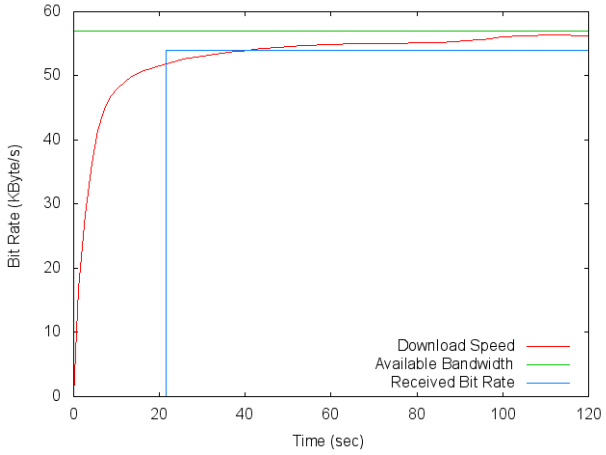

Figure 6. Layered Piece-Picking Series: Scenario 1

In Figure 6 the results for the first scenario are presented. As the available download bandwidth allows to download

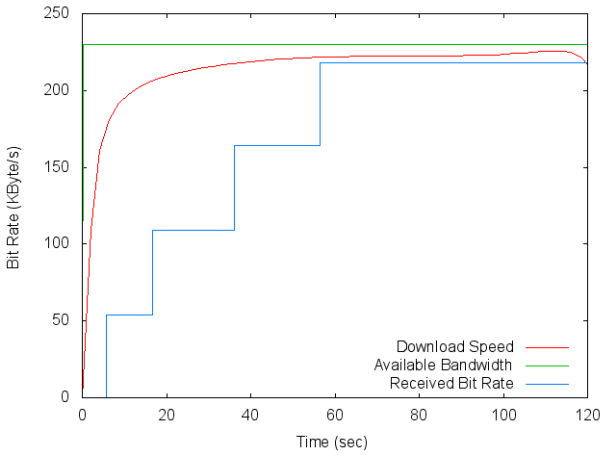

Figure 7. Layered Piece-Picking Series: Scenario 2

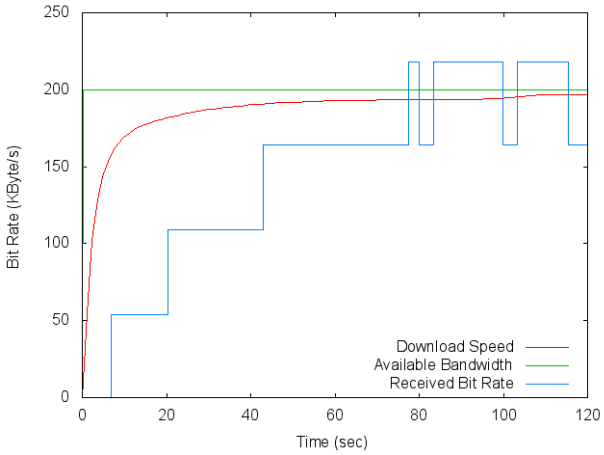

Figure 8. Layered Piece-Picking Series: Scenario 3

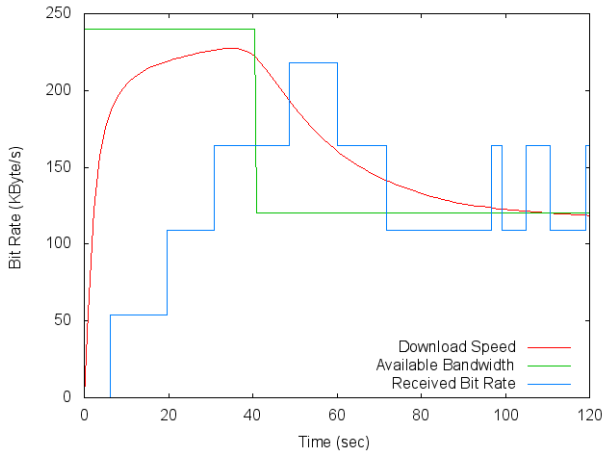

Figure 9. Layered Piece-Picking Series: Scenario 4

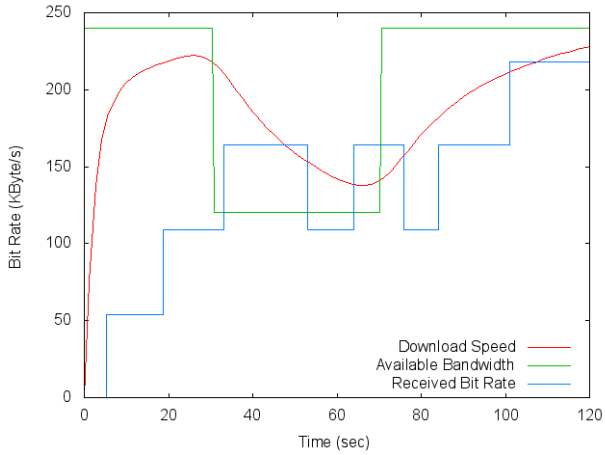

Figure 10. Layered Piece-Picking Series: Scenario 5

only the base layer, the behaviour of the piece-picking algorithm is very simple. Only base layer pieces are downloaded and the playback of the base layer is started after the initialization phase.

Figure 7 illustrates the behaviour in case of the second scenario. As mentioned before, a switch to a higher layer is performed as soon as the first piece of the layer is downloaded. Thus, the leecher peer starts with the playback of the base layer and gradually increases the quality until the highest layer is reached. The playback remains at best quality until the end of the scenario.

An interesting behaviour emerging from the configuration of the piece-picking algorithm for this test series can be seen in Figure 8. Having a download rate higher than the first three layers combined, the piece-picking algorithm will try to download the highest layer whenever possible. As the available bandwidth is not sufficient to constantly download all layers, the playback quality switches frequently. If a constant playback quality is preferred, the piece-picking algorithm can be configured accordingly.

In Figure 9 the fourth scenario is presented. At the beginning of the scenario the playback quality is gradually increased, similar to the behaviour in the second scenario. However, as soon as the highest playback quality is achieved, one seeder leaves the swarm and causes a decrease of the available bandwidth. Thus, the leecher peer cannot download the higher layer pieces anymore and decreases the playback quality. Nevertheless, the playback never stops and the piece-picking algorithm stabilizes after some time and downloads the best possible quality for the given bandwidth until the end of the scenario (between two and three layers, similar to the behaviour in scenario 3 but with less available bandwidth). The results of this test scenario show that our implementation and is robust to departing peers and can keep the video playback going, as long as sufficient seeding peers to download at least the base layer quality remain in the swarm.

Figure 10 illustrates the results for the final scenario. At

the beginning the playback is again gradually increased. However, one of the seeding peers leaves the swarm after 30 seconds and the piece-picking algorithm adjusts and downloads the best possible quality for the new bandwidth conditions (between two and three layers). After the joining of the new seeder, the playback quality is increased to the best possible quality, which can now be downloaded due to the improved bandwidth conditions. The results of this scenario show that our system is robust to churn and reacts suitably to leaving and joining seeders.

Another interesting observation of this test series is the initial playback delay when streaming layered content using our NextShare system. The test results of all scenarios show that the playback will start no more than ∼25 seconds after the leeching peer joins the swarm, assuming a download rate at least as high as the base layer's bit rate (otherwise no real-time playback is possible). Additionally, the playback of the base layer can start as fast as after five seconds, if the bandwidth conditions are good. This reduces the initial playback delay greatly compared to the single layer implementation of our system.

B. Quality Comparison Test Series

In this test series a comparison of the single layer and multi layer implementations of our system is performed. The goal of this test series is to compare the received quality for both approaches. For the evaluation process the previously described scenarios 2-5 are investigated.

The following graphs represent the results for this test series. Each graph illustrates the received video quality in PSNR for the single and multi layer implementations. The PSNR for a received piece is calculated by averaging the PSNR of all 64 frames contained within a piece for the multi layer implementation (the piece structure is described in Section III). For the single layer approach, one piece contains in average 16 frames providing similar quality as all four layers of the multi layer approach (the same piece size is used for both implementations). The received PSNR is calculated by summating the PSNR of the frames in the received pieces and dividing it by the total number of frames for the time slot.

In Figure 11 the results for the second scenario are displayed. During the initial phase where the quality is gradually increased, the received quality is clearly better for the multi layer implementation, as the base layer already provides a decent quality, while the single layer implementation only receives a part of the frames. However, when the highest quality playback is achieved (download of all pieces), the single layer approach shows slightly better results. This is due to the fact that a better PSNR can be achieved with single layer encoding at the same bit rate, as the multi layer encoding has a certain overhead in terms of coding efficiency (8.6 % for the example test sequence).

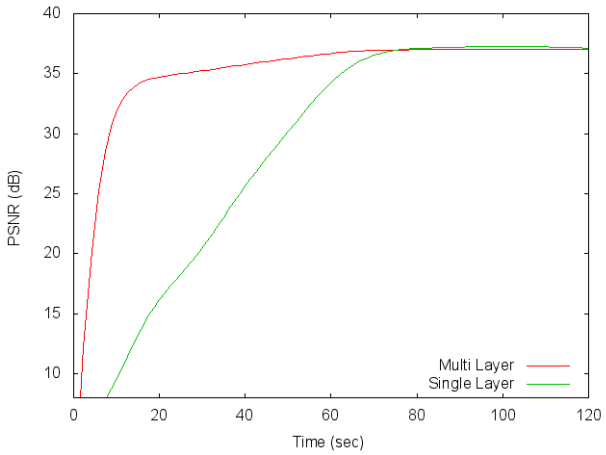

Figure 11. Quality Comparison Series: Scenario 2

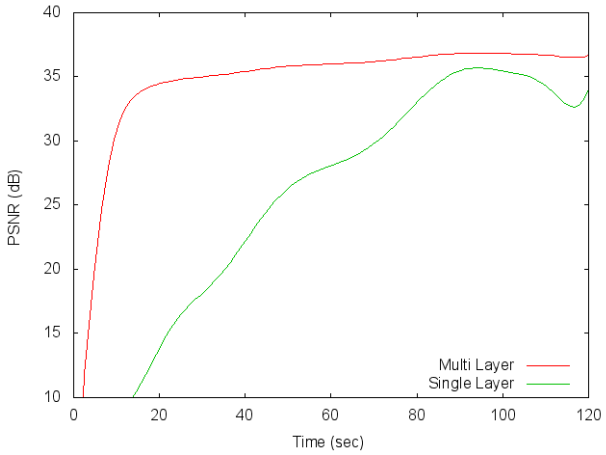

Figure 12. Quality Comparison Series: Scenario 3

Figure 12 shows the results for the third scenario. As the available bandwidth is not sufficient to constantly download all pieces, the single layer implementation has always a lower average PSNR than the multi layer implementation.

The results of scenario four in Figure 13 are similar. As one of the seeders leaves the swarm after 40 seconds, the average PSNR for the single layer approach remains rather low, while the multi layer approach can fall back to the lower layers and in comparison does not lose much in terms of average PSNR.

In Figure 14 the results for the fifth scenario are illustrated. Due to the leaving of one of the seeders after 30 seconds the average received quality of the single layer approach stays rather low. However, after the joining of the new seeder the quality for both implementations is increased and at the end of the scenario the quality for the single layer approach is slightly better, as all pieces are received in time.

Overall, the multi layer implementation has shown clearly better performance in terms of received quality during this

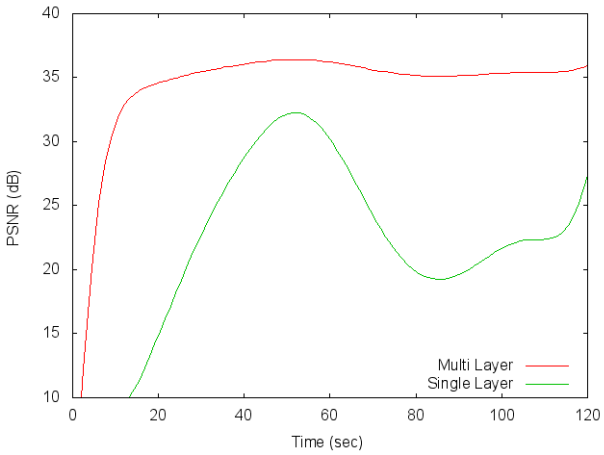

Figure 13. Quality Comparison Series: Scenario 4

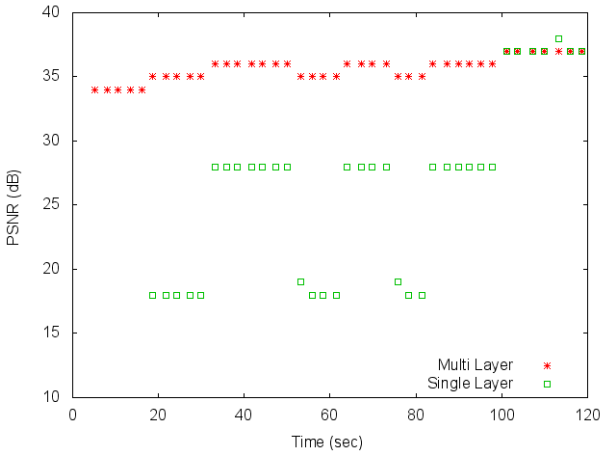

Figure 14. Quality Comparison Series: Scenario 5

test series. The single layer approach can only provide better quality if there is always more than sufficient bandwidth to download all pieces in time and no fluctuations occur. Additionally, for this evaluation we have only considered rather high bandwidth scenarios where most of the pieces can be downloaded in time. In low bandwidth scenarios the performance of the single layer approach gets even worse compared to the multi layer approach (only a smaller part of the frames is received in time).

VII. Future Work

Although our framework already provides a full solution for the streaming of scalable content over P2P networks, there are still open possible modifications that could enhance the viewing experience for the user.

Firstly, the current approach only utilizes the layered scalability provided by the SVC standard, i.e., the scalability in terms of resolution and fidelity. As a further step an additional support of temporal scalability would be desirable. Such a support could be realized by reordering the frames according to the hierarchical prediction structure and by storing the frames for different temporal levels in different pieces. However, this would increase the complexity of the piece-picking process and the trade-off between the increased complexity and the enhanced viewing experience needs to be investigated.

Another desirable further step would be the support of medium-grain scalability. However, such a support would require some changes to our architecture. As the piece picking algorithm currently works on piece level, where each piece contains a number of NALUs, and the medium-grain scalability allows to change the quality on NALU level, a new algorithm has to be designed to allow the partial download of pieces (i.e., to work on chunk level, with the need to investigate compatibility with existing clients). The other possibility would be to switch to dynamic piece size and map each NALU to a different piece. However, such a small piece such would be a bad choice with regard to the overhead of the piece sharing and a dynamic piece size would again break the compatibility with existing clients.

Furthermore, additional evaluations of our system will be performed to test the system's performance in large scale trials of our project's living lab [2]. Additionally, different configurations of the piece-picking algorithm will be evaluated to determine in which situations which piece-picking and quality switching strategies are preferable.

VIII. Conclusion

In this paper a framework for the provision of scalable content in a fully distributed P2P system was presented. In Section II, the selection of the scalable layers was explained and the mapping of the layers to units and subsequently Bittorrent pieces was illustrated. The main reasons for the choice of 64 frames per unit were to achieve a good coding efficiency while still maintaining the flexibility to quickly switch between layers if the network conditions change. In Sections III and IV the producer- and consumer-site architecture were presented. The producer-site architecture includes the encoding of the SVC bitstream and splitting of the bitstream into layers, the packetizing of the base layer and the audio stream into a suitable container format, the creation of the scalability metadata based on the SEI information at the beginning of the bitstream, and the ingest of the content into the core of the P2P system. On the consumer-site, the modules for retrieving and consuming the content were presented. The retrieved content is provided to the demuxer utilizing an HTTP socket and the demuxed audio and video streams are forwarded to the media consumption solution using RTP. The de-packetizing and decoding of the SVC content is subsequently performed by our customized SVC tools utilizing the scalability information provided by a suitable SDP document.

In the evaluation section, the performance of our P2P system supporting scalable content was evaluated. Two test series were performed in order to evaluate the performance of our piece-picking algorithm and to compare the new multi layer implementation to the already existing single layer implementation in our system. The results of the first test series show that the layered piece-picking algorithm can efficiently utilize the available bandwidth after the initialization phase. Additionally, the layered implementation greatly reduces the start-up delay and is robust to churn. The second test series showed that the multi layer implementation shows a clearly better performance than the single layer implementation in all situations where the bandwidths is restricted or bandwidth fluctuations occur.

The presented system is to our knowledge the first open-source P2P system with full SVC support. Additionally, it is fully compatible with third-party media consumption solutions due to utilizing the HTTP, RTP, and SDP protocols for providing the content from the P2P system to the media consumption modules. Thus, the system is easy to use and customize for interested users.

ACKNOWLEDGMENT

This work is supported in part by the European Commission in the context of the P2P-Next project (FP7-ICT-216217) [3].

REFERENCES

[1] Bittorrent protocol 1.0. URL: http://www.bittorrent.org. Last accessed 15-July-2011.

[2] The P2P-Next living lab. URL: http://livinglab.eu. Last accessed 15-July-2011.

[3] The P2P-Next project. URL: http://www.p2p-next.org. Last accessed 15-July-2011.

[4] *JSVM 9.15 Software Manual*, 2009.

[5] ISO/IEC 13818-1. *Generic coding of moving pictures and associated audio information: Systems*, 2000.

[6] ISO/IEC 14496-1. *MPEG-4 Part 14: MP4 File Format Version 2*, 2003.

[7] N. Capovilla, M. Eberhard, S. Mignanti, R. Petrocco, and J. Vehkapera. An architecture for distributing scalable content over peer-to-peer networks. In *Second International Conferences on Advances in Multimedia (MMEDIA)*, pages 1–6, 2010.

[8] P. Baccichet et al. Low-delay peer-to-peer streaming using scalable video coding. In *Packet Video*, pages 173–181, 2007.

[9] R. Finlayson. A more loss-tolerant RTP payload format for MP3 audio. RCF 3119, 2001.

[10] Z. Liu, Y. Shen, K. W. Ross, S. S. Panwar, and Y. Wang. Layerp2p: Using layered video chunks in P2P live streaming. *IEEE Transactions on Multimedia*, 11(7):1340–1352, August 2009.

[11] J. J. D. Mol, J. A. Pouwelse, M. Meulpolder, D. H. J. Epema, and H. J. Sips. Give-to-Get: Free-riding-resilient video-on-demand in P2P systems. In *Multimedia Computing and Networking*, volume 6818, San Jose, USA, 2008.

[12] R. Rejaie and A. Ortega. Pals: Peer-to-peer adaptive layered streaming. In *NOSSDAV Proc.*, pages 153–161, New York, NY, USA, 2003.

[13] T. Schierl and S. Wenger. Signaling media decoding dependency in the session description protocol. RCF 5538, 2009.

[14] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, September 2007.

[15] A. Sentinelli, L. Celetto, D. Lefol, C. Palazzi, G. Pau, T. Zahariadis, and A. Jari. A survey on P2P overlay streaming clients. In *Towards the Future Internet - A European Research Perspective*, pages 273–282, 2009.

[16] T. Stockhammer, M.M. Hannuksela, and T. Wiegand. H.264/AVC in wireless environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):657–673, July 2003.

[17] M. Uitto and J.Vehkaperä. Spatial enhancement layer utilisation for SVC in base layer error concealment. In *Mobimedia '09 Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, London, United Kingdom, 2009.

Federation Establishment Between CLEVER Clouds Through a SAML SSO Authentication Profile

Antonio Celesti, Francesco Tusa, Massimo Villari and Antonio Puliafito
Dept. of Mathematics, Faculty of Engineering, University of Messina
Contrada di Dio, S. Agata, 98166 Messina, Italy.
e-mail: {acelesti,ftusa,mvillari,apuliafito}@unime.it

***Abstract*—Cross-Cloud federation implies the establishment of a trust context between cloud platforms acting on different administrative domains and located in different places. The main advantage of federation is that clouds can set interdomain communications so that they can benefit of new business opportunities such as the enlargement of their virtual resources capability. The process of federation set up can be schematized in three subsequent phases: Discovery, Match-Making, and Authentication. In this work, considering several clouds based on both the CLEVER architecture and a Cross-Cloud Federation Manager module, responsible for the accomplishment of the three phases, we focus on the authentication phase required for a secure interaction between different CLEVER domains. More specifically, we designed a SAML SSO profile for a generic three-tier cloud architecture, showing the way in which it can be applied in different CLEVER-based clouds for the establishment of trusted interdomain communications in order to "lend" and "borrow" virtualized resources.**

***Keywords*-Cloud Computing; Federation; Authentication; CLEVER; XMPP; SAML.**

I. Introduction

Cloud computing brings a new level of efficiency in delivering services, i.e., Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), representing a tempting business opportunity for ICT operators of increasing their revenues.

Currently, the cloud computing scenario includes hundreds of independent, heterogeneous, private/hybrid clouds but, many business operators have predicted that the process toward an interoperable federated cloud scenario will begin shortly. According to Gartner [1], the evolution of the cloud computing market is hypothesized in three subsequent stages: stage 1 "Monolithic" (now), cloud services are based on proprietary architectures - islands of cloud services delivered by mega-providers (this is what Amazon, Google, Salesforce and Microsoft look like today); stage 2 "Vertical Supply Chain", over time, some cloud providers will leverage cloud services from other providers. The clouds will be proprietary islands yet, but the ecosystem building will start; stage 3 "Horizontal Federation", smaller, medium, and large providers will federate horizontally themselves to gain: economies of scale, an efficient use of their assets, and an enlargement of their capabilities. For simplicity, in the rest of the paper, with terms such as "cross-cloud federation", "federation in cloud computing", or "cloud federation" we will refer to the above mentioned "Horizontal Federation".

In our previous work [2], we described how to set up an interoperable heterogeneous cloud environment in a Horizontal Federation, where clouds can cooperate together accomplishing trust contexts and providing new business opportunities. In that work, we proposed a three-phase cross-cloud federation model where the federation establishment between clouds passes through three main phases: *discovery*, the cloud looks for other available clouds; *match-making*, the cloud selects between the discovered clouds the ones, which fit as much as possible its requirements; *authentication*, the cloud establishes a trust context with the selected clouds.

The authentication phase poses many serious problems in the cross-cloud federation establishment due to the need for each cloud of managing a huge number of credentials depending on the security mechanisms employed in each infrastructure. In fact, a cloud should be able to authenticate itself with other heterogeneous clouds regardless their security mechanisms, performing the log-in once, gaining the access to all the required resources. We identified such a problem as *Cloud Single-Sign On (SSO) Authentication* and we addressed it designing a new Security Assertion Markup Language (SAML) [3] profile, defining the steps needed for a secure cloud SSO authentication.

The current open source cloud implementations lack of modularity in their architecture, hence, it could be very difficult to modify them or integrating new features. For these reasons, at the Multimedia and Distributed Systems Laboratory (MDSLab) of the University of Messina, we have been developing a new Virtual Infrastructure Manager named CLoud-Enabled Virtual EnviRonment (CLEVER) [4].

In this work, starting from the analyzed issues regarding the Horizontal Federation, and considering our proposed solution for enabling the SSO authentication using SAML, we apply our idea to a concrete scenario including CLEVER-based clouds. CLEVER well meets the requirement of a cloud scenario whose actors intend to establish a Horizontal Federation: although it specifically aims at the design of a management layer for the administration of cloud infrastructures, differently from other existing open source middleware, it also provides simple and easily accessible interfaces

for enabling the interaction of different "interconnected" clouds.

The paper is organized as follows. Section II describes the state of the art of cloud federation. In Section III, we provide a detailed analysis of cloud federation requirements, introducing the concept of *home cloud* and *foreign cloud*. In section IV, we introduce our three-phase federation model, also discussing the Cross-Cloud Federation Manager (CCFM) module. In Section V, we focus on the authentication phase and in particular on the cloud SSO authentication problem, running through the SAML CCAA-SSO profile. Section VI provides the details about the CLEVER architecture, pointing out its main features. Section VII provides a detailed description of the SAML CCAA-SSO profile applied to a federation scenario including CLEVER-based clouds. Conclusions and lights to the future are summarized in Section VIII.

II. Related Work and Background

Cloud Computing is emerging as a promising paradigm able to provide a flexible, dynamic, resilient and cost effective infrastructure for both academic and business environments.

The paradigm is rather different if compared with the previous one, that is Grid computing, as it was remarked by Ian Foster, the father of Grid, in his work [5]. In particular consolidated research topics are appearing hard to face on cloud computing especially on the area of security.

As it is recently reported in [6], the authors have underlined that security and privacy in Cloud represent of the main challenges. They remarked in cloud environments it is necessary to deal with: authentication and identity management, trust management, policy integration, access control and accounting, secure service management, privacy and data protection, semantic heterogeneity. They recognized that cloud computing are becoming multi domain environments in which each domain can use different security, privacy, and trust requirements and potentially employ various mechanisms, interfaces, and semantics. Service-oriented architectures are naturally relevant technology to facilitate such a multi-domain formation through service composition and orchestration. They asserted that it is important to leverage existing research on multi-domain policy integration and the secure-service composition to build a comprehensive policy-based management framework in cloud computing environments.

In our point of view the scenario appears much more complex if we make a binding between the heterogeneity of environments with the possibility of federating them. The concept of federation has always had both political and historical implications: the term refers, in fact, to a type of system organization characterized by a joining of partially "self-governing" entities united by a "central government". In a federation, each self-governing status of the component entities is typically independent and may not be altered by a unilateral decision of the "central government". More specifically, looking at the political philosophy, the federation refers to the form of government or constitutional structure known as federalism and can be considered the opposite of the "unitary state". The components of a federation are in some sense "sovereign" with a certain degree of autonomy from the "central government": this is why a federation can be intended as more than a mere loose alliance of independent entities [7].

Considering the federation perspective in cloud computing environments, new terms are also been coined as Intercloud ("Think of the existing cloud islands merging into a new, interoperable Intercloud where applications can be moved to and operate across multiple platforms..." [8]) or Cross-cloud ("For the benefit of human society and the development of cloud computing, one uniform and interoperable Cross-cloud platform will surely be born in the near future..." [9]). Nowadays, cloud federation is becoming a topic more and more debated within both the scientific and ICT industry worlds [10]. In fact, as discussed in [11], it brings many new business advantages for the enhancement of cloud providers' profit. In such a perspective, new paradigms allowing providers to avoid the limitation of owning only a restricted amount of resources are rising.

Nevertheless, a few works are available in literature related to federation in cloud computing environments. The main reason is that several pending issues concerning security and privacy still have to be addressed, and a fortiori, is not clear what cloud federation actually means and what the involved issues are [12].

Nowadays, the latest trend to federate applications and service oriented architectures (SOAs) over the Internet is represented by the Identity Provider/Service Provider (Id-P/SP) model [13]. Examples are the aforementioned SAML, OpenID [14], Shibboleth [15] and Cardspace [16]. Such solutions, considered alone, do not solve the cloud federation issues. In fact, the federation problem in cloud computing is greater than the one in traditional systems. The main limit of the existing federation solutions is that they are designed for static environments requiring a priori policy agreements, whereas clouds are high-dynamic and heterogeneous environments, which require particular automatic security and policy arrangements. Keeping in mind the cloud federation perspective, several security issues are already picked out. Interoperability in federated heterogeneous cloud environments is faced in [9], in which the authors propose a trust model where the trust is delegated between trustworthy parties, which satisfy certain constrains.

Nevertheless, such works do not fully clarify what it is really meant with the term cloud federation. Basically, it is not fully evaluated when, why, and how a cloud federation should be established and what the impact over the existing infrastructure, the involved architectural issues, and the

security concerns are. Therefore, we think a cloud federation model addressing architectural and security issues, also with implementation practice compliant with existing cloud infrastructures, is strongly needed.

III. Cross-Cloud Federation Analysis: our Reference Scenario

In this Section we try to clarify ideas concerning the general concept of cross-cloud federation. In order to identify requirements and goals, we propose a possible resource provisioning scenario where clouds might benefit of federation advantages. Cloud Computing relies its computational capabilities exploiting the concept of "virtualization". This technology has re-emerged in recent years as a compelling approach of increasing resource utilization and reducing IT services costs. The common theme of all virtualization technologies is hiding the underlying infrastructure by introducing a logical layer between the physical infrastructure and the computational processes. The virtualization is being possible thanks to Virtualization Machine Monitors (VMMs commonly known as "hypervisors"), i.e. processes that run on top of a given hardware platform, control and emulate one or more other computer environments (virtual machines). Each of these virtual machines, in turn, runs its respective "guest" software, typically an operating system, executed as if it is installed on a stand-alone hardware platform.

Private clouds hold their own virtualization infrastructure where several virtual machines are hosted to provide services to their clients. When argument of discussion is cloud federation the skeptics could ask: why should a cloud federate itself with other clouds? In our opinion, the answer is simple: cloud federation brings new business opportunities. In fact, in a scenario of "cross-cloud federation", each cloud operator is able to transparently enlarge its own virtualization resources amount (i.e., increasing the number of instantiable virtual machines and therefore cloud services) asking computing and storage capabilities to other clouds.

According to our analysis, within the above mentioned scenario we distinguish among cloud's client, home cloud and foreign cloud:

- **Cloud's client**. An IT company, organization, university, generic single end-user ranging from desktop to mobile users or a cloud provider using the *aaS supplied by a target "home cloud" according to a pay-per-use model.
- **Home cloud**. A cloud provider which receives *aaS instantiation requests by its clients. Each home cloud for the arrangement, composition, and delivery of such services can use the computing and storage resources of its own virtualization infrastructure along with the resources borrowed by foreign clouds according to a pay-per-use model.
- **Foreign cloud**. A cloud provider which lends its storage and computing resources to home clouds according to a pay-per-use model. More specifically, a foreign cloud reserves part of its own virtualization infrastructure for a home cloud, so that the home cloud can logically count on an elastic virtualization infrastructure whose capabilities are greater than the capabilities of its own physical virtualization infrastructure. Therefore even thought the virtual environments and services of a home cloud are logically placed in its virtualization infrastructure, in reality they can be physically placed in parts of the virtualization infrastructure lent by foreign clouds. A cloud provider could be at the same time both home cloud and/or foreign cloud.

There is not limit to the possible business scenarios which can take place in a cross-cloud federation environment. Such scenarios include, for example, virtualization capability enlargement, resource optimization, provisioning of distributed IaaS, PaaS, and SaaS spread over different clouds, power saving and so on.

In order to better explain the idea of cross-cloud federation, let us consider as reference the "virtualization capability enlargement" scenario depicted in Figure 1. When a home cloud realizes that its virtualization infrastructure has saturated its capabilities, in order to continue providing services to its clients (i.e., other clouds, enterprises, generic end users, etc), it decides to federate itself with foreign clouds A and B. The home cloud, besides hosting virtualization resources inside its own virtualization infrastructure, is also able to hosting virtual machines inside the foreign clouds A and B virtualization infrastructures, enlarging the amount of its available virtualization resources (See Figure 1, bottom part). Therefore, although the virtualization resources rent to the home cloud are physically placed within the virtualization infrastructures of foreign clouds A and B, they are logically considered as resources indeed hosted within the home cloud virtualization infrastructure. Despite the obvious advantages, the implementation of such cross-cloud federation scenario is not at all trivial. The main reason is that clouds are more complicated than traditional systems and the existing federation models are not applicable. In fact, while clouds are typically heterogeneous and dynamic, the existing federation models are designed for static environments where it is needed an a priori agreement among the parties to make up the federation. Keeping in mind the aforementioned scenario, we think cloud federation needs to meet the following requirements: *a) automatism and scalability*, a home cloud, using discovery mechanisms, should be able to pick out the right foreign clouds which satisfies its requirements reacting also to cloud changes; *b) interoperable security*, it is needed the integration of different security technologies, for example, permitting a home cloud to be able to join the federation without changing its security policies. In the "interoperable security" context we identify: *1) SSO authentication*, a home cloud should

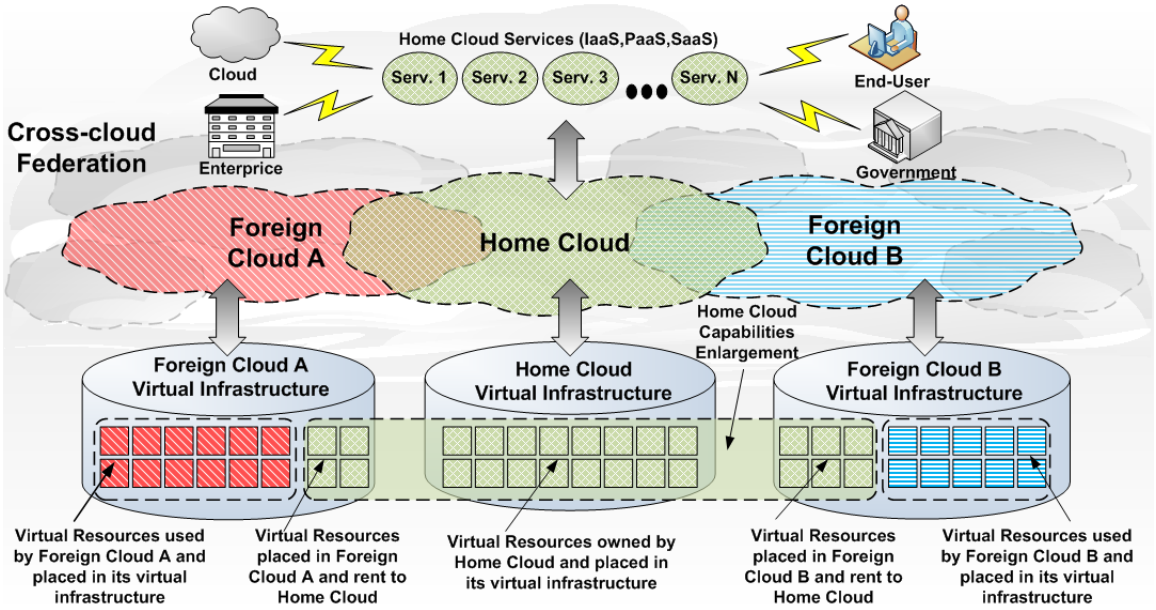

Figure 1. Cross-Cloud Scenario: basic for heterogeneous and federated clouds.

be able to authenticate itself once gaining the access to the resources provided by federated foreign clouds belonging to the same trust context without further identity checks; *2) digital identities and third parties*, each home cloud should be able to authenticate itself with foreign clouds using its digital identity guaranteed by a third party. This latter feature is more challenging because it implies a cloud has to be considered as a subject uniquely identified by some credentials.

IV. Cross-cloud federation: Architectural Overview

In Section III, we described the concept of federation and its bindings with the cloud. In the following, we provide a detailed description of the approach used to address the cross-cloud federation issues. Considering the requirements of automatism, scalability and interoperability previously stated, our solution tries to answer all such issues. Describing the federation process we point out three main different phases: *discovery*, *match-making* and *authentication*. These phases are opportunely explained in the following.

A. The Three-Phase Cross-Cloud Federation Model and the Cross-Cloud Federation Manager

In order to identify the main components constituting a cloud and better explain the federation idea on which our work is based, we are considering the internal architecture of each cloud as the three-layered stack [17] presented schematically in Figure 2.

B. Architectural Overview

Starting from the bottom, we can identify: *Virtual Machine Manager*, *Virtual Infrastructure (VI) Manager* and *Cloud Manager*. VI Manager is a fundamental component of private/hybrid clouds acting as a dynamic orchestrator of Virtual Environments (VEs), which automates VEs setup, deployment and management, regardless of the underlying Virtual Machine Manager layer (i.e., Xen, KVM, or VMware). The Cloud Manager layer is instead able to transform the existing infrastructure into a cloud, providing cloud-like interfaces and higher-level characteristics for security, contextualization and VM disk image management.

In a cloud architecture designed according to the aforementioned three-layered stack, all the cloud components and their respective functions are clearly defined and separated, thus introducing simplicity and efficiency when the cloud middleware has to be modified or new features have to be added. In our work, we exploited such modular characteristics of the layered cloud architecture, and introduced a new component within the Cloud Manager layer (depicted in the top part of Figure 2), named *Cross-Cloud Federation Manager* (CCFM). The CCFM has been conceived for enabling each cloud to perform all the operations needed to pursue the target of the federation establishment.

The cross-cloud scenario we are considering can be seen as an highly dynamic environment: new clouds, offering different available resources and different authentication mechanisms could appear, while others could disappear. Taking into account such dynamism, when a home cloud needs to "lease" external resources from a foreign cloud, the first step the home cloud will perform refers to the

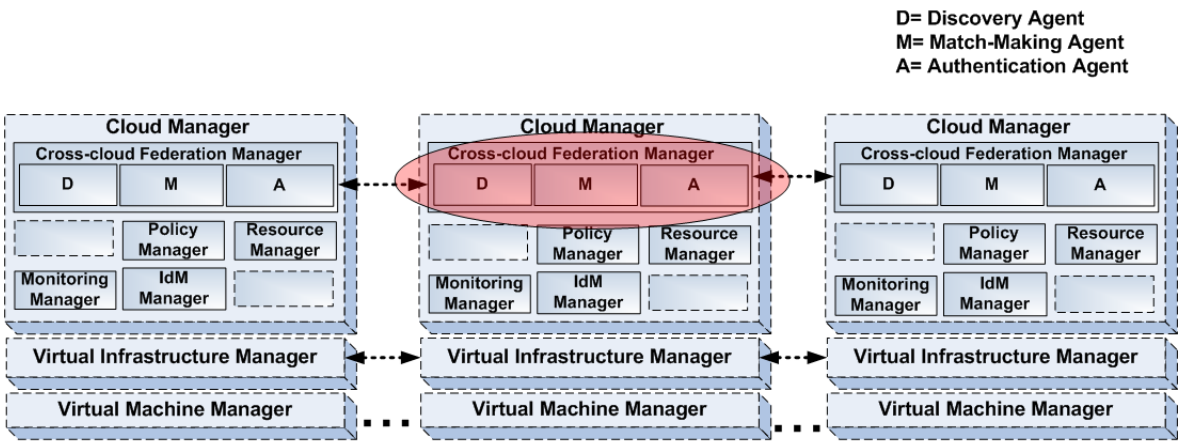

Figure 2. General federated three-tier cloud architectures.

discovery (phase 1) of the foreign cloud, which properly *matches* (phase 2) its requirements (both in terms of available resources and supported authentication mechanisms). Once these two steps have been performed, and the best foreign cloud has been found, in order to establish a secure interaction between the home cloud and the selected foreign cloud, an *authentication* (phase 3) process will begin.

The CCFM module represents the main "actor" in our three-phase federation model. In our design, it consists of three different subcomponents (agents) each addressing a different phase of the federation model:

- The *discovery agent* manages the discovery process among all the available clouds within the dynamic environment. Since its state is pretty flexible and dynamic, the discovery process has to be implemented in a totally distributed fashion: all the discovery agents must communicate exploiting a p2p approach.
- The *match-making agent* accomplishes the task of choosing the more convenient foreign cloud, evaluating all the parameters regarding the QoS, available resources and available authentication mechanisms. By means of specific algorithms, this agent is able to evaluate from all the available (discovered) clouds, the ones that best "fit" the requirements (e.g. the load capacity in terms of resources leasing and the supported authentication methods) of its home cloud.
- The *authentication agent*, cooperating with third parties trusted entities, takes part in the creation of a security context between home and foreign clouds. When the authentication phase begins, the home cloud authentication agent contacts its "peer" on the foreign cloud: the authentication process between such agents (and thus the clouds) will be lead exchanging authentication information in form of meta-data, also involving trusted third parties in the process. The Authentication Agent communicates both with other peers and third parties via web service interfaces.

The accomplishment of the authentication process, carried out by the authentication agents of both home and foreign clouds, leads to the establishment of a secure and direct connection between the related VI Manager Layer of the same clouds. As consequence, the home cloud will be able to instantiate (or migrate) Virtual Resources (VMs) on the Foreign Cloud in a secure environment. The concept of migration can be seen as the opportunity to move the Virtual Machines not only in intra-site domain but also to transfer them on federated inter-site domains. In this case the migration might occur across subnets, among hosts that do not share storage and across administrative boundaries.

Although in Section IV-C we'll describe the three phases needed to pursue the cloud federation, the main scope of this paper refers to the solution of the cloud SSO authentication problem.

In our work, the practical solution to overcome the authentication problem is the introduction the well-known concept of Identity Provider (IdP) along with a new SAML profile (further details are presented in Section V).

C. The Three-Phase Cross-Cloud Federation Process

In this Section, we provide a more detailed description of the three-phase cross-cloud federation model considering the scenario represented in Figure 3. As depicted, such a scenario includes both home clouds and foreign clouds, which are represented according to the three-tier architecture already discussed. More specifically, each cloud platform is composed as follows: The highest stack level includes the CCFM module, which comes with its own agents (discovery, match-making and authentication), instead the underlying layer includes a generic VI Manager. Instead, for our dissertation it is not relevant the specific solution employed at the lowest layer.

We remarked the need of providing a global authentication mechanism exploitable from all the entities belonging to the

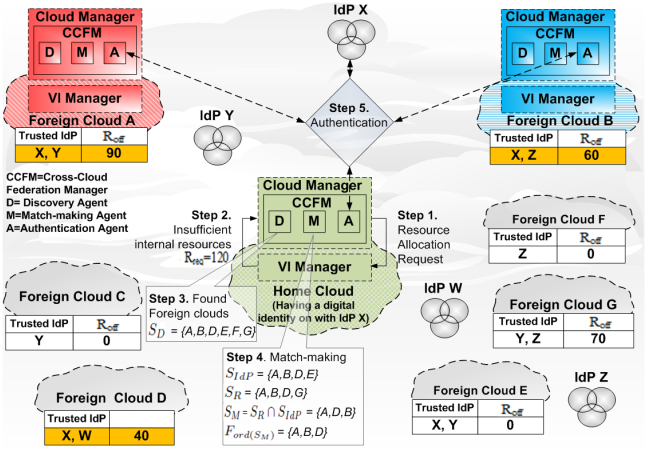

Figure 3. Example of cross-cloud federation establishment.

cloud federation. In Figure 3, together with home clouds and foreign clouds, IdPs are also depicted. An IdP is a provider of digital identity representing a trusted third party, which provides authentication services to its clients. In such scenario we assume each home cloud must have one digital identity at least on one IdP (even though many cloud digital identities may exist on different IdPs), whereas each foreign cloud must be trusted or compliant with one or more IdPs. Before explaining our motivation to the introduction of IdP within our scenario, we provide a description of the three phases needed to achieve the cloud federation.

In the scenario of federation establishment depicted in Figure 3, during the step 1, the home cloud manager layer receives a request for services from its clients and sends a resource allocation request (i.e. virtual machines) to the underlying VI manager layer. In step 2 the home cloud VI manager, evaluating its instantaneous workload, replies to the request notifying it has not enough resources. In step 3 (the discovery phase) the home cloud manager decides to ask for resources to foreign clouds: the resource request is forwarded to the CCFM, which, by means of its discovery agent, will begin the *discovery* process to obtain a list of all the available foreign clouds. The discovery phase can exploit whatever p2p approach to achieve the complete list of cloud providers. Each discovered foreign cloud is associated to a set of meta-data describing several cloud information: the amount and type of the resources available for leasing, the offered SLA level and the supported IdP(s). In this particular example, the agent has found the set of discovered foreign clouds $S_D = \{A, B, C, D, E, F, G\}$.

In step 4 the *match-making* phase begins: the match-making agent of the home cloud selects from the set of discovered foreign clouds S_D the ones, which fits its requirements. The adopted criteria to perform the selection is based on two different evaluation tasks: in the first one, starting from the foreign clouds set S_D, a new subset $S_R = \{A, B, D, G\}$ is obtained considering the foreign clouds better satisfying the home cloud request in terms of resources availability (CPU, RAM, storage) and QoS. In the second evaluation task, starting from the discovered foreign clouds set S_D, the match-making agent selects the subset of foreign clouds $S_{IdP} = \{A, B, D, E\}$, having trusted relationship(s) with the IdP(s) on which the home cloud already has a digital identity. In this example foreign clouds A, B, D, and E are trusted with the IdP X, which provides authentication services to the home cloud guaranteeing for its digital identity. The subsequent operation accomplished by the match-making agent refers to the definition of the set of match-made clouds $S_M = S_R \cap S_{IdP}$.

We now define the metrics R_{req} and $R_{off}(F_i)$ representing respectively a measure of the resources requested by the home cloud, and a measure of the resources offered by the foreign cloud F_i. The value of the metric is obtained evaluating different parameters such as CPU, RAM, storage and QoS for both R_{req} and $R_{off}(F_i)$. In order to identify which foreign clouds fit the home cloud requirements, the match-making agent achieves a list of preferred foreign clouds $F_{ord(S_M)} = \{F_1, F_2, \ldots, F_n\}$ considering the set S_M and ordering its element by the R_{off} value, in a descending order.

Considering the example depicted in Figure 3, $S_M = \{A, D, B\}$ and $F_{ord(S_M)} = \{A, B, D\}$. The match-making agent has to consider the resources provided by the first k foreign clouds of $F_{ord(S_M)}$ to satisfy the condition $R_{req} \leq \sum_{i=1}^{k} R_{off}(F_i), 1 \leq k \leq n$ (in the scenario depicted in Figure 3, we assume $k = 2$ and consequently both foreign clouds A and B will be chosen to establish the federation).

In step 5 (authentication phase), in order to establish a federation with foreign cloud A and B, a cloud SSO *authentication* process has to be started by the home cloud. Such process will involve: the authentication agent of the home cloud, the corresponding peers of the foreign cloud A and B and the IdP X (trusted with A and B, on which the home cloud has a digital identity) where the home cloud performs a SSO log-in. Once the home cloud and foreign cloud A authentication agents establish a trust context, their respective underlying VI manager layers setup a low-level trust context allowing the cross-cloud resource provisioning. Therefore, the home cloud VI manager will be able to instantiate virtual resources on the foreign cloud VI manager. Even if cross-cloud federation has to be established also with foreign cloud D, no further authentication tasks would be needed because foreign cloud D has already a trusted relationship with IdP X.

As can be perceived, the employment of the IdPs presents some advantages well fitting our cross-cloud federation scenario: even though each cloud has its internal security mechanisms, whatever the *foreign cloud* is, regardless of its authentication mechanisms, by means of IdPs a *home cloud* will be able to authenticate itself with other foreign

clouds already having a trust relationship, exploiting the well-known concept of SSO. The resource provisioning in cross-cloud federation may be solved establishing trust relationships between the clouds using several IdPs containing the credentials of the cloud asking for resources. Section V better describes the steps involved in phase 5, pointing out the technologies employed to implement the authentication and the set of information exchanged between the involved entities. The same Section describes our new SAML profile designed to accomplish the cloud SSO authentication in a federated scenario.

V. The Authentication Phase Using SAML

In this Section, after a brief description of the SAML standard, we focus on the authentication phase (step 5 of Figure 3) of our three-phase cross-cloud federation model performed by the Authentication Agent. More specifically, using the SAML technology we propose a new *Cross-Cloud Authentication Agent SSO Profile*, which describes the messages exchanging flow between a home cloud, foreign clouds and IdPs during the establishment of a trust context.

A. SAML Technology Overview

SAML is an XML-based standard for exchanging authentication and authorization assertions between security domains, more specifically, between an Identity Provider (IdP) (a producer of assertions) and a generic Service Provider (SP) (a consumer of assertions). SAML consists of: a subject, a person or a software/hardware entity that assumes a particular digital identity and interacts with an online application, composed of several heterogeneous systems; a SP or relying party, a system, or administrative domain, that relies on information supplied to it by the Identity Provider; an IdP or asserting party, a system, or administrative domain, that asserts information about a subject. In literature, such a model is also referred as IdP/SP.

The aim of SAML is enable a principal to perform SSO. This means a principal, by means of its IdP, must be able to authenticate itself once gaining the access to several trusted service providers which might use also different security technologies. SAML assumes the principal has enrolled in at least one identity provider offering SSO authentication services. The main advantage of SAML is it does not care how authentication services are implemented. In fact, the whole SAML authentication process of a subject on a service provider is performed by means of a set of messages exchanging SAML security assertions. Service providers, in order to authenticate a principal, merely relies on the assertion sent by the trusted IdP (on which the principal is enrolled).

SAML combines four key concepts: assertion, binding, protocol and profile. Assertion consists of a package of information that supplies one or more statements (i.e. authentication, attribute, and authorization decision) made by the IdP. Authentication statement is perhaps the most important meaning the IdP has authenticated a subject at a certain time. A Protocol (i.e. Authentication Request, Assertion Query and Request, Artifact Resolution, etc) defines how subject, service provider, and IdP might obtain assertions. More specifically, it describes how assertions and SAML elements are packaged within SAML request and response elements. A SAML binding (i.e. SAML SOAP, Reverse SOAP (PAOS), HTTP Redirect (GET), HTTP POST Binding, etc) is a mapping of a SAML protocol message over standard messaging formats and/or communications protocols. For example, the SAML SOAP binding specifies how a SAML message is encapsulated in a SOAP envelope. A profile (i.e. Web Browser SSO, Enhanced Client or Proxy (ECP), Single Logout, Attribute Profiles, etc) is a technical description of how a particular combination of assertions, protocols, and bindings defines how SAML can be used to address particular scenarios.

B. The SAML Cross-Cloud Authentication Agent SSO (CCAA-SSO) Profile

The authentication agent has been designed both to manage the digital identity of the home cloud and to perform authentication tasks sending/receiving authentication requests to/from foreign clouds, interacting with their respective peer modules. More specifically, the authentication agent does not directly manages the digital identity of the cloud, but uses one or more trusted IdPs acting as guarantor when the agent likes to authenticate the home cloud with other foreign clouds during the federation establishment.

To address the cloud SSO authentication problem, during the cross-cloud federation establishment we developed a new SAML profile named Cross-Cloud Authentication Agent SSO (CCAA-SSO). This profile was designed to enable a home cloud to perform SSO authentication on several foreign clouds both having a trusted relationship with the home cloud's IdP and regardless their security mechanisms. Such an authentication process is fundamental for the subsequent establishment of a secure interaction between the home cloud VI manager and one or more foreign cloud VI manager(s). Once the secure interaction has been established the home cloud is able to gain the access to the required resources offered by the foreign cloud. More specifically, the profile defines the messages exchange flow between the home cloud, the foreign clouds, and the IdP, solving the cloud SSO authentication problem. The implementation of the profile has been accomplished using and extending the java libraries of the OpenSAML project [18] for both the authentication agents and the IdP. In order to accomplish such tasks, the agent has been developed exposing a web service interface using the SOAP [19] technology, but nothing prevents the adoption of other web service technologies such as REST, JAX-RPC, or XML-RPC.

In a CCAA-SSO profile use case, both the home cloud

and the foreign cloud, by means of their own Authentication Agents, represent respectively the subject and the relying party, whereas the IdP acts as the third party asserting to a foreign cloud the trustiness of the home cloud identity. The CCAA-SSO profile has been designed as a combination of the following SAML elements: an assertion including an authentication statement, a request-response protocol, and a SAML SOAP Binding.

Considering the scenario already pointed out in Section IV-C, in the following we describe the authentication process previously marked as phase 5 keeping in mind our SAML CCAA-SSO profile considering a VI Manager. In Figure 4 is shown the flow of messages exchanged between the home cloud, the foreign cloud A and the IdP X, putting aside for the time being foreign cloud B. More specifically, inside each cloud both Authentication Agent and the VI Manager are involved in the process. In step 5.1 the Authentication

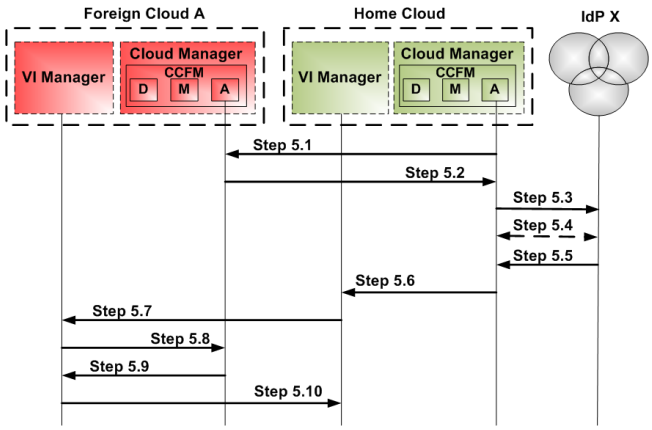

Figure 4. Sequence diagram describing the steps of the CCAA-SSO profile during the authentication of the home cloud with the foreign cloud A by means of the IdP X.

Agent, on behalf of the home cloud manager, forwards to the corresponding peer of the foreign cloud A a SOAP request for a set of virtual resources by means of a XML document. In step 5.2 the Authentication Agent of the foreign cloud A responds to the home cloud with a SAML authentication request enveloped in a SOAP message. In step 5.3 the Authentication Agent of the home cloud unpacks the authentication request received at step 5.2 and forwards it via SAML/SOAP message to the IdP X, making a SSO request. As a valid trust context does not exist, in step 5.4 the IdP X authenticates the home cloud using a given security technology (the independence from the security technology used by each cloud is accomplished). In step 5.5, as the home cloud identity is verified, the IdP X responds to the authentication request by means of a SAML/SOAP response message, signing it with its private key. In step 5.6 the Authentication Agent of the home cloud unpacks the authentication assertion received in step 5.5 and forwards it to the underlying VI Manager. In step 5.7 the VI manager of the home cloud sends the authentication assertion via SAML/SOAP to the corresponding peer of the foreign cloud A. In step 5.8 the VI manager of the foreign cloud B forwards the received authentication statement to its authentication agent, which proves its correctness verifying the digital sign using the public key of the IdP X (see step 5.5). In step 5.9 the VI manager of the foreign cloud B receives a notification about the authentication assertion validity and authenticate the home cloud VI Manager establishing a secure interaction. In step 5.10 the VI manager of the foreign cloud provides the resources requested by the home cloud at step 5.1

The authentication process of the home cloud with the foreign cloud B is analogous to the one already described for foreign cloud A, with one important difference: since the home cloud has already performed the authentication on the IdP X in the step 5.4, no further authentication is needed because a trust context already exists (the SSO is thus accomplished). Therefore, the SAML CCAA-SSO profile combines both security and flexibility ensuring cloud SSO authentication in cross-cloud federation environments between clouds representing a possible solution for secure federated cloud interactions.

In the Section VII the same sequence of steps involved in the CCAA-SSO profile will be considered again, in a new scenario where a new Virtual Infastructure Manager named CLEVER will be introduced: thanks to its features, the CLEVER middleware well fits the requirement of dynamic resource sharing that is a pillar of a federated cloud environment. All the details regarding this middleware will be discussed in the following Section (VI).

VI. CLEVER: A VIRTUAL INFRASTRUCTURE MANAGER

In this Section we will focus on the description of CLEVER, a VI Manager which aims to provide *Virtual Infrastructure Management* services and suitable interfaces at the *High-level Management* layer to enable the integration of high-level features such as Public Cloud Interfaces, Contextualization, Security and Dynamic Resources provisioning. After a brief description of both its architecture and its main features, the way by means it is possible to interconnect different CLEVER domains will be discussed.

A. CLEVER: an Architectural Overview

Looking at the existing middleware implementations, which act as *High-level Cloud Manager* [20], [21], it can be said that their architecture lacks modularity: it could be a difficult task to change these cloud middleware for integrating new features or modifying the existing ones. CLEVER instead intends granting an higher scalability, modularity and flexibility exploiting the plug-ins concept. This means that other features can be easily added to the middleware just introducing new plug-ins or modules within its architecture without upsetting the organization.

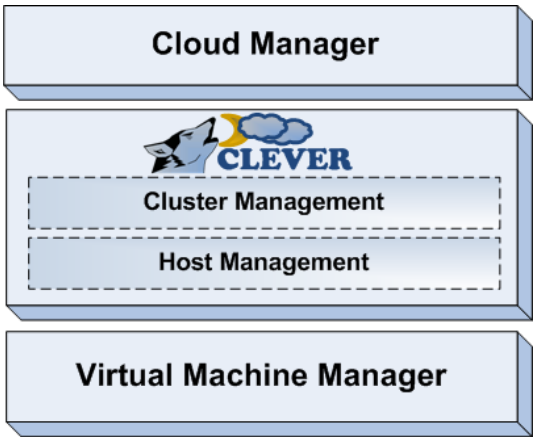

Figure 5. How the VI Manager layer is implemented in CLEVER: the cluster management and the host management

Furthermore, analysing the current existing middleware [22], [23], which deal with the *Virtual Infrastructure Management*, some new features could be added within their implementation in order to achieve a system able to grant high modularity, scalability and fault tolerance. The main idea on which CLEVER is based, finds in the terms flexibility and scalability its key-concepts, leading to an architecture designed to satisfy the following requirements: 1) *persistent communication* among middleware entities; 2) *transparency* respect to "user" requests; 3) *fault tolerance* against crashes of both physical hosts and single software modules; 4) heavy *modular design* (e.g. monitoring operations, managing of hypervisor and managing of VEs images will be performed by specific plug-ins, according to different OS, different hypervisor technologies, etc); 5) *scalability* and *simplicity* when new resources have to be added, organized in new hosts (within the same cluster) or in new clusters (within the same cloud); 6) automatic and optimal *system workload balancing* by means of dynamic VEs allocation and live VEs migration. The typical scenario where CLEVER could be deployed consists of a set of physical hardware resources (i.e., a cluster) where VEs are dynamically created and executed on the hosts considering their workload, data location and several other parameters. The basic operations our middleware should perform refer to: 1) Monitoring the VEs behavior and performance, in terms of CPU, memory and storage usage; 2) Managing the VEs, providing functions to destroy, shut-down, migrate and network setting; 3) Managing the VEs images, i.e., images discovery, file transfer and uploading.

Considering the concepts stated in [4] and looking at Figure 5, such features, usually implemented in the *Virtual Infrastructure Management* layer, can be further analyzed and arranged on two different sub-layers: *Host Management* (lower) and *Cluster Management* (higher).

Grounding the design of the middleware on such logical subdivision and taking into account the satisfaction of all the above mentioned requirements, the simplest approach to design our middleware is based on the architecture schema depicted in Fig. 6, which shows a cluster of n nodes (also an interconnection of clusters could be analyzed) each containing a *host level* management module (Host Manager). A single node may also include a *cluster level* management module (Cluster Manager). All these entities interact exchanging information by means of the *Communication System* based on the Extensible Messaging and Presence Protocol (XMPP) [24]. In particular, the main entities of CLEVER, communicates each other "talking" and exchanging messages within the same XMPP room, like in a traditional chat. As the Figure 6 shows, a CM coordinates the HMs of the whole cluster sending XMPP messages to them by means of a multi-user-chat.

Finally, the set of data necessary to enable the middleware functioning is stored within a specific *Database* deployed in a distributed fashion.

Figure 6 shows the main components of the CLEVER architecture, which can be split into two logical categories: software agents (typical of the architecture itself) and the tools they exploit. To the former set belong both *Host Manager* and *Cluster Manager*:

- **Cluster Manager (CM)** acts as an interface between the clients (software entities, which can exploit the cloud) and the HM agents. CM receives commands from the clients, performs operations on the HM agents (or on the database) and finally sends information to the clients. It also performs the management of VE images (uploading, discover, etc.) and the monitoring of the overall state of the cluster (resource usage, VEs state, etc.). Following our idea, at least one CM has to be deployed on each cluster but, in order to ensure higher fault tolerance, many of them should exist. A master CM will exist in active state while the other ones will remain in a monitoring state, although client messages are listened whatever operation is performed.
- **Host manager (HM)** performs the operations needed to monitor the physical resources and the instantiated VEs; moreover, it runs the VEs on the physical hosts (downloading the VE image) and performs the migration of VEs (more precisely, it performs the low level aspects of this operation). To carry out these functions it must communicate with the hypervisor, hosts' OS and distributed file-system on which the VE images are stored. This interaction must be performed using a plug-ins paradigm.

Regarding the tools such middleware components exploit, we can identify the *Distributed Database* and the *XMPP Server*:

- **Distributed Database** is merely the database containing the overall set of information related to the

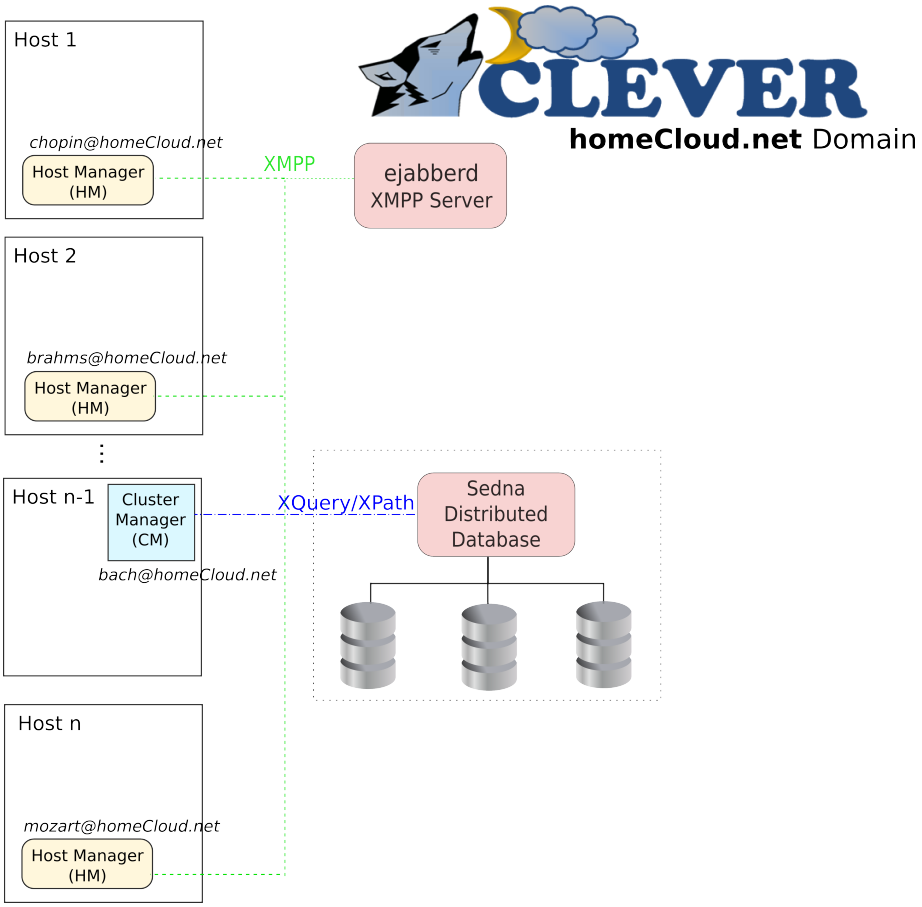

Figure 6. CLEVER reference scenario.

middleware (e.g. the current state of the VEs or data related to the connection existing on the Communication System). Since the database could represent a centralized point of failure, it has to be developed according to a well structured approach, for enabling fault tolerance features. The best way to achieve such features consists of using a Distributed Database. In the current CLEVER implementation, the database is based on the sedna native XML database system [25]. Sedna provides a full range of core database services (e.g., persistent storage, ACID transactions, security, indices, hot backup). Flexible XML processing facilities include W3C XQuery implementation, tight integration of XQuery with full-text search facilities and a node-level update language.

- **XMPP Server** is the "channel" used to enable the interaction among the middleware components. In order to grant the satisfaction of our requirements, it is able to offer: decentralization (i.e., no central master server should exist: such capability in native on the XMPP) in a way similar to a p2p communication system for granting fault-tolerance and scalability when new hosts are added in the infrastructure; flexibility to maintain system interoperability; security based on the use of channel encryption: since the XMPP Server also could exploit the distributed database to work, the solution enables an high fault tolerance level and allows system status recovery if a crash occurs. All these features are guaranteed in the current implementation by means of the employment of the Ejabberd XMPP server [24]. Ejabberd is a Jabber/XMPP instant messaging server, licensed under GPLv2 (Free and Open Source), written in Erlang/OTP. Among other features, ejabberd is cross-platform, fault-tolerant, clusterable and modular.

B. How CLEVER Supports Interdomain Communication

Even though the XMPP uses a client/server model there is not a central authoritative server. As anyone may run its own XMPP server on its own domain, it is the interconnection among these servers which makes up the XMPP network. Every user on the network has a unique Jabber ID (JID). To avoid requiring a central server to maintain a list of IDs, the JID is structured like an e-mail address with a user name and a domain name for the server where that

user resides, separated by an at sign (@). For example, considering the CLEVER scenario a CM could be identified by a JID *bach@homeCloud.net*, whereas a HM could be identified by a JID *liszt@foreignCloudA.net*: *bach* and *liszt* respectively represent the host names of the CM and the HM, instead *homeCloud.net* and *foreignCloudA.net* represent the domains of the cloud resources.

Let us suppose that *bach@homeCloud.net* wants to communicate with *liszt@foreignCloudA.net*, *bach* and *liszt*, each respectively, have accounts on the *homeCloud.net* and *foreignCloudA.net* XMPP servers. When *bach* wants to start the communication, a sequence of events is triggered:

1) *bach* sends its message to the *homeCloud.net* server
2) The *homeCloud.net* server opens a connection to the *foreignCloudA.net* server.
   a) If *homeCloud.net* server has successfully performed the authentication with *foreignCloudA.net* server, the messagge is forwarded.
   b) If *homeCloud.net* server has not successfully performed the authentication on *foreignCloudA.net* server, the message is dropped.
3) If 2a is verified, the *foreignCloudA.net* server checks to see if *liszt* is currently connected. If not, the message is stored for later delivery.
4) If *liszt* is online, the *foreignCloudA.net* server delivers the message to *liszt*.

A CLEVER cluster includes a set of HMs, orchestrated by a CM, all acting on a specific domain and connected to the same XMPP room. Each HM is deployed in a physical host and is responsible to manage its computing and storage resources according to the commands given by the CM. The idea of federation in CLEVER environments is founded on the concept that if a CLEVER cluster on a domain needs of external resources of other CLEVER clusters, acting on different domains, a sharing of resources can be accomplished, so that the resources belonging to a domain can be logically included in another domain. Within CLEVER this is straightforward by means of the built-in XMPP features.

Considering the aforementioned domains *homeCloud.net* and *foreignCloudA.net*, in scenarios without federation, they respectively include a different XMPP rooms (i.e., *cleverRoom@homeCloud.net* and *cleverRoom@foreignCloudA.net*) on which a single CM, responsible for the administration of the domain, communicates with several HMs, typically placed within the physical cluster of the CLEVER domain. Instead, considering a federated scenario among the two domains, if the CM *bach* of the *homeCloud.net* domain needs of external resources, it could invite within its *cleverRoom@homeCloud.net* room one or more HMs of the *foreignCloudA.net* domain. As previously stated, in order to accomplish such a task a trust relationship between the *homeCloud.net* and the *foreignCloudA.net* XMPP server has to be established. Such a concept will be better clarified in Section VII by means of a concrete use case.

VII. The CCAA-SSO Profile Applied to CLEVER

In Section V-B we described the authentication process in a generic cloud environment, pointing out the sequence of step involved in the CCAA-SSO profile. In the following, keeping in mind the aforementioned description of the CLEVER cloud middleware, we aim to discuss how the CCAA-SSO profile may be used in a cloud scenario where CLEVER acts as VI Manager.

This Section will provide more details about the XML documents exchanged among the actors of the profile and will clarify how the resources may be shared among different CLEVER domains exploiting the features offered by the XMPP. In order to describe the process, the same sequence of step already analyzed in Section V-B will be considered. In this description, for simplicity, instead of naming the steps as 5.1-5.10, the notation 1-10 will be employed.

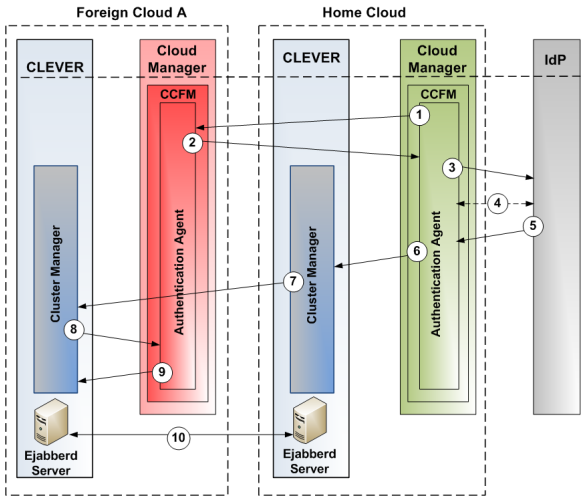

Figure 7. Sequence diagram describing the steps of the CCAA-SSO profile during the authentication between the CLEVER home cloud and the CLEVER foreign cloud A by means of the IdP X.

When the CM of the CLEVER site identified from the domain *homeCloud.net* realizes it has not enough resources to satisfy the stipulated SLA, it tries to ask further resources to an external cloud. These resources are managed from the CM exchanging XMPP messages on the room *cleverRoom@homeCloud.net*, in order to orchestrate the available physical resources (each physical host is managed by an HM).

Supposing that the discovery phase has been accomplished from the *homeCloud.net* and the cloud selected for the federation is *foreignCloudA.net*, the authentication phase can begin according to the steps already

reported in the Section V-B. In this CLEVER scenario, more specifically, during the step 1 the Authentication Agent, on behalf of the home cloud manager, forwards to the corresponding peer of the foreign cloud A a SOAP request (by means of a XML document) for a set of HMs that should join its *cleverRoom@homeCloud.net* to increase the available physical resources. In the SOAP request message reported below, such document is embedded inside the <ResourceType> element and is not depicted for briefness.

```
<?xml version="1.0" encoding="UTF-8"?>
<S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/
    envelope/">
   <S:Header/>
   <S:Body>
      <ns2:AA-ForeignCloud-A-ResReq xmlns:ns2="https://
          cloudA.net/SAML2/">
         <ResourceType>"XML resource description
             document"</ResourceType>
      </ns2:AA-ForeignCloud-A-ResReq>
   </S:Body>
</S:Envelope>
```

In step 2 the Authentication Agent of the foreign cloud A responds to the home cloud with a SAML authentication request containing an authentication query. Considering the underlying SAML/SOAP response, the authentication request is provided by means of the element <samlp:AuthnRequest...>.

```
<?xml version="1.0" encoding="UTF-8"?>
<S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/
    envelope/">
   <S:Body>
     <ns2:AA-ForeignCloud-A-ResReqResponse xmlns:ns2="
         http://webservices/">
       <return>
        <samlp:AuthnRequest xmlns:samlp="urn:oasis:names
            :tc:SAML:2.0:protocol" xmlns:saml="urn:
            oasis:names:tc:SAML:2.0:assertion" ID="cba2
            " Version="2.0" IssueInstant="2010-11-12T17
            :23:32Z" AssertionConsumerServiceIndex="0">
          <saml:Issuer>https://cloudA.net/SAML2</saml:
              Issuer>
          <samlp:NameIDPolicy
           AllowCreate="true"
           Format="urn:oasis:names:tc:SAML:2.0:nameid-
               format:transient"/>
        </samlp:AuthnRequest>
       </return>
     </ns2:AA-ForeignCloud-A-ResReqResponse>
   </S:Body>
</S:Envelope>
```

In step 3 the Authentication Agent of the home cloud unpacks the authentication request received at step 2 and forwards it via SAML/SOAP to the IdP X, making a SSO request. Since a valid trust context does not exist, in step 4 the IdP X authenticates the home cloud using a given security technology (the independence from the security technology used by each cloud is accomplished). In step 5, since the home cloud identity is verified, the IdP X responds to the authentication request by means of the following SAML/SOAP response, identified by the element <samlp:Response...>. Such element contains an assertion (see element <saml:Assertion...>) with an authentication statement (see element <saml:AuthnStatement...>) and has been signed by the IdP X using its private key (see elements <saml:Issuer> and <ds:Signature xmlns:ds="http//www.w3.org/2000/09/xmldsig#" >).

```
<?xml version="1.0" encoding="UTF-8"?>
<S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/
    envelope/">
<S:Body>
<ns2:IdpX-SSO-ServiceResponse xmlns:ns2="http://
    webservices/">
<return>
   <samlp:Response xmlns:samlp="urn:oasis:names:tc:SAML
       :2.0:protocol" xmlns:saml="urn:oasis:names:tc:
       SAML:2.0:assertion" ID="62af" InResponseTo="cba2"
        Version="2.0" IssueInstant="2010-11-12T17:23:34Z
       " Destination="https://cloudA.net/SAML2/SSO/SOAP
       ">
  <saml:Issuer>https://idpx.net/SAML2</saml:Issuer>
  <samlp:Status>
  <samlp:StatusCode Value="urn:oasis:names:tc:SAML:2.0:
      status:Success"/>
  </samlp:Status>
  <saml:Assertion xmlns:saml="urn:oasis:names:tc:SAML
      :2.0:assertion" ID="5c4e" Version="2.0"
      IssueInstant="2010-01-12T18:35:23Z">
    <saml:Issuer>https://idpx.net/SAML2</saml:Issuer>
    <ds:Signature xmlns:ds="http://www.w3.org/2000/09/
        xmldsig#">
     mgQpzczIazNLSIr8qp7mt0C8jWLBrRsIChVGDML44
     tfZDPCOZfGfbWNBy97ODoEvTtptJtpjp9NN
     JTSweVTofRcy8tHrvLuJnLmMmDbE50KsRoo+yA==
        z6h5g2KOdBkZS7g9w0TJFKlI/OJUOhyodpRr
     8XY9+h/4euIVxg5vXuD6PldBqWgKYtY84+910IP7TXQJS/
        cblOCIf2TdMo55vR0QGDYdBt2yRXd1
     wCUO93dtaSAF6WVid55JE4oraYFEFMfOmgQpzczIazNLSI
     r8qp7mt0C8jWLBrRsIChVGDML44tfZ
     hdttW0jOIazNLSIr8qp7mt0C8jWLBrRsIChVGDML4s+
        xEyyN4hrCEvz2hIcLYA5Q4B1HTKryMCw5
     PIJt0eaTeMicjAyrN+iynUjpW2uAgCvPYHbk4Le/i
    </ds:Signature>
    <saml:Subject>
      <saml:NameID Format="urn:oasis:names:tc:SAML:2.0:
          nameid-format:transient">
        5a42edc7-6439-4de9-12d2-836a74df279c
      </saml:NameID>
      <saml:SubjectConfirmation Method="urn:oasis:names:
          tc:SAML:2.0:cm:bearer">
        <saml:SubjectConfirmationData InResponseTo="dfa6"
            Recipient="https://cloudA.net/SAML2/SSO/
            SOAP" NotOnOrAfter="2010-11-12T17:24:34Z"/>
      </saml:SubjectConfirmation>
    </saml:Subject>
    <saml:Conditions NotBefore="2010-11-12T17:23:34Z"
        NotOnOrAfter="2010-11-12T17:24:34Z">
      <saml:AudienceRestriction>
        <saml:Audience>https://cloudA.net/SAML2</saml:
            Audience>
      </saml:AudienceRestriction>
    </saml:Conditions>
    <saml:AuthnStatement AuthnInstant="2010-11-12T17
        :24:50Z" SessionIndex="21a4">
      <saml:AuthnContext>
        <saml:AuthnContextClassRef>
          urn:oasis:names:tc:SAML:2.0:ac:classes:
              PasswordProtectedTransport
        </saml:AuthnContextClassRef>
      </saml:AuthnContext>
    </saml:AuthnStatement>
  </saml:Assertion>
 </samlp:Response>
</return>
</ns2:IdpX-SSO-ServiceResponse>
</S:Body>
</S:Envelope>
```

In step 6 the Authentication Agent of the home cloud

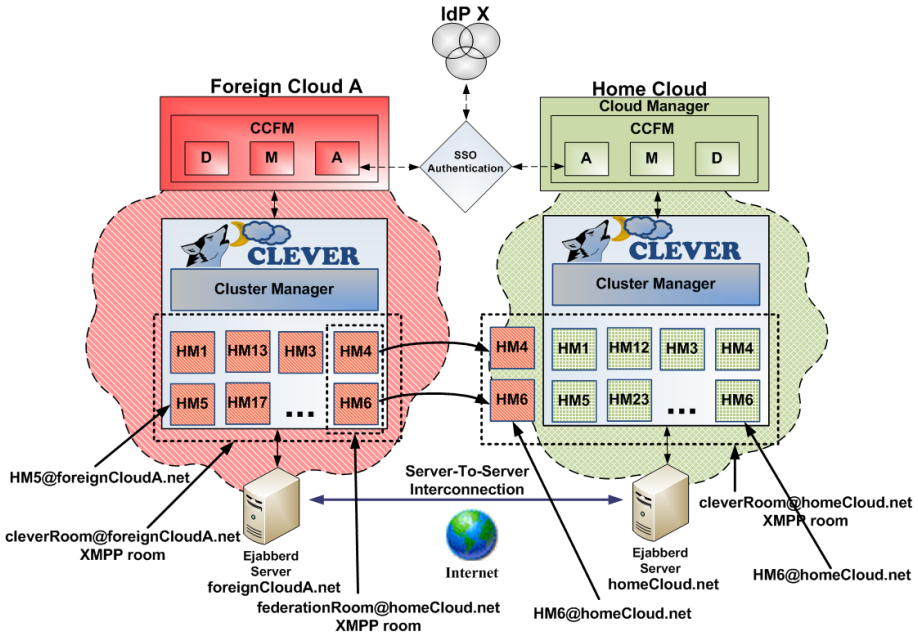

Figure 8. Example of resource sharing among a Home Cloud and a ForeignCloud in a federated scenario.

unpacks the authentication assertion received in step 5 and forwards it to the underlying CLEVER middleware where it is captured from the CM. This latter, in step 7, sends the authentication assertion via SAML/SOAP to the corresponding peer of the foreign cloud A. In step 8 the CLEVER CM of the foreign cloud B forwards the received authentication statement to its authentication agent, which proves its correctness verifying the digital sign using the public key of the IdP X (see step 5).

The Sequence of steps we reported above is represented on the Figure 8. More specifically, the authentication process based on SAML, the IdP and the Authentication Agent is represented on the top part of the Figure by means of the relation *SSO Authentication* which involves these three entities of the scenario.

In step 9 the CLEVER of the foreign cloud B receives a notification about the authentication assertion validity and authenticates the CLEVER home cloud establishing a secure interaction. This interaction leads to the establishment of a server-to-server connection among the homeCloud.net and foreignCloudA.net XMPP servers and therefore an interconnection among the considered CLEVER domains. This operation is represented in the Figure 8 with the double arrow joining the two Ejabberd Server of the two different domains homeCloud.net and foreignCloudA.net (Server-To-Server Interconnection).

In the step 10, the set of resources requested by the home cloud is allocated within the foreign cloud domain: as showed in Figure 8 a temporary XMPP room *federationRoom@foreignCloudA.net*, including two HMs (HM4 and HM6), is created. Furthermore, these HMs will be also included in the *cleverRoom@foreignCloudA.net* room, marked as "rented" as XMPP presence status. This allows the CLEVER CM of the foreign cloud to passively monitor the performance of the rented cluster resources (i.e., the rented HMs) although these are used by the home cloud.

In this way, the CLEVER CM of the foreign cloud is able to rent the HMs included within the *federationRoom@foreignCloudA.net* to the home cloud. As depicted in the Figure, the CLEVER CM of the home cloud can invite the HMs of the *federationRoom@foreignCloudA.net* room to join its *cleverRoom@homeCloud.net* room, so that it can enlarge its available resources for the instantiation of virtual machines.

VIII. Conclusions and Future Works

In this paper we focus on cross-cloud environments, debating the way in which different clouds acting on different administrative domains can establish federation relationships in order to lend and borrow virtualization resources. An approach for the federation establishment consisting of three phases (Discovery, Match-Making, and Authentication) was introduced, and keeping in mind a generic three-tier cloud architecture (from the bottom: VM Manager, VI Manager, and Cloud Manager), a Cross-Cloud Federation Manager (CCFM) responsible for the accomplishment of the aforementioned three phases has been designed and described.

More specifically, the designed CCFM consists of three software agents: Discovery, Match-Making, and Authentication, each one responsible for the accomplishment of a phase during the federation establishment process. In this

paper, we particularly focused on the Authentication Agent, designing a SAML CCAA-SSO profile for generic three-tier cloud architecture.

In our testbed, a use case of the application of the designed SAML CCAA-SSO profile has been performed using cloud platforms consisting on the CLEVER VI Manager and on an implementation of the Authentication Agent of the CCFM module. Details about the sequence of SAML messages exchanged between the involved cloud and the IdP have been provided also by means of several example of XML document captured during the authentication establishment.

Finally, a discussion on the way in which federated CLEVER-based clouds are able, after authentication, to lend and borrow virtualization resources (i.e., in the CLEVER terminology HMs) is provided, also discussing what it happens behind the scenes from the point of view of CLEVER.

In future works, as in a cross-cloud federated environment including hundreds of clouds, considering only one IdP is too much restrictive, we plan to study scenarios including different distributed IdPs also having trustiness relationships each other. In this way we will aim to evaluate the amount of authentications and IdP enrollments needed, either employing real testbeds or by means of a simulated environment, including hundreds of clouds dynamically joining and leaving federations.

REFERENCES

[1] T. Bittman, "The evolution of the cloud computing market," *Gartner Blog Network, http://blogs.gartner.com*, November 2008.

[2] A. Celesti, F. Tusa, M. Villari, and A. Puliafito, "Three-phase cross-cloud federation model: The cloud sso authentication," *Advances in Future Internet, International Conference on*, vol. 0, pp. 94–101, 2010.

[3] SAML V2.0 Technical Overview, OASIS, http://www.oasis-open.org/committees/ download.php/11511/sstc-saml-tech-overview-2.0-draft-10.pdf.

[4] F. Tusa, M. Paone, M. Villari, and A. Puliafito., "CLEVER: A CLoud-Enabled Virtual EnviRonment," in *15th IEEE Symposium on Computers and CommunicationsS Computing and Communications, 2010. ISCC '10. Riccione*, June 2010.

[5] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Environments Workshop, 2008. GCE '08*, pp. 1–10, 2008.

[6] H. Takabi, J. B. Joshi, and G.-J. Ahn, "Security and privacy challenges in cloud computing environments," *IEEE Security and Privacy*, vol. 8, pp. 24–31, 2010.

[7] Forum of Federations: http://www.forumfed.org/en/index.php.

[8] Sun Microsystems, Take your business to a Higher Level - Sun cloud computing technology scales your infrastructure to take advantage of new business opportunities, guide, April 2009.

[9] W. Li and L. Ping, "Trust model to enhance security and interoperability of cloud environment," in *Cloud Computing*, pp. 69–79, November 2009.

[10] R. Ranjan and R. Buyya, "Decentralized overlay for federation of enterprise clouds," Handbook of Research on Scalable Computing Technologies, 2010. pages: 191-217.

[11] Goiri, J. Guitart, and J. Torres, "Characterizing cloud federation for enhancing providers' profit," *Cloud Computing, IEEE International Conference on*, vol. 0, pp. 123–130, 2010.

[12] N. Leavitt, "Is cloud computing really ready for prime time?," *Computer*, pp. 15–20, January 2009.

[13] Liberty Alleance Project, http://projectliberty.org.

[14] OpenID Authentication 2.0, OpenID Foundation, http://openid.net/specs/openid-attribute-exchange-2_0.html.

[15] The Shibboleth, a Software of the Internet2 Initiative: http://shibboleth.internet2.edu/.

[16] Microsoft Windows Cardspace, http://netfx3.com/content/WindowsCardspaceHome.aspx.

[17] B. Sotomayor, R. Montero, I. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *Internet Computing, IEEE*, vol. 13, pp. 14–22, September 2009.

[18] OpenSAML, "Open source libraries in Java and C++ providing core message, binding, and profile classes for implementing applications based on SAML 1.0, 1.1, and 2.0", http://saml.xml.org/internet2-opensaml.

[19] "Web services security: Soap message security 1.0, oasis, http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-soap-message-security-1.0.pdf."

[20] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the Use of Cloud Computing for Scientific Workflows," in *SWBES 2008, Indianapolis*, December 2008.

[21] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The Eucalyptus Open-Source Cloud-Computing System," in *Cluster Computing and the Grid, 2009. CCGRID '09. 9th IEEE/ACM International Symposium on*, pp. 124–131, May 2009.

[22] OpenQRM, "the next generation, open-source Data-center management platform", http://www.openqrm.com/.

[23] B. Sotomayor, R. Montero, I. Llorente, and I. Foster, "Resource Leasing and the Art of Suspending Virtual Machines," in *High Performance Computing and Communications, 2009. HPCC '09. 11th IEEE International Conference on*, pp. 59–68, June 2009.

[24] Ejabberd, The Erlang Jabber/XMPP Daemon Community: http://www.ejabberd.im/.

[25] Sedna, Native XML Database System: http://modis.ispras.ru/sedna/.

Greening Ad Hoc Networks through Detection and Isolation of Defecting Nodes

Maurizio D'Arienzo
Dipartimento di Studi Europei e Mediterranei
Seconda Università di Napoli
maudarie@unina.it

Francesco Oliviero, Simon Pietro Romano
Dipartimento di Informatica e Sistemistica
Università degli Studi di Napoli "Federico II"
folivier@unina.it, spromano@unina.it

***Abstract*—Current ad hoc networks rely on a silent mutual agreement among nodes to relay packets towards the destinations. The effort made by each single node to serve the others is usually repaid with the chance to successfully set up its own traffic sessions. However, limited power, together with security concerns, can push certain nodes to refrain from cooperating. Such nodes will thus act as parasites, while the others will unawarely keep on trusting them for what concerns the mutual service agreement. In this paper we show how energy consumption in Ad Hoc Networks can be dramatically reduced if we stimulate cooperation by providing mechanisms for the detection and isolation of selfish nodes. We present a novel routing protocol exploiting a behavior-tracking algorithm based on game theory and allowing traffic to be forwarded only towards cooperative nodes. Through extensive simulations, we show how we can significantly reduce power wastage at the same time maximizing the delivery rate. Under this perspective, cooperation can definitely be seen as an incentive for all nodes, since it allows to optimize one of the most crucial parameters impacting the performance of ad hoc networks.**

***Index Terms*—Energy Efficiency, Fairness, Game Theory, Cooperation, Ad Hoc Routing Protocols.**

I. Introduction

Ad hoc networks are composed of several nodes with wireless connection capability. Differently from wired networks, in an ad hoc environment each node is an end system and a router at the same time. A transmission between a sender and a receiver happens with the help of one or more intermediate nodes that are requested to relay packets according to routing protocols designed for this kind of networks. A blind trust agreement among nodes makes it possible the right message forwarding. However, wireless nodes have often limited power resources, and some of them are asked to relay packets more frequently than they do with respect to their own relay requests. Thus, a good percentage of power is wasted to serve other nodes. Besides, the open nature of the current ad hoc network protocols raises some security concerns. In fact, although in the recent past there have been proposals of protocol modifications to enhance security, at present the aggregation of new nodes is usually uncontrolled and open to potential malicious users. In such a situation, a generic node of the network has to decide whether to trust or not to trust the other nodes. This obviously calls for a capability of each single node to somehow interpret (or, even better, predict) the behavior of the other nodes, since they represent fundamental *allies* in the data transmission process [1].

The situations in which a decision of a part depends on the predicted behavior of another part have been elegantly studied in game theory. Game theory has been already applied [2] [3] [4] to ad hoc networks with interesting results. The basic assumption is that all the players follow a rational behavior and try to maximize their payoff. The simplest games see the involvement of only two players who have to decide whether to cooperate or defect with the others. The best solution may not maximize the payoff, but can reach an *equilibrium* as proposed by Nash. One of the versions of this game is known as *prisoner's dilemma* and has an equilibrium in case both users decide to defect. This is true for the game played only one time while in its iterated version the situation is more complex and even cooperation can be convenient. In case of ad an hoc network, the player is a node that needs to cooperate with the others to send its traffic. However some nodes can decide to defect for a number of unspecified reasons and, as a first need, the other nodes should be informed of their behavior in order to react in the most appropriate way.

In this paper, we show how cooperation can be perceived by nodes as an incentive, thanks to the fact that it helps save the overall amount of energy needed for data transmissions. Differently from recent works proposed in the literature [5] [6], which aim at making the routing process become *natively* aware of the energy-related parameters, we herein propose a different approach, by leveraging cooperation in order to improve the overall energy efficiency of an ad hoc network without modifying the existing routing protocol. Our work is indeed complementary to the above mentioned proposals, in that it can co-exist with any routing protocol, be it legacy or energy-aware. We try and exploit a different perspective on energy efficiency, which is much more related to the behavioral patterns of the nodes rather than to the specific mechanisms and protocols adopted in the network.

Delving into some of the details of how we deal with the behavioral aspects of the problem at hand, we present in the paper an algorithm to identify and isolate defecting nodes. The algorithm takes inspiration from the results of game theory and keeps a local trace of the behavior of the other nodes. At the beginning the behavior of all the other nodes is unknown, but as soon as the first flows of traffic are exchanged among them, each node becomes gradually aware of the past behavior of the others, which can be either cooperative or defecting. Once the defecting nodes are identified, different countermeasures can

be adopted. The current version of the algorithm makes the decision of not relaying packets coming from defecting nodes as long as they do not cooperate, but other, less disruptive policies can be considered and included. The algorithm is implemented in an existing ad hoc routing protocol and is validated in the ns-2 simulator. The current experimental results highlight the induced reduction of throughput of defecting nodes.

The paper is organized in six Sections. Section II deals with both background information and related work. Section III presents the algorithm we designed to infer behavioral information about the network nodes, whose implementation is described in Section IV. Results of the experimental simulations we carried out are presented in Section V, while Section VI provides concluding remarks and proposes some directions of future work.

II. Background and Related Work

In this section we try to shade light on the context of our contribution, by properly defining the scope of our research, as well as its application to the wide set of *green networking* proposals that have recently come to the fore in the international research community. We start by proposing a bird's eye view on the most recent works that have focused on energy-aware routing in ad hoc networks. Then, we move the focus to the most important aspect of our contribution, namely cooperation. Indeed, as we already pointed out, cooperation is a fundamental subject of our recent research and is herein studied under one of its most challenging facets, i.e. its use as an incentive for all the nodes of the network, thanks to the significant performance improvements that it entails in terms of energy savings associated with data transmissions.

A. Energy-aware routing in ad hoc networks

Routing table computation performed by a routing protocol based on energy measures can improve efficiency in ad hoc networks. In this regard, a great amount of energy-aware routing protocols for ad hoc networks have been proposed in the last years ([7], [8], [9], [10], [11]). They can be roughly classified based on their specific goals. They can in fact try to: (i) minimize the total power needed to transmit packets; (ii) maximize the lifetime of every single node; (iii) minimize the total power needed to transmit packets at the same time maximizing the lifetime of every single node. Some interesting energy-efficient route selection schemes, falling in one of the previous categories, are presented in [7] and briefly described in the following.

Minimum Total Transmission Power Routing (MTPR) is a routing protocol aimed at minimizing overall power consumption in ad hoc networks. Given a source s and a destination d, we denote with P_r the total transmission power for a generic route r from s to d. P_r is the sum of the power consumed for the transmission between each pair of adjacent nodes belonging to r. MTPR selects the route r^* such that $r^* = min_{r \in R} P_r$, where R is the set containing all possible routes from s to d. A simple shortest path algorithm can be used to find this route.

Minimum Battery Cost Routing (MBCR) associates each node n_i in the network with a weight $f_i(c_i(t)) = 1/c_i(t)$, where $c_i(t)$ is the battery capacity level of n_i at time t. Given a source s and a destination d, if we say E_r the sum of the nodes weights of a generic route r from s to d, MBCR selects the route r^* such that $r^* = min_{r \in R} E_r$, where R is the set containing all possible routes from s to d. Such a scheme will always choose routes with maximum total residual energy.

With *Min-Max Battery Cost Routing* (MMBCR), starting from the above definition of $f_i(c_i(t))$, for each route r from a source s to a destination d, a cost is defined as $C_r(t) = max_{i \in r} f_i(c_i(t))$. The chosen route r^* verifies the relation $C_{r*}(t) = min_{r \in R} C_r(t)$. MMBCR safeguards nodes with low energy level because it selects the route in which the node with minimum energy has more energy, compared to the nodes with minimum energies of the other routes.

Conditional Max-Min Battery Capacity Routing (CMMBCR) proposes an approach based on both MTPR and MMBCR. Let us consider the node of a generic route r from a source s to a destination d, with lowest energy. Let also $m_r(t)$ be its energy, and R the set of all the routes from s to d. If some paths with $m_r(t)$ over a specific threshold exist in R, one of these will be chosen using the MTPR scheme. Otherwise, the route r^* satisfying the relation $m_{r*}(t) = max_{r \in R} m_r(t)$ will be selected. This scheme suffers from an unfair increment of the forwarding traffic towards nodes with more energy [10].

Minimum Drain Rate (MDR [9]) proposes a mechanism that takes into account node energy dissipation rate, thus avoiding the above problem. MDR defines for each node n_i a weight $C_i = RBP_i/DR_i$, where RBP_i is the residual battery power and DR_i the drain rate of n_i. Intuitively, DR_i represents the consumed energy per second in a specified time interval. Now, let C_r be the minimum weight of a generic route r from a source s to a destination d. MDR selects the route r^* such that $C_{r*} = max_{r \in R} C_r$. In this way, residual energy level, as well as the energy consumption rate due to the incoming traffic to be forwarded, are jointly taken into account.

As we already stated, in this paper we do not embrace an approach aimed at modifying routing in order to let it become energy-aware. We rather propose to induce network nodes to cooperate, by demonstrating that a cooperative behavior turns out to have a significant effect on performance, in terms of reduction of the energy needed for data transmissions. In the next subsection we then focus on the behavioral aspects related to cooperation of the nodes of an ad hoc network.

B. Cooperation in ad hoc networks

Cooperation of nodes involved in an ad hoc network is usually induced because the efforts related to the offered services are compensated with the possibility to request a service from the other nodes. However current ad hoc network protocols do not provide users with guarantees about the correct behavior of other nodes that can eventually decide to act as parasites. Several works have identified the problem of stimulating cooperation and motivating nodes towards a common benefit. The main solutions rely on a virtual currency

or on a reputation system, and more recently on game theory.

Virtual currency systems [12] [13] give well behaving users a reward every time they regularly relay a packet. They can then reuse the reward for their transmissions as long as they have a credit. The first issue of such systems is related to the need of a centralized server to store all the transactions among the users. Besides, the system is not completely fair with all nodes. Nodes placed at the boundary of the area are usually less involved in relay operations and then excluded from rewards, even if they are ready to be involved. Also, the messages regarding the transactions need to be secured in order to avoid spoofing of malicious nodes.

Reputation systems repeatedly monitor and build a map of trustworthy nodes on the basis of their behavior [4] [5] [6] [14] [15]. These systems distinguish between the *reputation*, which rates how well a node behaved, and *trust*, which represents how honest a node is. Most of these systems consider the reputation value as a metric of trust . A node is refrained to relay a packet coming from untrusted nodes, which are then excluded from the network operations. Several issues are related to the use of these systems. First, each node needs to maintain a global view of the reputation values with considerable caching. Some proposals keep local information, others disseminate reputations to other nodes, with an increased overhead due to the exchange of such messages. Reputation values can be modified, forged or lost during operations, and they can differ from node to node, which can bring to inconsistency, i.e. node X considers node Y trustworthy while node Z considers the same node untrustworthy.

To overcome some of these issues, it has been proposed to model the nodes taking part to an ad hoc network with game theory. There are different models studied in game theory and some of them have been already applied to ad hoc networks. In next subsections we review the most important models of game theory and some applications to ad hoc networks.

C. Game theory Basics

Game theory is a branch of applied mathematics that witnessed a great success thanks to the application of its results to a wide selection of fields, including social sciences, biology, engineering and economics. Game theory covers different situations of conflicts regarding, in a first attempt, two agents (or *players*), and in the generalized version, a population of players. Each of these players expects to receive a reward, usually named *payoff*, at the end of the game. The basic assumption is that all the players are self interested and rational: given a utility function with the complete vector of payoffs associated with all possible combinations, a rational player is always able to place these values in order of preference even in case they are not numerically comparable (e.g. an amount of money and an air ticket). This not necessarily means that the best value will be selected, since the final reward of each player is strongly dependent on the decision of the other players. Each player is then pushed to plan a *strategy*, that is a set of actions aiming at a total *payoff* maximization, provided that he is aware that the other players will try to do the same.

Games are now classified in several categories according to various properties. If the players tend to be selfish in the achievement of the best payoff, the game is classified as non cooperative rather than cooperative. When there is a common knowledge of the utility function for all players, the game is with complete information, otherwise it is considered incomplete. There are several real world examples that fall in one of these categories. Here we are mainly interested in the difference between strategic and extensive games. In strategic (also known as *static*) games the players make their decision simultaneously, without any knowledge of the others' intention. Even if the game is repeated, the players are still unaware of others' plans and do not have the chance to react to a previous action. This last opportunity is instead available in case of extensive games. Such games are played more than once and the players can evaluate what the others did at least during the last tournament, so that they can potentially decide to modify their strategy for the next move. Also, the payoff is cumulated at the end of each round rather than accounted for only once.

One of the fundamental problems of game theory is known as *prisoner's dilemma*, which can be represented in the matrix format of Fig. 1: two suspects of a crime are arrested and jailed in different cells with no chance to communicate between each other. They are questioned by the police and receive the same deal: if one confesses (*defect*) and the other stays silent (*cooperate*), the first is released, the second is convicted and goes to prison with a sentence of 10 years, the worst; if both stay silent (*cooperate*), they go to prison for only 1 year; if both testify against the other (*defect*) they go to prison with a sentence of 5 years. The situation in which they both stay silent (*cooperate*) is the more convenient to both of them; however, it was demonstrated that a rational behavior is to confess (*defect*) and receive the sentence of 5 years, and this situation represents the only equilibrium, as first introduced by Nash [16] [17]. Hence, the prisoner's dilemma falls in the field of strategic non-cooperative games.

In its basic form the prisoner's dilemma is played only once and has been applied to many real life situations of conflict, even comprising thorny issues of state diplomacy. Another version of the prisoner's dilemma is played repeatedly rather than a single time and is known as iterated prisoner's dilemma (ITD), which turned out to be a cooperative game under certain circumstances [18][19]. The goal of both players still is the maximization of their payoff, as the cumulated payoff earned at each stage. If the number of rounds is finite and known in advance, the strategy of always defecting is still the only situation of equilibrium and the game is still non-cooperative. However, in case the number of repetitions is infinite, it was demonstrated that the choice to always defect is not the only equilibrium as even the choice of cooperating may be an equilibrium. In this case, one of the strategies that let players maximize their payoff is the so-called *Tit for Tat* game, in which each player repeats the past behavior of the other player: a player is keen to cooperate if the other node behaved correctly the last time, otherwise it defects. If we

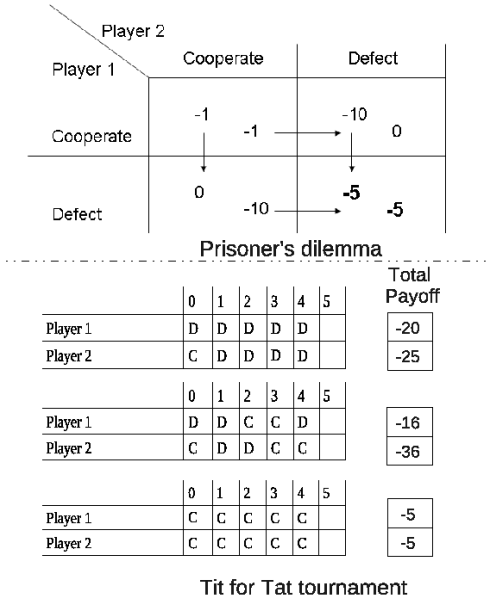

Fig. 1: The tit-for-tat strategy in action

consider the first five tournaments of a two players game, a player who defects (D) against a cooperative (C) player adopting the tit for tat strategy would play (D,D,D,D,D) and earn $(0, -5, -5, -5, -5) = -20$. If the first player decides to cooperate two times out of five (D,D,C,C,D), he would earn $(0, -5, -10, -1, 0) = -16$. In case he always cooperates, his payoff would be $(-1, -1, -1, -1, -1) = -5$, which is the best he can achieve. So, continued cooperation for the iterated prisoner's dilemma also yields the best payoff. Despite this benefit, the main result of the tit for tat strategy is that it stimulates the cooperation. We base our algorithm to mitigate the node selfishness on the results of this version of the game.

D. Game Theory applied to Ad Hoc Networks

One of the first proofs of the improvements produced by cooperation in ad hoc networks is presented in [2]. The authors provide a mathematical framework for studying the effects of cooperation in ad hoc networks. They first introduce a normalized acceptance rate (NAR) as the ratio between the successful relays provided to the others and the relay requests made by the node. Then they propose two models, namely GTFT (Generous Tit for Tat) and m-GTFT for the case of multiple players, to give the (rational) nodes the chance to make a decision concerning the possibility to cooperate or defect with other nodes, and they analytically demonstrate that these models represent a Nash equilibrium. In such a situation, a node does not improve its NAR to the detriment of the others. Also, at the opposite of reputation schemes, each node can maintain per session rather than per packet information, thus leading to a scalable solution.

In [20] the authors prove the selfishness property of the nodes in a MANET by using the Nash equilibrium theorem [16]. They define a generic model for node behavior that takes into account also energy consumption due to the transmission process. By adopting a punishment based technique they prove that it is possible to escape from the theoretically unique equilibrium point of non-cooperation and to enforce a cooperation strategy under specific conditions.

In [21] the authors also fucus on forwarding mechanisms. They provide a model for node behavior based on game theory in order to determine under which conditions cooperation with no incentives exists. They prove that network topology and communication patterns might significantly help enforce cooperation among nodes.

Game theory has been also used to improve routing algorithms in wireless networks. An actual implementation of a game theory model in the AODV routing protocol with two distinct approaches has been proposed in [3]. The first plays a deterministic tit for tat game and the second a randomized version of the same game deployed with a genetic algorithm. In both cases, they achieve better performance in terms of experienced delay and packet delivery ratio in case of cooperation of nodes. The models are tested in a simulated environment and rely on static distribution of nodes' behavior profiles while not supporting a mechanism for a dynamic adaptation to changed situations.

III. Algorithm description

In an ad hoc network, the number of nodes and links can change during time, so we consider the number of nodes $N(t)$ as a function of time t. We also define a dynamic array $C(t)$ of $N(t)$ elements for each node of the network. The generic element $c_i(t)$ of $C(t)$ assumes the values (UNKNOWN, COOPERATE, DEFECT) meaning that the behavior of node i at time t is respectively unknown, cooperative or non cooperative. At time $t = 0$ all the values are set to UNKNOWN, since at the beginning each node is not aware of the behavior of the other nodes.

Suppose the generic node s of the network needs to send some traffic to the destination d. The first task is to discover an available path, if it exists, to reach the destination. To this purpose, we consider a source based routing protocol capable of discovering a list $A(t)_{(s,d)i}$ $\forall i : 0 < i < P$ of P multiple paths. All the nodes in the list $A(t)_{(s,d)i}$ are considered under observation and marked as probably defecting in the array $C(t)$ unless a positive feedback is received before a timeout expires. The sender s starts sending his traffic along all the discovered paths. If the destination node generates D acknowledgement messages containing the list of all the nodes $L_{(s,d)i}$ $0 < i < D$ traversed, as it happens in some source based routing protocols, the sender s is informed about the behavior of intermediate nodes. For each acknowledgement message received, the sender s can make a final update of the array $C(t)$ by setting the matching elements $c_i(t)$ contained in the list $L_{(s,d)i}$ as cooperative. Notice that the last update overwrites the previous stored values and represents the most recent information concerning the behavior of a node. An example of the evolution of the described algorithm is presented in Fig. 2.

Given this algorithm, each node is aware of the behavior of other nodes and can react in the most appropriate way. For

$A(t)_{|A,F}$={A,C,E,F}
$A(t)_{|A,F}$={A,B,C,E,F}
$A(t)_{|A,F}$={A,C,E,F}
$A(t)_{|A,F}$={A,B,C,D,F}

Node	A	B	C	D	E	F
expected	-	D	D	D	D	-
final	-	U	U	U	U	-

U - Unknown
C - Cooperative
D - Defecting

Timeout expired

Node	A	B	C	D	E	F
expected	-	D	D	D	D	-
final	-	C	C	C	D	-

Fig. 2: Algorithm description

example, a node can refuse to relay packets of defecting nodes, or operate a selective operation like queuing their packets and serving them only if idle and not busy with the service requested by cooperative nodes. In this first proposal, we rely on the harsh policy of packet discarding, and this brings to the isolation of defecting nodes. However, a defecting node can even gain trust of other nodes if it starts to cooperate. The array $C(t)$ is not static over time and its values are continuously updated. In fact, due to the dynamic situation of ad hoc networks, the search of available paths is frequently repeated, and the list $A_{(s,d)}$ consequently updated. Hence, if a defecting node decides to cooperate, its identification address will be included in one of the acknowledgement messages $L_{(s,d)i}$ sent to the sender s and its aim to cooperate will be stored in the array $C(t)$.

The situation described here for the pair (s, d) is replicated for all the possible pairs of nodes that try to interact, but each node stores only one array $C(t)$ that is updated upon reception of any acknowledgement message, wherever it comes from. Furthermore, not all the packets relayed are checked in order to verify the nodes' behaviors, but only a sample of them, thus keeping the total overhead under control.

IV. An Ad Hoc Network Routing Protocol for Discovery of Defecting Nodes

The algorithm introduced in the previous section has been implemented in an existing source based routing protocol for ad hoc networks. We first modified this protocol to support the search of multiple paths, and then included the new algorithm for the identification of non cooperative nodes. In the next subsections we present the existing protocol, with respect to both its basic and newly added features.

A. Multipath source based routing in ad hoc networks

The dynamic configuration of an ad hoc network topology makes the routing protocols used in wired networks unsuitable for this kind of networks. Hence, several new protocols have been designed and made available to manage this collection of wireless nodes. To the purpose of identifying defecting nodes, an acknowledgment, or missed acknowledgement technique is needed. Among the many ad hoc routing protocols, AH-CPN (Ad Hoc Cognitive Packet Network) [22] is designed to support QoS and make an intense use of acknowledgement messages independently from the transport protocol in use. AH-CPN is the wireless version of CPN (Cognitive Packet Network) [23], a proposal for a self aware network architecture

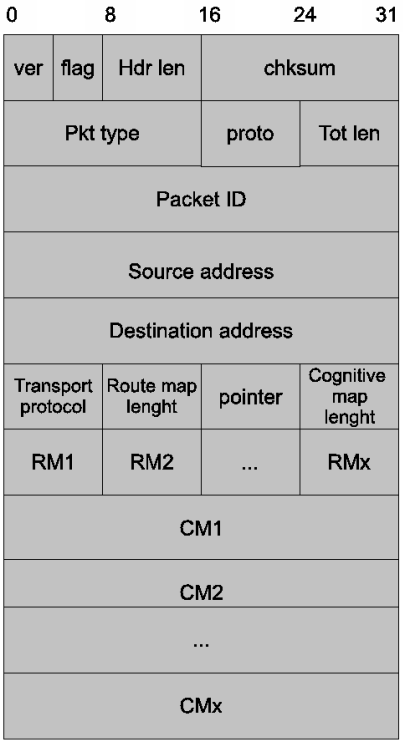

Fig. 3: The CPN header

with native support for QoS. Both in AH-CPN and CPN, the presence of a neural network engine enables to undertake dynamic and fast routing decisions as soon as a condition, like for example a congested link or a different user's requirement, has changed. An always active traffic of smart packets discovers new paths according to specific QoS goals, e.g. discovering paths that minimize the delay or maximize the throughput. This information is made available to the interested nodes that can send traffic along the defined path on a source based routing basis. The smart traffic keeps on looking for the specific goals, and in case a better path is found, the sender is informed and can update its routing path.

There are four different kinds of packets in AH-CPN, all sharing the same header (depicted in Fig. 3): Smart Packets (SP), Smart Acknowledgements (SA), Dumb Packets (DP), and Dumb Acknowledgements (DA).

Smart packets are those described at the beginning of this section. They are lightweight packets containing a QoS goal sent by a sender to a destination. These packets are routed with the Random Neural Network (RNN) [24] algorithm that runs on each node and which selects the next hop by taking into account the past behavior of the link. Every time a SP traverses a node the route map (RM) field is updated with the node's address. Once at the destination, a SA is generated and sent backwards along the RM received in the SP. Finally, the actual data can be sent across the network in a DP, which is prepared with the whole path copied in the RM field. Internal nodes relay DPs to the next hop excerpted from the RM field, and they add timestamp information useful to evaluate the round trip time (RTT) between each pair of nodes along the path. These RTT data are stored in special *mailboxes*

present in each node and provide the RNN algorithm with precious information concerning the past behavior of a link. Once the DP reaches its destination, a DA is sent along the reverse path. Notice that differently from IP networks, in CPN the acknowledgements are generated upon reception of each single packet, whatever the transport protocol is. This feature is helpful in the deployment of our algorithm to identify defecting nodes, as we will soon explain.

The basic CPN version looks for one available path, the best in terms of the requested QoS goal. We modified this protocol to search for multiple paths. To this purpose, SPs are initially sent via flooding to collect a certain number of available paths (up to a well defined threshold, which can be properly configured at setup time). To prevent loops, SPs are marked with an identification number ID, and those with the same ID touching a node for the second time are discarded. SPs reaching the same destinations with different contents for what concerns the routing map RM are considered valid, and SAs are sent backward to inform the sender. The sender collects the different SAs and updates its routing table. DPs are sent on a round robin basis. Once the available paths are discovered, the transmission of SPs is not terminated; it is rather repeated periodically for path maintenance, to check if the topology has changed, and in our case also to verify if there is a different configuration concerning the behavior of nodes.

B. Identification and isolation of defecting nodes

We provide the multipath source based routing protocol with the support for identification and isolation of defecting nodes. The array $C(t)$ is computed and stored at each node. Its dimension can change according to the number of nodes active in the ad hoc area. When node a needs to send traffic to node b, SPs are immediately sent in flooding. We make the assumption that non cooperative nodes try to cheat by forwarding inexpensive SPs, that do not carry any payload, while they do not relay DPs containing the real data. In case the non cooperative nodes decide to block the SPs forwarding, they are immediately discovered as non cooperative and have no chance to cheat. In this scenario, every time a SP traverses a node, its cognitive map is extended with the label of the visited node. Once at the destination, the complete cognitive map is copied into the DA and sent back to the sender along the reverse path. Obviously, this is repeated for all the discovered paths, so at the end of this process node a has a complete knowledge of all the available paths, including those comprising cheating nodes, and these are all stored in $A(t)_{(a,b)}$. At the time of the first transmission, the real data are packed in multiple DPs and sent along all the available paths on a round robin basis, but the interested cheating nodes will not relay them. Since in CPN a destination b must send an acknowledgement message DA whatever the transport protocol is, node a will receive only the DAs containing the successful paths, i.e. those without cheating nodes. This information, as described before, helps finalize the array $C(t)$ with the list of cooperative and defecting nodes, and the traffic is sent only along the path or the paths composed of cooperative nodes rather than towards all the available paths. When one of the cheating nodes requests the relaying of a message to node a, it is aware of his past behavior and can decide to drop all its packets, while it can regularly relay packets coming from cooperative users.

The situation concerning the cooperation and the selection of paths is not static and can change during time, so isolated nodes are not banned forever from the network. Although the traffic from a node is delivered only along paths composed of cooperative nodes, sending nodes keep on checking periodically the paths containing the defecting nodes. Should a defecting node decide to change its behavior and begin to cooperate, the routing protocol soon detects this change and admits again the node to the transmission of flows. This way, a node reacts following a *Tit for Tat* strategy.

V. Experimental results

In a wireless scenario, normal operation can often lead to a high level of iniquity. The introduction of a system able to detect defecting nodes can instead increase the fairness of a wireless ad hoc network in terms of both delivery ratio and energy consumption. To show these two aspects, we repeatedly ran two type of experiments in the ns-2 simulator.

In the first series of experiments we designed a scenario associated with several working conditions on a simple wireless testbed composed of 8 nodes (see Fig.4), labeled from 0 to 7. In such network we set up the following conditions: (i) node 3 defects all the time; (ii) the behavior of node 4 dynamically changes over time; (iii) all the other nodes are cooperative. The duration of the experiments is set to 10 minutes. The defection of a node means that the relay of traffic to serve other nodes is totally stopped, so the percentage of node 3's cooperation is always 0% (of the total time). As far as node 4 is concerned, five situations are considered, most of them offering the other nodes the chance to reply with a *tit for tat* strategy:

1) Node 4 never cooperates. Requests of relay are never forwarded, so the percentage of cooperation is 0%;
2) Node 4 follows a switching behavior: assuming that the time is divided in 4 equal slots of 150 seconds each, node 4 cooperates during the first 75 seconds of the second and fourth slot interval, then it defects all the time; the total percentage of cooperation is hence 25%;
3) Node 4 still switches its behavior: it defects during two slots and cooperates in the other two; in this case the total percentage of cooperation is thus 50%;
4) Node 4 switches its behavior in a way that is opposite to the one described in the second item of this list: node 4 defects during the first 75 seconds of the first and third slot interval, then it cooperates all the time; the total percentage of cooperation is hence 75%;
5) Node 4 always cooperates; all relay requests are served, for a final percentage of cooperation of 100%.

Two equal sessions of constant bit rate traffic are activated between node 4 and node 0 and node 1 and node 7, respectively at time 1.0 and at time 2.0. In the ideal situation of all

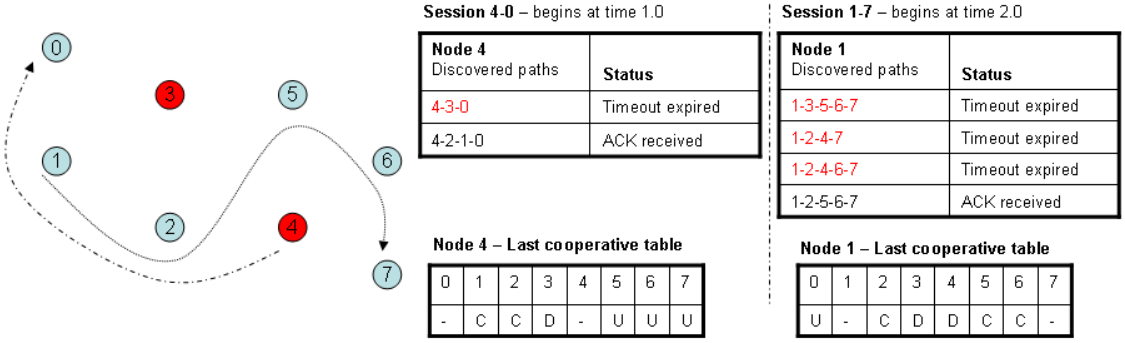

Fig. 4: The simulated testbed

cooperating nodes, the shortest paths would be $(4, 3, 0)$ and $(1, 2, 4, 7)$. However, node 3 is always defecting, so the path $(4, 3, 0)$ turns out to be unavailable and the traffic coming from node 4 is forced along the other available path $(4, 2, 1, 0)$. As long as node 1 does not generate traffic, it does not have the chance to track the behavior of node 4, so the relay requests coming from node 4 are regularly served. At time 2.0 node 1 begins the discovery of paths to reach node 7. Besides the other choices, the best path $(1, 2, 4, 7)$ is soon discovered and selected to immediately generate traffic. If node 4 follows a switching behavior, then node 1 has the chance to react in compliance with the *tit for tat* strategy. Notice that in case node 4 is in a defecting state, node 1 can still send traffic to the destination along the path $(1, 2, 5, 6, 7)$.

In Fig. 5 we report the delivery ratio of node i as the ratio $G_i(t) = r_i(t)/s_i(t)$ at the end of the experiment ($t = 10min$) between the number of bytes correctly received at destination $r_i(t)$ and the total number of bytes sent $s_i(t)$. The x axis represents the percentage of node 4's cooperation, the y axis is the final delivery ratio $dr_i(t)$. Initially (left part of the x axis in Fig. 5), node 4 is fully defecting; the same applies to node 3. Traffic from node 4 towards node 0 is regularly sent between time 1.0 and time 2.0 because node 1 did not generate any request and did not yet test the behavior of the other nodes. At time 2.0, however, node 1 tries to send traffic to node 7 and hence has the chance to verify the behavior of the other nodes. Among the other discovered paths, it realizes that paths comprising nodes 4 and 3 are not working, so as soon as the timeout expires it marks nodes 3 and 4 as defecting and immediately stops relaying traffic coming from node 4. The final delivery ratio dr_1 of node 1 is closer to the ideal value because the alternative path $(1, 2, 5, 6, 7)$ is soon discovered and used for the entire duration of the experiment. The delivery ratio dr_4 is instead severely reduced.

As node 4's percentage of cooperation increases up to 100%, the delivery ratio dr_4 also increases until it reaches a value close to dr_1 when there is full cooperation. Although node 3's defection makes the path $(4, 3, 0)$ unavailable, the routing protocol discovers the alternative path $(4, 2, 1, 0)$ composed of cooperative nodes, while the shortest $(1, 2, 4, 7)$ is regularly available in this case. This is the only situation in which node 4 maximizes its goodput. In the intermediate cases the trend is linear and clearly demonstrates the correct

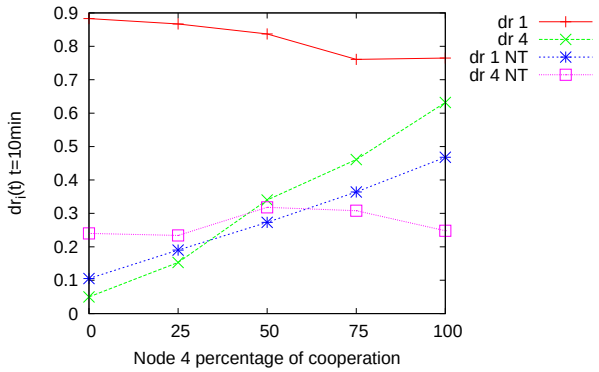

Fig. 5: Delivery ratio of Node 1 and Node 4

implementation of the *tit for tat* reaction mechanism, as node 1 cooperates only when node 4 does the same. Delivery ratio dr_1 remains more or less unaltered independently of node 4's behavior, thanks to the fact that node 1 has a chance to discover alternative cooperative paths.

We compared these results with the situation in which the nodes are unable to detect the defecting behavior. We mark these sessions with *NT* in the same Fig.5. The situation is now opposite to the previously analyzed case because delivery rate dr_4 outperforms dr_1 in the case of node 4's full defection. Node 1 is now unaware of node 4's defection; hence, while its traffic is not relayed, it regularly relays the incoming packets having node 4 as source. Anyway, both delivery ratios (dr_1 and dr_4) are lower than in the previous case. This time the lack of tracing of nodes defection affects even node 4's performance, because such node tries to forward traffic not only along the path $(4, 2, 1, 0)$ but also along the uncooperative path $(4, 3, 0)$, which explains the halved final delivery ratio.

We then evaluated the effective energy spent by node 1 and node 4 to successfully deliver their packets to the respective destinations. This energy is calculated as $E_{eff_i} = Ec_i * dr_i$, being Ec_i the energy consumed by node i and dr_i the delivery ratio computed as described above. Fig.6 illustrates the average effective energy consumed by node 1 and 4 in both situations of detection (active and inactive), as well as in all the aforementioned conditions of cooperation. Notice how the trace corresponding to the detection enabled is always lower compared to the case when detection is disabled. Besides, both traces decrement as the cooperation increases, and reach their

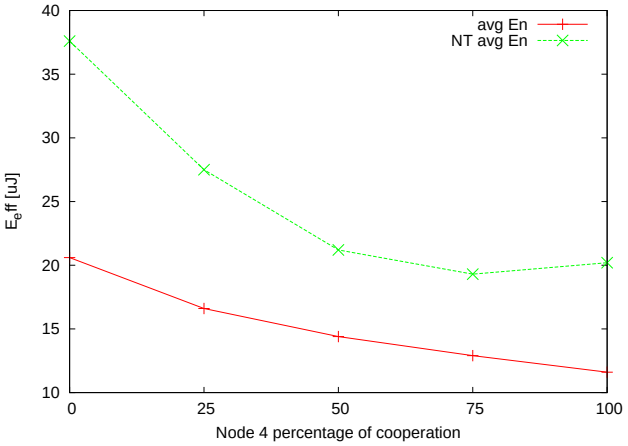

Fig. 6: Average effective energy

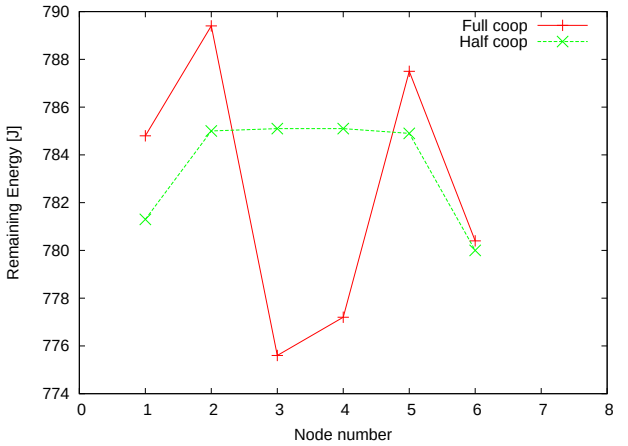

Fig. 7: Effect of defection on energy distribution

lowest values when cooperation is high.

In the second series of experiments we describe how an appropriate combination of defection and cooperation can yield a better distribution of the energy. We made use of the same 8 node testbed with two sessions of equal constant bit rate traffic between node 0 and 7 and vice versa, with a total duration of still $t = 10$ minutes. We evaluated the remaining energy at the end of the experiment for all the *inner* nodes labeled from 1 to 6 in the two different situations of (i) full nodes' cooperation and (ii) partial cooperation of node 3 and 4 for 50% of time. The final levels of energy are reported in Fig.7 showing a better balance when node 3 and 4 are semi-cooperating while keeping the same average consumption of all nodes, which is of 782.4 J with a variance of 31.6 in the former case, and of 783.5 J with a variance of 5.2 in the latter one.

VI. Conclusions

In this paper we showed how cooperation positively affects the performance of an ad hoc network, by helping reduce the overall energy consumption associated with data transmissions. We demonstrated through simulations that cooperation actually acts as an incentive for nodes, since it allows for a lower average energy expenditure with respect to the packets successfully delivered. We also studied the positive impact of cooperation on nodes' delivery ratio, which is considered a key performance indicator for any networked environment. Finally, we gave a first proof that if a subset of core nodes deliberately opts for defecting, the energy consumption can be better distributed among the nodes.

We do believe that the behavior-based approach that we presented in this work can be effectively exploited in a number of alternative scenarios, since it actually works along a dimension, which turns out to be complementary to other potential approaches, like, for example, ad-hoc designed energy-efficient routing paradigms.

This work is clearly a first step towards the study of cooperation effects in ad hoc networks. Among the numerous improvements that we identified and that represent directions of our future work, we firstly mention a more detailed analysis of the dependence of the performance improvements deriving from cooperation on the specific network topology taken into account. Apart from this, we also intend to study how the specific location of a node in the ad hoc network topology affects its performance and consequently its willingness to cooperate. This requires that a thorough analysis of the tradeoff between relaying other nodes' packets and sending one's own data is conducted.

References

[1] M. D'Arienzo, F. Oliviero, and S.P. Romano. Smoothing selfishness by isolating non-cooperative nodes in ad hoc wireless networks. In *Advances in Future Internet*, AFIN '10, pages 11–16, Washington, DC, USA, 2010. IEEE Computer Society.

[2] V. Srinivasan, P. Nuggehalli, C. F. Chiasserini, and R. R. Rao. Cooperation in wireless ad hoc networks. In *INFOCOM 2003.*, vol. 2, pp. 808–817, April 2003.

[3] K. Komathy and P. Narayanasamy. Trust-based evolutionary game model assisting aodv routing against selfishness. *J. Netw. Comput. Appl.*, 31(4), pp. 446–471, 2008.

[4] S. Marti, T. J. Giuli, K. Lai, and M. Baker. Mitigating routing misbehavior in mobile ad hoc networks. In *ACM MobiCom '00*, pp. 255–265, New York, USA, 2000.

[5] F. Olivero and S. P. Romano. A reputation-based metric for secure routing in wireless mesh networks. In *IEEE GLOBECOM 2008.*, pp. 1–5, December 2008.

[6] K. Mandalas, D. Flitzanis, G.F. Marias, and P. Georgiadis. A survey of several cooperation enforcement schemes for MANETs In *IEEE Int. Symp. on DOI* pp. 466 - 471, 2005

[7] C. K. Toh, "Maximum Battery Life Routing to Support Ubiquitous Mobile Computing in Wireless Ad Hoc Networks", IEEE Communications Magazine, 2001.

[8] P. Sondi and D. Gantsou, "Voice Communication over Mobile Ad Hoc Networks: Evaluation of a QoS Extension of OLSR using OPNET", Proceedings of AINTEC'09, Bangkok.

[9] D. Kim, J. J. Garia Luna Aceves, K. Obraczka,J. Cano and P. Manzoni, "Power-aware routing based on the energy drain rate for mobile ad hoc networks", in Proceedings of IEEE $11th$ International Conference on Computer Communications and Networks, Pages $562 - 569$, 2002.

[10] Floriano De Rango, Marco Fortino, "Energy efficient OLSR performance evaluation under energy aware metrics", in Symposium on Performance Evaluation of Computer and Telecommunication Systems, Pages $193 - 198$, 2009.

[11] S. Mahfoudh, P. Minet, "Survey of Energy Efficient Strategies in Wireless Ad Hoc and Sensor Networks", IEEE International Conference on Networking, Cancun, Mexico, Pages $1 - 7$ (2008).

[12] S. Zhong, Y. Yang, and J. Chen. Sprite: A simple, cheat-proof, credit-based system for mobile ad hoc networks. In *INFOCOM 2003.*, vol 3, pp. 1987 - 1997, 2003.

[13] L. Buttyán and J. P. Hubaux. Stimulating cooperation in self-organizing mobile ad hoc networks. *Mob. Netw. Appl.*, 8(5), pp. 579–592, October 2003.

[14] Y. Po-Wah and C.J. Mitchell. Reputation methods for routing security for mobile ad hoc networks. *Mobile Future and Symp. on Trends in Communications*, pp. 130-137, 2003.

[15] S. Buchegger and J.-Y. Le Boudec. Performance analysis of the confidant protocol. In *ACM MobiHoc '02*, pp. 226–236, New York, USA, 2002.

[16] J. Nash. Non-Cooperative Games. *The Annals of Mathematics.*, 54(2), pp. 286–295, 1951.

[17] J. F. Nash. Equilibrium Points in n-Person Games. In *Proceedings of the National Academy of Sciences of the United States fo America.*, 36(1), pp. 48–49, 1950.

[18] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1988.

[19] R. Axelrod and D. Dion. The further evolution of cooperation. *Science*, 242(4884), pp. 1385–1390, December 1988.

[20] A. Urpi, M. Bonuccelli, and S. Giordano. Modelling Cooperation in Mobile Ad Hoc Networks: A Formal Description of Selfishness. In *Proceedigns of Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks.*, 2003.

[21] M. Félegyházi, J. P. Hubaux, and Levente Buttyán. Nash Equilibria of Packet Forwarding Strategies in Wireless Ad Hoc Networks. *IEEE Transaction on Mobile Computing.*, 5(5), pp. 463–476, 2006.

[22] E. Gelenbe and R. Lent Power-aware ad hoc cognitive packet networks. *Ad Hoc Networks*, 2(3), pp. 205–216, July 2004.

[23] E. Gelenbe, R. Lent, and Z. Xu. Design and performance of cognitive packet networks. *Perform. Eval.*, 46(2-3), pp. 155–176, 2001.

[24] E. Gelenbe. Learning in the recurrent random neural network. *Neural Comput.*, 5(1), pp. 154–164, 1993.

User's Macro and Micro-mobility Study using WLANs in a University Campus

Miguel Garcia, Sandra Sendra, Carlos Turro, Jaime Lloret
Universitat Politècnica de València
Camino Vera s/n, 46022, Valencia, Spain
migarpi@posgrado.upv.es, sansenco@posgrado.upv.es, turro@cc.upv.es, jlloret@dcom.upv.es

***Abstract* — Wireless Local Area Networks are important and necessary in university campuses and large enterprise areas. Such networks allow us to have data connection anywhere without wires and many other benefits such as to obtain the location of the users. Moreover, they can be used to track the users. Tracking the user's mobility we can know which places are most visited, if people have to go to places that are far from their office, detect the best location for emergency points, etc. Moreover, we can study the mobility pattern of several users. According to this pattern, the network can use reconfiguration systems to reallocate resources and improve its connectivity. This paper shows us the case study of a university campus of two square kilometers and how we have taken advantage of the information gathered from the wireless network. Two studies (macro and micro mobility) have been done in order to make this survey. This approach can be used by the enterprises to optimize the sites to place their resources (network printers, servers, meeting rooms, etc.). Furthermore, the network's administrators can use these parameters to improve the network's behavior by providing a better connection, better roaming, etc.**

Keywords - People Mobility; macro-mobility; micro-mobility; people tracking; WLANs

I. INTRODUCTION

Nowadays, wireless local area networks are widely implemented. The public organisms such as universities, governments, etc. are well known examples. These wireless networks are usually based on the IEEE 802.11 b/g standard [1]. The standard presents many advantages. We can emphasize some of them:

- The use of a free band in the 2.4 GHz.
- Speeds up to 54 Mbps.
- The user's comfort is bigger than in wired networks.
- Wireless networks allow the access of multiple computers with a smaller infrastructure cost.
- The compatibility among different devices is very high, because of the organization Wi-Fi [2].

The IEEE 802.11 b/g networks present the intrinsic problems of any wireless technology. Some of them are:

- The wireless connections bandwidth is smaller than in wired connections.
- Wireless connections are more prone to be attacked because they can be accessed from anywhere, although there are several methods to encrypt the communication.
- The roaming can stop any communication between the devices of the network.
- WLAN is not compatible with other wireless technologies like Bluetooth [3], UMTS [4], etc.

The wireless LAN network is mainly used to transmit data, but there are many other applications. One of the most well known applications is the indoor positioning system. The localization is made using the access point's received signal strength and the use of different mathematical methods is possible [5]. There are many other applications such as providing connectivity in meetings, wireless VoIP, wireless IPTV and so on. Most of them are a service guided to end-user.

In this paper, we use the data obtained from the WLAN in order to study the mobility of the users. The roaming information can be used to know the behavior of the people in a place, to relocate the bandwidth and how they move from one building to another. This information will let us know the movement of the users, what buildings are most visited, etc.

We will study the mobility from several points of view: macro-mobility, that relates to the behavior of users between buildings and big areas, micro-mobility that relates to the movement of users inside a building and from that data we will extract the attractor points of the area, that are the focal points that have to be considered for the movement of people.

This paper is based on a previous work [6] presented by the same authors. In this paper we have added a micro mobility study and improved several main parts of that conference paper.

The paper is organized as follows. In Section 2, we discuss the related work about mobility tracked by wireless networks. Section 3 explains our university wireless network. Section 4 describes the steps performed to gather mobility data. Macro-mobility and people tracking measurements can be observed in Section 5. Section 6 is devoted to micro-mobility and what is the mobility profile of mobile users in our wireless network. Moreover, we show the behavior of a regular user. Finally, Section 7 presents our conclusions and future work.

II. RELATED WORK

We have found several works related to people mobility and tracking.

In [7], Z. Chen et al. presented a system that works like an indoor GPS. It uses RFID and provides directional instructions for users while tracking things. It is called DynaTrack and consists of three key parts which are the RFID tags and readers, database servers that hold information about things' location and the DynaTrack client

side interface. This system uses a dynamic or static system tags, depending on if it is an object or a person. Also, they tell us that if an object, initially labeled as static, begins to move, the system is able to change its address dynamically.

Another example of a tracking system is presented by J.G. Markoulidakis et al. in [8]. In this paper, we can see a new system based on Third Generation Mobile Telecommunication Systems (TGMTS) and the three basic types of mobility models that are appropriate for the full range of the TGMTS design issues. They propose a methodological modeling approach called Integrated Mobility Modeling Tool (IMMT). IMMT tries to improve some aspects of other systems like the validation of the theoretical input assumptions and analytical models or the effect of the mobility model accuracy.

The authors in [9] show the possibilities of utilizing RFID, Wi-Fi and Bluetooth wireless technologies to determinate their limitations in personnel/equipment tracking and mapping mine works of Pollyanna (underground mine in Oklahoma). Other wireless technologies, as the conventional satellite GPS technology, are not feasible there. They evaluate the advantages in the real-time location services (RTLS) technology to determine their applicability and limitations to underground mining at the Pollyanna.

In [10], B. Issac et al. presented a predictive mobility management system which could make mobility on an IEEE 802.11 network more proactive with minimum loss and delay, when compared to existing schemes. Their proposal is focused on WLAN installations within a restricted campus and to predict the mobility path of a mobile node and use that information to lessen the handoff delay.

A wireless indoor tracking system, based purely in software because no additional hardware is required, is described in [11]. It can be used to track and locate both moving and static WLAN-enabled devices inside a building. The system uses complex mathematic algorithms and determines the locations of the mobile devices according to the received signal strength from visible access points. The author categorizes the WLAN-based location determination algorithms, into two groups: deterministic and probabilistic algorithms. Finally, the paper is concluded making some reflections about the number of APs and their correct localization in order to obtain reliable results.

There are other works that show a study and even try to imitate the human behavior movements through simulations. One of them is the paper presented by T. Liu et al. in [12]. They present a model in order to mimic human movement behavior. It is built as a two-level hierarchy in which the top level is the global mobility model (GMM): a deterministic model that is used to create intercell movements and the bottom level is the local mobility model (LMM): a stochastic model with dynamically changing state variables to model intracell movement.

Another example of a human behavior simulation is given in the paper presented by C. Bettstetter [13]. It shows a model that can be used in simulations of mobile and wireless networks. He uses a combination of principles for direction and speed control to provide the movement of the users. It shows the calculation process to simulate changes of speed, stop-and-go behaviors or address control, among others.

In [14], the authors present a general methodology for obtaining the mobility information from wireless network traces, and for classifying mobile users and APs. In order to develop this methodology they use Fourier transform and Bayes' theory. The authors find some relations between several parameters, but in their study they say that the data is too variable because it depends on seasonal cycles, trend, regression term and irregular effects.

In [15], J. Gosh et al. analyze a yearlong wireless network users' mobility by tracing the data collected. They propose an efficient method to determine the main mobility profiles of a user using a mixture of Bernoulli's distribution. This method allows the authors to predict from 10% to 30% of the user's mobility.

In summary, the works presented previously carried out studies only related with users' tracking, except for [12] and [13] that develop simulation models of people's mobility. In all of these previous papers, the buildings are considered as passive objects (they do not give information). In contrast, in our work, each building is considered a group of several APs, and we use these groups to see the user's mobility around our university. With these data we could relocate some services and the displacement of the users would be more efficient.

III. WIRELESS NETWORK DESCRIPTION

The Universitat Politècnica de València (UPV) is distributed on three Campuses. One of them is located in Valencia and contains about 80% of the students and staff of the University. It has a dimension of about three kilometers long and one kilometer wide. There are two smaller campuses in the nearby cities Gandia and Alcoy. There are around 4,000 researchers and educational personnel, around 1,500 staff and around 36,000 students among the three campuses. The distribution of students in each faculty is shown in Table I.

On these Campuses, a wireless IEEE 802.11 b/g network is deployed. It comprises more than 575 access points to get full coverage, including not only the buildings and offices, but the surrounding gardens and open space between these buildings. So, any person in the UPV can roam seamlessly between any locations. The distribution of these access points is: 33 APs are in the Campus of Gandia, 42 APs are in the Campus of Alcoy and 500 APs are in the main Campus (Campus de Vera). The APs are installed to allow the users a continuous coverage as they roam throughout a facility. The coverage of each access point varies between 30 m at 54 Mbps and 137 m at 1 Mbps for indoor environments.

The access points are from Cisco Systems Inc. (models 1130, 1140 and 1300) and they are configured with three simultaneous SSIDs, one with VPN authentication, another one with 802.1X authentication and the last one interacted into the EDUROAM (European roaming project) for visitors. Any member of the University, and from others via EDUROAM, has free access to that wireless network.

TABLE I. PEOPLE REGISTERED IN EACH BUILDING

Building	People registered
E. Politecnica Superior de Alcoy	2298
E.T.S de Ingenieria de Informatica	3240
E.T.S. de Arquitectura	3858
E.T.S. de Gestión de la Edificación	2920
E.T.S. de Ingenieria del Diseño	4794
E.T.S. del Medio Rural y Enología	1074
E.T.S.I. de Agronomos	1841
E.T.S.I. de Caminos, Canales y Puertos	3145
E.T.S.I. de Telecomunicación	1409
E.T.S.I. de Geodesica, Cartografia y Topología	1027
E.T.S.I. de Industriales	3479
E. Politecnica Superior de Gandia	2320
F. de Administración y Dirección de Empresas	2271
F. de Bellas Artes	2334
Total	**36010**

IV. PEOPLE TRACKING MEASUREMENTS

It is quite complicated to predict if the students will visit more times some buildings than others. This study could be used to relocate some schools and services in order to obtain a more effective and efficient distribution or even to help planning the construction of a new university. In this section, the measurement process will be explained in order to analyze the number of users' change between buildings. Baseline measurement

In order to gather information from the wireless network, the SNMP agent was activated in all wireless APs using only the required messages. Every time a MAC address is associated to an AP, it sends a SNMP trap message to a central server. This information is stored in a database to be processed and analyzed.

First, APs are grouped according to the building where they are placed. This activity was not difficult because in our university each AP has a unique identifier, formed by the name from the building and the MAC address. All the APs in a building can be grouped easily using the same badge.

The database contains several tables in order to analyze the information. There is a table that stores the day and the month of the information jointly with the AP DNS name and the MAC that has been associated. Another table relates every access point with the building where it is placed. These tables allow us to make several queries such as:

- MACs registered
- Buildings with wireless access points
- MACs in every building
- MACs in every campus
- MACs that roam between buildings
- MACs that roam between buildings every day
- MACs that roam between buildings every month
- MACs associated to every AP during a day
- MACs associated to every AP during a month
- APs in every building
- APs where each MAC has been associated in a day
- APs where each MAC has been associated in a month

As we have said before, in this paper we are going to do a study where we analyze the macro-mobility and the micro-mobility. Knowing this starting point we will treat the data in different ways. Firstly we will group the associations that occurs in each building (macro-mobility), on the other hand we analyze the MAC addresses, which travel more and which are the access points with more visits. In order to process the information recollected in the database, we have used several SQL queries to extract the data needed and then we analyze the mobility of users.

Moreover, when we have selected the data, we have used a spread sheet to calculate some parameters as average number of visits in a building, percentage of visits, etc. With this spread sheet we have made some figures to represent better the information collected.

V. MOBILITY BEHAVIOUR OF USERS BETWEEN BUILINGS (MACRO-MOBILITY)

In this section, the macro-mobility in our campuses will be studied. In this part of the study we analyze the mobility between buildings and campuses, besides we try to describe the people behavior using these data.

A. Data processing

The people tracking measurement process is based on the number of people roaming among buildings. It is also measured where the people stop during a period of time. These measurements let us know the quantity of movements among all the buildings in the campus. In order to estimate the time that a user takes to go from the A building to another B building, we keep in mind that it could be the C building inside this itinerary. Roaming will exist among the A building, the C building and the B building, but the displacement will be considered from the building A to the B building. This study cannot be considered as a system of privacy intrusion, because this system does neither save a correspondence list of each person, nor their MAC addresses. Only the amount of movements is interesting. Moreover, the MACs used in this study are not real. The system changes a real MAC to another one (fictional) to preserve the privacy of people

Once all APs of each building have been grouped, the roaming among the APs of the same building will not be taken into account because these movements are inside the same building and the user does not move among buildings.

All these data has been stored in a database during a month to carry out this study. Firstly, the data have been purified because there was some information that was not useful to this study. These data have been taken daily and, therefore, we can show the information gathered during a regular day or show the information about monthly activity.

In order to process the data we have used a spreadsheet Excel 2007 with the NodeXL tool [16]. NodeXL is an extendible toolkit for network overview, discovery and exploration. The core of NodeXL is a special Excel 2007 workbook template that structures data for network analysis and visualization. Six main worksheets currently form the template. There are worksheets for “Edges”, “Vertices”, and “Images” in addition to worksheets for “Clusters,” mappings of nodes to clusters (“Cluster Vertices”), and a global overview of the network’s metrics (“Overall Metrics”).

As we will see in the following section, this software tool allows visualize the roaming among the buildings, the quantity of roaming made, filter the quantity of roaming, etc. NodeXL is a powerful tool that can help us analyze the behavior of the network. NodeXL aims to make analysis and visualization of network data easier by combining the common analysis and visualization functions with the familiar spreadsheet paradigm for data handling. The tool enables essential network analysis tasks and thus supports a wide audience of users in a broad range of network analysis scenarios.

B. Roaming during a day in the Vera Campus

Table II shows the numbers of changes among the buildings in one day. The rows represent the number of users that roam from that building to another. The situations where there is no mobility among a pair of buildings, that is, there is no MAC roaming between that two buildings in a day, are represented with a dash (-). The biggest value obtained in the user's mobility during a day is carried out from the building "E.T.S. de Gestión en la Edificación" and the "E.T.S. Ingeniería Informática" (4912 roamings). This is because of the buildings situation. The easiest way to access the "E.T.S. Ingeniería Informática" building is through the "E.T.S. de Gestión de la edificación" building. Furthermore, this last building is located in front of a tram stop, so it is an entry zone to this part of the university.

In Table II, it can be seen that "E.T.S. of Telecommunication" building has many roamings to other buildings. The reason is similar to the previous one, in front of this building there is also a tram stop and this building is located in the central area of the main campus. Among the "E.T.S. de Telecomunicación" and "E.T.S. de Caminos, Canales y Puertos" there are 2290 roamings in a day, this is because it is needed to cross the "E.T.S. de Telecomunicación" building to arrive to "E.T.S. de Caminos, Canales y Puertos" building. Another building that has a lot of roamings is the "E.T.S.I. Geodésica, Cartográfica y Topografía" building. In this case these roamings were caused because it is placed near the snack bar. This snack bar has wireless coverage thanks to the APs of the "E.T.S.I. Geodésica, Cartográfica y Topografía" building. We will see several movements related with this building in our studies. Lastly, among the "E.T.S.I. Caminos, Canales y Puertos" building and "E.T.S. Arquitectura" building there are 2202 roamings by day. These roamings could be due to:

a) The proximity between both buildings

b) The relationship of contents that are taught in both buildings. May be students and/or professors walk from one building to the other in order to carry out theoretical or practice classes.

C. Roaming during a month in the Vera Campus

Table III shows the roaming value carried out during a month among the buildings of the Vera Campus of the Universitat Politècnica de Valéncia. In this table, the data movements from one building to another, and vice versa, have been added. That is, we have not considered the direction of the roaming.

The maximum number of roamings in one month is carried out among the "E.T.S. de Gestión en la Edificación" building and "E.T.S. Ingeniería Informática" building. We explained why before. The number of roamings between "E.T.S. de Telecomunicación" and "E.T.S. de Caminos, Canales y Puertos" buildings was 26746. In this case roamings were due to the proximity of the buildings and because it is necessary to cross the building "E.T.S. de Telecomunicación" to arrive to "E.T.S. de Caminos, Canales y Puertos" when the people come from the tram.

There were also many user movements between "E.T.S. Arquitectura" building and the "E.T.S. de Telecomunicación" building (17704 user movements in a month).

The "E.T.S. Ingeniería del Diseño" and "E.T.S.I. Industriales" buildings have also a lot of roamings. These buildings have many users registered (see Table I). There are many roamings between these buildings due to the likeness of the studies. It seems that there are many subjects imparted by the same department, so there are professors moving between these buildings indistinctly.

"F. de Bellas Artes" building had less roamings. We think that it is because "fine arts" students do not use too much computers, laptops or mobile devices to connect to the wireless data network, as it happens with the students of the other buildings (this is a technical university).

Lastly, "E.T.S. Medio Rural y Enología", the "E.P.S. Alcoy", and "E.P.S. Gandia" have very few movements between them (there are 1350 roamings among "E.P.S. Alcoy" and "E.P.S. Gandia"). This is because these buildings belong to different campus located in different cities. There are users that one day can be in a campus and, after some hours, they are in another campus. In this case there is a hard-roaming because the user loses the connection during a large time because the user is travelling. If we take into account the buildings of the Vera Campus we observe that the most number of roamings are between the Vera Campus and the rest of campuses.

D. Roaming between Campuses

Figure 1 shows the values of the roamings carried out among the different campuses of the Universitat Politècnica de Valéncia during a month. These campuses are Escuela Politécnica Superior de Gandia (located in Gandia city), Escuela Politécnica Superior de Alcoy (located in Alcoy city), E.T.S. Medio Rural y Enología (located in one of the main avenues of Valencia City).

The number of movements in a month between the Gandia's campus and the Vera's campus are 5497. It is the highest value between campuses. There are quite a lot of movements between Gandia's campus and Vera's campus because they are relatively near (around 56 km.). There is a good public transport communication and many professors of Gandia's campus also work in Vera's campus. Moreover, a lot of lecturers of E.P.S. de Gandia have in Vera's campus their place, where they do their researches. The roamings between Alcoy's campus and Vera's campus is quite lower (1844 roamings), the reasons are very similar but there are fewer movements because in Alcoy there are less people.

TABLE II. ROAMING BETWEEN BUILDINGS IN ONE DAY.

	E. P. S. Alcoy	E.T.S Ingeniería Informática	E.T.S. Arquitectura	E.T.S. Gestion en la Edificación	E.T.S. De Ingeniería Del Diseño	E.T.S. Medio Rural Y Enología	E.T.S.I. Agronomos	E.T.S.I. Caminos, Canales y Puertos	E.T.S.I. Telecomunicación	E.T.S.I. Geodesica, Cartografica y Top.	E.T.S.I. Industriales	E. P. S. Gandia	Facultad De Administración y Dirección de Empresas	Facultad De Bellas Artes
E. P. S. Alcoy		—	—	—	—	—	7	—	—	—	—	—	—	—
E.T.S Ingeniería Informática	3		1296	—	428	—	385	564	—	—	—	—	113	40
E.T.S. Arquitectura	—	—		—	—	—	701	—	—	—	—	—	42	—
E.T.S. Gestion en la Edificación	15	4912	986		341	—	494	568	—	—	—	—	111	195
E.T.S. Ingenieria del Diseño	6	—	249	—		—	137	200	—	—	—	—	279	27
E.T.S. Medio Rural y Enologia	—	7	12	17	4		24	11	—	—	9	—	2	4
E.T.S.I. Agronomos	—	—	—	—	—	—		—	—	—	—	—	55	—
E.T.S.I. Caminos, Canales y Puertos	2	—	2202	—	—	—	248		—	—	—	—	47	49
E.T.S.I. Telecomunicación	2	663	1379	1307	458	20	218	2270		—	507	28	100	95
E.T.S.I. Geodesica, Cartografica y Topografia	4	94	50	92	304	1	47	40	64		50	2	1047	12
E.T.S.I. Industriales	12	491	742	537	475	—	965	434	—	—		5	76	72
E. P. S. Gandia	32	16	7	20	7	—	10	24	—	—	—		4	—
Facultad Administración y Dir. de Empresas	—	—	—	—	—	—	—	—	—	—	—	—		—
Facultad Bellas Artes	—	—	86	—	—	—	52	—	—	—	—	—	10	

TABLE III. ROAMING BETWEEN BUILDINGS IN A MONTH.

	E.T.S de Inge. de Informatica	E.T.S. de Arquitectura	E.T.S. de Gestión de la Edificación	E.T.S. de Ing. del Diseño	E.T.S. del Medio Rural y Enología	E.T.S.I. de Agronomos	E.T.S.I. de Caminos, Canales y Puertos	E.T.S.I. de Telecomunicación	E.T.S.I. de Geodesica, Cartografia y Top.	E.T.S.I. de Industriales	E. P. Superior de Gandia	F. de Adm.y Dirección de Empresas	F. de Bellas Artes
E. P. S. Alcoy	67	378	389	161	2	136	67	96	163	315	1350	70	2
E.T.S de Ingenieria de Informatica		15643	62768	5806	166	7906	7344	10429	1276	6704	845	1231	3448
E.T.S. Arquitectura			13226	4163	267	8591	15180	17704	652	9314	502	670	1164
E.T.S. de Gestión de la Edificación				5535	494	7906	7531	15996	1572	8389	698	1837	2431
E.T.S. de Ingenieria del Diseño					98	1914	3304	7644	3214	7804	334	3470	480
E.T.S. del Medio Rural y Enología						636	423	365	19	184	20	78	38
E.T.S.I. Agronomos							3104	3139	487	11437	572	965	708
E.T.S.I. de Caminos, Canales y Puertos								26746	464	5811	979	557	577
E.T.S.I. de Telecomunicación									958	7785	1054	1286	1181
E.T.S.I. de Geodesica, Cartografía y Top.										767	145	11086	161
E.T.S.I. Industriales											230	1162	796
E. P. S. Gandia												99	39
F. de Administración y Dir. de Empresas													168

The number of roamings among "E.T.S. Medio Rural y Enología" and Vera's campus is 2768, but there are very few movements between "E.T.S. Medio Rural y Enología" and Gandia's campus, and between "E.T.S. Medio Rural y Enología" and Alcoy's campus.

We can see in Figure 1 that Vera's campus is the campus that receives more visits. This result is a prospective fact because Vera's campus is the main campus of our university and most of the formalities, documentation procedures, applications and administrative issues have to be made there.

Lastly, we can see that 1350 movements per month are carried out among "E.P.S. de Gandia" and "E.P.S. de Alcoy". The main reason of these movements is the existence of many professors that teach classes in both campuses so they must move between them.

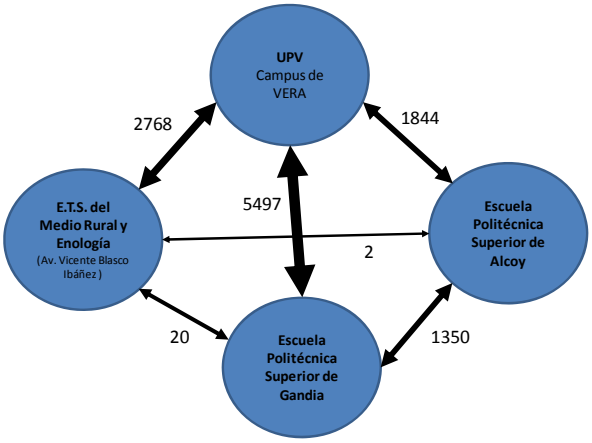

Figure 1. Roaming between Campuses

E. People Behavior

In this section, we will evaluate the users' movements by day in the Vera's campus. We will also analyze the number of changes carried out per user in a day.

In Figure 2, the 5 highest roaming values between buildings are shown. The situation of more mobility is given among "FI" and "GE". It has a value of 4912. In this figure, all of displacements shown have a higher value than 1834 movements/day. In this case, the movements are given among "IND"-"BIB", "DSIC"-"EI", "ARQ"-"BIB" and "CASALU"-"BIB". With these data we can obtain some information. E.g. the students of "E.T.S. Arquitectura" and "E.T.S.I. Industriales" visit the university library more times than the other students of the university. On the other hand, there are many movements among the university library building and the "Student's house" building (this building is used by the students to study, to connect to Internet and to develop any activity). This movement is due to the vicinity of buildings (see Figure 2) and many users that are in one of the buildings usually visit the other building. In Figure 3, the 10 highest roaming values are shown. In this case, we have the 5 previous movements (see Figures 2) and 5 more. These 5 new displacements are carried out among "GE"-"DSIC", "GE"-"ARQ", "FI"-"EI", "ARQ"-"ASIC" and "ARQ"-"CCP". The minimum number of roamings of all displacements seen in Figure 3 is 1483 per day. One of the buildings that had more movements is "ARQ" (E.T.S. Arquitectura). The main reason seems to be because some services are offered in this building. For example, this building has some snack bars, and there are some banks in the bottom plant. We can also find a hairdresser, bookstores, etc. and it can be found a great number of movements among this place and other buildings.

Figure 4 represents all the Vera's campus movements during one day. Almost all of buildings have users' mobility. We can state that the wireless network of our university is very robust. This network can support the mobility of all users giving the appropriate service.

Lastly, we analyzed the number of changes per person. This information is shown in Table IV.

Expression (1) is used to know the number of changes per person.

$$Changes/Person = \frac{Total\_changes\_between\_buildings}{registered\_people\_Building\_1 + registered\_people\_Building\_2} \quad (1)$$

The buildings that have the biggest number of movements per person are "E.T.S. de Gestión en la Edificación" and "F. Informática" buildings. They obtained a value of 0.797 movements per person. The movements among "IND"-"BIB" and "ARQ"-"BIB" also possess a high number of changes per person, 0.644 and 0.608 respectively.

VI. MOBILITY OF USERS INSIDE OF A BUILDING (MICRO-MOBILITY)

In the previous section we have analyzed the mobility between buildings and campuses. Now in this section we are going to study the mobility of the most mobile users from the point of view of small-scale mobility.

A. First steps in micro-mobility

In this subsection we are going to present the data from the point of view of micro-mobility. Firstly, we have selected the number of clients per day, which visit only one AP, two APs, and so on until 9 APs. This is depicted on table V.

TABLE IV. CHANGES/PERSON BETWEEN SOME BUILDINGS .

Buildings with Roaming	Registered People Building 1	Registered People Building 2	Changes	Changes/person
FI-GE	3240	2920	4912	0,797
ARQ-BIB	3858	—	2344	0,608
IND-BIB	3479	—	2239	0,644
CCP-ARQ	3145	3852	1772	0,253
FI-EI	3240	60	1744	0,528
GE-DSIC	2920	40	1705	0,576
FI-DSIC	3240	40	1484	0,452
EI-GE	60	2920	1380	0,381
TEL-ARQ	1409	3858	1347	0,256
CASALU-ARQ	—	3858	1308	0,339

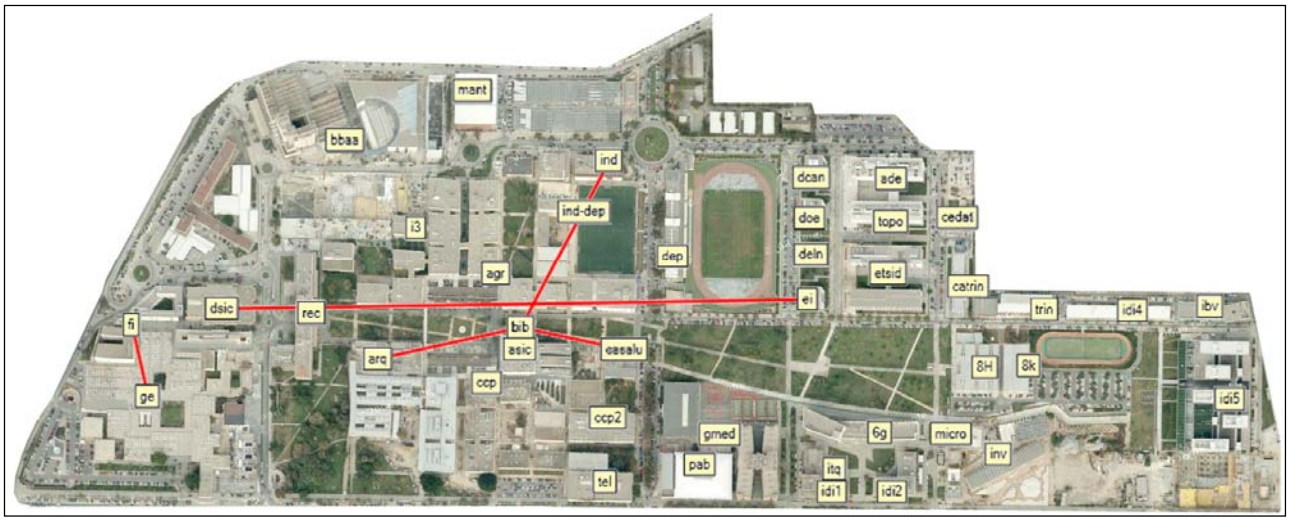

Figure 2. 5 highest roaming values in Vera's Campus

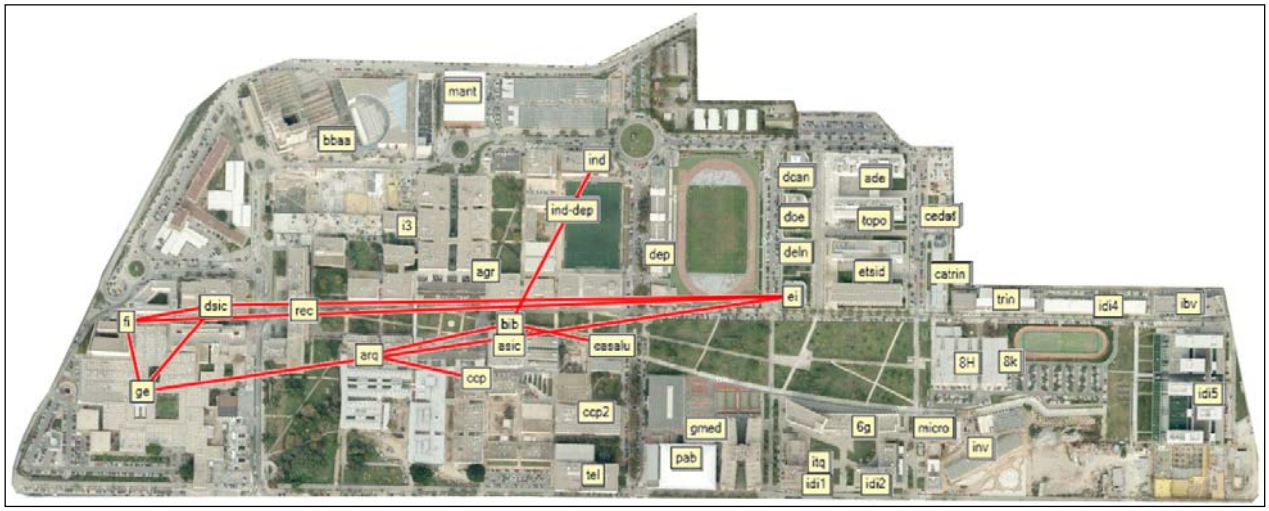

Figure 3. 10 highest roaming values in Vera's Campus

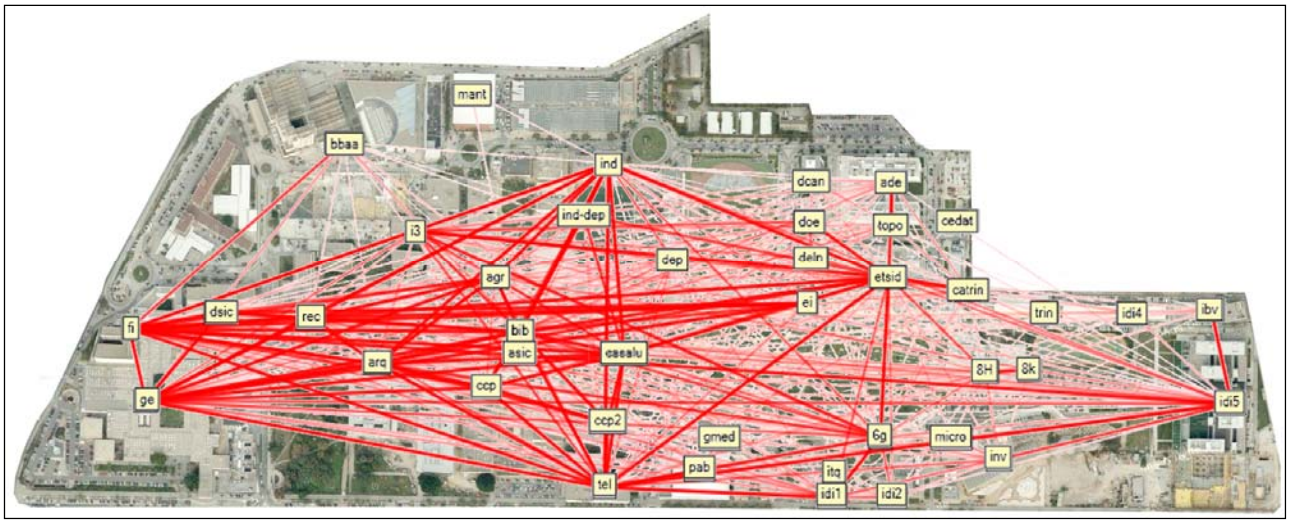

Figure 4. All roaming values in the Vera's Campus

These data were taken for a month. We assume that clients with 1 or 2 APs visited are statics because the wireless coverage in our campuses is nearly 100% so there are a lot APs to cover the maximum number of sites. Other users visit about 3 or 4 APs a day, meaning that they are limited mobile clients. From now on we will call that kind of clients as "laptop clients" because this type of users take their laptops and they go from their offices to another site to make a meeting. Finally we have the true mobile users, and we consider any user as a mobile user when a client has been detected in 5 or more than 5 different APs.

In Figure 5, it is represented the visits collected per day in each case. In this figure we can observe that all lines have the same pattern. We can extract several conclusions from this figure. First there are more static people than mobile people, because we can see that the line referenced to 1 AP visited per day is higher than others. Moreover, there are some days that break the normal pattern, this is due to these days are Sundays (on Sundays, faculties are close and only the library and the student house are open), e.g. the eighth day. Besides, between the 14th and the 20th day, there are few visits registered in APs, because for these days we had holiday time. As we can see, these maximum and minimum values are in all patterns, independently if we have a mobile or a static user. In mobile users these values are softer because we have fewer users.

Total numbers in this test can be viewed in the table V. In this table we shown the total visits that have only one AP, 2 AP and so on until 9 AP. Apart from that, we have calculated the percentage of these total visits and the percentage of visits when a client has visited more than *x* APs, being *x* the number of APs visited. In this table we can observer that the 58.6% of the users in our university are static. The 23.1% are laptop users and the rest are mobile users (18.3%). Although this percentage seems a bit low, in our community means that 7686 people are mobile users. So, it is important to know this data when we redesign the wireless network or when we try to implement roaming systems to improve the quality of the end-applications.

In order to analyze our collected data we need to select the correct data and the data that disturb our analysis. According to these criteria, the holiday days and Sundays we have to delete. In Figure 6 we represent a boxplot to see where the most important data are. This figure shows the maximums and minimums (start and end of lines) clients who have been associated to 1 AP, 2 AP, 3 AP, etc. In each AP, the green box represents that the most of data are between these intervals. For example, if we pay attention in the users who visit two APs, minimum value is 179 visits (it occurred the 20th day), the maximum value was obtained on the 2nd day and the number of visits was 1541. For this case the most of visits are between 812.5 and 1167. We can observe that the boxes are smaller according to the number of APs are higher, it is due to the number of mobile clients are lower than static or laptop clients. When we have clients, which have visited 7, 8 or 9 access points, the variability of their associations is very small compared with the static clients, it occurs for the same reason. In table V we can see that the most mobile clients (7, 8 or 9 APs visited) represent the 5.6 % of total registered users.

Finally in this subsection we want to show another figure, which represents the APs needed according to the different visits number (see Figure 7). In order to make this figure we have deleted the data, which could get worse our analysis, we have deleted the 2nd day and the holiday period (14th – 20th, both included). In this figure we have used the average number of visits to print the line called data and then we have made a mathematical approximation with the trend line.

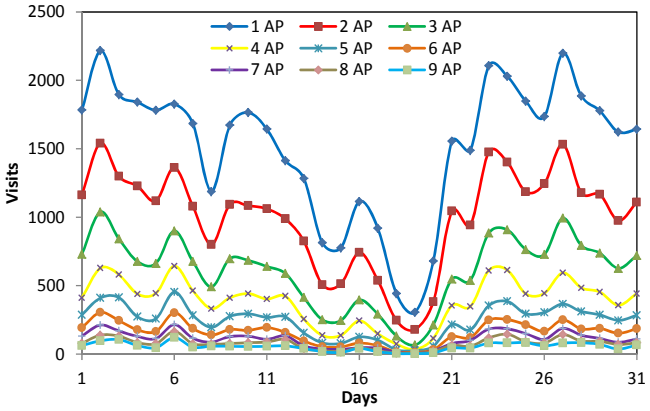

Figure 5. Visits per day for a month

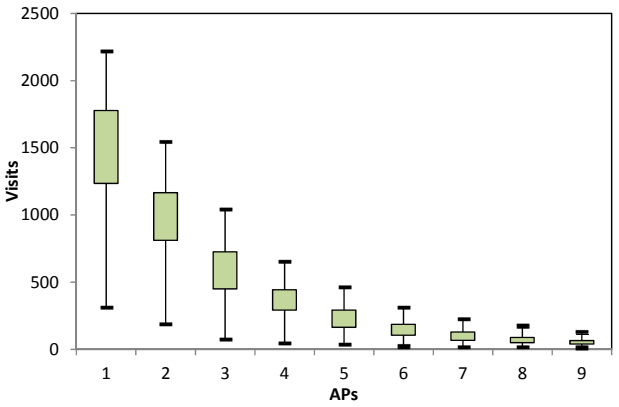

Figure 6. Boxplot of visits per AP.

TABLE V. TOTAL APS VISITED PER DAY FOR A MONTH.

APs visited per day	1 AP	2 APs	3 APs	4 APs	5 APs	6 APs	7 APs	8 APs	9 APs	> 10 APs
Total	46936	31010	18870	11900	7564	4904	3373	2393	1704	4408
Percentage	35,3%	23,3%	14,2%	8,9%	5,7%	3,7%	2,5%	1,8%	1,3%	3,3%
Total > x		64,7%	41,4%	27,2%	18,3%	12,6%	8,9%	6,4%	4,6%	3,3%

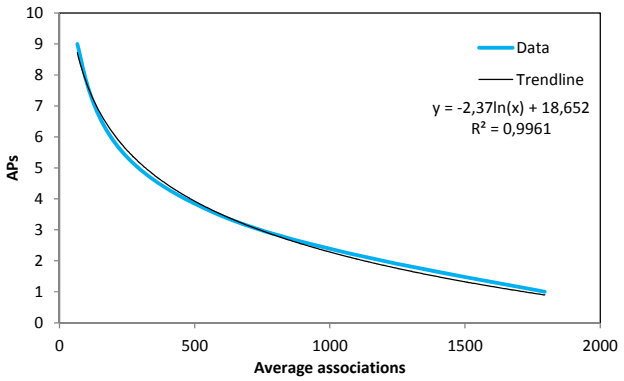

Figure 7. AP needed according to is tythe visits.

According to Figure 7, we can affirm that if we had few visits in our APs, we should configure our APs to manage mobile user correctly. However, if we had a lot of visits in our APs, we should configure our wireless network to provide the service to static users. So, the question is regarding to the number of visits, how I know if a user is mobile?

We relate the mobility of a user according to the different number of APs visited. When 1 or 2 APs are visited, we say that the user is static, but when 3 or 4 APs are visited, the user is light itinerant and when more APs visited, the user is mobile.

We have estimated relationship between the number of visits and the number of APs visited, in this scenario. This relationship is shown in expression 2. *APv* is the number of access points visited and *visits* are the different visits registered by the access point.

$$APv = -2.37 \ln(visits) + 18.652 \quad (2)$$

This expression has a R^2 equals to 0.9961. R2 means the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be defined by the model. In our case the 99.61% of data follow the model, but this model is calculated with average visits. For this reason, we have checked our real data with this model and we have obtained that the 84.48% of data follow this model. It is a good result and our model could be used for our university to make some changes in the network as we will see in the conclusions.

B. *Mobile users*

In this subsection we analyze the behavior of the most mobile users. In order to know who is he or she, we have defined several SQL queries (but always maintaining the anonymity of the persons). In order to have a better study, we have observed more data. In this case, the observation time period has been two months (March and April of the year 2010).

Figure 8 plots the visits that a regular user makes during 2 months. We have observed that during this time, this user does not have mobility every day, so this figure shows only the days that the user has had mobility.

According to Figure 8, we can state that the user has quite mobility, because in many cases it has had more than 10 visits per day (even up to 55 visits in one day), so it is an important value. It seems that this user works only with a mobile device and it is always connected to the wireless network, which allows us to register all the user's movements.

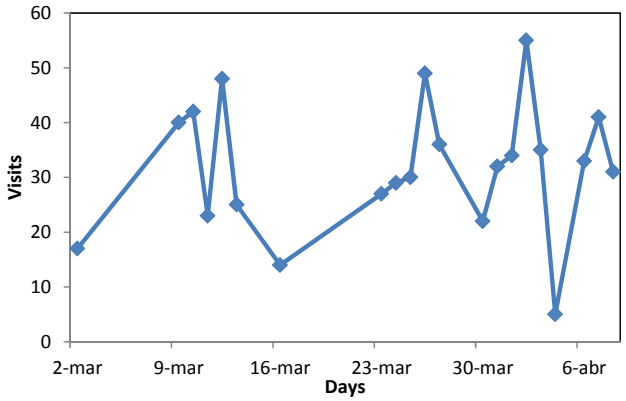

Figure 8. Visits of the most mobile user.

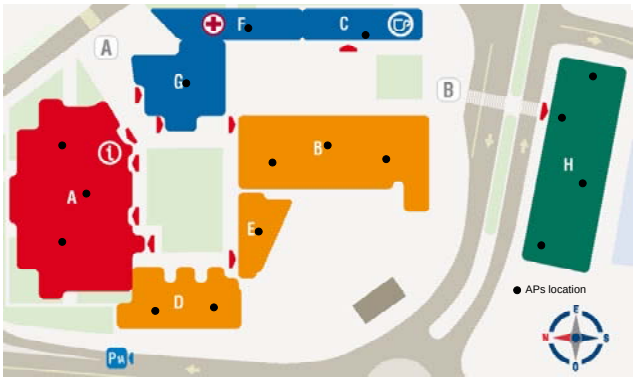

Figure 9. Gandia's Campus.

C. Micro-mobility in Gandia's Campus

Now, we are going to fix in a smaller place than Vera's Campus, this is Gandia's Campus. This campus is the second biggest one, if we pay attention to the number of people (according to Table I, 2039 people are registered).

We can see the campus in Figure 9. This figure represents several buildings; each one has its name according to a letter. In this map, we have located the access points (black points) approximately in the correct place. We say approximately because all buildings have several floors and in order to make a good design, sometimes the APs are situated in other places because we want more coverage or less interference.

Moreover, the name of each AP in our university helps us to know where it is more or less located. The names follow a structure. We are going to explain it with an example.

The access point called *ac1-gnd1a0e.net.upv.es* has the following explanation. First three letters *"ac1"* is the model of the access point used. Then, the following three ones, in this case *"gnd1"*, is the location. *"gnd1"* is Gandia, but this part of the name is the same for all names in Table IV. Then, there is the name of building and the floor, in this case *"a0"* means building A and floor 0. Next, there is the orientation: north = n, south = s, west = o and east = e. In the example we have an access point located in the east, *"e"*. Finally, we have *"net.upv.es"* means that this device belongs to UPV.

In order to analyze the movement of a user, we have selected the most mobile user of Gandia's Campus. We are going to see his movement using the Table VI. It shows us the access point name, where the user has visited, the day and the time.

In this case we can see that this user has quite mobility in few hours. The mobility of this user starts at 11:00. First he connects to *"ac1-gnd.1a0c.net.upv.es"*, so he is located on building A, in floor 0 and on the AP of the center. He moves to the east (*"ac1-gnd.1a0e.net.upv.es"*) on the same floor and then he goes up to the first floor. He stays there during an hour. Then, the user does the similar movement steps. At 13:34 he stays at the AP of the center of building A, in floor 0. Then, he goes to the east direction and to the first floor. At 14:42 he changes to the D building, firstly to the ground floor, and then to the first floor. Next, he goes to the E building, and he goes up to 6th floor (*"ac1-gnd1e6c.net.upv.es"*). Finally, he goes to the outside of the buildings (*"ac3-extgnd1.net.upv.es"*) until 16:05.

TABLE VI. MOBILITY IN GANDIA'S CAMPUS.

MAC address	Month	Day	Hour	AP name
000d.720c.fc37	Mar	5	11	ac1-gnd1a0c.net.upv.es
000d.720c.fc37	Mar	5	11	ac1-gnd1a0e.net.upv.es
000d.720c.fc37	Mar	5	11	ac1-gnd1a1e.net.upv.es
000d.720c.fc37	Mar	5	12	ac1-gnd1a0c.net.upv.es
000d.720c.fc37	Mar	5	13	ac1-gnd1a0c.net.upv.es
000d.720c.fc37	Mar	5	13	ac1-gnd1a1e.net.upv.es
000d.720c.fc37	Mar	5	14	ac1-gnd1d0c.net.upv.es
000d.720c.fc37	Mar	5	14	ac1-gnd1d1s.net.upv.es
000d.720c.fc37	Mar	5	14	ac1-gnd1e6c.net.upv.es
000d.720c.fc37	Mar	5	14	ac3-extgnd1.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1a1c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1a1e.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1a1o.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1c0c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1d2c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1e1c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1e5c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1e7c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1f1c.net.upv.es
000d.720c.fc37	Mar	5	16	ac1-gnd1g2se.net.upv.es
000d.720c.fc37	Mar	5	16	ac3-extgnd3.net.upv.es
000d.720c.fc37	Mar	5	17	ac1-gnd1d1n.net.upv.es
000d.720c.fc37	Mar	5	17	ac1-gnd1e1c.net.upv.es
000d.720c.fc37	Mar	5	17	ac1-gnd1e2c.net.upv.es
000d.720c.fc37	Mar	5	17	ac1-gnd1e3c.net.upv.es
000d.720c.fc37	Mar	5	17	ac1-gnd1e6c.net.upv.es

At 4 p.m he starts to move a lot. He is on the first floor of building A, then, he moved to building C. When he finished his activities on this building he went to building D, and then, he went to building E, staying in the first floor, the 5th floor and, finally, the 7th floor. Next, he went to the 1st floor of building F. Then he went to the 2nd floor of building G and, finally, he moved to the outdoor zone. Finally, at 5 p.m. he went from the 1st floor of building D to building E. In this building he moved from the 2nd, to the 3rd and, finally to the 6th floor. And, then, he finished his movements.

According to these movements we could indicate what kind of person has this behavior. In this case, we think that he could be a language lecturer because, first, he was on building A, where the classrooms are placed, maybe he is teaching his lessons. Then, he went to his office (language offices are on building E). From 14:42 to 16:05 he went to take his lunch outside of the university. Then, he went to building A. May be he enters to the university through this building. He went to his office again and, then, he went to building F (another place where are language offices) to see a colleague and they went to the bar (building C). Finally he came back to his office. With this type of micro-mobility we can observe the people behavior only by using wireless networks. When the same movements are done by many people at the same time, we can predict the needed resources for a period of time in a specific place. May be, the network will require more devices in an specific part of the network because of the number of connections during a limited period of time one day in a week.

VII. Conclusion and Future Work

In this paper, we presented a user mobility study based on the roaming of the MACs in the wireless network of the Universitat Politècnica of València. We made two types of studies. First we have analyzed the macro-mobility. In this case, the mobility between buildings and campuses has been studied. Second, we studied the micro-mobility that let us know the mobility inside buildings.

Using the measured data we can analyze the behavior of students and professors in the campus. This study let us build reallocation bandwidth scenarios to increase the comfort of the end user, i.e., if we note that a set of users go far to an area, we could determine in detail where will they go. May be this information could be used to reallocate services and departments in the campus by changing their place or putting a branch near the appropriate place.

Besides, we could see that making a detailed WLAN study (micro-mobility), and knowing behavior of some people allow us to know the type of person that is walking through the areas. This type of study could let us know where we should place the network printers, the coffee and drinks machines, and some internal services.

We have also shown that we can obtain some information about the user just studying his mobility profile. This kind of information could be used also for advertising services, direct marketing and similar tasks.

Finally, this analysis will be the basis for our future work on a dynamic management and control system of the wireless network. According to the user mobility, the system will be able to give more bandwidth in those areas where more users are, and the roaming system will be more efficient. Even we are thinking on talking with the transport services to let them know which is the mobility between our campuses in order to provide an adequate transport service.

References

[1] IEEE 802.11-2007 Standard. Available at http://standards.ieee.org/getieee802/download/802.11-2007.pdf

[2] Wi-Fi Alliance. Available at http://www.wi-fi.org/ [accessed, July 14, 2009]

[3] Bluetooth SIG. Specification of the Bluetooth system. Specification Version 1.1 (Feb. 2002). Bluetooth SIG.

[4] E. Dahlman, B. Gudmundson, M. Nilsson and J.Sköld, "UMTS/IMT-2000 based on Wideband CDMA", IEEE Communications Magazine, vol. 36, pp 70-80, September 1998.

[5] M. Garcia, J. Tomás, F. Boronat and J. Lloret, "The Development of Two Systems for Indoor Wireless Sensors Self-location". Ad Hoc & Sensor Wireless Networks, vol. 8, num. 3-4, pp. 235-258. 2009.

[6] Sandra Sendra, Miguel Garcia, Carlos Turro, Jaime Lloret, "People Mobility Behaviour Study in a University Campus Using WLANs". Ubicomm 2009. Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp.124-129, Sliema, Malta. October 2009.

[7] Z Chen and W Chew, "DynaTrack: Dynamic Directional Instructions for Tracking Assets and People in Office Environment", Available in http://hci.stanford.edu/srk/cs377a-mobile/project/final/chen-chew.doc

[8] J.G.Markoulidakis, G.L.Lyberopoulos, D.F.Tsirkas, E.D.Sykas, "Mobility Modeling in Third Generation Mobile Telecommunication Systems", IEEE Personal Communications, vol. 4, Issue. 4, pp 41-56. Aug. 1997.

[9] G. Radinovic and K. Kim, " Feasibilty study of RFID/ Wi-Fi / BlueTooth wireless tracking system for underground mine mapping – Oklahoma", In procedings of "Incorporating Geospatial Technologies into SMCRA Business Processes", Atlanta, GA, March 25 – 27, 2008.

[10] B. Issac, K. Hamid and C.E.Tan, "Wireless Mobility Management with Prediction, Delay Reduction and Resource Management in 802.11 Networks", IAENG International Journal of Computer Science, Vol.35, Issue 3. August 2008.

[11] R. Zhou, " Wireless Indoor Tracking System (WITS)", In proceedings of 12th in a series of conferences within the framework of the European University Information Systems Organisation (EUNIS), Tartu, Estonia. 28-30 June, 2006

[12] T. Liu, P. Bahl and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks",IEEE Journal Selected Areas in Communications, Vol. 16, No. 6, pp. 922-936. August 1998.

[13] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects", In proceedings of ACM SIGMOBILE Mobile Computing and Communications Review (MC2R), vol.5 no.3, p.55-66. Rome, Italy. 16-21 July 2001.

[14] Minkyong Kim and David Kotz. Periodic properties of user mobility and access-point popularity. Personal Ubiquitous Comput. Vol. 11, isue 6, pp. 465-479. August 2007.

[15] Joy Ghosh, Matthew J. Beal, Hung Q. Ngo, and Chunming Qiao. On profiling mobility and predicting locations of wireless users. In Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality (REALMAN '06). Florence, Italy. 26 May 2006.

[16] M. Smith, B. Shneiderman, N. Milic-Frayling, E.M. Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, "Analyzing (Social Media) Networks with NodeXL", In Proceedings of the Fourth International Conference on Communities and Technologies, C&T '09. The Pennsylvania State University, USA. June 25-27, 2009.

Web 2.0 Data: Decoupling Ownership from Provision

Mark Wallis, Frans Henskens, Michael Hannaford
Distributed Computing Research Group
University of Newcastle
Newcastle, Australia
Email: mark.wallis@uon.edu.au, frans.henskens@newcastle.edu.au, michael.hannaford@newcastle.edu.au

***Abstract*—Current Internet trends have caused us to outgrow existing online data storage paradigms. This paper presents an extended model for distributed online data storage. The model addresses issues of data duplication, data freshness and data ownership, while facilitating two modes of data access - direct and indirect. Direct data access is implemented using advanced handoff techniques while indirect access is implemented using robust server-to-server protocols that enforce strict policies on data management. Results are presented that compare this solution to existing technologies and an example migration path is described for existing Web 2.0 applications wishing to adopt this new paradigm.**

***Keywords*-distributed, storage, personal data, data ownership**

I. Introduction

The current data storage model for Web 2.0 applications defines data stores managed by the web application owner. Web 2.0 has increased the popularity of user-generated content, which has placed massive quantities of data into these stores. Users are often forced into signing EULAs that restrict their ownership of this data. In fact, having the application provider manage this data has resulted in problems with data ownership, data freshness and data duplication. Previous work [1], [2] introduced the concept of a model that vests content storage with the original content generator. This allows the content generator to retain ownership while supporting most Web 2.0 applications ability to function as they do currently. The model creates the possibility of a single-version-of-the-truth for user-generated content.

The Distributed Data Storage Application Programming Interface (DDS API) allows web applications to seamlessly present user-generated content to a 3rd party user. Using this API, 3rd-party users interact directly with both the web application and the DDS systems hosting the end user's content. Data owners can manage their data elements by interacting with a DDS directly while also exposing this data to web applications via a publish/subscribe model. This paper presents version 2 of the Distributed Data Service (DDS) API. In version 2, information can be accessed via a pass-through mode, which allows 3rd-party web browsers direct access to remote DDSs. Additionally, web applications may request information directly from DDS services under strict contractual arrangements before enriching the data and providing it to the end user. This direct access is protected by strict electronic contracts and dynamic handshaking.

The justification for a distributed storage model is based on the concept that in a Web 2.0 application the majority of information is generated in a distributed fashion by end users. The term 'Web 2.0' is a business generated term, which can be traced back to 2005 when O'Reilly first defined the concept of web generations [3]. At the time, Web 2.0 was identified as any web application that matched the following criteria:

- The application represents a service offering and is not pre-packaged software.
- The application data evolves as the service is used. This is in contrast to applications in which static data is generated solely by the application owner.
- A framework is provided that supports and encourages user submission of software enhancements. These submissions are generally in the form of plug-ins or extensions to the web application.
- Evolution of the application is driven by the end user as well as the application owner.
- The application interface supports interaction from multiple client devices such as mobile phones and PDAs.
- The application provides a lightweight, yet dynamic, user interface.

A high-level overview of the Web 2.0 design is shown in Figure 2.

A key criterion of interest to this research is that evolution of the website is tied to the degree of user interaction. This is driven directly by the fact that the primary service provided by a Web 2.0 site generally relies heavily on user-generated content. The more that data owners interact with the website, the greater the experience of all users. This paradigm has proven popular because website owners are no longer solely responsible for content generation. The fact that the amount and richness of the data provided by data owners can be directly tied to the success of a website only exacerbates the problem of data ownership, as the contained data becomes an important asset for the website owners. Without this asset, they would have substantially less to offer to their user base.

This paper formally defines the interface used by the distributed data service - the DDS API version 2. Section II provides some background information on where this

research sits in our larger work. Section III follows with an overview of the problems which the DDS addresses. Section IV provides an overview of existing approaches to this problem. Section V formally defines the DDS APIv2 as a specification that can be used to implement distributed data systems using a distributed content management model. It encompasses all phases of the data management including the insertion and retrieval of data by 3rd parties. Section VI reviews the strict contractual model used in enriched-mode. Section VII describes the proof-of-concept implementation provided by this paper, which includes multi-language and multi-platform components. Sections VIII and VIII-C present performance metrics and provide a comparison of the DDS solution against other existing technologies. Section IX provides an overview of how the solution scales before Section X presents an overview and conclusion.

A preliminary version of this work has been reported [1].

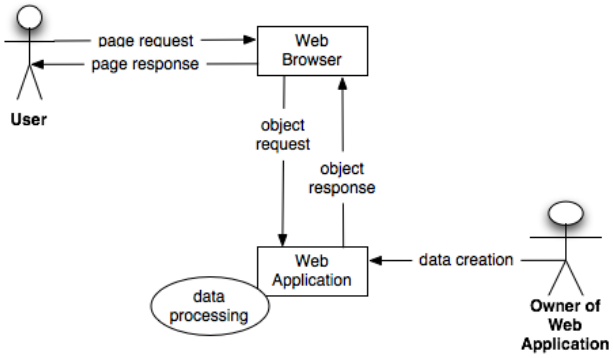

Figure 1. Web 1.0 model

II. Background

Web 1.0, as shown in Figure 1 web sites comprise service providers creating and presenting static content for consumption through users' web browsers. The move to Web 2.0 has seen an increase in focus on user-generated content and dynamic user interfaces. Support for these concepts was never built into the original web browser design, which was focused primarily on static content [3]. Security was also added as an afterthought, quite unsuccessfully, as can be seen from the large number of security alerts related to web browser technology [4]. While attempts have been made to improve the browsers, often using complete rewrites [5], it has become apparent that a re-design of the way we interact online is required.

The advent of Cloud Computing [6] has brought about change to the server-side of the Web. The deployment of Web 2.0 applications in the Cloud has raised concerns to do with data security and regional regulations [7]. Vendor lock-in has also become a major issue [8] with vendors refusing to share common APIs.

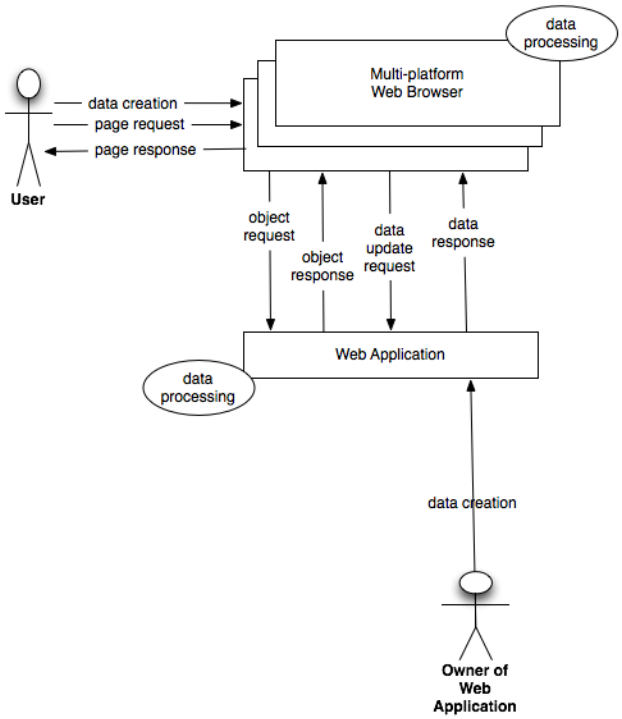

Figure 2. Web 2.0 model

The authors believe that the roles of Internet-connected computers and software should be re-examined. This re-examination suggests that the concept of the Cloud should be expanded to encompass both server and client side resources [9]. Developers can then use component-based software engineering tools to develop applications built from various components that execute in a distributed fashion [10].

The distributed component model raises the question of where data should be stored on the Web. The DDS API allows data storage to be distributed across the Internet while decoupling ownership and management of data from the provision of web applications. The data owner retains access and control over security of their data while continuing to allow the current generation of web applications to function with minimal change.

III. Problem Description

The problems with the existing storage models in Web 2.0 can be viewed from two aspects - the web application and the end user. From a web application viewpoint, monolithic storage structures have resulted in the following key problems:

1) Web applications must host and manage large-scale database systems to store all the user-generated content in their data centre.
2) Since all the data is stored by the web application the hosting network must bear the full load of transferring

that data between the monolithic storage structure and the user-base.

3) Web application providers must deal with complex privacy and regulation issues stemming from the fact that access to user-generated content can sometimes be restricted by privacy laws.
4) Monolithic storage models introduce large single-points-of-failure. In a Web 2.0 model, the risk is that a single Web application hosting valuable data may go offline, either temporarily or permanently.

Three key problems affecting the end user are data ownership, data freshness and data duplication [2].

1) Data ownership is the most important issue of the three. This relates to the fact that 3rd party web applications can currently place restrictions on the usage of a user's data just because they store it locally on their servers. Data owners are forced to agree to EULAs to use services that can take away the owner's rights to their own data, despite them owning the original data [11].
2) Data freshness refers to the situation in which a user provides data to one or more web applications and that data changes at a later date. The original data provided to the web applications then becomes stale, unless the user is able to recall and update all web applications to whom the data has been provided. The manual user update would be performed on a per-web-application basis, possibly dealing with access issues such as expired login or forgotten credentials.
3) Data duplication refers to the situation in which multiple web applications store copies of the same piece of data. While the associated implementations seems trivial when considering such pieces of data as a single postal address, the issue expands dramatically when dealing with multimedia data such as image, music and video libraries.

These issues have all developed through the increased use of SAAS and Web 2.0 architectures. SAAS (Software as a Service) is a model where software is provided as an online service as opposed to a distributed executable piece of code. While the benefits of a SAAS model are well documented, it is the combined use of SAAS with the increased level of user-content being stored online, which has led to the above problems. The model presented in this paper resolves these issues while also maintaining the benefits of SAAS and the Web 2.0 model.

IV. Literature Review

Web applications treat user-data as as a commodity [12] and lack the motivation to relinquish control. However, it is in the user's best interest to control their data and privately manage the three key issues of data duplication, data freshness and data ownership.

Much can be gained by reviewing distributed storage in general. Technologies such as CORBA [13] have long leveraged distributed storage concepts. Such ideas apply across a wide range of technologies from Grid Computing [14] to Distributed Operating Systems [15]. The Internet provides a global transport mechanism and is therefore a perfect environment for deployment of distributed systems. Ongoing advances in communications technology can only assist.

Distributed Storage in Cloud Computing [16] provides storage transparently across multiple devices and sites. End users, even application developers, are unaware of where data is stored. The Cloud approach provides developers with access to large managed data storage operations requiring minimal effort. Ongoing concerns include trust, security and regulatory obligations [7].

Research relating to storage on the client side has focused on Local Storage [17], as introduced in HTML5. Local storage allows web applications to utilise basic key/value style storage, which is implemented within the users web browser. This represents an evolution from Cookies [18]; the focus on provision of enriched user interfaces with decreased bandwidth requirements. HTML5 continues to be an emerging technology, and criticism of its viability to solve the current issues with Web 2.0 are common [19].

Software engineering, operating system and Cloud Computing research combine to define the concept of a Web Operating System. This concept was originally introduced as a way of deploying an operating system that was capable of managing distributed resources [20]. As it evolved the focus moved to provision of remote online desktops [21] and bridging the gap between the Operating System and the Semantic Web [22]. The primary driving force in the marketplace is currently Google with their Chrome OS, a replacement for traditional operating systems. Chrome OS promotes access to web applications using the web browser itself as a large component of the operating system [23]. Increasing the scope of the web browser's involvement in application execution has been introduced in many forums. The Super Browser [24] focuses in this area and expands the design of the classic web browser beyond the requirements of Web 2.0.

Analysis of the above technologies has identified the need for a consolidated approach to storage across web applications. This becomes especially important when we expand the scope of web applications to support components executing on both the client side and the server side. This paper introduces a model that addresses this need.

V. Model: DDSv2 API

The model presented in Figures 3 through 7 addresses the identified problems of data ownership, freshness and duplication. This is accomplished by introducing additional technology that allows storage of data to be offloaded

from the web application. Instead, the storage is made the responsibility of 3rd party storage providers. These providers lease storage services to individual data owners, and act as a 'single version of the truth' provider for that piece of data. Data-owners can either be corporations, business groups, public entities or even individuals. Enhancements to the standard web browser design allows this data to be accessed seamlessly for integration into displayed web pages. An API is established to govern communication between the various actors in the model. These web browser enhancements are a stepping stone between existing browser technologies and a complete Super-Browser implementation [25].

The model definition can be broken into three main areas: storage, access and presentation.

A. Storage

The first phase of distributing data storage involves the data owner subscribing to a distributed data service (DDS). This service is responsible for storing the owner's data elements and is located either in-house or outsourced to a specialised data storage provider. The subscription procedure is shown in Figure 3.

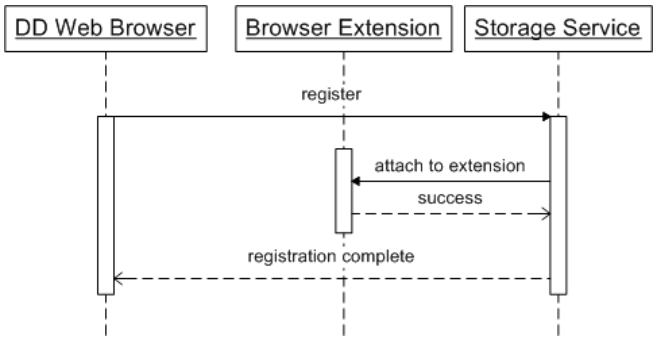

Figure 3. SS Registration

To aid integration of the DDS into the data owner's web experience a new module is introduced called the 'DDS browser extension'. This extension executes within the data owner's web browser, and is aware of the new distributed data service model. When a Web 2.0 site requests content from the data owner, he or she provides a link into the DDS in which the data is stored. Later, when the data is needed for display as part of a page generated by the site, the displaying browser instance uses the embedded link to directly retrieve the data from the owners's DDS. In current pilot implementations of the model, the module is implemented using the browser-extension technologies provided by most modern web browsers [26]. Future Super-Browser implementations will see the module as a distinct component executing within the web browser virtual machine [25]. Communication between the browser extension and the storage service is achieved by inserting a *DDS-StorageService-AttachRequest* into the HTTP response. This request is transparently inspected on-the-fly by the browser extension, which allows the extension to re-write parts of the response HTML dynamically. For the data-owner, the extension re-writes HTML form fields with links that activate a browsing interface for selecting data objects. For the end user, the extension re-writes links into a remote DDS with data returned from that specific DDS.

Implementation of the storage service registration process and the browser extension is implementation specific, though the API for linking the two components together is the first aspect defined by the DDS API. As long as implementations maintain the API for the DDS series of request/response messages, interoperability is maintained. The security of the connection between the browser extension and storage service is also implementation specific, but it is assumed that SSL encryption would be used at the tunnel level and basic authentication would be used to authenticate the end user to the storage service.

Phase two of the process, presented in Figure 4, involves the data owner publishing content to the storage service. Again, the implementation is not restricted by the API as long as each piece of stored data is given a unique identifier that is global in that data owner's domain. The unique identifier comprises three components:

[name]:[path]@[system]

The *name* component represents an identifier for each specific piece of data (for example, credit_card). The *path* component supports a hierarchical storage structure allowing a data owner wishing to store various groupings of data (for example, a data owner may have separate sets of 'personal' and 'business' data). The *system* component is a unique identifier for the specific distributed data service. It is generally a DNS name referencing the DDS itself. To reduce vendor lock-in it is recommended that data owners implement their own DNS pointers so that migration from one DDS to another does not result in the reissue of unique identifiers due to a change in the related system component. Using DNS for the system component also solves the issues of locating a specific entities DDS.

The data format proposed by this design is based on a published set of XML schema's that represent each type of data stored by the DDS. While enforcing data format appears restrictive, it is necessary to ensure that interoperability is promoted between web applications and storage services. Such generic interoperability is one of the main requirements for a solution that does not promote vendor lock-in.

B. Access

Once the data owner has uploaded data into their DDS, the next stage is to allow web applications to subscribe to this data. This is presented in Figure 5. During the registration process for a DDS-enabled web application, the web browser extension inspects the HTTP response traffic from

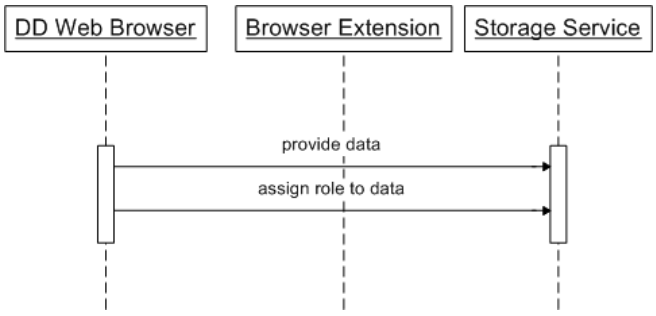

Figure 4. SS Information Upload

the application and detects a *DDS-Application-Subscribe-AuthRequest* request. This request, and the corresponding response generated by the browser extension, is the basis for establishing a trust relationship between the web application and the distributed data service. The request/response messages are built on extensions to SAML [27] and act as a basis for exchanging public-key cryptography credentials between the end user and the web application.

Once a relationship is established, as shown in Figure 6, the data owner establishes a link between their data element and the web application. This link is akin to the data owner uploading content to the web application in a standard Web 2.0 scenario, except that in the DDS design the data owner provides the unique identifier of the data as opposed to uploading the content itself.

The browser extension again plays a key role, ensuring that web applications can support both DDS-enabled and classic Web 2.0 clients. Additional DDS-enabled attributes are inserted into HTML *input* tags to inform the browser extension that the data owner's interface should be modified to accept a unique ID value rather than actual data. This linkage maybe optionally implemented by showing a pop-up window containing an index of the data stored in the DDS to allow the data owner to select individual pieces of data graphically, rather than manually entering the unique ID of the data.

Once the link is established between a piece of data referenced by the web application and the storage location for that data in a DDS, the web application is free to request the data directly from the DDS. This is useful in the circumstance where the web application is still required to store a subset of data locally in order to provide a service. For example, in the case of an image, the web application may request and store meta-data relating to the size of the image to assist in rendering pages during the presentation stage. This also opens up the possibility of web applications temporarily caching data, if permitted by the data owner policy as described in section VI.

C. Presentation

The final stage of the design is the presentation stage. This stage can execute in two modes. The original DDS v1 provided a single pass-through mode of operation, where the web application hands off responsibility for requesting user content to the web browser. The web browser receives a handoff request and communicates with the various DDSs directly.

An additional mode of operation has been introduced in DDS v2; it allows web applications to operate as clients of a DDS. This 'enriched' mode is useful when the web application is providing a value-added service that cannot be provisioned by the client.

Examples of where each of these modes of operation would be implemented are given in section VII.

1) Pass-through Presentation: The presentation stage in pass-through mode is depicted in Figure 7. Here we define how the data is transparently presented to end users. Again the browser extension plays a key role. In this instance the extension operates as a 3rd party and does not have any direct relationship to the DDS of the rendered data. For example, an end user may access a web application and request to view an image collage built from images stored in multiple DDSs owned by multiple data owners.

In this case the web application, instead of returning raw data, will return a *DDS-Present-DataRequest*. This call contains security information exchanged between the web application and DDS during the initial authentication request. This security information is protected using public key cryptography to ensure that it cannot be abused to falsify links between a DDS and unauthorised web applications and clients. The trust relationship enforced in this case is between the web application and the DDS, hence the DDS itself does not need to be aware of all the end users who can render the data linked to a specific web application. The *DDS-Present-DataRequest* message triggers a handoff of the user from the web application to the DDS, allowing the browser extension to request and render the data directly from the DDS, under the web application's instruction.

Under the DDS model, clients are required to perform additional processing to pass-through and cater for the DataRequest messages. Actual data transfer and rendering functions remain largely unaffected other than the fact that the web browser, on average, would be compiling single pages from multiple data sources. These data sources would be a combination of static data from the web application and dynamic data sourced from one or more DDS systems. Web page rendering engines in modern web browsers already support rendering a single page from multiple components so actual page rendering will appear identical to the end user when compared with current solutions. The rendering engine needs to be mindful of handling 'partial' data outages where a subset of DDS's are unavailable.

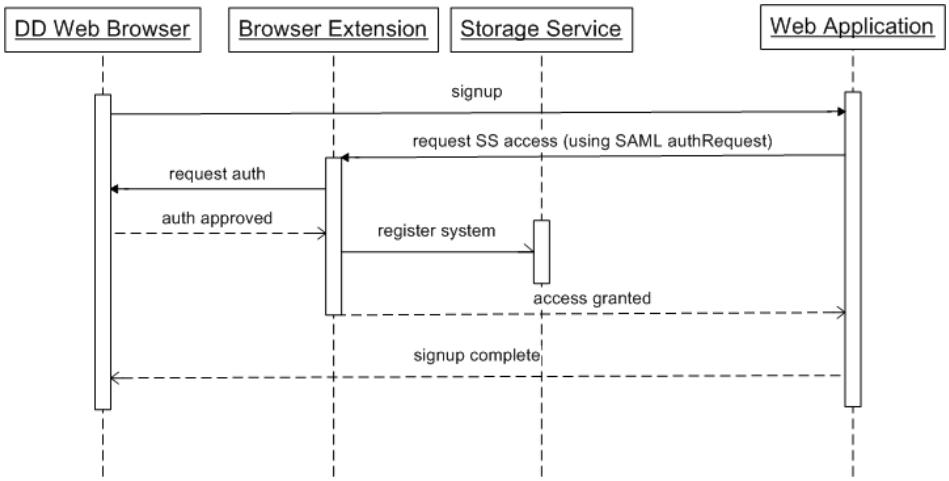

Figure 5. Web Application Registration

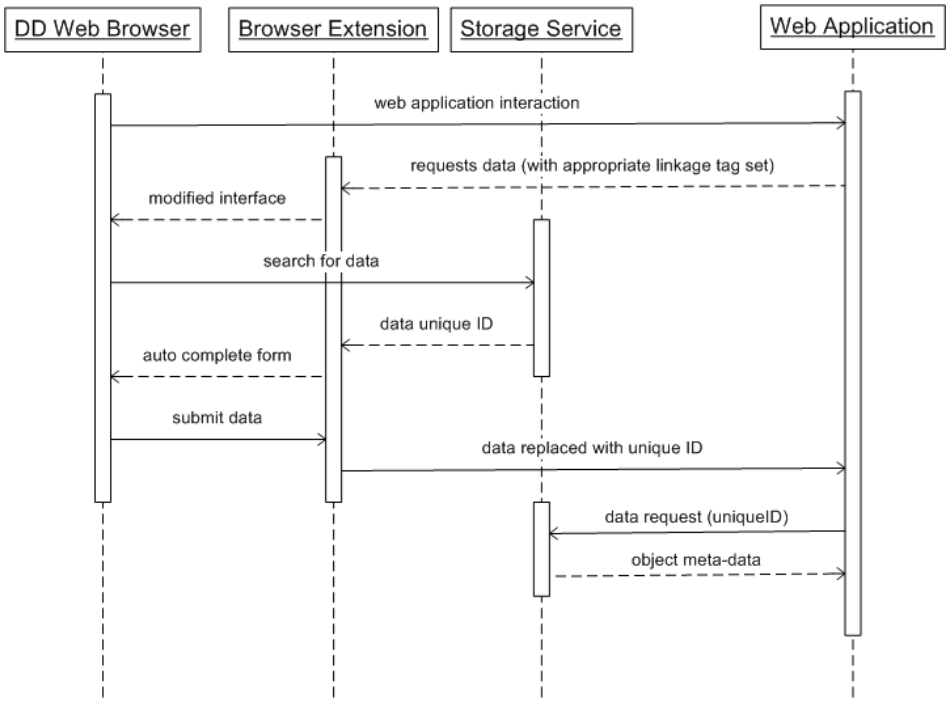

Figure 6. Data Linkage

2) Enriched Presentation: The presentation stage in enriched mode is depicted in Figure 8. A web application requests information directly from a DDS, with the intent to enrich the content before presenting it to the end user.

The same *DDS-Present-DataRequest* call is used as seen in pass-through mode. The request is passed directly from the web application to the DDS instead of being passed through the end users' web browser.

A benefit of this model is that end user web browsers are not required to perform any additional processing functions. With this in mind, enriched mode can be seen as a migration strategy or fall-back scenario. In cases where a web browser does not support the DDS API the content can be proxied through the web application using enriched mode.

VI. Policies and Security

One of the key enhancements to v2 of the DDS API is the introduction of policies. Policies are defined by both web applications and data owners. Both policies must be compatible for a data linkage to occur. Policies enforce rules that define what web applications can do with user data once they have successfully requested a link to that

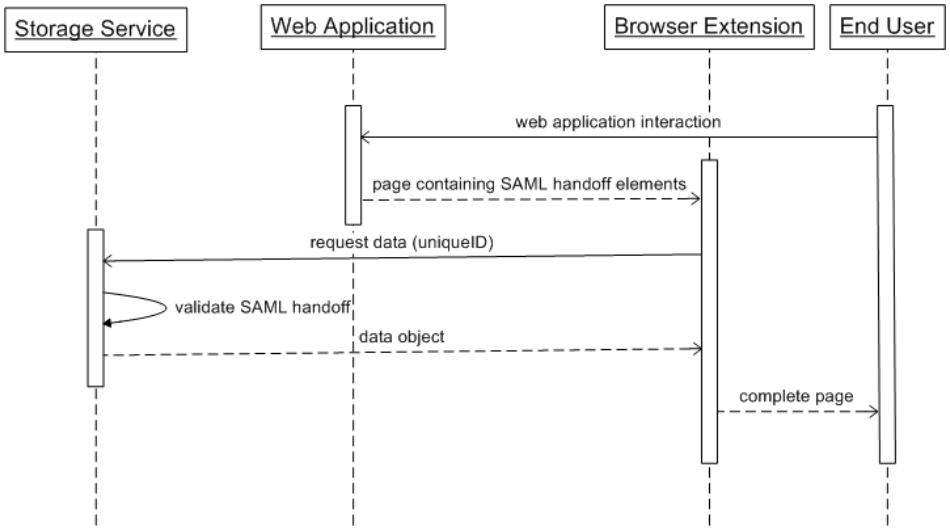

Figure 7. Data Presentation - Pass through

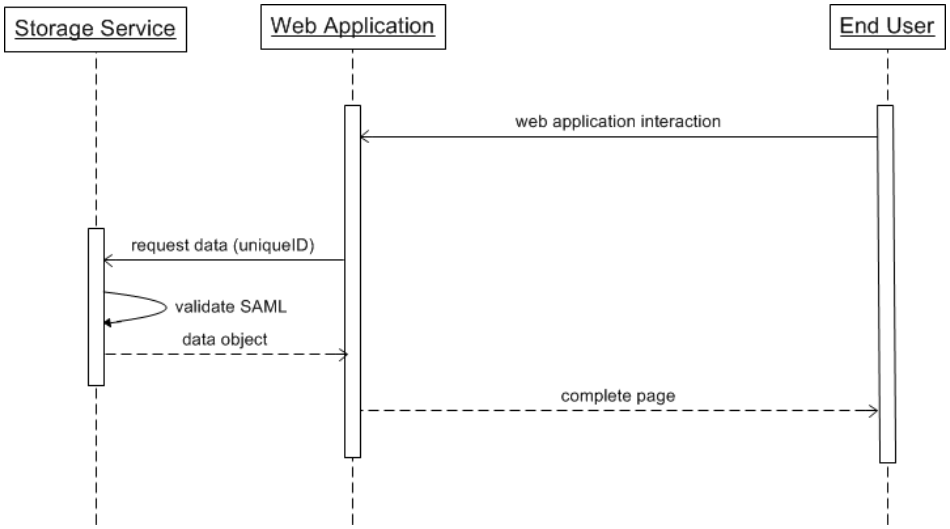

Figure 8. Data Presentation - Enriched

data. In pass-through mode the policy defines the operation required to allow end users' web browsers to obtain data after a successful handoff. The introduction of enriched mode requires web applications to state their intentions before directly accessing DDS data.

Web Application policies are presented to the data owner during the data linkage phase. They can range from a free-text description stating the applications intent to a strict XML schema stating common use-cases such as;

- Transparent pass-through only.
- Retrieve, manipulate, transfer.
- Retrieve and cache for a pre-determined period.
- Retrieve, store and transfer ownership.
- Forward to 3rd party.

If the free-text form of policy is attached to the data linkage request then the data-owner will be presented with the free-text to review before accepting the linkage. If the XML schema approach is used then the policy comparison can be processed automatically using a policy defined by the data-owner. During DDS registration the data owner can create their own policy stating which use-cases they would accept for the data they store in the DDS. This can then be compared automatically to the policy presented by the web application. If the policies fail to match then the data-owner will be given a choice whether to override the policy matching or deny the data linkage request.

Accepted policies form a contract between the data owner and web application. This can be viewed as a capability [28] in a distributed system model. If a web applications policy changes in the future all capabilities must be revoked or

multiple policy versions must be maintained.

These policies form a key foundation for limiting the scope of data freshness, data duplication and data ownership problems. They provide the means for Web Applications to support enhanced user control of the data, which is incorporated into their application. This is in contrast to the existing Web 2.0 model where the data owner is forced to agree to terms dictated by the web application owner, relinquishing all rights to their data. Using the XML schema approach a data owner can publish what policy requirements they enforce on their data in a parseable fashion. This allows for policies to be negotiated and automatically enforced without user interaction. A well defined schema of common use-cases has been defined as part of this API.

In cases where multiple web applications are providing the same service the end user can compare policies and decide which provider offers the most compatible set of requirements. In situations where a web application provider is operating as a monopoly the benefit of policies is centered on providing end users a clear definition how their data will be treated without having to manually read a complex EULA. The authors recognise that a web application may require a user to relinquish rights to their data, however the DDS v2 provides alternatives that previously did not exist.

If the data owner at any stage wishes to revoke access to their data then they can access the DDS and revoke that web applications linkage. This can be done for specific pieces of data or all data stored in the DDS. Is required, the DDS will then perform a server-to-server call to the web application informing it that the data linkage has been invalidated. This will ensure that the web application will not try and handoff a request to that data in future sessions.

VII. Prototype and Use-Case Experiments

A proof-of-concept prototype was developed to demonstrate all of the components described in section V. The prototype comprises the following:

- A skeleton distributed data service implemented in Java and utilising the Amazon S3 service for data storage.
- A proof-of-concept DDS-enabled web application, implemented in PHP, which is capable of subscribing to a DDS and linking image and postal address data.
- A Mozilla Firefox web browser extension implemented in Javascript to provide the data owner and end user experience.

Multiple programming languages were selected for the prototype to demonstrate the programming language-agnostic nature of the DDS API.

The prototype has been used to demonstrate three use cases. These use-cases are presented below.

A. Use Case 1 - basic pass through

The domain of this use-case is an in-house group address-book application with which users can create a profile and upload their office address and a profile photo. The address and photo data are stored in the data owners DDS. The following is an example runtime flow from the prototype application. It describes a user linking some data and a second user in turn rendering that data.

- User A (Bob) accesses the website for his DDS of choice and begins the registration process.
- DDS(Bob) sends an attachment request message (through the HTML response) that is detected by his web browser extension.
  - *DDS:DDS-StorageService-AttachRequest() to Extension(Bob)*
- Extension(Bob) requests Bob's approval to attach to the DDS and sends a successful response message.
  - *Extension(Bob):DDS-StorageService-AttachResponse(SUCCESS) to DDS(Bob)*
- Bob then continues to interact with the website presented by the DDS to upload his office address and profile image data.
- Bob now accesses the website for the address book application (WebApp) and begins the registration process.
- The Web Application sends an attach request to the browser extension in Bob's browser. The request is signed with the web application's private key and includes a copy of the web application's public key for verification.
  - *WebApp:DDS-Application-Subscribe-AuthRequest(publickey(WebApp),WebApp) to Extension(Bob)*
- The browser extension requests Bob's authorisation to allow that web application to subscribe to data within his DDS and sends back a response. The browser also forwards the request on the DDS so the DDS can locally register the request.
  - *Extension(Bob):DDS-Application-Subscribe-AuthResponse(SUCCESS) to WebApp*
- The web application then allows Bob to upload an image. Attached to the standard INPUT HTML element an additional DDS-Enabled="true" attribute is included. This instructs Extension(Bob) to render that input element as a DDS data-lookup field.
- Bob selects his profile image from the pop-up DDS interface and the browser extension provides the unique ID of the data ID(image) back to the web application for storage.
- User B (Alice) now accesses the website for the address book application and asks to view Bob's profile.
- The web application inserts a data request message into the HTML response that is received by Extension(Alice).
  - *WebApp:DDS-Present-DataRequest(publickey(WebApp),ID(image)) to Extension(Alice)*

- Alice's web browser extension then establishes a direct connection to Bob's DDS using the system code provided in ID(image) and forwards on the data request.
- DDS(Bob) authenticates the request by validating the signature of the message using the publickey(WebApp) established during the authentication request stage.
- DDS(Bob) then returns the image for rendering by Extension(Alice).

B. Use Case 2 - enhanced pass through

This use-case presents the model where data enrichment is required but the client-side is capable of providing the processing. A classic example of this is providing a map of a users address.

The DDS registration and web application phases are the same as presented in the first use case. The data linkage and presentation layers change as follows;

- The web application allows Bob to upload his address. Attached to the standard INPUT HTML element an additional DDS-Enabled="true" attribute is included. This instructs Extension(Bob) to render that input element as a DDS data-lookup field.
- Bob selects his address from the pop-up DDS interface and the browser extension provides the unique ID of the data ID(address) back to the web application for storage.
- User B (Alice) now accesses the website for the address book application and asks to view Bob's profile.
- The web application inserts a data request message into the HTML response that is received by Extension(Alice).
  - *WebApp:DDS-Present-DataRequest(publickey(WebApp), ID(address)) to Extension(Alice)*
- Alice's web browser extension then establishes a direct connection to Bob's DDS using the system code provided in ID(address) and forwards on the data request.
- DDS(Bob) authenticates the request by validating the signature of the message using the publickey(WebApp) established during the authentication request stage.
- DDS(Bob) then returns the address for rendering by Extension(Alice).
- The address is passed by Extension(Alice) to the web browser where it is received by client-side code such as a Javascript library
- The Javascript library takes the address information and passes it to a 3rd party web service, which returns a graphical representation of the address as a map.

In the above use-case it is expected that DDS(Bob) would have enforced policy on the request when it saw that the WebApp was requesting the linkage. The DDS would be aware that the data would be eventually on-forwarded to a 3rd party (the map generating service) by inspecting the requested policy. This would allow Bob to limit the scope-of-use of his address information.

C. Use Case 3 - server-side enrichment

This use-case presents an example where pass-through is not sufficient and the web application requires direct access to the end users' data. In this scenario the user stores their address information and the web application needs to directly access the DDS so that it can obtain content for a mailing label.

Again, the DDS registration and web application registration phases are the same as presented in the second use case. The data linkage phase and presentation phases change as follows;

- User A (Bob) returns to the website where he has previous registered and purchases an item.
- The web application requests that Bob provides his address. Attached to the standard INPUT HTML element an additional DDS-Enabled="true" attribute is included. This instructs Extension(Bob) to render that input element as a DDS data-lookup field.
- Bob selects his address from the pop-up DDS interface and the browser extension provides the unique ID of the data ID(address) back to the web application for storage.
- The web application processes the request and readies the item for shipment. The application then makes a request directly to DDS(Bob) requesting the address.
  - *WebApp:DDS-Present-DataRequest(publickey(WebApp), ID(address)) to DDS(Bob)*
- DDS(Bob) authenticates the request by validating the signature of the message using the publickey(WebApp) established during the authentication request stage.
- DDS(Bob) then returns the address for use by the web application.

In this case, there is no 3rd party and the web application itself requires access to the information stored in the DDS. The web application does not permanently store the address information as it may become stale, hence ID(address) is stored instead, and is used to re-request for future orders.

VIII. System Evaluation

A. Functional Requirements

The model presented in this paper sets out to solve multiple issues stemming from the traditional monolithic storage approach used by web applications on the Internet.

Distributing data element storage greatly reduces the resource requirements of web applications. Storage requirements will decrease to only those needed to store meta data on the web application itself rather than the user-generated content. Bandwidth requirements for the web application will drop as the application will only be returning basic

HTML, CSS, script and meta-content such as logos and branding. All user-generated content will be directly transferred to the end user from the related DDS systems. Lastly, the web application provider will no longer be required to meet varying privacy legislation requirements as they will not be directly storing any user's personal data. This requirement is instead offloaded to the DDS providers, which can operate in the same geographical region (and hence be subject to the same legislature) as the data owner.

From the end user perspective, the model addresses the data freshness issue by replacing the N multiple copies of a data element with N links that point to a single instance of the data stored in the data owner's DDS. These N links are abstracted using DNS technology to ensure that a user can migrate from one DDS to another without invalidating the links. This removes the danger of a system accessing obsolete versions of data by creating a single version of the truth for every data element in the system.

The model also addresses the data duplication-created storage wastage issue. This is true provided the number of bytes used to store a link is less than the number of bytes used to store the actual data. With this assumption we achieve a reduction of the storage requirements in a single system from (M * objects) to (N * objects) where N is the size of a link and M is the average size of stored objects. A web application would only be required to store the [system] component of the link once per user.

Most importantly, the issue of data ownership is also addressed. Use of owners' data was previously dictated by web applications, and was typically enforced by end user licensing 'agreements'. If a user wished to use a particular web application they had no choice but to accept the EULA. With the described DDS model, the user has more freedom. It is safe to assume that the DDS itself may also enforce a EULA on the end user, but in this situation the user has the buying power to procure services from another DDS provider that requests a less restrictive license.

Security of the DDS system is provided in multiple layers. All communication between the data owner and the DDS can be protected using such existing technologies as SSL. Basic authentication would suffice when the transport layer is protected. The link established by the data owner with the web application forms the basis of an authentication token, which is then used to authenticate end users through the web application into the data owners DDS. This handoff is protected using the SAML handoff framework. All security assertions as signed with the DDS validating the signature as belonging to the data owner. This allows the DDS to ensure that any request for data coming from an end user, through a specific web application, has been authorised by the data-owner.

B. Comparative Evaluation

While the authors could not find any other model specifically targeting the core issues of this paper, there are systems that are similar in nature to the DDS.

1) CMS: Parallels can be drawn between the DDS and Content Management Systems [29]. The DDS can be viewed as a personal CMS that allows its content to be seamlessly embedded into 3rd party web applications. The DDS provides distributed storage of user content for web applications that previously relied on monolithic storage repositories.

Current CMS solutions do not scale to the level required to implement an Internet-wide distributed storage solution due to their own reliance on monolithic storage structures.

2) CDN: Content Delivery Networks [30] are distributed storage networks that allow companies to host data objects on 3rd party networks. This allows them to take advantage of geo-location based load balancing and link peering to achieve reduced bandwidth costs. The typical CDN solution is similar to the delivery paradigm in the DDS model except that CDNs do not currently provide a seamless way for data owners to push content into the network and have that content transparently accessed by authorised web applications. CDNs are static in nature, and do not scale to the dynamic features that the DDS model provides.

3) Cloud SSP: The presented design ties directly into the realms of Cloud Computing [16], Service-Orientated architectures [31] and SAAS (Software-as-a-service) [32]. In a sense, a DDS can be seen as a SSP (Storage Service Provider) in a Storage-as-a-service [33] cloud component that allows other web applications to publish and subscribe to data within the Cloud. The DDS model described in this paper, however, provides the necessary additional access and presentation layers on-top of the storage to ensure that the user experience is seamless.

Cloud computing can play an important part in the design and hosting of the DDS storage system itself. As the DDS API does not explicitly define the internal design of the DDS, the vendor is free to, for example, use Cloud Computing technologies, this providing a DDS solution that benefits from the dynamic scalability and per-usage business models that the Cloud provides.

C. Performance

The paradigm shift described in this paper dictates a movement of data storage away from classic monolithic storage, towards a distributed network of data storage services. As such, performance has been analysed to identify overheads introduced by the additional access and presentation complexity. While a small constant overhead was identified due to the web application-to-DDS handoff requirements, performance increases were also identified in the following areas:

- Speed improvements under high-load situations due to the reduced data transfer requirements of web applica-

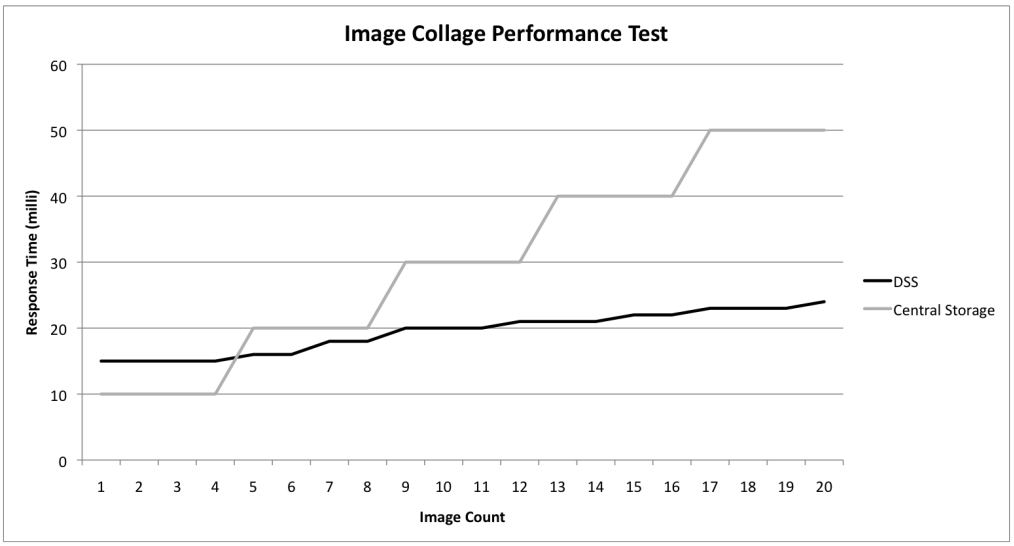

Figure 9. Performance results of the Collage Test showing pipeline enhancements

tions. Instead of the web application being responsible for provision of all the displayed data, the data transfer requirements are shared between the web application and the various linked distributed data services. This has the potential to reduce the network load of web applications.

- Client geo-locality can be utilised when users access data that is geographical in nature and when the required distributed data services are located closer to the client than to the web application. For example, a user browsing images of their friends on a social networking site would experience improved performance if the DDSs for their friends were less network hops away than the social networking web application itself.
- Speed improvements were identified in cases in where a single webpage is built of multiple separately loaded elements, a popular model in systems that rely heavily on user-generated content. In the general Web 2.0 case, browser pipelining restrictions limit the number of simultaneous network requests to any server. By distributing data storage the impact of these restrictions is reduced. Figure 9 shows the performance improvements for the case in which a single page is built from multiple data elements, each separately sourced. The experiment was performed with a pipelining restriction of four simultaneous connections per server. The results show a marked improvement using the DDS system.

Performance of the DDS system can be adversely affected by poor connectivity and bandwidth to specific DDS nodes. The open market for storage services will assist in driving competition between storage providers to reduce this risk. Obviously there are upfront costs involved in the integrating web applications with the DDS API, but these are offset by reduction in ongoing costs bandwidth and storage costs.

D. Backwards Compatibility and Migration Path

With the introduction of enriched mode a clear migration path has been presented for web applications taking on the new storage paradigm. A browser check can be performed during session initiation that informs the web browser if the end user is using a browser capable of the DDSv2 API. If not, enriched mode can be used to proxy the data to the end user. ”Pass-through” rights can be enforced via policy and users using older browsers will still see the benefits of this approach. Network limitations come into play in this scenario as data is proxied through the remote web application rather than being accessed directly by the end users’ web browser from the various distributed storage services.

IX. Scalability

The DDS system scales exceptionally well due to the decentralised nature of the data storage. Each user is free to choose their own DDS host(s). As the user base grows, the number of DDSs linked in a web application also grows. Each unique DDS can execute within a Cloud Computing environment, hence internal scalability is also supported in the situation where large numbers of data owners choose to use the same DDS (for example, all users of a particular University may choose to use a University-hosted DDS solution).

The DDS solution also scales into the corporate space where each corporate entity could host their own DDS. This would allow the employees of a company to share data internally, as well as externally through restricted publish/subscribe functions. The openness of the DDS API

allows corporate entities to protect their data by controlling which web applications subscribe to specific pieces of data.

From an end user perspective, the DDS system scales in the same fashion as a traditional client/server model. The transparent nature in the way the DDS browser extension provides visibility of DDS-stored information ensures that the end users' experience remains unaltered. From a connection viewpoint, the constant overhead described in the performance review above has no effect on the solutions ability to scale when compared to traditional approaches.

X. Conclusion

This paper addresses three concerns resulting from the growing popularity of Web 2.0 applications by formally defining a new paradigm for the distributed storage of data on the Internet. The standard for web applications has evolved, from static pages comprising a limited number of elements to complex pages rendered from a large numbers of elements. Web 2.0 has seen a trend towards bandwidth intensive elements originally generated by end users. As the user take-up of Web 2.0 applications continues, it is sensible to adopt a distributed approach that parallels the way content is originally generated.

Problems caused by the usage of monolithic data storage features have been mitigated by adopting a distributed storage approach. Moving from monolithic to distributed structures is a proven technique for sharing load that has been used extensively in other areas such as Cloud Computing [16] and Transaction Management [24].

The key issue of data ownership is addressed for end users by ensuring that storage is the responsibility of distributed data service(s) directly engaged by them. DDS providers are liable to data owners, not to web applications, and hence data owners have control over use of their data. Data ownership is clear-cut because owners are responsible for both storage of, and access to, the data.

Data freshness is addressed using a publish/subscribe model and an enhanced SAML-based handoff model for data presentation. The data rendered in web pages is always the freshest version because it is sourced directly from the data owner's DDS. Data duplication is also addressed by removing the need for data to be stored by web applications. Appropriate web application registration and linking reduces the number of copies of any piece of data to a single instance stored in the DDS.

Policies implemented during the web application and DDS handshaking ensure that end users are aware of how their data will be treated by that web application. The policies provide a level of protection which does not exist in the current Web 2.0 design. They provide an avenue to comparing multiple web service vendors from a data policy perspective.

Current modelling and experiments show that overall system performance is comparable to the existing Web 2.0 paradigm in the general case, with minor constant overhead caused by the handoff procedures. When a web application renders a page containing multiple data elements from multiple DDS repositories, we observe a performance improvement compared to existing technology due to the bypassing of web browser pipelining restrictions.

Comparisons made against similar systems show that the new paradigm can greatly increase the quality and protection of data in a Web 2.0 space. For the DDS model to become widely utilised the DDS API will need to be adopted as a standard.

References

[1] M. Wallis, F. A. Henskens, and M. R. Hannaford, "A distributed content storage model for web applications," in *The Second International Conference on Evolving Internet (INTERNET-2010)*, 2010.

[2] ——, "Publish/subscribe model for personal data on the internet," in *6th International Conference on Web Information Systems and Technologies (WEBIST-2010)*. INSTICC, April 2010.

[3] T. O'Reilly. (2005) What is web 2.0. [Online]. Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. (cited June 2010)

[4] WebDevout, "Web browser security statistics, http://www.webdevout.net/browser-security," *WebDevout*, November 2009. [Online]. Available: http://www.webdevout.net/browser-security

[5] C. Reis, A. Barth, and C. Pizano, "Browser security: lessons from google chrome," *Commun. ACM*, vol. 52, no. 8, pp. 45–49, 2009.

[6] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599 – 616, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6V06-4V47C7R-1/2/d339f420c2691994442c9198e00ac87e

[7] H. Newman, "Why cloud storage use could be limited in enterprises," *Enterprise Storage Forum*, 2009.

[8] M. Brandel, "The trouble with cloud: Vendor lock-in," *CIO.com*, 2009.

[9] M. Wallis, F. A. Henskens, and M. R. Hannaford, "Expanding the cloud: A component-based architecture to application deployment on the internet," in *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 2010.

[10] ——, "Component based runtime environment for internet applications," in *IADIS International Conference on Internet Technologies and Society (ITS 2010)*, 2010.

[11] AFP. (2009) About-facebook: backflip on data ownership changes. [Online]. Available: http://www.smh.com.au/articles/2009/02/19/1234632933247.html. (cited June 2010)

[12] T. Lee, "Data ownership might not work for social networking sites," *techdirt*, 2008. [Online]. Available: http://www.techdirt.com/articles/20080516/1124101137.shtml

[13] Object Management Group, *Common Object Request Broker Architecture: Core Specification*, March 2004.

[14] B. Jacob and et al, *Introduction to Grid Computing*. IBM Redbooks, 2005.

[15] A. S. Tanenbaum and R. V. Renesse, "Distributed operating systems," *ACM Comput. Surv.*, vol. 17, no. 4, pp. 419–470, 1985.

[16] G. Boss, P. Malladi, D. Quan, L. Legregni, and H. Hall. (2007, October) Cloud computing. [Online]. Available: http://download.boulder.ibm.com/ibmdl/pub/software/dw/wes/hipods/Cloud_computing_wp_final_8Oct.pdf. (cited June 2010)

[17] I. Hickson, *HTML 5 Editor's Draft*, w3c editor's draft 2011 ed., W3C, January 2011. [Online]. Available: http://dev.w3.org/html5/spec/Overview.html

[18] D. Raggett, A. L. Hors, and I. Jacobs, *HTML 4.01 Specification*, w3c recommendation december 1999 ed., W3C, December 1999. [Online]. Available: http://www.w3.org/TR/REC-html40/

[19] P. Krill, "W3c: Hold off on deployment html5 in websites," *InfoWorld*, 2010. [Online]. Available: http://www.infoworld.com/d/developer-world/w3c-hold-html5-in-websites-041

[20] A. Vahdat, P. Eastham, C. Yoshikawa, E. Belani, T. Anderson, and D. Culler, "Webos: Operating system services for wide area applications," in *Proceedings of the Seventh Symposium on High Performance Distributed Computing*, 1998. [Online]. Available: http://citeseer.ist.psu.edu/61096.html

[21] R. MacManus, "What is a webos?" *ZDNet Tech Update*, 2006. [Online]. Available: http://blogs.zdnet.com/web2explorer/?p=178

[22] L. Dignan, J. Perlow, and T. Steinert-Threlkeld, "From semantic web (3.0) to the webos (4.0)," *ZDNet Tech Update*, 2007. [Online]. Available: http://blogs.zdnet.com/BTL/?p=4499

[23] S. Pichai and L. Upson, "Introducing the google chrome os," *Google Blog*, 2009.

[24] F. A. Henskens and M. G. Ashton, "Graph-based optimistic transaction management," *Journal of Object Technology*, vol. 6, no. 6, pp. 131–148, July/August 2007.

[25] F. A. Henskens, "Web service transaction management," *International Conference on Software and Data Technologies (ICSOFT)*, July 2007.

[26] K. Feldt, *Programming Firefox: Building Rich Internet Applications with XUL (Paperback)*. O'Reilly Media, Inc, April 2007.

[27] OASIS, "Security assertion markup language (saml) v2.0 technical overview," Working Group, Tech. Rep., 2007.

[28] F. A. Henskens, J. Rosenberg, and J. L. Keedy, "A capability-based distributed shared memory," in *Proceedings of the 14th Australian Computer Science Conference*, 1991.

[29] A. Mauthe and P. Thomas, *Professional Content Management Systems: Handling Digital Media Assets*. Wiley, 2004.

[30] M. Hofmann, *Content Networking: Architecture, Protocols and Practice*. Morgan Kaufmann Publishers, 2005.

[31] M. Bell, *Introduction to Service-Oriented Modeling, Service-Oriented Modeling: Service Analysis, Design, and Architecture*. Wiley and Sons, 2008.

[32] K. Bennett, P. Layzell, D. Budgen, P. Brereton, L. Macaulay, and M. Munro, "Service-based software: the future for flexible software," in *Seventh Asia-Pacific Software Engineering Conference (APSEC'00)*, vol. 17th, 2000, p. 214.

[33] J. Foley, "How to get started with storage-as-a-service," *InformationWeek Business Technology Network*, 2009.

Management-aware Inter-Domain Routing for End-to-End Quality of Service

Mark Yampolskiy[1,2,4], Wolfgang Hommel[2,4], Vitalian A. Danciu[3,4], Martin G. Metzker[3,4], Matthias K. Hamm[4]
[1] *German Research Network (DFN), Alexanderplatz 1, 10170 Berlin, Germany*
[2] *Leibniz Supercomputing Centre (LRZ), Boltzmannstr. 1, 85748 Garching, Germany*
[3] *Ludwig-Maximilians-Universität Munich (LMU), Oettingenstr. 67, 80538 Munich, Germany*
[4] *Munich Network Management (MNM) Team, Oettingenstr. 67, 80538 Munich, Germany*
myy@dfn.de, hommel@lrz.de, danciu@nm.ifi.lmu.de, metzker@nm.ifi.lmu.de, hamm@mnm-team.org

***Abstract*—Services sensitive to network quality converge onto general-purpose data networks which, in contrast to special-purpose (e.g., public telephony) networks, lack built-in quality control functions needed by many applications, like Internet telephony or video conferencing. High-volume, high-performance applications such as those in Grid and Cloud computing may be too important for customers to rely on mere promises of network quality, while at the same time requiring connections traversing multiple network operators' domains. Thus, in addition to end-to-end QoS assurances, customers of these applications demand management functionality for those connections made available to them. Traditional routing procedures are insufficient to select paths according to these requirements, as they rely on evaluation of only one parameter (e.g., hop count), while QoS parameters alone will account for multiple independent metrics.**

We present a solution that addresses these issues by combining a routing procedure, a common set of QoS operations, and an information model for the representation of connection properties within and across administrative domains.

***Keywords*-end-to-end; quality of service (QoS); inter-domain routing; network management**

I. Introduction

Today's convergent or converged networks are intended to support a growing number of major services with highly varying requirements on the transport system. Such services include end-user facing services, such as voice and video telephony and their conferencing counterparts, but also high-capacity interconnects between the scientific sites of grid installations or the provisioning of cloud services, co-provided in different administrative domains, as well as connections between parts of virtual private networks (VPN links).

The inter-networking layer (i.e., the Internet protocols) does not support quality management inherently. Instead, many different Quality of Service (QoS) schemes have been implemented by different network operators. They do assure a stated quality of the *transport*, but typically omit customer-facing *management* capabilities. At the same time, the scope of network management is limited to single administrative domains.

Nevertheless, customers of important and expensive applications require network management facilities as part of the service being provided. Their capabilities range from read-only inspection functions (e.g., performance monitoring) to functions that alter the state of the network (e.g., adjustments of communication channel parameters).

Such capabilities are readily provided within single domains, but *inter-domain* communication channels (i.e., connections spanning multiple autonomous administrative domains) will require the co-operation of all domains in order to achieve *end-to-end quality guarantees* as well as an aggregate management function presented to the customer as part of the service. Such communication channels, provided as a service to a (paying) customer, we call *concatenated services* (CS).

A. Concatenated services

Combining the outlined demand and focus, our work specializes on the development of a solution for concatenated services, which are probably the most challenging type of point-to-point connections with respect to planning and operation. The following properties are characteristic for CS [24]:

- **User perspective:** a guarantee for the E2E quality of the connection and its management is required;
- **Service composition:** the E2E service is composed of horizontally (i.e., at the same network layer) concatenated connection segments, which are realized by different SPs;
- **Organizational relationships:** all SPs involved in the service's provisioning are independent organizations and are considered equal partners.

Due to high complexity of such connections, some scenarios exhibit unacceptable connection planning and establishment delays, especially when preparing a connection with non-trivial QoS requirements. In some cases, the planning phase, i.e., the identification of path segments that adhere to such requirements, may take up to several weeks (e.g., [36]). This is due to the lack of standardisation and automation of a planning process spanning multiple administrative domains. Each leg of the route must be negotiated with the owner/operator, including the accessibility of a suitable next hop and the QoS and management requirements for the connection. While this concerns long-term connections (i.e., of longer duration than the planning phase), it is obvious that such planning times cannot be always tolerated.

The algorithms underlying common routing protocols (link state, distance vector) are not applicable, as they rely on fulfilment of the optimality criterion. In contrast, the set

of routing metrics dictated by QoS parameter thresholds not necessarily holds this criterion, leading to undecidable choices. For example, assuming the requirements for a connection were minimum bandwidth b and maximum delay d with $(b, d) = (1Mbit, 100ms)$, the choice between two alternatives $(1Mbit, 50ms)$ and $(2Mbit, 75ms)$ cannot be made by "shortest" path semantics alone: the first alternative is better in terms of delay, while the second one is better in terms of bandwidth.

A standardisation solution may equally prove unfeasible: Allowing the requester of the connection (user, customer) to specify the kinds and values of the QoS parameters obviates the use of a linear projection function that might calculate "best" values from values of known, i.e., pre-defined, QoS parameters.

In addition, several connection topologies are conceivable. Even though the necessity for QoS assurance exists for different connection types, in the presented work we focus specifically on dedicated point-to-point connections. A discussion of application areas and aspects of different connection types, e.g., point-to-multipoint, is explicitly omitted. Furthermore, from the customer point of view as well as from the perspective of services built on top of network connections, the quality of the connection is important, but not the technology used for its realization. As the bridging between different network layers and technologies is very well understood and is broadly applied, e.g., in the Internet, we assume in our work that network connections are realized on the same ISO/OSI reference model network layer, and consider in our further discussion only the quality of connections and connection segments.

Finally, when devising a routing procedure for use across multiple independent carriers, acceptance of the procedure is crucial for its success: providers may reside in different legal domains, they may have different levels of interest in providing such services (depending on their business model, the state and load of their network, their management capabilities, etc.), and they may have different views on sharing the network management information required by the service (i.e., the managed connection) that is to be provided.

Consider the example sketched in Figure 1. SP_1's customer requires an end-to-end link between a start end-point within SP_1's domain and another end-point (target) within the domain of a different service provider SP_3, which is not a neighbour of SP_1. The customer specifies certain properties with regard to the quality of the transport (minimum bandwidth, maximum delay) as well as requirements on the management of the link during its operation phase (monitoring, constraints on maintenance windows). Via an agreed-upon customer service management (CSM) interface the customer formulates a request to SP_1 to aggregate and to deliver such an end-to-end link.

The customer's knowledge of the topology shown in the upper part of the figure is limited to the end-points in the domains of SP_1 and SP_3, respectively. Each service providers' knowledge is limited to its own domain, and

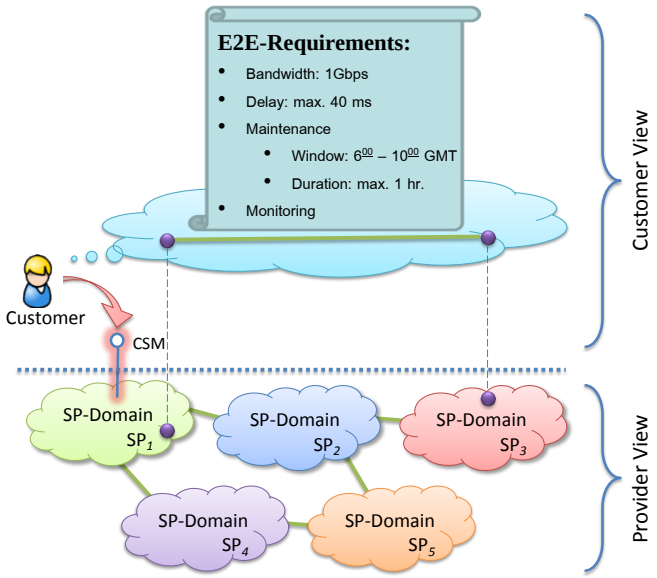

Figure 1. Example request for a concatenated service with QoS and management requirements

the identity of their neighbour SPs. Starting with only this information, the service provider SP_1 needs to find a route to the target end-point in SP_3's domain, such that the aggregate link adheres to the specification of bandwidth and delay and at the same time fulfils the management requirements specified by the customer. Thus, to deliver the link according to specification, SP_1 needs the cooperation of the other SPs in order to select the parts of the route (hops) to the remote end-point and to ascertain the quality and management properties of each part of the link. It is understood that during operation/use of the link each SP will be responsible for the parts of the link that pertains to its own domain. The monitoring has to reflect information from all participating domains in order to provide an end-to-end view to SP_1's customer, and all participating domains need to comply with the customer's requirement on maintenance windows.

In this work, we address the setup of such multi-part connections that involve multiple services providers, each of which is responsible only for parts of the overall network connection. In particular, we focus on the provisioning of high-quality, managed communication channels across multiple autonomous administrative domains; we specifically include operations support in our QoS considerations. The idea underlying this work is to establish transport quality and management properties at the time of routing.

B. Line switching

Experience gained with circuit and packet switched networks has provided deep insights into details about which technical and organizational measures are needed for quality assurance, and about which real-world challenges have to be overcome.

Circuit-switching technology, which is used, e.g., in *Public Switched Telephone Networks* (PSTN), has proven

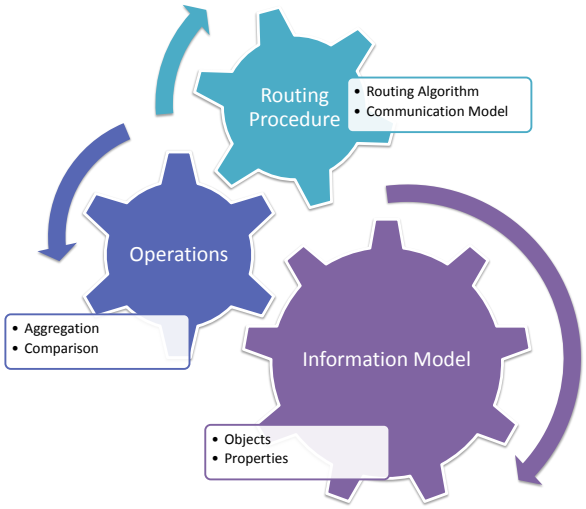

Figure 2. Three inter-operating conceptual building blocks

to be a viable solution for quality assurance, but at the same time it is not truly optimal regarding resource utilization. Packet-switched networks on the contrary have turned out to utilize the available resources much better, but at the same time the parallel communication flows can interfere with each other and consequently affect the quality of each other, for example due to overload packet drops.

Experience gathered in both types of networks types shows that the quality assurance first and foremost requires a thorough planning of the resources needed to meet the quality goals. Furthermore, the desired quality can only be achieved if the necessary amount of resources is allocated to each connection instance. Finally, resource reservation alone does not necessarily guarantee that the goals are met during the service instance operation. Regarding this aspect, current best practices suggest the establishment of management procedures for all life cycle phases of the service instances.

Based on this experience, we focus on concatenated services realized via *line switching* concepts, because this technique has proven to be a viable solution for quality assurance. As line switching can be emulated in a packet switched network, e.g., applying a combination of MPLS and RSVP in IP networks, our choice of the switching paradigm imposes no practical limitation regarding its application to the infrastructure used by real-world network operators.

C. Contribution

In this article, we explain the inter-operation of three conceptual building blocks that address the problem at hand, as sketched in Figure 2: a procedure relying on a management-aware inter-domain routing algorithm, which has been presented in [1], a dedicated information model (IM) for network connections and their QoS-specific properties. This IM, first described in [3], defines the structure and semantics of the information necessary to the routing algorithm. The last building block is a generic QoS function scheme, i.e., a collection of operations which is needed during the routing in order to operate on customer- and user-specific QoS parameters (cf. [2]). During the defined routing procedure, the connection parts are selected and the required quality of all parts is defined. Further, during routing the management functionality from all involved domains can be integrated into the management functionality of a service instance. As the later aspect of routing as well as the interplay of all solution building blocks has not been published so far, we discuss it in more detail.

We summarise the requirements on such an approach in the Section II and proceed to outline our approach in Section III. Thereafter, we describe each of the three building blocks in Sections V, VI, and VII, respectively. Subsequently, in Section VIII, we illustrate the inter-operation of the building blocks by an example before discussing the properties and limitations of the approach in Section IX. We conclude the article with ideas for further investigations in Section X.

II. Requirements

The challenges which have to be overcome in order to realize CS are manyfold. Partially they are caused by CS characteristic properties.

A. Fundamental assumptions

The case described in the Introduction implies that our approach is supported by a set of assumptions, as follows:

a) SPs are independent organisations; they are not obliged to participate in the provisioning of a service/connection.
b) A service instance (connection) cannot be realised by one SP alone.
c) Information about the network topology within an SP's domain is not available.
d) Reliable information about an SP's network capabilities is not available.
e) Management functions and management information are not offered by SPs without prior agreement.
f) Multiple independent QoS parameters will be specified for a connection according to the user/customer demands.
g) There is no fixed set of QoS parameters that will be requested by customers; and there is no fixed combination of parameters.
h) There is no fixed set of management operations that will be requested by customers.
i) Requirements on an existing service instance may change during its life-time.

An important challenge for our approach was to avoid a violation of any one of these assumptions. For this reason, the approach must fulfil the requirements formulated in the following.

B. Collection of requirements

The foremost requirements are those pertaining to the aim of our work:

1) The approach shall determine a path across multiple autonomous SPs' domains.
2) The approach shall ensure that QoS parameter thresholds are respected for the end-to-end service.
3) The approach shall ensure that management capabilities are provided for the end-to-end service.
4) Where necessary, management information shall be aggregated in a manner opaque to the customer.
5) Multiple independent QoS parameters shall be specifiable for a service instance.
6) Multiple independent management functions shall be specifiable for a service instance.

In addition, several requirements originate from the settings of concatenated services.

Participation of multiple SPs: Globalization of business and research collaborations has the consequence that the communication partners, which are using the same network connection, can be spread over the entire world. Due to economic and often also legal reasons, such connections usually cannot be realized by only a single *Service Provider* (SP). Similar to services in the Internet and in PSTN areas, multiple SPs have to be involved in the realization of a single network connection, leading to the following requirements:

7) No requirements shall be imposed on items within SPs' domains.
8) The approach shall not be dependent on the number of SPs that co-provide the service.
9) The approach shall not be dependent on global up-to-date information or on a central instance.

Customer-specific QoS-combinations: The quality of the customer-faced services in general depends on the different quality parameters of the underlying connections. Furthermore, in general it does not only depend on a single QoS parameter, but rather on the service-specific combination of multiple independent QoS parameters. For example, a video streaming service might be insensitive to delay and jitter as long as the connection bandwidth enables the pre-buffering of content that is not yet displayed to the user. However, for an internet-telephony service, the delay and jitter of the underlying connection might lead to a negative user experience. During video-telephony and -conferencing, also the synchronization between video and audio signals becomes important. In summary:

10) The customer shall have only one point of contact, her original SP.
11) The customer shall be given means for managing the service, once established.
12) The customer shall be allowed to specify QoS parameters and management capabilities.

Highly dynamic environment: The rapid development of new services as well as the very high dynamics of changes is characteristic for modern IT services. As the quality requirements on the underlying network connections might differ between services, also the high adaptivity and extensibility of solutions becomes a critical success factor. Especially the extensibility of support for new QoS parameters is needed, which have not been considered before:

13) The approach shall be extensible regarding the QoS parameters supported.
14) The approach shall be extensible regarding the management capabilities provided to customers.
15) Service planning shall be automatable.

III. Approach Outline

We argue that user- and customer-tailored requirements for connection services can be only truly fulfilled, when already considered during the ordering process. Our approach is a routing process, consisting of three conceptual parts: a *routing procedure*, a *generic function scheme* and an *information model* (see section I-C). A routing procedure satisfying the requirements we laid out in Section II takes into account all known existing connections with their properties and end-to-end requirements when finding a path between two end-points.

By employing line switching to realize concatenated services, our algorithm may regard the path through an SP's domain as an atomic link, defined by its ingress and egress points. Based on this decision, our algorithm does not require knowledge of every SP's network topology. This reduces the problem-space and leaves it entirely to the SP to realize the link. As a result, our information model may describe links, properties and functions in a technology agnostic way. Abstracting from components and technologies, it is easier to model inter-domain links, where the end-points are in domains of different SPs.

Connection requirements directly affect the tasks of the routing procedure, which in the case of line switching are: the selection of the path between the two end-points, the designation of all connection segments and interconnecting all points along the chosen path [35]. *QoS routing* (also referred to as *QoS-aware routing*) extends this definition by taking into account end-to-end user requirements. QoS requirements are defined for all connection parts and have to be guaranteed by all SPs, in order to meet the given end-to-end goals.

As discussed in the requirements section, for true quality assurance the management of service instances by the customer has to be considered. Therefore in our work we introduce *management-aware routing* as a QoS-aware routing with additional tasks, namely:

- the definition of management functionality, which have to be provided for each involved connection segment and
- the integration of specified functionalities into multi-domain management functionality of the whole service instance.

As management processes are specified as an interaction between roles with specific responsibilities, the assignment of roles to the involved SPs must be considered as an integral part of the overall management-aware routing procedure. The mentioned routing types and their tasks are depicted in Figure 3.

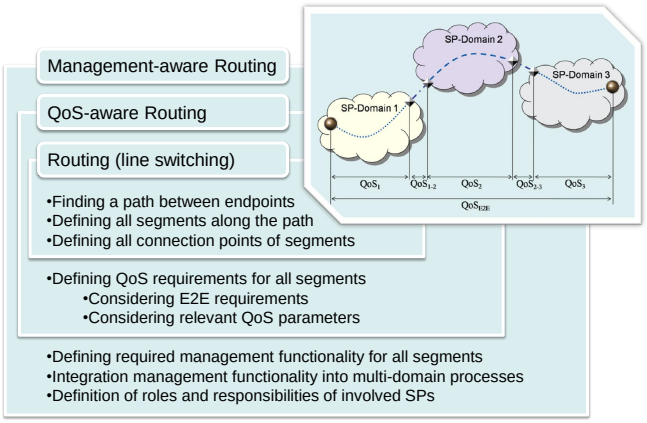

Figure 3. Routing types and corresponding tasks

In our approach the algorithm is designed to split the planning and evaluation of (sub-)paths into multiple smaller tasks, which we see as the key to having a generic set of functions. By breaking down these two problems into smaller parts, our algorithm can be seen as a framework where each metric may provide its own functions and rule sets. Our goal is to describe this generic approach, not to provide an explicit formulation of QoS-requirements and metrics. An example of how our concept may be instantiated will be given in Section VIII.

A. Routing procedure

The routing procedure consists of a *routing algorithm*, which is broadly seen as the very core of every routing approach, and *communication patterns* in the form of a communication model. The purpose of a routing algorithm is to choose between different available connection segments and select those, which will form a path between two endpoints. The details of our algorithm can be found in [1]. In this paper our focus lies on the interaction between components of proposed solution.

Our routing procedure is dedicated to find and establish a path in advance, with an explicit planning phase, before user data can be sent.

As our approach shall not require SPs to share their entire topology knowledge with each other, there needs to be interaction between SPs so that viable paths can be found, requested and managed. Communication patterns describe how routing instances interact with each other and when information is to be requested from another SP, as we aim to keep information exchange to a minimum. Communication relationships also describes how management functions of involved connection segments will be connected and consolidated in order to be offered to the customer.

Basically, we distinguish between *source routing* and *routing by delegation*. While source routing in its extreme form means that the entire routing is performed by the customer's service provider, routing by delegation allows an SP to ask another SP to find a (sub-) path for the remaining part of the route. Through the combination of these two techniques the routing procedure can adapt to policy constraints, concerning the information exchange of SPs.

B. Operations on properties

A generic function scheme is required to specify and evaluate boundaries and optimality criteria for different kinds of quality requirements. The example in Section I-A includes *bandwidth* and *maintenance window* as quality requirements on the requested connection. Bandwidth is limited by the capacities of involved physical links. The link with the *lowest* capacity determines the *maximum* bandwidth achievable on a path. Requirements on maintenance windows are less straight forward than on bandwidth. In the case of maintenance windows *upper and lower boundaries* for maintenance planning as well as the duration of maintenance work are specified. Clearly, an algorithm needs a different set of operations to reason on maintenance windows in contrast to bandwidth. The generic function scheme provides a common interface, so that specialized operations for each parameter may be accessed by the algorithm in a unified manner.

Thus the main goal of our function scheme is to enable the algorithm to take all possible kinds of metrics into account. The most important part of this task is defining the semantics of our proposed generic functions to ensure compatibility of metrics and interoperability between communication partners. Another part of this task is providing a method for representing QoS-requirements in our information model.

C. Information model

Existing information models (IM) often neglect the importance of inter-domain connections. Even though inter-domain connections are comparatively small parts of the overall path, their quality is indispensable for the computation of the overall end-to-end quality. Due to very restrictive SP policies, each SP usually has access only to its end of an inter-domain connections. There are cases in which it is possible to derive the properties of inter-domain links from two partial views onto the same physical link, like available resources on that link. For such cases our IM provides the possibility to describe partial views of different SPs and means to correlate these views.

We designed our information model to describe link properties similarly to connection requirements, which allows us to use the same data model for both. This eases interaction between peer entities as a conversion from requirements to resources is not required, thus leaving less room for misunderstandings and different interpretations. Also, to uniform information exchange, our information model includes management functions so that they can be treated like any other QoS-parameter in the routing procedure.

IV. Related Work

Our presented work covers many aspects in order to provide end-to-end links with user-defined quality-requirements and management-functions. Even though most of the problems outlined in Section II have been

addressed in previous work, none of them covers the topic in a full extend.

Originally, physical line switching technologies were used in PSTN networks. In ATM, logical line switching in combination with resource reservation has been implemented. As these technologies were tailored to fit the needs of the voice-only telephony, they do not offer the required flexibility of novel services provided over, for example, the Internet. In this area of (virtual) circuit switching, the Dynamic Circuit Network (DCN) cooperation is probably the most advanced research project. The main focus of this project is on the dynamic provisioning of dedicated paths, literally guaranteeing bandwidth to customers [4], [5]. Among others, projects like OSCARS [6], DRAGON [7], Phosphorus [8], and the Géant-developed AutoBAHN [9] are involved in this cooperation and several successful demonstrations have been presented to the research community [10]. Another approach to on-demand circuit provisioning is the DICE (DANTE–Internet2–CANARIE–ESnet) [11] collaboration of European and North American research networks, where an architectural concept, based on the experience gained in previous projects has already been elaborated and published [12].

The mentioned projects and collaborations are focused primarily on two aspects:

1) various techniques and technologies for dynamic circuit switching and resource reservation within a single administrative domain and

2) interoperability between the developed management systems as well as between networking technologies used in these domains in order to automatically switch multi-domain network connections.

Plans for the consideration of QoS parameters are limited to sole connection propert – its bandwidth. Management aspects in dynamically provisioned circuits are limited to circuit monitoring, mostly reusing praxis approved concepts of the Géant E2E Link Monitoring System (E2Emon) [13]. DICE plans to extend this monitoring concept with a combination of measurements at different network layers and with different monitoring techniques.

An advanced example of network connections with customer-tailored properties is the Géant E2E Links service (also referred to as Géant Lambda) [16], [17]. Among other scenarios, Géant E2E Links have been used to realize challenging connections for the international research project Large Hadron Collider (LHC) [14] and for the Grid cooperation DEISA [15]. These links were established considering multiple QoS parameters as well as management aspects, like inter-domain trouble-shooting procedures or the coordination of maintenance windows among multiple SPs. The biggest drawback of this approach is very long connection establishment time, as these links are planned and set up manually. As planning of Géant E2E Links might consider potential possible connections for which infrastructure still have to be procured, installed, and configured, estimating the time necessary to find and set up a path between two endpoints is very difficult. In the case of Géant E2E Links, the time between circuit ordering and service production start – influenced by different factors – can vary between a few weeks and several months [36].

Management of end-to-end multi-domain network connections is mostly limited to monitoring in the form of technology specific solutions, e.g., the well-established monitoring in SDH networks [21]. Another, still emerging technique is OAM (Operations, Administration, and Management) for Ethernet [22], which aims at a wider scope than merely end-to-end monitoring. For example, failure signaling has already become part of OAM for Ethernet, which allows for quicker detection of and reaction to component failures along the entire path. The more generic multi-domain solutions, like the E2E Monitoring System (E2Emon) for Géant E2E Links, are currently limited to a project-specific combination of technology-agnostic QoS parameters [13].

The routing is mostly covered in the graph theory. In a graph that models a network as vertices and edges, QoS-parameters on path segments are usually represented as edge-weights. Established routing algorithms are based on Bellman's optimality principle and find paths between two vertices. Bellman's optimality principle is valid for the class of functional equations of finding the maximum, the minimum or the k-th element [25]. As this class requires (at least) a partial ordering of the domain, these algorithms are not applicable for multi-dimensional search-spaces and especially multi-weighted graphs. The different weights on a single edge cannot always be merged to a single one. The example in Section I-A includes requirements on bandwidth and maintenance windows. As bandwidth requirements cannot be converted to maintenance window requirements and vice versa, a partial ordering over both weights cannot be found and Bellman's optimality principle cannot be applied here.

Proposed algorithms that are capable of searching high-dimensional spaces, e.g., SAMCRA [27] or H_MCOP [28], require full information about network topology, link properties and connection properties. This make them inapplicable for multi-domain routing, as it collides with the restrictive information policies of most SPs and maximises search-space.

Routing algorithms operate on information about available connections or connectivity possibilities. One of the biggest problems of information models (IM) for network description is data-hiding. For instance, the Common Information Model (CIM) [31] focuses on the description of relations between services and underlying network technologies. While the CIM is a very good basis for managing single networks it cannot be used for our purposes. Modelling a connection segment with QoS-parameters requires modelling of the entire network the segment resides in. In order to comply with the strict information policies in multi-domain environments, networks need to be modelled abstractly with hardly any information. This cannot be accomplished using the CIM and other information models.

Furthermore, almost none of the IMs used for network descriptions associates multiple abstract properties with connections. One noticeable exception is the ITU-T recommendation G.805 for the description of optical transport networks [30], where multiple technical connection properties can be associated with a single connection. In the recommendation these are only technical parameters that are needed to interconnect the segments, e.g., multiplexing of different channels. Especially this recommendation does not consider connection's quality and management functionality properties.

V. Inter-Domain Management-aware Routing

The main task of routing is always path selection. As mentioned at the beginning of this paper, in order to guarantee end-to-end requirements, the required properties of all segments have to be planned. This includes considering the management functionality needed in further phases of the service instance's life cycle. Our proposal for routing and defining of quality targets for the chosen segments is presented in Section V-A. As has already been pointed out, an integral part of *management-aware routing* is the integration of management functionality provided by involved domains into integrated overall management functionality for the whole service instance. We analyse these integration aspects in Section V-B.

A. Connection planning

As the routing procedure operates on knowledge about available potential connection segments, we introduce our information representation to the extent we deem needed to understand our routing proposal. An elaborated discussion of our information model and an elaboration of our decisions will be given in Section VII.

The routing procedure we propose operates on semi-global knowledge about available connections. All this information is represented at the abstraction level of an SP's organizational domain. This means that we look only at connections between network equipment installed at administrative edges of provider networks. In general, all connections which can be realized either within a single SP's domain or interconnecting two neighboring SPs are potential *connection segments* of an end-to-end connection. As from an SP's point of view every realized connection segment is a provided service, we refer to an endpoint of a segment as *Service Connection Point* (SCP). We choose this abstraction in order to be independent of network technologies and of application areas in which our routing concept can be applied. A SCP can be mapped to various "edge-components", from a logical UNI/NNI interface when looking at paths through the Internet, to a Point of Presence (POP) when planning backbone connections.

Semi-global knowledge means that for routing we need information over multiple, but not necessarily all, SP-domains. This knowledge can be extended on demand, by requesting additional information from other SPs. Consequently, during routing we distinguish between already known connection segments and existing but not (yet) known connection segments. The same applies to SCPs of those segments. Further, some of the known connection segments can be considered by the routing algorithm as a part of the planned route.

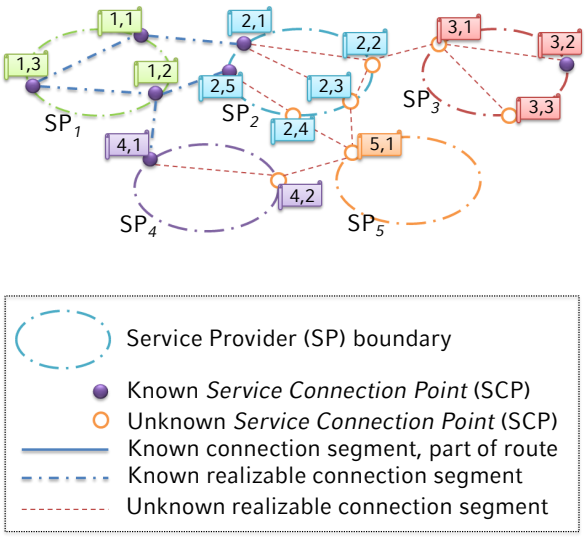

Figure 4. Semi-global knowledge about available connection segments at the beginning of routing [1]

Corresponding to the example presented at the begin of this paper, Figure 4 presents a possible initial semi-global knowledge of SP_1. In this figure, SCPs are labeled with IDs. These IDs consist of two numbers, where the first number identifies the SP to which a SCP belongs and the second number is a sequential numbering of SCPs within an SP. Please note that we have chosen this ID representation only for the sake of better readability and reference in this paper.

Due to various factors, e.g., steady ordering of new and decommissioning of existing end-to-end connections, the resources available to possible end-to-end connections changes constantly. The consequence is that an SP cannot expect to have up-to-date knowledge about all available connection segments. At the same time every SP always has exact knowledge about its own network equipment and available capacities. We further assume that every SP can obtain information about capacities available to interconnections with neighboring SPs. We see this as a common situation for most SPs, as such information is required for proper management of the own network infrastructure. Further, we assume that every SP can determine the quality properties that can be guaranteed over all available own links.

For instance, in Figure 4 SP_1 is in charge of routing between SCPs with IDs "1,3" and "3,2", it can choose which connection segments should be considered as a part of the path. When selecting these segments SP_1 relies on knowledge about available capacities and own preferences regarding which neighbouring SP to use. At the same time SP_1 has to consider customer requirements, i.e., for all considered paths the realizable connection properties must

be computed and compared with the user-provided end-to-end requirements. If, at least, one of the end-to-end requirements cannot be fulfilled, another (for the SP lesser preferable) alternative must be considered in the same way.

In the case when SP_1 selects the way through SCPs "1,1" and "2,1", it also implies that the next segment is realized by SP_2. In [1] we discuss advantages and drawbacks of different routing strategies. Based on this discussion, we recommend the *source routing* strategy as long as possible. This means that, in this particular case, despite the fact that the next segment is realized by SP_2, the selection of this segment for the end-to-end connection should be done by SP_1. This in turn means that SP_1 has to obtain actual information from SP_2, containing the "remote" SCPs through which SP_2 is willing to realize connection segments and which properties can be guaranteed for these segments.

We see two main advantages in using source routing strategies: 1) avoiding (or reduction of) nested communication relationships speeds up the overall communication processes and decision propagation, and 2) retrieved information about available connection segments and their properties can be reused, if alternatives have to be examined. Nevertheless, employing source routing requires high trust relationships between communicating SPs. Especially in big open provider collaborations, e.g., the Internet or the PSTN-network, this is not always the case. To meet this problem we propose *on-demand delegation* of the routing task for the remaining part of a path. This is based on the praxis approved theory, that SPs are more likely to trust topologically closer SPs and especially all neighbouring SPs.

Similar to a CSM interface (see Section I), which is commonly referred to as the management interface between provider and customer, we introduce the *Domain Service Management* (DSM) interface as a means for communication between independent SPs interested in a collaborative realization of end-to-end network connections. Communication between SPs should always be performed through the DSM interface.

As mentioned in Section III, obtaining information about SPs' available capacities, as well as making decisions about segments from outside an SP will most probably violate domain policies. In order to avoid or at least minimize violations of these policies, we propose to specify always strict boundaries for requests. When requesting a segment, the requesting SP provides information about the remaining part of the end-to-end path, i.e., the two determining SCPs between which the path still has to be established, and about the connection properties the end-to-end path has to comply to. Such procedures allow queried SPs to provide only few alternatives matching the specified restrictions, as opposed to providing full and unreflected information about an SPs infrastructure. Furthermore, we propose that the queried SPs provide these alternatives in the most-to-least preferable order. Even though misuse cannot be fully avoided, we propose that the SP responsible for routing (in the example SP_1) tries the alternatives in the specified order.

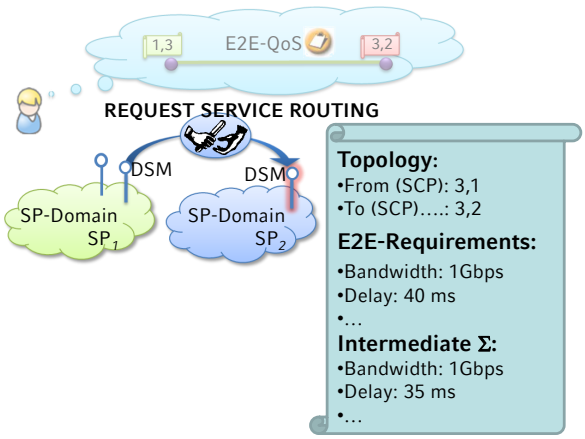

Figure 5. On-demand delegation of the routing task [1]

In our example, if SP_3 refuses to provide information about available connection segments, SP_1 can delegate the routing task to SP_2 (see Figure 5). In this case topology restrictions, customer specific end-to-end requirements as well as properties realizable by the found partial path are passed to SP_2. Advantages and disadvantages of *routing-by-delegation* are opposite to those of source routing: information about SPs' available capacities are hidden and the selection of the most preferable route is guaranteed. On the other hand, communication becomes to be nested and information about previously checked alternate routes cannot be re-used, which might lead to redundant checks.

Our proposed routing procedure incorporates not only the principles for the selection of the next connection segment, but also communication between SPs needed for information requests and delegation of the routing task. These advanced considerations make it impossible to describe our procedure as a pure pseudo code alone. Instead we define a state-machine (see Figure 6), showing planning-phases as states and outcomes as transitions. Beginning with an intermediate SCP that is at the end of the path considered so far, the next connection segment is determined. If required information is missing, a planner enters an information-retrieval phase (A2) where the missing information is requested from a SP. After calculating the properties of the entire path, including the new next segment, these properties are evaluated against end-to-end requirements. If at least one of specified requirements is violated, an alternative segment has to be considered. If requirements are fulfilled, the distant SCP of the new segment is considered as a new intermediate SCP and the planner returns to the starting state (A1). We need to support the case that an information request is rejected. In this case, the routing task is delegated by a *FINDROUTE*-request to the last SP in the considered path (A5). Our procedure terminates, i.e., reaches an accepted end-state, either when the desired endpoint is reached or no more alternative connection segments are available.

As it is very common for the description of an algorithm to discuss its runtime analysis, please note that the main

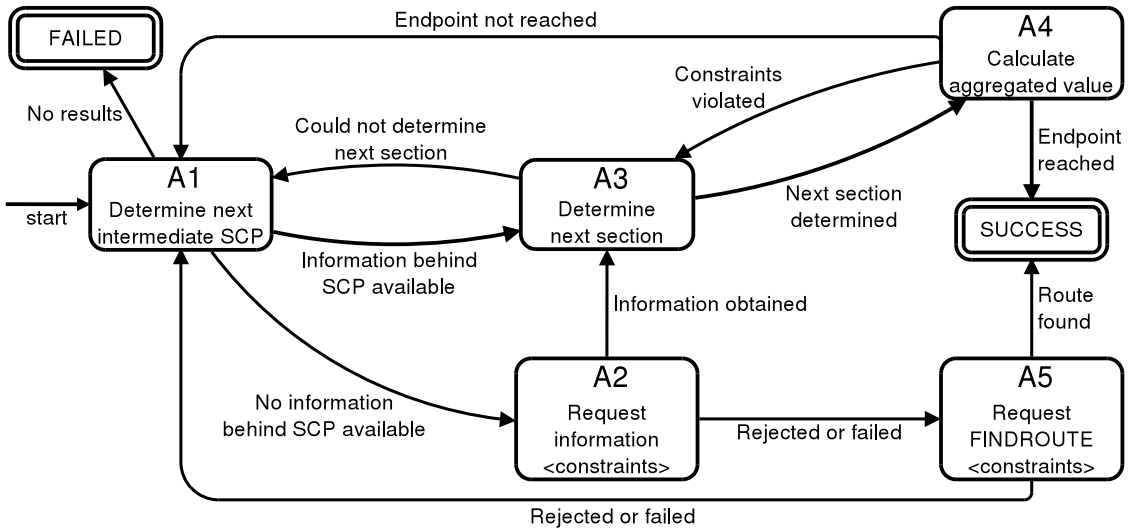

Figure 6. Routing processes

purpose of any algorithm is to achieve its goals. Therefore the quality of an algorithm should not only be evaluated through the performance of an implementation, but primarily through the quality and the properties of the yielded result. In order to provide such an evaluation of our proposal, we point at the order in which connection segments are considered for a new route. Regardless whether source routing or routing by delegation is used, consideration of available segments is identical to the Depth First Search procedure. In this procedure, alternatives are considered in an SP-specific order of preference. A criterion for considering an alternative is not fulfillment of at least one specified end-to-end requirement. Consequently, the proposed procedure is nothing else but an inter-domain policy-based routing procedure considering end-to-end requirements.

B. Tying management functionality together

Experience shows that quality guarantees are only possible to hold if the planned quality targets are leveraged by management procedures. The management of a service instance is in turn only possible if the management of all its integral parts is possible. Therefore two of the central purposes of management-aware routing are the definition of management functionality of each involved connection segment and its integration into the overall management of new multi-domain service instances.

In a multi-domain environment, direct access to the network infrastructure of any SP is in general not possible, as this is very likely to violate provider policies. Therefore the integration of management functionality means negotiation of *communication interfaces* and *responsibility areas* as well as *communication relationships* with all involved SPs. The possibility to negotiate communication relationships and responsibility areas for each connection should influence the acceptance of our approach positively, as we give SPs a framework through which they can control information- and command-flows. Negotiating DSM interfaces is an essential part of integrating management functionality and should be performed during the negotiation of other relevant connection segment properties like QoS parameters and management functionality.

Very important for tying together management functionality is the possibility to specify one or more DSM addresses by information requests. Representation of information by inter-domain communication will be discussed in Section VII. DSM address is presented in class COMMUNICATIONDSM (see Figure 11). This enables SPs to specify multiple communication interfaces for different purposes, e.g., regular information requests, monitoring or event-triggered notifications. Concerning the confirmation of requests, SPs providing connection segments should also specify communication interfaces for their management functionalities. Furthermore, this will guarantee flexibility to SPs to separate interfaces for service instance negotiation and operation, as well as to specify different interfaces for different customers and/or service instances.

Every SP providing a connection segment as a part of overall end-to-end service, is responsible to ensure the agreed quality targets are met and to provide management functionality through the negotiated interfaces. For the integration of management functionality we define that each SP is in charge, i.e., responsible, for the area for which it has performed the routing task. To illustrate this we refer to the example in Figure 4, outlined before. In the example the final path is going through the SCPs with IDs ”1,3”, ”1,1”, ”2,1”, ”2,2”, ”3,1”, and ”3,2”. Since routing has been delegated to SP_2, the example has multiple responsibility areas, depicted in Figure 7.

We propose to distinguish between two types of responsibility delegation. Following the first option, which we call *FullProxy*, in our example SP_2 would build an abstraction layer, hiding the entire remaining part of the path from SP_1. This means that both, build-up and tying of management functionality, is hidden by the proxy-SP and SP_1 has no direct control over them. The main advantage

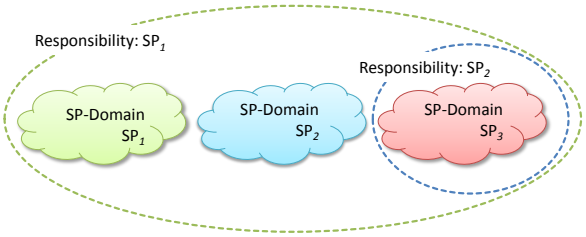

Figure 7. Responsibilities for function integration in example

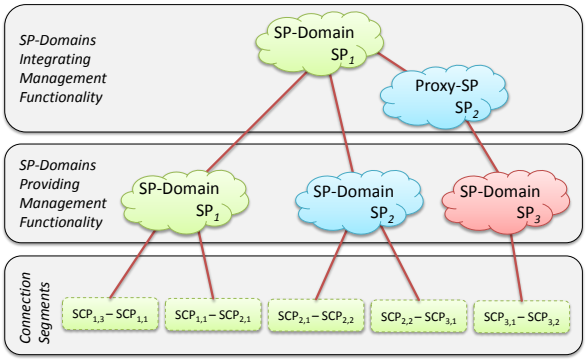

Figure 8. Communication relationships by *FullProxy* option

of this option is the best possible compliance with restrictive SP policies. Therefore we see this option (similar to the applicability of the *routing-by-delegation*) as an always feasible one. The disadvantages of this option are illustrated in Figure 8. The most apparent disadvantage of this option is the increase of communication hops between SPs, due to nested communication relationships, which will lead to increased reaction times. Another significant disadvantage is the increase of intermediate processing by proxy-SPs. Finally we want to point out the inevitable deviation of reaction time by connection parts with direct connection of management functionality, e.g., negotiated using *source routing*, and connection parts with nested integration of management functionality, e.g., negotiated using *routing-by-delegation* with *FullProxy* option. Such deviation not only decreases the overall end-to-end performance, but also might require special treatment with respect to the realization of such a solution.

Therefore we see the necessity to introduce a second delegation option, which we call *TransparentProxy*. Using this option it is possible to signal the proxy-SP that the communication interfaces for management functionality specified by the requester should be re-used during reservation and ordering of the delegated part of the path. The biggest advantage of this option is the possibility to keep communication relationships as flat as possible. As SP-policies might restrict acceptable communication relationships, e.g., only to neighboring SPs, this option may not always be feasible. Therefore, we see the *FullProxy* option as a fallback solution, similar to *routing-by-delegation* being a fall-back solution for the case when *source routing* is not applicable.

VI. Generic QoS-Operations

During the routing two operations are needed: 1) the values of connection segments considered as a part of the route should be aggregated; and 2) the aggregated value should be compared with the E2E requirements. All these operations have to be performed on the properties (and property combinations), which are relevant for the particular customer request. In Section VII we will show, how various connection properties (and their combinations) can be described in a similar fashioned way. Unfortunately, the similar fashioned description alone is not enough for the similar fashioned operations on those properties. The reason is the various semantics of the values related to various connection properties. For example, operations needed for aggregation and comparison of *delay* values are not applicable for other very important QoS parameter – *bandwidth*. The same can be said about operations on other properties, e.g., on *maintenance window*.

In order to deal with this problem, in [2] we have defined a generic function schema, which enables treatment of various connection properties and their combinations in a similar fashioned way. The basic idea is that all parameters are distinguished bases on global unique IDs. Further, all IDs for every supported qualitative and quantitative QoS parameter as well as for management functionality and its property should be defined in a registration tree. This opens possibilities to specify for every property ID a set of functions needed for operations on it (see Figure 9). The functions _AGGREGATE_LINKS and _ORDER_COMPARE are used to aggregate and compare values of a single property. These functions operate on properties of connection segments, and are therefore relevant for the routing procedure described in Section V. The reasons for the remaining three functions are slightly more complex. Due to restrictive SP policies, access to the network equipment is generally impossible from outside of one's own SP. Function _AGGREGATE_LINKPARTS is needed for computation of properties by segments interconnecting two neighbor SPs (so called *Inter-Domain Links*). This function is needed because in general no direct access to the network infrastructure of the neighbor SP is possible, and consequently both SPs have only own (partial) knowledge about properties, which can be guaranteed by a particular Inter-Domain Link. The remaining two functions (_SELECT_BEST and _SELECT_WORST) are needed for operation on value ranges, which can be specified by SP as realizable connection property. Value ranges can be used instead of specifying multiple connection segments between same two SCPs with associated fixed values. For a more elaborated discussion about the needs and application areas of these functions please see [2] and [3].

The association of these functions with property IDs enables the similar fashioned operations on various properties. This means that during the routing for every relevant property the corresponding functions can be obtained based on property ID. Subsequently, these functions can be used to operate on the values of the properties, which

have been specified for connection segments.

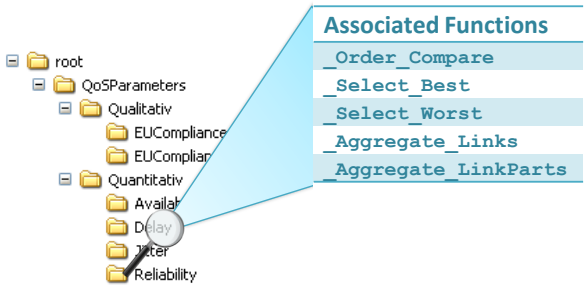

Figure 9. Association of functions with a property ID in the registration tree [2]

The described functions are always associated with a single property. In order to operate on the customer specific property combinations, we derive operations on multiple properties based on single-property operations. The aggregation of multiple properties is straight forward and is defined as a combination of per-property aggregations. The comparisons of property combinations is, however, more complicated. The reason is the possibility that by comparison of two value-combinations some of the properties in the first and some other of the second combination are better. This situation is reflected in the definition for comparison of value combinations (see Figure 10). The "$\prec$" symbol is used here to denote which value is better.

$$\overrightarrow{Compare}(\vec{U},\vec{V}) ::= \begin{cases} =, \text{ if } & \forall 1 \le i \le m : u_i = v_i \\ \prec, \text{ if } & \forall 1 \le i \le m : (u_i \prec v_i \\ & \quad \vee\ u_i = v_i) \wedge \\ & \exists 1 \le j \le m : u_j \prec v_j \\ \succ, \text{ if } & \forall 1 \le i \le m : (u_i \succ v_i \\ & \quad \vee\ u_i = v_i) \wedge \\ & \exists 1 \le j \le m : u_j \succ v_j \\ \neq, \text{ if } & \exists 1 \le i \le m : u_i \prec v_i \wedge \\ & \exists 1 \le j \le m : u_j \succ v_j \end{cases}$$

Figure 10. Comparison of property-combinations [2]

Even though in the presented paper these functions have been motivated by their necessity during the routing, their application area is much broader. For instance, they can be used for the monitoring of already established connections and the comparison of monitored values against designated targets.

The definition of similar fashioned operations on connection properties has several advantages. Among the most important is the easiness of extensibility for the support of new connection properties. Furthermore, by extending property support no changes or adaptation in implemented routing algorithm will be needed. And finally, if a global registration tree is used among all SPs for property IDs and function definition, operations on property values and their combinations will be identically among all SPs.

VII. Information Model

The basis for every routing is the knowledge about available connection segments and their properties. As described in Section V, during the routing missing information has to be requested from the other SPs. In order to comply with SP's information policies, the requested information has to be restricted strictly to the needed information. Figure 11 describes how the requested properties of the segment can be specified. In order to support user-specific E2E requirements, various combinations of relevant properties can be used within the RequestedProperties class. We distinguish between tree kinds of properties: Qualitative QoS parameters, quantitative QoS parameters, and management functionality (reflected correspondingly in classes QualitativeQoS, QuantitativeQoS, ManagementFunctionality). The distinction between various properties is made based on globally-unique IDs. This is a basis for a similar-fashioned treatment of various parameters, as it is described in Section VI. With management functionality class combination of properties (class Property) can be associated. This class is used for specification of different management function specific properties, e.g., polling interval for monitoring. Similarly to QoS parameters, the distinction between properties is made based on its IDs. Finally, values (class AssociatedValue) can be associated with quantitative QoS parameters and with properties of management functionality.

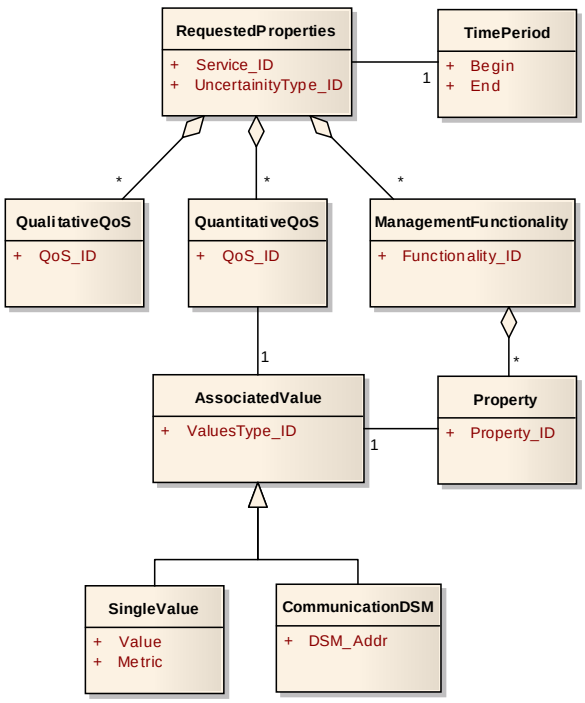

Figure 11. Specifying properties by service instance requests

If SP which can provide connection segments accept information request, it has to provide own up-to-date information. In our work we assume that every SP has exact knowledge about its own network equipment and

available capacity. We further assume that every SP can obtain information about capacity available for interconnection with the neighboring SPs. We see this as a very realistically assumption for most SPs, as such information – among other – is required for proper management of the own network infrastructure. Further we assume that every SP can determine the quality properties which can be guaranteed over available own connections.

The description of information is significantly more complex then information request, as it also should describe various topological properties (see Figure 12). As has been discussed at the beginning of this article, an expectation that SPs will share the information needed for management of the own infrastructure (i.e., detailed network topology representation as well as capacity and usage of particular physical network connections) can be seen as a very unrealistic one. One of the most critical aspects in a multi-domain environment is very restrictive information policies. The omission of the collision with those policies is the main reason, why the proposed routing procedure operates on information at the abstraction level of an SP's organizational domain. Such a representation hides most realization aspects, like network topology, which strongly complies with the mentioned information policies.

For the description of available connections three classes are used: COMPOUNDLINK, COMPONENTLINK, and COMPONENTLINKPART. The class COMPOUNDLINK is used as a wrapper for multiple alternative connection segments connecting the same two SCPs. Such a situation can occur, for instance, due to alternative physical connections available between these SCPs. Even though such details are hidden from SP's outsiders, they can result in connection segments with different properties. The class COMPONENTLINK, which represents the connection segments, is associated with the class LINKPROPERTIES, which specify segment's properties. The structuring of connection segment properties is identical to the one of information request (see Figure 11). Finally, the class COMPONENTLINKPART represents the SP's view at an Inter-Domain Link. As discussed in Section VI, every SP has only information about quality guarantees realizable at the SP-facing side of an inter-domain connection. Therefore, the whole quality of such connections has to be calculated from two views on it of involved SPs. For this purpose function _AGGREGATE_LINKPARTS has to be associated with connection properties (see Section VI).

Every of these three connection classes interconnect two SCPs, which contain their globally-unique IDs needed for segment stitching as described above. Further, every SCP is associated with a SP Domain in which it is located. The reason is the necessity of SP responsible for routing to communicate with those SPs, which can provide connection segments. During the routing, considering a new connection segment as a part of the route implies that the next connection segment should be adjacent to the distant SCP of the considered route. Information about such an adjacent segment can only be obtained from an SP owning the mentioned SCP. In close collaborations of very limited amount of SPs it can be assumed that every SP knows the interface to communicate with every other SP. In open and/or highly dynamic collaborations this is not the case. We however assume that every SP can maintain up-to-date information about communication interfaces of all its neighbors. Therefore by specifying the domain owning the SCP, a communication interface DSM_ADDR of this SP should always be provided.

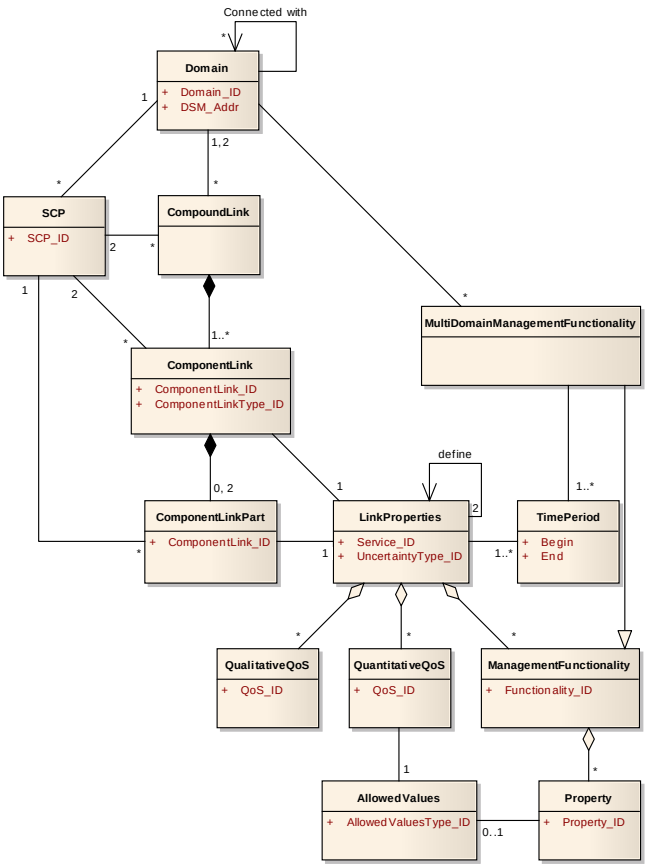

Figure 12. IM for available connections and properties of a single SP [3]

Finally, the MULTIDOMAINMANAGEMENTFUNCTIONALITY class has to be mentioned. This class is derived from MANAGEMENTFUNCTIONALITY, which enables the specification of multi-domain management functionality in a similar fashioned way as it is done by the management functionality of connection segments. This description is needed by the delegation of multi-domain management functionality, as it will be described in Section IX-A.

The inter-domain routing procedure, which we have described in Section V, operates on a semi-global knowledge, i.e., the knowledge spanning over multiple – but not necessarily all – SP domains. At the same time, each SP is only able to provide information about connection segments in which realization it is involved. This raises the question, how such single-domain views of different SPs can be combined to semi-global multi-domain view?

If every SP can provide its view at the connection

segments it is involved in, deriving a multi-domain view from a multiple single-domain views can be realized relatively easily. We propose that every SP, when specifying segments, always provides the IDs of two SCPs at its ends. It is necessary that these IDs are globally unique. Only then the SP responsible for routing can ”stitch” these segments at the SCPs with identical IDs together (see Figure 13).

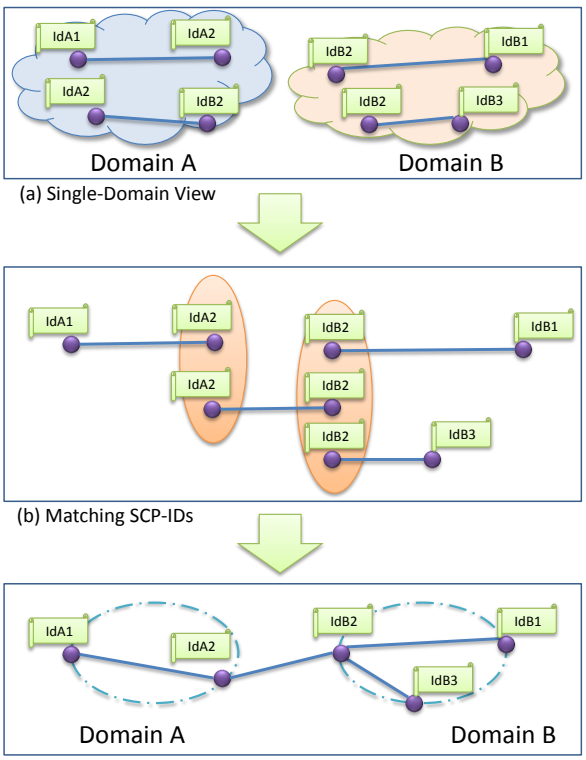

Figure 13. Deriving a multi-domain view from single-domain views [3]

A more elaborated discussion about details of information model as well as about reasons for the selected representation can be found in [3].

VIII. Putting it all together

In in Sections V, VI, and VII we have described the three building blocks of our proposal. In order to illustrate how elaborated concepts can be applied, we reuse the example outlined at the beginning of this article. We recall that in the example the customer has contacted SP_1 and has requested a connection between SCPs with IDs ”1,3” and ”3,2” with following properties:

- Bandwidth: 1 Gbps
- Delay: max 40 ms
- Maintenance
  - Window: 6^{00} – 10^{00} GMT
  - Duration: max 1 hr.
- Monitoring

From the customer’s point of view, SP_1 is the sole provider of the E2E connection with the mentioned properties. For the planning of a corresponding multi-domain path SP_1 needs routing functionality. Furthermore, in order to comply with the outlined parameters also multi-domain *Monitoring* functionality is needed. As SP_1 can provide both functionalities, it can start the process of finding the route fulfilling all end-to-end customer requirements.

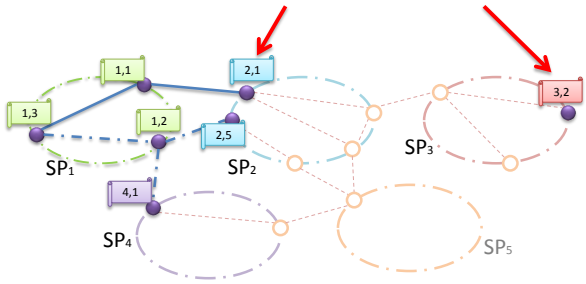

Figure 14. Global view at the available services

Figure 14 presents the situation after SP_1 chose the path through its own network. It still has to find the route between SCPs ”2,1” and ”3,2”, but has no up-to-date knowledge about connections available in SP_2. Consequently a request for information should be send to SP_2. As proposed in Section V-A, this information request should be accompanied by the topology- and property-restriction. We recall that the main motivation of restricting requested information was the necessity to comply with very restrictive information policies. An additional positive effect is the significant reduction of connection segments, which have to be considered for the route. This in turn speeds up the overall routing process.

```
<RequestedProperties Service_ID="CS" UncertainityType_ID="Guaranteed">
  <TimePeriod Begin="Now" End="OpenEnd"/>

  <QuantitativeQoS QoS_ID="Bandwidth">
    <AssociatedValue ValueType_ID="SingleValue" Value="1" Metric="Gbps"/>
  </QuantitativeQoS>

  <QuantitativeQoS QoS_ID="Delay">
    <AssociatedValue ValueType_ID="SingleValue" Value="140" Metric="ms"/>
  </QuantitativeQoS>

  <ManagementFunctionality Funcitonality_ID="Maintenance">
    <Property Property_ID="Window">
      <AssociatedValue ValueType_ID="ValueRange" Min="0300" Max="0700" Metric="GMT"/>
    </Property>
    <Property Property_ID="Duration">
      <AssociatedValue ValueType_ID="SingleValue" Value="1" Metric="hr"/>
    </Property>
  </ManagementFunctionality>

  <ManagementFunctionality Funcitonality_ID="Monitoring">
    <Property Property_ID="CommunicationDSM">
      <CommunicationDSM DSM_Addr="dsm.domainSP1.com/mon/instance932/"/>
    </Property>
  </ManagementFunctionality>

</RequestedProperties>
```

Figure 15. Specifying needed properties in information requests

In Section VII we have presented how requested properties can be structured (see Figure 11). For communication between SPs, this structure has to be encoded. Currently, application-layer protocols are often built on top of SOAP communication. As the information in such communication is encoded in a human-readable XML format, this suits very well for illustration of concepts. Figure 15 presents how the example specific information restrictions can be represented in XML according to the proposed structure. Please note that the proposed property structure can be encoded differently.

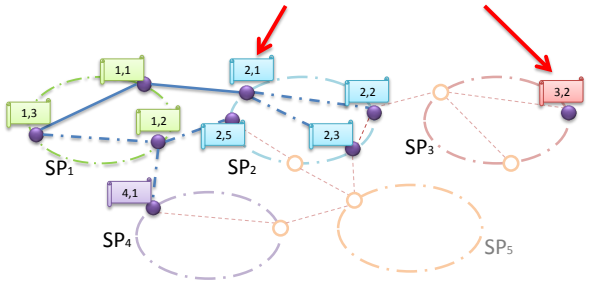

Figure 16. Global view at the available services

The structuring of the requested information has been described in Section VII (see especially Figure 13). By information request SP_1 also specify its DSM address for monitoring functionality. This is a crucial aspect in order to tie functionality together, as it has been described in Section V-B. The representation in XML format can be done similarly to the presented request. Please note that regarding the answer, SP_2 should provide own DSM-address(es) through which the monitoring functionality of available connection segments can be reached during instance operation. These addresses might vary between defferent service instances due to various SP-internal reasons.

From	To	Route-Part?	Property	Value
$SCP_{1,3}$	$SCP_{1,1}$	✓	Bandwidth	1 Gbps
			Delay	19 ms
			Maintenance Window	06:00-10:00 GMT
			Maintenance Duration	1 hr
$SCP_{1,3}$	$SCP_{1,2}$	-	Bandwidth	1 Gbps
			Delay	15 ms
			Maintenance Window	06:00-10:00 GMT
			Maintenance Duration	1 hr
$SCP_{1,1}$	$SCP_{2,1}$	✓	Bandwidth	1 Gbps
			Delay	4 ms
			Maintenance Window	06:00-10:00 GMT
			Maintenance Duration	1 hr
$SCP_{1,2}$	$SCP_{2,5}$	-	Bandwidth	1 Gbps
			Delay	3 ms
			Maintenance Window	07:00-09:00 GMT
			Maintenance Duration	1 hr
$SCP_{1,2}$	$SCP_{4,1}$	-	Bandwidth	1 Gbps
			Delay	6 ms
			Maintenance Window	06:00-07:00 GMT
			Maintenance Duration	1 hr
$SCP_{2,1}$	$SCP_{2,3}$	-	Bandwidth	1 Gbps
			Delay	20 ms
			Maintenance Window	06:00-08:00 GMT
			Maintenance Duration	40 min
$SCP_{2,1}$	$SCP_{2,2}$	-	Bandwidth	1 Gbps
			Delay	10 ms
			Maintenance Window	07:00-09:00 GMT
			Maintenance Duration	1 hr

Table I
KNOWLEDGE TABLE OF SP_1

According to the routing process specified in Figure 6, up to now we have selected SCP ”2,1” as the next intermediate SCP (action A1), have recognized that some information is missing, and therefore requested this information (action A2). Now we have to determine, which connection segment should be considered as next section in the path (action A3). Figure 16 depicts graphically the knowledge of SP_1 after SP_2 has provided the requested information. Please note that the information about irrelevant connection segments has not been provided at all. In order to illustrate the follow-up actions of the defined routing procedure, in Table I we present the knowledge base available to SP_1 at this point of procedure. It contains imaginary values associated with properties of available segments as well as the flag whether the particular segment is considered as a part of the route or not. Please note that for SP_2 there is no need to provide information which is not fulfilling the restriction.

According to the information provided by SP_2, two connection segments are available. The most preferable one is the connection between SCPs ”2,1” and ”2,3”. According to the defined routing procedure, the SP responsible for routing has to consider available connection possibilities in the most-to-less preferable order of the SP providing them. Therefore SP_1 has first to consider this connection possibility and therefore calculate the aggregated value of the partial route (action A4 in Figure 6).

In order to calculate the path properties, aggregation functions for the relevant properties have to be accessed first. As described in Section VI, these functions are associated with the property IDs. These functions can be used for semantic-aware operations on the values of particular connection properties. This means that aggregation function is:

- *min* for the Bandwidth
- *addition* for Delay
- *intersection* for Maintenance window
- *max* for Maintenance duration
- Logical *AND* for availability of Monitoring

Using these functions and values from the knowledge table we can calculate the properties of the intermediate path going through SCPs ”1,3”, ”1,1”, ”2,1”, and ”2,3” (Action A4 in Figure 6). Now we have to compare the calculated multi-property value with the end-to-end customer requirements. This is done again based on the functions defined for the particular properties (see Section V). The result of the comparison will be $\neq$, as the all properties except delay are better than the requirements, but the requirement for delay cannot be fulfilled by this path. Consequently, the next alternative connection segment, e.g., between SCPs ”2,1” and ”2,2” has to be considered according. As the path going through SCPs ”1,3”, ”1,1”, ”2,1”, and ”2,2” hold the requirements, its last SCP ”2,2” is chosen as next intermediate SCP (action A1) and the outlined procedure can be repeated.

Let us assume that with the above outline procedure the path fulfilling all end-to-end requirements and going through SCPs ”1,3”, ”1,1”, ”2,1”, ”2,2”, and ”3,1” have been found (see Figure 17). This means that the following segment(s) can be only provided by SP_3. As the information about available connection segments is missing, SP_1 has to request this information from SP_3 (action A2 in Figure 6). Let us further assume that SP_3 rejects this information request due to some internal reasons. For this case our proposal foresees a delegation (action A5) of a

routing task to the last SP in the already found path, i.e., in this case to SP_2. Graphically this situation is depicted in the Figure 5.

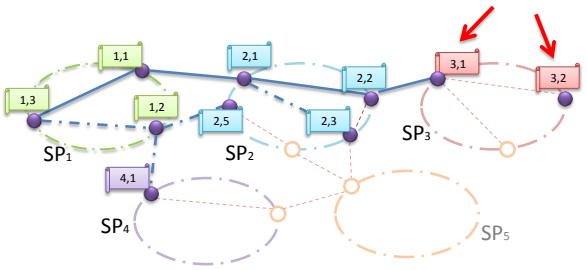

Figure 17. Global view at the available services

If SP_2 accepts the request for routing delegation, it takes over the routing for the remaining part. It receives information about end-to-end requirements, values which can be guaranteed by the intermediate route and two SCPs between which it has to perform routing. SP_2 choses SCP "3,1" as its intermediate SCP (action A1) and proceeds as outlined before. If the end point is reached, SP_2 reports the properties of the whole found path to the requester of the routing task.

Please note that in order to tie management functionality together, SP_2 has to specify to SP_3 the communication DSM (see Section V-B). In the case of FULLPROXY delegation, SP_2 specifies its own address. In this case, if the route could be found, it also has to report its own DSM to SP_1. The provider SP_2 should further aggregate monitoring information of all following segments in its own responsibility area. By this delegation option, SP_2 provides SP_1 exactly with the aggregated view.

Regarding the delegation by using the TRANSPARENT-PROXY option, during the routing in its own responsibility area, SP_2 should specify for monitoring purposes the DSM address of SP_1. Also, concerning the acceptance of SP_3 to provide the last segment, SP_2 should report the DSM of SP_3's monitoring instance to SP_1. In this case SP_1 is only responsible for routing but not for aggregation of monitoring information.

IX. DISCUSSION

In this section we will outline our thoughts about aspects going beyond the scope of the presented paper, but highly related to the presented solution. At first we will discuss the delegation of management tasks to the trusted third party SPs. Then in Section IX-B we will present our thoughts about reservation and ordering of planned route. Section IX-C is dedicated to outline our view at the acceptance of the presented solution by SPs. Finally, in Section IX-D we will present positioning of our solution among other existing alternatives.

A. Delegating Multi-Domain Management Tasks

The discussion up to now has implicitly presumed that every SP-domain approached by the customer as well as every proxy-SP can take over the whole functionality, like routing and integration of needed management functionality. The practicability of such solution depends on different factors. Most important are the complexity and costs for the development and maintenance of such solution by every SP. To a large extent this might be influenced by the utilization of the multi-domain functionality through requests for new and operation of established E2E connections.

In practice, two concepts have been often successfully combined in order to cope with high complexity and costs: specialization and sharing. A per se example is the DNS service, which is realized by specialized SPs and shared among SPs specializing on, e.g., content provisioning. Therefore we propose to support the delegation of multi-domain management tasks. As most common multi-domain management tasks, the following should be mentioned: routing, monitoring, accounting, or outage localization. Figure 18 depicts the delegation of the routing task. Instead of performing routing by themselves, as this was defined in our original proposal, provider SP_1 can delegate this task to SP_R which should follow the management-aware routing procedure. Please note that even if the trustworthiness of such specialized multi-domain management service providers might be very high, especially in large collaborations it does not necessary eliminate the necessity of task-delegation to proxy-SPs.

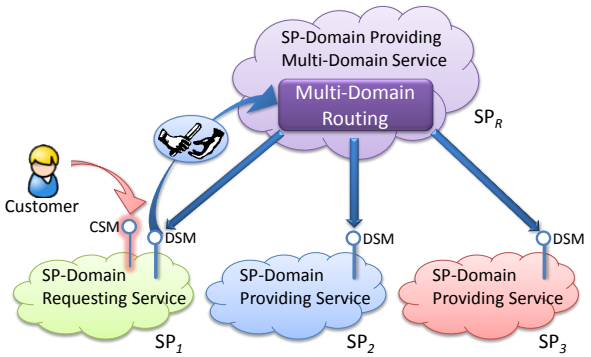

Figure 18. Delegating multi-domain functionality to a third party

B. Route Reservation and Ordering

In packet switching, like in IP-networks, the routing is an all-including procedure, as it is combined with the packet forwarding to the next hop which will be then in charge for further routing and forwarding. In line switching, the situation is more complex and routing can be seen as a general plan which still has to be implemented. If the plan is not sufficient, than even perfect implementations of it will not lead to the desired result; but even if the plan is perfect, it can be undermined by its bad implementation. Therefore we see the described management-aware routing as the most critical – but not sole – prerequisite for provisioning of multi-domain network connections compliant with user-specific E2E requirements.

As we consider a multi-domain environment, the information which is used during the routing is not only

semi-global, but also can be considered as not 100% up-to-date. This is mainly caused by concurrent requests, which can come from (in general various) SPs responsible for routing of different connections. On the one hand, various connection segments considered for these connections might be realized upon the same physical infrastructure. On the other hand, as a means for reduction of capital expenditure for infrastructure, every SP is interested in maximizing resource usage. Consequently, in the time between the information request and the actual ordering of the selected connection segments, the needed physical resources might be assigned to other service instances. In other words, we face the multi-domain concurrency for the same limited resources.

Various elaborated techniques are available for dealing with the concurrency issue. The most common way to avoid conflicts and/or minimize deadlocks is the reservation of the resources before their ordering. This technique can be also applied to concatenated services at the SP abstraction layer (see Figure 19). Every node in Figure 19 represents the state of one service part, e.g., a connection segment, from the perspective of SPs responsible for the provisioning of service instance. As the communication between different SPs can only be performed via some communication protocol, the transitions arrows between states are described with different kinds of requests. Please note that these names only reflect the semantics of requests and do not imply any suggestions for protocol implementation.

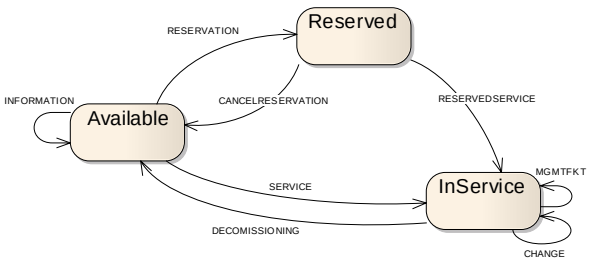

Figure 19. State transition at multi-domain abstraction level

Even though the state transition by itself is quite common, two additional aspects have to be specified: 1) behavior of SPs regarding requests for state transition of a single service part; and 2) the way to apply such requests to multiple service parts.

Concerning the definition of SP's behavior we propose to reuse elements of bilateral and trilateral negotiation models, as they are described in [34]. More precisely, if SP_{RS} (RS for "Requesting Service") requests from SP_{PS} (PS for "Providing Service") the reservation of connection segment with a certain quality, SP_{PS} should be able to confirm the reservation with equal or smaller quality as requested. This should cover the case if SP_{PS} in the time between information- and reservation-requests has already reserved some needed resources for a different service instance. After the reservation is confirmed, for each reserved service part, SP_{RS} should be able to request its ordering. In this case, however, SP_{PS} should be either able to provide the quality equal or better than the one confirmed by reservation, or report the failure of ordering.

The outcome of the management-aware routing procedure is a combination of connection segments along the route associated with the apparently feasible QoS thresholds. As all SPs providing those segments are also known, simultaneous communication with all those SPs would be possible. Concerning the reservation of service segments we however argue against the simultaneous communication with all SPs. In this case, reduced quality by reservation of a sole segment would automatically lead to a violation of the E2E user requirements. Instead, we propose to reserve segments sequential one-by-one from one endpoint to another one. In the case that SP_{PS} confirms reduced qualities, SP_{RS} can try to balance the performance degradation in one segment by increasing the planned thresholds of remaining segments. For this purpose information obtained during the routing by information requests can be reused. If the balancing is not possible, a re-routing for the remaining part of the path can be performed. If even this fails, SP_{RS} should cancel reservations of all connection segments which have been already successfully reserved. In order to minimize reservation time, messages about cancelling reservation can be sent simultaneously to all relevant SPs.

If reservation of all segments has been successful and E2E requirements are met, requests for ordering of these segments can be sent to corresponding SPs. As – corresponding to the proposed behavior – SPs may only provide better quality than the confirmed by the reservation, requests for ordering can be sent simultaneously to all SPs. If at least one of SPs cannot fulfill its commitment, it will report failure to SP_{RS}. In this case SP_{RS} has to cancel reservations of all remaining segments immediately.

Leveraging the proposed procedure for reservation and ordering of connection segments along the found route we score two goals: 1) first of all the fulfillment of E2E requirements; and 2) minimizing the time needed for resource reservation before a new service instance can become operational.

C. Examine SP-Acceptance for Elaborated Solution

As discussed in Sections I and II, the most critical requirement for an inter-domain routing approach is its acceptance by SPs. In our proposal, this real world requirement has been reflected by considering SP interests and restrictions.

As a proof of concept we refer to our experience within the context of the ***I**nformation **S**haring across **H**eterogeneous **A**dministrative **Re**gions* (I-SHARe) activity [32]. This activity has been established in Géant in order to foster information exchange during the manual planning and operation of E2E Links. During the first phase of this activity, requirements and constraints for the (back then) upcoming tool had to be captured and categorized. This has provided us with a deep insight into demands and concerns of network service providers.

First of all the sharing of detailed network information

(such as network topology or overall available capacity) is commonly seen as unacceptable. At the same time it is acceptable for SPs to exchange information about connection segments at SP abstraction level, which are possible for a particular planned E2E Link. Also exchange of more detailed information with neighboring SPs is very common by connection planning. The information model described in Section VII directly reflect these aspects. As our proposal strives for automatic route planning, we expect that multiple alternative service parts should be specified in the order of the SPs' preferences. Primarily, it should simplify information management and also reduce the necessary number of communication steps. Several interviews with operators involved in the planning of Géant E2E Links have shown the acceptance of SPs to provide information about multiple service parts during the routing process, even it is not used in the established manual processes.

By gathering of established in Géant manual route planning processes one further very important constraint has been identified. The only true concern that has been repeatedly mentioned by operations of different SPs is the compliance of the service part choice to the SP preferences. The enforcement of this aspect has been reflected in both routing modes outlined in Section V.

As our proposal can provide routing planning under consideration of SP constraints, we are confident that our solution can be broadly accepted by SPs. As it also considers the customer-specific E2E requirements and management functionality, our solution opens for SPs possibility for delivering of high-quality network services to the customers.

D. Application areas and possible adaptations

Compared to the established connectivity-oriented routing approaches, the proposed solution requires significantly more information, computations, and – which is the most time-consuming part – inter-domain communication. This statement is applicable not only to routing, but also to the subsequent reservation and ordering processes. Therefore, it cannot be considered as a routing procedure for a mass service, as it is the case by IP routing in Internet. Instead we suppose the application area of the developed algorithm to be in the middle-scale niche between mass services, which are focused on a pure connectivity with best-effort quality, and carrier-grade connections, which are mainly manually planned very-long-term connections with dedicated quality specified in contracts (see Figure 20).

The proposed solution can provide near-real-time planning and ordering of a route with customer-tailored E2E connection properties. This solution is applicable to scenarios like video-on-demand, video-conferencing, on demand connectivity for Grid or cloud collaborations, and other areas in which customers demand (and often are willing to pay) not only for the pure connectivity but also for the connection quality. Especially this is applicable for scenarios like video-conferencing, where bad quality

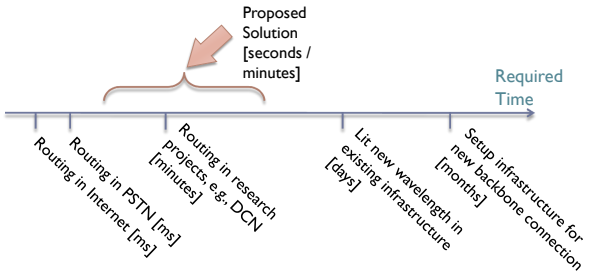

Figure 20. Positioning of proposed solution

of one parameter (e.g., jitter) cannot be covered by another one (e.g., bandwidth).

Depending on the offered service and/or the specifics of the SP cooperation, the proposed algorithm can be modified in order to improve its scalability. Particularly, we would like to outline the following two cases:

- If the connection service is only offered with QoS parameters that do not require E2E consideration of all involved parts, e.g., bandwidth or data encryption, a *routing-by-delegation* approach can generally be used. Especially in combination with the simultaneous resource reservation, it can prove to be very scalable. A very good example for this strategy can be seen in telephone connections, which offer constant bit rate and low jitter.
- In the case of a small and especially very tight SP cooperation, the routing instance can be centralized. In this case only this instance performs the whole E2E routing for all new connection requests, which neglects the concurrence between simultaneously ordered service instances. Consequently, the service part reservation can be omitted or performed simultaneously. A similar strategy with a two-level routing instances is used, e.g., in the DCN cooperation. This approach corresponds to the *source routing* approach (see Section V), in which the routing task is always delegated to a central instance. The applicability of this approach, however, depends directly on the willingness of SPs to provide complete information about all available service parts to the routing instance and to accept its inter-domain manager role.

A further simplification can be applied for operations on supported properties. If only few properties have to be supported and no extension of support is planned for the future, the usage of a registration tree for the definition of operations on properties can be omitted. This should significantly improve the performance of the implemented solution. On the other hand, support of not yet considered properties becomes an issue again.

As a conclusion, the proposed solution is applicable in the most challenging case of an open SP collaboration as well as with a variety of customer-specific QoS parameters.

X. Future work

As a mean to quality assurance for Concatenated Services, we have elaborated a novel conceptual solution consisting of three technology-independent building blocks. As the next logical step we see the evaluation of existing technologies, whether and how they can be used to implement the proposed solution. The desired outcome of this evaluation is a proposal for implementation of the presented solution and its building blocks. In the case if not all concepts can be implemented, the missing functionality can be seen as requirements for future technology development.

The proposed solution covers only the connection planning aspect during the ordering phase of service instance life cycle. In order to support SP collaboration during further phases, we plan to elaborate a proposal for inter-domain management processes. For instance, the methodologies for E2E monitoring have to be defined, which is needed for both monitoring of commitments' fulfillment and for problem localization. We see such processes as an essential part of the overall solution for CSs.

As the occurrence of outages cannot be fully excluded, strategies for handling them have to be elaborated. Therefore we plan to investigate the applicability of self-adaptation techniques as a strategy for the multi-domain compensation of a single-domain quality reduction or outages of service parts.

An additional aspect, which should not be neglected, is the consideration of security in multi-domain environments. Therefore the integration of federated identity and trust management into the management of CSs will be a part of our future research.

As stated at the beginning of this article, our work has focused on dedicated point-to-point connections. For the future, we also plan to evaluate the applicability of the developed solutions to the establishment and management of point-to-multipoint as well as multipoint-to-multipoint network connections with quality assurance.

Acknowledgment

The authors wish to thank the members of the Munich Network Management Team (MNM Team) for helpful discussions and valuable comments on previous versions of this paper. The MNM Team directed by Prof. Dr. Dieter Kranzlmüller and Prof. Dr. Heinz-Gerd Hegering is a group of researchers at Ludwig-Maximilians-Universität München, Technische Universität München, the University of the Federal Armed Forces and the Leibniz Supercomputing Centre of the Bavarian Academy of Science. See http://www.mnm-team.org.

References

[1] Yampolskiy, M., Hommel, W., Lichtinger, B., Fritz, W., Hamm, M.K. Multi-Domain End-to-End (E2E) Routing with multiple QoS Parameters. Considering Real World User Requirements and Service Provider Constraints. In *Proceedings The Second International Conference on Evolving Internet, 2010. INTERNET 2010,* Valencia, 2010.

[2] Yampolskiy, M., Hommel, W., Schmitz, D., Hamm, M.K. Generic Function Schema for Operation on multiple Network QoS-Parameters. Submitted to *The Second International Conference on Advanced Service Computing, 2010. SERVICE COMPUTATION 2010,* Lisbon, 2010.

[3] Yampolskiy, M., Hommel, W., Marcu, P., Hamm, M.K. An information model for the provisioning of network connections enabling customer-specific End-to-End QoS guarantees. In *Proceedings 7th IFIP/IEEE International Conference on Services Computing, 2010. SCC 2010,* Miami, 2010.

[4] Dynamic Circuit Network, Publications. http://www.internet2.edu/network/dc/publications.html [10 May 2010].

[5] Internet2 How to Connect: Internet2's Dynamic Circuit Network. http://www.internet2.edu/pubs/DCN-howto.pdf White paper 2008.

[6] ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS), Publications. http://www.es.net/OSCARS/ [10 May 2010].

[7] Dynamic Resource Allocation via GMPLS optical Networks (DRAGON), Publications. http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/Publications [10 May 2010].

[8] Phosphorus, Publications. http://www.ist-phosphorus.eu/publications.php [10 May 2010].

[9] Automated Bandwidth Allocation across Heterogeneous Networks (AutoBAHN), Bandwidth on Demand Deliverables. http://www.geant2.net/server/show/nav.1914 [10 May 2010].

[10] Joint Techs Winter 2009 TAMU DCN Demo. https://spaces.internet2.edu/display/DCN/Joint+Techs+Winter+2009++TAMU+DCN+Demo [28 May 2010].

[11] DICE (DANTE-Internet2-CANARIE-ESnet) Collaboration. http://www.geant2.net/server/show/conWebDoc.1308 [28 May 2010].

[12] Proposed architecture for inter-domain circuit monitoring. http://code.google.com/p/perfsonar-ps/wiki/CircuitMonitoring [28 May 2010].

[13] Yampolskiy, M., Hamm, M.K. Management of Multidomain End-to-End Links; A Federated Approach for the Pan-European Research Network Géant 2. In *Proceedings 10th IFIP/IEEE International Symposium on Integrated Network Management, 2007. IM'07,* Munich, 2007; 189-198.

[14] LHC - The Large Hadron Collider Homepage. http://lhc.web.cern.ch/lhc/ [10 May 2010].

[15] DEISA - Distributed European Infrastructure for Supercomputing Applications Homepage. http://www.deisa.eu/ [10 May 2010].

[16] Aptet, E., Chevers, J, Garcia Vidondo, M., Tyley, S. *Géant Deliverable DS2.0.3,3: Report on GÉANT2 Ad-vanced Services - Lambdas and Switched Optical* Géant2 technical paper 2009.

[17] Schauerhammer, K., Ullmann, K. Operational Model for E2E links in the NREN/GÉANT2 and NREN/Cross-Border-Fibre supplied optical platform. www.geant2.net/upload/pdf/GN2-06-119-OPConcept.pdf Géant2 technical paper 2006.

[18] Office of Governance Commerce (OGC) *IT Infrastructure Library (ITIL)*. London, 2007.

[19] TeleManagement Forum (TMF) *New Generation Operations Systems and Software (NGOSS)*. 2004.

[20] Hamm, M.K. *Eine Methode zur Spezifikation der IT-Service-Managementprozesse Verketteter Dienste* 2009;

[21] Khurram, K. (Editor) *Optical Networking Stan-dards: A Comprehensive Guide for Professionals* (1st edn). 2006;

[22] Brockners, F. Ethernet OAM Overview: Making Ethernet Manageable. In *Proceedings 1. DFN Forum Kommunikationstechnologien,* Müller P, Neumair B, Dreo Rodesek G (eds). Kaiserslautern, 2008; 35-44.

[23] Ziegelmann, M. *Constrained shortest paths and related problems* 2004;

[24] Hamm, M. K., Yampolskiy, M. IT Service Management verketteter Dienste in Multi-Domain Umgebungen. Modellierung und Teilaspekte. *PIK Praxis der Informationsverarbeitung und Kommunikation* 2008; **31**(2):82-89.

[25] Bellman, R. The theory of dynamic programming. In *Proceedings of the National Academy of Sciences of the United States of America,* 1952; 716-719.

[26] Jaffe, J.M. Algorithms for finding paths with multiple constraints. *Networks* 1984; **14**(1):95-116.

[27] Kuipers, F.A. *Quality of service routing in the internet: Theory, complexity and algorithms* 2004;

[28] Feng, G. A multi-constrained multicast QoS routing algorithm. *Computer Communications* 2006; **29**(10):1811-1822.

[29] Castineyra, I. and Chiappa, N. and Steenstrup, M. The Nimrod routing architecture. RFC-1992 [1996].

[30] ITU-T Generic functional architecture of transport networks. ITU-T: G.805 [2000].

[31] TeleManagement Forum (TMF) Common Information Model (CIM). [2009].

[32] Hamm, M. K., Yampolskiy, M., Hanemann, A. I-SHARe. *DFN Mitteilungen* 2008; **74**:39-41.

[33] De Marinis, E. and Hamm, M. and Hanemann, A. and Vuagnin, G. and Yampolskiy, M. and Cesaroni, G. and Thomas, S.-M. *Deliverable DS3.16.1: Use Cases and Requirements Analysis for I-SHARe* Géant2 technical paper 2008.

[34] Steinmetz, R. *Multimedia-Technologie. Grundlagen, Komponenten und Systeme* 2000;

[35] M. Yampolskiy. *Maßnahmen zur Sicherung von E2E–QoS bei Verketteten Diensten*. Ph.D. Thesis, Ludwig–Maximilians–Universität München, December 2009.

[36] Géant SA2-T1 working group *Static Dedicated Wavelength Service: Joint Infrastructure and Operational Level Agreement* Géant2 technical paper, Draft 2010.

Benefits of Virtual Network Topology Control based on Attractor Selection in WDM Networks

Yuki Minami*, Yuki Koizumi*, Shin'ichi Arakawa*, Takashi Miyamura†, Kohei Shiomoto† and Masayuki Murata*
*Graduate School of Information Science and Technology
Osaka University
Osaka 565-0871, Japan
{y-minami, ykoizumi, arakawa, murata}@ist.osaka-u.ac.jp
†NTT Network Service Systems Laboratories
NTT Corporation
Tokyo 180-8585, Japan
{miyamura.takashi, shiomoto.kohei}@lab.ntt.co.jp

***Abstract*—Virtual Network Topology (VNT) is one efficient way to transfer IP packets over wavelength-routed optical networks. In recent years various new services have emerged, and IP traffic has been highly fluctuated. Therefore, adaptability against changes of traffic is one of the most important characteristics to accommodate the IP traffic efficiently. To achieve the adaptability, we have proposed a VNT control method using an attractor selection model. In this paper, we investigate the adaptability of our VNT control method via computer simulations. Simulation results on various physical topologies indicate that our VNT control method can successfully adapt to changes of traffic around twice higher variance comparing with conventional VNT control methods. We also demonstrate that our VNT control method achieves one-tenth of control duration.**

***Keywords*-WDM; Virtual Topology Control; Virtual Topology Reconfiguration; Attractor Selection; Internet Protocol;**

I. INTRODUCTION

The rapid growth in the number of users and in the number of multimedia services is dramatically increasing traffic volume on the Internet. Wavelength division multiplexing (WDM) offers high-capacity data transmission by multiplexing optical signals into a fiber. With the optical cross-connects (OXCs) that switches the optical signals in all-optical domain offers the wavelength-routing. That is, sets of optical transport channels, called lightpaths, are established between nodes. Since the Internet protocol (IP) is emerging as a dominant technology, the ability to carry the IP traffic efficiently is an important issue to enjoy the WDM-based optical networks. One approach to accommodate IP traffic on WDM networks is to configure a virtual network topology (VNT), which consists of lightpaths and IP routers, through the wavelength-routing.

Many approaches to accommodate traffic demand by configuring VNTs have been investigated. One of approaches is that VNTs are statically constructed to efficiently accommodate one or multiple traffic demand matrices [2–5]. These approaches inherently assume that the traffic demand matrices are available before the VNT is constructed or assume that changes in the traffic demand matrices are predictable. However, the approaches cannot efficiently handle unexpected changes in traffic demand matrices since VNTs are configured for a certain set of traffic demand matrices. For example, the emergence of new services, such as peer-to-peer networks, voice over IP, and video on demand causes large fluctuations on traffic demand in networks [6], which makes the existing VNT control mechanisms insufficient to accommodate the traffic demand. Koizumi et. al [7] points out that, when there are overlay networks on top of the network controlled by the VNT control mechanism, traffic demand fluctuates greatly and changes in traffic demand are unpredictable. Therefore, VNT control methods that are adaptive to the traffic changes become important to avoid traffic congestions and to use network resources efficiently.

Recently, the dynamic VNT control that dynamically reconfigures VNTs based on their detection of degraded performance or periodic measurements of the network status without a priori knowledge of future traffic demand has been proposed [8,9]. In Ref. [8], the authors propose VNT control by assuming that traffic demand is changing gradually with a period of more than several hours. The method rely on the traffic demand matrices, and therefore the method cannot apply to VNT controls with short intervals of reconfigurations since traffic demand matrices is difficult to obtain with in a short period of time. In Ref. [9], the authors consider an hour-order traffic change and propose a VNT control method to adapt to the change. The method measures the traffic volume on each lightpath for short period and reconfigures VNT by using traffic demand matrices estimated from the measurement. The VNT control method again relies on the traffic demand matrices and the authors therefore try to estimate the traffic demand matrices more accurately. However, it requires several traffic measurements to estimate the traffic demand matrix. Thus the method cannot estimate traffic demand matrices correctly with a short period of time,

and cannot be applied when the traffic demand is highly fluctuated.

We therefore developed a VNT control method that is adaptive against changes in network environment without using traffic demand matrices [10, 11]. Our method uses an attractor selection that models behavior where living organisms adapt to unknown changes in their surrounding environments and recover their conditions. The fundamental concept underlying the attractor selection is that a system is driven by stochastic and deterministic behavior, and these are controlled by simple feedback of current condition. This characteristic is one of the most important differences between the attractor selection and other existing heuristic algorithms and optimization approaches. Our method measures only the traffic load on each lightpath (hereafter, we call the traffic load as link utilization). The quantity of information on the link utilization is less than that obtained from traffic demand matrices, but information about the link utilization is retrieved directly using, for example, SNMP (Simple Network Management Protocol).

Koizumi et al. [10,11] demonstrated that the VNT control based on attractor selection could reconfigure VNT with fast reaction and adaptation against changes in traffic demand. However, the simulation conditions are limited. In their simulation, they used a 19-node physical topology and evaluate with the randomly changing traffic demand. In the paper, only the concept of VNT control based on attractor selection was demonstrated, and did not understand well for the benefit of the method against the existing heuristic VNT control methods. In this paper, we evaluate the adaptability of the VNT control method against unknown and/or unexpected changes in surrounding environments and the range of network resource amount the method needs. We conduct simulations with various changes in traffic demand and physical topologies, and quantitatively show that the VNT control method can adapt to more various traffic changes than the existing heuristic methods.

The rest of this paper is organized as follows. Section II shows our network model. Section III briefly explains attractor selection and Section IV briefly explains our VNT control based on attractor selection. Section V shows the evaluation results and the performance our VNT control method. Finally, we conclude this paper in Section VI.

II. Network Model and Related Works

In this section, we describe the network model that we will use for the VNT configuration. Each node in the physical topology has IP routers and OXCs (Figure 1) and nodes are connected with optical fibers (Figure 2). The OXCs consist of three main blocks: input section, non-blocking optical switches, and output section. In the input section, optical signals are demultiplexed into W fixed wavelengths. Then, each wavelength is transferred to an appropriate output port by the non-blocking optical switch. Finally, at the output

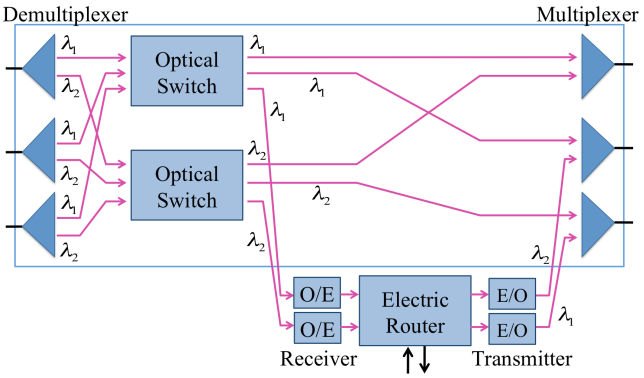

Figure 1. Node architecture

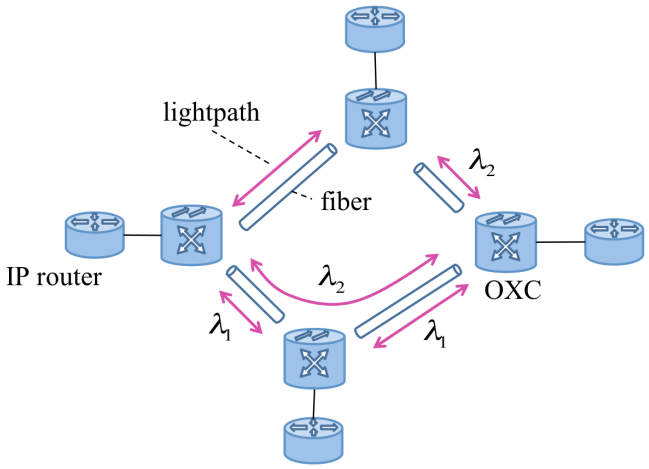

Figure 2. WDM network

section, each wavelength is multiplexed again and sent to the next node. By configuring the switches along a path, a lightpath is configured and a particular wavelength is carried from a transmitter at a node to a receiver at the other node without any electronic processing.

VNT is then constituted by setting up lightpaths on top of the physical topology (Figure 3). The actual traffic of the upper layer protocol, such as the IP, is carried on the constructed VNT. When a lightpath terminates at this node, the IP packets on the lightpath are converted to electrical signals and forwarded to the electronic router. When a lightpath begins at this node, IP packets from the electronic router are transmitted over the lightpath after being converted to optical signals. If lightpath is configured between all node-pairs, we do not need IP's packet processing in the VNTs. However, since the number of transmitters/receivers is limited, we should properly configure and reconfigure VNT.

In recent years, GMPLS (Generalized Multi-Protocol Label Switching), that is the technology to set and release lightpaths constructing VNT, is standardizing [12]. While international standard is being developed, advances in op-

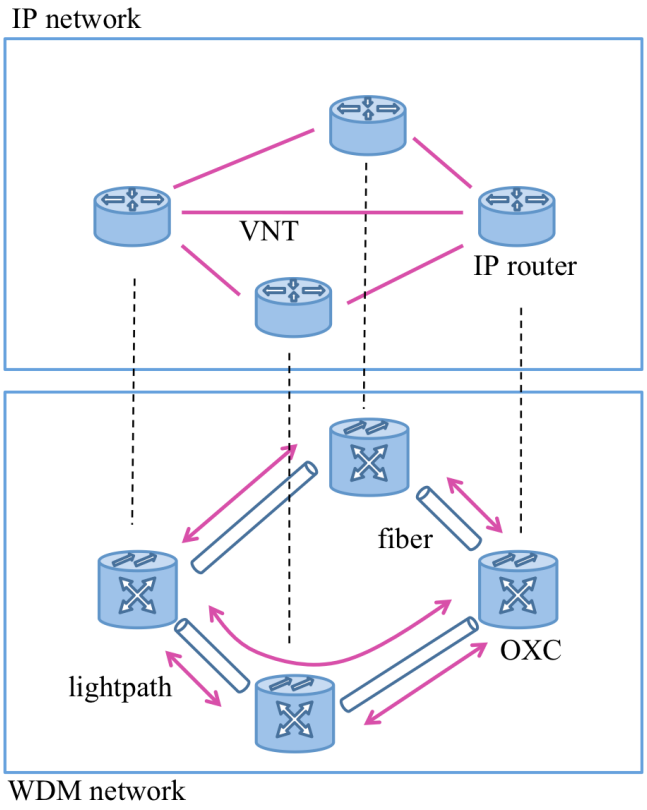

Figure 3. VNT configuration in IP-over-WDM network

tical technologies, such as optical switch with fast switching time, adaptability against changes of traffic demand becomes one of the important characteristics for the VNT controls. For example, in Ref. [8], authors try to keep the link utilizations between an upper limit and a lower limit against the traffic that grows during peak hours and falls during off-peak hours. The approach is different from previous studies that redesign the virtual topology according to an obtained traffic demand matrix through traffic measurements [4]. Ohsita et al. developed a VNT control method with an estimation of traffic demand matrices, and show that the method works with an hour-order traffic change [9]. More recently, Koizumi et al. demonstrated a VNT control based on the attractor selection model, which does not use the traffic demand matrix for VNT controls and thus achieves more shorter intervals of VNT reconfigurations [10, 11].

III. Attractor Selection

We first explain an attractor selection model that describes biological activities in a cell. The model is developed for a cell biology and please refer to Ref. [13] to understand the biological context of the model. This section explains an overview of attractor selection model, which will be use to explain the VNT control based on the attractor selection model in Section IV.

A. Outline of Attractor Selection

The attractor selection model represents metabolic reactions controlled by gene regulatory networks in a cell. Figure 4 illustrates a schematic representation in a cell. Each gene in the gene regulatory network has an expression level of proteins and deterministic and stochastic behaviors in each gene control the expression level. An attractor selection model is consists of regulatory behaviors having attractor which is determined by activation and inhibition between each genes, growth rate as feedback of the current condition of the network, and noise, which is stochastic behavior.

Attractors are a part of the equilibrium points in the solution space in which the current condition is preferable. The basic mechanism of an attractor selection consists of two behaviors: deterministic and stochastic behaviors. When the current condition is suitable for the current environment, i.e., the system state is close to one of the attractors, deterministic behavior drives the system to the attractor.

When the current condition is poor, stochastic behavior dominates over deterministic behavior. While stochastic behavior is dominant in controlling the system, the system state fluctuates randomly due to noise and the system searches for a new attractor. When the current condition has recovered and the system state comes close to an attractor, deterministic behavior again controls the system. These two behaviors are controlled by simple feedback of the current condition in the system. In this way, attractor selection adapts to environmental changes by selecting attractors using stochastic behavior, deterministic behavior, and simple feedback. In the following section, we introduce attractor selection that models the behavior of gene regulatory and metabolic reaction networks in a cell.

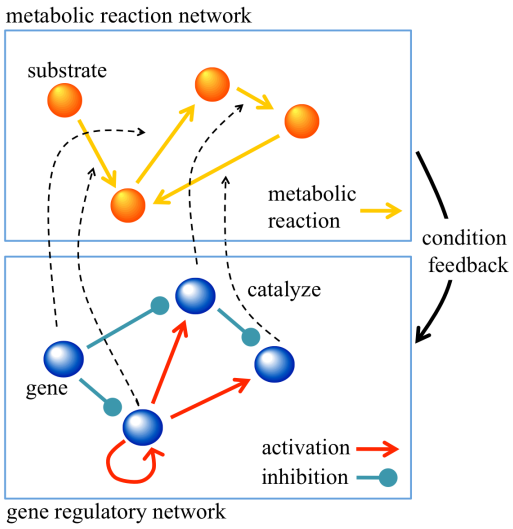

Figure 4. Networks in a cell

B. Mathematical Model of Attractor Selection

The internal state of a cell is represented by a set of expression levels of proteins on n genes, $(x_1, x_2, ..., x_n)$, and concentrations of m metabolic substrates, $(y_1, y_2, ..., y_m)$. The dynamics of the expression level of the protein of the i-th gene, x_i, is described as

$$\frac{dx_i}{dt} = f\left(\sum_{j=1}^{n} W_{ij}x_j - \theta\right) \cdot v_g - \cdot x_i v_g + \eta \qquad (1)$$

The first and second terms at the right hand side represent the deterministic behavior of gene i, and the third term represents stochastic behavior.

In the first term, the regulation of protein expression levels on gene i by other genes are indicated by regulatory matrix W_{ij}, which takes 1, 0, or -1, corresponding to activation, no regulatory interaction, and inhibition of the i-th gene by the j-th gene. The rate of increase in the expression level is given by the sigmoidal regulation function, $f(z) = 1/(1 + e - \mu z)$, where $z = \Sigma W_{ij}x_j - \theta$ is the total regulatory input with threshold θ for increasing x_i, and μ indicates the gain parameter of the sigmoid function.

The second term represents the rate of decrease in the expression level on gene i. This term means that the expression level decreases depending on the current expression level. The last term, η, represents molecular fluctuations, which is Gaussian white noise. Noise η is independent of production and consumption terms and its amplitude is constant. The change in expression level x_i is determined by deterministic behavior and stochastic behavior η. The deterministic and stochastic behaviors are controlled by growth rate v_g, which represents the conditions of the metabolic reaction network. In the metabolic reaction network, metabolic reactions, which are internal influences, and the transportation of substrates from the outside of the cell, which is an external influence, determine the changes in concentrations of metabolic substrates y_i. The metabolic reactions are catalyzed by proteins on corresponding genes. The expression level decides the strength of catalysis. A large expression level accelerates the metabolic reaction and a small expression level suppresses it. In other words, the gene regulatory network controls the metabolic reaction network through catalyses. Some metabolic substrates are necessary for cellular growth. Growth rate v_g is determined as an increasing function of the concentrations of these vital substrates. The gene regulatory network uses v_g as the feedback of the conditions on the metabolic reaction network and controls deterministic and stochastic behaviors. If the concentrations of the required substrates decrease due to changes in the concentrations of nutrient substrates outside the cell, v_g also decreases. By decreasing v_g, the effects that the first and second terms have on the dynamics of x_i decrease, and the effects of η increase relatively. Thus, x_i fluctuates randomly and the gene regulatory network searches for a new attractor. The fluctuations in x_i lead to changes in the rate of metabolic reactions via the catalyses of proteins.

When the concentrations of the required substrates again increase, v_g also increases. Then, the first and second terms again dominate the dynamics of x_i over stochastic behavior, and the system converges to the state of the attractor. In next section, we describe our VNT control method based on attractor selection.

IV. VNT Control Based on Attractor Selection

In this section, we briefly explain VNT control methods based on attractor selection. Attractors are a part of the equilibrium points in the solution space in which the current condition is preferable. In our VNT control method, we regard the attractor as VNT, and then select it based on the attractor selection model.

A. VNT Control Method

In the cell, the gene regulatory network controls the metabolic reaction network, and the growth rate, which is the status of the metabolic reaction network, is recovered when the growth rate is degraded due to changes in the environment. In our VNT control method, the main objective is to recover the performance of the IP network by appropriately constructing VNT when performance is degraded due to changes in traffic demand. Therefore, we interpret the gene regulatory network as a WDM network and the metabolic reaction network as an IP network (Figure 5).

Outline of our VNT control method is as follows:

Step. 1 Measure the link utilization via SNMP (Simple Network Management Protocol).

Step. 2 Determine growth rate from the link utilization. Growth rate express if IP network is in good condition or not. We describe detail of how to determine growth rate in Section II-C. Note that the degree of influence of deterministic behaviors and stochastic behaviors is determined by the growth rate.

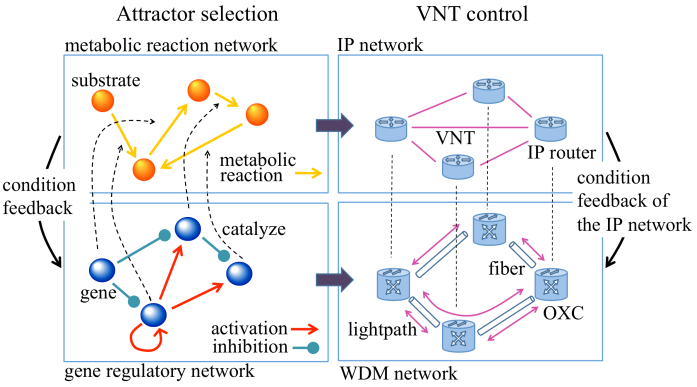

Figure 5. VNT control based on the attractor selection

Step. 3 The number of lightpaths is determined based on the expression level of each gene. Then, the VNT is reconfigured. We describe how to decide the number of lightpaths in Section II-C.

Step. 4 Transfer the IP traffic over the newly constructed VNT. Consequently the link utilization changes again, so we repeat these steps again.

B. Interaction in VNT Control

This section describes our VNT control method in detail. We consider the dynamical system that is driven by the attractor selection. We place genes on every source-destination pair (denote p_{ij} for nodes i and j) in the WDM network, and the expression level of the genes $x_{p_{ij}}$ determines the number of lightpaths on between nodes i and j. To avoid confusion, we refer to genes placed on the network as *control units* and expression levels as *control values*. The dynamics of $x_{p_{ij}}$ is defined by the following differential equation,

$$\frac{dx_{p_{ij}}}{dt} = v_g \cdot f\left(\sum_{p_{sd}} W(p_{ij}, p_{sd}) \cdot x_{p_{sd}} - \theta_{p_{ij}}\right) - v_g \cdot x_{p_{ij}} + \eta \quad (2)$$

where η represents Gaussian white noise, f is the sigmoidal regulation function, and v_g is the growth rate. v_g indicates the condition of the IP network.

The number of lightpaths between node pair p_{ij} is determined according to value $x_{p_{ij}}$. We assign more lightpaths to a node pair that has a high control value than a node pair that has a low control value. $\theta_{p_{ij}}$ in the sigmoidal regulation function f is the threshold value to control the number of lightpaths.

The regulatory matrix W represents relations of the activation and inhibition between control units. Each element in the regulatory matrix, denoted as $W(p_{ij}, p_{sd})$, represents the relation between node pair p_{ij} and p_{sd}. The value of $W(p_{ij}, p_{sd})$ takes a positive number α_A, zero, or a negative number α_I, each corresponding to activation, no relation, and inhibition of the control unit on p_{ij} by the control unit on p_{sd}. For example, if the lightpath on p_{ij} is activated by that on p_{sd}, increasing $x_{p_{sd}}$ leads to increasing x_{ij}. That is, node pair p_{sd} increases the number of lightpaths on p_{ij} in our VNT control method. In our method, we define α_A as $\alpha_A = 1.08N/N_A$, α_I as $\alpha_I = 1.08N/N_I$, where N is the number of control units, N_A is the number of control units that is activated, and N_I is the number of control units that is inhibited.

We consider three motivations for defining the regulatory matrix in WDM networks. First, when we assign a new lightpath to detour traffic from node i to j for substitute of another lightpath, the traffic passing from node i to j will be transmitted by the new lightpath. Therefore, the control units on each node pair along the route of the lightpath from node i to j activate the control unit on p_{ij}. Next, we consider the situation where a path on the IP network uses the lightpaths on p_{ij} and p_{sd}. In this case, some traffic on p_{ij} is also transported on p_{sd}. If the number of lightpaths on p_{ij} is increased, the number of lightpaths on p_{sd} should also be increased to transport IP traffic efficiently. Therefore, the control units on p_{ij} and p_{sd} activate each other. Finally, we consider the situation that node pairs share a certain fiber. Here, if the number of lightpaths on one node pair increases, the number of lightpaths on the other node pairs should decrease because of limitations on wavelengths. Therefore, the control unit on p_{ij} is inhibited by the control unit on p_{sd} if lightpaths between these node pairs share the same fiber.

The growth rate indicates the current condition of the IP network, and the WDM network seeks to optimize the growth rate. In our VNT control method, we use the maximum link utilization on the IP network as a metric that indicates the current condition of the IP network. To retrieve the maximum link utilization, we collect the traffic volume on all links and select their maximum value. This information is easily and directly retrieved by SNMP. Hereafter, we will refer to the growth rate defined in our VNT control method as *activity*. Figure 6 indicates the function determining the activity. The activity must be an increasing function for the goodness of the current condition of networks. We therefore convert the maximum link utilization on the IP network, $u_{\max}$, into the activity, v_g, by the following equation.

$$v_g = \begin{cases} \frac{\gamma}{1+\exp(\delta \cdot (u_{\max} - \zeta))} & \text{if } u_{\max} \geq \zeta \\ \frac{\gamma}{1+\exp(\delta' \cdot (u_{\max} - \zeta))} & \text{if } u_{\max} < \zeta \end{cases} \quad (3)$$

Here, γ is the parameter that scales v_g and δ represents the gradient of this function. The constant number, ζ, is the threshold for the activity. If the maximum link utilization is more than threshold ζ, the activity approaches 0 due to the poor condition of the IP network. Then, the dynamics of our VNT control method is governed by noise and search for a new attractor. If the maximum link utilization is less

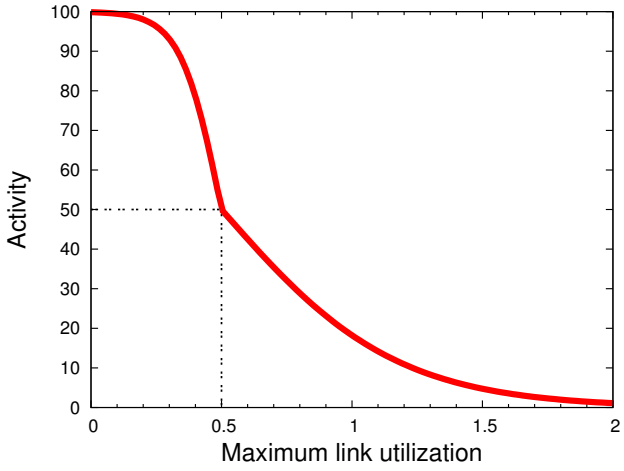

Figure 6. Activity function

than ζ, we increase the activity. Then the system is driven by deterministic behavior and the system will be stable.

C. The Number of Lightpaths

The number of lightpaths between node pair p_{ij} is calculated from $x_{p_{ij}}$ that is the expression level of gene placed for p_{ij}. To simplify the model of our VNT control method, we assume that the number of wavelengths on optical fibers will be sufficient and the number of transmitters and receivers of optical signals will restrict the number of lightpaths between node pairs. Each node has P_R receivers and P_T transmitters. We assign transmitters and receivers to lightpaths between p_{ij} based on $x_{p_{ij}}$ normalized by the total control values for all the node pairs that use the transmitters or the receivers on node i or j. The number of lightpaths between p_{ij}, $G_{p_{ij}}$, is determined as

$$G_{p_{ij}} = \min\left(\lfloor P_R \cdot \frac{x_{p_{ij}}}{\sum_s x_{p_{sj}}} \rfloor, \lfloor P_T \cdot \frac{x_{p_{ij}}}{\sum_d x_{p_{id}}} \rfloor\right) \quad (4)$$

Since we adopt the floor function for converting real numbers to integers, each node has residual transmitters and receivers. We assign one lightpath in descending order of $x_{p_{ij}}$ while the constraint on the number of transmitters and receivers is satisfied. Note that other constraints derived from physical resources can easily be considered for determining $G_{p_{ij}}$. For instance, when we pose a constraint on the number of wavelengths on fibers, we assign wavelengths on fibers through which the lightpath passes based on $x_{p_{ij}}$ normalized by the sum of expression levels on the corresponding fiber.

V. Performance Evaluation

We next evaluate the adaptability of our VNT control method against changes of traffic demand via computer simulations. For comparison purpose, we first introduce an existing heuristic method in Section V-A and then present some simulation results.

A. Existing Heuristic Method

Ref. [8] proposed a heuristic VNT control method, which we will refer to "ADAPTATION". ADAPTATION aims at achieving adaptability against changes in traffic demand. This method reconfigures VNTs according to the link utilization and the traffic demand matrix. ADAPTATION has a lower limit and an upper limit for link utilization and reconfigure VNT to put link utilization in the region. ADAPTATION measures the actual link utilization every 5 minutes and adds a new lightpath to the current VNT when congestion occurs. This method places a new lightpath on the node pair with the highest traffic demand among all node pairs that use the congested link.

ADAPTATION uses the information of traffic demand matrix to identify the node pair that has the largest traffic demand. However, collecting the information of traffic demand matrix is difficult in general because measurements of individual flows in a real-time manner are required. In this paper, we use the tomogravity method [14] that estimates the traffic demand matrix based on the information of link utilization, and we apply the estimated traffic demand matrix to the ADAPTATION. Note that both our VNT control method and ADAPTATION use only the information of link utilization that we can get easily by SNMP to calculate the activity of the IP network, but our VNT control method does not estimate the traffic demand matrix.

B. Simulation Conditions

We use the European Optical Network (EON) topology as shown as shown in Figure 7. The EON topology has 19 nodes and 39 bidirectional fibers. Each node has eight transmitters and eight receivers.

We focus on changes in traffic demand in the IP network as the environmental changes. For the evaluation, we prepare the traffic demand matrices where traffic demand from node i to j, d_{ij}, follows a lognormal distribution. We set the variance of logarithm of d_{ij} to be σ^2 and with the mean to be 1. Then, we change the σ^2 to evaluate the adaptability against the changes of network environments. Each traffic demand matrix is normalized such that the total amount of traffic, $\sum_{p_{ij}} d_{p_{ij}}$, is the same and is set to 10 in a unit of bandwidth of lightpaths.

In the simulation, Our VNT control method collects information about the link utilization every 5 minutes by SNMP. The parameter settings of our VNT control method are shown in Table I. η used in Equation 2 follows normal distribution with variance of 0.2 and the mean of 0.

For the parameter settings for the ADAPTATION method, we set the lower limit to 0.1 and the upper limit to 0.5. ADAPTATION measures the actual link utilization every 5 minutes and control VNT in the simulation.

C. Simulation Results

We first show the maximum link utilization dependent on time in Figure 8. In obtaining the figure, we set σ^2 to

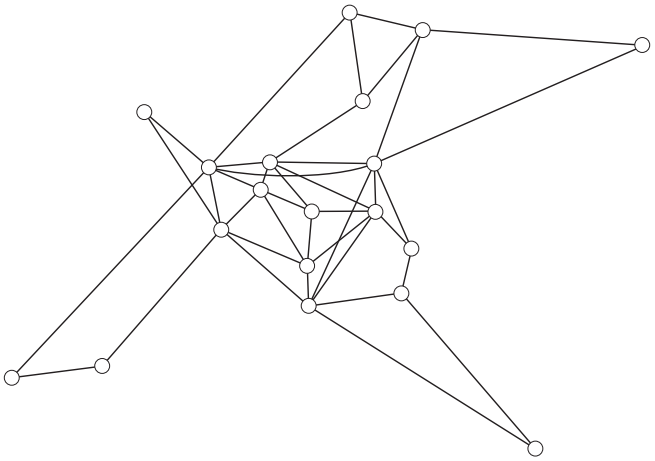

Figure 7. EON topology

Table I
PARAMETER SETTINGS OF OUR VNT CONTROL METHOD.

Name	Values
γ	100
δ	13
δ'	3
ζ	0.5
μ	1

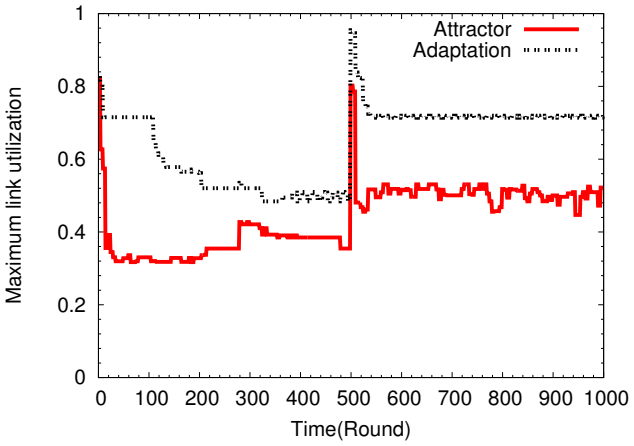

Figure 8. Changes of maximum link utilization

2.0 and change the traffic demand at time 500 by setting the different value of random seed for d_{ij}. In both methods, the maximum link utilizations gradually decrease after the change of traffic demand occurs, while our VNT control method sharply decreases the maximum link utilization. In the figure, our VNT control method successfully decreases the maximum link utilization to be lower than 0.5, while the ADAPTATION cannot decrease. We regard that the VNT control is successful when the maximum link utilization is decreased to less than 0.5. Otherwise the control is fail.

We evaluate the success rate of VNT reconfigurations by changing the parameter σ^2 from 0 to 2.4, and conducting the simulation 100 times for each value of σ^2. The results are shown in Figure 9 where the horizontal axis represents the value of σ^2 and the vertical axis represents the average of success rate.

We observe that our method achieves 100% success rate when σ^2 is less than 1.1. Comparing with the results of the ADAPTATION method, our virtual topology control can successfully adapt changes of traffic demand around twice higher variance comparing with the ADAPTATION method. In both methods, the success rate more decreases as σ^2 takes larger values. However, when σ^2 is 2.4, the success rate of our method is higher than 80%, while that of the ADAPTATION method decreases significantly.

We next discuss the control duration, defined as the time from when the traffic change occurs to when the maximum link utilization becomes less than 0.5. Figure

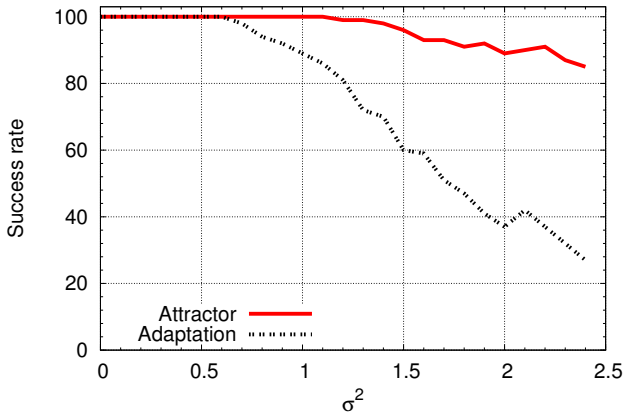

Figure 9. Success rate of VNT reconfigurations in EON topology

10 shows the average and 90% confidence interval of the control duration dependent on σ^2. For calculating the control duration, we use only the cases when VNT reconfigurations are successful. We observe that our method achieves lower control duration comparing with the ADAPTATION method. As the σ^2 increases, the difference between our method and ADAPTATION method increases. Looking at the results when σ^2 is 2.4, the averaged control duration of the ADAPTATION method is 90 minutes, while the averaged control duration of our method is only 30 minutes. More importantly, the confidence interval of the ADAPTATION is wide: the interval ranges from 30 minutes to 150 minutes. However, results of our method are ranging from 5 minutes to 60 minutes.

A disadvantage of our method is shown in Figure 11. The figure shows the maximum value of control durations. When σ^2 is 1.1 and 1.5, the control duration of our method

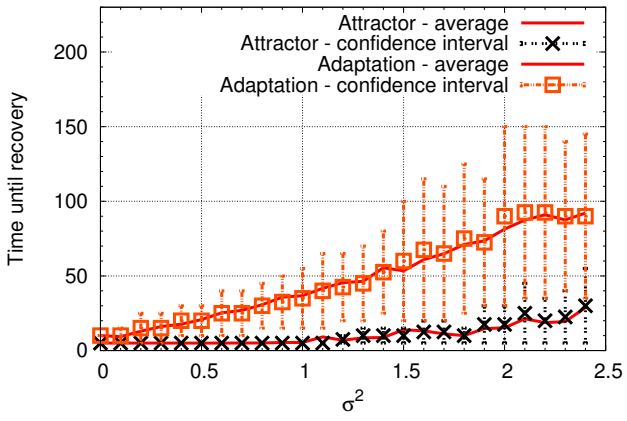

Figure 10. Average and 90% confidence interval of control duration in EON topology

is larger than that of the ADAPTATION method because of the stochastic behavior of our method; the noise term in Equation 2 does not work well in some cases. Note however that the success rate of our method is higher than that of the ADAPTATION method when σ^2 is 1.1 and 1.5.

We next show the results of our method in the Abilene topology (Figure 12). Figure 13 shows the success rate in Abilene topology. We can see that our method achieves 100% success rate when σ^2 is less than 2.0 and our method keep high rate compared with ADAPTATION. Looking at the Figure 14 that show the time until recovery, we again observe that our method reconfigure the VNT with fast reaction; the 90% confidential interval ranges from 5 minutes to 20 minutes.

We also conduct simulations for larger physical topology having 100 node and 200 bidirectional fibers that are connected randomly. Results are summarized in Table II. In the simulation, the total traffic volume is set to 30 in a unit of bandwidth of lightpaths. We also set the number of transmitters/receivers on each node to be 24. With these parameter settings, our method decreases the maximum link utilization as shown in Figure 15 and the success rate is higher than 90% as shown in Table II. Note that when the number of transmitters/receivers is too small for the physical topology, the number of attractors, i.e., the number of VNT candidates, is also small. In this case, the VNT control based on the attractor selection is difficult to search for a new attractor through a noise term in Equation 2.

To see the effect of number of transmitters/receivers more clearly, we conduct a set of simulations for the EON topology by changing the number of transmitters and receivers and evaluate the success rate for each VNT control method. We prepare three traffic scenarios shown in Table III by changing traffic volume and σ^2. In scenario 1, we set the traffic volume to the same as above simulation and σ^2 to 1.0. In scenario 2, we set the traffic volume to 1.5 times larger than scenario 1, but use the same value for σ^2. For scenario 3, we increase both the traffic volume and the σ^2. We generate a traffic demand matrix based on the parameters for each traffic scenario and then evaluate whether VNT control methods successfully adapt to the traffic demand. The other simulation conditions are the same as the simulation conditions of the EON topology (See Section V-B).

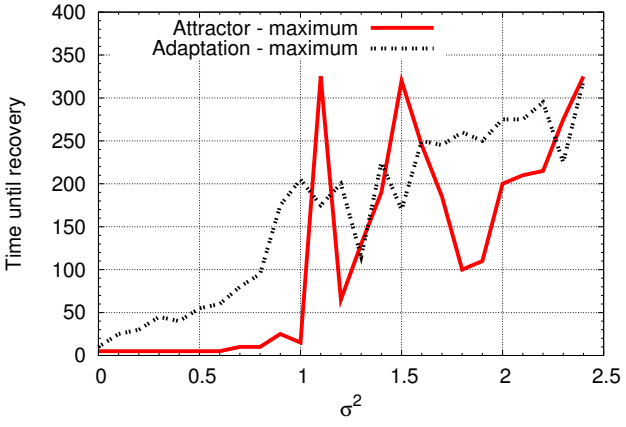

Figure 11. The maximum control duration in EON topology

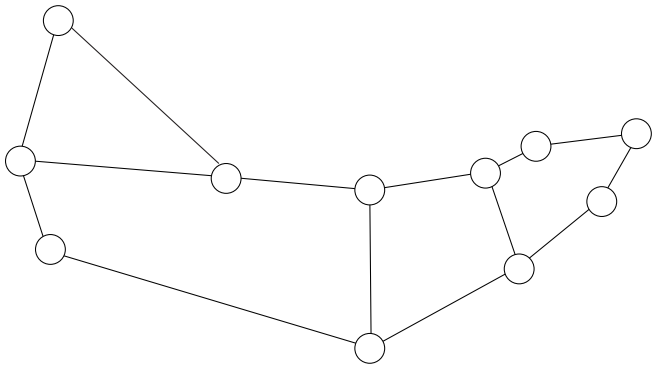

Figure 12. Abilene topology

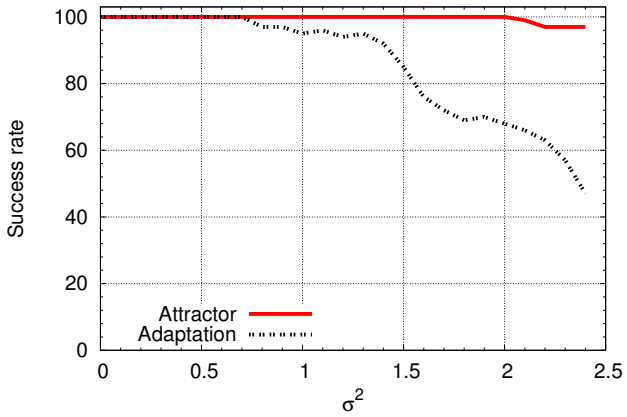

Figure 13. Success rate of VNT reconfigurations in Abilene topology

Table II
SUCCESS RATE OF VNT RECONFIGURATIONS IN 100-NODE TOPOLOGY

σ^2	Success Rate
1.3	100
1.4	100
1.5	100
1.6	100
1.7	100
1.8	100
1.9	100
2.0	100
2.1	100
2.2	98
2.3	98
2.4	97

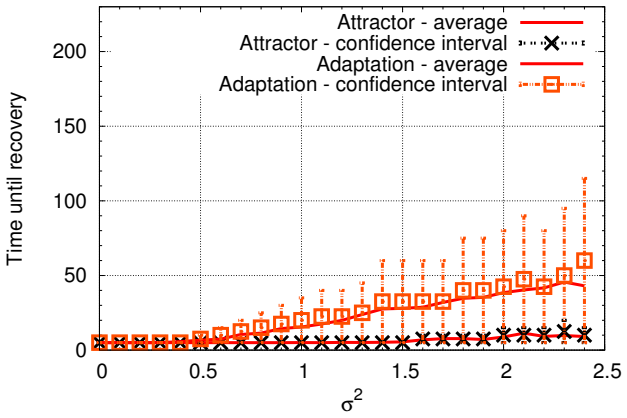

Figure 14. Average and 90% confidence interval of control duration in Abilene topology

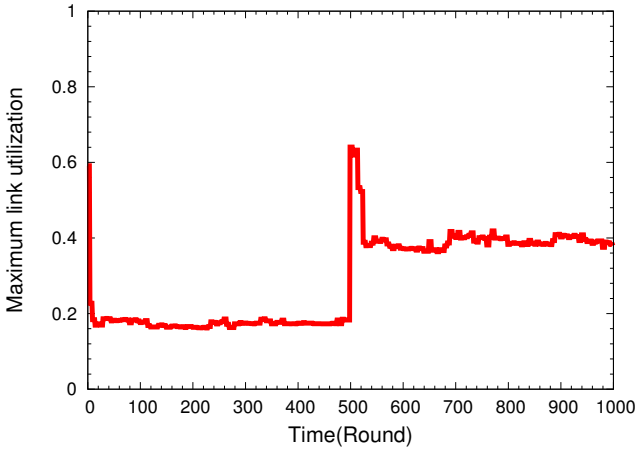

Figure 15. Attractor - Changes maximum link utilization in 100-nodes topology

Table IV summarizes the success rate of VNT reconfigurations. In the table, we introduce the results of MLDA (Minimum-delay Logical topology Design Algorithm) [4] which is a heuristic and centralized VNT control method on the basis of a given traffic demand matrix. MLDA simply places lightpaths between nodes in descending order of traffic demand. We show the results of MLDA with the actual traffic demand matrix to see how transmitters and receivers are effectively used in VNT control methods. From the table, we observe that each VNT control method have significantly low success rate with the small number of transmitters/receivers. This means that enough number of transmitters/receivers is crucial for VNT controls in changing network environments.

Table III
THREE TRAFFIC SCENARIOS IN THE EON TOPOLOGY

	Total traffic volume (relative to Figure 9)	σ^2
Scenario 1	1.0	2.0
Scenario 2	1.5	1.0
Scenario 3	1.5	2.0

The benefit of heuristic VNT control methods appears with the small number of transmitters/receivers; ADAPTATION and MLDA have a higher success rate than ATTRACTOR when the number of transmitters/receivers is 5 or 6 for traffic scenario 1, but the differences are marginal. More importantly, the benefit of MLDA disappears as the number of transmitters/receivers increases; the success rate of ATTRACTOR and MLDA is mostly the same for traffic scenarios 1 and 2. Looking at the case of traffic scenario 3, we observe a disadvantage of MLDA. That is, the results of MLDA show poor success rate comparing with the results of ATTRACTOR. The reason is that when the variance of traffic demand matrix increases, the heuristic behind MLDA fails. ADAPTATION has a much lower success rate than the other VNT control methods due to the estimation error in obtaining the traffic demand matrix from the information of link utilization.

In summary, heuristic VNT control methods based on the traffic demand matrices have a capability to obtain good VNTs with the small number of transmitters/receivers. Therefore, the heuristic VNT control methods are useful for the network with gradual change in traffic demand matrices. However, for the network having large fluctuations on traffic demand, enough number of transmitters/receivers is crucial. With this case, our VNT control based on attractor selection achieves good adaptability to the traffic change and lower control durations.

VI. CONCLUSION

Adaptability against changes of traffic demand is one of the important characteristics. In this paper, we evaluated the adaptability of VNT control based on the attractor selection. Simulation results with various physical topologies and traffic demand matrices showed that our VNT control method could successfully adapt changes of traffic around twice higher variance comparing with existing heuristic method. We also demonstrated that our VNT control method achieves short control duration of VNT reconfiguration in most cases. We then evaluated the success rate with different number of transmitters/receivers for three traffic scenarios, and compared our VNT control methods with existing two heuristic VNT control methods. The results indicate that existing methods have a higher success rate than our VNT control method when the number of transmitters/receivers is small, but its differences are marginal. To achieve an adaptive VNT controls to the changes of network environments, enough number of transmitters/receivers is crucial. With this case, our VNT control based on attractor selection achieves good adaptability to the traffic change and lower control durations.

Table IV
SUCCESS RATE (IN PERCENTAGE)

	Number of transmitters/receivers								
	1	2	3	4	5	6	7	8	9
scenario 1 ATTRACTOR	0	0	0	0	6	12	52	89	96
scenario 1 ADAPTATION	0	0	0	0	7	23	30	37	51
scenario 1 MLDA	0	0	0	0	7	17	52	85	92
scenario 2 ATTRACTOR	0	0	0	0	0	0	17	82	100
scenario 2 ADAPTATION	0	0	0	0	0	2	8	18	32
scenario 2 MLDA	0	0	0	0	0	0	23	79	97
scenario 3 ATTRACTOR	0	0	0	0	0	0	5	44	96
scenario 3 ADAPTATION	0	0	0	0	0	0	0	2	7
scenario 3 MLDA	0	0	0	0	0	0	4	20	42

ACKNOWLEDGMENT

This work is partly supported by SCOPE (Strategic Information and Communications R&D Promotion Programme) operated by Ministry of Internal Affairs and Communications of Japan and by Grant-in-Aid for Scientific Research (B) 22300023 of the Ministry of Education, Culture, Sports, Science and Technology in Japan.

REFERENCES

[1] Y. Minami, Y. Koizumi, S. Arakawa, T. Miyamura, K. Shiomoto, and M. Murata, "Adaptive virtual network topology control in WDM-based optical networks," in *Proceedings of INTERNET 2010*, pp. 49–54, Sept. 2010.

[2] S. Arakawa, M. Murata, and H. Miyahara, "Functional partitioning for multi-layer survivability in IP over WDM networks," *IEICE Transactions on Communications*, vol. 83, pp. 2224–2233, Oct. 2000.

[3] N. Ghani and S. Wang, "On IP-over-WDM integration," *IEEE Communications Magazine*, vol. 38, pp. 72–84, Mar. 2000.

[4] R. Ramaswami, K. Sivarajan, I. Center, and Y. Heights, "Design of logical topologies for wavelength-routed optical networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 840–851, June 1996.

[5] F. Ricciato, S. Salsano, A. Belmonte, and M. Listanti, "Off-line configuration of a MPLS over WDM network under time-varying offered traffic," in *Proceedings of INFOCOM*, pp. 57–65, June 2002.

[6] Y. Liu, H. Zhang, W. Gong, and D. Towsley, "On the Interaction Between Overlay Routing and Underlay Routing," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2005)*, vol. 4, pp. 2543–2553, Mar. 2005.

[7] Y. Koizumi, S. Arakawa, and M. Murata, "Stability of virtual network topology control for overlay routing services," *OSA Journal of Optical Networking*, vol. 7, pp. 704–719, July 2008.

[8] A. Gencata and B. Mukherjee, "Virtual-topology adaptation for WDM mesh networks under dynamic traffic," *IEEE/ACM Transactions on Networking*, vol. 11, pp. 236–247, Apr. 2003.

[9] Y. Ohsita, T. Miyamura, S. Arakawa, S. Ata, E. Oki, K. Shiomoto, and M. Murata, "Gradually reconfiguring virtual network topologies based on estimated traffic matrices," *IEEE/ACM Transactions on Networking*, vol. 18, pp. 177–189, Feb. 2010.

[10] Y. Koizumi, T. Miyamura, S. Arakawa, E. Oki, K. Shiomoto, and M. Murata, "Application of attractor selection to adaptive virtual network topology control," in *Proceedings of BIONETICS*, pp. 1–8, Nov. 2008.

[11] Y. Koizumi, T. Miyamura, S. Arakawa, E. Oki, K. Shiomoto, and M. Murata, "Robust virtual network topology control based on attractor selection," in *Proceedings of ONDM*, pp. 123–128, Feb. 2009.

[12] K. Shiomoto, "Requirements for GMPLS-based multi-region and multi-L." RFC 5212, July 2008.

[13] C. Furusawa and K. Kaneko, "A generic mechanism for adaptive growth rate regulation," *PLoS Computational Biology*, vol. 4, p. e3, Jan. 2008.

[14] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale IP traffic matrices from link loads," *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, pp. 206–217, June 2003.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601