CTRQ 2017

The Tenth International Conference on Communication Theory, Reliability, and Quality of Service

ISBN: 978-1-61208-550-0

April 23 - 27, 2017

Venice, Italy

CTRQ 2017 Editors

Richard Li, Huawei Technologies, USA
Carla Merkle Westphall, University of Santa Catarina, Brazil
Eugen Borcoci, Politehnica University of Bucharest, Romania

CTRQ 2017

Forward

The Tenth International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ 2017), held between April 23-27, 2017 in Venice, Italy, continued a series of events focusing on the achievements on communication theory with respect to reliability and quality of service. The conference also brought onto the stage the most recent results in theory and practice on improving network and system reliability, as well as new mechanisms related to quality of service tuned to user profiles.

The processing and transmission speed and increasing memory capacity might be a satisfactory solution on the resources needed to deliver ubiquitous services, under guaranteed reliability and satisfying the desired quality of service. Successful deployment of communication mechanisms guarantees a decent network stability and offers a reasonable control on the quality of service expected by the end users. Recent advances on communication speed, hybrid wired/wireless, network resiliency, delay-tolerant networks and protocols, signal processing and so forth asked for revisiting some aspects of the fundamentals in communication theory. Mainly network and system reliability and quality of service are those that affect the maintenance procedures, on the one hand, and the user satisfaction on service delivery, on the other hand. Reliability assurance and guaranteed quality of services require particular mechanisms that deal with dynamics of system and network changes, as well as with changes in user profiles. The advent of content distribution, IPTV, video-on-demand and other similar services accelerate the demand for reliability and quality of service.

The conference had the following tracks:

- Quality and Reliability
- IONCOMM: Identity Oriented Networks-based Infrastructure and Communications
- Internet of Things - Recent Trends, Technologies and Techniques
- Reliability and Maintenance

We take here the opportunity to warmly thank all the members of the CTRQ 2017 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to CTRQ 2017. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the CTRQ 2017 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that CTRQ 2017 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of communication theory, reliability and quality of service. We also hope that Venice, Italy

provided a pleasant environment during the conference and everyone saved some time to enjoy the unique charm of the city.

CTRQ 2017 Committee

CTRQ Steering Committee
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Pål Ellingsen, Bergen University College, Norway
Wojciech Kmiecik, Wroclaw University of Technology, Poland
Leyre Azpilicueta, Tecnológico de Monterrey, Mexico

CTRQ Industry/Research Advisory Committee
Carlos Kavka, ESTECO SpA, Italy
Daniele Codetta Raiteri, Università del Piemonte Orientale, Italy
Kiran Makhijani, Huawei Technologies, USA

CTRQ 2017
Committee

CTRQ Steering Committee

Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Pål Ellingsen, Bergen University College, Norway
Wojciech Kmiecik, Wroclaw University of Technology, Poland
Leyre Azpilicueta, Tecnológico de Monterrey, Mexico

CTRQ Industry/Research Advisory Committee

Carlos Kavka, ESTECO SpA, Italy
Daniele Codetta Raiteri, Università del Piemonte Orientale, Italy
Kiran Makhijani, Huawei Technologies, USA

CTRQ 2017 Technical Program Committee

Mazin Alshamrani, MoHaj, Saudi Arabia / University of South Wales, UK
Leyre Azpilicueta, Tecnológico de Monterrey, Mexico
Dirk Bade, University of Hamburg, Germany
Jasmina Barakovic Husic, BH Telecom, Joint Stock Company / University of Sarajevo, Bosnia and Herzegovina
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Christos Bouras, University of Patras - Computer Technology Institute & Press «Diophantus», Greece
Daniele Codetta Raiteri, Università del Piemonte Orientale, Italy
Manfred Droste, Universität Leipzig, Germany
Pål Ellingsen, Bergen University College, Norway
Andras Farago, University of Texas at Dallas, USA
Gianluigi Ferrari, University of Parma, Italy
Tulsi Pawan Fowdur, University of Mauritius, Mauritius
Borko Furht, Florida Atlantic University, USA
Julio César García Álvarez, Universidad Nacional de Colombia, Colombia
Rita Girao-Silva, University of Coimbra / INESC-Coimbra, Portugal
Apostolos Gkamas, University Ecclesiastical Academy of Vella of Ioannina, Greece
Teresa Gomes, University of Coimbra, Portugal
Teodor Lucian Grigorie, University of Craiova, Romania
Ilias Iliadis, IBM Research - Zurich, Switzerland
Mohsen Jahanshahi, Islamic Azad University, Tehran, Iran
Alexey Kashevnik, SPIIRAS, Russia
Sokratis K. Katsikas, Norwegian University of Science & Technology (NTNU), Norway

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Reliability Study of Multilayer Multistage Interconnection Networks Equipped with Internal Path Redundancy

Eleftherios Stergiou [(1)], Dimitrios Liarokapis[(1)] , Euripidis Glavas[(1)]
[(1)] Dept. of Computer Engineering, TEI of Epirus
Arta, Greece
e-mail: ster@teiep.gr, dili@teiep.gr, eglavas@teiep.gr

Georgios E. Rizos[(2)], D. C. Vasiliadis[(2)]
[(2)]Network Operations Center, TEI of Epirus
Arta, Greece
e-mail: georizos@teiep.gr, dvas@teiep.gr

***Abstract*—Multilayer multistage interconnection networks have introduced multiple parallel layers for enhancing performance metrics over traditional multistage interconnection networks. In this work, we propose an innovative multilayer multistage interconnection network fabric, which improves the switch efficiency by allowing multiple internal paths. The proposed fabric has demonstrated improved performance metrics compared to other existing fabrics and acceptable reliability in terms of fault tolerance. This type of fabric can considerably enhance the connection points of a modern network and improve its data flow. The configuration of the network allows for a conflict drop resolution mechanism or a classic backpressure blocking mechanism. This novel fabric can also handle multicast traffic, hotspot traffic or a combination of them.**

Keywords- Reliability; Multilayer Multistage Interconnection Networks; Quantitative Analysis; Multistage Architecture; Performance Evaluation.

I. INTRODUCTION

Multilayer Multistage Interconnection Networks (MLMINs) are devices that improve performance metrics when transferring data. The use of MLMINs avoids the necessity for the crossbar type of interconnection device, which is expensive to construct. However, although this shared-bus type of switching device is of low cost, it has low performance and therefore there is a need for interconnection devices, which strike a balance between efficient performance, reliability and reasonable cost. MLMINs were proposed by Tutsch and Hommel [1], after it was found in various studies that more switching power was needed in the last stages of a multistage interconnection network (MIN) than in the first stages [1]-[5].

The performance of MLMINs has been studied thoroughly by Garofalakis et al. [3] [6].

A typical MLMIN consists of an $N \times N$ MIN and $L = \log_k N$ stages and $k \times k$ or $k \times n$ switching elements (SEs), where k and n are the number of inlets/outlets of the SEs.

A typical MLMIN has two seriatim segments: the single-layer segment followed by the multilayer segment.

In the first segment, each stage consists of (N/k) SEs of size $k \times k$, while the last stage consists of SEs with size $k \times n$, where $n \geq k$.

In the second stage, however, the multilayer segment consists of $k \times k$ SEs in parallel rows. If the single layer segment has S stages, the second part has $(L-S)$ stages. The size of S is a matter of engineering choice, and depends on the degree of reduction to be implemented in the last stages of construction.

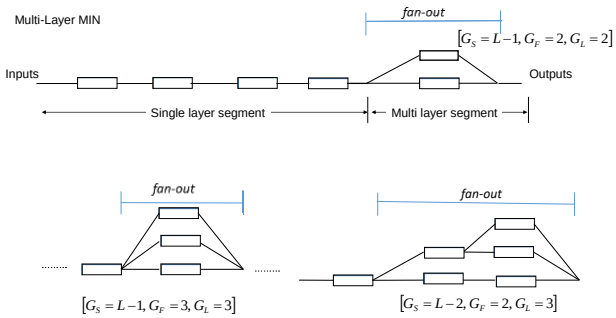


Figure 1. Schematic view of several MLMINs and corresponding definitions

In addition to the above, in order to accurately determine the structure of a MLMIN, three further parameters are required [3].

The first parameter is called the "Start replication factor" (G_S), and denotes the stage number at which the replication of layers starts. The second parameter is known as the "Growth factor" (G_F), and denotes the number of layers that can be developed at one stage by each SE. Finally, the third parameter is the "Layer limit factor" (G_L), and this denotes the maximum number of layer replications. Structures which can be described by the above three parameters are known as semi-layer MINs [6].

Figure 1 illustrates several MLMINs and gives the three factors that describe them. Of all the possible types of multilayer MINs, semi-layer MINs are the most suitable for use in the information technology industry due to the simplicity of their construction compared to other multilevel devices.

 ISBN: 978-1-61208-550-0

In addition to their performance characteristics, their fault tolerance or reliability is of interest to the scientific community, since the requirements of modern applications involve a constant demand for improved solutions [7, 9-10, 12-13].

In this work, in order to improve the behaviour of MLMINs even further, the number of possible paths between each pair (input-output) is multiplied. In this architecture, in cases of internal blocking, the system has the ability to send the packet via another, 'parallel', alternative route. This operation reduces the stacking of packets in internal buffers, increasing forwarding speed.

A series of multiplexers (MUXs) at the beginning of the fabric is introduced to implement this concept. This mechanism can dissipate the packets using multiple paths. The diffusion of packets increases (since more parallel paths are used) as the blocking phenomenon develops.

This forms the basic concept underlying the current work. To implement this idea, a new architecture is introduced which significantly raises both performance metrics and the reliability factor to acceptable values. Moreover, this system provides the ability to handle special load cases such as hotspot or multicast traffic.

Hence, the novel contribution of this work can be briefly summarized as:

- A new, productive and therefore promising MLMIN architecture is introduced with finite buffers, multiple internal routes and MUXs.
- A simple but detailed study of the proposed fabrics is carried out, showing that these fabrics have many benefits; the results from the proposed fabric in terms of metrics, the reliability factor and packet latency are encouraging.
- A comparative study of other similar modern architectures indicates that the new fabric outperforms these in permutation capability and has an acceptable level of reliability.

The remainder of this paper is organized as follows: In Section II, the proposed MLMIN fabric is introduced and several details of its operation (e.g., internal paths and routing) are presented. In Section III, the basic definitions are given and the necessary analysis carried out. In Section IV, the reliability and performance capability are analysed, and several numerical outcomes are described for the implemented MLMINs in terms of network size. Finally, in Section V, the conclusions and anticipated future work are presented.

II. THE PROPOSED FABRIC

A. Proposed fabric: Semi-layer MIN with internal path redundancy

To establish internal paths and to disperse the load uniformly, a preamble block composed of MUXs is located in front of the MIN. Thus, the proposed MIN consists of three segments in sequence, as presented below (Fig. 2).

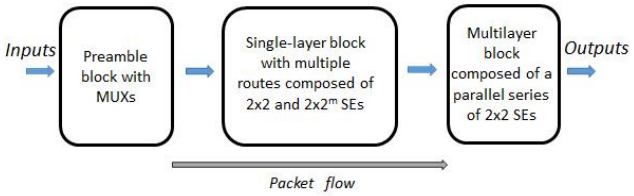


Figure 2: Fabric with internal path redundancy consisting of three blocks

1) Preamble of the MIN segment: In the preamble block, a column of $m \times 1$ MUXs are used to connect the inputs of the first stage of the MIN. Thus, a $N \times N$ network needs $(3N/2)$ multiplexers at the input stage (See Figure 3). This part is introduced because we want to have the possibility to make diffusion packages to alternative paths, when the first of them are busy.The main body of the MLMIN follows the preamble segment.

2) First segment of the MIN (single-layered segment): The first segment of the MIN contains only a single layer, which employs multiple internal paths (see Figure 3).

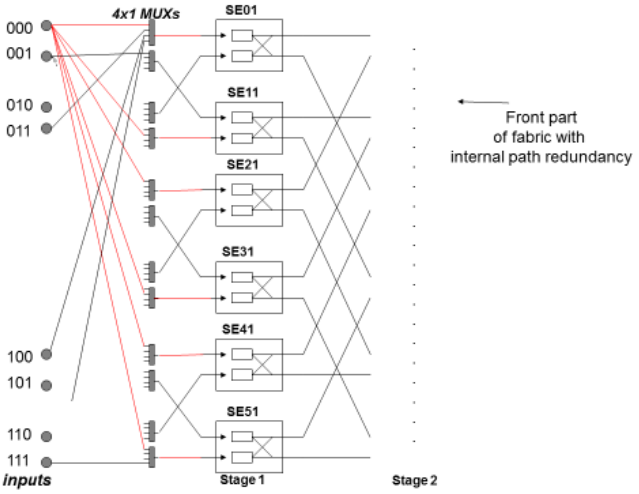


Figure 3: Detail of the connection between the preamble and single-layer segments, shown here for a network of diameter 3

A network with size $N \times N$ has $(\log_2 N - S)$ stages with $(3 \cdot N/4)$ switches per stage, where S is the stage number of the single-layer segment. This part consists of 2×2 switches, except for the final stage, which employs 2×2^m switches, where $m = 1, 2, \ldots$ depends on the fan-out configuration to which they are connected (Figure 4).

3) Second segment of the MIN (multilayered segment or fan-out): The multilayered segment or fan-out is accommodated in the second segment (Fig. 4). In the case of a double layer (one stage multiplied) the fan-out has a quantity $(2N/2) = N$ of 2×2 switches, while the triple-layered (one stage multiplied) network is composed of $(3N/2)$ 2×2 switches.

The number of SEs in a $N \times N$ network with one final stage multiplied can be calculated as $\left(\frac{3 \cdot N}{4}\right) \cdot \left(\log_2 N - 1\right) + N$, for a double-layered network, $\left(\frac{3 \cdot N}{4}\right) \cdot \left(\log_2 N - 1\right) + \frac{3}{2} \cdot N$, for the corresponding fabric with a triple-layered network and so on.

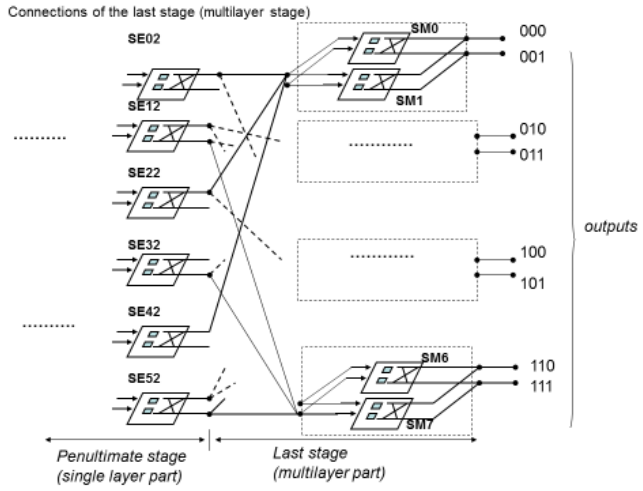


Figure 4: Detail of the connection between single layer and fan-out, shown here for a network of diameter 3

For simplicity, the proposed device is referred to here as SemiL (Semi-Layer MIN with multiple internal paths).

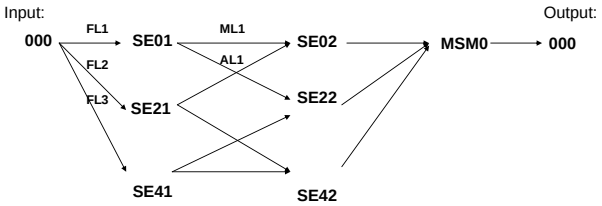
B. Multiple paths between two points

Fig. 5 shows all the possible routes between two points of the MIN (inlet: 000; outlet: 000), e.g., in a fabric with eight inputs and equal outputs. All alternative paths have the same length, which means that while the network is operating in real time, many equivalent actions for bypassing blocks are generated when faults arise in the system.

In standard $N \times N$ MLMINs, there are $2^{\log_2 N}$ distinct routes from one input point to all outputs, and thus a total of possible routes of $N \cdot 2^{\log_2 N}$ distinct paths. However, the equivalent novel MLMIN fabric, which contains internal redundancy paths, provides $\frac{3 \cdot N}{4} \cdot N \cdot 2^{\log_2 N}$ distinct paths. Hence, for example, a classic MLMIN with $N = 8$ provides 64 distinct paths, while the equivalent novel MLMIN fabric provides six times more (6x64) distinct paths.

C. Marking paths

Marking paths and establishing a hierarchy among the different paths aids in the implementation of an automatic routing mechanism. Taking an arbitrary input-output pair, then according to Fig. 5, packets can be routed from an input (e.g., 000) via three different SEs (SE01, SE21 and SE41). These links can be considered as the first link (FL) of the path.



1-path: 000 →SE01 →SE02 →MSM0 → 000

2-path: 000 →SE01 →SE22 →MSM0 → 000

3-path: 000 →SE21 →SE02 →MSM0 → 000

4-path: 000 →SE21 →SE42 →MSM0 → 000

5-path: 000 →SE41 →SE22 →MSM0 → 000

6-path: 000 →SE41 →SE42 →MSM0 → 000

Six different internal routes between a pair input-output.

Figure 5: Example of multiple paths for an input-output pair (000, 000) shown here for a network of diameter 3

Thus, FL1, FL2 and FL3 links are marked. Each request at Stage 1 has to choose one of these three links. Subsequently, from Stage 1 to the next stage, each request has two available links to choose from. The first is marked as the main link (ML), while the other can be marked as the auxiliary link (AL).

The same pattern of marking of links is continued for all the following stages of the single-layer segment. As the number of stages is increased, the number of distinct paths is multiplied. Thus, we have links marked MLn and ALn where n ϵ[1, +∞].

As a whole, this approach leads to distinct routes which constitute a hierarchy at every stage. Fig. 5 shows all the possible paths between an arbitrary input-output pair. In addition, it depicts the intermediate nodes (e.g., switches) that are involved in a specific input-output pair.

D. Routing technique and blocking mechanism

The clocking of the internal fabric guides the whole routing operation. The packets are forwarded synchronously at every stage in a parallel manner.

In MINs, routing and forwarding decisions are made independently for each packet by computation for each outgoing link. When a packet enters the fabric, it receives the 'routing address' (RA). The RA has as many bits as the number of stages in the MIN. The bits of the RA guide the packet-forwarding through the MIN, from SE to SE.

The RA is used by a packet in order to reach its unicast destination (including in the case of hotspot traffic). Moreover, in the case of multicast distribution, beyond the RA, a 'multicast tree number address' (MTNA) is assigned to each packet. Both types of address have as many bits as the number of stages in the MIN. The multicast address

 ISBN: 978-1-61208-550-0

indicates at which stage a packet has to act as a multicast operation (generating a copy packet), thus creating the so-called 'multicast tree' if this mechanism is implemented in the fabric.

The multistage construction described above can work either with backpressure or with a drop resolution mechanism. For simplicity, the accommodation of the backpressure technique in the system is considered here. This backpressure mechanism is appropriate in a network where systems employ non-deterministic protocols such as the UDP protocol; however, the back-pressing operation is completely unsuitable for deterministic protocols, such as the TCP protocol. Nowadays, although the backpressure mechanism is considered responsible for phenomena such as packet looping, large packet delays, and scalability, it continues to be implemented due to a lack of credible alternatives.

For the selection of an internal path which is not overburdened, the concept of backpressure weight per link may be introduced at the start of the device and constantly calculated for each input-output pair. A backpressure weight is an operation (function) based on local queue conditions, and gives information about the link state.

In contrast, the drop resolution mechanism, also known as the relaxing blocking mechanism, may be used when the switching element suffers from blocking; it sends back a request releasing a signal to the sender along a specific established route. In this case, this path is discarded at the end of the time cycle. The blocked request, now lost, repeats the same process (starting again from the beginning) until a new path is established.

III. BASIC DEFINITIONS AND ESSENTIAL ANALYSIS OF THE FABRIC

- *Probability of packet arrivals* (λ) represents the offered load at an input; in our experiment, the probability of packet arrivals is ranked from 0.1 to 1.0, with steps equal to 0.1. This probability can be expressed as λ_{Norm}.

  The arrival process for packets at the output queues of the first stage of the network is given by a binomial distribution $bin(c, \lambda / c)$, where λ is the fixed probability of a packet being generated by a processor at each cycle.
- The *arrival process of packets* at the output queues of Stage (i) (for $i = 2$ to $i = S - 1$ of the network, where S is the number of stages in the single-layer segment) is approximated by a binomial distribution $bin(c, u^{(i-1)} / c)$, where $u^{(i-1)}$ is the *utilization* of a queue (buffer) of Stage $(i-1)$, which we assume plays the role of the fixed probability of packets generated by processors at each cycle, feeding Stage (i). A similar forward technique is applied at the fan-out.
- *Buffer size* (b) represents the maximum number of packets that can be held by an input buffer of an SE.
- *Reliability* (r) of a component represents the probability of a switch component being operational. Each switch component has its own reliability. The reliability symbol can be denoted by $r_{type\,n}$, where type signifies the switch type and n depicts the number of gates of the switch. Thus, for example, r_{MUXn} and r_{DEMUXn} signify the reliability of the nx1 multiplexer and the 1xn de-multiplexer respectively, and r_{SE2x2}, r_{SE2x4} and r_{SE2x6} represent the reliabilities of SE_{2x2}, SE_{2x4} and SE_{2x6} correspondingly.
- *Average throughput* (T_{avg}) signifies the average number of packets accepted by all destinations per network cycle. Because the last stage of a single layer is never blocked, the *average throughput* can be expressed as $T_{avg} = u^{(s)}$, where u is the average utilization of the S stages.

  Moreover, T_{Norm} is the normalized throughput appearing at the end of the fabric.
- *Normalized packet latency* (D_{Norm}), of a MIN with L stages is defined as the ratio of average latency (D_{avg}) to the minimum packet delay. As minimum packet delay we consider the minimum number of time slots needed by a packet to be transmitted to its destination, i.e. when packets don't face any blocking during their routes. For L-stage MINs the minimum delay is equal to (L) time slots. Thus $D_{Norm} = D_{avg} / L$
- *The normalized packet loss probability* $\lambda_{loss(Norm)}$ of typical MLMINs (with S stages in a single layer segment), with or without multiple internal routes that apply the backpressure blocking mechanism at the inputs of the MINs before they enter the fabrics, can be expressed as

$$\lambda_{loss(Norm)} = \lambda^{(0)}_{loss(Norm)} = \lambda_{Norm} - T_{Norm} \quad (1)$$

  where λ_{Norm} depicts the normalized packet arrivals at the beginning of the fabric. This can also be written

$$\sum_{i=1}^{S-1} \lambda^{(i)}_{loss(Norm)} = 0 \quad (2)$$

  where $\lambda^{(i)}_{loss(Norm)}$ depicts the probability of packet loss in a queue at the i^{th} intermediate stage of a MIN.
- *Reliability analysis*: the reliability of a fabric is the ability to overcome all unexpected circumstances. Here, the point-to-point reliability (p-t-p reliability), also known as *terminal reliability*, is used. With p-t-

p reliability, the probability of at least one fault-free path existing between an input-output pair is considered. The reliability between an arbitrary given pair (inlet-outlet) is dependent on all the switch elements involved, insofar as the failure of each element implies the failure of the current routing action [11]-[13]. Supposing there is a switching subsystem with a series of n switch components, each with a reliability of r_i; then, the reliability of this subsystem can be calculated as

$$R_{n\,Switches\,in\,Series} = \prod_{i=1}^{n} r_i \quad (3)$$

On the other hand, for a subsystem with n parallel switch components, at least one of these must be active in order for this subsystem to operate successfully. In this case, the reliability of the subsystem is calculated as

$$R_{n\,Parallel\,Switches} = 1 - \prod_{i=1}^{n} (1 - r_i) \quad (4)$$

IV. RELIABILITY AND PERFORMANCE ANALYSIS

A. Reliability of the proposed fabric

For reliability analysis of the novel fabric, p-t-p reliability is used. As discussed above, in p-t-p reliability the probability of at least one fault-free path existing between an input-output pair is considered [See Figure 6]. The *total p-t-p reliability* R_{fabric} of an arbitrary input-output pair can be expressed by Equation (5)

$$R_{fabric} = R_{preamble} \times R_{SLMIN} \times R_{MLMIN} \quad (5)$$

Given the probability of a multiplexer and a SE being in normal operation (i.e., r_{MUX} and r_{SE} respectively are known), then the *total reliability of the fabric*, R_{fabric}, can be calculated as

$$R_{fabric} = r_{MUX} \times r_{SE}^{S} \times \left(1 - \prod_{i=1}^{L-S} (1 - r_{SE})^{G_L}\right) \quad (6)$$

,where S is the length of a single-layer segment and G_L is the maximum number of layer replications. The values in Equation (6) are higher than the corresponding values of the conventional equivalents such as banyan- or delta-type MINs with the same *network size* $\left(R_{MIN} = r^{\log_2 N}\right)$.

The *p-t-p reliability* for a route within this novel fabric (2SemiL and 3SemiL cases with two and three layers respectively) can be illustrated as in Figure 7.

For a given probability of a multiplexer and a SE being in normal operation (r_{MUX} and r_{SE} respectively) the R_{fabric} of a SemiL network with fan-out in the final stage can be expressed as

$$R_{SemiL} = r_{MUX} \cdot r_{SE2x2}^{(\log N - 2)} \cdot r_{SE2xn} \cdot \left(1 - \left(1 - r_{SE2x2}\right)^{G_L}\right) \quad (7)$$

,where G_L depicts the maximum number of layer replications, G_L has values of 2 and 3 for 2SemiL and 3SemiL networks respectively, and n in r_{SExn} is equal to 4 and 6 respectively.

The *p-t-p reliability* has been given in earlier studies in the literature for various types of MIN architectures (e.g., Pars, Augmented Shuffle Exchange Network (ASEN) and Augmented Baseline Network (ABN) [12] [14][15].

For Pars networks, the *reliability* is given in [8] as follows

$$R_{Pars} = r_{MUX} \cdot r_{SE2x2}^{(\log_2 N - 2)} \cdot \left(1 - \left(1 - r_{SE2X2} \cdot r_{DEMUX}\right)^2\right) \quad (8)$$

Figure 7 shows the total input-output reliability versus the reliability (r) of each component for the novel fabric (SemiL networks) with $G_L = 3$.

With adjacent groups of bars corresponding to various values of a component's reliability r, the graph shows a gradual increase in the *total p-t-p reliability* for networks with numbers of inlets/outlets N gradually increasing (where $N = 2^k$, $k = 3,\ldots,8$).

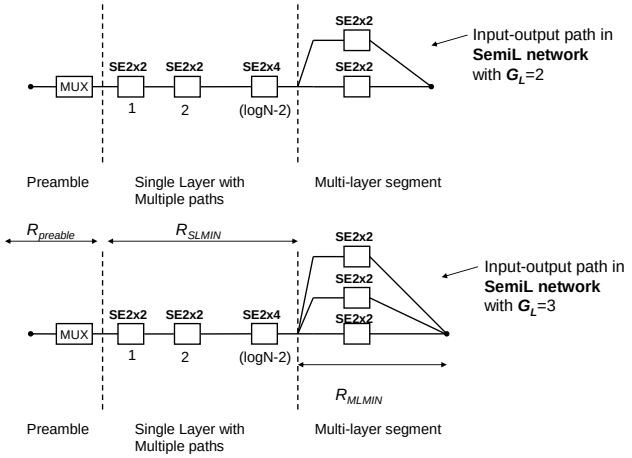


Figure 6: Typical point-to-point paths in 2 and 3SemiL networks

For low values of component factor r (e.g., r=0.9) it can clearly be seen that the total p-t-p reliability for networks with high network diameter (e.g., size N=256) is greatly reduced.

When the factor r tends to 1, all the fabrics, regardless of network diameter (size), tend to have the maximum p-t-p reliability of value 1.

 ISBN: 978-1-61208-550-0

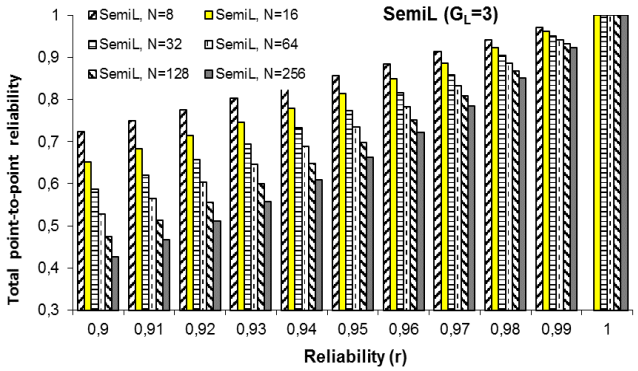


Figure 7: Total path reliability versus reliability of components (r) for various network diameters

Fig. 8 shows the total input-output reliability for various types of MIN. With adjacent groups of bars corresponding to given values of the component reliability (r), the two left-hand bars of each group represent the total p-t-p reliability of MLMINs (SeLMINs) that do not use internal paths and have replication factors of 3 and 2 respectively.

The two right-hand bars of each group depict the total p-t-p reliability of MLMINs (SemiLs) that include internal paths and have replication factors of 3 and 2 respectively. The central bar represents the p-t-p reliability of a Pars network.

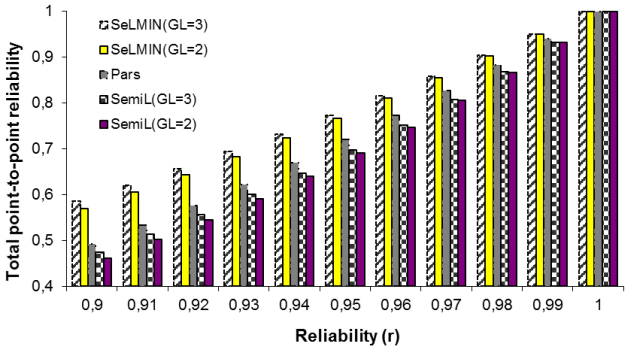


Figure 8: Total path reliability for various types of MIN versus reliability of components (r)

It can be seen that the novel fabric is not the most effective in terms of p-t-p reliability. Fortunately, this weakness is improved if the selected components have a reliability tending to a value of 1.

B. *Performance capability*

In most cases, the performance of MINs is estimated using analytical approaches [2]; in the remaining cases, simulation is used [4]-[6]. In this study, network performance was calculated using simulation. For this simulation, a special-purpose simulator was developed to evaluate the overall network performance of the MLMINs with internal path redundancy as well as without redundancy paths. This tool was developed in C++, and is capable of operating under various configuration scenarios and handling uniform traffic. It operates with various input parameters such as buffer size, number of inlets/outlets, offered load, number of stages, and number of layers in the last segment of the fabric. Each SE was modelled using an array of non-shared buffer pairs of queues. Each queue operates on a FIFO (first input, first output) principle and each buffer is considered to be empty initially. In the same way, the simulation was carried out for single-layer MINs, and the results were used for comparison with the corresponding fabrics with redundant paths.

All simulation experiments were performed at packet level, assuming fixed-size packets transmitted in fixed-size timeslots, where the timeslot cycle is defined as the time required by a packet to be transferred from one stage to the next. All packet contentions, which occur when two packets claim the same next point, are resolved randomly.

Figure 9 represents the increments of *normalized packet latency* of various MLMIN constructions with a network size of 8 (that is, a network diameter of value 3) and various values of uniform type *offered load* (varying from 10–100%).

The two upper solid curves depict the *normalized latency* of a single-layer MIN (SiLMIN) with *buffer sizes* of 2 and 3 respectively. The dotted curves represent the same quantity for semi-layer MINs (SeLMINs) without internal path redundancy, with *buffer sizes* of 2 and 3 respectively and a *replication factor* of 3.

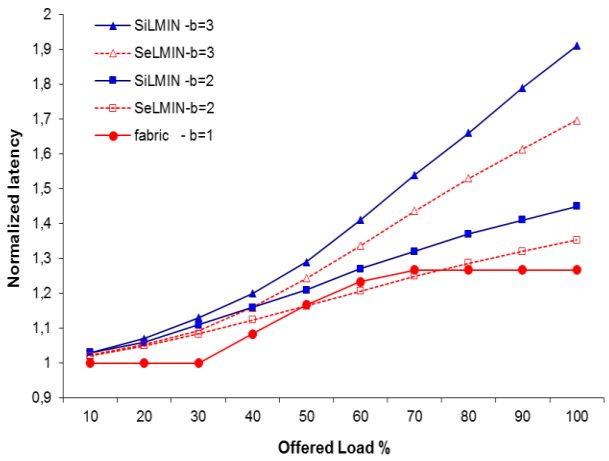


Figure 9: Normalized packet latency versus scalable offered load for various MIN architectures with a network diameter of 3

Finally, the lowest solid curve depicts the packet latency of the novel fabric (with internal paths redundancy, also referred to here as SemiL) with buffer size equal to 1 and replication factor of 3.

This diagram demonstrates that when the *offered load* exceeds a value of 70%, the novel fabric presents the smallest *packet latency* in comparison with other devices which are equivalent in terms of *network size*. This improvement in latency can be explained as follows. When a heavy load is required to be transferred, this leads to a large

 ISBN: 978-1-61208-550-0

number of blocks (and particularly in the last stages). As a consequence, this leads the fabric to use a greater number of alternative routes, thus relieving the blocking phenomenon. The conclusion which can clearly be drawn from this plot is that the novel fabric is the most efficient choice in terms of packet delay, in comparison with other similar architectures that do not include internal paths.

V. CONCLUSION

An innovative MLMIN is presented in this paper. The proposed fabric is composed of a preamble containing MUXs, a single-layer MIN with internal path redundancy implementation, and a fan-out in the final segment. This structure is composed of 2×2 and 2×2^m switches, depending on the replication factor. Multiple paths for each input-output pair are provided by the fabric. This allows network connections to handle the traffic more efficiently. The current preliminary but careful study indicates that the proposed MLMIN architecture outperforms conventional MLMINs in terms of performance metrics and offers acceptable values of reliability and fault tolerance.

This work has many possibilities for extension. For example, this system requires study under conditions of hotspot traffic or a multicast service, and should also be evaluated in terms of performance metrics, priorities and cost metrics when operating with a backpressure or relaxing mechanism.

REFERENCES

[1] D. Tutsch and G. Hommel, “MLMIN: A Multicore Processor and Parallel Computer Network Topology for Multicast”, Computers and Operations Research Journal. Elsevier, Vol. 35, No. 12, December 2008, pp. 3807-3821

[2] J. Garofalakis and E. Stergiou, “An Analytical performance model for multistage interconnection networks with blocking”, Proc. Of CNSR 2008, May(2008), Proc. pp. 373-381

[3] J. Garofalakis and E. Stergiou, “Mechanisms and analysis for supporting multicast traffic by using Multilayer Multistage Interconnection Networks”, International Journal of Network Management, Volume 21, Issue 2, 2011, pp. 130-146

[4] D. C. Vasiliadis, G. E. Rizos, and C. Vassilakis, “Performance Analysis of blocking Banyan Switches”, Proceeding of the IEEE sponsored International Joint Conference on Telecommunications and Networking CISSE 06(2006), pp. 107-111

[5] D. C. Vasiliadis, G. E. Rizos, C. Vassilakis, and E. Glavas, “Performance Evaluation of Multicast Routing over Multilayer Interconnection Networks”, Proceedings of the Fifth Advanced International Conference on Telecommunications (AICT 2009), pp. 395 - 403

[6] E. Stergiou and J. D. Garofalakis, “A Simulation study for optimizing the performance of Semi-layered Delta Networks”, Proceeding of 1st International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH - 2011), 29-31 July, Noordwijkerhout, Netherlands 2011, pp. 257–265

[7] V. P. Bhardwaj and N. Nitin, "Message broadcasting via a New Fault Tolerant Irregular Advance Omega Network in Faulty and Non faulty Network Environments." Journal of Electrical and Computer Engineering 2013 (2013), pp. 1–17

[8] F. Bistouni and M. Jahanshahi, “Pars network: A multistage interconnection network with fault-tolerance capability”, Journal of Parallel and Distributed Computing, Elsevier, Volume 75, January 2015, pp. 168–183

[9] F. Bistouni and M. Jahanshahi, "Analyzing the reliability of shuffle-exchange networks using reliability block diagrams." Reliability Engineering and System Safety 132 (2014): pp. 97-106

[10] M. Jahanshahi and F. Bistouni, "A new approach toimprove reliability of the multistage interconnection networks." Computers and Electrical Engineering 40.8 (2014): pp.348-374

[11] M. Jahanshahi and F. Bistoumi, “Improving the reliability of the Benes network for use in large-scale- systems”, Microelectronics Reliability, 55.3(2015), pp. 679-695

[12] S. Rajkumara and N. K. Goyala, “Review of Multistage Interconnection Networks Reliability and Fault-Tolerance”, Technical Review, Taylor and Francis, 26 Oct 2015, pp. 223-230

[13] N. A. M. Yunus, M. Othman, Z. M. Hanapi, and K. Y. Lun, “Reliability Review of Interconnection Networks”, IETE Technical Review, Taylor and Francis, 26 Jan 2016, pp. 596-606.

 ISBN: 978-1-61208-550-0

Reliability and Quality of Service of an Optimized Protocol for Routing in VANETs

Samira Harrabi
ENSI University/COSMOS laboratory
Mannouba, Tunisia
Email: samira.harrabi@gmail.com

Ines Ben Jaffar
ESCT University
Mannouba,Tunisia
Email: ines.benjaafar@gmail.com

Khaled Ghedira
ISG University
Tunis, Tunisia
Email: khaled.ghedira@anpr.tn

***Abstract*—Vehicular Ad hoc NETworks (VANETs) are a special kind of Mobile Ad hoc NETworks (MANETs), which can provide scalable solutions for applications such as traffic safety, internet access, etc. To properly achieve this goal, these applications need an efficient routing protocol. Yet, contrary to the routing protocols designed for the MANETs, the routing protocols for the VANETs must take into account the highly dynamic topology caused by the fast mobility of the vehicles. Hence, improving the MANET routing protocol or designing a new one specific for the VANETs are the usual approaches to efficiently perform the routing protocol in a vehicular environment. In this context, we previously enhanced the Destination-Sequenced Distance-Vector Routing protocol(DSDV) based on the Particle Swarm Optimization (PSO) and the Multi-Agent System (MAS). This motivation for the PSO and MAS comes from the behaviors seen in very complicated problems, in particular routing. The main goal of this paper is to carry out a performance evaluation of the enhanced version in comparison to a well-known routing protocol which is the Intelligent Based Clustering Algorithm in VANET (IBCAV). The simulation results show that integrating both the MAS and the PSO is able to guarantee a certain level of quality of service in terms of loss packet, throughput and overhead.**

Keywords–VANET; MAS; PSO; Routing; Quality of service; Routing protocol.

I. INTRODUCTION

In recent years, the progresses in wireless mobile networks have permitted the emergence of a new type of networks, named Vehicular Ad hoc NETworks (VANETs). The VANETs arose from a special form of Mobile Ad hoc NETworks (MANETs) [1]. This particular kind of networks is developed as a main component of Intelligent Transportation Systems (ITS) in order to enhance driving, passengers safety and comfort [2].

The VANETs are formed by vehicles equipped with On Board Units (OBU), and a fixed infrastructure called Road Side Units (RSU). Both units have wireless communication abilities. In fact, the OBUs can communicate with each other as well as with the RSUs in an ad hoc way. Principally, as depicted in Figure 1, there are two types of communications modes in vehicular networks which are: Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I).

Although the VANETs have a lot of similarities with the MANETs as their low bandwidth, their short radio transmission range, and their omni-directional broadcast in most scenarios, they differ from ad hoc networks in numerous ways. Indeed, the vehicular networks are characterized by the rapid changes in communications links.

In addition, frequent disconnections between nodes can occur due to low density [3]. Therefore, designing an efficient protocol for routing in vehicular networks seems to be a key challenge created by the above properties [4][5]. Moreover, applying the MANETs routing protocols in vehicular environments is inefficient [6], since these approaches do not take the above-mentioned characteristics into account. Thus, modifying these methods or proposing new protocols specific for the VANETs are the usual solutions to efficiently resolve the routing challenge in the VANETs. Aiming to solve the routing problem in vehicular networks, we formerly enhanced in [7][8] the DSDV protocol based on the PSO and the MAS. The improved version is called *PSO-C-MADSDV*.

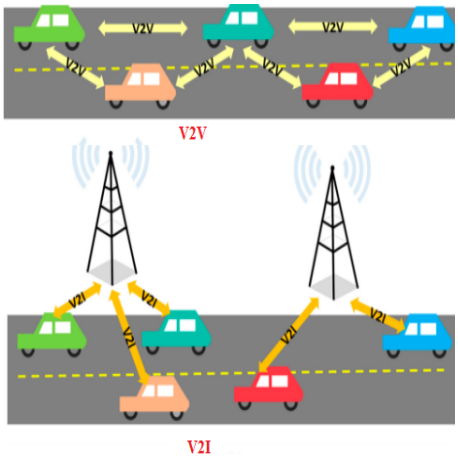


Figure 1. Communications mode in vehicular networks.

The remainder of this paper is structured as follows. The Section II underlines and describes the challenge in routing for the VANETs. Besides, it proves the limitation of applying the MANETs protocols for vehicular scenarios. The Section III presents some related works that deal with routing in vehicular networks. Also, it sums our proposed PSO-C-MADSDV routing method. Finally, in Section IV, we present the simulations results obtained regarding packets losses, throughput and overhead. There is a comparison between the PSO-C-MADSDV and the IBCAV protocols. At last, the section V gives conclusions and future works that may arise.

II. ISSUES OF ROUTING IN VANETS

Routing is defined as the task of forwarding a data packet from a source node to its destination. Sometimes, this process requires multi-hop forwarding nodes. To this end, finding the

 ISBN: 978-1-61208-550-0

routes to deliver the packets to their destination is the role and the responsibility of routing protocols. In general, an efficient routing protocol is one that is able to forward packets with a short rate of dropped packets and provide a minimal amount of the overhead.

Unlike the routing protocols designed for the MANETs, the routing protocols for the VANETs must principally take into consideration the highly dynamic topology [9]. Consequently, applying traditional MANET routing protocols in vehicular networks is inefficient. Hence, modifying or improving the MANET protocols is the usual requirement to resolve expeditiously the routing challenge in the VANETs. In fact, to better understand this challenge brought by the VANETs, it is necessary first to analyse the specific features of these networks.

The VANETs are a very dynamic environment since they are formed with vehicles that join and leave the network all the time. Even though they have many similarities with the MANETs, like their short radio transmission range and their low bandwidth, they possess some particular characteristics making them different from the ad hoc networks in several aspects.

Actually, the VANETs are characterized firstly by their quick changes in network topology. Secondly, the link between vehicles may be interrupted frequently mainly because of the low density of vehicles. Finally, out of the networking aspects, the different applications that are expected to run over the VANET make it a unique environment. Also, they pose interesting questions related to the protocol design. Consequently, in the literature there are different ways to address the routing challenges in the VANETs.

III. RELATED WORK

Recently, numerous studies reported in the literature have dealt with routing in the VANETs. As discussed above, the specific characteristics make routing a big challenge that requires to be solved in the vehicular environments. Indeed, the MANETs protocols have proved that their performance is poor in the VANETs [10][11].

The key problem with these routing methods ((Ad hoc On-demand Distance Vector (AODV)[12], Dynamic Source Routing(DSR)[13], etc.) in VANETs scenarios is the route instability. In fact, due to the high mobility of vehicles, the paths that have been established as fixed succession nodes, can be interrupted frequently. As a result, this interruption increases overhead, minimizes the rate of delivery ratios, and growths as well the delays of transmission data.

As illustrated in Figure 2, when the vehicle V1 moves out of the transmission range of the source node, the path (Vs, V1, Vd) created at time t will be broken at the instant t+Dt. To solve this problem, an alternative solution is given by the geographical routing protocols (e.g., Greedy Perimeter Stateless Routing (GPSR)[14]). This category does not establish routes, but it utilizes the geographical position of the destination node and its neighbor to deliver data.

Differently from the node-centric routing, the geographical routing approaches have the advantage that any mobile node ensuring progress to the given destination can be applied to forward data.

Thus, in Figure 3, to deliver data to the destination node Vd, the node V2 can be used instead of the node V1. Even with a better route stability, the geographical routing methods

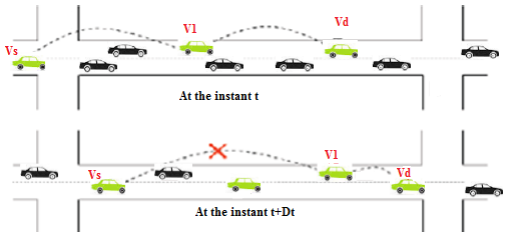


Figure 2. Mobility problem in VANETs.

do not perform well in scale scenarios [10][15]. In this case, their main problem is that many times are needed to look for a next hop (i.e., a node closer to the destination than the current node).

Accordingly, since in the VANETs the vehicle movements are more constrained on roads rather than a geographical region [10], the wrong road routes that do not lead to the destination can be selected. In addition, packets can be transferred to dead ends providing unnecessary and extra traffic overhead in the network as well as longer delays for packets.

Instead of routing data on the dotted route, geographical forwarding delivers data to V1 and V2, following the shortest geographical route from Vs to Vd on a dead path, as shown in Figure 3.

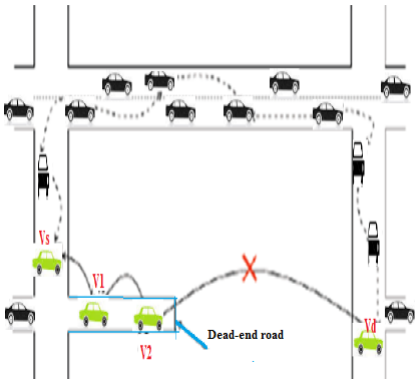


Figure 3. Drawback of geographical forwarding approach.

To address this limitation, numerous road-based routing protocols [10][11][15][16] have been proposed. These methods forward packets based on the shortest road path between the source node and destination. However, [10][16] did not take into account the vehicular traffic flow.

As demonstrated in Figure 3, it is possible that the paths segments on the shortest roads are empty or have network fragmentations. To alleviate this issue, other routing approaches were given in [15][17][18][19]. The purpose of these projects was to use some historical data concerning average daily/hourly vehicular traffic flows.

Unfortunately, this data was not an accurate indicator of the actual road traffic conditions, as events such as road constructions or traffic redirection were not rare. In order to solve the routing challenge in the VANETs, other studies were published [20][21][22][23]. The idea of those related works was to improve the MANETs protocols to make them suitable in a vehicular environment. In this context, we focused in

 ISBN: 978-1-61208-550-0

[7][8] on enhancing the DSDV protocol.

In fact, during the process of designing and deploying a VANET, various questions must be answered that pertain to protocol performance and usefulness. For instance, when designing a routing protocol, a key question is: How can we integrate the VANETs features (road topology, real-time road traffic flow, presence of building, etc.) for better performance? What is the best way to integrate them? All these questions and many more require knowledge of the topological characteristics of the VANET, which are addressed in [7].

The responses have been given based on a multi-agent system approach. A MAS is composed of a collection of autonomous software agents which are capable of completing desired goals cooperatively. The basic attributes of an agent that are considered typical are autonomy, learning and cooperation [24]. These properties imply that agents are capable of executing independently from any other control and possibly asynchronously, discover relevant knowledge from the environment and other agents that may help in attaining the desired goals, and work cooperatively and competitively with other agents. In addition, when performing message routing, a key question is: Which are the highest-quality vehicles? The forwarding process would lead to an optimal communication cost with a minimal number of rebroadcasts so as to reduce latency and packet loss.

Particle Swarm Optimization (PSO) [25] is a stochastic optimization technique, inspired by the idea of a flock of birds moving towards a final goal through cooperation as well as independent exploration. The underlying phenomenon of PSO is that knowledge is optimized by social interaction in the population. The PSO searches for the optimal solutions by updating the velocity and the position of each particle.

Motivated by the performance of the PSO algorithm, in [8], our published research work mainly concentrated on optimizing the routing quality of service. Therefore, we have made an attempt to enhance the routing performance in terms of throughput, packet loss and overhead based on the clustering approach [26]. This technique helps the protocol to minimize the messages count and to increase the network connectivity. It also makes the communication more secure and more stable.

Nevertheless, the previous paper did not compare the PSO-C- MADSDV to any routing protocol specifically designed for the VANETs to see which manner (modifying the traditional methods or proposing new protocols) is the most efficient to deal with the routing issue.

IV. STUDY AND PERFORMANCE COMPARISON

In this section, we investigate the routing protocols for the VANETs. To evaluate the performance of our proposed approach and to demonstrate the usefulness of the agent technology, as well as the PSO algorithm, we chose to compare our method with the Clustering Algorithm in the VANET (IBCAV) [27].

The IBCAV seeks to enhance the routing performance in the VANETs by employing inter-layered methods, as well as the awareness of the network traffic flow. It combines several factors using a smart method on the basis of an artificial neural network. The clustering technique was also applied. In fact, the cluster size, speed and density of vehicles are the metrics taken to form a cluster. For a header selection, the IBCAV combines the factors utilizing the Genetic Algorithm (GA)[28].

The selected protocols were evaluated through simulation using some performance metrics. Hence, in this part, we first present the used metrics. Second, we analyze the obtained results.

A. Analyzed Metrics

In a highly mobile environment as the VANETs, characterized by frequent topology changes, the major routing problem is the breaking of links, which can cause packet loss. The metrics used to assess the performance are the following:

- **Rate of dropped packet**: It presents the number of the data packets having failed to reach the destination.
- **Throughput**: It sums the data packets produced by each source node, counted by kbit/s.
- **Routing overhead**: This metric is utilized to measure the effectiveness of the routing protocol. Indeed, it is determined as the total number of additional routing packets per the number of unique data packets received at destinations. Moreover, this parameter counts the extra traffic produced by the protocol for successfully transmitted packets.

B. Simulation Results

This section makes an attempt to evaluate the performance of the PCO-C-MADSDV and the IBCAV over low, medium and high density with a node mobility speed of 30m/s. The evaluation is done using the DARS simulator (Dynamic Ad-Hoc Routing Simulator)[29]and the JADE framework [30]. The simulation parameters are listed in Table 1.

TABLE I. simulation parameters.

Parameter	*value*
Transmission rate	54Mbps.
Simulation time	50s.
Playground Dimensions	1300m x 700m.
Routing protocols	PSO-C-MADSDV and IBCAV.
Transmission range	150m.
Number of nodes	30.
Mobility Model	Random Waypoint Model[31]
MAC layer	8012.11p

We first present the obtained results in terms of dropped packet rate. After that, we analyze the performance of both routing methods in terms of throughput. Finally, we demonstrate the impact of nodes density on previous protocols according to the routing overhead.

- **Rate of dropped packet**: The graph in Figure 4 demonstrates the obtained results regarding the average of packet loss ratio. As it can be seen, the number of dropped packets in both approaches with low density (10 to 20) is nearly the same and slightly goes up with the increase in vehicles density. However, in medium and high density the IBCAV protocol drops much more packets compared to the PSO-C-MADSDV.
  For example, at 30 vehicles, the IBCAV suffers a loss of 8.12%, whereas our approach suffers a loss of 5.2%.
  In addition, the best behavior of the PSO-C-MADSDV is more noticeable when the number of vehicles grows to reach 50 nodes. In this scenario, the IBCAV protocol drops about 21% of the delivered packets while the PSO-C- MADSDV is more efficient and loses 17%.

 ISBN: 978-1-61208-550-0

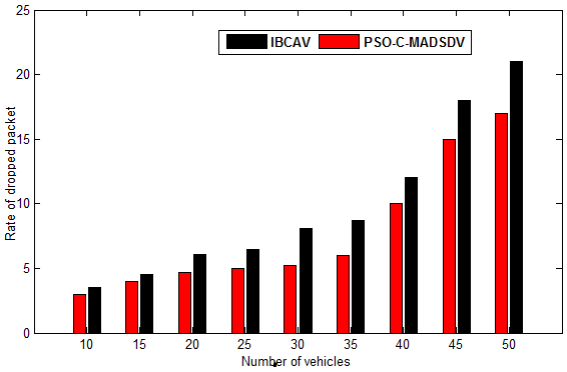


Figure 4. Analysis of dropped packet rate.

To sum up, for all scenarios, the PSO-C-MADSDV outperforms the IBCAV protocol. This is thanks to the PSO algorithm which converges quickly to the best and optimal solution. As a consequence, the probability of producing path breakages decreases. This increase is also guaranteed by the benefits of the agent paradigm, more particularly the autonomy that makes it possible to establish a link despite the topological change.

- **Throughput** : Figure 5 depicts the corresponding throughput obtained for both IBCAV and PSO-C-MADSDV protocols. From the plotted results, we can observe that the PSO-C-MADSDV achieves greater throughput compared to the IBCAV scheme, especially with high-density scenarios.

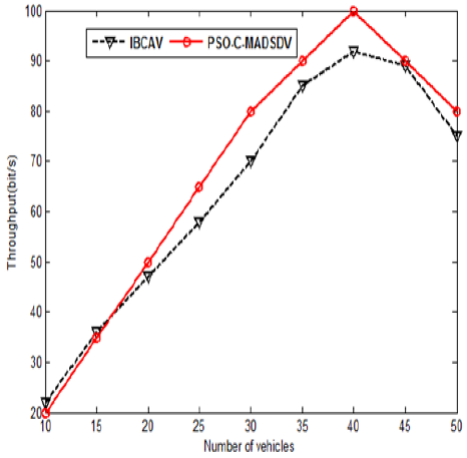


Figure 5. Analysis of throughput.

The main reason for this behavior is that the PSO-C-MADSDV does not require extra time to look for the paths. Whereas, for the IBCAV, there is some spent time in which the protocol does not forward packets. Consequently, the throughput declines.

- **Routing overhead**: Considering the obtained results indicated in Figure 6, we can see that the IBCAV protocol produces the highest rate of routing traffic into the network compared to the PSO-C- MADSDV. This observation is valid for all density levels.

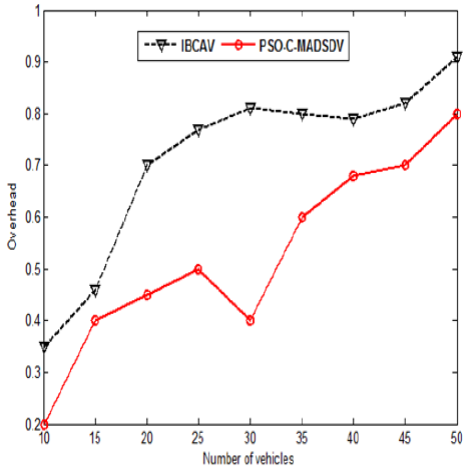


Figure 6. Analysis of overhead.

The reason for this superiority is the cluster technique used by the PSO-C- MADSDV, which can it more stable against the link failure compared to the IBCAV. This makes it more efficient as it avoids sending unnecessary packets.

V. Conclusion and Future Work

Thanks to the advances in wireless technology, it is possible to form a network using vehicles, called VANET. It is a particular class of the MANET. Nevertheless, the high dynamic nature of the vehicular network, caused by the high speed of vehicles, makes it different from the MANET and various challenges arise, especially the routing issue. Therefore, to solve this problem, it is essential to design a new protocol taking the mobility model into account or to improve the MANETs routing protocols to suit the VANETs nature.

In this paper, an attempt has been made to provide a comparative analysis of two routing protocols, which are the PSO-C-MADSDV and the IBCAV. The first one is an enhanced version of the traditional MANETs protocol, whereas the IBCAV is a specific routing scheme designed for the VANET. The key aim of our comparative study is to identify the way that has a better performance, taking place in highly mobile environment of the VANET.

For the simulation results, we can observe that our proposed PSO-C-MADSDV approach outperforms the IBCAV in terms of throughput, packet drop and routing overhead. Hence, considering the obtained results, we can conclude that the designed protocols for routing in the VANETs need to be improved to adapt well in some real-time scenarios.

As a future plan, we envision to evaluate the PSO-C-MADSDV in different scenarios (i.e., city, urban, rural, etc.) to test the impact of varying the communication environment on its performance.

References

[1] S. Yousefi, M. Siadat Mousavi, and M. Fathy, Vehicular Ad Hoc Networks (VANETs): Challenges and Perspectives, 6th International Conference on ITS Telecommunications Proceedings, 2006, Pp 761-766.

[2] H. H. Hartenstein and K.P. Laberteaux, A tutorial survey on vehicular ad hoc networks", IEEE Communications Magazine, vol. 46, no. 6, June 2008, pp. 164-171.

[3] Y. Kim, Security Issues in Vehicular Networks,"pp.468-472,IEEE 2013.

 ISBN: 978-1-61208-550-0

[4] J. J.Blum , A.Eskandarian, and L. J. Hoffman, " Challenges of intervehicle ad hoc networks, IEEE Trans. Intelligent Transportation Systems, Vol. 5, No. 4, 2004, pp.347-351.

[5] S. Harrabi, I. Ben Jaafar, and K. Ghedira, Routing Challenges and Solutions in Vehicular Ad hoc Networks, Sensors and Transducers (ISSN: 2306-8515, e-ISSN 1726-5479), Vol. 206, Issue 11,November 2016, pp31-42.

[6] P. Ranjan and K. K. Ahirwar, Comparative Study of VANET and MANET Routing Protocols, in Proceedings of the International Conference on Advanced Computing and Communication Technologies (ACCT 11), January 20-22, 2011, pp. 517-523.

[7] S. Harrabi, W.Chainbi, and K.Ghdira, A multi-agent proactive routing protocol for Vehicular Ad-Hoc Networks. Proc.of The 2014 International Symposium on Networks, Computers and Communications (ISNCC 2014) , Hammamet- Tunisia, 17-19 June 2014. http://dx.doi.org/10.1109/SNCC.2014.6866523.

[8] S. Harrabi, I. Ben Jaafar, and K. Ghedira, Novel Optimized Routing Scheme for VANETs, The 7th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2016), September 19-22, 2016, London, United Kingdom.

[9] Ch. S. Raju, M. Sailaja, Ch. Balaswamy, Adaptability of MANET Routing Protocols for VANETS, International Journal of Advanced Research in Computer and CommunicationEngineering, 2, 7, July 2013, pp. 2823-2829.

[10] C. Lochert et al. A routing strategy for vehicular ad hoc networks in city environments, in Proceedings IEEE Intelligent Vehicles Symposium, Columbus, OH, USA, June 2003, pp. 156-161.

[11] V. Naumov and T. Gross, Connectivity-aware routing (CAR) in vehicular ad hoc networks, in Proceedings IEEE International Conference on Computer Communications, Anchorage, AK, USA, May 2007, pp. 1919-1927.

[12] C. E. Perkins and E. M. Royer, Ad hoc on-demand distance vector routing, in Proceedings 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, USA, February 1999, pp. 90-100.

[13] D. B. Johnson and D. A. Maltz Dynamic source routing in ad hoc wireless networks, Mobile Computing, vol. 353, no. 5, pp. 153-161, 1996.

[14] B. Karp and H. T. Kung, GPSR: greedy perimeter stateless routing for wireless networks, in Proceedings 6th International Conference on Mobile Computing and Networking, Boston, MA, USA, August 2000, pp. 243-254.

[15] T. Li, S. K. Hazra, and W. Seah, A position-based routing protocol for metropolitan bus networks, in Proceedings IEEE 61st Vehicular Technology Conference VTC-Spring, Stockholm, Sweden, June 2005, pp. 2315-2319.

[16] J. Tian, L. Han, K. Rothermel, and C. Cseh, Spatially aware packet routing for mobile ad hoc inter-vehicle radio networks, in Proceedings IEEE Intelligent Transportation Systems, Shanghai, China, October 2003, pp. 1546-1551

[17] M. Jerbi, R. Meraihi, S.-M. Senouci, and Y. Ghamri-Doudane, GyTAR: improved greedy traffic aware routing protocol for vehicular ad hoc networks in city environments, in Proceedings 3rd ACM International Workshop on Vehicular Ad Hoc Networks (VANET), Los Angeles, CA, USA, September 2006, pp. 88-89.

[18] H. Wu, R. Fujimoto, R. Guensler, and M. Hunter, MDDV: A Mobility-Centric Data Dissemination Algorithm for Vehicular Networks, in Proceedings 1st ACM International Workshop on Vehicular Ad Hoc Networks (VANET). Philadelphia, PA, USA: ACM, October 2004, pp. 47-56.

[19] J. Zhao and G. Cao, Vadd: Vehicle-assisted data delivery in vehicular ad hoc networks, IEEE Transactions on Vehicular Technology, vol. 57, no. 3, May 2008, pp. 1910-1922.

[20] T. Kaur and A. K. Verma, Simulation and Analysis of AODV routing protocol in VANETs, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012, pp. 293-301.

[21] C. Perkins, E. Belding-Royer, and S. Das, RFC 3561-ad hoc on-demand distance vector (AODV) routing, Internet RFCs,2003 pp. 138.

[22] A. Moravejosharieh, H. Modares, R.Salleh, and E. Mostajeran, Performance Analysis of AODV, AOMDV, DSR, DSDV Routing Protocols in Vehicular Ad Hoc Network, International Science Congress Association, Vol. 2(7), July (2013),pp.66-73.

[23] H. Saini and R.Mahapatra, Implementation and Performance Analysis of AODV Routing Protocol in VANETs, International Journal of Emerging Science and Engineering (IJESE) ISSN: 23196378, Volume-2, Issue-3, January 2014, pp. 24-29.

[24] M. Wooldridge and N.R.Jennings, Intelligent agents: Theories, Architectures and Languages, January 1995. Lecture Notes in Artificial Intelligence, vol. 890, ISBN 3-540-58855-8.

[25] R. C. Eberhart and J.Kennedy, Particle Swarm Optimization, In Proc. of IEEE International Conference on Neural Networks, 1995, pp. 1942-1948.

[26] T. Priyanka and T. P. Sharma, A Survey On Clustering Techniques Used In Vehicular Ad Hoc Network, Proceedings of 11th IRF International Conference, 15th June-2014, Pune, India, ISBN: 978-93-84209-27-8

[27] M. Mottahedi, S.jabbehdari, and S.Adabi, IBCAV: intelligent based clustering algorithm in vanets, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, (January 2013), pp. 538-543.

[28] D.Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989, pp. 1-25.

[29] DARS, URL : http://dynamic-ad-hoc-routing-simulator-dars.soft112.com/ [accessed: 2017-01-14.

[30] F. Bellifemine, A. Poggi, and G. Rimassa,Developing multi- 365 agent systems with a FIPA-compliant agent framework. Software-Practice and experience. Vol. 31 Issue 2, Pp 103-128, February 2001.

[31] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, A performance comparison of multi-hop wireless ad hoc network routing protocols, in Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking(Mobicom98), ACM, October 1998.

 ISBN: 978-1-61208-550-0

Regional Comparisons of Critical Telecommunication Infrastructure Resiliency Based on Outage Data

Andrew P. Snow, John C. Hoag
School of Information and Telecommunication Systems
Ohio University
Athens, Ohio USA
Email: snowa@ohio.edu; hoagj@ohio.edu

William A. Young
Department of Management and Strategic Leadership
Ohio University
Athens, Ohio USA
Email: youngw1@ohio.edu

Gary R. Weckman, Naga T. Gollamudi
Department of Industrial and Systems Engineering
Ohio University
Athens, Ohio USA
Email: weckmang@ohio.edu, ng746812@ohio.edu

***Abstract*— The resiliency of telecommunication infrastructure by US Federal Emergency Management Agency (FEMA) region, is presented, based on almost 9,000 telecommunication outages over a 14 year period. Executive policy organizations in the US have described a resilient infrastructure as one that can minimize the magnitude and/or duration of service disruptions. To that end an empirical assessment of resiliency is made by region using telecommunication outages, each of which has duration and magnitude (the number of users affected by the outage). Wireline central offices are essential to telecommunication infrastructure, as they house local switching, transmission, and user access infrastructure for both voice and data services, including the Public Switched Telephone System (PSTN), the Internet, mobile communications, and emergency communication. Central office resiliency is studied by proxy, examining local telephone switch outages in those offices over 14 years. Regional comparisons are first made using classic time-series-of-event reliability techniques, allowing reliability trend comparisons. Next, outage cause comparisons are made. Then, a resiliency metric is presented that allows a fair comparison between regions differing in population. Marked differences in resiliency trends are apparent in some FEMA regions.**

Keywords- Resiliency; FEMA; telecommunication outage; critical infrastructure.

I. INTRODUCTION

The resilience of telecommunication services and capabilities are important to any nation. In the US, the Department of Homeland Security (DHS) states, in reference to critical infrastructure of all types, that resilience is

> *"….the ability to adapt to changing conditions and withstand and rapidly recover from disruption due to emergencies. Whether it is resilience towards acts of terrorism, cyber attacks, pandemics, and catastrophic natural disasters, our national preparedness is the shared responsibility of all levels of government, the private and nonprofit sectors, and individual citizens."* [1]

Additionally, The National Infrastructure Advisory Council (NIAC) was created as a Federal Advisory Committee to advise the President and the Secretary of Homeland Security on all areas of critical infrastructure. NAIC further refines the definition of effective infrastructure resilience to include measurable attributes:

> *"Infrastructure resilience is the ability to reduce the magnitude and/or duration of disruptive events. The effectiveness of a resilient infrastructure or enterprise depends upon its ability to anticipate, absorb, adapt to, and/or rapidly recover from a … disruptive event."* [2]

This research presents a resilience metric that include not only the magnitude and duration (called "impact") of telecommunication outages, but also frequency of these outages, on a regional basis. The regional paradigm chosen for this research is defined by FEMA, who

> *"…coordinates the federal government's role in preparing for, preventing, mitigating …. responding to, and recovering from all domestic disasters, whether natural or man-made, including acts of terror."* [3]

In performance of this mission, FEMA has divided the US and its territories into ten regions, shown in Figure 1, which present a useful and practical way to study critical telecommunication infrastructure resiliency across the US.

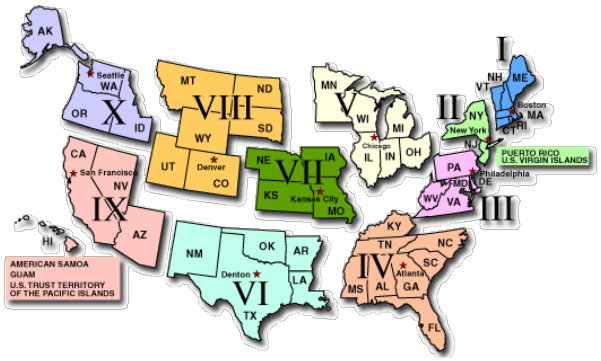


Figure 1. FEMA regions [4]

This paper addresses telecommunication infrastructure resiliency based on a 14 year record of U.S. PSTN local telecommunication switch outage data. In Section II Background, the role of local switches as service access points in the PSTN is covered. The importance of Central Offices where local switches, mobile switching, and internet access equipment is housed is discussed. Also, a description of the outage data is presented. In Section III, Methods and Results, regional differences in outage causality, reliability and resiliency methods and results are presented. Lastly, in Section IV Summary, major findings, research limitations, implications, and future research are discussed.

II. BACKGROUND

A. Central Office Buildings Serve More Than the PSTN

Central Offices house not only local voice telephone switches, but also other important network elements such as mobile communication backhaul and Internet access/transport equipment. For instance, it is not uncommon for a Mobile Switching Center (MSC) circuit switch to be located in a Central Office building. Alternately, the local exchange carrier often provide a wireless carrier Layer 1/2 connectivity between its Base Station Controllers (BSC) and its MSCs, which might be tens to hundreds of miles away. In these cases, the Central Office building is very important to wireless voice and data services.

Additionally, there are often optical SONET add/drop multiplexers in Central Offices that not only provide trunking between PSTN switches, but also digital internet trunks. At Layer 2/3, central offices involved in backhaul could be forwarding or aggregating Metro Ethernet virtual circuits per VLAN, interconnecting virtual circuits to Multiprotocol Label Switching (MPLS). Also, Central offices house important Internet assets such as routers and DSL internet access equipment. So, a range of services from Metropolitan Ethernet, MPLS, multiplexers and DSL can be adversely affected by the same factor that causes a local telecommunication switch in a Central Office to fail [5].

For these reasons, PSTN local switch outages that are caused by external circumstances, potentially affecting all electronics in a Central Office building, can be an indicator of PSTN, Mobile, and Internet telecommunication infrastructure resilience. Such local switch outage causes include those induced by external power outages, building damage, massive line cuts, and acts of god.

B. Local Telecommunication Switches

The PSTN is a complex system composed of a switching subsystem, a signaling subsystem, and a transmission subsystem. The switching subsystem routes voice calls throughout the PSTN network. The signaling subsystem coordinates call initiation, maintenance, and termination. The transmission subsystem provides physical links between switches so end-to-end voice circuit connections can be made. The signaling and transmission subsystems are not part of the research in this paper. The switching subsystem consists of local exchange switches (local switches), tandem switches, and international gateway access switches (see Figure 2). Only the local exchange switching subsystem is investigated in this study. There are three types of local switches: standalone, host, and remote. Less common are some tandem switches that also have access lines, but they are very small percentage of all switches in this study with outages. Importantly, best practices requires E911 call centers to connect to two tandems – however, as many centers are far away from tandems, some local switches are configured to act as tandems for E911 calls [6]. The importance of local switches using circuit switched technologies should not be underestimated. Even though the PSTN is migrating to voice over internet protocol (VoIP), the migration will take many years, and local switches will be in service for many years [7]. In 2011, although there were 32 million VoIP subscribers, there were 117 million subscribers connected over local loops to circuit switched local switches [8]. Also consider the work of Lyons, et al, where the economic impact of telecommunication outages were empirically assessed for local exchange outages. The economic loss estimates are based on actual business and residential demographics, including residential service and manufacturing. Economic loss estimates ranged from €370,000 to €1.1 million per day, for seven local exchange outages in Ireland [9].

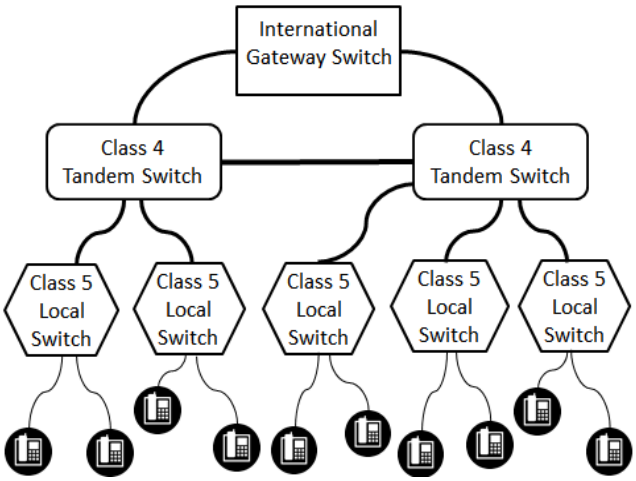


Figure 2. The PSTN switching subsystem.

C. Local Telecommunication Switch Outage Data

This study investigates 8,975 local telecommunication switch outages in the U.S. of at least 2 minutes in duration for a 14-year period (1996-2009) and considers only totally failed switches rather than partially failed switches (partially failed switches were not reported). Scheduled outages are not included because of their small duration – only the 8,875 outages caused by failures are studied here. Scheduled outages are not considered. This outage data was reported to the Federal Communications Commission (FCC) and obtained from [10]. Unfortunately, after 2009, the FCC stopped requiring carriers to report this data. Carriers classified each local switch outage incident using one of fifteen FCC defined cause codes. In this research, by combing

 ISBN: 978-1-61208-550-0

similar cause codes reported to the FCC into categories, we reduced the fifteen causes reported to the FCC down to five causality categories, similar to what was previously done in [11]:

- Human Procedural Errors: Procedural errors made in installation, maintenance or other activities by Telco employees, contractors, or vendors.
- HW and SW Design errors: Software or hardware design errors made by the switch vendor prior to installation.
- Hardware Errors: A random hardware failures, which causes the switch to fail.
- External Circumstances: An event not directly associated with the switch, which causes it to fail or be isolated from the PSTN.
- Other/unknown: A failure for which the cause was not ascertained by the carrier.

As each reported switch outage includes date, time, duration, magnitude, and location, important reliability and resiliency analysis can be performed.

III. OUTAGE ANALYSIS METHODS AND RESULTS

A. *Regional Local Switch Causality Differences*

To see to what extent switch outage cause categories might differ across regions, histograms were created for major cause categories. Each histogram shows the percentage of outages due to a particular causal category across the ten regions, two of which are shown in Figure 3. These categories, their composition, and the distribution of failures to each category are shown in Table I.

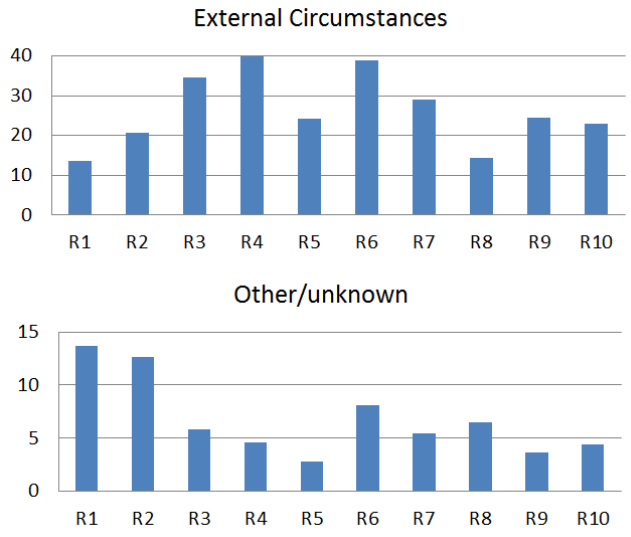


Figure 3. Regional causal percentage histogram examples.

TABLE I. LOCAL SWITCH OUTAGE FREQUENCY BY CAUSE

OUTAGE CATEGORY	Frequency	Percent
Human Procedural Error	1,394	16%
HW or SW Design Error	1,214	14%
Random HW Failure	2,951	33%
External Circumstances	2,900	32%
Other/Unknown	516	6%
Total	8,975	100%

For each cause category, the question of interest is whether there is a statistical difference in causal category percentages across the regions. The following hypotheses were used to test this notion for each histogram:

H_0: Cause category percentages is uniformly distributed across the 10 regions

H_a: Cause category percentage is not uniformly distributed across the 10 regions

The method used to test the hypotheses is the Chi-squared test, where the expected values are the average percentage across the 10 regions for each cause category, and the observed values are the actual percentages across the regions in a histogram. The results are shown in Table II, where we accept differences across the regions in the "External Circumstances" and the "Other/Unknown" cause categories.

TABLE II. REGIONAL OUTAGE CAUSALITY DIFFERENCES

CAUSE CATEGORY	ACCEPT	P-VALUE	CONCLUSION
External Circumstances	Ha	0.0005	Different across regions
Random HW Failures	H_0	0.4599	No difference across regions
Human Procedure Errors	H_0	0.5655	No difference across regions
HW or SW Design Errors	H_0	0.2599	No difference across regions
Other/Unknown	Ha	0.0340	Different across regions

These results are useful, because they indicate that "External Circumstance" outages are not uniform across the regions. The makeup of external circumstances are the type of events that potentially affect all central office telecommunications equipment/services, rather than just the PSTN local telecommunication switch. For instance, the vast majority of external circumstance local switch outages are due to environmental reasons such as FCC cause codes described by "acts of god", "external power failure", "environmental" and "massive transmission facility loss". These type of outages are very likely to also affect all other communications services in the central office building such as mobile switches, transmission and internet equipment. Note that Regions 4 and 6 have the highest percentages of outages due to external circumstances

 ISBN: 978-1-61208-550-0

B. Regional Local Switch Reliability Differences

Reliability is a study of times-to-failure (ttf), or said another way, the study of failure arrival process. If a failure rate is constant, the failure process is a stationary failure process, and Mean-Time-to-Failure (MTTF) can be calculated. If the times-to-failure are independent and exponentially distributed, the process is called a Homogeneous Poison Process (HPP). If the ttf's are not exponentially distributed, but still independent, the failure process is a renewal process (RP), and the distributions can be fitted to other distributions such as Weibull, Gamma, or other distributions.

Cumulative failure count versus time plots are often used to assess whether the arrival rate is constant, in that a straight line is apparent. However, if the plot is concave (bending down), reliability growth is indicated as the failures are decreasing over time. Conversely, if the plot is convex (bending up), reliability deterioration is indicated as the time between failures is decreasing. In these cases, the failure process is non-stationary and distributions cannot be used and MTTF cannot be calculated, as it is a function of time. However, if the bending is smooth and steady (monotonically increasing or decreasing), these processes can be modeled as Non-homogeneous Poison Processes (NHPP), also known as "doubly-stochastic processes" as the arrivals are random and the rate of arrivals is changing over time. If the changes are not smooth, these processes can often be analyzed in a piecewise fashion over time. The reliability from cumulative plots can easily be assessed visually, and if the changes appear subtle, analytical methods can be used to arrive at the degree of statistical significance of trends (using such tests as the Laplace trend, Lewis-Robinson, Mann, or the MIL-Handbook tests [12].

In this research, to compare the reliability of local switch outages by region, cumulative plots were made for each region and visually assessed for reliability growth, constancy, or deterioration. In no instances were the visual presentations subtle, so no formal statistical trend tests are necessary. An example of cumulative outage plots vs. years are shown in Figure 4. Note that the number of outages differs, however this is not important at this stage of the comparison, as we are looking for differences in trends for each region. A normalized resiliency comparison will be made later in the paper.

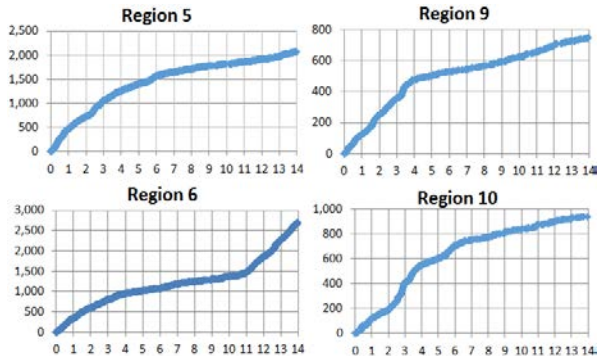


Figure 4. Regional switch cumulative failures vs. years

Region 5 exhibits classic reliability growth, which could by modeled by an NHPP. However, Region 6 exhibits two distinct piecewise regions – steady reliability growth over about 11 years, followed by reliability deterioration over the last 3 years. Region 9 also exhibits two distinct piecewise regions – fairly constant reliability for about 4 years followed by 10 years of fairly constant reliability at a much lower failure rate. Region 10 indicates reliability growth, but is not as smooth as smooth an improvement as Region 5. A summary of the visual assessments for each region is shown in Table III.

TABLE III. LOCAL SWITCH RELIABILITY TRENDS BY REGION

Reg.	Rel. Trend (Imp. - Improvement; Det. – Deterioration)
1	Monotonic Imp.
2	Monotonic Imp.
3	Monotonic Imp.
4	Monotonic Imp. for 13 years, steep Det. for last year
5	Monotonic Imp
6	Imp. for 11 years, steep Det. for last 3 years
7	Monotonic Imp.
8	Det. years 0-4, marked Imp. years 4-14
9	Constant years 0 to 4, marked Imp. years 4-14
10	Monotonic Imp.

At this point, it is useful to present the cumulative failure plot for all 8,975 switch failures in the U.S., as seen in Figure 5. Note that the overall trend is decreasing, as indicated by the Laplace Trend Test statistic U, where U is like a Z-score where at a value greater than +1.96, we accept the hypothesis of reliability growth at a critical value of 0.05. However, the trend is seen to be monotonically improving up to year 11, after which it starts to monotonically deteriorate. It appears that the reason for this deterioration is due to Regions 4 and 6. This observation is corroborated by the external circumstance frequencies for Regions 4 and 6, which is in Figure 3.

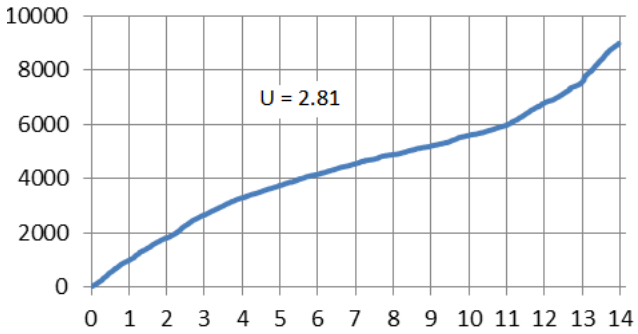


Figure 5. Local switch cumulative failure plot vs. years for the U.S.

C. Regional Local Switch Resiliency Differences

Earlier, we pointed out NAIC's interest to be minimizing the magnitude and/or the duration of impacts to critical infrastructure. So any resiliency measure must account for these factors. There has been past work using these two variables in assessing impact of telecommunication outages.

 ISBN: 978-1-61208-550-0

McDonald introduced the User Lost Erlang (ULE) as an impact metric for large-scale outages, given by:

$$ULE = log_{10}(Magnitude) \quad (1)$$

where magnitude is the subscribers impacted. So the impact of an outage affecting 100,000 lines would be 5 ULE. McDonald figured such a metric would be easy to use and be understandable by the public, similar to the logarithmic Richter scale for the intensity of earthquakes [13]. The disadvantage of the ULE is that only outage magnitude is taken into account -- duration is not. As outages have both size and duration, the ULE was not adopted or used, but did establish the need for an outage metric.

The FCC introduced use of the Lost Line Hour (LLH) metric, which is the product of the number of subscribers lost times the duration in hours, and also the lost line minute. For instance, a 100,000 line switch out for ½ hour represents 50,000 LLH. Although a straightforward metric incorporating both size and duration, the LLH does not include blocked calls. Also, the LLH is not logarithmic, a feature that nicely accommodates very long or large outages.

In the U.S., Committee T1, published an American National Standards Institute (ANSI) sponsored metric called the Outage Index (OI). This metric mapped duration and magnitude to weightings, which are logarithmic-like. The carrier industry adopted the OI metric. Analysis of the OI indicated a network administrator bias, as the index was sensitive to large size outages, but insensitive to long duration outages [14]. As an example, in [15] it was demonstrated that if a local switch with 10,000 lines experiences an outage of 24 hours duration the OI is 0.529, while an 8-day outage for the same switch is 0.532. The other disadvantage of OI is that the values are not intuitive, for example, what does an outage with an OI of 1.24 mean with respect to impact? Lastly, in [16], resiliency was more recently defined as the fraction of subscribers deriving successful service. Although an interesting metric, the number of users impacted is not apparent from say 0.99 resiliency factor.

Snow and Weckman recently introduced a novel resiliency metric for local switch outages in [17]. The metric is referred to as OI_{dbK}, includes both duration and magnitude (represented by LLH), is logarithmic, and intuitive as it is referenced to a baseline outage of 1000 LLH:

$$OI_{dbK} = 10log_{10}\left[\frac{LLH}{1000}\right] \quad (2)$$

Like the well-known *dbm*, a power referenced to 1 milliwatt in communications engineering, a doubling is about 3db while a halving is about -3db. Additionally, a tenfold increase is 10db and decrease by a tenth is -10db. Below are a few examples of OIdbK:

- OI_{dbK} = 0 corresponds to 1,000 LLH, as log of 1 is 0
- OI_{dbK} = -3 represents a halving, or 500 LLH
- OI_{dbK} = 20 corresponds to two orders of magnitude above 1,000 LHH, or 100,000 LLH
- OI_{dbK} = 23 is s a doubling above 20, or 200,000 LLH.

This new metric tames wide swings of LLH, and give an intuitive reference when doing time series plots and regression of outage resilience over time. Of course, if desirable, we can also have OI_{dbM}, which references the severity to one million LLH. Additionally, with outliers controlled, linear regression can be used to assess trends in resilience. An example of the utility of OIdbK is seen by referring to Figure 6 and Figure 7, where LLH and OIdbK for impact due to external circumstances are plotted The LLH plot appears unremarkable while the OIdbK plot indicates relative values of impact and a clear upward trend. In fact, statistically significant linear regression results for local switch external circumstance OIdbK indicate a 10.6 db increase per 10 years, which represents just over a 10 fold increase in LLH. [17].

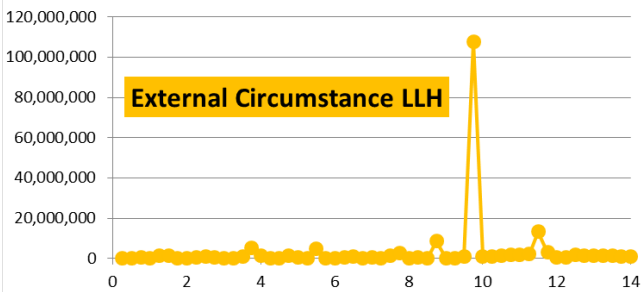


Figure 6. Quarterly LLH vs. years for ext. circumstance outages

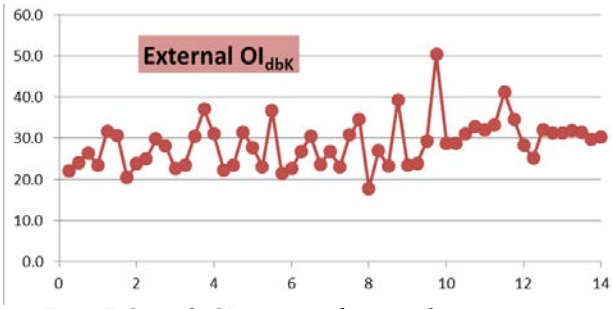


Figure 7. Quarterly OI_{dbk} vs. years for external circumstance outages

FEMA regions are different in a number of geographical factors, such as climate, area, and population. Where there are more people, there are more switches, so we expect more failures and more impact on resilience due to switch outages. There is a large range in population over the regions, as seen in Table IV.

TABLE IV: REGIONAL POPULATION AND OUTAGES

Region	Population (Millions)	Outages
1	19.0	139
2	31.6	301
3	28.3	519
4	55.3	2,300
5	50.7	1,299
6	34.4	2,139
7	13.1	1,065
8	9.7	246
9	44.0	488
10	11.6	479

 ISBN: 978-1-61208-550-0

For a fair comparison of regional resilience, in this research we modify the OIdbK by weighting LLH by population in millions, and reference the value to 1000 LLH per 1 Million people:

$$OI_{dbK/M} = 10log_{10}\left[\frac{LLH}{Pop\_Mill}/\frac{1000LLH}{1\ Mill}\right] \quad (3)$$

where Pop_Mill is average regional population (in Millions, e.g., 19.0 for Region 1) over the study period. The average population was used because there was little percentage change in regional population over the study period. This new metric has all the advantages of OI_{dbK} in addition to being scaled to the number of people in the region.

Comparative regional resilience examples are shown in Figure 8 and Figure 9, due to all outages. In Figure 8, upward impact over time indicates resiliency deterioration in both Regions 4 and 6. The deterioration in these regions are very similar, however note the large outliers in years 6 and 10 for Region 6. Examples of resiliency growth and constancy are shown in Figure 9. In Region 1, relatively constant resiliency is indicated, while in Region 1 strong resiliency growth is seen by the dramatic downward trend in outage impact.

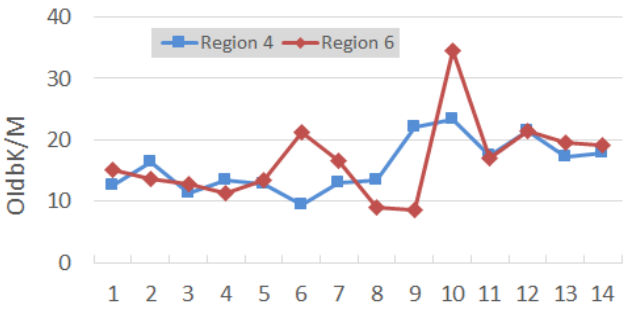


Figure 8. Yearly outage impact for regions 4 and 6

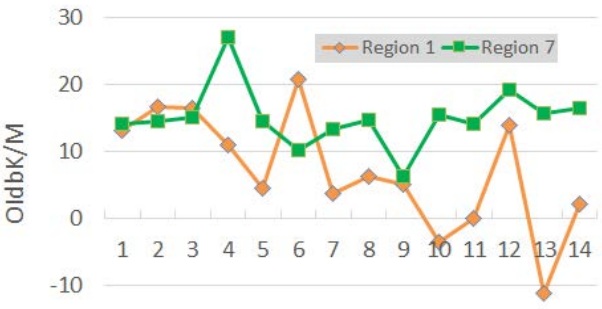


Figure 9. Yearly outage impact for regions 1 and 7

Although there is not room in this paper to show all ten regional impact plots, a qualitative description of FEMA regional resilience is given in Table V. In this table, regional resiliency is described as either improving, constant or deteriorating. Trend descriptions are based on regression lines for each regions outage impact plot.

Table V. REGIONAL RESILIENCY TRENDS

Reg.	Resiliency	Description based on Regression
1	Improving	Starting at 16db, and dropping to 0db over 14-years (from 40K to 1K LLH per mill pop.)
2	Improving	Starting at 16db, and dropping to 6db over 14 years (from 40K to 4K LLH per mill pop.)
3	Deteriorating	Starting at 10db, and increasing to 16db over 14-years (from 10K to 40K LLH per mill pop)
4	Deteriorating	Constant first 8-years at 13db, up to 19db over 6-years (from 20K to 80K LLH per mill pop.)
5	Constant	Constant at about 10db, or about 10K LLH per mill in pop.; some large variances
6	Deteriorating	Starting at 13db, and increasing to 18db over 14-years (from 20K to 80K LLH per mill pop.)
7	Constant	Constant at about 15db, or about 40K LLH per mill in pop.
8	Improving	Starting at 10db, and dropping 10db over 14 years, a drop of 10,000 LLH per mill pop.
9	Constant	Constant at about 10db, or about 10K LLH per mill in pop.; some large variances
10	Constant	Constant at about 10db, or about 10K LLH per mill in pop.; some large variances

IV. SUMMARY

A. *Major Findings*

The major findings of this research on FEMA regional local telecommunication switch outages from 1996 to 2009 are:

- From a causality perspective, there are statistically significant differences in external circumstance outages across the regions. Additionally, histograms indicate the largest differences are due to Regions 4 and 5. External circumstance local switch outages are also likely to affect other telecommunication sector capabilities such as mobile and internet.
- Over a 14-year period, the arrival process of outages in each region are non-stationary processes for which time-to-failure distributions and MTTF metrics are not feasible. However, there are clear instances for some regions where there are different processes over the entire period, which can be segmented and analyzed separately. In some instances they are piecewise linear (constant reliability) where MTTF can be calculated.
- Most regions experienced dramatic reliability growth in local switch reliability over the 14-year period, although two regions (Regions 4 and 5) experienced initial reliability growth for most of the time period but severe reliability deterioration towards the latter part of the 14 years.

 ISBN: 978-1-61208-550-0

- From a resiliency perspective a recently introduced resiliency metric was successfully modified to weight impact by regional population, offering a fairer way to compare geographically different regions. The metric also conforms to NAIC desires, as it accounts for both magnitude and duration.
- These results indicate that empirical methods and metrics are very useful in understanding the impact of outages to critical infrastructure, and that resilience is best understood when coupled with reliability, or the arrival rate of outages, which are in fact resiliency deficits.

B. Research Limitations

As the quantitative research presented here is based on data reported by carriers, it is not known how consistent the reporting was over a 14-year period by each carrier, and how similar the capabilities of carriers to capture outages, and accurately report the size, duration, and cause of outages. Additionally, only complete switch failures are reported and partial switch outages were excluded from reporting requirements. Also, this research was limited to using lost line hours in its resiliency metrics, as no reporting of blocked calls was required of carriers. Lastly, actual impact of local switch external circumstances outages on mobile communication and Internet services/infrastructure in the same Central Office buildings cannot be quantified by these results.

C. Future Work and Policy Implications

More research is required to develop better resilience measures across all sectors of the telecommunication industry, in addition to metrics that can be linked or include economic impact. Additionally, the results in this paper indicate that local switch outages might serve as a "canary in the mine shaft" with respect to telecommunication infrastructure resiliency, due to the plethora of other telecommunication service sector equipment residing in PSTN Central Office buildings. In retrospect, these results indicate that the FCC's discontinuance of local switch outage reporting after 2009 might be unfortunate, as an insightful reliability and resiliency bellwether seems to have been lost.

V. ACKNOWLEDGEMENT

This work is based on an unpublished paper presented at the 13th Annual Conference on Telecommunications and Information Technology, ITERA 2015 [18]. The original work has been corrected, refined and augmented here.

REFERENCES

[1] Quote retrieved December 2016 from http://www.dhs.gov/national-infrastructure-advisory-council

[2] NAIC Charter November 2013 [Online] available from http://www.dhs.gov/publication/niac-charter

[3] Quote retrieved December 2016 from https://www.fema.gov/about-agency

[4] Figure retrieved December 2016 from https://www.fema.gov/risk-mapping-assessment-planning/regional-contact-information

[5] Author discussions with engineers from several Local Exchange Carriers, March 1-8, 2015.

[6] A. P. Snow, A. Shyirambere, J. Arauz, G. R. Weckman. "A Reliability and Survivability Analysis of US Local Telecommunication Switches." International Journal On Advances in Telecommunications 6, no. 3 and 4 (2013): 81-97.

[7] J. Gillan, and D. Malfara, The transition to an all-IP network: a primer on the architectural components of IP interconnection, National Regulatory Research Institute, May 2012.

[8] FCC, Local Telephone Competition: Status as of December 31, 2010, Industry Analysis and Technology Division, Wireline Competition Bureau, October 2011.

[9] S. Lyons, E. Morgenroth, R. Tol, "Estimating the value of lost telecoms connectivity", Electronic Commerce Research and Applications 12, 2013, pp. 40–51.

[10] FCC Report 43-05, ARMIS Service Quality Report Table IVa, retrived December 2016 from http://transition.fcc.gov/wcb/armis/ September 2012.

[11] A. P. Snow, J. Arauz, G. R. Weckman, and A. Shyirambere. "A Reliability and Survivability Analysis of Local Telecommunication Switches Suffering Frequent Outages." In ICN 2013, The Twelfth International Conference on Networks, pp. 209-216. 2013.

[12] D.M Louit, R. Pascual, A. K. S. Jardine, "A Practical Procedure for the Selection of Time-to-failure Models Based on the Assessment of Trends in Maintenance Data", Reliability Engineering and System Safety 94 (2009) 1618-1628.

[13] J. C. McDonald, "Public network integrity-avoiding a crisis in trust" Journal on Selected Areas in Communications, IEEE Journal, Volume: 12 , Issue: 1, 1994.

[14] A. P. Snow, "A Survivability Metric for Telecommunications: Insights and Shortcomings", IEEE Computer Society, Proceedings, Information Survivability Workshop – ISW'98, 1998, pp. 135-138.

[15] A. P. Snow and Y. Carver, "Carrier-industry, fcc and user perspectives of a long duration outage: challenges in characterizing impact", T1A1.2/99- 026, Contribution to Committee T1 – Telecommunications, Boulder Colorado, 1999.

[16] M. Omer, R. Nilchiani, and A. Mostashari, "Measuring the resilience of the global internet infrastructure system", 3rd Annual IEEE Systems Conference, March 2009, pp. 152-162.

[17] A. P. Snow and G. R. Weckman, "Trends in Local Telecommunication Switch Resiliency." In ICN 2014, The Thirteenth International Conference on Networks, pp. 178-184. 2014.

[18] A. P. Snow, G. R. Weckman, N. Gollamudi, J. C. Hoag, W. A. Young, "A Resiliency Assessment of Critical Telecommunication Infrastructure by FEMA Region: Empirical Metrics and Trends.", 13th Annual Conference on Telecommunications and Information Technology, ITERA 2015. Unpublished.

 ISBN: 978-1-61208-550-0

A Universal Mechanism to Handle ION Packets in SDN Network

Lixuan Wu[1], Jiang Liu[1,2] , Tao Huang[1,2] , Weihong Wu[1] , Bin Da[3]

[1]State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
[2]Beijing Advanced Innovation Center for Future Internet Technology
Beijing, China
[3] Network Technology Laboratory, 2012 Laboratories
Beijing Huawei Digital Technologies Co., Ltd.
Beijing, China

Emails: {shinning@bupt.edu.cn , liujiang@bupt.edu.cn, htao@bupt.edu.cn, 344259446@qq.com, dabin@huawei.com}

***Abstract*—Identity Oriented Networks (ION) provide mechanisms for scalability, mobility and operations across heterogeneous entities by disseminating unique identity of end points from their position in the network. However, current devices cannot parse the newly added ID field in 3.5 layer. This paper puts forward a universal Software Defined Networking (SDN)-based packet processing mechanism to parse information in high layers and exchange redundant information in low layers with key information in high layers at the entrances of network. Thus, the key field in high layer is visible in low layer and can be parsed by current protocol and routing devices. The article takes GPRS Tunneling Protocol (GTP) packets for example to explain the packet processing method. Besides, delay caused by the packet processing module is measured and an experiment is made to verify that parsing high-layer field succeeds and strategies on high-layer field can be made.**

Keywords-ION; SDN; OpenFlow; GTP; TEID.

I. INTRODUCTION

With the development of network technologies and user requirements, mobility has been a significant trend. Firstly, the number of mobile devices including M2M modules keeps increasing and is expected to be 11.6 billion by 2020. Secondly, mobile traffic has grown faster year by year. Mobile traffic content also tends to carry more video traffic including streaming video, which requires high bandwidth transmission capabilities. Thirdly, offload mobile traffic (traffic from dual-mode devices over Wi-Fi or small-cell networks) is taking more and more proportion of the whole traffic. In 2015, mobile offload traffic exceeded cellular traffic for the first time. In general, more devices, more traffic and more offload connectivity pattern should be taken into consideration in next-generation network.

In 5G era, network has five performance requirements: 1) bandwidth and speed throughput: 10Gps; 2) latency: less than 1ms; 3) scale: 10-100 times than Long Term Evolution (LTE) [1]; 4) session continuity: ubiquitous; 5) mobility speed: 500km/h. However, current LTE architecture has many constraints. The handoff delay and latency is noticeable. Due to requirement of global IP addresses, multi-homing features make IP addresses aggregation difficult and lead to large RT on routers.

In the context, Identity Oriented Networks (ION) has been put forward to improve mobility performance. The fundamental premise of ION is the one that provides mechanisms for scalability, mobility and operations across heterogeneous entities by disseminating unique identity of end points from their position in the network. A 3.5 layer is added between the IP layer and the TCP/UDP layer. An identity field is carried in the new layer to identify a node, an app or anything. ID can be bond with an IP address locator to complement forwarding. Thus, network has the following improvements: 1) native mobility; 2) ID-based Apps; 3) multi-homing ID with global scope; 4) context awareness based on ID profile; 5) ID-based security.

Although ION improves network performance in many ways, current switching and routing devices can only parse information no higher than the IP layer. The long-term trend is to develop new devices that are capable to parse the 3.5 ID layer information. However, it will take a long time to update all routing devices and Capital Expenditure (CAPEX) should also be considered. To solve the issue, this article put forward a universal SDN-based packet processing mechanism to parse information in high layers and exchange redundant information in low layers with key information in high layers. Thus, the key field in high layer is visible in low layer and can be parsed by current protocol and routing devices.

The paper is organized as follows. Section II introduces related knowledge of the example case. Section III specifically explains the SDN-based packet processing mechanism to parse the high-layer field incompatible with OpenFlow protocol and utilizes the field to control network at a smaller granularity. Section IV measures the delay caused by the packet processing module and verifies the success of parsing high-layer field and making strategies based on the high-layer field. Finally, conclusions are presented in Section V.

 ISBN: 978-1-61208-550-0

II. RELATED KNOWLEDGE

Since ION is in conceptual phase, this paper takes GTP packets for example to explain the packet processing of parsing high-layer information and utilizing high-layer information to handle packets. The augment supporting the analogy is that GTP packets face the similar problem in incapability of recognizing high layer information when being processed by OpenFlow [2] devices.

GTP is an IP-based communication protocol in Evolved Packet Core (EPC) [3]. The protocol consists of GTP-C, GTP-U and GTP', which are respectively used in the GTP control plane, GTP user plane and charging data transmission. Since this paper studies packet processing in data plane, GTP-U is the core concern. GTP-U protocol stack is shown in Figure 1. For uplink packets, the radio layer ends in eNodeB. ENodeB encapsulates GTP packets and establishes the GTP tunnel to S-GW. S-GW establishes the GTP tunnel to P-GW and P-GW decapsulates GTP packets and forwards it to Internet. For downlink packets, the processing procedures are reversed. There is a key field named Tunnel endpoint identifier (TEID) in GTP-U header. TEID is used to multiplex different connections in the same GTP tunnel. A GTP connection is uniquely confirmed by a source tunnel IP, a destination tunnel IP and a TEID. In this article, GTP packets refer to GTP-U packets without special statement.

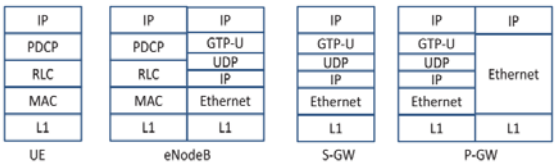


Figure 1. GTP-U protocol stack

Software-defined networking (SDN) [4] is an emerging solution for fine-grained control and management of networks. It separates the control plane (SDN controller) and data plane (switching and routing devices) of network. OpenFlow is a generally accepted southbound protocol between the controller and switching devices. There is a widely agreed trend that SDN should be integrated into 5G core network. However, GTP-U header locates over UDP layer. In current OpenFlow protocol, only information below layer 4 (including layer 4) can be parsed. Therefore, TEID is invisible in OpenFlow and GTP connection could not be recognized by OpenFlow switches. To address the problem, a packet processing method is put forward to parse the high layer field and exchange it with redundant fields in low layer.

III. PACKET PROCESSING IMPLEMENTATION

By analyzing the packet transmission within GTP tunnels, it could be found that forwarding decisions are made depending on the destination tunnel endpoint IP address. Therefore, during transmission in GTP tunnels, the source IP address is redundant information. Besides, both source IP address field and TEID field have a length of 32 bit in common. As a result, exchanging TEID field with source IP address field at the tunnel entrance would expose useful TEID field when forwarding the GTP packets in GTP tunnel without influencing normal processing mechanisms of GTP packets. Moreover, OpenFlow devices work normally complying with OpenFlow protocol since TEID field is already in layer 3 and could be parsed by OpenFlow protocol.

A. Architecture

In Evolved Packet Core (EPC) architecture, control functions and forwarding functions of S/P-GW are coupled, which restricts core network flexibility. Thus, separating control functions and forwarding functions of S/P-GW is a main trend of 5G mobile core network developments. Base on that, this article designs the core network architecture in Figure 2. This architecture retains most structure of the current EPC architecture. The major differences are: 1) separating the control functions and forwarding functions of S/P-GW; 2) introducing SDN controller to cooperate with S/P-GWc to manage the network between S-GWs and P-GWs; 3) data plane are OpenFlow-enabled and comply with the control of SDN controller.

The data plane consists of S/P-GW's user plane devices and OpenFlow devices, while the control plane consists of SDN controller, S/P-GW's control plane, Mobility Management Entity (MME), Home Subscriber Server (HSS) and Policy Control and Charging Rules Function (PCRF). In this case, MME works the same in EPC. It manages mobility, chooses S-GW for user equipment (UE) and establishes the GTP tunnel between eNodeB and S-GW. The control functions of S/P-GWs operate over SDN controller and communicate with it by JSONRPC messages to swap UE information and S/P-GW information. The S/P-GW control plane strategies are implemented in coordination with MME, HSS and PCRF, including user IP allocation and traffic flow template (TFT) assignment. SDN controller controls data plane devices with OpenFlow protocol and manage TFT with S/P-GW. S/P-GWs provide terminal of GTP tunnels and anchor GTP tunnels during handoff. Applying SDN to manage the transmission network makes it convenient to realize overhead control and routing optimization. Besides, this architecture is compatible with current 3GPP [5] standards, which is the smooth evolution for mobile core network to integrate SDN.

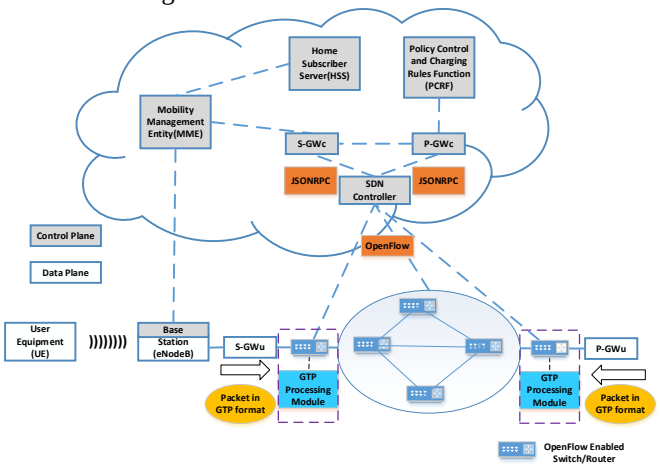


Figure 2. SDN-based mobile core network architecture

B. *Data Plane Implementation*

To parse the high layer information of TEID, a GTP processing module is attached to a logical port of an OpenFlow device at the entrances of S/P-GWs, cooperating with OpenFlow pipeline processing to handle GTP packets. The OpenFlow device with GTP processing module can be regarded as an extension of gateway functions. For uplink GTP packets at S-GWs and downlink GTP packets at P-GWs, TEID field and source IP address field should be exchanged to make the TEID field visible to SDN controller and OpenFlow devices between S-GWs and P-GWs. Thus, for downlink GTP packets at S-GWs and uplink GTP packets at P-GWs, TEID field and source IP address field should be exchanged to restore the previous packets.

The OpenFlow devices utilize OpenFlow pipeline processing to handle packets. There are at least five flow tables in the OpenFlow device. Table 0 has the highest priority. It matches GTP packets (udp_port is 2152) whose in_port is not connected to the GTP processing module and then sends them to the GTP processing module. Table 1 has the second highest priority. It matches GTP packets whose in_port is connected to GTP processing module and sends them to Table 2, otherwise it sends them to Table 3. Table 2 has the third highest priority. It matches GTP packets whose source MAC address is the S/P-GW to which it connects and sends them to Table 4, otherwise it outputs them to the connected S/P-GW. Table 3 forwards regular packets except GTP packets and Table 4 forwards GTP packets to be sent to the core network. The GTP processing module implements the exchange of high-layer TEID field and low-layer source IP field. Firstly, it parses the high-layer TEID field. Then, it exchanges TEID field with the low-layer source IP field. Lastly, it sends the processed packet back to the logical port where the packet comes. The OpenFlow multi-stage flow tables at entrances of S/P-GWs are shown in Figure 3.

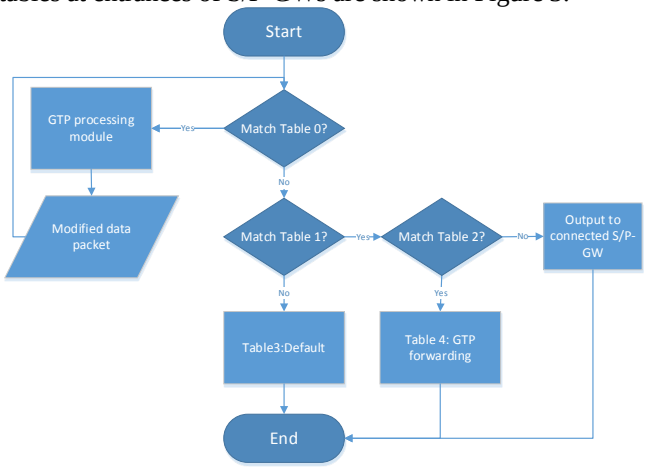


Figure 3. OpenFlow multi-stage flow tables at entrances of S/P-GW

After the packet processing at the entrance of core network, OpenFlow devices and SDN controller could have access to TEID field of GTP packets in the network between S-GWs and P-GWs. Thus, strategies based on TEID field are easy to implement.

C. *Control Plane Implementation*

As a result of exchange processing at GTP tunnel entrances, traffic control (routing optimization, Qos, etc.) at the granularity of GTP connections could be realized without changing working principles of OpenFlow devices and SDN controller.

Varieties of applications could be developed and run over SDN controller. Those applications leverage TEID field in layer 3 and make specific decisions on different TEIDs. An example of routing optimization based on TEID is given in Figure 4. Leveraging the advantage of the global view of SDN controller, different routing planning can be easily implemented.

Considering that different GTP connections have different routing demands, three routing planning modules are added to SDN controller. One implements shortest path planning, one implements maximum bandwidth path planning and another implements minimum delay path planning. When the controller receives packet-in messages, it matches TEID field and invokes corresponding path planning module for different TEIDs. The designed route is distributed to the network by flow-mod messages.

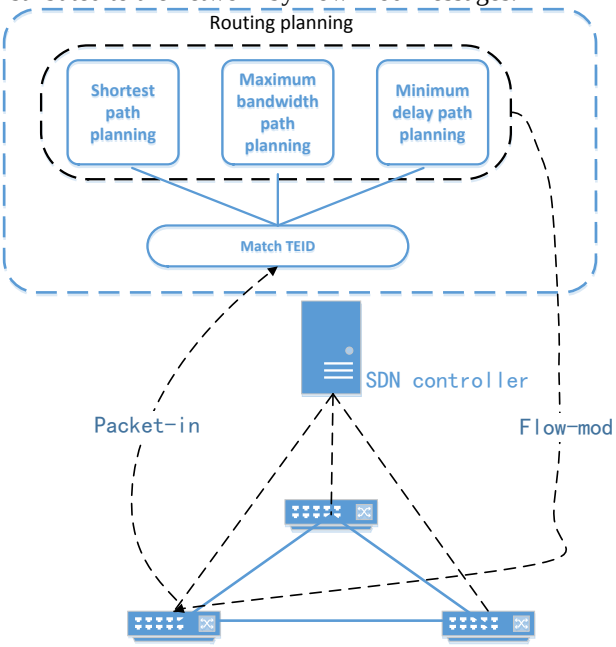


Figure 4. Routing planning example based on TEID

Routing modules based on different demands can be developed by similar methods.

IV. EVALUATION

Introducing GTP processing module into the network would cause extra delay during packet transmission. Thus, evaluation on delay should be taken into consideration. Besides, whether traffic control at the granularity of TEID is realized or not is also to be proved.

A. *Experimental Setup*

The experimental system is set on Ubuntu 14.04 LTS. Mininet [6] 2.3.0d1 was utilized to simulate the network and

 ISBN: 978-1-61208-550-0

Ryu [7] was utilized as an SDN controller. The topology is shown in Figure 5. SGW1 and PGW1 are two virtual machines which generate GTP packets. MS1 and MS2 are set to send GTP packets to a logical port attached by the GTP processing module. MH1 and MH2 are two network namespaces and work as GTP processing modules, respectively attached to the logical port of MS1 and MS2. S1, S2, S3, S4 and S5 work as switches in the core network.

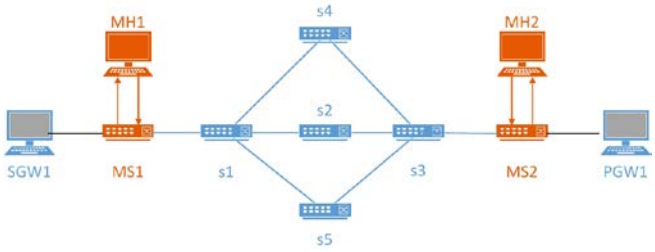


Figure 5. Experiment topology

- SGW1 and PGW1

OpenGGSN-0.91 [8] is used to build GTP tunnels. The PGW1 serves as a GGSN node that has one tunnel IP, and the SGW1 serves as an SGSN node that has several tunnel IPs. The TEID of each tunnel is allocated by orders. For the purpose of a simple design, the IP address of SGW1 and PGW1 are 10.0.0.1 and 10.0.0.2, and the MAC addresses are 00:00:00:00:00:01 and 00:00:00:00:00:02.

- MH1 and MH2

MH1 and MH2 work as the GTP processing module, aiming at exchanging source IP field with TEID field. There are two network devices in MH1, including eth0 and etSGW1. EtSGW1 works in a promisc mode to get each packet it receives. And eth0 works in multicast mode and has an IP address of 10.0.0.3 and a MAC address of 00:00:00:00:00:03. A python application named gtp_modify.py is running on MS1 to realize the GTP process module's functions. The application firstly captures a GTP packet from etSGW1, and then it cuts out the data segment and parses TEID field by counting the bit position. After getting the TEID, it exchanges the TEID field with source IP address field, and finally sends the packet back to the network through eth0.

Except that MH2's eth0 has an IP address of 10.0.0.4 and a MAC address of 00:00:00:00:00:03, the other settings and working pattern of MH2 are the same as MH1's.

- MS1 and MS2

MS1 is located between SGW1 and core networks, connected to MH1. And MS2 is located between PGW1 and core network, connected to MH2. The flow tables in MS1 and MS2 are described in Section III. GTP-U packets that should be sent into or out of core network will be delivered to GTP processing module.

- Ryu controller

Ryu serves as a SDN controller and is mainly responsible to set flow tables to OpenFlow switches. It will create a default route at exactly the time when the network is built, so that some control message like GTP-C can be forwarded successfully.

B. *Result Analysis*

In order to verify traffic control abilities at the granularity of GTP connections, there are three GTP connections in our scenario. The PGW1 has one tunnel IP of 192.168.0.1, while SGW1 has three tunnel IPs: 192.168.0.2 for tunnel1, 192.168.0.3 for tunnel2 and 192.168.0.4 for tunnel3. Besides, tunnel1's TEID is 0x00000001, and tunnel2's is 0x00000002, and tunnel3's is 0x00000003. Iperf [9] is utilized to generate traffic flows. Flow1 is sent at a speed of 100kbps in tunnel1, while flow2 is at a speed of 80kbps in tunnel2 and flow3 is at a speed of 60kbps in tunnel3. Based on TEID and destination IP address, different routes are designed. Flow1 is designed to go through s1->s4->s3, while flow2 is designed to go through s1->s2->s3 and flow3 is designed to go through s1->s5->s3. Traffic bandwidth is measured in each route, and the delay caused by MH1's GTP processing module is also detected.

Figure 6 shows the delay caused by GTP processing module in MH1. It can be seen that the delay range from 46ms to 98ms. The average delay is approximately 75ms. The delay value is a bit large. However, the delay will only be generated at the entrance of the whole network and the value is relatively stable regardless of the network scale. When network scale expands and the whole delay increases, the delay caused by GTP processing module tends to have a smaller influence to the whole network. Besides, the delay can be diminished by promoting the hardware performance.

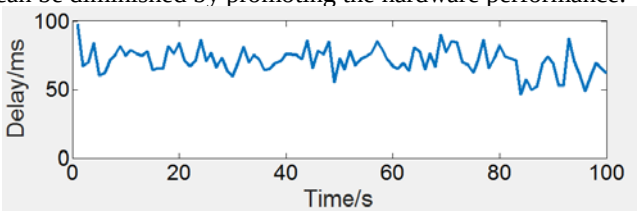


Figure 6. Delay caused by GTP processing module in MH1

Figure 7 shows traffic bandwidth in route s1->s4->s3, s1->s2->s3 and s1->s5->s3. It can be seen that the average bandwidth on the three routes are approximately 100kbps, 80kbps and 60kbps. It matches the bandwidth of flow1, flow2 and flow3, which illustrates the route design based on TEID is realized. Besides, Wireshark [10] is also utilized to capture packets at switch s2, s4 and s5. The results show that only packets with TEID 1 appear at switch s4, while only packets with TEID 2 appear at switch s2 and only packets with TEID 3 appear at switch s5. It further verifies that traffic control at the granularity of GTP connections is realized.

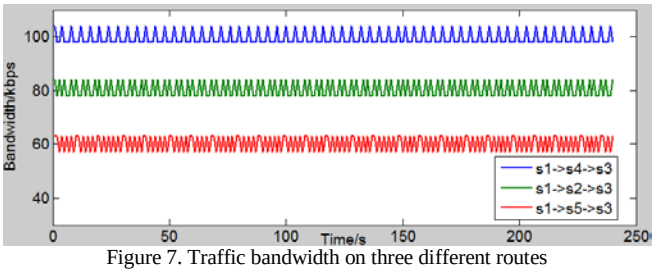


Figure 7. Traffic bandwidth on three different routes

Results show that traffic control abilities at the granularity of GTP connections are achieved.

V. CONCLUSIONS

This paper designs a universal SDN-based mechanism to handle packets containing high-layer field incompatible with OpenFlow protocol, without extending protocols or upgrading devices. An example of handling GTP packets is shown to explain the processing mechanism. Besides, the processing delay is measured and an experiment verifies that TEID field is successfully parsed and strategies at the granularity of TEID can be made. By utilizing the packet processing mechanism, ION packets can be transferred in SDN network and strategies at the granularity of ID field can be made. The mechanism could be universally utilized to tackle the problem of parsing a relatively small field incompatible with OpenFlow protocol as a temporary solution.

ACKNOWLEDGMENT

The authors thank Fei Yang and several engineers from Huawei for their valuable knowledge support.

REFERENCES

[1] "LTE", [Online]. Available from: http://www.3gpp.org/technologies/keywords-acronyms/98-lte 2017.04.01

[2] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, et al, "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, 2008, pp. 69-74

[3] "EPC", [Online]. Available from: http://www.3gpp.org/technologies/keywords-acronyms/100-the-evolved-packet-core 2017.04.01

[4] N. McKeown, "Software-defined networking," INFOCOM keynote talk, vol. 17, no. 2, 2009, pp. 30-32.

[5] "3GPP", [Online]. Available from: http://www.3gpp.org/ 2017.04.01

[6] "Mininet", [Online]. Available from: http://mininet.org/ 2017.04.01

[7] "Ryu", [Online]. Available from: https://osrg.github.io/ryu/ 2017.04.01

[8] "OpenGGSN", [Online]. Available from: https://sourceforge.net/projects/ggsn/ 2017.04.01

[9] "iPerf", [Online]. Available from: https://iperf.fr/ 2017.04.01

[10] "Wireshark", [Online]. Available from: http://wireshark.com/ 2017.04.01

 ISBN: 978-1-61208-550-0

A ID/Locator Separation Prototype Using Drone for Future Network

Shoushou Ren, Yongtao Zhang
2012, Network Technology Lab, Huawei Technologies Co., Ltd., Beijing, China
E-mail: {renshoushou, zhangyongtao3}@huawei.com

***Abstract*—The routing and addressing system of today's Internet is facing serious scaling problems, which are mainly caused by the overloading of IP address semantics. To address this problem, several recent schemes have been proposed to replace the IP namespace with separation of namespaces for identities and locators. ID Oriented Networks (ION) is one such mechanism. In this paper, a drone prototype based on ION implementation is described. An ID-to-ID communication between a moving drone and a stationary endpoint is demonstrated. ION protocol primitives are defined along with packet format, encapsulation/decapsulation, as well as the handover process. The results obtained from the prototype of ION show that the ID-to-ID communication continues to works well and is not interrupted when the location of the drone changes. This prototype shows that the basic idea of ID/Locator separation is a feasible and positive way to solve the scaling issue in the current Internet Protocol.**

Keywords-drone; identifier; locator; handover.

I. INTRODUCTION

It has been widely recognized that today's Internet routing and addressing system is facing serious scaling problems [1][2]. A common consensus is that this scaling issue is mainly caused by the overloading of Internet protocol (IP) address semantics [3]. That is, an IP address represents not only the location but also the identity of a host. Therefore, several new schemes [4], such as the Locator/ID Separation Protocol (LISP) [5] and Host Identity Protocol (HIP) [6][7][8], have been proposed to replace the IP namespace in today's Internet with a locator namespace and an identity namespace. In these schemes, a locator namespace consists of *locators* that represent the attachment point of hosts in the network, while the identity namespace consists of *identifiers* (ID), also known as endpoint identities (EIDs) that represent unique identities of hosts. When IDs are separated from their network attachment position information, packets destined for IDs are generally forwarded with the default routing method by using the locators as IPs. By decoupling an identifier from its locator, changes to a host's location become transparent to the upper layers above including TCP.

Consider the communication between two User Equipments (UEs) in the ION network. Each UE only needs to know the other's ID before the connection is established, since only the ID can tell them *who* the correspondent ID is. While the locator is only used for packet forwarding in the internet and it may change according to different access gateways. Thus, the communication is called an ID-to-ID communication.

In this paper, we present a drone prototype which is realized based on the basic idea of ION. The drone has a unique and fixed ID when flying across different access gateways. While its locator changes when it flies across the network accessing different gateways. Our prototype ensures that the drone can establish an ID-to-ID connection with the remote ground station, which is also an ID aware host. Moreover, when the drone accesses different gateways, the ID-to-ID communication between the drone and the ground station is continuously maintained even when the drone's locator changes.

The rest of this paper is structured as follows. In Section II, we introduce the basic framework of the Identity Oriented Network. In Section III, we describe the topology of the drone prototype and introduce the main entities in the prototype. In Section IV, the detail designs of our prototype are presented, including the id packet format, packet encapsulation and decapsulation, as well as the handover process. At last, we conclude this paper in Section V.

II. IDENTITY ORIENTED NETWORKS

Based on the idea of Identity and Location separation, ION framework is briefly described in Figure 1 and the details are out of scope for this paper. Since identity and locators are separated, ION expands network layer concept to accommodate ID in the following manner.

- *ID layer* is a distributed function responsible for ID management and authentication services.
- *Mapping system:* An ID/location resolution system is introduced which maintains mappings between a host and its location.
- *ID based connection:* In order to inter-connect two endpoints independent of network address an ID aware socket connection.

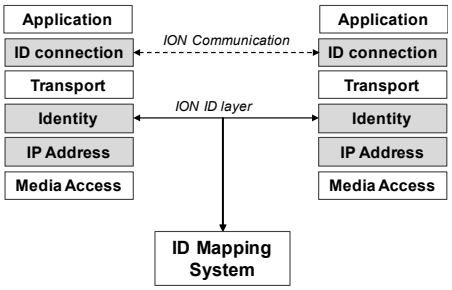


Figure 1. Brief framework of ION.

 ISBN: 978-1-61208-550-0

ION architecture enhances traditional network layer with identity awareness. Some advantages ION scheme include (a) communication of non-IP devices such as IoT, (b) a smoother and seamless location agnostic mobility and (c) cross-silo communication across applications working with same network entities. Please refer to Next Generation Protocols (NGP) paper for further details on ION [9].

III. TOPOLOGY OF THE DRONE PROTOTYPE

The topology of our drone prototype is depicted in Figure 2, which mainly consists of following five entities:

- *Universal Access Gateway (UAG):* The UAG is the edge access gateway in the ION architecture. The UAG is in charge of locator assignment, locator management and access control. When a UE, such as the drone in our prototype, is online and accesses to a UAG first time, the UAG assigns an IPv6 address as locator to it. Then the UAG registers the ID/Locator mapping item of the UE to the mapping system and caches the item until the UE leaves. The UAG can support the wired access as well as wireless access of UEs. UAGs also perform packet forwarding function as traditional gateways. Three UAGs are deployed in our prototype and the drone flies randomly in the area covered by the three UAGs.
- *Access Point (AP):* Traditional APs. The drone access to the UAG via an AP. Only one AP is deployed under each UAG for the case of layer-3 handover [10] [11], which will be further explained in the next section.
- *Drone:* The drone is an ID aware host with a unique and fixed ID. When it accesses a UAG, a locator will be assigned, which is used to locate where it is. The drone is equipped with a camera for shooting real-time video when flying across different UAGs. It is controlled by the ground station and the video will be transmitted to the ground station via ID-to-ID communication.
- *Ground Station (GS):* the GS, which is also an ID aware host, is the controller of the drone. It receives and displays the video shot by the drone.
- *ID-Locator Mapping System (ILMS):* The ILMS stores all the ID/Locator mapping items that have been registered. Once a UE is assigned a locater or by its access UAG, the ID/Locator item will be registered or updated to the ILMS. If a UE wants to communicate with other ID hosts, their locators can also be retrieved from the ILMS.

Note that ID of hosts may be set before leaving the factory or assigned after that by some organizations. In our prototype, we use the IPv6 address those are with prefix *2F00::* as IDs. The goal of our prototype is: 1) realize an ID-to-ID communication between the drone and the remote GS; 2) when the drone's locator changes while roaming across different UAGs, the ID-to-ID communication could be kept continuous

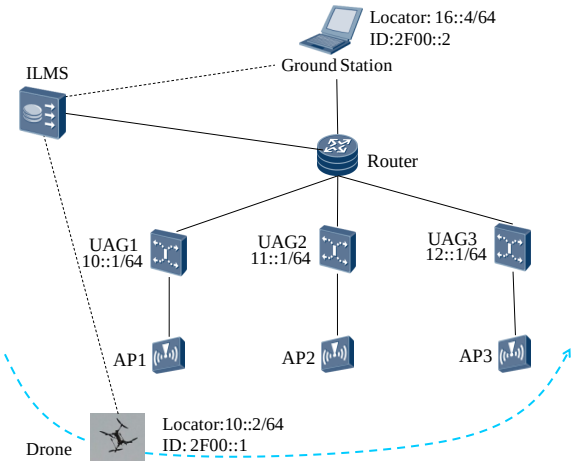


Figure 2. Topology of the drone prototype.

IV. PROTOCOL PRINCIPLE

Some new protocol principles are designed to realize the ID-to-ID communication between the drone and GS.

A. Packet Format

The main change in ID packet lies in the IP-layer header. The tuple <*src_ip, dst_ip*> in a normal IP packet is replaced by a new header of tuple <*src_id, dst_id, src_loc, dst_loc*> in the id packet, which is shown in Figure 3. In this prototype, the IP address in the normal IP packets has the same meaning with the locator in id packets.

B. Packet Encapsulation

The packet encapsulation process of id packet in the id-to-id communication is depicted in Figure 4.

When a packet is generated by the TCP layer, it will be first checked by an *is_ID()* function to determine whether it belongs to an ID-to-ID communication based on its *src_ip* and *dst_ip*, which can be found in the 5-tuple of TCP sockets. If the *src_ip* or *dst_ip* is with IPv6 prefix *2F00*, the packet will be further encapsulated into an id packet by the *id_out()* function. Otherwise, the packet will be sent to the *dst_ip* as a normal IP packet.

If the *2F00* prefix is detected, the drone tries to get the locator of the GS in its own cache and the UAG's cache. If fails, a request will be sent to the ILMS for the retrieval of GS's locator according to its id. Then, the normal packet will be encapsulated as Figure 3 shows. The drone's locator, i.e., the *src_loc*, is assigned when it accesses a UAG. The *dst_loc* is retrieved from caches or from the ILMS. Since we use the ipv6 address with prefix *2F00* as id, the *src_id* in id packet is the same with *src_ip* in the normal packet, and the *dst_id* in id packet is the same with *dst_ip* in the normal packet.

 ISBN: 978-1-61208-550-0

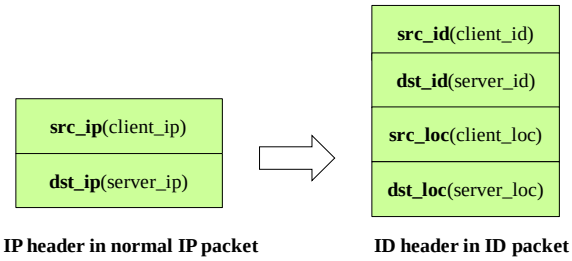


Figure 3. Changes of IP header in ID packet.

At last, the encapsulated ID packet will be sent to the access AP and UAG. The access UAG just treats the locator as the normal IP and forwards all packets as usual according to its routing table.

C. Packet Decapsulation

The decapsulation process of ID packets is shown in Figure 5. Once a packet is received by the hardware, it will be send to the IP layer and checked by the *is_ID()* function to determine whether it's an id packet or not. If the packet is a normal packet, it will be sent to the TCP layer directly. Otherwise, it will be treated as an id packet and further decapsulated by the *id_in()* function. The *id_in()* strips the locator header, *src_loc* and *dst_loc* fields. Then the stripped packet will be sent to the TCP layer as a normal packet.

It should be noted that in this prototype, the ID hosts (i.e., the drone and the GS) are designed to be aware of ID/locator separation. The locator header of ID packets is encapsulated and decapsulated at the drone for realization convenience. In fact, the ID/locator separation network can also be designed as that the hosts are completely unaware of ID/locator separation. This can be realized by embedding the encapsulation and decapsualation of id packets into gateways rather than hosts.

D. Handover

When the drone moves outside the range of its access AP, a handover process must be handled. Since the layer-2 handover [12][13] doesn't lead to changes of locator, we only consider the layer-3 handover in this prototype. Only one AP is deployed under each UAG, which means when the drone flies across different APs, its locator will change, leading to a layer-3 handover.

The layer-3 handover process is detailed in Figure 6.

Step 0: the drone, with id *2F00::1* and locator *10::2* assigned by UAG, communicate with the GS, whose id is *2F00::2*, via UAG1.

Step 1: The drone probes the signal strength of the access AP. Once it detects the signal strength is lower than a threshold, the handover process will be activated. Then the drone sends a handover notification to UAG1.

Step 2: Upon receiving the notification, UAG1 will send a confirm information to the drone and starts to caches packets with *dst_loc* or *des_ip* equals to *10::2*.

Step 3: After receiving the confirmation from UAG1, the drone disconnects from the UAG1-AP and tries to connect the AP under UAG2. If success, the drone will get a new locator *11::2*, which is assigned by UAG2. Then the drone uses the new locator to notify the ILMS as well as the GS that its locator has changed from *10::2* to *11::2*. The ILMS and the GS then update their mapping item related to id *2F00::1* and return the confirmation to the drone that its locator has been updated. At the same time with sending the locator update notification, the GS will also send its new locator to UAG1, notifying UAG1 that it has successfully finished the handover and requests for the cached packets. Upon receiving the notification, UAG1 also sends a confirmation to the drone.

Step4: With the same id *2F00::1* and the new locator *11::2*, the drone continues the id-to-id communication with the GS. The packets in fly will also be tunneled to the drone according to the new locator.

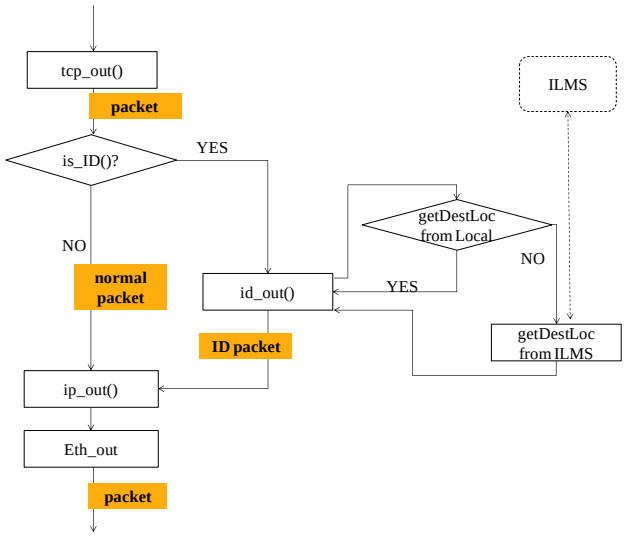


Figure 4. Packet encapsulation process.

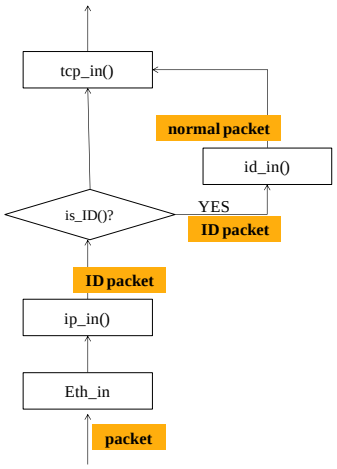


Figure 5. Packet decapsulation process.

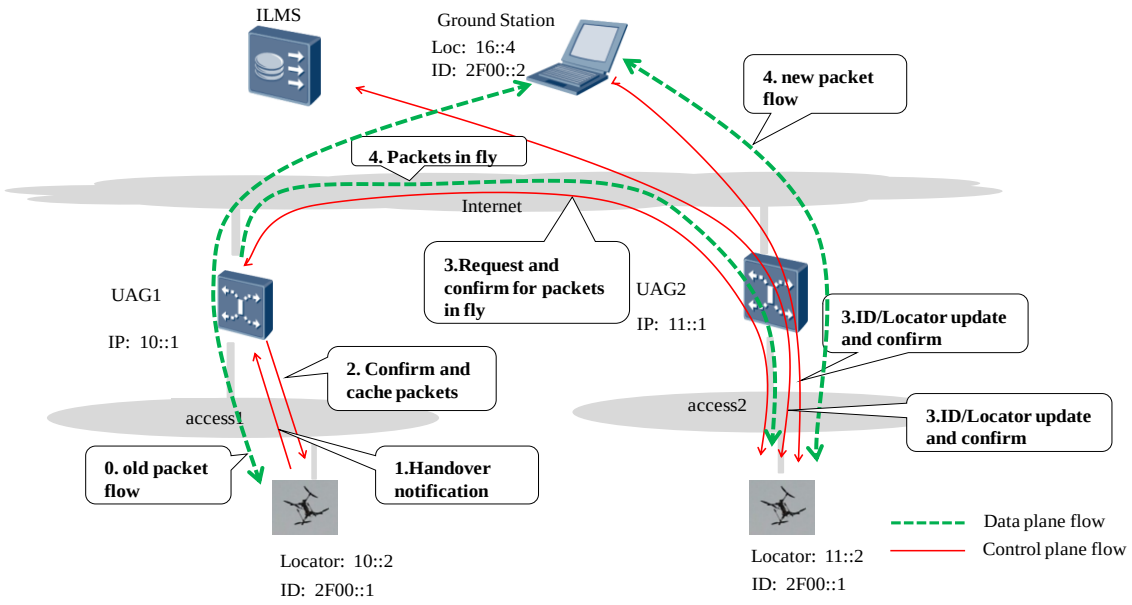


Figure 6. Handover process in id-to-id communication.

From the view of the GS, the corresponding node in the ID-to-ID communication is always the drone during the handover process. Thus, changes of the drone's location is transparent to the upper layers above including TCP/IP, and the ID-to-ID connection can be kept continuous.

V. CONCLUSION

In this paper, we presented a drone prototype based on the idea of ID/Locator separation in the ION. ID is designed as the only identifier of hosts, while the locator is only used for routing and packet forwarding. ID-to-ID communication is realized between the drone and the ground station. We also proposed some protocol principles to define the format, as well as encapsulation/decapsulation of id packets. The handover process is also designed.

The basic idea of ID/Locator separation is now widely accepted by researchers and Internet organizations such as IETF. This prototype shows that this basic idea is a feasible and positive way to solve the scaling issue in the current Internet Protocol.

REFERENCES

[1] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and Principles of Internet Traffic Engineering". IETF Internet Standard, RFC 3272, May 2002.

[2] BGP Routing Table Analysis Reports, http://bgp.potaroo.iiet/, 2012.

[3] D. Meyer, L. Zhang, and K. Fall, "Report from the IAB Workshop on Routing and addressing". IETF Internet Standard, RFC4984, September 2007.

[4] R. Koodli, Ed., "Fast Handovers for Mobile IPv6", IETF Internet Standard, RFC4086, July 2005.

[5] D. Farinacci, V. Fuller, D. Meyer, and D.Lewis, "The Locator/ID Separation Protocol (LISP)," IETF Internet Standard, RFC6830, January 2013.

[6] R. Moskowitz and P. Nikander, "Host Identity Protocol (HIP) Architecture", IETF Internet Standard, RFC 4423, May 2006.

[7] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, "Host Identity Protocol, IETF Internet Standard", RFC5201, April 2008.

[8] Henderson, T. R., Ahrenholz, J. M., and Kim, J. H., "Experience with the host identity protocol for secure host mobility and multihoming," In IEEE Wireless Communications and Networking, pp. 2120-2125 ,2003.

[9] DGS/NGP-004, "Next Generation Protocols: Evolved Architecture for mobility using Identity Oriented Networks".

[10] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," IETF RFC 3775, June 2004.

[11] R. Koodli, "Fast Handovers for Mobile IPv6," IETF RFC 4068, July 2005.

[12] H. Soliman, C. Castelluccia, K. El Marlki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility management," IETF RFC 5380, Oct. 2008.

[13] H. Y. Jung, H. Soliman, S. J. Koh, and J. Y. Lee, "Fast Handover for Hierarchical MIPv6," IETF Internet Draft, April 2005.

 ISBN: 978-1-61208-550-0

Enabling Advanced Network Services in the Future Internet Using Named Object Identifiers and Global Name Resolution

Shreyasee Mukherjee, Parishad Karimi, Dipankar Raychaudhuri
WINLAB, Rutgers University
North Brunswick, New Jersey, USA.
Email: {*shreya, parishad, ray*}*@winlab.rutgers.edu*

Francesco Bronzino
Inria
Paris, France
Email: *francesco.bronzino@inria.fr*

***Abstract*—This paper presents the concept of named object identifiers as the architectural foundation for realizing advanced services, mobility and security in the future Internet. The proposed named object approach uses unique identifiers for service definition and end-to-end message delivery, and can be added as a new layer on top of the IP architecture in a backward-compatible manner. The proposed ID-based service layer requires control plane support in the form of a global name resolution service (GNRS) for dynamic binding of names to network addresses. The requirements for a generalized and flexible name resolution service are discussed considering both functional and performance aspects. Several proposed realizations of the name resolution service are described, including DMap and Auspice used in the MobilityFirst future Internet architecture. In conclusion, examples are given for some new services supported by the proposed identifier-based architecture and specific identifier-based protocol designs, such as mobility, multi-homing, multicast and context-based services.**

Keywords–Future Internet Architecture (FIA); Named services; Name resolution.

I. INTRODUCTION

The current Internet architecture, which was designed with fixed hosts in mind, uses IP address to identify both the users, as well as their location. This overloading of the namespace, also called location-identity conflation [1], makes deploying basic services such as mobility or multi-network access, challenging. End-to-end protocols such as, TCP are tied to the IP address of an interface which changes as an end-point moves, causing transport and application sessions to break. Group based communication to Internet-of-Things or anycast based cloud service access are important use cases which currently require overlay networking solutions above the IP layer and could benefit from improved network layer services. We believe that it is timely to consider evolving the IP architecture to support location-independent identifier-based communications between ”named objects” in order to realize significant service flexibility and security benefits [1]–[3]. Separation of names or identities (IDs) from network address/locator has been proposed in multiple architectures to facilitate location-independent communication [2]–[6]. Note that while some architectures use *names* to denote network-attached objects, others use identities (IDs), both of which can be loosely defined as a string defining a communicating end-point, and, for the purpose of this paper, we use them interchangeably. Assigning long-lasting unique IDs to different

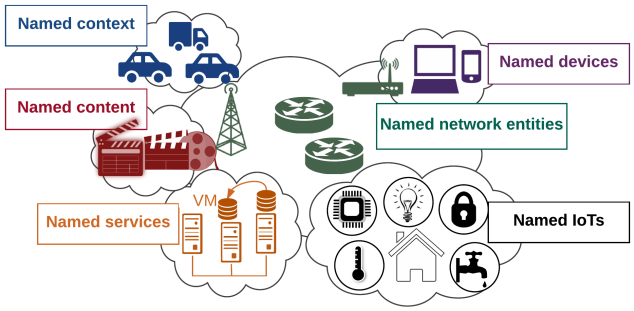


Figure 1. Named-object abstraction: names can be assigned to any network connected entity

network entities ranging from end-points to contents will allow for native support of services on top of any underlying routing mechanism, including IP. As shown in Figure 1, names are quite versatile, and can be used to identify any network-attached object, from traditional end-user devices to IoT groups (for example, sensors in a smart home), specific content in information-centric networks, named network entities (such as access points and routers within a domain) and context, (for example, vehicles on the New Jersey Turnpike between exits 9 and 10).

The naming layer will be placed between the network and transport layer, alleviating the need for those layers to inefficiently support different services like mobility [7] and multihoming [8]. The packet header will include both the ID and the routable address(es), allowing for address-based data traversal within the routers. This will provide a backward-compatible solution, which can be incrementally-deployable on top of the current IP-based Internet. The control plane that enables routing of ID-based packets consists of a globally reachable name resolution service, which will provide name to address mapping to end-points, first-hop routers or any core router depending on the service.

In this paper, we first discuss the necessity of such a global name resolution service (Section II). Then we identify the design requirements for realization of a generalized distributed name resolution service for ID-based networks (Section III). Next we describe two in-network and one overlay name resolution services developed for the MobilityFirst architecture [4] and highlight the set of requirements they fulfill (Section IV). Finally, we explain how such a generic name resolution service

 ISBN: 978-1-61208-550-0

TABLE I. COMPARISON OF EXISTING AND PROPOSED MAPPING RESOLUTION SYSTEMS

System	Mapping resolution	Implementation
Domain Name System (DNS)	URL → IP addr	Dedicated servers in application layer
Network address translation (NAT)	IP addr → IP addr	In-network at NAT-capable routers
MobilityFirst GNRS	GUID name → network addr	In-network routers
LISP ALT	IP addr name → IP addr location	Dedicated overlay servers
Serval	Service ID → sock/addr	In-network at Serval-capable routers

can enable basic mobility or session continuity as well as advanced services, such as multi-network access, large scale multicast and context aware services (Section V).

II. NAME RESOLUTION SERVICES

Clearly, the use of identifiers implies the need for an efficient resolution system that can provide fast and efficient identity to location translation for all such named objects. In the current Internet, two similar resolution systems exist. A distributed globally available Domain Name System (DNS) translates identities (URLs) to obtain network locations (IP addresses) [9] and NAT capable routers locally translate private IP addresses to public IP addresses. However, a key drawback of existing systems is that they were designed based on the notion that most entries will either be static or change at a relatively slow time-scale. Even though DNS has historically evolved significantly from the time it was based on text files to sophisticated hierarchically distributed resolvers, it still lacks the support for the requirements of next generation networks, i.e. a distributed mapping infrastructure that can scale to orders of magnitude higher update rates with orders of magnitude of lower user-perceived latency. Alternatively new designs for distributed resolution services have been proposed for a variety of device, content and service oriented communication, most notably the global name resolution service (GNRS) in MobilityFirst [10]–[12], the distributed overlay ALT servers in LISP [13] as well as distributed translation of service identifiers to interfaces in Serval [14].

Table I summarizes the basic design choices and namespace translation of the aforementioned resolution systems. As shown, each of these resolution systems have their own implementation logic or APIs, and reside in different layers of the internet architecture. While they each focus on slightly different objectives, we believe that it is useful to look into the fundamental requirements for identity-based networks and propose a generic name resolution service with a unified control plane that allows interoperability in the data plane of existing and proposed ID-based architectures.

III. FUNCTIONAL REQUIREMENTS

In this section we describe the key requirements from a resolution system to enable ID oriented future services.

A. Low Update and Response Latency

User perceived latency plays a crucial role in the quality of experience of any digital commercial service subscribers. As reported by VMware, typical network latency of about 100 milliseconds is considered acceptable for the usage of

TABLE II. BASIC RESOLUTION SYSTEM STRUCTURE AND API SEMANTICS

insert(ID, <value set>, opts)
update(ID, <value set>, opts)
query(ID)
delete(ID)

(a) API semantics

Key	Value	Metadata
ID	<values>	Opts
...	...	...

(b) Database structure

their office productivity services [15]. In 2006, Amazon found that every 100 millisecond of added latency reduces sales by 1%. Considering that Amazon's total revenue in 2006 was 19B+, this would have amounted to a loss of 190M per 100 milliseconds of added latency [16]. Future 5G applications such as vehicle-to-vehicle safety messaging or real-time mobile control may require less than one millisecond network latency to be deployable in practice [17], which mandates future name resolution systems to be able to return up-to-date responses within milliseconds.

Note that the network latency includes both the time to update a mapping and the time to return a correct response to a querying entity for the mapping. Therefore, the resolution system should be physically distributed with the distribution optimized to find the sweet spot of minimizing lookup latency and update latency. That is, an ideal resolution system should have potentially all mappings in close network-proximity to both the entities making inserts and queries. While this could be practically hard to achieve in some cases, specially if the entities are topologically far away, it brings forth interesting challenges on how to optimize distribution based on the identity of the service itself; for example, vehicle to vehicle safety communication ($\sim$1 millisecond) vs. locally popular content caching (10s of milliseconds) vs. globally available personal cloud storage($\sim$100 milliseconds).

B. Storage and Load Scalability

There are approximately 4.9 billion global mobile data users and according to a recent study, over 20% of these users currently change network addresses over 10 times a day [18]. Cisco has predicted that by 2021, the number of mobile users will go up to 5.5 billion, whereas the total number of mobile connected devices could be as high as 12 billion [19]. Even if the mobile data users follow certain predictable patterns of mobility, this growth in the number of mobile objects will generate in the order of 10s of billions of daily updates. This in turn would require further resources and create additional workload for the common name resolution infrastructure.

To address these scalability, DNS currently relies heavily on caching of mapping entries through its hierarchy (local name servers, authoritative name servers, top level domains) to help reduce both system load and client-perceived latency. However, handling mobility at this scale requires up-to-date responses, which makes caching ineffective (near-zero TTLs). As a resut, the load and client-perceived latency increase with the mobility rate. Therefore the proposed resolution system should be able to scale to orders of magnitude higher storage and load scalability than existing systems.

C. Extensibility and Flexibility

In order to simplify the deployment of a range of ID based services, the resolution system should be flexible enough to

 ISBN: 978-1-61208-550-0

TABLE III. SUMMARY OF REQUIREMENTS FROM A NAME RESOLUTION SYSTEM

Requirements	Goals
Latency	Low(<1ms) to medium(100ms) based on service req.
Scalabilty	Very high workload(>100B updates per day)
	Moderately high storage based on distribution
Implementation	In-network and distributed
Semantics	Flexible and scalable information schema
	<key, value> pair + supplementary information (optional)
	Standardized APIs
Security	Attack resilient, access control, flexible policies
	Optional confidential info, private instantiations

store multiple kinds of mapping (key→values). For example, it could capture relationships like grouping between names by providing name-to-name mapping and recursive resolution of names. This would not only enable name based multicast communication but also allow a richer information schema to be mapped onto names and then stored in the same resolution system, as explained further in Section V.

The syntax and semantics should also be flexible enough to support a range of existing and future name based architectures [4]–[6], which could all utilize the resolution system as a common control plane, accessible through well-defined standardized APIs. Therefore, the structure of the database itself should not be bound to the structure of the names. For example, HIP [5] and MobilityFirst [4] use flat names, whereas LISP [6] which utilizes IP addresses, has hierarchical names. The database should also allow extensible fields or some form of optional information to be stored per mapping as meta-data, which could be essential for certain kinds of service deployments.

Table IIIa highlights the basic API that includes semantics for inserting a new entry, updating an existing entry, as well as querying and explicitly deleting of an entry. Although time-to-live (TTL) based delete could be performed, similar to DNS, we believe that TTL based designs make it difficult to handle fast mobility as well as temporary disconnections prevalent in wireless access scenarios. Table IIIb further shows the database structure with fields for inserting the ID as the key, a set of values and optional meta-data.

D. Security and Reliability

The resolution service serves as a database, mapping IDs to the location of network-attached objects (which may be correlated to physical locations). Its central role in providing such name resolution entails security and privacy as important design considerations. Local or private instantiations and confidential mappings should also be provisioned for. However, there should not be a single root of trust and strict hierarchical distributions, since, database placement should be optimized based on the service requirements, which in most cases is not closely tied to autonomous systems and network hierarchy. It is also important to allow access control and flexible policy support to prevent malicious usage of the infrastructure [20].

Table III summarizes the broad set of functional requirements for a generic name resolution system for ID-oriented communication in the future internet.

IV. GLOBAL NAME RESOLUTION INFRASTRUCTURE FOR MOBILITYFIRST

MobilityFirst relies heavily on the name resolution service for advanced network-layer functionalities. This reliance necessitates high performance from the resolution service, which depends on resolving identifiers to dynamic attributes in a fast, consistent, and cost-effective manner at Internet scale. Keeping the above requirements in mind, the project has looked into alternative designs of the name resolution system [10]–[12], including both in-network and overlay designs. The MobilityFirst namespace is flat, with globally unique identifiers (GUIDs) that can be assigned to any network-attached object, from individual devices to groups, network routers and services, as shown earlier in Figure 1. These GUIDs are 160 bits and derived from public keys, hence they are self-certifiable and cryptographically secure. Routing is based on network addresses with the name resolution system storing up-to-date mappings between the GUID and its corresponding network addresses. The data packets are also self-sustained and carry both the GUID and the routing address in the header. This ensures in-transit packets to be rebinded by any router along the path, to a new network address through a mapping re-query, as and when required (for example, during mobility). All of the three designs, that is, DMap [10], Auspice [11] and GMap [12], support the basic APIs for insert, update and querying an entry based on GUIDs and have similar database structures with globally distributed implementation and no centralized root of trust.

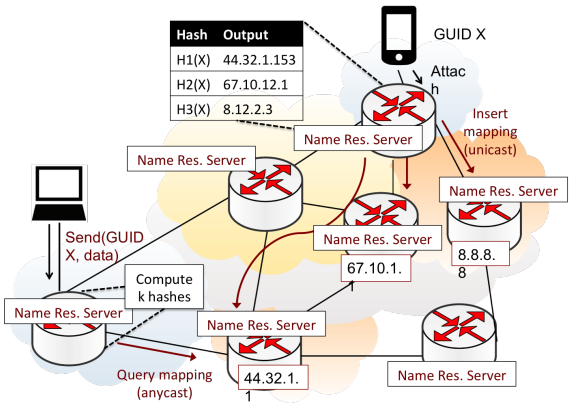


Figure 2. DMap based insertion and lookup of GUID X to locator mapping

DMap: The direct mapping (DMap) design was the first proposed implementation, which is an in-network approach, wherein every autonomous system (AS) in the global network participates in a hashmap based name resolution service in order to share the workload of hosting GUID to network address mapping. Figure 2 provides an overview of how DMap distributes each GUID mapping across K replica servers in the internet. Assuming the underlying routing to be stable and all networks to be reachable, DMap hashes every GUID to K network addresses (which are IP addresses in this example) and then stores the mapping at those K addresses. Every time the mapping changes, K update messages are sent to each of the servers at these locations. Correspondingly, every query for the current mapping of the GUID is anycasted to the nearest of the K locations, as shown.

DMap is the simplest of the three designs and it manages workload balance across all the ASes efficiently. Since uniform hash functions decide *where* a mapping is stored, basic DMap implementation is not suitable for geographically optimized

 ISBN: 978-1-61208-550-0

mapping placement based on service requirements. However the focus of this work was on providing a globally available mapping system with high availability, and moderate latencies, making it ideal to handle basic mobility and services with medium latency requirements. Detailed internet scale simulation of DMap shows that with 5 replicas per GUID, the 95^{th} percentile latency is around 86 milliseconds [10], which is reasonable for most user-mobility centric applications.

Auspice: The main design goal of Auspice, which uses an overlay approach above the network layer, is to provide an automated infrastructure for the placement of geo-distributed name resolvers in order to reduce update and query latencies to tens of milliseconds [11]. The two main components of Auspice are the *replica controllers*, which determine the number and geo-location of name resolvers, and the *name resolvers (active replicas)*, which are responsible for maintaining the identifiers attributes and replying to user-request read or write operations. Each name is associated to a fixed K number of replica-controllers and a variable number of active replicas of the corresponding resolver.
Auspice performs per GUID optimized replica placement with the replica controllers aggregating update and query frequency to compute popularity and hence number of replicas of the mapping required and where to place them. Although the mapping infrastructure is distributed, Auspice is an overlay implementation and does not require in-network routers to participate in sharing the workload. The database design is also more generic with key as the GUID and the mapping being expressed as a <type, length, value> field. Therefore, Auspice can store arbitrary strings as a value mapped onto a GUID. Auspice also takes into account the resource and latency trade off in its optimization for replica management. So if more resources are available, it can decide to disseminate more replicas per GUID and hence reduce overall lookup latency. Detailed comparative evaluation shows that Auspice with 5 replicas is comparable to commercially deployed UltraDNS (16 replicas) and with 15 replicas has 60% lower latency than UltraDNS. Auspice with 5 replicas is also 1.0 to 24.7 secs lower than three top-tier managed DNS service providers for propagating updates globally.

GMap: Finally GMap [12] is an updated version of DMap, in which the GUID→address mapping is distributed hierarchically considering geo-location and local popularity. For each GUID, similar consistent hash functions are used to assign resolution servers. However for each mapping, the servers are categorized into local, regional and global sets, based on geo-locality. Each mapping now gets replicated into K1 local servers, K2 regional servers and K3 global servers. Therefore, unlike Auspice, GMap does not require per-GUID replica optimization, but still achieves better latency than DMap, at the cost of higher storage workload, due to increased number of replicas per GUID. In addition, GMap allows temporary in-network caching of the mapping along the route between a resolution server and a querying entity, to ensure future mapping requests for the same GUID to be resolved faster. Internet-scale simulations show GMap to achieve similar latency goals of tens of milliseconds as Auspice but with lower complexity and computation overhead. Table IV summarizes the key features of each of the designs.

TABLE IV. SUMMARY OF MOBILITYFIRST NAME RESOLUTION SERVICE IMPLEMENTATIONS

	Auspice	**GMap**	**DMap**
Implementation	Overlay	In-network	In-network
Algorithm type	Demand-aware replicated state machine	Distributed hash table	Distributed hash table
Record content	GUID to arbitrary number of values	GUID to arbitrary values (recursively other GUIDs or Network Addresses)	GUID to up to 5 NAs, each with an expiration time and prioritization weight
Name server placement	Geo-located based on requests	Geo-located based on physical location of the GUID	Not Geo-located, except 1 local mapping
Number of replicas per GUID	Based on recent demand and update frequency	Fixed number; each GUID has K1 local, K2 regional, K3 global replicas	Fixed number: each GUID has K global, 1 local replicas
Caching	No caching; load balancing by adjusting number of name servers	Caches response along the path from querying entity and name server	Future work

V. NAME BASED SERVICES

In this section, we explain how a range of services, namely mobility, multihoming, multicast and context-aware services can be supported efficiently, using the concept of "named object" identity within the network.

A. Host and Network Mobility

Due to the rapid proliferation of mobile users, ranging from cellphones to drones, mobility should be treated as a first-class service. One of the most significant use cases for future networks is supporting mobile data services on a fast scale, like authentication and dynamic mobility, involving both micro-level handoff and macro-level roaming. The current approaches for mobility support such as mobile IP [7] suffer from routing inefficiency (in terms of latency, overhead and congestion at service gateways), due to triangular routing through an anchor point. Mobility can be handled better within a name-based architecture which is facilitated by a name resolution service meeting the functional requirements discussed in Section III.

- Baseline: This is the simplest case where on delivery failure, the packet is re-sent from the original sender's location.
- Re-bind (also called "late binding"): When a delivery fails, the name resolution service is queried for an updated location and the packet is forwarded from the current network address, instead of the original sender's location.
- Last Known: This is an extension to the 're-bind' case. The main difference is observed when the user is disconnected and the current location is not available in the name resolution service. In such a case while the 're-bind' scheme holds the packet, waiting for a location update, the 'last known' scheme forwards the packet to the last known location in the GNRS. We expect the user to be closer to his previously known location when compared to the location of the sender.
- Ideal: This scheme represents best possible scenario. Using prediction schemes with the information available in the name resolution service it is possible to get closer to the performance of the ideal case.

 ISBN: 978-1-61208-550-0

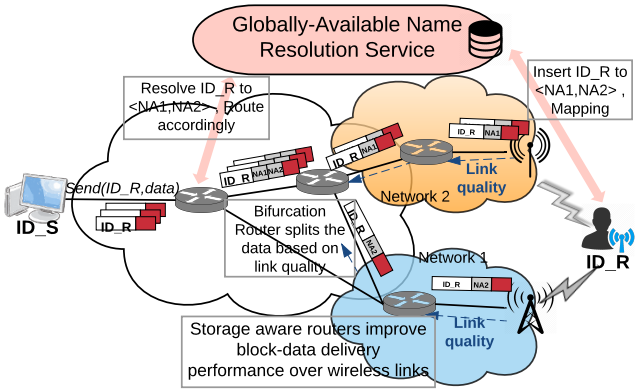


Figure 3. Overview of multihoming supported by a globally-available name resolution service

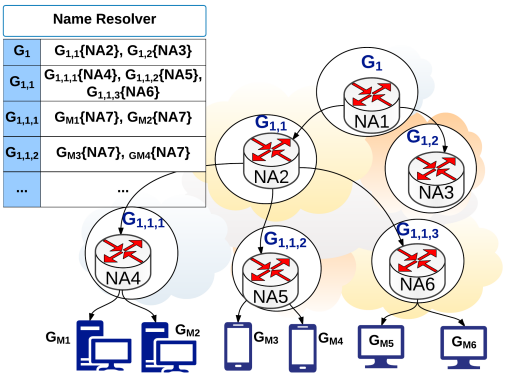


Figure 4. Name based multicast with recursive name lookups using the name resolution service

B. Multi-homed traffic engineering

Multihoming can natively be supported by a name-based architecture. A multi-homed device is simultaneously attached to more than one service network. The separation of names and addresses allows for a device or group name to be bound to a dynamic set of multiple network addresses, denoting the points of attachment of the device to the network. In-network multipath routing is enabled using a global name resolution service as follows: the first hop router that receives a packet destined to another endpoint's name queries the name resolution service for the locations of that name. After receiving the reply from the service, the first hop router appends all the network addresses associated with the receiver's ID to the packet. As data packets traverse through the core network, the routers forward the packets until a branching point is reached. This branching point is the router that faces different next hops for the various network addresses and can dynamically change in case of mobility. This bifurcation router can be programmed to schedule the data on each path according to link quality metrics or policies inserted by the multi-homed endpoints. The link quality metric can utilize cross-layer information from link layer protocols or a feedback mechanism from edge wireless networks. This service can be enabled on top of IP as well, with some limitations on performance, considering the lack of path quality information in current mainstream network and link layer protocols. An overview of how a distributed name resolution service which serves the functional goals discussed earlier can facilitate multihoming is shown in Figure 3.

These approaches have been shown to boost the performance of multihoming compared with current end-to-end approaches such as MPTCP [21], [22]. The extensible fields as metadata for each identifier in the name resolution service can further allow for storing fine-grained expressive policy information about the multi-path connection, e.g., prefer WiFi to LTE; or use WiFi for delay-tolerant downloads and LTE for delay-sensitive traffic, etc.

C. Large-scale multicast

Internet applications like video streaming, online gaming and social networks, e.g. Twitter, often require dissemination of the same piece of information to multiple consumers at the same time. While multicast routing protocols have long been available, most of these applications rely on unicast based solutions without support from the network. Using appropriate multicast routing solutions would help, however, existing network-layer multicast solutions (e.g., PIM-SM [23], MOSPF [24]) have not been widely adopted, mainly because issues with scalability and coordination across multiple domains. In view of the shortcomings of existing schemes, a network-layer multicast solution, that utilizes the named-object abstraction was designed as part of the MobilityFirst project. In this design, names are used to identify a multicast group, as well as the multicast tree itself and can be stored and managed in a distributed fashion through the name resolution infrastructure. As shown in Figure 4, a multicast service-manager computes the multicast tree and assigns GUIDs to each of the branching routers of the tree. This tree is then stored in the resolution service in a recursive manner, wherein each branching router maps onto the next set of downstream branching points along with their network addresses, with the leaves of the tree being the names of the actual devices subscribed to the multicast group. Data packets are sent encapsulated from one branching point to the next, with the outer header containing the GUID and network addresses of the branching router, whereas the inner header containing GUID of the multicast group. Our detailed simulations in [25] show that name-based multicast scales elegantly as the group size and network size increase compared to inter-domain IP multicast [26].

D. Next-generation context-aware services

Finally, using the same name abstraction and the name resolution service, we would like to highlight how a rich set of context-aware services can be supported. Figure 5 shows one such context, where a survivor wants to send a message to "firemen dealing with incident X". As shown in the figure, the information layer is very rich and can include a complicated graph of relationships, including incident hierarchy (all incidents→incident X→X Fire), geographical hierarchy (US→<NJ, CA>), responder relationships (first responders→<police, firemen>) and so on. However, these can be mapped onto a flat naming plane using GUIDs through an object resolution service, as shown. Therefore, the information schema can be flat with no relationships (for example, individual devices), strictly hierarchical (for example, content names in a content centric network [27]) or a mix of all of the above (such as Wikipedia categories [28]), but can still be efficiently mapped into a flat namespace, by cleanly separating the information-space from the namespace. Next the name

 ISBN: 978-1-61208-550-0

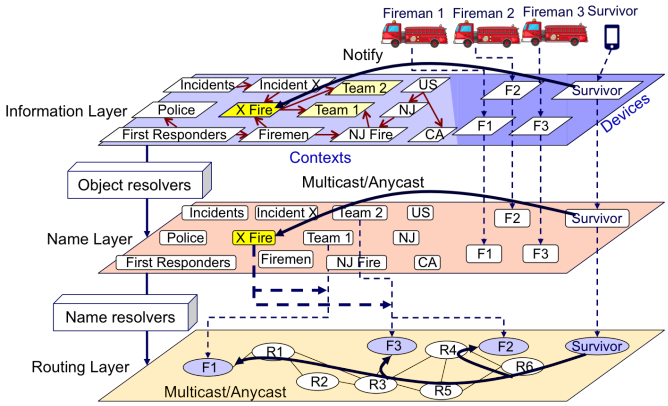


Figure 5. Context-aware services: Mapping a message, *Send* ("Fireman dealing with incident X", "Help message") from a survivor using names

resolution service can be updated to map these GUIDs to network addresses or recursively to other GUIDs.

The end-points do not need to be aware of the separation or the relationships, which can be handled by specific service managers related to each service. For example, in Figure 5, a disaster-management service manager, can determine the information schema, assign GUIDs and update the object resolvers and the name resolvers, such that when a survivor sends a contextual message (send to all all firemen handling incident X), the application on the end-host maps this context to an appropriate GUID and the network in-turn maps this GUID to an appropriate set of network addresses and anycasts or multicasts the message based on the service requirements. Ongoing work in MobilityFirst is focused on efficient design of the object resolvers and the name assignment services for enabling efficient contextual delivery use-cases. [29]

VI. Conclusion

This paper identifies the key set of requirements for a generic resolution service as a unified control plane for identifier-based architectures. Three alternative implementations of a global name resolution infrastructure were described and compared in terms of their design choices and trade-offs. Finally, the paper explained how advanced services such as mobility, multihoming and multicast and context-aware services can be supported using named-object service abstractions along with an efficient name resolution service.

Acknowledgment

The authors would like to thank Dr. Jiachen Chen, WINLAB, Rutgers University for his help with the figures and crucial feedback. This research was supported by the NSF Future Internet Architecture (FIA) grant CNS-134529.

References

[1] J. Saltzer, "On the Naming and Binding of Network Destinations." RFC 1498, 1993.

[2] D. Clark, R. Braden, A. Falk, and V. Pingali, "FARA: Reorganizing the addressing architecture," in ACM SIGCOMM CCR, 2003.

[3] H. Balakrishnan, K. Lakshminarayanan, S. Ratnasamy, S. Shenker, I. Stoica, and M. Walfish, "A layered naming architecture for the internet," in ACM SIGCOMM CCR, 2004.

[4] A. Venkataramani et al., "Mobilityfirst: A mobility-centric and trustworthy internet architecture," SIGCOMM CCR, 2014.

[5] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, "Host Identity Protocol." RFC 5201, 2008.

[6] D. Farinacci, D. Lewis, D. Meyer, and V. Fuller, "The Locator/ID Separation Protocol (LISP)." RFC 6830, 2013.

[7] C. E. Perkins, "Mobile ip," IEEE Communications Magazine, 1997.

[8] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses." RFC 6824, 2015.

[9] P. Mockapetris, "Domain Names - Concepts and Facilities." RFC 1034, 1987.

[10] T. Vu et al., "Dmap: a shared hosting scheme for dynamic identifier to locator mappings in the global internet," in IEEE ICDCS, 2012.

[11] A. Sharma et al., "A global name service for a highly mobile internetwork," in ACM SIGCOMM Computer Communication Review, 2014.

[12] Y. Hu, R. D. Yates, and D. Raychaudhuri, "A Hierarchically Aggregated In-Network Global Name Resolution Service for the Mobile Internet," in WINLAB TR 442.

[13] V. Fuller, D. Farinacci, D. Meyer, and D. Lewis, "Lisp alternative topology (lisp+ alt)." RFC 6836, 2013.

[14] E. Nordström et al., "Serval: An end-host stack for service-centric networking," in Proc. of USENIX NSDI, 2012.

[15] "VMware View 5 with PCoIP, Network Optimization Guide White Paper," www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/view/vmware-view-5-pcoip-network-optimization-guide-white-paper.pdf, 2011 [accessed: 2017-02].

[16] G. Linden, "Make Data Useful," www.gduchamp.com/media/StanfordDataMining.2006-11-28.pdf, 2006 [accessed: 2017-02].

[17] "5G:A Technology Vision," www.huawei.com/5gwhitepaper, 2013 [accessed: 2017-02].

[18] Z. Gao, A. Venkataramani, J. F. Kurose, and S. Heimlicher, "Towards a Quantitative Comparison of Location-Independent Network Architectures ," in Proc. of ACM Sigcomm, 2014.

[19] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021," 2017.

[20] X. Liu, W. Trappe, and J. Lindqvist, "A policy-driven approach to access control in future internet name resolution services," in Proc. of ACM MobiArch, 2014.

[21] P. Karimi, I. Seskar, and D. Raychaudhuri, "Achieving high-performance cellular data services with multi-network access," in IEEE Globecom, 2016.

[22] S. Mukherjee, A. Baid, I. Seskar, and D. Raychaudhuri, "Network-assisted multihoming for emerging heterogeneous wireless access scenarios," in Proc. of IEEE PIMRC, 2014.

[23] D. Farinacci et al., "Protocol independent multicast-sparse mode (PIM-SM): Protocol specification," in RFC2362, 1998.

[24] J. Moy, "Multicast extensions to OSPF," in IETF RFC 1584, 1994.

[25] S. Mukherjee, F. Bronzino, S. Srinivasan, J. Chen, and D. Raychaudhuri, "Achieving Scalable Push Multicast Services Using Global Name Resolution," in Proc. of IEEE Globecom, 2016.

[26] D. Meyer and B. Fenner, "Multicast source discovery protocol (MSDP)," in RFC 3618, 2003.

[27] V. Jacobson et al., "Networking named content," in Proceedings of emerging networking experiments and technologies. ACM, 2009.

[28] "Wikipedia Categories," http://en.wikipedia.org/wiki/Help:Categories, [accessed: 2017-02].

[29] J. Chen, M. Arumaithurai, X. Fu, and K. Ramakrishnan, "CNS: Content-oriented notification service for managing disasters," in Proc. of ACM ICN, 2016.

 ISBN: 978-1-61208-550-0

Cross-Silo and Cross-Eco IoT Communications with ID Oriented Networking (ION)

Bin Da, Richard Li, Xiaofei Xu, Xiaohu Xu
NGIP Laboratory, Beijing Huawei Digital Technologies Co., Ltd.
No.156 Beiqing Street, Haidian District, Beijing, P.R.China, 100095
Email: {dabin, renwei.li, xuxiaofei, xuxiaohu}@huawei.com

***Abstract*—This paper reviews basic IoT architectures, the corresponding evolution at different stages, and presents generalized IoT interoperations under the trend of cross-silo and cross-ecosystem communications. In line with these trends and requirements, ID Oriented Networking, with the detailed background and implementation framework, is elaborated, which contributes to achieve unified IoT communications in future networks. Specifically, ION has the following key components: Network Mapping System, ID Management System, and ID Relationship Management System. And additionally, ION is able to naturally support universal mobility of IoT terminals and enhance intrinsic security of IoT networks, while also can facilitate internetworking of all virtual and physical things over distinct domains, for a fully connected world. At the end of this paper, the merits, challenges and future work of ION are briefly discussed as well.**

Keywords- Internet of Things; IoT; Identifier Locator Split; ID Oriented Networking; ION; Cross-Silo; Cross-Eco.

I. INTRODUCTION

The Internet of Things (IoT) originates from RFID (Radio Frequency IDentification) and relevant technologies in 1980s, which is formally coined as IoT in 1999 [1]. Since then, the IoT paradigm has evolved in several generations, from the vast usage of tagged things and sensor networks [2], to ubiquitously connected smart things over Internet [2][3], and to recently proposed socialized and cloudified internet of things [4][5]. Along with such evolution direction, IoT is envisioned to become a global infrastructure that is able to interconnect everything in the world, which finally fulfills the objective of Everything as a Service (EaaS) [6].

As surveyed in the literature [2]-[7], the essential components of IoT should consist of: physical things with unique identifiers (IDs) for data capturing and local storage; routing mechanism for remote storage and processing; protocols for interoperability and service provision; and trustworthiness among things for security and privacy. Recently, virtualized entities become a prominent feature or candidate component of IoT's further evolution, which associates the Real World Objects (RWOs) with Virtual World Objects (VWOs) [8], for improved communication response and efficiency. All these components are widely practiced in various scenarios such as wearables, smart home, smart city, connected cars, supply chain, cyber physical system, and so forth.

Furthermore, lots of IoT Alliances or Groups have been emerging in the past few years, such as oneM2M established in 2012, Thread launched in 2014, and Open Connectivity Foundation (OCF) newly formed in 2016 [7]. The typical feature, in the infancy of these alliances or groups, is to unite distinctly siloed IoT enabling technologies for achieving full interoperability, inside their respective ecosystems. However, for IoT to be consumed in a ubiquitous manner and be always accessible, these ecosystems are also required to communicate with each other. Henceforth, referred to as cross-ecosystem or cross-eco, this paper concentrates on providing mechanisms to enable cross-silo and cross-eco communications, in a fully connected IoT world.

In line with aforementioned application scenarios and tendencies, the vision of a smart world can be imagined, where cross-silo and cross-eco interconnections become pervasive as a hidden infrastructure. As a result, for achieving this vision, this paper introduces the concept of ID Oriented Networking (ION), and its specific usage for globally unified IoT communications, while all IoT terminals are assumed to be with intelligence in a foreseen trend.

The remainder is organized as follows: in Section II, the IoT architectures are briefly reviewed, with current IoT interoperation status. Then, the ION is elaborated in detail in Section III, with the corresponding building blocks, implementation framework, essential merits, and key features for IoT interoperability. Afterwards, Section IV discusses the challenges and future work, and Section V finally concludes this paper.

II. ARCHITECTURE, EVOLUTION AND INTEROPERATION

This section firstly reviews traditional IoT architectures which are prominent in industry and academy, then the evolution directions of IoT in the past decades are described. After which, a summary for the current status of generalized IoT interoperations is presented. Finally, the trend of cross-silo and cross-eco IoT communications is highlighted.

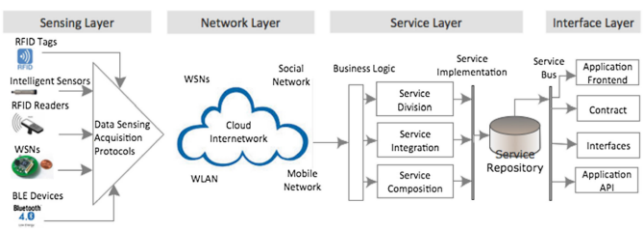


Figure 1. Service Oriented Architecture (SOA).

A. IoT Architecture Review

Traditionally, the Service Oriented Architecture (SOA) and its variants are designed for IoT [2], which generally has

four layers (Sensing, Network, Service, and Interface layers). As shown in Fig.1 [2], the sensing layer normally contains a variety of hardware objects (e.g., RFID tags, sensors and actuators), for acquiring data. The network layer practically facilitates the data transfer over wired or wireless networks. In addition, the service layer generates and manages services whenever required. Lastly, the interface layer presents universal methods that are used by specific applications. Besides this fundamental SOA design, there proposed lots of IoT architectures, with distinct focuses on the applied scenarios [7]. Among which, IoT-A reference model forms a sophisticated architecture, with hundreds of practical IoT requirements into consideration [9].

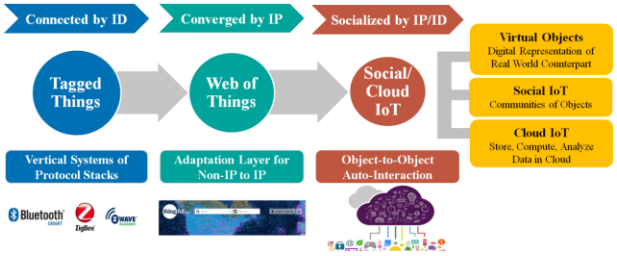


Figure 2. Evolution of Internet of Things (IoT).

B. IoT Evolution

As aforementioned, the term of Internet of Things (IoT) formally emerges in 1999 [1], which mainly builds on the previously developed RFID technologies. Since then, IoT's connotation has been continuously expanding, in particular, the corresponding IoT evolution is briefed in Fig. 2, in line with our analysis. The first generation of IoT is early contributed by the RFID technology, which connects things by RFID tags and transfers data relevant to the things being tagged, for generating meaningful information flows (e.g., for Supply Chain Management). Furthermore, other IoT connectivity technologies are devised [7], including Bluetooth, ZigBee, Z-Wave and so forth, for satisfying vertical applications, as exemplified by connected things indoor or outdoor. However, these distinguished verticals cannot be operable with each other, since they are addressed by different identifiers and interconnected by different mechanisms or protocols [2]. Thus, in the second generation, different vertical technologies usually resort to a gateway for protocol translation, so as to enable cross-silo IoT communications at local scale. Recently, adaption layers are developed (e.g., 6LoWPAN) for further extending vertical IoT domains to be connected with the Internet, which becomes Web of Things after Non-IP and IP convergence [3]. As a result, siloed IoT enabling technologies are able to be interconnected globally via the Internet.

There is also a tendency towards evolution of IoT to be socialized and cloudified. Accordingly, socialized means IoT terminals tend to establish social links just as humans do [4], cloudified implies to build virtual counterparts of physical things in the cloud and to be equipped with cloud computing technologies [5]. The rationales behind such a tendency are multifold: Thing-to-Thing connections are expected to far exceed Human-to-Human connections in near future; Thing-to-Thing connections are also becoming more intelligent and autonomous, with little or no human intervention; Moreover, the data associated with ubiquitously intelligent things and their interconnections will continue exponentially increasing, which finally leads to a large share of IoT data in the cloud. Thus, everything will be intelligent to smartly associate themselves with other things, for on-demand requirements in various applications, which may even resemble human-to-human interactions to formulate thing-to-thing communities with autonomy. The Social IoT (SIoT) is then proposed [4], for systematically describing thing-to-thing relationships and interactions, along with some essential functionalities. Similarly, cloudified IoT solutions are also implemented by different platforms, for integrated data analytics and management over IoT entities [5].

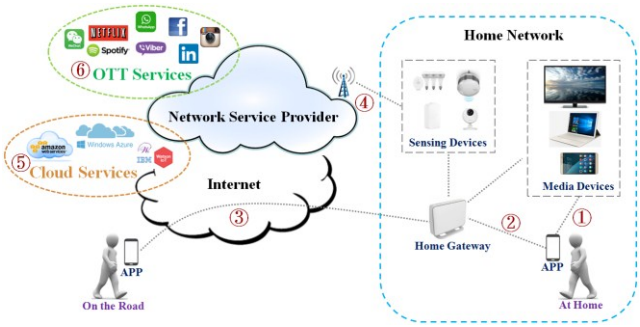


Figure 3. Generalized IoT Interoperations.

C. Current State of Generalized IoT Interoperations

As previously described for the trend of IoT evolution, IoT-based interactions tend to become pervasive, anytime and anywhere, which is not merely for simple data collection but for meaningful service-oriented control. However, in real word, there exist entirely different demands for and types of IoT interoperations, in the manners of Thing-to-Thing and Human-to-Thing connections.

To be more specific, some IoT terminals with fixed positions serve for collecting data locally or remotely, which is usually for centralized data storage and analytics along with few associated actuations (e.g., sensor configuration in remote metering). However, in lot of scenarios, especially with the involvement of human and smart things (i.e., generalized IoT terminals with intelligence), the nearby or remote interactions are aimed for achieving certain services. Without loss of generality, a smart device as the intelligent IoT terminal is exemplified in Fig. 3, for human-involved control on potential interconnections with terminals in a home network, which normally needs an application running in the device as control interface. In Fig. 3, six service-oriented control manners are illustrated with sequential numbers, which are briefly explained below one by one:

① *Direct Point-to-Point (P2P) Operation*: In the first case, the device is able to directly control the surrounding devices via P2P connections, which may use Wi-Fi Direct, Bluetooth Low Energy (BLE), Near Field Communications (NFC) [7]. In this P2P mode, the signal flows are not redirected from any other third parties, and are often used for content sharing, direct actuation, wireless payment.

 ISBN: 978-1-61208-550-0

② *Interaction via Adjacent Gateway*: The device also can go through a nearby gateway to control other IoT terminals, while the gateway practically shields the difference among different IoT technologies such as Wi-Fi, ZigBee, Z-Wave. As in Fig. 3, this is a typical scenario for smart home or smart office. Note that, for Wireless Sensor Network (WSN), there still uses a gateway for collecting data from all sensor nodes through specific IoT enablers, however the interactions are much less as compared to smart home case.

③ *Remote Operation via Internet*: Besides the above two cases for proximity control, the device is able to operate remote IoT terminals through traditional Internet, such as turning on the air conditioner at home on the road.

④ *Interaction via Operator's Gateway*: With the arising of Low Power Wide Area Network (LPWAN) technologies in recent years, the device is entitled to directly connect various IoT terminals at home via the operator's gateway at a remote distance (e.g., 10km away from home), which resembles the home gateway in the second case but with much longer operation distance. Note that these LPWAN technologies are also known as cellular IoT enablers, which are embodied by NB-IoT, LoRa, SigFox and so forth [10].

⑤ *Remote Operation via Cloud*: The cloud service is now integrated with IoT technologies at different levels, which is in line with three cloud types in particular, known as public, private, and hybrid services. These cloud-based services enable centralized control over IoT terminals with data view and data analytics, regardless of the distance. As a result, the device can easily control remote IoT terminals through the cloud services, with hidden underlying IoT technologies.

⑥ *Operation via Over The Top (OTT) Applications*: The human social network applications, like WeChat, Whatsapp, Facebook, are penetrating into all domains of our daily life, including controlling IoT terminals as well. For instance, WeChat is able to perform wireless payment and remote control over smart devices now, and Facebook also can adjust IoT terminal behavior through aforementioned Web of Things. This type of OTT-based operation is actually built upon individual vertical ecosystems with hybrid usage of previous cases.

Note that all the above six interoperation manners continuously generate data, which fully demonstrates the IoT's demand of being integrated with advanced cloud services, forming Cloud of Things (CoT) [5].

D. Trend of Cross-Silo and Cross-Eco IoT Communications

In Section I, the fundamental concept of cross-silo and cross-eco IoT communications is briefly introduced. In this sub-section, a more detailed view is presented in Fig. 4, for elaborating such trend for IoT interoperability. Specifically, in Fig. 4 - (a), it shows the current status of IoT industrial layers with protocol stack, from which, it can be observed that there generally exist two types of IoT channels. One type covers relatively long distance, such as LoRa, SigFox, NB-IoT, which are known as LPWAN. Meanwhile, the others target on short distance connectivity like Bluetooth, ZigBee, Z-Wave. Obviously, these distinct IoT enabling technologies result in siloed operations in various applicable scenarios. Thus, to eliminate the underlying differences below Transport layer, an Adaptation layer can be utilized to link Non-IP and IP enablers with the Internet, the cloud or simply the centralized applications for achieving cross-silo IoT communications. As shown in Fig. 4 – (b), in line with previous philosophy, lots of ecosystems are established accordingly, such as Apple HomeKit, Google Weave, Open Interconnect Consortium, AllSeen Alliance and so forth, for IoT interoperations in Application layer or in Cloud. However, these independently formulated ecosystems become individual bigger silos at their infancies. As a result, for fulfilling the vision of complete interoperability of IoT, the trend of cross-eco communications is arising recently, which is diversely through merging, liaison, asset transfer, or interworking protocols as explicitly illustrated in Fig. 4 – (c) for exemplifying the newly formed Open Connectivity Foundation (OCF). Note that Huawei has established its own IoT ecosystem, which consists of OpenLife Platform, HiLink Protocol, LiteOS, and IoT chipsets.

Based on the observations in Fig. 4, we have proposed a generalized type of internetworking denoted as ION, which adopts the identifier locator split framework and constructs an additional layer below Transport layer for horizontally universal connections, including cross-silo/eco IoT cases. The following section will introduce the details of ION.

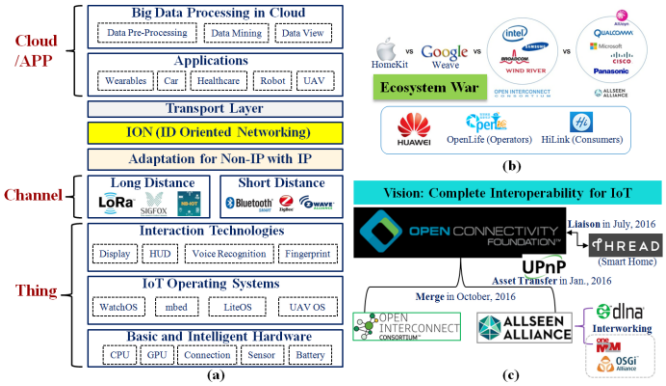


Figure 4. Trend of Cross-Silo and Cross-Eco IoT Communications.

III. ID ORIENTED NETWORKING

This section introduces ID Oriented Networking (ION) concept and architecture in detail, with the background, implementation framework, essential merits, and relevance to cross-silo and cross-eco IoT communications.

A. ION Background

Future networks need to satisfy many demanding requirements such as high throughput, extremely low latency, flexible mobility, intrinsic security, networking automation, and so forth. Recently, at the European Telecommunications Standards Institute (ETSI) Next Generation Protocol (NGP) forum [11], Huawei introduced IP2020 which aims to meet these requirements for various future life scenarios (e.g., autonomous driving, tactile internet, AR/VR). IP2020 is a holistic solution that includes a high-throughput transport

 ISBN: 978-1-61208-550-0

layer, Self-X networking automation, intrinsic network security and ID Oriented Networking.

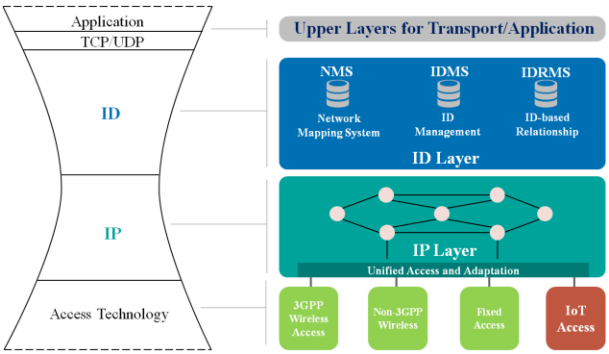


Figure 5. ID Oriented Networking (ION) Overview.

B. *ION Overview*

As shown in Fig. 5, ION follows the idea of Identifier (ID) and Locator Split (ILS) in general [12]. As is well known, the traditional Internet Protocol (IP) address assumes overloaded semantics of being both endpoint identifier and routing locator. Over past years, several proposals have been formulated to decouple the IP into two layers, which contributes to ID and IP layers as shown in Fig. 5 [11]. The IP layer aligns with the successful Internet practices to establish global reachability while ID layer performs functions essential for an endpoint's identity. The ID layer in ION framework has three components: Network Mapping System (NMS) for translating ID to locator whenever queried; ID Management System (IDMS) for centralized or distributed management of universal identifiers; ID Relationship Management System (IDRMS) for maintaining proper relations among ID-labeled physical or virtual entities. In addition, the data or information associated with all these indentified entities should be managed as well, which might resort to cloud-based solutions for vast data storage and analytics and is currently beyond the scope of ION.

As previously mentioned, the idea behind ILS is not novel for usage in ION, which can be observed in many existing ILS research [12]. In the literature, identifiers could be categorized into three classes: IDs over pure IP addresses having different connotations, as in LISP and ILNP; flat IDs based on PKI with self-certifying features, including HIP and MobilityFirst; hierarchical or hybrid IDs, as designed in RANGI [13]. Moreover, for translating ID to locator, many mapping systems are proposed accordingly, such as RVS for HIP and GNRS for MobilityFirst, while our previous work presents a comprehensive summary as well [14].

C. *ION Implementation Framework (IONIF)*

In this sub-section, an ION Implementation Framework (IONIF) towards globally unified IoT communications is elaborated. IONIF is the realization of ION architecture, which integrates ID management, NMS, and IP reachability, to deliver ID aware networks. Applications benefit from ID aware transport using ID-oriented API, and the enabled sockets are location agnostic and can preserve end-to-end connections even the underlying locator layer attributes change. As in Fig. 6, the IONIF has four layers, which are locator, ID, ID-oriented socket API, and application layers, which comply with the previously layered ION overview.

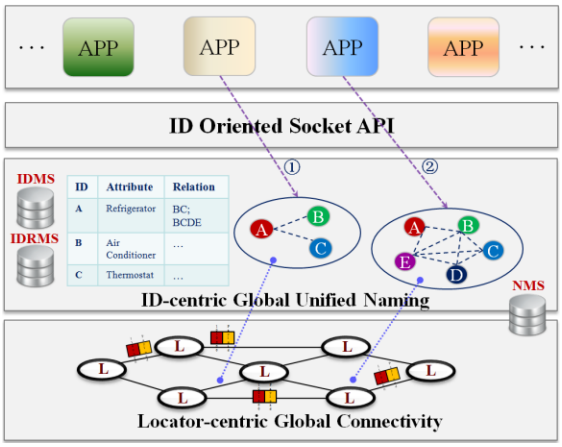


Figure 6. ION Implementation Framework (IONIF).

More specifically, the locator layer aims at achieving global connectivity via locator-based addressing and routing. As shown in Fig. 5, the most promising candidate of such global connectivity locator should be IP and its variants, which specify the destinations for packet-based deliveries [2]. The ID layer, as the core of IONIF, presents unique features for building flexible on-demand relationships horizontally, and satisfying upper layers' demands vertically. With the assistance of a global-scale ID Management System (IDMS), a worldwide unified ID management can be realized, which potentially supports distinct ID formats as well. Along with IDMS, on-demand relationships are created and managed according to specific application requirements in ID Relationship Management System (IDRMS), in a proactive or reactive manner. Furthermore, IDMS and IDRMS could be integrated to manage the identifiers with their respective semantic attributes and relationships, such as ID 'A' indicates a refrigerator associating with other entities in Fig.6.

Furthermore, as previously observed in Fig. 5, the access and IoT hardware heterogeneity has been shielded by the function of unified access and adaptation, thus the ID layer is able to enable Radio Access Technology (RAT) agnostic functions such as ID-based access control, ID-enabled privacy protection, ID-aware AAA, and other policies. In addition, for properly locating communication endpoints and supporting RAT-agnostic mobility management, NMS is dynamically used to map identifiers to locators. The NMS may be maintained by dedicated organizations, working in centralized or hierarchical decentralized manner, resembling the traditional DNS or some new design paradigms [14]-[16].

Above the ID layer, there exists an ID-oriented socket API for ID-aware data transmissions, which provides the interface to application layer and has adaptability to lower-layer ID-based on-demand relationships. Moreover, in application layer, individual applications may request to establish tailored relationships for their operating things through this ID-oriented API. For example, a control application for home appliances, including refrigerator, air conditioner, thermostat etc., requires to build a same-owner

 ISBN: 978-1-61208-550-0

relationship among these appliances belonging to different manufacturers. Note that some fundamental Thing-to-Thing relationships are well investigated in Social IoT (SIoT) [4], which can be referred to for relation establishment in IONIF.

As illustrated in Fig. 6, the horizontal relationship for an IoT community may further embrace a new feature of automatic relation-aware self-expansion, which determines useful and useless relationships for upper layer services by dynamically enrolling new members or removing existing members in a relational cluster. For instance, the relation formulated by pointer ① for one application could be expanded to a renewed relation initiated by the same application, through involving new members with updated relations. Alternatively, individual applications with different services may also activate distinct relationships having partially shared members, as pointed by ① and ② for two applications sharing three common members. For IONIF, these horizontal ID-based relationships are maintained in IDRMS, being assisted by IDMS.

Currently, the IONIF is still under development and refinement, and its core implementation components presented in this sub-section are able to accelerate global connectivity for unified IoT communications in near future. In which, Thing-to-Thing relations are maintained just like human society, and these things' relations are expected to be further intertwined with human behavior and services.

D. ION for IoT Interoperations

Based on above description, IONIF shows the potential for the future IoT interoperations in Fig. 7, other than integrated operations in the application layer. Previously, the IoT evolution has shown the trend towards cross-silo and cross-eco communications. In near future, with the help of ION, a unified IoT cross operation could be easily built upon ID layer, facilitating all the actions demonstrated in Fig. 3. In particular, regardless of IoT enabling technologies (e.g., Bluetooth, Z-Wave, LoRa, etc.), the universal adaptation layer normalizes the data transmitted among different IoT verticals, and further enables the connection with IP layer for global reachability. In addition, IDs defined in ION can persistently label all communication endpoints, without considering their specific routing locations. Note that the dynamic binding from ID to IP for smooth data transmission could be at the level of individual things supporting IP or at the level of IoT gateways with local locators other than IP.

As a result, the heterogeneity of IoT technologies become hidden beneath ID layer, as in Figures 5 and 6, and upper services can request any type of on-demand relationship over IoT terminals, which fully satisfies the future trend.

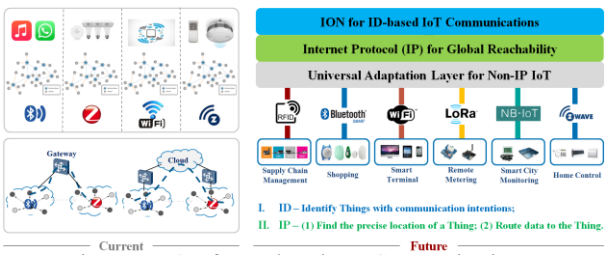


Figure 7. ION for ID-based IoT Communications.

E. ION Merits

As shown in Figures 5-7, the merits of introducing ION for IoT interoperations, other than the present IoT integration in application layer, are multifold, which are concisely summarized below:

- *Labels of Communication Entities*: Currently, the communication identifiers of IoT terminals are in different formats in individual IoT enabling technologies. Thus, a unified identifier naming paradigm is highly desirable in the ID layer as shown in Figures 5-7, for consistent labeling of communicating IoT entities across various domains or ecosystems. Although this unified identifier could be a long-term multilateral effort, there exist a few trials of promoting such type of identifier [13], [15]-[17]. For example, a PKI-based hash value in binary format may be used as a unique 128-bit identifier as suggested in RANGI scheme [13].
- *Intrinsic Security*: Besides the link-level security in pairing stage of individual IoT technologies (e.g., Bluetooth and ZigBee), an intrinsic level of security built upon Identity Based Signature (IBS) scheme is under development, for the purpose of enforcing future mobility security, network trust, and identity and key management [11]. As expected, authentication before establishing a transport layer connection may close many security holes nowadays in TCP/IP protocols, while further reducing the burden of deep packet inspection and the consequent overhead [18].
- *Mobility Support for IoT Terminals*: ION largely follows the ILS paradigm, as described in previous sections, it thus naturally supports mobility of IoT terminals. Note that the fundamental principle behind mobility support is consistent communication identifiers regardless of location changes [12].
- *Social Community of Things*: This is a prominent feature in line with IoT evolution in Section II, as Thing-to-Thing interconnections become pervasive. Based on the observation in Fig. 6, a social community with on-demand relationships among smart things could be established upon specific service requests. Meanwhile, such social community can be managed similarly as human society, with dynamic enrollment or removal and intelligent interaction, for achieving valued-added functions in autonomy.

IV. FUTURE WORK

For ION utilization in large scale, many challenges are inevitable in front, and are briefly discussed in this section, for the purpose of future work.

As highlighted in Fig. 6 for IONIF, two essential elements of ION, i.e., IDMS and IDRMS, are logically intertwined for managing the universal identifiers and their on-demand relationships. However, unifying distinguished identifier formats of various IoT regimes under a single framework may take unexpected effort to achieve, and the consensus over atomic relationship definition for IoT terminals may encounter similar difficulty. Thus, the ID format definition with various ID support should be revisited,

 ISBN: 978-1-61208-550-0

along their potentially dynamic relationships in a socialized community. Meanwhile, an extended universal adaptation plane might be utilized to bridge existing siloed identifier domains, based on the adaptation layer shown in Fig. 7, which also needs a further study.

As noticed, ION naturally support mobility due to constant communication identifiers, however, the mapping from identifier to locator may take additional time and becomes a new bottleneck. Thus, the NMS in ION should be further explored to fully support mobility of IoT terminals in distinct scenarios, which may accommodate all current ID formats in a unified way. Accordingly, an IDEAS group has been recently formulated in IETF, with the target of new mapping system design with novel principles and proof-of-concept verifications for ILS schemes in general [19]. As a result, a generalized IoT mobility may be enhanced through a united endeavor over NMS design in near future.

Furthermore, the security imposed by ION over IoT communications should be well designed so as to enable all-round protections. As aforementioned, IoT security can be boosted after the introduction of ION in ID layer for future networks. However, since the interconnections and the accompanying data with the IoT terminals continue to be dramatically increasing, the security in every phase could be threatened, which occurs either in cross layer or in a hybrid manner. Thus, formulating a holistic security scheme, with consideration of identification, authentication, integrity, privacy, trust, safety, reliability, responsiveness, immunity, autonomy and so forth, is always a challenging work for candidate research [18].

ION socket, for broadly enabling ION implementation, may require modifications on host side. Thus, the problem of smooth adoption of ION in large scale, with minimum impacts on other layers is worthwhile to be further examined. The viable solutions might be either through a middleware for properly linking legacy and ION-based transmissions, or through an ID-aware socket that understands intrinsic connotations when legacy and ION-based IDs are actually utilized.

As previously observed in IoT evolution, integrating IoT technologies with cloud computing is also a desirable trend for ION to serve IoT practices with hugely manageable data behind identifiers. Thus, hierarchical cloud enabled (i.e., fog/edge/core clouds) IoT under ION framework is a valuable extension as well.

In summary, for achieving unified IoT communications, the functional components and key enabling technologies under the proposed ION framework are of importance for future refinement and study.

V. CONCLUSION

In this paper, the basic IoT architectures with its evolution stages are firstly introduced, which is followed by the driving forces and trends for cross-silo and cross-eco IoT interoperations. Subsequently, ID Oriented Networking (ION) with the corresponding background, core functional components, and implementation framework are elaborated. Finally, the merits and future work are briefly discussed.

Overall, a smart world with unified communications under ION framework is imaginable, where generalized intelligent things in all types are agilely interconnected for providing integrated services to numerous local and global demanders.

REFERENCES

[1] Kevin Ashton, "That 'internet of things' thing," RFID Journal, pp. 97-114, 2009.

[2] S. Li, D. Li, and S. Zhao, "The internet of things: a survey," Information Systems Frontiers, pp. 243-259, 2015.

[3] D. Zeng, S. Guo, and Z. Cheng, "The web of things: a survey," Journal of Communications, pp. 424-438, 2011.

[4] L. Atzori, A. Iera, G. Morabito, et al, "The social internet of things (SIoT) - when social networks meet the internet of things: concept, architecture and network characterization," Computer Networks, pp. 3594-3608, 2012.

[5] S. Distefano, G. Merlino, and A. Puliafito, "Enabling the cloud of things," in Proceedings of IEEE Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012, pp. 858-863.

[6] A. Botta, W. Donato, et al, "Integration of cloud computing and internet of things: a survey," Future Generation Computer Systems, pp. 684-700, 2016.

[7] A. Al-Fuqaha, M. Guizani, et al, "Internet of things: a survey on enabling technologies, protocols, and applications," IEEE Communications Surveys & Tutorials, pp. 2347-2376, 2015.

[8] I. Farris, R. Girau, et al, "Social virtual objects in the edge cloud," IEEE Cloud Computing, pp. 20-28, 2015.

[9] IoT-A Reference Model: http://www.iot-a.eu

[10] J. P. Bardyn, T. Melly, et al, "IoT: The era of LPWAN is starting now," in Proceedings of IEEE European Solid-State Circuits Conference (ESSCIRC), Oct. 2016, pp. 25-30.

[11] Huawei IP 2020 Project Introduction is available: http://www.layer123.com/download&doc=Huawei-1016-Renwei-Towards_2020-Challenges

[12] W. Ramirez, X. Masip-Bruin, et al, "A survey and taxonomy of ID/Locator Split Architectures (ILSA)," Computer Networks, pp. 13-33, 2014.

[13] Y. Jia, X. Lu, et al, "A novel host mobility support method in IPv4/IPv6 network of RANGI architecture," in Proceedings of IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012, pp. 2083-2087.

[14] Bin Da, X. Xu, K. Bi, and X. Zheng, "DNS with mapping service in identifier locator split architecture," in Proceedings of 22nd APCC Conference, 2016, pp. 470-475.

[15] A. Sharma, X. Tie, et al, "A global name service for a highly mobile internetwork," ACM SIGCOMM Computer Communication Review, vol. 44, no. 4, pp. 247-258, 2014.

[16] V. P. Kafle, Y. Fukushima, and H. Harai, "ID-based communication for realizing IoT and M2M in future heterogeneous mobile networks," in Proceedings of IEEE International Conference on Recent Advances in IoT, 2015.

[17] Felice Armenio, Henri Barthel, et al, "The EPCglobal architecture framework," 2005.

[18] J. Granjal, E. Monteiro, and J. S. Silva, "Security for the internet of things: a survey of existing protocols and open research issues," IEEE Communications Surveys & Tutorials, pp. 1294-1312, 2015.

[19] IETF ID EnAbled networkS (IDEAS) Mailing List: https://mailarchive.ietf.org/arch/search/?email_list=ideas

 ISBN: 978-1-61208-550-0

Reliability Assessment of Erasure Coded Systems

Ilias Iliadis and Vinodh Venkatesan
IBM Research – Zurich
8803 Rüschlikon, Switzerland
Email: {ili,ven}@zurich.ibm.com

***Abstract*—Replication is widely used to enhance the reliability of storage systems and protect data from device failures. The effectiveness of the replication scheme has been evaluated based on the Mean Time to Data Loss** (MTTDL) **and the Expected Annual Fraction of Data Loss** (EAFDL) **metrics. To provide high data reliability at high storage efficiency, modern systems employ advanced erasure coding redundancy and recovering schemes. This article presents a general methodology for obtaining the** EAFDL **and** MTTDL **of erasure coded systems analytically for the symmetric, clustered and declustered data placement schemes. Our analysis establishes that the declustered placement scheme offers superior reliability in terms of both metrics. The analytical results obtained enable the derivation of the optimal codeword lengths that maximize the** MTTDL **and minimize the** EAFDL**.**

Keywords–Reliability metric; MTTDL; EAFDL; RAID; MDS codes; Information Dispersal Algorithm; Prioritized rebuild.

I. INTRODUCTION

The reliability of storage systems is affected by data losses due to device and component failures, including disk and node failures. Permanent loss of data is prevented by deploying redundancy schemes that enable data recovery. However, additional device failures that may occur during rebuild operations could lead to permanent data losses. Over the years, several redundancy and recovery schemes have been developed to enhance the reliability of storage systems. These schemes offer different levels of reliability, with varying corresponding overheads due to the additional operations that need to be performed, and different levels of storage efficiencies that depend on the additional amount of redundant (parity) data that needs to be stored in the system.

The effectiveness of the redundancy schemes has been evaluated predominately based on the Mean Time to Data Loss (MTTDL) metric. Closed-form reliability expressions are typically obtained using Markov models with the underlying assumption that the times to component failures and the rebuild times are independent and exponentially distributed [1-13]. Recent work has shown that these results also hold in the practical case of non-exponential failure time distributions. This was achieved based on a methodology for obtaining MTTDL that does not involve any Markov analysis [14]. The MTTDL metric has been used extensively to assess tradeoffs, to compare schemes and to estimate the effect of various parameters on system reliability [15-18].

To cope with data losses encountered in the case of distributed and cloud storage systems, data is replicated and recovery mechanisms are used. For instance, Amazon S3 is designed to provide 99.999999999% (eleven nines) durability of data over a given year [19]. Similarly, Facebook [20], LinkedIn [21] and Yahoo! [22] consider the amount of data lost in given periods. To address this issue, a recent work has introduced the Expected Annual Fraction of Data Loss (EAFDL) metric [23]. It has also presented a methodology for deriving this metric analytically in the case of replication-based storage systems, where user data is replicated r times and the copies are stored in different devices. As an alternative to replication, storage systems use advanced erasure codes that provide a high data reliability as well as a high storage efficiency. The use of such erasure codes can be traced back to as early as the 1980s when they were applied in systems with redundant arrays of inexpensive disks (RAID) [1][2]. The RAID-5, RAID-6 and replication-based systems are special cases of erasure coded systems. State-of-the-art data storage systems [24][25] employ more general erasure codes, where the choice of the codes used greatly affects the performance, reliability, and storage and reconstruction overhead of the system. In this article, we focus on the reliability assessment of erasure coded systems and how the choice of codes affects the reliability in terms of the MTTDL and EAFDL metrics.

The MTTDL of erasure coded systems has been obtained analytically in [26]. To reduce the amount of data lost, it is imperative to assess not only the frequency of data loss events, which is obtained through the MTTDL metric, but also the amount of data lost, which is expressed by the EAFDL metric [23]. The EAFDL and MTTDL metrics provide a useful profile of the size and frequency of data losses. Towards that goal, we present a general framework and methodology for deriving the EAFDL analytically, along with the MTTDL, for the case of erasure coded storage systems. The model developed captures the effect of the various system parameters as well as the effect of various codeword placement schemes, such as clustered, declustered, and symmetric data placement schemes. The results obtained show that the declustered placement scheme offers superior reliability in terms of both metrics. We also investigate the effect of the codeword length and identify the optimal values that offer the best reliability.

The remainder of the paper is organized as follows. Section II describes the storage system model and the corresponding parameters considered. Section III presents the general framework and methodology for deriving the MTTDL and EAFDL metrics analytically for the case of erasure coded systems. Closed-form expressions for the symmetric, clustered, and declustered placement schemes are derived. Section IV compares these schemes and establishes that the declustered placement scheme offers superior reliability. Section V presents a thorough comparison of the reliability achieved by the declustered placement scheme under various codeword configurations. Finally, we conclude in Section VI.

II. STORAGE SYSTEM MODEL

The storage system considered comprises n storage devices (nodes or disks), with each device storing an amount c of data,

 ISBN: 978-1-61208-550-0

TABLE I. NOTATION OF SYSTEM PARAMETERS

Parameter	Definition
n	number of storage devices
c	amount of data stored on each device
l	number of user-data symbols per codeword ($l \geq 1$)
m	total number of symbols per codeword ($m > l$)
(l, m)	MDS-code structure
s	symbol size
k	spread factor of the data placement scheme
b	reserved rebuild bandwidth per device
$1/\lambda$	mean time to failure of a storage device
s_{eff}	storage efficiency of redundancy scheme ($s_{\text{eff}} = l/m$)
U	amount of user data stored in the system ($U = s_{\text{eff}}\, n\, c$)
$\tilde{r}$	minimum number of codeword symbols lost that lead to an irrecoverable data loss ($\tilde{r} = m - l + 1$ and $2 \leq \tilde{r} \leq m$)
$1/\mu$	time to read (or write) an amount c of data at a rate b from (or to) a device ($1/\mu = c/b$)

such that the total storage capacity of the system is $n\,c$. Modern data storage systems use various forms of data redundancy to protect data from device failures. When devices fail, the redundancy of the data affected is reduced and eventually lost. To avoid irrecoverable data loss, the system performs rebuild operations that use the data stored in the surviving devices to reconstruct the temporarily lost data, thus maintaining the initial data redundancy.

A. Redundancy

According to the erasure coded schemes considered, the user data is divided into blocks (or symbols) of a fixed size (e.g., sector size of 512 bytes) and complemented with parity symbols to form codewords. In this article, we consider (l, m) maximum distance separable (MDS) erasure codes, which are a mapping from l user data symbols to a set of m ($> l$) symbols, called a codeword, in such a way that any subset containing l of the m symbols of the codeword can be used to decode (reconstruct, recover) the codeword. The corresponding storage efficiency, s_{eff}, is given by

$$s_{\text{eff}} = \frac{l}{m}\ , \tag{1}$$

such that the amount of user data, U, stored in the system is given by

$$U = s_{\text{eff}}\, n\, c \ = \ \frac{l\, n\, c}{m}\ . \tag{2}$$

The notation used is summarized in Table I. The parameters are divided according to whether they are independent or derived, and are listed in the upper and the lower part of the table, respectively.

The m symbols of each codeword are stored on m distinct devices, such that the system can tolerate any $\tilde{r} - 1$ device failures, but a number of $\tilde{r}$ device failures may lead to data loss, with

$$\tilde{r} = m - l + 1\ . \tag{3}$$

From the preceding, it follows that

$$1 \leq l < m \qquad \text{and} \qquad 2 \leq \tilde{r} \leq m\ . \tag{4}$$

Examples of MDS erasure codes are the following:

Replication: A replication-based system with a replication factor r can tolerate any loss of up to $r - 1$ copies of some data, such that $l = 1$, $m = r$ and $\tilde{r} = r$. Also, its storage efficiency is equal to $s_{\text{eff}}^{(\text{replication})} = 1/r$.
RAID-5: A RAID-5 array comprised of N devices uses an $(N - 1, N)$-MDS code, such that $l = N - 1$, $m = N$ and $\tilde{r} = 2$. It can therefore tolerate the loss of up to one device, and its storage efficiency is equal to $s_{\text{eff}}^{(\text{RAID-5})} = (N - 1)/N$.
RAID-6: A RAID-6 array comprised of N devices uses an $(N - 2, N)$-MDS code, such that $l = N - 2$, $m = N$ and $\tilde{r} = 3$. It can therefore tolerate a loss of up to two devices, and its storage efficiency is equal to $s_{\text{eff}}^{(\text{RAID-6})} = (N - 2)/N$.
Reed–Solomon: It is based on (l, m)-MDS erasure codes.

B. Symmetric Codeword Placement

We consider a placement where each codeword is stored on m distinct devices with one symbol per device. In a large storage system, the number of devices, n, is typically much larger than the codeword length, m. Therefore, there exist many ways in which a codeword of m symbols can be stored across a subset of the n devices. For each device in the system, let its *redundancy spread factor* k denote the number of devices over which the codewords stored on that device are spread [26]. In a symmetric placement scheme, the $m - 1$ symbols of each codeword corresponding to the data on each device are *equally* spread across $k - 1$ other devices, such that these devices alltogether form a group of k devices. It also holds that the $m - 2$ codeword symbols corresponding to the codewords shared by any two devices within this group are equally spread across $k - 2$ other devices, and so on. Consequently, all the symbols of each codeword in the system are contained within such a group, which implies that the system is effectively comprised of n/k disjoint groups of k devices. Each group contains an amount U/k of user data, with the corresponding codewords placed on the corresponding k devices in a distributed manner. We proceed by considering the clustered and declustered placement schemes, which are special cases of symmetric placement schemes for which k is equal to m and n, respectively.

1) Clustered Placement: In this placement scheme, the n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to the *clustered* placement, each codeword is stored across the devices of a particular cluster. In such a placement scheme, it can be seen that no cluster stores the redundancies that correspond to data stored on another cluster. The entire storage system can essentially be modeled as consisting of n/m independent clusters. In each cluster, data loss occurs when $\tilde{r}$ devices fail successively before rebuild operations complete successfully.

2) Declustered Placement: In this placement scheme, all $\binom{n}{m}$ possible ways of placing m symbols across n devices are equally used to store all the codewords in the system. This is a symmetric placement scheme, in that for any given device, the same number of the codeword symbols that correspond to the codewords on that device are contained in each set of any other $m - 1$ devices.

C. Codeword Reconstruction

When storage devices fail, codewords lose some of their symbols and this leads to a reduction in data redundancy. The system attempts to maintain its redundancy by reconstructing the lost codeword symbols using the surviving symbols of the affected codewords.

 ISBN: 978-1-61208-550-0

1) Exposure Levels and Amount of Data to Rebuild: At time t, let $D_j(t)$ be the number of codewords that have lost j symbols, with $0 \leq j \leq \tilde{r}$. The system is at exposure level e $(0 \leq e \leq \tilde{r})$, where

$$e = \max_{D_j(t)>0} j. \tag{5}$$

In other words, the system is at exposure level e if there are codewords with $m-e$ symbols left, but there are no codewords with fewer than $m-e$ symbols left in the system, that is, $D_e(t) > 0$, and $D_j(t) = 0$ for all $j > e$. These codewords are referred to as the *most-exposed* codewords. Let the number of most-exposed codewords when entering exposure level e be denoted by C_e, $e = 1, \ldots, \tilde{r}$. At $t = 0$, $D_j(0) = 0$ for all $j > 0$ and $D_0(0)$ is the total number of codewords stored in the system. Device failures and rebuild processes cause the values of $D_1(t), \cdots, D_{\tilde{r}}(t)$ to change over time, and when a data loss occurs, $D_{\tilde{r}}(t) > 0$. Device failures cause transitions to higher exposure levels, whereas rebuilds cause transitions to lower ones.

In this article, we will derive the reliability metrics of interest using the direct path approximation, which considers only transitions from lower to higher exposure levels [14][26][27]. This implies that each exposure level is entered only once.

2) Prioritized or Intelligent Rebuild: At each exposure level e, the *prioritized or intelligent* rebuild process attempts to bring the system back to exposure level $e-1$ by recovering one of the e symbols that each of the most-exposed codewords has lost, that is, by recovering a total number of C_e symbols. Let A_e denote the amount of data corresponding to the C_e symbols and let s denote the symbol size. Then, it holds that

$$A_e = C_e\, s\ . \tag{6}$$

The notation used is summarized in Table II. For an exposure level e $(< \tilde{r})$, A_e represents the amount of data that needs to be rebuilt at that exposure level. In particular, upon the first-device failure, it holds that

$$A_1 = c\ . \tag{7}$$

D. Rebuild Process

During the rebuild process, a certain proportion of the device bandwidth is reserved for data recovery, with b denoting the actual reserved rebuild bandwidth per device. The rebuild bandwidth is usually only a fraction of the total bandwidth available at each device; the remainder is used to serve user requests. Let us denote by b_e $(\leq b)$ the rate at which the amount A_e of data that needs to be rebuilt at exposure level e is written to selected device(s). In particular, let us denote by $1/\mu$ the time required to read (or write) an amount c of data from (or to) a device, given by

$$\frac{1}{\mu} = \frac{c}{b}\ . \tag{8}$$

E. Failure and Rebuild Time Distributions

In this work, we assume that the lifetimes of the n devices are independent and identically distributed, with a cumulative distribution function $F_\lambda(.)$ and a mean of $1/\lambda$. We further consider storage devices with failure time distributions that belong to the large class defined in [14], which includes real-world distributions, such as Weibull and gamma, as well as exponential distributions. The storage devices are *highly reliable* when the ratio of the fixed time $1/\mu$ to read all contents of a device (which typically is on the order of tens of hours) to the mean time to failure of a device $1/\lambda$ (which is typically at least on the order of thousands of hours) is small, that is, when

$$\frac{\lambda}{\mu} = \frac{\lambda\, c}{b} \ll 1\ . \tag{9}$$

TABLE II. NOTATION OF SYSTEM PARAMETERS AT EXPOSURE LEVELS

Parameter	Definition
e	exposure level
C_e	number of most-exposed codewords when entering exposure level e
R_e	rebuild time at exposure level e
$P_{e\to e+1}$	transition probability from exposure level e to $e+1$
$\tilde{n}_e$	number of devices at exposure level e whose failure causes an exposure level transition to level $e+1$
α_e	fraction of the rebuild time R_e still left when another device fails causing the exposure level transition $e \to e+1$
V_e	fraction of the most-exposed codewords that have symbols stored on another of the $\tilde{n}_e$ devices
A_e	amount of data corresponding to the C_e symbols $(A_e = C_e\, s)$
b_e	rate at which recovered data is written at exposure level e

According to [14][26], when the cumulative distribution function F_λ satisfies the condition

$$\mu \int_0^{1/\mu} F_\lambda(t)\, dt \ll 1, \quad \text{with } \frac{\lambda}{\mu} \ll 1\ , \tag{10}$$

the MTTDL reliability metric of replication-based or erasure coded storage systems tends to be insensitive to the device failure distribution, that is, the MTTDL depends only on its mean $1/\lambda$, but not on its density $F_\lambda(.)$. In [23], it was shown that this also holds for the EAFDL metric in the case of replication-based storage systems, and in this article, we will show that this is also the case for erasure coded systems.

III. DERIVATION OF MTTDL AND EAFDL

We briefly review the general methodology for deriving the MTTDL and EAFDL metrics presented in [23]. This methodology does not involve any Markov analysis and holds for general failure time distributions, which can be exponential or non-exponential, such as the Weibull and gamma distributions.

At any point of time, the system can be thought to be in one of two modes: normal mode and rebuild mode. During normal mode, all data in the system has the original amount of redundancy and there is no active rebuild process. During rebuild mode, some data in the system has less than the original amount of redundancy and there is an active rebuild process that is trying to restore the lost redundancy. A transition from normal mode to rebuild mode occurs when a device fails; we refer to the device failure that causes this transition as a *first-device* failure. Following a first-device failure, a complex sequence of rebuild operations and subsequent device failures may occur, which eventually leads the system either to an irrecoverable data loss (DL) with probability P_{DL} or back to the original normal mode by restoring initial redundancy, which occurs with probability $1 - P_{\text{DL}}$. The MTTDL is then given by [23]

$$\text{MTTDL} \approx \frac{1}{n\, \lambda\, P_{\text{DL}}}\ . \tag{11}$$

 ISBN: 978-1-61208-550-0

Let H denote the corresponding amount of data lost conditioned on the fact that a data loss has occurred. The metric of interest, that is, the Expected Annual Fraction of Data Loss (EAFDL), is subsequently obtained as the ratio of the expected amount of data lost to the expected time to data loss normalized to the amount of user data:

$$\text{EAFDL} = \frac{E(H)}{\text{MTTDL} \cdot U} , \tag{12}$$

with the MTTDL expressed in years. Let us also denote by Q the unconditional amount of data lost upon a first-device failure. Note that Q is unconditional on the event of a data loss occurring in that it is equal either to H if the system suffers a data loss prior to returning to normal operation or to zero otherwise, that is,

$$Q = \begin{cases} H , & \text{if DL} \\ 0 , & \text{if no DL} . \end{cases} \tag{13}$$

Therefore, the expected amount of data lost, $E(Q)$, upon a first-device failure is given by

$$E(Q) = P_{\text{DL}} \, E(H) . \tag{14}$$

From (11), (12) and (14), we obtain the EAFDL as follows:

$$\text{EAFDL} \approx \frac{n \, \lambda \, E(Q)}{U} , \tag{15}$$

with $1/\lambda$ expressed in years.

A. Reliability Analysis

From (11) and (15), it follows that the derivation of the MTTDL and EAFDL metrics requires the evaluation of the P_{DL} and $E(Q)$, respectively. These quantities are derived by considering the direct path approximation [12][13][27], which, under conditions (9) and (10), accurately assesses the reliability metrics of interest [14][23].

Next, we present the general outline of the methodology in more detail.

1) Direct Path to Data Loss: Consider the direct path of successive transitions from exposure level 1 to $\tilde{r}$. In [12][13][27], it was shown that P_{DL} can be approximated by the probability of the direct path to data loss, $P_{\text{DL,direct}}$, that is,

$$P_{\text{DL}} \approx P_{\text{DL,direct}} = \prod_{e=1}^{r-1} P_{e \to e+1}, \tag{16}$$

where $P_{e \to e+1}$ denotes the transition probability from exposure level e to $e+1$. The above approximation holds when storage devices are highly reliable, that is, it holds for arbitrary device failure and rebuild time distributions that satisfy conditions (9) and (10). In this case, the relative error tends to zero as λ/μ tends to zero [14].

As the direct path to data loss dominates the effect of all other possible paths to data loss considered together, it follows that the amount of data lost H can be approximated by that corresponding to the direct path:

$$H \approx H_{\text{direct}} . \tag{17}$$

Also, from (13) and (17) it follows that

$$Q \approx \begin{cases} H_{\text{direct}} , & \text{if DL follows the direct path} \\ 0 , & \text{otherwise} . \end{cases} \tag{18}$$

Consequently, to derive the amount of data lost, it suffices to proceed by considering the H and Q metrics corresponding to the direct path to data loss.

Note that the amount of data lost, H, is the amount of user data stored in the most-exposed codewords when entering exposure level $\tilde{r}$, which can no longer be recovered and therefore is irrecoverably lost. As the number of these codewords is equal to $C_{\tilde{r}}$ and each of these codewords contains l symbols of user data, it holds that

$$H = C_{\tilde{r}} \, l \, s \overset{(6)}{=} l \, A_{\tilde{r}} . \tag{19}$$

2) Amount of Data to Rebuild and Rebuild Times at Each Exposure Level: We now proceed to derive the conditional values of the random variables of interest given that the system goes through this direct path to data loss. Let R_e denote the rebuild times of the most-exposed codewords at each exposure level in this path, and let α_e be the fraction of the rebuild time R_e still left when another device fails causing the exposure level transition $e \to e+1$. In [28, Lemma 2], it was shown that, for highly reliable devices satisfying conditions (9) and (10), α_e is approximately uniformly distributed between zero and one, that is,

$$\alpha_e \sim U(0,1), \quad e = 1, \ldots, \tilde{r} - 1 . \tag{20}$$

Let $\vec{\alpha}$ denote the vector $(\alpha_1, \ldots, \alpha_{\tilde{r}-1})$, $\vec{\alpha}_e$ the vector $(\alpha_1, \ldots, \alpha_e)$, $\vec{C}_e$ the vector $(C_1, \cdots, C_e)$ and $\vec{A}_e$ the vector $(A_1, \cdots, A_e)$. Clearly, for the rebuild schemes considered, the fraction α_e of the rebuild time R_e still left also represents the fraction of the most-exposed codewords not yet recovered upon the next device failure. Therefore, the number of most-exposed codewords that are not yet recovered is equal to $\alpha_e \, C_e$. Clearly, the fraction V_e of these codewords that have symbols stored on the newly failed device depends on the codeword placement scheme. Consequently, the number of most-exposed codewords when entering exposure level $e+1$ is given by

$$C_{e+1} = V_e \, \alpha_e \, C_e , \quad e = 1, \ldots, \tilde{r} - 1 , \tag{21}$$

and by virtue of (6), the corresponding amount of data that is not yet rebuilt is given by

$$A_{e+1} = V_e \, \alpha_e \, A_e , \quad e = 1, \ldots, \tilde{r} - 1 , \tag{22}$$

with V_e depending only on the placement scheme.

Repeatedly applying (22) and using (7) yields

$$A_e = c \prod_{j=1}^{e-1} V_j \, \alpha_j . \tag{23}$$

Remark 1: From (23), it follows that the expected amount of data to be rebuilt at each exposure level does not depend on the duration of the rebuild times.

At exposure level 1, according to (7), the amount A_1 of data to be recovered is equal to c. Given that this data is recovered at a rate of b_1 and that the time required to write an amount c of data at a rate of b is equal to $1/\mu$, it follows that the rebuild time R_1 is given by

$$R_1 = \frac{b}{b_1} \cdot \frac{1}{\mu} . \tag{24}$$

 ISBN: 978-1-61208-550-0

As the rebuild times are proportional to the amount of data to be rebuilt and are inversely proportional to the rebuild rates, it holds that

$$\frac{R_{e+1}}{R_e} = \frac{A_{e+1}}{A_e} \cdot \frac{b_e}{b_{e+1}}, \quad e \geq 1 . \tag{25}$$

Using (22), (25) yields

$$R_{e+1} = V_e \, \alpha_e \, \frac{b_e}{b_{e+1}} \, R_e , \quad e = 1, \ldots, \tilde{r} - 2 , \tag{26}$$

or

$$R_e = G_{e-1} \, \alpha_{e-1} \, R_{e-1} , \quad e = 2, \ldots, \tilde{r} - 1 , \tag{27}$$

where

$$G_e \triangleq \frac{b_e}{b_{e+1}} \, V_e , \quad e = 1, \ldots, \tilde{r} - 2 . \tag{28}$$

Repeatedly applying (27) and using (28) yields

$$R_e = \frac{b_1}{b_e} \, R_1 \prod_{j=1}^{e-1} V_j \, \alpha_j , \quad e = 1, \ldots, \tilde{r} - 1 . \tag{29}$$

Let $\tilde{n}_e$ be the number of devices at exposure level e whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $e+1$. Subsequently, the transition probability $P_{e \to e+1}$ from exposure level e to $e+1$ depends on the duration of the corresponding rebuild time R_e and the aggregate failure rate of these $\tilde{n}_e$ highly reliable devices, and is given by [14]

$$P_{e \to e+1} \approx \tilde{n}_e \, \lambda \, R_e , \quad \text{for } e = 1, \ldots, \tilde{r} - 1 . \tag{30}$$

Substituting (29) into (30) yields

$$P_{e \to e+1}(\vec{\alpha}_{e-1}) \approx \tilde{n}_e \, \lambda \, \frac{b_1}{b_e} \, R_1 \prod_{j=1}^{e-1} V_j \, \alpha_j . \tag{31}$$

3) Estimation of P_{DL}: Consider the direct path $1 \to 2 \to \cdots \to \tilde{r}$ of successive transitions from exposure level 1 to $\tilde{r}$. For ease of reading, we denote the successive transitions from exposure level e to $\tilde{r}$ by $e \to \tilde{r}$. We first evaluate P_{DL}, the probability of data loss. From (16), and using the fact that α_e does not depend on $R_1, \alpha_1, \cdots, \alpha_{e-1}$, it follows that

$$\begin{aligned}
P_{\text{DL}} \;&\approx\; P_{1 \to \tilde{r}} = P_{1 \to 2} P_{2 \to \tilde{r}} = P_{1 \to 2} E_{\alpha_1}[P_{2 \to \tilde{r}}(\alpha_1)] \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\alpha_1) P_{3 \to \tilde{r}}(\alpha_1)] \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\alpha_1) E_{\alpha_2 | \alpha_1}[P_{3 \to \tilde{r}}(\alpha_1, \alpha_2)]] \\
&= \cdots \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\vec{\alpha}_1) E_{\alpha_2}[P_{3 \to 4}(\vec{\alpha}_2) \cdots \\
&\qquad \cdots E_{\alpha_{\tilde{r}-2}}[P_{\tilde{r}-1 \to \tilde{r}}(\vec{\alpha}_{\tilde{r}-2})] \cdots] \\
&= E_{\vec{\alpha}_{\tilde{r}-2}}[P_{1 \to 2} P_{2 \to 3}(\vec{\alpha}_1) \cdots P_{\tilde{r}-1 \to \tilde{r}}(\vec{\alpha}_{\tilde{r}-2})] \\
&= E_{\vec{\alpha}_{\tilde{r}-2}}\left[\prod_{e=1}^{\tilde{r}-1} P_{e \to e+1}(\vec{\alpha}_{e-1})\right] = E_{\vec{\alpha}_{\tilde{r}-2}}[P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2})] ,
\end{aligned} \tag{32}$$

where

$$P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \triangleq \prod_{e=1}^{\tilde{r}-1} P_{e \to e+1}(\vec{\alpha}_{e-1}) , \tag{33}$$

with

$$P_{1 \to 2}(\vec{\alpha}_0) \triangleq P_{1 \to 2} . \tag{34}$$

Substituting (31) into (33), and using (30) and (34), yields

$$P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \approx (\lambda \, b_1 \, R_1)^{\tilde{r}-1} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, (V_e \, \alpha_e)^{\tilde{r}-1-e} . \tag{35}$$

Unconditioning (35) on $\vec{\alpha}_{\tilde{r}-2}$, and given that the elements of $\vec{\alpha}_{\tilde{r}-2}$ are independent random variables approximately distributed according to (20) such that $E(\alpha_e^k) \approx 1/(k+1)$, (32) yields

$$P_{\text{DL}} \approx (\lambda \, b_1 \, R_1)^{\tilde{r}-1} \, \frac{1}{(\tilde{r}-1)!} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, V_e^{\tilde{r}-1-e} . \tag{36}$$

Using (8) and (24), (36) yields

$$P_{\text{DL}} \approx (\lambda \, c)^{\tilde{r}-1} \, \frac{1}{(\tilde{r}-1)!} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, V_e^{\tilde{r}-1-e} . \tag{37}$$

4) Estimation of $E(Q)$: We now proceed to evaluate $E(Q)$, the expected amount of data lost. Considering the *direct path* $1 \to 2 \to \cdots \to \tilde{r}$ of successive transitions from exposure level 1 to $\tilde{r}$, it follows from (18) and the fact that α_e does not depend on $R_1, \alpha_1, \cdots, \alpha_{e-1}$, that

$$\begin{aligned}
E(Q) \;&\approx\; P_{1 \to 2} E(Q | 1 \to 2) \\
&= P_{1 \to 2} E_{\alpha_1 | R_1}[E(Q | \alpha_1)] \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\alpha_1) E(Q | \alpha_1, 2 \to 3)] \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\alpha_1) E_{\alpha_2 | \alpha_1}[E(Q | \alpha_1, \alpha_2)]] \\
&= \cdots \\
&= P_{1 \to 2} E_{\alpha_1}[P_{2 \to 3}(\vec{\alpha}_1) E_{\alpha_2}[P_{3 \to 4}(\vec{\alpha}_2) \cdots \\
&\qquad \cdots P_{\tilde{r}-1 \to \tilde{r}}(\vec{\alpha}_{\tilde{r}-2}) E_{\alpha_{\tilde{r}-1}}(Q | \vec{\alpha}_{\tilde{r}-1})] \cdots] \\
&= E_{\vec{\alpha}_{\tilde{r}-1}}[P_{1 \to 2} P_{2 \to 3}(\vec{\alpha}_1) \cdots P_{\tilde{r}-1 \to \tilde{r}}(\vec{\alpha}_{\tilde{r}-2}) \\
&\qquad\qquad E(Q | \vec{\alpha}_{\tilde{r}-1})] \\
&\overset{(17)(18)}{=} E_{\vec{\alpha}_{\tilde{r}-1}}\left[\left(\prod_{e=1}^{\tilde{r}-1} P_{e \to e+1}(\vec{\alpha}_{e-1})\right) E(H | \vec{\alpha}_{\tilde{r}-1})\right] \\
&\overset{(33)}{=} E_{\vec{\alpha}_{\tilde{r}-1}}[P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \, E(H | \vec{\alpha}_{\tilde{r}-1})] \\
&\overset{(19)}{=} E_{\vec{\alpha}_{\tilde{r}-1}}[P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \, E(l \, A_{\tilde{r}} | \vec{\alpha}_{\tilde{r}-1})] \\
&= E_{\vec{\alpha}_{\tilde{r}-1}}[P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \, l \, E(A_{\tilde{r}} | \vec{\alpha}_{\tilde{r}-1})] \\
&= E_{\vec{\alpha}_{\tilde{r}-1}}[G(\vec{\alpha}_{\tilde{r}-1})] ,
\end{aligned} \tag{38}$$

where

$$G(\vec{\alpha}_{\tilde{r}-1}) \triangleq l \, P_{\text{DL}}(\vec{\alpha}_{\tilde{r}-2}) \, E(A_{\tilde{r}} | \vec{\alpha}_{\tilde{r}-1}) . \tag{39}$$

Using (23) and (35), (39) yields

$$G(\vec{\alpha}_{\tilde{r}-1}) \;\approx\; l \, c \, (\lambda \, b_1 \, R_1)^{\tilde{r}-1} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, (V_e \, \alpha_e)^{\tilde{r}-e} . \tag{40}$$

Unconditioning (40) on $\vec{\alpha}_{\tilde{r}-1}$, and given that the elements of $\vec{\alpha}_{\tilde{r}-1}$ are independent random variables approximately distributed according to (20) such that $E(\alpha_e^k) \approx 1/(k+1)$, (38) yields

$$E(Q) \approx l \, c \, (\lambda \, b_1 \, R_1)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, V_e^{\tilde{r}-e} . \tag{41}$$

Using (8) and (24), (41) yields

$$E(Q) \approx l \, c \, (\lambda \, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, V_e^{\tilde{r}-e} . \tag{42}$$

 ISBN: 978-1-61208-550-0

5) Evaluation of $E(H)$: The expected amount $E(H)$ of data lost conditioned on the fact that a data loss has occurred is obtained from (14) as the ratio of $E(Q)$ to $P_{\rm DL}$. Consequently, using (37) and (42), it follows that

$$E(H) = \frac{E(Q)}{P_{\rm DL}} \approx \left(\frac{l}{\tilde{r}} \prod_{e=1}^{\tilde{r}-1} V_e \right) c \, . \tag{43}$$

Remark 2: From (43), it follows that the expected amount of data lost conditioned on the fact that a data loss has occurred does not depend on the duration of the rebuild times.

6) Evaluation of MTTDL *and* EAFDL*:* Substituting (37) into (11) yields

$$\text{MTTDL} \approx \frac{1}{n\,\lambda} \, \frac{(\tilde{r}-1)!}{(\lambda\, c)^{\tilde{r}-1}} \prod_{e=1}^{\tilde{r}-1} \frac{b_e}{\tilde{n}_e} \, \frac{1}{V_e^{\tilde{r}-1-e}} \, . \tag{44}$$

Substituting (2) and (42) into (15) yields

$$\text{EAFDL} \approx m\,\lambda\,(\lambda\, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \prod_{e=1}^{\tilde{r}-1} \frac{\tilde{n}_e}{b_e} \, V_e^{\tilde{r}-e} \, . \tag{45}$$

B. Symmetric Scheme

Here, we consider the case where the redundancy spread factor k is in the interval $m < k \leq n$. As discussed in Section II-C2, at each exposure level e, the *prioritized* rebuild process recovers one of the e symbols that each of the C_e most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols. Thus, there are C_e symbols to be recovered in total, which corresponds to an amount A_e of data. For the symmetric placement discussed in Section II-B, these symbols are recovered by reading $(m - \tilde{r} + 1)\, C_e$ symbols, which corresponds to an amount $(m - \tilde{r} + 1)\, A_e$ of data, from the $k - e$ surviving devices in the affected group. Note that these are precisely the devices at exposure level e whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $e + 1$. Consequently, it holds that

$$\tilde{n}_e^{\rm sym} = k - e \, . \tag{46}$$

Furthermore, the recovered symbols are written to the spare space of these devices in such a way that no symbol is written to a device in which another symbol corresponding to the same codeword is already present. Owing to the symmetry of the symmetric placement, the same amount of data is being read from each of the $\tilde{n}_e$ devices. Similarly, the same amount of data is being written to each of the $\tilde{n}_e$ devices. Consequently, the total read/write rebuild bandwidth b of each device is split between the reads and the writes, with the read rate being equal to $(m - \tilde{r} + 1)\, b/(m - \tilde{r} + 2)$ and the write rate being equal to $b/(m - \tilde{r} + 2)$. Therefore, the total write bandwidth, which is also the rebuild rate b_e, is given by

$$b_e^{\rm sym} = \frac{\tilde{n}_e^{\rm sym}}{m - \tilde{r} + 2} \, b \, , \quad e = 1, \ldots, \tilde{r} - 1 \, . \tag{47}$$

Once all lost codeword symbols have been recovered, they are transferred to a new replacement device.

When the system enters exposure level e, the number of most-exposed codewords that need to be recovered is equal to C_e, $e = 1, \ldots, \tilde{r}$. Upon the next device failure, the expected number of most-exposed codewords that are not yet recovered is equal to $\alpha_e\, C_e$. Owing to the nature of the symmetric codeword placement, the newly failed device stores codeword symbols corresponding to only a fraction

$$V_e^{\rm sym} = \frac{m - e}{k - e} \, , \quad e = 1, \ldots, \tilde{r} - 1 \, . \tag{48}$$

of these most-exposed, not yet recovered codewords.

Substituting (46), (47) and (48) into (44), (45) and (43), and using (3) yields

$$\text{MTTDL}_k^{\rm sym} \approx \frac{1}{n\,\lambda} \left[\frac{b}{(l+1)\,\lambda\, c} \right]^{m-l} (m-l)! \prod_{e=1}^{m-l} \left(\frac{k-e}{m-e} \right)^{m-l-e} \, , \tag{49}$$

$$\text{EAFDL}_k^{\rm sym} \approx \lambda \left[\frac{(l+1)\,\lambda\, c}{b} \right]^{m-l} \frac{m}{(m-l+1)!} \prod_{e=1}^{m-l} \left(\frac{m-e}{k-e} \right)^{m-l+1-e} \, , \tag{50}$$

and

$$E(H)_k^{\rm sym} \approx \left(\frac{l}{m-l+1} \prod_{e=1}^{m-l} \frac{m-e}{k-e} \right) c \tag{51}$$

$$= \frac{l\,(m-1)!\,(k-m+l-1)!}{(m-l+1)\,(k-1)!\,(l-1)!} \, c \, . \tag{52}$$

Note that for a replication-based system, for which $m = r$ and $l = 1$, (49) and (50) are in agreement with (42.b) and (43.b) of [23], respectively.

Remark 3: From (49), (50), and (51), it follows that $\text{MTTDL}_k^{\rm sym}$ depends on n, but $\text{EAFDL}_k^{\rm sym}$ and $E(H)_k^{\rm sym}$ do not.

Remark 4: From (49), (50), and (51), it follows that, for $m - l = 1$, $\text{MTTDL}_k^{\rm sym}$ does not depend on k, whereas for $m - l > 1$, $\text{MTTDL}_k^{\rm sym}$ is increasing in k. Also, for $m - l \geq 1$, $\text{EAFDL}_k^{\rm sym}$ and $E(H)_k^{\rm sym}$ are decreasing in k. Consequently, within the class of symmetric placement schemes considered, that is, for $m < k \leq n$, the $\text{MTTDL}_k^{\rm sym}$ is maximized and the $\text{EAFDL}_k^{\rm sym}$ and the $E(H)_k^{\rm sym}$ are minimized when $k = n$.

C. Clustered Placement

As discussed in Section II-B1, in the clustered placement scheme, the n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to the *clustered* placement, each codeword is stored across the devices of a particular cluster. At each exposure level e, the rebuild process recovers one of the e symbols that each of the C_e most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols. Note that the remaining symbols are stored on the $m - e$ surviving devices in the affected group. As these are precisely the devices at exposure level e whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $e + 1$, it holds that

$$\tilde{n}_e^{\rm clus} = m - e \, . \tag{53}$$

 ISBN: 978-1-61208-550-0

The rebuild process in clustered placement recovers the lost symbols by reading $m-\tilde{r}+1$ symbols from $m-\tilde{r}+1$ of the $\tilde{n}_e$ surviving devices of the affected cluster. The lost symbols are computed on-the-fly and written to a spare device using the rebuild bandwidth at a rate of b. Consequently, it holds that

$$b_e^{\text{clus}} = b\,, \quad e = 1, \ldots, \tilde{r}-1\,. \tag{54}$$

When the system enters exposure level e, the number of most-exposed codewords that need to be recovered is equal to C_e, $e = 1, \ldots, \tilde{r}$. Upon the next device failure, the expected number of most-exposed codewords that have not yet been recovered is equal to $\alpha_e\, C_e$. Clearly, all these codewords have symbols stored on the newly failed device, which implies that

$$V_e^{\text{clus}} = 1\,, \quad e = 1, \ldots, \tilde{r}-1\,. \tag{55}$$

Substituting (53), (54) and (55) into (44), (45) and (43), and using (3) yields

$$\text{MTTDL}^{\text{clus}} \approx \frac{1}{n\,\lambda}\left(\frac{b}{\lambda\,c}\right)^{m-l} \frac{1}{\binom{m-1}{l-1}}\,, \tag{56}$$

$$\text{EAFDL}^{\text{clus}} \approx \lambda\left(\frac{\lambda\,c}{b}\right)^{m-l}\binom{m}{l-1}\,, \tag{57}$$

and

$$E(H)^{\text{clus}} = \frac{l}{m-l+1}\,c\,. \tag{58}$$

Note that for a replication-based system, for which $m = r$ and $l = 1$, (56), (57) and (58) are in agreement with (42.a), (43.a) and (39.a) of [23], respectively.

D. Declustered Placement

As discussed in Section II-B, the declustered placement scheme is a special cases of a symmetric placement scheme in which k is equal to n. Consequently, for $k = n$, (49), (50) and (51) yield

$$\text{MTTDL}^{\text{declus}} \approx \frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)! \prod_{e=1}^{m-l}\left(\frac{n-e}{m-e}\right)^{m-l-e}\,, \tag{59}$$

$$\text{EAFDL}^{\text{declus}} \approx \lambda\left[\frac{(l+1)\,\lambda\,c}{b}\right]^{m-l}\frac{m}{(m-l+1)!} \prod_{e=1}^{m-l}\left(\frac{m-e}{n-e}\right)^{m-l+1-e}\,, \tag{60}$$

and

$$E(H)^{\text{declus}} \approx \left(\frac{l}{m-l+1}\prod_{e=1}^{m-l}\frac{m-e}{n-e}\right)c \tag{61}$$

$$= \frac{l\,(m-1)!\,(n-m+l-1)!}{(m-l+1)\,(n-1)!\,(l-1)!}\,c\,. \tag{62}$$

Note that for a replication-based system, for which $m = r$ and $l = 1$, (59), (60) and (61) are in agreement with (36.b), (37.b) and (39.b) of [23], respectively.

IV. Optimal Placement

Here, we identify which of the placement schemes considered offers the best reliability in terms of the MTTDL, EAFDL and $E(H)$ metrics. From Remark 4, it follows that the placement that maximizes MTTDL and minimizes EAFDL and $E(H)$ is either the clustered $(k = m)$ or the declustered one $(k = n)$. We therefore proceed by comparing these two schemes when $m < n$, or, by also using (4), when

$$1 \le l < m \qquad \text{and} \qquad 1 \le m-l < m < n\,. \tag{63}$$

A. *Maximizing* MTTDL

From (56) and (59), it follows that

$$\frac{\text{MTTDL}^{\text{declus}}}{\text{MTTDL}^{\text{clus}}} \approx \left(\frac{1}{l+1}\right)^{m-l}(m-l)!\binom{m-1}{l-1} \prod_{e=1}^{m-l}\left(\frac{n-e}{m-e}\right)^{m-l-e}\,. \tag{64}$$

Remark 5: From (64), it follows that the placement that maximizes MTTDL does not depend on λ, b and c.

Depending on the values of m and l, we consider the following three cases:

1) $m-l=1$*:* For $m-l=1$, (64) yields

$$\frac{\text{MTTDL}^{\text{declus}}}{\text{MTTDL}^{\text{clus}}} \approx \frac{m-1}{m} < 1\,. \tag{65}$$

2) $m-l=2$*:* For $m-l=2$, (64) yields

$$\frac{\text{MTTDL}^{\text{declus}}}{\text{MTTDL}^{\text{clus}}} \approx \frac{(m-2)(n-1)}{(m-1)^2} \begin{cases} <1 & \text{for } n = m+1 \\ >1 & \text{for } n \ge m+2\,. \end{cases} \tag{66}$$

3) $m-l \ge 3$*:* For $m-l \ge 3$, (64) can be written as follows:

$$\frac{\text{MTTDL}^{\text{declus}}}{\text{MTTDL}^{\text{clus}}} \approx \frac{m-1}{l+1}\cdots\frac{l+1}{l+1}\,\frac{l}{l+1}\,\frac{n-m+l+1}{l+1} \left(\frac{n-m+l+2}{l+2}\right)^2 \prod_{e=1}^{m-l-3}\left(\frac{n-e}{m-e}\right)^{m-l-e}\,. \tag{67}$$

Using (63), (67) yields

$$\frac{\text{MTTDL}^{\text{declus}}}{\text{MTTDL}^{\text{clus}}} > \frac{l}{l+1}\,\frac{n-m+l+1}{l+1}\left(\frac{n-m+l+2}{l+2}\right)^2 \ge \frac{l}{l+1}\,\frac{l+2}{l+1}\left(\frac{l+3}{l+2}\right)^2 = \frac{l\,(l+3)^2}{(l+1)^2\,(l+2)} = \frac{2[l^2+2(l-1)+1]}{(l+1)^2\,(l+2)} + 1 > 1\,. \tag{68}$$

Remark 6: From the preceding, it follows that the MTTDL is maximized by the declustered placement scheme, except in the cases of $m-l=1$ and of $m-l=2$ with $n=m+1$, where it is maximized by the clustered placement scheme.

 ISBN: 978-1-61208-550-0

B. *Minimizing* EAFDL

From (57) and (60), it follows that

$$\frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \approx (l{+}1)^{m-l}\ \frac{(l-1)!}{(m-1)!}\ \prod_{e=1}^{m-l}\left(\frac{m-e}{n-e}\right)^{m-l+1-e}. \tag{69}$$

Remark 7: From (69), it follows that the placement that minimizes EAFDL does not depend on λ, b and c.

Depending on the values of m and l, we consider the following two cases:

1) $m-l=1$*:* For $m-l=1$, (69) yields

$$\frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \approx \frac{m}{n-1}\ \begin{cases} =1 & \text{for } n=m+1 \\ <1 & \text{for } n\geq m+2\,. \end{cases} \tag{70}$$

2) $m-l\geq 2$*:* For $m-l\geq 2$, (69) can be written as follows:

$$\frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \approx \frac{l+1}{m-1}\,\frac{l+1}{m-2}\,\cdots\frac{l+1}{l}\ \frac{l}{n-m+l}\ \prod_{e=1}^{m-l-1}\left(\frac{m-e}{n-e}\right)^{m-l+1-e}. \tag{71}$$

Using (63), (71) yields

$$\frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \ <\ \frac{l+1}{n-m+l} \leq \frac{l+1}{(m+1)-m+l} = 1\,. \tag{72}$$

Remark 8: From the preceding, it follows that the EAFDL is minimized by the declustered placement scheme. In particular, when $m-l=1$ and $n=m+1$, the clustered and declustered placement schemes yield the same EAFDL.

C. *Minimizing* $E(H)$

From (58) and (61), and using (63), it follows that

$$\frac{E(H)^{\text{declus}}}{E(H)^{\text{clus}}} \approx \prod_{e=1}^{m-l}\frac{m-e}{n-e}\ <1\,. \tag{73}$$

Remark 9: From (73), it follows that the declustered placement minimizes $E(H)$ for any λ, b, c.

D. *Synopsis*

We summarize our findings regarding the reliability offered by the data placement schemes as follows. Independently of the device reliability characteristics and mean expressed by $1/\lambda$, the reserved rebuild bandwidth b and the device capacity c, the declustered placement scheme minimizes the expected amount of data lost when loss occurs. Also, for $m-l=1$, the clustered placement scheme maximizes the MTTDL, but the declustered placement scheme minimizes the EAFDL. However, for $m-l\geq 2$, and for practical values of n and m, the declustered placement scheme maximizes the MTTDL and at the same time minimizes the EAFDL.

V. Reliability Comparison

Here, we assess the relative reliability of the declustered placement, which according to Remarks 6, 8 and 9 is the optimal one, under various codeword lengths m. We perform a fair comparison by considering systems with the same amount of user data, U, stored under the same storage efficiency, s_{eff}. From (2), it follows that the number of devices n is fixed. Also, from (1) it follows that

$$m-l\ =\ (1-s_{\text{eff}})\,m\ =\ h\,m\,, \tag{74}$$

where h is given by

$$h\ \triangleq\ 1-s_{\text{eff}} \tag{75}$$

and is fixed.

Using (74) to substitute l in (59) and (60) yields

$$\text{MTTDL}^{\text{declus}} \approx \frac{1}{n\,\lambda}\left[\frac{b}{[(1-h)m+1]\,\lambda\,c}\right]^{hm}(hm)!\ \prod_{e=1}^{hm}\left(\frac{n-e}{m-e}\right)^{hm-e}, \tag{76}$$

and

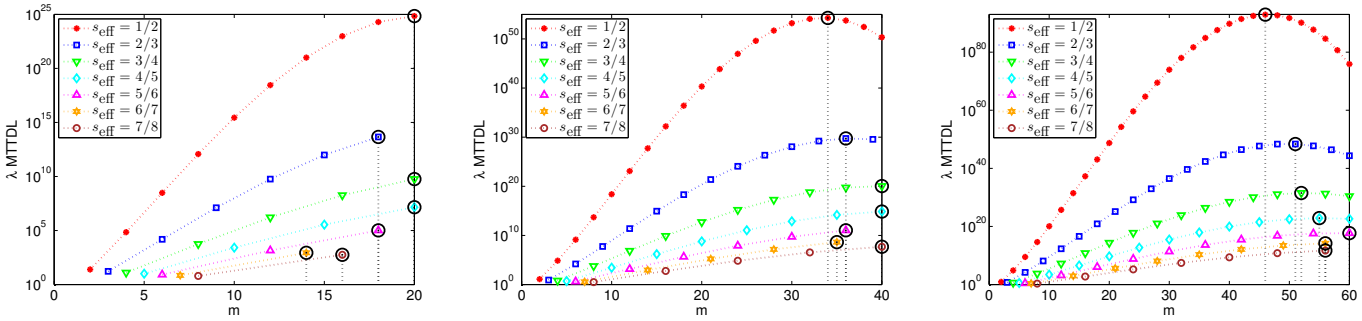
$$\text{EAFDL}^{\text{declus}} \approx \lambda\left[\frac{[(1-h)m+1]\,\lambda\,c}{b}\right]^{hm}\frac{m}{(hm+1)!}\ \prod_{e=1}^{hm}\left(\frac{m-e}{n-e}\right)^{hm+1-e}. \tag{77}$$

As discussed in Section III-A, the direct-path-approximation method yields accurate results when the storage devices are highly reliable, that is, when the ratio λ/μ of the mean rebuild time $1/\mu$ to the mean time to failure of a device $1/\lambda$ is very small. We proceed by considering systems for which it holds that $\lambda/\mu=\lambda\,c/b=0.001$. The combined effect of the number of devices and the system efficiency on the $\lambda\,\text{MTTDL}^{\text{declus}}$ measure is obtained by (76) and shown in Figure 1 as a function of the codeword length. The values for the storage efficiency are chosen to be fractions of the form $z/(z+1)$, $z=1,\ldots,7$, such that the first point of each of the corresponding curves is associated with the single-parity $(z,z+1)$-erasure code, and the second point of each of the corresponding curves is associated with the double-parity $(2z,2z+2)$-erasure code. We observe that the MTTDL increases as the storage efficiency s_{eff} decreases. This is because, for a given m, decreasing s_{eff} implies decreasing l, which in turn implies increasing the parity symbols $m-l$ and consequently improving MTTDL.

Let us now consider the single-parity codewords, which correspond to the first points of the curves. As s_{eff} increases, so do m and l, which results in a decreasing MTTDL for these codewords. This is due to the fact that as m increases, there are l data symbols, that is, more data symbols associated with each parity. This is in accordance with the results presented in Figure 2 of [26]. We observe that the same applies for the double-parity codewords, which correspond to the second points of the curves.

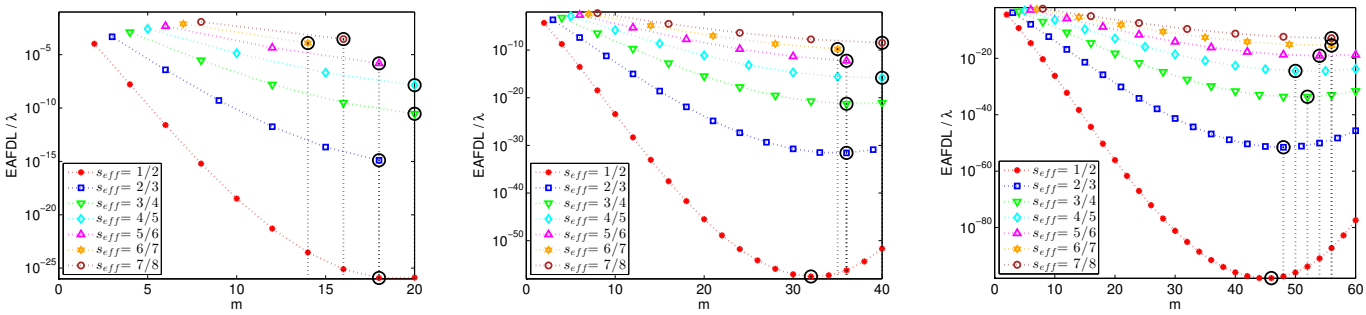
The combined effect of the number of devices and the system efficiency on the $\text{EAFDL}^{\text{declus}}/\lambda$ measure is obtained

 ISBN: 978-1-61208-550-0



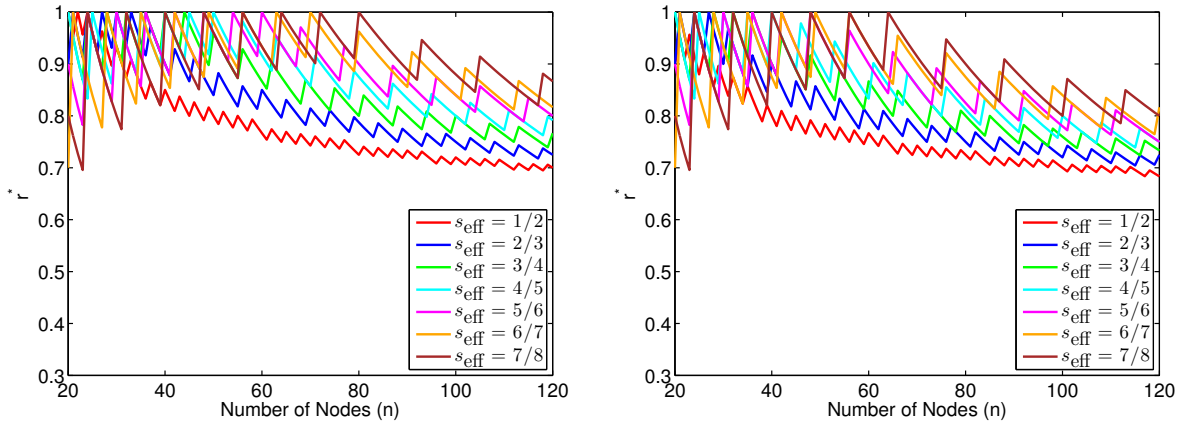
(a) $n = 20$ (b) $n = 40$ (c) $n = 60$

Figure 1. MTTDL$^{\text{declus}}$ vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 7/8$; $\lambda/\mu = 0.001$.



(a) $n = 20$ (b) $n = 40$ (c) $n = 60$

Figure 2. EAFDL$^{\text{declus}}$ vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 7/8$; $\lambda/\mu = 0.001$.



(a) MTTDL (b) EAFDL

Figure 3. r^* vs. number of devices for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 7/8$; $\lambda/\mu = 0.001$.

by (77) and shown in Figure 2 as a function of the codeword length. We observe that the EAFDL increases as the storage efficiency s_{eff} increases. Also, as s_{eff} increases, the EAFDL for the single-parity codewords, which correspond to the first points of the curves, also increases. We observe that the same applies for the double-parity codewords, which correspond to the second points of the curves.

We now proceed to identify the optimal codeword length, m^*, that maximizes the MTTDL or minimizes the EAFDL for a given storage efficiency. The optimal codeword length is dictated by two opposing effects on reliability. On the one hand, large values of m imply that codewords can tolerate more device failures, but on the other hand result in a higher exposure degree to failure as each of the codewords is spread across a large number of devices. In Figures 1 and 2, the optimal values, m^*, are indicated by the circles, and the corresponding codeword lengths are indicated by the vertical dotted lines. We observe that for small values of n, it holds that $m^* = n$, whereas for large values of n it holds that $m^* < n$. By comparing Figures 1 and 2, we deduce that in general the optimal codeword lengths for MTTDL and EAFDL are similar, but not identical. Let us define by r^* the ratio of m^* to n, that is

$$r^* \triangleq \frac{m^*}{n} \,. \tag{78}$$

The r^* values for various values of the system efficiency and for the two metrics of interest are shown in Figure 3. From Figures 3(a) and (b) we deduce that the optimal codeword lengths for MTTDL and EAFDL are similar, and for some values of n even identical. It can be proved that as n grows to infinity, the r^* values for MTTDL and EAFDL approach a common value that depends on s_{eff} and is roughly equal to 0.6.

VI. Conclusions

We considered the Mean Time to Data Loss (MTTDL) and the Expected Annual Fraction of Data Loss (EAFDL) reliability metrics of storage systems using advanced erasure codes. A methodology was presented for deriving the two metrics analytically. Closed-form expressions capturing the effect of various system parameters were obtained for the symmetric, clustered and declustered data placement schemes. We established that the declustered placement scheme offers superior reliability in terms of both metrics. Subsequently, a thorough comparison of the reliability achieved by the declustered placement scheme under various codeword configurations was conducted. The results obtained show that the optimal codeword lengths for MTTDL and EAFDL are similar and, as the system size grows, approach a common value that depends only on the storage efficiency.

Extending the methodology developed to derive the reliability of erasure coded systems under arbitrary rebuild time distributions and in the presence of unrecoverable latent errors is a subject of further investigation.

References

[1] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 1988, pp. 109–116.

[2] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson, "RAID: High-performance, reliable secondary storage," ACM Comput. Surv., vol. 26, no. 2, Jun. 1994, pp. 145–185.

[3] M. Malhotra and K. S. Trivedi, "Reliability analysis of redundant arrays of inexpensive disks," J. Parallel Distrib. Comput., vol. 17, Jan. 1993, pp. 146–151.

[4] W. A. Burkhard and J. Menon, "Disk array storage system reliability," in Proceedings of the 23rd International Symposium on Fault-Tolerant Computing, Jun. 1993, pp. 432–441.

[5] K. S. Trivedi, Probabilistic and Statistics with Reliability, Queueing and Computer Science Applications, 2nd ed. New York: Wiley, 2002.

[6] Q. Xin, E. L. Miller, T. J. E. Schwarz, D. D. E. Long, S. A. Brandt, and W. Litwin, "Reliability mechanisms for very large storage systems," in Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST), Apr. 2003, pp. 146–156.

[7] T. J. E. Schwarz, Q. Xin, E. L. Miller, D. D. E. Long, A. Hospodor, and S. Ng, "Disk scrubbing in large archival storage systems," in Proceedings of the 12th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Oct. 2004, pp. 409–418.

[8] S. Ramabhadran and J. Pasquale, "Analysis of long-running replicated systems," in Proc. 25th IEEE International Conference on Computer Communications (INFOCOM), Apr. 2006, pp. 1–9.

[9] B. Eckart, X. Chen, X. He, and S. L. Scott, "Failure prediction models for proactive fault tolerance within storage systems," in Proceedings of the 16th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2008, pp. 1–8.

[10] K. Rao, J. L. Hafner, and R. A. Golding, "Reliability for networked storage nodes," IEEE Trans. Dependable Secure Comput., vol. 8, no. 3, May 2011, pp. 404–418.

[11] J.-F. Pâris, T. J. E. Schwarz, A. Amer, and D. D. E. Long, "Highly reliable two-dimensional RAID arrays for archival storage," in Proceedings of the 31st IEEE International Performance Computing and Communications Conference (IPCCC), Dec. 2012, pp. 324–331.

[12] I. Iliadis and V. Venkatesan, "An efficient method for reliability evaluation of data storage systems," in Proceedings of the 8th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2015, pp. 6–12.

[13] ——, "Most probable paths to data loss: An efficient method for reliability evaluation of data storage systems," International Journal on Advances in Systems and Measurements, vol. 8, no. 3&4, Dec. 2015, pp. 178–200.

[14] V. Venkatesan and I. Iliadis, "A general reliability model for data storage systems," in Proceedings of the 9th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2012, pp. 209–219.

[15] A. Dholakia, E. Eleftheriou, X.-Y. Hu, I. Iliadis, J. Menon, and K. Rao, "A new intra-disk redundancy scheme for high-reliability RAID storage systems in the presence of unrecoverable errors," ACM Trans. Storage, vol. 4, no. 1, May 2008, pp. 1–42.

[16] A. Thomasian and M. Blaum, "Higher reliability redundant disk arrays: Organization, operation, and coding," ACM Trans. Storage, vol. 5, no. 3, Nov. 2009, pp. 1–59.

[17] K. M. Greenan, J. S. Plank, and J. J. Wylie, "Mean time to meaningless: MTTDL, Markov models, and storage system reliability," in Proceedings of the USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage), Jun. 2010, pp. 1–5.

[18] I. Iliadis, R. Haas, X.-Y. Hu, and E. Eleftheriou, "Disk scrubbing versus intradisk redundancy for RAID storage systems," ACM Trans. Storage, vol. 7, no. 2, Jul. 2011, pp. 1–42.

[19] "Amazon Simple Storage Service." [Online]. Available: http://aws.amazon.com/s3/ [retrieved: March 2017]

[20] D. Borthakur et al., "Apache Hadoop goes realtime at Facebook," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 2011, pp. 1071–1080.

[21] R. J. Chansler, "Data availability and durability with the Hadoop Distributed File System," ;login: The USENIX Association Newsletter, vol. 37, no. 1, 2013, pp. 16–22.

[22] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies (MSST), May 2010, pp. 1–10.

[23] I. Iliadis and V. Venkatesan, "Expected annual fraction of data loss as a metric for data storage reliability," in Proceedings of the 22nd Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2014, pp. 375–384.

[24] C. Huang et al., "Erasure coding in Windows Azure Storage," in Proceedings of the USENIX Annual Technical Conference (ATC), Jun. 2012, pp. 15–26.

[25] "IBM Cloud Object Storage." [Online]. Available: www.ibm.com/cloud-computing/products/storage/object-storage/how-it-works/ [retrieved: March 2017]

[26] V. Venkatesan and I. Iliadis, "Effect of codeword placement on the reliability of erasure coded data storage systems," in Proceedings of the 10th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2013, pp. 241–257.

[27] V. Venkatesan, I. Iliadis, C. Fragouli, and R. Urbanke, "Reliability of clustered vs. declustered replica placement in data storage systems," in Proceedings of the 19th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Jul. 2011, pp. 307–317.

[28] V. Venkatesan and I. Iliadis, "Effect of codeword placement on the reliability of erasure coded data storage systems," IBM Research Report, RZ 3827, Aug. 2012.

 ISBN: 978-1-61208-550-0

Modelling of Cascading Effect in a System with Dependent Components via Bivariate Distribution

Hyunju Lee[1] and Ji Hwan Cha[2]
Department of Statistics
Ewha Womans University
Seoul, Rep. of KOREA
e-mail[1]: hyunjlee@ewhain.net
e-mail[2]: jhcha@ewha.ac.kr

***Abstract*— A cascading failure is a failure in a system of interconnected parts, in which the breakdown of one element can lead to the subsequent collapse of the others. Cascading effect is quite common in power grids and can also frequently occur in computer networks (such as the Internet). In this paper, we consider systems composed of components having interdependence and cascading effect. For this, based on the notion of conditional failure rate, a new bivariate distribution for modelling the lifetimes of dependent components is constructed. A comparison of systems having interdependence and cascading effect with those having independent components is also performed.**

Keywords-modelling of cascading effect; bivariate distribution; dependence; load-sharing components; parallel system

I. INTRODUCTION

Dependent random quantities can frequently be encountered in practice and they have been modelled by using bivariate distributions. In the literature, various specific parametric models for bivariate distributions have been suggested and studied [1]-[4]. A good review on the modelling of multivariate survival models can be found in [5]. An excellent encyclopedic survey of various bivariate distributions can be found in [6].

In this paper, we propose a new general class of dependent distributions, which is different from the previous models in [1]-[6]. The structure of this paper is as follows. In Section 2, taking into account the physics of failure of items and the interrelationship between them, we propose and discuss a general methodology for constructing a general class of bivariate distributions. In Section 3, based on the developed class, we study the lifetimes of systems having interdependence and cascading effect. Finally, in Section 4, concluding remarks are given.

II. MODELLING BIVARIATE DISTRIBUTION

Suppose that the system is composed of two components: component 1 and component 2 and they start to operate at time $t=0$. The original lifetimes of components 1 and 2, when they start to operate, are described by the corresponding failure rates $\lambda_1(t)$ and $\lambda_2(t)$, respectively. These original lifetimes of components 1 and 2 are denoted by X_1^* and X_2^*, respectively, assuming that X_1^* and X_2^* are stochastically independent.

Similar to [1]'s model, we consider the practical situation when the failure of one component increases the stress of the other component, which results in the shortened residual lifetime of the remaining component. Under this type of dependency, we denote the corresponding eventual lifetimes of components 1 and 2 by X_1 and X_2, respectively.

Define $\Psi_1(t)=1$ ($\Psi_2(t)=1$) if component 1 (component 2) is functioning at time t, whereas $\Psi_1(t)=0$ ($\Psi_2(t)=0$) if component 1 (component 2) is at failed state at time t. For notational convenience, let $\tilde{i}=2$ when $i=1$; whereas $\tilde{i}=1$ when $i=2$. Then, we assume the conditional failure rate of component i is given by:

$$\begin{aligned} & r_i(t \mid \Psi_{\tilde{i}}(s)=1, 0\le s\le t) \\ &\equiv \lim_{\Delta t\to 0}\frac{1}{\Delta t}P(t<X_i\le t+\Delta t \mid \Psi_{\tilde{i}}(s)=1, 0\le s\le t, X_i>t) \\ &= \lambda_i(t),\ \ i=1,2, \quad (1) \end{aligned}$$

and

$$\begin{aligned} & r_i(t \mid \Psi_{\tilde{i}}(s)=1, 0\le s<u; \Psi_{\tilde{i}}(s)=0, u\le s\le t) \\ &\equiv \lim_{\Delta t\to 0}\frac{1}{\Delta t}P(t<X_i\le t+\Delta t \mid \\ & \Psi_{\tilde{i}}(s)=1, 0\le s<u; \Psi_{\tilde{i}}(s)=0, u\le s\le t, X_i>t) \\ &= \alpha_i(t-u)\lambda_i(t),\ t\ge u,\ i=1,2, \quad (2) \end{aligned}$$

where $\alpha_i(w)\ge 1$, for all $w\ge 0$, $i=1,2$.

In the following discussions, the notations $S(x_1,x_2)$, $f(x_1,x_2)$ will be used to denote the joint survival function and the joint probability density function (pdf) of X_1 and X_2, respectively. We will now suggest the joint distribution of X_1 and X_2 under the assumed model. The proof can be found in [7].

Result 1. Under the conditional failure rate model stated in (1)-(2), the joint survival function $S(x_1,x_2)$, for $0<x_1<x_2$, is given by:

$$S(x_1,x_2)$$

$$= \int_{x_1}^{x_2} \lambda_1(u) \exp\left(- \int_0^{x_2-u} \alpha_2(w)\lambda_2(u+w)dw \right) \times \exp\left(- \int_0^{u} \lambda_1(w) + \lambda_2(w)\, dw \right) du + \exp\left(- \int_0^{x_2} \lambda_1(w) + \lambda_2(w)\, dw \right), \text{ for } 0 < x_1 < x_2, \quad (3)$$

and $S(x_1, x_2)$, for $0 < x_2 \le x_1$, can be obtained symmetrically by replacing x_1, x_2, $\lambda_1(\cdot)$, $\lambda_2(\cdot)$, $\alpha_1(\cdot)$, $\alpha_2(\cdot)$ in the right-hand side of (3) with respective opposite components. The corresponding joint pdf, for $0 < x_1 < x_2$, is given by:

$$f(x_1, x_2) = \lambda_1(x_1)\lambda_2(x_2)\alpha_2(x_2 - x_1) \times \exp\left(- \int_0^{x_1} \lambda_1(w) + \lambda_2(w)\, dw \right) \exp\left(- \int_0^{x_2-x_1} \alpha_2(w)\lambda_2(x_1+w)\, dw \right), \text{for } 0 < x_1 < x_2,$$

and $f(x_1, x_2)$, for $0 < x_2 \le x_1$, can also be obtained symmetrically.

III. System Reliability

In this section, we study the lifetimes of systems when the baseline distributions are Weibull (see [7]). Let $\lambda_i(t) = \mu_i \gamma_i (\mu_i t)^{\gamma_i - 1}$, $t \ge 0$, $i = 1,2$, and $\alpha_i(t) = \alpha_i t + 1$, $\alpha_i > 0$, $i = 1,2$. In this case, from Result 1, the joint survival function of $S(x_1, x_2)$ and the joint pdf $f(x_1, x_2)$ can be obtained.

Suppose that the system is a series system. Then, the lifetime of the system is $T_S = \min\{X_1, X_2\} = \min\{X_1^*, X_2^*\}$. Thus, in this case, the lifetime of a system having interdependence components is the same as that of a system having independent components. Thus, this case is not relevant to the dependence structure of the proposed model.

We now assume that the system is a parallel system. In this case, the lifetime of a system having interdependence components is $T_S = \max\{X_1, X_2\}$. In order to obtain the distribution of T_S, define $T_1 = \min\{X_1, X_2\}$ and $T_2 = \max\{X_1, X_2\}$. Then, the joint pdf of (T_1, T_2), $g(t_1, t_2)$, is given by:

$$g(t_1, t_2) = f(t_1, t_2) + f(t_2, t_1), \; t_1 \le t_2,$$

and thus, the pdf of T_S is given by:

$$f_{T_S}(t) = \int_0^t g(t_1, t) dt_1, \; t \ge 0.$$

Cleary, the pdf of lifetime of a system having independent components can be obtained by setting $\alpha_i = 0$, $i = 1,2$. The graphs of the survival functions of the system having interdependence components when $\mu_i = 0.5$, $\gamma_i = 3$, $\alpha_i = 2$, $i = 1,2$, and the system having independent components are given in Figure 1.

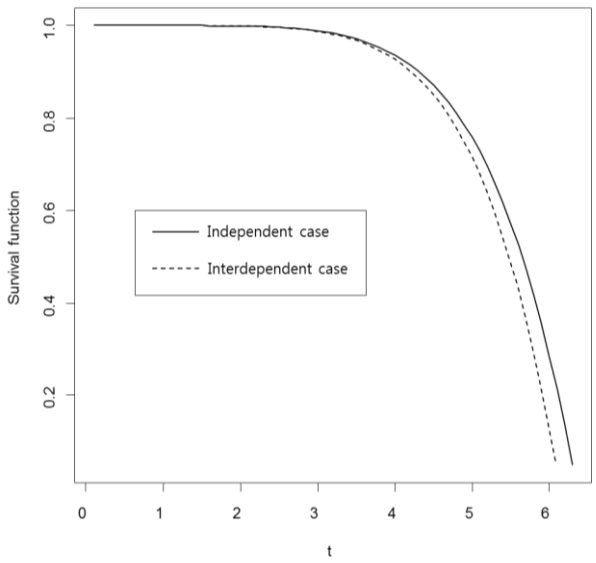


Figure 1. The survival functions.

As can be seen from Figure 1, the survival function for the interdependent case is smaller than that for the independent case. For brevity, we have just considered two simple cases (series and parallel systems). More complex cases could be considered in similar ways.

IV. Concluding Remarks

There have been studies on real systems with several interdependent components and with propagated and cascading effects (see, e.g., Mo et al. [8] and Bobbio et al. [9]). In this paper, a general methodology for constructing new general classes of bivariate distributions has been suggested. Based on the proposed class, the lifetimes of systems having interdependence and cascading effect have been studied. Based on the developed class, numerous bivariate distributions can further be generated and new issues on the estimation and testing of the model parameters should be discussed in the future studies. In this paper, our discussions are mainly focused on generating bivariate models. However, a similar approach could be applied to generate a new class of multivariate distributions. More detailed discussion on this issue is also given in Lee and Cha [7].

Acknowledgment

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0093827). This work was also supported by the National Research Foundation of Korea

(NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B2014211).

REFERENCES

[1] J. E. Freund, "A bivariate extension of the exponential distribution," Journal of the American Statistical Association, Vol. 56, pp. 971-977, 1961.

[2] A. W. Marshall and I. Olkin, "A multivariate exponential distribution," Journal of the American Statistical Association , Vol. 62, pp. 30-44, 1967.

[3] H. W. Block and A. P. Basu, "A continuous bivariate exponential extension," Journal of the American Statistical Association , Vol. 69, pp. 1031-1037, 1974.

[4] M. Shaked, "Extension of the Freund distribution with applications in reliability theory," Operations Research, Vol. 32, pp. 917-925, 1984.

[5] P. Hougaard, "Modelling multivariate survival," Scandinavian Journal of Statistics, Vol. 14, pp. 292-304, 1987.

[6] N. Balakrishnan and C. Lai, "Continuous bivariate distributions (second edition)," New York: Springer, 2009.

[7] H. Lee and J. H. Cha, "On construction of general classes of bivariate distributions," Journal of Multivariate Analysis, Vol. 127, pp. 151-159, 2014.

[8] Y. Mo, L. Xing, F. Zhong and Z. Zhang, "Reliability evaluation of network systems with dependent propagated failures using decision diagrams," IEEE transactions on Dependable and Secure Computing, Vol. 13, pp. 672-683, 2016.

[9] A. Bobbio, D. C. Raiteri, L. Portinale and S. Montani, "A dynamic Bayesian network based framework to evaluate cascading effects in a power grid," Engineering Applications of Artificial Intelligence. Vol. 25, pp. 683-697, 2012.

 ISBN: 978-1-61208-550-0