CTRQ 2018

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

ISBN: 978-1-61208-629-3

April 22 - 26, 2018

Athens, Greece

CTRQ 2018 Editors

Ilias Iliadis, IBM Zurich Research Laboratory, Switzerland
Kiran Makhijani, Huawei Technologies, USA

CTRQ 2018

Forward

The Eleventh International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ 2018), held between April 22, 2018 and April 26, 2018 in Athens, Greece, continued a series of special 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.

We take here the opportunity to warmly thank all the members of the CTRQ 2018 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 who dedicated their time and effort to contribute to CTRQ 2018. 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 2018 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that CTRQ 2018 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of communication theory, reliability and quality of service. We also hope that Athens, Greece, provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

CTRQ 2018 Chairs

CTRQ Steering Committee

Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Pål Ellingsen, Bergen University College, Norway
Wojciech Kmiecik, Wroclaw University of Technology, Poland

CTRQ 2018
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 2018 Technical Program Committee
Bassant Abdelhamid, Ain Shams University, Cairo, Egypt
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
Safdar Hussain Bouk, DGIST, Daegu, Korea
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
Sudharman K. Jayaweera, University of New Mexico Albuquerque, USA
Alexey Kashevnik, SPIIRAS, Russia
Sokratis K. Katsikas, Norwegian University of Science & Technology (NTNU), Norway
Carlos Kavka, ESTECO SpA, Italy

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 of Erasure Coded Systems under Rebuild Bandwidth Constraints

Ilias Iliadis
IBM Research – Zurich
8803 Rüschlikon, Switzerland
Email: ili@zurich.ibm.com

***Abstract*—Modern storage systems employ erasure coding redundancy and recovering schemes to ensure high data reliability at high storage efficiency. The widely used replication scheme belongs to this broad class of erasure coding schemes. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss** (MTTDL) **and the Expected Annual Fraction of Data Loss** (EAFDL) **metrics. To improve the reliability of data storage systems, certain data placement and rebuild schemes reduce the rebuild times by recovering data in parallel from the storage devices. It is often assumed though that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. In large-scale data storage systems, however, the network bandwidth is constrained. This article obtains the** MTTDL **and** EAFDL **of erasure coded systems analytically for the symmetric, clustered, and declustered data placement schemes under network rebuild bandwidth constraints. The resulting reliability degradation is assessed and the results obtained establish that the declustered placement scheme offers superior reliability in terms of both metrics. Efficient codeword configurations that achieve high reliability in the presence of network rebuild bandwidth constraints are identified.**

Keywords–Storage; Reliability; Data placement; MTTDL; EAFDL; RAID; MDS codes; Information Dispersal Algorithm; Prioritized rebuild; Repair bandwidth; Network bandwidth constraint.

I. INTRODUCTION

In today's large-scale data storage systems, data redundancy is introduced to ensure that data lost owing to device and component failures can be recovered. Appropriate redundancy schemes are deployed to prevent permanent loss of data and, consequently, enhance the reliability of storage systems. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss (MTTDL) [1-20] and, more recently, the Fraction of Data Loss Per Year (FDLPY) [21] and the equivalent Expected Annual Fraction of Data Loss (EAFDL) reliability metrics [22-24]. Analytical reliability expressions for the MTTDL were obtained predominately using Markovian models, which assume that component failure and rebuild times are independent and exponentially distributed. In practice though, these distributions are not exponential. To cope with this issue, system reliability was assessed in [16][18][23][24] using an alternative methodology that does not involve any Markovian analysis and considers the practical case of non-exponential failure and rebuild time distributions. Moreover, the misconception reported in [25] that MTTDL derivations based on Markovian models provide unrealistic results was dispelled in [26] by invoking improved MTTDL derivations that yield satisfactory results, and also by drawing on prior work that analytically obtains MTTDL without involving any Markovian analysis.

Earlier works have predominately considered the MTTDL metric, whereas recent works have also considered the EAFDL metric [22][23][24]. The introduction of the latter metric was motivated by the fact that Amazon S3 considers the durability of data over a given year [27], and, similarly, Facebook [28], LinkedIn [29] and Yahoo! [30] consider the amount of data lost in given periods.

To protect data from being lost and improve the reliability of data storage systems, replication-based storage systems spread replicas corresponding to data stored on each storage device across several other storage devices. To improve the low storage efficiency associated with the replication schemes, erasure coding schemes that provide a high data reliability as well as a high storage efficiency are deployed. Special cases of such codes are the Redundant Arrays of Inexpensive Disks (RAID) schemes, such as RAID-5 and RAID-6, that have been extensively deployed in the past thirty years [1][2].

State-of-the-art data storage systems [31-34] employ more general erasure codes that affect the reliability, performance, and the storage and reconstruction overhead of the system. In this article, we focus on the reliability assessment of erasure coded systems in terms of the MTTDL and EAFDL metrics. These metrics were analytically derived in [23] for the symmetric, clustered, and declustered data placement schemes under the assumption that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. For instance, in the case of a declustered placement, redundant data associated with the data stored on a given device is placed across all remaining devices in the system. In this way, the rebuild process can be parallelized, which in turn results in short rebuild times. The restoration time can be minimized provided there is sufficient network rebuild bandwidth available. In large-scale data storage systems though, the network bandwidth is constrained.

The effect of network rebuild bandwidth constraints on the reliability of replication-based storage systems was studied in [8][15]. It was found that spreading replicas over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process, led to system reliability being significantly reduced. The reliability of erasure coded systems in the absence of bandwidth constraints was assessed in [23]. The MTTDL and EAFDL metrics were obtained analytically for the symmetric, clustered, and declustered data placement schemes based on a general framework and methodology. In this article, we recognize that this methodology also holds in the case of network rebuild bandwidth constraints and apply it to derive enhanced closed-form reliability expressions for the MTTDL

 ISBN: 978-1-61208-629-3

and EAFDL metrics for these placement schemes in the presence of such rebuild bandwidth constraints. Subsequently, we provide insight into the effect of the placement schemes and the impact of the available network rebuild bandwidth on system reliability. The validity of this methodology for accurately assessing the reliability of storage systems was confirmed by means of simulation in several contexts [14-16, 18, 22]. It was demonstrated that the theoretical predictions for the reliability of systems comprised of highly reliable storage devices match well with the simulation results obtained. Consequently, the emphasis of the present work is on the theoretical assessment of the effect of network rebuild bandwidth constraints on the reliability of erasure coded systems. Also, this work extends the reliability results obtained in [15] for the special case of replication-based storage systems to the more general case of erasure coded systems.

The remainder of the article is organized as follows. Section II describes the storage system model and the corresponding parameters considered. Section III presents the adaptation of a general framework and methodology for deriving the MTTDL and EAFDL metrics analytically for the case of erasure coded systems under network rebuild bandwidth constraints. Closed-form expressions for the symmetric, clustered, and declustered placement schemes are derived. Section IV presents numerical results demonstrating the effectiveness of the erasure coding redundancy schemes for improving the system reliability. It also assesses the sensitivity to the network rebuild bandwidth constraints under various codeword configurations. Section V provides a discussion on the applicability of the results obtained. Finally, we conclude in Section VI.

II. Storage System Model

Modern data storage systems use erasure coded schemes 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. We proceed by briefly reviewing the basic concepts of erasure coding and data recovery procedures of such storage systems. To assess their reliability, we consider the model used in [23], and adopt and extend the notation. More precisely, the storage system considered comprises n storage devices (nodes or disks), with each device storing an amount c of data, such that the total storage capacity of the system is $n\,c$.

A. Redundancy

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. We consider (m, l) 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, having the property 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

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

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 $(m > l)$
(m, l)	MDS-code structure
k	spread factor of the data placement scheme, or group size (number of devices in a group)
b	reserved rebuild bandwidth per device
$B_{\max}$	maximum network rebuild bandwidth
$F_\lambda(.)$	cumulative distribution function of device lifetimes
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 $2 \leq \tilde{r} \leq m)$
N_b	maximum number of devices from which rebuild can occur at full speed in parallel $(N_b = B_{\max}/b)$
B_{eff}	effective network rebuild bandwidth
$1/\mu$	time to read (or write) an amount c of data at a rate b from (or to) a device $(1/\mu = c/b)$
$1/\lambda$	mean time to failure of a storage device $(1/\lambda = \int_0^\infty [1 - F_\lambda(t)]dt)$

Consequently, 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 $\tilde{r}$ 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 $(N, N - 1)$ 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 $(N, N - 2)$ 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 (m, l) MDS erasure codes.

B. Symmetric Codeword Placement

According to a symmetric codeword placement, each codeword is stored on m distinct devices with one symbol per device. In a large storage system, the number of devices, n, is usually much larger than the codeword length, m. Therefore, there are 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, the *redundancy spread factor* k denotes the number of devices over which the codewords stored on that device are spread [18]. The system effectively comprises n/k

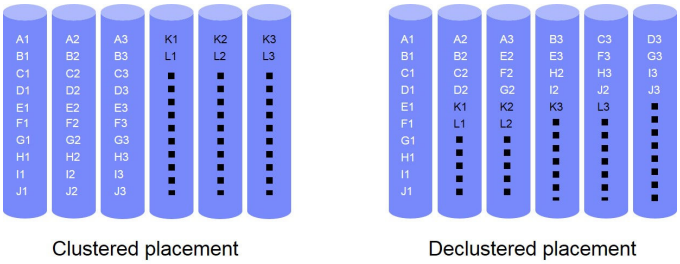


Figure 1. Clustered and declustered placement of codewords of length $m = 3$ on $n = 6$ devices. X1, X2, X3 represent a codeword (X = A, B, C, . . . , L).

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. Each codeword is placed entirely in one of the n/k groups. Within each group, all $\binom{k}{m}$ possible ways of placing m symbols across k devices are equally used to store all the codewords in that group.

In such a symmetric placement scheme, within each of the n/k groups, the $m-1$ codeword symbols corresponding to the data on each device are *equally* spread across the remaining $k-1$ devices, the $m-2$ codeword symbols corresponding to the codewords shared by any two devices are equally spread across the remaining $k-2$ devices, and so on. Note also that the n/k groups are logical and therefore need not be physically located in the same node/rack/datacenter.

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. This results in n/m groups for clustered and one group for declustered placement schemes.

1) Clustered Placement: 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, as shown in Figure 1. 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, as shown in Figure 1.

The clustered and declustered placement schemes represent the two extremes in which the symbols of the codewords associated with the data stored on a failing device are spread across the remaining devices and hence the extremes of the degree of parallelism that can be exploited when rebuilding this data. For declustered placement, the symbols are spread equally across *all* remaining devices, whereas for clustered placement, the symbols are spread across the smallest possible number of 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.

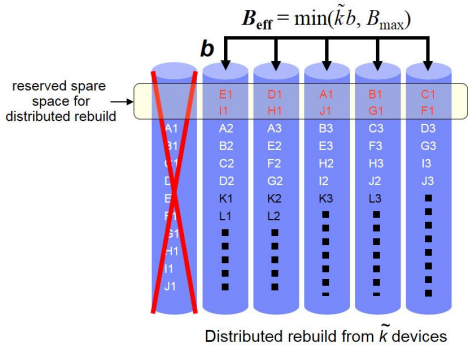


Figure 2. Rebuild under declustered placement.

When a declustered placement scheme is used, as shown in Figure 2, spare space is reserved on each device for temporarily storing the reconstructed codeword symbols before they are transferred to a new replacement device. The rebuild process used to restore the data lost by failed devices is assumed to be both *prioritized* and *distributed*. As discussed in [23], a prioritized (or intelligent) rebuild process always attempts to first rebuild the *most-exposed* codewords, namely, the codewords that have lost the largest number of symbols. The prioritized rebuild process recovers one of the symbols that each of the most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols. In a distributed rebuild process, the codeword symbols lost by failed devices are reconstructed by reading surviving symbols from a number, say $\tilde{k}$, of surviving devices, and storing the recovered symbols in the reserved spare space of the $\tilde{k}$ surviving devices, as shown in Figure 2.

A certain proportion of the device bandwidth is reserved for data recovery during the rebuild process, with b denoting the actual reserved rebuild bandwidth per device. This bandwidth is usually only a fraction of the total bandwidth available at each device, with the remaining bandwidth being used to serve user requests. Thus, the lost symbols are rebuilt in parallel using the rebuild bandwidth b available on each surviving device. During this process, it is desirable to reconstruct the lost codeword symbols on devices in which another symbol of the same codeword is not already present. Assuming that the system is at exposure level u (as described in Section II-D below) b_u $(\leq b)$ denotes the rate at which the amount of data that needs to be rebuilt (repair traffic) is written to selected device(s). In particular, $1/\mu$ denotes the time required to read (or write) an amount c of data from (or to) a device, given by

$$\frac{1}{\mu} = \frac{c}{b} \,. \tag{5}$$

In a distributed rebuild process involving $\tilde{k}$ devices, the total network bandwidth required to perform rebuild at full speed is $\tilde{k}\,b$. Let $B_{\max}$ $(\geq b)$ denote the maximum available network bandwidth for rebuilds. Then, the effective network rebuild bandwidth used by rebuilds, $B_{\text{eff}}(\tilde{k})$, cannot exceed $B_{\max}$ and is therefore given by

$$B_{\text{eff}}(\tilde{k}) = \min(\tilde{k}\,b, B_{\max}) = \min(\tilde{k}, N_b)\,b \,, \tag{6}$$

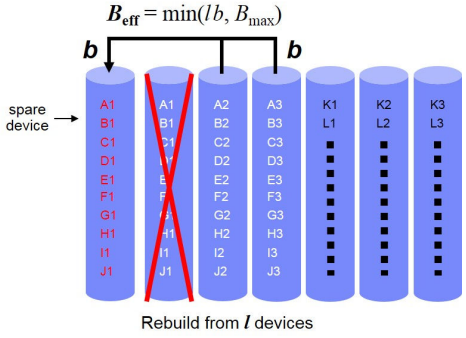


Figure 3. Rebuild under clustered placement.

where N_b specifies the effective maximum number of devices from which rebuild can occur in parallel at full speed, and is given by

$$N_b \triangleq \frac{B_{\max}}{b} . \tag{7}$$

Note that N_b may not be an integer; it only represents the *effective* maximum number of devices from which distributed rebuild can occur at full speed. Substituting $b = c\mu$ into (6), we get

$$B_{\text{eff}} = \min(\tilde{k}, N_b)\, c\mu . \tag{8}$$

A similar reconstruction process is used for other symmetric placement schemes within each group of k devices, except for the clustered placement. When clustered placement is used, the codeword symbols are spread across all $k = m$ devices in each group (cluster). Therefore, reconstructing the lost symbols on the surviving devices of a group will result in more than one symbol of the same codeword on the same device. To avoid this, the lost symbols are reconstructed directly in spare devices as shown in Figure 3. In these reconstruction processes, decoding and re-encoding of data are assumed to be done on the fly and so the time taken for reconstruction is equal to the time taken to read and write the required data to the devices. Note also that alternative erasure coding schemes have been proposed to reduce the amount of data transferred over the storage network during reconstruction (see [35][36] and references therein).

D. Exposure Levels and Amount of Data to Rebuild

At time t, $D_j(t)$ denotes the number of codewords that have lost j symbols, with $0 \le j \le \tilde{r}$. The system is at exposure level u $(0 \le u \le \tilde{r})$, where

$$u = \max_{D_j(t)>0} j. \tag{9}$$

The system is at exposure level u if there are codewords with $m - u$ symbols left, but there are no codewords with fewer than $m - u$ symbols left in the system, that is, $D_u(t) > 0$, and $D_j(t) = 0$, for all $j > u$. These codewords are referred to as the *most-exposed* codewords. 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. Let t_u denote the time of the first transition from exposure level $u - 1$ to exposure level u, and t_u^+ the instant immediately after t_u. Then, the number, C_u, of most exposed codewords when entering exposure level u, $u = 1, \ldots, \tilde{r}$, is given by $C_u = D_u(t_u^+)$.

Analytic expressions for the reliability metrics of interest were derived in [23], using the direct path approximation, which considers only transitions from lower to higher exposure levels [14][16][18]. This implies that each exposure level is entered only once.

E. Failure and Rebuild Time Distributions

We adopt the model and notation considered in [24]. The lifetimes of the n devices are assumed to be independent and identically distributed, with a cumulative distribution function $F_\lambda(.)$ and a mean of $1/\lambda$. Real-world distributions, such as Weibull and gamma, as well as exponential distributions that belong to the large class defined in [16] are considered. The storage devices are characterized to be *highly reliable* in that the ratio of the mean 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,

$$\frac{\lambda}{\mu} = \frac{\lambda\, c}{b} \ll 1 . \tag{10}$$

We consider storage devices whose the cumulative distribution function F_λ satisfies the condition

$$\mu \int_0^{1/\mu} F_\lambda(t)\, dt \ll 1, \quad \text{with } \frac{\lambda}{\mu} \ll 1 , \tag{11}$$

such that the MTTDL and EAFDL reliability metrics of erasure coded storage systems tend to be insensitive to the device failure distribution, that is, they depend only on its mean $1/\lambda$, but not on its density $F_\lambda(.)$[23].

III. Derivation of MTTDL and EAFDL

The MTTDL metric assesses the expected amount of time until some data can no longer be recovered and therefore is irrecoverably lost, whereas the EAFDL metric assesses the fraction of stored data that is expected to be lost by the system annually. The MTTDL$(B_{\max})$ and EAFDL$(B_{\max})$ metrics are derived as a function of $B_{\max}$ based on the framework and methodology presented in [23]. More specifically, this methodology uses the direct path approximation and does not involve any Markovian analysis. It holds for general failure time distributions, which can be exponential or non-exponential, such as the Weibull and gamma distributions that satisfy condition (11). Note that this framework is general in that it also applies in the case where the network rebuild bandwidth is constrained. The only parameters that are affected by the network rebuild bandwidth constraint are the rebuild rates and, accordingly, those parameters that depend on them, such as the rebuild times. Analytic expressions for the two metrics of interest were derived in [23, Equations (44) and (45)] as follows:

$$\text{MTTDL}(B_{\max}) \approx \frac{1}{n\,\lambda}\, \frac{(\tilde{r}-1)!}{(\lambda\, c)^{\tilde{r}-1}} \prod_{u=1}^{\tilde{r}-1} \frac{b_u(B_{\max})}{\tilde{n}_u}\, \frac{1}{V_u^{\tilde{r}-1-u}} , \tag{12}$$

and

$$\mathrm{EAFDL}(B_{\max}) \approx m\,\lambda\,(\lambda\,c)^{\tilde{r}-1}\;\frac{1}{\tilde{r}\,!}\;\prod_{u=1}^{\tilde{r}-1}\frac{\tilde{n}_u}{b_u(B_{\max})}\;V_u^{\tilde{r}-u}\;, \tag{13}$$

where $\tilde{n}_u$ represents the number of devices at exposure level u whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $u+1$. Also, V_u represents the fraction of the most-exposed codewords at exposure level u that have symbols stored on a newly failed device that causes the exposure level transition $u \rightarrow u+1$. Note that this fraction depends only on the codeword placement scheme. As mentioned in the preceding, b_u, the rate at which the amount of data that needs to be rebuilt at exposure level u is written to selected device(s), depends on $B_{\max}$, the maximum network rebuild bandwidth.

Remark 1: From [23, Equation (43)], it follows that the expected amount $E(H)$ of data lost, given that a data loss has occurred, does not depend on b_u and therefore is not affected by the maximum network rebuild bandwidth. Consequently, this reliability metric is not considered in this article.

Remark 2: The analytic expressions for the MTTDL and EAFDL reliability metrics were derived in [23] in the absence of network rebuild bandwidth constraints. Consequently, they correspond to the case of $B_{\max} = \infty$, with the two metrics being denoted by MTTDL(∞) and EAFDL(∞), respectively.

From (12) and (13), it follows that

$$\frac{\mathrm{MTTDL}(B_{\max})}{\mathrm{MTTDL}(\infty)} = \frac{\mathrm{EAFDL}(\infty)}{\mathrm{EAFDL}(B_{\max})} = \theta\;, \tag{14}$$

where θ represents the *reliability reduction factor* that assesses the reliability degradation due to a network rebuild bandwidth constraint, and is given by

$$\theta \;\triangleq\; \prod_{u=1}^{\tilde{r}-1}\frac{b_u(B_{\max})}{b_u(\infty)}\;. \tag{15}$$

Remark 3: From (15), and given that $b_u(B_{\max})$ decreases as $B_{\max}$ decreases, it follows that θ decreases as $\tilde{r}$ increases and $B_{\max}$ decreases.

A. Symmetric Placement

We consider the case where the redundancy spread factor k is in the interval $m < k \leq n$. As discussed in [23, Section III-B], at each exposure level u, the *prioritized* rebuild process recovers one of the u symbols that each of the most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols from the $\tilde{n}_u$ surviving devices in the affected group. According to [23, Equation (46)], it holds that

$$\tilde{n}_u^{\mathrm{sym}} = k - u\;. \tag{16}$$

Furthermore, in the absence of a network rebuild bandwidth constraint, the total write bandwidth, which is also the rebuild rate b_u, is given by [23, Equation (47)]

$$b_u^{\mathrm{sym}}(\infty) = \frac{\tilde{n}_u^{\mathrm{sym}}}{m-\tilde{r}+2}\;b \overset{(16)}{=} \frac{(k-u)\,b}{m-\tilde{r}+2}\,,\quad u = 1,\ldots,\tilde{r}-1\;. \tag{17}$$

In the presence though of a network rebuild bandwidth constraint, $B_{\max}$, and according to (6), with $\tilde{k} = \tilde{n}_u = \tilde{n}_u^{\mathrm{sym}}$, the rebuild rate b_u is given as a function of $B_{\max}$ by

$$b_u^{\mathrm{sym}}(B_{\max}) = \frac{B_{\mathrm{eff}}(\tilde{n}_u)}{m-\tilde{r}+2} = \frac{\min(\tilde{n}_u\,b, B_{\max})}{m-\tilde{r}+2} = \frac{\min(\tilde{n}_u, N_b)\,b}{m-\tilde{r}+2} \overset{(16)}{=} \frac{\min(k-u, N_b)\,b}{m-\tilde{r}+2}\,,\;\text{ for }\; u = 1,\ldots,\tilde{r}-1\;. \tag{18}$$

Substituting (17) and (18) into (15) yields

$$\theta^{\,\mathrm{sym}} = \prod_{u=1}^{\tilde{r}-1}\frac{\min(k-u, N_b)}{k-u}\;. \tag{19}$$

Note that when $N_b \geq k-1$, the system reliability is not affected because all rebuilds are performed at full speed, and therefore the θ factor is equal to one. However, when $N_b < k-1$, it may not be possible for some of the rebuilds to be performed at full speed, and therefore the factor θ will be less than one, which affects the system reliability. Consequently, the reliability reduction factor, θ, depends on the *bandwidth constraint factor*, ϕ, given by

$$\phi \;\triangleq\; \min\left(\frac{N_b}{k}, 1\right) \overset{(7)}{=} \min\left(\frac{B_{\max}}{k\,b}, 1\right)\;,\;\text{ with } 0 \leq \phi \leq 1\;. \tag{20}$$

From (19) and (20), and recognizing that $\min(k-u, N_b) = \min(\min(k-u, k), N_b) = \min(k-u, \min(k, N_b)) = \min(k\min(1, N_b/k), k-u) = \min(k\,\phi, k-u)$, it follows that

$$\theta^{\,\mathrm{sym}} = \prod_{u=1}^{\tilde{r}-1}\min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right)\;. \tag{21}$$

Using (3) and (21), and the fact that MTTDL(∞) and EAFDL(∞) are given by [23, Equations (49) and (50)], respectively, (14) yields

$$\mathrm{MTTDL}_k^{\mathrm{sym}}(B_{\max}) \approx \frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)!\prod_{u=1}^{m-l}\left(\frac{k-u}{m-u}\right)^{m-l-u}\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right)\;, \tag{22}$$

and

$$\mathrm{EAFDL}_k^{\mathrm{sym}}(B_{\max}) \approx \lambda\left[\frac{(l+1)\,\lambda\,c}{b}\right]^{m-l}\frac{m}{(m-l+1)!}\prod_{u=1}^{m-l}\left(\frac{m-u}{k-u}\right)^{m-l+1-u}\Bigg/\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right)\;, \tag{23}$$

where $B_{\max}$ is expressed via ϕ given by (20).

Note that for a replication-based system, for which $m = r$ and $l = 1$, and by virtue of (19) and (21), (22) is in agreement with Equation (24) of [15], with $c/b = 1/\mu$.

Remark 4: From (22) and (23), it follows that MTTDL$_k^{\mathrm{sym}}$ depends on n, but EAFDL$_k^{\mathrm{sym}}$ does not.

Remark 5: From (22) and (23), and for any value of ϕ, it can be proved that for $m-l \geq 2$, MTTDL$_k^{\mathrm{sym}}$ is increasing in k. It can also be proved that for any $m-l \geq 1$, EAFDL$_k^{\mathrm{sym}}$ is not increasing in k. Consequently, within the class of symmetric placement schemes considered, that is, for $l+1 < m < k \leq n$, the MTTDL$_k^{\mathrm{sym}}$ is maximized and the EAFDL$_k^{\mathrm{sym}}$ is minimized by the declustered placement scheme, that is, when $k = n$.

 ISBN: 978-1-61208-629-3

B. Clustered Placement

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 u, the rebuild process recovers one of the u symbols that each of the C_u most-exposed codewords has lost by reading $m-\tilde{r}+1$ of the remaining symbols. Note that the remaining symbols are stored on the $m-u$ surviving devices in the affected group. According to [23, Equation (53)], it holds that

$$\tilde{n}_u^{\text{clus}} = m - u \; . \tag{24}$$

In the case of clustered placement, the rebuild process recovers the lost symbols by reading l symbols from l of the $\tilde{n}_u$ surviving devices of the affected cluster. In the absence of a network rebuild bandwidth constraint, the symbols are read at a rate of b from each of the l devices, such that the effective network rebuild bandwidth is equal to $B_{\text{eff}} = l\,b$. Subsequently, the lost symbols are computed on-the-fly and written to a spare device at a rate of $B_{\text{eff}}/l = b$. Consequently, it holds that

$$b_u^{\text{clus}}(\infty) = b\,, \quad u = 1, \ldots, \tilde{r}-1 \; . \tag{25}$$

In the presence though of a network rebuild bandwidth constraint, B_{max}, the effective network rebuild bandwidth is equal to $B_{\text{eff}} = \min(l\,b, B_{\text{max}})$, which implies that the lost symbols are written to a spare device at a rate of B_{eff}/l. Thus, the rebuild rate b_u is given as a function of B_{max} by

$$b_u^{\text{clus}}(B_{\text{max}}) = \frac{B_{\text{eff}}(B_{\text{max}})}{l} = \frac{\min(l\,b, B_{\text{max}})}{l} = \frac{\min(l, N_b)\,b}{l}\,, \quad \text{for } u = 1, \ldots, \tilde{r}-1 \; . \tag{26}$$

Substituting (25) and (26) into (15) yields

$$\theta^{\,\text{clus}} = \left(\frac{\min(l, N_b)}{l}\right)^{\tilde{r}-1} \; . \tag{27}$$

As $l < m$, it holds that $\min(l, N_b) = \min(\min(l, m), N_b) = \min(\min(N_b, m), l) = \min(m \min(N_b/m, 1), l) = \min(m\phi, l)$, where, analogously to (20), and with $k = m$,

$$\phi \triangleq \min\left(\frac{N_b}{m}, 1\right) \overset{(7)}{=} \min\left(\frac{B_{\text{max}}}{m\,b}, 1\right)\,, \quad \text{where } 0 \le \phi \le 1 \; . \tag{28}$$

Consequently, (27) yields

$$\theta^{\,\text{clus}} = \min\left(\frac{m}{l}\,\phi\,, 1\right)^{\tilde{r}-1} \; . \tag{29}$$

Remark 6: From (29), it follows that for $m\,\phi/l \ge 1$ or, equivalently, for $\phi \ge s_{\text{eff}} = l/m$, $\theta^{\,\text{clus}}$ is equal to one, which implies that the bandwidth constraint does not affect the system reliability.

Using (3) and (29), and the fact that MTTDL(∞) and EAFDL(∞) are given by [23, Equations (56) and (57)], respectively, (14) yields

$$\text{MTTDL}^{\text{clus}}(B_{\text{max}}) \approx \frac{1}{n\,\lambda}\left(\frac{\min(m\,\phi,\, l)\,b}{l\,\lambda\,c}\right)^{m-l} \frac{1}{\binom{m-1}{l-1}}\,, \tag{30}$$

$$\text{EAFDL}^{\text{clus}}(B_{\text{max}}) \approx \lambda\left(\frac{l\,\lambda\,c}{\min(m\,\phi,\, l)\,b}\right)^{m-l} \binom{m}{l-1}\,, \tag{31}$$

where B_{max} is expressed via ϕ given by (28).

Remark 7: Note that as far as the data placement is concerned, the clustered placement scheme is a special case of a symmetric placement scheme for which k is equal to m. However, its reliability assessment cannot be directly obtained from the reliability results derived in Section III-A for the symmetric placement scheme by simply setting $k = m$. The reason for that is the difference in the rebuild processes. In the case of a symmetric placement scheme, recovered symbols are written to the spare space of existing devices, whereas in the case of a clustered placement scheme, recovered symbols are written to a spare device. This results in different rebuild bandwidths, which are given by (17) and (25), respectively.

C. Declustered Placement

The declustered placement scheme is a special case of a symmetric placement scheme in which k is equal to n. Consequently, for $k = n$, (22) and (23) yield

$$\text{MTTDL}^{\text{declus}}(B_{\text{max}}) \approx \frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l} (m-l)! \prod_{u=1}^{m-l}\left(\frac{n-u}{m-u}\right)^{m-l-u} \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right)\,, \tag{32}$$

and

$$\text{EAFDL}^{\text{declus}}(B_{\text{max}}) \approx \lambda\left[\frac{(l+1)\,\lambda\,c}{b}\right]^{m-l} \frac{m}{(m-l+1)!} \prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Bigg/ \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right)\,, \tag{33}$$

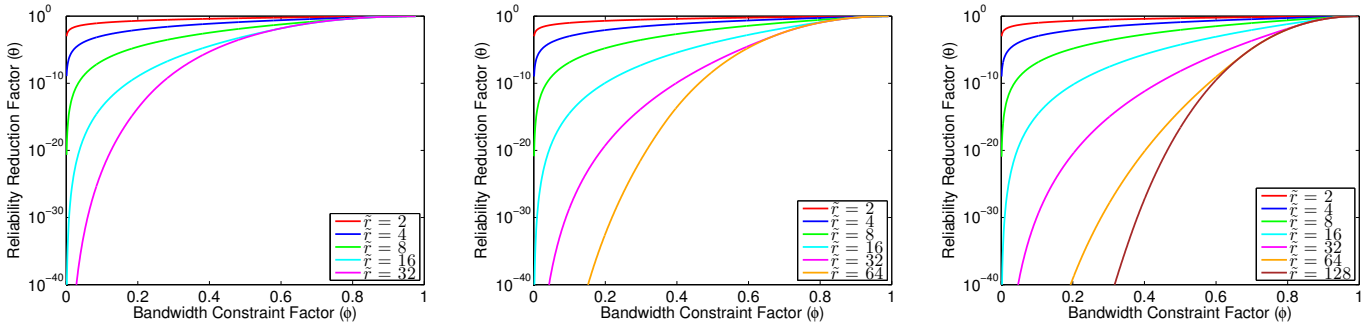
where B_{max} is expressed via ϕ given by (20) with $k = n$.

IV. Numerical Results

First, we assess the reduction in reliability owing to bandwidth constraints. The reliability reduction factor, θ, is obtained by (21) and (29) for the symmetric and clustered placements, respectively, and shown in Figures 4 and 5 as a function of the bandwidth constraint factor. For a symmetric placement scheme, Figure 4 demonstrates that as the group size k increases, the reliability reduction factor θ decreases and the magnitude of the reduction is more pronounced for larger values of $\tilde{r}$. Clearly, if codewords are spread over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process, the system reliability is affected and a drastic reliability degradation occurs as the system size increases. In contrast, according to Remark 6, the reliability of a clustered placement scheme remains unaffected for $\phi \ge l/m = (m - \tilde{r} + 1)/m$. This is due to the fact that the effective rebuild bandwidth is significantly smaller because the rebuilds are not distributed, but performed directly on a spare device. However, as Figure 5 demonstrates, for $\phi < l/m$, the reliability reduction factor drops sharply, especially for large values of $\tilde{r}$.

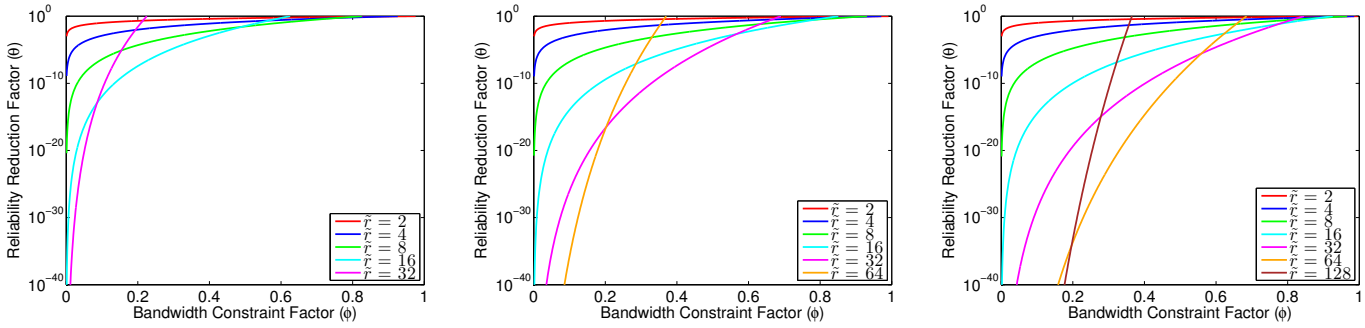
Next, we consider a storage system of a given size and assess its reliability for various codeword configurations, storage efficiencies, and network rebuild bandwidth constraints.

 ISBN: 978-1-61208-629-3



(a) $k = 40$ (b) $k = 100$ (c) $k = 200$

Figure 4. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; symmetric placement.



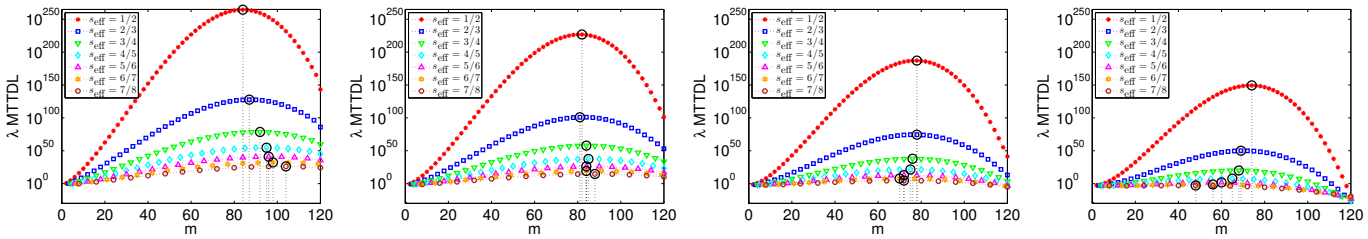
(a) $m = 40$ (b) $m = 100$ (c) $m = 200$

Figure 5. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; clustered placement.

In particular, we consider a system containing 120 devices under a declustered placement scheme ($k = n = 120$), which according to Remark 5 is the optimal one within the class of symmetric schemes. The amount of user data stored, U, is determined by the storage efficiency, s_{eff}, via (2). As discussed in Section II-E, the analytical reliability results obtained are accurate 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 $\lambda/\mu = \lambda\, c/b = 0.001$.

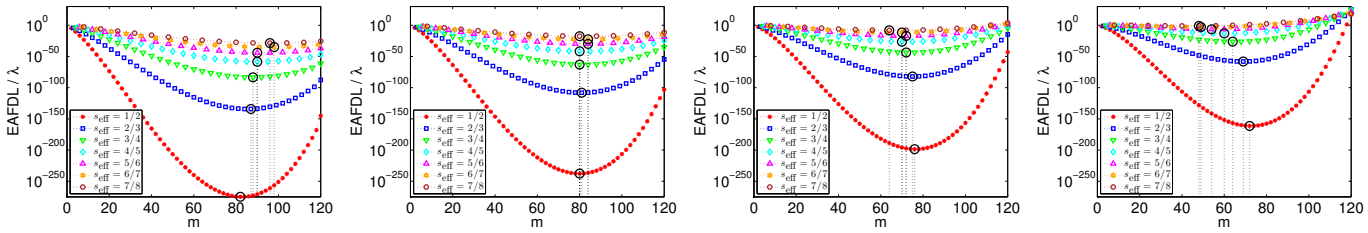
The combined effect of the network rebuild bandwidth constraint and the system efficiency on the normalized λ MTTDL measure is obtained by (32) and shown in Figure 6 as a function of the codeword length. In particular, when the codeword length is equal to the system size ($m = k = n$), the placement becomes clustered and the normalized λ MTTDL measure is obtained by (30). Three cases for the network rebuild bandwidth constraint were considered: $\phi = 1$ corresponds to the case where there is no network rebuild bandwidth constraint given that $N_b \geq k = 120$ or, equivalently, $B_{\max} \geq k\, b = 120\, b$; $\phi = 0.1$ and $\phi = 0.01$ correspond to the cases where $N_b = 0.1\, k = 12$ and $N_b = 0.01k = 1.2$ or, equivalently, $B_{\max} = 0.1\, k\, b = 12\, b$ and $B_{\max} = 0.01\, k\, b = 1.2\, b$, respectively. 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.

For all values of ϕ considered, 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 the MTTDL. Furthermore, for a given storage efficiency, s_{eff}, the MTTDL decreases by orders of magnitude as the maximum network rebuild bandwidth decreases. We now proceed to identify the optimal codeword length, m^*, that maximizes the MTTDL for a given bandwidth constraint and storage efficiency. The optimal codeword length is dictated by two opposing effects on reliability. On the one hand, larger values of m imply that codewords can tolerate more device failures, but on the other hand, they result in a higher exposure degree to failure as each of the codewords is spread across a larger number of devices. In Figure 6, the optimal values, m^*, are indicated by the circles, and the corresponding codeword lengths are indicated by the vertical dotted lines. By comparing Figures 6(a), (b), and (c), we deduce that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the maximum MTTDL value of $4{\times}10^{78}$ is obtained when $m = m^* = 92$. However, in the case of $\phi = 0.1$, the maximum MTTDL value of $6{\times}10^{57}$ is obtained for $m^* = 84$. The reason for the reduction of the optimal codeword length is due to the fact that for a given value of s_{eff} and as m increases, so does $\tilde{r}$, which, according to Remark 3, results in a smaller reliability reduction factor. Thus, the reliability reduction factor corresponding to $m = 92$ is smaller than the one corresponding to $m = 84$, which in turn causes the MTTDL for $m = 92$ to no longer be optimal as it becomes smaller than the one for $m = 84$. Note that for $m = 84$ and $s_{\text{eff}} = 3/4$, from (1) and (3), it follows that $l = 63$ and $\tilde{r} - 1 = 21$. From (21), and given



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 6. Normalized MTTDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 7. Normalized EAFDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$.

that $u \leq \tilde{r} - 1 = 21 \ll k = 120$, such that $\phi/(1 - u/k) \approx \phi$, it now follows that $\theta \approx \phi^{\tilde{r}-1} = 0.1^{21} = 10^{-21}$, which implies that the reliability is reduced by 21 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the maximum MTTDL values of $6{\times}10^{37}$ and $8{\times}10^{19}$ are obtained for $m^* = 76$ and $m^* = 68$, respectively.

The combined effect of the network rebuild bandwidth constraint and the system efficiency on the normalized $\text{EAFDL}^{\text{declus}}/\lambda$ measure is obtained by (31) and (33), and shown in Figure 7 as a function of the codeword length. We observe that the EAFDL increases as the storage efficiency s_{eff} decreases. Furthermore, for a given storage efficiency, s_{eff}, the EAFDL increases by orders of magnitude as the maximum network rebuild bandwidth decreases. Similarly to the case of MTTDL, by comparing Figures 7(a), (b), and (c), we observe that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the minimum EAFDL value of $4{\times}10^{-84}$ is obtained when $m = m^* = 88$. However, in the case of $\phi = 0.1$, the minimum EAFDL value of $9{\times}10^{-64}$ is obtained for $m^* = 80$, which implies that the reliability is reduced by 20 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the minimum EAFDL values of $2{\times}10^{-44}$ and $6{\times}10^{-27}$ are obtained for $m^* = 72$ and $m^* = 64$, respectively. By comparing Figures 6 and 7, we deduce that in general the optimal codeword lengths m^*_{MTTDL} (for MTTDL) and m^*_{EAFDL} (for EAFDL) are similar.

Reducing $B_{\max}$ or, equivalently, ϕ, affects the optimal codeword length as follows.

Proposition 1: For any storage efficiency s_{eff}, and for both reliability metrics, the optimal codeword length m^* decreases as ϕ decreases.

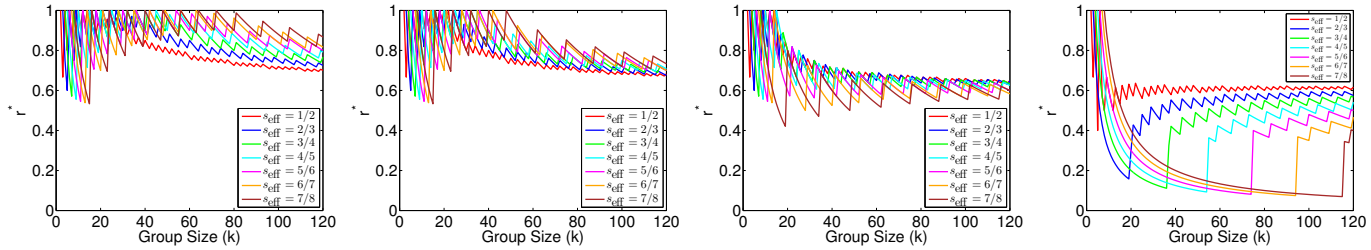
Proof: Consider two bandwidth constraint factors ϕ_1 and ϕ_2 with $\phi_1 > \phi_2$. Let m_1^* and m_2^* be the corresponding optimal codeword lengths for the MTTDL metric. We shall now show that $m_1^* \geq m_2^*$.

As m_1^* is the optimal codeword length for ϕ_1, it holds that $\text{MTTDL}(\phi_1, m) \leq \text{MTTDL}(\phi_1, m_1^*)$ for all $m \geq m_1^*$. Also, from (1) and (3), it holds that $\tilde{r} = (1 - s_{\text{eff}})\, m + 1$, which implies that as m increases, so does $\tilde{r}$. From (15), it follows that $\theta^{(2)}/\theta^{(1)} = \prod_{u=1}^{\tilde{r}-1} \frac{b_u(\phi_2)}{b_u(\phi_1)}$, which, owing to the fact that $b_u(\phi_2) \leq b_u(\phi_1)$ $\forall u$, decreases as $\tilde{r}$ or, equivalently, m increases. Consequently, $\theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)}$ for all $m \geq m_1^*$. Also, from (14), it follows that $\text{MTTDL}(\phi_2{,}m)/\text{MTTDL}(\phi_1{,}m) = \theta_m^{(2)}/\theta_m^{(1)}$ for all values of m. From the preceding, it follows that $\text{MTTDL}(\phi_2{,}m)/\text{MTTDL}(\phi_1{,}m) = \theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)} = \text{MTTDL}(\phi_2{,}m_1^*)/\text{MTTDL}(\phi_1{,}m_1^*) \leq \text{MTTDL}(\phi_2{,}m_1^*)/\text{MTTDL}(\phi_1{,}m)$ for all $m \geq m_1^*$. Thus, $\text{MTTDL}(\phi_2{,}m) \leq \text{MTTDL}(\phi_2{,}m_1^*)$ for all $m \geq m_1^*$, which in turn implies that $m_2^* \leq m_1^*$. The proof for EAFDL is similar to that for MTTDL and is therefore omitted. ■

From (22) and (23), it follows that the optimal codeword length depends on k and ϕ, but not on the storage system size, n. To investigate the behavior of the optimal codeword length, m^*, as the group size, k, increases, we proceed by considering the normalized optimal codeword length r^*, namely, the ratio of m^* to k:

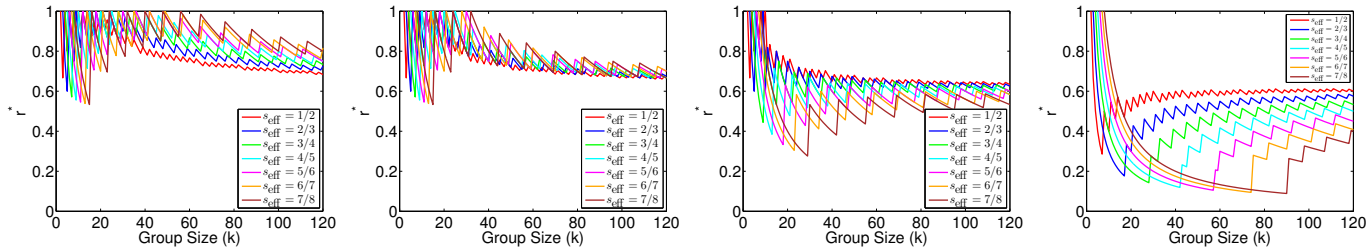
$$r^* \triangleq \frac{m^*}{k}\,. \tag{34}$$

The r^* values for the MTTDL and EAFDL metrics are shown in Figures 8 and 9, respectively, for various storage efficiencies. According to Proposition 1, for any storage efficiency s_{eff} and for any given group size k, the optimal codeword lengths and, consequently, the r^* values decrease as ϕ decreases. Also, when the bandwidth constraint factor ϕ is small, the r^* values first decrease and then gradually increase as k increases. The initial decrease is due to the fact that the optimal codeword length m^* remains fixed and equal to $z + 1$, which is the minimum possible codeword length for the storage efficiency fractions $z/(z + 1)$, $z = 1, \ldots, 7$. For example, in the case of $s_{\text{eff}} = 7/8$ and $\phi = 0.001$, $m^* = 8$ for $k < 115$ in the case of MTTDL, or for $k < 90$ in the case of EAFDL, as



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

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



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 9. r^* for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$.

shown in Figures 8(d) and 9(d), respectively. However, it can be proved that as k increases further, the r^* values for MTTDL and EAFDL approach a common value that depends only on the storage efficiency, s_{eff}, but not on the bandwidth constraint factor, ϕ, and are in the interval $[e^{-1/2} = 0.606, 0.648]$.

V. Discussion

Although erasure coding schemes provide a high data reliability at a high storage efficiency, the rebuild process involves I/O operations and network transfers that increase the consumption of device and network bandwidth. In particular, large MDS codes pose a challenge on the usage of network resources given that a lost symbol is recovered via an (m, l) erasure code through the transfer of a large number of l symbols from l surviving devices over the network. Consequently, recovering large amounts of data results in additional traffic over increased time periods, which has an impact on the latency of the foreground workload and therefore affects system performance. This issue, also known as the *repair bandwidth problem*, has prompted the development of alternative erasure coding schemes that aim at reducing the amount of data transferred over the storage network during reconstruction (see [35][36] and references therein). They can, however, result in higher amounts of data being read from the surviving devices and therefore in longer rebuild times. The effect of these methods on system reliability is beyond the scope of this paper and is a subject of further investigation.

The analytical findings of this work are relevant for the case of large data centers employing erasure coding where the excessive rebuild traffic competes with the huge amount of traffic generated by the frequent access of a large number of storage devices. To ensure a desired performance level, the network bandwidth devoted to the repair traffic needs to be contained. For small values of ϕ and k, a small codeword length should be selected, as discussed in Section IV. For large values of k, the codeword length should still be kept relatively small for performance reasons. This is in agreement with the practical values given in [36] for the various parameters considered. In particular, to keep the storage overhead low, the storage efficiency should be chosen in the range of 0.66 to 0.75.

VI. Conclusions

Data storage systems use erasure coding schemes to recover lost data and enhance system reliability. Network rebuild bandwidth constraints, however, may degrade reliability. A general methodology was applied for deriving the Mean Time to Data Loss (MTTDL) and the Expected Annual Fraction of Data Loss (EAFDL) reliability metrics analytically. Closed-form expressions capturing the effect of a network rebuild bandwidth constraint were obtained for the symmetric, clustered and declustered data placement schemes. We established that the reliability of storage systems is adversely affected by the network rebuild bandwidth constraints. The declustered placement scheme was found to offer superior reliability in terms of both metrics. An investigation of the reliability achieved by this scheme under various codeword configurations was subsequently conducted. The results obtained demonstrated that both metrics are optimized by similar codeword lengths. For large storage systems that use a declustered placement scheme, the optimized codeword lengths are about 60% of the storage system size, independently of the network rebuild bandwidth constraints. The analytical reliability expressions derived can be used to identify redundancy and recovery schemes, as well as data placement configurations that can achieve high reliability. The results obtained can also be used to adapt the data placement schemes when the available network rebuild bandwidth or the number of devices in the system changes so that the system maintains a high level of reliability.

Extending the methodology developed to derive the reliability of erasure coded systems under bandwidth constraints

 ISBN: 978-1-61208-629-3

for arbitrary rebuild time distributions and in the presence of unrecoverable latent errors is a subject of further investigation. Also, owing to the parallelism of the rebuild process, the model considered yields very small rebuild times for large system sizes. Taking into account the fact that the rebuild times cannot be smaller than the actual failure detection times requires a more sophisticated modeling effort, which is also part of future work.

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] Q. Lian, W. Chen, and Z. Zhang, "On the impact of replica placement to the reliability of distributed brick storage systems," in Proc. 25th IEEE International Conference on Distributed Computing Systems (ICDCS), Jun. 2005, pp. 187–196.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] V. Venkatesan, I. Iliadis, and R. Haas, "Reliability of data storage systems under network rebuild bandwidth constraints," in Proceedings of the 20th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Aug. 2012, pp. 189–197.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] ——, "Most probable paths to data loss: An efficient method for reliability evaluation of data storage systems," Int'l J. Adv. Syst. Measur., vol. 8, no. 3&4, Dec. 2015, pp. 178–200.

[21] S. Caron, F. Giroire, D. Mazauric, J. Monteiro, and S. Pérennes, "P2P storage systems: Study of different placement policies," Peer-to-Peer Networking and Applications, Mar. 2013, pp. 1–17.

[22] 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.

[23] ——, "Reliability assessment of erasure coded systems," in Proceedings of the 10th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2017, pp. 41–50.

[24] ——, "Reliability evaluation of erasure coded systems," Int'l J. Adv. Telecommun., vol. 10, no. 3&4, Dec. 2017, pp. 118–144.

[25] J. G. Elerath and J. Schindler, "Beyond MTTDL: A closed-form RAID 6 reliability equation," ACM Trans. Storage, vol. 10, no. 2, Mar. 2014, pp. 1–21.

[26] I. Iliadis and V. Venkatesan, "Rebuttal to 'Beyond MTTDL: A closed-form RAID-6 reliability equation'," ACM Trans. Storage, vol. 11, no. 2, Mar. 2015, pp. 1–10.

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

[28] 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.

[29] 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.

[30] 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.

[31] D. Ford et al., "Availability in globally distributed storage systems," in Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2010, pp. 61–74.

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

[33] S. Muralidhar et al., "f4: Facebook's Warm BLOB Storage System," in Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2014, pp. 383–397.

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

[35] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network coding for distributed storage," Proc. IEEE, vol. 99, no. 3, Mar. 2011, pp. 476–489.

[36] M. Zhang, S. Han, and P. P. C. Lee, "A simulation analysis of reliability in erasure-coded data centers," in Proceedings of the 36th IEEE Symposium on Reliable Distributed Systems (SRDS), Sep. 2017, pp. 144–153.

 ISBN: 978-1-61208-629-3

Reliability Measure for a System Operating under Random Environment

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

***Abstract*—In this paper, we consider a system operating in a random external shock process. The underlying system performance is modelled by a quality (output) function which is decreasing due to degradation. Shocks affect the failure rate of the system directly and, at the same time, they additionally decrease the quality function. Expectations (unconditional and conditional on survival) and variability of this time-dependent quality function are analyzed.**

Keywords-quality characteristic; random environment; shock process; intensity process; variability measure.

I. INTRODUCTION

The performance of various engineering systems is often characterized not only by reliability characteristics, but also by characteristics of performance (output). For instance, a quality function for production systems can be described by the production rate, i.e., the number of items produced in a unit interval of time. For navigation systems, this quality is characterized by the accuracy of navigation parameters such as heading, altitude and longitude. It is well understood that most engineering systems are deteriorating in some stochastic sense and deterioration affects not only reliability indices but also the quality of performance [1][2]. [1] and [2] have considered mostly deterministic quality function. However, the quality or performance of a system should depend on random operational environment. In this regard, in this paper, we will consider stochastic quality functions.

In this paper, we study the reliability measure for a system operating in a random environment. The random environment is modeled by a process of external shocks. We suggest a novel approach in shocks modeling when shocks have a double effect, i.e., they act directly on the failure rate (more precisely, on the corresponding failure rate process) that characterizes the time to failure of a system and, at the same time, on the quality function as well. For example, for a network system, if a shock (e.g., external attack) occurs, the susceptibility to a failure of the network increases and, at the same time, the performance of the network decreases. To account for this complex influence and to obtain explicit expressions for characteristics of interest, we derive the necessary conditional and unconditional average characteristics under the assumption of the Non-homogeneous Poisson Process (NHPP) of shocks. Specifically, we obtain the expectation and the variance of the quality function of a system on condition that a system is operable at a given instant of time and without this condition.

In Section 2, we introduce the model studied in this paper. Furthermore, the unconditional and conditional expected quality functions are derived. In Section 3, the unconditional and conditional variability measures are obtained. Finally, in Section 4, we provide a brief conclusion.

II. EXPEXTED QUALITY OF THE SYSTEM

Assume that a non-repairable system is operating in a random environment modeled by the NHPP of shocks $\{N(t), t \geq 0\}$ with the rate of occurrence $\lambda(t)$, where $N(t)$ is the number of shocks by time t. Define its lifetime by the following conditional failure rate (intensity process) [3]

$$\lambda_t = r_0(t) + \eta N(t), \tag{1}$$

where $r_0(t)$ is the baseline failure rate of a system that is operating in the absence of shocks and $\eta > 0$ is a constant jump in the failure rate on occurrence of each shock. Thus, each shock increases λ_t in each realization of this stochastic process by the same deterministic value.

Let $Q(t)$ be a deterministic quality or performance function of an operating system, which is monotonically decreasing [1]. Moreover, assume also that the quality or performance is decreasing on each shock. To account for this effect of the shock process in a consistent way, we assume that the quality at time t under a shock process is given by the following stochastic process

$$\widetilde{Q}(t) = Q(t) \prod_{i=1}^{N(t)} \exp\{-\psi(T_i)\}, \tag{2}$$

where $\psi(t) > 0$ is a deterministic function and $0 \leq T_1 \leq T_2 \leq \ldots$ are the sequential arrival times of shocks in the NHPP.

Let $I(t)$ denote the corresponding indicator of the system state (1 if the system is operating at time t and 0 if it is in the state of failure). Our first measure of interest is

$$Q_E(t) = E[\widetilde{Q}(t) \cdot I(t)], \tag{3}$$

 ISBN: 978-1-61208-629-3

which is the expectation of the quality function of a system at time t (assuming that the quality is 0 when a system is in the state of failure). Note that when $\widetilde{Q}(t) \equiv 1$, for all $t \geq 0$, $Q_E(t)$ in (3) is the usual 'reliability function'.

Result 1. The expected quality function $Q_E(t)$ is given by

$$Q_E(t) = Q(t)\exp\left\{-\int_0^t r_0(u)du\right\}\exp\left\{-\int_0^t \lambda(u)du\right\}\cdot \times \exp\left\{\int_0^t \exp\{-\eta t - \psi(x) + \eta x\}\lambda(x)dx\right\}\cdot$$

Proof. It can be shown that the joint distribution of $(T_1, T_2, ..., T_{N(t)}, N(t))$ is given by

$$\left(\prod_{i=1}^{n}\lambda(t_i)\right)\exp\left\{-\int_0^t \lambda(u)du\right\},\ 0 \leq t_1 \leq t_2 \leq ... \leq t_n \leq t, n = 0,1,2,...,$$

and taking expectation of $[\widetilde{Q}(t)\cdot I(t)]$ with respect to this distribution yields the desired result.

In many instances and especially when considering characteristics of quality in a population of systems, it could be more interesting and practically sound to obtain the expected quality for systems that are 'operating at time t'. Hence, our second measure of interest is the following conditional expectation:

$$Q_{ES}(t) = E[\widetilde{Q}(t) \mid T > t], \quad (4)$$

where T is the system lifetime and "S" in $Q_{ES}(t)$ stands for "survived".

Result 2. The conditional expected quality function $Q_{ES}(t)$ is given by

$$Q_{ES}(t) = Q(t) \times \exp\left\{\int_0^t \exp\{-\eta t - \psi(x) + \eta x\}\lambda(x)dx - \int_0^t \exp\{-\eta(t-x)\}\lambda(x)dx\right\}.$$

Proof. It is similar to the proof of Result 1.

III. Variability in Quality of The System

Note that the quality of a system $\widetilde{Q}(t) = Q(t)\prod_{i=1}^{N(t)} \exp\{-\psi(T_i)\}$ and the conditional quality of a system $(\widetilde{Q}(t) \mid T > t)$ are stochastic processes. In the previous section, we have considered expectations of these quality measures as important reliability characteristics of a system. In this section, we will discuss the time-dependent variability of the quality, which can be represented by the variance or the conditional variance at each time instant. Thus, we now define the following measures for variability of quality.

$$VQ_E(t) = Var[\widetilde{Q}(t)I(t)],$$

and

$$VQ_{ES}(t) = Var[\widetilde{Q}(t) \mid T > t].$$

These measures are obtained in the following result.

Result 3. The variability measures $VQ_E(t)$ and $VQ_{ES}(t)$ are given by

$$VQ_E(t) = Q(t)^2 \exp\left\{-\int_0^t r_0(u)du\right\}\exp\left\{-\int_0^t \lambda(u)du\right\} \times \exp\left\{\int_0^t \exp\{-\eta t - 2\psi(x) + \eta x\}\lambda(x)dx\right\} - Q(t)^2 \exp\left\{-2\int_0^t r_0(u)du\right\}\exp\left\{-2\int_0^t \lambda(u)du\right\} \times \exp\left\{2\int_0^t \exp\{-\eta t - \psi(x) + \eta x\}\lambda(x)dx\right\},$$

and

$$VQ_{ES}(t) = Q(t)^2 \times \exp\left\{\int_0^t \exp\{-\eta t - 2\psi(x) + \eta x\}\lambda(x)dx - \int_0^t \exp\{-\eta(t-x)\}\lambda(x)dx\right\} - Q(t)^2 \exp\left\{2\int_0^t \exp\{-\eta t - \psi(x) + \eta x\}\lambda(x)dx - 2\int_0^t \exp\{-\eta(t-x)\}\lambda(x)dx\right\},$$

respectively.

Proof. It is similar to the proof of Result 1.

Note that $VQ_E(t)$ represents the unconditional variation, whereas $VQ_{ES}(t)$ provides the conditional variation.

IV. Conclusion

In this paper, we have considered a system operating under a random Poisson shock process. Each shock affects the failure rate of the system and the quality of the system simultaneously. Under the suggested model, the unconditional and conditional expected quality functions have been derived. Furthermore, the unconditional and conditional variability measures have also been obtained. This paper extends the previous works [1][2] by considering stochastic quality functions, which is practically meaningful generalization.

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] M. Finkelstein, "The performance quality of repairable systems," Quality and Reliability Engineering International, Vol. 19, pp. 67-72, 2003.

[2] M. Finkelstein, "Failure rate modeling for reliability and risk," London: Springer, 2008.

[3] J. H. Cha and J. Mi, "Study of a stochastic failure model in arandom environment," Journal of Applied Probability, Vol. 44, pp. 151-163, 2007.

 ISBN: 978-1-61208-629-3

A Survey of Internet Protocol and Architectures in the Context of Emerging Technologies

Kiran Makhijani, Renwei Li, Alexander Clemm, Uma Chanduri, Yigzhen Qu, Lin Han

Future Networks, America Research Center
Huawei Technologies Inc., Santa Clara, CA, USA
email: {kiran.makhijani, renwei.li, alexander.clemm, uma.chunduri, yingzhen.qu, lin.han}@huawei.com

***Abstract*—The Internet technologies need an overhaul to support next-generation of applications requiring communications between machines and humans. This paper is a survey of the state of current internetworking architecture and its engineering properties. The purpose of this paper is to highlight the aging of original design goals and motivations. We aim to formulate a new set of guidelines that maybe used to postulate design principles of the new network architectures.**

Keywords–Internet architecture, Internet Protocol, Routing, Switching; Ossification, layering.

I. INTRODUCTION

The Internet has grown remarkably since its foundational work was published as *A Protocol for Packet Network Intercommunication* [1]. This specification was developed into Transmission Control Protocol/Internet Protcol (TCP/IP) in compliance with the Internet design principles [2]. While the Internet has proven to scale and support diverse set of applications and users, the more recent technological advancements such as Machine to Machine (M2M) communications, connected or live Augmented Reality/Virtual Reality (AR/VR), Vehicle to Anything (V2X) communications etc., impose new requirements on connectivity that did not exist before. The applications based on these technologies are far more stringent about both network resource constraints and packet delivery guarantees. The current architecture lacks several artifacts to guarantee support for real-time, low latency and reliable services. In this regard, several new network architectures have been proposed with different motivations; however, none of them have been attentive to strict quality of service constraints.

In this paper, we systematically analyse effects of current architectural and engineering design choices (both adversely and favorably) that can be used to understand specific gaps in the context of emergent applications. These effects are identified as: 1) the commercial effect, 2) layering, 3) addressing, 4) Ossification, and 5) services; They will be discussed in detail to highlight their influence and stronghold on the current state of the Internet. In this study we show that the current principles of inter-networking are not sustainable to serve applications built for the use of emerging technologies. The paper further aims to achieve the following:

(a)

1) Briefly describe use cases catogarized as emerging applications.
2) Provide an analysis of original design principles and corresponding engineering effects.
3) Guidelines to be taken under consideration when designing new or evolving the current Internet architecture.

The paper is organized as follows: Section II briefly mentions future network architectures related work, while Section III starts with the background and motivation for this paper, in Section IV we analyse the original concept and design goals of Internet architecture. Section V is a discussion on the engineering effects of the Internet and their analysis in context of emerging applications. Section VI proposes properties to be taken in to account for designing new internet architecture. In Section VII, we expose the factors that will drive the need for new Internet architectures. We conclude with a summary of this survey in Section VIII.

II. RELATED WORK

This paper primarily analyses several published works of Kahn, Cerf and Clark. Their insights and reflections on the design of the Internet have been taken into consideration in the context when analysing the current state of Internet.

The discussion for new architecture has come up several times. In fact, immediately following the Internet impact, guidelines for the future network architecture were produced in RFC1287 [3]. It revealed several interesting shortcomings relating to addresses, multi-protocol architectures, traffic control and security. It also mentioned that service awareness was necessary in general and specifically for voice, video and teleconferencing type of applications.

There has been continuous effort in building next generation internet encompassing from evolutionary to clean-slate approaches [4]. More recently, some of the large-scale future internet initiatives eXpressive Internet Architecture (XIA), Future Internet Research and Experimentation (FIRE), Named Defined Networks (NDN), Software Defined Networking (SDN), etc. [5] have been proposed to solve known problems. None of these initiatives can be qualified as either failed or successful projects since they did not get deployed and tested in live environments. In principle, the network community understands a need to upgrade the Internet architecture and design, however, none of the efforts have been able to stir a serious interest from commercial sector. Several federated and national initiatives such as Future Internet Architecture (FIA) [6], 4WARD [7], AKARI [8], Study Group 13, Future Networks (SG13) [9] and many more do not transition from research to commercial mainstream even

 ISBN: 978-1-61208-629-3

after having undergone thorough experimentation (FIRE [10], Global Environment for Network Innovations (GENI) [11]).

An obvious reason is the growth in Internet and its ability to absorb many motivations of new architectures. Another possible reason may be that the new solutions focus on a particular problem-domain instead of taking the holistic approach along the lines of design principles. Our contribution focuses on support for communication aspects of current and future technological advances in medicine, manufacturing, city planning and automating vehicles etc as a driver to review current Internet design.

III. BACKGROUND

Clark's Internet design philosophy serves as the guiding principles of the Internet architecture [2]. According to Kahn [12], the reference architecture and TCP/IP as an implementation are often used interchangeably, but that was not the intent. The reference design of the Internet was a logical framework for interconnection of independent networks and TCP/IP is one such instance that implemented it. Kahn also admits that the reference architecture itself does not assume the idea of linking different networks together will result into a single system. The vision was to foster multiple implementations serving different systems from the same abstract architecture. With the TCP/IP, this generality of the design was lost which prevents the evolution of Internet from its current state without disruption [4]. The TCP/IP resists change and on-boarding new services to support new applications is a difficult task.

Traditionally, services with special constraints in networks concern with delivery of data through Quality Of Service (QoS) parameters that are represented by coarse grained means of allocating network resources (e.g., buffers and bandwidth) using code points [13] [14] on per hop or end-to-end [15] basis. For example, a service characteristic such as lower latency may be marked to code-point that indicate 'real-time' traffic. In contrast, the emergent services for scenarios such as M2M, V2X, AR/VR communications are associated with extremely strict resource constraints and absolute guarantees of QoS in the network. For example, industrial automation relies on M2M communication to achieve reliable interaction between different type of machines with a fine-grained granularity in delay variation. Any failure to deliver data in precise time-interval could cause machines go into stall mode halting the over all production. Similarly, in V2X scenarios, the infrastructure should be able to gather live information from multiple sources such as approaching signals, road conditions, and other vehicles to make real-time decisions about public safety and streamlining traffic flow; while ensuring the decision is fed to an autonomous vehicle instantaneously; any delay makes information stale and unusable. Rest of the paper collectively refers to these use cases as *emerging applications*.

IV. FOUNDATIONS OF THE INTERNET ARCHITECTURE

The primary goal of the reference internet architecture is to provide an effective technique to multiplex packet switched data over interconnected networks. There were seven additional goals (see [2]) that had to be met at the time of the internetworking design. While these design principles are generally accepted, as times change and technologies evolve, some of the original principles cannot be followed as is.

The first goal, '*Internet should continue to provide communication service...*' is about suvivability and fate-sharing. In networks, fate-sharing suggests that it is acceptable to lose the state information of an entity, if the entity itself is lost. This principle entirely takes away the responsibility of reliability in the network which will require some knowledge of relevant state.

There is an indirect consequence of this principle, that the network is stateless with no knowledge besides forwarding information of an entity. While, it is true that maintaining an overall state of all the sessions in the Internet is unmaintainable; there are specific scenarios where it provides resilience, robustness through faster recovery and security. There is also a question of what determines that an entity is lost. Whether a session was withdrawn gracefully or due to failure such as congestion or packet loss in the network can not be determined by the network itself. In industrial interent, M2M communication scenarios require bounded latency and are sensitive to delays, such connections benefit from having state in the network. Relaxing this fate-sharing principle will help determine fate of an entity. In Internet of Things (IoT) communications entities would goto sleep mode but may still have associated active state in the network for high reliability scenarios. Therefore, future internet design goal may consider fate-sharing to be optional or need-basis; Certain type of services, such as those requiring zero packet loss, in-network stateful buffering can help trigger retransmissions from a nearer hop without involving end hosts.

Additionally, it is noted that the statelessness is already diminishing in the Internet to a certain extent due to ever-growing use of middle boxes that are largely stateful. The middle boxes are generally considered to compromise network transparency and break End-to-End (E2E) principle. Yet, in practice they bring a lot of value to commercial enterprises by performing Network Address Translation (NAT), firewall and similar functions.

The second goal *it should support, at the transport service level, a variety of types of services* manifested in to not making any underlying assumption about the services in the datagrams. Unfortunately, this behavior does not translate well in TCP/IP. In the context of telecommunications and data communications convergence, voice service in a telecom network outperforms Voice over IP (VoIP). Support for real-time applications needing low latency still cannot be assured. This is due to lack of service awareness about the packets as it is transmitted through the network. Clark had a broader view about the structure of datagrams as building blocks that provide pieces of information about services and corresponding resource requirements in such a manner that each datagram is a self-describing construct. This behavior rightfully, was too complex for that time and did not make it to TCP/IP. While Type Of Service (TOS) in IP is available, it is a) too generalized for emerging services characterization, b) in practice, the interpretation and scope is always within an internal network and has no significance in internetworking.

The third and seventh goals, *'the architecture must permit distributed management of its resources*' and '*it must be accountable'* are somewhat related to the cost. Distributed management of network resources is realized through control plane protocols. In this regard, the composition of services and

 ISBN: 978-1-61208-629-3

allocation of network resources, has been a difficult problem. This is because of the trusted domain concept and establishment of trust between transit networks happens outside the Internet Protocol (IP). In the absence of seventh goal (accounting), there is no distributed way to convey explicit business value in datagrams to obtain resources from transiting networks. It would have been a simpler problem to solve if the structure of datagram had permitted for presence of accounting information. Hence, the third and seventh goals of original design have remained unfulfilled.

The fourth goal, *architecture must be cost effective* pertains to inefficiencies in packet transmission that are incurred either due to header overheads or retransmissions. In the context of IoT type of devices, large header related inefficiencies become even more prominent and are handled through header compression schemes [16]. Back then (1980s), a retransmission rate of 1 in 100 was tolerable. However, it is now unacceptable for AR/VR applications that tolerate loss of 1 in 10000 packets [17], using current TCP throughput computations for a 15200 Mbps stream with delay tolerance of 0.106 ms. For such applications, retransmissions and packet loss have to be absolutely avoided with the assumptions that end users are willing to incur the cost of sych services.

Essentially, many of Clark's goals are not sufficient to meet present-time service requirements as discussed in this section and a reformation of the design principles are necessary. This can possibly be achieved by reviving the concept of datagrams carrying relevant information for use in the networks. Emerging applications are in need of in-network state, service awareness and resource control. The cost effectiveness varies for different application environments and business demands. To this effect Internet being cost-effective cannot be a fundamental design goal and applications should have choice to opt for premium services.

V. Engineering Inertial Effects

Over multiple decades, a lot of engineering effort has gone into keeping the Internet stable and allowing it to scale. The resulting Internet is rigid, that resists changes necessary in the context of a wide variety of modern applications. The structure of the present day Internet can be described through a set of inertial effects since they provide means to maintain status quo while avoiding substantial changes. An exploration of these effects will reveal the trade-offs between their strengths and shortcomings which may further help design next-generation architectures.

A. Commercial Effect

Commercial aspect of the Internet manifested into three characteristics viz. explosion of routing table, proliferation of private networks through tunnels, and a surge in non-default forwarding of the traffic.

Firstly, as new websites or corporate sites are added, replaced or merged, the global routing table is affected and often misconfigured routes lead to the instability in Internet backbones. This can become a cause of major outages on regular basis [18]. To minimize global routing updates, techniques like damping [19] are employed. However, this method has limitations such as loss of connectivity due to suppression of correct updates, missing routes and the configuration complexity [20].

Secondly, commercialization also created a business case for multi-site private networks. It became necessary for any corporation to isolate and protect its digital assets when traversing through the public Internet. Interestingly, the original TCP/IP design was a single system with no notion of network-to-network communication. To implement private networks tunnels are deployed emulating a network as a host (through a tunnel endpoint). It then requires a complex instrumentation and an entirely independent stack of protocols [21].

Thirdly, the measurement studies in [22] and related work [23] discovered E2E path anomalies, i.e., not all packets between the same source and destination were subjected to the same path (non-default routing). This is because various commercial features and business reasons have different service requirements that are fulfilled by operator driven configurations and/or route-policies.

The inertial aspect of this effect is that the current architecture implicitly resists change of any kind in favor stability, overrides the notion of single system (through Virtual Private Networks (VPN)s) and overrides default routing. A variety of services requiring real-time, high bandwidth, zero packet loss etc. resort to tricks such as path computations, route policies and complex configurations just to get close to meeting their service level objectives. They take away the dynamic nature of forwarding which is not desirable traits of emerging applications where seamless connectivity and ubiquitous mobility are essential requirements.

B. Layering and E2E Effect

The layering principle is honored when a) separation of layers is not compromised or violated (layer independence), b) there exists minimal layer crossing (services provided to next higher layer only). The E2E principle creates transparency; i.e., the network bears no knowledge of the contents and remains non-discriminant about the applications. Both layering and E2E principles are the foundation of the Internet and a consequence of how TCP/IP got implemented.

Layering gets breached in the form of tunnels, overlays, NATs, etc. through port blocking or filtering techniques. It is well known that many in-production routers do not allow traffic other than TCP and UDP [24] to pass through. This is an obvious violation but is necessary for Internet Service Provider (ISP)s to protect their network against spurious attacks. Layering helps scale different type of services but there are a few drawbacks as well. Firstly, multiple levels of encapsulations lead to bloating of the proverbial narrow waist in hourglass structure of TCP/IP. Secondly, the encapsulated layer comes with its own protocol, control and corresponding management entities, thereby increasing network complexity.

From an architectural standpoint, layer abstraction is a powerful concept, but in practice, it has resulted into a mechanism to hide deficiencies in the structure of datagrams that do not carry sufficient control information. Similarly, E2E effect has manifested into dumb networks and intelligent end points making is impossible for networks to make informed decisions unless middleboxes are deployed.

 ISBN: 978-1-61208-629-3

C. Network Ossification Effect

Ossification suggests both long-term survivability and as a consequence rigidity in the network. It is believed that both network [25] and transport [26] layers resist adoption of new technologies. The ossification is a consequence of gradual building of resiliency and stability in TCP/IP technology over a long period of time (also alluded to in Section V-A).

SDN has been positioned to mitigate effects of ossification and has produced several changes in network control through programmability. However, SDN also brings an increased complexity and scalability limitations due to central control for programming the networks. In contrast, transport ossification is mainly a side-effect of use of middle boxes to bypass lack of modularity in transport layer for service customizations. In an E2E client server communication, any change to TCP, needs coordination with all client instances. Over time, the structural uniformity has become more rigid and most customizations happen over HTTP instead (e.g., DASH, session management).

The ossification of both network and transport layer implies they cannot be changed. SDN only deals with the control plane programmability, however, dataplane flexibility is extremely desirable for M2M communications requiring low latencies. The emerging applications scenarios are exactly the kind where application level session management will be inefficient and impractical, instead a network assisted packet processing techniques will be necessary.

D. Addressing Effect

The Internet is ubiquitous and homogeneous because of a uniform network addressing scheme in which a host is understood in identical manner in each network. There are two major factors regarding host addresses. First factor pertained to the size; it was recognized that the 32-bit width will be too small to cover every host on the Internet [3] [27]. Second factor related to its structure limitations; that the early binding aspects of an application and address turned out to be limiting the functions of mobility, device portability and multi-homing. This is also characterzed as location and identifier separation concept.

E. Services Effect

Even though variety of services were anticipated in the architecture, the earlier ones were primarily texts or static digital image formats. With digitization of audio and video, many new applications started to emerge and Internet was suited for many of those. For example, early attempts to implement VoIP, which is a circuit-switched telecommunication service, over a packet switched network were suboptimal, because the goal of the internet was to support best-effort communication, but voice performs the best as a circuit switched application. The QoS markings that exists today are coarse-grained, therefore, only a very narrow category of services can be supported.

Today, there are even more variety of such services but with even more stringent requirements. M2M communications take humans out of the loop; an application relies on each machine-entity to function properly, respond, generate and process events according to prescribed behavior. Any delay, loss of packets in network could be misinterpreted as failures, causing system to take serious recovery actions leading to loss of productivity.

F. Summary of Inertial Effects

The runtime state of the Internet is a consequence of above mentioned effects that emerged from the TCP/IP implementation. Due to limitation of space, and non-technical aspects that vary for different countries, we do not discuss governance aspect here; however, it is also relevant in shaping the Internet. For interested reader we offer the following references [28] [29].

The Internet is 'robust yet fragile' [30]. It is well-engineered in responding to predicable events, but unable to handle unanticipated circumstances. Layering, addressing and commercialization effects are virtue of the principle of keeping network layer simple. The mechanisms adopted were mainly based on conservative and cost-effective design choices with the goal of scaling the Internet. Ironically, this has brought comlpexity elsewhere in management, operations and orchestration functions of the networks [31] [32]. In constrast, Ossification and services effects aimed to minimize variations in the network thereby lacking customization mechanisms that are needed for finer granularity of control for certain classes of applications.

The TCP/IP is an over simplified instance of the original design and trade-offs made several decades ago were well-suited for applications of that time. Not only that the state of art routers and network nodes are more powerful now but the networks play a much bigger role in all aspects. The emerging applications driven by M2M communications will expose the above mentioned limitation even further. A new balance has to be struck between preserving the stability aspect of the Internet and yet allow it to evolve for those applications.

VI. Properties for Transition from Classical to New Internet

We have discussed the architectural aging and engineering effects in previous sections and call the structure as classical Internet.

The Internet is diverifying in all facets, a new Internet architecture must be defined to be simultaneously public and private, secure and open, social and commercial as well as both human and machine centric. In the context of discussions in previous sections, the properties for new network architecture are proposed.

A. Multi-Instance Architecture

The idea of multi-instanced Internet should be explored. Where an instance could be a special purpose and means to connect with other instances if necessary. This could serve new generation of technologies with specific type of network resources better. This has already been noted as fragmented Internet in [33] (as a warning, not a feature). Often general-purpose solutions suffer from performance and complexity. In contrast special purpose networks can be more efficient but limited. A single system is automatically prone to be rigid and conservative, being a single point of failure. Having multiple instances allow features to be experimented and withdrawn.

 ISBN: 978-1-61208-629-3

RFC1958 [27] discusses the possibility of at least two network protocols to be in use to support gradual transition. Prevailing encapsulation-based mechanisms suffer from bloating (Section V-B). This approach could provide with flexible and dynamic bindings to information, in a tunnel-free manner.

B. Distributing Complexity

Section IV expains that the fate-sharing principle has led to stateless design of the networks. The connectivity scenarios are evolving rapidly as a large number of endhosts (e.g., wearables, sensors, appliances, etc.) are far less powerful than the routers and switches. It makes sense for network to be more aware of device behavior and the services they require by distributing some of communication processing from end-points into the networks. Several emerging applications (for example, V2X, industrial automation and remote control etc.) have strict service level criteria for normal operation in terms of bounded latency or committed bandwidth. Without direct sharing of such information, the network design becomes inefficient as an operator would need to understand requirements of each application and setup resources accordingly through central control on per hop basis. Also this is possible with SDN paradigm, maintenence of per flow state is non-trivial. In current architecture, the networks have evolved in a manner that the intelligence lies with the applications or end points. New design should identify mechanisms that distribute intelligence into the networks, as in service awareness, runtime state or behavior. This In effect, distributes complexity partially from the end points in the network. Such considerations are manadatory to meet service level objectives of absolute guarantes .

Many technological advances have happened in the hardware of network devices. The Network Processing Units (NPU)s, Ternary Content Addressable Memories (TCAM)s and port Application Specific Interface Circuit (ASIC)s are much faster than before. New architecture can use advances in hardware to their advantage while exploring solutions for emerging applications.

C. Self-Sufficing Datagrams

The datagram provides a basic building block out of which a various types of service can be implemented. The notion of a datagram carrying service-centric information can be an extremely powerful concept to address several control and management inefficiencies in network. A datagram should be a self-sufficient, self-describing entity comprising of user payload and control information about the flow or application it is part of. Obviously, it comes at a cost of additional bits on wire but a sensible structure and the framework could be deployed. The information must be network centric and should be detached from the transport aspects. As mentioned before, hardware advancements can be used to deploy efficient processing in the networks.

D. Flexible Address Structure

While the uniformity of addresses need to be preserved for the sake of reachability, a variable structure that is more sensitive for IoT devices should be supported. This has already been proposed in [34].

These are the main recommendations and can be used as guiding principles for new architectures to make distinction between the things in networks that should change (e.g., services, resources), and the things that provide stability (e.g., uniformity of addresses and layering). Other guidelines are possible based on specific choice of architectures but we only mention the ones that can be added to any next-generation architecture proposal.

VII. Factors Driving the need for New Internet

As mentioned earlier in Introduction, the need to change is driven by applications. Had the communications remained web-based online transactions or consuming streaming media, the current Internet works just fine. However, a ubiquity of connectivity is emerging in several aspects. It is required to think of the Internet as a fabric that interconnects humans, services, sensors, devices etc.

It is well known that IoT space will grow to billions of devices and each of them will have varying characteristics such as identity (corresponding connectivity address and gateway), its functionality (what purpose is it used for), energy efficiency (to help determine right type of transport mechanisms) to state a few. This high level of diversity is compounded by the volume of data that is produced at varying intervals. As a result current foundations of transport protocols do not apply to IoT, new techniques to efficiently transfer information and yet reducing or eliminating setup times are needed. 4

A fully automated vision of Industry 4.0 takes IoT to next level in terms of M2M communications. Today manifacturing networks are proprietory and purpose built. Looking ahead, there are three factors that will drive integration of Industrial network into mainstream Internet, a) combining Information Technology (IT) and Operation Technology (OT), b) use of common technologies, c) resource assurances. Even through the manufacturing in a factory is automated, OT and IT are managed as two separate networks. This requires a human to integrate results from information technologies in to operations. To achieve complete automation, IT and OT must be combined so that the results from complex analytics can be fed into command center. Secondly, the investments in infrastructure can be reduced by using standard technologies, which will not only incentivise manifactureres to automate at large scale but also allow with modern cloud based infrastructure solutions.

Inspite of the above compelling factors, it is a difficult task to change the incumbent Internet owing to its success. Today, it is so big that even minor outages are unacceptable. The Internet is IP based and most of the standardization is driven by Internet Engineering Task Force (IETF). These standards can only afford to bring segmented improvements in a particular focus area such as operations, routing, transport etc. Architectural changes related discussions often happen at other Standards and Development Organization (SDO)s such as European Telecommunications Standards Institute (ETSI), International Telecommunications Union (ITU) involved in study and evaluation of new network architectures. Perhaps a close coordination among these SDOs will be necessary to further the design of new architecture.

 ISBN: 978-1-61208-629-3

VIII. Conclusion

In this paper, we surveyed two related topics; first, the design decisions that led to current network architecture and were reasonable at that time. The second topic reflects upon the consequences of first in terms of its inertial effects that makes it difficult for the Internet to evolve from its current state. We also establish that the foundations of Internet architecture have been strong and were well engineered. However, in the context of emerging applications and new types of communications, some of the principles are outdated and must be revisited. This can be achieved either through a new or evolved architectural principles that balance both incumbant stability and adoption of new features. Looking ahead at emerging applications, Internet as a single system will be difficult to scale, it is simpler to evolve and adopt in multi instance environments. Finally, to future-proof new architecture and design, datagram building blocks are the key. They should be allowed to evolve, be extensible and support flexible mechansims for variety of applications.

References

[1] V. Cerf and R. Kahn, "A protocol for packet network intercommunication," *IEEE Trans. Commun.*, vol. 22, pp. 637–648, May 1974.

[2] D. D. Clark, "The Design Philosophy of the DARPA Internet Protocols," *ACM Symposium proceedings on Communications architectures and protocols*, vol. 18, pp. 106–114, Aug. 1988.

[3] D. D. Clark, L. Chapin, V. Cerf, R. Braden, and R. Hobby, "RFC 1287, Towards the Future Internet Architecture," *Internet Engineering Task Force (IETF)*, Dec. 1991.

[4] N. McKeown and B. Girod, "Clean-Slate Design for the Internet, Whitepaper Version 2.0," *A Research Program at Stanford University*, 2006.

[5] J. Pan, S. Paul, and R. Jain, "A survey of the research on future internet architectures," *IEEE Communications Magazine*, vol. 49, p. 38, July 2011.

[6] "NSF Future Internet Architecture Project." http://www.nets-fia.net, 2010. Last accessed 14 April 2018.

[7] "EU, Future Internet Initiative, FP7 Project." "http://www.4ward-project.eu". Last accessed 14 April 2018.

[8] "AKARI Architecture Design Project." http://www.nict.go.jp/publication/shuppan/kihou-journal/journal-vol62no2/journal-vol62no2-03-00.pdf, 2006.

[9] "Study Group 13 - Future networks." https://www.itu.int/en/ITU-T/about/groups/Pages/sg13.aspx. Last accessed 14 April 2018.

[10] "Future Internet Research and Experimentation." https://www.ict-fire.eu/fire, 2010. Last accessed 14 April 2018.

[11] "Global Environment for Network Innovations." http://www.geni.net, 2010.

[12] V. Cerf and R. E. Kahn, "Assessing the Internet: Lessons Learned, Strategies for Evolution, and Future Possibilities," *ACM Turing award lectures*, 2004.

[13] K. Nichols, S. Blake, F. Baker, and D. Black, "RFC 2474, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," *Internet Engineering Task Force (IETF)*, Dec. 1998.

[14] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "RFC 2475, An Architecture for Differentiated Services," *Internet Engineering Task Force (IETF)*, Dec. 1998.

[15] J. Wroclawski, "RFC 2210, The Use of RSVP with IETF Integrated Services," *Internet Engineering Task Force (IETF)*, Dec. 1997.

[16] J. Hui and P. Thubert, "RFC 6282, Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks," *Internet Engineering Task Force (IETF)*, Sept. 2011.

[17] R. Li, A. Sutton, Ed., "Next Generation Protocols - Market Drivers and Key Scenarios, Future Internet architecture." "http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp17_Next_Generation_Protocols_v01.pdf", Oct. 2016.

[18] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Backbone Failures," *Digest of Papers. Twenty-Ninth Annual International Symposium on Fault-Tolerant Computing (Cat. No.99CB36352)*, pp. 278–285, 1999.

[19] C. Villamizar, R. Chandra, and R. Govindan, "RFC 2439, BGP Route Flap Damping," *Internet Engineering Task Force (IETF)*, Nov. 1998.

[20] J. Vahapassi, "Internet Routing Stability, ICL Data Oy." http://keskus.hut.fi/opetus/s38130/k00/Papers/Topic17-Stability.doc.

[21] E. Rosen and Y. Rekhter, "RFC 4364, BGP/MPLS IP Virtual Private Networks (VPNs)," *Internet Engineering Task Force (IETF)*, Feb. 2006.

[22] M. Canbaz, K. Bakhshaliyev, and M. H. Gunes, "Analysis of path stability within autonomous systems," p. 38, 2017.

[23] A. Zakaria, M. Alsarayreh, I. Jomhawy, and M. Rabinovich, "Internet Path Stability: Exploring the Impact of MPLS Deployment," *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, 2016.

[24] "Port Blocking, Broadband Internet Technical Advisory Group Technical Working Group Report." "https://www.bitag.org/documents/Port-Blocking.pdf", Aug. 2013. Last accessed 14 April 2018.

[25] National Research Council, "Looking Over the Fence at Networks, A Neighbors View of Networking Research," *"National Academy Press"*.

[26] J. S. Turner and D. E. Taylor, "Diversifying the Internet," *GLOBECOM 05. IEEE Global Telecommunications Conference*, vol. 2, pp. 1–6, – 760, 2005.

[27] B. Carpenter, ed., "Architectural Principles of the Internet," *Internet Engineering Task Force (IETF)*, June 1996.

[28] L. Solum and M. Chung, "The layers principle: Internet architecture and the law," *Ssrn Electronic Journal*, vol. 179, no. 1, p. 38, 2003.

[29] J. Rexford and C. Dovrolis, "Future internet architecture," *Communications of the ACM*, vol. 53, pp. 36–40, Sept. 2010.

[30] R. Bush and D. Meyer, "RFC 3439, Some Internet Architectural Guidelines and Philosophy," *Internet Engineering Task Force (IETF)*, Dec. 2002.

[31] M. Behringer and G. Huston, "A Framework for Defining Network Complexity," *Internet Engineering Task Force (IETF)*, Nov. 2012.

[32] M. H. Behringer, "Classifying Network Complexity," *ACM Workshop on Rearchitecting the Internet, ReArch 09*, p. 13, Nov. 2009.

[33] J. W. Drake, V. G. Cerf, and W. Kleinwchter, "Future of the Internet Initiative White Paper, Internet Fragmentation: An Overview," *World Economic Forum*, p. 80, 2016.

[34] P. Esnault, Ed., "GR NGP 004 - Evolved Architecture for mobility using Identity Oriented Networks." "http://www.etsi.org/deliver/etsi_gr/NGP/001_099/004/01.01.01_60/gr_NGP004v010101p.pdf", Jan. 2018.

 ISBN: 978-1-61208-629-3