Performance Evaluation of Packet Re-ordering on Concurrent Multipath Transmissions for Transport Virtualization

Thomas Zinner*, Phuoc Tran-Gia

University of Wuerzburg,
Institute of Computer Science, Chair of Communication Networks
Am Hubland, 97074 Wuerzburg, Germany
E-mail: zinner,trangia@informatik.uni-wuerzburg.de
*Corresponding author

Kurt Tutschku

University of Vienna,
Institute of Computer Science, Chair of Future Communication,
Universitaetsstrasse 10/T11, 1090 Vienna, Austria
E-mail: kurt.tutschku@univie.ac.at

Akihiro Nakao

University of Tokyo,
Graduate School of Interdisciplinary Information Studies
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
National Institute for Communication and Information Technology,
1-33-16, Hakusan, Bunkyo-ku, Tokyo, 113-0001, Japan.
E-mail: nakao@iii.u-tokyo.ac.jp

Abstract:

From the viewpoint of communication networks *Network Virtualization (NV)* extends beyond pure operational issues and addresses many impasses of the current Internet. The idea of *Transport Virtualization (TV)* progresses the capabilities of NV and enables the independence from a specific network transport resource. The independence is achieved by pooling multiple transport resources and selecting the best resources for exclusive or concurrent use. However, the application and selection of concurrent paths is rather complex and introduces inevitable packet re-ordering due to different stochastic delay characteristics on the used paths. Packets arriving at the destination out-of-order have to be stored in a re-sequencing buffer before reassembled packets are forwarded to the application. We provide a simulation framework based on discrete event simulation which allows an evaluation of the re-sequencing buffer occupancy. Further, we perform an analysis of the fundamental behaviors and factors for packet re-ordering in concurrent multipath transmissions.

Keywords: Network Virtualization, Transport Virtualization, Simulative Performance Evaluation, Concurrent Multipath Transmission, Re-sequencing Buffer

Reference to this paper should be made as follows: Zinner T., Tutschku K., Nakao A. and Tran-Gia, P. (xxxx) 'Performance Evaluation of Packet Reordering on Concurrent Multipath Transmissions for Transport Virtualization', *Int. J. Communication Networks and Distributed Systems*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Thomas Zinner studied computer science and physics at the University of Wuerzburg, Germany. He received his diploma degree in computer science in 2007. Since then he has been a researcher at the Institute of Computer Science and pursuing his PhD. His current research focuses on Quality of Experience for Video Streaming - especially scalable video codecs - in combination with performance evaluation and network virtualization techniques.

Kurt Tutschku holds the Chair of "Future Communication" (endowed by Telekom Austria) at the University of Vienna since September 2008. Before that, he was an Expert Researcher at the NICT Tokyo, Japan. Until December 2007 he was an Assistant Professor at the Department of Distributed Systems, University of Wuerzburg. There he led the department's group on Future Network Architectures and Network Management. He received his diploma and doctoral degree in Computer Science from University of Wuerzburg in 1994 and 1999 respectively and completed his Habilitation ("'State Doctoral Degree"') at the University of Wuerzburg in 2008. His main research interests include future generation communication networks, Quality-of-Experience, and the modeling and performance evaluation of future network control mechanisms and services in the emerging Future Internet, particular of P2P overlay networks.

Akihiro Nakao received his B.S.(1991) in Physics, M.E.(1994) in Information Engineering from the University of Tokyo. He was at IBM Yamato Laboratory, at Tokyo Research Laboratory and at IBM Texas Austin from 1994 till 2005. He received M.S.(2001) and Ph.D.(2005) in Computer Science from Princeton University. He has been teaching as an Associate Professor in Applied Computer Science, at Interfaculty Initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, the University of Tokyo since 2005. He has also been an expert visiting scholar/a project leader at National Institute of Information and Communications Technology (NICT) since 2007.

Phuoc Tran-Gia is professor at the Institute of Computer Science at the University of Wuerzburg, Germany. Previously he was at academia in Stuttgart, Siegen (Germany) as well as industry at Alcatel (software development System 12), IBM Zurich Research Laboratory (Zurich, Switzerland, architecture and performance evaluation of communication networks). He is consultant and cooperation project leader with Siemens (ICN Board, Munich, ICM Berlin), Nortel (Texas), T-Mobile International (Bonn), France Telecom (Belfort), European Union (European Science Foundation, Brussels), and is coordinating the G-Lab project National Platform for Future Internet Studies.

1 Introduction

Today, data transport is achieved mainly by the Internet Protocol (IP). Its routing feature is used to accomplish scalability in the connection of local subnets and end systems. The IP protocol assumes that the interconnecting nodes (i.e., the routers) and the actual transport resources (i.e., the links) are stable and change only in case of failures. The Future Internet will consist of interconnected subnets and interconnecting nodes as well. However, the availability of these resources is expected to be provided for a certain lifetime and their capabilities are highly variable. Networks that are based on the model of temporal *leasing* of variable resources are referred to as *federated networks*, cf. Peterson et al. (2009*b*), and technologies that enable the safe sharing of such network resources are denoted as Network *Virtualization (NV)* mechanisms.

A major task when creating federated networks for data transport is the selection of resources. The selection has to consider the temporal availability of the resources. Therefore, a measurement-based scheme for selecting these resources is needed. This scheme should adapt the federated transport network to the variable demand for resources as well as to the currently available resources. Of course, the selection scheme has to scale sufficiently such that it can be applied in large networks.

The concept of *Transport Virtualization* (TV) enhances the capabilities of future networks. Tutschku et al. (2009*b*) introduced TV as an alternative mode of Network Virtualization. While NV typically facilitates the sharing of resources, TV creates virtual resources (e.g., virtual links) based on the aggregation of resources. The simplest form of TV is achieved by collecting multiple transport resources (even from different virtual networks or providers) and selecting the best resources for exclusive or concurrent use. The type of the resources used for aggregation does not matter; they can be either of physical or of virtual nature.

The use of *concurrent multipath (CMP)* transmissions, i.e., of parallel usage of different paths, will bring exceptional advantages to networks, such as higher throughput and increased resilience. However, CMP will introduce additional complexity which has to be understood. First, CMP transmission will inevitably introduce out-of-order packets due to different stochastic packet delay characteristics on the paths. The re-ordering can be compensated by buffering at the destination, possibly leading to increased end-to-end delay but still being transparent to the transport protocol. Second, the different stochastic delay processes on the paths can amend each other in their negative effects on out-of-order packets. Third, the strength and occurrences of these combination effects are highly non-intuitive.

Thus, an analysis of the fundamental behaviors and factors for packet re-ordering in CMP transmissions has to be carried out and is presented in this paper. The analysis applies analytical techniques as well as event-based simulations due to the complexity of the CMP mechanisms. Therefore, we also discuss how to improve event-based simulation in order to achieve correct insights.

The paper is structured as follows. First, we will briefly explain the idea of Transport Virtualization. Then, we detail how to implement TV using a CMP transmission mechanism and discuss the path selection for such a mechanism. After that, we introduce the CMP mechanism and its packet handling. Furthermore, we outline the analytical and simulative performance models used in this study. In particular, we discuss how to model and simulate packet delay realistically, i.e., packet re-ordering occurs not on a path. Finally, we provide a case study on the *re-sequencing buffer occupancy probability*

distribution for different path delays. The investigation of this probability distribution gives insights in how to select the set of paths used in a TV mechanism using CMP transfer. The paper is concluded by a brief summary.

2 Transport Virtualization

One of the main benefits of virtualization techniques is the abstraction of computer resources. An operating system, for example, may provide virtual memory to an application program. This memory gives the program the feeling that it can use a contiguous working memory, while in fact it may be physically fragmented and may even overflow on to disk storage. The actual location of the data in the physical memory does not matter and is hidden. Thus, this virtualization technique makes the resource ”memory” independent from its physical ”location”.

The idea of Transport Virtualization extends the concept of NV by transferring the feature of location independence to transport resources. TV is motivated partly by the abstraction introduced in P2P-based content distribution networks (CDNs). Advanced P2P CDNs, such as eDonkey or BitTorrent, apply the concept of *multi-source download (MSD)* where different peers are *pooled*. Upon a download, a peer receives multiple parts of a file in parallel from different peers. As a result, the downloading peer does not rely any more on a single peer which provides the data, and the reliability is increased. The providing peers are typically selected such that the throughput is optimized. An appealing feature of the MSD concept is that the actual physical location of the file does not matter. Thus, these P2P CDNs can be viewed as an abstract and almost infinite storage for the data files. Thus, an abstraction of a storage resource, similar to the example of virtual memory, is achieved.

The above outlined concept of a virtual storage resource is now transferred to the area of data transport. *Transport Virtualization* can be viewed as an abstraction concept for data transport resources. Hereby, the physical location of the transport resource does not matter as long as this resource is accessible. In TV an abstract data transport resource can be combined from one or more physical or overlay data transport resources. Such a resource can be, e.g., a leased line, a wavelength path, an overlay link, or an IP forwarding capability to a certain destination. These resources can be used preclusive or concurrently and can be located even in different physical networks or administrative domains. Thus, an abstract transport resource exhibits again the feature of location independence.

3 Implementing TV using Concurrent Multipath Transfer in Advanced Routing Overlays

A scalable approach for routing overlays is the concept of *one-hop source routing* which is presented by Gummadi et al. (2004), and is implemented by the *SORA architecture*, cf. Lane & Nakao (2007). Hereby, the user data is forwarded to a specific intermediate node, denoted as SORA router, which then relays the traffic to its destination using ordinary IP routing. The details of this architecture is discussed by Lane & Nakao (2007, 2008) and Tutschku et al. (2009*a*).

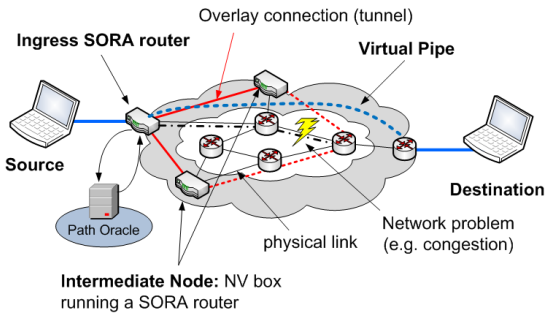


Figure 1 Simplified SORA Architecture

The TV in the considered SORA architecture is achieved by a concurrent multipath transfer mechanism. The mechanism combines multiple overlay paths (even from different overlays) into a single virtual high-capacity pipe. The combined paths are used in parallel by sending data packets concurrently on different overlay paths. This principle is also know as *striping*.

The SORA architecture itself is depicted in Figure 1. The paths which form the virtual pipe are chosen by the ingress router out of large number of potential paths, cf. Lane & Nakao (2008), Tutschku et al. (2009*a*). The *path oracle* discovers the available paths, i.e. the components of the abstract data transport resource. Current discussions, e.g., by Vinay et al. (2007), suggest that a path oracle can be provided by the network operator or by other institutions. Altogether, the CMP mechanism combined with the path oracle facilitates an abstraction of a data transport resource. Instead of using a single fixed data transport resource, the system relies now on location independent, multiple and varying resources.

An important question for TV is the selection of pooled resources, i.e., the selection of potential paths. Typically, a *good* path has a short transmission delay, like discussed by Lane & Nakao (2008), and as a result, the mean path delay is an initial candidate as selection criterion. However, the selection of concurrent paths is rather complex. CMP transmission will inevitably introduce packet re-ordering due to different stochastic packet delay characteristics on the different paths, stochastic delay processes can amend each other in their negative effects (see Section 5), and moreover, the strength and occurrences of these combination effects are highly non-intuitive. Therefore, a simple stochastic characterization of a path, such as the mean delay, is not expected to be sufficient. A path can exhibit a highly varying delay. The delay is composed of the propagation delay as well as queuing and processing delays at each router along the path. The path delay is therefore a complex stochastic process which can be measured and, based on the measurements, a delay distribution can be computed. We will not discuss how to obtain and how to measure such a delay distribution in this paper. Instead, we want to describe the influence of different delay distributions on TV when multiple paths are used.

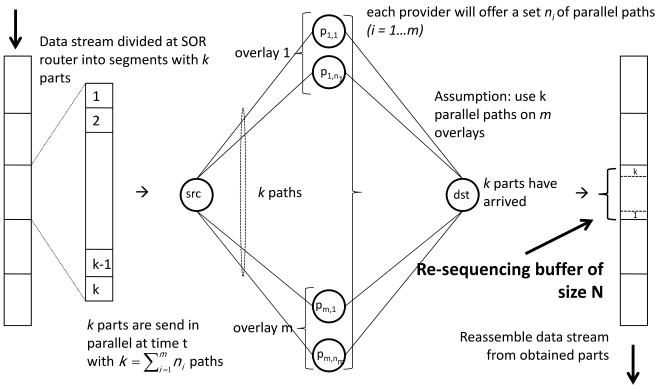


Figure 2 Transmission Mechanism

4 Mechanisms and Performance Models

Next, we will outline the performance models for the suggested striping mechanism for CMP transmission in TV. We start with the striping mechanism itself and then detail the analytical and simulative performance models. In particular, we discuss how to model and simulate packet delay realistically, i.e., without packet re-ordering on a path.

4.1 Striping Mechanism

Figure 2 shows a detailed model of the striping mechanism suggested for use in CMP transmission for TV. The data stream is divided into segments which are split into k smaller parts. The k parts are transmitted by the set of paths, i.e., in parallel on k different overlay links. The receiving router reassembles these parts. The parts can arrive at the receiving router after different time intervals since they experience varying delays. Therefore, it is possible that parts arrive ”out-of-order”. It should be noted that part re-ordering can only happen between different paths. The order of parts on a path is maintained since packets typically cannot overtake each other.

Part or packet re-ordering due to multipath transmissions may have a severe impact on the application performance. In order to reduce this effect, the receiving router maintains a finite re-sequencing buffer. However, when the re-sequencing buffer is filled and the receiving router is still waiting for parts, part loss can still occur. This loss of parts is harmful for the application and should be minimized. Therefore, an important objective in the operation of the system is to minimize the re-sequencing buffer occupancy. This can be achieved by a selection of paths with appropriate delay characteristics.

4.2 Analytical Model

Initial analytical and approximative methods to estimate the re-sequencing buffer occupancy in case of multipath downloads or transmissions already exist in literature. A first analytical model can be found in Nebat & Sidi (2006). The suggested performance model is already very powerful and allows the computation of the re-sequencing

buffer occupancy. However, its only applicable in a limited set of scenarios. A deep investigation of different scheduling mechanisms, adapted to the path delay for instance, is not feasible. Further, this model allows only a loose lower bound approximation of the end-to-end delay. To overcome these shortcomings we present a simulation framework which allows an investigation of the resequencing buffer occupancy. The model can easily be extended in order to examine scheduling mechanisms or the end-to-end delay.

4.3 Simulation Model

The behavior of the re-sequencing buffer occupancy is investigated by time discrete, event based simulation. The simulation model assumes a continuous data stream. The stream is divided into parts which are sent in parallel on several paths. In our evaluation we consider transmissions over either two or three paths. The delay on the paths is modeled by discrete delay distributions with a resolution of one time unit. A packet is transmitted every time unit on a path.

In order to achieve a realistic behavior of different path delays on the resequencing buffer, we have to ensure that no packet re-ordering on a single path can occur. This means that whenever a random delay for a path is generated, the previous delay has to be considered within the generation algorithm. Hence, a current path delay is at least as large or equal than a previous path delay minus the interim time between the previous and the current packet. Furthermore, the relative frequency of all delays on a path has to converge against the given delay distribution for that path. We will discuss in the next subsection how we can generate packet delay distributions which fulfill these requirements.

4.4 Modeling Packet Delay

Before we describe the suggested packet delay model, we outline the impact of an inappropriate delay model on the buffer occupancy.

In order to show this effect, we investigate the influence of a concurrent transmission over two paths with equal bandwidth. We assume that both paths have the same delay distribution, which is set to a truncated gaussian-like distribution, each with mean value $\mu = 50$ time units and a standard deviation $\sigma = 20$ time units. The duration of the simulation run is 1 million time units, and we conducted 5 runs with different seeds. We investigate two scenarios: a) the delay distribution allows ”no packet re-ordering on a single path” and b) the delay distribution permits ”packet re-ordering on a single path”. Thus, we are able to investigate the impact of these scenarios with equal path delay distributions on the re-sequencing buffer occupancy.

Fig. 3 shows the re-sequencing buffer occupancy distribution for the two scenarios. The distribution of the buffer occupancy differs significantly for scenario a) dependent (”no packet re-ordering on a single path”) and b) independent packet delays (”packet re-ordering on a single path”). For an independent random delay (”packet re-ordering on a single path”), the re-sequencing buffer occupancy for the given input distributions follows a gaussian distribution. On the contrary, in case of a dependent random number generation, the re-sequencing buffer occupancy is much smaller. We can conclude that with independent packet delays, the buffer occupancy will be significantly higher than in case of dependent packet delays. Hence, packet re-ordering on a single path has to be avoided in order to get accurate meaningful results of the buffer occupancy.

Now we explain how we generate a dependent packet delay distribution. The applied path delay model is based on the delay model introduced in Nebat & Sidi (2006) and extended for the use in simulations. First, we describe the basic necessary conditions for the delay distribution to ensure that packets arrive at their destination in the same order they were transmitted. The complete derivation can be found in Nebat & Sidi (2006). The following notations are used:

- d_i: the delay experienced by packet i
- Δ_i^t: the inter-departure time for packets $i-1$ and i

Consequently, there are only two restrictions for the experienced delay of packet i:

1. $d_i \geq 0$, this means that delay is always a positive value, and
2. $d_i \geq d_{i-1} - \Delta_i^t$, denoting the fact, that packet i can not overtake packet $i-1$. Nevertheless, they may arrive at the destination simultaneously.

Therefore, d_i can be any probabilistic function $f(d_{i-1}, \Delta_i^t)$ that satisfies

$$f(d_{i-1}, \Delta_i^t) \geq \begin{cases} d_{i-1} - \Delta_i^t, d_{i-1} \geq \Delta_i^t \\ 0 \qquad d_{i-1} < \Delta_i^t \end{cases} \tag{1}$$

To ensure the delay does not diverge, we need a stability constraint on $f(d_{i-1}, \Delta_i^t)$.

For the special case where packets are transmitted at a constant rate, i.e., every k time units we have a constant inter-departure time $\Delta_i^t = l \ \forall i$, (1) becomes

$$f(d_{i-1}, \Delta_i^t) = f(d_{i-1}) \geq \begin{cases} d_{i-1} - l, d_{i-1} \geq l \\ 0 \qquad d_{i-1} < l \end{cases}.$$

The introduction of l is an extension to the model presented in Nebat & Sidi (2006) and enables us to adjust the path delay in a higher resolution.

For simplicity, let us assume that the delay is an integer value expressed as multiples of a time unit. Consequently, $f(d_{i-1})$ can be any integer that satisfies

$$d_i = f(d_{i-1}) \geq \begin{cases} d_{i-1} - l, d_{i-1} > l \\ 0 \qquad d_{i-1} \leq l \end{cases} \tag{2}$$

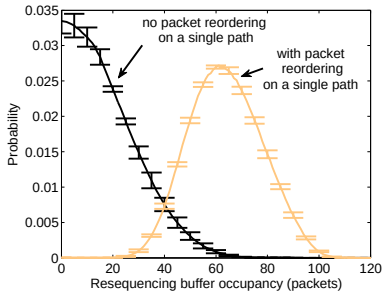


Figure 3 Buffer occupancy with and without packet re-ordering within a path on a 99.9% confidence level

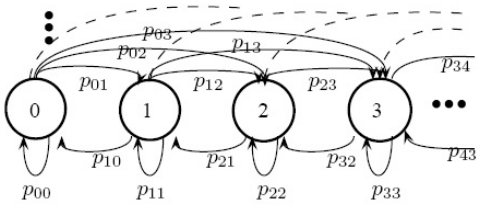


Figure 4 Markov-chain representing the delay, $l = 1$

Any probabilistic function that corresponds to f in (2) can be represented by a Markov-chain, similar to Fig. 4, which is an example for $l = 1$, i.e., packets are sent every time unit. Here, state i corresponds to delay of i time units and the arrows correspond to the transitions among states with the respective probabilities. These transition probabilities can be written as a transition probability matrix P consisting of the elements $p_{i,j}$. For a finite maximum delay d_n, the transition probability matrix can be written as:

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \cdots & p_{0,n-1} & p_{0,n} \\ p_{1,0} & p_{1,1} & p_{1,2} & \cdots & p_{1,n-1} & p_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{l,0} & p_{l,1} & p_{l,2} & \cdots & p_{l,n-1} & p_{l,n} \\ 0 & p_{l+1,1} & p_{l+1,2} & \cdots & p_{l+1,n-1} & p_{l+1,n} \\ 0 & 0 & p_{l+2,2} & \cdots & p_{l+2,n-1} & p_{l+2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & p_{n,n-1} & p_{n,n} \end{pmatrix}.$$

The matrix P consists of four parts. On the main diagonal are the probabilities to remain in the current delay state. On the right side of the main diagonal are the transition probabilities which increase the current delay state. The left side of the main diagonal illustrates a decrease in the delay state. The maximum decrease in the delay state depends on the interdeparture time. Thus, not all states smaller than the current state can be reached. This is expressed by $p_{i,j} = 0 \; \forall i < j - l - 1$.

For a given delay distribution d, we have to solve the fix-point equations

$$d = d \cdot P, \tag{3}$$

so that the resulting transition probability matrix, and thus the represented Markov-chain is irreducible and aperiodic. That way, we ensure that the delay process is recurrent. The transition probability matrix P can be used in our simulation model to assure that 1) the packet delay follows the given delay distribution d and that 2) no packet re-ordering occurs by a transmission on a single path. Before we can use P, we still have to solve Equation 3, which is described in the next subsection.

4.5 Computing the Transition Matrix by Using Linear Programming

In this subsection we describe our approach to determine a transition probability matrix P which fulfills the fix-point Equation 3. This equation is under-determined, i.e., we have to determine $\frac{1}{2} \cdot (n^2 + 2nl - l^2 + n - l)$ parameters with $n+1$ equations, whereas $n+1$ is the size of the delay state space d. For that, we model the problem as a Linear Program (LP) and solve this program with ILOG CPLEX, cf. IBM (2009). In addition to the previously introduced variables, we define the following variables:

- c_1: vector with lower bounds for values on the main diagonal of matrix P
- c_2: vector with upper bounds for values on the main diagonal of matrix P

Algorithm 1 Determine the transition matrix

Maximize

$$f(P) = \sum_{i=0}^{n} \sum_{j=0}^{n} p_{i,j} \tag{4}$$

Subject to

$$\sum_{i=0}^{n} p_{i,j} \cdot x_i = x_j \ \ \forall j; \tag{5}$$

$$\sum_{j=0}^{n} p_{i,j} = 1 \ \ \forall i \tag{6}$$

$$p_{i,j} = 0, \ \ i < j - l - 1; \tag{7}$$

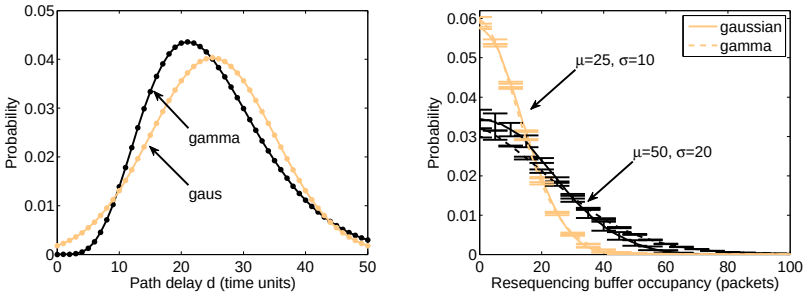
$$0 \ < \ p_{i,j} < 1, \ \ j - l < i < j \tag{8}$$

$$c_1 \ < \ p_{i,j} < c_2, \ \ i = j \tag{9}$$

$$0 \ < \ p_{i,j} < 1, \ \ i > j \tag{10}$$

The LP depicted by Algorithm 1 aims to compute the transition probability matrix. This is expressed by Equation 4. The constraints on the variables are described in the following:

- Equation 5 describes the $n+1$ fix point Equations derived from Equation 3.
- The sum of each line in P has to satisfy the normalizing condition, i.e., the sum over all probabilities is equal to 1, which is expressed by Equation 6.
- Equation 7 illustrate the transition probabilities in the lower left part of P. These probabilities have to be zero in order to ensure that the next delay value is not smaller than the current delay minus the inter-departure time at the source.
- The probabilities between the main diagonal and the zero values are depicted by Inequalities 8. These probabilities denote a slow delay decrease of an inter-departure time. Thus, we can assure no packet re-ordering.



(a) Gaussian and gamma distributions

(b) Re-sequencing buffer occupancy for similar path delay distributions

Figure 5 Similar path delay distributions

- The main diagonal of the matrix, depicted by the Inequalities 9, indicate that the delay remains constant. In order to avoid the trivial solution of the problem, the identity matrix I, these probabilities have to be smaller than 1, which is expressed by c_2.
- An increase of the delay between the departure of two packets is illustrated by the Inequalities 10.

5 Re-sequencing Buffer Occupancy - A Case Study

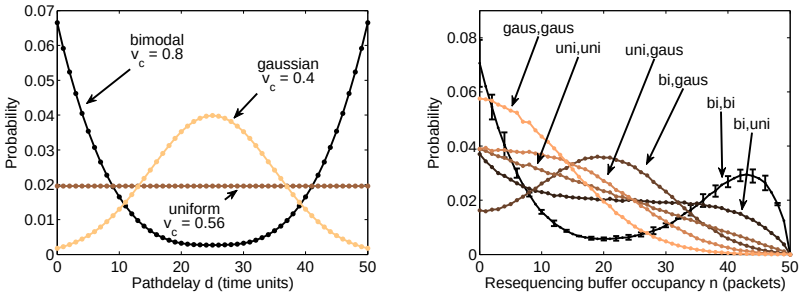
In this case study, we want to investigate the resequencing buffer occupancy for different path delay distributions. The results presented next are of theoretical nature since they are obtained by the consideration of abstract models. However, they are intended to give a deeper insight into the practical question of how to select appropriate paths.

First, we will examine the system behavior in case of similar path delay distributions with equal mean values and equal standard deviations. After that, we investigate diverse path delay distributions, i.e., delay distributions with equal mean delay values but different standard deviations. In Subsection 5.3 we examine the impact of different path delay parameters for one distribution type and show results for transmissions via two and three concurrent paths. A short validation of the simulation model is given in Subsection 5.4. For that we use an analytical model, cf. Nebat & Sidi (2006), originally developed for parallel multi-download.

5.1 Robustness of the Buffer Occupancy in Case of Similar Path Delay Distributions

As similar distribution types, we choose a truncated gaussian (labeled *gaus*) and a truncated gamma (*gamma*) distribution. The probability mass functions are depicted in Figure 5(a).

The mean delay and the standard deviation value for each of the depicted distributions are respectively $E[d] = \mu = 50$ and $\sigma = 20$. For these distributions, we also investigate value pairs of $\mu = 50, \sigma = 10$ and $\mu = 25, \sigma = 10$. The delay values range



(a) Path delay distributions with equal mean delay and different variations

(b) Re-sequencing buffer occupancy in case of nonsimilar path delay distributions

Figure 6 Nonsimilar path delay distributions

between $d_{min} = 0$ to $d_{max} = 100$. We choose these distributions in order to evaluate the system behavior under similar conditions. We start with the investigation of a transmission over two equal paths and compare two similar path delay distribution types. The results of these experiments are illustrated on a 95% confidence level as Probability Mass Functions (PMFs) in Figure 5(b). We used $\mu = 50, \sigma = 20$ and $\mu = 25, \sigma = 10$ as parameters for the delay distributions. For both parameter sets, the difference between two equal gamma distributed path delays and two gaussian distributed path delays are negligible. The difference can be explained by the influence of higher statistical moments than the variance, like skewness and kurtosis for instance. We can conclude that these similar distribution types have, for equal average and variance, a similar influence on the buffer occupancy. Therefore, the buffer occupancy shows a good robustness, i.e. invariance, within this family of packet delay distributions. Hence, the distribution type with least computation requirements can be used.

5.2 Buffer Occupancy in Case of Diverse Path Delay Distributions

Now, we want to examine the resequencing buffer in case of different types of delay distributions. Therefore, three different path delay distributions are considered for the paths: a truncated gaussian (label *gaus*), a uniform (*uni*), and a bimodal distribution (*bi*). The PMFs of the distributions are depicted in Figure 6(a). The mean delay for each distribution is $\mu = 25$, whereas the delay ranges from $d_{min} = 0$ to $d_{max} = 50$. The coefficient of variation c_v varies between $c_v = 0.4$ for the gaussian distribution and $c_v = 0.8$ for the bimodal distribution. We decided to investigate these distributions in order to evaluate the system behavior under highly different conditions, e.g., gaussian vs. bimodal delay. We conduct an investigation of two concurrent paths. The buffer occupancies for different combinations of delay distributions are depicted in Figure 6(b). The y-axis denotes the probability of the packets stored in the re-sequencing buffer, assigned on the x-axis. For the sake of clarity we plotted only the (bi,bi) buffer occupancy distribution with confidence intervals for a confidence level of 95%.

For the case of two gaussian delay distributions, the buffer occupancy is left leaning and higher buffer occupancies are not very likely. However, for two bimodal delay distributions a large fraction of the probability mass covers a buffer occupancy bigger than 30 packets. Thus, we can conclude that paths with bimodal delay should be avoided

since the probability for a heavy loaded re-sequencing buffer is higher. It should be noted that the maximum buffer occupancy in the investigated scenario is $o_{max} = 50$.

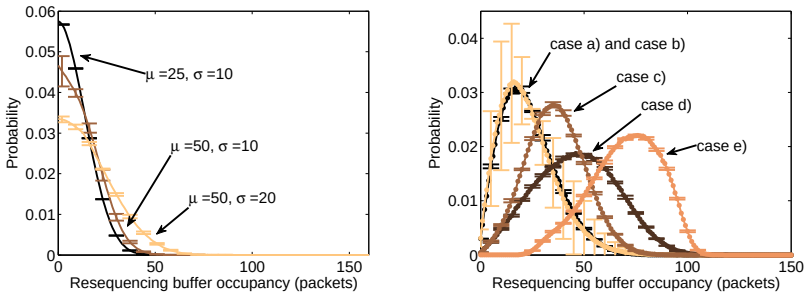
5.3 Impact of Different Path Delay Parameters on the Buffer Occupancy and Insights to Path Selection

In this study, we investigate the criteria for path selection. Intuitively a path selection algorithm should select those paths which provide the shortest delay. Typically this is important for traffic from an interactive application with realtime constraints like video streaming or VoIP. For the investigation, we consider truncated gaussian-like delay distributions with different mean packet delays $\mu = 50$ and $\mu = 25$ and different standard deviations $\sigma = 20$, $\sigma = 10$, and $\sigma = 5$. We start with a concurrent transmission over two paths. The influence on the resequencing buffer is depicted on a $95\%-$confidence level in Figure 7(a) as PMF. It can be seen that in case of $\sigma = 10$ the buffer occupancy is almost independent from the mean value. For $\sigma = 20$ the distribution of the buffer occupancy gets lower and expands comparing to $\sigma = 10$. We can conclude that in case of a transmission over two paths with equal distributions the buffer occupancy depends mainly on the standard deviation.

In a next step, we investigate the system behavior in case of a transmission over three paths. The results for five different scenarios are depicted as PMF on a $95\%-$confidence level in Figure 7(b). The scenarios are a) all paths with same parameters $\mu = 50$, $\sigma = 10$, b) all paths with same parameters $\mu = 25$, $\sigma = 10$, c) two paths with parameters $\mu = 50,\ \sigma = 10$, one path with $\mu = 25,\ \sigma = 10$, d) the reverse of c), and e), similar to d), but two paths with parameters $\mu = 25,\ \sigma = 5$.

A close look at Figure 7(b) reveals that a) and b) have similar buffer distributions. This indicates that the pure delay has no impact on the buffer occupancy, which has already been shown for the corresponding two path buffer occupancy. For the cases c) and d), it is remarkable that case c), with two high path delays, performs better in terms of buffer occupancy than case d). Intuitively, more paths with lower delay should result in a better performance, but this is obviously not true as shown in case d).

Let us consider this behavior in detail. We assume a single packet from the high delay path which is effectively much overdue. Until the arrival of this overdue packet, the low delay paths can easily increase the occupancy of the re-sequencing buffer. Thus,



(a) Different normal distributed path delays

(b) Re-sequencing buffer occupancy in case of 3 concurrent paths

Figure 7 Case study for different normal distributed path delays

the buffer can be filled quickly by the low delay paths. This example shows that the high delay path becomes more dominant over low delays paths in terms of buffer occupancy. The selection of the paths should level the variation of the range of mean delays. Recent findings in VoIP systems, cf. Hoßfeld et al. (2008), show that a constant delay has no significant impact on the Quality of Experience of the application, as long as the delay is below a certain threshold. In such a case, the user does not notice whether the transmission is conducted via a path with a high or a low delay. Thus, it might be better in TV to choose a path with a higher mean delay in order to relieve the resequencing buffer and avoid packet loss.

Finally, for the cases d) and e) we can see that in case e) the system performs worse in terms of buffer occupancy than in case d). Here the standard deviation on the smaller paths is lower compared to case d) and thus high and low delays are unlikely.

Figure 7(b) gives also deep insights into path selection strategies for CMP transfer. This is of particular interest when the selection is done in federated networks since the selection has to be done carefully on small time scales. The results show, that a naive selection of two paths with low path delay, i.e. case d) is not the best choice in terms of suppressing packet re-ordering. Less re-ordering occurs in case c), when the average path delay for the third path is leveled.

5.4 Comparison of Simulative and Analytical Model

In order to validate the simulative model we use the analytical model presented by Nebat & Sidi (2006). The analytical model was developed to describe the influence of multi-source downloads on the re-sequencing buffer. Thus, it can be easily enhanced to the presented concurrent multipath transfer architecture. The analytical and the simulative model are compared in Figure 5.4 for a transmission over two concurrent paths. For that, we used selected path delay distribution pairs presented in Figure 6(a) and Figure 6(b). The x-axis denotes the re-sequencing buffer occupancy, the y-axis the corresponding probability. It can be seen that the results for the re-sequencing buffer occupancy are similar. The simulative results on a confidence level of 95% always contain the results computed with the analytical model. Thus, we can conclude that our simulative approach is valid and that we can use it for a deeper analysis of the re-sequencing buffer occupancy.

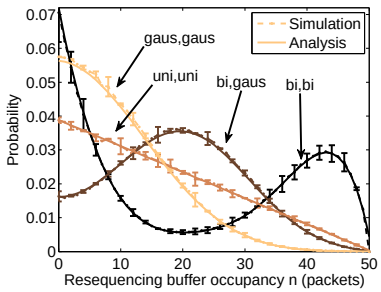


Figure 8 Comparison of analytical and simulative model

6 Related Work

Multipath transmission mechanisms have been suggested for IP networks for a while. The Stream Control Transmission Protocol (SCTP) as presented by Ong & Yoakum (2002) enables multi-homing but does not facilitate CMP transfer and its advantages. Iyengar et al. (2006) investigated that typical problems with packet reordering occur. While Iyengar et al. (2006) proposed to enhance SCTP such that it can react to packet re-ordering, this paper presents a mechanism to suppress reordering as far as possible. As a result, transport protocols like SCTP or TCP do not have to be modified. The *pTCP* proposal from Hsieh & Sivakumar (2002) and the *multiPath TCP* proposal from Shakkottai et al. (2006) exploit concurrent transmissions. However, they focus on flow control and the coordination among flows. Our work complements these studies since we are aiming at the selection of paths.

Path selection is often investigated on network layer in context of multipath routing, cf. Gojmerac et al. (2008). We amend these works by considering the transmission mechanism on transport layer.

A major work, which addresses the network and transport layer concurrently, is the DaVinci architecture, presented by He et al. (2008). In DaVinci, the paths are selected such that the creation of bottlenecks is avoided. The investigation in our paper goes beyond the results presented in the DaVinci architecture. The DaVinci proposal considers mainly the mean path delay. However, as we have seen in Section 5, this assumption might be not sufficient. Our paper discusses the fundamental performance issues in selecting paths according to their detailed statistical characteristics.

Finally, three efforts of IETF working groups (WGs) should be mentioned. The Application-Layer Traffic Optimization (ALTO) WG, cf. Peterson et al. (2009*a*), aims at providing P2P CDNs with information to perform better-than-random initial peer selection. This selection protocol might be extended to path selection. The Low Extra Delay Background Transport (LEDBAT) WG, cf. Shalunov & Sridhavan (2009), investigates multipath TCP connections in order to better saturate bottlenecks. At last, Transport Area WG discusses currently the combination of BGP and multipath TCP, cf. Polk & Fairhurst (2009).

7 Conclusion

In this paper we explain the idea of *Transport Virtualization (TV)*, which provides a location independent abstraction for data transport resources and outlined the TV concept by the example of *concurrent multipath (CMP)* transmission in one-hop source routing overlays. CMP transport has many appealing advantages such as higher throughput and increased resilience. However, its application increases also the complexity of the system. In particular, the use of concurrent paths introduces inevitable out-of-order packet delivery.

We discuss an important performance issue of CMP transmission, the *re-sequencing buffer occupancy probability distribution* under the influence of the delay distribution on the used paths in the CMP mechanism. Due to the complexity of this mechanism, we combine analytical and simulative techniques in order to investigate the re-sequencing buffer occupancy.

The simulation of the mechanism has to ensure that no overtaking of packets occurs on a path. Therefore, the simulation has to model this behavior by generating an appropriate delay time series. We ensure this for given delay distributions by solving a fix-point equation using Linear Programming. An algorithm how to do this is described in this paper.

With respect to the CMP transmission, it turns out that different stochastic delay processes can amend each other in their negative effects on the packet reordering, leading to a higher re-sequencing buffer occupancy. Also, the strength and occurrences of these combination effects are highly variable.

The investigation gives also deep insights into path selection strategies for CMP transfer. This is of particular interest in federated networks, when path selection on small time scales is considered. The results show that a naive selection of using the paths with the lowest delay is not the best choice in terms of suppressing packet re-ordering. Less re-ordering occurs when the variation path delay for the third path is leveled.

Future work will be devoted to measurement-based validation of the presented performance model for CMP transfer. In addition, these measurements are needed to identify the time accuracy of the delay distributions and the time scales required for the path selection.

Acknowledgments

The authors would like to thank Dirk Staehle and Tobias Hoßfeld for the fruitful discussions and their support during the course of this work. The sponsorship of this research by the European FP7 Network of Excellence "Euro-NF" through the Specific Joint Research Project "Multi-Next" is thankfully acknowledged.

References

Gojmerac, I., Reichl, P. & Jansen, L. (2008), 'Towards low-complexity internet traffic engineering: The adaptive multi-path algorithm', *Computer Networks: The International Journal of Computer and Telecommunications Networking* **52**(15), 2894–2907.

Gummadi, K. P., Madhyastha, H. V., Gribble, S. D., Levy, H. M. & Wetherall, D. (2004), Improving the reliability of internet paths with one-hop source routing, *in* 'OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation', USENIX Association, Berkeley, CA, USA, pp. 13–13.

He, J., Zhang-Shen, R., Li, Y., Lee, C.-Y., Rexford, J. & Chiang, M. (2008), Davinci: dynamically adaptive virtual networks for a customized internet, *in* 'CONEXT '08: Proceedings of the 2008 ACM CoNEXT Conference', ACM, New York, NY, USA, pp. 1–12.

Hoßfeld, T., Hock, D., Tran-Gia, P., Tutschku, K. & Fiedler, M. (2008), Testing the IQX hypothesis for exponential interdependency between qos and qoe of voice codecs iLBC and g.711, *in* '18th ITC Specialist Seminar on Quality of Experience', Karlskrona, Sweden.

Hsieh, H.-Y. & Sivakumar, R. (2002), pTCP: An end-to-end transport layer protocol for striped connections, *in* 'ICNP '02: Proceedings of the 10th IEEE International Conference on Network Protocols', IEEE Computer Society, Washington, DC, USA, pp. 24–33.

IBM (2009), 'Ilog cplex', http://www.ilog.com/products/cplex/.

Iyengar, J., Amer, P. D. & Stewart, R. (2006), 'Concurrent multipath transfer using sctp multihoming over independent end-to-end paths', *IEEE/ACM Transactions on Networking* **14**(5), 951–964.

Lane, J. R. & Nakao, A. (2007), SORA: A shared overlay routing architecture, *in* 'Proceedings of the 2nd International Workshop on Real Overlays And Distributed Systems (ROADS)', ACM, Warsaw, Poland.

Lane, J. R. & Nakao, A. (2008), Best-effort network layer packet reordering in support of multipath overlay packet dispersion, *in* 'IEEE GLOBECOM 2008: Global Telecommunications Conference, 2008', IEEE, pp. 2457–2462.

Nebat, Y. & Sidi, M. (2006), 'Parallel downloads for streaming applications: a resequencing analysis', *Performance Evaluation* **63**(1), 15–35.

Ong, L. & Yoakum, J. (2002), 'RFC 3286: an introduction to the stream control transmission protocol (SCTP)', http://www.ietf.org/rfc/rfc3286.txt.

Peterson, J., Gurbani, V. & Marocco, E. (2009*a*), 'Charter of the application-layer traffic optimization (alto) working group', http://www.ietf.org/proceedings/75/alto.html.

Peterson, L., Sevinc, S., Lepreau, J., Ricci, R., Wrocalwski, J., Faber, T., Schwab, S. & Baker, S. (2009*b*), 'Slice-Based Facility Architecture', http://svn.planet-lab.org/attachment/wiki/GeniWrapper/sfa.pdf.

Polk, J. & Fairhurst, G. (2009), 'Charter of the transport area working group (TSVWG)', http://www.ietf.org/dyn/wg/charter/tsvwg-charter.html.

Shakkottai, H. H. S., Hollot, C. V., Srikant, R. & Towsley, D. (2006), 'Multi-path TCP: A joint congestion control and routing scheme to exploit path diversity on the internet', *IEEE/ACM Transactions on Networking* **14**, 1260–1271.

Shalunov, S. & Sridhavan, M. (2009), 'Charter of the low extra delay background transport (ledbat) working group'.

Tutschku, K., Zinner, T., Nakao, A. & Tran-Gia, P. (2009*a*), 'Network virtualization: Implementation steps towards the future internet', *ECEASST*.

Tutschku, K., Zinner, T., Nakao, A. & Tran-Gia, P. (2009*b*), Re-sequencing buffer occupancy of a concurrent multipath transmission mechanism for transport system virtualization, *in* 'Proc. of the 16. Kommunikation in verteilten Systemen 2009 (KiVS 2009)', Kassel.

Vinay, A., Feldmann, A. & Scheideler, C. (2007), 'Can ISPs and P2P systems co-operate for improved performance?', *SIGCOMM Computer Communication Review* **37**(3), 29–40.