MINIMIZING QUEUING CONSTRAINS IN SERVICE DELIVERY ON ENTERPRISE COMMUNICATION&COMPUTING PLATFORM

N. Kryvinska[1], C. Strauss[1], P. Zinterhof[2]
[1]University of Vienna, Vienna, Austria; [2]University of Salzburg, Salzburg, Austria

Corresponding Author: N. Kryvinska, University of Vienna, Department of e-Business, Faculty of Business, Economics and Statistics
Bruenner Str. 72, A - 1210, Vienna, Austria; natalia.kryvinska@univie.ac.at

Abstract. The key objective of the enterprise communication&computing platform is to deliver services at anytime, anywhere, to any device [1]. The service providers are in critical need for the agile service delivery, to succeed in highly concurrent e-services business marketplace. Thus, regardless of many similarities between the service delivery models used by different providers, the customization is highly desired to enable communications within diverse operations and business support systems (OSS/BSS) and niche-market developments [2]. We take also into consideration that an analytical performance evaluation is crucial for the justification of the effectiveness of the modelling of different operational conditions in delivering of high-quality services [3]. Consequently, in this paper, we apply a mathematical model into the monitoring of services delivery in two interconnected systems in tandem on the enterprise communication&computing platform. We consider here a queuing network model used to represent a series of 2 single-server queues, each with unlimited waiting space and the FIFO service discipline [4]. And, we develop our model in order to obtain feasible values of main performances features.

1 Introduction

The Internet Service Providers (ISPs) and Mobile Virtual Network Operators (MVNOs) as well as IT departments of large companies face various problems when designing, building/implementing, and engineering/managing their enterprise infrastructures. They struggle mostly because of inflexible legacy networks in the face of accelerating technological growth and increasing business and customer demands. Integrated infrastructure solutions are extremely complex to design, support, and manage. Besides, the enterprise communication&computing platform must be able to adapt quickly, to support new technologies that enable rapidly evolving business processes [1].

The solution for all the difficulties enterprise communication&computing platform faces cannot be simple and whole, done in once. It can be efficient only if it is modular and includes differentiated features. As an answer to this complex dilemma - we split different problems into different sub-areas, and apply certain customized solutions to the certain kind of challenges.

Besides, the services and providers continue to converge and bring in many different types of services, such as wireless, broadband media, and voice ones. Plus the rising security and compliance pressures are forcing companies to re-evaluate their traditional network architectures in order to enforce stronger access control and auditing policies without inhibiting the levels of flexibility and agility.

Moreover, computing centres are now distributed across the world, as there are devices and clients that they support. And, the situation totally cannot be comparable with the past, where computing centres typically supported limited number long-term services, for instance - a single server ran the same application for its lifecycle of operation. The little changes were made, networks were static, often flat to ease administration, and network security was confined to just a few points within the implementation.

But, designing networks and architectures for millions of users, transactions, and diverse content types requires an approach that differs from that of standard infrastructure designs. These infrastructures must support convergent services, with high levels of reliability and performance, while maintaining manageability as well as security. The computing architectures and applications must be designed to support ubiquitous access. Flexibility is required to support new or evolving business and technical requirements with agile respond to business conditions and competitive pressures. The “always-on” access has become a requirement whether services are accessed over the Internet, intranet or extranet [1, 2].

2 Analytical Model to Monitor Services Delivery over 2 Interconnected Systems in Tandem on the Enterprise Communication&Computing Platform

The most critical factor in the real-time services delivery is the response time or delay. Therefore, we examine here a mathematical model in order to apply it into the monitoring of end-to-end delay in the services delivery over two interconnected systems in tandem.

Interconnected systems in tandem (Figure 1) received significant attention in literature because of their pervasiveness and implication in real life. For example, Avi-Itzhak [5] studied the system with arbitrary input and regular service times [6]. Other related work, in the sense that it focuses on the response time as opposed to the joint queue length, is done by Knessl and Tier [7]. They have studied the first two moments of the response time in an open two-node queuing network with feedback for the case with an exponential processor sharing (PS) node and a FIFO node, while the arrivals at the PS node are Poisson. Chao and Pinedo [8] examined the case of two tandem queues with batch Poisson arrivals and no buffer space in the second queue. They allowed the service times to be general and obtained the expected time in system [6].

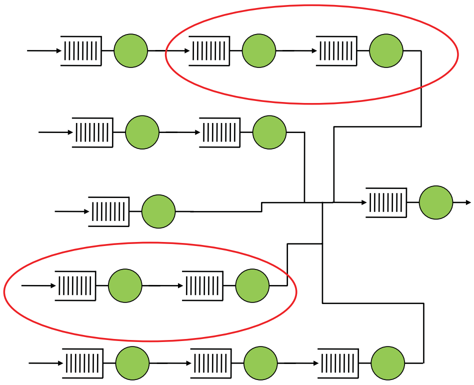

Figure 1. An Example of Interconnected Systems in Tandem.

2.1 Modelling Assumptions and Model Description

We begin with a description of the model under consideration. We have observed that in almost earlier research, it was assumed that the arrival process is a Poisson process. For other arrival processes it is seldom possible to find an exact expression for the mean waiting time except in the case where the holding times are exponentially distributed. In general it is assumed, that either the arrival process or the service process should be Markovian. For GI/G/1 queuing system it is in the state to give the theoretical boundary for the mean waiting time and response time [9].

2.2 Performance of Analysis

The example for a non-Poisson arrival process is the queuing system GI/M/1, where the distribution of the inter-arrival times is a general distribution, given by the density function *f(t)* [9]. The probability, that arriving customer finds the server busy - θ is not the same as the server utilization – ρ for GI/M/1 because of the general pattern of arrivals. Only the random arrivals have $\theta = \rho$. The value of θ can be obtained from formula.

$$\theta = f^{*}(s)\left(\frac{1-\theta}{T_S}\right); \qquad 0 \le \theta < 1. \tag{1}$$

where *f*(s)* is the Laplace-Stieltjes transform of the pdf of inter-arrival times. In some cases formula (1) can be solved analytically, but in general a numerical procedure is required.

In Figure 2 is presented probability that arriving customer finds the server busy for different distributions of inter-arrival time [10].

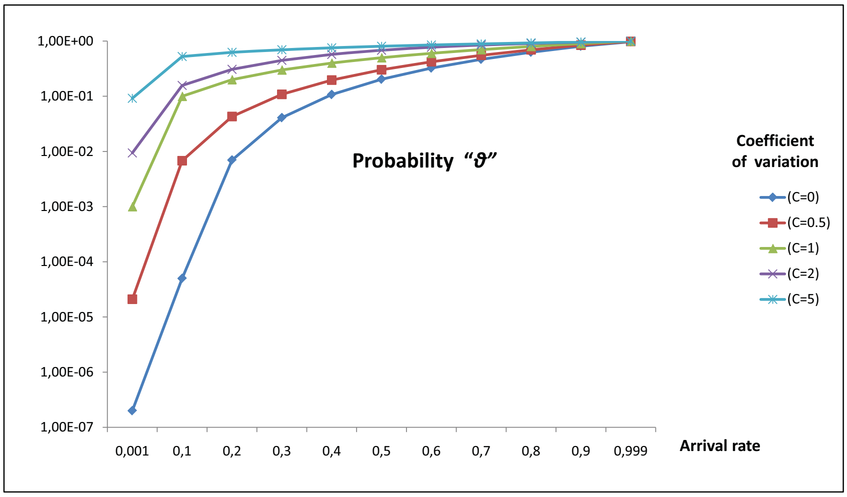

Figure 2. The probability θ that arriving customer finds the server busy for different distributions of inter-arrival time.

Next, we calculate the average waiting time using the formula (2). The waiting time for the G/M/1 queuing system has a modified exponential distribution, and it is described in formula (3).

$$T_W = \frac{\theta T_S}{1-\theta}. \qquad (2) \qquad\qquad p(T_W < 1) = 1 - \theta e^{-\frac{t}{T}}. \qquad (3)$$

It is necessary to notice, that the average waiting time increases when the arrival pattern becomes more irregular. Figure 3 shows average waiting time for different values of C_A^2. The effect of increased variance in the inter-arrival time is apparent, and is very marked at high utilizations [11].

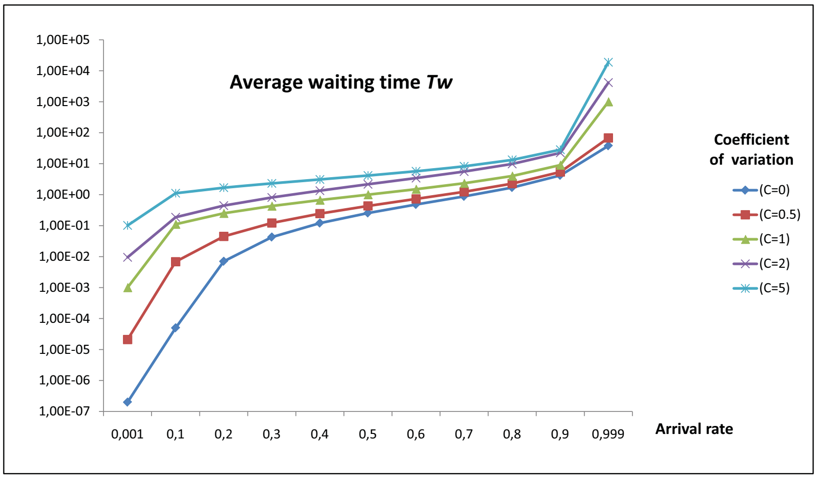

Figure 3. The average waiting time.

The average time in system for the G/M/1 queuing system has the pattern (Figure 4) presented in formula (4).

$$T = \frac{T_S}{1-\theta}. \qquad (4) \qquad\qquad p(T < t) = 1 - e^{-\frac{t}{T}} \qquad (5)$$

The service time has an exponential distribution, formula (5) [11].

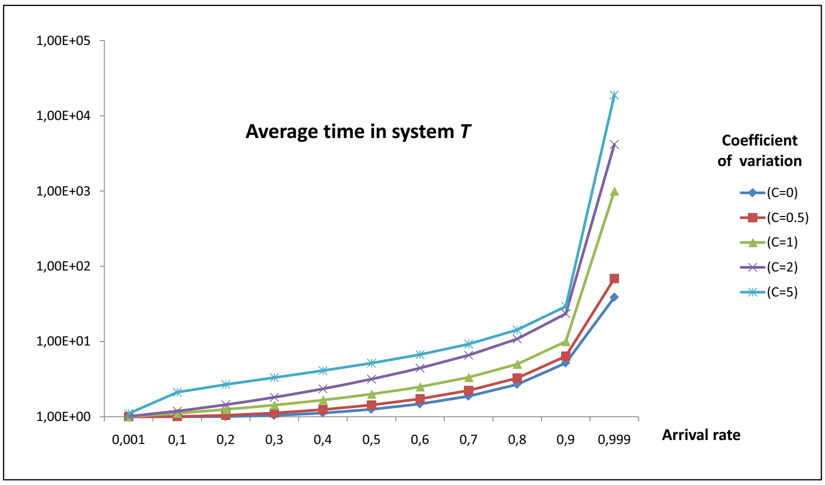

Figure 4. The average time in system, equal standard deviation.

3 Practical Usage of the Proposed Model

We introduce and analyse here architecture, namely “Unified Network Architecture”, built onto the mathematical model, for the monitoring services delivery in two interconnected systems in tandem on the enterprise communication&computing platform. The “Unified Network Architecture” (Figure 5) can provide a carrier-ready transport platform for packet tandem applications with proven PSTN compatibility, the rich set of transport options, high availability and improved tandem functioning. This architecture provides fast call transfer, which facilitates revenue-sustaining long-distance services. The “Unified Network Architecture” allows service providers to deliver new enhanced services and applications, while easily adding trunking capacity, provisioning new routes, easing data congestion points.

Proficient of supporting multiple applications from the same platform, the “Unified Network Architecture” allows carriers to deploy a single solution for IP or PSTN tandem functions, including:

- long distance tandem services - to reduce carriers capital costs in communicating voice calls over long distance trunks between their voice switches.
- SIP-based applications - to build upon a SIP-based architecture, the solution supports third-party vendors to deliver a wide variety of new services.
- business VPNs - for carriers to offer business customers services such as Centrex - either add voice to existing IP VPNs or set up initial VPNs with voice and link branch offices with low-cost service [12].

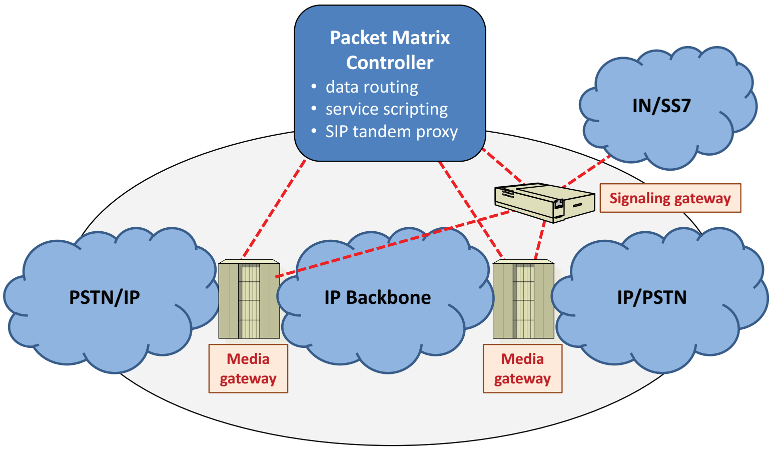

Figure 5. The “Unified Network Architecture”, tandem-based.

4 Conclusions

We applied in this paper a mathematical model for the monitoring of services delivery in two interconnected systems in tandem on the enterprise communication&computing platform. We considered here a queuing network model used to represent a series of 2 single-server queues, each with unlimited waiting space and the FIFO service discipline. We examined and developed our model in order to obtain feasible values of main performances features.

We started with a description of the model under consideration. Next, we obtained the performance measure parameters of the GI/M/1-type Markov model. A computational model the response time is presented in Section 2, while Section 3 addresses the practical usage of our model. We demonstrated also the strength of our model through the practical implementation examples.

5 References

[1] Lofstrand, M., and Carolan, J.: *Sun’s Pattern-Based Design Framework: the Service Delivery Network*, Sun BluePrints™ OnLine, Sun Microsystems, September 2005.

[2] White Paper: *Enabling Service Delivery Using the Microsoft Connected Services Framework*, Microsoft Corporation, January 2005.

[3] Mun, Y.: *Performance Analysis of Banyan-Type Multistage Interconnection Networks Under Nonuniform Traffic Pattern*. The Journal of Supercomputing, Volume 33, Number 1, July 2005, pp. 33-52(20).

[4] Glynn, P. W., and Whitt, W.: *Departures from Many Queues in Series*. The Annals of Applied Probability. Volume 1, Number 4, 1991, pp. 546-572.

[5] Avi-Itzhak B.: *A sequence of service stations with arbitrary input and regular service times*, Management Science, 11 (5), 1965, pp.565–571.

[6] Van Houdt, B., Attahiru Sule Alfa.: *Response time in a tandem queue with blocking, Markovian arrivals and phase-type services*. Operations Research Letters 33, 2005, pp.373 – 381.

[7] Knessl, C., Tier, C.: *Approximation to the moments of the sojourn time in a tandem queue with overtaking*. Stochastic Models 6 (3), 1990, pp.499–524.

[8] Chao, X., Pinedo, M.: *Batch arrivals to a tandem queue without an intermediate buffer*. Stochastic Models 6 (4), 1990, pp.735–748.

[9] Iversen, V. B.: *Fundamentals of Teletraffic Engineering*. 2001. Online: www.tele.dtu.dk/teletraffic

[10] Hashida, O., Ueda, T., Yoshida, M., and Murao Y.: *Queueing Tables*. The Electrical Communication Laboratories Nippon Telegraph and Telephone Public Corporation, Tokyo, Japan, 1980.

[11] Tanner, M.: *Practical Queueing Anylysis*. IBM McGraw-Hill Series, 1995.

[12] Application Note: *Packet Tandem Applications*, Convergent Networks, January, APP-TAN-0703, 2003.