Numerical analysis of infinite server queues with correlated arrivals

Tamás Éltető[1] and Miklós Telek[2]

[1] Ericsson Research Budapest, Hungary
tamas.elteto@ericsson.com
[2] Dept. of Telecommunications, Technical University of Budapest
telek@hit.bme.hu

Abstract. This paper investigates the effect of correlation in the arrival process of infinite server queueing systems by numerical analysis. The arrival process is described by a Markov Arrival Process (MAP) and the holding time can have arbitrary distribution with finite mean. We provide an analytic expression for the moment-generating function of the stationary distribution of customers and based on that we approximate the distribution numerically.
We also investigate the concept of "critical time-scale" in MAP/G/∞ queueing systems. Numerical examples investigate the cases when the time scale of correlations in the arrival process are below and above the "critical time-scales". We show that in the presence of long lasting correlations, the system can be approximated by the mixture of independent short-range dependent subsystems.

Keywords: Markov arrival process, infinite server queue, MAP/G/∞ queueing system, numerical analysis, long range dependent (LRD), short range dependent (SRD).

1 Introduction

In this paper we present a method for analysing a MAP/G/∞ queueing system with which we investigate the correlation effects of the arrivals. The analysis of infinite server system was addressed assuming phase-type renewal arrival processes in [12]. This result was generalized assuming batch Markov renewal processes in [9]. The analysis of infinite server systems with batch Markov arrival process (BMAP) was considered in [10].

[12] developed a system of differential equations for the generating function of the number of customers and differential equations for its moments. [10] provides the differential equations for the generating function of the BMAP/G/∞ system and gives numerically feasible formulae for phase-type service time. For general service times, both papers recommend the numerical solution of the system of differential equations.

The approach proposed in this paper is essentially different from the ones of [12] and [10]. Following this new approach we obtain the known analytic results for general service time distribution, and, as a direct consequence of the approach, we develop analytical

results and feasible computational method for the case of discrete service time distribution[3].

We also propose a numerically feasible method for computing the distribution of the number of customers in the system based on the probability generator matrix[4]. Indeed it is a numerical inversion method which does not pose any restrictions on the service time distribution.

The proposed numerical techniques are used in a telecommunication example, where the traffic flows arrive in a non-Poisson manner. Such problems are also addressed in e.g. [1, 4] and references therein. There are two important assumptions in our examples:

- The arrival process is independent of the number of customers/flows in the system.
- The holding time of the customers/flows depends on the arrival process only.

With this example, we investigate the concept of "critical time-scale". It has been shown that the Internet traffic exhibit correlations for several time-scales which has significant implications on the traffic statistics [6].

[14] argues that in finite single server queueing systems the correlation in the arrival process (e.g. packets, cells arrival) is important in the time-scale of queueing and the correlation above this time-scale does not affect the network performance significantly.

Our example shows a scenario in which such "critical time-scale" exists in the case of infinite server queueing systems as well. A similar concept has been used in [5] in a measurement-based admission control context. In that paper the authors argue that a measurement-based admission control algorithm could take into consideration slow changes in the traffic while fast changes can be dealt with overprovisioning.

The examples indicate that the MAP/G/∞ system with long lasting correlations in the arrivals can be decomposed. Above the critical time scale, the correlations can dealt with by independent subsystems with short range dependence.

The rest of the paper is organized as follows. Section 2 presents the basic modeling concept. This concept is used for the analysis of MAP/D$_n$/∞ queue in Section 3 and MAP/G/∞ queue in Section 4. Further properties of MAP/G/∞ queues related with their numerical analysis are presented in Section 5. Section 6 presents an example which is used to evaluate the effect of arrival process correlation and also to investigate the accuracy of numerical methods. The paper is concluded in Section 7.

2 Notations and preliminaries

The moment-generating function of the number of arrivals of a MAP with matrices $\mathbf{D_0}$ and $\mathbf{D_1}$ is

$$P_{i,j}(z,t) = E(z^{N(t)}|J(0)=i, J(t)=j)P(J(t)=j|J(0)=i),$$

where $N(t)$ denotes the number of arrivals in $(0,t)$ and $J(t)$ is the state of the background Markov chain at time t. The $\mathbf{P}(z,t) = \{P_{i,j}(z,t)\}$ matrix is given, e.q., in [8]: $\mathbf{P}(z,t) = e^{(\mathbf{D_0}+z\mathbf{D_1})t}$.

[3] [12] mentions the case of discrete service times as an example, but does not make further considerations.
[4] Defined in Theorem 3.6.2 of [8] on page 78.

In order to handle time dependent MAPs we introduce the $\mathbf{P}(z,t,T) = \{P(z,t,T)_{i,j}\}$ matrix as the moment-generating function of the number of arrivals in the (t,T) interval, where

$$P_{i,j}(z,t,T) = E(z^{N(t,T)}|J(t)=i, J(T)=j)P(J(T)=j|J(t)=i), \tag{1}$$

and $N(t,T)$ denotes the number of arrivals in the (t,T) interval. Using this definition one can investigate the basic properties of piece-wise constant time-dependent MAP processes.

Let MAP1 and MAP2 be two MAP processes on the same state-space. MAP1 is characterized by $(\mathbf{D_0^{(1)}}, \mathbf{D_1^{(1)}})$ and MAP2 by $(\mathbf{D_0^{(2)}}, \mathbf{D_1^{(2)}})$. In the (s,t) interval, the arrivals are generated by MAP1 and in the (t,u) interval the arrivals are generated by MAP2 such that the initial state of MAP2 at t is the same as the final state of MAP1 at t, i.e., $J_1(t) = J_2(t)$ (Figure 1).

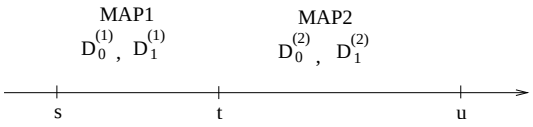


Fig. 1. Piece-wise constant time-dependent MAP

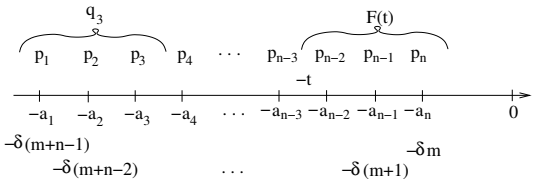


Fig. 2. Discrete equidistance service time distribution

Theorem 1. *The moment-generating function of the number of arrivals in the (s,u) interval is*

$$\mathbf{Z}(z,s,u) = \mathbf{Z_1}(z,s,t) \cdot \mathbf{Z_2}(z,t,u),$$

where $\mathbf{Z_1}(z,s,t) = e^{(\mathbf{D_0^{(1)}} + z\mathbf{D_1^{(1)}})(t-s)}$ *and* $\mathbf{Z_2}(z,t,u) = e^{(\mathbf{D_0^{(2)}} + z\mathbf{D_1^{(2)}})(u-t)}$.

3 The MAP/$\mathbf{D_n}/\infty$ queueing system

Consider an infinite-server queueing system where the customers arrive to the system according to a $(\mathbf{D_0}, \mathbf{D_1})$ MAP and stay for a random amount of time X. The service time is a finite discrete random variable with probability distribution: $P(X = a_i) = p_i$, $i = 1, \ldots, n$, where $a_1 > a_2 > \ldots > a_n > 0$. For notational convenience we introduce $a_{n+1} = 0$ (Figure 2).

Theorem 2. *The moment-generating function of the stationary number of customers in the system (N) is*

$$E(z^N) = \pi \left(\prod_{i=1}^{n} e^{(\mathbf{D} + (z-1)q_i\mathbf{D_1})(a_i - a_{i+1})} \right) \mathbb{1}, \tag{2}$$

where $\mathbf{D} = \mathbf{D_0} + \mathbf{D_1}$, π *is the stationary distribution of the Continuous-time Markov Chain (CTMC) with generator* $\mathbf{D}$, $q_i = \sum_{j=1}^{i} p_j$, $\mathbb{1}$ *is the column vector of ones, and the product is ordered.*

In the special case when the service time is a discrete equidistance distribution between $m\delta$ and $(n+m-1)\delta$, such that $a_i = (n+m-i)\delta$ (Figure 2) we have $F(t) = P(X \leq t) = \sum_{j=n+m-\lfloor t/\delta \rfloor}^{n} p_j$, $q_i = 1 - F((n+m-i-1)\delta)$ and the matrix of moment-generating function becomes

$$\mathbf{P}(z) = \prod_{i=1}^{n} e^{(\mathbf{D}+(z-1)q_i\mathbf{D_1})\delta} = \prod_{i=1}^{n} e^{(\mathbf{D}+(z-1)(1-F((n+m-i-1)\delta))\mathbf{D_1})\delta} . \quad (3)$$

4 The MAP/G/∞ queueing system

In this section we generalize the result of the previous section to analyze the MAP/G/∞ queue. This model is already analyzed in [10], however, the moment-generating function below can be obtained as a limit of Equation 3.

Theorem 3. *The moment-generating function of the number of customers in MAP/G/∞ queue with holding time distribution $F(x)$ with support on the (t,T) interval is*

$$\begin{aligned}\mathbf{P}(z,t,T) = \mathbf{I} + \int_t^T (\mathbf{D} + (z-1)(1-F(y_1))\mathbf{D_1})dy_1 + \\ \int_t^T (\mathbf{D} + (z-1)(1-F(y_1))\mathbf{D_1}) \int_t^{y_1} (\mathbf{D} + (z-1)(1-F(y_2))\mathbf{D_1})dy_2 dy_1 + \ldots\end{aligned} \quad (4)$$

Corollary 1. *Due to Theorem 3 and (3) the moment-generating function $\mathbf{P}(z,t,T)$ can be expressed as the product of "sub moment-generating functions": $\mathbf{P}(z,t,T) = \mathbf{P}(z,t^*,T) \cdot \mathbf{P}(z,t,t^*)$, where $t^* \in (t,T)$.*

The moment-generating function of the number of customers with unbounded positive holding time distribution is obtained as $\mathbf{P}(z,0,\infty) = \lim_{T\to\infty} \mathbf{P}(z,0,T)$. Based on $\mathbf{P}(z,t,T)$ the probability of having n customers in the system, $\mathbf{V}_n(t,T)$, can be obtained using the residue theorem.

Corollary 2. *Since $\mathbf{P}(z,t,T)$ is the moment-generating function of $\mathbf{V_n}(t,T)$, $\mathbf{V_n}(t,T)$ is the residue of the function $\frac{1}{z^{n+1}}\mathbf{P}(z,t,T)$ in $z=0$.*

$$\mathbf{V}_n(t,T) = \frac{1}{2\pi i} \oint \frac{1}{z^{n+1}} \mathbf{P}(z,t,T)dz. \quad (5)$$

where i is the imaginary unit.

5 Numerical analysis of the distribution of customers

We define the following relation of random variables X and Y, and their moment-generating functions.

Definition 1. *$X \trianglelefteq Y$ iff $P(X \leq x) \geq P(Y \leq x)$ for $\forall x$ (i.e., $F_X(x) \geq F_Y(x)$) and $X(z) \trianglelefteq Y(z)$ iff $P(X \leq x) \geq P(Y \leq x)$ for $\forall x$ where $X(z) = E(z^X)$ and $Y(z) = E(z^Y)$ are the moment-generating functions of X and Y, respectively.*

Note that

- $X_1 \trianglelefteq Y_1$ and $X_2 \trianglelefteq Y_2$ implies $X_1 + X_2 \trianglelefteq Y_1 + Y_2$,
- $X_1(z) \trianglelefteq Y_1(z)$ and $X_2(z) \trianglelefteq Y_2(z)$ implies $X_1(z)X_2(z) \trianglelefteq Y_1(z)Y_2(z)$ and
- $X \trianglelefteq Y$ implies that the nth ordinary and the nth factorial moments of X are not greater than the ones of Y, i.e., $E(X^n) \leq E(Y^n)$ and $E(X(X-1)\dots(X-n+1)) \leq E(Y(Y-1)\dots(Y-n+1))$.

Let $\mathbf{P}(z,t,T)$ and $\mathbf{P}^*(z,t,T)$ be the generator matrices of the MAP/G/∞ systems with service times $F(t)$ and $F^*(t)$.

Theorem 4. *If $F(t) \geq F^*(t)$ then $\mathbf{P}(z,t,T) \trianglelefteq \mathbf{P}^*(z,t,T)$.*

Note that Thoerem 4 and its proof is also valid for any stationary arrival process.

Theorem 5. *$\mathbf{P}(z,t,T)$ is bounded by the following integrals of matrix exponential expressions*

$$\underline{\mathbf{P}}(z,t,T) \trianglelefteq \mathbf{P}(z,t,T) \trianglelefteq \overline{\mathbf{P}}(z,t,T), \tag{6}$$

where $t = t_0 < t_1 < \ldots < t_{N-1} < t_N = T$

$$\underline{\mathbf{P}}(z,t,T) = \prod_{i=0}^{N} \check{\mathbf{P}}(z, t_{N-i}, t_{N-i+1}), \quad \overline{\mathbf{P}}(z,t,T) = \prod_{i=0}^{N} \hat{\mathbf{P}}(z, t_{N-i}, t_{N-i+1}),$$

and

$$\begin{aligned}\check{\mathbf{P}}(z,t_k,t_{k+1}) = e^{\mathbf{D}(t_{k+1}-t_k)} + (z-1)(1-F(t_{k+1}))\int_{t_k}^{t_{k+1}} e^{\mathbf{D}(t_{k+1}-y_1)}\mathbf{D_1}e^{\mathbf{D}(y_1-t_k)}dy_1 + \\ (z-1)^2(1-F(t_{k+1}))^2\int_{t_k}^{t_{k+1}} e^{\mathbf{D}(t_{k+1}-y_1)}\mathbf{D_1}\int_{t_k}^{y_1} e^{\mathbf{D}(y_1-y_2)}\mathbf{D_1}e^{\mathbf{D}(y_2-t_k)}dy_2dy_1 + \ldots\end{aligned} \tag{7}$$

$$\begin{aligned}\hat{\mathbf{P}}(z,t_k,t_{k+1}) = e^{\mathbf{D}(t_{k+1}-t_k)} + (z-1)(1-F(t_k))\int_{t_k}^{t_{k+1}} e^{\mathbf{D}(t_{k+1}-y_1)}\mathbf{D_1}e^{\mathbf{D}(y_1-t_k)}dy_1 + \\ (z-1)^2(1-F(t_k))^2\int_{t_k}^{t_{k+1}} e^{\mathbf{D}(t_{k+1}-y_1)}\mathbf{D_1}\int_{t_k}^{y_1} e^{\mathbf{D}(y_1-y_2)}\mathbf{D_1}e^{\mathbf{D}(y_2-t_k)}dy_2dy_1 + \ldots\end{aligned} \tag{8}$$

Theorem 5 gives upper and lower bound when the service time distribution has a finite support. In case infinite support the theorem gives lower bound, while an approximate upper bound is given by Theorem 7.

Theorem 6. *The moment-generating function fulfills the following bound*

$$\|\mathbf{P}(z,t,T)\| \leq e^{|z-1|\|\mathbf{D_1}\|\int_t^T (1-F(y))dy} \tag{9}$$

Theorem 7. *Computing the probability masses with (5) the error caused by a finite truncation of the service time distribution at T is bounded by*

$$\|\mathbf{V_n}(0,\infty) - \mathbf{V_n}(0,T)\| \leq lD^n\epsilon(z,T), \tag{10}$$

where $\epsilon(z,T) = \left(e^{|z-1|\|\mathbf{D_1}\|\int_T^\infty (1-F(y))dy} - 1\right) e^{|z-1|\|\mathbf{D_1}\|E(X)}$ where l is the length of the closed contour around 0 on which the residue is calculated and $1/D$ is the minimum distance of the closed contour from 0. If the residue is calculated on circle with radius r then $l = 2r\pi$ and $D^n = r^n$.

6 Numerical example

The assumption of the Poisson customer/flow arrival is considerably general since this is the limiting process in telecommunication systems when arrivals occur from a large population especially in scenarios where there is significant network traffic [2, 7]. However, there are cases, where the arrival process is not Poisson. This is the case, for example, in the common model of web browsing where the download of the objects of web pages involve several transfers [17].

It is also possible that even the starts of the user activities are correlated. Consider a web-based customer service scenario. The customers get "one-time passwords", e.g., via their mobile phones and start their sessions using these passwords. If several users get their identifiers at the same time, then the session starts are not Poisson.

In the examples, the the arrivals will be modelled by Markov-modulated Poisson Process (MMPP), because the modulation of the arrival rate allows to investigate the concept of *critical time-scale* defined in [14] and used for example in [5]. We evaluate two of such examples. Both examples models a system, where voice (video, etc.) calls are routed through an egress node. The arrival of the calls is modulated by some external effects.

It will be shown that if the time scale of the call arrival rate change is below the holding time of the calls then the system cannot be decomposed into subsystems. However, if the routing changes happen above the range of call durations, the system can be well approximated by a probability mixture. Moreover, it will be shown that the distribution of the call holding time changes the critical time-scale.

The first example is a multiservice system focusing on the traffic of voice calls. Normally, all calls are allowed and routed towards a switch, but sometimes the calls are blocked. The decision is made in a filter depending on the available resources to set up new calls. The time-scale of the decisions are a few seconds.

In the second example, two areas are interconnected via a microwave link that transmits the calls. However, the link might become unavailable due to environmental conditions in which case a backup satellite link is used. The calls are rerouted to the satellite link, but only a few new calls are allowed. The time-scale here is in the order of 100-1000s.

Figure 3 shows the structure of the call arrivals. The calls from Area A always arrive to the switch according to a constant rate Poisson process. Area B generates calls at a constant rate too but these calls are sometimes blocked. This is modeled by a two-state Markov chain (see Figure 4). The arrival rate from Area A is $\lambda_A = 0.1\ 1/s$ and from Area B is $\lambda_B = 9.9\ 1/s$. The call duration is $E(X) = 100s$.

We consider two service time distributions: hyper-exponential and Weibull. There are three main reasons for choosing these distributions:

1. The Weibull distribution is not Phase-type distribution therefore it seems to be rather difficult to calculate the factorial moments.
2. The asymptotical decay of the chosen Weibull distribution is slower than exponential.
3. The two considerably different distributions show the stationary distribution of the system is sensitive to the holding time in the present example.

In the case of the Weibull distribution the result is compared to simulation output and to the upper and lower bounds.

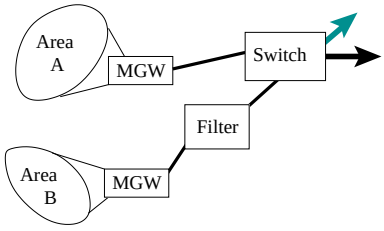


Fig. 3. The routes of the calls in the two examples

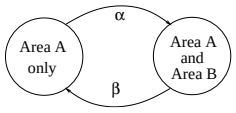


Fig. 4. The routed areas to Egress node 1 in the different MMPP states

The numerical procedure is implemented in Octave 2.1.40. The exact factorial moments were calculated by Maple 5.00. The simulation was done on a tailor-made simulator. The runnning time of the numerical calculations was 15 hours on a desktop with 2.5 GHz CPU, 1 GByte RAM and Red-Hat linux system. Inverting the generating function could be done parallel so this can be further reduced. The first 20 factorial moments were calculated in 24 hours on a SUN Enterprise 420R computer with 450MHz CPU, 2GByte RAM.

6.1 Hyper-exponential service time distribution

The call duration distribution is hyper-exponential: $F(x) = 1 - \gamma_1 e^{-\mu_1} - \gamma_2 e^{-\mu_2}$, where $1/\mu_1 = 50s$, $\gamma_1 = 0.947$, $1/\mu_2 = 1000s$, $\gamma_1 = 0.053$. Figure 5 shows the stationary distributions of the system. The state durations in the MMPP were varied from $1/\beta = 1s$ to $10000s$ and $1/\alpha = 1.5s$ to $15000s$. When the arrival rate changes very frequently (i.e. within seconds, Figure 5a), the correlations are very short and a unimodal distribution develops.

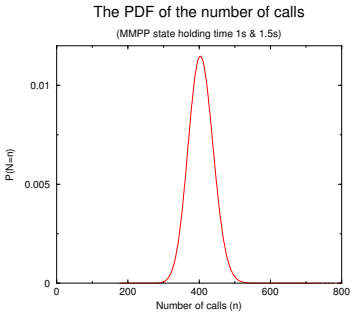
In Figure 5b and c the correlations in the arrival process start to have effects and a bi-modal distribution develops. However, the system cannot be decomposed into one where the calls arrive only from Area A and another where the calls arrive both from Area A and Area B. The difference be seen in Figure 5c, where the mixture of two Poisson distributions representing the decomposition is shown.

Figure 5d shows a scenario in which the range of the correlations is well above the range of the call holding time. In this case the decomposition is possible.

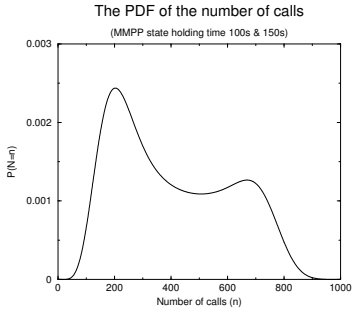
In order to validate the numerical computation method Table 1 compares some lower and upper bounds for the factorial moments with the analytical calculation. The relative errors are at most 5-6%.

	lower b.	exact.	upper b.	lower b.	exact.	upper b.	lower b.	exact.	upper b.
n	$1/\beta = 1s, 1/\lambda = 1.5s$			$1/\beta = 100s, 1/\lambda = 150s$			$1/\beta = 1000s, 1/\lambda = 1500s$		
1	4.06 $e02$	4.07 $e02$	4.09 $e02$	4.06 $e02$	4.07 $e02$	4.09 $e02$	4.06 $e02$	4.07 $e02$	4.09 $e02$
2	1.66 $e05$	1.67 $e05$	1.68 $e05$	2.07 $e05$	2.08 $e05$	2.09 $e05$	2.86 $e05$	2.87 $e05$	2.88 $e05$
10	1.51 $e26$	1.56 $e27$	1.60 $e26$	9.71 $e27$	9.87 $e27$	1.00 $e28$	1.04 $e29$	1.06 $e29$	1.08 $e29$
15	2.19 $e39$	2.30 $e39$	2.39 $e39$	2.48 $e42$	2.54 $e42$	2.61 $e42$	7.50 $e43$	7.70 $e43$	7.93 $e43$
20	3.54 $e52$	3.76 $e52$	3.97 $e52$	7.53 $e56$	7.78 $e56$	8.06 $e56$	5.88 $e58$	6.10 $e58$	6.34 $e58$

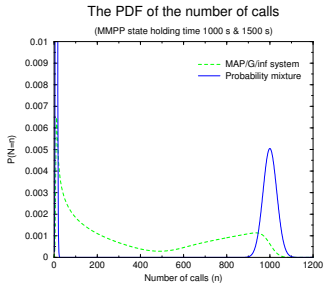
Table 1. Comparison of the first 20 factorial moments from exact and numerical calculation



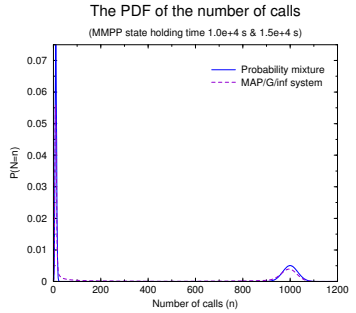
a) $(1/\beta = 1s, 1/\alpha = 1.5s)$



b) $(1/\beta = 100s, 1/\alpha = 150s)$



c) $(1/\beta = 1000s, 1/\alpha = 1500s)$



d) $(1/\beta = 10000s, 1/\alpha = 15000s)$

Fig. 5. The distributions of the number of calls with different parameters

6.2 Weibull service time distribution

Figure 6 presents the distributions of the system with call duration: $F(x) = 1 - e^{-(x/\theta)^\eta}$, where $\eta = 0.986$, $\theta = 94.8$. Figure 6 presents the numerical and discrete event simulation results.

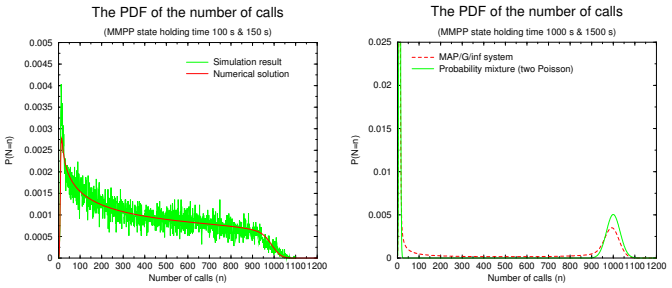


Fig. 6. The distributions of the number of calls with Weibull service time distribution $(1/\beta = 100, 1/\alpha = 150)$ and $(1/\beta = 1000, 1/\alpha = 1500)$

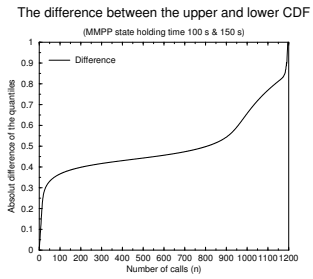


Fig. 7. The difference of bounds with Weibull service time distribution $(1/\beta = 100, 1/\alpha = 150)$

The bounds presented in Section 5 give a fairly low approximation error. T is chosen to be 465000 seconds. The overall error is 10^{-19} for $\mathbf{V}_n$. The probability distribution is discretised: 40000 points are distributed in an equi-probabilistic way, while 5000 points are spaced evenly from $1337s$ to $100000s$ and the last point is at $465000s$.

Figure 7 shows the absolute difference of the upper and lower bounds. The differences of bounds' p-quantiles versus the p-quantile of the lower bound.

7 Conclusion

A new analysis approach is presented to analyse MAP/G/∞ queues. This approach couples the elementary properties of inhomogeneous MAPs and infinite server queues. The obtained transform domain expression (4) is known from [10], but additionally the applied approach provides new insights to the queueing behaviour (e.g., Theorem 4), which we utilized for developing an accurate numerical analysis procedure. The proposed numerical analysis method is applicable with general service time distribution and supplemented with a detailed numerical error analysis.

The numerical analysis makes it possible to establish upper and lower bounds for the exact distribution function and factorial moments. In this way, the approximation can be validated by comparing these upper and lower bounds as it was shown in the numerical examples for both the moments and the quantiles.

Based on the proposed method we investigate the effect of correlated input processes. We found that the short range correlation (with respect to the service time) in the arrival process of infinite server queues does have an effect on the distribution of customers, but with the longer scale correlation the system can be decomposed into independent subsystems.

From these examples we conclude that in infinite server queues the long-range correlation of the arrival process need not necessarily be considered. It is enough to substitute the model with long-range dependence with subsystems having only short range dependence, which can be described by smaller MAP/G/∞ models and the final result can be obtained as the probability mixture of the subsystems' distribution.

Moreover, the comparison of the hyper-exponential and Weibull distributed holding times shows that the critical time-scales can differ though the average holding times are the same. The critical time-scale on which the system can be decomposed into subsystems is in the order of 10^4 seconds when the holding time hyper-exponential while it is approximately one order of magnitude smaller (10^3 seconds) with Weibull distributed holding time.

References

1. C. Barakat, P. Thiran, G. Iannaccone, C. Diot, P. Owezarski, "A flow-based model for Internet backbone traffic",IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, 2002.
2. J. Cao, W. S. Cleveland, D. Lin, D. X. Sun, "On the nonstationarity of Internet traffic", ACM SIGMETRICS Performance Evaluation Review, v.29 n.1, p.102-112, June 2001.
3. Tamás Éltető, Miklós Telek, "Numerical analysis of infinite server queues with correlated arrivals", Technical Report, http://www.szit.bme.hu/~eltetot/inf_srv_rep.ps.gz
4. S. Ben Fredj, T. Bonald, A. Proutiere, G. Regnie, J. Roberts, "Statistical Bandwidth Sharing: A Study of Congestion at Flow Level", ACM SIGCOMM, August 2001.

5. M. Grossglauser. "A time-scale decomposition approach to measurement-based admission control", IEEE/ACM Transaction on networking, vol. 11/4, August 2003.
6. Rachid El Abdouni Khayari, Ramin Sadre, Boudewijn R. Haverkort, "A Validation of the Pseudo Self-Similar Traffic Model", International Conference on Dependable Systems and Networks (DSN'02), June 23 - 26, Washington, D.C., USA, 2002.
7. M. F. T. Karagiannis, M. Molle, A. Broido, "A Nonstationary Poisson View of Internet Traffic", in Proceedings of IEEE INFOCOM, Mar. 2004.
8. G. Latouche and V. Ramaswami. *Introduction to Matrix-Analytic Methods in Stochastic Modeling*. Series on statistics and applied probability. ASA-SIAM, 1999.
9. L. Liu, J. G. C. Templeton, "The $GR^{X_n}/G_n/\infty$ system: system size", Queueing Systems 8 (1991), 323-356.
10. Hiroyuki Masuyama and Tetsuya Takine. Analysis of an Infinite-Server Queue with Batch Markovian Arrival Streams, *Queueing Systems*, vol. 42, no.3, pp. 269-296, 2002.
11. R. German and C. Lindemann, "Analysis of stochastic Petri nets by the method of supplementary variables," *Performance Evaluation*, vol. 20, pp. 317–335, 1994.
12. V. Ramaswami and M. Neuts "Some explicit formulas and computational methods for infinite-server queues with phase-type arrival", *J. Appl. Prob.*, vol. 17, 498-514, 1980.
13. B. Ryu and S. B. Lowen. "Point process models for self-similar network traffic, with applications", Stochastic Models, vol. 14, 735-761, 1998.
14. Bong K. Ryu, Anwar Elwalid . "The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities," SIGCOMM'96, 1996.
15. L. Takács. "On Erlang's formula", The Annals of Mathematical Statistics, vol. 40, 71-78, 1969.
16. Tadafumi Yoshihara, Shoji Kasahara and Yutaka Takahashi. "Practical Time-Scale Fitting of Self-Similar Traffic with Markov-Modulated Poisson Process", *Telecommunication Systems*, vol. 17/1-2, 185-211, 2001.
17. A.Vidacs, J. Barta, Zs. Kenesi and T. Elteto. "Measurement-Based WWW User Traffic Model for Radio Access Networks", 7th International Workshop on Mobile Multimedia Communications, Tokyo, Japan, 2000.