A Matching Model for MAP-2 using Moments of the Counting Process

Suyong Eum, *RMIT University, Melbourne, Victoria 3000, Australia*
Richard Harris, *Massey University, Palmerston North, Private Bag 11 222, New Zealand*
Irena Atov, *Massey University, Palmerston North, Private Bag 11 222, New Zealand*

Keywords: Markovian Arrival Process, Matching Model, Counting Process, SNMP Measurement Data

1. Introduction

In the context of network capacity planning and management, traffic matrix (TM) estimation plays a crucial role, as it provides the means for computing the volume of traffic that flows between all pairs of nodes in a network. Since the TM estimation problem was first addressed in [13], various techniques for its solution have been introduced based on observed traffic counts on every link in the network. In operational IP networks, these link counts can be obtained from readily available link load measurement data provided by SNMP. If a TM is obtained systematically over a number of fixed time periods, the components of these traffic matrices form a counting process. In existing statistical approaches for TM estimation, this counting process has been modelled by conventional traffic models such as Poisson [13], and Gaussian [5], since these traffic models have attractive theoretical properties. These approaches, however, have been shown to have limited accuracy and cannot be generally applied to practical IP networks [11]. Thus, increasingly, there is a need for new methods that make more accurate assumptions about the traffic characteristics of the flows between OD pairs.

In [6] we proposed a new method for Traffic Matrix (TM) estimation in IP networks that provides an improvement on accuracy over existing methodologies. The proposed method solves the TM estimation problem by dividing it into two main subproblems: 1) First, from a multiple set of link counts obtainable from a consecutive set of SNMP measurements and a given routing matrix, obtain a series of traffic matrices by applying a standard deterministic approach. The components of these matrices represent estimates of the volumes of flows being exchanged between all pairs of nodes at specific measurement points and they form stochastic counting processes, respectively. 2) Then, a Markovian Arrival Process of order two (MAP-2) is applied to model the counting processes formed from this series of estimated traffic matrices.

We now comment on the choice of MAP-2 process for modelling components of the traffic matrix: In order to obtain efficient algorithms for capacity planning and management, the network planner has to consider analytically tractable traffic descriptors. Tractability also has to be preserved in the parameter matching procedures, required for characterisation of the traffic inputs based on real traffic measurement data. In [2] Markovian Arrival Processes (MAPs) have been successfully applied in the context of IP network planning and it has been shown that these traffic descriptors provide a very good balance between modelling accuracy (i.e., the capability of capturing burstiness and correlation characteristics of real IP traffic) and efficiency – as provided by powerful matrix-analytic computational procedures for queueing systems. Furthermore, to support this planning model in [3], a moment based matching method was developed to approximate bursty and correlated traffic inputs by a MAP-2 process. It has been demonstrated that the MAP-2, together with their proposed matching method, can model bursty and correlated traffic more accurately than conventional MMPP-2 models with respect to the observed queueing behaviour. However, to populate this model one needs to collect empirical data of inter-arrival times on links in the network and this is more difficult to apply in our proposed TM framework [6], since this relies on link counts being provided as input.

In this paper, we develop a matching model for a MAP-2 process that is purely based on *the moments of the counting processes*. Although the model is more generally applicable, our efforts were driven by the need to provide a solution for Step 2 of the above-mentioned TM estimation framework. In this context, we focus on capturing the dynamics of large aggregates of traffic, rather than those of individual flows. To validate the accuracy of the proposed matching model, we employ a simulation study by using a synthetic correlated traffic, as well as real network traffic obtained from two backbone networks as input into a single-sever queue, served in a first-in-first-out (FIFO) fashion. The queue waiting time and queue length performance characteristics were measured when both the original and the approximate MAP-2 processes are fed into the queue in order to see if the approximate model matches well the impact that real traffic has on a queueing performance. The results showed a good accuracy of the proposed model across a range of traffic load scenarios.

The rest of this paper is organized as follows: In Section 2, first the MAP process is introduced with its properties. Then, the proposed matching model for a MAP-2 based on the moments of a counting process is presented. In Section 3, a simulation scenario is presented for evaluation of the accuracy of the proposed matching model along with the validation results. Finally, concluding remarks are presented in Section 5.

2. AN APPROXIMATE MODEL FOR THE CORRELATED TRAFFIC

A MAP is a process which counts transitions of a *finite* continuous-time Markov chain (CTMC) with m states. The size m is called the order of the MAP, and determines the dimensions of matrices D_0 and D_1:

$$D_0 = \begin{bmatrix} -q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & -q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & -q_{mm} \end{bmatrix}, \quad D_1 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \tag{1}$$

where $q_{ii} = \sum_{j=1,j\neq i}^{m} q_{ij} + \sum_{j=1}^{m} a_{ij}$. D_0 and D_1 represent the transition rates of the MAP process and define the infinitesimal generator $Q = D_0 + D_1$. D_0 has negative diagonal elements and non-negative off-diagonal elements governing the transitions of the arrival process which do not produce an arrival and D_1 is a non-negative matrix, with elements governing the transitions of the arrival process which produce an arrival. MAP-2 process is uniquely defined by six parameters because the negative diagonal elements of D_0 are expressed using the remaining parameters: $q_{11} = q_{12} + a_{11} + a_{12}$ and $q_{22} = q_{21} + a_{21} + a_{22}$. The steady state probability vector π for a MAP-2 process is defined as follows.

$$\pi(1,1) = \frac{a_{21} + q_{21}}{a_{21} + q_{21} + a_{12} + q_{12}}, \quad \pi(1,2) = \frac{a_{12} + q_{12}}{a_{21} + q_{21} + a_{12} + q_{12}} \tag{2}$$

The six parameter characterisation of a MAP-2 process permits finer control for capturing the properties of the original process that heavily influence the performance of a queueing system, as opposed to four parameter characterisation (MMPP-2) or one (Poisson) and, thus, it can achieve a much closer match to the original process.

We denote the counting process by $N(t)$, representing the number of arrivals in the interval (0,t]. Our proposed matching model for a MAP-2 process uses the following six characteristic parameters of the counting process $N(t)$ *viz*:
1. Mean arrival rate - $E[N(t)]$,
2. Index of dispersion for counts - $I(t)$,
3. Limiting index of dispersion for counts - $I(\infty)$,
4. Squared coefficient of variation of the counts - c^2,
5. Covariance of the counts in intervals $(0,t]$ and $(t,2t]$ - $C(t)$,
6. Third moment of the counts - $\mu_3(t)$.

In [8] [12], the mean $E[N(t)]$ and the variance $Var[N(t)]$ of the number of arrivals in $(0,t]$ for a MAP process were given by:

$$\begin{aligned} E[N(t)] &= \pi D_1 e \cdot t \\ Var[N(t)] &= (1 + 2\pi D_1 e)E[N(t)] - 2\pi D_1(e\pi + Q)^{-1}D_1 e \cdot t - 2\pi D_1(I - \mathbf{e}^{Qt})(e\pi + Q)^{-2}D_1 e\,. \end{aligned} \tag{3}$$

where e, **e**, and I are a column vector of ones of appropriate dimension, the base of the natural logarithm, and the identity matrix, respectively. Thus, for the ratio of the variance and the mean of the number of arrivals in

$(0, t]$, namely the index of dispersion for counts, one obtains:

$$I(t) = \frac{Var[N(t)]}{E[N(t)]} = 1 + 2\pi D_1 e - \frac{2\pi D_1 (e\pi + Q)^{-1} D_1 e}{\pi D_1 e} - \frac{2\pi D_1 (I - \mathbf{e}^{Qt})(e\pi + Q)^{-2} D_1 e}{\pi D_1 e t} \ . \quad (4)$$

For $t \to \infty$, the limiting index of dispersion for counts becomes:

$$I(\infty) = \lim_{t\to\infty} \frac{Var[N(t)]}{E[N(t)]} = 1 + 2\pi D_1 e - \frac{2\pi D_1 (e\pi + Q)^{-1} D_1 e}{\pi D_1 e} \ . \quad (5)$$

The squared coefficient of variation of $N(t)$ is given by:

$$\begin{aligned} c^2 &= \frac{E[(N(t))^2]}{(E[N(t)])^2} = \frac{Var[N(t)] + (E[N(t)])^2}{(E[N(t)])^2} \quad (6) \\ &= \frac{(1 + 2\pi D_1 e)\pi D_1 e \cdot t - 2\pi D_1 (e\pi + Q)^{-1} D_1 e \cdot t - 2\pi D_1 (I - \mathbf{e}^{Qt})(e\pi + Q)^{-2} D_1 e + (\pi D_1 e \cdot t)^2}{(\pi D_1 e \cdot t)^2} \ . \end{aligned}$$

The covariance of $N(t)$ in intervals (0,t] and (t,2t] is given by [12]:

$$C(t) = \pi (M_1(t))^2 e - (\pi M_1(t) e)(\pi \mathbf{e}^{Qt} M_1(t) e) \quad (7)$$

where $M_1(t)$ represents the first moment matrix of the counts during an interval $(0, t]$ and is given by:

$$M_1(t) = E[N(t)]e\pi + (e\pi - Q)^{-1} D_1 e\pi + e(\pi D_1 (e\pi - Q)^{-1}) - 2\frac{E[N(t)]}{t_0} e\pi$$

Lastly, the third moment of the counts, $\mu_3(t)$, can be easily derived from the third factorial moment, $\nu_3(t)$, as shown below:

$$\mu_3(t) = E[(N(t) - E[N(t)])^3] = E[(N(t))^3] - 3E[(N(t))^2]E[N(t)] + 2(E[N(t)])^3 \ . \quad (8)$$

In [9], the third factorial moment, $\nu_3(t)$, for the counting process was given in terms of the six parameters of MAP-2, as follows:

$$\nu_3(t) = E[N(t)(N(t) - 1)(N(t) - 2)] = E[(N(t))^3] - 3E[(N(t))^2] + 2E[N(t)] \quad (9)$$

$$E[(N(t))^3] = \nu_3(t) + 3E[(N(t))^2] - 2E[N(t)] \quad (10)$$

After some algebraic transformation, for the third moment of the counts one obtains:

$$\begin{aligned} \mu_3(t) &= \nu_3(t) + 3E[(N(t))^2] - 2E[N(t)] - 3E[(N(t))^2]E[N(t)] + 2(E[N(t)])^3 \quad (11) \\ &= \nu_3(t) + 3((1 + 2\pi D_1 e)\pi D_1 e \cdot t - 2\pi D_1 (e\pi + Q)^{-1} D_1 e \cdot t \\ &\quad -2\pi D_1 (I - \mathbf{e}^{Qt})(e\pi + Q)^{-2} D_1 e + (\pi D_1 e \cdot t)^2) - 2(\pi D_1 e \cdot t) \\ &\quad -3((1 + 2\pi D_1 e)\pi D_1 e \cdot t - 2\pi D_1 (e\pi + Q)^{-1} D_1 e \cdot t \\ &\quad -2\pi D_1 (I - \mathbf{e}^{Qt})(e\pi + Q)^{-2} D_1 e + (\pi D_1 e \cdot t)^2)(\pi D_1 e \cdot t) + 2(\pi D_1 e \cdot t)^3 \ . \end{aligned}$$

The above six characteristics (i.e., the values on the left-hand side of equations (3), (4), (5), (6), (7) and (11)) can be easily derived from SNMP-based measurement data. The expression of these six characteristics and their respective measurement values form a nonlinear system of equations which can be solved to obtain the unknown parameters of the MAP-2 process. Note that, in order to solve this system, one has to also specify a time point t for the interval of the counts. In [7], the choice of the time t for their matching model for MMPP-2 was discussed in detail. They chose a time point based on the best fit to the variance-time curve. We have followed the same approach for obtaining a good choice for the time t in our model.

The closed-form expressions of the chosen six characteristics of MAP-2 are quite complex and do not lend themselves easily to efficient inversion procedures for derivation of the MAP-2 parameter set. Therefore, for the solution of the nonlinear system of equations, denoted as $\mathbb{C}_i = \mathbb{F}_i(\mathbf{P})$ where $i = (1, 2, \ldots, 6)$, we use the approach of solving the following minimisation problem:

$$\textbf{Minimise} \quad \sum_{i=1}^{6} [\mathbb{F}_i(\mathbf{P}) - \mathbb{C}_i]^2, \qquad \text{s.t.} \quad \mathbf{P} \geq 0 \tag{12}$$

where $\mathbb{C}_i$ represent the statistical characteristics collected from the original process, and $\mathbf{P} = (p_1, p_2, \ldots, p_6)$ is a vector of the six unknown parameters defining the approximate MAP-2 process. For the solution of the above minimisation problem we used Matlab optimisation toolbox and we focused on finding a local minimum rather than employing a global solution approach. Searching for a global minimum would significantly increase the computational effort and, for the purposes of our model, this is regarded impractical [1]. Specifically, the nonlinear optimisation program was solved using a Matlab function called `fminsearch`. The function is used for solving unconstrained nonlinear optimisation problems and it starts its search using a solution from an initial estimate given as an argument. The general approach we employed is to gradually increment the initial estimate by a small step size, and use it with `fminsearch` until a local minimum is found.

3. RESULTS AND DISCUSSION

To validate the accuracy of the proposed matching model a purpose built queueing simulator was developed using C++. Evaluation of the performance of the matching model is performed by applying the following three-step procedure. First, we perform a simulation study of a single server FIFO queue (with Erlang-2 service time distribution) by feeding various traffic arrival processes and measure the corresponding queueing performance characteristics. Then, we run the MAP-2 matching model to derive the respective approximate processes for each of the original traffic inputs and feed these into the queueing simulator. Finally, we compare the queue waiting time and queue length statistics obtained from the simulator with both the original and approximate processes in order to examine the capability of the MAP-2 matching model to deliver the target queuing performance.

In our evaluation method, we first generate synthetic "real-network like" correlated traffics by feeding a mixture of n MAP-2 processes (n=2,4,8) as component arrival processes. This was motivated by the illustration in [1], that a superposition of four MAP-2 processes can capture well second-order self-similar behaviour over several time scales. Ten different MAP-2 component processes were used, each with varying burstiness and correlation parameters, to create 20 different test scenarios. Appropriate scaling method was applied to vary the traffic utilization at the queue for eight different values starting from 0.25 to 0.95 for a step size of 0.05. In addition, we use a real traffic data obtained from two backbone networks. Specifically, we use data traces monitored from an Internet uplink operated by a major ISP in New Zealand [4], as well as from a backbone link of the WIDE national network in Japan [10]. The first trace is from the Auckland-II data set generated by NLANR MOAT and the WAND research groups. This data trace was obtained from an OC-3 ATM link, which connects The Auckland University to the global Internet, and the duration of this trace run is three hours. The second data trace is obtained from the WIDE backbone traffic on an OC-12 link, during the daily peak hours from about 14:00 on January 1, 2005, lasting for 15 minutes. Both packet traces include GPS time stamps in each packet header allowing us to compute the inter-arrival time of packets flowing on these links at fine time granularities.

The first set of results pertaining to the correlated traffic test scenarios are summarised in Table 1 as averages over each set of traffic inputs at a given link utilization value. The table shows the relative percentage errors (RE) for the waiting time and queue length statistics obtained from the approximate model with respect to the ones incurred by the real traffic as a function of the link utilization (in the range $0.25 - 0.95$) for 2, 4 and 8 superimposed MAP-2 processes, respectively, The REs are defined as $\mathrm{RE} = 100(\mathrm{approximate} - \mathrm{real})/\mathrm{real}$ and the average absolute RE (ARE) for each link load, respectively, are computed as $\mathrm{ARE} = \sum_{\mathrm{i}=1}^{20} |\mathrm{RE_i}|/20$.

The results demonstrate a very close match of the performance statistics when both the correlated traffic and

[1]Note that traditionally local minimum solution approach has been applied in parameter estimation techniques for MMPP based on maximum likelihood methods.

the approximated MAP-2 traffic are fed into the queueing system.

	2*MAP2			4*MAP2			8*MAP2		
ρ	E[W]	V[W]	E[Q]	E[W]	V[W]	E[Q]	E[W]	V[W]	E[Q]
0.25	3.64	7.00	6.52	0.22	15.11	1.92	1.50	0.17	1.76
0.35	3.90	3.30	5.47	1.02	20.47	3.21	2.66	4.26	2.94
0.45	2.89	5.91	3.47	0.65	23.54	2.74	2.46	15.01	2.72
0.55	1.95	6.91	2.19	0.60	20.51	2.84	3.29	26.03	3.40
0.65	0.66	8.05	0.22	0.57	13.86	1.87	5.23	21.70	5.27
0.75	0.15	2.13	1.44	0.46	8.60	0.28	0.85	26.80	1.07
0.85	2.51	6.83	5.15	0.79	21.55	2.14	0.30	8.32	0.23
0.95	5.25	2.59	8.71	0.87	17.74	3.03	4.88	2.31	5.30
ARE	2.62	5.34	4.15	0.62	17.67	2.25	2.65	13.08	2.84

Table 1: RE(%) for the mean and the variance of the queue waiting time and the mean queue length

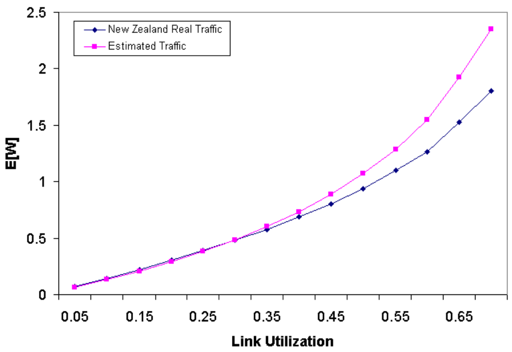

Figure 1: The mean queue waiting time for Auckland II traffic trace. The ARE(%) across the (0.05-0.7) utilisation is 11.44%.

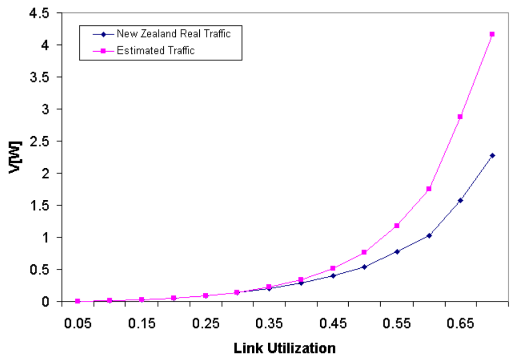

Figure 2: The variance of queue waiting time for Auckland II traffic trace. The ARE(%) across the (0.05-0.7) utilisation is 33.13%.

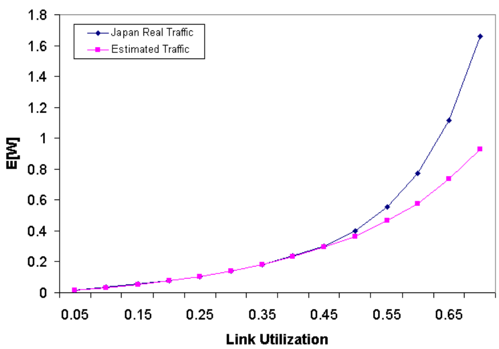

Figure 3: The mean queue waiting time for WIDE traffic trace. The ARE(%) across the (0.05-0.7) utilisation is 10.72%.

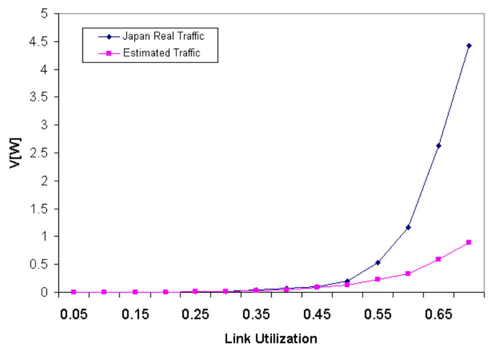

Figure 4: The variance of queue waiting time for WIDE traffic trace. The ARE(%) across the (0.05-0.7) utilisation is 41.46%.

To verify whether our model performs well with actual Internet traffic, in the second set of test scenarios in our evaluation study, we considered real IP traffic as input processes into the analyzed queueing system. Figures 1, 2, 3, and 4 give a summary of the queue waiting time results obtained from data traces Auckland II (the first two) and WIDE (the last two), respectively. It can be seen, that the model matches well the impact that the real traffic has on a queueing performance across traffic intensities that fall in the range (0.05-0.5) and for traffic intensities above 0.5 the accuracy of the model starts to decrease slightly. Note, however, that across the whole range (0.05-0.7) the accuracy of the prediction of the mean waiting time is quite good with ARE = 11.44% and 10.72% for the two traffic traces, respectively. Given that operational networks are designed and managed with a maximum link utilization as a constraint and its value is often in the 0.7 range, we conclude that the proposed model is capable of capturing and conveying the queueing performance of the original processes well in practical applications.

4. CONCLUSION

A new matching method has been developed based on the moments of the counting process, which takes six characteristics of an aggregate traffic and translates them into an equivalent MAP-process by solving a minimisation problem. This model was specifically built for use in the context of traffic matrix estimation and, consequently, it takes at input traffic count statistics which can be readily obtained from SNMP system measurement data.

The accuracy of the proposed model has been evaluated using a queueing simulation approach using two types of input traffics: synthetic correlated traffic and real traffic obtained from two ISP networks. The simulation results demonstrated that the proposed matching model is capable of capturing and conveying the queueing performance of the original processes well. Good accuracy was maintained across the range of link traffic intensities typical of operational networks (i.e., up to 70% link loads). Although the accuracy of the model is satisfactory, the nonlinear optimisation program it deploys requires a high computational effort. Further work will focus on refining this solution method so that a more efficient inversion algorithm can be developed for estimation of the MAP-2 parameter set.

References

[1] A. Andersen and B. Nielsen. A Markovian Approach for Modelling Packet Traffic with Long-Range Dependence. *IEEE Journal of Selected Areas in Communications.*, 16(5), June 1998.

[2] I. Atov. *Design of IP Networks with End-to-End Performance Guarantees*. PhD thesis, RMIT University, Melbourne, Australia, 2003.

[3] I. Atov and R. J. Harris. Approximate Model for Merging Markovian Arrival Processes. *NEW2AN'04*, Feb 2004.

[4] Auckland II Data Set, Available:, http://pma.nlanr.net/traces/long/auck2.html/.

[5] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. *J. Amer Stat Assoc*, pages 1063–1075, 2000.

[6] S. Eum, R. J. Harris, and I. Atov. Traffic Matrix Estimation with Markov Arrival Process of the Order of Two (MAP-2). In *International Teletraffic Congress, (ITC20) [Accepted for publication]*, Ottawa, Canada, June 2007.

[7] H. Heffes and D.M. Lucantoni. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE JSAC*, 4(6):856–868, 1986.

[8] A. Heindl. *Traffic-Based Decomposition of General Queueing Networks with Correlated Input Processes*. PhD thesis, Technische Universität, Berlin, German, 2001.

[9] K. Kawanishi. On the Counting Process for a Class of Markovian Arrival Processes with an Application to a Queueing System. *Queueing Syst. Theory Appl*, 49:93–122, 2005.

[10] MAWI Working Group Traffic Archive, Available:, http://tracer.csl.sony.co.jp/mawi/.

[11] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. *ACM SIGCOMM*, August 2002.

[12] S. Narayana and M.F. Neuts. The First Two Moments Matrices of the Counts for the Markovian Arrival Process. *Stochastic Models*, 8:459–477, 1992.

[13] Y. Vardi. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *J. Amer Stat Assoc*, pages 365–377, 1996.