dMAPAR-HMM: Reforming Traffic Model for Improving Performance Bound with Stochastic Network Calculus

Qingqing Yang, Xi Peng, Huiwen Yang, Gong Zhang, and Bo Bai
Theory Lab, Central Research Institute, 2012 Labs, Huawei Technologies Co. Ltd.
Hong Kong SAR, China

***Abstract*—A popular branch of stochastic network calculus (SNC) utilizes moment-generating functions (MGFs) to characterize arrivals and services, which enables end-to-end performance analysis. However, existing traffic models for SNC cannot effectively represent the complicated nature of real-world network traffic such as dramatic burstiness. To conquer this challenge, we propose an adaptive spatial-temporal traffic model: dMAPAR-HMM. Specifically, we model the temporal on-off switching process as a dual Markovian arrival process (dMAP) and the arrivals during the on phases as an autoregressive hidden Markov model (AR-HMM). The dMAPAR-HMM model fits in with the MGF-SNC analysis framework, unifies various state-of-the-art arrival models, and matches real-world data more closely. We perform extensive experiments with real-world traces under different network topologies and utilization levels. Experimental results show that dMAPAR-HMM significantly outperforms prevailing models in MGF-SNC.**

***Index Terms*—Network Traffic Modeling, Stochastic Network Calculus, Performance Evaluation, Performance Bound**

I. Introduction

Performance analysis, especially the evaluation of performance bound, is essential for supporting sufficient quality of service (QoS) for delay-sensitive network applications. Network calculus (NC) provides an ascendant methodology focusing on computing the performance bounds, i.e., the delay and backlog bounds. There exist two branches in NC: deterministic network calculus (DNC) [1], [2], [3] and stochastic network calculus (SNC) [4], [5], [6], [7], [8], [9], [10], [11], [12]. DNC relies on deterministic models to analyze the strict worst-case performance, which usually results in quite low network utilization. However, most commercial network services allow statistical multiplexing of traffic flows to enhance network utilization, and prefer probabilistic QoS metrics. Then SNC has been proposed by considering statistical behaviors to calculate the probabilistic performance bounds at a known small violation probability. Figure 1 shows the framework of MGF-based SNC. Network traffic characteristics are extracted by the feature extraction module, and then are used by the performance evaluation module based on the MGF-SNC framework. The results of performance evaluation will be fed back to the center controller for further planning and optimization.

978-3-903176-58-4 ©2023 IFIP

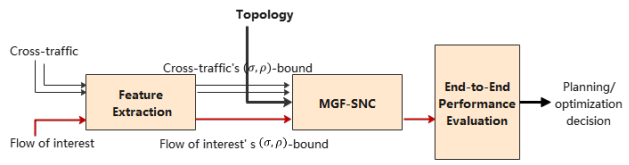


Figure 1: The workflow of end-to-end performance evaluation in MGF-SNC.

To investigate traffic flows, the analysis of inter-arrival times (IATs) and packet sizes is the basic step. If they are found to follow certain probability distributions, the corresponding stochastic model of these flows can be determined. In MGF-SNC, packet sizes (or arrivals) and IATs are commonly assumed to conform to conventional distributions, including the Normal, Exponential, et.al. However, our analysis and extensive past attempts [13], [14], [15], [16], [17], [18] reveal that none of the conventional distributions are accurate enough for real network flows.

As early as the 1970's, [19] reported a noticeable behavior of traffic, which is called "burstiness" defined by peak to average transmission rate. It implies that an on-off switching pattern widely exists in real traffic flows. The early literature quantitatively describes real flows from a fractal perspective [13], [20], [21], [22] and has been subsequently applied by [16], [23], [24], [15]. The fractal models convincingly reveal traffic characteristics, such as self-similarity (SS) and long-range dependence (LRD). An important issue with them is that their MGFs grow super-linearly with time and hence the underlying SNC results for MGFs are not directly applicable [25].

To tackle the challenge in traffic modeling, we introduce a novel spatial-temporal model: dMAPAR-HMM, as illustrated in Figure 2. Inspired by the definition of modulation in telecommunications, we regard the arrival process as a transmission signal that can be demodulated into a carrier signal and an input signal of positive impulses. Specifically, the carrier signal is modeled by a dual Markovian arrival process (dMAP) to depict the (temporal) on-off switching process observed in real traffic traces, and the input signal is modeled by an autoregressive hidden Markov model (AR-HMM) for the

representation of the (spatial) data amount during *on* phases. We show that various state-of-the-art (SOTA) arrival models used in SNC can be readily represented by the proposed dMAPAR-HMM model. With traffic traces from real networks, we demonstrate that dMAPAR-HMM can accurately depict multi-dimensional multi-order traffic characteristics.

For providing performance analysis, we develop the $(\sigma(\theta), \rho(\theta))$-envelope of the MGF of dMAPAR-HMM, which can fit in with the framework of MGF-SNC. Though here we focus on deriving the delay bound with dMAPAR-HMM, the results can be extended without much effort to the backlog bound. Besides the $(\sigma(\theta), \rho(\theta))$-envelope of MGF, more general envelopes, such as stochastically bounded burstiness envelope, could also be considered similarly [26]. By extensive experiments under different network topologies and utilization levels, we show that dMAPAR-HMM can significantly boost the effectiveness of MGF-SNC, which means that both the tightness and the reliability of the delay bound are enhanced.

The rest of the paper is organized as follows. In Section II, the spatial-temporal model dMAPAR-HMM is introduced. In Section III, we present the interface of the proposed traffic model to MGF-SNC for network traffic. In Section IV, the feasibility of dMAPAR-HMM is examined with real-world traces. Section V concludes the article.

II. Network Traffic

In this section, we introduce the proposed traffic model dMAPAR-HMM in details by showing how it captures the temporal and spatial dynamics of real-world traffic. We also reveal that many famous models are correlated to dMAPAR-HMM, and show that our model favors the retention of critical traffic features.

A. Traffic Modeling

We divide the time into discrete intervals to conform to the SNC approach. We use $A(s,t)$ for the amount of data traffic arriving in the interval $(s,t]$, i.e.,

$$A(s,t) = \sum_{k=s+1}^{t} a_k,$$

where we adopt the convention $A(t,t) = 0$. For ease of expression, we use $A(t)$ to represent $A(0,t)$. In addition, we use a_k to represent data traffic arriving during the k-th timeslot, that is, $a_k = A(k-1,k) = A(k) - A(k-1)$.

In the real world, traffic flows exhibit on-off switching patterns. Moreover, arrivals of all timeslots are difficult to be entirely modeled as a single stochastic process. Therefore, we are motivated to parse the traffic flow and build an appropriate model for each component. In this paper, we present a discrete-time spatial-temporal model dMAPAR-HMM for traffic flows delivered in real networks. As shown in Figure 2, a traffic flow is discretized into a time series $\{a_t\}, t = 1, 2, \cdots$, based on a constant timeslot interval Δt. The time series $\{a_t\}$ can be considered to be formed by modulating an input signal of positive pulses by a carrier signal with an on-off switching scheme and unit amplitude. The input signal, i.e., the

Symbol	Meaning
$\mathcal{S}^{\text{on}}$	state space of MAP^{on}
$\mathcal{S}^{\text{off}}$	state space of MAP^{off}
$\mathcal{N}$	state space of X
$\mathcal{S}$	state space of Z
Q	transition probability matrix of the carrier signal in the embedded state space
P	transition probability matrix of X
T	transition probability matrix of Z
X	hidden state of the input signal
Z	hidden state of the dMAPAR-HMM signal
Δt	timeslot interval for discretization
$\circ$	Hadamard product
$\otimes$	Kronecker product
I_n	Identity matrix of size n

Table I: Frequently used symbols in this paper.

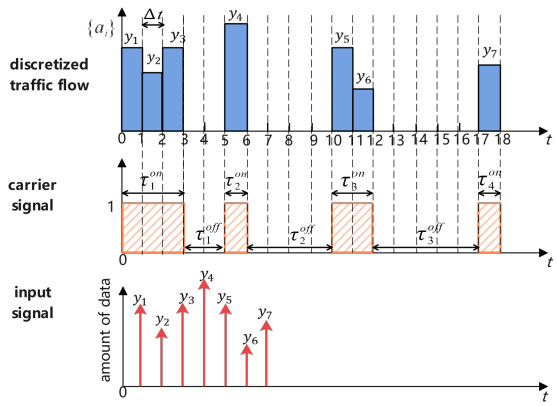


Figure 2: **dMAPAR-HMM.** Mapping the discrete-time observations into a spatial-temporal model, regarding the temporal on-off switching process as the carrier signal, and the data transmitted in on phases as the input signal.

spatial model, $\{y_1, y_2, \cdots\}$, contains the amount of non-zero arrivals along $\{a_t\}$. The carrier signal, i.e., the temporal model, represents the temporal feature by recording the IATs of both on and off phases, i.e., $\{\tau_1^{\text{off}}, \tau_2^{\text{off}}, \cdots\}$ and $\{\tau_1^{\text{on}}, \tau_2^{\text{on}}, \cdots\}$, which are supplementary to each other. We will subsequently discuss the temporal and spatial models in details. Table I lists the frequently used symbols in this paper.

1) Temporal Model: dMAP: In general, the temporal feature of a flow is quite challenging to characterize, and we propose a novel dMAP to enable an effective modeling methodology. The dMAP model consists of two independent continuous-time (or discrete-time) MAPs that randomly switch between each other. The two MAPs stand for the inter-arrival processes of the off and on phases, respectively. We choose MAPs for temporal modeling for two reasons. First, MAP is regarded as one of the most expressive models for inter-arrival processes. Second, it is analytically tractable by using matrix-analytic methods. The dMAP model further enhances the representation capacity of the conventional MAP, and allows parallel computing for parameter estimation of two MAPs. In our case, the carrier signal is controlled by dMAP, where $\{\tau_1^{\text{off}}, \tau_2^{\text{off}}, \cdots\}$ and $\{\tau_1^{\text{on}}, \tau_2^{\text{on}}, \cdots\}$ are modeled by two independent MAPs – $\text{MAP}^{\text{off}}(m_1)$ and $\text{MAP}^{\text{on}}(m_2)$ – with

Case	1-step transition probability
	$(1,i,j) \curvearrowright (1,k,j)$: $\mathrm{MAP}^{\mathrm{off}}$ is frozen in state j, and $\mathrm{MAP}^{\mathrm{on}}$ makes a hidden transition from state i to k.
	$(1,i,j) \curvearrowright (0,i,k)$: $\mathrm{MAP}^{\mathrm{on}}$ is frozen in state i, and $\mathrm{MAP}^{\mathrm{off}}$ makes an observable transition from state j to k.
	$(0,i,j) \curvearrowright (0,i,k)$: $\mathrm{MAP}^{\mathrm{on}}$ is frozen in state i, and $\mathrm{MAP}^{\mathrm{off}}$ makes a hidden transition from state j to k.
	$(0,i,j) \curvearrowright (1,k,j)$: $\mathrm{MAP}^{\mathrm{off}}$ is frozen in state j, and $\mathrm{MAP}^{\mathrm{on}}$ makes an observable transition from state i to k.

Table II: State transition of the carrier signal.

$$
Q = \begin{array}{c|cccc|cccc}
 & (0,0,0) & (0,0,1) & (0,1,0) & (0,1,1) & (1,0,0) & (1,0,1) & (1,1,0) & (1,1,1) \\
\hline
(0,0,0) & 1-\nu_0\Delta t & \nu_{01}\Delta t & & & \frac{\nu'_{02}(\nu_{02}+\nu_{03})\Delta t}{\nu'_{02}+\nu'_{03}} & & \frac{\nu'_{03}(\nu_{02}+\nu_{03})\Delta t}{\nu'_{02}+\nu'_{03}} & \\
(0,0,1) & \nu_{10}\Delta t & 1-\nu_1\Delta t & & & & \frac{\nu'_{02}(\nu_{12}+\nu_{13})\Delta t}{\nu'_{02}+\nu'_{03}} & & \frac{\nu'_{03}(\nu_{12}+\nu_{13})\Delta t}{\nu'_{02}+\nu'_{03}} \\
(0,1,0) & & & 1-\nu_0\Delta t & \nu_{01}\Delta t & \frac{\nu'_{12}(\nu_{02}+\nu_{03})\Delta t}{\nu'_{12}+\nu'_{13}} & & \frac{\nu'_{13}(\nu_{02}+\nu_{03})\Delta t}{\nu'_{12}+\nu'_{13}} & \\
(0,1,1) & & & \nu_{10}\Delta t & 1-\nu_1\Delta t & & \frac{\nu'_{12}(\nu_{12}+\nu_{13})\Delta t}{\nu'_{12}+\nu'_{13}} & & \frac{\nu'_{13}(\nu_{12}+\nu_{13})\Delta t}{\nu'_{12}+\nu'_{13}} \\
\hline
(1,0,0) & \frac{\nu_{02}(\nu'_{02}+\nu'_{03})\Delta t}{\nu_{02}+\nu_{03}} & \frac{\nu_{03}(\nu'_{02}+\nu'_{03})\Delta t}{\nu_{02}+\nu_{03}} & & & 1-\nu'_0\Delta t & & \nu'_{01}\Delta t & \\
(1,0,1) & \frac{\nu_{12}(\nu_{02}+\nu_{03})\Delta t}{\nu_{12}+\nu_{13}} & \frac{\nu_{13}(\nu_{02}+\nu_{03})\Delta t}{\nu_{12}+\nu_{13}} & & & & 1-\nu'_0\Delta t & & \nu'_{01}\Delta t \\
(1,1,0) & & & \frac{\nu_{02}(\nu'_{12}+\nu'_{13})\Delta t}{\nu_{02}+\nu_{03}} & \frac{\nu_{03}(\nu'_{12}+\nu'_{13})\Delta t}{\nu_{02}+\nu_{03}} & \nu'_{10}\Delta t & & 1-\nu'_1\Delta t & \\
(1,1,1) & & & \frac{\nu_{12}(\nu'_{12}+\nu'_{13})\Delta t}{\nu_{12}+\nu_{13}} & \frac{\nu_{13}(\nu'_{12}+\nu'_{13})\Delta t}{\nu_{12}+\nu_{13}} & & \nu'_{10}\Delta t & & 1-\nu'_1\Delta t
\end{array} \tag{1}
$$

a switching scheme.

Let us denote MAP(m) for an m-dimensional MAP which is governed by an underlying continuous-time Markov chain (CTMC) with state space $\mathcal{S} = \{0, \cdots, m-1\}, m \geq 1$. Let $\mathbf{C}_0$ and $\mathbf{C}_1$ denote the matrices of state transition rates with zero and one arrival, respectively. The following restrictions apply to the $\mathbf{C}_i$:

$$
\begin{aligned}
0 \quad &\leq \quad [\mathbf{C}_1]_{ij} \quad < \infty \\
0 \quad &\leq \quad [\mathbf{C}_0]_{ij} \quad < \infty \quad i \neq j \\
&\quad\;\; [\mathbf{C}_0]_{ii} \quad < 0 \\
&(\mathbf{C}_0 + \mathbf{C}_1)\mathbf{1} \quad = \mathbf{0},
\end{aligned}
$$

where $\mathbf{1}$ denotes an all-one column vector with an appropriate dimension. In total, there are $2m^2 - m$ free parameters. The model parameters can be estimated using the algorithms listed in [17], [27]. To accommodate to the SNC framework, we shall discretize $\mathrm{MAP}^{\mathrm{off}}(m_1)$ and $\mathrm{MAP}^{\mathrm{on}}(m_2)$ with Δt. The timeslot interval Δt should be small enough so that the probability of having two or more events occurring within Δt can be neglected. Consider a continuous-time MAP(2) with the matrix representation

$$
\mathbf{C}_0 = \begin{bmatrix} -\nu_0 & \nu_{01} \\ \nu_{10} & -\nu_1 \end{bmatrix} \quad \text{and} \quad \mathbf{C}_1 = \begin{bmatrix} \nu_{02} & \nu_{03} \\ \nu_{12} & \nu_{13} \end{bmatrix},
$$

where ν_i is the sum rate at which the process makes a transition from state i, and ν_{ij} is the rate at which the process makes a transition from state i into state j. The transition probability matrices of the corresponding discrete-time MAP are

$$
\mathbf{D}_0 = \begin{bmatrix} 1-\nu_0\Delta t & \nu_{01}\Delta t \\ \nu_{10}\Delta t & 1-\nu_1\Delta t \end{bmatrix} \text{ and } \mathbf{D}_1 = \begin{bmatrix} \nu_{02}\Delta t & \nu_{03}\Delta t \\ \nu_{12}\Delta t & \nu_{13}\Delta t \end{bmatrix},
$$

and the row sum of $\mathbf{D}_0 + \mathbf{D}_1$ is always one. For space considerations, we will use a set of model parameters $\mathbf{D} = \{(\nu_i, \nu_{i,0}, \cdots, \nu_{i,i-1}, \nu_{i,i+1}, \cdots, \nu_{i,2m-1})_{i=0,\cdots,m-1}; \Delta t\}$ to represent the discrete-time MAP.

As aforementioned, the inter-arrival processes of the on and off phases are governed by $\mathrm{MAP}^{\mathrm{off}}(m_1)$ and $\mathrm{MAP}^{\mathrm{on}}(m_2)$ with state spaces $\mathcal{S}^{\mathrm{off}} = \{0, \cdots, m_1 - 1\}$ and $\mathcal{S}^{\mathrm{on}} = \{0, \cdots, m_2 - 1\}$, respectively. Let us use a tuple $(o, s^{\mathrm{on}}, s^{\mathrm{off}})$ to represent the state of the carrier signal at timeslot t, with $o \in \{0, 1\}$ indicating the physical state of the carrier signal (0: off and 1: on) and $(s^{\mathrm{on}}, s^{\mathrm{off}}) \in \mathcal{S}^{\mathrm{on}} \times \mathcal{S}^{\mathrm{off}}$ the hidden state of the carrier signal, and introduce a mapping ι_1 to embed the higher-order state space into a 1-order one:

$$
\iota_1(o, s^{\mathrm{on}}, s^{\mathrm{off}}) = o * m_1 * m_2 + s^{\mathrm{on}} * m_1 + s^{\mathrm{off}}.
$$

Given the matrix representation of $\mathrm{MAP}^{\mathrm{on}}$ and $\mathrm{MAP}^{\mathrm{off}}$, the 1-step transition probability matrix Q of the carrier signal can be put into a compact form,

$$
Q = \left(\begin{array}{c|c} Q_{\mathrm{off}\curvearrowright\mathrm{off}} & Q_{\mathrm{off}\curvearrowright\mathrm{on}} \\ \hline Q_{\mathrm{on}\curvearrowright\mathrm{off}} & Q_{\mathrm{on}\curvearrowright\mathrm{on}} \end{array} \right),
$$

where $Q_{\mathrm{off}\curvearrowright\mathrm{off}}$, $Q_{\mathrm{off}\curvearrowright\mathrm{on}}$, $Q_{\mathrm{on}\curvearrowright\mathrm{off}}$, and $Q_{\mathrm{on}\curvearrowright\mathrm{on}}$ are four $(m1 * m2) \times (m1 * m2)$ block matrices, governing the on-off switching process of the carrier signal. All possible state transitions of Q are summarized in Table II.

As an illustration, we show how to compute Q in the following example. Suppose we have $\mathrm{MAP}^{\mathrm{off}}(2)$ with $\mathbf{D}^{\mathrm{off}} = \{\nu_0, \nu_{01}, \nu_{02}, \nu_{03}; \nu_1, \nu_{10}, \nu_{12}, \nu_{13}; \Delta t\}$ and $\mathrm{MAP}^{\mathrm{on}}(2)$ with $\mathbf{D}^{\mathrm{on}} = \{\nu'_0, \nu'_{01}, \nu'_{02}, \nu'_{03}; \nu'_1, \nu'_{10}, \nu'_{12}, \nu'_{13}; \Delta t\}$. The 1-step transition probability matrix Q is shown in Eqn. (1).

2) Spatial Model: AR-HMM: We investigate the models for characterizing the spatial feature and find that the autoregressive hidden Markov model (AR-HMM) [30] is suitable for this job. Suppose X is a finite state homogeneous Markov chain. Let $\mathcal{N} = \{0, 1, \cdots, N-1\}$ be the state space of X. The observable process y has the form

$$
y_{t+1} = \mu(X_t) + \sum_{i=1}^{p} \phi_i(X_t) y_{t+1-i} + \sigma(X_t)\epsilon_{t+1}.
$$

Model	Architecture
MAP [17], [27], [28]	disable $\mathrm{MAP}^{\mathrm{on}}$ + AR-HMM($p=0, N=1, \mu=1, \sigma=0$)
BMAP [29]	disable $\mathrm{MAP}^{\mathrm{on}}$ + AR-HMM($p=0, N=m_1\times\{\text{maximal batch size}\}, \sigma=0$)
Poisson [11], [10]	disable $\mathrm{MAP}^{\mathrm{on}}$ + $\mathrm{MAP}^{\mathrm{off}}(m_1=1)$ + AR-HMM($p=0, N=1, \mu=1, \sigma=0$)
cPoisson	disable $\mathrm{MAP}^{\mathrm{on}}$ + $\mathrm{MAP}^{\mathrm{off}}(m_1=1)$ + AR-HMM($p=0, N=1$)
MMOO [11], [10]	$\mathrm{MAP}^{\mathrm{off}}(m_1=1)$ + $\mathrm{MAP}^{\mathrm{on}}(m_2=1)$ + AR-HMM($p=0, N=1, \sigma=0$)
MMP [9]	disable $\mathrm{MAP}^{\mathrm{on}}$ & $\mathrm{MAP}^{\mathrm{off}}$ + AR-HMM($p=0$)
AR(p) [9]	disable $\mathrm{MAP}^{\mathrm{on}}$ & $\mathrm{MAP}^{\mathrm{off}}$ + AR-HMM($N=1$, residual="normal")
Normal [9]	disable $\mathrm{MAP}^{\mathrm{on}}$ & $\mathrm{MAP}^{\mathrm{off}}$ + AR-HMM($p=0, N=1$, residual="normal")
Exponential [11], [10], [9]	disable $\mathrm{MAP}^{\mathrm{on}}$ & $\mathrm{MAP}^{\mathrm{off}}$ + AR-HMM($p=0, N=1$, residual="exponential")

Table III: **Popular arrival models widely used in MGF-SNC.** The SOTA arrival models used in SNC can be readily represented by the proposed dMAPAR-HMM model.

Algorithm 1: Online EM algorithm for AR-HMMs.

Input :
- Chain state space $\mathcal{N}=\{0,1,\cdots,N-1\}$;
- Array of initial probabilities $\Pi_0=(\pi_0,\cdots,\pi_{N-1})$ such that π_i stores the probability that $X_0=i$;
- Number of time lags p of the observable process;
- Sequence of observations $(y_0,y_1,\cdots)$.

Output: The state transition matrix P, and the model parameters Θ.

foreach *time t* **do**
  /* E-step. */
  foreach *state $i,j\in\mathcal{N}$* **do**
    $\Xi_t \leftarrow$ Information matrix at time t;
    $b_t \leftarrow \Xi_t\Pi_{t-1}$;
    $Q \leftarrow P^T$;
    $O[:,i] \leftarrow Q\Xi_t O[:,i] + b_t[i]Q[:,i]$;
    $J[:,i,j] \leftarrow Q\Xi_t J[:,i,j] + b_t[i]Q[j,i]\mathbf{e}_j$;
    foreach *lag $l,r\in\{0,1,\cdots,p\}$* **do**
      $F[:,i,l] \leftarrow Q\Xi_t F[:,i,l] + b_t[i]Q[:,i]y_{t-l}$;
      $H[:,i,l,r] \leftarrow Q\Xi_t H[:,i,l,r] + b_t[i]Q[:,i]y_{t-l}y_{t-r}$;
    end
  end
  $\Pi_t \leftarrow Qb_t$.

  /* M-step. */
  foreach *state $i,j\in\mathcal{N}$, lag $l\in\{1,\cdots,p\}$* **do**
    $P[i,j] \leftarrow \frac{J[i,j]}{O[i]}$;
    $\mu[i] \leftarrow \frac{F[i,0]-\sum_{l=1}^{p}\phi[l,i]F[i,l]}{O[i]}$;
    $\sigma[i] \leftarrow \frac{\mu[i]^2O[i]+H[i,0,0]+\sum_{l=1}^{p}\sum_{r=1}^{p}\phi[l,i]\phi[r,i]H[i,l,r]}{O[i]} - \frac{2\mu[i]F[i,0]-2\sum_{l=1}^{p}\phi[l,i]H[i,0,l]}{O[i]}$;
    $\phi[l,i] \leftarrow \frac{H[i,0,l]-\sum_{r\neq l}\phi[r,i]H[i,l,r]}{H[i,l,l]}$.
  end
end

Namely, the $(t+1)_{th}$ observation is influenced by the previous p observations and by the state of X at the previous time t (i.e., the reaction to X_t is not instantaneous).

The key components of an AR-HMM are: (1) N: the number of states of the hidden Markov chain; (2) P: the state transition matrix of the hidden Markov chain; (3) p: the number of time lags of the observable process; and (4) $\Theta=\{(\mu_0,\cdots,\mu_{N-1}),(\phi_{i,0},\cdots,\phi_{i,N-1})_{i=1,\cdots,p},(\sigma_0,\cdots,\sigma_{N-1})\}$: the model parameters of the AR(p)-HMM(N). Suppose that the noise $\{\epsilon_t\}$ is a sequence of i.i.d. random variables (which are, in particular, independent of the X_t). Given $\{y_1,\cdots,y_t\}$, N and p can be chosen according to the AIC/BIC information criteria [31], [32] as well as cross-validation, and P and Θ can be estimated by using the EM-algorithm [33], [34]. Define

$$A(X_t)=\begin{bmatrix}\phi_1(X_t) & 1 & & \\ \phi_2(X_t) & 0 & 1 & \\ \vdots & & \ddots & \\ \phi_p(X_t) & \cdots & & 0\end{bmatrix},\quad C_{(t,t)}=\begin{bmatrix}\phi_1(X_t)\\ \vdots\\ \phi_p(X_t)\end{bmatrix}$$

and $C_{(s,t)}=A(X_s)C_{(s+1,t)}$, $\forall s<t$. For stationary purpose, $\lim_{n\to\infty}C_{(t-n,t)}=0$. When $p=1$ and $N=1$, it is equivalent to $|\phi|<1$.

3) Observation Process: dMAPAR-HMM: Let us denote

$$O_t=1_{\{\text{timeslot } t \text{ is in the on phase}\}} \text{ and } N(t)=\sum_{i\le t}O_i.$$

The observation process a can then be written as $a_t=O_t\cdot y_{N(t)}, t=1,2,\cdots$. The random variable O_t indicates the physical state of the carrier signal at time t, and $y_{N(t)}$ is the amplitude of the input signal at time t.

Let Z be a discrete-time Markov chain whose state space is $\mathcal{S}=\{\iota(o,s^{\mathrm{on}},s^{\mathrm{off}},i)|(o,s^{\mathrm{on}},s^{\mathrm{off}},i)\in\{0,1\}\times\mathcal{S}^{\mathrm{on}}\times\mathcal{S}^{\mathrm{off}}\times\mathcal{N}\}$, where $\iota(o,s^{\mathrm{on}},s^{\mathrm{off}},i)=N*\iota_1(o,s^{\mathrm{on}},s^{\mathrm{off}})+i$. The transition probability matrix of Z is given by

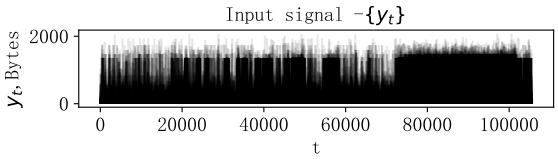
$$T=\left(\begin{array}{c|c}Q_{\mathrm{off}\curvearrowright\mathrm{off}}\otimes I_N & Q_{\mathrm{off}\curvearrowright\mathrm{on}}\otimes P\\ \hline Q_{\mathrm{on}\curvearrowright\mathrm{off}}\otimes I_N & Q_{\mathrm{on}\curvearrowright\mathrm{on}}\otimes P\end{array}\right). \tag{2}$$

Thereby a_t can be rewritten as

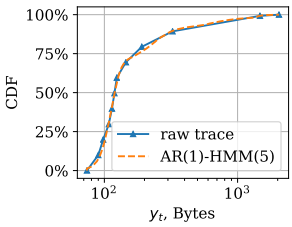
$$\begin{aligned}a_t=\ & \breve{\mu}(Z_{t-1},Z_t)+\sum_{i=1}^{p}\breve{\phi}_i(Z_{t-1},Z_t)y_{N(t-1)+1-i}\\ & +\breve{\sigma}(Z_{t-1},Z_t)\epsilon_t,\end{aligned}$$

where $\breve{g}(Z_{t-1},Z_t)=1_{\{Z_t\ge m_1*m_2*N\}}g_{Z_{t-1}\%N}$, $g\in\{\mu,\phi_1,\cdots,\phi_p,\sigma\}$, and $\%$ is the modulo division.

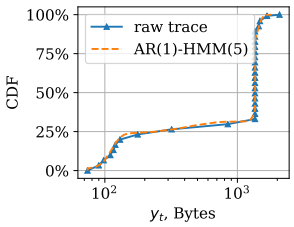
It is worth noticing that many well-known arrival models, such as MAP, batch MAP (BMAP) et.al., are special cases



(a) Spatial feature of the discretized WIDE trace with (src IP, dst IP, protocol)=(203.76.247.112, 204.57.206.30, IPv4) captured at 2022-04-13 08:00. The x-coordinate t is the index of the time series $\{y_t\}$.

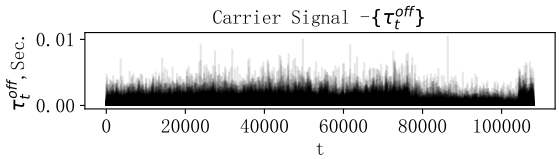


(b) CDF of y_t, $t \in [0, 1e4]$.

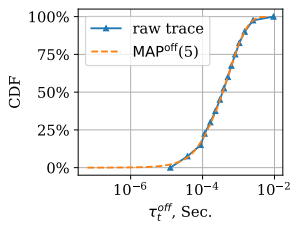


(c) CDF of y_t, $t \in [9e4, 1e5]$.

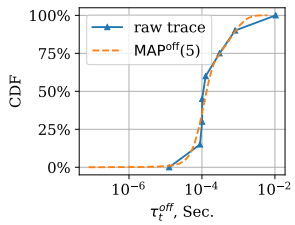
Figure 3: **Spatial feature extraction.** AR-HMM can capture the dynamics of the input signal.



(a) Temporal feature of the discretized WIDE trace as shown in Figure 3. The x-coordinate t is the index of the time series $\{\tau_t^{\text{off}}\}$.



(b) CDF of τ_t^{off}, $t \in [0, 1e4]$



(c) CDF of τ_t^{off}, $t \in [9e4, 1e5]$

Figure 4: **Temporal feature extraction.** MAP can capture the dynamics of the carrier signal.

of the proposed dMAPAR-HMM model (see Table III for more details). Another benefit of dMAPAR-HMM is that it is additive when the residuals conform to additive white Gaussian noise. Namely, if each individual flow is fitted by dMAPAR-HMM, the aggregate flow can also be modeled by dMAPAR-HMM. In view of its strong generalization, dMAPAR-HMM can be used to model various real-world traffic flows, which is, however, a great challenge to many existing models.

B. The Applications in Traffic Feature Extraction

To verify the effectiveness of dMAPAR-HMM in real-world traffic feature extraction, we have conducted experiments on different kinds of traces, such as the multimedia streaming and video conference collected from Huawei commercial and campus networks, as well as open-source WIDE traces from MAWI[1], which is a mixture of commodity applications and research experiments. Due to space limitations and data openness, herein we show the results with typical traffic flows from the WIDE traces.

To accommodate the SNC framework, we adopt a constant timeslot interval of $\frac{mean(IATs)}{40}$ to discretize the time. We fit the discrete-time processes with dMAPAR-HMM and generate synthetic traces using the well-fitted model. Specifically, we use Algorithm 1 to fit the spatial model, and the algorithms listed in [17], [27] to fit the temporal model. As we can see from Figures 3 and 4, the proposed model can well capture the dynamics of the input and the carrier signals. We compare the real and synthetic traces by showing the sample path as well as the accumulated arrivals in Figure 5, which indicates that our model can efficiently learn the arrival patterns from the input traffic. We further compare these traces in terms of the Hurst exponent H and the coefficient of variation CV (see Table IV for more details). It can be concluded from the experimental results that the proposed model can exactly capture the various characteristics of traffic from the decomposed spatio-temporal signals.

[1]The dataset is available at http://mawi.wide.ad.jp/.

III. Interface to MGF-SNC

In this section, we deduce an interface for the spatio-temporal model to MGF-SNC.

An arrival process $A(s,t)$ is defined as being (σ_A, ρ_A)-constrained if for all $\theta > 0$, there exist $\sigma_A(\theta)$ and $\rho_A(\theta) \in \mathbb{R}_+ \cup \{+\infty\}$ such that the MGF of $A(s,t)$ satisfies

$$\mathbb{E}[e^{\theta A(s,t)}] \leq e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))}.$$

This is related to the theory of effective bandwidth that is defined as

$$\frac{1}{\theta(t-s)} \ln \mathbb{E}[e^{\theta A(s,t)}]. \tag{3}$$

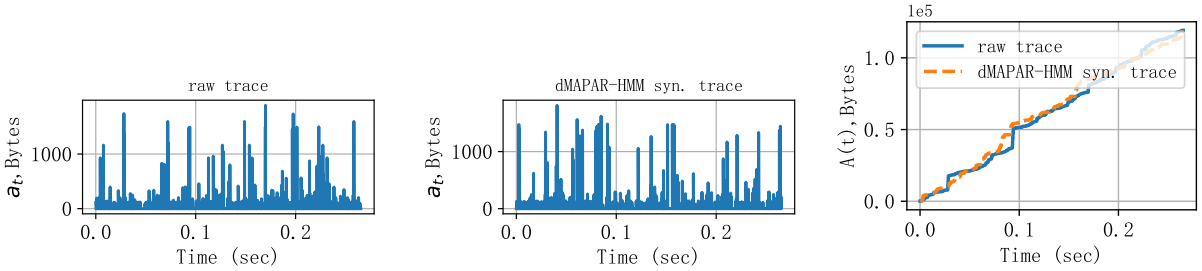
The effective bandwidth characterizes the statistical behavior of traffic arrivals and has be applied to the derivation of closed-form solutions for end-to-end performance bounds [6].

A. MGF of dMAPAR-HMM

Consider an arrival process with dMAPAR-HMM arrivals. Let $V(s,t)$ be the vector that stands for $[\mathbb{E}[e^{\theta A(s,t)}|Z_s = k]]_{k\in\mathcal{S}}$, which is the MGF of $A(s,t)$ conditional on the state of Z_s at time s. By definition, we have $V_k(t,t) = 1, \forall k \in \mathcal{S}$. For all $s \leq t$, we have the following result for the MGF.

Theorem 1: For all $s \leq t$, we have

$$V_k(s,t) = \mathbb{E}\Big[e^{\sum_{i=1}^{p}[\theta_i(s,t)-\theta]y_{N(s)+1-i}} \times \prod_{l=s+1}^{t} M(\theta_1(l,t); Z_{l-1}, Z_l)|Z_s = k\Big],$$



(a) (b) (c)

Figure 5: **Sample paths and cumulative arrivals of raw and synthetic traces.** The figures are, from left to right: (a) discretized WIDE trace; (b) synthetic trace with well estimated dMAPAR-HMM; and (c) accumulative arrival (bytes) over a 0.25-second window.

WIDE Trace captured at 2022-04-13 08:00	FlowID	H (*empirical*)	H (*estimated*)	CV (*empirical*)	CV (*estimated*)
(src IP, dst IP, protocol)=(203.76.247.112, 204.57.206.30, IPv4)	flow1	0.2977	0.2978	11.43	11.18
(src IP, dst IP, protocol)=(13.212.219.174, 163.194.192.18, UDP)	flow2	0.2739	0.2790	21.59	22.27
(src IP, dst IP, protocol)=(131.142.41.211, 144.195.33.147, TCP)	flow3	0.3085	0.3055	27.04	26.20

Table IV: Comparison between raw and synthetic traces.

where $\theta_i(s,t) = \theta\varphi_i(s,t)$, with

- $\varphi_{p+1}(s,t) = 1$
- $\varphi_i(t,t) = 1$
- $\varphi_i(s-1,t) = (1-O_s)\varphi_i(s,t) + O_s[\varphi_1(s,t)\breve{\phi}_i(Z_{s-1},Z_s) + \varphi_{i+1}(s,t)]$

$\forall i = 1,\cdots,p$ and $M(\cdot; Z_{l-1}, Z_l)$ is the MGF of $\breve{\mu}(Z_{l-1},Z_l) + \breve{\sigma}(Z_{l-1},Z_l)\epsilon_l$ given Z_{l-1}, Z_l.

Proof 1: The theorem can be easily proved by induction. Interest readers can refer to the arXiv link for the details.

B. Asymptotic Effective Bandwidth

As to Theorem 1, we first note that $O_s = 1_{\{Z_s \geq m_1 * m_2 * N\}}$. If we let $\varphi(s,t) = [\varphi_1(s,t),\cdots,\varphi_p(s,t)]^T$, $b(Z_s) = [0,0,\cdots,O_s]^T$ and

$$A(Z_{s-1},Z_s) = \begin{bmatrix} 1-O_s+\breve{\phi}_1(Z_{s-1},Z_s) & O_s & & \\ \breve{\phi}_2(Z_{s-1},Z_s) & 1-O_s & O_s & \\ \vdots & & \ddots & \\ \breve{\phi}_p(Z_{s-1},Z_s) & \cdots & \cdots & 1-O_s \end{bmatrix},$$

we would have

$$\begin{aligned} & \varphi(s-1,t) \\ = \;& b(Z_s) + A(Z_{s-1},Z_s)\varphi(s,t) \\ = \;& b(Z_s) + A(Z_{s-1},Z_s)[b(Z_{s+1}) + A(Z_s,Z_{s+1})\varphi(s+1,t)] \\ = \;& \cdots \\ = \;& \varphi(s-1,t-1) + A(Z_{s-1},Z_s)\cdots A(Z_{t-2},Z_{t-1}) \\ & \quad \times[b(Z_t) + A(Z_{t-1},Z_t)\mathbf{1} - \mathbf{1}] \\ \to \;& \varphi(s-1,t-1), \quad \text{as } t-s \to \infty. \end{aligned} \tag{4}$$

The last step of Eqn. (4) used the stationary condition of AR-HMM, which implies that $\{\varphi(s,t)\}_t$ converges as t tends to infinity. Let $\lim_{t-s\to\infty}\varphi(s,t) = W_s$, the following results hold for the effective bandwidth.

Theorem 2: Let $\boldsymbol{v} = [\mathbb{E}[W_s|Z_s = k]]_{k\in\mathcal{S}}$, A be the matrix $[A(Z_{s-1} = i, Z_s = j)]_{i,j\in\mathcal{S}}$ with its element belonging to $\mathbb{R}^{p\times p}$, and $B = [[0,\cdots,0]_{p\times 1},\cdots,[0,\cdots,0]_{p\times 1}, [0,\cdots,1]_{p\times 1},\cdots,[0,\cdots,1]_{p\times 1}]^T$. We then have

1) $\boldsymbol{v} = (I_{|\mathcal{S}|\times p} - T\circ A)^{-1}TB$;
2) Denote $\phi_{\max} = \max|\phi_{i,j}|$ and let $\boldsymbol{\pi} = [\pi_i]_{i\in\mathcal{S}}$ be the stationary probability vector of Z, where π_i stands for the stationary probability of state i. We then have

$$\lim_{\phi_{\max}\to 0} 1/\theta(t-s)(\ln\mathbb{E}[e^{\theta A(s,t)}] - \ln\boldsymbol{\pi}\mathrm{R}\Gamma^{t-s}(\theta)\mathbf{1}) = 0,$$

where

$$\begin{cases} \mathrm{R} = \mathrm{diag}([\mathbb{E}[e^{\sum_{k=1}^{p}\theta(v_{i,k}-1)y_{N(s)+1-k}}|Z_s = i]]_{i\in\mathcal{S}}) \\ \Gamma(\theta) = [\psi_{i,j}(\theta)]_{i,j\in\mathcal{S}}\circ T \\ \psi_{i,j}(\theta) = \begin{cases} e^{\theta\breve{\mu}_{i,j}v_{j,1}+\frac{1}{2}\theta^2\breve{\sigma}^2_{i,j}v^2_{j,1}}, & \epsilon_t \sim \mathcal{N}(0,1), \\ \dfrac{e^{\theta\breve{\mu}_{i,j}v_{j,1}}}{1-\theta\breve{\sigma}_{i,j}v_{j,1}}, & \epsilon_t \sim \mathrm{Exp}(1). \end{cases} \end{cases}$$

Proof 2: The proof can be found in the arXiv.

Based on Theorem 2 and Eqn. 3, we can obtain the asymptotic effective bandwidth. Since $\Gamma(\theta)$ is diagonalizable in $\mathbb{C}$, there exists a diagonal matrix D and a matrix U such that $\Gamma(\theta) = UDU^{-1}$. Hence,

$$\begin{aligned} \boldsymbol{\pi}\mathrm{R}\Gamma^{t-s}(\theta)\mathbf{1} &= \boldsymbol{\pi}\mathrm{R}UD^{t-s}U^{-1}\mathbf{1} \\ &= \textstyle\sum u_i D_{ii}^{t-s} \\ &= \max|D_{ii}|^{t-s}\textstyle\sum u_i(\frac{D_{ii}}{\max|D_{ii}|})^{t-s} \\ &\leq \max|D_{ii}|^{t-s}\textstyle\sum u_i \\ &= e^{\ln\sum u_i + \ln\max|D_{ii}|(t-s)}, \end{aligned}$$

with the notation $u_i = (\boldsymbol{\pi}\mathrm{R}U)_i(U^{-1}\mathbf{1})_i$.

The asymptotic effective bandwidth can then be programmatically calculated with

$$\begin{aligned} \tfrac{1}{\theta(t-s)}\ln\mathbb{E}[e^{\theta A(s,t)}] &\sim \tfrac{1}{\theta(t-s)}\ln\boldsymbol{\pi}\mathrm{R}\Gamma^{t-s}(\theta)\mathbf{1} \\ &\leq \sigma_A(\theta)/t - s + \rho_A(\theta), \end{aligned} \tag{5}$$

where $\sigma_A(\theta) = \ln\sum u_i/\theta$ and $\rho_A(\theta) = \frac{\ln\max|D_{ii}|}{\theta}$.

Algorithm 2: Feature Extraction

Input : flow, Δt, θ
Output: (σ, ρ)-bound

```
def FeatureExtractor(flow, Δt, θ):
    /* from continuous-time to
       discrete-time system.                    */
    {a_t} ← discretize the traffic flow with Δt ;
    (Q, P, Θ) ← fit dMAPAR-HMM with {a_t};
    /* calculate the input flow's
       (σ(θ), ρ(θ))-bound with the
       extracted features.                      */
    T ← Eqn.(2);
    π ← πT = π, ∑ π_i = 1;
    R(θ) ← Theorem 2 ;
    Γ(θ) ← Theorem 2 ;
    (σ_A(θ), ρ_A(θ)) ← Eqn. (5).
    return (σ_A(θ), ρ_A(θ))
```

IV. End-to-End Performance Evaluation

We have established a spatial-temporal model for network traffic, and introduced a ready-to-use interface for it to the MGF-SNC in previous sections. In this section, we shall use it directly to conduct end-to-end network analysis. Alike arrival process, a service process $S(s,t)$ is said to be (σ_S, ρ_S)-constrained if for all $\theta > 0$, there exist $\sigma_S(-\theta)$ and $\rho_S(-\theta) \in \mathbb{R}_+ \cup \{+\infty\}$ such that

$$\mathbb{E}[e^{-\theta S(s,t)}] \leq e^{\theta(\sigma_S(-\theta)-\rho_S(-\theta)(t-s))}.$$

There are three atomic operations underlying the MGF-SNC analysis framework [35]: *Aggregation*, *Left-over* and *Concatenation*. With these tools, we can reduce any feed-forward to a single flow - single server topology. From SNC theory, we know that, if arrivals and service are (σ, ρ)-constrained, the stochastic delay bound which is violated at most with probability ε is given by [26]

$$T_\varepsilon = \inf_{\theta \in \{\theta | \rho_A(\theta) < \rho_S(-\theta)\}} \Big\{ \frac{\sigma_A(\theta) + \sigma_S(-\theta)}{\rho_S(-\theta)} - \frac{\log(\varepsilon(1 - e^{\theta(\rho_A(\theta)-\rho_S(-\theta))}))}{\theta \rho_S(-\theta)} \Big\}. \tag{6}$$

A. Experimental Setup

Let us consider the topology shown in Figure 6. The switches support two queues. Flow 1 enters the highest priority queue. Flow 2 and 3 enter the lowest priority queue. The queuing mode is by Strict Priority where the priority sets the order in which queues are served, starting with the highest priority queue and going to the next lower queue only after the highest queue has been transmitted. Suppose that switches in this network are work-conserving servers with constant service rate c, i.e., $\sigma_S(-\theta) = 0$ and $\rho_S(-\theta) = c$. For any flow of interest, we can use Eqn. (6) to compute the end-to-end performance bounds with S being substituted by the end-to-end service S_{e2e}. To illustrate the importance of feature extraction in performance analysis, we take 6 well-known arrival models to benchmark the delay bound, which are: (1) independent normally distributed increments (Normal); (2) independent exponentially distributed increments (Exponential); (3) compounded Poisson (cPoisson); (4) 1-order autoregressive process (AR1); (5) Markov-modulated on-off (mmoo); and (6) Markov-modulated process (mmp). Comparisons are made against the ground truth which is obtained from the discrete-event simulation (DES) [36].

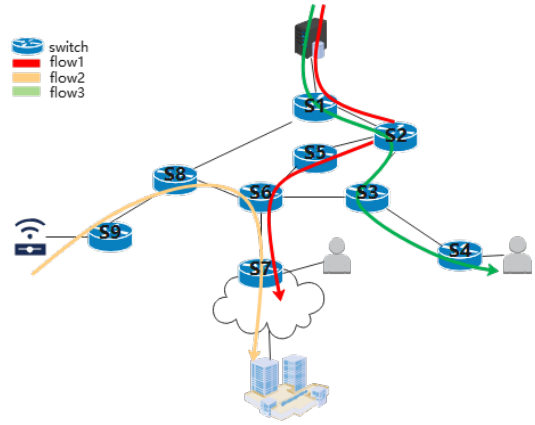


Figure 6: **Network topology of a small-scale commercial campus.** The flows are of typical backbone network traffic as shown in Table IV.

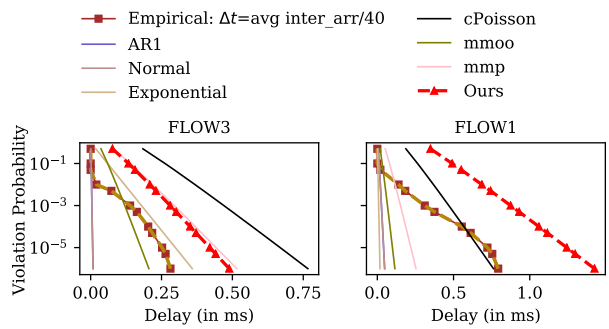


Figure 7: Performance evaluation results on the single flow - single server topology with $c = 100Mbps$.

B. Single Server Analysis

We begin our analysis with the single flow - single server topology. This is the most important topology, since we can reduce any feed-forward to this network by using the end-to-end service. Another benefit of it is that we can eliminate interference from other sources and focus on the affect of arrivals on performance evaluation. In Figure 7, we test the impact of arrival models in feature extraction. As we can see from this figure, if we use the Exponential model to extract traffic features in MGF-SNC, we will get a good estimate of the delay bound with flow 3. But when we move over to flow 1, we cannot even get a reliable estimate. The same argument applies to the mmp model and the cPoisson model.

C. Network Analysis

We continue with the network topology and adopt the "pay multiplexing only once" (PMOO) [12] algorithm to perform the end-to-end network analysis. Let us consider flow 1. The overall service offered to flow 1 can be described by the end-to-end service

$$S_{e2e}(s,t) = S_1 \otimes S_2 \otimes S_5 \otimes S_6 \otimes S_7(s,t),$$

where $\otimes$ is the *min-plus* convolution operator. The end-to-end service is not influenced by the cross-flows, and the only uncertainty to T_ε is from the arrivals. We test the impact of traffic models in Figure 8a. The first line of Figure 8a displays the result with $c = 100Mbps$, and the second line, $c = 35Mbps$. We see the same phenomenon here: the reliability is not guaranteed when we choose an inappropriate model to do feature extraction. Take the cPoisson model as an instance. It can provide accurate results in high-bandwidth scenarios, whereas it behaves dramatically poor as bandwidth decreases. MGF-SNC with the other arrival models might severely underestimate the delay bound. By contrast, dMAPAR-HMM can accurately capture the various characteristics of traffic from the decomposed spatio-temporal signals and hence significantly boost the effectiveness of MGF-SNC, which means that both the tightness and the reliability of the delay bound are enhanced (see also Figure 8b for the stochastic delay bound with violation probability $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-5}$, respectively).

Different from flow 1, flow 2 and 3's end-to-end services are influenced by the cross-flows. When the flow of interest merges with a cross-flow, its service might be strongly reduced. Compared to its deterministic counterpart, the impact of cross-traffic in SNC can be stronger, as computation operations like leftover service or deconvolution require several stochastic inequalities, which in many cases leads to loose bounds. Since our contribution is not on the service modeling, we will not pay a lot of discussion on how to reduce pessimism brought by SNC in complex topologies and leave this to future work. What we want to address here is that even if more advanced modeling of cross-flow is employed, MGF-SNC with existing arrival models might still underestimate the delay bound (see the last two columns of Figure 8a). The proposed dMAPAR-HMM model outperforms popular arrival models and shows robustness under different scenarios. The main credit for that goes to the versatility of dMAPAR-HMM which contributes to better extracting critical traffic features from network flows. However, the other arrival models may not work well. As network utilization level grows, complicated characteristics of network traffic gradually play a more influential role. Existing models are too short-sighted to capture such characteristics of real traffic and hence cannot provide an appropriate estimation for performance bound. The proposed dMAPAR-HMM accurately captures such characteristics, and hence improves performance analysis.

V. Conclusion

In this paper, we revisit the traffic modeling problem in SNC and propose a spatial-temporal model dMAPAR-HMM for real-world network traffic. We dissect various existing arrival models and reveal that dMAPAR-HMM unifies these models. Then we derive the MGF-bound of dMAPAR-HMM which can be directly integrated into the framework of MGF-SNC. Extensive experiments with real traffic traces demonstrate that dMAPAR-HMM can accurately represent multi-dimensional multi-order characteristics of current network traffic. Experiments also show that MGF-SNC with dMAPAR-HMM achieves a tighter and more robust performance bound, while the existing arrival models are either overly optimistic for accurate performance evaluation or sensitive to specific traffic types and scenarios. In a nutshell, dMAPAR-HMM provides a fine-grained QoS guarantee and facilitates effective network planning. We envision that dMAPAR-HMM will motivate further exploration of traffic modeling for improving SNC.

References

[1] R. Cruz, "A calculus for network delay. I. network elements in isolation," IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 114–131, 1991.

[2] ——, "A calculus for network delay. II. network analysis," IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 132–141, 1991.

[3] C. S. Chang, R. L. Cruz, J. Y. Le Boudec, and P. Thiran, "A min, + system theory for constrained traffic regulation and dynamic service guarantees," IEEE/ACM Transactions on Networking, vol. 10, no. 6, pp. 805–817, 2002.

[4] C. S. Chang, "Performance guarantees in communication networks," Springer Science & Business Media, 2000.

[5] F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," IEEE Transactions on Information Theory, vol. 52, no. 6, pp. 2300–2312, 2006.

[6] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in Proceedings of the 14th IEEE International Workshop on Quality of Service, 2006, pp. 261–270.

[7] Y. Jiang and Y. Liu, "Stochastic network calculus," Springer, vol. 1, 2008.

[8] F. Ciucu and J. Schmitt, "Perspectives on network calculus: No free lunch, but still good value," in Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. Association for Computing Machinery, 2012, p. 311–322.

[9] F. Poloczek and F. Ciucu, "Scheduling analysis with martingales," Performance Evaluation, vol. 79, pp. 56–72, 2014.

[10] P. Nikolaus, J. Schmitt, and M. Schuetze, "h-Mitigators: Improving your stochastic network calculus output bounds," Computer Communications, vol. 144, pp. 188–197, 2019.

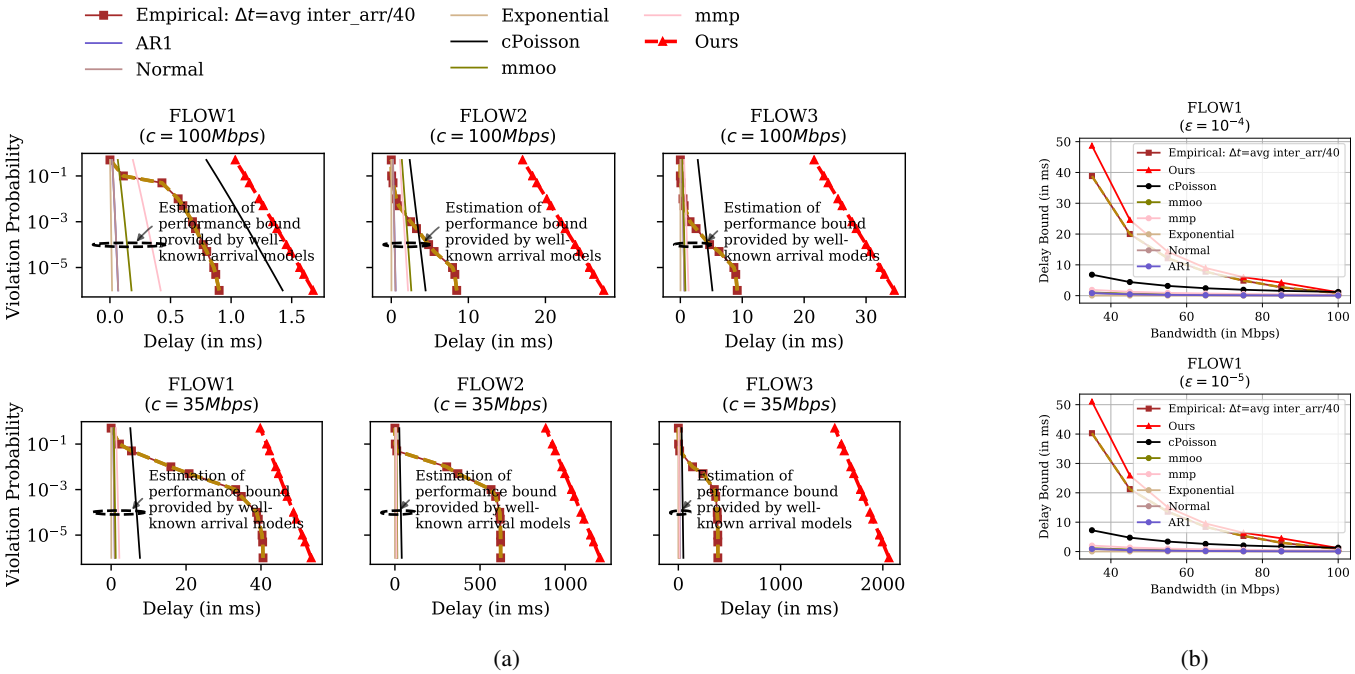
[11] P. Nikolaus and J. Schmitt, "Improving delay bounds in the stochastic network calculus by using less stochastic inequalities," in Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools. Association for Computing Machinery, 2020, p. 96–103.

[12] ——, "On per-flow delay bounds in tandem queues under (in)dependent arrivals," in Proceedings of IFIP Networking Conference (IFIP Networking) and Workshops, 2017, pp. 1–9.

[13] W. Leland, W. Willinger, M. Taqqu, and D. Wilson, "On the self-similar nature of ethernet traffic," ACM SIGCOMM Computer Communication Review, vol. 25, no. 1, p. 202–213, 1995.

[14] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, "A nonstationary Poisson view of Internet traffic," in Proceedings of IEEE INFOCOM, vol. 3, 2004, pp. 1558–1569.

[15] M. Li and S. Lim, "Modeling network traffic using generalized Cauchy process," Physica A: Statistical Mechanics and its Applications, vol. 387, no. 11, pp. 2584–2594, 2008.



(a) (b)

Figure 8: (a) End-to-end performance analysis of the network topology; and (b) stochastic delay bound with violation probability $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-5}$, respectively.

[16] M. Li, "Multi-fractional generalized Cauchy process and its application to teletraffic," Physica A: Statistical Mechanics and its Applications, vol. 550, p. 123982, 2020.

[17] X. Peng, F. Zhang, L. Chen, and G. Zhang, "A MAP-based performance analysis on 5G-powered cloud VR streaming," in Proceedings of IEEE International Conference on Communications, 2021, pp. 1–6.

[18] H. Jiang and C. Dovrolis, "Why is the Internet traffic bursty in short time scales?" SIGMETRICS Perform. Eval. Rev., vol. 33, no. 1, p. 241–252, jun 2005.

[19] F. Tobagi, M. Gerla, R. Peebles, and E. Manning, "Modeling and measurement techniques in packet communication networks," in Proceedings of the IEEE, vol. 66, 1978, pp. 1423–1447.

[20] M. Li and S. Chen, "Fractional Gaussian noise and network traffic modeling," in Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational Science, vol. 1. World Scientific and Engineering Academy and Society, 2009, p. 34–39.

[21] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," IEEE Transactions on Communications, vol. 43, no. 2/3/4, pp. 1566–1579, 1995.

[22] V. Paxson and S. Floyd, "Wide area traffic: the failure of poisson modeling," IEEE/ACM Transactions on Networking, vol. 3, no. 3, pp. 226–244, 1995.

[23] M. Li, "Modified multifractional Gaussian noise and its application," Physica Scripta, vol. 96, no. 12, p. 125002, 2021.

[24] ——, "Generalized fractional Gaussian noise and its application to traffic modeling," Physica A: Statistical Mechanics and its Applications, vol. 579, p. 126138, 2021.

[25] J. Liebeherr, A. Burchard, and F. Ciucu, "Delay bounds in communication networks with heavy-tailed and self-similar traffic," IEEE Transactions on Information Theory, vol. 58, no. 2, pp. 1010–1024, 2012.

[26] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," IEEE Communications Surveys Tutorials, vol. 17, no. 1, pp. 92–105, 2015.

[27] G. Horváth and H. Okamura, "A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure," in Proceedings of Computer Performance Engineering: 10th European Workshop. Springer Berlin Heidelberg, 2013, pp. 119–133.

[28] G. Casale, E. Zhang, and E. Smirni, "Trace data characterization and fitting for markov modeling," Performance Evaluation, vol. 67, no. 2, pp. 61–79, 2010.

[29] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling ip traffic using the batch markovian arrival process," Performance Evaluation, vol. 54, no. 2, pp. 149–173, 2003.

[30] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," Econometrica: Journal of the econometric society, pp. 357–384, 1989.

[31] H. Akaike, "A new look at the statistical model identification," IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.

[32] G. Schwarz, "Estimating the dimension of a model," The Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.

[33] A. Tenyakov, R. Mamon, and M. Davison, "Filtering of a discrete-time hmm-driven multivariate ornstein-uhlenbeck model with application to forecasting market liquidity regimes," IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 6, pp. 994–1005, 2016.

[34] D. Zhu, J. Lu, W. Ching, and T. Siu, "Discrete-time optimal asset allocation under higher-order hidden Markov model," Economic Modelling, vol. 66, pp. 223–232, 2017.

[35] C. S. Chang, Performance guarantees in communication networks. Springer Science & Business Media, 2000.

[36] B. Li, L. Chen, and X. Peng, "ns.py," June 2022. [Online]. Available: https://github.com/TL-System/ns.py

Ethics

This work does not raise any ethical issues.