Improving the Mean-Field Fluid Model of Processor Sharing Queueing Networks for Dynamic Performance Models in Cloud Computing

Johan Ruuskanen, Tommi Berner, Karl-Erik Årzén, and Anton Cervin
Department of Automatic Control, Lund University, Sweden
{johan.ruuskanen, tommi.nylander, karl-erik.arzen, anton}@control.lth.se

ABSTRACT

Resource management in cloud computing is a difficult problem, as one needs to balance between adequate service to clients and cost minimization in a dynamic environment of interconnected components. To make correct decisions in such an environment, good performance models are necessary. A common modeling methodology is to use networks of queues, but as these are prohibitively expensive to evaluate for many applications, approximation methods for key metrics are frequently employed. One such method—that provides both transient solutions and short, scalable computation times—is the fluid model, which approximates the dynamics of the mean queue lengths using a system of ordinary differential equations. However, finding a fluid model that can adequately approximate an arbitrary queueing network is in general difficult. In the paper, we extend the state of the art with the following three contributions. First, we show that for any mixed multiclass queueing network of processor sharing and delay queues with phase-type service time distributions, such a fluid model can be found via the mean-field approximation. Furthermore, we propose an improved model based on smoothing of the processor share function that improves the performance of certain systems. Finally, using the smoothed mean-field model, we introduce an accurate closed-form approximation of the response time CDF over any subset of classes and queues.

Keywords

Queueing network; Processor sharing; Mean-field approximation; Fluid model; Response time approximation.

Acknowledgement

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors are members of the ELLIIT Excellence Center.

1. INTRODUCTION

Multiclass queueing networks in which only some request paths of classes are subjected to external arrivals are known as *mixed queueing networks* [Baskett et al. 1975]. These networks are of interest to study in performance modeling of cloud systems, as they can be used to model advanced behaviour such as interactions between synchronous and asynchronous calls to software layers. As exact analysis of queuing networks is often intractable and also prohibitively expensive to evaluate by simulation for many real-time applications, different approximation techniques of relevant performance metrics are often employed.

In certain cases, it is possible to translate the queueing network to a CTMC from which a fluid model can be obtained via the mean-field approximation,

$$\tfrac{d}{dt}\boldsymbol{x} = F(\boldsymbol{x}), \quad \boldsymbol{x}(0) = \boldsymbol{X}(0) \tag{1}$$

where $\boldsymbol{X}$ is the request population vector in the network, $\boldsymbol{x}(t)$ an approximation of $\mathbb{E}[\boldsymbol{X}(t)]$ and $F(\boldsymbol{X})$ the drift function of the corresponding CTMC. We will refer to such a fluid model as a *mean-field fluid model*. Considering multiclass networks, in [Pérez and Casale 2017] it was shown that for closed networks, Kurtz's theorem holds for any combination of multiclass *processor-sharing* (PS) or *delay* (INF) queues with general service times in the form of *phase-type* (PH) distributions.

This paper is an extended abstract of [Ruuskanen et al. 2021]. We refer to the original paper for proofs, more related work, further motivations, additional contributions and multiple experimental evaluations.

2. MIXED PS QUEUEING NETWORK

We built on the results of [Pérez and Casale 2017] and show that the results regarding Kurtz's theorem can be extended to mixed networks. Further, a compact nonlinear matrix form of the corresponding drift function is developed.

Before continuing, some definitions are needed. Let $\mathcal{Q}$ be the set of queues, where each queue i has $\mathcal{C}_i$ classes and where each class $r \in \mathcal{C}_i$ has a PH distributed service time with $\mathcal{S}_{i,r}$ phase states. Further, let $\mathcal{C}$ and $\mathcal{S}$ be the sets of all classes and phase states in the network. The PH distributions are parametrized by $\Psi^{i,r} \in \mathbb{R}^{|\mathcal{S}_{i,r}|\times|\mathcal{S}_{i,r}|}$, $\psi^{i,r} \in \mathbb{R}_+^{|\mathcal{S}_{i,r}|\times 1}$ and $\zeta^{i,r} \in \mathbb{R}_+^{|\mathcal{S}_{i,r}|\times 1}$ describing; the transition rates between the transient states, from the transient states to the absorbing state, and the entrance probabilities respectively. Further, let $P \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{C}|}$ be a substochastic matrix where $P_{i,j}^{r,s}$ defines the transition probability from $(i,r) \in (\mathcal{Q},\mathcal{C}_i)$ to $(j,s) \in (\mathcal{Q},\mathcal{C}_i)$. If $\sum_{s,j} P_{i,j}^{r,s} < 1$ then there is a possibility for a request to depart the network once completed at (i,r). Finally, let each class be subjected to the possiblity of Poisson arrivals, whose rates are given by $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{C}|\times 1}$.

If we assume that $\boldsymbol{X} \in \mathbb{Z}_+^{|\mathcal{S}|}$ is a vector of the population in each phase state, and stack the parameter matrices for the phase-type distributions into the following block diagonals

$$\begin{aligned}\boldsymbol{\Psi}\big(\in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}\big) &= \operatorname{diag}\big(\Psi^{1,1}, \Psi^{1,2}, \Psi^{1,3}, \ldots\big)\\ \boldsymbol{B}\big(\in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{C}|}\big) &= \operatorname{diag}\big(\psi^{1,1}, \psi^{1,2}, \psi^{1,3}, \ldots\big)\\ \boldsymbol{A}\big(\in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{C}|}\big) &= \operatorname{diag}\big(\zeta^{1,1}, \zeta^{1,2}, \zeta^{1,3}, \ldots\big)\end{aligned}$$

and form

$$\boldsymbol{P} = \begin{bmatrix} P_{1,1}^{\cdot,\cdot} & P_{1,2}^{\cdot,\cdot} & \cdots \\ P_{2,2}^{\cdot,\cdot} & P_{2,2}^{\cdot,\cdot} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{2}$$

obtaining the drift function can be simplified as follows.

Theorem 1. *The drift function of the mixed queueing network of PS and INF queues can be expressed as*

$$F(\boldsymbol{X}) = \boldsymbol{W}^T\theta(\boldsymbol{X}) + \boldsymbol{A}\boldsymbol{\lambda} \tag{3}$$

Proof. See [Ruuskanen et al. 2021, Th. 1]. □

where $\boldsymbol{W} = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{P}\boldsymbol{A}^T$, and $\theta(\boldsymbol{x})$ a function s.t. $\forall(i,r,a) \in (\mathcal{Q}, \mathcal{C}_i, \mathcal{S}_{i,r})$

$$\begin{aligned}\theta_{i,r,a}(\boldsymbol{X}) &= X_{i,r,a}g_{i,r,a}(\boldsymbol{X})\\ g_{i,r,a}(\boldsymbol{X}) &= \frac{\min\Big(k_i,\ \sum_{s\in\mathcal{C}_i}\sum_{b\in\mathcal{S}_{i,s}} X_{i,s,b}\Big)}{\sum_{s\in\mathcal{C}_i}\sum_{b\in\mathcal{S}_{i,s}} X_{i,s,b}}\end{aligned} \tag{4}$$

where $g_{i,r,a}(\boldsymbol{X})$ is known as the processor share.

Furthermore, the corresponding solution of the mean-field fluid model $\boldsymbol{x}(t)$ can be shown to be the limit of a scaled version of the mixed network. Introduce the sequence $\{\boldsymbol{X}^{(v)}\}_{v\geq 1}$, where $\boldsymbol{X}^{(v)}(0) = v\boldsymbol{X}(0)$, $\boldsymbol{k}^{(v)} = v\boldsymbol{k}$, and $\boldsymbol{\lambda}^{(v)} = v\boldsymbol{\lambda}$, then

Theorem 2. *For any $\delta > 0$,*

$$\lim_{v\to+\infty} \mathbb{P}\left\{\sup_{t\leq T}\left|\frac{\boldsymbol{X}^{(v)}(t)}{v} - \boldsymbol{x}(t)\right| > \delta\right\} = 0 \tag{5}$$

Proof. See [Ruuskanen et al. 2021, Th. 3]. □

3. SMOOTHED MEAN-FIELD MODEL

The convergence result in Theorem 2 does not state how accurate the approximation is, and mean-field approximations are in general known to suffer from approximation errors when the system size is small. To improve accuracy in these cases, we created a novel computationally cheap improvement for mean-field fluid models of mixed PS networks.

The main problem for these mean-field models lies in that

$$\mathbb{E}\left[\theta_a\left(\boldsymbol{X}\right)\right] \neq \theta_a\left(\mathbb{E}\left[\boldsymbol{X}\right]\right) \tag{6}$$

due to the minimum function and the fraction in $g_a\left(\boldsymbol{X}\right)$. To improve similarity between the quanitites in (6), we created a smoothing over the minimum using the inverse p-norm

$$\begin{aligned}\hat{g}_a\left(\boldsymbol{x}, p_{Q(a)}\right) &= \frac{1}{\left(1 + \left(k_{Q(a)}^{-1}\sum \boldsymbol{x}_{Q(a)}\right)^{p_{Q(a)}}\right)^{1/p_{Q(a)}}}\\ p_{Q(a)} &> 0,\ \forall a \in \mathcal{S}\end{aligned} \tag{7}$$

which gives the following smoothed mean-field fluid model

$$\begin{aligned}\frac{d\boldsymbol{x}}{dt} &= \boldsymbol{W}^T\hat{\theta}\left(\boldsymbol{x}, \boldsymbol{p}\right) + \boldsymbol{A}\boldsymbol{\lambda}\\ \boldsymbol{x}(0) &= \boldsymbol{X}(0), \quad \boldsymbol{p}(t) > \boldsymbol{0}\end{aligned} \tag{8}$$

The smoothing is dependent on the parameter $\boldsymbol{p}$, but as (7) can be shown to be monotone in $p_{Q(a)}$, the optimal $\boldsymbol{p}^*$ for a stationary system can be quickly found from data.

4. CLOSED-FORM APPROXIMATION OF RESPONSE TIME CDF

In performance models of cloud systems, different response time percentiles are often of interest to capture. Retrieving estimates from a fluid model are however difficult, and current techniques rely on either Chebyshev bounds, or by solving an extended fluid model. However, we showed that it is possible to create accurate closed-form approximations.

The main idea is that we can obtain the time-dynamics of the probability of finding a specific request at a certain phase state in $(i,r) \in (\mathcal{Q}, \mathcal{C}_i)$ via the Chapman-Kolmogorov equation. The resulting equation however becomes intractable, but assuming that every request recieves the mean processor share, and approximating the mean processor share using $\hat{g}\left(\boldsymbol{x}, \boldsymbol{p}\right)$ we end up with the following linear ODE

$$\mathbb{E}\left[\dot{\boldsymbol{\pi}}(t)\right] \approx \left(\Psi^{i,r}\right)^T \hat{g}_i\left[\boldsymbol{x}^*\left(\boldsymbol{p}\right), \boldsymbol{p}\right]\mathbb{E}\left[\boldsymbol{\pi}(t)\right], \quad \boldsymbol{\pi}(0) = \zeta^{i,r} \tag{9}$$

which yields a closed-form approximative solution to $\boldsymbol{\pi}(t)$. Here $\boldsymbol{x}^*(\boldsymbol{p})$ is the stationary solution to (8) given $\boldsymbol{p}$. As the response time CDF can be shown to be given as the complement of the probability that the request remains in (i,r), an approximation can be obtained as

$$\Phi_{i,r}\left(t \mid \boldsymbol{p}\right) \approx 1 - \boldsymbol{\pi}(0)^T \exp\left[\hat{g}_i\left[\boldsymbol{x}^*(\boldsymbol{p}), \boldsymbol{p}\right]\Psi^{i,r}t\right]\mathbb{1} \tag{10}$$

We further approximate the response time CDF over almost any subset of classes $\mathcal{C}_R \subset \mathcal{C}$. Let $\beta := \mathbb{R}_+^{|\mathcal{C}|\times 1}$ be the probability of entering specific states in $\mathcal{C}_R$ at $t = 0$, i.e., $\sum\beta = 1$, $\beta_r \geq 0$ if $r \in \mathcal{C}_R$ else $\beta_r = 0$, and introduce the matrix $\boldsymbol{P}_R := \mathbb{R}_+^{|\mathcal{C}|\times|\mathcal{C}|}$ which only includes the class-to-class transition probabilities between classes in $\mathcal{C}_R$, i.e. $\forall r, s \in \mathcal{C}$ $(\boldsymbol{P}_R)_{r,s} = \boldsymbol{P}_{r,s}$ if $r, s \in \mathcal{C}_R$ otherwise $(\boldsymbol{P}_R)_{r,s} = 0$. Then,

$$\Phi_{\mathcal{C}_R}\left(t \mid \beta\right) \approx 1 - \beta^T\boldsymbol{A}^T \exp\left[D^{\hat{g}\left[\boldsymbol{x}^*(\boldsymbol{p}), \boldsymbol{p}\right]}\boldsymbol{W}_R t\right]\mathbb{1} \tag{11}$$

where $\boldsymbol{W}_R = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{P}_R\boldsymbol{A}^T$ and $D^{\boldsymbol{x}}$ the diagonal matrix s.t. $D_{aa}^{\boldsymbol{x}} = x_a$ $\forall a \in \mathcal{S}$. Further,

Theorem 3. *$\Phi_{\mathcal{C}_R}\left(t \mid \beta\right)$ is a valid CDF for all valid choices of $\mathcal{C}_R$ and flow matrices defined for the drift in Theorem 1.*

Proof. See [Ruuskanen et al. 2021, Th. 7] □

5. REFERENCES

F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. 1975. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM* 22, 2 (April 1975), 248–260. https://doi.org/10.1145/321879.321887

Juan F. Pérez and Giuliano Casale. 2017. Line: Evaluating Software Applications in Unreliable Environments. *IEEE Transactions on Reliability* 66, 3 (2017), 837–853.

Johan Ruuskanen, Tommi Berner, Karl-Erik Årzén, and Anton Cervin. 2021. Improving the mean-field fluid model of processor sharing queueing networks for dynamic performance models in cloud computing. *Performance Evaluation* 151 (2021), 102231. https://doi.org/10.1016/j.peva.2021.102231