Technische Universiteit Eindhoven University of Technology

EURANDOM

Corrected-PH approximations for the workload of the MAP/G/1 queue with heavy-tailed services

Eleni Vatamidou, Ivo Adan, Maria Vlasiou, and Bert Zwart

Introduction

What makes the evaluation of the workload in a MAP/G/1 queue (FIFO) a challenging problem:

- significant correlations between arrivals, and
- heavy-tailed service times.

Here, we construct very accurate approximations of the workload with a small relative error.

Model description

Parameters related to arrivals:

- MArP with N states
- $\mathbf{P}$ (trans. prob.), $\boldsymbol{\pi}$ (stat. dist.)
- $\boldsymbol{\Lambda}$ (exp. trans. rates), $\mathbf{Q}$ (prob. real arriv. cust.)

Parameters related to service times:

- $G_\epsilon(t) = (1-\epsilon)F_p(t) + \epsilon F_h(t)$, $\epsilon \in [0,1)$, service time of real customer (mixture motivated by statistical analysis)
- $\tilde{\mathbf{G}}_\epsilon(s) = \tilde{G}_\epsilon(s)\mathbf{Q} + (\mathbf{I}-\mathbf{Q})$ (LST), μ_ϵ (mean)

Stability condition:

$$\boldsymbol{\pi}\left(\boldsymbol{\Lambda}^{-1} - \mu_\epsilon \mathbf{Q}\right)\mathbf{e} > 0,$$

LT of steady-state workload:

$$\tilde{\boldsymbol{\Phi}}_\epsilon(s) = [\tilde{\phi}_{\epsilon,1}(s), \ldots, \tilde{\phi}_{\epsilon,N}(s)].$$

Theorem. *There exists* $\mathbf{u}$ *s.t. [2]*

$$\tilde{\boldsymbol{\Phi}}_\epsilon(s)\left(\tilde{\mathbf{G}}_\epsilon(s)\mathbf{P}\boldsymbol{\Lambda} + s\mathbf{I} - \boldsymbol{\Lambda}\right) = s\mathbf{u}_\epsilon,$$

$$\tilde{\boldsymbol{\Phi}}_\epsilon(0)\mathbf{e} = 1.$$

Target: total workload

$$\tilde{v}_\epsilon(s) = \frac{s \cdot \mathbf{u}_\epsilon \cdot \text{adj}\left(\tilde{\mathbf{G}}_\epsilon(s)\mathbf{P}\boldsymbol{\Lambda} + s\mathbf{I} - \boldsymbol{\Lambda}\right)\mathbf{e}}{\det\left(\tilde{\mathbf{G}}_\epsilon(s)\mathbf{P}\boldsymbol{\Lambda} + s\mathbf{I} - \boldsymbol{\Lambda}\right)}.$$

Approach

Perturbation on a PH base model. Steps:

1. PH approximation as base.
   (a) Set $\epsilon = 0$ & $\tilde{F}_p(s) = q_n(s)/p_m(s)$, (polynomials $q_n(s), p_m(s)$).
   (b) Evaluate $\mathbf{u}$, determinant and adjoint of $\tilde{\mathbf{G}}(s)\mathbf{P}\boldsymbol{\Lambda} + s\mathbf{I} - \boldsymbol{\Lambda}$.
   (c) For $Re(y_j), Re(x_j) > 0$, $j = 1, \ldots, mr$,
   $$\tilde{v}(s) = \frac{\mathbf{ue}\prod_{j=1}^{mr}(s+y_j)}{\prod_{j=1}^{mr}(s+x_j)}.$$
   (d) $\mathbb{P}(V > t) = \mathcal{L}^{-1}\{\tilde{v}(s)\}$.
2. Take $G_\epsilon(t) = F_p(t) + \epsilon(F_h(t) - F_p(t))$ and find the perturbed parameters (ϵ pert. param.).
3. Find $\tilde{v}_\epsilon(s)$, by keeping only up to ϵ-order terms, i.e.,
   $$\tilde{v}_\epsilon(s) = \tilde{v}(s) + \epsilon\tilde{v}(s)k(s) + O(\epsilon^2),$$
   $k(s)$ is well-defined on $\mathbb{R}^+$.

Results

Definition. CORRECTED-PH APPROXIMATIONS

$$\begin{aligned}\mathbb{P}(\hat{V}_\epsilon > t) &:= \mathbb{P}(V > t) + \epsilon\frac{1}{\mathbf{ue}}\Big(\alpha\mathbb{P}(V{>}t)\\ &+ \beta\left(\mu_p\mathbb{P}(V{+}B^e{>}t) - \mu_h\mathbb{P}(V{+}C^e{>}t)\right)\\ &+ \gamma\left(\mu_p\mathbb{P}(V{+}V'{+}B^e{>}t) - \mu_h\mathbb{P}(V{+}V'{+}C^e{>}t)\right)\\ &+ \sum_{j=1}^{mr}\alpha_j\mathbb{P}(V{+}E_{y_j}{>}t)\\ &+ \sum_{j=1}^{mr}\beta_j\left(\mu_p\mathbb{P}(V{+}B^e{+}E_{y_j}{>}t)\right.\\ &\left.- \mu_h\mathbb{P}(V{+}C^e{+}E_{y_j}{>}t)\right)\\ &+ \sum_{j=1}^{mr}\gamma_j\left(\mu_p\mathbb{P}(V{+}V'{+}B^e{+}E_{y_j}{>}t)\right.\\ &\left.- \mu_h\mathbb{P}(V{+}V'{+}C^e{+}E_{y_j}{>}t)\right)\Big).\end{aligned}$$

Example. Erlang-2 arrival process with rate 5. Service times mixture of an Exp(3) distribution and a heavy-tailed one [1] with $\tilde{F}_h(s) = 1 - \frac{s}{(2+\sqrt{s})(1+\sqrt{s})}$ for $\epsilon = 0.1$.

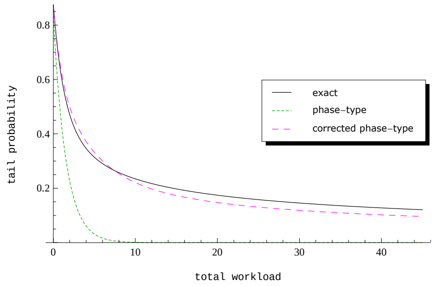

Figure 1: Comparison of exact total workload with approximations, for load 0.875

Conclusions

The approximations provide:

- correct tail behavior, and
- small absolute and relative error.

Future work: tandem queues.

References

[1] ABATE, J. AND WHITT, W. (1999). Explicit $M/G/1$ waiting-time distributions for a class of long-tail service-time distributions. *Oper.Res.Lett.* **25**, 25–31.

[2] ADAN, I.J.B.F. AND KULKARNI, V.G. (2003). Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Syst. Theory. App.* **45**, 113–134.