A Fluid Model for Many-Server Queues with Time-Varying Arrivals and Phase-Type Service Distribution

Yunan Liu, Ward Whitt
Department of Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027-6699
{yl2342,ww2040}@columbia.edu

We introduce and analyze a deterministic fluid model that serves as an approximation for the $G_t/PH/s_t + GI$ many-server queueing model, which has a general time-varying arrival process (the G_t), a phase-type (PH) service distribution (the PH), a time-dependent number of servers (the s_t) and allows abandonment from queue according to a general abandonment distribution (the $+GI$). We provide an efficient algorithm, using matrix-analytic methods, to compute all standard transient (time-dependent) performance measures, such as the fluid content in queue, potential waiting time, abandonment and service-completion rate, etc.

Time-varibility of the model data. An important and realistic feature of this $G_t/PH/s_t + GI$ model is the time-varibility of its arrival rate and staffing function. Unlike most textbook queueing models, customers arrive at real service systems with a time-varying stream, which has significant variations over the day. To cope with the nonstationary arrival pattern and to stabilize and control the system performance, staffing functions (number of servers) are designed to be time-dependent as well.

Non-Markovian probability structure. We make a significant step beyond queueing models with Markovian probability structures by considering non-exponential service and patience distributions. The extension is necessary because (i) the transient system dynamics depends heavily on these distributions beyond the means; (ii) statistical analysis showed that these distributions can be far from exponential in real service systems. Since queues with general service distributions are difficult to analyze, the PH assumption becomes a reasonable balancing point between model generality and computation tractability. On the one hand, the PH distribution is mathematically tractable using matrix-geometric methods, see [2]; on the other hand, this class of distributions can be used to approximate any distributions supported on $(0, \infty)$. Here we provide examples of fitting PH distributions to Log-normal distributions, using the EM algorithm developed in [1].

Transient dynamics. Based on solving a finite-dimensional ordinary differential equation (ODE), we develop an efficient algorithm to compute the standard time-dependent performance functions in a finite time interval. Solving this ODE, we obtain $\mathbf{B}(t)$, an n-dimensional row vector of the fluid in service, where n is the total number of service phases. This algorithm characterizes the performance separately in two system regimes: overloaded (OL) intervals and underloaded (UL) intervals. We provide regime switching criteria, with which the computation alternates between OL and UL intervals until reaching the end of the time horizon.

Steady-state performance. When all model parameters are constants, we establish a steady state for the $G/PH/s+GI$ stationary model and demonstrate the convergence to this steady state as $t \to \infty$. The result extends [5] and [3]. In [5] the steady state of the $G/GI/s + GI$ model was first established; in [3] the convergence to the steady state of the $G/M/s + GI$ model was developed.

A network of queues. Extending [4] where the $(G_t/M/s_t+GI)^m/M_t$ fluid network with an exponential service distribution was treated, we hereby generalize the analysis from the $G_t/PH/s_t+GI$ single-queue model to the $(G_t/PH/s_t+GI)^m/M_t$ network with a single customer class, multiple service pools, and a Markovian routing (the M_t). This network model has m queues or stations, each of which is a $G_t/PH/s_t+GI$ model. What is difficult here is that the total arrival rate at each queue is not part of the model data because it is the sum of the rates of external arrivals and feedbacks from other queues. To resolve this problem, we consider a bigger vector $\hat{\mathbf{B}}(t) \equiv (\mathbf{B}_1(t), \ldots, \mathbf{B}_m(t))$, where $\mathbf{B}_i(t)$ is an n_i-dimentional row vector of the fluid in service at queue i and n_i is the total number of service phases at queue i, $i = 1, 2, \ldots, m$. We provide another ODE based algorithm to compute $\hat{\mathbf{B}}(t)$ and other standard performance measures.

1. REFERENCES

[1] S. Asmussen. Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[2] G. Latouche and V. Ramaswami. Introduction to matrix analytic methods in stochastic modeling. *Society for Industrial and Applied Mathematics*, 1999.

[3] Y. Liu and W. Whitt. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with customer abandonment. *Queueing Systems*, 67(2):145–182, 2011.

[4] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, forthcoming 2011.

[5] W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54:37–54, 2006.