A Simple Diffusion Limit for Flows in Stable Queueing Networks

Yoni Nazarathy[1], Gideon Weiss[2]

[1] *Swinburne University of Technology, Australia*

[2] *The University of Haifa, Israel*

La Trobe Statistics Colloquium,

June 17, 2011

Overview

Part 1: A brief survey on queueing networks

Part 2: Network model

Part 3: A simple diffusion limit for flows

Part 1: A brief survey on queueing networks

A single server queue

Arrivals at rate λ

Services with mean μ^{-1}

$Q(t)$ is number of customers at time t

If $\lambda < \mu$ there is enough capacity: $Q(t)$ "stable"

The $M/M/1$ queue:

Arrival process is Poisson(λ)

Service times are i.i.d. exponential(μ)

"stable" implies that the Markov Chain, $Q(t)$ is positive recurrent

$\lim_{t\to\infty} P\big(Q(t) = k\big) = \big(1 - \frac{\lambda}{\mu}\big)\big(\frac{\lambda}{\mu}\big)^k, k = 0, 1, \ldots$

Output process is Poisson(λ)

Two Queues in Tandem

Arrival process is Poisson(λ)

Service times at server $i = 1, 2$ are i.i.d. exponential(μ_i)

Second server is also $M/M/1$.

If $\lambda < \mu_i$, $i = 1, 2$ then,

$$\lim_{t\to\infty} P\big(Q_1(t) = k_1, Q_2(t) = k_2\big) = \big(1 - \frac{\lambda}{\mu_1}\big)\big(\frac{\lambda}{\mu_1}\big)^{k_1}\big(1 - \frac{\lambda}{\mu_2}\big)\big(\frac{\lambda}{\mu_2}\big)^{k_2}$$

This is called "Product Form"

Jackson Networks

Jobs leaving queue i are routed to queue j w.p. p_{ij} or leave system w.p. $1 - \sum_{j=1}^{N} p_{ij}$. Routing is independent (Bernoulli).

Assume the sub-stochastic matrix $P = \big(p_{ij}\big)$ is stable, then $(I - P')^{-1}$ exists

Exogenous arrival rates: α_i

Assume "open"

Flow rates: $\lambda_i = \alpha_i + \sum_{j=1}^{N} \lambda_j p_{ji}$ or, $\lambda = (I - P')^{-1}\alpha$

Jackson (1956, 1963). If $\lambda_i < \mu_i$ for all nodes, then "Product Form":

$$\lim_{t\to\infty} P\big(Q_1(t) = k_1, \ldots, Q_N(t) = k_N\big) = \prod_{i=1}^{N} \big(1 - \frac{\lambda_i}{\mu_i}\big)\big(\frac{\lambda_i}{\mu_i}\big)^{k_i}.$$

Some Extensions have Product Form

- State dependent service rates in queues
- Closed networks (sometimes also called Jackson)
- Other service policies (PS, LCFS...)
- Many other extensions...

Modifications Often Break Down the Product Form

- A tandem queue with non-exponential service times
- A simple re-entrant line - the service policy plays a key role

Jackson like networks with general renewal exogenous arrival and general service times are sometimes called Generalized Jackson Networks:

- Not product form so no explicit results
- Stability Results
- Diffusion approximations when all nodes are at high load
- Parametric decomposition approximations

Parametric Decompositions

Ward Whitt 1983, The Queueing Network Analyzer.

The basic idea is to assume arrival processes are renewal process with specified rate and variance (squared coefficient of variation (SCV))

- Step 1: Compute $\lambda = (I - P')^{-1}\alpha$
- Step 2: Compute c^2 (vector of SCVs) of flows by means of a **non-rigorous** approximation
- Step 3: Assume that flows are renewal processes with (λ, c^2) and approximate queue lengths at each node as a GI/G/1 queue

Yields exact results in Poisson arrivals/Exponential service case - otherwise an approximation without many theoretical bounds and asymptotics

Some improvements (by Whitt and others) attempt to incorporate correlations of flows in calculation of c^2. Our contribution is in finding exact asymptotic correlations...

Parametric Decomposition: Calculation of SCVs

Each queue first collects a superposition of flows (Merging), then averages variability with that of service times and then splits flows and "distributes" the variability

Merging:

$$c^2_{\text{arrival}} = \sum_{j=0}^{N} \frac{\lambda_j p_{ji}}{\lambda_i} c_j^2$$

In the queue:

$$c^2_{\text{departure}} = \left(\frac{\lambda_i}{\mu_i}\right)^2 c^2_{\text{service}} + \left(1 - \left(\frac{\lambda_i}{\mu_i}\right)^2\right) c^2_{\text{arrival}}$$

Splitting:

$$c^2 = p\, c^2_{\text{total}} + (1 - p)$$

Mean Queue Length (Approximation for GI/G/1 Queue):

$$\frac{\lambda_i/\mu_i}{1 - \lambda_i/\mu_i} \frac{c^2_{\text{arrival}} + c^2_{\text{service}}}{2}$$

Part 2: Network model (the we consider)

Classes/Nodes/Queues: $i = 1, ..., N$ with a non-idling service policy

Input data realizations (primitives):

Arrival counts: $A_i(t)$

(Non-stop) Service counts: $S_i(t)$

Routing counts: $\Phi_{ij}(k)$, $\Phi_{i0}(k) = k - \sum_{j=1}^{N} \Phi_{ij}(k)$

Initial Conditions: $Q_i(0)$ (we'll set $= 0$ for simplicity)

Resulting processes:

Queue levels: $Q_i(t)$

Cumulative work: $T_i(t)$

Departure counts: $D_{ij}(t)$

Dynamics:

$$\begin{aligned} D_{ij}(t) &= \Phi_{ij}\big(S_i\big(T_i(t)\big)\big) \\ Q_i(t) &= Q_i(0) + A_i(t) + \sum_{j=1}^{N} D_{ji}(t) - \sum_{j=0}^{N} D_{ij}(t) \end{aligned}$$

Assumptions on primitives

Driving independent random variables:

- Inter-arrival times: $a_i(j), j = 1, 2, \ldots$ $E[a_i(1)] = \alpha_i^{-1}$. SCV is $d_i^2 < \infty$
- Service durations: $s_i(j), j = 1, 2, \ldots$ $E[s_i(1)] = \mu_i^{-1}$. SCV is $c_i^2 < \infty$
- Multinomial $(1, p_{i\cdot})$ random vectors $(\phi_{i\cdot})$ with $P(\text{moving to} j) = p_{ij}$

Resulting processes:

- $A_i(t) = \sup\{\ell : \sum_{j=1}^{\ell} a_i(j) \leq t\}$
- $S_i(t) = \sup\{\ell : \sum_{j=1}^{\ell} s_i(j) \leq t\}$
- $\Phi_{ij}(k) = \sum_{\ell=1}^{k} \phi_{ij}(\ell)$

Fluid and Diffusion Scalings

- Fluid scalings: $\bar{U}^n(t) = \frac{U(nt)}{n},\ n = 1, 2, \ldots$
- Fluid limit (FSLLN): $\bar{U}(t) = \lim_{n\to\infty} \bar{U}^n(t)$ (u.o.c)
- Diffusion scalings: $\hat{U}^n(t) = \frac{U(nt) - \bar{U}(nt)}{\sqrt{n}},\ n = 1, 2, \ldots$
- Diffusion limit (FCLT): $\hat{U}^n(t) \Rightarrow \hat{U}(t)$ (in function space)

In case of discrete parameter processes (such as Φ), use $U(t) = U(\lfloor t \rfloor)$

FSLLNs for primitives: $\bar{A}_i(t) = \alpha_i t$, $\bar{S}_i(t) = \mu_i t$, $\bar{\Phi}_{ij}(t) = p_{ij} t$

FCLTs for primitives: $\hat{A}_i(t) = BM(\alpha_i d_i^2)$, $\hat{S}_i(t) = BM(\alpha_i c_i^2)$,
$\hat{\Phi}_{i\cdot}(t) = BM(\Sigma_i)$

We further assume (under stability) $\bar{T}_i(t) = \theta_i := \lambda_i/\mu_i$

Part 3: A Simple Diffusion Limit for Flows

Desired: The law of $\hat{D}(t)$

Result: $\hat{D}(t)$ is a multi-dimensional Brownian motion with covariance matrix represented in terms of λ, d^2 and P (not c^2)

Implication: In case flows are renewal (and under technical Uniform Integrability assumptions), exact SCVs are immediately obtained - essentially provides an alternative/additional building block for parametric network decomposition methods

Note: Flows are typically not-renewal - and it is not clear if using exact asymptotic variance rates improves over previous parametric decomposition methods

Future (student) work: Integrate our exact asymptotic variability results with parametric decomposition methods that take variability of service times into account

Lemma: The following $(N+2)N$ equations hold:

$$\begin{aligned}
\hat{D}_{ij}^n(t) &= \hat{\Phi}_{ij}^n(\bar{S}_i^n(\bar{T}_i^n(t))) + p_{ij}\hat{S}_i^n(\bar{T}_i^n(t)) + p_{ij}\mu_i\hat{T}_i^n(t), \\
&\quad \text{for } i = 1, \ldots, N, \quad j = 0, \ldots, N. \\
\hat{Q}_i^n(t) &= \hat{A}_i^n(t) + \sum_{j=1}^{N} \hat{D}_{ji}^n(t) - \sum_{j=0}^{N} \hat{D}_{ij}^n(t), \\
&\quad \text{for } i = 1, \ldots, N.
\end{aligned}$$

Proof: Use,

$$D_{ij}(nt) = \Phi_{ij}(S_i(T_i(nt))) = \Phi_{ij}(n\bar{S}_i^n(\bar{T}_i^n(t)))$$

and

$$\bar{D}_{ij}(t) = \bar{\Phi}_{ij}(\bar{S}_i(\bar{T}_i(t))) = p_{ij}\mu_i\theta_i t,$$

and manipulate using definitions of fluid and diffusion scalings.

Lemma: Denote: $\tilde{\Phi}_{ij}^n(t) = \hat{\Phi}_{ij}^n(\bar{S}_i^n(\bar{T}_i^n(t)))$ and $\tilde{S}_i^n(t) = \hat{S}_i^n(\bar{T}_i^n(t))$ then,

$$\begin{bmatrix} \hat{D}^n(t) \\ \hat{T}^n(t) \end{bmatrix} = G \begin{bmatrix} \tilde{\Phi}^n(t) \\ \tilde{S}^n(t) \\ \hat{A}^n(t) \end{bmatrix} + F\hat{Q}^n(t),$$

with the matrixes G and F are given in terms of problem data and $\lambda = (I - P')^{-1}\alpha$.

Proof: Manipulate equations in previous lemma (structure of matrixes is a bit messy)

Theorem:

Assume that $\hat{Q}^n(t) \Rightarrow 0$ and $\lim_{n\to\infty} \bar{T}_i^n(t) = \theta_i t$, then,

$$\begin{bmatrix} \hat{D}^n(t) \\ \hat{T}^n(t) \end{bmatrix} \to BM\Big(G\,\Sigma\,G'\Big),$$

$$\text{with,}\;\; \Sigma = \begin{bmatrix} \lambda_1\Sigma_1 & & & & \\ & \ddots & & & \\ & & \lambda_N\Sigma_N & & \\ & & & \text{diag}(\lambda_i c_i^2) & \\ & & & & \text{diag}(\alpha_i d_i^2) \end{bmatrix},$$

and Σ_i being an $N+1$ dimensional covariance matrix of $\phi_{i,0}, \ldots, \phi_{i,N}$.

Proof: Simply apply functional CLT for renewal processes and the continuous mapping theorem to equations of previous lemma. Primitives Φ, S and A converge to BM, Q converges to 0 by assumption (stability)

Wrap Up

Our results immediately gives the marginal asymptotic variance of flows (or departures). It is also evident (from the structure of A) that the variability of services does not affect flows.

Given suitable UI conditions (technical) we get the asymptotic variance rates of the flows.

Assuming (typically falsely) that the flows are renewal processes, these are also the SCV's.

Hence we have an alternative parametric network decomposition method with **exact** asymptotic variances.

The practicality of this simple result (for better decomposition approaches) remains open

References

- G. Bolch, S. Greiner, H. de Meer and K. S. Trivedi, Queueing Networks and Markov Chains, 1998.
- H. Chen and D. D. Yao, Fundamentals of Queueing Networks, 2001.
- Yoni Nazarathy and Gideon Weiss, A Simple Diffusion Limit for Flows in Stable Queueing Networks, in preparation.
- Ahmad Al-Hanbali, Michel Mandjes, Yoni Nazarathy and Ward Whitt. The asymptotic variance of departures in critically loaded queues. Advances in Applied Probability, 43(1):243-263, 2011.