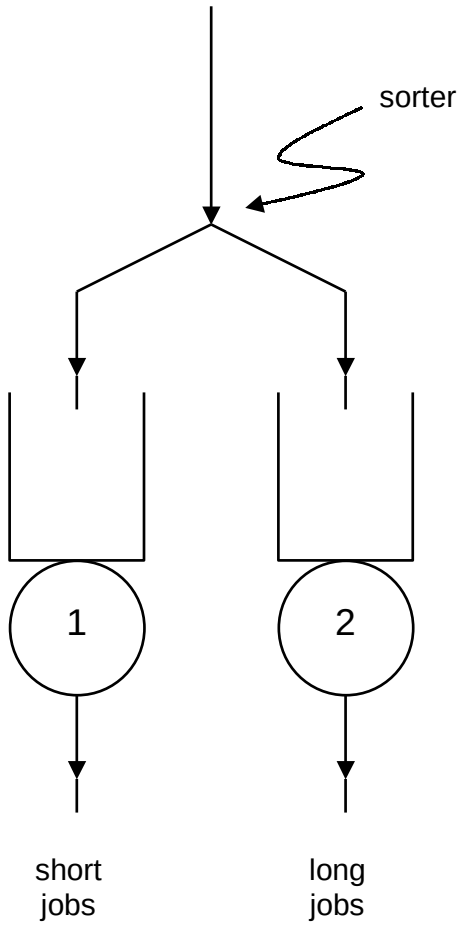
Heavy-Traffic Approximations for Many-Server and Multi-Server Systems

Peter W. Glynn

Stanford University

1. Sorting Jobs by Size



- processing times are iid with common distribution F.
- server one serves at rate a, the other at rate b $(a + b) = 1$.
- sorter directs jobs shorter than q to 1, larger than q to 2.
- what is the optimal split point q?

Heavy-Traffic Analysis

$$\begin{aligned} \lambda &= \text{arrival rate of jobs} \\ E\,V &= \text{mean processing time for typical job} \end{aligned}$$

Assume $\rho \stackrel{\triangle}{=} \lambda\, EV < 1$ (but close to one)

Mathematically, we formulate this as: Consider a family of queues, indexed by ρ, for which:

- Processing times in ρ'th system, namely $V_1(\rho)$, $V_2(\rho)$, $\ldots$, are i.i.d. with
$$V_i(\rho) = \rho V_i \qquad (E\,V_i = 1)$$
- Arrival process $N = (N(t) : t \geq 0)$ is independent of processing times.
- $\varepsilon^{-1/2}\,(\varepsilon N(t/\varepsilon) - t) \Longrightarrow c\,B(t)$,
so $c^2 = \sigma^2$ (generalized squared coefficient of variation)

The natural "split point" $q = q(\rho)$ splits the traffic so that both queues are equally loaded:

$$\frac{\lambda \int_0^{q(\rho)} x f_\rho(x) dx}{a} = \frac{\lambda \int_{q(\rho)}^{\infty} x f_\rho(x) dx}{b}$$

Then,

$$q(\rho) = \rho\, q$$

where

$$\int_0^q x f(x) dx = a\, E\, V$$

But we can do better by perturbing the split point:

$$\begin{aligned} q^*(\rho) &= q(\rho) + r^*(1-\rho) \\ &= q\,\rho + r^*(1-\rho) \end{aligned}$$

What is the optimal choice of r^*?

For r fixed (i.e. $q(\rho) = q + r(1-\rho)$)
we have that:

$$(1-\rho)\, Q_\rho(\infty) \Longrightarrow Z_1(\infty) + Z_2(\infty)$$

where,

$$Z_1(\infty) \stackrel{\mathcal{D}}{=} m_1(r)\, \mathsf{exp}_1(1)\,,$$

$$Z_2(\infty) \stackrel{\mathcal{D}}{=} m_2(r)\, \mathsf{exp}_2(1)\,,$$

Here,

$$m_1(r) = \frac{\mathsf{var}\left[V\, I(V \le q)\right] + a^2c^2}{2\,a\,\left(a - rq\, f(q)\right)}\,,$$

$$m_2(r) = \frac{\mathsf{var}\left[V\, I(V > q)\right] + b^2c^2}{2\,b\,\left(b + rq\, f(q)\right)}$$

To minimize

$$EQ_\rho(\infty)$$

(or, equivalently, to minimize $EW_\rho(\infty)$),
minimize over r.

$$r^* = \frac{\beta^{1/2} a \; + \; \alpha^{1/2} b}{(\alpha^{1/2} + \beta^{1/2}) \; q \, f(q)}$$

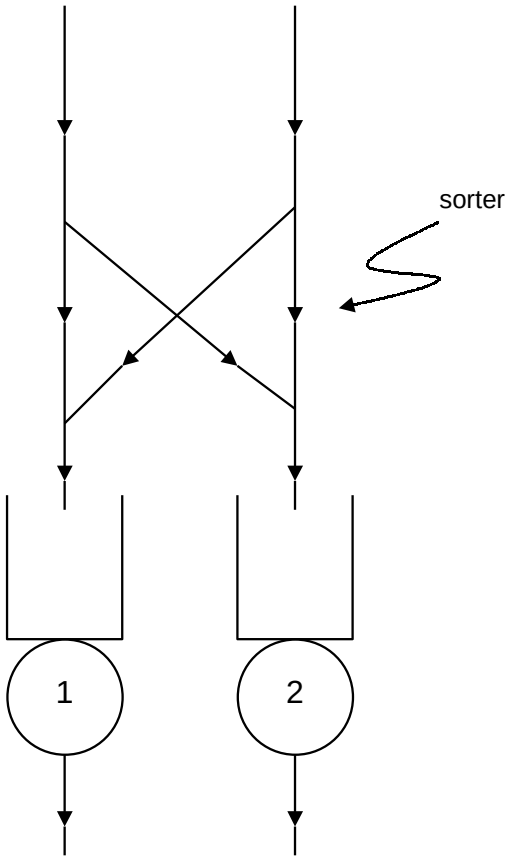
where

$$\alpha = \frac{\text{var}\left[V \; I(V \leq q)\right] + a^2 c^2}{2a}$$

$$\beta = \frac{\text{var}\left[V \; I(V > q)\right] + b^2 c^2}{2b}$$

This is easy to translate into an approximation for the optimal "split point" for a given "real-world" system (with no ρ-parametrization).

2. Sorting Jobs by Originating Population



- processing times are i.i.d. mixture

$$F = \lambda_1 F_1 + \lambda_2 F_2$$

- sorter allocates fraction p_i of population F_i to server i
- The optimal fractions p_1 and p_2 are the solution to a simple system of linear equations that involve all the first and second moment data in this problem (i.e. the mean and covariance structure of the arrival streams for the two job classes, as well as the means and variances of the processing times for the two streams).

Multi-server queues

vs

Many-server queues

vs

Infinite-server queues

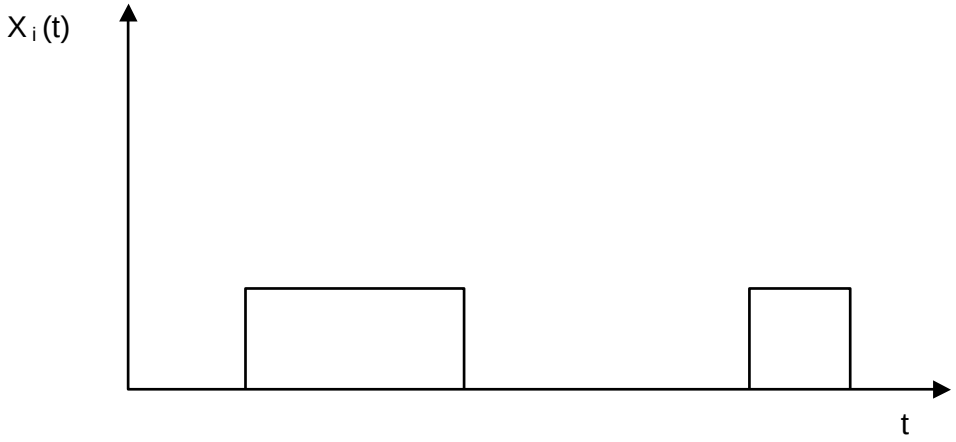
- Many-server queues and infinite-server queues are of principal interest when offered load is high
- When offered load is high and system is rarely congested, structure of two models is closely related (both quantitatively and qualitatively)
- Infinite-server systems (and related models) are of interest in their own right (e.g. electricity demand, warranty analysis, etc.)

Infinite server queues

- $M/G/\infty$ available in analytical closed form
- $G/G/\infty$ can be analyzed when arrival rate is high
- In many application settings, we are dealing with "time-of-day" effects, "seasonality" effects, etc
- Need extension to non-stationary setting
- Extension to $M(t)/G/\infty$ known
- We study non-stationary $G/G/\infty$ when arrival rate is high

Asymptotic Regime 1

- n individual independent sources
- Each individual source takes form



- $Q_n(t) = \sum_{i=1}^{n} X_i(t)$
- Gaussian limit as $n \to \infty$:

$$n^{1/2}\,(Q_n(\cdot) - n\,q(\cdot)) \Longrightarrow Z(\cdot)$$

in $D[0,\infty)$, where $Z = (Z(t) : t \geq 0)$ is a non-stationary Gaussian process that inherits the covariance structure of $X = (X(t) : t \geq 0)$,

i.e. $\mathrm{cov}(Z(s), Z(t)) = \mathrm{cov}(X_i(s), X_i(t))$.

- Large deviations for tail of $Q_n(t)$ follows from corresponding theory for sums of i.i.d. r.v.'s

$$\mathrm{P}\left(Q_n(t) > n(q(t) + \varepsilon)\right) \sim n^{-1/2}\,a(t)\,\exp(-n\,b(t))$$

Asymptotic Regime 2 (Borovkov, Whitt, G, etc.)

- single source running at rate n

$$\frac{N_n(\cdot) - n\,m(\cdot)}{\sqrt{n}} \Longrightarrow \xi(\cdot)$$

in $D[0,\infty)$, where $\mathsf{P}\left(\xi \in C[0,\infty)\right) = 1$

- at arrival times, an independent copy of the process $C(\cdot)$ is initiated, and

$$\mathsf{P}\left(C_i \in \cdot\,|N_n\right) = \mathsf{P}\left(\cdot\,|A_i^n\right)$$

where

$$\begin{aligned} h_1(t,s) &= E\left[C_i(t)\,|A_i^n = s\right] \\ h_2(t_1,t_2,s) &= E\left[C_i(t_1)C_i(t_2)\,|A_i^n = s\right] \\ h_3(t_1,t_2,s) &= h_2(t_1,t_2,s) - h_1(t_1,s)h_1(t_2,s) \end{aligned}$$

and

$$\sup_{s\geq 0} E\left[|C_i|^3(\infty)\,|A_i^n = s\right] < \infty$$

- Set

$$\Gamma_n(t) = \sum_{i=1}^{N_n(t)} C_i(t - A_i^n)$$

Then,

$$\frac{\Gamma_n(\cdot) - n\gamma(\cdot)}{n^{1/2}} \Longrightarrow Y(\cdot)$$

in $D[0, \infty)$, where

$$Y(t) = \theta(t) + \int_0^t h_1(t-s, s)\, d\xi(s)$$

Here, $\theta = (\theta(t) : t \geq 0)$ is a mean-zero Gaussian process (independent of ξ) with covariance

$$E\theta(t)\theta(t+u)$$

$$= \int_0^t h_3(t-s, t+u-s, s)\, dm(s)$$

and

$$\gamma(t) = \int_0^t h_1(t-s, s)\, dm(s)\,.$$

see Sun – G 04

Also, we have large deviations for the tail of $\Gamma_n(\cdot)$

Assume:

- $\frac{1}{n} \log E \exp\left(\sum_{i=1}^{m} \theta_i \left[N_n(t_i) - N_n(t_{i-1})\right]\right)$

$$\longrightarrow \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \kappa(\theta_i, s) ds$$

- $\varphi_C(\theta, t, s) = E\left[exp(\theta C_i(t)) \,|A_i^n = s\right] < \infty$

Then

$$\frac{1}{n} \log \mathsf{P}\left(\Gamma_n(t) > n(\gamma(t) + \varepsilon)\right) \longrightarrow -b(t)$$

Under additional assumptions,

$$\mathsf{P}\left(\Gamma_n(t) > n(\gamma(t) + \varepsilon)\right) \sim n^{-1/2}\, a(t) \exp(-nb(t))$$

Sun – G 04

Some Modeling Remarks

- Conventional heavy-traffic limit theory
  (s servers; s fixed)

  For "short-range dependent traffic" and finite variance processing times, system is approximated by reflecting Brownian motion.

  i.e. behavior depends on small number of parameters (mean, variance).

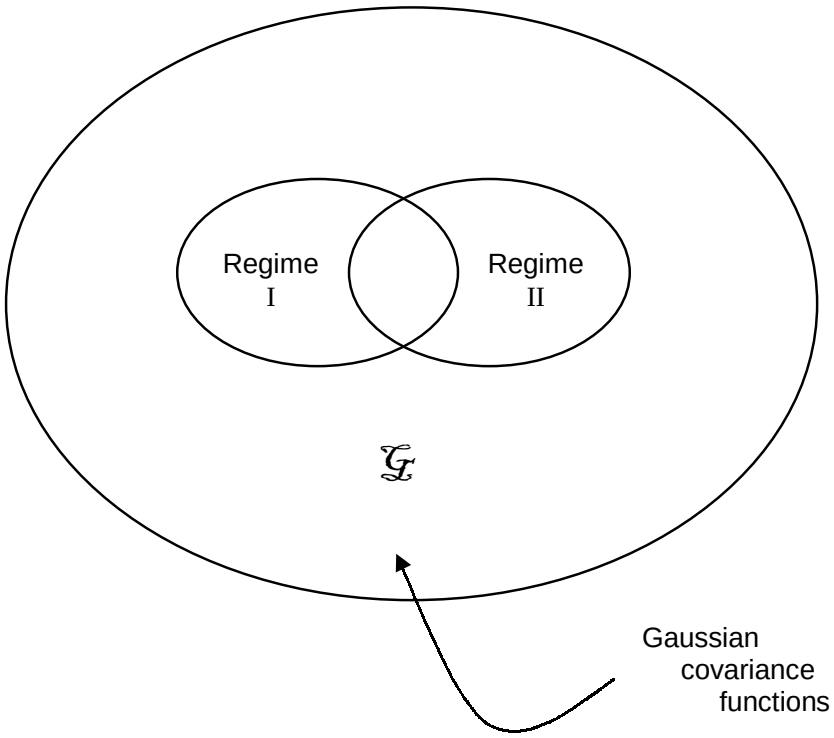
- Heavy-traffic limits for many-server and infinite server systems

  Basic Lesson: In great generality, system can be approximated as a functional of a Gaussian process.

  Can we assume that the covariance function is of a particular form?

  e.g. Borovkov covariance function essentially determined by distribution of processing times.

- But, this simplification is misleading!
- Many source asymptotic is equally compelling as an explanation for heavy-traffic (and has different form of covariance function)



- Claim: $\mathcal{G}$ is the "closure" of set of covariance functions obtainable as limits of reasonable queueing models.
- Implication: No justification for assuming simplified covariance function (in many application settings).

However, the Gaussian limit already is a significant simplification.

Multi-Server Setting:

Presence of multiple servers need not significantly improve performance (in terms of tail asymptotics)

correlated traffic

Question of model robustness