Some Comments Related to Limit Processes and Optimization for QED Queues

Marty Reiman
Bell Labs, Lucent Technologies
Murray Hill, NJ

M/M/N Diffusion Limit

Consider a family of M/M/N queues with arrival rates λ_N and service rate μ, and define $\rho_N = \lambda_N / N\mu$.

Let $Q_N(t) = \#$ of customers in the system at time t, $t \geq 0$, and let

$$\hat{Q}_N(t) = N^{-1/2} \left[Q_N(t) - N \right].$$

Theorem (Halfin and Whitt, 1981).
If $\sqrt{N}(1 - \rho_N) \to \beta$, with $-\infty < \beta < \infty$, and $\hat{Q}_N(0) \xrightarrow{d} \hat{Q}(0)$, then $\hat{Q}_N \xrightarrow{d} \hat{Q}$ in $D[0, \infty)$, where $\{\hat{Q}(t),\ t \geq 0\}$ is a diffusion on $\mathbb{R}$ with infinitesimal drift $m(x)$ given by

$$m(x) = \begin{cases} -\mu\beta \quad , & x \geq 0 \\ -\mu(x + \beta), & x \leq 0 \end{cases}$$

and infinitesimal variance $\sigma^2 = 2\mu$.

Comparison of Two Heavy Traffic Regimes

'Classical'	**QED (Halfin-Whitt)**
$\rho \to 1$ with N fixed	$\rho \to 1, N \to \infty, \sqrt{N}(1-\rho) \to c$
queue length not centered	queue length is centered
limit process is reflected	limit process is unconstrained
limit drift is constant	limit drift is state dependent
time in service: $O(n^{-1}) \to 0$	time in service is $O(1)$

The 'M' Model

Consider a call center handling 2 skills and having 3 agent pools.

We assume:

- Call arrival processes are Poisson, with rates λ_i, $i = 1, 2$
- There are N_1 Type 1 agents, who can only serve skill 1
- There are N_2 Type 2 agents, who can only serve skill 2
- There are N_0 Type 0 agents, who can serve both skills
- Service times are exponentially distributed, with rates μ_i, $i = 1, 2$ (which depend only on skill, not agent)

Combined Staffing & Scheduling Problem

Given staffing costs $c_i > 0, i = 0, 1, 2$ and waiting costs $d_j > 0, j = 1, 2$, choose staffing levels N_0, N_1, N_2 and a non-preemptive (and non-anticipating) scheduling policy π to minimize

$$\sum_{i=0}^{2} c_i N_i + \sum_{j=1}^{2} d_j E \int_0^\infty e^{-\gamma t} Y_j^\pi(t) dt$$

where $Y_j^\pi(t)$ is the number of skill j customers waiting in the queue under the scheduling policy π.

Assume that $c_0 > max(c_1, c_2)$: flexible servers cost more.

The QED Regime for the M Model

Consider a family of systems, indexed by n, where μ_i are held fixed and

$$N_i(n) = \alpha_i n + o(\sqrt{n}), \quad i = 0, 1, 2,$$

$$\lambda_j(n) = \bar{\lambda}_j n + \beta_j \sqrt{n} + o(\sqrt{n}), \quad j = 1, 2,$$

with $0 < \alpha_i < \infty, 0 < \bar{\lambda}_j < \infty$ and $-\infty < \beta_j < \infty$.

QED conditions:

$$\bar{\lambda}_i > \alpha_i \mu_i, \quad i = 1, 2,$$

and

$$\frac{\bar{\lambda}_1}{\mu_1} + \frac{\bar{\lambda}_2}{\mu_2} = \alpha_0 + \alpha_1 + \alpha_2.$$

Theorem (Atar, 2005). The asymptotically optimal waiting cost

$$\inf_{\pi \in \Pi} \sum_{j=1}^{2} d_j E \int_0^\infty e^{-\gamma t} \hat{Y}_j^\pi(t) dt$$

does not depend on $(\alpha_0, \alpha_1, \alpha_2)$ as long as $\alpha_i > 0, i = 0, 1, 2,$ and the QED conditions are satisfied.

An Open Problem

The waiting cost does not depend on $(\alpha_0, \alpha_1, \alpha_2)$, but the staffing cost $n \sum \alpha_i c_i$ clearly does.

Recall that $c_0 > max(c_1, c_2)$, so to reduce staffing costs we take $\alpha_0 \to 0$.

But $\alpha_0 = 0$ is not covered by extant theory.

In particular, a reasonable guess is that the asymptotically optimal staffing level satisfies

$$N_i(n) = \frac{\bar{\lambda}_1}{\mu_1} n + \eta_i \sqrt{n} + o(\sqrt{n}), \quad i = 1, 2$$

and

$$N_0(n) = \eta_0 \sqrt{n} + o(\sqrt{n}).$$

with $0 < \eta_0 < \infty$ and $-\infty < \eta_i < \infty,\ i = 1, 2.$

Open problem: Solve for the optimal scheduling policy for a given (η_0, η_1, η_2), determine the associated waiting cost and optimize over (η_0, η_1, η_2).