Asymptotic output variance of service systems stabilized by loss

D. J. Daley
Department of Mathematics and Statistics
The University of Melbourne

(Joint work with Yoni Nazarathy, ex Swinburn Univ Tech)
(walking distance from home)
(and Johan van Leeuwaarden, Eurandom and Eindhoven TU)

YN's August 2010 seminar
2011 QUESTA 'problem' papers by YN and DJD

[NW08] = YN + Weiss, 2008, QUESTA)

Questions are of ‘Mathematical Interest’

Curiosity driven research

‘Toy models’

(focus on key assumptions)

(number-crunching construction of Year 12 aggregate)

(bibliometric measures: ‘impact factors’)

[NW08] (QUESTA 2008) BRAVO effect:

Balancing **R**educes **A**symptotic **V**ariance of **O**utputs

M/M/1/K, Buffer of size K, Stationary

Arrivals are Poisson at rate λ,

Service times i.i.d. exponential at rate μ,

$$N_{\text{dep}}(0,t] \qquad \mathrm{E}\big(N_{\text{dep}}(0,t]\big) = \begin{cases} \lambda t & \text{if } \lambda < \mu, \\ \mu t & \text{if } \lambda \geq \mu. \end{cases}$$

$$\operatorname{var} N_{\text{dep}}(0,t] \sim \begin{cases} \lambda t & \text{if } \rho < 1, \\ \mu t & \text{if } \rho > 1. \end{cases}$$

$$\mathcal{D}_{\mathrm{M/M/1/}K} := \lim_{t\to\infty} \frac{\operatorname{var} N_{\text{dep}}(0,t]}{\mathrm{E}\big(N_{\text{dep}}(0,t]\big)}$$

$$\lim_{K\to\infty} \mathcal{D}_{\mathrm{M/M/1/}K} = \begin{cases} 1 & \text{if } 0 < \rho < \infty \text{ except for} \\ \frac{2}{3} & \text{if } \rho = 1. \end{cases}$$

(Q.1): WHY the discontinuity at $\rho := \lambda/\mu = 1$?

(Q.2): Is there similar behaviour with s servers ?

(either $s \geq 2$ or $s \uparrow$ or $s \to \infty$)

[NW08] includes a graph for systems $\mathrm{M/M/}s/(K-s)$ $(K\uparrow)$

'correct' family is for systems $\mathrm{M/M/}s/\sqrt{s}$ $(s\uparrow)$

Theoretical Physicist:
‘Examine a system at its critical point(s)’

Branching process: (biological processes)
Describe both sub- and super-critical behaviour
Transition regime when mean offspring ≈ 1

In GI/GI/s, $\rho = 1$ is critical point.

System ‘dull’ for $\rho > 1$.

GI/GI/s/K has critical point $\rho = 1$:

Demarcation point between two stable phases.

$\rho = 1$: critical pt. when arrival rate λ, service rate μ/s,

$$\rho = \lambda/[s \cdot \mu/s] \; = \; \frac{\text{arrival rate}}{\text{total service rate}}$$

Why $\dfrac{2}{3}$? Grimmett: ‘Magic’ constants
(ratio of ‘small’ integers)

As $\rho \uparrow 1$, $\mathcal{D}_{\mathrm{GI/GI/1}} \to 2\Big(1 - \dfrac{2}{\pi}\Big)$

$\mathcal{D}_{\ldots}$ is second-order rate: $\dfrac{\text{variance}}{\text{mean}}$

For s-server system, have s servers each working at rate μ/s, so system is 'balanced' when $\lambda = s[\mu/s] = \mu$. Write $\rho = \lambda/\mu$ as in 1-server case, so $\rho = 1$ for 'balance'.

Variance of N_{dep} has same asymptotics as for 1-server case except for $\rho \approx 1$:

In M/M/s/K with $\rho = 1$, when $s, K \to \infty$ in such a way that $K/\sqrt{s} \to \eta$ for some $0 < \eta \le \infty$,

$$\lim_{s,K\to\infty} \mathcal{D}_{\mathrm{M/M/}s/K} = \tfrac{2}{3} - L(\eta)$$

for a function $L(\eta) \to 0$ as $\eta \to \infty$ [e.g. fix s at some finite integer ≥ 1).

When $\rho = 1 - \beta/\sqrt{s}$ and $s \to \infty$, there is non-trivial limit behaviour but the limit $f(\eta, \beta)$ say is no longer $\frac{2}{3}$.

Why $K = O(\sqrt{s}\,)$? ?

(1) This is 'QED' regime: 'Quality and Efficiency Driven' (high utilization of servers with low probability of any appreciable waiting time) (Halfin & Whitt, c.1981)

(2) Find $K, s \to \infty$ such that both
$P_-^{(s)} := \Pr\{\text{arriving customer has no wait}\}$
$1 - P_-^{(s)} = \Pr\{\text{all servers busy}\}$
have positive limits . . . (solution: $K = O(\sqrt{s}\,)$).

[Re (Q.1): $\rho = 1 - \beta/\sqrt{s}$: limit discty vanishes at finer scale.]

(Q.2): We can find (expressions for) $\lim_{s,K\to\infty} \mathcal{D}_{\mathrm{M/M}/s/K}$.

Let $\{\pi_j\}$ be the stationary distribution of the system-size process

$$\pi_j = \Pr\{Q(t) = j\} \qquad \text{(all } t\text{)}.$$

$Q(t)$: birth–death process on state space $\{0, 1, \ldots, s + K\} = \{0, 1, \ldots, J\}$.

We need formulae for second moments in terms of birth and death rates . . .

[NW08] has an expansion for $\lim_{t\to\infty} \operatorname{var} N_{\mathrm{dep}}(0, t]/\mathrm{E}N_{\mathrm{dep}}(0, t]$ in $\mathrm{M/M}/1/K$ that comes from birth–death process expression due to Ward Whitt.

[NW08]'s formula:

$$(1 - \pi_J)(\mathcal{D} - 1) = -2\pi_J \sum_{i=0}^{J} P_i\Bigl(1 - \frac{\pi_J}{\pi_i}\,P_i\Bigr) \qquad (*)$$

where $P_i = \pi_0 + \cdots + \pi_i$. [NB: RHS $\geq -\frac{1}{2}$]

Limit relations (not given today) follow from $(*)$ via

(a) the sum equals $\int_{A_s} g_s(u)\, p_s(\mathrm{d}u)$ for appropriate simple functions g_s, atomic measures p_s and sets (intervals) A_s; look for weak cgce of measures and uniform cgce of functions.

(b) stationary probabilities π_i are like Poisson probabilities — use local CLT for individual terms or CLT for Poisson distribution. EXCEPT: Need rate of convergence so use Berry–Esseen CLT to get next order term, or for local CLT, use Feller's (1950/60/68) cgce of binomial probabilties to normal density yielding *uniform* bounds on error terms.

(c) $(*)$ is discrete sum over increasing number of terms — use cgce of discrete sums to limit as for a Riemann integral.

CONSERVATION arguments.

Output = Arrivals – lost customers

(Q.3): Diffusion approximations ? ? ?

$X_s(t) = \dfrac{Q_s(t) - s}{\sqrt{s\rho}}$ converges to a process-limit ($s \to \infty$).

How do we use this to give $\mathcal{D}$ (or appropriate analogue) ? ?

[Recall: $N_{\text{dep}}(\cdot)$ consists of

(a) downwards-only

(b) jumps of unit size . . .

So, approximation needs to be rectifiable (Brownian paths are not)]