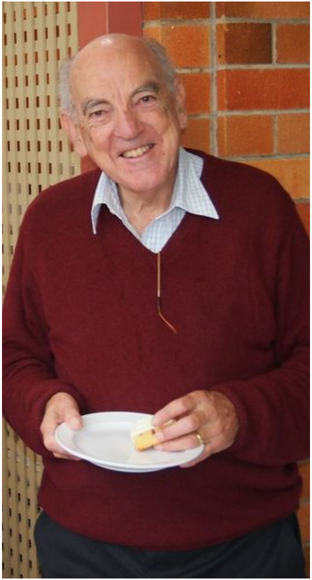
BRAVO for QED Queues

Yoni Nazarathy,
The University of Queensland

Joint work with

Daryl J. Daley, The University of Melbourne,
Johan van Leeuwaarden, EURANDOM, Eindhoven University of Technology.

Applied Probability Society Conference,
Costa Rica,
July, 2012.



Daryl J. Daley

Queues and Counting Processes:

$$Q(t) = Q(0) + \Big(A(t) - L(t)\Big) - \Big(R(t) + D(t)\Big)$$

Queues and Counting Processes:

$$Q(t) = Q(0) + \Big(A(t) - L(t)\Big) - \Big(R(t) + D(t)\Big)$$

Why analysis of the **output** counting process $D(t)$?

- Orders
- Production
- Arrival process to a downstream queueing system

Queues and Counting Processes:

$$Q(t) = Q(0) + \Big(A(t) - L(t)\Big) - \Big(R(t) + D(t)\Big)$$

Why analysis of the **output** counting process $D(t)$?

- Orders
- Production
- Arrival process to a downstream queueing system

Daryl Daley, "*Queueing Output Processes*", Advances in Applied Probability, 1976.

Queues and Counting Processes:

$$Q(t) = Q(0) + \Big(A(t) - L(t)\Big) - \Big(R(t) + D(t)\Big)$$

Why analysis of the **output** counting process $D(t)$?

- Orders
- Production
- Arrival process to a downstream queueing system

Daryl Daley, "*Queueing Output Processes*", Advances in Applied Probability, 1976.

Some performance measures of interest

- The law of $\{D(t),\ t \geq 0\}$
- $\mathbb{E}[D(t)]$, $\mathrm{Var}\big(D(t)\big)$
- $\lambda^* := \lim_{t\to\infty} \frac{\mathbb{E}[D(t)]}{t}$, $\quad \overline{V} := \lim_{t\to\infty} \frac{\mathrm{Var}\big(D(t)\big)}{t}$, $\quad \mathcal{D} := \frac{\overline{V}}{\lambda^*}$
- Asymptotic normality: $D(t) \sim \mathcal{N}\Big(\lambda^* t,\ \overline{V} t\Big)$, large t

Outline

- A (new) formula for asymptotic variance of outputs, $\mathcal{D} := \frac{\overline{V}}{\lambda^*}$
- Single servers (older BRAVO results)
- Many server scaling (new BRAVO results)

Asymptotic Variance of Outputs

Finite Birth-Death Asymptotic Variance

- Irreducible birth-death process on finite state space
- Birth rates: $\lambda_0, \ldots, \lambda_{J-1}$
- Death rates: $\mu_1, \ldots, \mu_J$
- Stationary distribution: $\pi_0, \ldots, \pi_J$
- $D(t)$ is number of downward transitions (deaths) during $[0, t]$, each "filtered" independently with state-dependent probabilities, $q_1, \ldots, q_J$.
- e.g. The departure process (served customers) in M/M/s/K+M systems

Of interest:

$$\mathcal{D} = \frac{\overline{V}}{\lambda^*} = \lim_{t\to\infty} \frac{\mathrm{Var}\big(D(t)\big)}{\mathbb{E}[D(t)]}$$

Finite Birth-Death Asymptotic Variance Formula

Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2013

$$\mathcal{D} := \lim_{t\to\infty} \frac{\mathrm{Var}\big(D(t)\big)}{\mathbb{E}[D(t)]} = 1 - 2\sum_{i=0}^{J}(P_i - \Lambda_i^*)\Big(q_{i+1} - \frac{\lambda^*}{\pi_i\lambda_i}(P_i - \Lambda_i^*)\Big),$$

with,

$$P_i := \sum_{j=0}^{i} \pi_j, \qquad \lambda^* := \sum_{j=1}^{J} \mu_j q_j \pi_j, \qquad \Lambda_i^* := \frac{\sum_{j=1}^{i} \mu_j q_j \pi_j}{\lambda^*}.$$

Note: In Weiss, Y.N. 2008, similar expression for case $q_i \equiv 1$

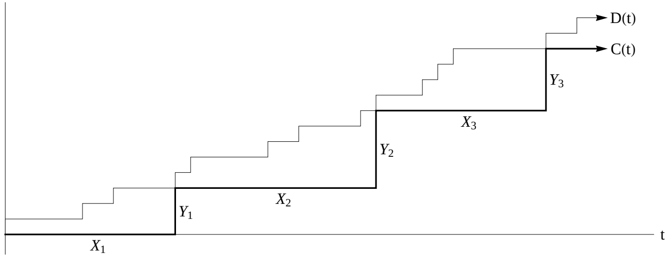
Note: In case $\lambda_i \equiv \lambda$, $q_i \equiv 1$:

$$\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\Big(1 - \pi_J\frac{P_i}{\pi_i}\Big)$$

Idea of Renewal Reward Derivation

"Embed" $D(t)$ in a Renewal-Reward Process, $C(t)$

1. $(X_n, Y_n) \equiv$ (busy cycle, number served) in cycle n
2. $N(t) = \inf\{n : \sum_{i=1}^{n} X_i > t\}$, $C(t) = \sum_{i=1}^{N(t)} Y_i$
3. Asymptotic variance rates of $C(t)$ and $D(t)$ are equal
4. Known:
   – Asymptotic variance rate of $C(t)$ is $\frac{1}{\mathbb{E}[X]}\mathrm{Var}\left(Y - \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}X\right)$
   – Systems of equations for 1'st, 2'nd and cross moments of X and Y



Single Server BRAVO
(older results)

M/M/1/K Queue

Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

Using $\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J\frac{P_i}{\pi_i}\right)$:

M/M/1/K Queue

Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

Using $\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J\frac{P_i}{\pi_i}\right)$:

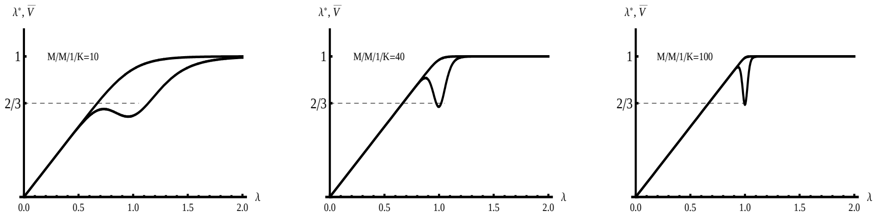
$$\mathcal{D} = \begin{cases} 1 + o_K(1), & \lambda \neq \mu, \\ \frac{2}{3} + o_K(1), & \lambda = \mu. \end{cases}$$

M/M/1/K Queue

Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

Using $\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J \frac{P_i}{\pi_i}\right)$:

$$\mathcal{D} = \begin{cases} 1 + o_K(1), & \lambda \neq \mu, \\ \frac{2}{3} + o_K(1), & \lambda = \mu. \end{cases}$$

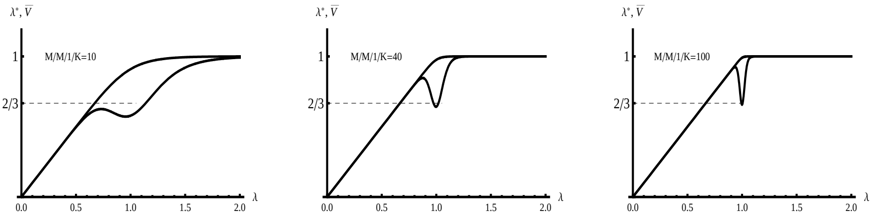


M/M/1/K Queue

Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

Using $\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J \frac{P_i}{\pi_i}\right)$:

$$\mathcal{D} = \begin{cases} 1 + o_K(1), & \lambda \neq \mu, \\ \frac{2}{3} + o_K(1), & \lambda = \mu. \end{cases}$$



We call this **BRAVO**:

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

When $K = \infty$, the formula for $\mathcal{D}$ does not hold. In this case,

$$\mathcal{D} = \begin{cases} 1, & \lambda \neq \mu, \\ ?, & \lambda = \mu. \end{cases}$$

A guess is $\frac{2}{3}$, since for $K < \infty$, $\mathcal{D} = \frac{2}{3} + o_K(1)$...

M/M/1 Queue

When $K = \infty$, the formula for $\mathcal{D}$ does not hold. In this case,

$$\mathcal{D} = \begin{cases} 1, & \lambda \neq \mu, \\ ?, & \lambda = \mu. \end{cases}$$

A guess is $\frac{2}{3}$, since for $K < \infty$, $\mathcal{D} = \frac{2}{3} + o_K(1)$...

Theorem:
Ahmad Al-Hanbali, Michel Mandjes, Y. N., Ward Whitt, 2011

For the M/M/1 queue with $\lambda = \mu$ and arbitrary initial conditions of $Q(0)$ (with finite second moments),

$$\mathcal{D} = 2\left(1 - \frac{2}{\pi}\right) \approx 0.727.$$

Proof based on analysis of classic Laplace transform of generating function of $D(\chi)$ where χ is an exponential random variable.

G/G/1 Queue

Moving away from the memory-less assumptions,

$$\mathcal{D} = \begin{cases} c_a^2, & \lambda < \mu, \\ ?, & \lambda = \mu, \\ c_s^2, & \lambda > \mu. \end{cases}$$

For M/M/1 it was $2(1 - \frac{2}{\pi})$...

G/G/1 Queue

Moving away from the memory-less assumptions,

$$\mathcal{D} = \begin{cases} c_a^2, & \lambda < \mu, \\ ?, & \lambda = \mu, \\ c_s^2, & \lambda > \mu. \end{cases}$$

For M/M/1 it was $2(1 - \frac{2}{\pi})$...

Theorem:
Ahmad Al-Hanbali, Michel Mandjes, Y.N., Ward Whitt, 2011

For the G/G/1 queue with $\lambda = \mu$, arbitrary finite second moment initial conditions $\big(Q(0), V(0), U(0)\big)$, and finite fourth moments of the inter-arrival and service times,

$$\mathcal{D} = (c_a^2 + c_s^2)\Big(1 - \frac{2}{\pi}\Big).$$

Proof based on diffusion limit of $(D(n\cdot) - \lambda n\cdot)/\sqrt{\lambda n\cdot}$ as $n \to \infty$ (Iglehart and Whitt 1971). Fourth moments are a technical condition used in establishing uniform integrability.

G/G/1/K Queue

$$\mathcal{D} = \begin{cases} c_a^2 + o_K(1), & \lambda < \mu, \\ ?, & \lambda = \mu, \\ c_s^2 + o_K(1), & \lambda < \mu. \end{cases}$$

For $M/M/1/K$ it was $\frac{2}{3} + o_K(1)$, for $G/G/1$ it was $(c_a^2 + c_s^2)(1 - \frac{2}{\pi})$...

G/G/1/K Queue

$$\mathcal{D} = \begin{cases} c_a^2 + o_K(1), & \lambda < \mu, \\ ?, & \lambda = \mu, \\ c_s^2 + o_K(1), & \lambda < \mu. \end{cases}$$

For $M/M/1/K$ it was $\frac{2}{3} + o_K(1)$, for $G/G/1$ it was $(c_a^2 + c_s^2)(1 - \frac{2}{\pi})$...

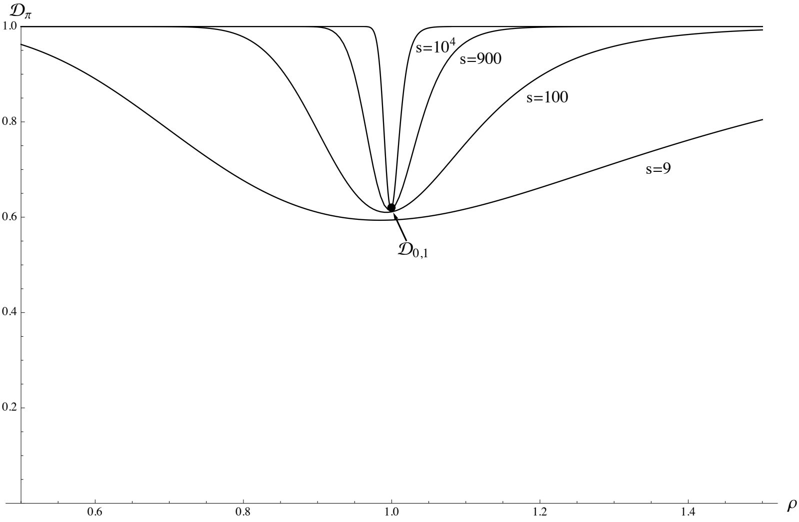
Conjecture (numerically tested), Y.N., 2011

For the G/G/1/K queue with $\lambda = \mu$ and arbitrary initial conditions and light-tailed service and inter-arrival times,

$$\mathcal{D} = (c_a^2 + c_s^2)\frac{1}{3} + O(\frac{1}{K}).$$

Numerical verification done by representing the system as PH/PH/1/K MAPs

Many Servers


$\mathcal{D}_\pi$
1.0
0.8
0.6
0.4
0.2
0.6
0.8
1.0
1.2
1.4
ρ
s=10^4
s=900
s=100
s=9
$\mathcal{D}_{0,1}$

Quality and Efficiency Driven (QED) Scaling Regime

A sequence of systems

Consider a sequence of $M/M/s/K$ queues with increasing $s = 1, 2, \ldots$ and with $\rho_s := \frac{\lambda}{s\mu}$ and K_s such that,

$$(1 - \rho_s)\sqrt{s} \to \beta \in (-\infty, \infty)$$
$$\frac{K_s}{\sqrt{s}} \to \eta \in (0, \infty)$$

So for large s:

$$\rho_s \approx 1 - \beta/\sqrt{s}$$
$$K_s \approx \eta\sqrt{s}$$

Halfin, Whitt, 1981, Garnett, Mandelbaum, Reiman 2002, Borst, Mandelbaum, Reiman, 2004, Whitt, 2004, Pang, Talreja, Whitt, 2007, Janssen, van Leeuwaarden, Zwart, 2011, Kaspi, Ramanan 2011, first session of this morning....

Favorable QED Properties

- Probability of delay converges to a value $\in (0, 1)$
- Mean waiting times are typically $O(s^{-1/2})$
- Large queue lengths almost never occur
- Quick mixing times
- In applications: Call-centers (etc...) describes behavior well and allows for asymptotic approximate optimization of staffing etc...
- How about BRAVO?

BRAVO for QED Queues

M/M/s/K QED BRAVO

Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2013

Consider QED scaling with $\beta \neq 0$:

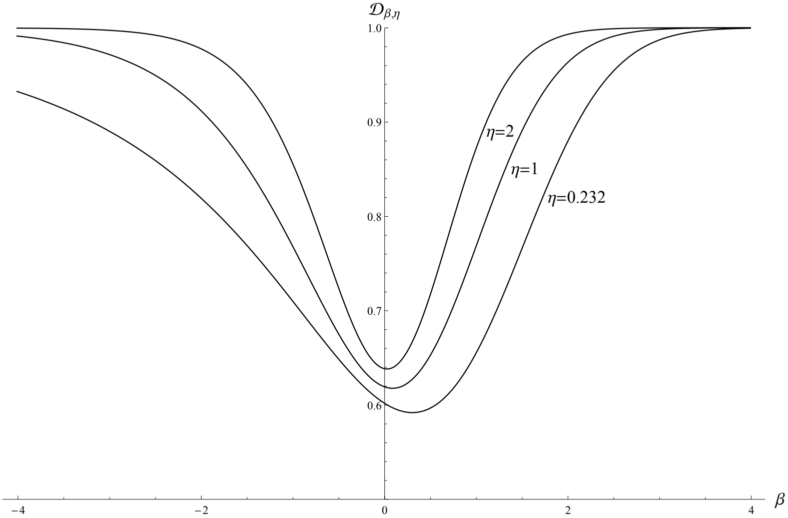
$$\mathcal{D}_{\beta,\eta} := \lim_{s,K\to\infty} \lim_{t\to\infty} \frac{Var\big(D(t)\big)}{\mathbb{E}\big(D(t)\big)},$$

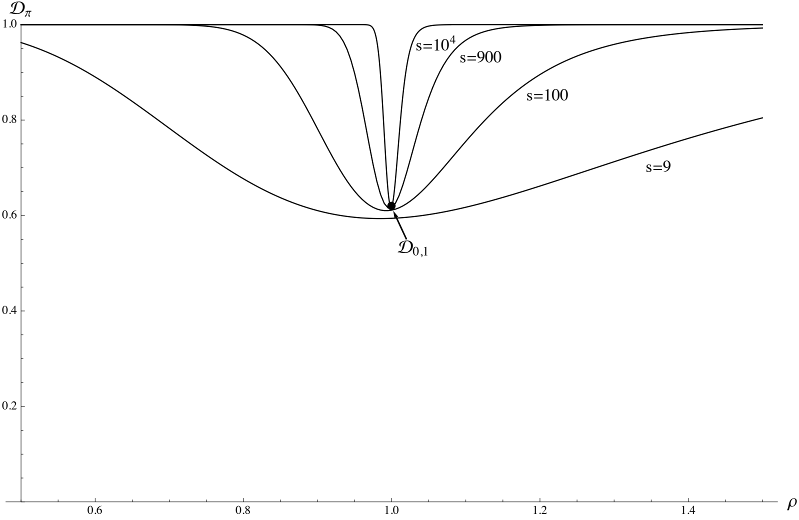
$$\mathcal{D}_{\beta,\eta} = 1 - \frac{2\beta^2 e^{-\beta\eta} h^2}{\phi(\beta)} \int_{-\beta}^{\infty} \left(1 - \beta e^{-\beta\eta} h \frac{\Phi(-u)}{\phi(u)}\right) \Phi(-u)\, du$$
$$+ 2e^{-\beta\eta} h(1 + e^{-\beta\eta} h)\Big(1 - \beta\eta - e^{-\beta\eta} + (1 - 2\beta\eta e^{-\beta\eta} - e^{-2\beta\eta})h\Big)$$

where

$$h = \lim_{s\to\infty} \frac{\mathbb{P}\big(Q_s \geq s\big)}{1 - e^{-\beta\eta}} = \frac{1}{1 - e^{-\beta\eta} + \frac{\beta\Phi(\beta)}{\phi(\beta)}}$$

BRAVO viewed through the QED lens





$\mathcal{D}_\pi$
1.0
0.8
0.6
0.4
0.2
s=10^4
s=900
s=100
s=9
$\mathcal{D}_{0,1}$
0.6
0.8
1.0
1.2
1.4
ρ

M/M/s/K QED BRAVO with $\rho \equiv 1$ ($\beta = 0$)

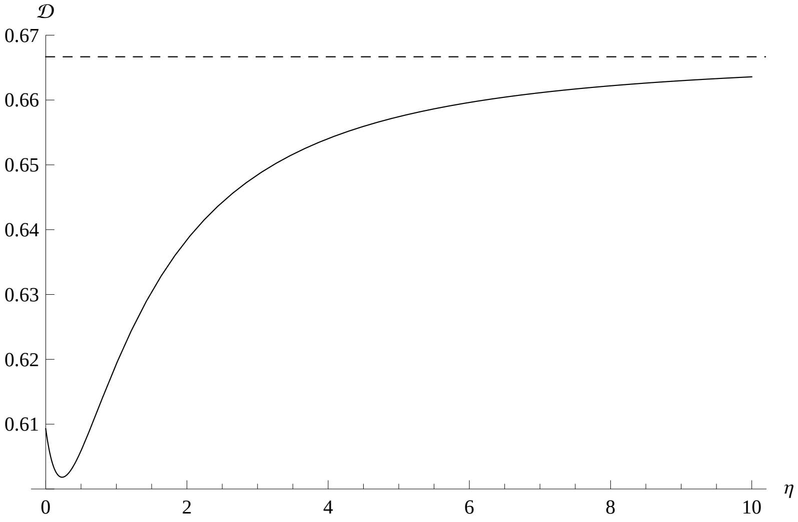
Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2013

$$\frac{K_s}{\sqrt{s}} \to \eta \in (0, \infty)$$

$$\mathcal{D}_{0,\eta} := \lim_{s,K\to\infty} \lim_{t\to\infty} \frac{Var\big(D(t)\big)}{\mathbb{E}\big(D(t)\big)},$$

$$\mathcal{D}_{0,\eta} = \frac{2}{3} - \frac{\left(6 - \frac{3\pi}{2}\right)\eta - \frac{1}{2}\pi\sqrt{\frac{\pi}{2}} + 3\sqrt{2\pi}(1 - \log 2)}{3\left(\eta + \sqrt{\frac{\pi}{2}}\right)^3}.$$

M/M/s/$\lfloor \eta\sqrt{s} \rfloor$ $s \to \infty$ at $\rho \equiv 1$ ($\beta = 0$)



Summary

Known BRAVO constants:

- Single server finite buffer: $2/3$
  (for G/G replace 2 by $c_a^2 + c_s^2$)
- Single server infinite buffer $2(1 - 2/\pi)$:
  (for G/G replace 2 by $c_a^2 + c_s^2$)
- Memoryless many servers finite buffer: $\mathcal{D}_{0,\eta} \in [0.6, 2/3]$

Not yet known:

- Memoryless many servers infinite buffer.
- Many servers without memoryless assumptions
- Systems with reneging or other customer loss mechanisms

Other questions: How can BRAVO be harnessed in practice?
Why does BRAVO occur?

References

- Daryl J. Daley, Johan van Leeuwaarden and Y.N., "*BRAVO for QED Finite Birth-Death Queues*", preprint.
- Daryl Daley, "*Revisiting queueing output processes: a point process viewpoint*", Queueing Systems, 68, pp. 395–405, 2011.
- Y.N., "*The variance of departure processes: puzzling behavior and open problems*", Queueing Systems, 68, pp. 385–394, 2011.
- Ahmad Al-Hanbali, Michel Mandjes, Y.N. and Ward Whitt, "*The asymptotic variance of departures in critically loaded queues*", Advances in Applied Probability, 43, pp. 243–263, 2011.
- Y.N. and Gideon Weiss, "*The asymptotic variance rate of the output process of finite capacity birth-death queues*", Queueing Systems, 59, pp. 135–156, 2008.