BRAVO for QED Queues

Yoni Nazarathy, The University of Queensland,

Joint work with

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

Technion Operations Research and Statistics Seminar, December 17, 2012.

1

- Queueing Systems and the Variance of Output Processes
- Previous BRAVO Results
- The Halfin-Whitt (or QED) Regime
- BRAVO for QED Queues

Queueing Systems and the Variance of Output Processes

G/G/s/K + G Queueing Systems

Jobs arrive randomly to a server (first 'G'), enter if Q(t) < s + K, otherwise are lost. Upon entering, the jobs queue up with impatience (+G) and may renege, are served by one of s servers and require service for a random durations (second 'G'). After service, jobs depart.

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

G/G/s/K + G Queueing Systems

Jobs arrive randomly to a server (first 'G'), enter if Q(t) < s + K, otherwise are lost. Upon entering, the jobs queue up with impatience (+G) and may renege, are served by one of s servers and require service for a random durations (second 'G'). After service, jobs depart.

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

Processes with specified laws (inputs)

- A(t) arrival counting process
- The random service durations
- The random patience durations
- Initial conditions

G/G/s/K + G Queueing Systems

Jobs arrive randomly to a server (first 'G'), enter if Q(t) < s + K, otherwise are lost. Upon entering, the jobs queue up with impatience (+G) and may renege, are served by one of s servers and require service for a random durations (second 'G'). After service, jobs depart.

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

Processes with specified laws (inputs)

- A(t) arrival counting process
- The random service durations
- The random patience durations
- Initial conditions

Resulting counting processes

- L(t) jobs arriving to a full system
- R(t) reneging jobs (leaving due to impatience)
- D(t) completed jobs

Analysis of Output Processes: D(t)

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

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

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

Analysis of Output Processes: D(t)

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

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.

Analysis of Output Processes: D(t)

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

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 $\{D(t), t \ge 0\}$
- $\mathbb{E}[D(t)]$, Var(D(t))

•
$$\lambda^* := \lim_{t \to \infty} \frac{\mathbb{E}[D(t)]}{t}, \quad \overline{V} := \lim_{t \to \infty} \frac{\operatorname{Var}(D(t))}{t}, \quad R := \frac{\overline{V}}{\lambda^*}$$

• Sometimes: asymptotic normality $D(t) \sim \mathcal{N} \Big(\lambda^* t, \ \overline{V} t \Big)$

The M Case

If the arrival process, $\{A(t), t \ge 0\}$ is Poisson and the service and impatience times are i.i.d. exponential sequences, we have the M/M/s/K + M model. (Termed 'Erlang-A' when $K = \infty$).

In this case Q(t) is a birth-death continuous time Markov Chain on the states $\{0, 1, 2, ..., s + K\}$, with birth and death rates,

$$\lambda_i = \lambda \mathbf{1}_{\{i < s + K\}}, \qquad \mu_i = \mu(i \wedge s) + \gamma(i - s)^+.$$

The constants $\lambda > 0$, $\mu > 0$ and $\gamma \ge 0$ correspond to the three *M*'s.

The M Case

If the arrival process, $\{A(t), t \ge 0\}$ is Poisson and the service and impatience times are i.i.d. exponential sequences, we have the M/M/s/K + M model. (Termed 'Erlang-A' when $K = \infty$).

In this case Q(t) is a birth-death continuous time Markov Chain on the states $\{0, 1, 2, ..., s + K\}$, with birth and death rates,

$$\lambda_i = \lambda \mathbf{1}_{\{i < \mathbf{s} + \mathbf{K}\}}, \qquad \mu_i = \mu(i \wedge \mathbf{s}) + \gamma(i - \mathbf{s})^+.$$

The constants $\lambda > 0$, $\mu > 0$ and $\gamma \ge 0$ correspond to the three *M*'s.

Resulting Counting Processes

- The overflow process, L(t) is a renewal process.
- The reneging and output processes, R(t), D(t) are non-renewal (unless s = 0 and K = 1)
- All these processes are Markovian Arrival Processes (MAPs)

Asymptotic Variance in the M Case

$$\overline{V} := \lim_{t \to \infty} rac{\operatorname{Var}(D(t))}{t}, \qquad R := rac{\overline{V}}{\lambda^*}.$$

Theorem: Y.N., Weiss, 2008

Consider an irreducible birth-death process on the finite state space $\{0, 1, \ldots, J\}$ with constant birth rates λ , death rates, μ_1, \ldots, μ_J and arbitrary initial distribution. Let D(t) count downward transitions (deaths) during [0, t]. Then,

$$R = 1 - 2 \frac{\pi_J}{1 - \pi_J} \sum_{i=0}^{J} P_i \left(1 - \pi_J \frac{P_i}{\pi_i} \right),$$

with,

$$\pi_{i} = \frac{\lambda^{i} \prod_{j=1}^{i} \mu_{j}^{-1}}{\sum_{j=0}^{J} \lambda^{j} \prod_{k=1}^{j} \mu_{k}^{-1}}, \qquad P_{i} := \sum_{j=0}^{i} \pi_{j}.$$

2008 derivation based on MAPs and relations to MMPPs for which there is an explicit formula (for the birth-death case). Newer derivation based on more elementary renewal-reward approach.

Previous BRAVO Results

This is the case, $s = 1, K < \infty, \gamma = 0$. π_i evaluates to a geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$.

Using
$$R = 1 - 2 rac{\pi_J}{1 - \pi_J} \sum_{i=0}^{J} P_i \Big(1 - \pi_J rac{P_i}{\pi_i} \Big)$$
:

This is the case, $s = 1, K < \infty, \gamma = 0$. π_i evaluates to a geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$.

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

$$R = \left\{ egin{array}{cc} 1+o_{K}(1), & \lambda
eq \mu, \ rac{2}{3}+o_{K}(1), & \lambda=\mu. \end{array}
ight.$$

This is the case, $s = 1, K < \infty, \gamma = 0$. π_i evaluates to a geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$.

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

$$R = \left\{egin{array}{cc} 1+o_{\mathcal{K}}(1), & \lambda
eq \mu, \ rac{2}{3}+o_{\mathcal{K}}(1), & \lambda=\mu. \end{array}
ight.$$



This is the case, $s = 1, K < \infty, \gamma = 0$. π_i evaluates to a geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$.

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

$$R = \left\{egin{array}{cc} 1+o_K(1), & \lambda
eq\mu,\ rac{2}{3}+o_K(1), & \lambda=\mu. \end{array}
ight.$$



We call this BRAVO:

Balancing Reduces Asymptotic Variance of Outputs

When $K = \infty$, the formula for R does not hold. In this case,

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

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

When $K = \infty$, the formula for R does not hold. In this case,

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

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

Theorem: Ahmad Al-Hanbali, Michel Mandjes, **Y.N.**, Ward Whitt, 2011 For the M/M/1 queue with $\lambda = \mu$ and arbitrary initial conditions

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

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

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

M/G/1 Queue - Relating BRAVO to Y-Intercept

Theorem: Yoav Kerner and Y.N., 2009

For M/G/1 with $\lambda < \mu,$ starting empty and third service moment finite:

$$Var(D(t)) = \lambda t - (1 - L_0) \frac{\rho}{(1 - \rho)^2} + o_t(1).$$

• L_0 expression of ρ and first 3 service moments

• For
$$M/M/1$$
, $L_0 = 0$

Example: Look at the curves Var(D(t))

- M/M/1 starting empty, $\mu =$ 1, $\rho =$ 0.991, 0.993, 0.995, 0.997
- Linear asymptote slope \approx 1, y-intercepts $\approx -10^4, -2\times 10^4, -4\times 10^4, -10^5$
- Simulate D(t): 3×10^4 repetitions, 10^5 time units, sample variance every 1000









G/G/1 Queue

Moving away from the memory-less assumptions,

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

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

G/G/1 Queue

Moving away from the memory-less assumptions,

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

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 (Q(0), V(0), U(0)), and finite fourth moments of the inter-arrival and service times,

$$R = (c_a^2 + c_s^2) \left(1 - \frac{2}{\pi}\right).$$

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

G/G/1/K Queue

Breaking the exponentially assumption,

$$R = \left\{ egin{array}{ll} c_a^2 + o_{\mathcal{K}}(1), & \lambda < \mu, \ ?, & \lambda = \mu, \ c_s^2 + o_{\mathcal{K}}(1), & \lambda < \mu. \end{array}
ight.$$

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

Breaking the exponentially assumption,

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

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,

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

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

Multi-servers and Reneging (s>1 and/or $\gamma>0$)

M/M/c/K (different notation in this slide, $\mu = 1$)



G/G/s Queue

Theorem:

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

For the G/G/s queue with $\lambda = \mu$, arbitrary finite second moment initial conditions, finite second moments of the inter-arrival and service times and technical uniform integrability assumptions,

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

As with the ${\sf G}/{\sf G}/1$ case, the limiting process is,

$$\inf_{s\in[0,t]} \{c_a^2 B_1(s) + c_s^2 B_2(t-s)\},\$$

the proof is the same, yet uniform integrability only established in special cases.

Handling Reneging

Theorem*: Daryl Daley, Johan van Leeuwaarden, Y.N. 2012

Consider an irreducible birth-death process on the finite state space $\{0, 1, \ldots, J\}$ with constant birth rates λ , death rates, μ_1, \ldots, μ_J and arbitrary initial distribution. Let D(t) count downward transitions (deaths) during [0, t], each "filtered" independently with state-dependent probabilities, p_1, \ldots, p_J . Then,

$$R = 1 - 2\frac{1}{\lambda^*}\pi'_+(Z - \lambda^*I)W^{-1}(\mathbf{p} \bullet \boldsymbol{\mu} - \lambda^*\mathbf{1}).$$

with **p** and μ vectors of p_i and μ_i reps. and,



* Cleaner expression (based on explicit inverse of W) in progress...

Idea of Renewal Reward Derivation

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

- $(X_n, Y_n) \equiv$ (busy cycle, number served) in cycle *n*
- **2** $N(t) = \inf\{n : \sum_{i=1}^{n} X_i > t\}, \ C(t) = \sum_{i=1}^{N(t)} Y_i$
- Solution Can show variance rates of C(t) and D(t) are equal

4 Known:

- Variance rate of C(t) is $\frac{1}{E[X]}$ Var $\left(Y - \frac{E[Y]}{E[X]}X\right)$ - Systems of equations for 1'st, 2'nd and cross moments of X and Y





$$K = 100, \quad \mu = 1, \quad \gamma \in \{0, 0.02, 0.1, 0.5\}.$$

λ

The Halfin-Whitt (or QED) Regime

Quality and Efficiency Driven (QED) Scaling Regime

A sequence of systems: $QED(\alpha, \beta)$ scaling

Consider M/M/s/K + M queues with $\rho_s := \frac{\lambda}{s\mu}$ such that,

$$(1-\rho_s)\sqrt{s} \to \beta \in (-\infty,\infty).$$

If $K < \infty$,

$$\frac{K}{\sqrt{s}} \to \eta :=: \frac{\sqrt{\pi/2}}{\alpha} \in (0,\infty).$$

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 ...

The key QED Property (of 'M' systems)

 $\{(Q(t) - s)/\sqrt{s}, t \ge 0\}$ converges weakly to $\{X(t) \land \eta, t \ge 0\}$, a diffusion process with infinitesimal mean $m(x) = -\beta\mu - \mu x$ for x < 0 and $m(x) = -\beta - (\gamma/\mu)x$ for x > 0 and infinitesimal variance $\sigma^2(x) = 2$.

Analyzing systems through the QED Lens:

- \bullet 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 (convergence to stationary distribution)
- 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 Asymptotic Variance under QED

Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2012

$$R \xrightarrow{s,K\to\infty} 1 - a_{\alpha,\beta}b_{\alpha,\beta}c_{\alpha,\beta} - d_{\alpha,\beta},$$

where,

$$\begin{split} \mathbf{a}_{\alpha,\beta} &:= e^{-\frac{\beta}{\alpha}\sqrt{\pi/2}}, \\ b_{\alpha,\beta} &:= \left(e^{\frac{1}{2}\beta^2} \Phi(\beta)\sqrt{2\pi} + \frac{1-\mathbf{a}_{\alpha,\beta}}{\beta}\right)^{-1}, \\ c_{\alpha,\beta} &:= \int_{-\infty}^{\beta} \Phi(u) \left(1 - \mathbf{a}_{\alpha,\beta}\phi(\beta)\frac{\Phi(u)}{\phi(u)}\right) du, \\ d_{\alpha,\beta} &:= 2\mathbf{a}_{\alpha,\beta}b_{\alpha,\beta} \left(1 + \frac{\mathbf{a}_{\alpha,\beta}b_{\alpha,\beta}}{\beta}\right) \left(\left(\frac{\sqrt{2/\pi}}{\alpha} - \frac{1-\mathbf{a}_{\alpha,\beta}}{\beta}\right)\left(1 + \frac{\mathbf{a}_{\alpha,\beta}b_{\alpha,\beta}}{\beta}\right) + \left(\frac{\sqrt{2/\pi}}{\alpha} - \frac{1-\mathbf{a}_{\alpha,\beta}}{\beta\mathbf{a}_{\alpha,\beta}}\right)\left(\frac{\mathbf{a}_{\alpha,\beta}b_{\alpha,\beta}}{\beta}\right)\right), \end{split}$$

with,

$$\phi(u):=rac{e^{-u^2/2}}{\sqrt{2\pi}},\qquad,\Phi(u):=\int_{-\infty}^u\phi(t)dt.$$

M/M/s/K Asymptotic Variance under QED with $ho \equiv 1$

Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2012

For M/M/s/K with $\rho_s \equiv 1$ and,

$$\frac{K}{\sqrt{s}} \to \frac{\sqrt{\pi/2}}{\alpha} \in (0,\infty).$$

$$\xrightarrow{s,K\to\infty} 1 - \frac{1+3\alpha}{\alpha} - \frac{\alpha}{\alpha} \frac{2-\log(1-\alpha)}{\alpha}$$

$$R \xrightarrow{s, K \to \infty} 1 - \frac{1 + 3\alpha}{3(1 + \alpha)^3} - \frac{\alpha}{1 + \alpha} \frac{2 - \log 2}{\pi}$$

Proof based on expansions of

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

BRAVO for QED M/M/s/K with $\rho = 1$



Wrap-Up

• The *R*-formula applied to QED–M/M/s/K yields clean asymptotics

- The *R*-formula applied to QED–M/M/s/K yields clean asymptotics
- The "extended" formula for systems with reneging still does not...?

- The *R*-formula applied to QED–M/M/s/K yields clean asymptotics
- The "extended" formula for systems with reneging still does not...?
- \bullet When analyzing actual systems, balancing α and β in the approximation is delicate

- The *R*-formula applied to QED–M/M/s/K yields clean asymptotics
- The "extended" formula for systems with reneging still does not...?
- \bullet When analyzing actual systems, balancing α and β in the approximation is delicate
- Limiting processes of D(t) under QED scaling may be fruitful for G/G cases...

References on Variance of Outputs and BRAVO

- Daryl Daley, Johan van Leeuwaarden and **Y.N.**, "*BRAVO for QED Finite Birth-Death Queues*", working paper.
- Y.N., "Diffusion Parameters of Flows in Stable Queueing Networks", working paper.
- Yoav Kerner and **Y.N.**, "On The Linear Asymptote of the M/G/1 Output Variance Curve", working paper.
- 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.

Extras

PH/PH/1/K BRAVO (K = 40)



M/M/1/K – Correlation Between $D(\cdot)$ and $L(\cdot)$



















