BRAVO for QED Queues

(and more stuff about variance of output counts)

Yoni Nazarathy

The University of Queensland

Rigorous Systems Group Seminar

Caltech

July 2012

Daryl J. Daley

Johan van Leeuwaarden

Ahmad Al-Hanbali

Michel Mandjes

Ward Whitt

Sophie Hautphenne

Yoav Kerner

Peter Taylor

Werner Scheinhardt

Gideon Weiss

Minimal Background: Queueing Models in this Talk

- Birth-Death queues:
  M/M/1, M/M/1/K, M/M/s/K, M/M/s/K+M ...
- General service times: G/G/1, G/G/1/K ...
- Open stable Jackson networks
- Open stable generalized Jackson networks

A basic packet conservation equation

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

Stochastic Counting Processes and their Variance

- Poisson processes:

$$\mathbb{E}[N(t)] = \mathrm{Var}\big(N(t)\big) = \lambda t$$

- Renewal processes:

$$\mathbb{E}[N(t)] \sim \lambda t \qquad \mathrm{Var}\big(N(t)\big) \sim \lambda c^2 t$$

- Counting processes resulting from queues?

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

The Output Process $D(\cdot)$

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

The Output Process $D(\cdot)$

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

Why analysis of the output counting process $\{D(t), t \geq 0\}$?

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

The Output Process $D(\cdot)$

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

Why analysis of the output counting process $\{D(t), t \geq 0\}$?

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

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

The Output Process $D(\cdot)$

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

Why analysis of the **output** counting process $\{D(t), t \geq 0\}$?

- 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

Talk 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)
- Something else: Asymptotic variance in queueing networks

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 (processed packets) 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 $q_i \equiv 1$

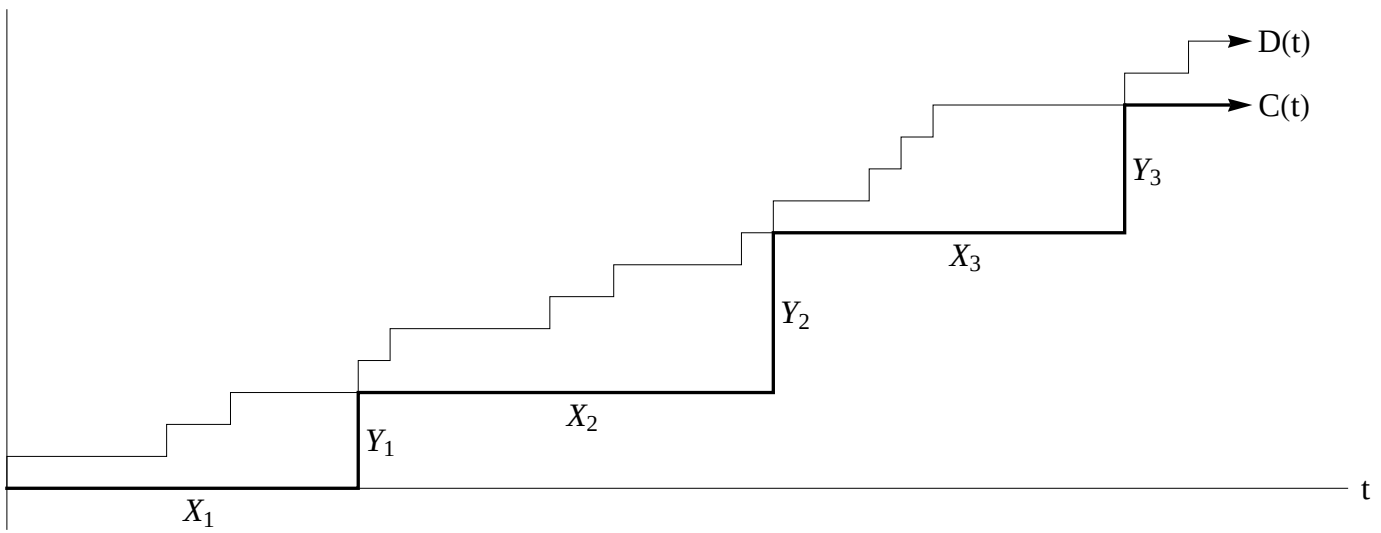
Note: In case $\lambda_i \equiv \lambda$, $q_i \equiv 1$:

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

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. $N(t) = \sup\{n : \sum_{i=1}^{n} X_i \leq t\}$, $C(t) = \sum_{i=1}^{N(t)} Y_i$
3. Asymptotic variance rates of $C(t)$ and $D(t)$ are equal
4. Known:
   – Asymptotic variance rate of $C(t)$ is $\frac{1}{\mathbb{E}[X]}\text{Var}\left(Y - \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}X\right)$
   – 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 $\mathcal{D} = 1 - 2\frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J\frac{P_i}{\pi_i}\right)$:

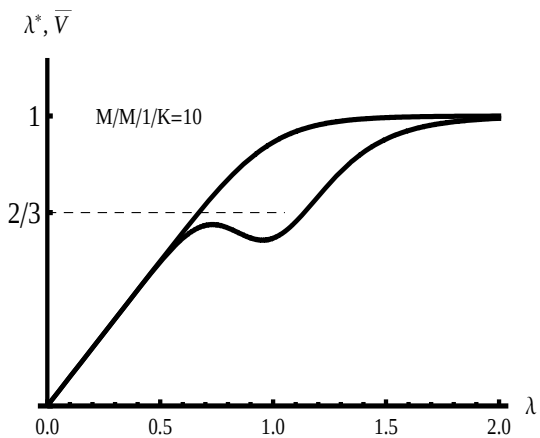
$$\mathcal{D} = \begin{cases} 1 + o_K(1), & \lambda \neq \mu, \\ \frac{2}{3} + o_K(1), & \lambda = \mu. \end{cases}$$

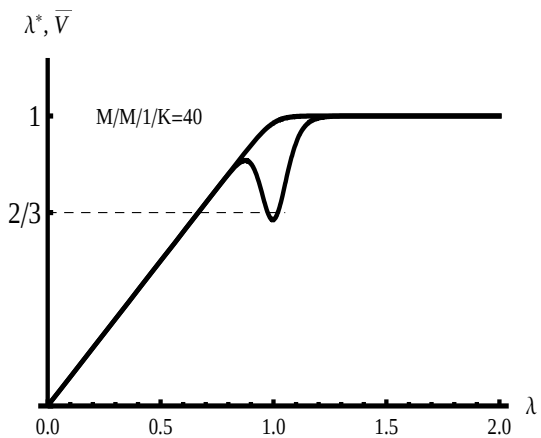
M/M/1/K Queue

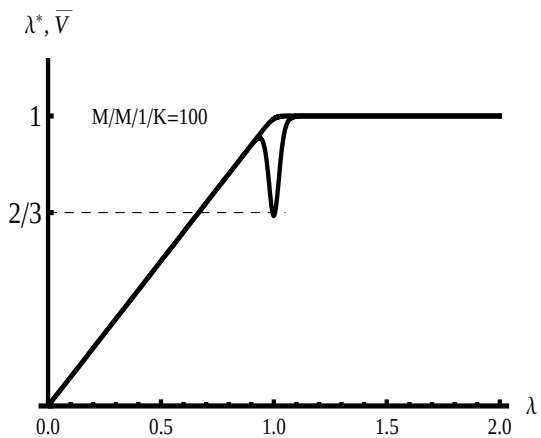
Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

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

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





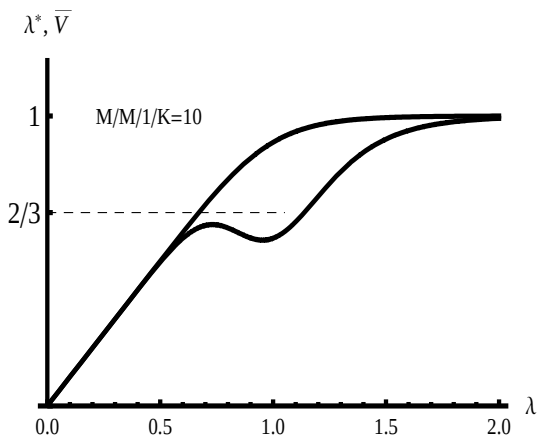


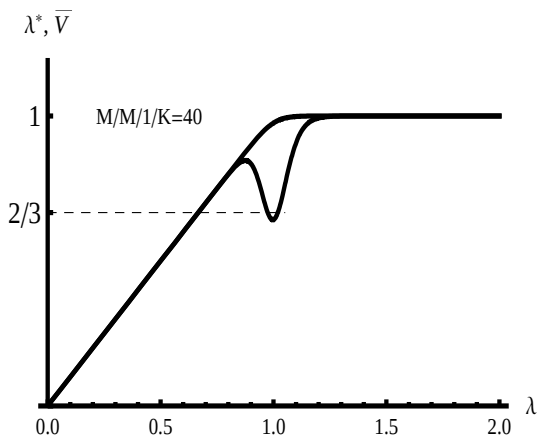
M/M/1/K Queue

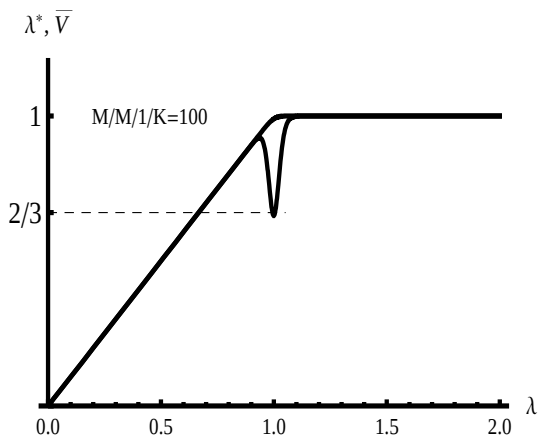
Here π_i is truncated geometric distribution when $\lambda \neq \mu$ and a uniform distribution when $\lambda = \mu$

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

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

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

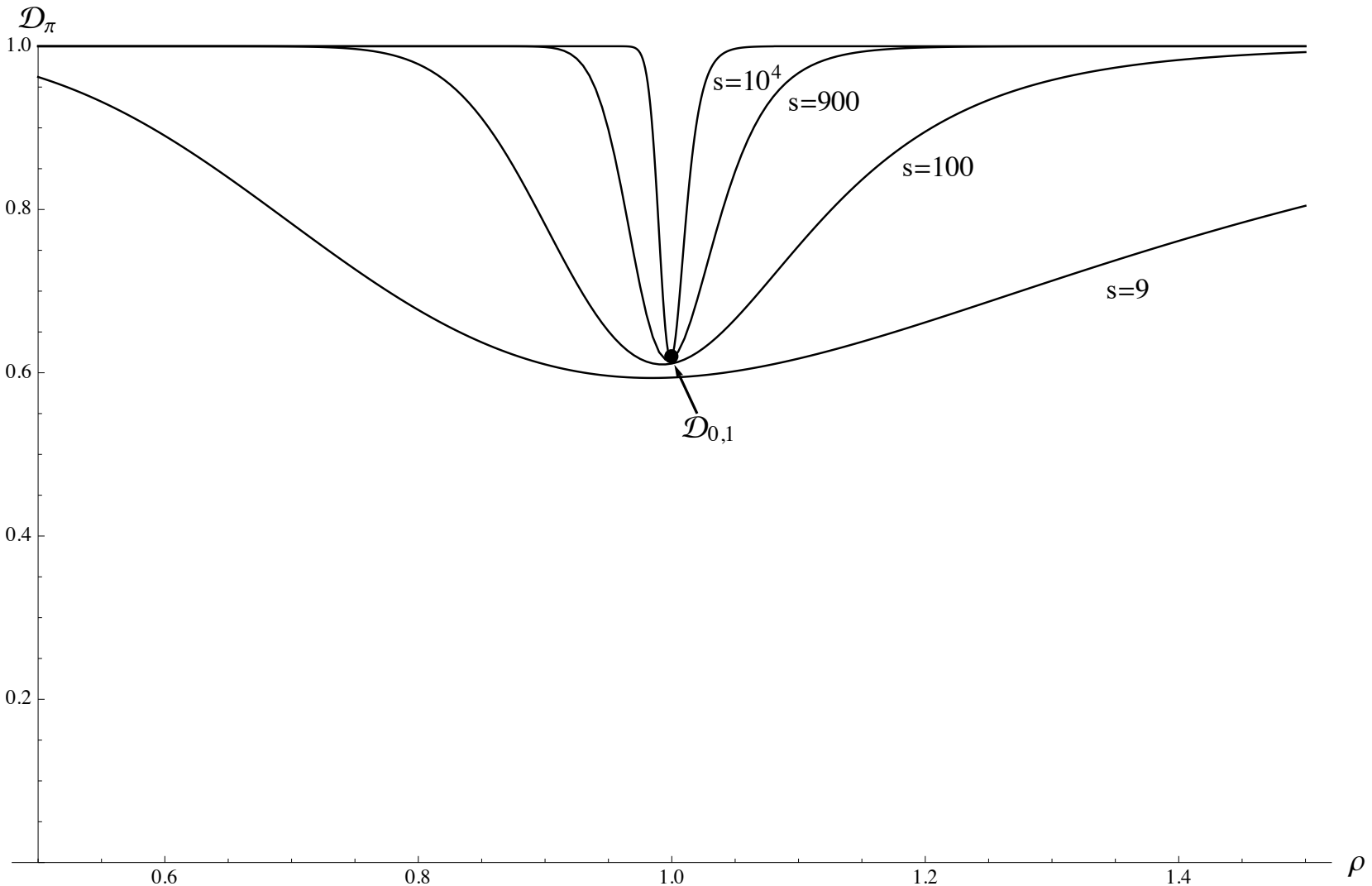
For the G/G/1/K queue with $\lambda = \mu$ and arbitrary initial conditions and light-tailed service and inter-arrival times,

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

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

Many Servers

M/M/s/$\lfloor\sqrt{s}\rfloor$



Quality and Efficiency Driven (QED) Scaling Regime

A sequence of systems

Consider a sequence of $M/M/s/K$ queues with increasing $s = 1, 2, ...$ 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...

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 $\beta \neq 0$:

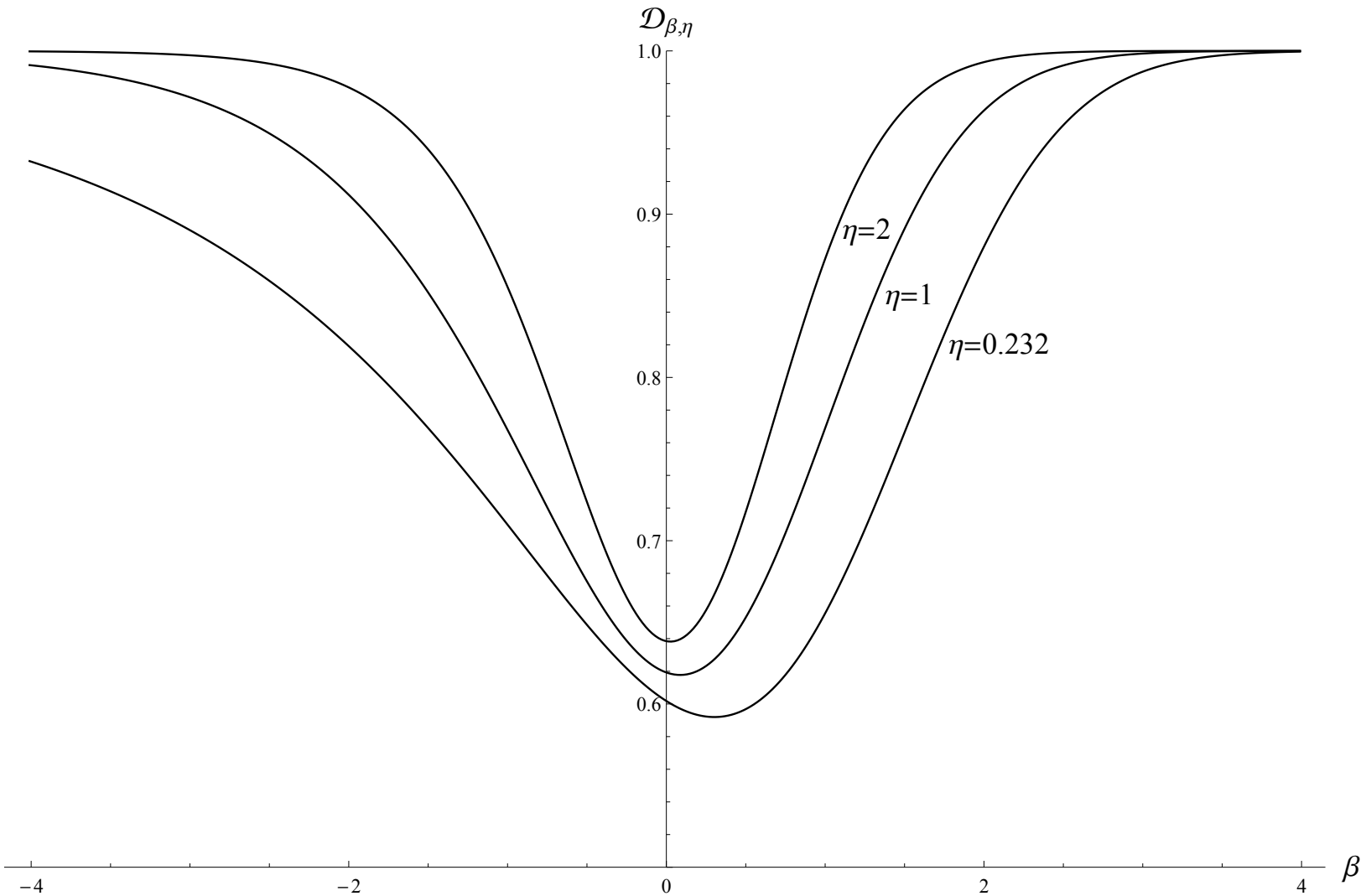
$$\mathcal{D}_{\beta,\eta} := \lim_{s,K\to\infty} \lim_{t\to\infty} \frac{Var\big(D(t)\big)}{\mathbb{E}\big(D(t)\big)},$$

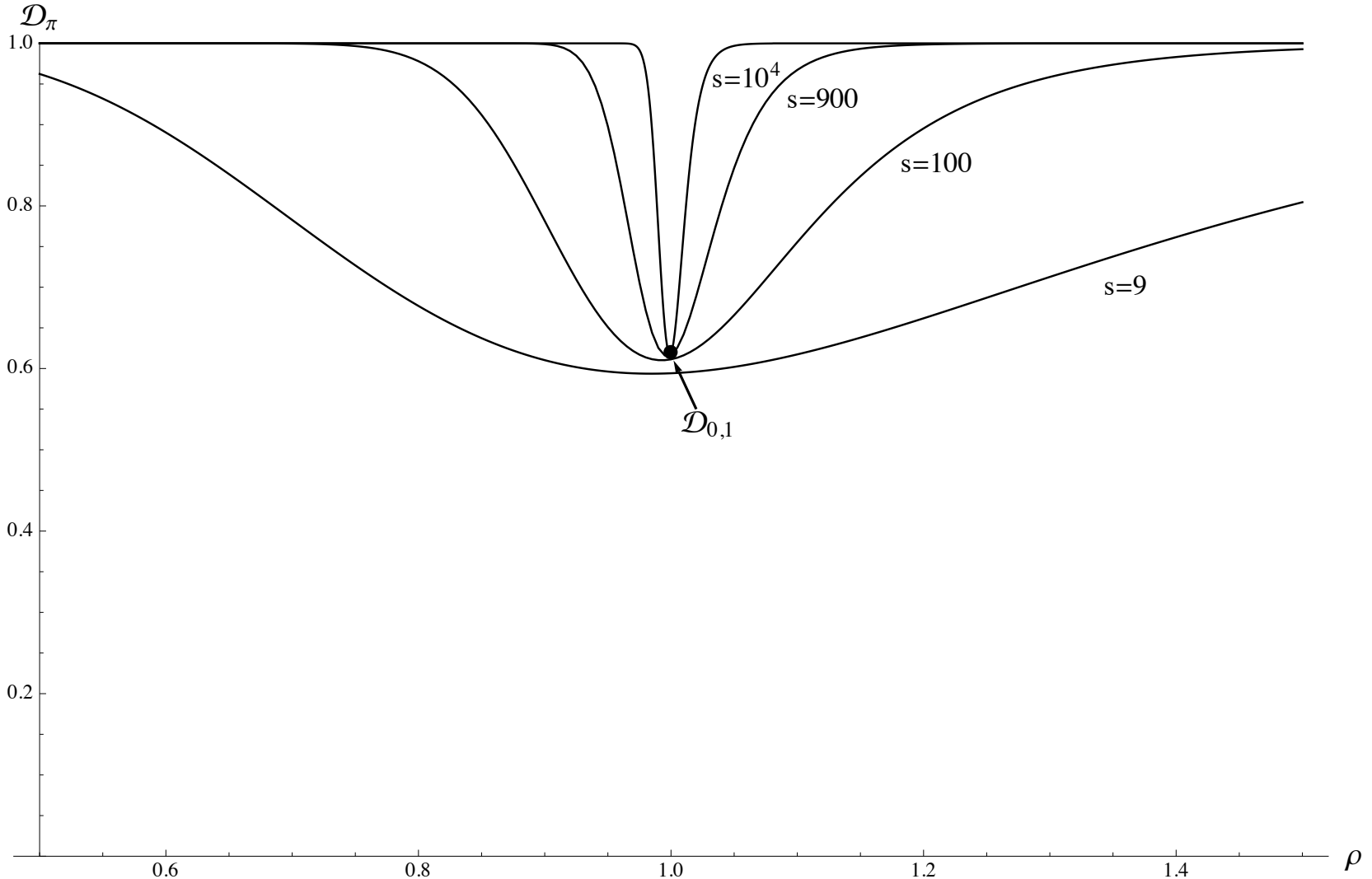
$$\mathcal{D}_{\beta,\eta} = 1 - \frac{2\beta^2 e^{-\beta\eta} h^2}{\phi(\beta)} \int_{-\beta}^{\infty} \Big(1 - \beta e^{-\beta\eta} h \frac{\Phi(-u)}{\phi(u)}\Big)\, \Phi(-u)\, du$$
$$+ 2e^{-\beta\eta} h(1 + e^{-\beta\eta} h)\Big(1 - \beta\eta - e^{-\beta\eta} + (1 - 2\beta\eta e^{-\beta\eta} - e^{-2\beta\eta})h\Big)$$

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

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

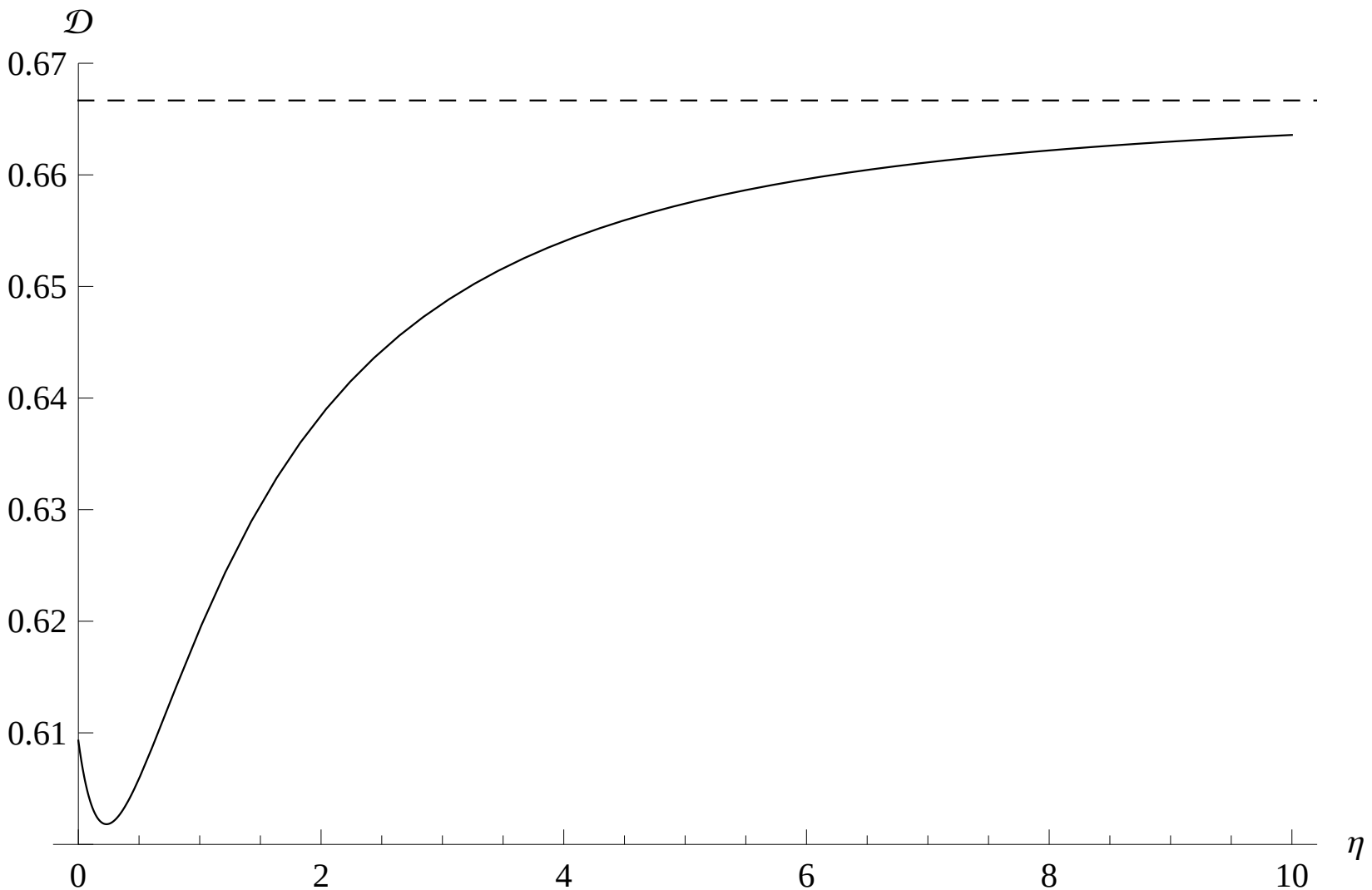
Theorem: Daryl Daley, Johan van Leeuwaarden, Y.N. 2013

Assume $\rho \equiv 1$ and $\frac{K_s}{\sqrt{s}} \to \eta \in (0, \infty)$. Then

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

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

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



Idea of BRAVO QED Proofs

Use

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

Using QED scaling:

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

"simply evaluate" the limit,

$$\lim_{s,K\to\infty} \frac{\pi_J}{1-\pi_J}\sum_{i=0}^{J} P_i\left(1 - \pi_J\frac{P_i}{\pi_i}\right).$$

Intermediate Summary: BRAVO

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 packet loss mechanisms

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

Further properties of $\mathrm{Var}\big(D(t)\big)$

The Stable M/G/1 Queue

Theorem: Sophie Hautphenne, Yoav Kerner, Y. N., Peter Taylor, 2013

Consider the stable M/G/1 queue with finite third service moment, parameterized by (*arrival rate, load, scv, skewness*)$= (\lambda, \rho, c^2, \gamma)$.

Stationary version:

$$\mathrm{Var}\big(D(t)\big) = \lambda t + L_e \frac{\rho}{(1-\rho)^2} + o(1),$$

$$L_e = \frac{(3c^4 - 4\gamma c^3 + 6c^2 - 1)\rho^3 + (4\gamma c^3 - 12c^2 + 4)\rho^2 + (6c^2 - 6)\rho}{6}.$$

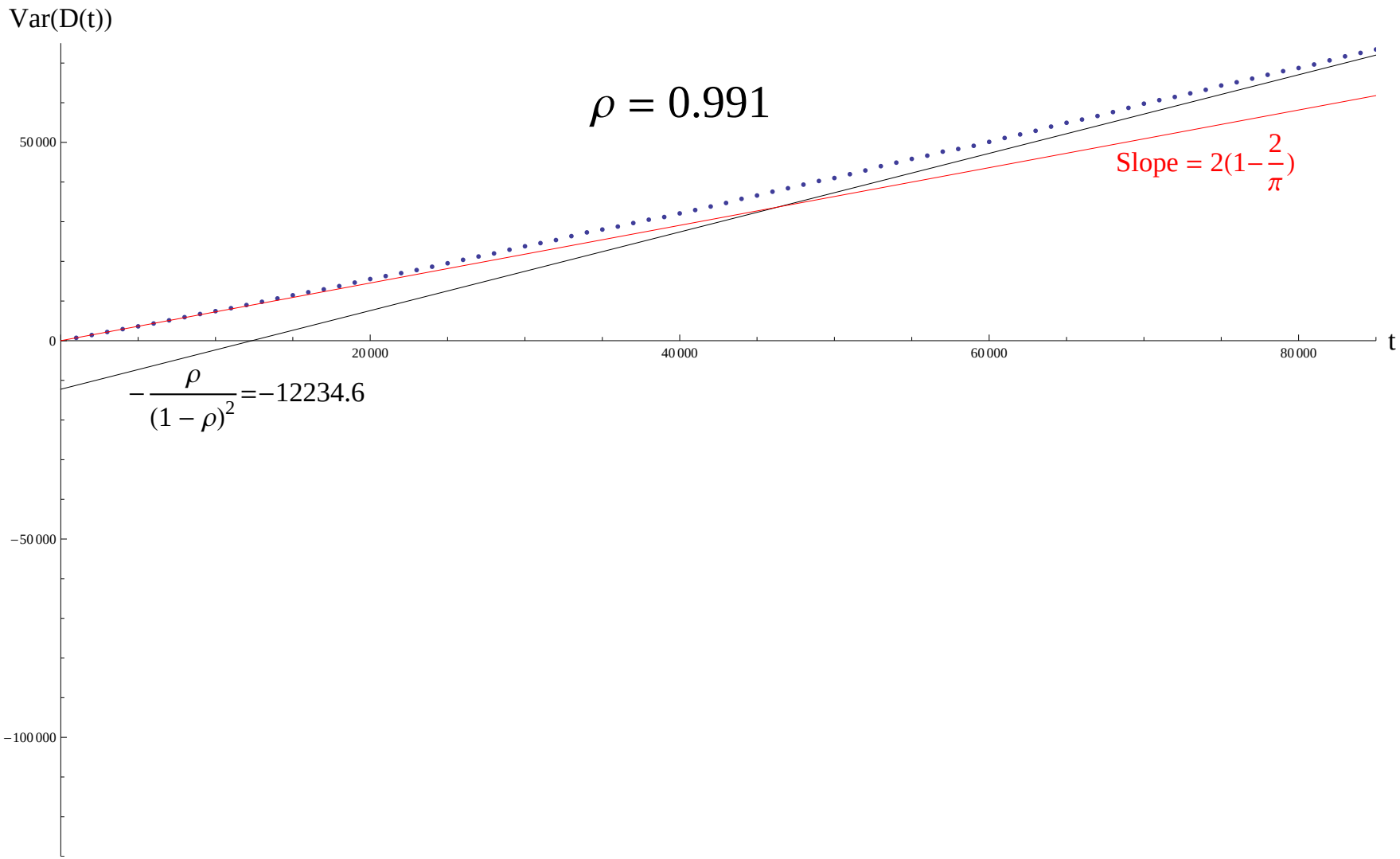
Starting empty version:

$$\mathrm{Var}\big(D(t)\big) = \lambda t - (1 - L_0) \frac{\rho}{(1-\rho)^2} + o(1),$$

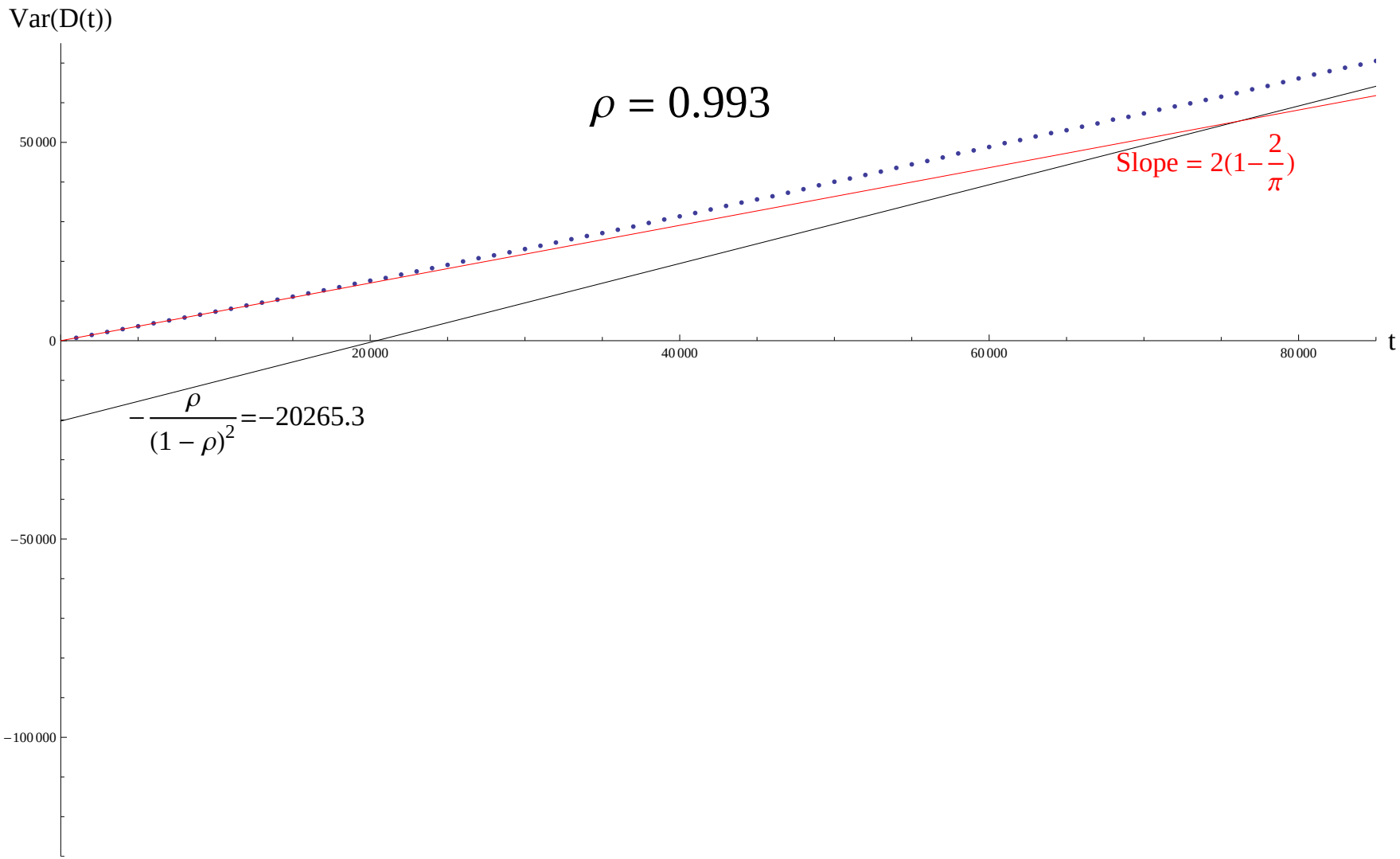
$$L_0 = \frac{(3c^4 - 4\gamma c^3 + 6c^2 - 1)\rho^3 + (4\gamma c^3 - 6c^2 - 2)\rho^2 - (6c^2 - 6)\rho}{12}.$$

M/M/1: $c^2 = 1, \gamma = 2$. $L_e = 0$, $L_0 = 0$.

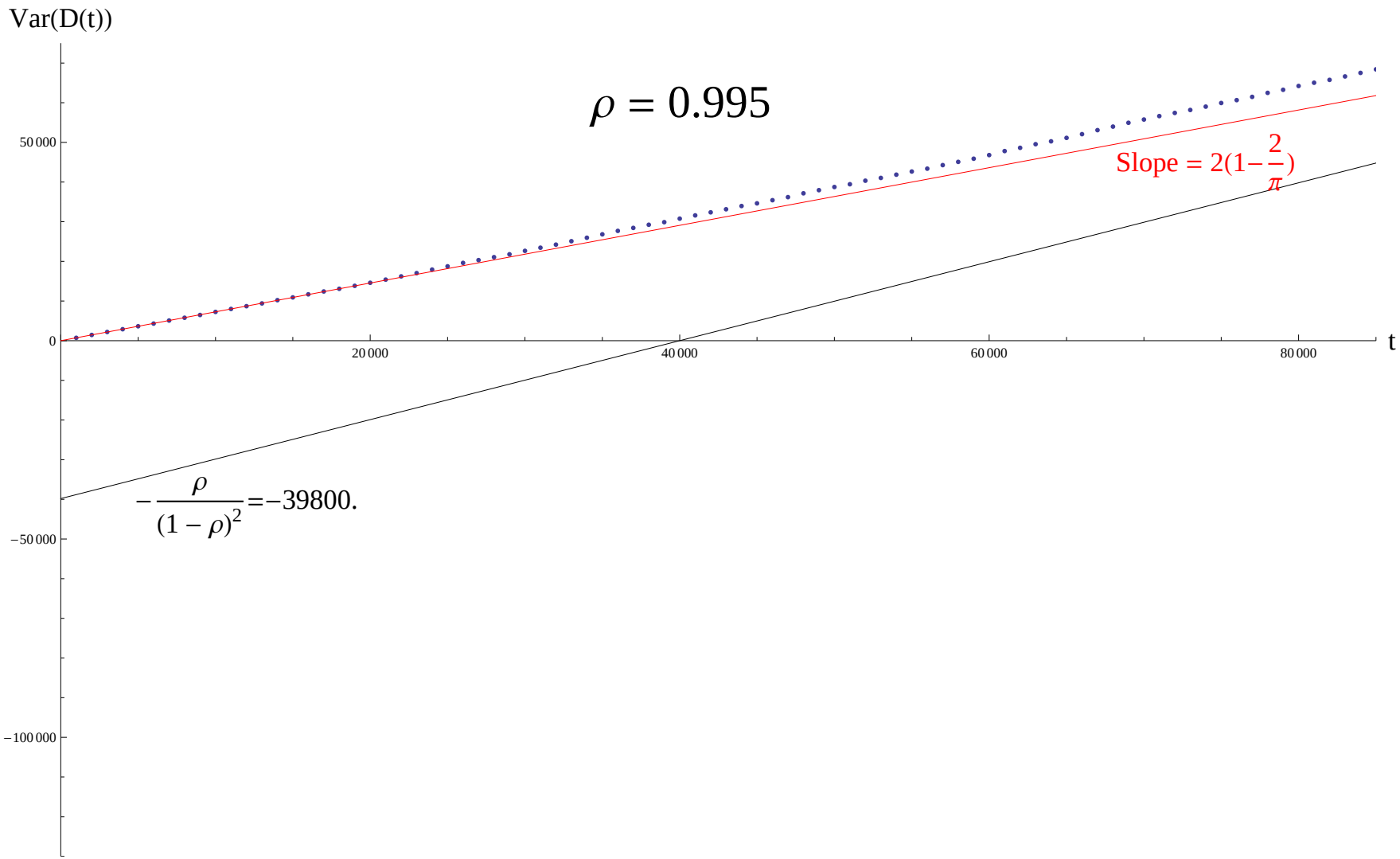
M/M/1 Queue


Var(D(t))
$\rho = 0.991$
Slope = $2(1-\frac{2}{\pi})$
50 000
0
−50 000
−100 000
20 000
40 000
60 000
80 000
t
$-\frac{\rho}{(1-\rho)^2} = -12234.6$

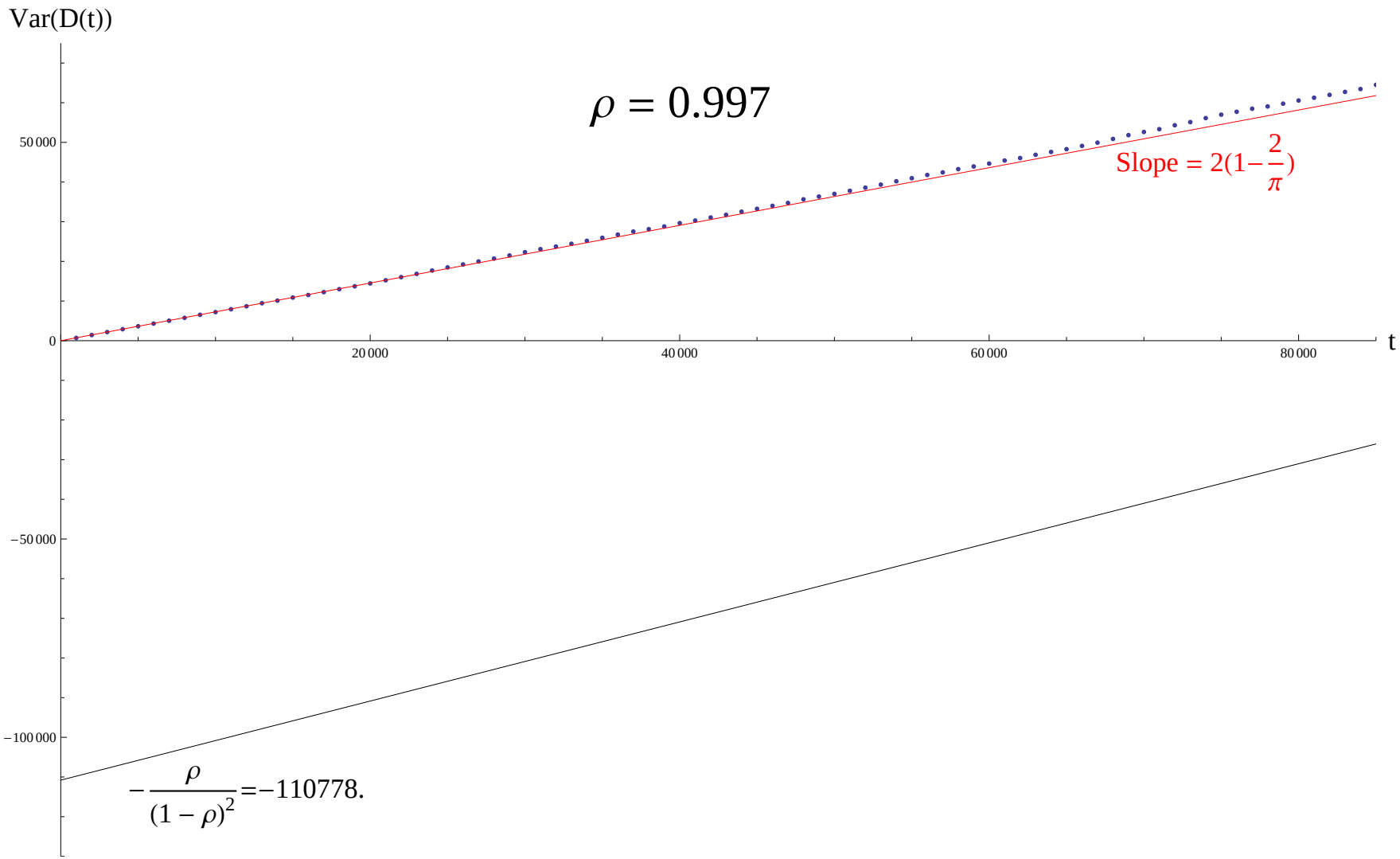
M/M/1 Queue


Var(D(t))
$\rho = 0.993$
Slope = $2(1-\frac{2}{\pi})$
50 000
0
−50 000
−100 000
20 000
40 000
60 000
80 000
t
$-\frac{\rho}{(1-\rho)^2} = -20265.3$

M/M/1 Queue


Var(D(t))
$\rho = 0.995$
Slope = $2(1-\frac{2}{\pi})$
50 000
0
−50 000
−100 000
20 000
40 000
60 000
80 000
t
$-\frac{\rho}{(1-\rho)^2} = -39800.$

M/M/1 Queue


Var(D(t))
$\rho = 0.997$
Slope = $2(1-\frac{2}{\pi})$
50 000
0
−50 000
−100 000
20 000
40 000
60 000
80 000
t
$-\frac{\rho}{(1-\rho)^2} = -110778.$

Asymptotic Variance of Flows in General Open Stable Queueing Networks

Stable Open Queueing Networks

Generalized Jackson network with external arrival vector α, routing matrix P and service capacity vector μ.

Stable if $\nu := (I - P')^{-1}\alpha < \mu$.

Counting processes: $E(\cdot)$ are entrances to nodes (exogenous and endogenous). $D_{i,j}(\cdot)$, are packets passed from node i to j.

Stable Open Queueing Networks

Generalized Jackson network with external arrival vector α, routing matrix P and service capacity vector μ.

Stable if $\nu := (I - P')^{-1}\alpha < \mu$.

Counting processes: $E(\cdot)$ are entrances to nodes (exogenous and endogenous). $D_{i,j}(\cdot)$, are packets passed from node i to j.

Simple:

$$\nu_i := \lim_{t\to\infty} \frac{\mathbb{E}[E_i(t)]}{t}.$$

Stable Open Queueing Networks

Generalized Jackson network with external arrival vector α, routing matrix P and service capacity vector μ.

Stable if $\nu := (I - P')^{-1}\alpha < \mu$.

Counting processes: $E(\cdot)$ are entrances to nodes (exogenous and endogenous). $D_{i,j}(\cdot)$, are packets passed from node i to j.

Simple:

$$\nu_i := \lim_{t\to\infty} \frac{\mathbb{E}[E_i(t)]}{t}.$$

Our contribution: Computable exact formulas for:

$$\sigma_{i,j} := \lim_{t\to\infty} \frac{Cov\Big(E_i(t), E_j(t)\Big)}{t}, \qquad \sigma_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{Cov\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)}{t}.$$

Main Queueing Network Result

Theorem: Werner Scheinhardt, Y. N. 2013

Define,

$$\Sigma^{(D)} := H\,\Sigma^{(F)}\,H', \qquad \Sigma^{(E)} := \Big(BH + [I \;\; 0]\Big)\Sigma^{(F)}\Big(BH + [I \;\; 0]\Big)'.$$

Here $\Sigma^{(F)}$ is the covariance matrix of the primitive input sequences (arrivals and routing) and the matrices B and H are easily constructed with only the inversion $(I - P')^{-1}$.

Under general assumptions of stable (single or multi-class) queueing networks, the processes $D(\cdot)$ and $E(\cdot)$ converge weakly to Brownian motions with the above covariance matrices.

Further,

$$\sigma_{i_1 \to j_1, i_2 \to j_2} = \Sigma^{(D)}_{(i_1-1)K+j_1,\ (i_2-1)K+j_2}, \qquad \sigma_{i,j} = \Sigma^{(E)}_{i,j}.$$

Note: σ's do not depend on the service variance. Compare with QNA (Queueing Network Analyzer, Whitt 80's).

Wrap Up

Is all this useful?

- BRAVO: Can it be incorporated in system planning, estimation, or control?
- Network Flows: Can heuristic queueing network decomposition schemes be improved?
- Other uses?

References

- Daryl J. Daley, Johan van Leeuwaarden and Y.N., "*BRAVO for QED Finite Birth-Death Queues*", preprint.
- Sophie Hautphenne, Yoav Kerner, Y.N., Peter Taylor, "*The Second Order Terms of the Variance Curves for Some Queueing Output Processes*", preprint.
- Y. N., Werner Scheinhardt, "*Diffusion Parameters of Flows in Stable Queueing Networks*", preprint.
- 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.