# Balancing Reduces Asymptotic Variance of Outputs

Yoni Nazarathy
EURANDOM and Department of Mechanical Engineering, TU/e Eindhoven,
CWI Amsterdam

Contains joint work with
Ahmad Al-Hanbali, Yoav Kerner, Gideon Weiss and Ward Whitt

Queueing Colloquium, CWI Amsterdam
September 28, 2009

# Outline

# Outline

# Asymptotic Variance Rate

- $D(t) \equiv$ Number of outputs during $[0, t]$
- Often, $\exists \overline{V}: \ \text{Var}\big(D(t)\big) = \overline{V}\, t + o(t)$
- $\overline{V} \equiv$ Asymptotic Variance Rate

## Examples

- $D(t)$ Poisson with rate $\lambda$: $\ \overline{V} = \lambda$
  — E.g. stationary M/M/1 with arrival rate $\lambda$
- $D(t)$ renewal with inter-renewal mean $1/\lambda$ and SCV $c^2$: $\ \overline{V} = \lambda c^2$
  — E.g. M/M/1/1 (no waiting room) with arrival = service rate = $\lambda$
  – $D(t)$ renewal with Erlang(2,$\lambda$)
  – $\text{Var}\big(D(t)\big) = \frac{1}{4}\lambda t + \frac{1}{8} - \frac{1}{8}e^{-2\lambda t}$ (stationary case)
- M/G/1 with infinite variance Pareto service times:
  $\overline{V}$ does not exist (Daley, Vesilo, 1997)

# Underloaded Queues ($\lambda < \mu$): Typically $\overline{V} = \overline{V}_{\text{in}}$

## GI/G/1

Proof:

1. Assume $Q(0) = 0$: $D(t) = A(t) - Q(t)$
2. $\text{Var}\big(D(t)\big) = \text{Var}\big(A(t)\big) + \text{Var}\big(Q(t)\big) - 2\text{Cov}\big(A(t), Q(t)\big)$
3. Assume $\text{Var}\big(Q(t)\big) < \infty$ and $\overline{V}$ exists:
   $\text{Cov}\big(A(t), Q(t)\big) \leq \sqrt{\text{Var}\big(A(t)\big)\text{Var}\big(Q(t)\big)} = O(\sqrt{t})$
4. Divide by $t$ and take $t \to \infty$

## GI/G/1/K

Intuition:

1. For $K$ large and $\lambda \ll \mu$, queue is almost never full
2. Similar to GI/G/1 ($\overline{V} \approx \overline{V}_{\text{in}}$)

# Overloaded Queues ($\lambda > \mu$): Typically $\overline{V} = \overline{V}_{\text{service}}$

## GI/G/1

Intuition:

1. After a finite time, $\tau$, queue never empties again
2. For $t > \tau$: output is a renewal process of services

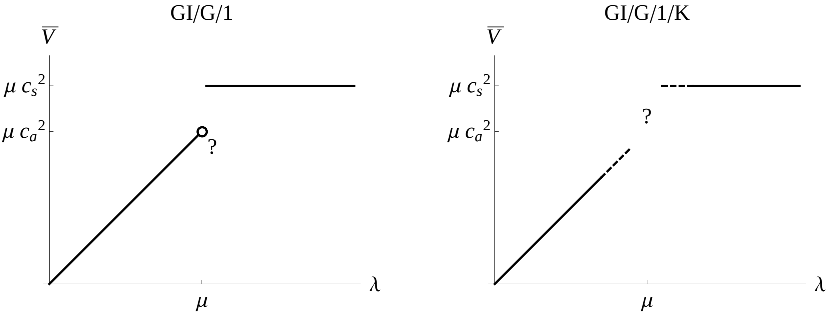
## GI/G/1/K

Intuition:

1. For $K$ large and $\lambda \gg \mu$, queue is almost never empty
2. Output is "almost" renewal ($\overline{V} \approx \overline{V}_{\text{service}}$)

# Intermediate Summary

$$\overline{V}_{\text{GI/G/1}} = \begin{cases} \lambda c_a^2 & \lambda < \mu \\ ? & \lambda = \mu \\ \mu c_s^2 & \lambda > \mu \end{cases}, \qquad \overline{V}_{\text{GI/G/1/K}} = \begin{cases} \lambda c_a^2 & \lambda \ll \mu \\ ? & \lambda \approx \mu \\ \mu c_s^2 & \lambda \gg \mu \end{cases}$$



# Flow Rate and IDC

- Flow rate:

$$\lambda^* = \lim_{t\to\infty} \frac{E[D(t)]}{t}$$

- Index of dispersion of counts (IDC):

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

## Examples

- IDC of a Poisson process is 1
- IDC of a renewal process is $c^2$

## For GI/G/1 and GI/G/1/K with $\lambda \neq \mu$:

The IDC is "renewal like"

# Preview of Results

We shall show:

**BRAVO**: **B**alancing **R**educes **A**symptotic **V**ariance of **O**utputs

| M/M/1 | GI/G/1 | M/M/1/K | GI/G/1/K |
|---|---|---|---|
| $\lambda 2\left(1-\frac{2}{\pi}\right)$ | $\lambda(c_a^2+c_s^2)\left(1-\frac{2}{\pi}\right)$ | $\lambda\frac{2K^2+K}{3(K+1)^2}$ | $\lambda\Big((c_a^2+c_s^2)\frac{1}{3}+o_K(1)\Big)$ |

## Notes

- When $\lambda = \mu$ the IDC is NOT "renewal like"
- $\lim_{K\to\infty} \overline{V}_{\text{GI/G/1/K}} \neq \overline{V}_{\text{GI/G/1}}$
- Currently: Complete proofs for M/M cases only
- Currently: Lacking good intuition for BRAVO

# Outline

# M/M/1 BRAVO

## Theorem

*For the M/M/1 queue with $\lambda = \mu$:*

$$\overline{V} = \lambda 2\Big(1 - \frac{2}{\pi}\Big)$$

## Notes

- $2\big(1 - \frac{2}{\pi}\big) = 0.72676\ldots$
- Explicit expression for $\mathrm{Var}\big(D(t)\big)$ in terms of Bessel functions
- From this explicit expression:

$$\mathrm{Var}\big(D(t)\big) = \overline{V}\,t + \frac{1}{\sqrt{\pi}}(\lambda t)^{1/2} + \big(\frac{1}{4} - \frac{1}{2\pi}\big) + o_t(1)$$

# M/M/1 BRAVO: Outline of Derivation

- Denote: $X_\alpha \sim \exp(\alpha)$, $\phi_\alpha(z) := E[z^{D(X_\alpha)}] = \int_0^\infty \alpha e^{-\alpha t} E[z^{D(t)}]dt$
- Denote $\phi^1_\alpha := \frac{\phi'_\alpha(1)}{\alpha}$ and $\phi^2_\alpha := \frac{\phi''_\alpha(1)}{\alpha}$:

$$E[D(t)] = \mathcal{L}^{-1}(\phi^1_\alpha), \quad E[D(t)^2] = \mathcal{L}^{-1}(\phi^2_\alpha + \phi^1_\alpha)$$

- Known:
  — $\phi_\alpha(z) = \frac{\alpha}{\mu(1-z)+\alpha}\left(1 + r_\alpha(z)\frac{1-z}{z-r_\alpha(z)}\right)$, $r_\alpha(z) = \frac{\lambda+\mu+\alpha-\sqrt{(\lambda+\mu+\alpha)^2-4\lambda\mu z}}{2\lambda}$
  — Inverse transforms of $\phi^1_\alpha$ and $\phi^2_\alpha$
- Set $\lambda = \mu = 1$ and combine:

$$\begin{aligned}\mathrm{Var}\big((D(t)\big) &= \frac{1}{4}e^{-4t}\Big(e^{4t}(1+8t) - (1+4t)^2 I_0(2t)^2 - 4e^{2t}tI_1(2t) \\ &\quad -16t^2 I_1(2t)^2 - 4tI_0(2t)\Big(e^{2t} + (2+8t)I_1(2t)\Big)\Big),\end{aligned}$$

with $I_j(2t) = \sum_{n=0}^\infty \frac{t^{j+2n}}{(j+n)!\cdot n!}$

# M/M/1 BRAVO: Relation to the Y-Intercept

## Theorem

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

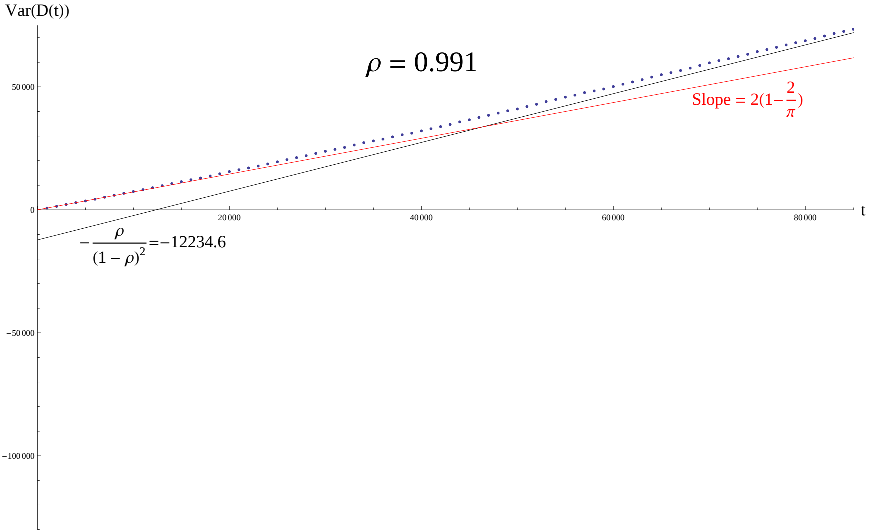
$$Var\big(D(t)\big) = \lambda t - (1 - L_0)\frac{\rho}{(1-\rho)^2} + o_t(1)$$

- *$L_0$ expression of $\rho$ and first 3 service moments*
- *For M/M/1, $L_0 = 0$*

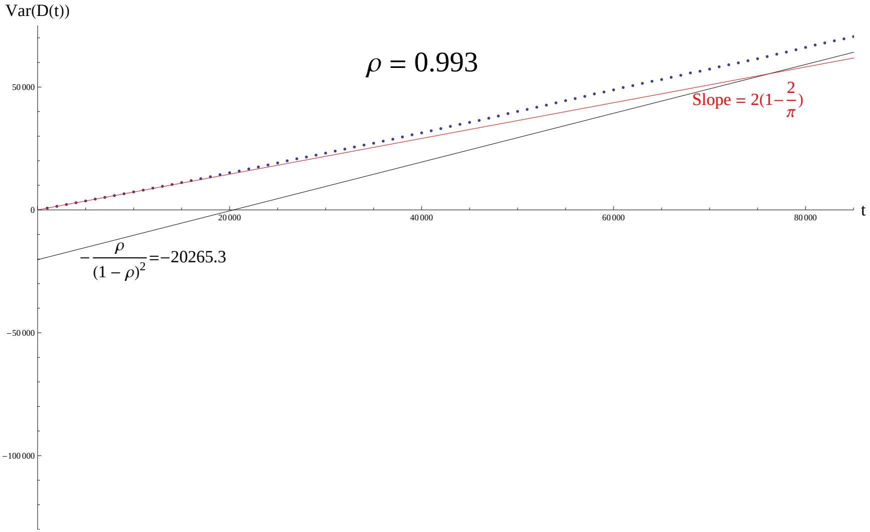
## Example: Look at the curves $\text{Var}\big(D(t)\big)$

- 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 $D(t)$:
  $3 \times 10^4$ repetitions, $10^5$ time units, sample variance every 1000

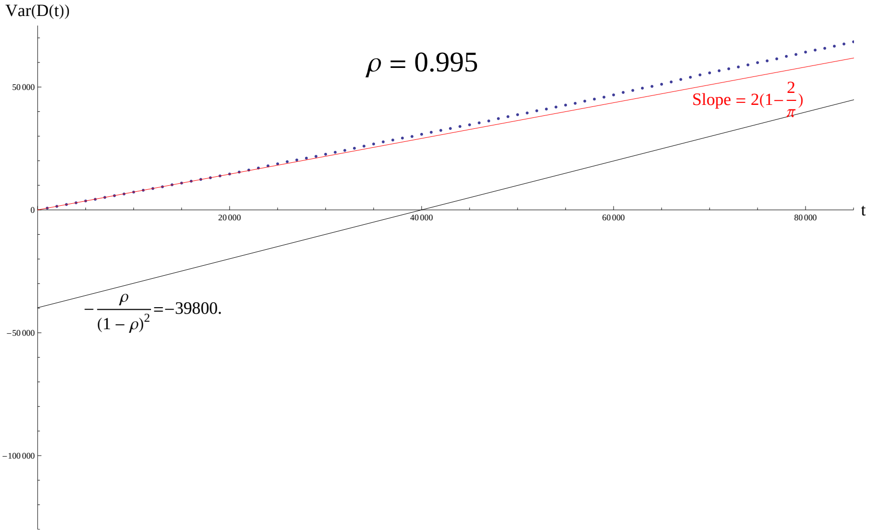
# M/M/1 BRAVO: Relation to the Y-Intercept



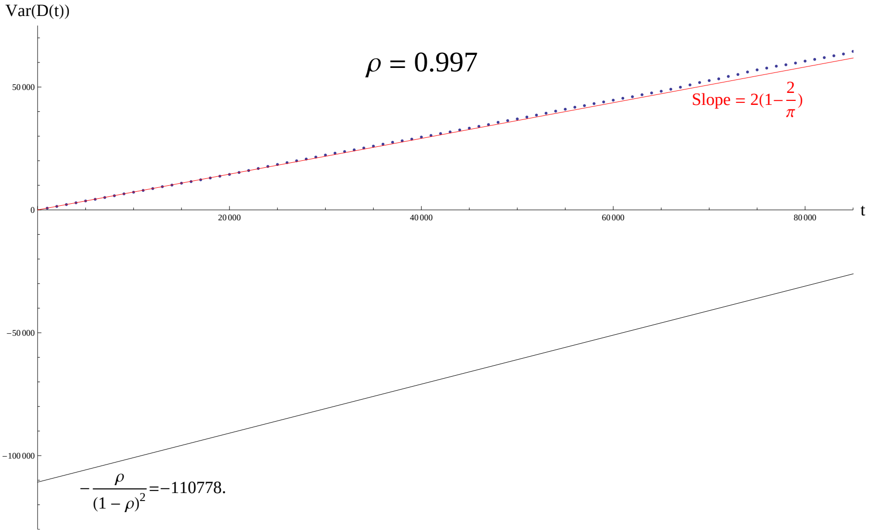
# M/M/1 BRAVO: Relation to the Y-Intercept



# M/M/1 BRAVO: Relation to the Y-Intercept



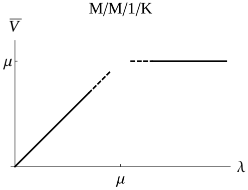
# M/M/1 BRAVO: Relation to the Y-Intercept



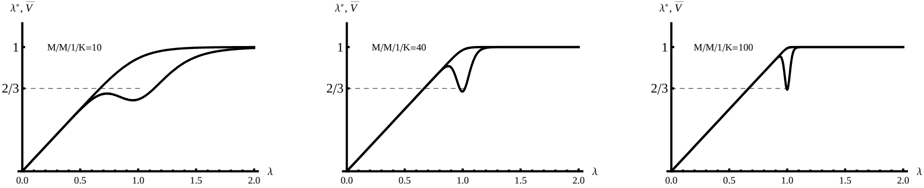
# Outline

# M/M/1/K BRAVO

We expect similar to M/M/1 when $\lambda \ll \mu$ or $\lambda \gg \mu$:




For $\lambda \approx \mu$ we get, $\overline{V} \approx \lambda \frac{2}{3}$:



# M/M/1/K BRAVO

## Theorem

*For the M/M/1/K queue with $\lambda = \mu$:*

$$\overline{V} = \lambda \frac{2K^2 + K}{3(K+1)^2}$$

## Notes

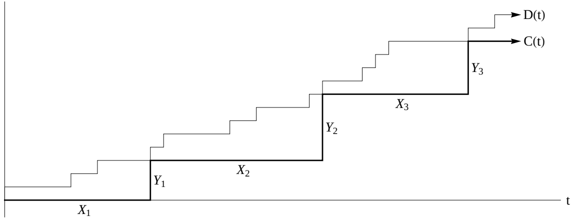
- Simple formula for any finite birth-death queue (depends on $\pi$)
- For "monotonic" finite birth-death queues, $IDC < 1$
- BRAVO also appears in M/M/c/K queues (IDC approximately minimized when $\lambda = \mu c$)
- For stationary M/M/1/K, $\lambda = \mu$:

$$\text{Var}\big(D(t)\big) = \overline{V}\, t + \frac{1}{180} \frac{7K^4 + 28K^3 + 37K^2 + 18K}{(K+1)^2} + o_t(1)$$

# M/M/1/K BRAVO Derivation (New Simple)

## "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. Variance rates of $C(t)$ and $D(t)$ are equal
4. Known:
   — Variance rate of $C(t)$ is $\frac{1}{E[X]} \mathrm{Var}\left(Y - \frac{E[Y]}{E[X]} X\right)$
   — For M/M/1/K: Explicit joint transform of $(X, Y)$



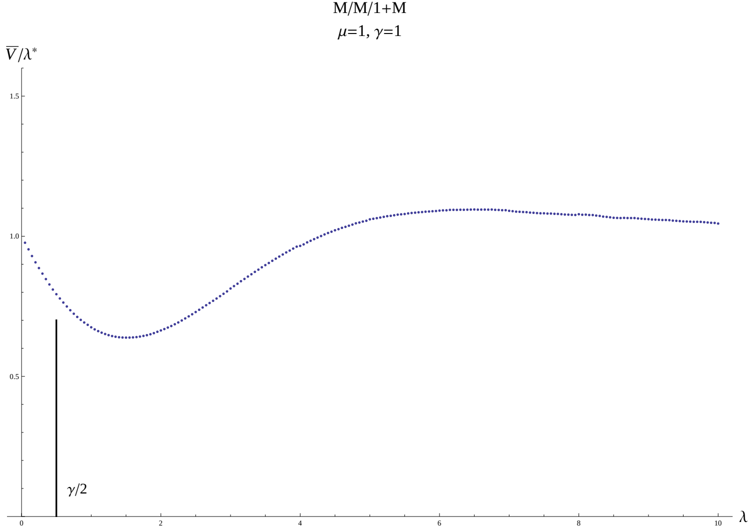
# Estimate $\overline{V}$ by Regenerative Simulation

- Assume $\mathrm{Var}(D(t)) = \overline{V}t + \overline{B} + o_t(1)$. Bias of $\frac{\mathrm{Var}(D(T))}{T}$ is $\overline{B}/T$
- Alternative: estimate moments of $(X, Y)$ and plug in $\frac{1}{E[X]}\mathrm{Var}\left(Y - \frac{E[Y]}{E[X]}X\right)$

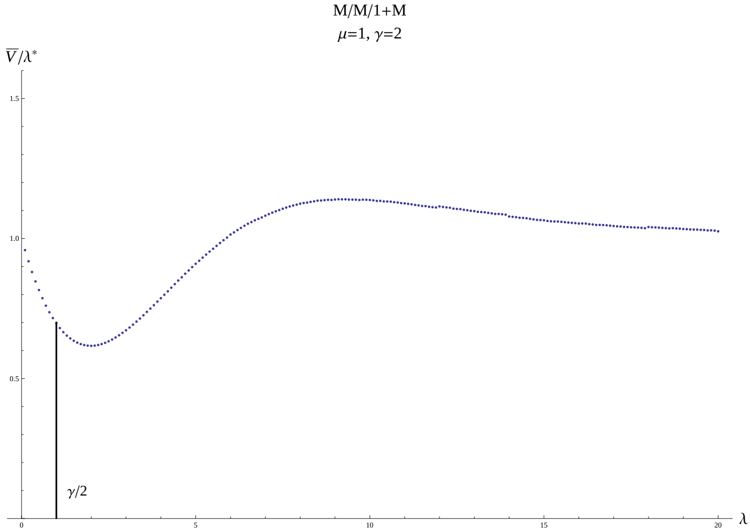
## Example: M/M/1+M

- Arrivals: $\lambda$, service: $\mu$, patience: $\gamma$
- $D(t)$ is a "thinned" version of $A(t)$
  — For i.i.d. thinning, IDC is 1
  — What about this case...
- Simulation details:
  — $\mu = 1$, $\gamma = 1, 2, 4, 8, \ldots, 256$
  — $\lambda$ varies over 200 points in $(0, \gamma/20)$
  — Regeneration state: approximately maximize $\pi_i$
  — $2 \times 10^6$ seeds, keep seed constant while varying $\lambda$
  — Running time: 15 minutes

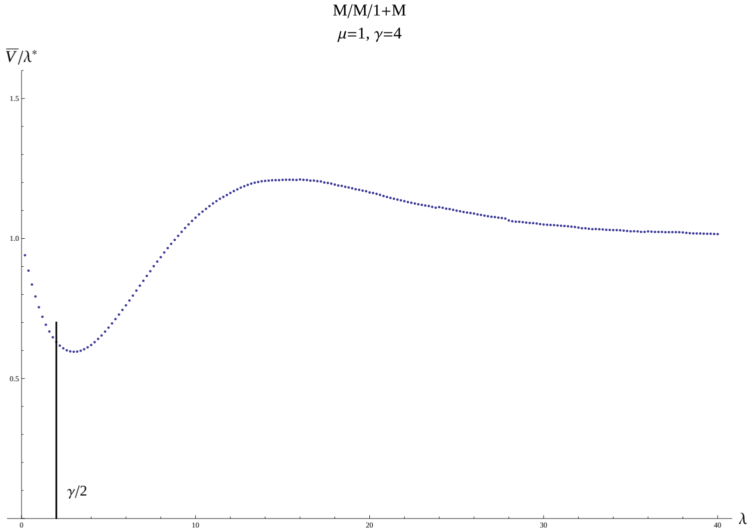
# Regenerative Simulation for M/M/1 + M



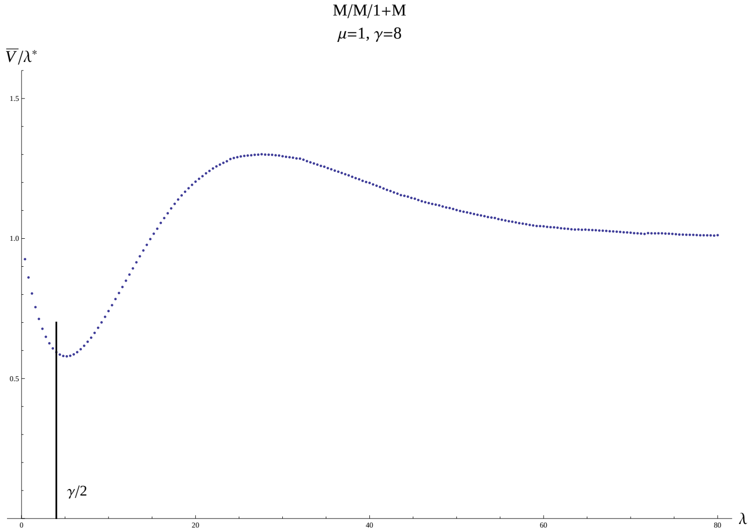
# Abandonments: M/M/1 + M



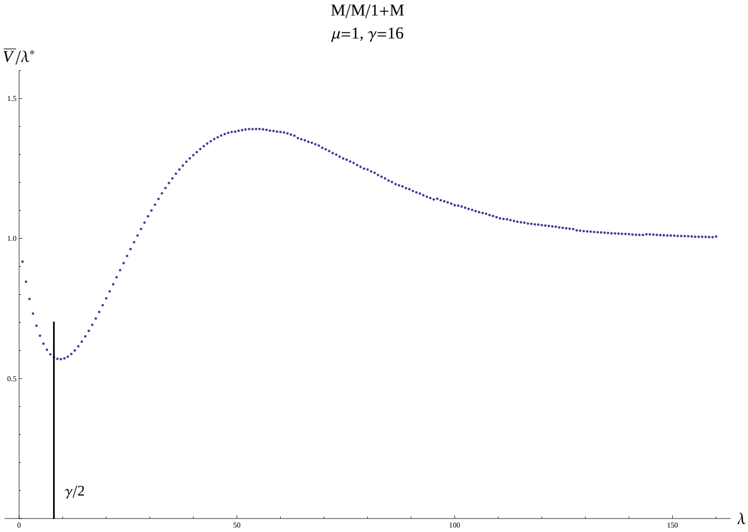
# Abandonments: M/M/1 + M



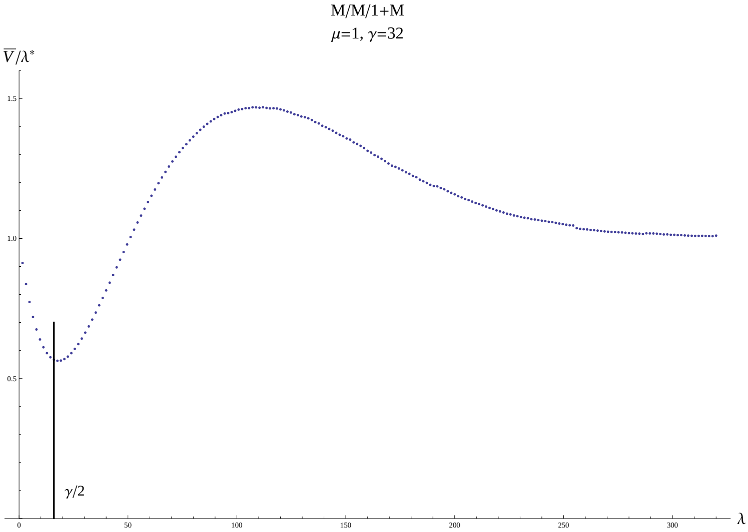
# Abandonments: M/M/1 + M



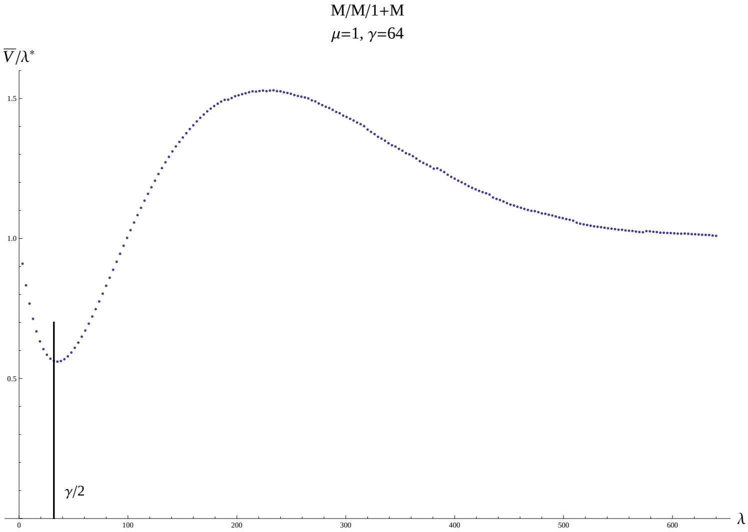
# Abandonments: M/M/1 + M



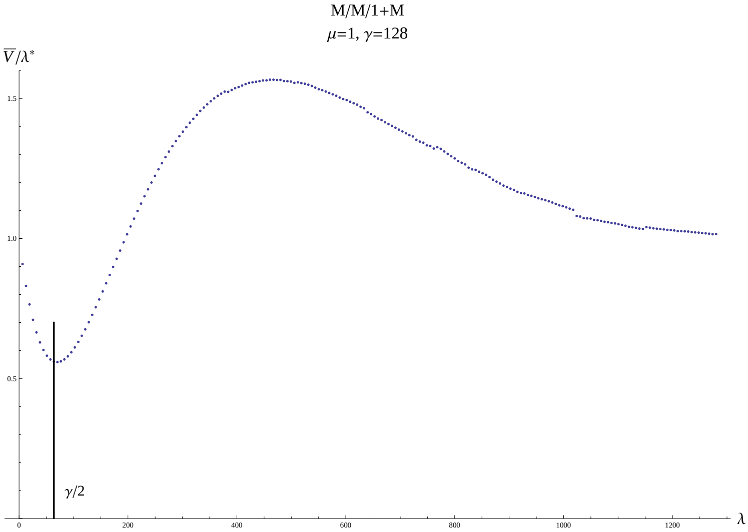
# Abandonments: M/M/1 + M



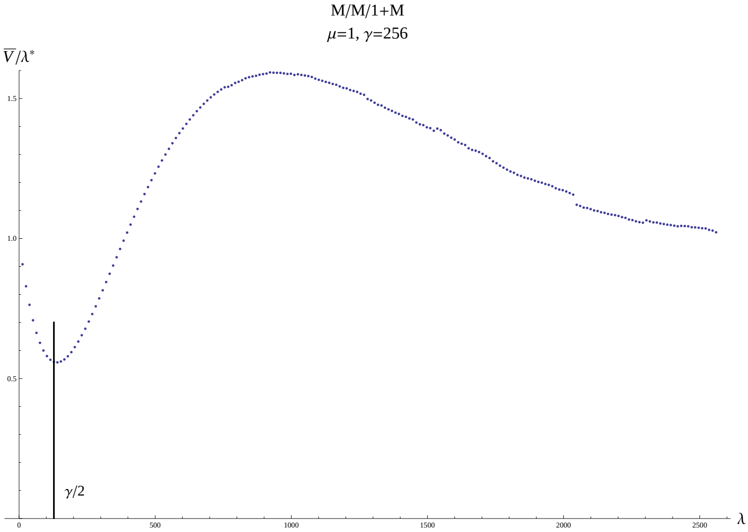
# Abandonments: M/M/1 + M



# Abandonments: M/M/1 + M



# Abandonments: M/M/1 + M



# Outline

# GI/G/1 BRAVO

## Claim

For GI/G/1 with $\lambda = \mu$:

$$\overline{V} = \lambda(c_a^2 + c_s^2)(1 - \frac{2}{\pi})$$

## Notes

- No full proof (yet)
- Moment conditions
- Holds also for $GI/G/c$

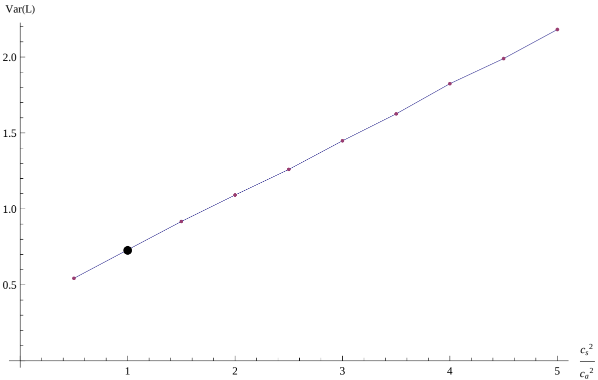
# GI/G/1 BRAVO

## Outline of basis for claim

1. Iglehart and Whitt (1970):
   — GI/G/c, $\lambda = \mu$
   — $\hat{D}^n(t) := (D(nt) - \lambda nt)n^{-1/2}$
   — $B_1(\cdot), B_2(\cdot)$ independent Brownian motions
   — Continuous mapping: $g(U, V)(t) := V(t) + \inf_{0 \le s \le t}\{U(s) - V(s)\}$
   — $\hat{D}^n(\cdot) \Rightarrow g(c_a B_1(\cdot), c_s B_2(\cdot))$
2. $\frac{D(t)-\lambda t}{\sqrt{\lambda t}} \Rightarrow c_a \inf_{0 \le t \le 1}\{B_1(t) + \sqrt{\frac{c_s^2}{c_a^2}} B_2(1-t)\}$
   — Uniform integrability conditions for first two moments
   — $\overline{V} = \lim_{t\to\infty} \frac{\mathrm{Var}\left(D(t)\right)}{t} = \lambda c_a^2 \mathrm{Var}\left(\inf_{0 \le t \le 1}\{B_1(t) + \sqrt{\frac{c_s^2}{c_a^2}} B_2(1-t)\}\right)$
3. Brownian motion simulations show variance is affine in $\frac{c_s^2}{c_a^2}$
4. M/D/1 result obtained directly: $\overline{V} = \lambda\left(1 - \frac{2}{\pi}\right)$
5. M/D/1, M/M/1 and affinity in $\frac{c_s^2}{c_a^2}$ imply the result

# GI/G/1 BRAVO

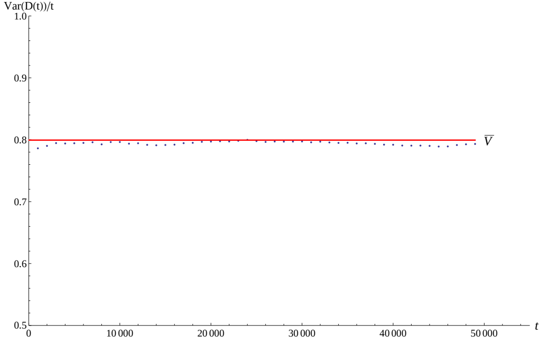
- Brownian motion simulations of $\inf_{0 \le t \le 1}\{B_1(t) + \sqrt{\frac{c_s^2}{c_a^2}} B_2(1-t)\}$
- Variance estimates as a function of $\frac{c_s^2}{c_a^2}$



# GI/G/1 BRAVO

## Explicit Simulation of $D(t)$, Taking Sample Variance

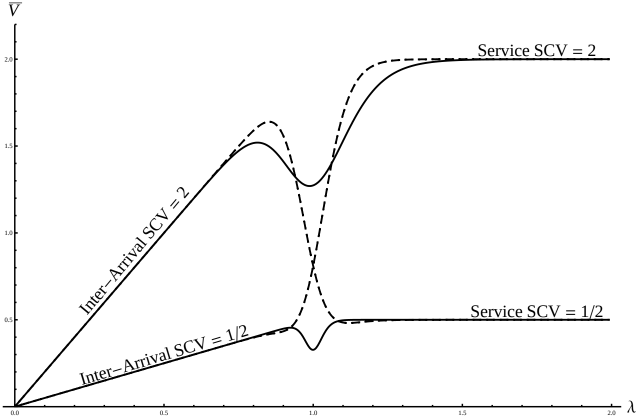
- Poisson arrivals ($c_a^2 = 1$), hyper-exp service with $c_s^2 = 1.2$
- $\overline{V} = (c_a^2 + c_s^2)\left(1 - \frac{2}{\pi}\right) = 0.80$
- $10^5$ repetitions, $50,0000$ time units, sample variance every 1000



# Outline

# GI/G/1/K BRAVO

Evaluating $\overline{V}$ using Markov Arrival Processes (Outputs of PH/PH/1/40):



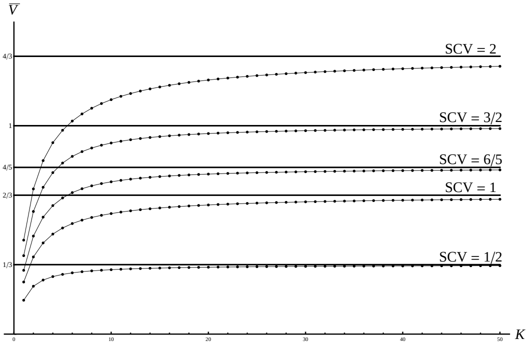

Clearly there is a BRAVO effect. What is the magnitude?

# GI/G/1/K BRAVO

## We Think:

$$\overline{V} = \lambda(c_a^2 + c_s^2)\frac{1}{3} + o_K(1)$$

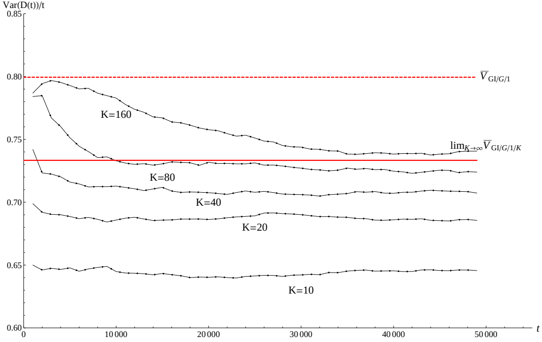
Numerically verifying for $c_a = c_s$:




Note: we also verified for $c_a \neq c_s$

## Explicit Simulation of $D(t)$, Taking Sample Variance

- Poisson arrivals ($c_a^2 = 1$), hyper-exp service with $c_s^2 = 1.2$
- $\overline{V} \approx (c_a^2 + c_s^2)\frac{1}{3} = 0.73$
- $10^5$ repetitions, $50,0000$ time units, sample variance every 1000



# Outline

# Discussion

**Question**

What "causes" BRAVO?

**Answer**

...

**Question**

Applications to operations management?

**Answer**

...

Thank you
for your attention