

Output processes of Queues, their Asymptotic Variance, and BRAVO

Yoni Nazarathy

The University of Queensland and Accumulation Point

Joint work with Daryl J. Daley and Jiesen Wang

Forty Years of Queueing Systems

Eurandom, Eindhoven

27–29 May 2026

Outline

- 1 Daryl Daley
- 2 BRAVO
- 3 Current Contribution
(Daley, N., Wang)

Daryl Daley

In memory of Daryl J. Daley (1939–2023)



*Photo in Melbourne, March 2023,
working on this paper!*

Daryl passed away on 16 April 2023, before this paper appeared in print.

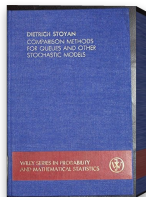
It is a privilege to present joint work with him at this celebration of forty years of *Queueing Systems*.

“An internationally acclaimed researcher in applied probability and a gentleman of great kindness.”

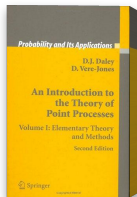
— Peter Taylor, JAP obituary (2023)

About Daryl

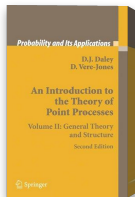
- Born Coburg, Victoria, 4 April 1939; Senior Fellow and later Professor (from 2005) at the Australian National University; Honorary Professorial Associate at the University of Melbourne from 2008.
- Doctoral student of David Kendall (Pembroke College, Cambridge, 1967).
- Foundational contributions across queueing, renewal theory, point processes, epidemic models, and stochastic geometry.
- Recipient of the Pitman Medal of the Statistical Society of Australia (2006).
- Hiker, musician, education researcher, family man, and an incredible conversationalist.



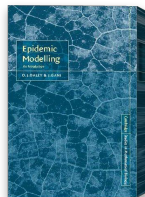
*Comparison
Methods for Queues
and Other
Stochastic Models*[†]
D. Stoyan,
translated and
edited by Daley
1983



*An Introduction to
the Theory of Point
Processes*
Vol. I, with
D. Vere-Jones
2003 (first ed. 1988)



*An Introduction to
the Theory of Point
Processes*
Vol. II, with
D. Vere-Jones
2008



*Epidemic
Modelling: An
Introduction*
with J. Gani
1999

[†] *Translated and substantially revised from D. Stoyan, Qualitative Eigenschaften und Abschätzungen stochastischer Modelle — a notable East–West collaboration of its time.*

Daryl's footprint in *Queueing Systems*

Across the first 40 years of QUESTA, four papers by Daryl:

1994 *Conditions for finite moments of waiting times in G/G/1 queues*
with R. D. Foley and T. Rolski. **QUESTA**.

2001 *The busy period of the M/GI/ ∞ queue*
QUESTA.

2011 *Revisiting queueing output processes: a point process viewpoint*
QUESTA.

2026 *Second order properties of thinned counts in finite birth–death processes*
with N. and J. Wang. **QUESTA**. — *the subject of this talk.*

Daryl on queues: beyond *Queueing Systems*

A short, thematic tour — selected from a much longer list.

- **Customer impatience & waiting times.** 1964 (JAMS); 1965, 1969 (JAP); with Foley & Rolski 1992 (AAP).
- **Single-server queue properties (1960s–70s).** serial-correlation of waiting times 1968 (*J. Austral. Math. Soc.*); stochastic monotonicity 1968 (*Z. Wahrsch.*); moment inequalities 1977 (*Z. Wahrsch.*).
- **Output & departure processes.** 1968 (*Ann. Math. Statist.*); 1974 in *Math. Methods in Queueing Theory* (LNEMS 98); 1976 (AAP, *Queueing output processes*).
- **Busy periods.** 1969 (JAP, parts I & II w/ Jacobs); 2004 with Baltrūnas & Klüppelberg (*Stoch. Proc. Appl.*).
- **Light-traffic approximations.** with Rolski: 1984 (MOR), 1991 (MOR), 1992 (AAP), 1994 (SPA).
- **Comparison & optimality of disciplines.** with Rolski 1984 (JAP); 1987 (SPA).
- **Long-range dependence in queues.** with Vesilo 1997 (SPA), 2000; with Rolski & Vesilo 2002 (AAP), 2007.

BRAVO

What is BRAVO?

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

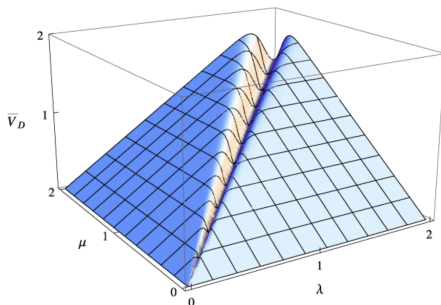
For an output counting process $D(t)$ define the asymptotic variance rate and the asymptotic index of dispersion of counts

$$\overline{V}_D := \lim_{t \rightarrow \infty} \frac{\text{Var } D(t)}{t}, \quad \mathcal{D} := \lim_{t \rightarrow \infty} \frac{\text{Var } D(t)}{\mathbb{E} D(t)}.$$

Reminder. In a stationary (stable) Markovian M/M/s, the output is a Poisson process at rate λ , so $\overline{V}_D = \lambda$ and $\mathcal{D} = 1$.

Reminder. For a renewal process with inter-event time T , $\mathcal{D} = c_T^2 := \text{Var}(T) / (\mathbb{E} T)^2$.

BRAVO picture #1

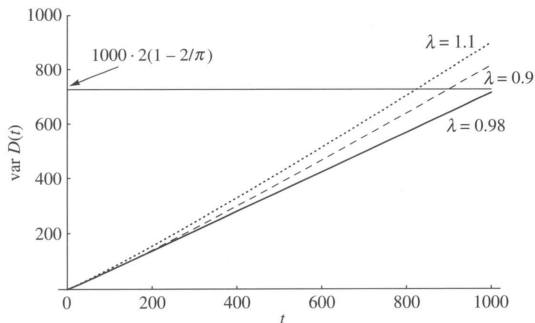


$\bar{V}_D$ for the M/M/1/40 output vs. (λ, μ) . N. & Weiss, *QUESTA*, 2008.

- Far from balance ($\rho \neq 1, \rho := \lambda/\mu$): $\mathcal{D} \approx 1$ (essentially Poisson).
- At balance ($\rho = 1$): $\mathcal{D} \rightarrow \mathcal{B} < 1$, the BRAVO constant. The value of $\mathcal{B}$ depends on the model.
- For the M/M/1/ K queue at $\rho = 1$, $\mathcal{B} = 2/3$ as the buffer size $K \rightarrow \infty$.

See also Hautphenne, Kerner, N. & Taylor, *EJOR*, 2015 — for further analysis via Markovian Arrival Processes (MAPs).

BRAVO picture #2



Al Hanbali, Mandjes, N. & Whitt, AAP, 2011.

- $\text{Var } D(t)$ vs t for the M/M/1 queue, $\mu = 1$.
- $\lambda = 0.9$ and $\lambda = 1.1$ rays: linear growth at rate $\bar{V}_D$ predicted by the model.
- $\lambda = 0.98$ (near-critical): variance flattens *below* the rays.
- Line at $1000 \cdot 2(1 - 2/\pi)$: the asymptotic BRAVO level for M/M/1 at $\rho = 1$.

Infinite-buffer single-server: $2(1 - 2/\pi)$

Critically loaded GI/G/1 (work output or departure counts at $\rho = 1$):

$$\mathcal{D} \longrightarrow \mathcal{B} = (1 - 2/\pi) (c_a^2 + c_s^2),$$

where c_a, c_s are the CVs of inter-arrival and service times. M/M/1 specialises to $2(1 - 2/\pi) \approx 0.7268$.

- Showed BRAVO is a *regime* phenomenon, not a finite-buffer artefact, and revealed the squared-CV structure. *Al Hanbali, Mandjes, N. & Whitt, AAP, 2011.*
- Finite-buffer counterpart conjectured: $\mathcal{B} = (c_a^2 + c_s^2)/3$ (recovers $2/3$ for M/M). *N., QUESTA, 2011.*
- Probabilistic mechanism: the critical busy period is regularly varying of index $\frac{1}{2}$. Proves GI/G/1 BRAVO under $2+\varepsilon$ moments (vs. $4+\varepsilon$ before). *N. & Palmowski, QUESTA, 2022.*

Many-server QED: a different universal constant

Many-server M/M/ n queue with n servers and a finite buffer of size K_n , taken to the Halfin–Whitt (QED) regime

$$(1 - \rho_n)\sqrt{n} \rightarrow \beta, \quad K_n/\sqrt{n} \rightarrow \kappa.$$

In this scaling $\mathcal{D} \rightarrow \mathcal{B}(\beta, \kappa)$, an explicit function of the standard-normal Φ, ϕ . At the balanced point

$$\mathcal{B} \approx 0.6093.$$

- BRAVO is a family of regime-specific constants, not one number.
Daley, N. & van Leeuwaarden, AAP, 2015.
- Subsequent correction to the integral expressions (numerical value 0.6093 unchanged).
Nadarajah & Pogány, ORL 46, 2018.

BRAVO and heavy traffic

The critically-loaded departure process. For GI/G/1 at $\rho = 1$, the *departure* process (not the queue length, not the arrival process) carries the BRAVO signature. Under diffusion scaling the departure process converges to

$$\hat{D}(t) = \inf_{0 \leq s \leq t} \{c_a^2 B_1(s) + c_s^2 B_2(t-s)\},$$

with B_1, B_2 independent Brownian motions.

Al Hanbali, Mandjes, N. & Whitt, AAP, 2011.

Heavy-traffic three-scale picture. At queue depth $\sim n$, $\text{Var } D(t)$ has three regimes: *sub-diffusive* ($t \ll n$), *diffusive* ($t \asymp n$, Poisson-like), *super-diffusive* ($t \gg n$, capped by the server). BRAVO is the *gap* between the latter two.

Glynn & R.J. Wang, SPA, 2023.

BRAVO constants at a glance

Model / regime	$\mathcal{B}$	Source
M/M/1/ K , $\rho = 1$, $K \rightarrow \infty$	$2/3$	NW 2008
GI/G/1/ K (conjecture)	$(c_a^2 + c_s^2)/3$	N. 2011
M/M/1, $\rho = 1$	$2(1 - 2/\pi) \approx 0.7268$	AhMNW 2011
GI/G/1, $\rho = 1$	$(1 - 2/\pi)(c_a^2 + c_s^2)$	AhMNW 2011; NP 2022
M/M/ n with buffer K_n , Halfin–Whitt	≈ 0.6093	DNvL 2015; NP 2018
Birth–death with thinned counts	general formula	<i>this talk</i>

Two recurring themes:

- (i) BRAVO is a *regime* effect, alive only near criticality;
- (ii) each model class has its own $\mathcal{B}$, but they share some root causes:
 - critical busy-period variance amplification;
 - avoidance of modulation at the boundaries;
 - heavy-traffic connection between regimes.

Frankly — I still don't fully understand it.

Current Contribution (Daley, N., Wang)

Setup: thinned counts in a birth–death process

- $Q(t)$: irreducible CTMC on $\mathbb{J} = \{0, 1, \dots, J\}$ with birth rates λ_i , death rates μ_i ; stationary distribution π_i .
- Each birth at state i is *independently retained* with prob. q_i^+ ; each death from state i with prob. q_i^- .
Convention: $q_0^- = q_J^+ = 0$.
- $N_q(0, t]$ counts the retained transitions in $(0, t]$.
- Object of study: the Asymptotic Index of Dispersion of counts

$$\mathcal{D} = \lim_{t \rightarrow \infty} \frac{\text{Var } N_q(0, t]}{\mathbb{E} N_q(0, t]}.$$

- *State-dependent* thinning is the new feature — $q_i^\pm$ may take any value in $[0, 1]$, not just $\{0, 1\}$.

Why state-dependent thinning?

Motivating example: M/M/s/K+M with reneging.

- Customers in queue abandon at rate γ each.
- In state i the total “death” rate is $\mu_i = \mu \min(i, s) + \gamma \max(i - s, 0)$.
- We want to count *served* customers, not abandoners. Each death is a real service with probability

$$q_i^- = \frac{\mu \min(i, s)}{\mu_i},$$

which depends on the state i . Set $q_i^+ \equiv 0$.

Prior variance results (N.–Weiss 2008; Daley–N.–van Leeuwen 2015) cover only $q \in \{0, 1\}$, i.e. no reneging.

Main result

Let $P_k = \sum_{i \leq k} \pi_i$, and set

$$\bar{\lambda} = \sum_{i=0}^{J-1} \pi_i \lambda_i (q_i^+ + q_{i+1}^-), \quad \Lambda_k = \frac{1}{\bar{\lambda}} \left(\pi_k \lambda_k q_k^+ + \sum_{i=0}^{k-1} \pi_i \lambda_i (q_i^+ + q_{i+1}^-) \right).$$

Both $\{P_k\}$ and $\{\Lambda_k\}$ are CDF vectors on $\mathbb{J}$.

Theorem

For any irreducible birth–death Q on finite $\mathbb{J}$,

$$\mathcal{D} = 1 + 2 \sum_{k=0}^{J-1} R_k, \quad R_k = (P_k - \Lambda_k) \left(\frac{\bar{\lambda} (P_k - \Lambda_k)}{\pi_k \lambda_k} + q_k^+ - q_{k+1}^- \right).$$

Quadratic part of R_k is ≥ 0 ; the sign-changing $q_k^+ - q_{k+1}^-$ piece is what lets $\mathcal{D}$ dip below 1 (the BRAVO regime).

Corollary: the complete pure birth/death case

Specialise to the cases without state-dependent thinning:

- *Complete pure birth-counting*: $q_i^+ \equiv 1, q_i^- \equiv 0$.
- *Complete pure death-counting*: $q_i^+ \equiv 0, q_i^- \equiv 1$.

In either case, set $\bar{\lambda}^* = \sum_{j=0}^{J-1} \pi_j \lambda_j$ and

$$\Lambda_k^- = \frac{1}{\bar{\lambda}^*} \sum_{i=0}^{k-1} \pi_i \lambda_i, \quad \Lambda_k^+ = \frac{1}{\bar{\lambda}^*} \sum_{i=0}^k \pi_i \lambda_i.$$

Corollary

For both the complete pure birth- and complete pure death-counting cases,

$$\mathcal{D} = 1 + 2\bar{\lambda}^* \sum_{k=0}^{J-1} \frac{(P_k - \Lambda_k^-)(P_k - \Lambda_k^+)}{\pi_k \lambda_k}.$$

Recovers the formula from N. & Weiss, *QUESTA*, 2008.

What the formula generalises

- **Complete pure death-counting** ($q^+ \equiv 0, q^- \equiv 1$): R_k collapses to $\bar{\lambda}(P_k - \Lambda_k^-)(P_k - \Lambda_k^+) / (\pi_k \lambda_k)$, recovering *N.-Weiss (2008)* for finite-buffer birth-death queues — previously proved via MAP/MMPP machinery; here it falls out of an elementary renewal-reward argument.
- **Complete pure birth-counting** is symmetric.
- **Infinite state space.** Under positive recurrence define

$$\mathcal{D}_\infty := 1 + 2 \sum_{k=0}^{\infty} R_k.$$

Conjectured to equal the true asymptotic index of dispersion of counts under suitable regularity (e.g. $\sum_k (1 - P_k)^2 / (\pi_k \lambda_k) < \infty$); tested against classical results on a panel of examples.

Regeneration. At $\mathcal{T} = \{t : Q(t-) = 0, Q(t) = 1\}$ the cycles (X_i, Y_i) — idle time + busy period, with the retained events inside — are i.i.d.

Brown–Solomon (1974) for renewal-reward with finite second moments:

$$\mathcal{D} = \frac{\mathbb{E}[Y^2] - 2R \mathbb{E}[XY] + R^2 \mathbb{E}[X^2]}{\mathbb{E}[Y]}, \quad R = \frac{\mathbb{E}[Y]}{\mathbb{E}[X]} = \bar{\lambda}.$$

Reduction. Computing $\mathcal{D}$ reduces to five busy-period moments starting from state 1:

$$\tau_1^{(1)}, \tau_1^{(2)}, n_1^{(1)}, n_1^{(2)}, c_1^{(1)}.$$

τ — busy-period length; n — number served in a busy period; c — cross moment.

Derivation continued

$$\phi_k(s, z) = \mathbb{E}[e^{-s\tau_k} z^{n_k} \mid Q(0) = k].$$

First-step analysis:

$$(s + r_k)\phi_k(s, z) = \lambda_k(zq_k^+ + 1 - q_k^+)\phi_{k+1}(s, z) + \mu_k(zq_k^- + 1 - q_k^-)\phi_{k-1}(s, z).$$

Differentiating in s, z at $(0, 1)$ yields five tridiagonal recursions (all with the *same* left-hand side form):

$$\begin{aligned} r_k \tau_k^{(1)} - \lambda_k \tau_{k+1}^{(1)} - \mu_k \tau_{k-1}^{(1)} &= 1, \\ r_k n_k^{(1)} - \lambda_k n_{k+1}^{(1)} - \mu_k n_{k-1}^{(1)} &= \lambda_k q_k^+ + \mu_k q_k^-, \\ r_k c_k^{(1)} - \lambda_k c_{k+1}^{(1)} - \mu_k c_{k-1}^{(1)} &= n_k^{(1)} + \lambda_k q_k^+ \tau_{k+1}^{(1)} + \mu_k q_k^- \tau_{k-1}^{(1)}, \\ r_k \tau_k^{(2)} - \lambda_k \tau_{k+1}^{(2)} - \mu_k \tau_{k-1}^{(2)} &= 2\tau_k^{(1)}, \\ r_k n_k^{(2)} - \lambda_k n_{k+1}^{(2)} - \mu_k n_{k-1}^{(2)} &= \lambda_k q_k^+ + \mu_k q_k^- + 2(\lambda_k q_k^+ n_{k+1}^{(1)} + \mu_k q_k^- n_{k-1}^{(1)}). \end{aligned}$$

Derivation continued

Let $\mathbf{W}$ be the $J \times J$ tridiagonal matrix ($r_k = \lambda_k + \mu_k$):

$$\mathbf{W} = \begin{pmatrix} r_1 & -\lambda_1 & 0 & \cdots & 0 \\ -\mu_2 & r_2 & -\lambda_2 & \cdots & 0 \\ 0 & -\mu_3 & r_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\lambda_{J-1} \\ 0 & \cdots & 0 & -\mu_J & r_J \end{pmatrix}.$$

Lemma

$$(\mathbf{W}^{-1})_{ij} = \sum_{k=1}^{\min(i,j)} \pi_j / (\pi_k \mu_k).$$

We have not seen this before (has anyone?) — it was “guessed” via trial and error and numerical inspection around 2013. Today with AI, such “guessing” is becoming much easier.

Derivation: assembling the result

In matrix form $\mathbf{W}\mathbf{x} = \text{RHS}$, and the explicit inverse yields closed-form moment expressions.

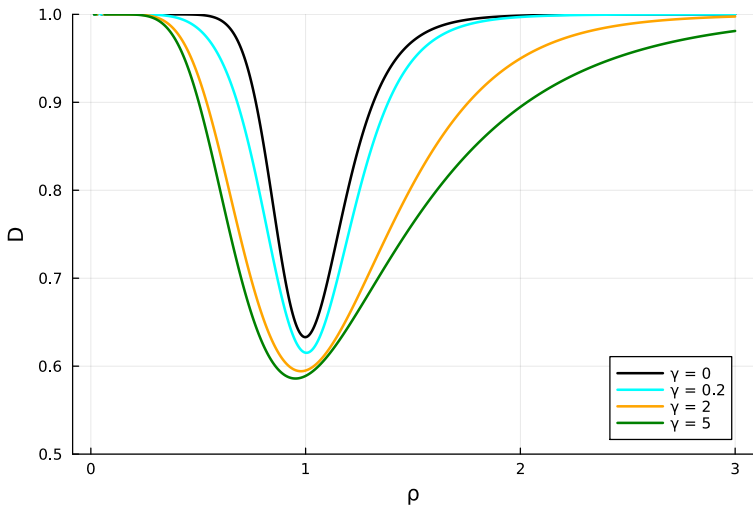
Plugging into Brown–Solomon and carefully extracting the required moments from the recursion (also considering the idle periods), a fair bit of algebra gives:

$$\mathcal{D} = 1 + 2 \sum_k R_k.$$

Note. The counting process is a MAP and admits matrix expressions, but we could not use those expressions directly to obtain this result — numerical MAM computations do, however, help verify it.

Related recent work: Asanjarani, Hautphenne & N., Performance Evaluation, 2026.

Example 1: M/M/s/K+M with reneging



$K = 20, s = 10, \mu = 1$, for several abandonment rates γ .
BRAVO *persists* under reneging.

Example 2: J animals fighting at the billabong



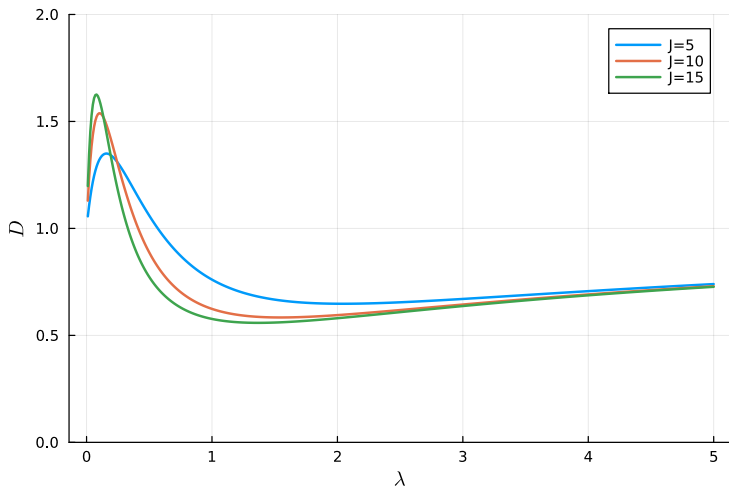
*Photo: Berrinba Wetlands billabong;
Wikimedia Commons, CC BY-SA 4.0.*



*Eindhoven born,
Australian grown,
Israeli heart.*

- Population of J animals.
- State i : number currently *at* the billabong.
- Each absent animal arrives at rate λ :
 $\lambda_i = (J - i)\lambda$.
- Each present animal leaves at rate μ :
 $\mu_i = i\mu$.
- On arrival, the newcomer interacts (fights) with one of the i already there with probability $q_i^+ = i/(i + 1)$.
- We count interactions; departures not counted ($q_i^- = 0$).

Example 2: finite population, $\mathcal{D}$ crosses 1



$\lambda_i = (J - i)\lambda$, $\mu_i = i\mu$ ($\mu = 1$), $q_i^+ = i/(i + 1)$, $q_i^- = 0$. Not a queueing output process.
 $\mathcal{D}$ is both sub- and supra-Poisson in different regimes; $\mathcal{D} \rightarrow 1$ as $\lambda \rightarrow \infty$.

Validating $\mathcal{D}_\infty$ on infinite-state examples

- **M/M/1 busy-cycle renewal** ($q_1^- = 1$, all other $q = 0$):

$$\mathcal{D}_\infty = 1 - \frac{2\rho(1 - 2\rho)}{1 - \rho}.$$

Matches the classical M/G/1 busy-period calculation.

- **M/M/1, constant double thinning** ($q_i^+ \equiv q^+$, $q_i^- \equiv q^-$):

$$\mathcal{D}_\infty = 1 + H(q^-, q^+),$$

where H is the harmonic mean. Thinning births *or* deaths alone keeps $\mathcal{D} = 1$; thinning both breaks Poisson cleanly.

Thank you

Questions?