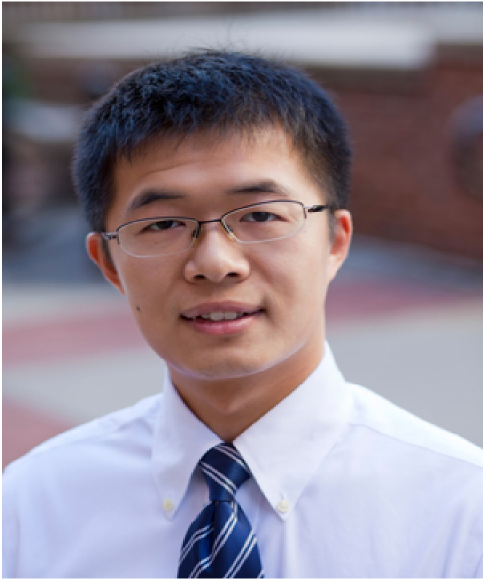
A Robust Queueing Network Analyzer (RQNA)

Based on the Index of Dispersion for Counts (IDC)

Ward Whitt and Wei You, Columbia University

Operations Day, NYU, March 2, 2018

Joint Work with Doctoral Student Wei You



Papers for the Talk

1. W. You, WW, **Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues**, *Operations Research*, online 2017.
2. W. You, WW, **Heavy-Traffic Limit of the GI/GI/1 Stationary Departure Process and its Variance Function**, *Stochastic Systems*, 2018?
3. W. You, WW, **A Robust Queueing Network Analyzer based on Indices of Dispersion**, in preparation.
4. W. You, WW, **Algorithms to Compute the Index of Dispersion of a Stationary Point Process**, in preparation.

Motivating Papers

1. WW, **The Queueing Network Analyzer**, *Bell System Tech. J.* 62, 9 (1983) 2779-2815.
2. K. W. Fendick, WW, **Measurements and approximations to describe offered traffic and predict the average workload in a single-server queue**, *Proc. IEEE* 77, 1 (1989) 171-194.
3. C. Bandi, D. Bertsimas, N. Youssef, **Robust Queueing Theory**, *Operations Research*, 63, 3 (2015) 676-700.

The Workload (Virtual Waiting Time) at One Queue

standard $G/G/1$ reverse-time construction:

Let $Z(t)$ be the workload at time 0, starting empty at time $-t$. Let $A(s)$ count the arrivals over $[-s, 0]$ and index the service times V_k backwards from time 0. Then the input, net-input and workload processes are, respectively,

$$
\begin{aligned}
Y(s) &\equiv \sum_{k=1}^{A(s)} V_k, \quad N(s) \equiv Y(s) - s, \quad s \geq 0, \quad \text{and} \\
Z(t) &\equiv \sup_{0 \leq s \leq t} \{N(s)\}, \quad t \geq 0. \quad \textbf{(a supremum)} \\
&\rightarrow Z \equiv \sup_{s \geq 0} \{N(s)\} \quad \text{as} \quad t \rightarrow \infty \quad \text{(a random variable).}
\end{aligned}
$$

The Stationary Workload with Scaling

For $\{V_k\}$ stationary with $E[V_k] = 1$, $A(t)$ a stationary point process on $\mathbb{R}$ with $E[A(t)] = 1$, and $0 < \rho < 1$, let

$$(A_\rho(s), Y_\rho(s), N_\rho(s)) \equiv (A(\rho s), Y(\rho s), Y(\rho s) - s), \quad s \geq 0,$$

$$Z_\rho \equiv \sup_{s \geq 0} \{N_\rho(s)\}. \quad \text{(a random variable)}$$

robust approximation for $E[Z_\rho]$: (Below we will use $b = \sqrt{2}$.)

$$\mathbf{Z}_\rho^* \equiv \sup_{s \geq 0} \{x : [0, \infty) \to \mathbb{R} : x(s) \leq E[N_\rho(s)] + b\sqrt{Var(N_\rho(s))}\}$$

$$= \sup_{\mathbf{s} \geq \mathbf{0}} \{-(\mathbf{1} - \rho)\mathbf{s} + \mathbf{b}\sqrt{\mathbf{Var}(\mathbf{N}_\rho(\mathbf{s}))}\}$$

for M/G/1: $$= \sup_{s \geq 0} \{-(1 - \rho)s + b\sqrt{\rho s(1 + c_s^2)}\} = \frac{b^2 \rho (1 + c_s^2)}{4(1 - \rho)}.$$

Partially Characterizing Variability Independent of Scale

- **for a nonnegative random variable X:** mean $E[X]$ and scv

$$c_X^2 \equiv \frac{Var(X)}{E[X]^2} \qquad (c_{bX}^2 = c_X^2 \text{ for } b > 0)$$

- **for a stationary point process $A(t)$:** mean and IDC

$$I_c(t) \equiv I_{c,A}(t) \equiv \frac{Var(A(t))}{E[A(t)]} \quad (I_{c,bA}(t) = I_{c,A}(t) \text{ for } b > 0)$$

- **for the input process $Y(t) \equiv \sum_{k=1}^{A(t)} V_k$:** mean and IDW

$$I_w(t) \equiv I_{w,A,V}(t) \equiv \frac{Var(Y(t))}{E[V_k]E[Y(t)]} \ (I_{w,b_1A,b_2V}(t) = I_{w,A,V}(t) \text{ for } b_i > 0)$$

Fendick&WW(1989): Relating the IDW to the Workload

- **normalized mean workload**

$$c_Z^2(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]} = \frac{2(1-\rho)E[Z_\rho]}{\rho}$$

(scaled to have nondegenerate limit as $\rho \downarrow 0$ and as $\rho \uparrow 1$

- **Key Idea:** $\mathbf{c_Z^2(\rho) \approx I_w(t_\rho)}$,

where the time t_ρ might possibly (unresolved) satisfy a **variability fixed-point equation**, e.g. from (15) of KW89,

$$t_\rho = \frac{\rho^2 I_w(t_\rho)}{(1-\rho)^2}.$$

Robust Approximation in terms of the IDW and IDC

robust approximation for $E[Z_\rho]$:

$$
\begin{aligned}
Z_\rho^* &= \sup_{s\geq 0}\left\{-(1-\rho)s+\sqrt{2Var(N_\rho(s))}\right\} \quad (b=\sqrt{2}) \\
&= \sup_{\mathbf{x}\geq \mathbf{0}}\left\{-(\mathbf{1}-\rho)\mathbf{x}/\rho+\sqrt{\mathbf{2xI_w(x)}}\right\} \quad (x\equiv \rho s) \\
&= \frac{\rho v}{2(1-\rho)} \quad \text{for} \quad I_w(x)=v, \quad x\geq 0\,(\text{for some constant } v).
\end{aligned}
$$

For $G/GI/1$ model, the indices of dispersion are related by

$$
I_w(x)=I_c(x)+c_s^2 \quad \text{where } I_c \text{ is IDC of } A(t)\text{, which is 1 if Poisson.}
$$

Hence, we focus on ways to calculate and approximate the IDC.

The Queueing Network Analyzer (QNA)

- WW, **The Queueing Network Analyzer**, *Bell System Tech. J.* 62, 9 (1983) 2779-2815.

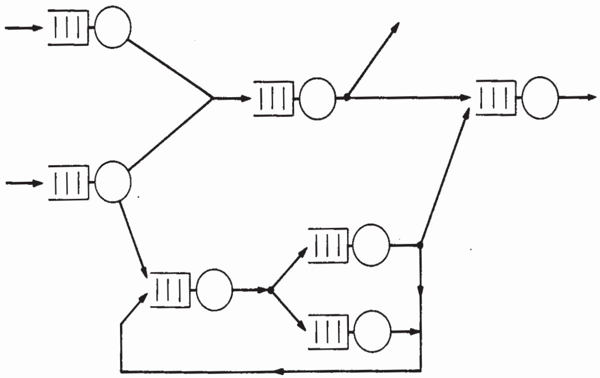
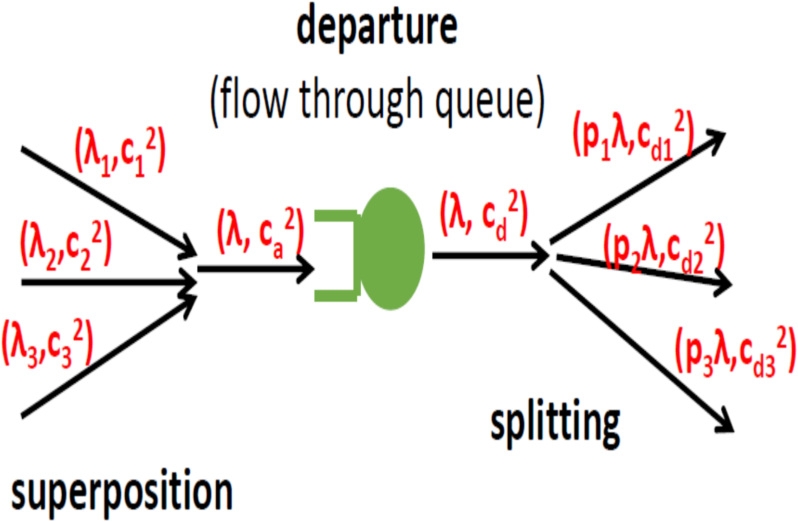


Fig. 1—An open network of queues.

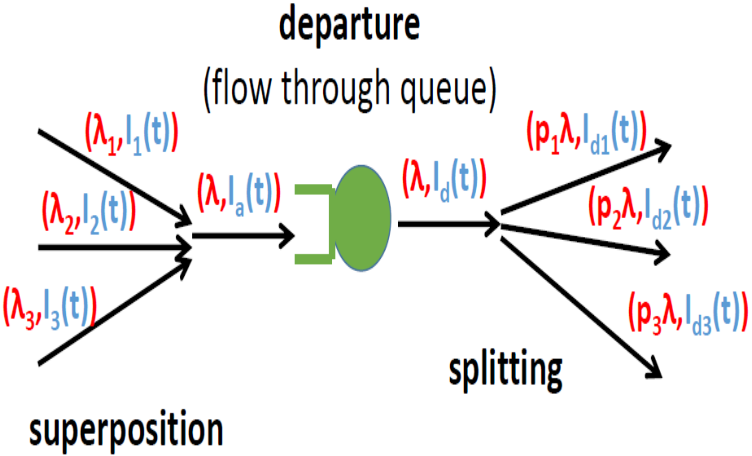
QNA Model Assumptions (restricted)

1. single-server FIFO queues with unlimited waiting space
2. mutually independent exogenous arrival processes, one per queue
3. mutually independent sequences of i.i.d. service times, one per queue
4. Markovian routing (with eventual departure)
5. arrival processes, service times and routing mutually independent
6. service times at queue j have finite mean m_j and scv $c^2_{s,j}$
7. stationary arrival process at queue j with rate $\lambda_{0,j}$
8. arrival process at queue j satisfying a FCLT with Brownian limit
    - arrival processes could be renewal, but need not be.

The Three Network Operations



The Three NEW Network Operations



The Network Operations (two are exact)

1. Superposition of Independent Streams:

$$I_{a,i}(t) = \sum_{j=0}^{k} (\lambda_{a,j,i}/\lambda_i) I_{a,j,i}(t), \quad t \geq 0.$$

2. Independent Splitting

$$I_{a,j,i}(t) = p_{j,i} I_{d,j}(t) + (1 - p_{j,i}), \quad t \geq 0.$$

Approximating the Departure IDC

$$
\begin{aligned}
I_d(t) &\approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t), \quad \text{where} \\
w_\rho(t) &\equiv w^*((1-\rho)^2 \lambda t / \rho c_x^2), \quad t \ge 0, \quad \text{and} \\
w^*(t) &\equiv 1 - \frac{1 - c^*(t)}{2t} \quad \text{for} \quad c^*(t) \equiv cov(R_e(0), R_e(t)) \\
&= \frac{1}{2t}\left((t^2 + 2t - 1)\left(1 - 2\Phi^c(\sqrt{t})\right) + 2\phi(\sqrt{t})\sqrt{t}\,(1+t) - t^2\right)
\end{aligned}
$$

for $c_x^2 \equiv c_a^2 + c_s^2$, $R_e(t)$ stationary canonical (drift -1, variance 1) RBM, Φ is cdf and ϕ pdf of $N(0,1)$.

Based on HT FCLT for stationary departure process from a $GI/GI/1$ queue.

The Departure Process IDC: Comparison with Simulation

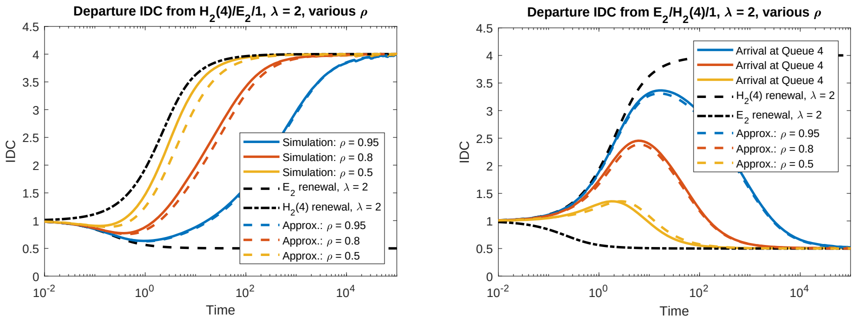


Figure: The departure IDC from $H_2(4)/E_2/1$ (left) and $E_2/H_2(4)/1$ (right) with $\lambda = 2$ and $\rho = 0.5, 0.8, 0.95$ together with reference IDCs for the $H_2(4)$ and E_2 renewal processes, in broken black lines.

Five Queues in Series: Comparison with Simulation

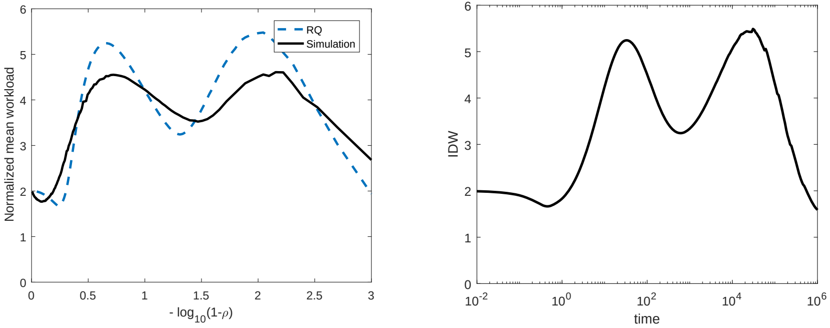


Figure: Simulation estimate of the normalized workload $c_Z^2(\rho)$ at the last queue compared to the RQ approximation $c_{Z^*}^2(\rho)$ (left) and the IDW at the last queue over the interval $[10^{-2}, 10^5]$ in log scale (right).

The Example with Four Internal Modes

There are five queues in series, denoted by

$$E_{10}/H_2(10)/1 \to \cdot/E_{10}/1 \to \cdot/H_2(10)/1 \to \cdot/E_{10}/1 \to \cdot/M/1,$$

where E_{10} is Erlang (sum of 10 i.i.d. exponentials) having scv $1/10$, while $H_2(10)$ is a hyperexponential (mixture of two exponentials) with scv $c^2 = 10$ and balanced means. The traffic intensities decrease:

$$\rho_1 = 0.99 > \rho_2 = 0.98 > \rho_3 = 0.70 > \rho_4 = 0.50.$$

The external arrival rate is set as $\lambda_1 = 1$, so at queue k, $E[V^{(k)}] = \rho_k$. We look at the IDC of the arrival process at the last M queue and the performance there as a function of the mean service time ρ there, $0 < \rho < 1$.

The End

Backup Slides

More References

Partially Characterizing The Variability of Flows

1. H. Heffes, **A class of data traffic proocesses - covariance function characterization and related queueing results** *Bell System Tech. J.* 59, 6 (1980) 997-929.
2. WW, **Approximating a Point Process by a Renewal Process: The View Through a Queue, An Indirect Approach**, *Management Science*, 27, 6 (1981) 619-636.
3. WW, **Approximating a Point Process by a Renewal Process: Two Basic Methods**, *Operations Research*, 30, 1 (1982) 125-147.
4. K. W. Fendick, V. Saksena, WW, **Dependence in packet queues**, *IEEE Trans. Commun.* 37, 11 (1989) 1173-1183.

The Basic Indices of Dispersion: IDC and IDI

1. D. R. Cox, P. A. W. Lewis, **The Statistical Analysis of Series of Events**, Methuen, London, 1966. (Section 4.5)
2. H. Heffes, D. Lucantoni **A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance** *IEEE J. Sel. Areas Commun.* SAC4, 6 (1986) 856-868.
3. K. Sriram, WW, **Characterizing superposition arrival processes in packet multiplexers for voice and data** *IEEE J. Sel. Areas Commun.* SAC4, 6 (1986) 833-846.

The Index of Dispersion for Work: IDW

1. K. W. Fendick, WW, **Measurements and approximations to describe offered traffic and predict the average workload in a single-server queue**, *Proc. IEEE* 77, 1 (1989) 171-194. (Also see references there to work by Heffes, Lucantoni, Neuts, Saksena, Sriram and others.)
2. K. W. Fendick, V. R. Saksena, WW, **Investigating Dependence in Packet Queues with the Index of Dispersion for Work**, *IEEE Transactions on Communications*, 39, 8 (1991) 1231-1244.

Traffic Rate Equations (exact)

$$\lambda_i = \lambda_{o,i} + \sum_{j=1}^{J} \lambda_{j,i} = \lambda_{o,i} + \sum_{i=1}^{J} \lambda_j p_{j,i},$$

Explaining the IDW scaling, I: $M/GI/1$

- **for the input process $Y(t) \equiv \sum_{k=1}^{A(t)} V_k$:** mean and IDW

$$I_w(t) \equiv I_{w,A,V}(t) \equiv \frac{Var(Y(t))}{E[V_k]E[Y(t)]} \qquad (I_{w,b_1A,b_2V}(t) = I_{w,A,V}(t))$$

- random sum, where A is Poisson and independent of i.i.d. $\{V_k\}$:

$$\begin{aligned} E[Y(t)] &= E[\sum_{k=1}^{A(t)} V_k] = E[A(t)]E[V] \\ Var(Y(t)) &= E[A(t)]E[V^2] = E[A(t)]E[V]^2(c_V^2 + 1) \\ I_w(t) &= c_V^2 + 1 = c_V^2 + I_c(t). \end{aligned}$$

Explaining the IDW scaling, II: $G/GI/1$

Assuming that $\{V_k\}$ is i.i.d. and independent of general stationary $A(t)$, by the conditional variance formula,

$$\begin{aligned} Var(Y(t)) &= \lambda t Var(V) + E[V]^2 Var(A(t)) \\ &= \lambda t E[V]^2 c_V^2 + E[V]^2 \lambda t I_{c,A}(t). \end{aligned}$$

By the stationarity, $E[Y(t)] = \lambda E[V]t$ and

$$I_w(t) \equiv \frac{Var(Y(t))}{E[Y(t)]E[V]} = c_V^2 + I_{c,A}(t) \quad (I_{w,b_1A,b_2V}(t) = I_{w,A,V}(t))$$

Explaining the IDW scaling, III (i): FCLT for random sums

Let random elements in the function space D^2 be defined for the partial sums on interarrival and service times by

$$\left(\hat{\mathbf{S}}_n^a(t), \hat{\mathbf{S}}_n^s(t)\right) \equiv n^{-1/2}\left(\left[S_{\lfloor nt \rfloor}^a - \lambda^{-1} nt\right], \left[S_{\lfloor nt \rfloor}^s - mnt\right]\right), \quad t \geq 0.$$

As in Donsker's theorem (Thm 4.3.2 of WW02), we assume that

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s\right) \Rightarrow (\sigma_a B_a, \sigma_s B_s) = \left(\lambda^{-1} c_a B_a, m c_s B_s\right) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty,$$

where B_a and B_s are (possibly dependent) standard BMs.

Explaining the IDW scaling, III (ii): FCLT for random sums

Let random elements in the function space D^2 be defined by

$$\left(\hat{\mathbf{N}}_n(t), \hat{\mathbf{Y}}_n(t)\right) \equiv n^{-1/2}\left([N(nt) - \lambda nt], [Y(nt) - \lambda mnt]\right), \quad t \geq 0.$$

Then, by Corollaries 7.3.1 and 13.3.2 in WW02,

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s, \hat{\mathbf{N}}_n, \hat{\mathbf{Y}}_n\right) \Rightarrow \left(\lambda^{-1} c_a B_a, m c_s B_s, \sqrt{\lambda} c_s B_a, \sqrt{\lambda} m (c_a B_a + c_s B_s)\right)$$

in D^4 as $n \to \infty$ for B_a and B_s above.

Explaining the IDW scaling, III (iii): random sums

Under associated uniform integrability, as $n \to \infty$,

$$
\begin{aligned}
Var(\hat{Y}_n(t)) &\to \lambda m^2 Var(c_a B_a(t) + c_s B_s(t)) \\
&= \lambda m^2 t(c_a^2 + c_s^2 + 2t^{-1} c_a c_s Cov(B_a(t), B_s(t))) \\
\text{so} \quad \frac{Var(\hat{Y}_n(t))}{\lambda m^2 t} &\to c_a^2 + c_s^2 + 2t^{-1} c_a c_s Cov(B_a(t), B_s(t)),
\end{aligned}
$$

which is independent of λ and m. Thus, in a stationary setting,

$I_{w,n}(t) \to I_w(t)$, where $I_{w,b_1 A, b_2 V}(t) = I_{w,A,V}(t)$ for $b_i > 0$, $i = 1, 2$.