A Robust Queueing Network Analyzer Based on Indices of Dispersion

Wei You
(joint work with Ward Whitt)

Columbia University

INFORMS 2018, Phoenix

November 6, 2018

Motivation

- Many complex service systems can be modeled as open queueing networks (OQN)
- The estimation of performance measures
  - important in many applications;
  - theoretical analysis is limited;
  - approximation remains an important tool.
- In this work we propose a fast and accurate Robust Queueing Network Analyzer (RQNA) to approximation performance measures in single-server OQNs.

Background - Previous Approximation Algorithms

Decomposition approximation methods

- Motivated by product-form solutions of Jackson Networks.
- Treat stations as independent single-server queues.

Examples

- The Queueing Network Analyzer (QNA) by Whitt (1983),
  - approximates each station by a GI/GI/1 queue.
- Kim (2011a, 2011b)
  - approximates each station by an MMPP(2)/GI/1 queue (Markov-Modulated Poisson Process);

Background - Previous Approximation Algorithms

Approximations using Reflected Brownian Motion (RBM)

- Approximate the steady-state queue length distribution by the stationary distribution of the limiting RBM;
- numerically calculate the steady-state mean of the RBM.

Examples

- QNET by Harrison and Nguyen (1990) for OQNs and by Dai and Harrison (1993) for CQNs;
- Sequential bottleneck decomposition (SBD) by Dai, Nguyen and Reiman (1994).

Background - Recent Developments

Recent Developments

- Interpolation method (IR) by Wu and McGinnis (2014).
- (Parametric) Robust Queueing (RQ) by Bandi et al. (2015).
- (Non-parametric) RQ by Whitt and You (2018a).

In this talk,

- non-parametric Robust Queueing Network Analyzer (RQNA) for OQNs.

Dependence in Queues

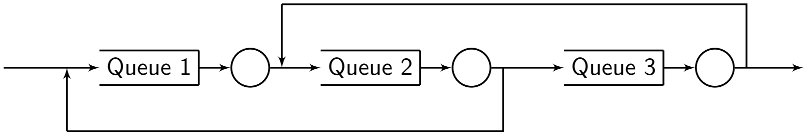


Figure: A three-station example.

Dependence rises naturally in queueing network:

- Dependence within/between the flows[1]:
  - introduced by departure, splitting and superposition;
  - also by customer feedback.

[1] arrival processes, departure process, etc.

Dependence in Queues

Dependence has significant impact on performance measures

- Dependence can have complicated temporal structure.
- The **level of impact** will depend on both the temporal structure and the traffic intensity.
- Parametric methods (QNA, QNET, parametric RQ) using first two moments to describe variability may fail.

3 Stations with Feedback

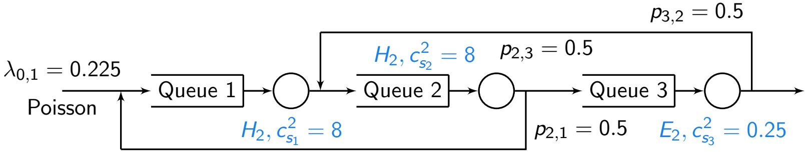


Table: The steady-state mean waiting time.

r = 0.5, (third parameter of H2 dist, weight on one mean)				
Queue	ρ	Simu	QNET	SBD
1	0.9	31.22	35.9 (15%)	26.0 (-17%)
2	0.675	8.32	10.2 (23%)	11.1 (33%)
3	0.45	2.00	1.89 (5.5%)	1.94 (3%)
Total		138.7	161.3 (16%)	135.3 (-2.5%)

r = 0.99, (third parameter of H2 dist, weight on one mean)				
Queue	ρ	Simu	QNET	SBD
1	0.9	27.67	35.9 (30%)	26.0 (-6.0%)
2	0.675	2.67	10.2 (282%)	11.1 (316%)
3	0.45	0.56	1.89 (236%)	1.94 (245%)
Total		103.8	161.3 (55%)	135.3 (30%)

Indices of Dispersion for Counts (IDC)

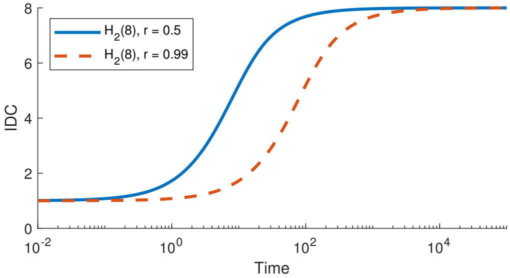
Indices of dispersion can describe the temporal structure.

- Fendick and Whitt (1989) first applied it to queueing approximation.

Definition from Cox and Lewis (1966)

$$I_a(t) \equiv Var(A(t))/E[A(t)], \quad t \geq 0,$$

where $A(t)$ is any stationary point process.



Indices of Dispersion for Counts (IDC)

Theorem (renewal process characterization theorem)

A renewal process $A(t)$ with positive rate λ is fully characterized by the IDC of its equilibrium (stationary) version $A_e(t)$:

$$I_a(t) \equiv Var(A_e(t))/E[A_e(t)].$$

- RQ-IDC, and so RQNA-IDC, utilize much more information of the underlying distribution;
- potentially more accurate and adaptive to complex distributions.

Robust Queueing for Single-Server Queues

- Let Z be the steady-state mean workload (virtual waiting time) of a single-server queue.

RQ for the workload in Whitt and You (2018a)

$$Z \approx Z^* \equiv \sup_{N \in \mathcal{U}} \sup_{0 \leq s \leq \infty} \{N(s)\},$$

where

$$\mathcal{U} = \left\{ N : N(s) \leq -(1-\rho)s + \sqrt{2\rho s (I_a(s) + c_s^2)/\mu},\ s \geq 0 \right\}.$$

and $I_a(t)$ is the IDC of the arrival process.

Robust Queueing for continuous-time workload

- Let Z be the steady-state mean workload (virtual waiting time) of a single-server queue.

The RQ-IDC algorithm

$$Z \approx Z^* = \sup_{0 \leq s \leq \infty} \left\{ -(1-\rho)s + \sqrt{2\rho s (I_a(s) + c_s^2)/\mu} \right\}.$$

- RQ-IDC converts the arrival IDC and the squared coefficient of variation (scv) of the service distribution into an approximation of the steady-state mean workload.
- I_a is defined for the stationary arrival process;

Generalization to RQNA

To extend RQ to RQNA, we need the IDC of the total arrival process

- for external flows, i.e., service processes and external arrival processes
  - **calculated** in special cases[2] (e.g. renewal process);
  - **estimated** by simulation or from data;
- for internal flows, i.e., internal arrival processes and departure processes.
  - **approximated** by RQNA.

[2] by numerically inverting the Laplace Transform

Generalization to RQNA: Internal Flows

The total arrival process at any queue:

- **superposition** of external arrival and **splitting** of **departure** processes.

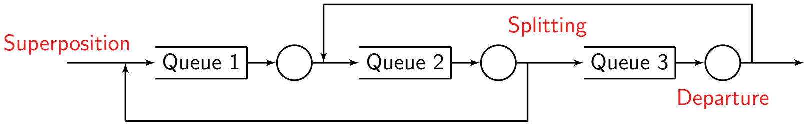


Figure: A three-station example.

The IDC Equations

Notations

- $I_{a,i}$: IDC of the total arrival process at station i;
- $I_{s,i}$: IDC of the service process at station i;
- $I_{d,i}$: IDC of the total departure process at station i;

The Departure Equation

$$I_{d,i}(t) \approx w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(t), \qquad \text{(Dep)}$$

where w_i is a weight function with explicit expression.

- Departure IDC is a convex combination;
- Supported by Heavy-traffic (HT) limit for the **stationary departure process** $\Rightarrow$ asymptotically exact.

The IDC Equations

One more notation

- $I_{a,i,j}$: IDC of the flow from station i to station j;

The **Splitting** and **Superposition** Equation

$$I_{a,i,j}(t) \approx p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t) \qquad \text{(Spl)}$$

$$I_{a,i}(t) \approx \sum_{j=0}^{K} (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t) + \beta_i(t) \qquad \text{(Sup)}$$

where $\alpha_{i,j}$ and β_i are correction term with explicit expression and $\lambda_{j,i} = p_{j,i}\lambda_j$ is the rate of the flow from i to j.

- Red terms recovers **independent** splitting or superposition.
- Blue term models **dependence** in the splitting or superposition operation.
- Supported by Heavy-traffic (HT) limit for the **stationary flows** in OQN.

The IDC Equations

In summary, the IDC equations are

$$I_{d,i}(t) = w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(\rho t), \qquad \text{(Dep)}$$

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t), \qquad \text{(Spl)}$$

$$I_{a,i}(t) = \sum_{j=0}^{K} (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t) + \beta_i(t). \qquad \text{(Sup)}$$

In matrix notation, we have

$$\mathbf{I}(t) = \mathbf{M}(t)\mathbf{I}(t) + \mathbf{b}(t).$$

- For each fixed t, the IDC equations form a system of linear equations;
- The IDC equations have unique solution if every customer eventually leave the system.

3 Stations with Feedback

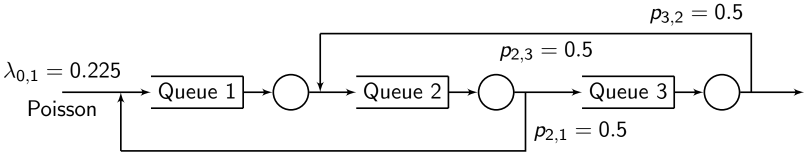


Figure: A three-station example.

Table: Traffic intensity.

Case	ρ_1	ρ_2	ρ_3
1	0.675	0.900	0.450
2	0.900	0.675	0.900
3	0.900	0.675	0.450
4	0.900	0.675	0.675

Table: Variability (squared coefficient of variation, scv) of service-time distributions.

Case	$c^2_{s,1}$	$c^2_{s,2}$	$c^2_{s,3}$
A	0.00	0.00	0.00
B	2.25	0.00	0.25
C	0.25	0.25	2.25
D	0.00	2.25	2.25
E	8.00	8.00	0.25

3 Stations with Feedback

Table: A comparison of four approximation methods to simulation for the total sojourn time in the three-station example.

Case		Simu	QNA	QNET	SBD	RQNA
A	1	40.39	20.5 (-49%)	diverging	43.0 (6.4%)	44.8 (11.0%)
	2	59.58	36.0 (-40%)	56.7 (-4.9%)	58.2 (-2.4%)	69.3 (16.4%)
	3	40.72	24.0 (-41%)	38.7 (-5.0%)	40.2 (-1.3%)	43.3 (6.3%)
	4	42.12	26.2 (-38%)	41.8 (-0.7%)	42.7 (1.3%)	41.2 (-2.2%)
B	1	52.40	42.0 (-20%)	52.6 (0.4%)	50.2 (-4.2%)	53.1 (1.4%)
	2	91.52	94.1 (2.8%)	83.7 (-8.5%)	95.3 (4.1%)	94.5 (3.2%)
	3	61.68	72.2 (17%)	61.9 (0.4%)	60.9 (-1.3%)	60.5 (-1.9%)
	4	63.34	75.8 (20%)	64.1 (1.3%)	64.7 (2.1%)	62.4 (-1.4%)
C	1	44.24	31.3 (-29%)	37.0 (-16%)	47.1 (6.4%)	42.1 (-4.8%)
	2	92.42	87.4 (-5.4%)	91.2 (-1.4%)	91.6 (-0.8%)	96.0 (3.8%)
	3	44.26	33.2 (-25%)	44.0 (-0.7%)	45.0 (1.7%)	44.0 (-0.6%)
	4	50.20	41.4 (-18%)	51.1 (1.7%)	52.2 (4.0%)	45.9 (-8.6%)
E	1	134.4	265 (97%)	155 (15%)	116 (-14%)	120 (-11%)
	2	213.1	308 (45%)	228 (7.1%)	206 (-3.3%)	173 (-19%)
	3	138.7	244 (76%)	161 (16%)	135 (-2.5%)	136 (-2.0%)
	4	155.1	252 (63%)	168 (8.2%)	147 (-5.0%)	148 (-4.8%)

3 Stations with Feedback

- Case E3:

$$(\rho_1, \rho_2 . \rho_3) = (0.9, 0.675, 0.45)$$
$$(c_{s_1}^2, c_{s_2}^2 . c_{s_3}^2) = (8, 8, 0.25)$$

Table: A comparison of six approximation methods to simulation for the sojourn time at each station of the three-station example.

Case E3, r = 0.5				
Queue	Simu	QNET	SBD	RQNA
1	31.22	35.9 (15%)	26.0 (-17%)	26.0 (-17%)
2	8.32	10.2 (23%)	11.1 (33%)	11.8 (42%)
3	2.00	1.89 (5.5%)	1.94 (3%)	0.93 (-54%)
Sum	138.7	161.3 (16%)	135.3 (-2.5%)	136.1 (-1.9%)

Case E3, r = 0.99				
Queue	Simu	QNET	SBD	RQNA
1	27.67	35.9 (30%)	26.0 (-6.0%)	26.0 (-6.0%)
2	2.67	10.2 (282%)	11.1 (316%)	6.03 (125%)
3	0.56	1.89 (236%)	1.94 (245%)	0.50 (-11%)
Sum	103.8	161.3 (55%)	135.3 (30%)	112.1 (8%)

Thanks!

References

References on Robust Queueing:

[BBY15] C. Bandi, D. Bertsimas, and N. Youssef, Robust Queueing Theory, *Operations Research*, 2015.

[WY18a] W. Whitt, W. You, Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues, *Operations Research*, 2018.

[WY18c] W. Whitt, W. You, A Robust Queueing Network Analyzer Based on Indices of Dispersion, submitted to INFORMS Journal on Computing, 2018.

[WY18e] W. Whitt and W. You, The Advantage of Indices of Dispersion in Queueing Approximations, submitted to *Operations Research Letters*, 2018.

[WY17] W. Whitt, W. You, Time-Varying Robust Queueing, submitted to *Operations Research*, 2017.

References on queueing network approximations:

[DH93] J. G. Dai and J. M. Harrison, The QNET method for two-moment analysis of closed manufacturing systems, *Annals of Applied Probability*, 1993.

[FW89] K. W. Fendick, W. Whitt, Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue, *Proceedings of the IEEE*, 1989.

[HN90] J. M. Harrison, V. Nguyen, The QNET Method for Two-Moment Analysis of Open Queueing Networks, *Queueing Systems*, 1990.

[SW86] K. Sriram, W. Whitt, Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data, *IEEE Journal on Selected Areas on Communications*, 1986.

[SW90] S. Suresh, W. Whitt, The Heavy-Traffic Bottleneck Phenomenon in Open Queueing Networks, *Operations Research Letters*, 1990.

[WM12] K. Wu, L. McGinnis, Interpolation Approximations for Queues in Series, *IIE Transactions*, 2012.

[WW82] W. Whitt, Approximating a Point Process by a Renewal Process: Two Basic Methods, *Operations Research*, 1982.

[WW83] W. Whitt, The Queueing Network Analyzer, *Bell System Technical Journal*, 1983.

[ZHS05] Q. Zhang, A. Heindl, E. Smirni, Characterizing the BMAP/MAP/1 Departure Process via the ETAQA Truncation, *Stochastic Models*, 2005.

References

References on HT limits:

[GZ06] D. Gamarnik, A. Zeevi, Validity of heavy traffic steady-state approximations in generalized Jackson Networks, *The Annals of Applied Probability*, 2006.

[IW70] D.L. Iglehart, W. Whitt, Multiple Channel Queues in Heavy Traffic II: Sequences, Networks and Batches. *Advanced Applied Probability*, 1970.

[Loy62] R. M. Loynes, The Stability of A Queue with Non-independent Inter-arrival and Service Times, *Mathematical Proceedings of the Cambridge Philosophical Society*, 1962.

[WY18b] W. Whitt, W. You, Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function, *Stochastic Systems*, 2018.

[WY18d] W. Whitt and W. You, Heavy Traffic Limits for the Stationary Flows in Generalized Jackson Networks, submitted to *Stochastic Systems*, 2018.

References on departure processes:

[D76] D. Daley, Queueing Output Processes, *Advances in Applied Probability*, 1976.

[B56] P. Burke, The Output of a Queuing System, *Operations Research*, 1956.

[G99] D. Green, Departure Processes from MAP/PH/1 Queues, thesis, 1999.

[T62] L. Takács, Introduction to the Theory of Queues, *Oxford University Press*, 1962.

[HKNT13] S. Hautphenne, Y. Kerner, Y. Nazarathy, P. Taylor, The Second Order Terms of the Variance Curves for Some Queueing Output Processes, `arXiv:1311.0069`, 2013.

[W84] W. Whitt, Approximations for Departure Processes and Queues in Series, *Naval Research Logistics Quarterly*, 1984.

Other Performance Measures

$$Z_\rho^* = \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + \sqrt{2\rho s I_w(s)/\mu} \right\}.$$

This RQ formulation give approximation of the mean steady-state workload. For other performance measures, we have

- Mean steady-state waiting time:

$$E[W] \approx \max\{0, Z^*/\rho - (c_s^2 + 1)/2\mu\}.$$

  - obtained by Brumelle's formula:

$$E[Z] = \rho E[W] + \rho \frac{E[V^2]}{2\mu} = \rho E[W] + \rho \frac{(c_s^2 + 1)}{2\mu}.$$

- Mean steady-state queue length, by Little's law,

$$E[Q] = \lambda E[W] = \rho E[W].$$

3 Stations with Feedback

Table: A comparison of six approximation methods to simulation for the sojourn time at each station of the three-station example.

Case D1, r = 0.5					
Queue	Simu	QNA	QNET	SBD	RQNA
1	1.478	1.24 (-16%)	1.48 (0.1%)	1.47 (-0.5%)	1.69 (14%)
2	10.22	13.9 (36%)	10.6 (3.7%)	10.4 (1.8%)	10.4 (1.8%)
3	1.563	1.53 (-2.1%)	1.54 (-1.5%)	1.59 (1.7%)	1.53 (-2.1%)
Sum	57.42	71.4 (24%)	58.8 (2.4%)	58.2 (1.4%)	58.7 (2.2%)
Case D1, r = 0.99					
Queue	Simu	QNA	QNET	SBD	RQNA
1	1.145	1.24 (8.3%)	1.48 (29%)	1.47 (28%)	1.28 (12%)
2	10.15	13.9 (37%)	10.6 (4.4%)	10.4 (2.5%)	10.4 (2.5%)
3	1.119	1.53 (37%)	1.54 (38%)	1.59 (42%)	1.28 (14%)
Sum	55.26	71.4 (29%)	58.8 (6.4%)	58.2 (5.3%)	57.0 (3.1%)

The Heavy-Traffic Bottleneck Phenomenon

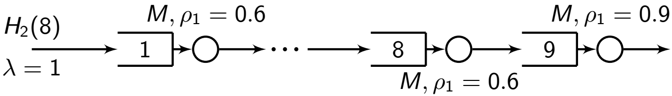
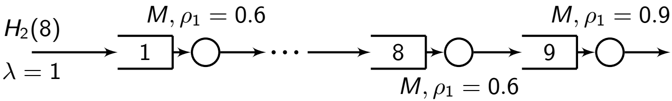


Figure: The heavy-traffic bottleneck example in Suresh and Whitt (1990).

Arrival Process		$H_2, c_a^2 = 8$ $r = 0.5$	$H_2, c_a^2 = 8$ $r = 0.95$
Queue 8	Simulation	1.44	0.92
	M/M/1	0.90 (-38%)	0.90 (-2.1%)
	QNA	1.04 (-28%)	1.04 (13%)
	SBD	1.01 (-29%)	1.01 (10%)
Queue 9	Simulation	29.15	8.94
	M/M/1	8.1 (-72%)	8.1 (-9.4%)
	QNA	8.9 (-69%)	8.9 (-0.4%)
	SBD	36.5 (25%)	36.5 (308%)

Table: Mean steady-state waiting times at Queue 8 and 9, compared with M/M/1 values and approximations.

The Heavy-traffic Bottleneck Phenomenon



Arrival Process		$H_2, c_a^2 = 8$ $r = 0.5$	$H_2, c_a^2 = 8$ $r = 0.99$
Queue 8	Simulation	1.44	0.92
	M/M/1	0.90 (-38%)	0.90 (-2.1%)
	QNA	1.04 (-28%)	1.04 (13%)
	SBD	1.01 (-29%)	1.01 (10%)
	IR	1.20 (-17%)	1.20 (7.1%)
	RQ	1.27 (-12%)	0.92 (-0.5%)
Queue 9	Simulation	29.15	8.94
	M/M/1	8.1 (-72%)	8.1 (-9.4%)
	QNA	8.9 (-69%)	8.9 (-0.4%)
	SBD	36.5 (25%)	36.5 (308%)
	IR	21.1 (-28%)	21.1 (136%)
	RQ	37.0 (27%)	16.5 (84%)

10 Stations with Feedback

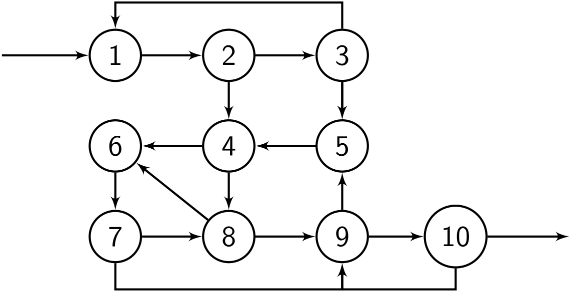


Figure: A ten-station with customer feedback example.

- The traffic intensity vector is $(0.6, 0.4, 0.6, 0.9, 0.9, 0.6, 0.4, 0.6, 0.6, 0.4)$.
- The scv's at these stations are $(0.5, 2, 2, 0.25, 0.25, 2, 1, 2, 0.5, 0.5)$

10 Stations with Feedback

Table: A comparison of five approximation methods to simulation for the mean steady-state sojourn times at each station.

Q	Simu	QNA	QNET	SBD	RQ	RQNA
1	0.99	0.97 (-2.8%)	1.00 (0.2%)	1.00 (0.4%)	0.97 (-2.0%)	1.00 (0.4%)
2	0.55	0.58 (6.0%)	0.56 (2.6%)	0.55 (0.2%)	0.55 (-0.1%)	0.56 (1.4%)
3	2.82	2.93 (4.2%)	2.90 (3.2%)	2.76 (-2.0%)	2.96 (5.0%)	2.75 (-2.5%)
4	1.79	1.34 (-25%)	1.41 (-21%)	1.76 (-1.6%)	2.34 (31%)	2.11 (18%)
5	2.92	2.49 (-15%)	2.44 (-17%)	2.81 (-3.6%)	3.77 (29%)	3.35 (15%)
6	0.58	0.64 (10%)	0.62 (7.4%)	0.59 (2.2%)	0.60 (3.8%)	0.49 (-16%)
7	0.24	0.24 (-1.7%)	0.26 (7.1%)	0.27 (11%)	0.23 (-3.0%)	0.24 (-1.3%)
8	0.58	0.64 (9.6%)	0.61 (4.6%)	0.60 (1.7%)	0.61 (3.9%)	0.59 (0.6%)
9	0.34	0.32 (-6.1%)	0.35 (2.0%)	0.43 (26%)	0.33 (-4.2%)	0.42 (21%)
10	0.29	0.30 (2.4%)	0.29 (1.4%)	0.28 (-1.7%)	0.28 (-1.5%)	0.26 (-8.7%)
sum	22.0	20.3 (-7.9%)	20.4 (-7.3%)	22.4 (1.7%)	26.1 (18%)	24.2 (9.9%)

Feedback Elimination

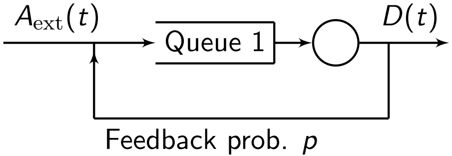


Figure: Immediate feedback.

- Normally, the immediate feedback returns the customer back to the end of the line at the same station.
- In the immediate feedback elimination procedure, the approximation step is to put the customer back at the head of the line.
  - The overall service time is then a geometric sum of the original service times.
- This does not alter the queue length process or the workload.

Feedback Elimination

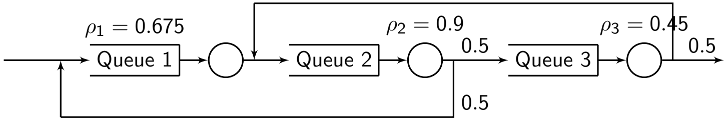


Figure: A three-station example.

For the general case,

- Near immediate feedback is defined as a feedback customer that does not go through a station with higher traffic intensity than the current station.
- For each station with feedback, we eliminate all near immediate feedback flows, the nadjust the service times just as in the single-station case.