

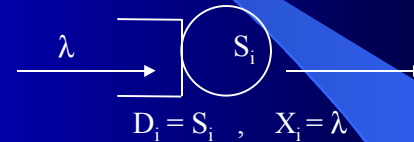
Ch.34 Mean-Value Analysis

- Analysis of QNMs (Queueing Network Models)
 - Operational Laws
 - Asymptotic Bounds
 - Balanced Job Bounds
 - Mean value Analysis
- MVA (Mean-value Analysis)
 - Single service center with open arrivals (M/M/1)
 - Open QNM
 - Closed QNM

Single Service Center (M/M/1)

- Derived Quantities

- $U_i(\lambda) = \lambda D_i = \lambda S_i$



- What will be the response time?

- $R_i(\lambda) = ???$

- Given $R_i(\lambda)$

- number of jobs in the system $Q_i(\lambda) = \lambda R_i(\lambda)$ (Little's Law)

- Given $Q_i(\lambda)$

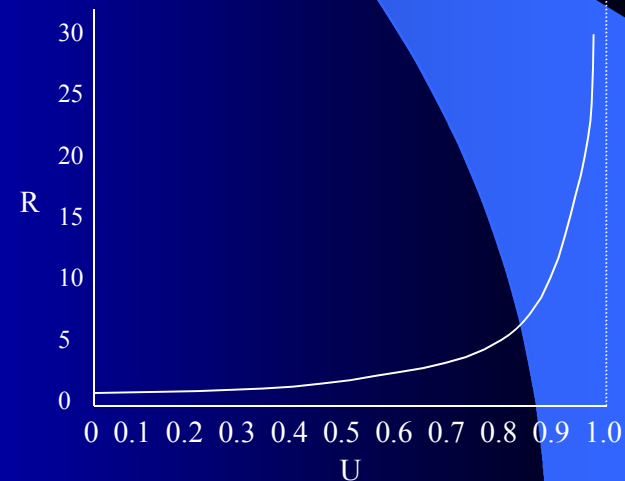
- $R_i(\lambda) = \text{my service} + \text{queueing delay (service time for jobs ahead of me)}$
 $= S_i + Q_i(\lambda)S_i$

Single Service Center (cont'd)

- Response time is
 - $R_i(\lambda) = S_i + Q_i(\lambda) S_i$
 $= S_i + \lambda R_i(\lambda) S_i$
 $\Rightarrow R_i(\lambda) = S_i / (1 - \lambda S_i)$
 $= S_i / (1 - U_i(\lambda))$

- Queue length
(including one in service)
 - $Q_i(\lambda) = \lambda R_i(\lambda)$
 $= \lambda(S_i + Q_i(\lambda)S_i)$
 $\Rightarrow Q_i(\lambda) = \lambda S_i / (1 - \lambda S_i)$
 $= U_i(\lambda) / (1 - U_i(\lambda))$

* $N = \rho / (1 - \rho)$



Open QNM

- On arrival at the i -th device,
 - The job sees Q_i jobs ahead (including the one in service)
 - Device queue length : $Q_i(\lambda) = U_i(\lambda)/(1-U_i(\lambda))$
 - Device response time : $R_i(\lambda) = S_i/(1-U_i(\lambda))$
 - For delay centers,
 - $R_i(\lambda) = S_i$
 - $Q_i(\lambda) = R_i X_i = S_i X V_i = X D_i = U_i(\lambda) < 1$
- Entire system
 - $N = Q_1 + Q_2 + \dots + Q_M$
 $= X_1 R_1 + X_2 R_2 + \dots + X_M R_M = X R$
 $\Rightarrow R = V_1 R_1 + V_2 R_2 + \dots + V_M R_M$

MVA for Closed QNM

- Basic relation

Why not N ? - consider when there is only one customer

- $R_i(N) = \text{my service} + \text{queueing delay (service time for jobs ahead of me)}$
 $= S_i + Q_i(N-1)S_i$
- Then, how to get $Q_i(N-1)$?

- Recursive algorithm

- Set $Q_i(0)=0$ for all devices (i)
- $R_i(1) = S_i + Q_i(0)S_i$
 $R(1) = V_1R_1(1) + V_2R_2(1) + \dots + V_MR_M(1)$
 $X(1) = N/[R(N)+Z] = 1/[R(1)+Z]$
 $Q_i(1) = X_i(1)R_i(1) = X(1)V_iR_i(1)$
- $R_i(2) = S_i + Q_i(1)S_i$

and repeat the above calculations until the desired population N

MVA for Closed QNM (cont'd)

- Question: CPU, Disk A & Disk B

- $D_{CPU} = 2, D_A = 3, D_B = 1$

- $N = 0$

- $Q_{CPU} = Q_A = Q_B = 0$

- $N = 1$

- $R_{CPU} = S_{CPU}(1 + Q_{CPU}) = 0.125(1 + 0) = 0.125$

- $R_A = S_A(1 + Q_A) = 0.3(1 + 0) = 0.3$

- $R_B = S_B(1 + Q_B) = 0.2(1 + 0) = 0.2$

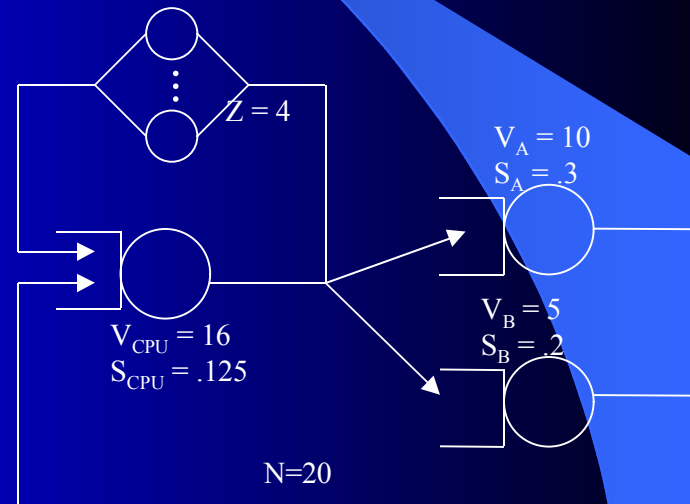
- $R = R_{CPU}V_{CPU} + R_A V_A + R_B V_B$
 $= 0.125 * 16 + 0.3 * 10 + 0.2 * 5 = 6$

- $X = N / (R + Z) = 1 / (6 + 4) = 0.1$

- $Q_{CPU} = X R_{CPU} V_{CPU} = 0.1 * 0.125 * 16 = 0.2$

- $Q_A = X R_A V_A = 0.1 * 0.3 * 10 = 0.3$

- $Q_B = X R_B V_B = 0.1 * 0.2 * 5 = 0.1$



MVA for Closed QNM (cont'd)

- Question: (continued)

- N=2

- $R_{\text{CPU}} = S_{\text{CPU}}(1 + Q_{\text{CPU}}) = 0.125(1 + 0.2) = 0.15$
 - $R_A = S_A(1 + Q_A) = 0.3(1 + 0.3) = 0.39$
 - $R_B = S_B(1 + Q_B) = 0.2(1 + 0.1) = 0.22$
 - $R = R_{\text{CPU}}V_{\text{CPU}} + R_A V_A + R_B V_B = 0.15 * 16 + 0.39 * 10 + 0.22 * 5 = 7.4$
 - $X = N / (R + Z) = 2 / (7.4 + 4) = 0.175$
 - $Q_{\text{CPU}} = X R_{\text{CPU}} V_{\text{CPU}} = 0.175 * 0.15 * 16 = 0.421$
 - $Q_A = X R_A V_A = 0.175 * 0.39 * 10 = 0.684$
 - $Q_B = X R_B V_B = 0.175 * 0.22 * 5 = 0.193$

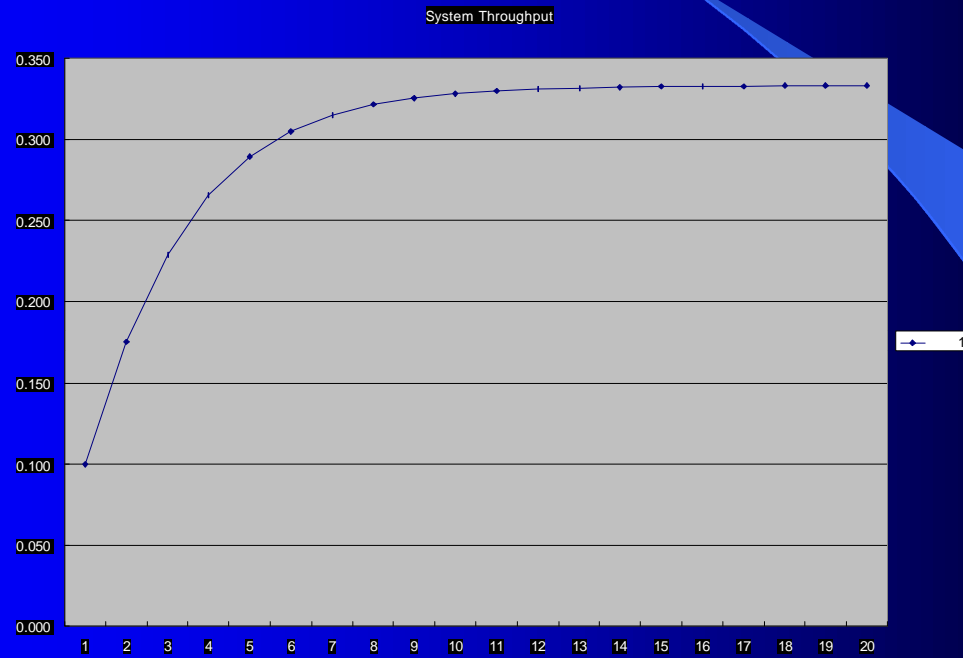
- N=3

-

MVA for Closed QNM (cont'd)

N	Response Time			System	Xput	Queue Length		
	CPU	Disk A	Disk B			CPU	Disk A	Disk B
0						0.000	0.000	0.000
1	0.125	0.300	0.200	6.000	0.100	0.200	0.300	0.100
2	0.150	0.390	0.220	7.400	0.175	0.421	0.684	0.193
3	0.178	0.505	0.239	9.088	0.229	0.651	1.158	0.273
4	0.206	0.647	0.255	11.051	0.266	0.878	1.721	0.338
5	0.235	0.816	0.268	13.256	0.290	1.088	2.365	0.388
6	0.261	1.009	0.278	15.659	0.305	1.275	3.081	0.424
7	0.284	1.224	0.285	18.216	0.315	1.433	3.858	0.449
8	0.304	1.457	0.290	20.888	0.321	1.564	4.684	0.466
9	0.321	1.705	0.293	23.647	0.326	1.670	5.551	0.477
10	0.334	1.965	0.295	26.470	0.328	1.752	6.450	0.485
11	0.344	2.235	0.297	29.340	0.330	1.816	7.374	0.490
12	0.352	2.512	0.298	32.245	0.331	1.865	8.318	0.493
13	0.358	2.795	0.299	35.176	0.332	1.901	9.276	0.496
14	0.363	3.083	0.299	38.126	0.332	1.928	10.245	0.497
15	0.366	3.374	0.299	41.089	0.333	1.948	11.223	0.498
16	0.369	3.667	0.300	44.064	0.333	1.963	12.207	0.499
17	0.370	3.962	0.300	47.045	0.333	1.974	13.195	0.499
18	0.372	4.259	0.300	50.032	0.333	1.981	14.187	0.499
19	0.373	4.556	0.300	53.022	0.333	1.987	15.181	0.500
20	0.373	4.854	0.300	56.016	0.333	1.991	16.177	0.500

MVA for Closed QNM (cont'd)



Approximate MVA

- MVA (Mean Value Analysis)
 - Applicable only if the network is a product form queueing network
=> it must satisfy the assumptions (Denning and Buzen)
: job flow balance, one-step behavior, device homogeneity
 - A recursive algorithm, which is time consuming : $O(MN)$
 $R(0) \rightarrow R(1) \rightarrow \dots \rightarrow R(N)$
- Approximate MVA
 - Schweitzer's approximation (1979)
 $R(N-1) \rightarrow R(N) \rightarrow R(N-1) \rightarrow R(N) \dots$ until converges
 - Assume that the queue length (at each device) increases proportionally as # jobs increases (in the network) => $Q_i(N)/N=c$ (constant for all N)
 $Q_i(N-1)/(N-1) = Q_i(N)/N$ or $Q_i(N-1)=Q_i(N) (N-1)/N$

Approximate MVA (cont'd)

- Approximate MVA : Schweitzer's approximation

- $Q_i(N-1) = Q_i(N) (N-1)/N$

- MVA algorithm can written as

- $R_i(N) = S_i(1+Q_i(N-1))$

- $= S_i(1+ Q_i(N) (N-1)/N)$

- $X(N) = N/(Z+R)$

- $= N/(Z+V_1R_1(N)+V_2R_2(N)+ \dots +V_MR_M(N))$

- $Q_i(N)=X(N)V_iR_i(N)$

- Calculate new $Q_i(N)$ repeatedly, and stop when it converges with the initial guess values for $Q_i = N/M$ for all i ($i=1,2,\dots,M$)

Approximate MVA (cont'd)

- Question: CPU, Disk A & Disk B

- Iteration 0

- $Q_{CPU} = Q_A = Q_B = N/M = 20/3 = 6.667$

- Iteration 1

- $R_{CPU} = S_{CPU}(1 + Q_{CPU}) = 0.125(1 + 6.667) = 0.958$

- $R_A = S_A(1 + Q_A) = 0.3(1 + 6.667) = 2.300$

- $R_B = S_B(1 + Q_B) = 0.2(1 + 6.667) = 1.533$

- $R = R_{CPU}V_{CPU} + R_A V_A + R_B V_B$
 $= 0.958 * 16 + 2.300 * 10 + 1.533 * 5 = 46$

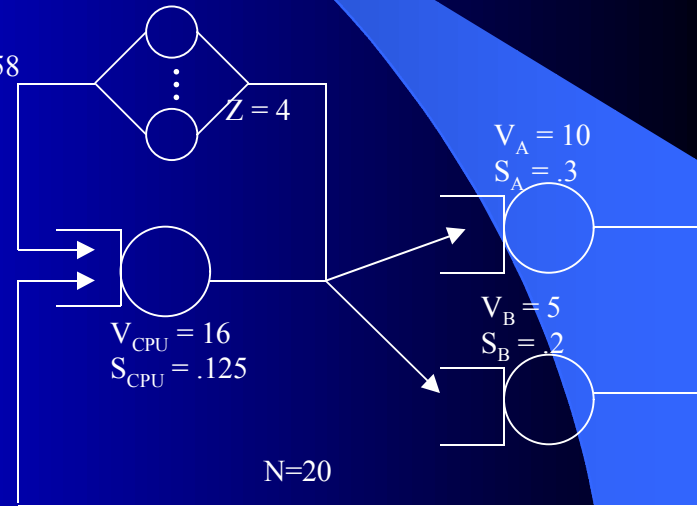
- $X = N / (R + Z) = 20 / (46 + 4) = 0.40$

- $Q_{CPU} = XR_{CPU}V_{CPU} = 0.40 * 0.958 * 16 = 6.133$

- $Q_A = XR_A V_A = 0.40 * 2.300 * 10 = 9.200$

- $Q_B = XR_B V_B = 0.40 * 1.533 * 5 = 3.067$

- ΔQ_{max}
 $= \max \{ |6.667 - 6.133|, |6.667 - 9.200|, |6.667 - 3.067| \}$
 $= 3.6$



Approximate MVA (cont'd)

Iteration	Response Time				Xput	Queue Length			Delta
	CPU	Disk A	Disk B	System		CPU	Disk A	Disk B	
0						6.667	6.667	6.667	
1	0.958	2.300	1.533	46.000	0.400	6.133	9.200	3.067	3.6
2	0.892	3.060	0.813	48.933	0.378	5.390	11.562	1.537	2.361713
3	0.799	3.769	0.507	53.003	0.351	4.484	13.222	0.890	1.660558
4	0.686	4.267	0.378	55.525	0.336	3.685	14.336	0.635	1.113393
5	0.586	4.601	0.327	57.013	0.328	3.072	15.081	0.536	0.74545
6	0.509	4.824	0.307	57.923	0.323	2.630	15.582	0.496	0.500668
7	0.454	4.975	0.299	58.502	0.320	2.323	15.918	0.479	0.336263
8	0.415	5.075	0.296	58.879	0.318	2.114	16.143	0.470	0.225285
9	0.389	5.143	0.294	59.128	0.317	1.973	16.294	0.466	0.150433
10	0.372	5.188	0.293	59.293	0.316	1.879	16.394	0.463	0.100133
11	0.360	5.218	0.293	59.403	0.315	1.816	16.460	0.462	0.066476
12	0.352	5.238	0.292	59.475	0.315	1.775	16.504	0.461	0.044044
13	0.347	5.251	0.292	59.523	0.315	1.747	16.534	0.460	0.029139
14	0.343	5.260	0.292	59.555	0.315	1.729	16.553	0.459	0.019258
15	0.341	5.266	0.292	59.576	0.315	1.717	16.566	0.459	0.012718
16	0.340	5.270	0.292	59.590	0.315	1.709	16.574	0.459	0.008395

Approximate MVA (cont'd)

- Questions

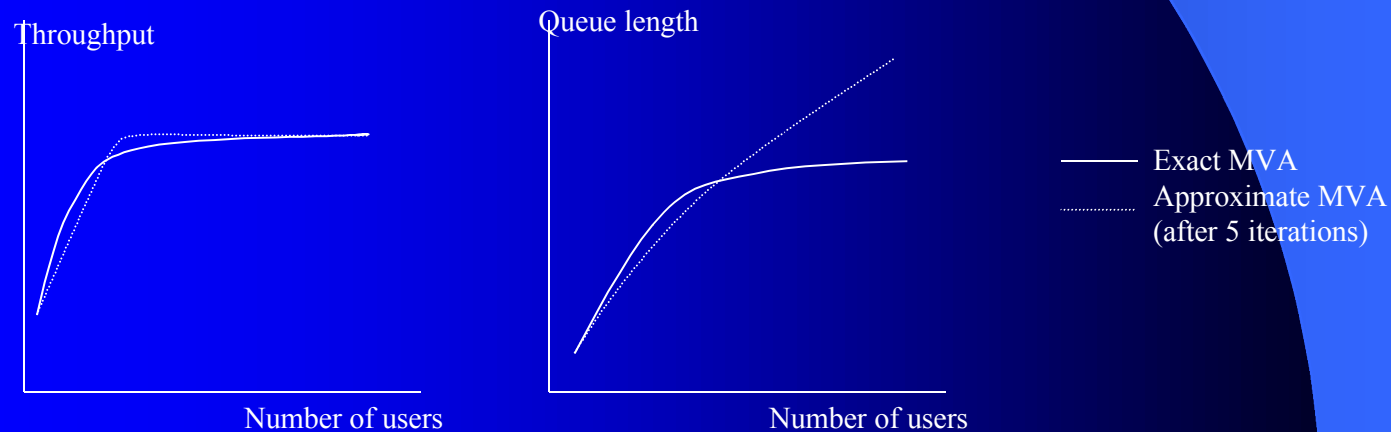
- Does it converge ?
- How accurate is it ?
- $Q_1 + Q_2 + \dots + Q_M = N$ at every iteration ?

- Answers

- Converges quickly in practice
- Small errors (5-10%) in practice
- $Q_1 + Q_2 + \dots + Q_M < N$: Where are the rest of the customers ?
 - $X(N) = N/(Z+R)$
 $= N/(Z+V_1R_1(N)+V_2R_2(N)+ \dots +V_MR_M(N))$
 - $Q_i(N)=X(N)V_iR_i(N)$
 - $Q_1 + Q_2 + \dots + Q_M = N/(Z+V_1R_1(N)+V_2R_2(N)+ \dots) * (V_1R_1(N)+V_2R_2(N)+ \dots) < N$
 - At delay center, $Q_i(N)=X(N)V_iR_i(N)= X(N)V_iZ_i$

Exact & Approximate MVA

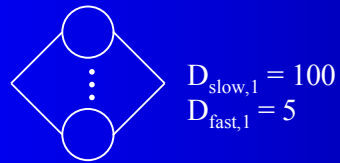
- Compare the two tables
 - Why they are different ? (not the same as textbook)
 - Even though throughput approaches steady-state value in a few steps but not the queue lengths => “when to stop” is important



* Multiple Class QNM

- Why do we have multiple classes of customers ?
 - Sometimes distinct customers have different behaviors
 - With single class QNM, they are averaged into a single number
 - It will be more accurate if those differences are reflected
- So, what's changed ?
 - Separate input & output measures per class
 - Input : arrival rate per class
 - Output : response time per class
- Study on
 - Fundamental Laws
 - Asymptotic Bounds
 - Mean Value Analysis

Fundamental Laws



$\lambda_{\text{slow}} = 0.5$



$\lambda_{\text{fast}} = 0.25$



- Little's Law still valid

$$U_{c,i} = X_c D_{c,i}$$

$$N_c = X_c R_c$$

- Example

$$U_{\text{slow},2} = X_{\text{slow}} D_{\text{slow},2}$$

$$= 0.5 * 1.2 = 60\%$$

$$U_{\text{fast},2} = X_{\text{fast}} D_{\text{fast},2}$$

$$= 0.25 * 1 = 25\%$$

$$U_2 = U_{\text{slow},2} + U_{\text{fast},2}$$

$$= 85\%$$

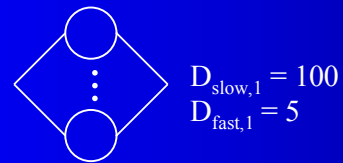
$$Q_{\text{slow},1} = X_{\text{slow}} R_{\text{slow},1}$$

$$= X_{\text{slow}} D_{\text{slow},1} \text{ (delay center)}$$

$$= 0.5 * 100 = 50$$

What will be the bottleneck center ?

Fundamental Laws (cont'd)



$\lambda_{slow} = 0.5$



$\lambda_{fast} = 0.25$



• Example

$$U_{slow,2} = 0.5 * 1.2 = 60\%$$

$$U_{fast,2} = 0.25 * 1 = 25\%$$

$$U_2 = U_{slow,2} + U_{fast,2} = 85\%$$

$$U_{slow,3} = 0.5 * 0.7 = 35\%$$

$$U_{fast,3} = 0.25 * 2 = 50\%$$

$$U_2 = U_{slow,3} + U_{fast,3} = 85\%$$

$$U_{slow,4} = 0.5 * 1.3 = 65\%$$

$$U_{fast,2} = 0.25 * 1 = 25\%$$

$$U_2 = U_{slow,2} + U_{fast,2} = 90\%$$

What is the bottleneck center ?

Asymptotic Bounds

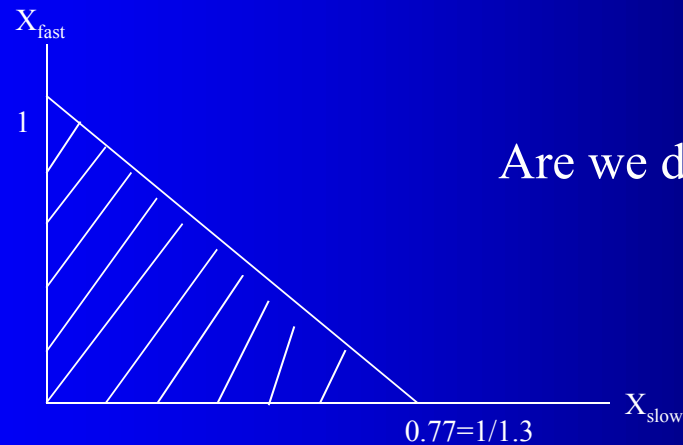
- Bounds come from Bottleneck Analysis

- $U_i = X D_i \Rightarrow X \leq 1/D_{\max}$

- $\Rightarrow X = X_{\text{slow}} + X_{\text{fast}} \leq ???$

- $\Rightarrow U_4 = X_{\text{slow}} D_{\text{slow},4} + X_{\text{fast}} D_{\text{fast},4} \leq 1$ (bottleneck device)

- $= 1.3 * X_{\text{slow}} + 1 * X_{\text{fast}} \leq 1$



Asymptotic Bounds (cont'd)

- What if $X_{\text{slow}}=0.25$ and $X_{\text{fast}}=0.4$?
 - $U_2 = U_{\text{slow},2} + U_{\text{fast},2} = 0.25*1.2+0.4*1 = 70\%$
 - $U_3 = U_{\text{slow},3} + U_{\text{fast},3} = 0.25*0.7+0.4*2 = 97.5\%$
 - $U_4 = U_{\text{slow},4} + U_{\text{fast},4} = 0.25*1.3+0.4*1 = 72.5\%$

