IEOR 4615: Service Engineering
Lecture 8, February 17, 2015

• Queueing Network Models (Tuesday)
  1. Classical Markov OQN Model (Sec. 7 of CTMC notes)
     1. Yet another CTMC, but multi-dimensional
     2. Product-Form Distribution
     3. Traffic Rate Equations
     1. Quantifying Variability and its Impact approximately
     2. The Parametric-Decomposition Approximation
     3. From Customer Routes to Markovian Routing

• Processing Networks (last Thursday & Friday)
  – Precedence Constraints, PERT/CPM
What do you see?

Need to have expectations
Experience and Models Help Greatly
Before
A Large Call Center
The Erlang A Model \((M/M/n+M)\)

\(M\) for “Markov” – special assumptions
This week: Hospitals

Intensive Care Unit (ICU)
Patient Flow in an Emergency Department
View as
An Open Network of Queues
Jackson (Open) Markov Queueing Network

(Section 7 in CTMC notes, handout and Courseworks)

\[ X_j(t) = \text{number of customers at station } j \text{ at time } t \]

\[(X_1(t), \ldots, X_m(t)) = \text{CTMC}\]

**Assumptions:**
- Customers come from outside and eventually leave
- Unlimited waiting room
- Number of stations is \( m \)
- External Poisson arrivals at each queue, rate \( \lambda_{e,j} \)
- Exponential service times at each station, rate \( \mu_j \)
- \( s_j \) servers at station \( j \)
- Markovian routing, \( (P_{i,j}) = P \)
Product-Form Steady-State Distribution

Steady state: convergence in distribution

$$(X_1(t), \ldots, X_m(t)) \to (X_1(\infty), \ldots, X_m(\infty)) \quad \text{as} \quad t \to \infty$$

Can find by $\alpha Q = 0$, but do not need to.

$$P(X_1(\infty) = j_1, \ldots, X_m(\infty) = j_m) = P(X_1(\infty) = j_1) \times \ldots \times P(X_m(\infty) = j_m)$$

Product-form = independence

where station $j$ is $M/M/s_j$ queue
Traffic Rate Equations

\[ \lambda_j = \lambda_{e,j} + \sum_{i=1}^{m} \lambda_i P_{i,j} \quad \text{for} \quad 1 \leq j \leq m \]

\[ \Lambda = \Lambda_e + \Lambda P \quad \text{(matrix version)} \]

\[ \Lambda = \Lambda_e (I - P)^{-1} \quad \text{(solution)} \]

traffic intensity at station \( j \): \( \rho_j = \frac{\lambda_j}{s_j \mu_j} \)

station \( j \) is \( M/M/s_j \) queue

With parameters: \( \lambda_j, \mu_j \) and \( s_j \)
The Queueing Network Analyzer (QNA)
(WW 1983, 1989 papers)

• **Approximate** Analysis of non-Markov OQN Model
• Characterize (quantify) *variability* at each queue
  – Arrival parameters: $\lambda$, $c_a^2$
  – Service parameters: $\mu$, $c_s^2$
  – $c^2 = \text{Variance/Mean}^2$
• Characterize (quantify) impact on performance
• Parametric-Decomposition Method
  – Network calculus for variability parameters
• Approximate Performance of GI/GI/s queues
Variability Parameters

For service times and interarrival times, think of IID random variables

squared coefficient of variation

\[ c^2 = c_X^2 = \frac{\text{Var}(X)}{E[X]}^2 \]

Exponential: \( c^2 = 1 \)
Deterministic: \( c^2 = 0 \)
Three Operations

(\lambda_1, c_1^2) 

(\lambda_2, c_2^2) 

(\lambda_3, c_3^2) 

superposition

flow through queue

(\lambda, c_a^2) 

(\lambda, c_d^2) 

splitting

(\lambda_1 \lambda, c_{d1}^2) 

(\lambda_2 \lambda, c_{d2}^2) 

(\lambda_3 \lambda, c_{d3}^2) 

(p_\lambda, c_{d1}^2) 

(p_\lambda, c_{d2}^2) 

(p_\lambda, c_{d3}^2)
From Routes to Markov Routing

• Start with routes for each class of customers

• Aggregate to convert to Markovian routing
  
  - $P_{ij}$ = proportion of departures from queue $i$ that go next to queue $j$
  
  - $(1/\mu_j)$ = average service time at queue $j$
  
  - $c^2_{sj}$ = appropriate scv
    
    • Second moment average of second moments
    
    • Variance = second moment – mean$^2$
    
    • Scv = variance/ mean$^2$
Example 1, p. 2791 of 1983 paper

- 2 queues and $r = 3$ routes

$$(n_k, \lambda_k, c_k^2; n_{k1}, \tau_{k1}, c_{s_{k1}}^2; \ldots; n_{kn_k}, \tau_{kn_k}, c_{s_{kn_k}}^2).$$

Here suppose that the $r$ vectors are:

$$(2, 2, 1; 1, 1, 1; 1, 3, 3)$$

$$(3, 3, 2; 1, 2, 0; 2, 1, 1; 1, 2, 1)$$

$$(2, 2, 4; 2, 1, 1; 1, 2, 1).$$
A Manufacturing Example
(from p. 1151 of Segal and WW 1989)

Table 1. Comparisons Between QNA and Simulation

<table>
<thead>
<tr>
<th>Model</th>
<th>Workstations</th>
<th>Products</th>
<th>Operations</th>
<th>Hops</th>
<th>WIP (Lots) Mean</th>
<th>Yield QNA/SIM</th>
<th>Interval (Time Units)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QNA</td>
<td>SIM</td>
<td>Mean</td>
</tr>
<tr>
<td>1</td>
<td>67</td>
<td>1</td>
<td>135</td>
<td>9</td>
<td>261.6</td>
<td>255.5</td>
<td>1.0045</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.2</td>
<td>32.4</td>
<td>4.81</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>1</td>
<td>108</td>
<td>15</td>
<td>41.8</td>
<td>40.5</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.7</td>
<td>12.4</td>
<td>2.8</td>
</tr>
</tbody>
</table>
But could have fork-join: Last Class Arrest-to-Arraignment Process Larson et al. (1993)