G/G/M QUEUEING NETWORKS MODEL WITH APPLICATION TO FAB DATASETS

Jinho Shin

KAIST
Industrial and systems engineering

Lab seminar, March 17, 2017
Contents

I. Problem description

II. Model formulation

III. Results of sensitivity analysis

IV. Concluding remarks and future works
Problem description

• Nature of the modern fab
  ▪ Reentrant process flow
  ▪ Deterministic (Most) & Probabilistic (Few) routing
  ▪ More than 30 different semiconductor products

• The heart of queueing network model
  ▪ Calculating the interarrival process variability, $C_a^2$

1) Decomposition with aggregation method (DWA)
   Toolset based $C_a^2$ calculation.

2) Decomposition without aggregation method (DWOA)
   ▪ DWA: Provides poor estimation for the networks with deterministic routing.
   ▪ DWOA: Able to separately approximate the departure variability from each operation at the same queue.


March 17, 2017, xS3D – 3
Problem description

- Preventive maintenance (PM) modeling
  - PMs modeled as classes of high priority non-preemptive (NPPR) customers.

- Issues on PM modeling
  - Reality Associated with a specific tool.
  - Model Served by any available tool.
Problem description

- Previous research

<table>
<thead>
<tr>
<th>Short title</th>
<th>Year</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• NPPR property for PM.</td>
</tr>
</tbody>
</table>

- Drawbacks
  [11]: Applicable for the systems with probabilistic routing.
  [2]: Simplified model - Aggregate all arrivals into a single process.
  [3]: Still, many modification points remain.

- Contribution
  1) We modified and improved ‘traffic variability’ and ‘waiting time approximation’ equations existed in [3].
  • We suggest the extension to existing approximations via decomposition without aggregation method.
  2) With the use of new approximation, various types of sensitivity analysis on total cycle time are conducted.
Experimentation

• Dataset description
  ▪ **Dataset 1**: MIMAC dataset #7
    - Deterministic routing system /* Publically available
  ▪ **Dataset 2**: Industry inspired fab dataset
    - Probabilistic & deterministic routing system /* Not publically available

• Simulation setup

  **Issue on simulation length**
  ▪ PM event arrivals can be very long. We need at least 100 events for each of them.
  ▪ Variation be larger as the toolset loading increases. Longer simulation required.

• Summary simulation analysis

<table>
<thead>
<tr>
<th>Category</th>
<th>Sensitivity analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Property</strong></td>
<td><strong>Bottleneck toolset loading</strong></td>
</tr>
<tr>
<td></td>
<td>Service distribution</td>
</tr>
<tr>
<td></td>
<td>Interarrival distribution</td>
</tr>
<tr>
<td>Various $\rho$ (%)</td>
<td>Various $Cs^2$ (Uniform)</td>
</tr>
<tr>
<td>Various $Ca^2$ (Gamma)</td>
<td></td>
</tr>
<tr>
<td>Dataset 1 &amp; 2</td>
<td>10 cases</td>
</tr>
<tr>
<td></td>
<td>(90, 91, 92, 93, 94, 95, 96, 97, 98, 99)</td>
</tr>
<tr>
<td></td>
<td>5 cases</td>
</tr>
<tr>
<td></td>
<td>(0.003, 0.030, 0.083, 0.163, 0.270)</td>
</tr>
<tr>
<td></td>
<td>6 cases</td>
</tr>
<tr>
<td></td>
<td>(0.0625, 0.125, 0.25, 0.5, 1, 2)</td>
</tr>
</tbody>
</table>
Results of sensitivity analysis

- Bottleneck queue loading, $\rho$

- Changing the lot arrival rate to vary the bottleneck queue loading.
- As loading increases, mean total cycle time also drastically increases.
- For dataset 2, prediction error is growing as loading increases.
- Dataset 1: Sim > Approx / Dataset 2: Approx > Sim
Results of sensitivity analysis

- Service time distribution, $Cs^2$

- The change of $Cs^2$ value shows a relatively minor effect.
- The mean total cycle time has little change (Increasing).
- Dataset 1: Sim > Approx / Dataset 2: Approx > Sim
- Nearly linear relationship between the $Cs^2$ and CT.
Results of sensitivity analysis

- Interarrival time distribution, $Ca^2$

- Nearly linear relationship observed between the $Ca^2$ and CT.
- Dataset 1: Sim > Approx / Dataset 2: Approx > Sim
- Still, our approximation follows well the tendency of simulation results.
Concluding remarks and future works

• G/G/m mean queueing approximation and numerical study
  – Modified G/G/m mean queueing network model using DWOA is proposed.
  – Simulation analysis on total CT conducted with two industry inspired datasets.
  – By varying 3 parameters in datasets, sensitivity analysis conducted.
    • The model has errors less than 5% in most cases.
    • The biggest error: 27.8%, 99% loading case of dataset 2.
    • For all, prediction errors are negative in deterministic routing system.
    • The errors are always positive in the system with probabilistic routing.
    • Generally, model follows well the tendency of simulation.

• Future works
  – Investigation on the systematic differences between the systems.
  – Improvement in the approximation quality for high loading cases.
  – Modification of the state of art batching approximation to our model.
G/G/M QUEUEING NETWORKS MODEL WITH APPLICATION TO FAB DATASETS

Q&A

Jinho Shin

KAIST
Industrial and systems engineering

Lab seminar, March 17, 2017
References


References


Appendix - Notations

- List of notations
  - Parameter sets
    \[ T = \{1, \ldots, W\} \]
    \[ O = \{1, \ldots, v\} \]
    \[ P = \{v + 1, \ldots, v + \theta\} \]
    \[ T_{k,O} \in O \]
    \[ T_{k,PM} \in P \]
    \[ T_k \]
  - Traffic rates related variables

- Above parameters define the toolsets, operations, PM type operations in network.
- By traffic equation, Mean total arrival rate of all customers to queue k also generated.
Appendix - Notations

- List of notations

- Traffic variability and duration related variables

\[
\begin{align*}
S_i & \quad C_{S_i} \\
D_i & \quad C_{D_i} \\
s(i) & \quad m_k \\
C_{a,PM_i} & \quad C_{a_i}^{EX} \\
C_{a_i} & \quad \rho \sigma (i), \rho \sigma (i)
\end{align*}
\]

Traffic variability equation
\[
C_{a_i}^2 = A_i + \sum_{j \in \mathbb{N}} (B_{ij}) C_{a_i}^2 + \sum_{j \in \mathbb{N}, \in \mathbb{N}} (C_{ij}) C_{a_j}^2
\]

- As the PM type operations only possess external arrival process, \(C_{a,PM_i}\) value is used.
- Based on the traffic variability equations, \(C_{a_i}\) value is generated for each operation.
Appendix - Notations

- List of notations
  - Total cycle time related variables

- Subject to generate the mean total CT of customer in the network.
- The variables for expected number of visits to each operation are defined.
Model formulation

- Variation of inter-arrival process

\[ C_{a_i}^2 = A_i + \sum_{j \in G_i \atop i \neq j} (B_{i,j})C_{a_j}^2 + \sum_{j \in G_i \atop i \neq j} \sum_{h \in T_j} (C_{j,h})C_{a_h}^2 \]

with \( A_i, B_{i,j}, C_{j,h} \) are given by:

\[
A_i = \frac{\lambda_i^{EX}}{\lambda_i}C_{a_i}^{EX \lambda_i} + \frac{1}{\lambda_i} \sum_{j \in G_i \atop i \neq j} \lambda_{j,i}q_{j,i} \left(1 - q_{j,i}\right) + \sum_{j \in G_i} \frac{\lambda_{j,i}q_{j,i}^2}{\lambda_i} \left( \frac{r_{\sigma(j)}^2}{\lambda_{j,i}} \right) \left( \frac{Cs_h^2}{\rho_{\sigma(h)}^2} \right) \left( \rho_{\sigma(j),i} \rho_{\sigma(j)} \right) + \sum_{j \in G_i} \frac{\lambda_{j,i}q_{j,i}^2}{\lambda_i} \left( \rho_{\sigma(j),i} \right) \left( 1 + \frac{Cs_j^2 - 1}{m_{\sigma(j)}^0.5} \right)
\]

- \( q_{j,i} \): Probability of customer routed from operation j to i
- \( \lambda_i (\lambda_f) \): Total arrival rate of customers to operation
- \( \rho_{\sigma(i),i} \): Loading due to operation i at queue \( \sigma(i) \)
- \( \rho_{\sigma(i)} \): Total loading of queue \( \sigma(i) \)
- \( \Gamma_{\sigma(f)} \): Total arrival rate of all customers to queue \( \sigma(f) \)
- \( Cs_h^2 \): SCV of the service time for operation h
- \( m_k \): Number of servers at queue k
- \( \sigma(i) \): Queue at which operation i performed

3) The service variability affected by other operations.
Model formulation

- Variation of inter-arrival process

\[
C_{a_i}^2 = A_i + \sum_{j \in G_i \atop i \neq j} (B_{i,j}) C_{a_j}^2 + \sum_{j \in G_i \atop i \neq j} \sum_{h \in T_j \atop h \neq j} (C_{j,h}) C_{a_h}^2
\]

with \(A_i, B_{i,j}, C_{j,h}\) are given by:

\[
B_{i,j} = \frac{q_{j,i} \lambda_j}{\lambda_i} \left[ \rho_{\sigma(j)}^2 \left( 1 - \frac{\rho_{\sigma(j),j}}{\rho_{\sigma(j)}} \right)^2 + (1 - \rho_{\sigma(j)}^2) \right]
\]

- \(q_{j,i}\): Probability of customer routed from operation \(j\) to \(i\)
- \(\lambda_i, \lambda_j\): Total arrival rate of customers to operation
- \(\rho_{\sigma(i),i}\): Loading due to operation \(i\) at queue \(\sigma(i)\)
- \(\rho_{\sigma(i)}\): Total loading of queue \(\sigma(i)\)
- \(\Gamma_{\sigma(j)}\): Total arrival rate of all customers to queue \(\sigma(i)\)

- \(\rho_{\sigma(i),i}\): Loading due to operation \(i\) at queue \(\sigma(i)\)
- \(\rho_{\sigma(i)}\): Total loading of queue \(\sigma(i)\)

- To adjust \(B_{i,j}\) in the case that operation \(h\) is serviced at the same queue as an operation \(j\).
Model formulation

- Total cycle time measure
  - Waiting time approximation

\[ \begin{align*}
W_{q_k} & \approx \frac{\left( \sum_{i \in T_k} \rho \sigma_{(i,i)} \right) \sqrt{m_k^{-1}}}{m_k^2} \frac{\sum_{i \in T_k, O} \left( c a_i^2 + C s_i^2 \right) (\lambda_i S_i^2) + \sum_{i \in T_{k,PM}} \left( C a_{PM_i} + C D_i \right) (\lambda_i, PM, D_i)}{2 \left( 1 - \sum_{i \in T_k} \rho \sigma_{(i,i)} \right) \left( 1 - \sum_{i \in T_{k,PM}} \rho \sigma_{(i,i)} \right)} \\
\end{align*} \]

- \( W_{q_k} \): the mean queueing delay time at queue \( k \) for (product) customers.

- Total cycle time approximation

\[ TC_i = \sum_{j \in O} \{ n_{i,j} S_j \} + \sum_{k \in T} \{ n_{i,k} W_{q_k} \}, \text{ for } \{ i : \lambda_i^{EX} > 0 \} \]

- Summation of all process times and delaying times in production line.
- The mean total cycle time is generated and be compared to simulation results.
Appendix: results of sensitivity analysis

- Simulation results summary (1)

<table>
<thead>
<tr>
<th>Sensitivity Analysis</th>
<th>Simulated &amp; Approximated mean total cycle time (h)</th>
<th>MIMAC dataset 7</th>
<th>Industry inspired fab dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
<td>Approximation</td>
<td>Simulation</td>
</tr>
<tr>
<td><strong>Sensitivity 1. Bottleneck queue loading</strong></td>
<td>Simulation</td>
<td>Approximation</td>
<td>Simulation</td>
</tr>
<tr>
<td>ρ = 90.0%</td>
<td>1606.27</td>
<td>1513.55</td>
<td>1384.58</td>
</tr>
<tr>
<td>ρ = 91.0%</td>
<td>1657.71</td>
<td>1560.58</td>
<td>1424.35</td>
</tr>
<tr>
<td>ρ = 92.0%</td>
<td>1731.66</td>
<td>1616.65</td>
<td>1462.27</td>
</tr>
<tr>
<td>ρ = 93.0%</td>
<td>1803.81</td>
<td>1685.28</td>
<td>1517.91</td>
</tr>
<tr>
<td>ρ = 94.0%</td>
<td>1946.82</td>
<td>1773.15</td>
<td>1592.06</td>
</tr>
<tr>
<td>ρ = 95.0%</td>
<td>2072.41</td>
<td>1891.30</td>
<td>1685.43</td>
</tr>
<tr>
<td>ρ = 96.0%</td>
<td>2243.70</td>
<td>2061.68</td>
<td>1813.47</td>
</tr>
<tr>
<td>ρ = 97.0%</td>
<td>2537.23</td>
<td>2338.07</td>
<td>2012.65</td>
</tr>
<tr>
<td>ρ = 98.0%</td>
<td>3021.92</td>
<td>2879.78</td>
<td>2381.75</td>
</tr>
<tr>
<td>ρ = 99.0%</td>
<td>4521.24</td>
<td>4492.57</td>
<td>3243.05</td>
</tr>
<tr>
<td><strong>Sensitivity 2. Service time distribution</strong></td>
<td>Simulation</td>
<td>Approximation</td>
<td>Simulation</td>
</tr>
<tr>
<td>Uniform ±10%</td>
<td>1606.27</td>
<td>1513.55</td>
<td>1384.58</td>
</tr>
<tr>
<td>Uniform ±30%</td>
<td>1611.72</td>
<td>1519.99</td>
<td>1425.72</td>
</tr>
<tr>
<td>Uniform ±50%</td>
<td>1631.22</td>
<td>1532.88</td>
<td>1491.52</td>
</tr>
<tr>
<td>Uniform ±70%</td>
<td>1649.16</td>
<td>1552.24</td>
<td>1548.24</td>
</tr>
<tr>
<td>Uniform ±90%</td>
<td>1677.24</td>
<td>1578.01</td>
<td>1621.14</td>
</tr>
<tr>
<td><strong>Sensitivity 3. Interarrival time distribution</strong></td>
<td>Simulation</td>
<td>Approximation</td>
<td>Simulation</td>
</tr>
<tr>
<td>Gamma, 16</td>
<td>995.89</td>
<td>898.45</td>
<td>1192.03</td>
</tr>
<tr>
<td>Gamma, 8</td>
<td>1013.93</td>
<td>939.46</td>
<td>1194.56</td>
</tr>
<tr>
<td>Gamma, 4</td>
<td>1078.33</td>
<td>1021.47</td>
<td>1216.49</td>
</tr>
<tr>
<td>Gamma, 2</td>
<td>1273.08</td>
<td>1185.50</td>
<td>1282.71</td>
</tr>
<tr>
<td>Gamma, 1</td>
<td>1606.27</td>
<td>1513.55</td>
<td>1384.58</td>
</tr>
<tr>
<td>Gamma, 0.5</td>
<td>2289.14</td>
<td>2169.66</td>
<td>1659.46</td>
</tr>
</tbody>
</table>
Appendix: results of sensitivity analysis

- Simulation results summary (2)

<table>
<thead>
<tr>
<th>Sensitivity Analysis</th>
<th>Mean total CT comparison (Difference, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MIMAC dataset 7</td>
</tr>
<tr>
<td><strong>Sensitivity 1. Bottleneck queue loading</strong></td>
<td></td>
</tr>
<tr>
<td>( \rho = 90.0% )</td>
<td>-5.77</td>
</tr>
<tr>
<td>( \rho = 91.0% )</td>
<td>-5.86</td>
</tr>
<tr>
<td>( \rho = 92.0% )</td>
<td>-6.64</td>
</tr>
<tr>
<td>( \rho = 93.0% )</td>
<td>-6.57</td>
</tr>
<tr>
<td>( \rho = 94.0% )</td>
<td>-8.92</td>
</tr>
<tr>
<td>( \rho = 95.0% )</td>
<td>-8.74</td>
</tr>
<tr>
<td>( \rho = 96.0% )</td>
<td>-8.11</td>
</tr>
<tr>
<td>( \rho = 97.0% )</td>
<td>-7.85</td>
</tr>
<tr>
<td>( \rho = 98.0% )</td>
<td>-4.70</td>
</tr>
<tr>
<td>( \rho = 99.0% )</td>
<td>-0.63</td>
</tr>
<tr>
<td><strong>Sensitivity 2. Service time distribution</strong></td>
<td></td>
</tr>
<tr>
<td>Uniform ( \pm 10% )</td>
<td>-5.77</td>
</tr>
<tr>
<td>Uniform ( \pm 30% )</td>
<td>-5.69</td>
</tr>
<tr>
<td>Uniform ( \pm 50% )</td>
<td>-6.03</td>
</tr>
<tr>
<td>Uniform ( \pm 70% )</td>
<td>-5.88</td>
</tr>
<tr>
<td>Uniform ( \pm 90% )</td>
<td>-5.92</td>
</tr>
<tr>
<td><strong>Sensitivity 3. Interarrival time distribution</strong></td>
<td></td>
</tr>
<tr>
<td>Gamma, 16</td>
<td>-9.78</td>
</tr>
<tr>
<td>Gamma, 8</td>
<td>-7.35</td>
</tr>
<tr>
<td>Gamma, 4</td>
<td>-5.27</td>
</tr>
<tr>
<td>Gamma, 2</td>
<td>-6.88</td>
</tr>
<tr>
<td>Gamma, 1</td>
<td>-5.77</td>
</tr>
<tr>
<td>Gamma, 0.5</td>
<td>-5.22</td>
</tr>
</tbody>
</table>