Cycle Time Approximations for the G/G/m Queue Subject to Server Failures and Cycle Time Offsets with Applications

James R. Morrison
Central Michigan University
Department of Engineering and Technology

Donald P. Martin
Productivity Engineering
IBM, Systems and Technology Group
Presentation Overview

- Standard approximations for the mean cycle time in a G/G/m queue
- Extensions to the G/G/m queue: Idle with WIP
- Extensions to the G/G/m queue: Cycle time offsets
- Extensions to the G/G/m queue: Defection of lots from failed servers
- Application to toolsets in IBM’s 200mm semiconductor fabricator
- Concluding remarks
The G/G/m Queue

- G/G/m queue contains
  - m equivalent tools
  - Random service times (General distribution with mean $1/\mu$)
  - Random time between lot arrivals (General distribution, mean $1/\lambda$)
  - System loading $\rho = \lambda / (m\mu)$ (Utilization of capacity)
Cycle Time Approximations

- Popular approximation for the G/G/m queue

\[ E(CT) \approx \frac{1}{\mu} + \frac{1}{\mu} \left( \frac{C_s^2 + C_A^2}{2} \right) \left( \frac{\rho^{-1+\sqrt{2m+2}}}{m(1-\rho)} \right) \]

Appears in the text
Factory Physics

Grows as loading increases

where

- \((1/\mu)\) is the mean service time
- \(C_s\) is the coefficient of variation of the service time (std/mean)
- \(C_A\) is the coefficient of variation of the interarrival time (std/mean)
- \(m\) is the number of servers
- System loading \(\rho = (\lambda / m \mu) < 1\)
The Approximation is Exact in Some Cases

- For the M/G/1 queue this expression is exact

\[ E(CT) = \frac{1}{\mu} + \frac{1}{\mu} \left( \frac{1 + C_s^2}{2} \right) \frac{\rho}{(1 - \rho)} \]

- \((1/\mu)\) is the mean service time
- System loading \(\rho = (\lambda / \mu) < 1\)
- \(C_s\) is the coefficient of variation of the service time

### Normalized Cycle Time Comparison: M/D/1 vs M/M/1

**Increasing variation or fewer tools**

**Loading (Utilization of Capacity)**

**Normalized Cycle Time**

0 0.2 0.4 0.6 0.8 1

0 1 2 3 4 5 6 7 8

Normalized Cycle Time
Approximations Incorporating Tool Failure

- A tool may be subject to random failures
  - Time to failure is exponentially distributed (mean $m_F$)
  - Time to repair is generally distributed ($m_R$, $\sigma_R$ and $C_R = \sigma_R/m_R$)
  - Mean availability is $A = m_F / (m_R + m_F)$
  - Production is resumed following repair

- Popular approximation which is exact for the M/G/1 queue with failures

\[
E(CT) \approx \left( \frac{1}{\mu^*} + \frac{1}{\mu^*} \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \rho^* \right) \frac{1}{1 - \rho^*}
\]

where

- $\mu^* = \mu A$
- $C_{S,E}^2 = C_S^2 + \left( 1 + C_R^2 \right)(1 - A)Am_R\mu$ - effective $C_S^2$
- $C_A = \sigma_A / \left( 1 / \mu \right)$, coefficient of variation of interarrival time
- System loading $\rho^* = \lambda / (\mu A) < 1$

Appears in the text Factory Physics

Service time is inflated by availability
Approximations for G/G/m Queue With Tool Failures

- Consider G/G/m queue with exponential time to failure for each tool
  - Time to repair is generally distributed ($m_R$, $\sigma_R$ and $C_R = \sigma_R/m_R$)
  - Mean availability is $A = m_F / (m_R + m_F)$
  - Lots remain with the failed server and production resumes upon repair

- Natural generalization suggested by the previous approximations

$$E(CT) \approx \frac{1}{\mu^*} + \frac{1}{\mu^*} \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \frac{\left( \rho^* \right)^{-1+\sqrt{2m+2}}}{m(1-\rho^*)}$$

where

- $\mu^* = \mu A$
- $C_{S,E}^2 = C_S^2 + (1 + C_R^2)(1 - A)Am_R\mu$ - effective $C_S^2$
- $C_A = \sigma_A / (1 / \mu)$, coefficient of variation of interarrival time
- System loading $\rho^* = \lambda / (m \mu A) < 1$
M/M/2 Queue Subject to Tool Failure

- Comparison of the approximation with exact results for the M/M/2 queue
  - Exponential repair
  - $m_F = 16$ hours
  - $m_R = 4$ hours
  - Process time $(1 / \mu) = 1$ hour

- The simpler intuitive Martin approximation is obtained by substituting

$$
\frac{(\rho^*)^m}{m(1 - \rho^*)} \rightarrow \frac{(\rho^*)^m}{1 - (\rho^*)^m}
$$

A Comparison of Approximate and Exact Cycle Time Performance:
M/M/2 Queue with Random Failure and Repair

MTTF = 16 h, MTTR = 4 h, process time = 1 h (all exponential)
Idle Tools in the Presence of WIP

• A tool may be idle even in the presence of WIP
  – Loading time
  – Operator unavailable

• Model the idle with WIP as a random addition to the process time (mean $\Omega$ and standard deviation $\sigma_{\Omega}$)

$$E(CT) \approx \frac{1}{\mu_e} + \frac{1}{\mu_e} \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \left( \rho^* \right)^{\sqrt{2(m+1)}-1} \frac{m(1-\rho^*)}{m(1-\rho^*)},$$

where $\rho^* = \frac{\lambda(\Omega + 1/\mu)}{(mA)} < 1$

$$C_{S,E}^2 = \frac{\sigma_s^2 + \sigma_{\Omega}^2}{[(1/\mu) + \Omega]^2} + (1 + C_R^2)(1 - A) \left[ \frac{m_R}{(1/\mu) + \Omega} \right]$$

$$\mu_e \equiv \left[ \frac{1}{\mu A} + \frac{\Omega}{A} \right]^{-1}$$

Loading is increased

Production speed is reduced
Cycle Time Offsets

• Common manufacturing events include:
  – Transport of lots from one toolset to another
  – Hold of lots pending resolution of a process concern
  – Post production delay

• Often independent of the queue at a particular toolset

\[
E(CT) \approx T + H + P + \frac{1}{\mu_e} \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \frac{(\rho^*)^{\sqrt{2(m+1)-1}}}{m(1 - \rho^*)},
\]

where

– \( T \) is the mean transport time for lots arriving to the toolset
– \( H \) is the mean time that lots are on hold before release to the queue
– \( P \) is the mean post production delay before transport to the next toolset
Loyalty to a Failed Tool

- Recall that lots were assumed to remain with a failed tool once they begin production at that tool.
- If the tools exhibit 80% availability the approximation yields:

\[ E(CT) \approx \frac{1.25}{\mu} + \frac{1.25}{\mu} \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \left( \rho^* \right)^{-1 + \sqrt{2m+2}} m(1 - \rho^*) \]

- Even in very low loading conditions \((\rho^* = 0)\):

\[ E(CT) \approx 1.25 \left( \frac{1}{\mu} \right) \]

- Inappropriate model for some toolsets as lots may defect from a failed server in favor of an available one!
Defection of Lots From a Failed Tool

- Suppose lots are allowed to defect to another tool in the event that their production is interrupted by tool failure.

- In very low loading conditions ($\rho^* = 0$)
  - Service time may continue uninterrupted if another tool is up.
  - Only if *all tools have failed* will the service be delayed.
  - Roughly expect (with deterministic repair times)

$$\lim_{\rho \to 0^+} E(CT) \approx \frac{1}{\mu} + (1 - A)^m \frac{m_R}{m + 1}$$

Probability that an arriving lot sees all tools in failure

Residual down time when all tools fail.
General Cycle Time Approximation

- For the G/G/m queue, incorporating
  - Failure prone tools with deterministic repair times
  - Idle with WIP
  - Cycle time offsets
  - Defection of lots from failed servers

\[
E(CT) \approx (T + H + P) \quad \text{Cycle time offsets}
\]

\[
+ (1 - A)^m \left( \frac{m_R}{m + 1} \right) \quad \text{All tools fail}
\]

\[
+ \left( \frac{1}{\mu} + \Omega \right) \quad \text{Process time}
\]

\[
+ \left( \frac{1}{\mu} + \Omega \right) \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \left( \rho^* \right)^{\sqrt{2(m+1) - 1}} \frac{m(1 - \rho^*)}{m(1 - \rho^*)} \quad \text{Queueing}
\]
Alternate Cycle Time Approximation

- For the G/G/m queue, incorporating
  - Failure prone tools with deterministic repair times
  - Idle with WIP
  - Cycle time offsets
  - Defection of lots from failed servers

\[ E(CT) \approx (T + H + P) \]

\[ + (1 - A)^m \left( \frac{m_R}{m + 1} \right) \]

\[ + \left( \frac{1}{\mu} + \Omega \right) \]

\[ + \left( \frac{1}{\mu} + \Omega \right) \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \left( \rho^* \right)^m \quad \text{All tools fail} \]

\[ + \left( \frac{1}{\mu} + \Omega \right) \left( \frac{C_{S,E}^2 + C_A^2}{2} \right) \left( \rho^* \right)^m \quad \text{Process time} \]

\[ 1 - \left( \rho^* \right)^m \quad \text{Intuitive & simpler queueing approximation} \]
Applying the Cycle Time Approximation: First Example

- For a tool set operating in IBM’s 200mm fabricator

- For a biweekly period:
  - Measure statistics \(C_A^2, A, \rho^*, \Omega, T, \ldots\)
  - Measure actual cycle time performance
  - Compare!

- Dominant factors
  - Cycle time offsets
  - Idle with WIP
  - Low loading
Applying the Cycle Time Approximation: Second Example

- For a tool set operating in IBM’s 200mm fabricator

- For a biweekly period:
  - Measure statistics ($C_A^2$, $A$, $\rho^*$, $\Omega$, $T$, …)
  - Measure actual cycle time performance
  - Compare!

- Dominant factors
  - Cycle time offsets
  - Idle with WIP
  - Multiplicity of tools
  - Low variability
Concluding Remarks

• Queueing models for manufacturing system performance evaluation

• Standard approximations for the mean cycle time in a G/G/m queue

• Extensions to the G/G/m queue
  – Idle with WIP
  – Cycle time offsets
  – Defection of lots from failed servers

• Application to toolsets in IBM’s 200mm semiconductor fabricator

• Future directions: Apply to all toolsets, rollup to fab performance curve, more rigorous lot defection analysis