Queueing Theory

Basic Queuing Relationships

Resident items

Waiting items

Residence time

General		Single Server	
$r = \lambda T_r$	Little's formula	$\rho = \lambda T_s$	
$w = \lambda T_w$	Little's formula	$r = w + \rho$	
$T_r = T_w + T_s$			

Single server Utilisation

System Utilisation

Multiserver
$\rho = \dfrac{\lambda T_s}{N}$
$u = \lambda T_s = \rho N$
$r = w + N\rho$

Little's formulae are the most important equation in queuing theory

M/G/1 … M/M/1 … M/D/1

(a) General Service Times (M/G/1)

$$A = \frac{1}{2}\left[1 + \left(\frac{\sigma_{T_s}}{T_s}\right)^2\right]$$

$$r = \rho + \frac{\rho^2 A}{1-\rho}$$

$$w = \frac{\rho^2 A}{1-\rho}$$

$$T_r = T_s + \frac{\rho T_s A}{1-\rho}$$

$$T_w = \frac{\rho T_s A}{1-\rho}$$

(b) Exponential Service Times (M/M/1)

$$r = \frac{\rho}{1-\rho} \qquad w = \frac{\rho^2}{1-\rho}$$

$$T_r = \frac{T_s}{1-\rho} \qquad T_w = \frac{\rho T_s}{1-\rho}$$

$$\sigma_r = \frac{\sqrt{\rho}}{1-\rho} \qquad \sigma_{T_r} = \frac{T_s}{1-\rho}$$

$$\Pr[R = N] = (1-\rho)\rho^N$$

$$\Pr[R \le N] = \sum_{i=0}^{N} (1-\rho)\rho^i$$

$$\Pr[T_R \le T] = 1 - e^{-(1-\rho)q T_s}$$

$$m_{T_r}(y) = T_r \times \ln\left(\frac{100}{100-y}\right)$$

$$m_{T_w}(y) = \frac{T_w}{\rho} \times \ln\left(\frac{100\rho}{100-y}\right)$$

(c) Constant Service Times (M/D/1)

$$r = \frac{\rho^2}{2(1-\rho)} + \rho$$

$$w = \frac{\rho^2}{2(1-\rho)}$$

$$T_r = \frac{T_s(2-\rho)}{2(1-\rho)}$$

$$T_w = \frac{\rho T_s}{2(1-\rho)}$$

$$\sigma_r = \frac{1}{1-\rho}\sqrt{\rho - \frac{3\rho^2}{2} + \frac{5\rho^3}{6} - \frac{\rho^4}{12}}$$

$$\sigma_{T_r} = \frac{T_s}{1-\rho}\sqrt{\frac{\rho}{3} - \frac{\rho^2}{12}}$$

Single server – queue size as function of σ

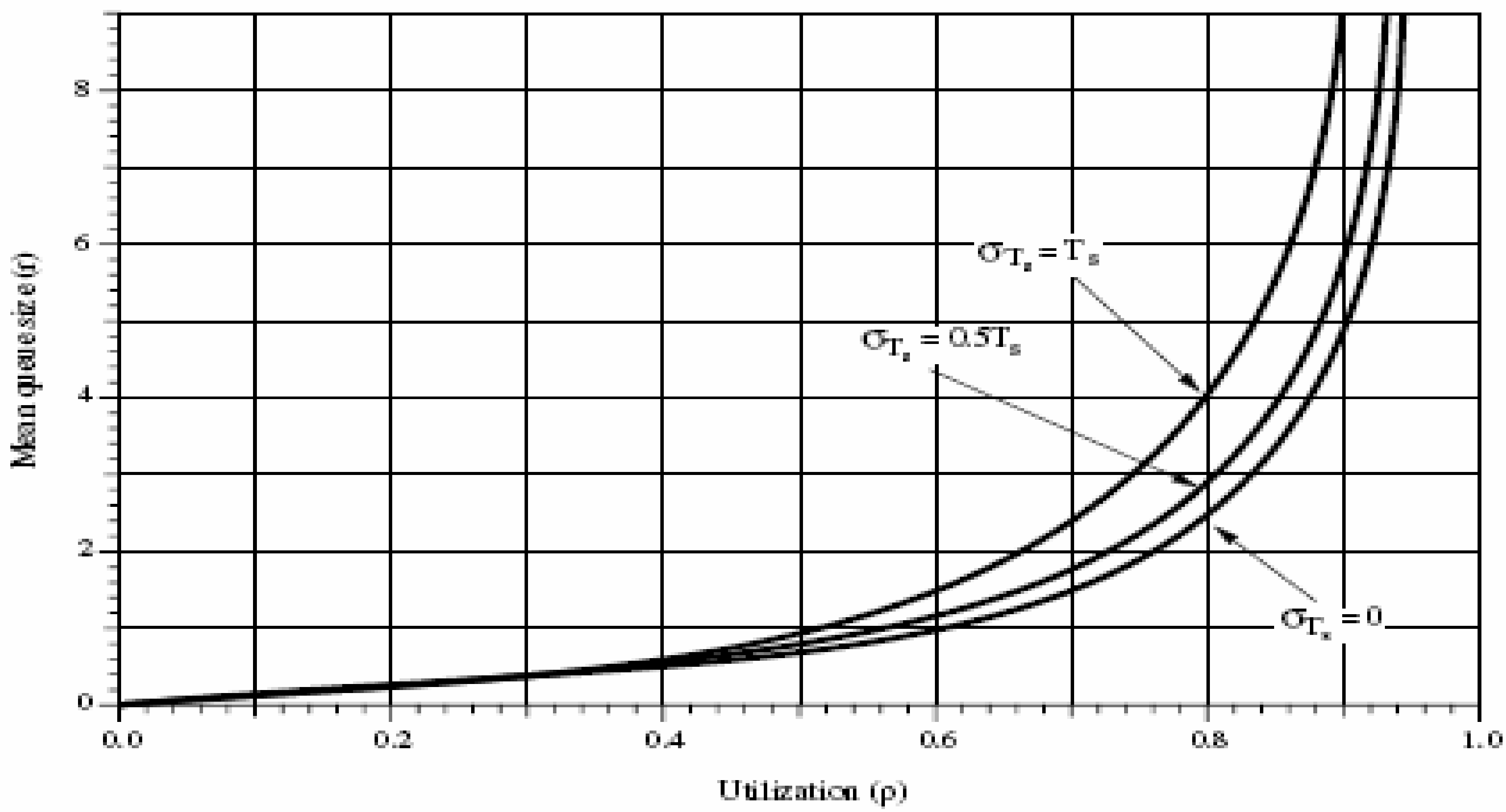


Figure 4 Mean Queue Size for Single-Server Queue

Single server – residency time as function of σ

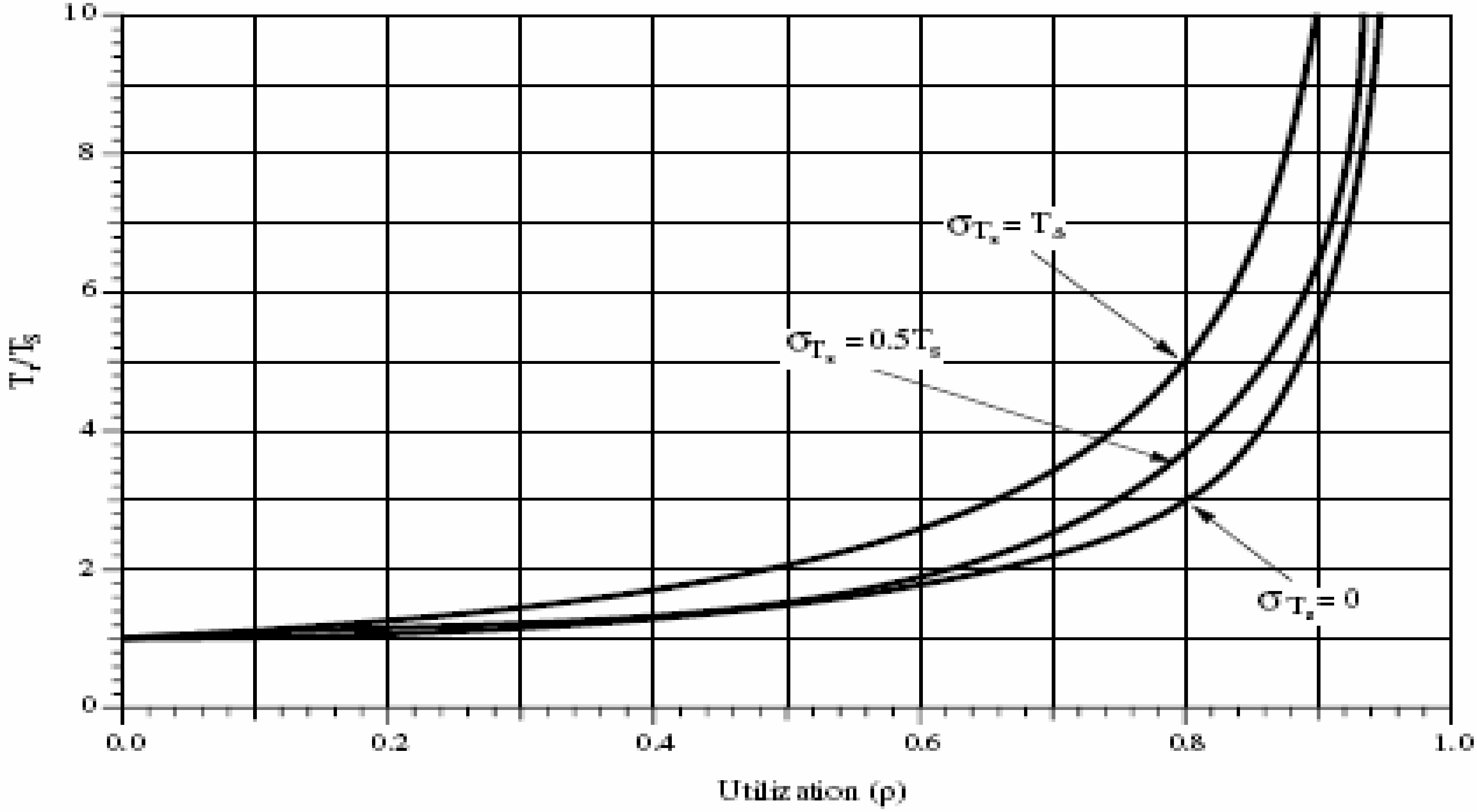


Figure 5 Mean Residence Time for Single-Server Queue

M/M/N Analysis

Assumptions:

1. Poisson arrival rate.
2. Exponential service times
3. All servers equally loaded
4. All servers have same mean service time
5. First-in, first-out dispatching
6. No items are discarded from the queue

$$K = \frac{\sum_{I=0}^{N-1} \frac{(N\rho)^I}{I!}}{\sum_{I=0}^{N} \frac{(N\rho)^I}{I!}}$$ Poisson ratio function

Erlang-C function = Probability that all servers are busy = $C = \frac{1-K}{1-\rho K}$

M/M/N Analysis

$$r = C\frac{\rho}{1-\rho} + N\rho \qquad w = C\frac{\rho}{1-\rho}$$

$$T_r = \frac{C}{N}\frac{T_s}{1-\rho} + T_s \qquad T_w = \frac{C}{N}\frac{T_s}{1-\rho}$$

$$\sigma_{T_r} = \frac{T_s}{N(1-\rho)}\sqrt{C(2-C) + N^2(1-\rho)^2}$$

$$\sigma_w = \frac{1}{1-\rho}\sqrt{C\rho(1+\rho-C\rho)}$$

$$\Pr\left[T_w > t\right] = C\mathrm{e}^{-N(1-\rho)t/T_s}$$

$$m_{T_w}(y) = \frac{T_s}{N(1-\rho)}\ln\left(\frac{100C}{100-y}\right)$$

$$T_d = \frac{T_s}{N(1-\rho)}$$

Variability

- **Definition**: Variability is anything that causes the system to depart from regular, predictable behavior.

- **Sources of Variability**:
  – setups
  – machine failures
  – materials shortages
  – yield loss
  – rework
  – operator unavailability
  - workpace variation
  - differential skill levels
  - engineering change orders
  - customer orders
  - product differentiation
  - material handling

Measuring Process Variability

$$t = \text{mean}$$

$$\sigma = \text{standard deviation}$$

$$c = \frac{\sigma}{t} = \text{coefficient of variation, CV}$$

$$c^2 = \frac{\sigma^2}{t^2} = \text{squared coefficient of variation, SCV}$$

Kendall's Classification

Characterization of a queueing station

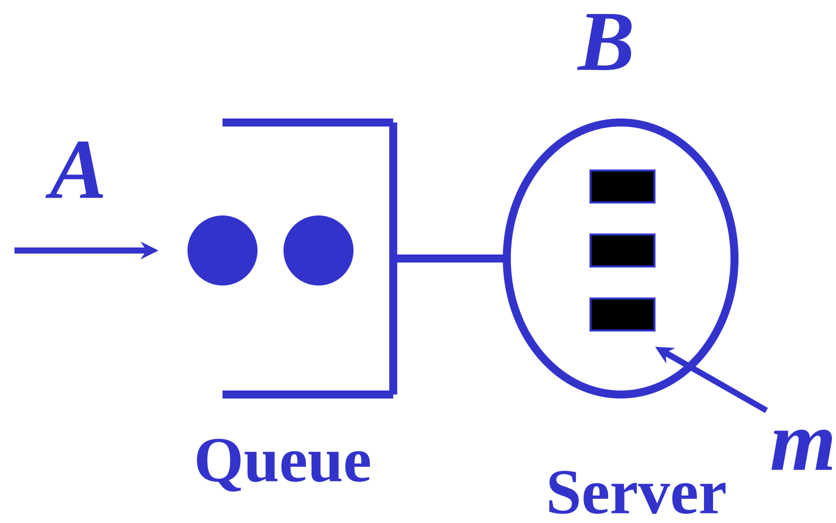
$$A \,/\, B \,/\, m \,/\, b$$

A: arrival process

B: service process

m: number of machines

b: maximum number of jobs
that can be in the system



M: exponential (Markovian) distribution

G: completely general distribution

D: constant (deterministic) distribution.

Queueing Parameters

r_a = the rate of arrivals in customers (jobs) per unit time

$t_a = 1/r_a =$ the average time between arrivals.

c_a = the CV of inter-arrival times.

m = the number of machines.

b = buffer size (i.e., maximum number of jobs allowed in system.)

t_e = mean effective process time.

r_e = the rate of the station in jobs per unit time = m/t_e.

c_e = the CV of effective process times.

u = utilization of station = r_a/r_e.

Queueing Measures

- **Measures**:

  T_q = the expected waiting time spent in queue.

  T = the expected time spent at the process center, i.e., queue time plus process time.

  N = the average jobs at the station.

  N_q = the expected jobs in queue.

- **Relationships**:

  $T = T_q + t_e$

  $N = r_a \times T$

  $N_q = r_a \times T_q$

- **Result**: If we know T_q, we can compute N, N_q, T.

The *G/G/1* Queue

- **Formula**:

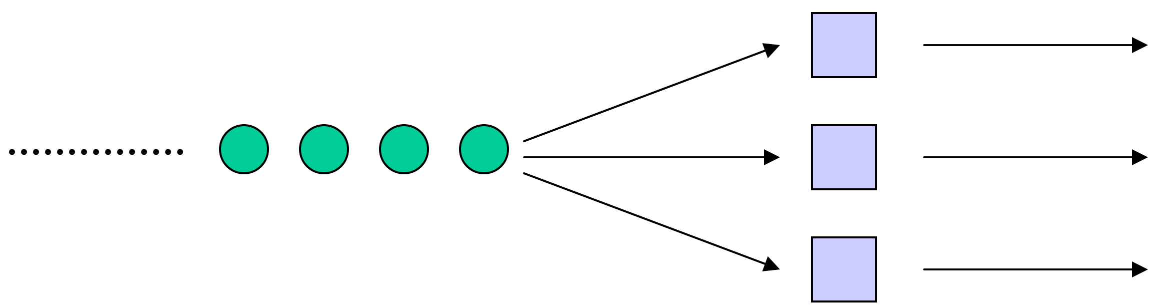
$$\mathrm{T}_q \approx \underbrace{\left(\frac{c_a^2 + c_e^2}{2}\right)}_{V} \underbrace{\left(\frac{u}{1-u}\right)}_{U} \underbrace{t_e}_{T} \qquad = \left(\frac{c_a^2 + c_e^2}{2}\right) \mathrm{T}_q(M/M/1)$$

- **Observations**:
  - Refer to as Kingman's equation or *VUT* equation.
  - Separate terms for variability, utilization, process time.
  - T_q (and other measures) increase with c_a^2 and c_e^2.
  - ***Variability causes congestion!***

The *M/M/m* Queue

- Systems with multiple machines in parallel.
- All jobs wait in a single queue for the next available machine.

$$\mathrm{T}_q(M/M/m) \approx \frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} t_e$$



The *G/G/m* Queue

- **Formula**:

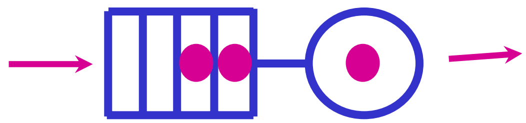
$$\mathrm{T}_q \approx \underbrace{\left(\frac{c_a^2 + c_e^2}{2}\right)}_{V} \underbrace{\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)}_{U} \underbrace{t_e}_{T} = \left(\frac{c_a^2 + c_e^2}{2}\right) \mathrm{T}_q(M/M/m)$$

- **Observations**:
  - Useful model of multi-machine workstations
  - *Extremely* general.
  - Fast and accurate.
  - Easily implemented in a spreadsheet (or packages).

Effects of Blocking

- ***VUT* Equation**:
  - characterizes stations with infinite space for queueing
  - useful for seeing what will happen to N, T without restrictions
- **But real world systems often constrain N**:
  - physical constraints
  - logical constraints
- **Blocking Models**:
  - estimate N and r_a for given set of rates, buffer sizes
  - much more complex than non-blocking (open) models, often require simulation to evaluate realistic systems

The *M/M/1/b* Queue



B buffer spaces

$$N(M/M/1/b) = \frac{u}{1-u} - \frac{(b+1)u^{b+1}}{1-u^{b+1}}$$

Goes to $u/(1 - u)$ as $b \to \infty$
Always less than N(*M/M/1*)

$$Thoughput(M/M/1/b) = \frac{1-u^b}{1-u^{b+1}} r_a$$

Goes to r_a as $b \to \infty$
Always less than Throughput(*M/M/1*)

$$T(M/M/1/b) = \frac{N(M/M/1/b)}{Throughput(M/M/1/b)}$$

Little's law

Variability Pooling

- Variability pooling: combine multiple sources of variability.
- Basic idea: the CV of a sum of independent random variables decreases with the number of random variables.
- Example:
  - Batch processing
  - Safety stock aggregation
  - Queue sharing

Conclusions

- Variability is a fact of life.
- There are many sources of variability in manufacturing systems.
- The coefficient of variation is a key measure of item variability.
- Variability propagates.
- Waiting time is frequently the largest component of the total time in the system.
- Limiting buffers reduces total time in the system at the cost of decreasing throughput.
- Variability pooling reduces the effect of variability.