MFS605/EE605
Systems for Factory Information and Control

Lecture 4

Fall 2005

Larry Holloway

Dept. of Electrical and Computer Engineering and

UK Center for Manufacturing

- References:
  - Modeling and Analysis of Mfg. Systems, Askin and Standridge (John Wiley, 1993)
  - Factory Physics, by Hopp and Spearman (McGraw Hill 1996)
  - Simulation Modeling and Analysis, 2nd ed. Law and Kelton, 1991 (McGraw Hill)
  - Probability and Random Processes for Electrical Engineering, 2nd edition, A. Leon-Garcia, 1994 (Addison-Wesley)
  - Nelson, “Stochastic Modeling, Analysis, and Simulation”, 1995
  - Hillier & Lieberman, “Intro to Oper. Research, 6th edition”, 1995

Randomness…(Review)

Sources of Variability:

- Natural variation
  - operator difference, tool wear, variation of material
- Breakdown/repair and other unexpected delays
- Setups and other irregular but expected delays
- Quality problems…

Problem: Variability in unbuffered lines lead to blocking and starvation, and thus loss of capacity

---

Random Processing Times without buffers

- Effect of Random Processing Times in balanced, *unbuffered* lines with no breakdowns (Askin/Standridge)
- CV is Coefficient of Variation = (std.dev.) / mean
- → *Reducing CV reduces our losses*

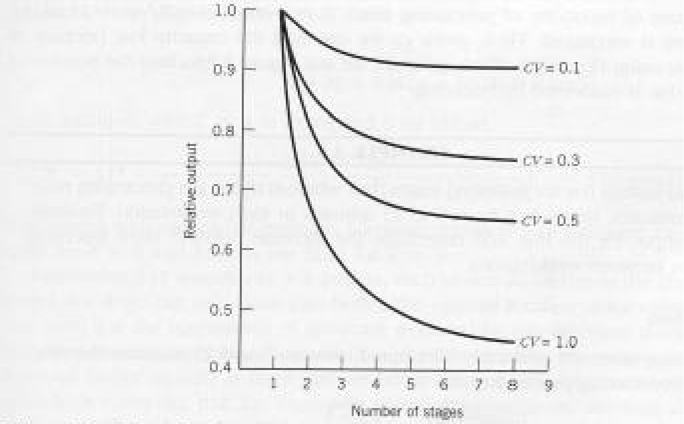


Figure 3.6 Effect of random processing time in balanced, unbuffered lines.

Recovery of lost output from buffering (review)

Buffers help reduce starvation and blocking due to time variations and disruptions.

→ Recovery of lost output, but at cost of adding WIP

→ Higher the variation, the more WIP required to recover

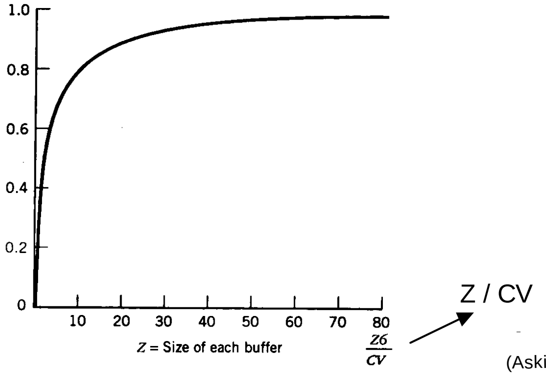
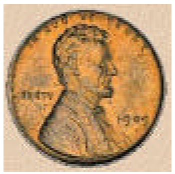


Figure 3.7 Proportion of lost output recovered by buffering in balanced lines.

(Askin and Standridge)

Example: Penny Fab 1

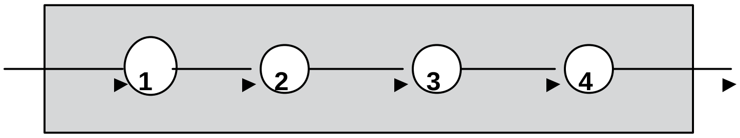
(from Hopp and Spearman)



Produce giant novelty pennies.

- **Process 1: Punch press cuts penny blanks**
- **Process 2: Stamping of images**
- **Process 3: Rim put on penny**
- **Process 4: Cleaning and deburring**

Suppose two hours per machine *(“perfectly balanced”)*
24 hours /day
r_b = ?
Raw Process Time, T_0 = ?
Critical WIP = ?



Best Case

Best case (zero variability) performance:

- **Minimum flow time for a given WIP given by:**

$$FT_{best} \begin{cases} = T_0 & \text{if } w < \text{critical} \quad \text{(no waiting)} \\ = w / r_b & \text{else} \end{cases}$$

Maximum Throughput rate for a given WIP given by:

$$rate_{best} \begin{cases} = w / T_0 & \text{if } w < \text{critical} \\ = r_b & \text{else} \end{cases}$$

This is the case for PennyFab 1
(perfectly balanced, no variability)

Worst Case (impractical)

Worst Case (impractical): Suppose average processing time per station was same, but actually all the time was spent with the first job.

Note: this is still not random, but we have picked the worst case possible

Example: 1st job = 8hrs; 2nd, 3rd, 4th take 0 hours.
Then: avg.=2 still, but 4th job must wait 8 hours at each station!

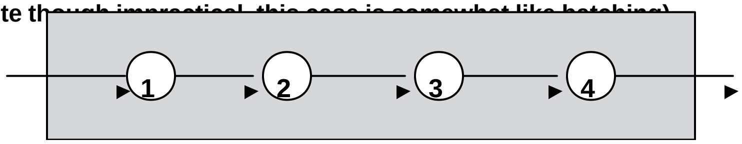
For given WIP w, define:

$FT_{worst} = w T_0$ **(waiting on all other WIP)**

$rate_{worst} = 1/T_0$

Example: For Penny Fab 1, we have to wait 8 hours each station
For w = 4, Flow time = 8 +8+8+8 = 32 hours = 4*8
Throughput rate = 4/32 = 1/8 jobs / hour

(Note though impractical, this case is somewhat like batching)



Practical Worst Case

- Real systems rarely operate like the best or worst case.
- To consider between the best and worst, we invoke probability.

- The derivation of this “practical worst case” will be done later. For now, we present the basic idea:
  - All stations perfectly balanced
  - All stations are single machines
  - Process times are random. In particular, assume that our system is “memoryless” – so the remaining time for a job at a station is independent of the time that the job has already been there.

- See graphs next page – derivations to be done with queueing material

---

WIP relationships (Hopp and Spearman, 1996)

- Flow time vs. WIP

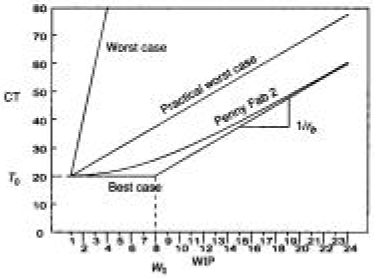


Figure 7.9
Cycle time versus WIP in Penny Fab Two

- rate vs. WIP

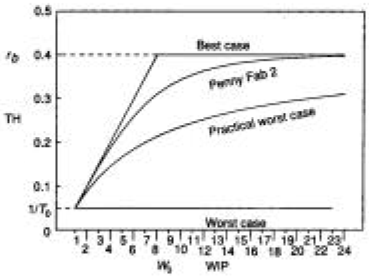


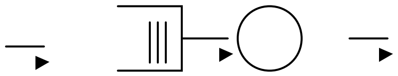
Figure 7.10
Throughput versus WIP in Penny Fab Two

Queueing models

- <u>**Queueing Models**</u>
  - **Probabilistic -- primarily steady state averages**
  - **Advantages:**
    - Analytical solution to problems with randomness & uncertainty
    - --> fast solution (*if solution exists or is known!)*
  - **Disadvantages:**
    - Requires simplifying assumptions
    - Best for smaller models
    - solutions not always possible
    - most suited for steady state analysis

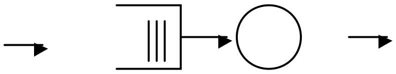
Queuing Theory: Basic Terms

- **Queuing Theory: Study of lines, waiting in lines**



- **Buffer or Queue**
- **Server -- service rate =** m
  - **machine**
  - **repair**
  - **(sometimes even parts)**

- **Customers --- arrival rate =** l
  - **parts waiting for machining**
  - **machines waiting for repair**
  - **...**



- **Issues: What is distribution of arrivals**
  - What is distribution of service time
  - What is policy for selecting next customer?

- **Queuing theory answers**
  - **expected number of customers in system**
  - **expected number of customers in queue**
  - **avg. customer waiting time in system**
  - **avg customer waiting time in queue**
  - **avg. utilization of server**

Queuing Theory

- **In this course, we focus *mostly* on simplest queue: M/M/1**
  - **M stands for *Markov* or *Memoryless***
  - **Notation: Arrival Process / Service Process / # of servers**

  - **M/M/1 -->**
    - Poisson arrivals: (exponential interarrival times)

      time distribution = $\lambda e^{-\lambda t}$ for t ³ 0

      λ = mean arrival rate (expected # arrivals / time)
    - Exponential service times

      time distribution = $\mu e^{-\mu t}$ for t ³ 0

      μ = mean service rate (expected # of customers completing service per unit time.

WIP relationships (review)

(Hopp and Spearman, 1996)

FIGURE 7.9

Cycle time versus WIP in Penny Fab Two

- Flow time vs. WIP

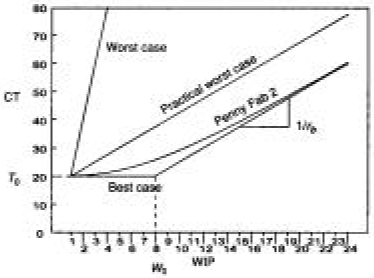
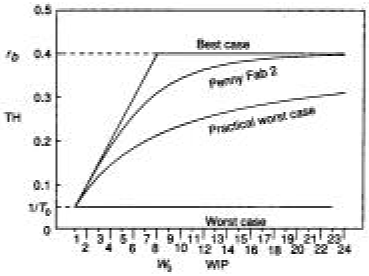


FIGURE 7.10

Throughput versus WIP in Penny Fab Two

- rate vs. WIP



"Practical Worst Case"

- **The "practical worst case" in previous diagrams corresponds to a balanced line of single machine stations with exponential process times.**

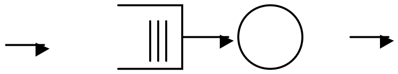
  (even worse can happen under bad batching)

- **"Exponential distribution" of times is "Markov" or "Memoryless".**
  - **This extreme randomness arises from:**
    - "random" arrivals (time to next is not given by time since prev.)
    - Highly variable service times – examples:
      - job shop or repair station
      - Setups

- **We will study models of behavior for this "practical worst case"**
  - → **Queueing models**
    - Provides analytical relationships of average waits, average lengths of queues, etc.
    - Results on only restricted classes of systems

Queuing Theory: Basic Terms

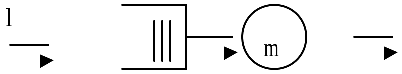
- **Queuing Theory: Study of lines, waiting in lines**



- **Buffer or Queue**
- **"Server" -- service rate =** μ
- **"Customers" --- arrival rate =** λ

In this course, we focus *mostly* on simplest queue: M/M/1
  - **(Notation: Arrival Process / Service Process / # of servers)**
  - **M is for "markov" or "memoryless"**
    - → Poisson arrivals (exponential inter-arrivals)
    - → Exponential service times
  - **"First Come First Served" (FIFO) queue discipline**

Utilization Assumption



- **We assume** λ/μ **< 1.**
  - **Why??**

$\rho = \lambda/\mu$ **is called *utilization***

Exponential Assumption

- **Exponential distributions give simplest queuing analysis**

- **Is exponential distribution reasonable?**

  For service time distribution:
  - **inappropriate if service same over most parts (as if same parts)**
  - **appropriate if variety of customers each requiring different operations**

  For Interarrival time distributions:
  - **Appropriate if customers appear in very random manner**
  - **Inappropriate if customers typically appear in groups, or if customers may postpone arrival based on length of queue**

Birth-Death Model

- **Quantities of interest in queuing theory:**

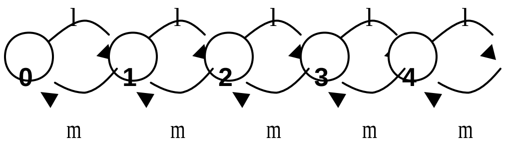
  P_n: probability of having *n* parts in system

  W: avg. wait in system

  L: avg. number of customers in system

  Consider steady-state averages

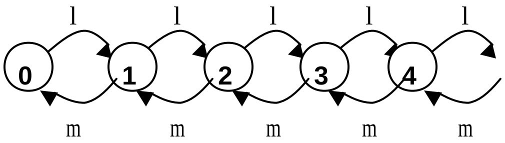
  Assume unbounded queue size



Consider steady-state averages: --> Freq. of entry to state same as frequency of departing from state.

Birth-Death Model

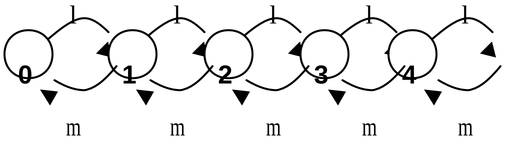
- Quantities of interest in queuing theory:
  - P_n: probability of having *n* parts in system
  - W: avg. wait in system
  - L: avg. number of customers in system
  - *Consider steady-state averages*
  - *Assume unbounded queue size*



Consider steady-state averages: --> Freq. of entry to state same as frequency of departing from state.

Deriving P_n from birth-death model

- Balance Equations: entry freq. = exit freq. for each state.



- Insert derivations

Results

- $P_0 = 1 - l/m = 1 - r$
- $P_n = (l/m)^n (1 - l/m) = r^n (1-r) = r^n P_0$ **(for n>0)**
- **Example 1: l = 5 parts/hr, m = 10 parts / hr.**
- **Example 2: l = 8 parts/hr, m = 10 parts / hr.**

Other questions:

- **What is probability that the machine is idle?**
- **What is the probability that there are more than 3 customers in the system?**

Avg. # of customers

- **L: = avg. number of customers *in the system***

$$L = E[n] = \sum_{n=0}^{\infty} nP_n$$

L = l / (m- l) = r / (1 - r)

Little’s Law

- **W = mean time spent in system**
- ***<u>Little’s law</u>*:**

L = $l_{\mathbf{eff}}$ **W**

mean number of customers in system =
effective arrival rate x mean wait time

mean # in “system” = arrival rate x ***mean wait in “system”***
mean # in queue = arrival rate x ***mean wait in queue***

Why the “effective arrival rate”?

- **Nice graphical proof of Little's Law found in Leon-Garcia, 2nd edition, pages 502-503.**

Mean Wait Time in System

- **W = mean wait time in system**

L = $\frac{\lambda}{(\mu-\lambda)} = \frac{\rho}{(1-\rho)}$

- **Little's law:** **L =** λ **W**
- **W = L /** λ **=** 1 **/** $(\mu - \lambda)$
- **Examples:**

- **Example: We have demand rate of 1 part per hour. We want average lead time per part under 2 hours. How fast a workstation do we need?**

Summary of results so far

M/M/1 queue:

- **exponential service times** **(service rate =** m **)**
- **Poisson arrivals** **(arrival rate =** l **)**
- **1 server, infinite buffer capacity**

- **P_0 = 1 -** $l/m = 1 - r$
- **P_n = (**l/m**)^n (1-** $l/m)$ $= r^n(1-r) = r^n$ **P_0** **(for n>0)**

- **L: = mean number of customers *in the system***

$$L = E[n] = \sum_{n=0}^{\infty} nP_n = \frac{l}{(m-l)} = \frac{r}{1-r}$$

- **W:= mean wait time in the system**

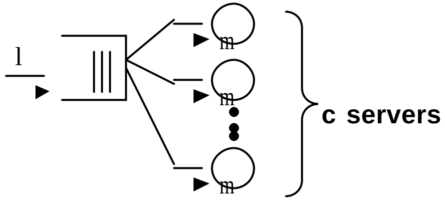
$$W = \frac{L}{l} = \frac{1}{(m-l)} = \frac{1}{m(1-r)}$$

Waiting time in Queue and # in Queue

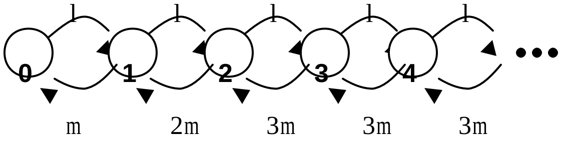
- W_Q : mean waiting time in just the queue
  = mean time in system -- mean time in service
  $W_Q = W - 1/m$
- L_Q : mean number of customers in the queue
- Use Little's Law:
  $L_Q = l\ W_Q$

Examples:

Multiple servers: M/M/c



- Utilization: $$r = \frac{l}{cm} < 1$$
- Basic intuition from birth-death graph (for c=3):



Basic results for M/M/c

- **(table 11.1 from Askin Standridge)**

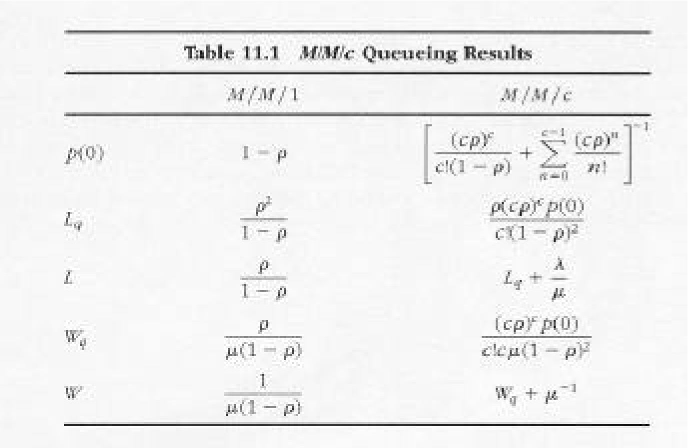


Table 11.1 *M/M/c* **Queueing Results**

	$M/M/1$	$M/M/c$
$p(0)$	$1-\rho$	$\left[\frac{(c\rho)^c}{c!(1-\rho)}+\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!}\right]^{-1}$
L_q	$\frac{\rho^2}{1-\rho}$	$\frac{\rho(c\rho)^c p(0)}{c!(1-\rho)^2}$
L	$\frac{\rho}{1-\rho}$	$L_q+\frac{\lambda}{\mu}$
W_q	$\frac{\rho}{\mu(1-\rho)}$	$\frac{(c\rho)^c p(0)}{c!c\mu(1-\rho)^2}$
W	$\frac{1}{\mu(1-\rho)}$	$W_q+\mu^{-1}$

Example of M/M/c

- **Suppose c=2, m = 5, l = 8**

$$r=\frac{l}{cm}=\frac{8}{10}$$

$$P_0=\left(\frac{(cr)^c}{c!(1-r)}+\sum_{n=0}^{c-1}\frac{(cr)^n}{n!}\right)^{-1}=\left(\frac{(1.6)^2}{2(.2)}+\left(1+\frac{1.6}{1}\right)\right)^{-1}=(9)^{-1}=0.1111$$

$$L_Q=\frac{r(cr)^c P_0}{c!(1-r)^2}=\frac{0.8(1.6)^2(0.111)}{2(.2)^2}=2.84$$

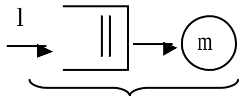
$$L=L_Q+\frac{l}{m}=2.84+1.6=4.44$$

$$W=\frac{L}{l}=\frac{4.44}{8}=0.555$$

$$W_Q=\frac{L_Q}{l}=\frac{4.44}{8}=0.355$$

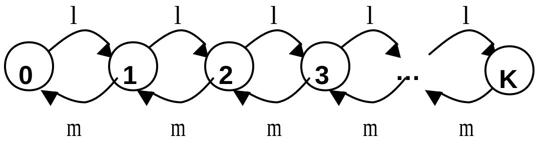
M/M/1/K

- **Finite capacity queues: State probabilities and # in system**



System capacity of K

  - **Birth-death model**



$$P_i = \rho^i P_0$$

$$1 = \sum_{i=0}^{K} P_i = \sum_{i=0}^{K} \rho^i P_0 \longrightarrow P_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

$$L = E[n] = \sum_{n=0}^{K} nP_n = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & for\ \rho \neq 1 \\ \dfrac{K}{2} & for\ \ \rho = 1 \end{cases}$$

- **M/M/1/K queues: mean wait in system:**
  - **Note some customers get turned away.**
  - **For Little's Law, we need *effective arrival rate***

    offered load is measure of demand on system: λ/μ
    carried load is actual demand met by system: λ_a/μ

$$\lambda_a = effective\ arrival\ rate = \lambda\ (1 - P_K)$$

$$W = \frac{L}{\lambda_a}$$

$$W_Q = W - \frac{1}{\mu}$$

$$L_Q = W_Q \lambda_a \qquad note\ it\ can\ be\ shown\ this\ is\ L_Q = L - (1 - P_0)$$

M/M/1/K example

- **Suppose K=6, m = 1/8, l = 1/10**

  P_i = ?

M/M/1/6	**M/M/1/infinity**
P0 = .25307	**P0=.2**
P1=.20246	**P1=.16**
P2=.16197	**P2=.128**
P3=.12957	**P3=.1024**
P4=.10366	**P4=.08192**
P5=.08293	**P5=.065536**
P6=.06634	**P6=.0524288**
Total P0 to P6: 1.0	**Total P0 to P6: .790284**

- **Key point: the limited size system turns away parts sometimes**

since the M/M/1/infinity system has probability (1-.79)= 21% of being in state beyond 6.

- V(N) = E(N) + },.2V(S) + ~ + },.4[E(S2)r 3(1 -p) 4(1 - p)2
- V(N) = E(N) + },.2V(5) + ~ + },.4[E(52)]2 3(1 -p) 4(1 -p)2 (11.9)

M/G/1 Queues

- **(from page 367, Askin and Standridge) – FCFS (FIFO)**

$$E(T) = E(S) + \frac{\lambda E(S^2)}{2(1-\rho)} \qquad (11.6)$$

$$V(T) = V(S) + \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2[E(S^2)]^2}{4(1-\rho)^2} \qquad (11.7)$$

$$E(N) = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \qquad (11.8)$$

$$V(N) = E(N) + \lambda^2 V(S) + \frac{\lambda^3 E(S^3)}{3(1-\rho)} + \frac{\lambda^4[E(S^2)]^2}{4(1-\rho)^2} \qquad (11.9)$$

- S is for service time (so E(S) = 1/μ)
- T is for throughput time (so E(T) = W)
- N is for number of jobs (so E(N) = L)

Note reducing E(S^2) reduces E(T) and E(N)

Note that E(T) and E(N) depend only of variance and mean of service time distribution, not its distributional form.

M/G/1 Queues

- **(from page 367, Askin and Standridge)**

$$E(T) = E(S) + \frac{\lambda E(S^2)}{2(1-\rho)} \tag{11.6}$$

$$V(T) = V(S) + \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2[E(S^2)]^2}{4(1-\rho)^2} \tag{11.7}$$

$$E(N) = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \tag{11.8}$$

Example from text: λ = 10/week. All distributions with same mean
Key points: the exponential distribution gives the worst case out the distributions with comparable means. It is thus useful as a practical worst case (maximum randomness) estimate.

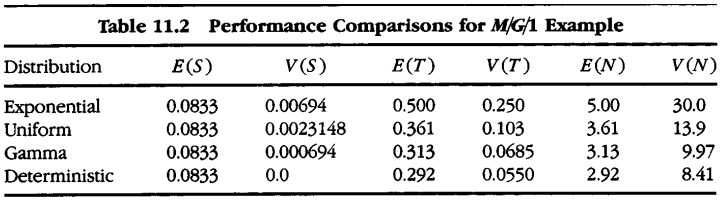


Table 11.2 Performance Comparisons for *M/G/*1 Example

Distribution	$E(S)$	$V(S)$	$E(T)$	$V(T)$	$E(N)$	$V(N)$
Exponential	0.0833	0.00694	0.500	0.250	5.00	30.0
Uniform	0.0833	0.0023148	0.361	0.103	3.61	13.9
Gamma	0.0833	0.000694	0.313	0.0685	3.13	9.97
Deterministic	0.0833	0.0	0.292	0.0550	2.92	8.41

Review:

- **M/M/1 queue:**
  L = $\lambda/(\mu-\lambda) = \rho/(1-\rho)$
  W = L / λ **=** $1/(\mu-\lambda)$
  $W_Q = W - 1/\mu$
  $L_Q = \lambda W_Q$

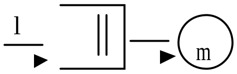
- **M/M/1/K**
  L =

  $\lambda_a = \lambda(1 - P_K)$
  $W = L / \lambda_a$
  $W_Q = W - 1/\mu$
  $L_Q = \lambda_a W_Q$

Effect of many small (many queue) vs. single machine

Effect of many small (multi-queue) vs. multi-server (same queue)

Summary



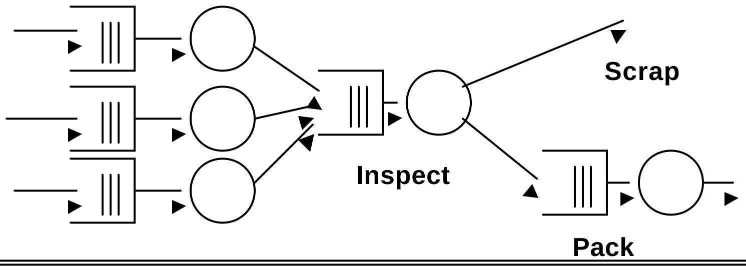
- M/M/1 Queues
- M/M/c Queues ---- c servers
- M/G/1 Queues ---- generalized service distributions
- M/M/1/K Queues ---- Limited capacity in system

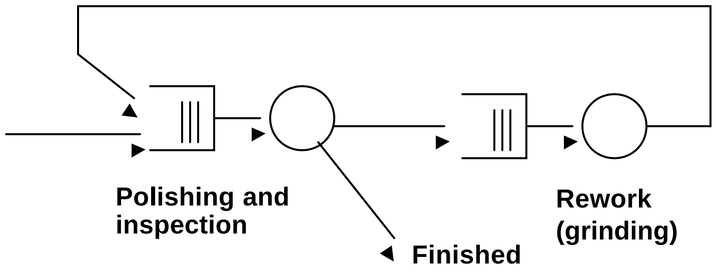
Systems with multiple queues

- Open networks vs. closed networks

- ***open Jackson networks:***
  - open network
  - Poisson external arrivals
  - exponential service times (possibly c>1)
  - unlimited buffers at each node
  - probabilistic routing
  - First come – First Serve
  - one type of customer

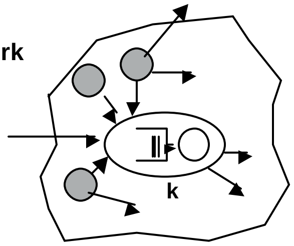
Examples





Open Jackson Networks

- **Why Jackson networks?**
  - **Probability of system state can be calculated from looking at the individual queues in the system.**
  - ***Under steady state conditions, each facility in a Jackson network behaves as if it were an <u>independent</u> M/M/s queuing system with effective arrival rate:***
    - $\lambda_k = a_k + \Sigma_{i=1...m} \lambda_i p_{ik}$
    - p_{ik} **probability of routing from i to k**
    - a_k **arrival rate in from outside**
    - λ_k **arrival rate at node k**
    - **m number of nodes in network**



Jackson Network Solution Procedure

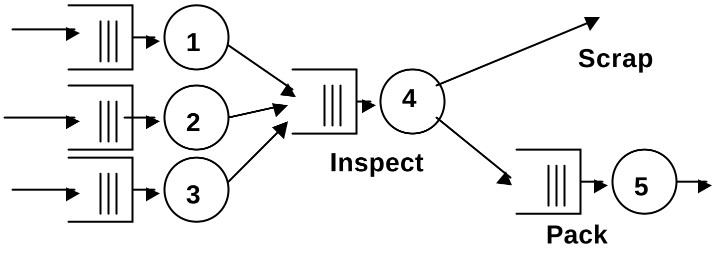
1. **Solve for arrival rates using**

$$\lambda_k = a_k + \Sigma_{i=1...m} \lambda_i p_{ik}$$

1. **Analyze each workstation as a M/M/1 or M/M/c queue**
   Use standard table of L, W, L_Q, W_Q formulas

2. **Combine results across workstations to get performance measures of the entire system.**

Queueing Network example 1

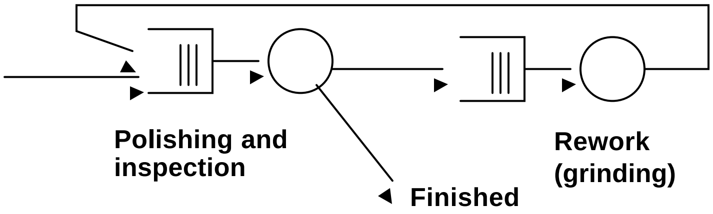


Scrap

Inspect

Pack

Queueing Network example 2

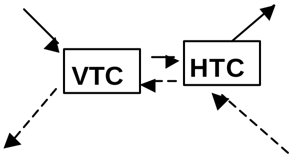


Polishing and inspection

Finished

Rework (grinding)

Queuing network example 3



	Part A	*Part B*
Path	VTC-HTC-exit	HTC-VTC-exit
Demand	Poisson 30/day	Poisson 60/day
VTC time (min.)	uniform(8,12)	uniform(8,12)
HTC time (min.)	uniform(10,14)	uniform(6,10)

Extensions

Networks with general distribution service times.