What's the Use of Stochastic Models of Manufacturing Systems?

John Buzacott
York University

TU Eindhoven
24 June 2010

Decisions about Manufacturing Systems

- Strategic
  - R&D Priorities
  - Investment
    - Capacity
    - Configuration
- Tactical
  - Inventory levels
  - Work force size
- Operational
  - Scheduling

R&D Priorities

- Issues
  - How to achieve a significant improvement in performance
    - Productivity?
    - Quality?
  - How to meet market needs better
    - Customization?
  - How to reduce and share risks
    - Demand?
    - Technology?
    - Finance?

Investment

- Capacity
- Configuration
- Costs
- Flexibility
  - Product life cycle
  - Demand
  - Disruptions
  - Finance

Inventories

- What to stock where
- How much to stock
- When and how to replenish
- How much storage space
- Who owns inventories
- What payment structure

Scheduling

- What job to do next on a machine
- What jobs to release when
- Moving jobs
- Machine, job and worker status:
  - how accurate and timely?

How do Manufacturing System Models Help?

- 60-70 years of development
- How do they help decision making?
- Are they used much?
- Opportunities met and missed

Manufacturing System Models

- Simulation
- Mathematical Models
  - Deterministic
  - Stochastic

Types of Mathematical Models

- Deterministic models
  - Usually Linear or Non-linear programming
  - Use to determine bottlenecks
  - Also use for investment analysis (higher level decisions)
- Stochastic models (represent variability)
  - Sample path models (find general impact of parameters - e.g., does performance improve?)
  - Queueing theory based models
    - Determine queue lengths (hence inventories and backlogs)
    - Models may be
      - Exact (give correct answer given assumptions)
      - Approximate (contain a step which violates an assumption)

Simulation vs Analytical Models

- Simulation:
  - Advantages:
    - Representation of detail of system design and operation
    - Ability to change parameter values
  - Disadvantages
    - Statistical aspects
      - Experimental design
      - Run length
    - Computing resources
    - Validation and verification of model

Simulation vs Analytical Models (cont.)

- Mathematical (queueing/stochastic and deterministic) models
  - Advantages
    - May give formula for performance
      - Ease of finding impact of changes
    - Insights into system behavior
    - Easier to explore impact of changes in system design
  - Disadvantages
    - (Usually) only simple systems can be modelled
    - Require considerable skill to develop
    - Time to develop usually quite uncertain

How Approaches Used

- Simulation: (*Industry preference*)
  - Detailed design
  - Exploring alternative operating procedures and maintenance strategies
  - Predicting performance
- Analytical Models: (*Academic preference*)
  - Preliminary design
  - Exploring different concepts
  - Gaining insight

Decisions about Manufacturing Systems – Typical Modelling Approaches

- Strategic
  - R&D Priorities (*economic models or stochastic models*) IIASA
  - Investment
    - Capacity (*stochastic programming*) Dofasco
    - Configuration (*simulation* or *stochastic models*) GM
- Tactical Stelco
  - Inventory levels (*stochastic optimization*) Shell
  - Work force size (*LP or MIP*) Hotpoint
- Operational
  - Scheduling (*simulation*) BTH/IBM Stelco

Summary 1

- Stochastic models focus on capturing impact of variability
- If they can be developed they are easy to apply and enable many alternative designs to be compared
- BUT they are not easy to develop and not all systems can be modelled
- Mostly only toy systems can be modelled
- Insights gained usually make it worthwhile using them

What Queueing Models Exist?

- Models of
  - Job shops
  - Flow lines
  - Closed loop manufacturing systems
  - Supply chains

  But in each case not all systems can be modelled and models do not allow all ways of system can be operated to be represented

Job Shop Models

- Job shops consist of
  - Machines
  - Jobs
- Jobs move between machines
  - Sequence of machines visited may differ between jobs
- Jobs wait in queues in front of each machine where they require processing

Reducing Model Complexity

- Aggregation of jobs
  - Different jobs have different routing
  - If keep track of every type of job model may have too much detail to handle
  - So replace multiple job types by a single (aggregated) job type
  - -> Probabilistic job routing
- Aggregation of machines
  - Replace a group of machines by a single equivalent machine
  - Service time at aggregated machine represents service times at individual machines
  - Sometimes this requires that the aggregated machine has a service rate that depends on the number of jobs in the queue at the machine

Illustration

- 3 machines in parallel
- Aggregate and replace 3 machines by a single equivalent machine
- Equivalent machine has a service rate depending on number in queue

Number of jobs at machines	Service rate
1	μ
2	2μ
3 or more	3μ

The Simplest Job Shop Model: Jackson queueing network

- Assumptions:
  - Number of machines = m
  - Poisson (random) arrivals of jobs to system (rate λ)
  - Routing depends only on current machine (probabilistic routing p_{ij}= prob. $i \Rightarrow j$)
  - Routing of external arrivals given by q_i
  - Exponential service time at machines (rate may be queue length dependent)(rate μ)
  - No limits on queue lengths at machines

Jackson Model Results: single server queues

- $P(n^1,n^2,\ldots,n^m)=\rho_1^{n1}\,\rho_2^{n2}\ldots\rho_m^{nm}P(0,0,\ldots,0)$ where $P(0,0,\ldots,0)=(1-\rho_1)\,(1-\rho_2)\ldots(1-\rho_m)$
- $\rho_i=\lambda_i/\mu_i$ where λ_i is determined by the equations

  $\lambda_i=q_i\lambda+\Sigma_j p_{ji}\lambda_j$ for i=1,2,..,m
- Note product form of solution, i.e., $P(n^1,n^2,\ldots,n^m)=P(n^1)P(n^2)\ldots P(n^m)$ with $P(n^i)$ same distribution of queue lengths as an M/M/1 queue

Networks: GI/GI/1 Approximation

- What if arrivals are not Poisson or service times not exponential?
- Approximation:
  - Assume product form solution still applies
  - Model each machine by a GI/GI/1 queue
    - Use GI/GI/1 approximation or bound to determine queue length
    - Distribution of service times (or at least 1st and 2nd moment) at machine known
    - Need to find distribution of interarrival times at machines
      - 1st moment: use same approach as Jackson network
      - 2nd moment:
        - » (1) need scv of departures from each machine
        - » (2) need scv of flow from machine j to machine I
        - » (3) need scv of superposition of arrival streams to machine i

2nd moment determination

- (1) scv of departures C_d^2 if mean waiting time $E[T]$

$$C_d^2 = C_a^2 + 2\rho(1-\rho) - E[T]2\lambda(1-\rho)$$

- (2) scv of selecting from departure stream with probability p_{ji}

$$C_a^2(j \to i) = 1 - p_{ji} + p_{ji}C_d^2$$

$$C_a^2(0 \to i) = 1 - q_i + q_i C_a^2$$

- (3) scv of combining arrival streams at i

$$\lambda_i C_{ai}^2 = \sum_{j=1,\, j\neq i}^{m} \lambda_j p_{ji} C_a^2(j \to i) + \lambda q_i C_a^2(0 \to i)$$

Application
When to use flow lines

- Compare two systems when there are two tasks X and Y to be done
  - (A) Flow line: machine 1 does task X, machine 2 does task Y
  - (B) Parallel system: machine 1 does tasks X and Y, machine 2 does tasks X and Y, jobs allocated randomly to machines
- Flow line:
  - Machine 1: $E[N_1] = \frac{C_a^2 + C_S^2}{2(1-\rho)}; \qquad C_d^2 = C_S^2$
  - Machine 2: $E[N_2] = \frac{C_S^2 + C_S^2}{2(1-\rho)}$
- Parallel system: $E[N^p] = 2\frac{(1/2 + C_a^2/2) + C_S^2/2}{2(1-\rho)}$
- Flow line better if $C_S^2 < 1/2$

Flow Line Models

- Jackson networks assume no limits on storage
  - Blocking occurs if storage limited
    - Job finished at machine but no space in downstream store
  - Flow line models should represent blocking
- Two main applications
    - Transfer lines: machine failure and repair
    - Flow lines with human operators
      - People show characteristic variability in times to perform repetitive tasks
        - » Variation within one person
        - » Variation between people

Simple two machine model of blocking

- Each machine takes an exponentially distributed time to perform task
  - Machine i (i=1,2) has parameter λ_i
- Store has capacity *z*,
- System state = *N*, $N \in \{I, 0, 1, ..., z, B\}$
- $$P\{N = j\} = \frac{r^{j+1}(1-r)}{1-r^{z+3}}, r = \lambda_1 / \lambda_2$$
- Throughput (maximum production rate) given by

$$TH = \lambda_1 \frac{(1-r^{z+2})}{1-r^{z+3}}, r \neq 1$$

$$TH = \lambda \frac{(z+2)}{z+3}, r = 1$$

More than 2 stages

- No longer get simple product form solution
- Good approximation - use the idea of a ***stopped arrival*** queue
- In a stopped arrival queue the arrival process is switched off, once the queue length reaches a critical number
- If time between arrivals not exponential, *stopped arrival* not the same as *lost arrival*
- With lost arrival distribution of time between admissions hard to determine, with stopped arrival know that once queue length falls below critical number time to next admission same as time to next arrival
- Stopped arrival queues are reversible, lost arrival queues are not. (*reversible*: same throughput if interchange arrival and service processes)

M/G/1/N Stopped Arrival Queues

- M/G/1/N stopped (and lost) arrival queues

$$p_N(n) = \begin{cases} \dfrac{p_\infty(n)}{1-\rho P_\infty(N)}, n = 0,1,\ldots,N-1; P_\infty(N) = \displaystyle\sum_{n=N}^{\infty} p_\infty(n) \\ \dfrac{(1-\rho)P_\infty(N)}{1-\rho P_\infty(N)}, n = N \end{cases}$$

GI/GI/1/N Stopped Arrival Queue Approximation

- (1) Use GI/GI/1/∞ approximation

$$p_\infty(n) = \sigma p_\infty(n-1), n = 2,3,...$$

also $p_\infty(0)=1-\rho$, so $p_\infty(1)= \rho(1-\sigma)$

and the average queue length

$$\hat{n} = \frac{\rho}{1-\sigma} \quad \text{or} \quad \sigma = \frac{\hat{n}-\rho}{\hat{n}}$$

Thus, given an approximation for $\hat{n}$ can find an approximation for the queue length distribution.

- (2) Assume queue length distributions in stopped arrival GI/GI/1/N queue and in GI/GI/1/∞ queue have same relationship as exists between a stopped arrival M/G/1/N queue and an M/G/1/∞ queue. Thus

$$p_N(n) = \begin{cases} \dfrac{1-\rho}{1-\rho^2\sigma^{N-1}} & n = 0 \\ \dfrac{\rho\sigma^{n-1}(1-\sigma)}{1-\rho^2\sigma^{N-1}} & n = 1,2,..,N-1 \\ \dfrac{(1-\rho)\rho\sigma^{N-1}}{1-\rho^2\sigma^{N-1}} & n = N \end{cases}$$

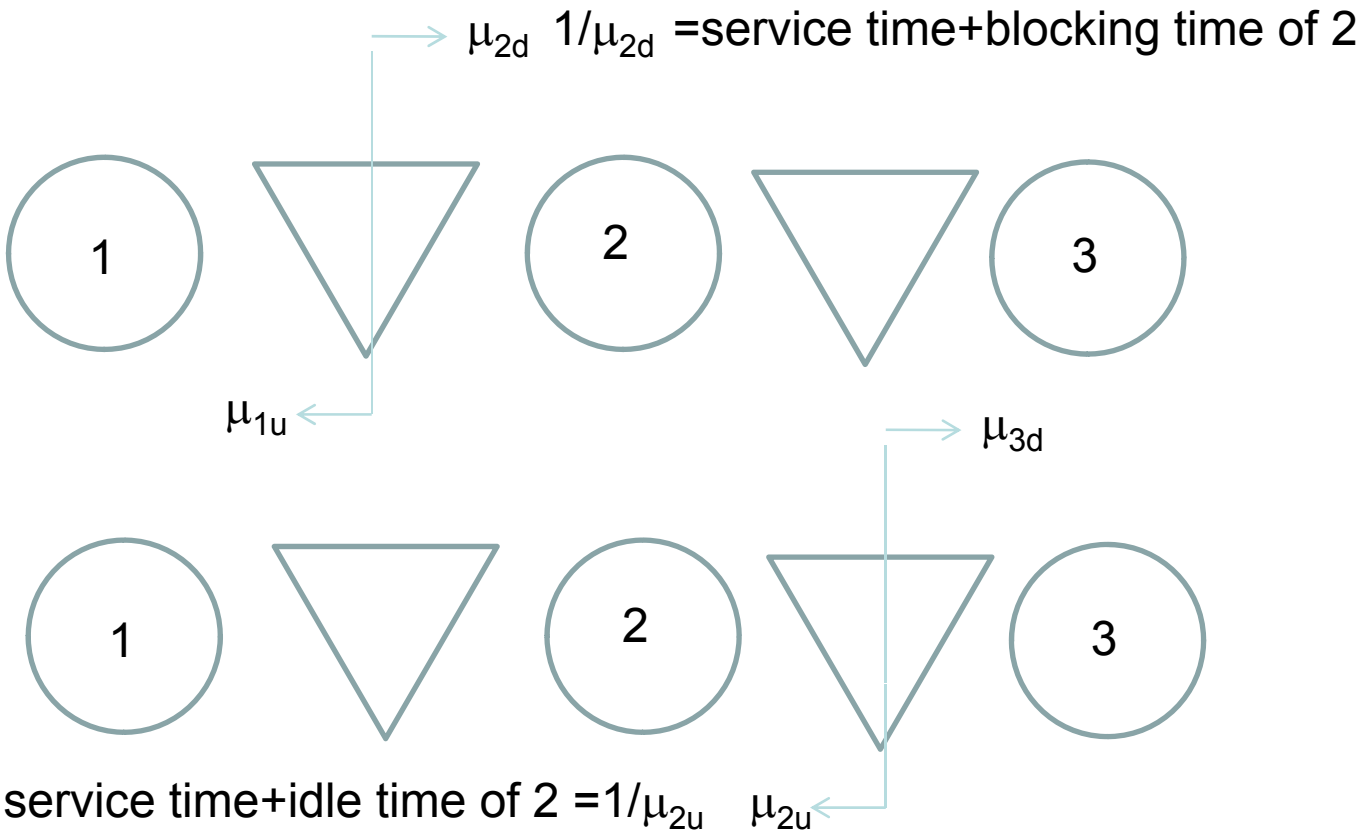
Throughput Approximation

- The throughput if ρ<1 is thus given by

$$TH = \frac{\lambda(1-\rho\sigma^{N-1})}{1-\rho^2\sigma^{N-1}}, \rho < 1$$

- If ρ>1, use reversibility and look at queue in opposite direction, so define ρ_R=1/ ρ and $\sigma_R = (\hat{n}_R - \rho_R)/\hat{n}_R$ where $\hat{n}_R$ is the average queue length in the reversed GI/GI/1/∞ queue.
- The throughput is then $TH = \frac{\mu(1-\rho_R\sigma_R^{N-1})}{1-\rho_R^2\sigma_R^{N-1}}, \rho > 1$
- If ρ=1, σ=1, so $TH = \frac{\lambda(1+(N-1)\nu}{2+(N-1)\nu}, \rho = 1, \nu = \lim_{\rho\to 1}\frac{d\sigma}{d\rho}$

Flow Line Approximation (Sevastyanov, Zimmern)



$$1/TH_i = 1/\mu_{iu} + 1/\mu_{id} - 1/\mu_l$$

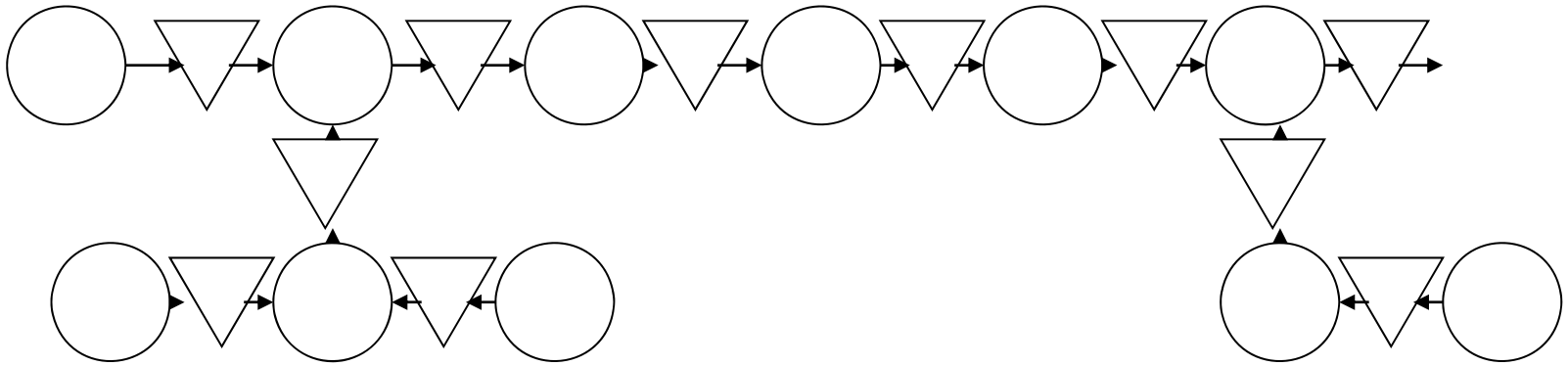
Flow Line Approximation

- An *m* stage line can be viewed as *m*-1 two stage lines by standing at each buffer (Zimmern 1956, Sevastyanov 1962) where buffers i=2,…,m
- Mean time between arrivals at buffer i =$1/\mu_{i-1u}$, mean time between departures buffer i=$1/\mu_{id}$,
- Time between departures from stage i=idle time+service time+blocking time, idle time+service time= $1/\mu_{iu}$, service time+blocking time=$1/\mu_{id}$, thus $1/TH_i = 1/\mu_{iu}+1/\mu_{id}-1/\mu_l$
- With $\rho_i = \mu_{i-1u}/\mu_{id}$ can determine TH_i (assuming know scv's)
- Also equations $TH = TH_{i,}$ for i=1,…,m-1
- Gives 3m-4 equations in 3m-4 unknowns so can solve recursively to determine TH and average inventory levels

Allowing for scv

- Two approaches
  - Simple: assume scv of upstream line from buffer i is scv of service time at stage i, scv of downstream line is scv of stage i+1.
  - More complicated: determine scv of (service time + idle time), scv of (service time + blocking time)
- Simple approach often quite adequate

Application-GM Car Body Assembly



A Complex Simulation of the Entire Car Plant; developed by combining simulation models of individual departments

Validation: Use stopped arrival model + some extensions to allow for assembly
While stopped arrival model is also an approximation, know that it is usually within 5% of actual, also stopped arrival model gives insights into role of each inventory bank

Closed Loop Systems

- Jackson network results:
  - If exponential service and random routing get product form solution for queue length distribution
  - Need visit ratio – define load/unload station as having visit ratio 1. For station j determine v_j by $$v_j = q_j + \sum_{i=1, i \neq j}^{m} v_i p_{ij}$$
  - Product form solution (K normalizing constant) $$p(n_0, n_1, .., n_m) = K \prod_{i=0}^{m} p_i(n_i)$$

$$p_i(n_i) = \left(\frac{v_i}{\mu_i} \right)^{n_i}$$

Solution using Mean Value Analysis

- Recursive approach for number of pallets (work carriers) – 1,2,..,n
- Based on observation that the probability of a system state seen by an arrival at station i is equal to the probability of that state in a system with one less pallet
- $E[T_i(l)]$ = time spent by a job at station i if l pallets in system

$$E[T_i(l)] = (E[N_i(l-1)]+1)/\mu_i$$

$$TH(l) = l/\left[\sum_{i=0}^{m} v_i E[T_i(l)]\right]$$

$$E[N_i(l)] = v_i TH(l) E[T_i(l)]$$

Extended Mean Value Analysis

- Approximation for non-exponential service times

$$E[T_i(l)] = \{E[N_i(l-1)] - v_i TH(l-1)E[S_i] + 1\}E[S_i] + v_i TH(l-1)E[S_i^2]/2, i = 0,\ldots,m$$

$$TH(l) = \frac{l}{\sum_{i=0}^{m} v_i E[T_i(l)]}$$

$$E[N_i(l)] = v_i TH(l)E[T_i(l)], i = 0,\ldots,m$$

Non exponential service times and multiple classes of customers

- Use Extended Multiclass MVA (approximation)
- *p* classes, one load/unload station+*m* machines

$$\lambda_i^{(r)}(\mathbf{l}) = v_i^{(r)} TH_i(\mathbf{l})$$

$$\lambda_i(\mathbf{l}) = \sum_{r=1}^{p} \lambda_i^{(r)}(\mathbf{l})$$

$$E[T_{i,s}(\mathbf{l}+\mathbf{e}_s)] = E[S_i^{(s)}] + \sum_{r=1}^{p} (E[N_{i,r}(\mathbf{l})] - \lambda_i^{(r)} E[S_i^{(r)}(\mathbf{l})]) E[S_i^{(r)}] + \lambda_i(\mathbf{l}) E[S_i(\mathbf{l})] \left(\frac{E[S_i^2(\mathbf{l})]}{2E[S_i(\mathbf{l})]} \right), i = 0,\ldots,m$$

$$E[S_i(\mathbf{l})] = \frac{1}{\lambda_i(\mathbf{l})} \sum_{r=1}^{p} \lambda_i^{(r)}(\mathbf{l}) E[S_i^r], i = 0,\ldots,m$$

$$E[S_i^2(\mathbf{l})] = \frac{1}{\lambda_i(\mathbf{l})} \sum_{r=1}^{p} \lambda_i^{(r)}(\mathbf{l}) E[(S_i^r)^2], i = 0,\ldots,m.$$

$$TH_s(\mathbf{l}+\mathbf{e}_s) = \frac{l_s + 1}{\sum_{i=0}^{m} v_i^{(s)} E[T_{i,s}(\mathbf{l}+\mathbf{e}_s)}$$

$$E[N_{is}(\mathbf{l}+\mathbf{e}_s)] = v_i^{(s)} TH_i(\mathbf{l}+\mathbf{e}_s) E[T_{i,s}(\mathbf{l}+\mathbf{e}_s)]$$

Results

- Errors usually about 5%
- Multiclass systems have some unusual properties:
  - TH for class *j* may decrease if increase number of pallets of class $k \neq j$
  - Total throughput can decrease if increase number of pallets of some class

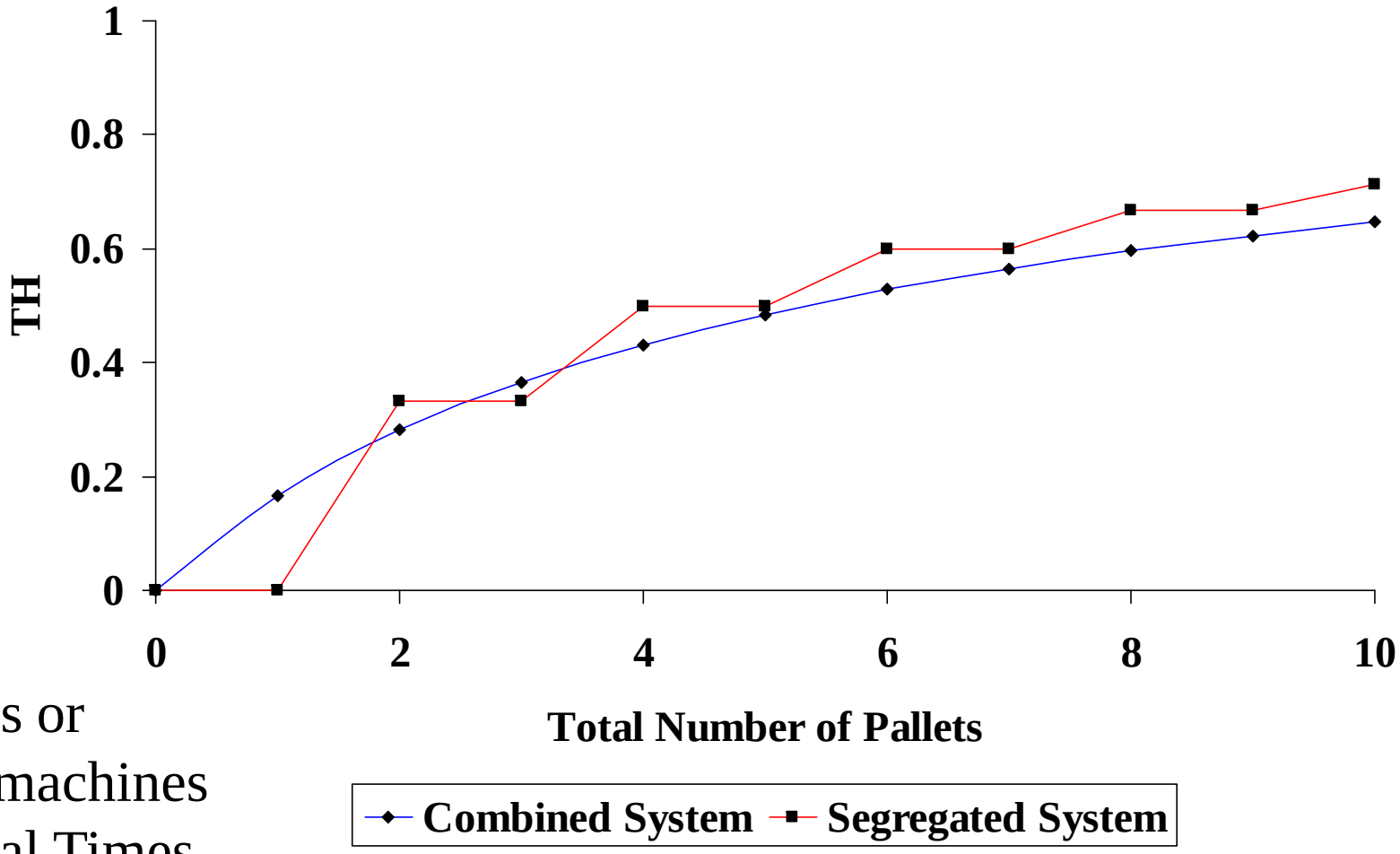
Application: Why the FMS concept did not meet its promise

- Around 1980 Flexible Manufacturing systems were viewed as the answer to manufacturing problems in batch production
- But experience showed that they did not meet the hype and promise – software problems, management problems, etc.
- But queueing models can show that the FMS concept is fundamentally flawed

Models

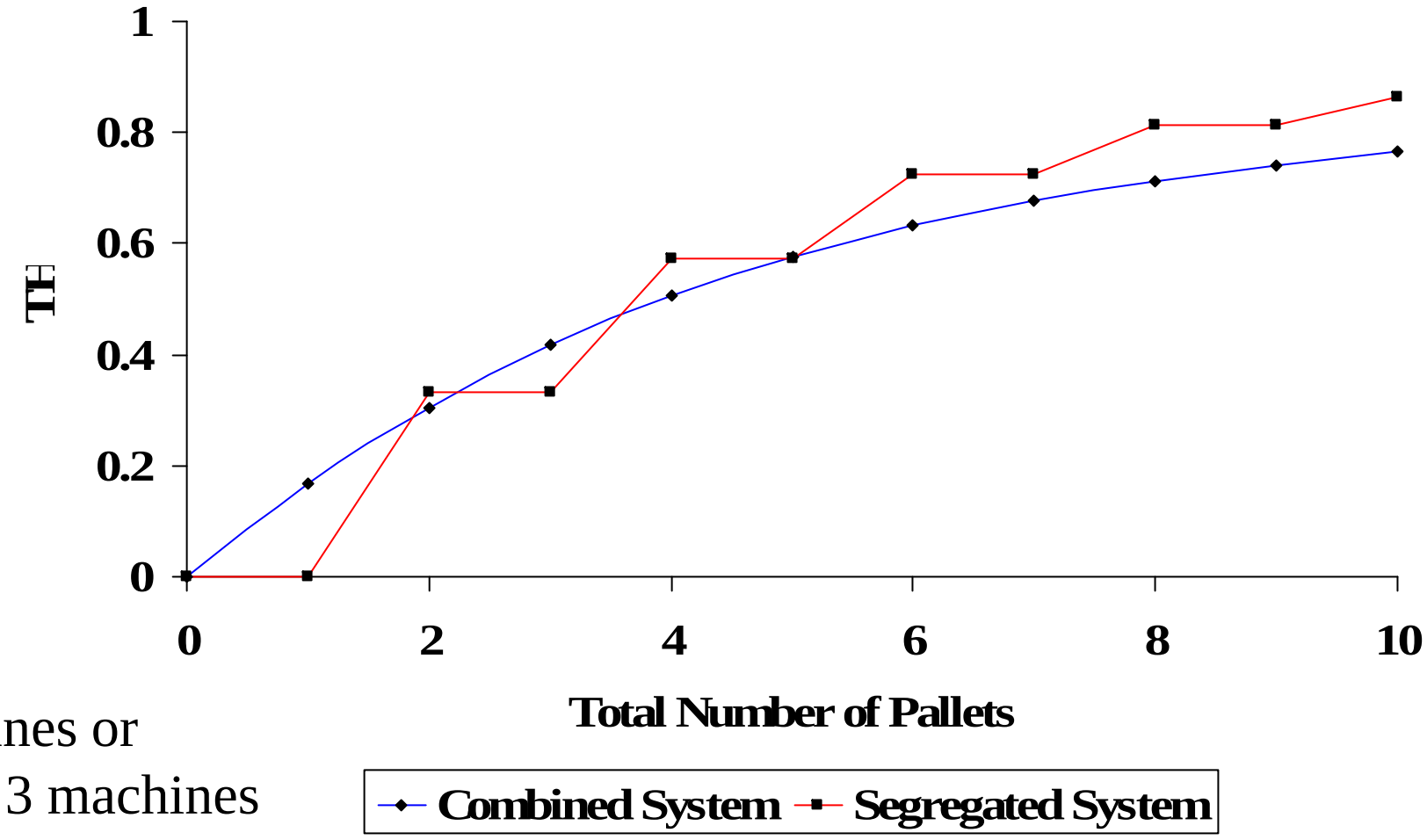
- Suppose FMS processes two job classes, one class has longer service times than the other, but scv of service times of the two classes are the same.
  - Mixing the two job classes increases the scv of the service time of a job at any machine
- Compare with two separate systems each dedicated to one class. Assume machines process at half the speed in the separate system

Combined or Segregated Systems



6 machines or
Two X 3 machines
Exponential Times

Combined or Segregated Systems



6 machines or
Two X 3 machines
Deterministic Times

Separate Pallets

- **Two job classes, 6 machines, deterministic times**
- Common pallets (10) TH = 0.766 (frac cl 1 =2/3)
- Separate pallets (4,6) TH = 0.671 (frac cl 1 = 0.466)
- Separate pallets (5,5) TH = 0.718 (frac cl 1 = 0.569)
- Separate pallets (6,4) TH = 0.770 (frac cl 1 = 0.668)

- **Two separate systems, each with 3 machines, times doubled**
- Each 5 pallets TH = 0.865 (frac cl 1 = 2/3)

Multiple Pallet Types

- Throughput penalty not necessarily large but getting pallet allocation right difficult
- Scheduling different pallet types also a concern
- Note that separate systems still much better

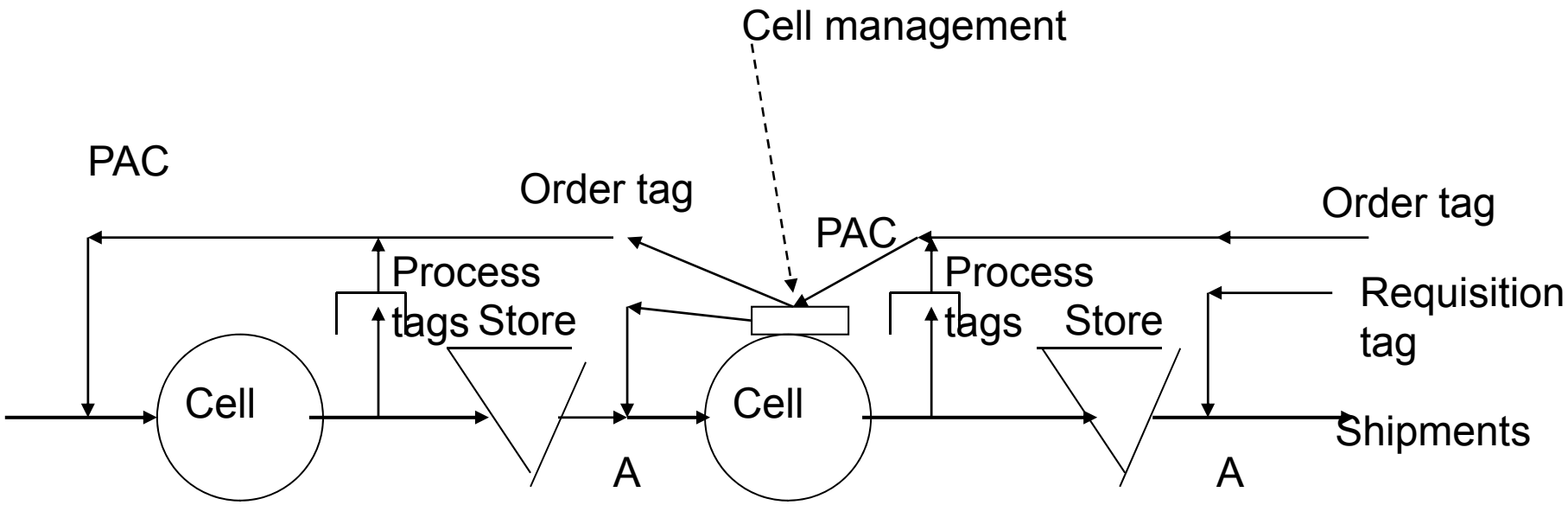
Model Implications

- Keep the range of service times as small as possible
- Small systems with very similar part types are better than large systems with diverse part types
- Multiple pallet types require complex scheduling and it is hard to get the number of pallets of each type correct
- -> Large FMS difficult to manage and schedule
- -> Small flexible cells inherently better

Modelling Supply Chains

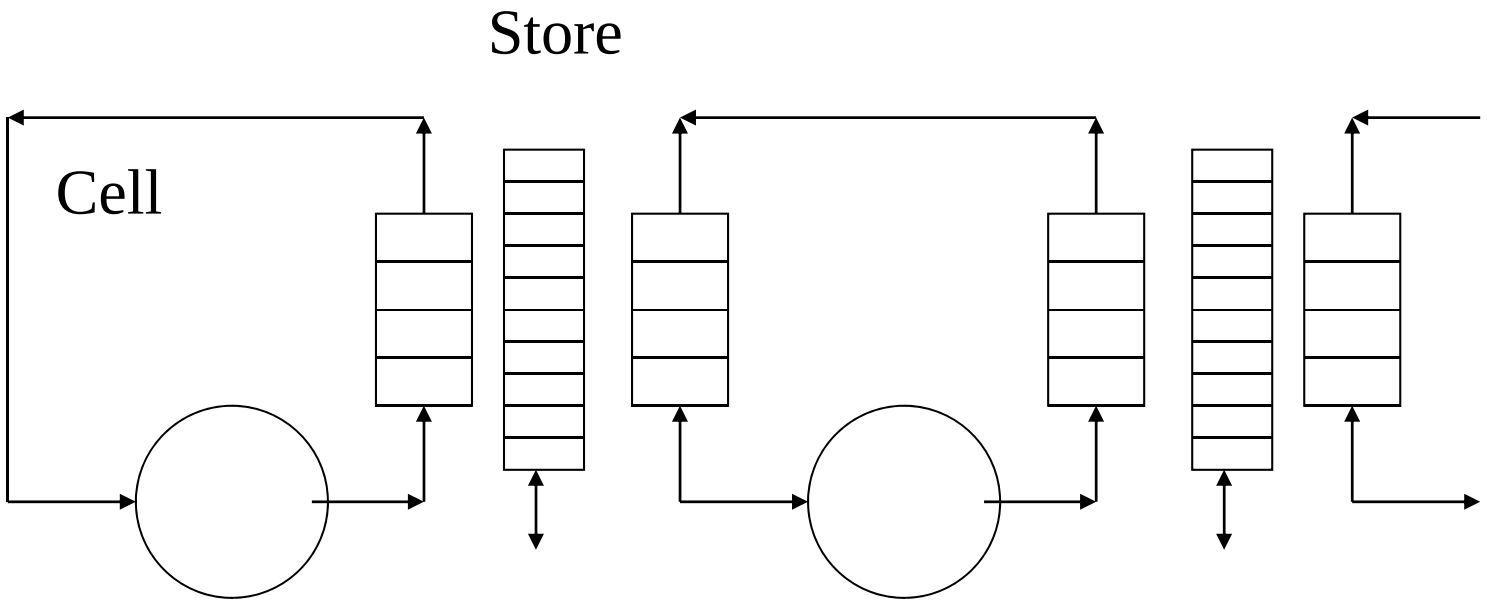
- Need to model flow of material AND flow of information
- Material processing occurs in cells
- Material stored between cells
- Flow determined by demands (orders, requisitions) or by forecasts of demand arrivals
- Models must represent both material flow and information flow

Cells in Series



Cell Management controls timing of requisition tag
(and also batching of demands)

(Approximate) Performance Models

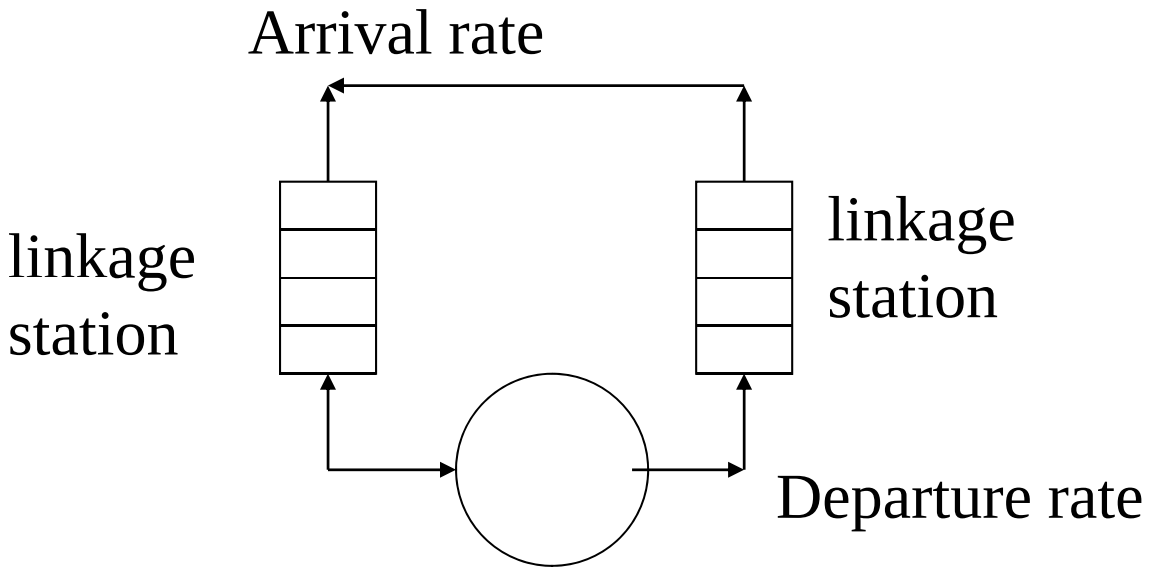


Two components to model:

A. Model of circulation in a cell

B. Model of variation of inventory position in a store

Cell Model

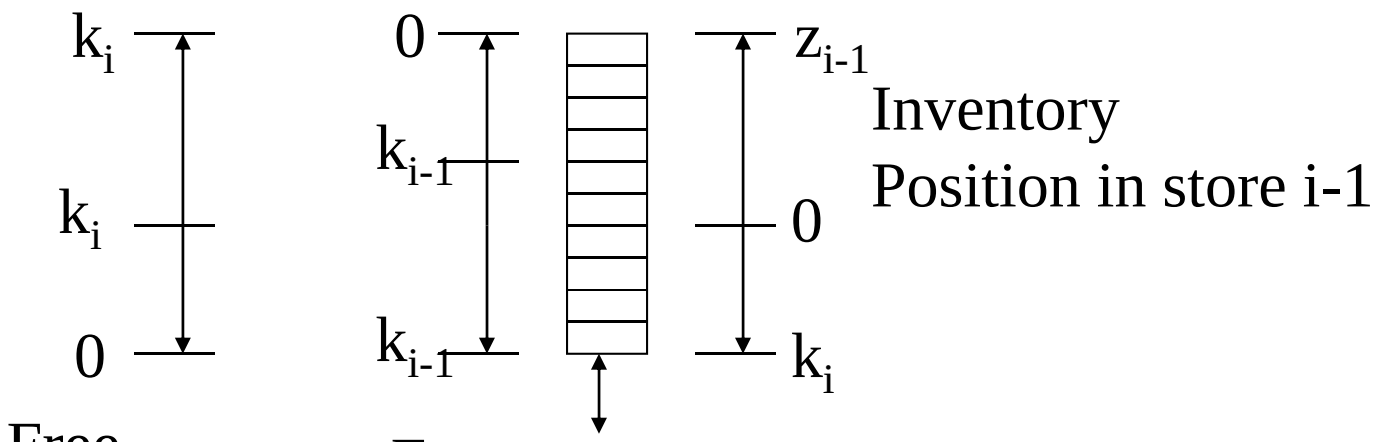


Use a closed queueing model with k_i customers

Slight difference: if no queue on arrival at linkage station

service time may be zero (zero at upstream if inventory position > 0 in store i-1, zero downstream if information position less than 0)

Store i-1 Model



Rate of increase of inventory position = departure rate from cell i-1
given no. of free cell i-1 Kanbans

Rate of decrease in inventory position = arrival rate of tags in cell i
given no. of free cell I Kanbans

Insights into MRP Controlled Systems

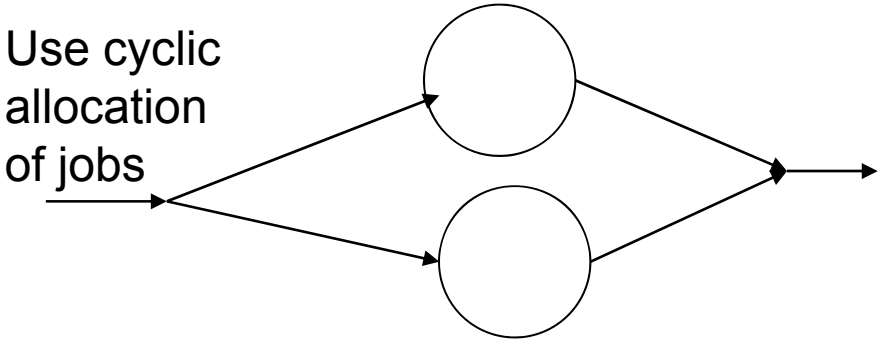
- If have perfect forecast safety time better than safety stock
- As forecast accuracy worsens, safety stock becomes better

Application: Mass Customization

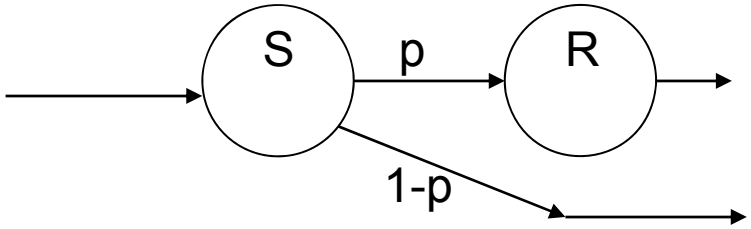
- Issues:
  - Organizing system
    - Tasks common to all products and tasks unique to an individual product
    - Should all tasks be done on a common system *or* only do the common tasks on a common system and do unique tasks separately
  - Make to stock and make to order
    - Which products should be made to stock and which made to order

Organizing System

- Option A: 2 parallel servers each taking S+pR



- Option B: 1 server takes S, other server takes R and gets fraction p of jobs



E[S]=pE[R]

Model Analysis

- A:

$$E[T] = E[S] + pE[R]$$

$$\text{var}\, T = \text{var}\, S + p\,\text{var}\, R + p(1-p)(E[R])^2$$

$$C_T^2 = \frac{C_S^2 + (C_R^2 + 1 - p)/p}{4}, if E[S] = pE[R]$$

$$E[N] = \frac{2(C_a^2/2 + C_T^2)}{2(1-\rho)}$$

- B:

$$E[N] = \frac{C_a^2 + C_S^2}{2(1-\rho)} + \frac{1 - p + pC_S^2 + C_R^2}{2(1-\rho)}$$

- A better than B: If all times deterministic when p>1/2
  If all times exponential when p>1/3
- Unusual tasks best segregated

Make to Stock and Make to Order

- Make to stock: Target z, single machine, exponential processing times

$$E[B] = \frac{\rho^{z+1}}{1-\rho}, \qquad E[I] = z - \frac{\rho(1-\rho^{z})}{1-\rho}$$

- Multiple products: product j performance as if service rate $\hat{\mu}_j = \mu - \sum_{i=1, i\neq j}^{m} \lambda_i, \qquad \hat{\rho}_j = \frac{\lambda_j}{\hat{\mu}_j}, \qquad E[B_j] = \frac{\hat{\rho}_j^{z_j+1}}{1-\hat{\rho}_j}$

- Suppose $\lambda_i = \gamma \lambda_{i-1}$
  then allocating a given total target optimally among 5 products to minimize total E[B] find that e.g if z=7 and ρ=0.8,
  individual targets are 4, 2,1,0,0

Insights from Models

- Mass customization means that variability increases.
- Cannot be avoided but its effects can be reduced by system configuration design and by optimizing the system operation
- Some customers and products may find service worse
- Low demand unique products can be quite disruptive
- To debate with marketing need to understand impact of variability

Optimal Control of Queueing Systems

- Sometimes it is valuable to look at optimal control models of queueing systems
- Usually use a Markov decision process model to determine the optimal action in a given state

Example: Improving Flow Line Throughput

- Set of tasks to be done, more tasks than workers

  Tasks: a->b->c->d
  Workers: 1 2

- (A) Allocate specific tasks to each worker, e.g., a b to 1, c d to 2.
- (B) Use optimal control to decide which task a worker does next (workers cannot overtake each other)
- Find that (B) gives better throughput than (A), particularly if workers have different speeds. Need substantially less inventory to reach a given throughput target.
- Not a simple rule to decide which task a worker does next so may be impractical
  If 1 much faster than 2 then 2 only does task d (means that slower worker often idle) but if 1 and 2 are not too different policies more complicated

- Implication: the advantages of teams (if they achieve dynamic allocation)
  the disadvantage (they use fixed task allocation)

Conclusions

- Queueing models are very useful for
  - Validating simulations
  - Verifying correctness of representation of system behavior and operation
  - Gaining insight into the effect of variability in
    - Service times
    - Arrivals
    - Routing
  - Seeing how changes in system structure and configuration can reduce impact of variability
  - Exploring new ideas for system design

How do Stochastic Models of Manufacturing Systems Help?

- 60-70 years of development
  - Diminished interest in last 10 years
  - Future?
- How do they help decision making?
  - Insight
- Are they used much?
  - Semiconductor industry
  - Auto industry
- Opportunities met and missed
  - FMS
  - Mass Customization
  - Services