Time Variability in Manufacturing Systems

2.810

Stanley B. Gershwin

November 1, 2004

Goals

- To explain important measures of system performance.
- To show the importance of random, potentially disruptive events in factories.
- To give some intuition about behavior of these systems.
- To describe some current tools and methods.

Quantity and Quality

- Quantity – how much and when.
- Quality – how well.

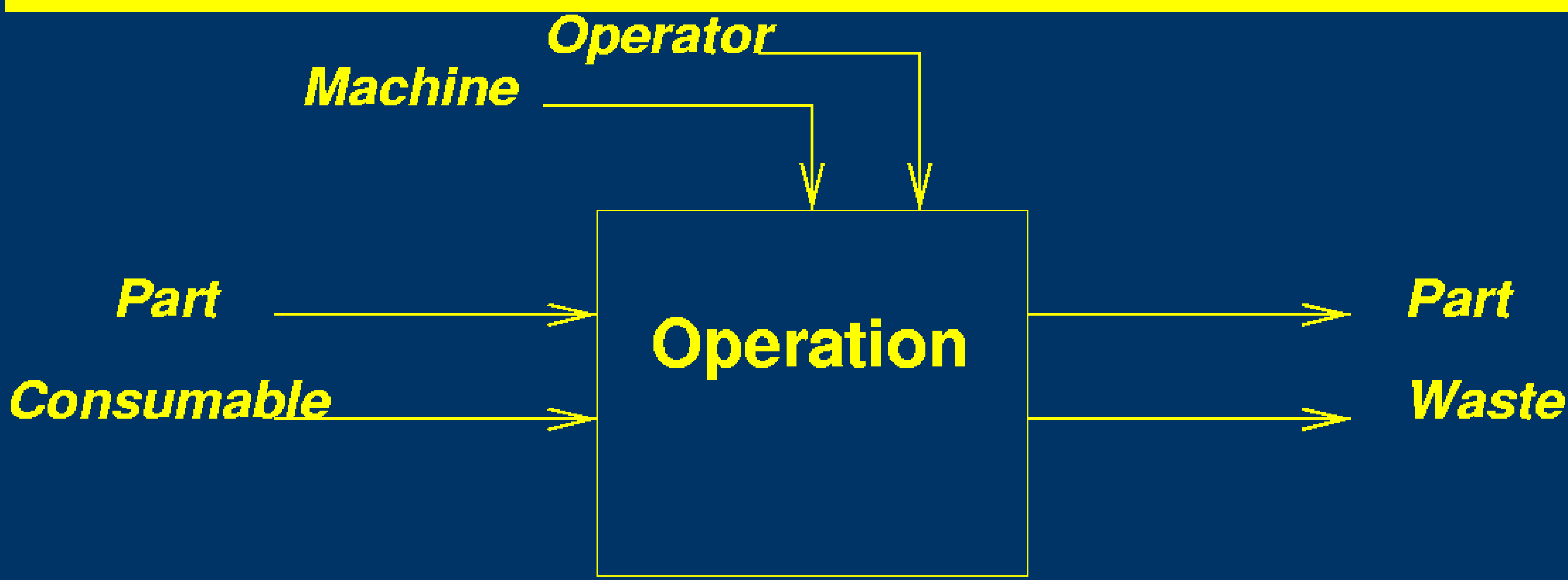
Here, we focus on *quantity*.

General Statement: Variability is the enemy of manufacturing.

Conflicting Manufacturing System Objectives

- Make to Stock:
  - items available when a customer arrives *(lots of inventory)*
  - high profits and low prices *(little inventory)*
- Make to Order:
  - early delivery promises *(unreliable promises or excess capacity)*
  - reliable delivery promises *(late promises or excess capacity)*
  - high profits and low prices *(no excess capacity)*

Operation



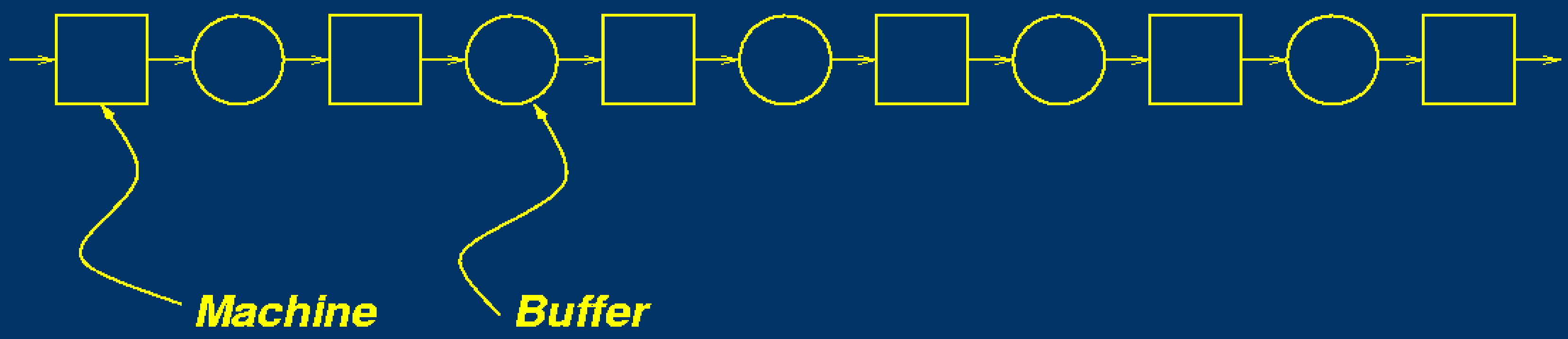
Nothing happens until everything is present.

Waiting

Whatever does not arrive last must wait.

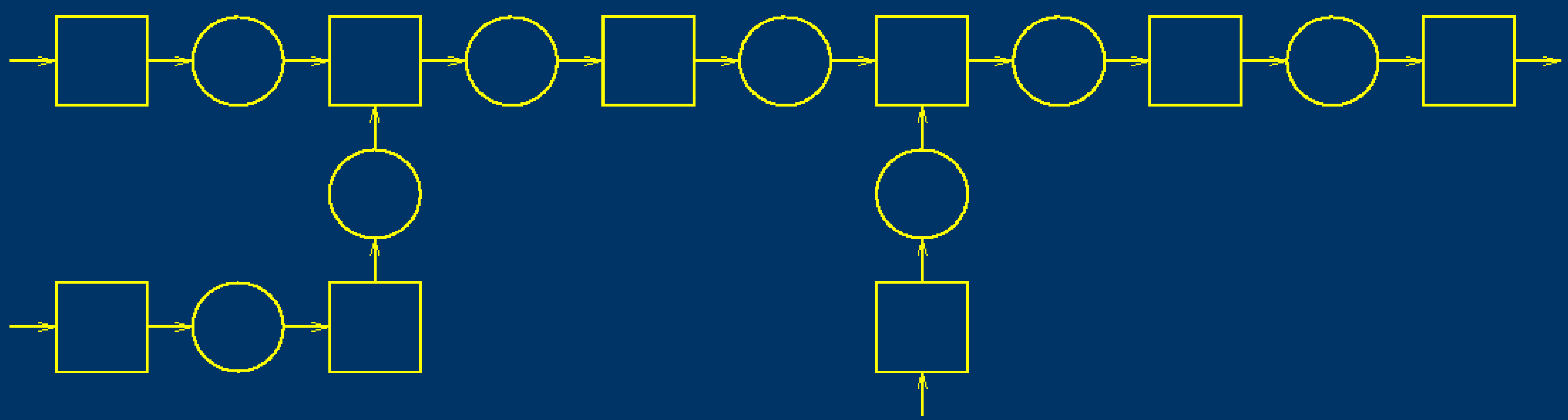
- *Inventory:* parts waiting.
- *Underutilization:* machines waiting.
- *Idle work force:* operators waiting.

Flow shop , *Flow line* , or *Production line:*



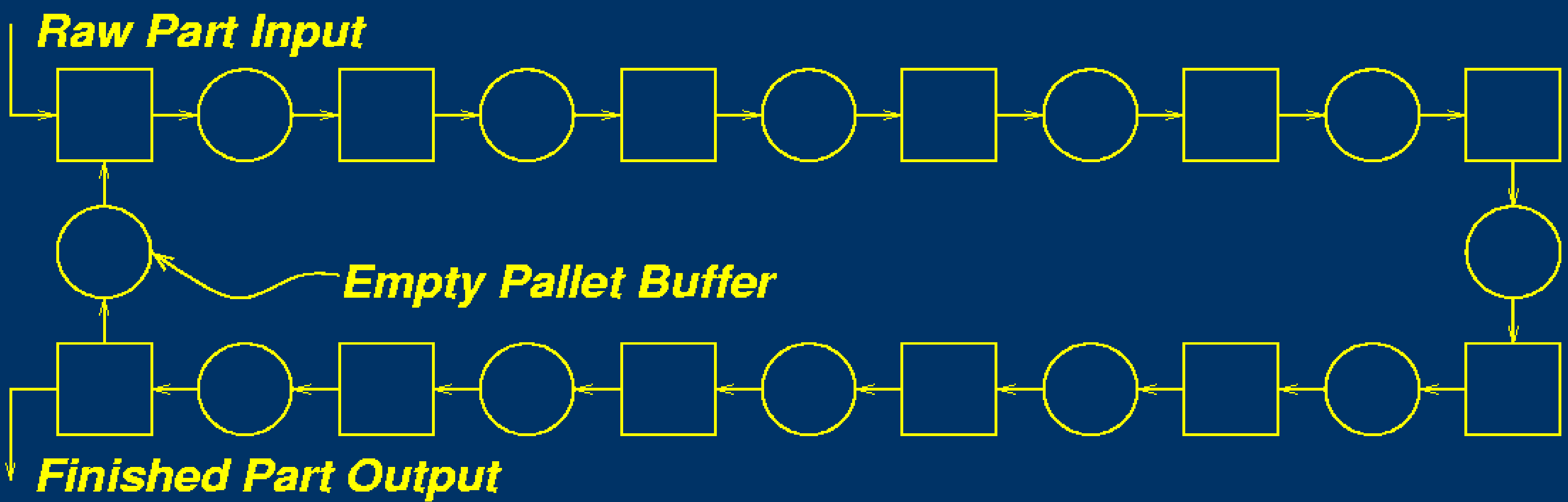
Traditionally used for high volume, low variety production.

Assembly system:



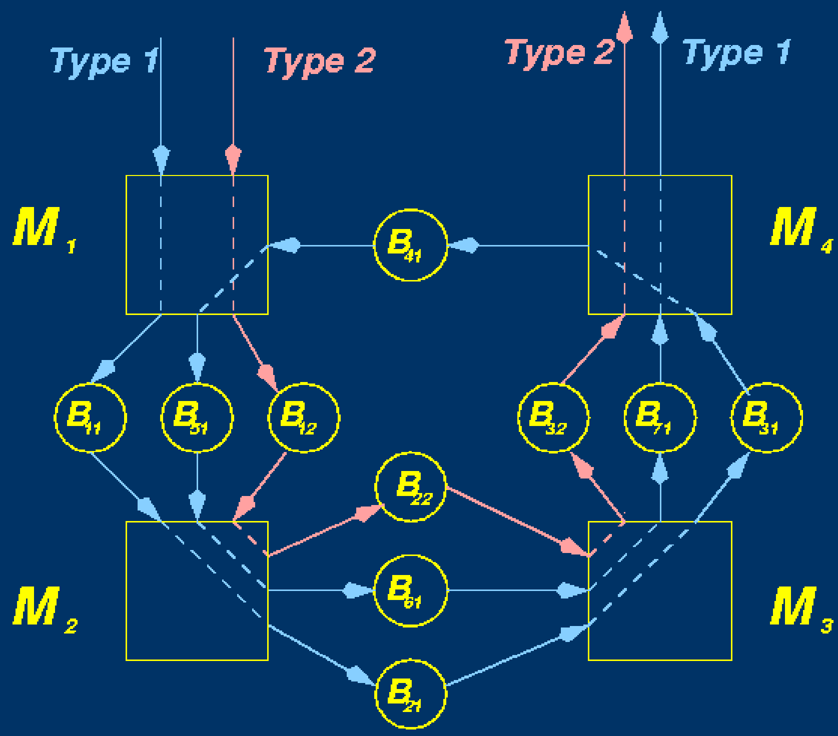
Assembly systems are *trees* , and may involve *thousands* of parts.

Closed pallet or fixture system:



Pallets or fixtures travel in a closed loop.

System with reentrant flow and two part types



Job Shop

- Machines not organized according to process flow.
- Often, machines grouped by department:
  - ⋆ mill department
  - ⋆ lathe department
  - ⋆ etc.
- Great variety of products.
- Different products follow different paths.
- Complex management.

Two Issues

- Efficient design of systems;
  - ⋆ Today's focus.
  - ⋆ *Actual* focus is on performance analysis – analysis precedes synthesis.
- Efficient operation of systems after they are built.

Time

- All factory performance measures are about time.
  - ⋆ *production rate:* how much is made in a given time.
  - ⋆ *lead time:* how much time before delivery.
  - ⋆ *cycle time:* how much time a part spends in the factory.
  - ⋆ *delivery reliability:* how often a factory delivers on time.
  - ⋆ *capital pay-back period:* the time before the company get its investment back.

Time, continued

- Time appears in two forms:
  - delay
  - capacity utilization
- Every action has impact on both.

Delay

- An operation that takes 10 minutes adds 10 minutes to the delay that
  - a workpiece experiences while undergoing that operation;
  - every other workpiece experiences that is waiting while the first is being processed.

Capacity Utilization

- An operation that takes 10 minutes takes up 10 minutes of the time of
  - ⋆ a machine
  - ⋆ an operator
  - ⋆ other resources?
- Since there are a limited number of minutes of each resource available, there are a limited number of operations that can be done.

Production Rate

- *Operation Time:* the time that a machine takes to do an operation.
- *Production Rate:* the average number of parts produced in a time unit. (Also called *throughput.*)

If nothing interesting ever happens (no failures, etc.),

$$\text{Production rate} = \frac{1}{\text{operation time}}$$

... but something interesting *always* happens.

Capacity

- *Capacity:* the maximum possible production rate of a manufacturing system, for systems that are making only one part type.
  - ⋆ *Short term capacity:* determined by the resources available right now.
  - ⋆ *Long term capacity:* determined by the average resource availability.
- Capacity is harder to define for systems making more than one part type. Since it is hard to define, it is *very* hard to calculate.

Queueing theory

- Simplest model is the $M/M/1$ queue:
  - ⋆ Exponentially distributed inter-arrival times — mean is $1/\lambda$; λ is *arrival rate* (customers/time).
  - ⋆ Exponentially distributed service times — mean is $1/\mu$; μ is *service rate* (customers/time).
  - ⋆ Infinite waiting area.
- Define $\rho = \lambda/\mu$.

Performance of $M/M/1$ queue

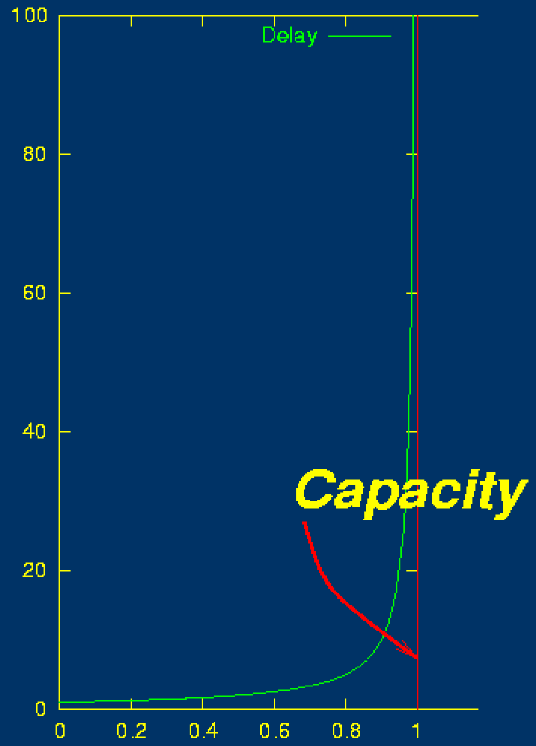
- Average number of customers in the system is

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \textit{if } \rho < 1$$

- *Little's law:* $L = \lambda W$, where W is the average customer's waiting time.
- Therefore

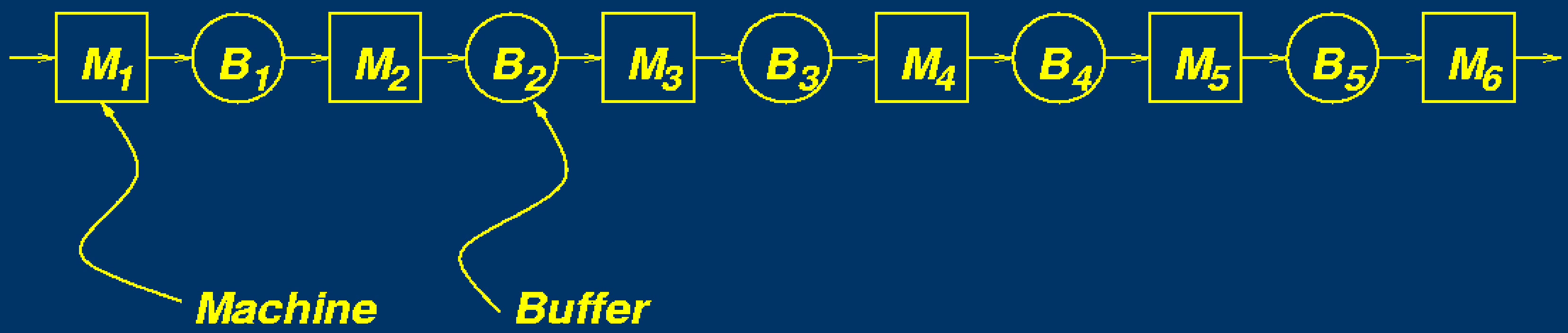
$$W = \frac{1}{\mu - \lambda} \textit{if } \rho < 1$$

Performance of $M/M/1$ queue



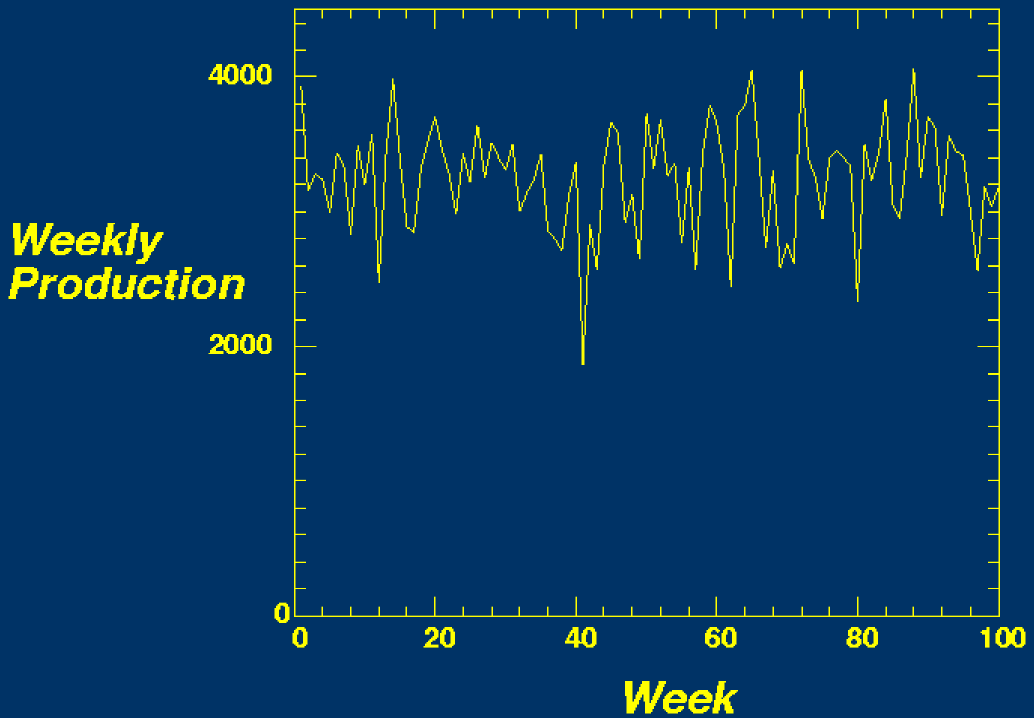
- μ is the *capacity* of the system.
- If $\lambda > \mu$, waiting time grows over time.

Production lines — single-part-type, multiple-stage systems



- Also called *transfer lines* , *flow lines* , etc.

Output Variability



Production output from a simulation of a transfer line.

Single Reliable Machine

- If the machine is perfectly reliable, and its average operation time is τ, then its maximum production rate is $1/\tau$.
- *Note:*
  - ⋆ Sometimes *cycle time* is used instead of *operation time*, but *BEWARE:* cycle time has two meanings!
  - ⋆ The other meaning is the time a part spends in a system. If the system is a single, reliable machine, the two meanings are the same.

Operation-Dependent Failures (ODFs)

- A machine can only fail while it is working.
- *IMPORTANT!* MTTF *must* be measured in working time!
- This is the usual assumption.

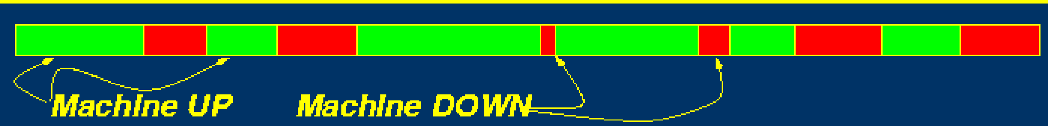
Single Unreliable Machine

- If the machine is unreliable, and
  - ⋆ its average operation time is τ,
  - ⋆ its mean time to fail is MTTF,
  - ⋆ its mean time to repair is MTTR,

then its maximum production rate is

$$\frac{1}{\tau}\left(\frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}\right)$$

Single Unreliable Machine — Proof



- Average production rate, while machine is up, is $1/\tau$.
- Average duration of an up period is MTTF.
- Average production during an up period is MTTF/τ.
- Average duration of up-down period: MTTF + MTTR.
- Average production during up-down period: MTTF/τ.
- Therefore, average production rate is (MTTF/τ)/(MTTF + MTTR).

Geometric Up- and Down-Times

- *Assumptions:* Operation time is constant (τ). Failure and repair times are *geometrically* distributed.
- Let p be the probability that a machine fails in during any given operation. Then $p = \tau/\text{MTTF}$.

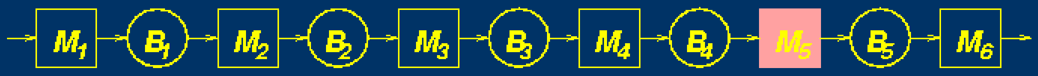
Geometric Up- and Down-Times, continued

- Let r be the probability that M gets repaired in during any operation time when it is down. Then $r = \tau/\text{MTTR}$.
- Then the *average production rate* of M is
$$\frac{1}{\tau}\left(\frac{r}{r+p}\right).$$
- *(Sometimes we leave out "average.")*

Production Rates

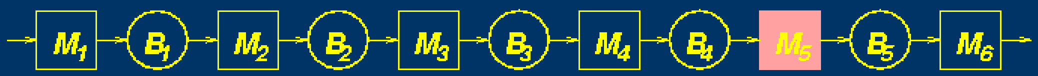
- So far, the machine really has *three* production rates:
  - ⋆ $1/\tau$ when it is up *(short-term capacity)* ,
  - ⋆ 0 when it is down *(short-term capacity)* ,
  - ⋆ $(1/\tau)(r/(r+p))$ on the average *(long-term capacity)* .

Infinite-Buffer Line



- The production rate of the line is the production rate of the *slowest* machine in the line — called the *bottleneck* .
- *Slowest* means least average production rate, where average production rate is calculated from one of the previous formulas.

Infinite-Buffer Line

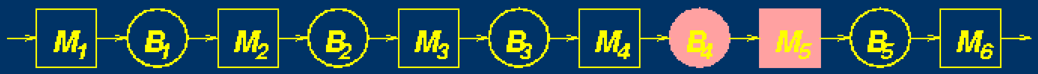


- Production rate is therefore

$$P = \min_i \frac{1}{\tau_i} \left(\frac{\text{MTTF}_i}{\text{MTTF}_i + \text{MTTR}_i} \right)$$

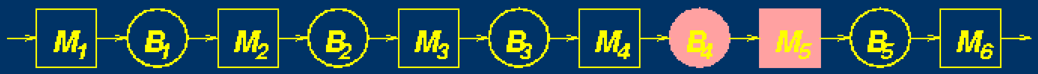
- and M_i is the bottleneck.

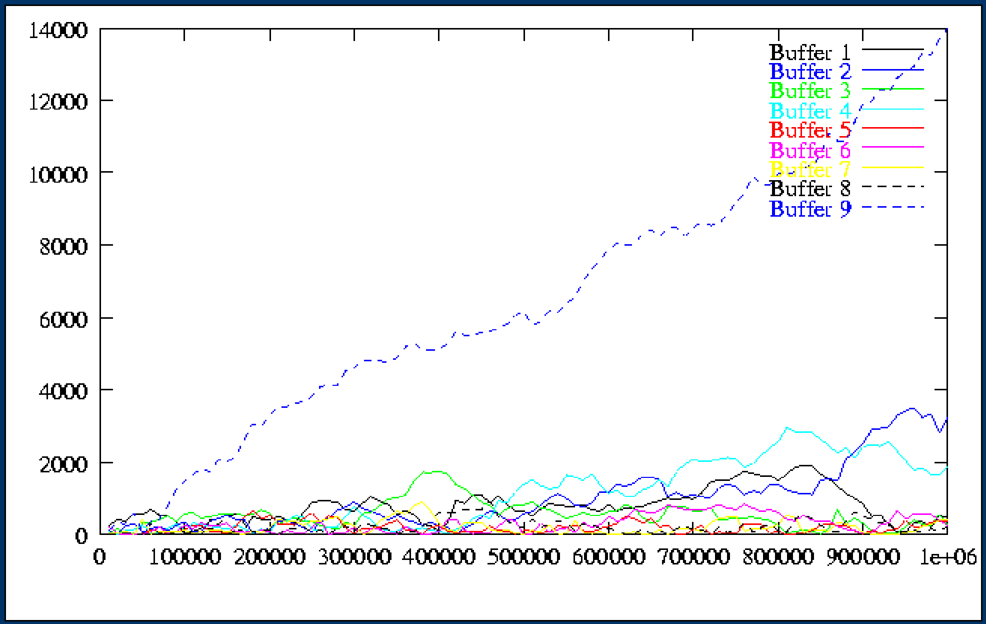
Infinite-Buffer Line



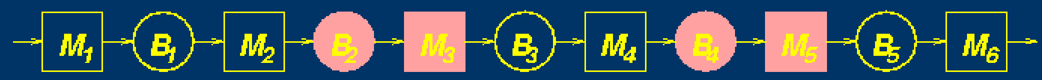
- The system is not in steady state.
- An infinite amount of inventory accumulates in the buffer upstream of the bottleneck.
- A finite amount of inventory appears downstream of the bottleneck.

Infinite-Buffer Line



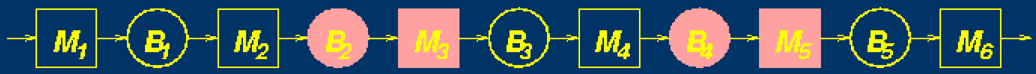


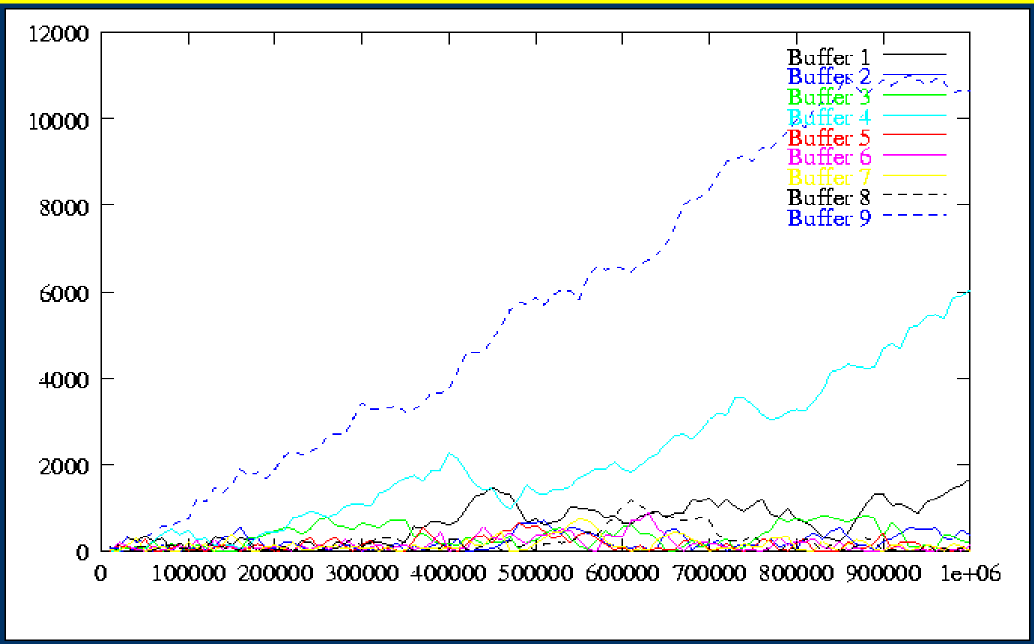
Infinite-Buffer Line



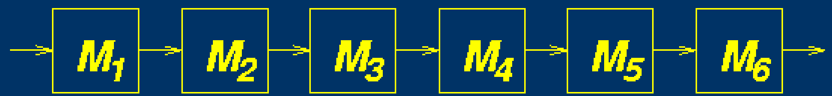
- The *second bottleneck* is the slowest machine upstream of the bottleneck. An infinite amount of inventory accumulates just upstream of it.
- A finite amount of inventory appears between the second bottleneck and the machine upstream of the first bottleneck.
- Et cetera.

Infinite-Buffer Line



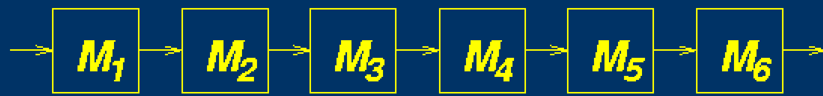


Zero-Buffer Line



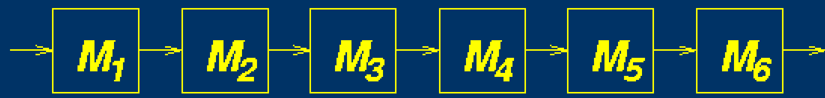
- If any one machine fails, or takes a very long time to do an operation, *all* the other machines must wait.
- Therefore the production rate is usually less — possibly much less – than the slowest machine.

Zero-Buffer Line



- *Example:* Constant, unequal operation times, perfectly reliable machines.
  - ⋆ The operation time of the line is equal to the operation time of the slowest machine, so the production rate of the line is *equal to* that of the slowest machine.

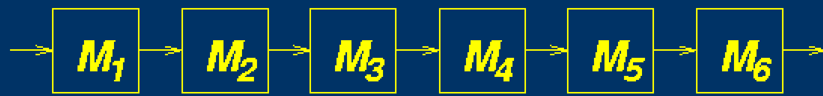
Zero-Buffer Line



Constant, equal operation times, unreliable machines.

- *Assumption:* Failure and repair times are *geometrically* distributed.
- Define $p_i = \tau/\text{MTTF}_i$ = probability of failure during an operation.
- Define $r_i = \tau/\text{MTTR}_i$ probability of repair during an interval of length τ when the machine is down.

Zero-Buffer Line

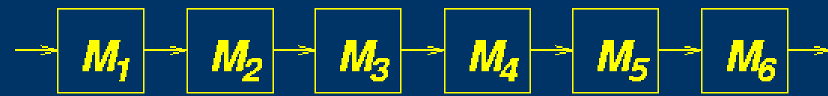


Constant, equal operation times, unreliable machines.

- Then the *isolated production rate* of a single machine M_i is

$$\frac{1}{\tau}\left(\frac{r_i}{r_i + p_i}\right) = \frac{1}{\tau}\left(\frac{1}{1 + \frac{p_i}{r_i}}\right).$$

Zero-Buffer Line



Constant, equal operation times, unreliable machines.

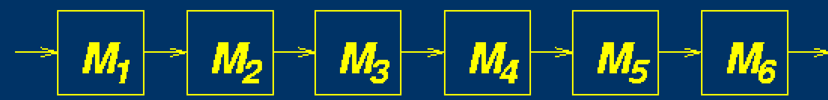
Buzacott's Zero-Buffer Line Formula:

Let k be the number of machines in the line. Then

$$P = \frac{1}{\tau} \frac{1}{1 + \sum_{i=1}^{k} \frac{p_i}{r_i}}$$

- Same as the earlier formula when $k = 1$.

Zero-Buffer Line

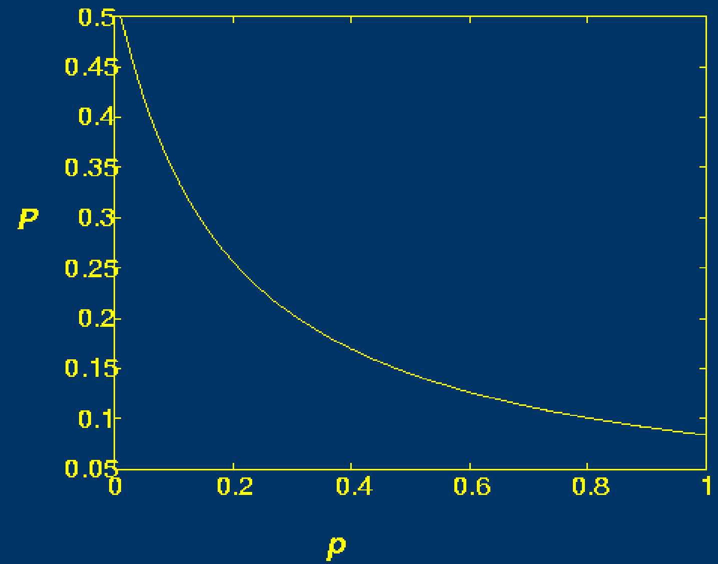


Constant, equal operation times, unreliable machines.

- Note that $\boldsymbol{P}$ is a function of the *ratio* $\mathbf{p}_i/\mathbf{r}_i$ and not $\mathbf{p}_i$ or $\mathbf{r}_i$ separately.
- The same statement is true for the infinite-buffer line.
- However, the same statement is *not* true for a line with finite, non-zero buffers.

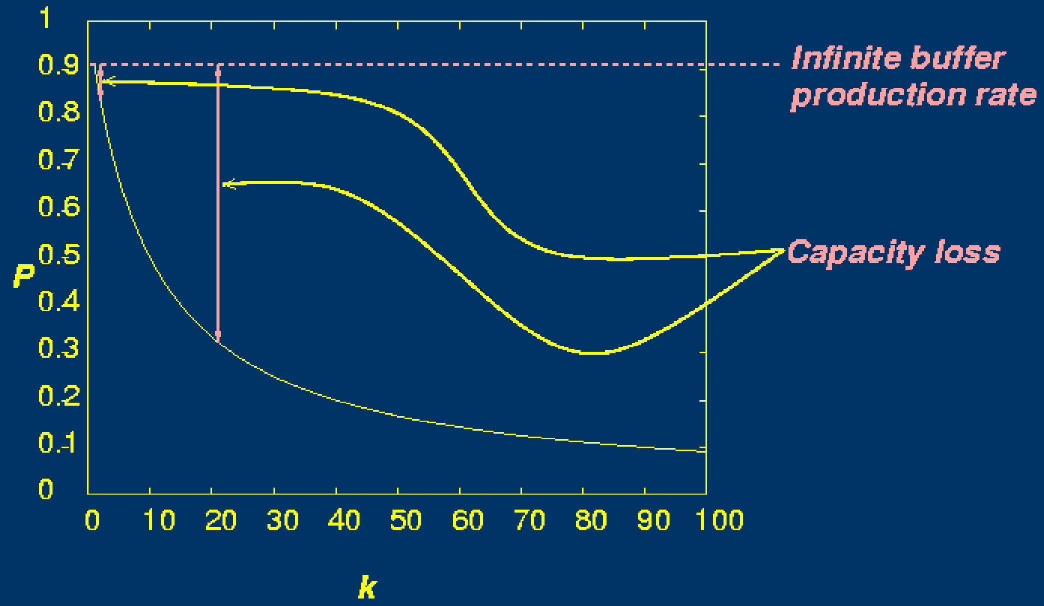
P as a function of p_i

All machines are the same except M_i. As p_i increases, the production rate decreases.

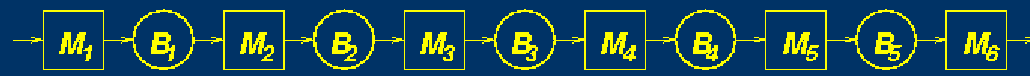


P as a function of k

All machines are the same. As the line gets longer, the production rate decreases.

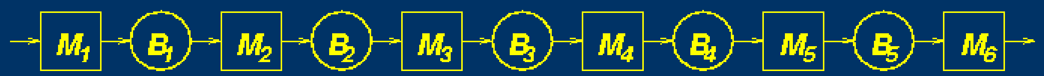


Finite-Buffer Lines



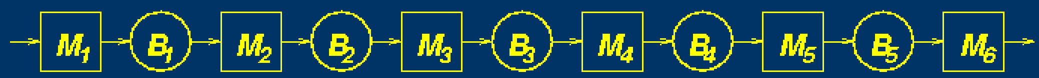
- Motivation for buffers: recapture some of the lost production rate.
- Cost
  - ★ in-process inventory/lead time
  - ★ floor space
  - ★ material handling mechanism

Finite-Buffer Lines



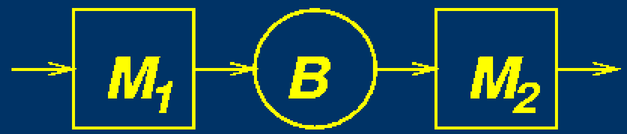
- Infinite buffers: no propagation of disruptions.
- Zero buffers: instantaneous propagation.
- Finite buffers: delayed propagation.
  - ⋆ New phenomena: blockage and starvation.

Finite-Buffer Lines



- Difficulty:
  - ★ No simple formula for calculating production rate or inventory levels.
- Solution:
  - ★ Simulation
  - ★ Analytical approximation

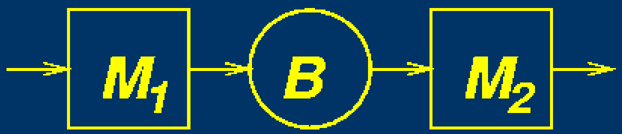
Two Machine, Finite-Buffer Lines



Several models available:

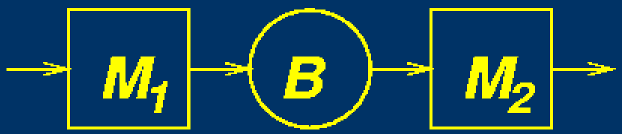
- *Deterministic processing time* , or *Buzacott model:* deterministic processing time, geometric failure and repair times; discrete state, discrete time.

Two Machine, Finite-Buffer Lines

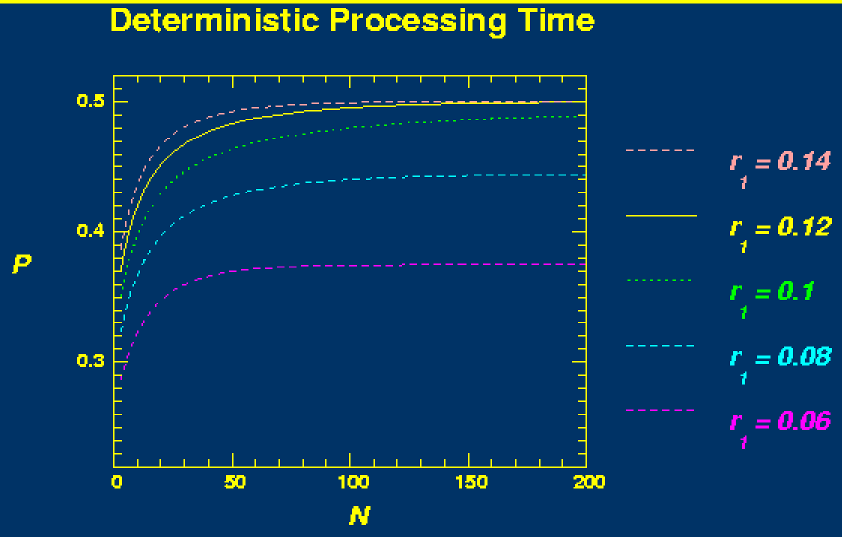


- *Exponential processing time:* exponential processing, failure, and repair time; discrete state, continuous time.
- *Continuous material,* or *fluid:* deterministic processing, exponential failure and repair time; mixed state, continuous time.

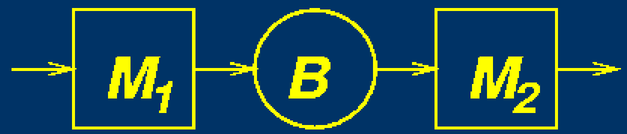
Two Machine, Finite-Buffer Lines



$$\tau = 1.$$
$$p_1 = .1$$
$$r_2 = .1$$
$$p_2 = .1$$

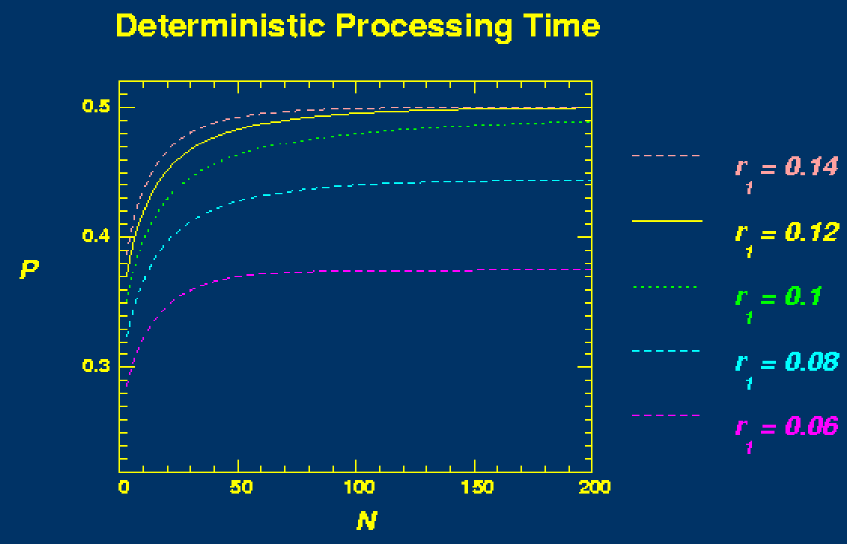


Two Machine, Finite-Buffer Lines

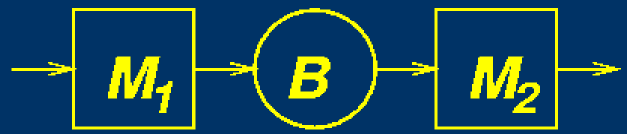


Discussion:

- Why are the curves increasing?
- Why do they reach an asymptote?
- What is P when $N = 0$?
- What is the limit of P as $N \to \infty$?
- Why are the curves with smaller r_1 lower?



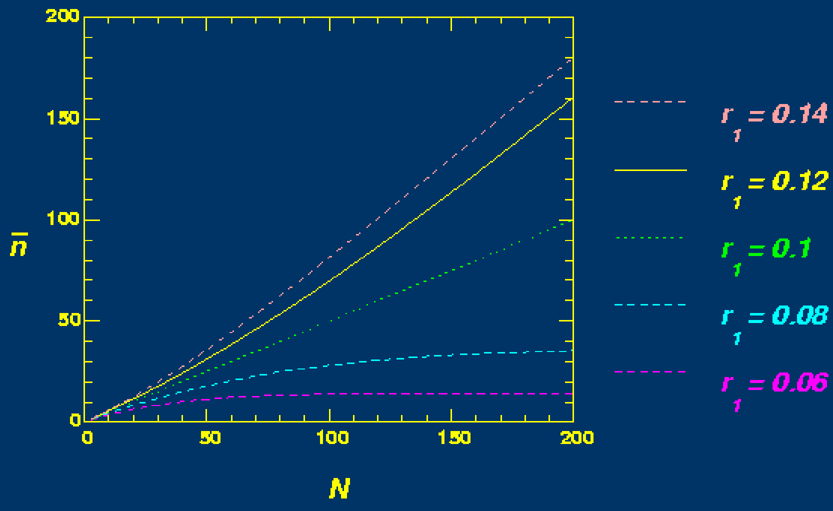
Two Machine, Finite-Buffer Lines



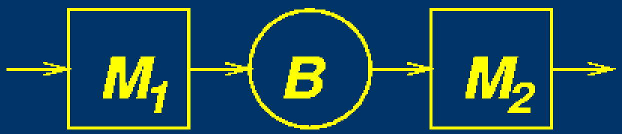
Discussion:

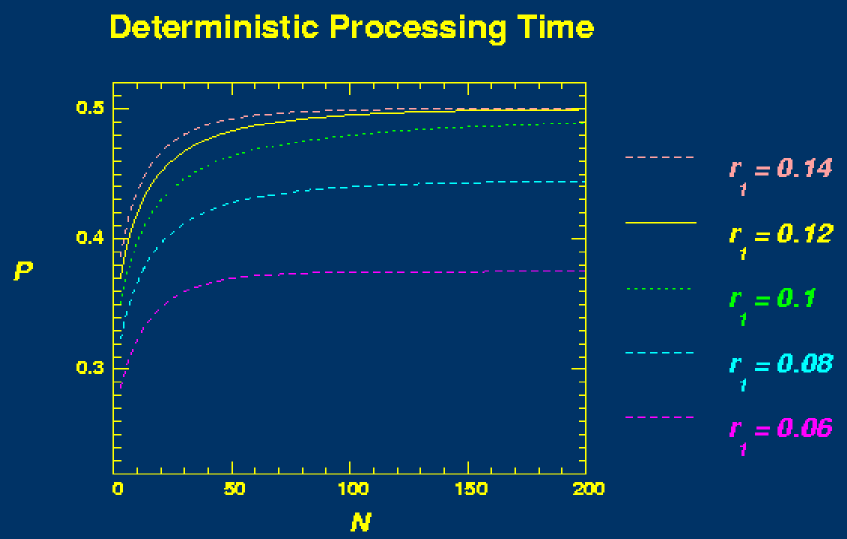
- Why are the curves increasing?
- Why *different* asymptotes?
- What is $\bar{n}$ when $N = 0$?
- What is the limit of $\bar{n}$ as $N \to \infty$?
- Why are the curves with smaller r_1 lower?

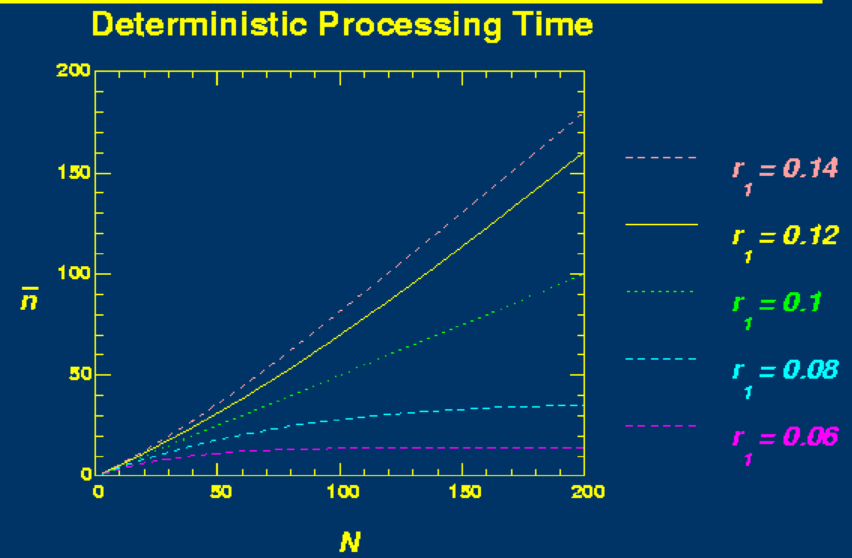
Deterministic Processing Time



Two Machine, Finite-Buffer Lines

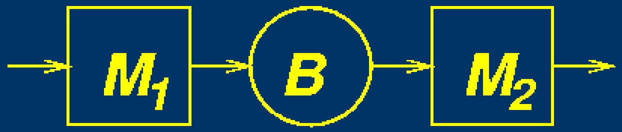




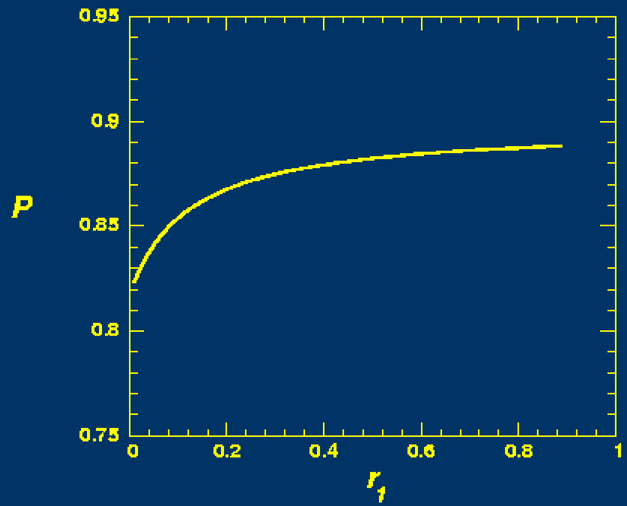


What can you say about the optimal buffer size?

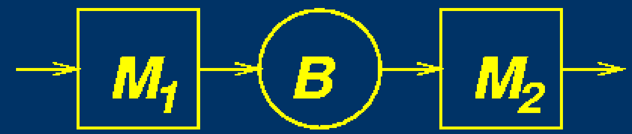
Should we prefer short, frequent, disruptions or long, infrequent, disruptions?



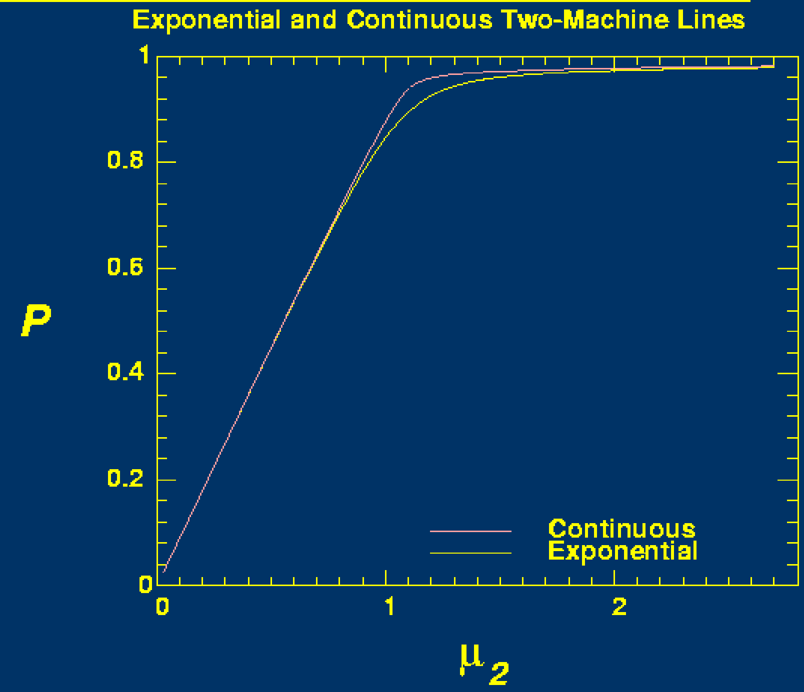
- $r_2 = 0.8$, $p_2 = 0.09$, $N = 10$
- r_1 and p_1 vary together and $\frac{r_1}{r_1+p_1} = .9$
- *Answer:* evidently, short, frequent failures.
- *Why?*



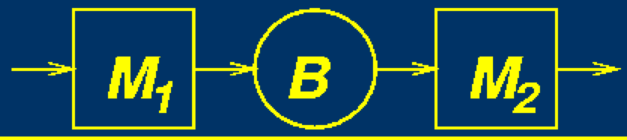
Two Machine, Finite-Buffer Lines

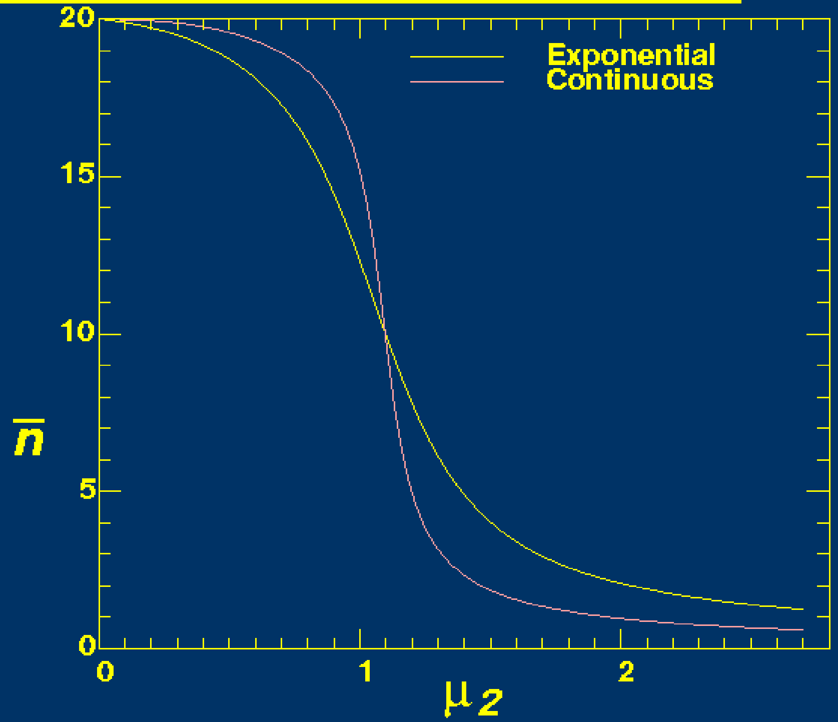


- $r_1 = 0.09$, $p_1 = 0.01$, $\mu_1 = 1.1$
- $r_2 = 0.08$, $p_1 = 0.009$
- $N = 20$
- *Explain the shapes of the graphs.*



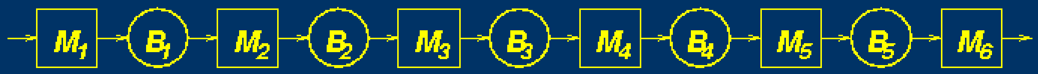
Two Machine, Finite-Buffer Lines





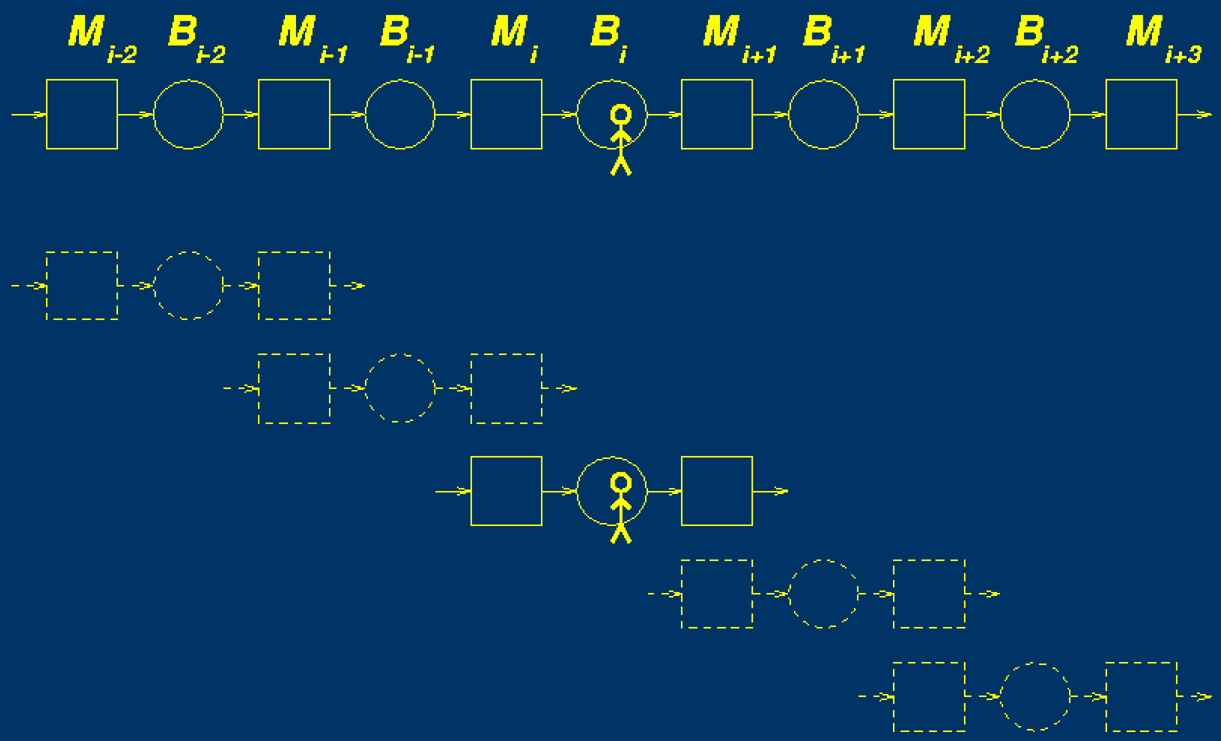
- *Explain the shapes of the graphs.*

Long Lines

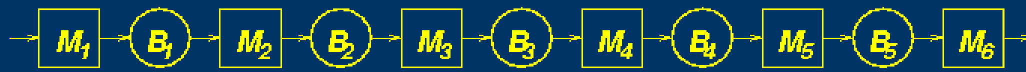


- Difficulty:
  - ★ No simple formula for calculating production rate or inventory levels.
  - ★ State space is too large for exact numerical solution.
  - ★ *Decomposition* seems to work successfully.

Decomposition

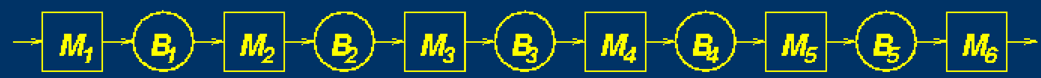


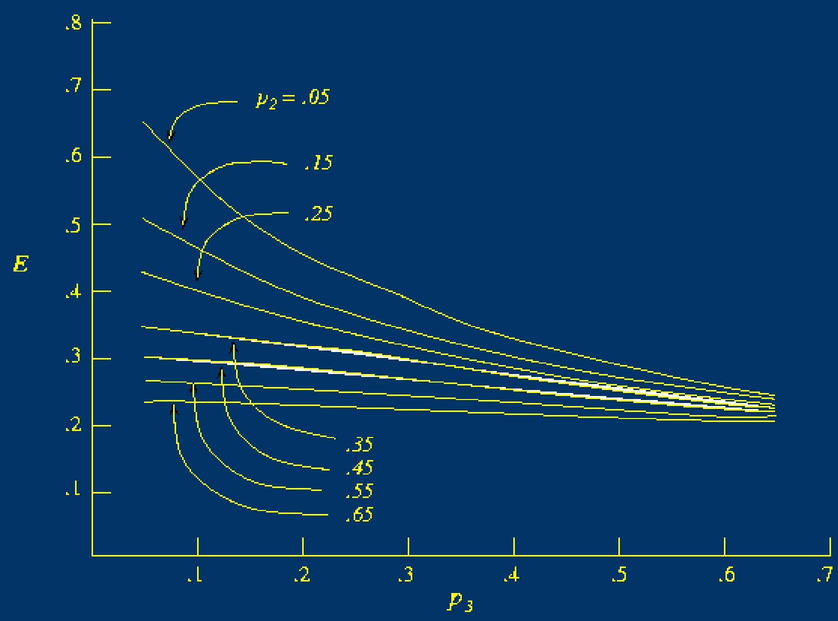
Decomposition



- Consider an observer in Buffer B_i.
  - ⋆ Imagine the material flow that the observer sees *entering* and *leaving* the buffer.
- We construct a two-machine line (ie, we find r_1, p_1, r_2, p_2, and N) such that an oberver in its buffer will see almost the same thing.
- The parameters are chosen as functions of the behaviors of the other two-machine lines.

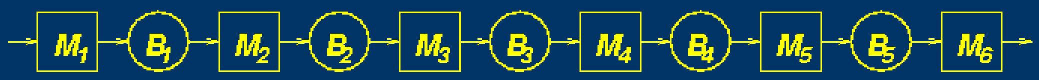
Example

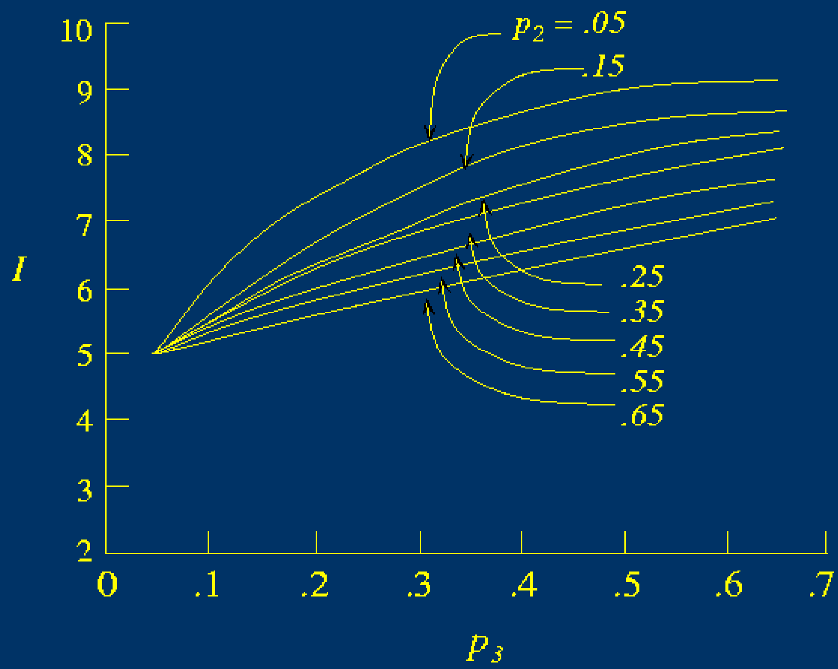




$$r_1 = r_2 = r_3 = .2$$
$$p_1 = .05$$
$$N_1 = N_2 = 5$$

Example



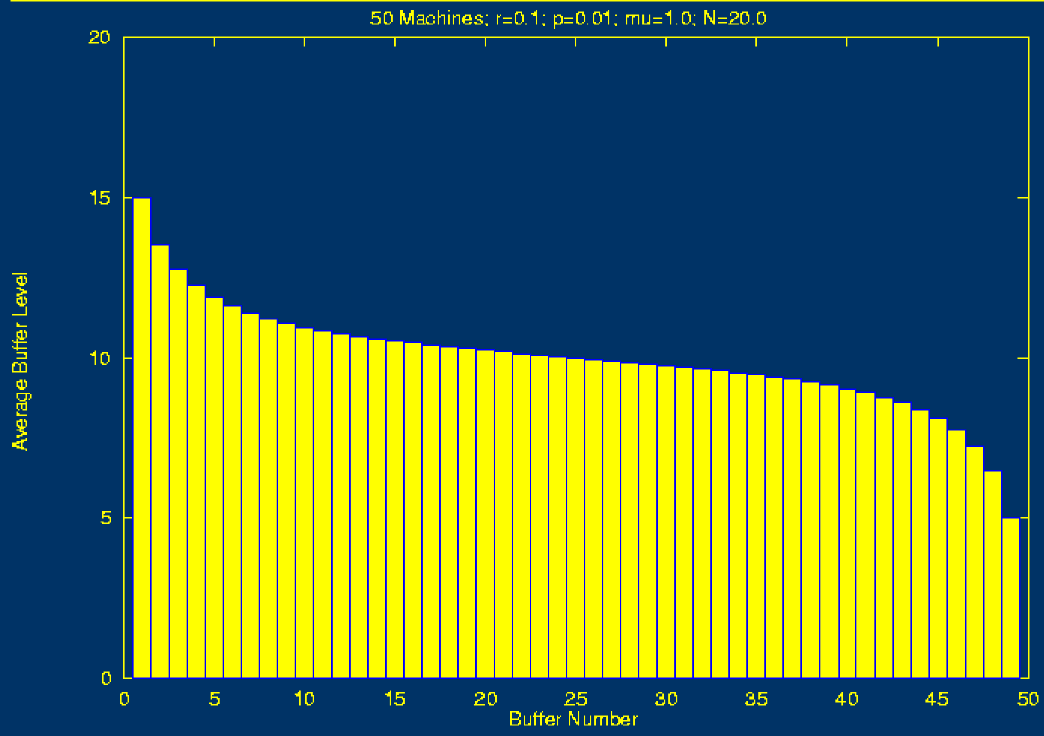


$$r_1 = r_2 = r_3 = .2$$

$$p_1 = .05$$

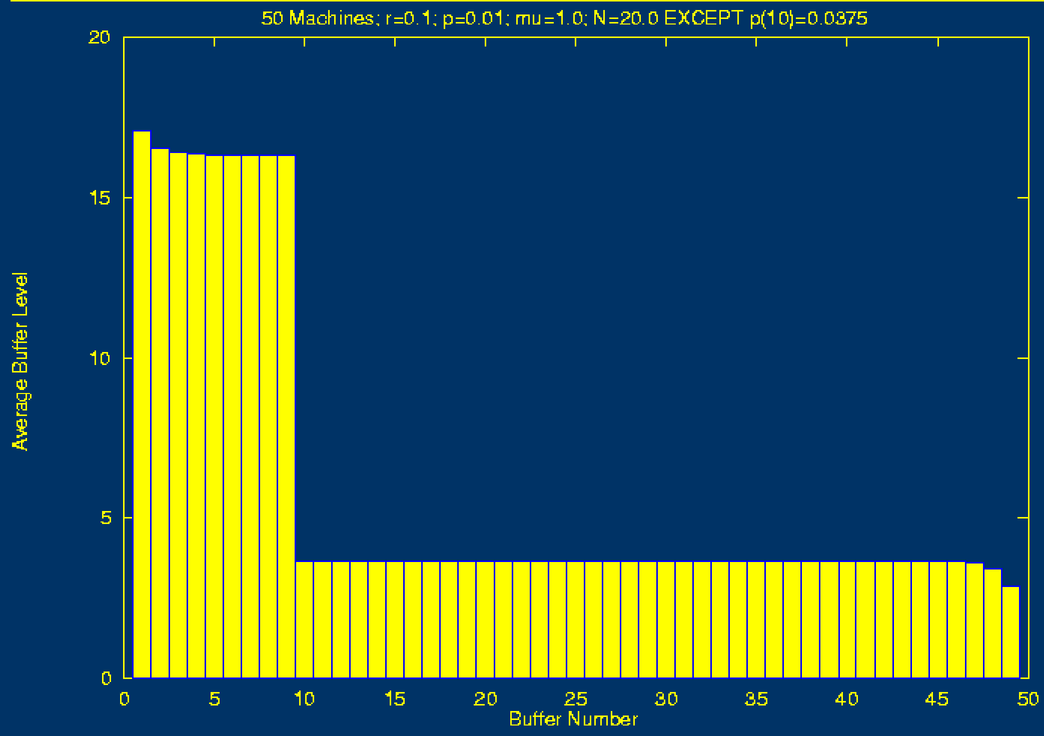
$$N_1 = N_2 = 5$$

Example



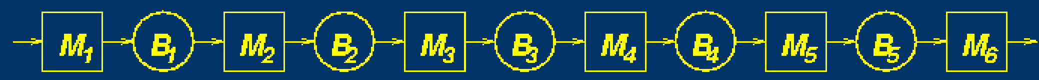
Distribution of material in a line with identical machines and buffers. *Explain the shape.*

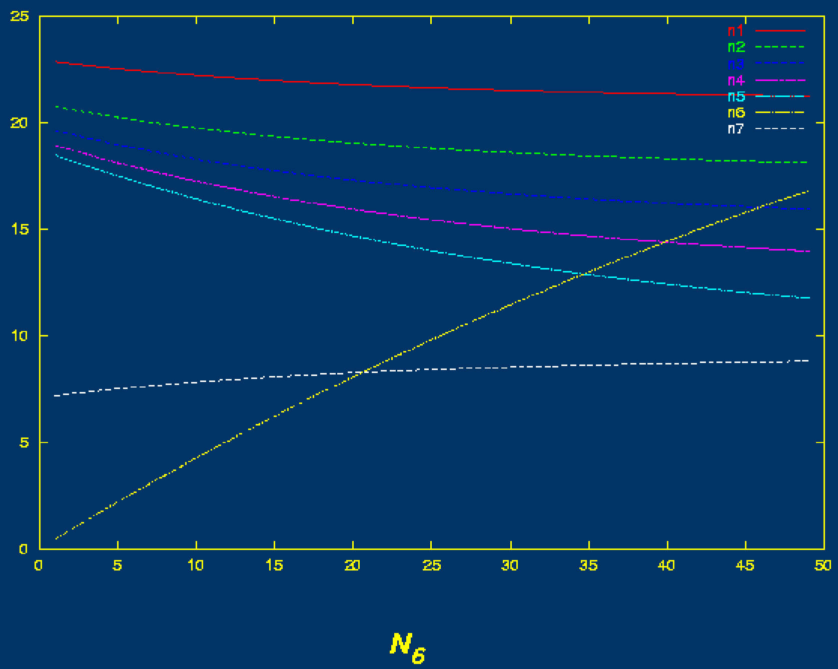
Example



Effect of a bottle-neck. Identical machines and buffers, except for M_{10}.

Example

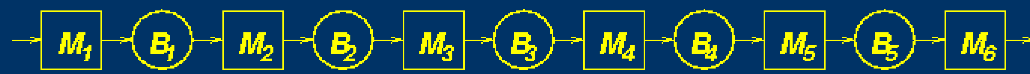


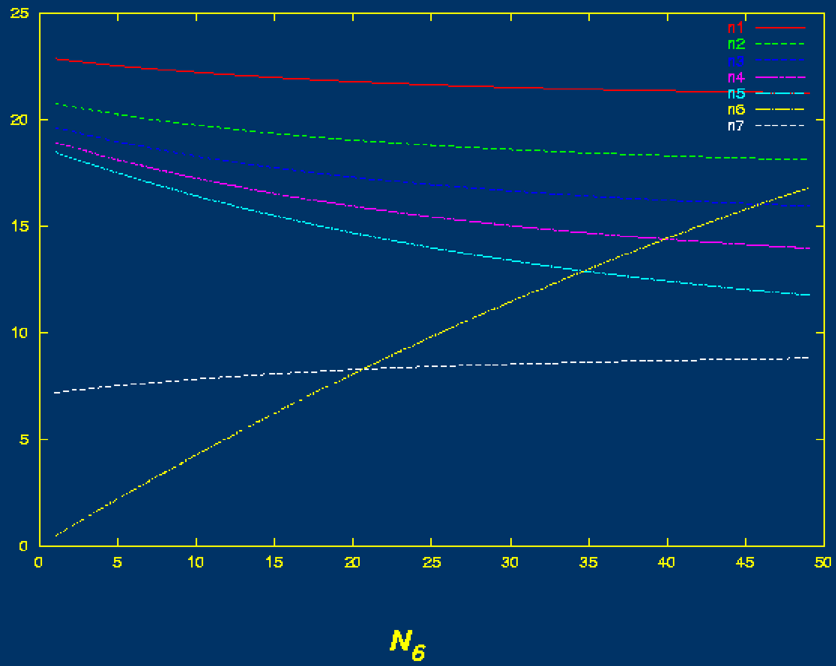


Continuous material model.

- Eight-machine, seven-buffer line.
- For each machine, $r = .075$, $p = .009$, $\mu = 1.2$.
- For each buffer *(except Buffer 6)*, $N = 30$.

Example

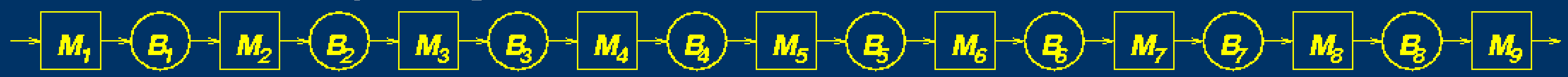




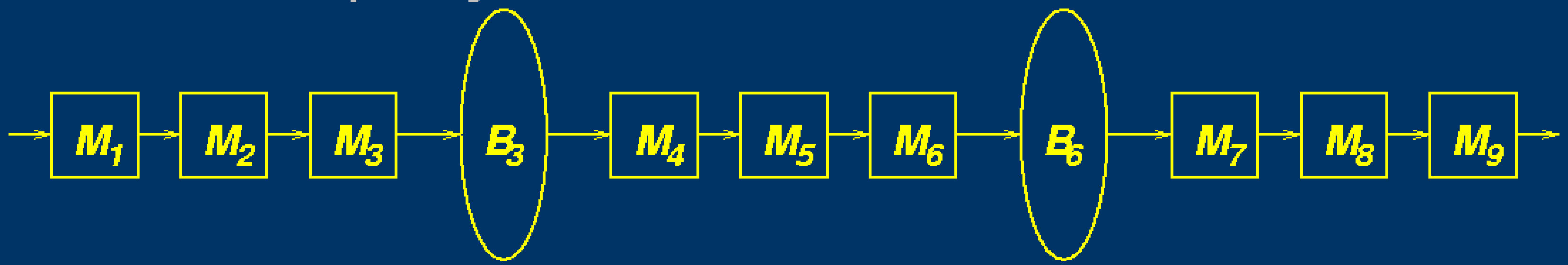
- Which $\bar{n}_i$ are decreasing and which are increasing?
- Why?

Example: *Which has a higher production rate?*

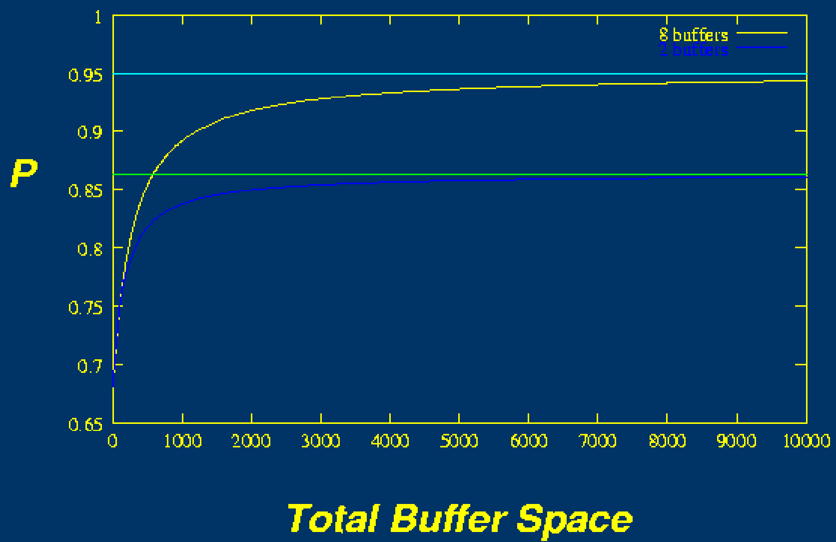
- 9-Machine line with two buffering options:
  - 8 buffers equally sized; and



  - 2 buffers equally sized.

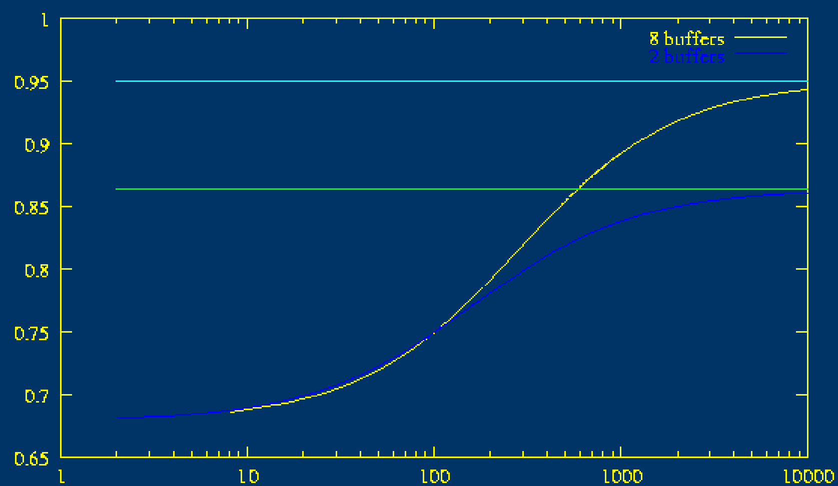


Example



- Continuous model; all machines have $r = .019$, $p = .001$, $\mu = 1$.
- What are the asymptotes?
- Is 8 buffers *always* faster?

Example



- *Is 8 buffers always faster?*
- Evidently not, but difference is not significant in systems with very small buffers.

Problem: Optimal buffer space distribution.

- Design the buffers for a 20-machine production line.
- The machines have been selected, and the only decision remaining is the amount of space to allocate for in-process inventory.
- *The goal is to determine the smallest amount of in-process inventory space so that the line meets a production rate target.*

Problem: Optimal buffer space distribution.

- The common operation time is one operation per minute.
- The target production rate is .88 parts per minute.

Problem: Optimal buffer space distribution.

- *Case 1* MTTF= 200 minutes and MTTR = 10.5 minutes for all machines ($P = .95$ parts per minute).

Problem: Optimal buffer space distribution.

- *Case 1* MTTF= 200 minutes and MTTR = 10.5 minutes for all machines ($\boldsymbol{P = .95}$ parts per minute).
- *Case 2* Like Case 1 except Machine 5. For Machine 5, MTTF = 100 and MTTR = 10.5 minutes ($\boldsymbol{P = .905}$ parts per minute).

Problem: Optimal buffer space distribution.

- *Case 1* MTTF= 200 minutes and MTTR = 10.5 minutes for all machines ($\mathbf{P = .95}$ parts per minute).
- *Case 2* Like Case 1 except Machine 5. For Machine 5, MTTF = 100 and MTTR = 10.5 minutes ($\mathbf{P = .905}$ parts per minute).
- *Case 3* Like Case 1 except Machine 5. For Machine 5, MTTF = 200 and MTTR = 21 minutes ($\mathbf{P = .905}$ parts per minute).

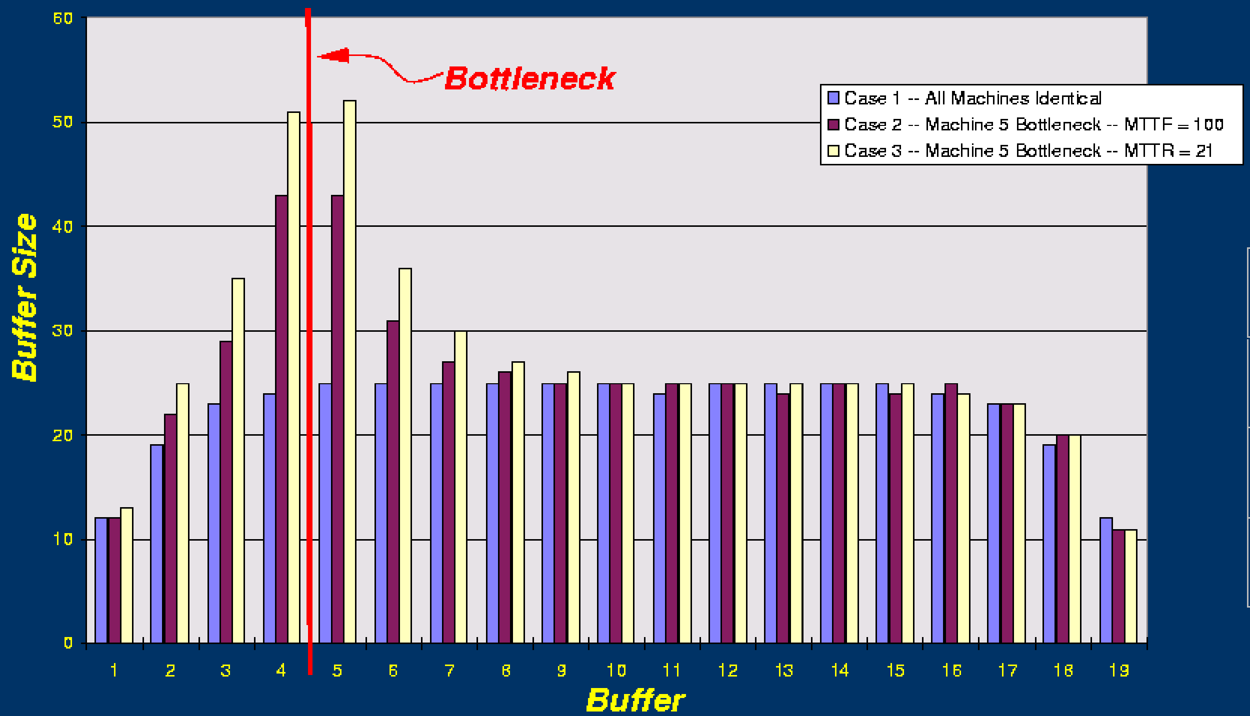
Problem: Optimal buffer space distribution.

Are buffers really needed?

Line	Production rate with no buffers, parts per minute
Case 1	.487
Case 2	.475
Case 3	.475

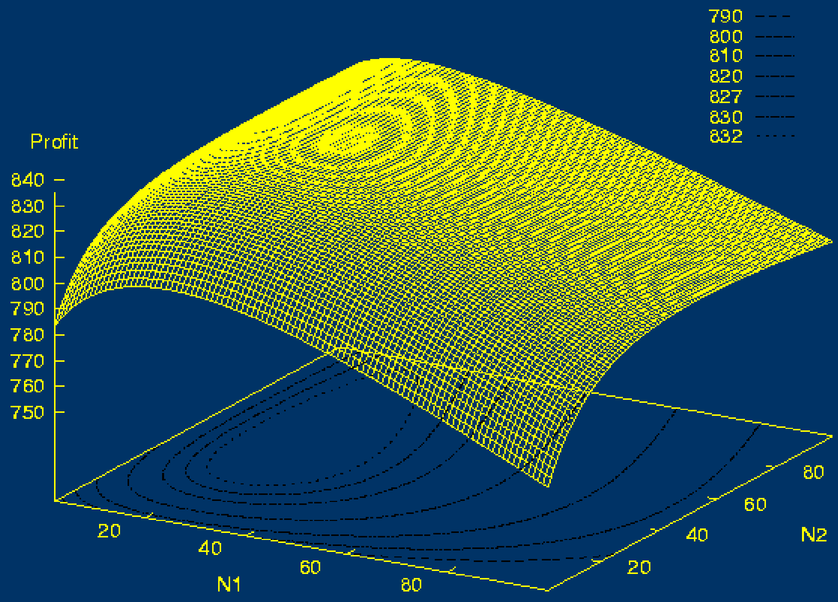
Yes. *How were these numbers calculated?*

Solution: Optimal buffer space distribution.



Line	Space
Case 1	430
Case 2	485
Case 3	523

Profit as a function of buffer sizes



- Three-machine, continuous material line.
- $r_i = .1$, $p_i = .01$, $\mu_i = 1$.
- $\Pi = 1000P(N_1, N_2) - (\bar{n}_1 + \bar{n}_2)$.