Performance of Computer System

Introduction to Queuing Model

Dr. Lina Xu
lina.xu@ucd.ie

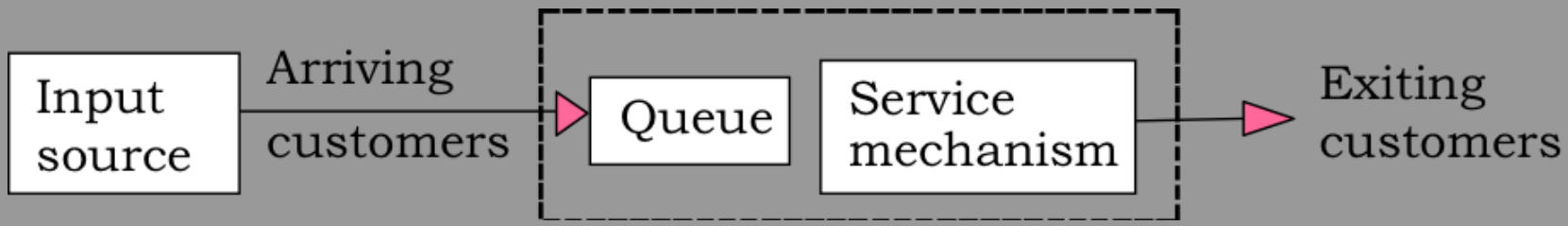
School of Computer Science,
University College Dublin

October 6, 2018

Overview on the Topics

- Basic structure and components
- Performance measures
- Steady-state analysis and Little's law
- Birth-death processes
- Single-server and multi-server examples
- Flow balance equations

Structure of Single Queuing Systems



- Customers need not be people; other possibilities include parts, vehicles, machines, jobs.
- Queue might not be a physical line; other possibilities include customers on hold, jobs waiting to be printed, planes circling airport.

Components of Model

Input Source

- The size of the "calling population" may be modeled as infinite or finite.
- Calculations are easier in the infinite case and in many cases this is a reasonable approximation (bank, pizza parlor, blood bank).

Queuing Discipline

- First-come first-served (FCFS or FIFO) is the most frequent assumption, but priority ordering is important in some settings

Service Mechanism

- One or more servers may be placed in parallel.

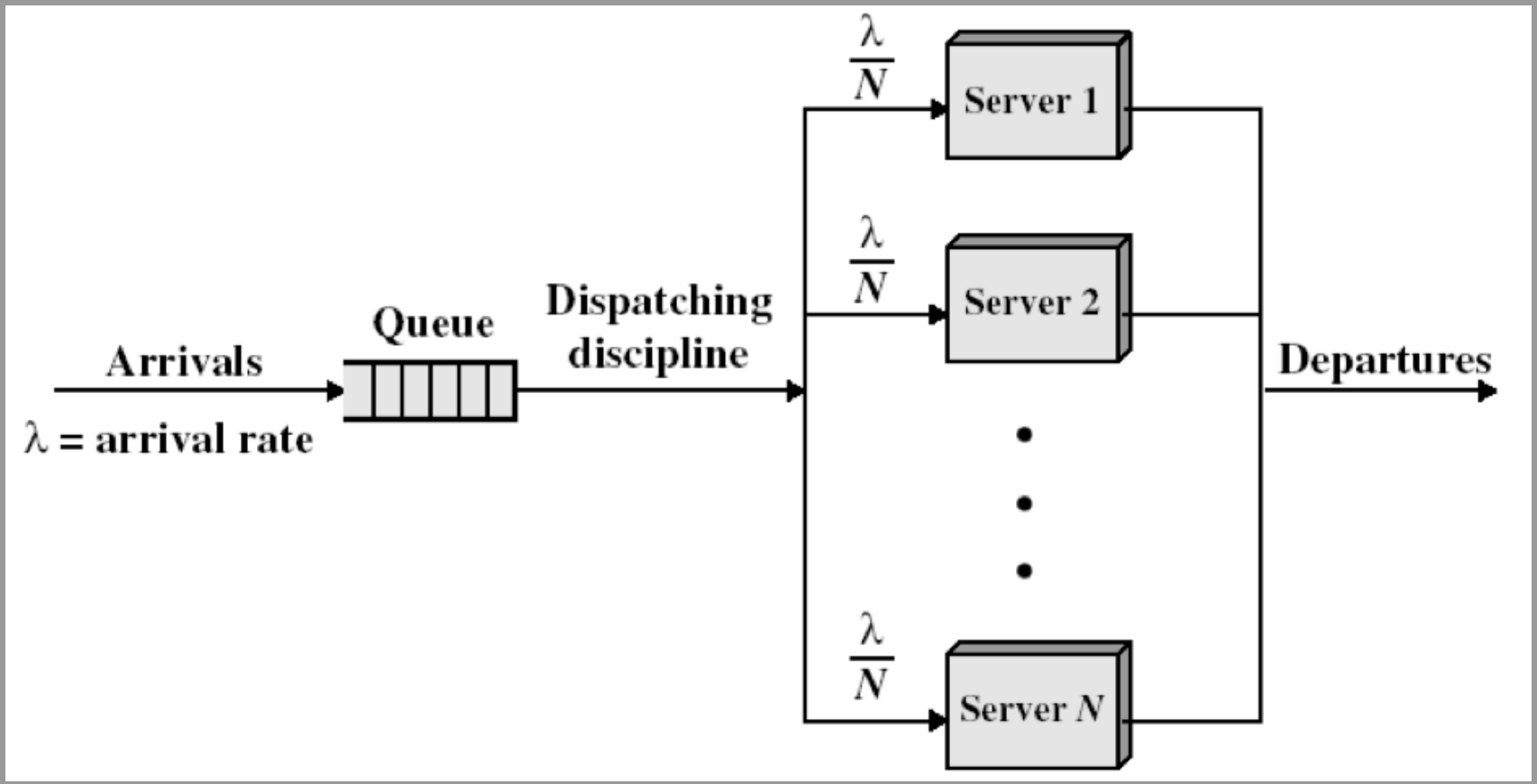
Queuing Applications

System	Arrival Process	Service Process
Bank	Customers Arrive	Tellers serve customers
Pizza parlor	Orders are phoned in	Orders are driven to customers
Blood bank	Pints of blood arrive via donation	Patients use up pints of blood
Shipyard	Damaged ships sent to shipyard for repair	Ships are repaired & return to sea
Printers	Jobs arrive from computers	Documents are printed

Typical Performance Questions

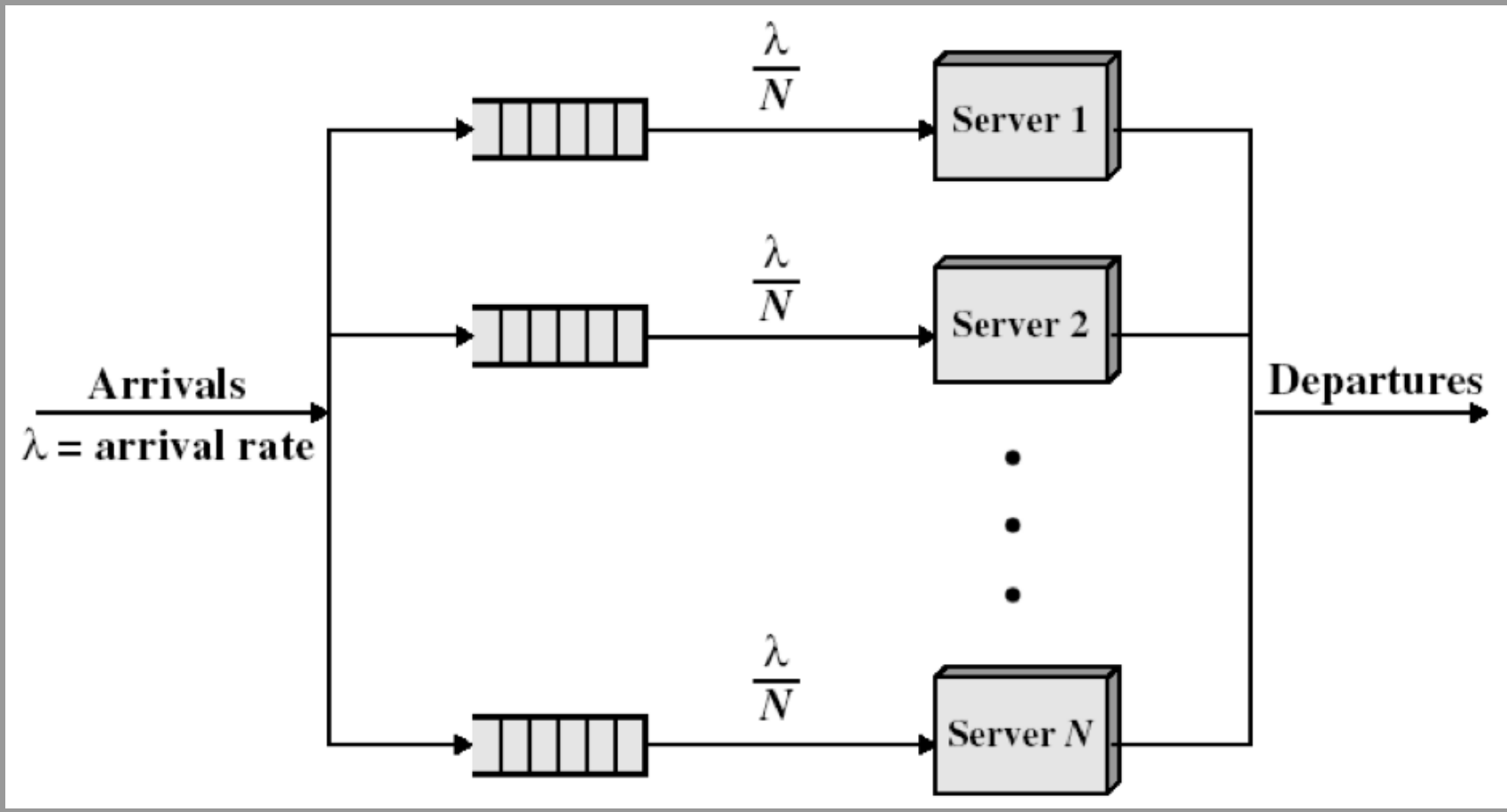
- What is the ...
  - average number of customers in the system?
  - average time a customer spends in the system?
  - probability a customer is rejected?
  - fraction of time a server is idle?
- These questions are aimed at characterizing complex systems.
- Analyses results are used to support decision-making.
- In queuing (and most analyses of complex stochastic systems), design takes the form of asking "what if" questions rather than trying to optimise the design.
  - How much the average waiting time can be reduced if the severing rate is doubled?
  - What is the average waiting time if the input rate is tripled?
- Queuing system can help to maximise system performance with limited resource.

Multiple Servers, Single Queue



- What is average wait time/jobs in the queue?
- What is average time in the system?

Multiple Servers, Single Queue



- What is average wait time/jobs in the queue?
- What is average time in the system?

Notation and Terminology

- $N(t)$ = Number of customers in the system at time $t \geq 0$
- $P_k(t)$ = probability exactly k customers in system at time t, given the number in system at time 0
- s = Number of parallel servers
- λ_k = mean arrival rate (expected number of arrivals per unit time)
- μ_k = mean service rate (expected number of departures per unit time)

Note: (Both λ_k and μ_k assume k customers are in system)

If...

- If λ_k does not depend on the number of customers in system, $\lambda_k = \lambda$.
- If there are s servers, each with the same service rate, then
  - $\mu_k = k\mu$ for $0 \leq k < s$.
  - $\mu_k = s\mu$ for $k \geq s$.
  - $s\mu$ is customer service capacity per unit time.
  - $\rho = \frac{\lambda}{s\mu}$, is utilisation factor (traffic intensity).
- The systems we study will have $\rho < 1$ because otherwise the number of customers in the system will grow without bound.
- We will be interested in the steady-state behaviour of queuing systems (the behaviour for t large).
- Obtaining analytical results for $N(t)$, $P_k(t)$, ... for arbitrary values of t (the transient behaviour) is much more difficult.

Notation for Steady-State Analysis

- π_k = probability of having exactly k customers in the system
- L = expected number of customers in the system
- L_q = expected queue length (doesn't include those being served)
- W = expected time in system, including service time
- W_q = expected waiting time in the queue (doesn't include service)

Little?s Law by John D.C. Little

- For any queuing system that has a steady state and has an average arrival rate of λ,
  - $L = \lambda W$
  - For example, if the average waiting time is 2 hours and customers arrive at a rate of 3 per hour then, on average, there are 6 customers in the system.
- Similarly,
  - $L_q = \lambda W_q$
- If $\mu k = \mu$ for all $k \geq 1$ then $W = W_q + \frac{1}{\mu}$, $\frac{1}{\mu}$ is the mean service time here.

Benefit of Little?s Law

- These three relationships allow us to calculate all four quantities L, L_q, W and W_q, once **one** of them is known
- $L = \lambda W$ requires no assumptions about arrival or service time distributions, the size of the calling population, or limits on the queue.

Little's Law—Example

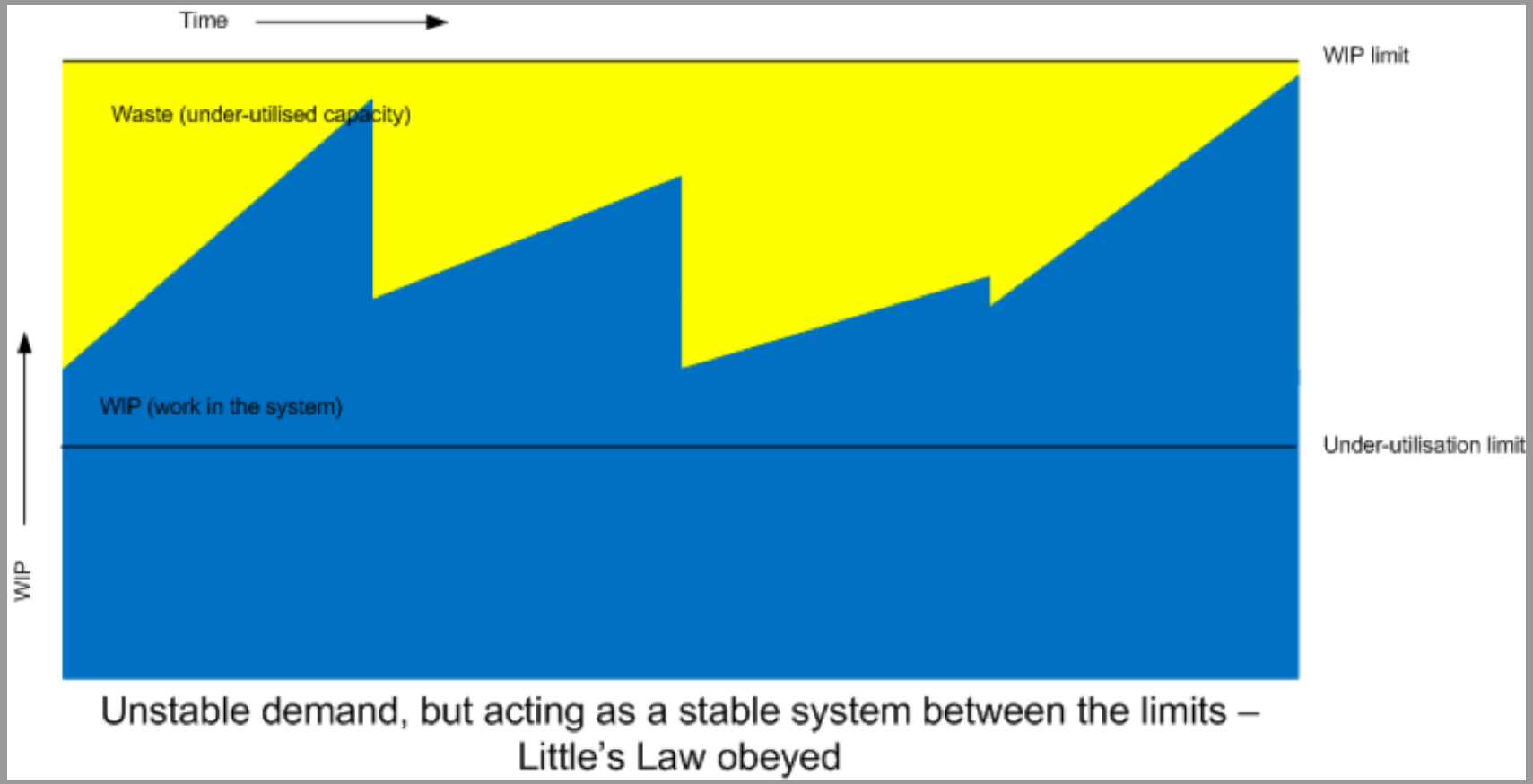
A monitor on a disk server showed that the average time to satisfy an I/O request was 100 milliseconds. The I/O rate was about 100 requests per second. What was the mean number of requests at the disk server?

- Using little's law,
- Mean number in the disk server=arrival rate $\times$ response time
  =(100 requests/second) $\times$ (0.1 second)
  =10 requests

Work In Process (WIP)

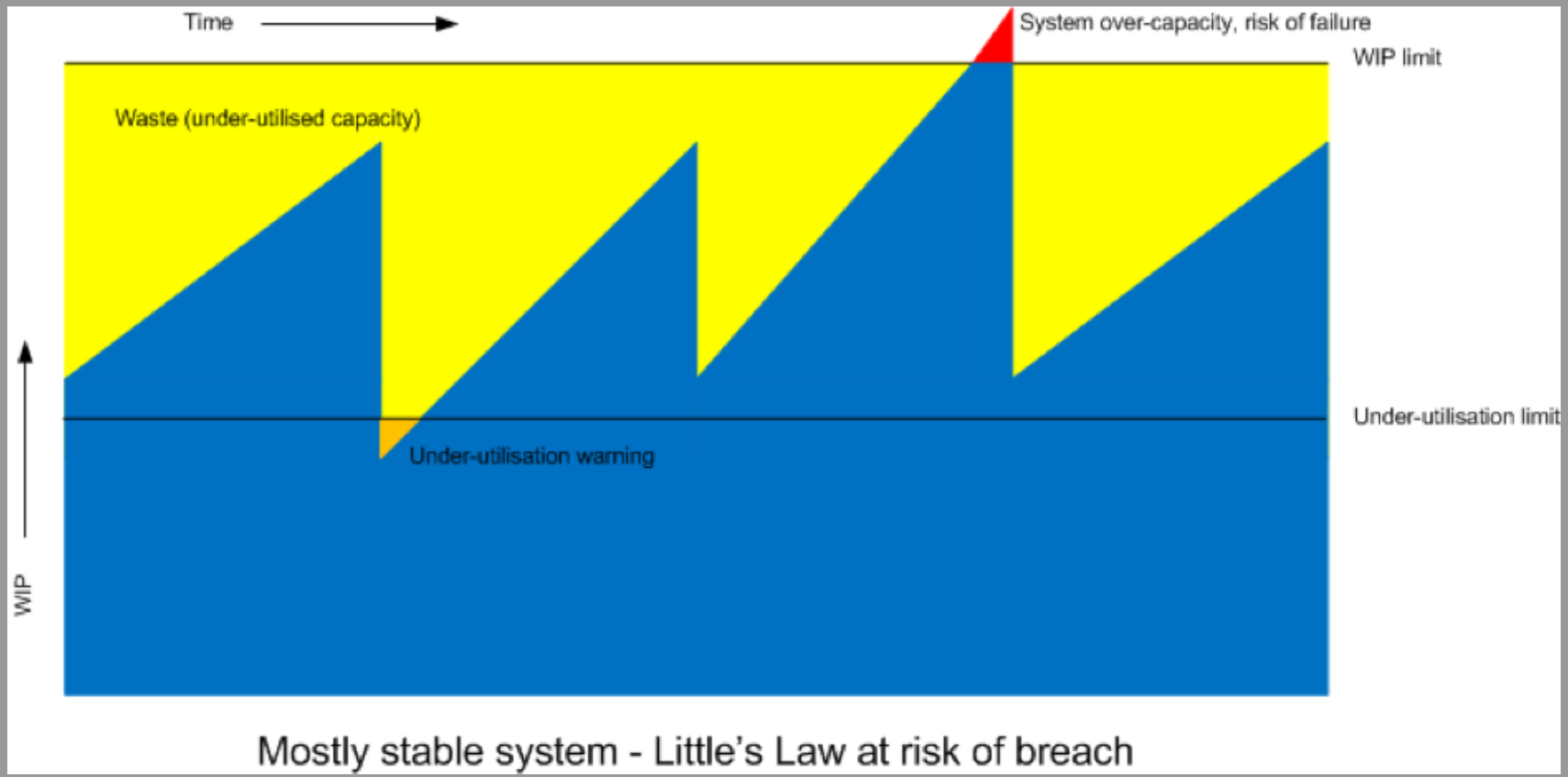
- Little?s law is also only applicable if your arrival rate (or at least the average) is equal to the departure rate.
- This can be more difficult to keep consistent, but the easiest way to deal with this is to alter your arrival rate to match the departure rate.
  - This means not accepting new projects or starting work on new tasks until a current one is completed.
- Work that has entered the development process but is not yet finished and available to a customer or user. Refers to all assets or work products of a product or service that are currently being worked on or waiting in a queue to be worked on.
- In order to maintain steady state, a WIP limit is defined before the bottleneck. This ensures the “system” (a particular step in the process) is protected from overload by limiting the arrival rate.

Work In Process (WIP)



Unstable demand, but acting as a stable system between the limits – Little's Law obeyed

Work In Process (WIP)



Work In Process (WIP) Example

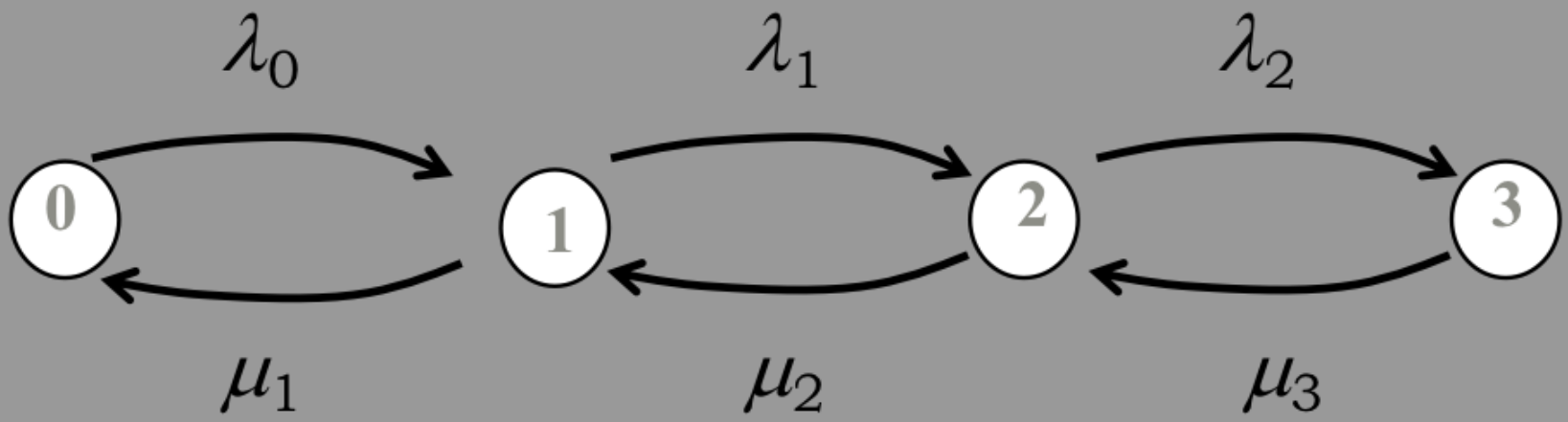
- Semiconductor devices are manufactured in extremely capital-intensive fabrication facilities. The manufacturing process entails starting with a silicon wafer and then building the electronic circuitry for multiple identical devices through hundreds of process steps.
- Suppose that the semiconductor factory starts 1,000 wafers per day, on average; this is the input rate. The start rate has remained fairly stable over the past 9 months. We track the amount of WIP inventory.
- The WIP varies between 40,000 and 50,000 wafers; the average WIP is 45,000 wafers.

What is λ, L and W?

Birth and Death Processes

- Applications in a variety of areas, but in queuing model, "birth" refers to the arrival of a customer while "death" refers to the departure of a customer.
- Recall, $N(t)$ = The number of customers in the system at time t.
- Assumptions
  - Given that $N(t) = k$, the Probability Density Function (PDF) governing the remaining time until the next birth (arrival) is $exp(\lambda_k)$, $k = 0, 1, 2,...$
  - Given that $N(t) = k$, the PDF governing the remaining time until the next death (service completion) is $exp(\mu_k)$, $k = 0, 1, 2,...$
  - All random variables are assumed to be independent

Evaluation Examples



- We will investigate steady-state (not transient) results for birth-death processes based on the
- **Expected rate in = Expected rate out** principle
- Let π_k = steady-state probability of being in state k.

Balance Equations: In = Out

Expected rate in = Expected rate out principle

- Flow into State 0 $\rightarrow \mu_1\pi_1 = \lambda_0\pi_0 \leftarrow$ Flow out of State 0
- Flow into State 1 $\rightarrow \lambda_0\pi_0 + \mu_2\pi_2 = \lambda_1\pi_1 + \mu_1\pi_1 \leftarrow$ Flow out of State 1
- Flow into State 2 $\rightarrow \lambda_1\pi_1 + \mu_3\pi_3 = \lambda_2\pi_2 + \mu_2\pi_2 \leftarrow$ Flow out of State 2
- ...
- Flow into State k $\rightarrow \lambda_{k-1}\pi_{k-1} + \mu_{k+1}\pi_{k+1} = \lambda_k\pi_k + \mu_k\pi_k \leftarrow$ Flow out of State k

Balance Equations: Deduction

$$\pi_1 = \frac{\lambda_0}{\mu_1}\,\pi_0\,, \qquad \pi_2 = \frac{\lambda_1}{\mu_2}\,\pi_1 + \frac{1}{\mu_2}\,(\mu_1\pi_1 - \lambda_0\pi_0) = \frac{\lambda_1}{\mu_2}\,\pi_1 = \frac{\lambda_1\lambda_0}{\mu_2\mu_1}\,\pi_0$$

$= \mathbf{0}$

$$\pi_k = \left(\frac{\lambda_{k-1}\ldots\lambda_0}{\mu_k\ldots\mu_1}\right)\pi_0$$

Let $\pi_k = C_k\pi_0$, $k = 1, 2, \ldots$ $C_0 = 1$, and $\Sigma_k\pi_k = 1$

so $(C_0 + C_1 + C_2 + \cdots)\pi_0 = 1$ or $\pi_0 = 1 / (1 + \Sigma_k C_k)$

Some Steady-State Results

Once we have calculated the π_k, we can find:

- $L = \sum_{k=1}^{\infty} k\pi_k$, expected number of customers **in the system**
- $L_q = \sum_{k=s+1}^{\infty} (k-s)\pi_k$, expected number of customers **in the queue**
- $L_s = L - L_q$, expected number of customers **in service**
- $\bar{\lambda} = \sum_{k=0}^{\infty} \lambda_k \pi_k$, average (effective) arrival rate
- $E = L_s/s$, efficiency of system (utilisation)

Queuing Example with 3 Servers

- Arrival rate: $\lambda = 5/\text{hr}$
- Service rate: $\mu = 2/\text{hr}$
- Arriving customer balks when 6 are in system.
- Steady-state probabilities:

State	0	1	2	3	4	5	6
Component	π_0	π_1	π_2	π_3	π_4	π_5	π_6
Probability	0.068	0.170	0.212	0.177	0.147	0.123	0.102

Determining System Characteristics

What is the probability that all servers are idle?

Determining System Characteristics

What is the probability that all servers are idle?

- $\Pr\{\text{all servers idle}\} = \pi_0 = 0.068$

Determining System Characteristics

What is the probability that a customer will not have to wait?

Determining System Characteristics

What is the probability that a customer will not have to wait?

- $\Pr\{\text{no wait}\} = \pi_0 + \pi_1 + \pi_2 = 0.45$

Determining System Characteristics

What is the probability that a customer will have to wait?

Determining System Characteristics

What is the probability that a customer will have to wait?

- $\Pr\{\text{wait}\} = 1 - \Pr\{\text{no wait}\} = 0.55$

Determining System Characteristics

What is the probability that a customer balks?

Determining System Characteristics

What is the probability that a customer balks?

- $\Pr\{\text{customer balks}\} = \pi_6 = 0.102$

Steady-State Measures for Example

Expected number in queue:

- $L_q = 1\pi_4 + 2\pi_5 + 3\pi_6 = 0.700$

Steady-State Measures for Example

Expected number in service:

- $L_s = \pi_1 + 2\pi_2 + 3(1 - \pi_0 - \pi_1 - \pi_2) = 2.244$

Steady-State Measures for Example

Expected number in the system:

- $L = L_q + L_s = 2.944$

Steady-State Measures for Example

Efficiency of the servers:

- $E = L_s/s = 2.244/3 = 0.748$ or 74.8%

Little's Law with Average Arrival Rate

Applying Little's Law with $\bar{\lambda}$, we can calculate

$$W = L / \bar{\lambda} \qquad \& \qquad W_q = L_q / \bar{\lambda}$$

Expected waiting time in the system

Expected waiting time in the queue

Results assume that steady state will be reached.

Example with 3 Servers (cont.)

To compute average waiting times we must first find the average arrival rate:

$$\lambda = \sum \lambda_k \pi_k \quad \text{where } \lambda_k = \lambda = 5 \ (k = 0,1,\ldots,5) \text{ and } \lambda_k = 0 \ (k > 5), \text{ simplifies to}$$

$$= \lambda(\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5)$$

$$= \lambda(1 - \pi_6) = 5(1 - 0.102) = 4.488 \ / \ \text{hour}$$

Consequently,

$$W_s = L_s / \bar{\lambda} = 0.5 \text{ hours}$$

$$W_q = L_q / \bar{\lambda} = 0.156 \text{ hours}$$

$$W = L / \bar{\lambda} = 0.656 \text{ hours}$$