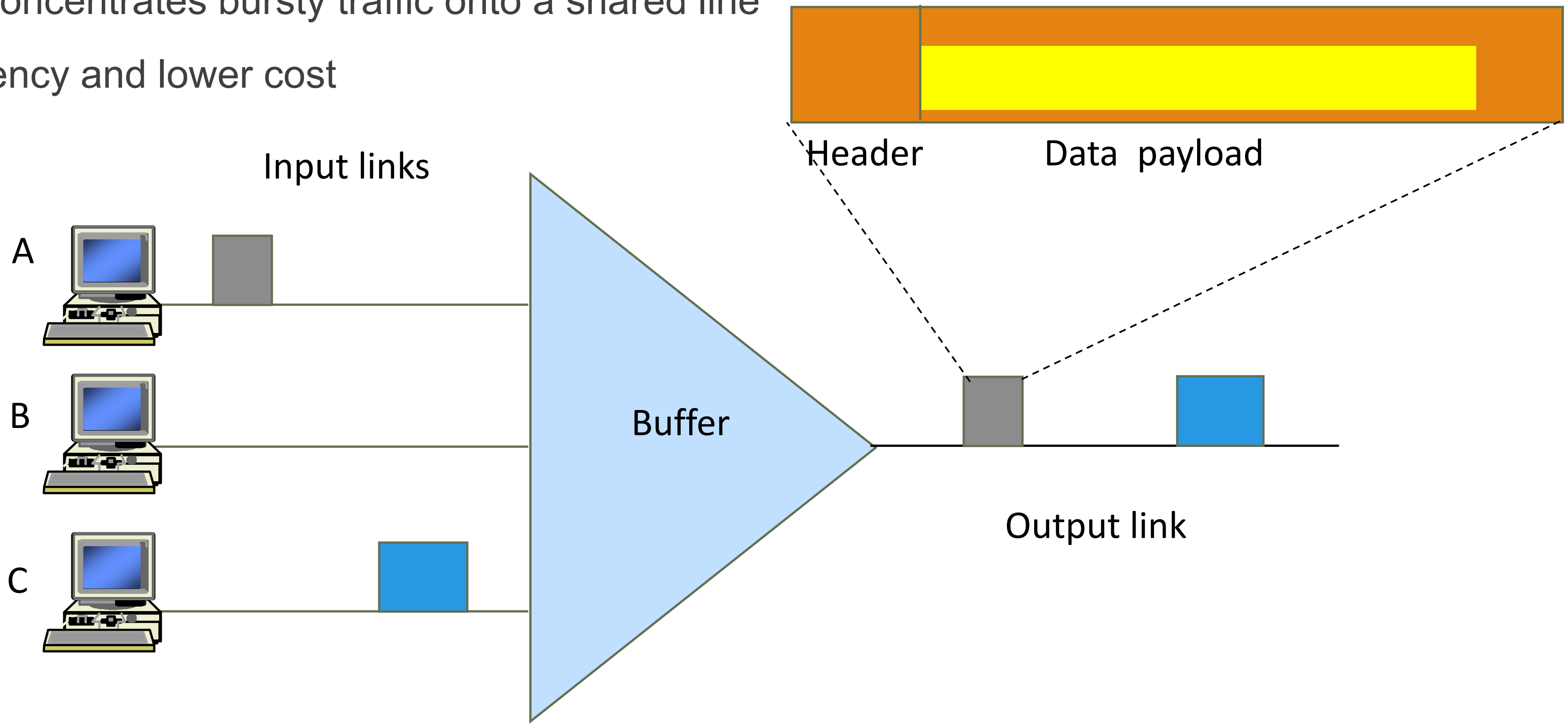
Statistical Multiplexing & Queuing Analysis

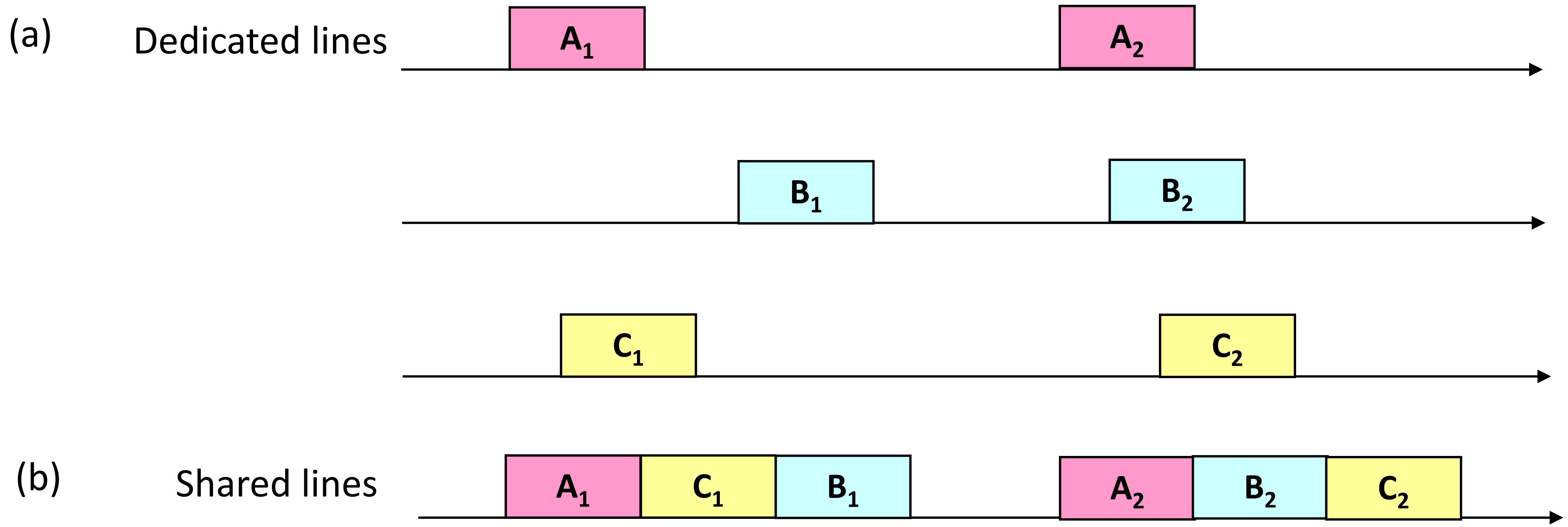
A/PROF. DUY NGO

Statistical Multiplexing

- Multiplexing concentrates bursty traffic onto a shared line
- Greater efficiency and lower cost

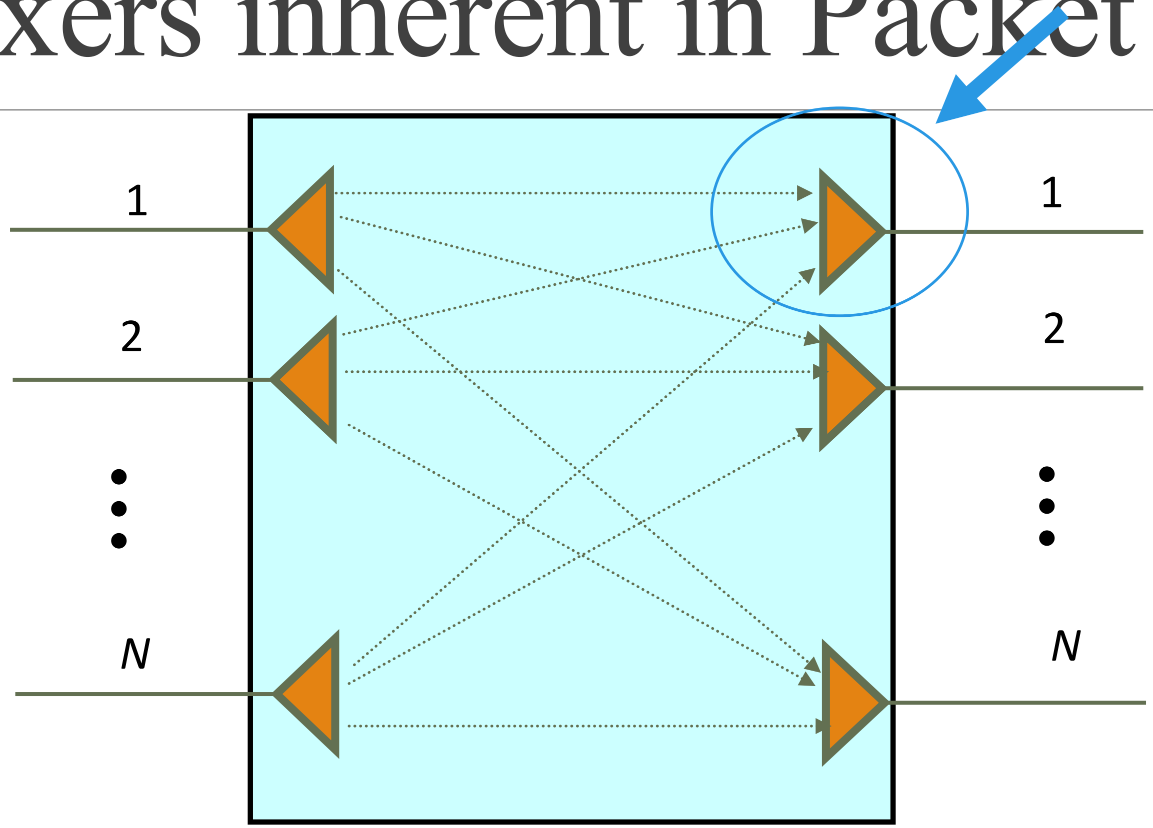


Tradeoff Delay for Efficiency



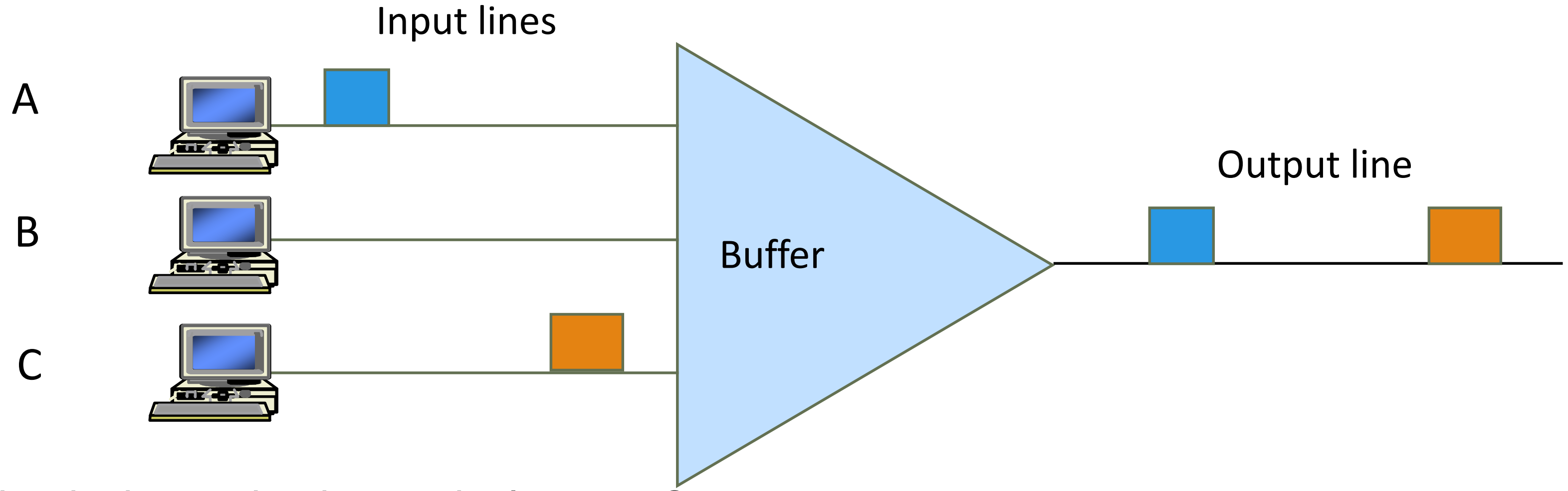
- Dedicated lines involve not waiting for other users, but lines are used inefficiently when user traffic is bursty
- Shared lines concentrate packets into shared line; packets buffered (delayed) when line is not immediately available

Multiplexers inherent in Packet Switches



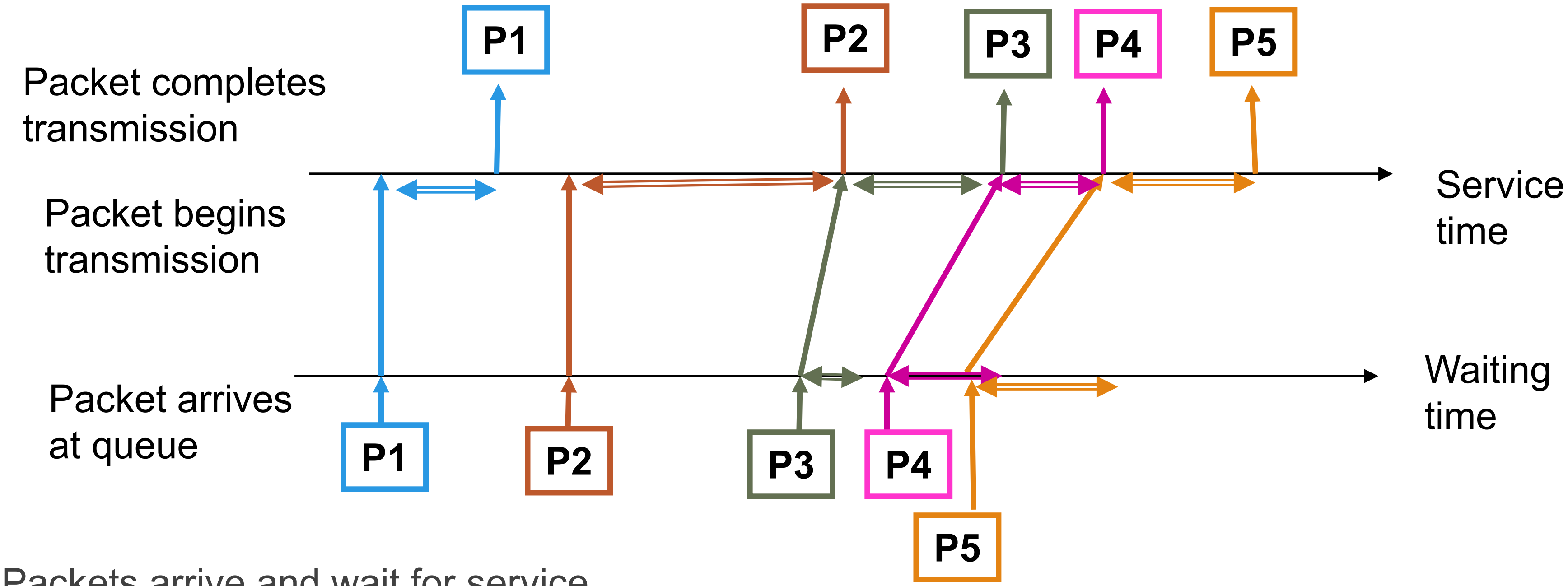
- Packets/frames forwarded to buffer prior to transmission from switch
- Multiplexing occurs in these buffers

Multiplexer Modeling



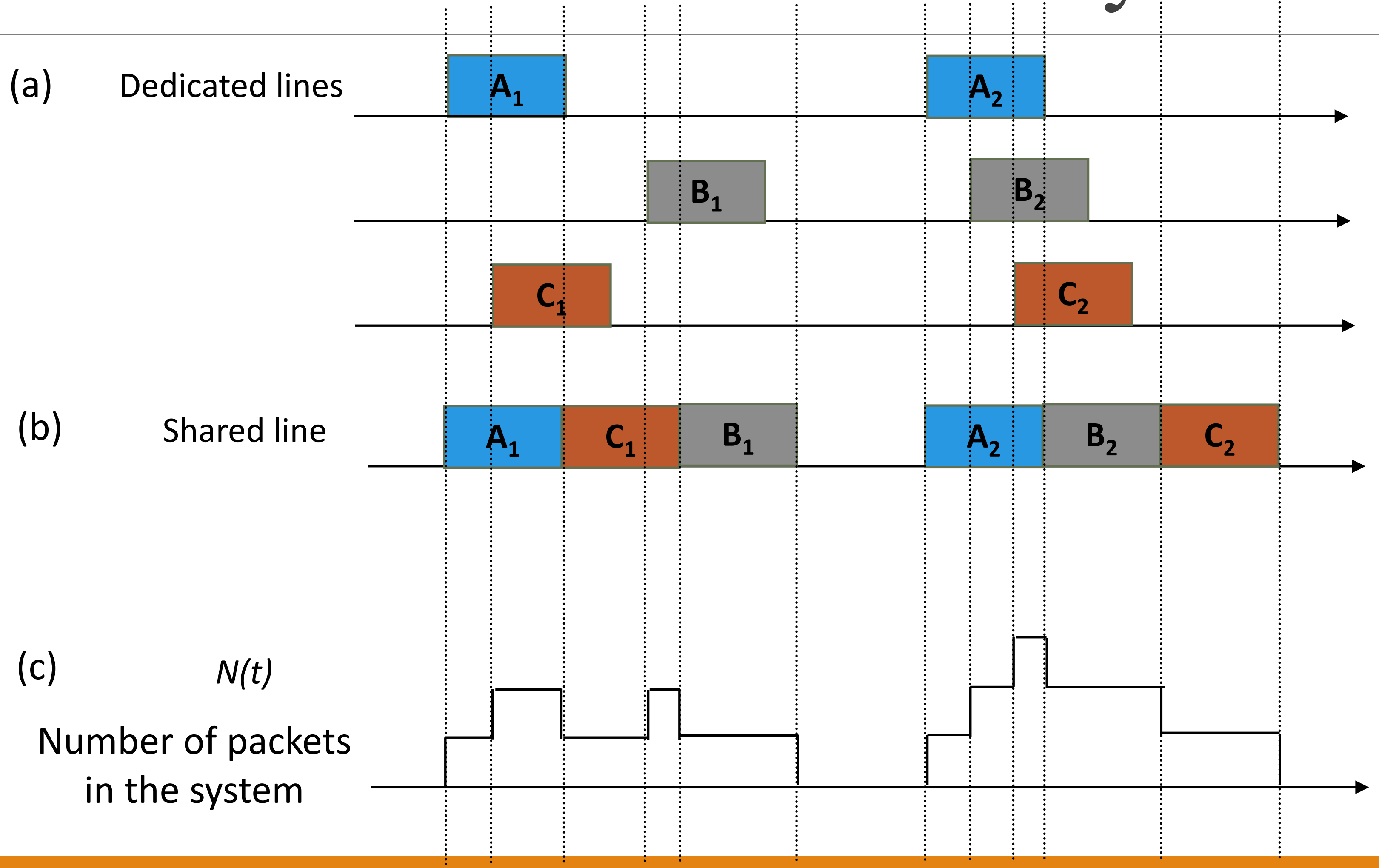
- Arrivals: what is the packet interarrival pattern?
- Service time: how long are the packets?
- Service discipline: what is order of transmission?
- Buffer discipline: if buffer is full, which packet is dropped?
- Performance measures:
- *Delay distribution; packet loss probability; line utilization*

Delay = Waiting + Service Times



- Packets arrive and wait for service
- Waiting time: from arrival instant to beginning of service
- Service time: time to transmit packet
- Delay: total time in system = waiting time + service time

Fluctuations in Packets in the System



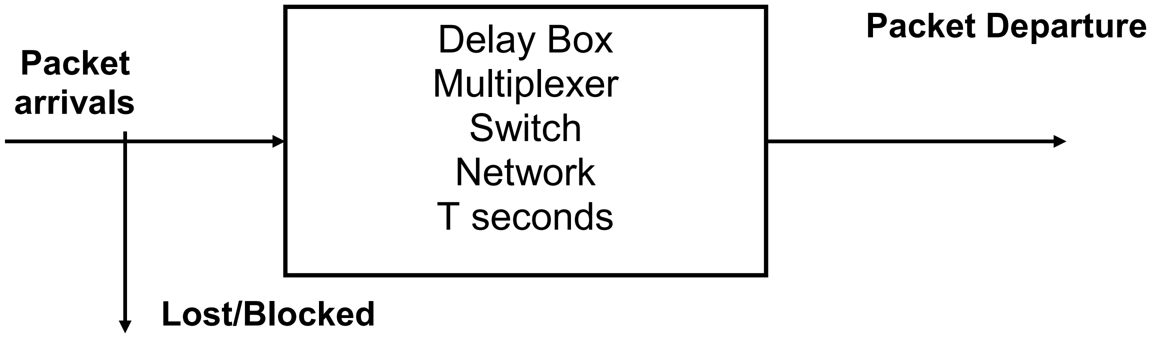
Packet Lengths & Service Times

- R bits per second transmission rate
- L = # bits in a packet
- $X = L/R$ = time to transmit ("service") a packet
- Packet lengths are usually variable
  - Distribution of lengths → dist. Of service times
  - Common models:
    - Constant packet length (all the same)
    - Exponential distribution
    - Internet measured distributions fairly constant

Basic Queuing Theory

- One of the most important measure of performance of a data network is the average delay required to deliver a packet from origin to the destination
- It is important to understand the nature and mechanism of delay, and the manner in which it depends on the characteristics of a network
- Queuing theory is the primary methodological framework for analyzing network delay.
- Queuing analysis mostly applies to a *delay* system where call requests are could be queued when a system is unable to offer any capacity

Queuing System



- We are interested in the following performance parameters:
  - Time spent in the system/queue
  - Number of packets in the system: *n(t)*
  - Fraction of arriving packets/calls that are lost or blocked
  - Average throughput

Arrival Rates and Traffic Load

- Let $A(t)$ be the number of packet arrivals at the system in the interval of 0 to t.
- Let $B(t)$ be the number of blocked packets and $D(t)$ be the number of departed packets
- The number of packets in the queue is $N(t) = A(t) - D(t) - B(t)$
- Assuming the system is empty at t=0, the long term (steady state) arrival rate is given by:

$$\lambda = \underset{t \to \infty}{Lt} \frac{A(t)}{t}$$

- The throughput of the system is equal to the long term departure rate, which is given by:

$$throughput = \underset{t \to \infty}{Lt} \frac{D(t)}{t} calls / \sec$$

Arrival Rates and Traffic Load

- The average no. of packets in the system is given by:

$$E[N] = \underset{t\to\infty}{Lt} \frac{1}{t} \int_0^t N(t')dt'$$

- The fraction of blocked packets is given by: $P_b = \underset{t\to\infty}{Lt} \frac{B(t)}{A(t)}$

- Long time arrival rate is given by:

$$\lambda = \underset{n\to\infty}{Lt} \frac{n}{\tau_1 + \tau_2 + \ldots + \tau_n} = \underset{n\to\infty}{Lt} \frac{1}{(\tau_1 + \tau_2 + \ldots + \tau_n)/n}$$

$$= \frac{1}{E[\tau]}$$

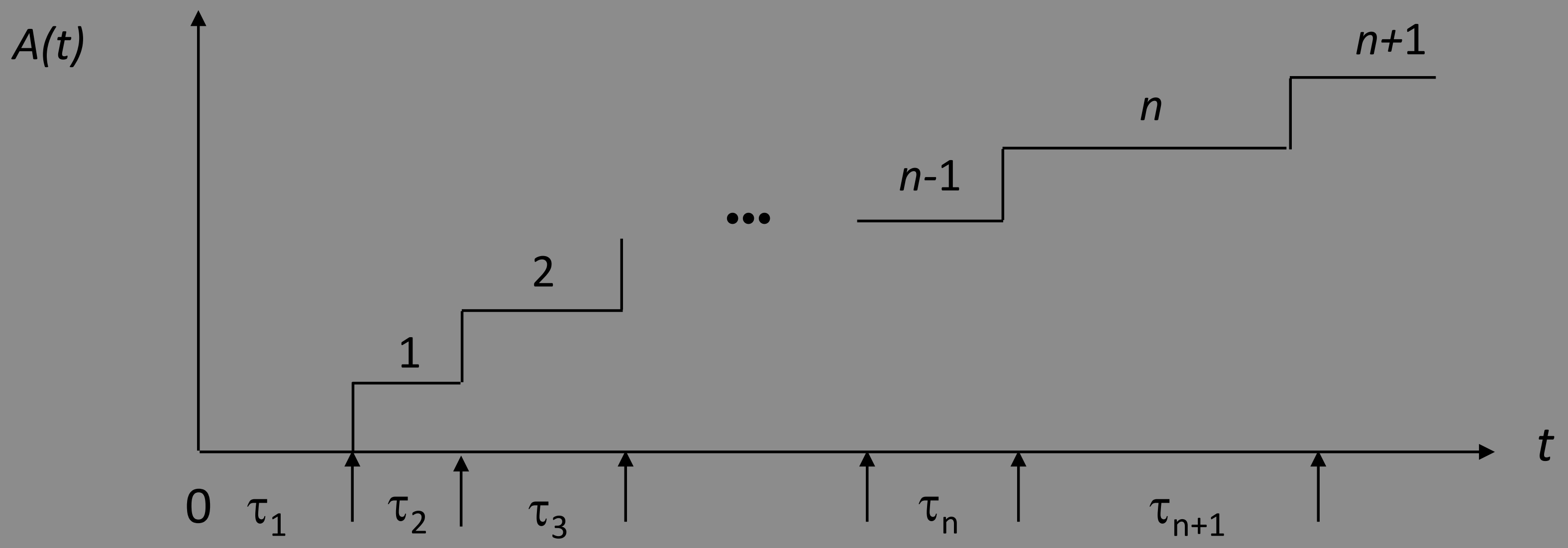
Little's Formula

- Considering a system where calls or packets (for data network) arrive in random to obtain network services. Service time of a packet is *L/C* where *L* is the packet length and *C* is the transmission rate (service rate).
- Little's theorem could be used to estimate following quantities:
  - Average no. of calls/packets in a system
  - Average delay to service a call or a packet
- Using the simplistic form of the little's theorem, number of calls in a system can be calculated:

$$N = \lambda T$$

- Probabilistic form is:

$$E[N] = \lambda E[T]$$



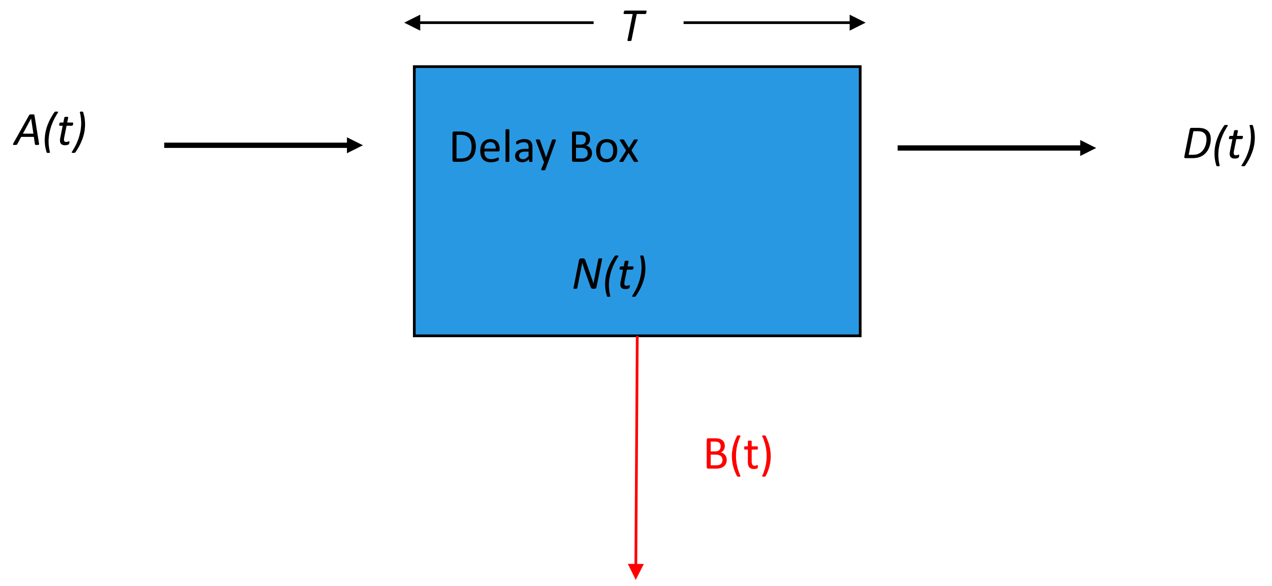
Time of nth arrival = $\tau_1 + \tau_2 + \ldots + \tau_n$

$$\text{Arrival Rate} = \frac{n \text{ arrivals}}{\tau_1 + \tau_2 + \ldots + \tau_n \text{ seconds}} = \frac{1}{(\tau_1+\tau_2+\ldots+\tau_n)/n} \longrightarrow \frac{1}{E[\tau]}$$

Arrival Rate = 1 / mean interarrival time

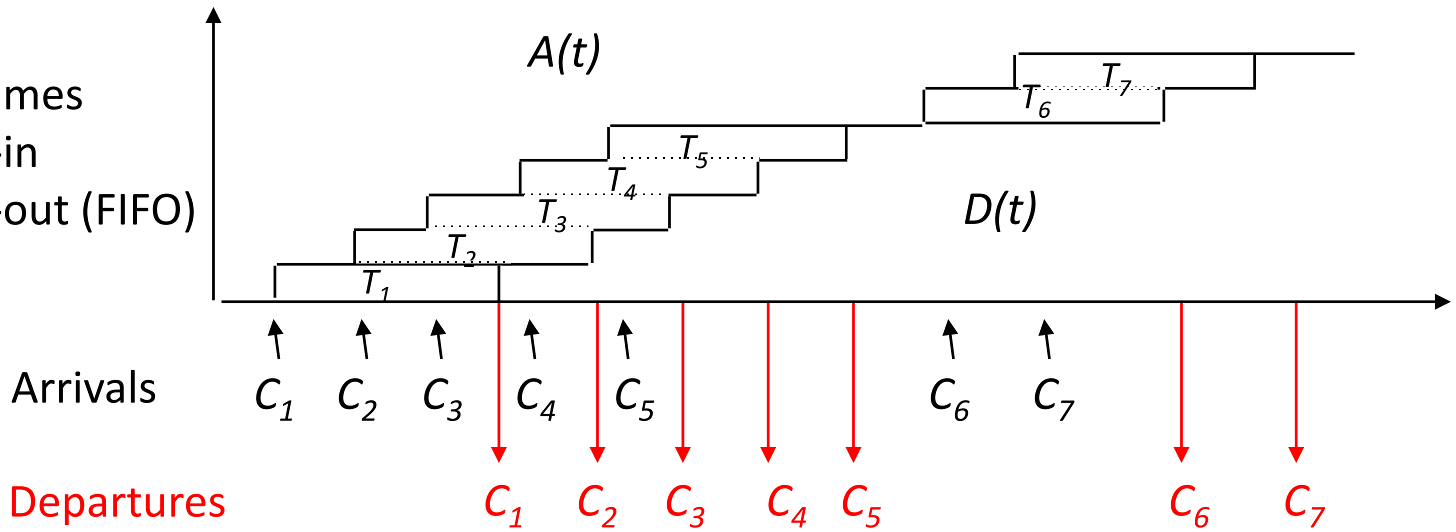
Basic Queuing Model

- A simple queue model



Packet Arrival/Departure: FIFO

Assumes first-in first-out (FIFO)



Little's Formula: FIFO System

- Time average number of calls in a system up to time is :

$$\frac{1}{t_o}\int_0^{t_o} N(t')dt' = \frac{1}{t_o}\left\{\sum_{j=1}^{A(t_o)} T_j\right\}$$

  - Dividing both sides by $A(t_o)$

$$\frac{1}{t_o}\int_0^{t_o} N(t')dt' = \frac{A(t_o)}{t_o}\left\{\frac{1}{A(t_o)}\sum_{j=1}^{A(t_o)} T_j\right\}$$

- The first term on the left side of the equation is the average arrival time and the second term is the expected time spent by calls

Little's Formula: FIFO System

- Considering a system where some calls could be blocked, the little's formula is modified as below, where P_b is the probability of blocking

$$E[N] = \lambda(1 - P_b)E[T]$$

- Little's formula can be extended for a network where a link consists of M multiplexers or switches

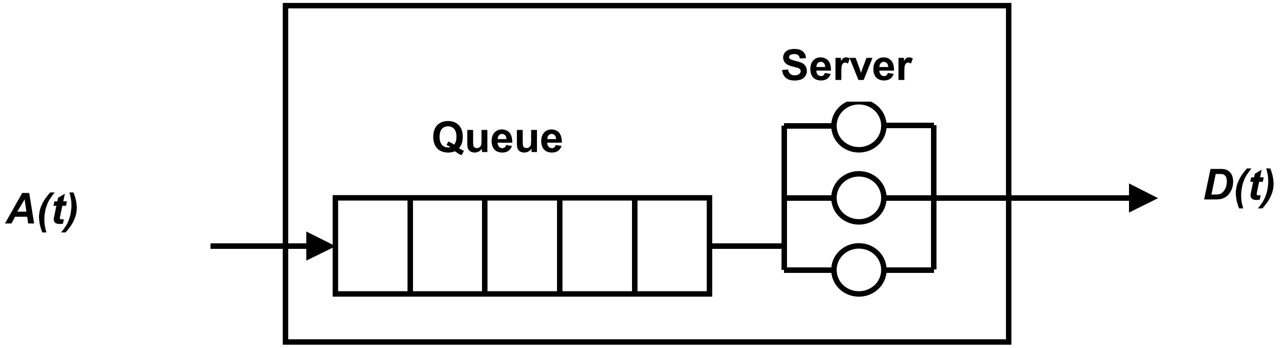
$$E[N_{net}] = \lambda_{net} E[T_{net}] = \sum \lambda_m E[T_m]$$

$$E[T_{net}] = E[N_{net}] \big/ \lambda_{net} = \frac{1}{\lambda_{net}} \sum_m \lambda_m E[T_m]$$

Basic Queuing Model

- Work done by A. K. Erlang, a famous Dutch telecommunications engineer lead to the fundamental development of models to analyse resources sharing systems such multiplexers, switches, etc.

- In a telecommunication system calls/packets arrives randomly and use resources for a random period of time. In a delay system when all system resources are busy, new arrived calls are kept in a 'queue' until a suitable resource is available.



Basic Queuing Model

- Arrival process is very important in communication network. Traffic arrival could be deterministic when the interarrival times are equal and constant
- Arrival process is considered to be exponential, if the inter-arrival times are exponential random variables with mean $E[\tau] = 1/\lambda$. Exponential process is described by the following equation.

$$P[\tau > t] = e^{-t/E[\tau]} = e^{-\lambda t}$$

- For exponential interarrival times, the number of arrivals *A(t)* in an interval of length *t* is given by a Poisson random variable with mean $E[A(t)]=\lambda t$:

$$P[A(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

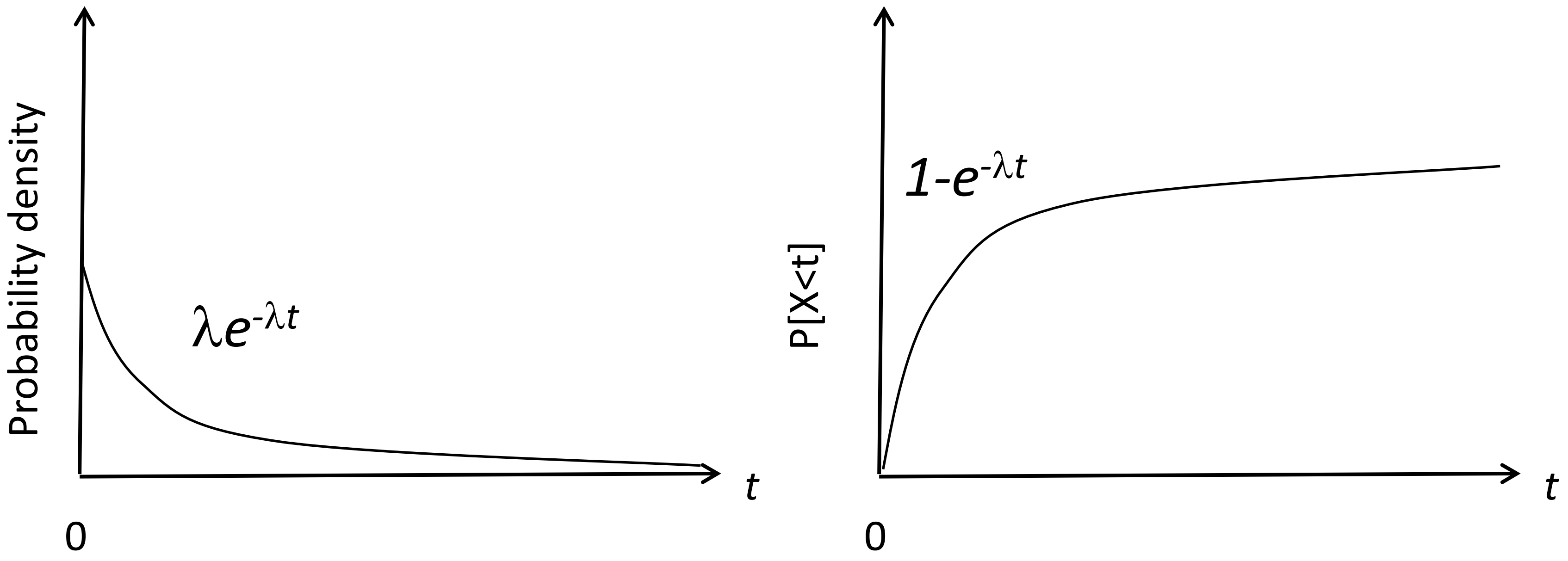
Poisson Arrivals

- Average arrival rate: λ packets per second
- Arrivals are equally-likely to occur at any point in time
- Time between consecutive arrivals is an exponential random variable with mean $1/\lambda$
- Number of arrivals in interval of time *t* is a *poisson* random variable with mean λt

$$P[\text{ k arrivals in t seconds}] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Exponential Distribution

$$P[X > t] = e^{-t/E[X]} = e^{-\lambda t} \quad \text{for } \; t > 0\,.$$



Performance of a Packet Statistical Multiplexer

- Let λ packets/sec be the average packet arrival rate to a multiplexer
- If $\lambda > \mu$, then the buffer build up and packets could be lost
- If $\lambda < \mu$, number of packets can fluctuate and transmission of long packets may cause buffer overflow
- Buffer overflow can be prevented by increasing the buffer size
- Load is defined as $\rho = \lambda/\mu$, when $\lambda < \mu$ then $\rho < 1$
- Statistical multiplexing technique can be analyzed for different queuing system
- Book describe the performance of a statistical multiplexer using M/M/1/K queuing system

Queuing System: Kendall's Notation

Input specifications:

- G: general (no assumptions)
- M: purely random

Service time distribution:

- G: general (no assumptions)
- M: negative exponential service time distribution
- D: constant

N: number of servers (finite number)

L: number of sources (finite length)

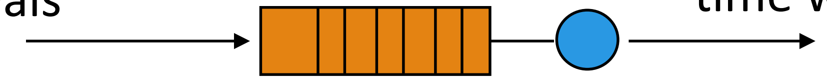
∞: queue length (infinite length)

M/M/1/K Queueing Model

Poisson Arrivals rate λ

$K - 1$ buffer

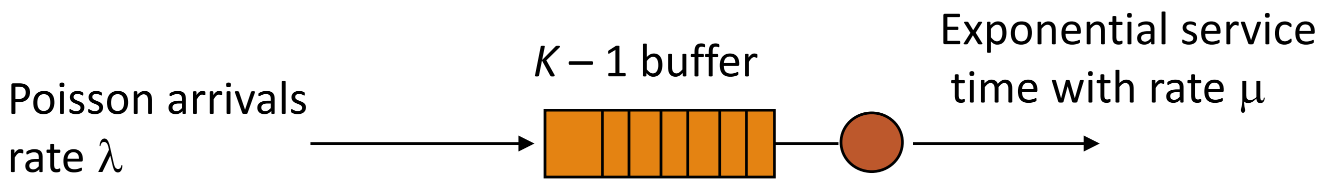
Exponential service time with rate μ



At most K *packets allowed in system*

- 1 packet served at a time; up to $K - 1$ can wait in queue
- Mean service time $E[X] = 1/\mu$
- Key parameter load: $\rho = \lambda/\mu$
- When $\lambda << \mu$ ($\rho \approx 0$), packets arrive infrequently and usually find system empty, so delay is low and loss is unlikely
- As λ approaches μ ($\rho \to 1$) , packets start bunching up and delays increase and losses occur more frequently
- When $\lambda > \mu$ ($\rho > 0$) , packets arrive faster than they can be processed, so most packets find system full and those that do enter have to wait about $K - 1$ service times

State Model of a Buffer



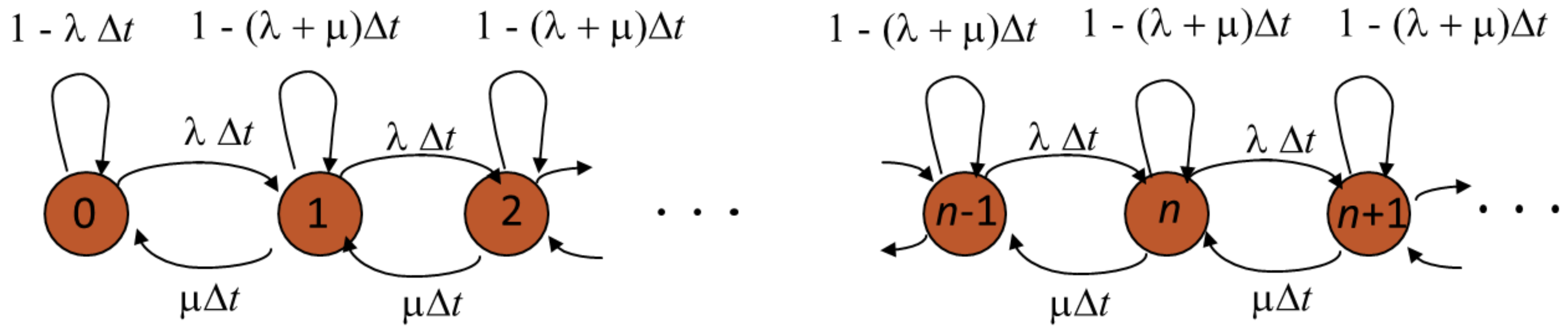


Figure A.9

M/M/1/K Performance Equations

Probability of Overflow:

$$P_{loss} = \frac{(1-\rho)\rho^{K}}{1-\rho^{K+1}}$$

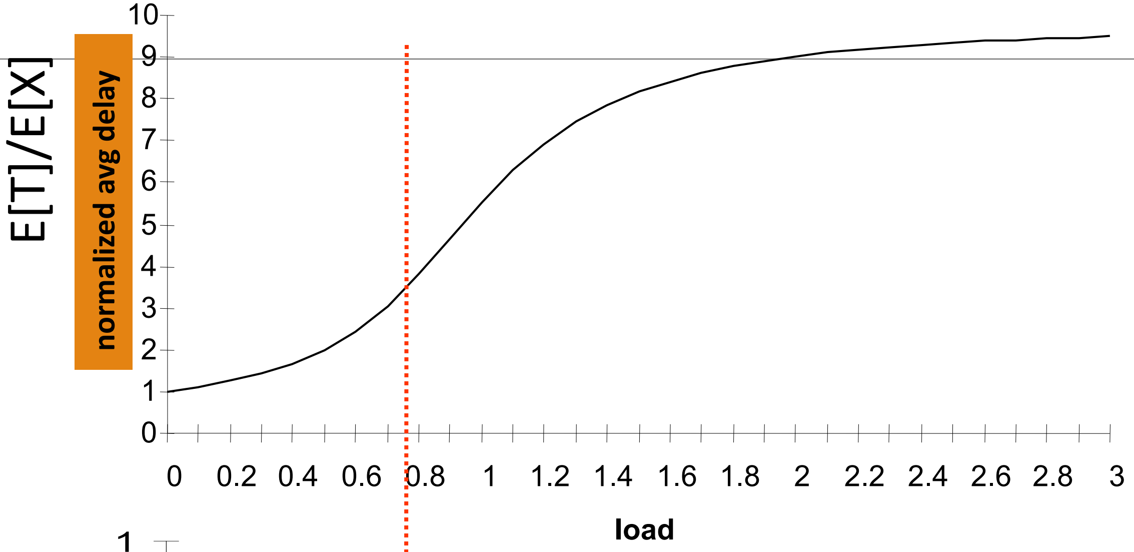
Average number of packets in the queue:

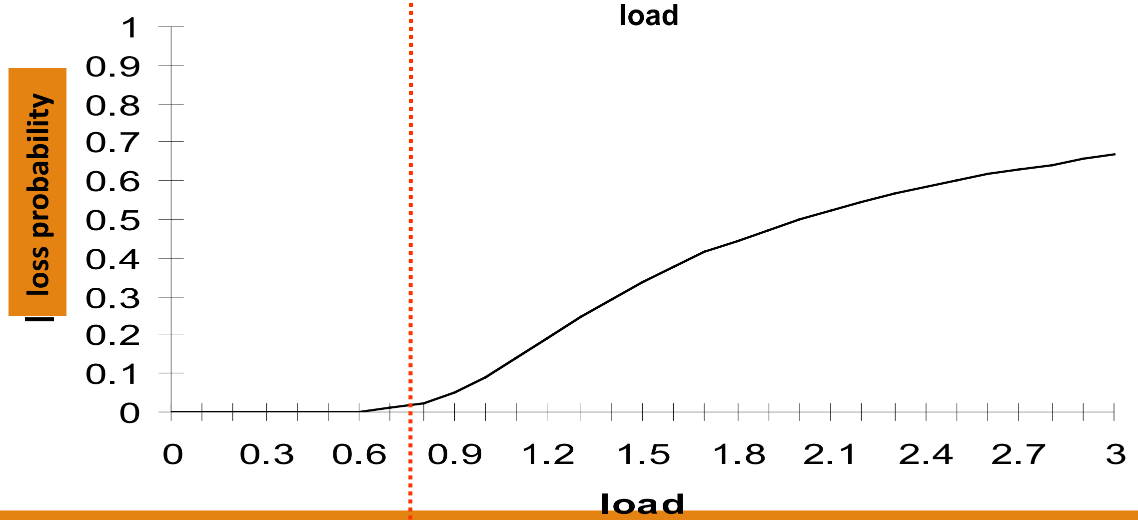
$$E[N] = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

Average packet delay

$$E[T] = \frac{E[N]}{\lambda(1-P_K)}$$

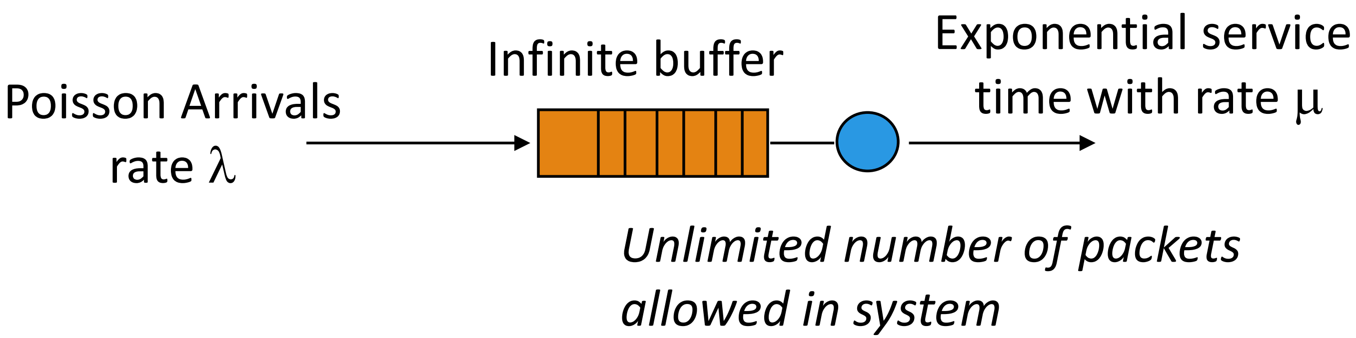
M/M/1/10 Queue Performance





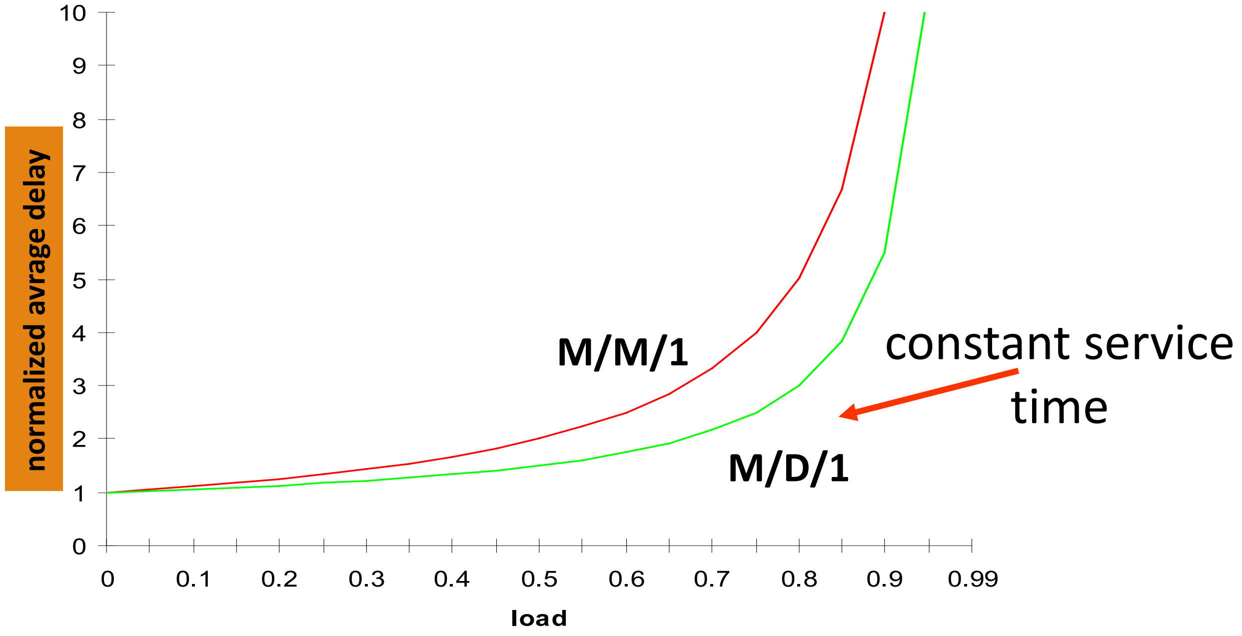
- Maximum 10 packets allowed in system
- Minimum delay is 1 service time
- Maximum delay is 10 service times
- At 70% load delay & loss begin increasing
- What if we add more buffers?

M/M/1 Queue Model



- P_b=0 Since packets are never blocked
- Average time in system E[T] = E[W] + E[X]
- When $\lambda << \mu$, calls/packets arrive infrequently and delays are low
- As λ approaches μ; packets start bunching up and average delays increase
- When $\lambda > \mu$; packets arrive faster than they can be processed and queue grows without bound (unstable)

Avg. Delay in M/M/1 & M/D/1 Systems



$$E[T_M] = \frac{1}{\lambda}\left[\frac{\rho}{1-\rho}\right] = \left[\frac{1}{1-\rho}\right]\frac{1}{\mu} = \left[\frac{\rho}{1-\rho}\right]\frac{1}{\mu} + \frac{1}{\mu} \quad \text{for M/M/1 model.}$$

$$E[T_D] = \left[1 + \frac{\rho}{2(1-\rho)}\right]\frac{1}{\mu} = \left[\frac{\rho}{2(1-\rho)}\right]\frac{1}{\mu} + \frac{1}{\mu} \quad \text{for M/D/1 system.}$$