ECE 158A: Lecture 4

Fall 2015

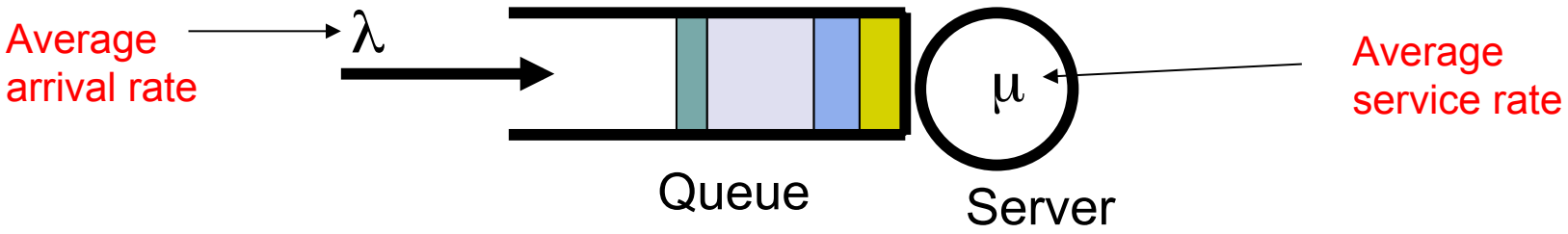
Queues Introduce Queuing Delays

- In Lecture 2, we saw that with a single link
  - packet delay = tx delay + propagation delay
- In a packet switch network
  - packet delay = tx delay + propagation delay + queuing delay
- In practice, the processing delay at each node has to be taken into account
- Queuing delay caused by bursts of packets reaching a server at the same time
- If a queue saturates, then packets are dropped and the packet delay becomes infinite

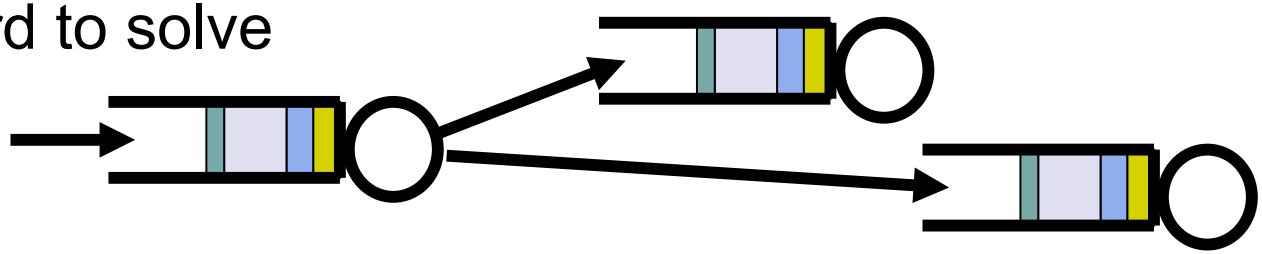
Queuing Theory

Queuing Theory

- Mathematical theory that studies queues
- Branch of operations research, applications include traffic control, networking, supply chains
- A single queue system is described by
  - Arrival process: Stochastic process describing packet arrivals
  - Service process: Stochastic process describing packet departures from queue (server processing time)

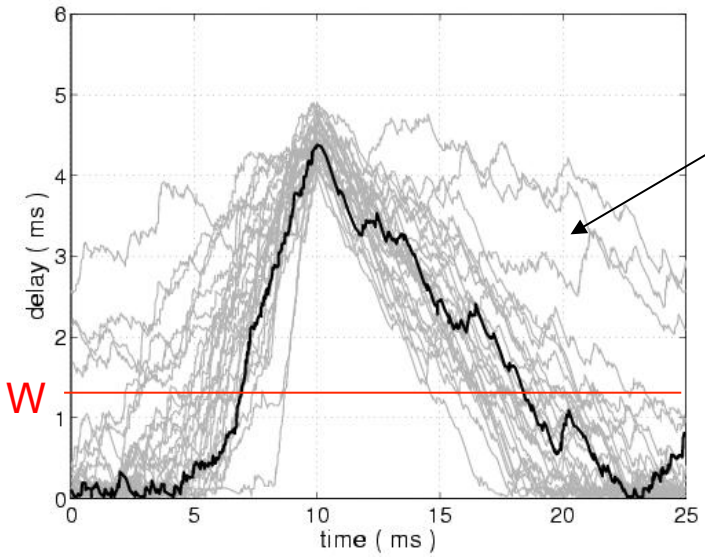


- Problems involving networks of queues are usually very hard to solve



Queuing Theory

- Measures of interest:
  - Stability (queue length remains finite)
  - Average waiting time in queue, denoted by W as in “wait”
  - Average number of packets in queue, denoted by L as in “length”



Typical queue dynamics, different realizations of the queue

Example M/M/1

- Arrival process is memoryless (M): No matter what, each second a new packet arrives with probability λ
- Service process is memoryless (M): If the queue in not empty, each second the server completes serving a packet with probability μ
- The server can process only 1 packet at the time (1)

- The service time of a packet varies:
  - 1 sec with probability μ
  - 2 secs with probability $(1-\mu)\mu$
  - 3 secs with probability $(1-\mu)^2\mu$
  - Etc
- What distribution is this? What is the average?

M/M/1: Basic Results

- The system is stable if $\lambda < \mu$. One can define the utilization as $\rho=\lambda/\mu$

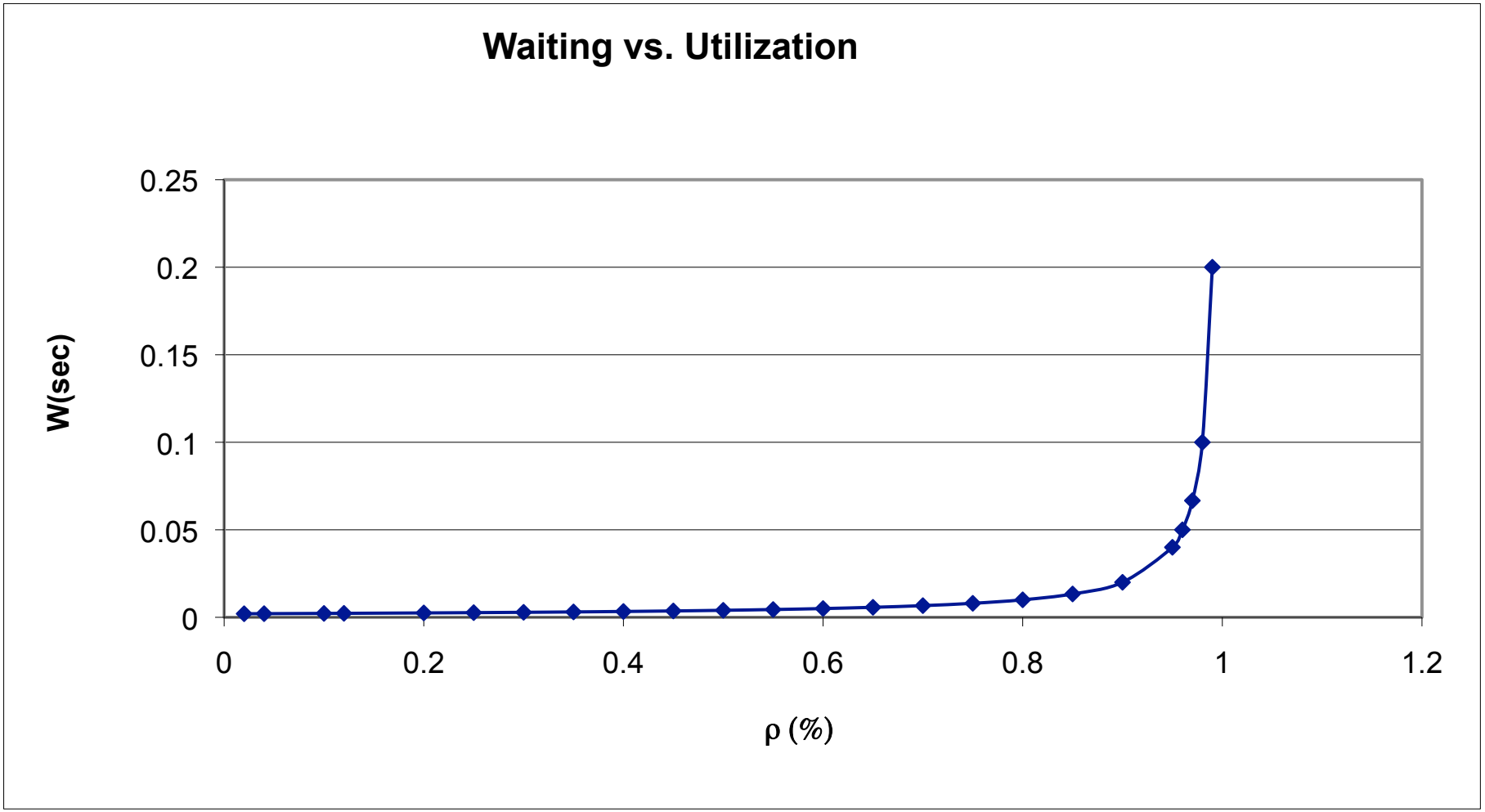
- The average time spend in queue (waiting in line + being served) is given by

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

- The average number of packets *L* waiting in queue (including the one being served) is given by

$$L = \frac{\lambda}{\mu - \lambda}$$

M/M/1: Basic Results



Little's Result (1961)

- General result in queuing theory (not limited to M/M/1 model)

$$L = \lambda W$$

- Example: Say that 1billion users send packets in the Internet at an average rate of 10MB per day. Say also that each packet spends 10ms in the Internet, on average. Then, the average number of bits in the Internet is

$$L = \left(10^9[\text{user}] \times \frac{8 \times 10^7[\text{bits/ user/ day}]}{60 \times 60 \times 24[\text{sec/ day}]}\right)\left(10^{-2}[\text{sec}]\right) = 9.25 \times 10^9[\text{bits}]$$

Problem on MM1 queue