Networking

Prof. Dr.-Ing. Holger Hermanns

Dependable Systems & Software
Saarland University

Tutorials

Summer 04 Tue, 10:30

Session D: Queueing Basics

Queueing Basics

Study chapter 3 of [Bertsekas/Galagher]

Our goal today:

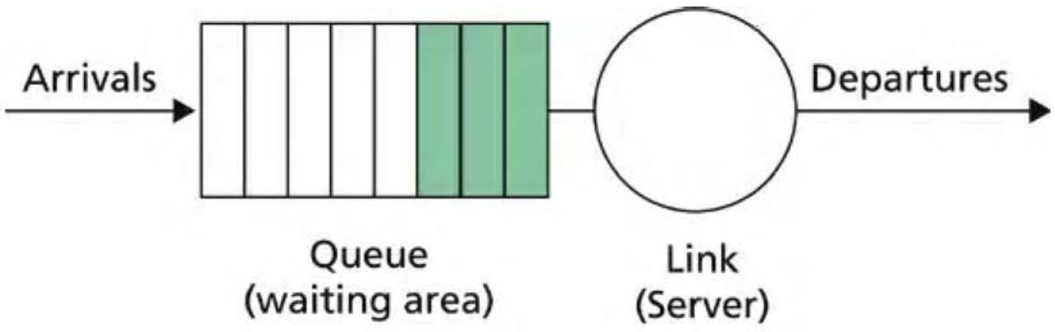
- develop an understanding of the fundamentals of capacity and delay behaviour of the Internet core

- approach:
  - straight into the heart of the matters

Overview:

- Basic Queueing model
- Little's law
- Markov everywhere: M/M/...
- From single queues to networks of queues
- Beyond Markov properties

Context

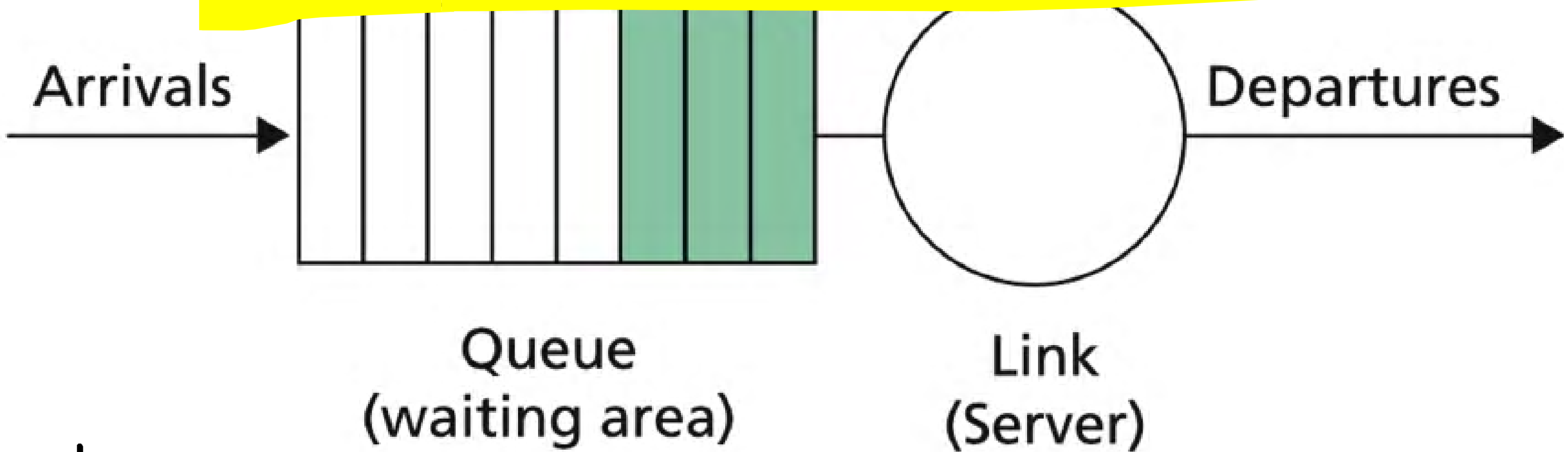


- Why study queues?
  - Framework for analyzing network queueing delay.
- Typical measures of interest:
  - The average number of customers in the system.
  - The average delay per customer.
- What do we need as input?
  - The customer arrival rate λ.
  - The customer service rate μ.

Think of:

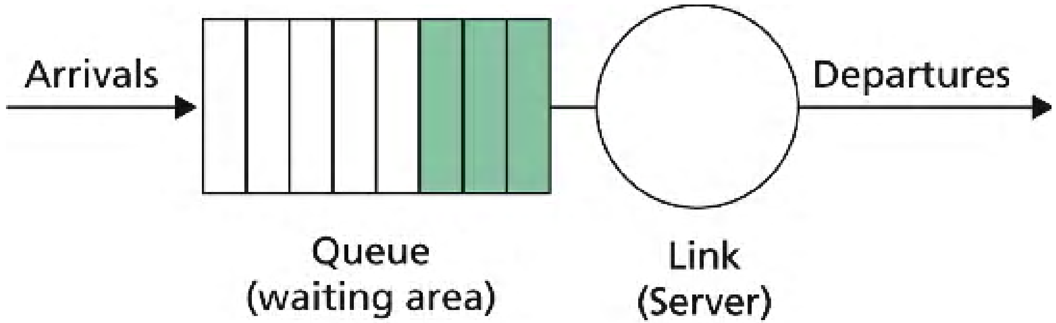
- Server: Router
- Customer: Packet
- λ arriving packets per second
- L bits per packet
- C bits per second (link bandwith)
- μ=L/C

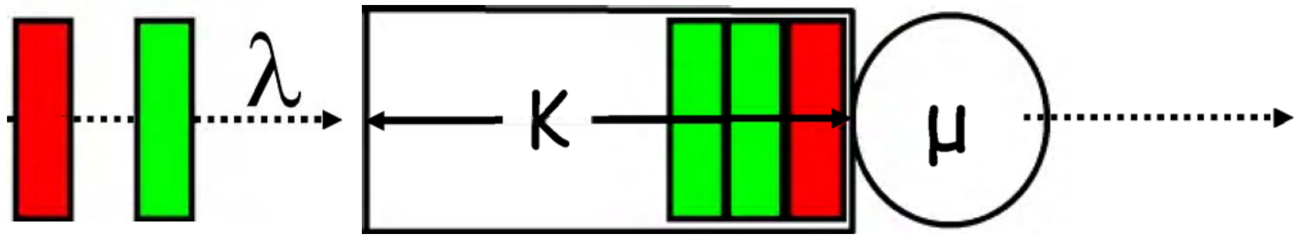
Queues in the Internet



- Router
  - e.g. buffering of incoming packets (waiting to be routed)
  - e.g. buffering of outgoing packets (waiting to get on the link)
- Client
  - e.g. buffering of streaming media
- Server
  - e.g. buffering requests to be processed

The generic model of a queue

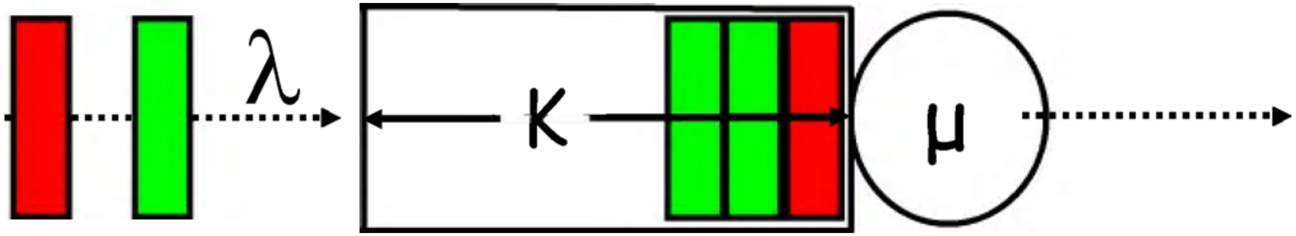




- Buffer of size K
  (# of customers in system)
- Customers arrive at rate λ
- Customers are served at rate μ

- λ and μ are average rates
  but we don't bother much about this for now

Queues: General Observations

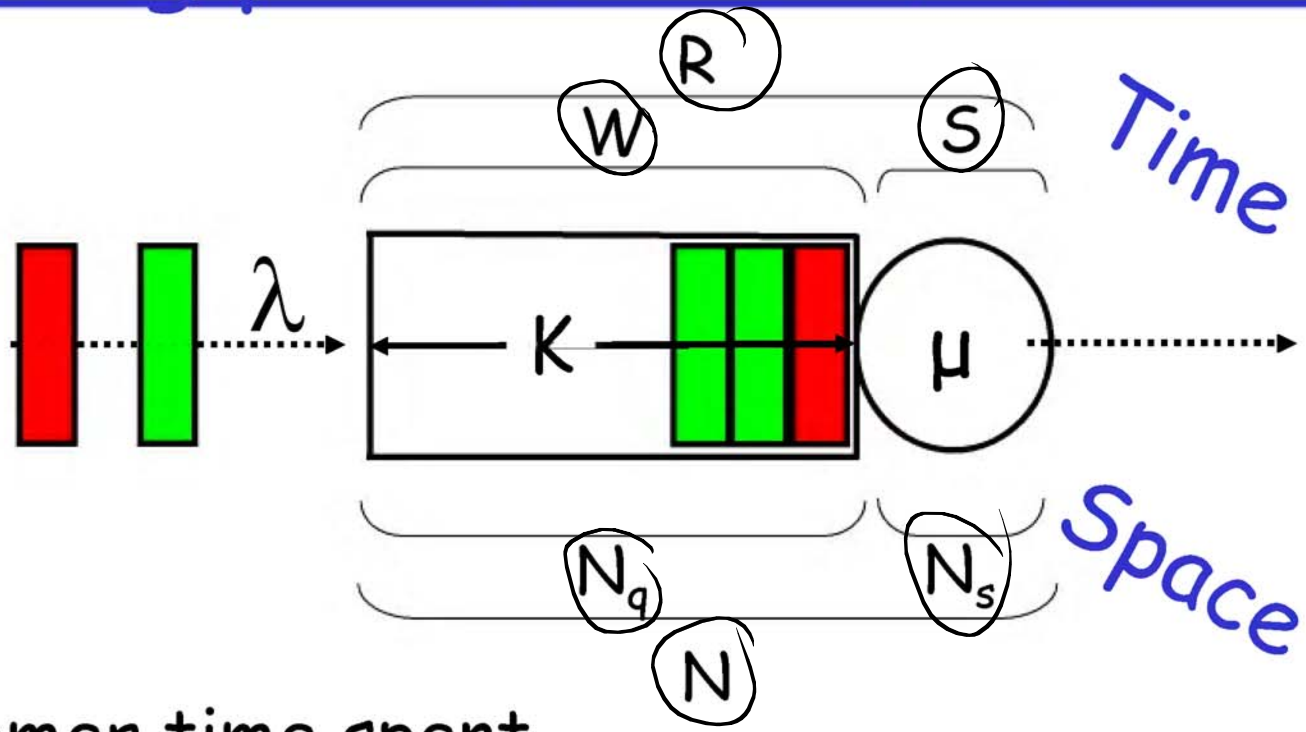


- **Increase in λ:**
  - more customers in queue (on average),
  - longer delays to get through queue;

- **Decrease in μ:**
  - longer delays to get processed,
  - leads to more customers in queue;

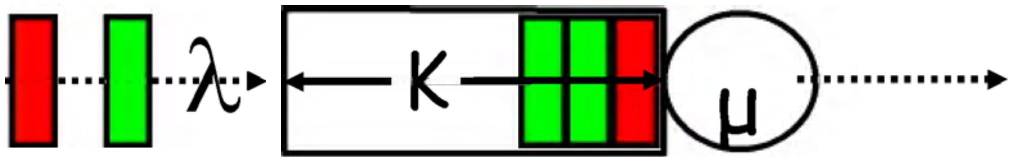
- **Decrease in K:**
  - customer drops more likely,
  - less delay for the "average" customer accepted into the queue.

Queueing parameters of interest



- Customer time spent
  - in queue: W (waiting time)
  - in service: S
  - in the complete system: R (response time)
- Number of customers
  - in queue: N_q
  - in service: N_s ('utilisation')
  - in the complete system: N

Little's Law



- Let c_i be the i^{th} customer arriving to the queue

- Let N_i be the number customers already in the queue when c_i arrives

- Let R_i be time spent by c_i in the system
  - both in the queue and while being served

- If $K = \infty$ (unlimited queue size) then

expectation (or 'average' if you like)

$$E[N] = \lim_{i\to\infty} E[N_i] = \lambda \lim_{i\to\infty} E[R_i] = \lambda E[R]$$

Little's Law in various flavours

On the long run ($t \rightarrow \infty$),

- System
  - E[N] number of customers in system, on average.
  - E[R] response time, on average.

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

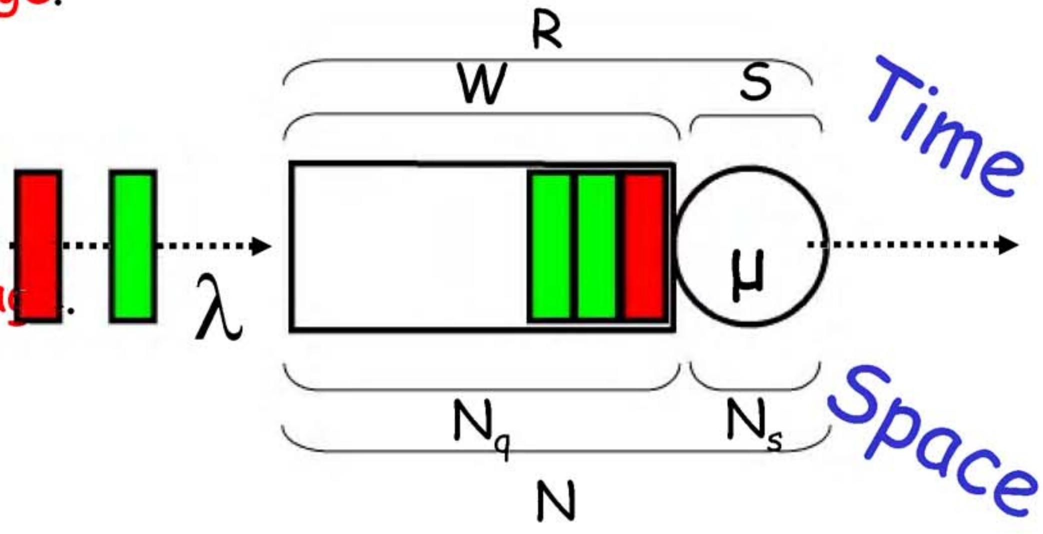
  3/sec

- Queue
  - $E[N_q]$ queue length, on average.
  - E[W] waiting time, on average.

  45 = $E[N_q] = \lambda \underbrace{E[W]}_{15\,sec}$

- Server
  - $E[N_s]$ utilisation on average.
  - E[S] servicing time on average.

  $E[N_s] = \lambda E[S]$



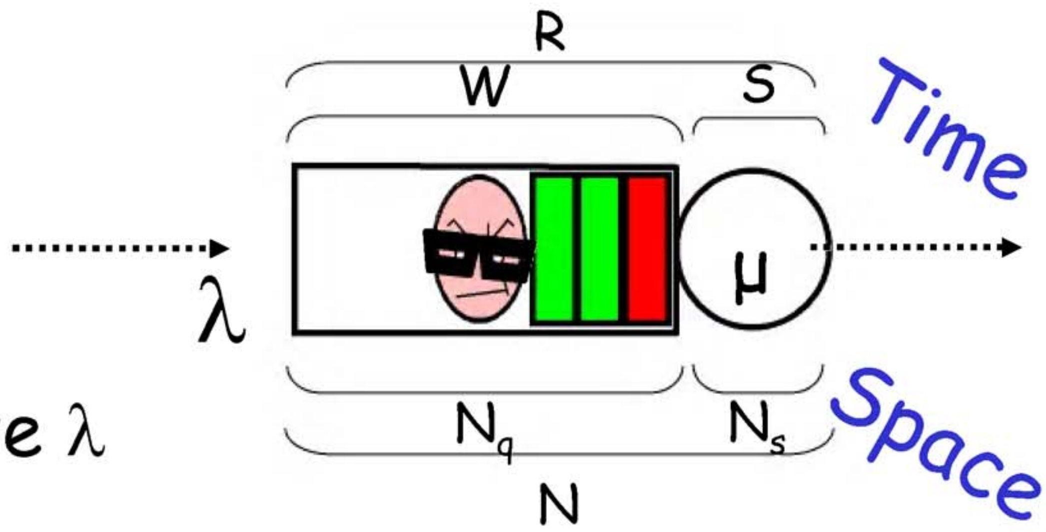
Little's Law

How to understand the law:

- An average customer  sees an average queue length $E[N_q]$.

- It takes him $E[W]$ on average to travel through the queue.

- When  leaves the queue, on average the queue still has length $E[N_q]$, being refilled with rate λ during $E[W]$.



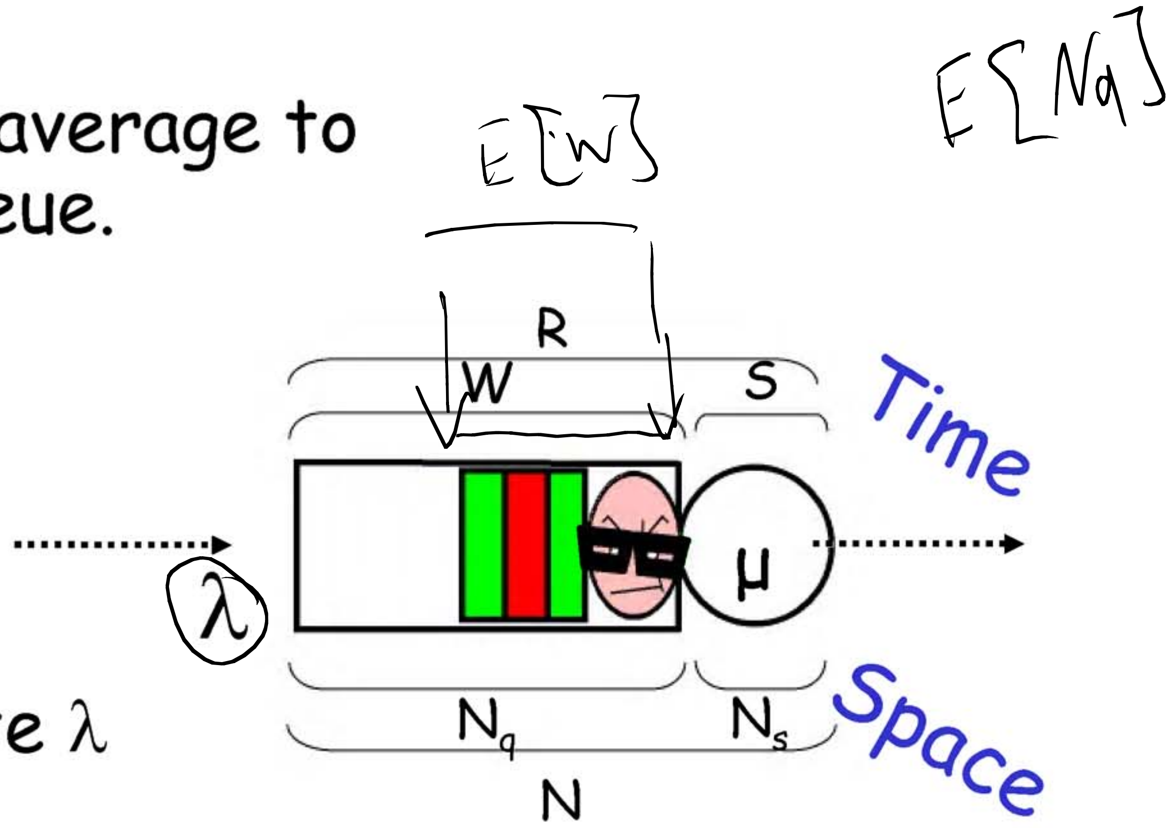
Little's Law

How to understand the law:

- An average customer  sees an average queue length $E[N_q]$.

- It takes him $E[W]$ on average to travel through the queue.

- When  leaves the queue, on average the queue still has length $E[N_q]$, being refilled with rate λ during $E[W]$.



Little's law

- Rel n some entity to
  - time
  - average ds in this entity

- No assumptions about

Very very general,
holds for (almost) all networks of queues etc etc.

- Only based on expected values,
  only provides long-run expectations

- Independent of
  - number of servers (could have more servers per queue),
  - used scheduling discipline (could serve in reverse order),
  - queue length etc etc.

An application of Little's law

- Consider a window flow control system for packet transmission, with window size W.

- There cannot be more than W packets in the session at any time, i.e. $W \geq E[N]$.

- According to Little

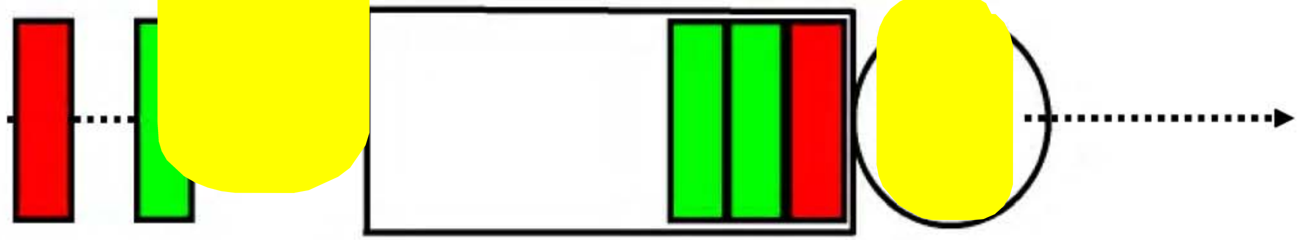
$$W \geq \lambda E[R]$$

where R is the response time of the system.

- Now, if congestion builds up and R increases, λ must eventually decrease.

Q: How?

The M/M/1 Queue



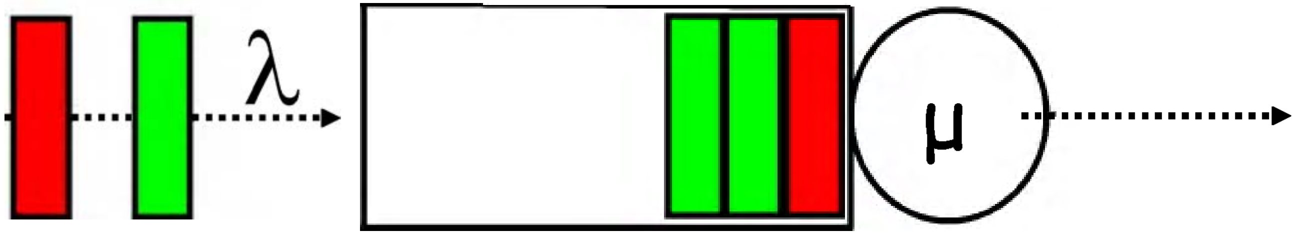
- So far we have been quite relaxed about the arrival time distributions and the service time distributions.

- Now let's study a specific case [illegible]

the so-called **M/M/1** queue.

This type of queue has been a classical model for telecommunication network dimensioning, and continues to serve as a model for Internet traffic -- though being somewhat less appropriate. (More on this later).

The M/M/1 Queue



- The M/M/1 queue is obtained by fixing
  - M: Markov (i.e. memoryless) arrival rate λ
  - M: Markov (i.e. memoryless) service rate μ
  - 1: a single service station.

both λ and μ are parameters of exponential distributions.

- For given λ and μ we know E[N]=λ E[R] etc.

- How do we determine E[N] or E[R]?

On the appropriateness of the exponential distribution

- Statistical independent behaviours naturally lead to the exponential distribution.
- Examples
  - [illegible]
  - cars passing bridges (except rush hour etc)
  - length of files on an arbitrary computers (prior to mp3)

  ...

usually referred to as 'Poisson' arrivals

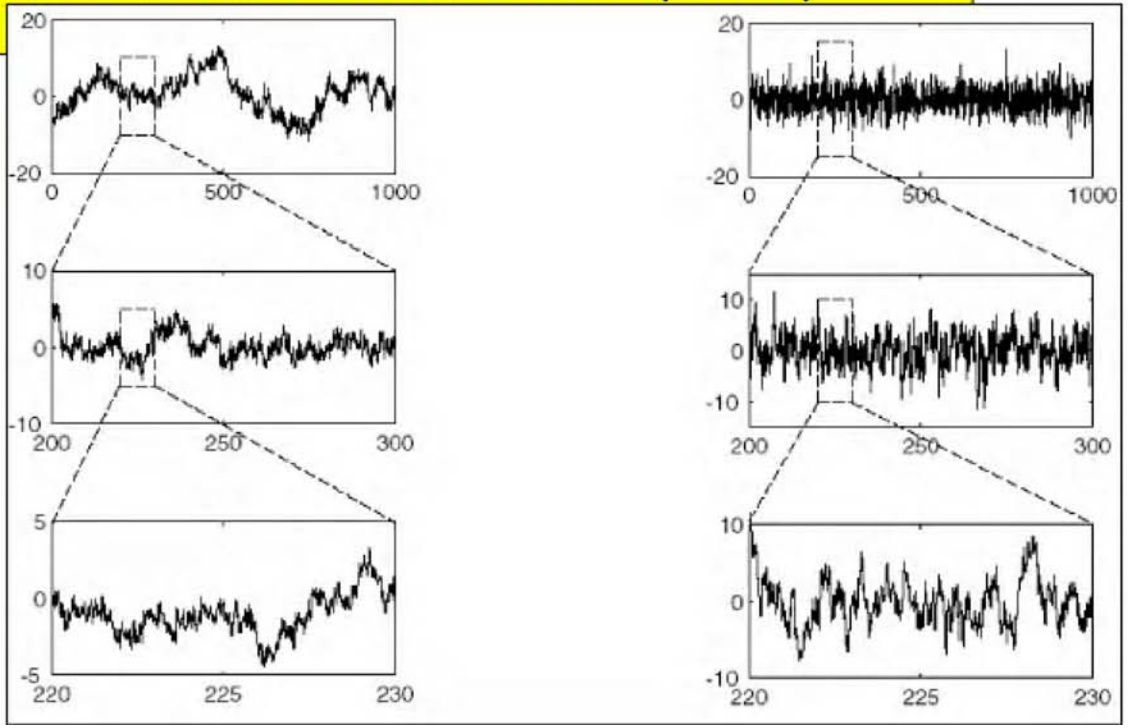
for reasons that you are invited to find out yourself

On the appropriateness of the exponential distribution

- Lucent Press Release, June 6, 2001
  http://www.lucent.com/press/0601/010606.bla.htlm
  Cao, Cleveland, Lin, and Sun, "*Internet traffic tends toward Poisson and Independent as the Load Increases*" (2002)

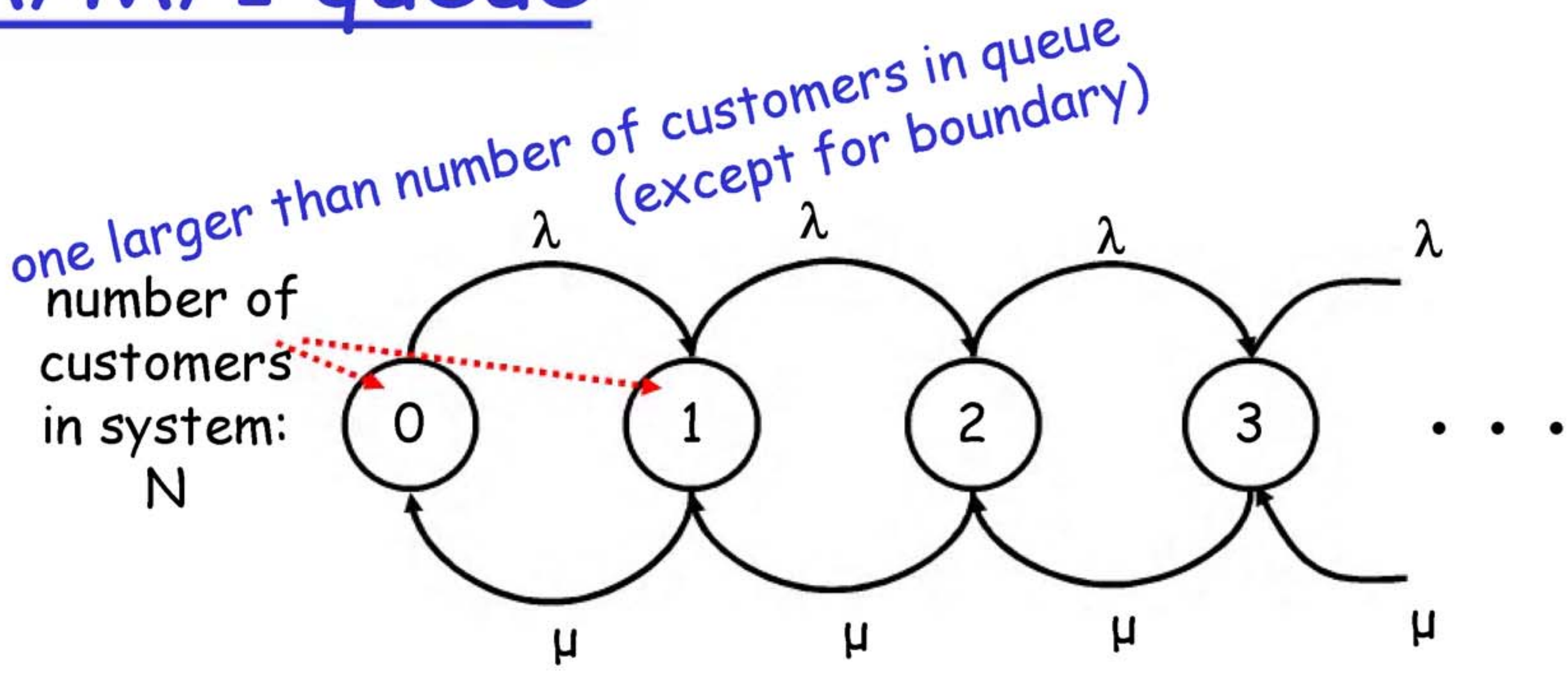
'Packet interarrival times (of aggregate traffic) become exponentially distributed and independent as the link load increases'

Not so sure about this!



heavy-tailed (self-similar) exponential

A state diagram for the M/M/1 queue



- Is a 'Continuous Time Markov Chain'
- $E[N] \equiv \sum_{n=0}^{\infty} n\, P(N=n)$

$$0 = -\lambda \pi_0 + \mu \pi_1 \Rightarrow \lambda \pi_0 = \mu \pi_1$$

$$(\lambda+\mu)\pi_1 = \lambda \pi_0 + \mu \pi_2$$

$$\pi_0 = \frac{\mu}{\lambda}\pi_1$$

$$(\lambda+\mu)\pi_{n+1} = \lambda \pi_n + \mu \pi_{n+2}$$

$$\pi_1 = \frac{\lambda}{\mu}\pi_0$$

$$(\lambda+\mu)\frac{\lambda}{\mu}\pi_0 = \lambda \pi_0 + \mu \pi_2$$

$$\left(\frac{(\lambda+\mu)\lambda}{\mu} - \frac{\lambda\mu}{\mu}\right)\pi_0 = \mu \pi_2$$

$$\frac{\lambda^2}{\mu}\pi_0 = \mu \pi_2$$

$$\pi_2 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$$

$$\pi_{n+1} = \left(\frac{\lambda}{\mu}\right)\pi_n = \left(\frac{\lambda}{\mu}\right)^{n+1} \cdot \pi_0$$

M/M/1 queue analysis

$$\rho = \frac{\lambda}{\mu}$$

$$\sum_{0}^{\infty} \pi_n = 1$$

Let's assume an infinite buffer capacity (we are not the only one to do so).

We are dealing with the following matrix

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -\lambda-\mu & \lambda & 0 & \cdots \\ 0 & \mu & -\lambda-\mu & \lambda & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

for which we need to solve

$$\vec{\pi} Q = 0$$

to obtain the steady-state state probabilities $P(N=n)=\pi_n$.

If $\lambda < \mu$ we get: $\pi_n = \rho^n (1-\rho)$

M/M/1 queue analysis

instable if $\lambda > \mu$

As long as $\lambda < \mu$, queue has the following (long-run) limit properties

- $P(N=n) = \rho^n(1-\rho)$
  - (indicates fraction of time spent with n customers in queue)
  - Utilization = 1 – P(N=0) = ρ

- Notation
  - $\rho = \lambda/\mu$
    - ratio of arriving/departing traffic ('traffic intensity')
  - N = # customers in system
  - R = customer time in system

- $E[N] \equiv \sum_{n=0}^{\infty} n\, P(N=n) = \rho/(1-\rho)$

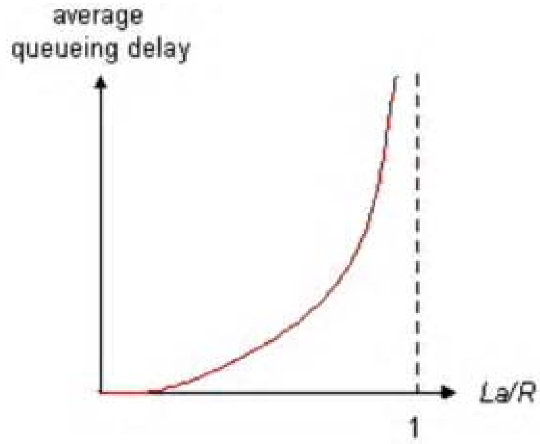
- $E[R] = E[N] / \lambda$ (Little)
  $= \rho/(\lambda\,(1-\rho)) = 1 / (\mu - \lambda)$

M/M/1 queue analysis

Queueing delay (revisited)

- R=link bandwidth (bps)
- L=packet length (bits)
- a=average packet arrival rate

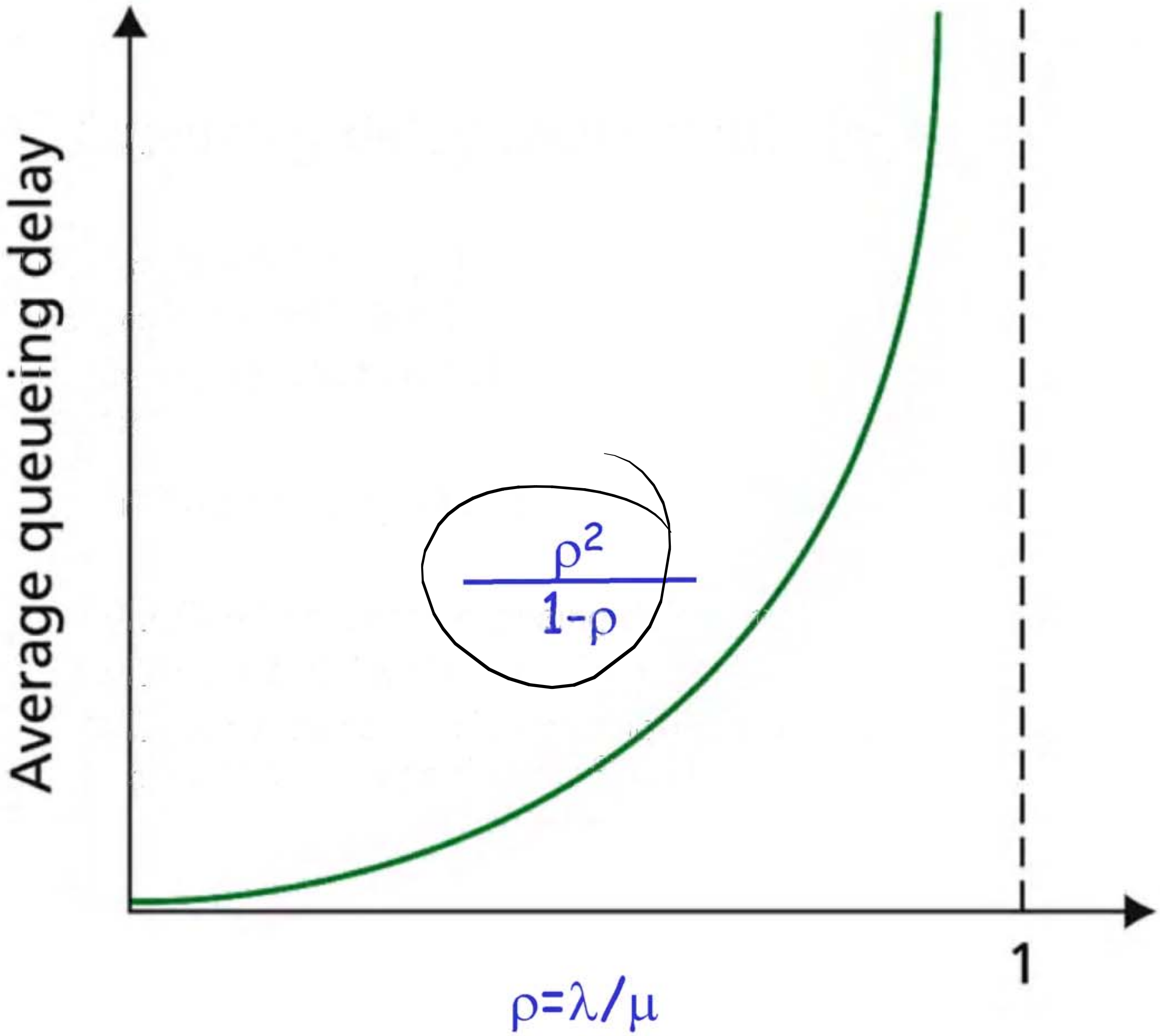
traffic intensity = La/R



- La/R ~ 0: average queueing delay small
- La/R -> 1: delays become large
- La/R > 1: more "work" arriving than can be serviced, average delay infinite!

M/M/1 queue analysis

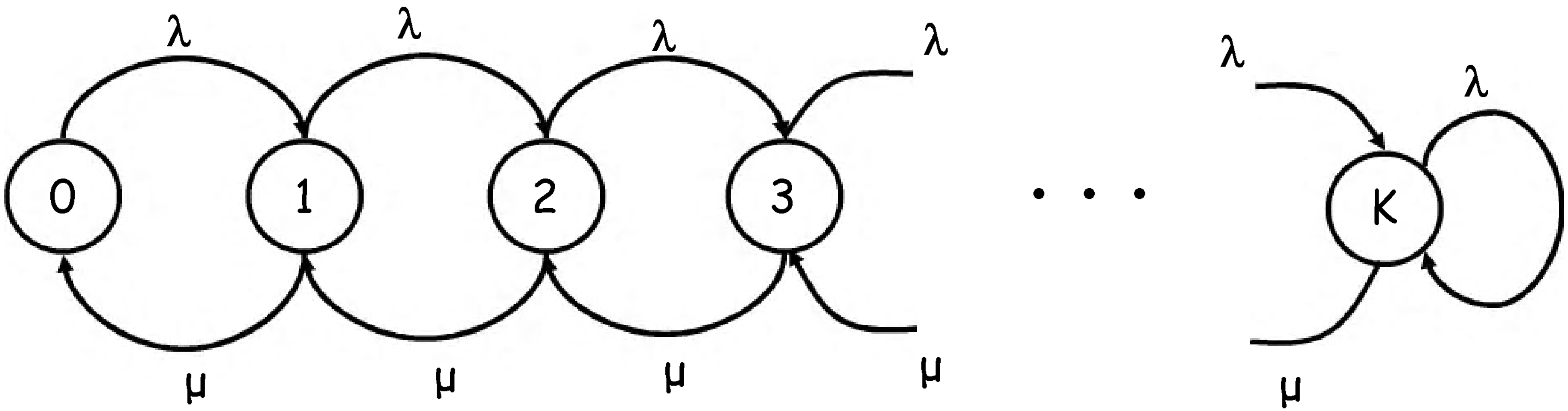
Q: Why this?



Bounded buffer M/M/1 queue

Note: $\rho \geq 1$ permitted

- Also can be modeled as a CTMC
  - requires K+1 states for a model (queue + server) that holds K packets
  - stay in state K upon a customer arrival



Bounded buffer queue properties

- $P(N=n) = \begin{cases} \rho^n(1-\rho) / (1 - \rho^{K+1}), & \rho \neq 1 \\ 1 / (K+1), & \rho = 1 \end{cases}$

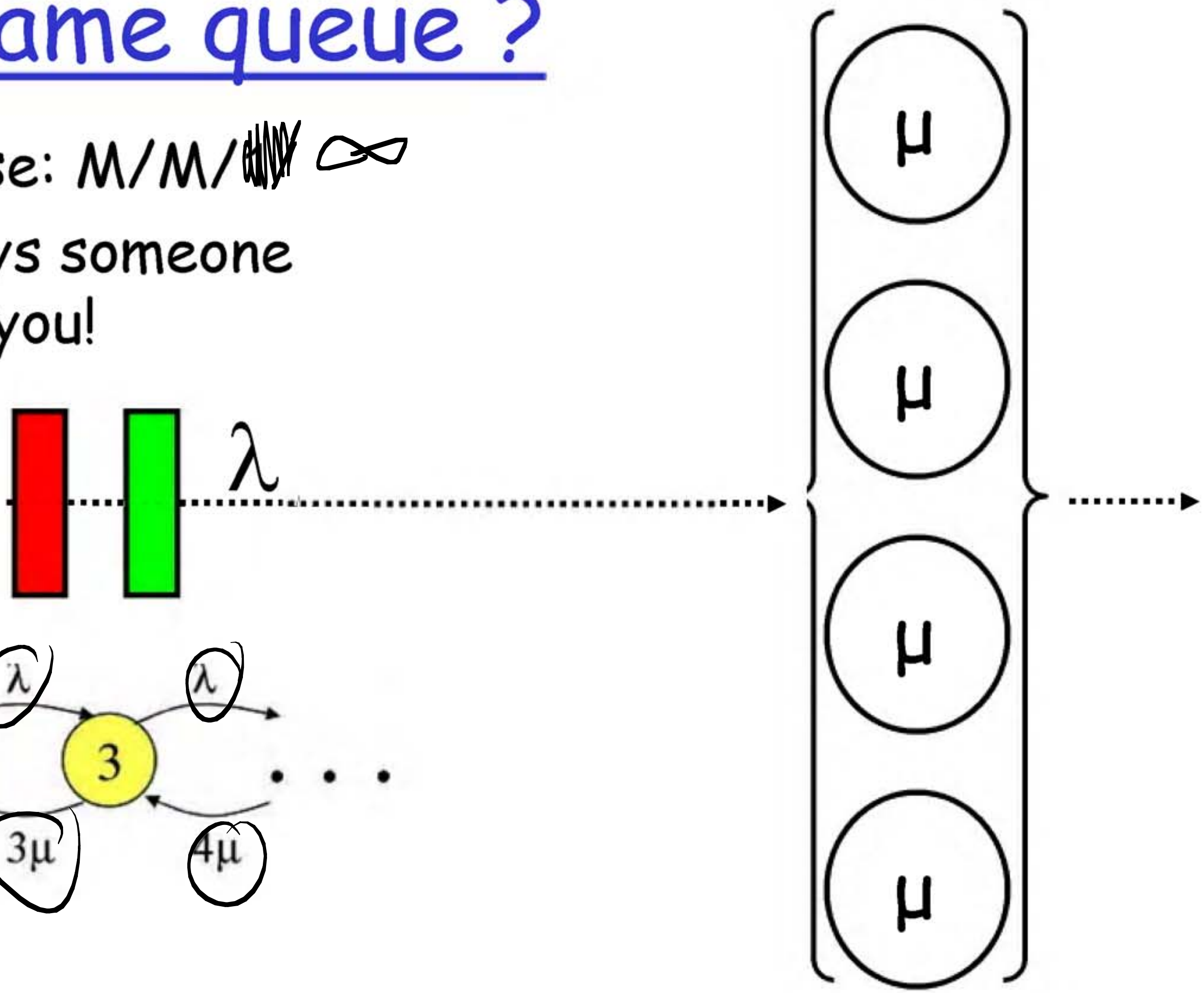
- $E[N] = \begin{cases} \rho/((1-\rho)(1 - \rho^{K+1})), & \rho \neq 1 \\ 1 / (K+1), & \rho = 1 \end{cases}$

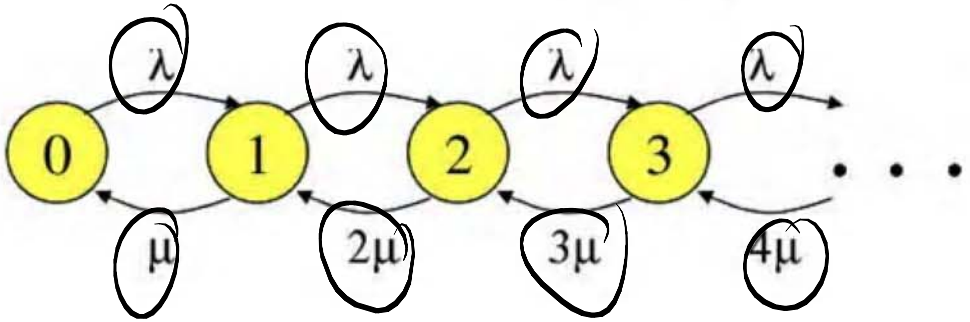
divide unbounded buffer values by $(1 - \rho^{K+1})$

- Utilisation = $1 - P(N=0) = \rho(1 - \rho^K)/(1 - \rho^{K+1})$

What if many servers serve the same queue ?

- The extreme case: M/M/∞
- No waiting, always someone around to serve you!



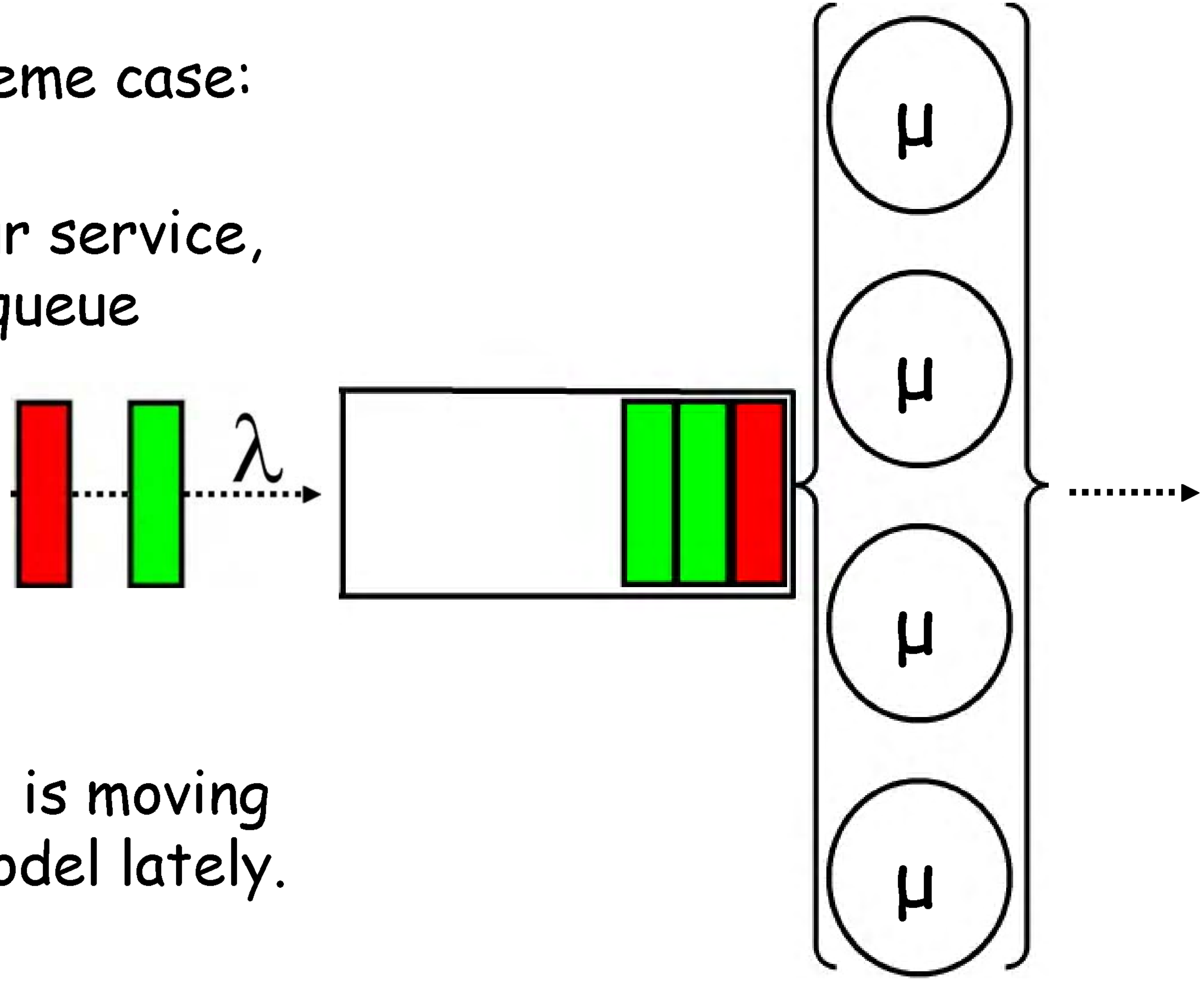


- Response time:

  $R=1/\mu$

What if many servers serve the same queue ?

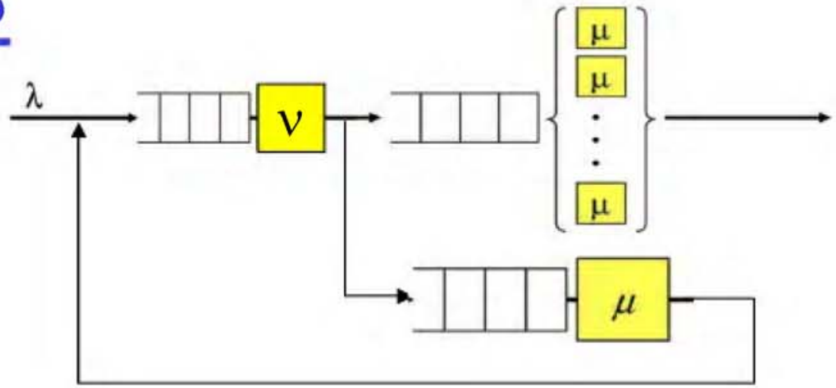
- The not-so extreme case:
- m servers at your service, serving a single queue



- *Deutsche Bahn* is moving towards this model lately.

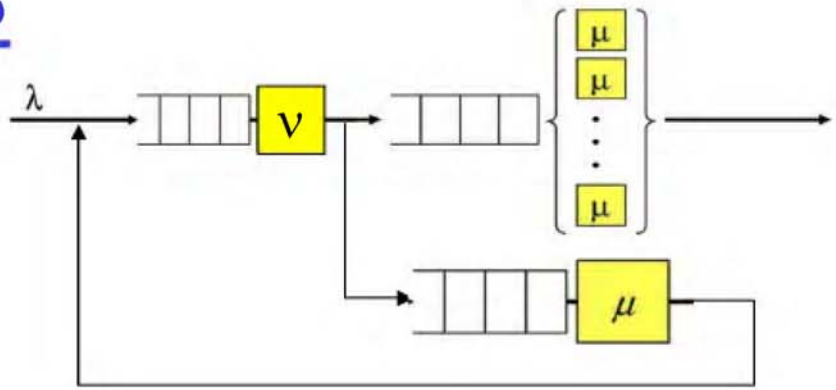
  Q: Why?

Networks of queues



- Queueing networks consist of
  - queues
  - routes of customers
- If all component queues are M/M/...,
  you now know how to analyse them!

Networks of queues



- Queueing networks consist of
  - queues
  - routes of customers
- If all component queues are M/M/...,
  you now know how to analyse them!
- A simple example:

  'Tandem' queue:
  - 1st: M/M/1 with queue size 3
  - 2nd: M/M/5 with queue size 3