Queueing Theory

IN4390 Quantitative Evaluation of Embedded Systems

Koen Langendoen

Queueing theory

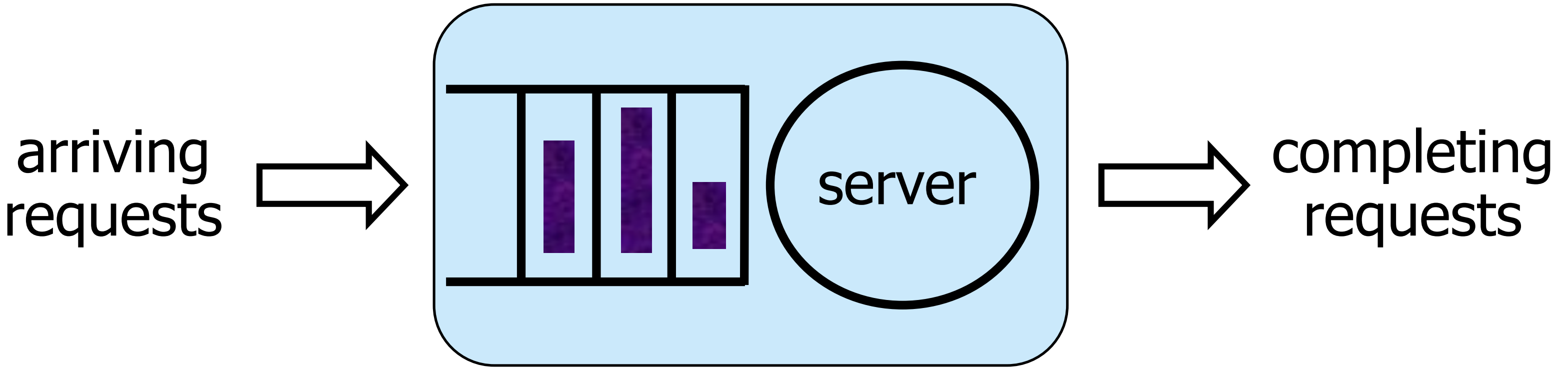
Yet another take at performance evaluation

- Measurements
  - DoE
  - **Operational Laws**
- Simulations
  - …
- Modeling
  - Petri nets
  - Markov modeling
  - **Queueing theory**

why bother?

A queueing system

Kendall notation (shortened)



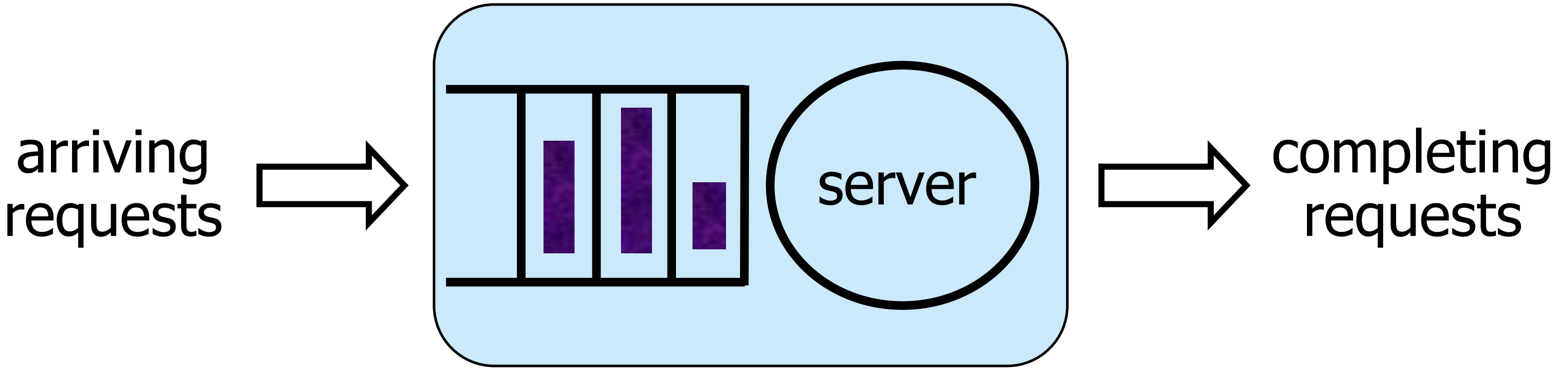
Characterized by **A/S/m**

- **A**: interarrival time distr.
- **S**: service time distr.
- **m**: #servers

The M/M/1 queue

The most popular model

Why oh why?



Characteristics

- **exponential** interarrival time
- **exponential** service time

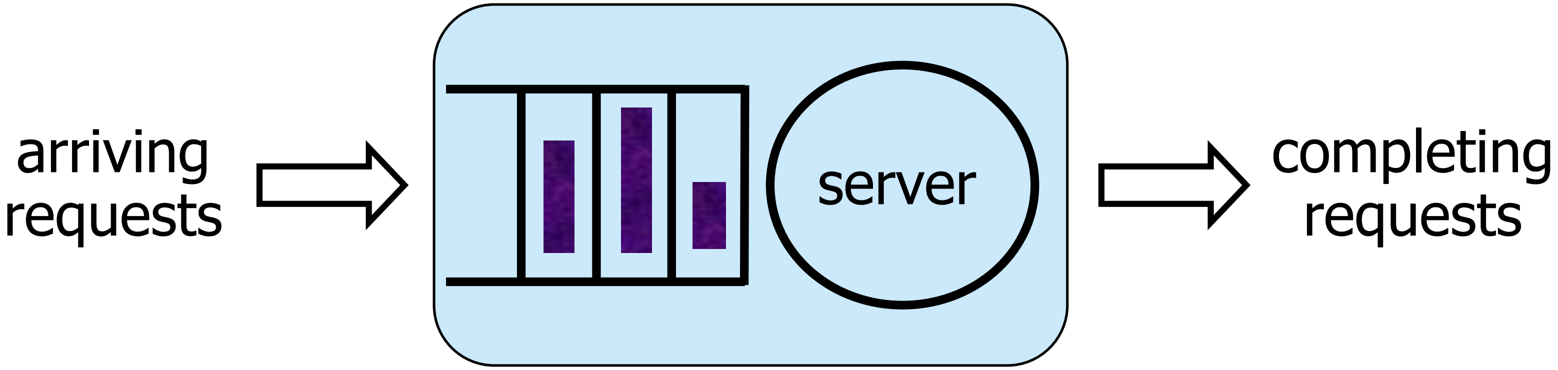
} realistic distr. with long tail

- **memoryless** is easy to analyze

The M/M/1 queue

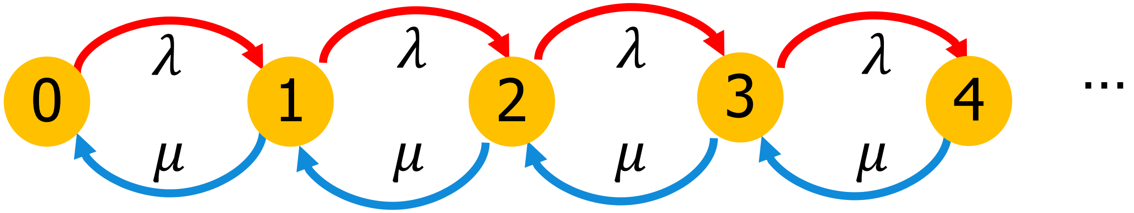
Connect to Markov models

Any suggestion?



- Sate encodes #requests in the system

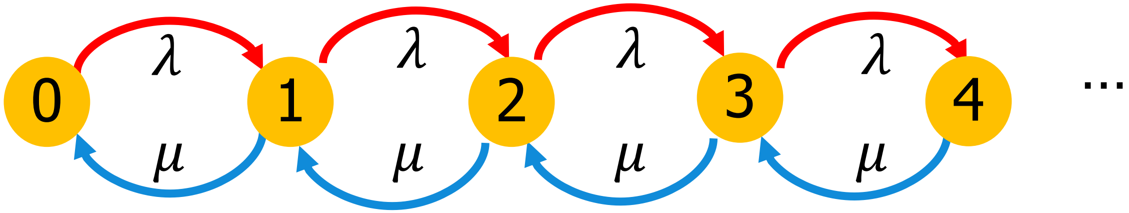
What about infinity?



- Analyze steady state

The M/M/1 queue

Mathematical analysis



Goal:

- A closed form expression of the probability of the number of jobs in the queue (P_i) given only λ and μ

To compute

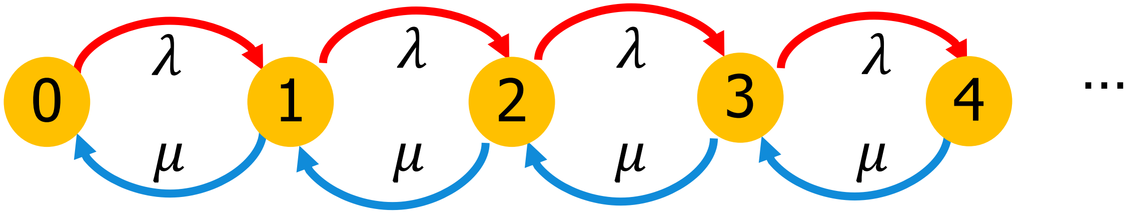
- #requests in the system (N)
- response time (R)

How?

How?

The M/M/1 queue

Mathematical analysis



Steady state ($\lambda < \mu$):

- Flows must be in equilibrium

What does that denote?

$\rho = \lambda/\mu$

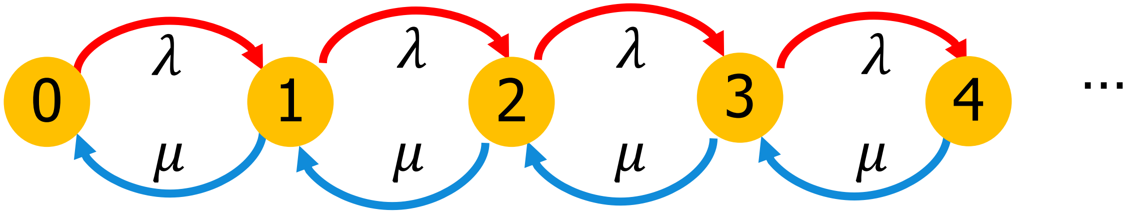
From left to right

$$\left.\begin{array}{l} \bullet\ \lambda P_0 = \mu P_1 \\ \bullet\ \lambda P_1 = \mu P_2 \\ \quad\vdots \\ \bullet\ \lambda P_{n-1} = \mu P_n \end{array}\right\} \quad \begin{array}{l} P_1 = \rho P_0 \\ P_2 = \rho^2 P_0 \\ \quad\vdots \\ P_n = \rho^n P_0 \end{array}$$

Also holds for n=0 ☺

The M/M/1 queue

Mathematical analysis



Steady state (ρ < 1):

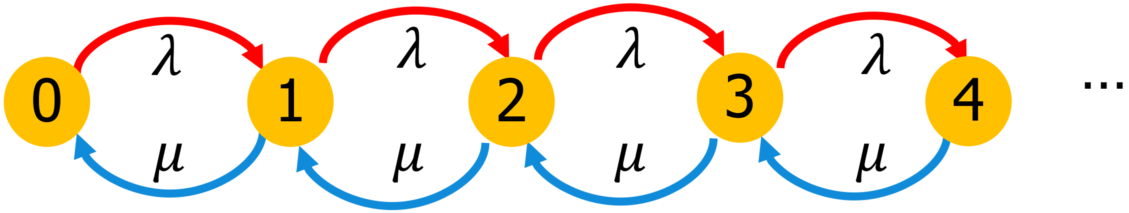
- Flows must be in equilibrium: $P_n = \rho^n P_0$
- Probabilities must sum to one:

Looks familiar?

$$\sum_{n=0}^{\infty} P_n = 1 \Rightarrow P_0 \boxed{\sum_{n=0}^{\infty} \rho^n} = 1 \Rightarrow P_0 \boxed{\frac{1}{1-\rho}} = 1$$

The M/M/1 queue

Mathematical analysis



Steady state (ρ < 1):

- Flows must be in equilibrium
- Probabilities must sum to one

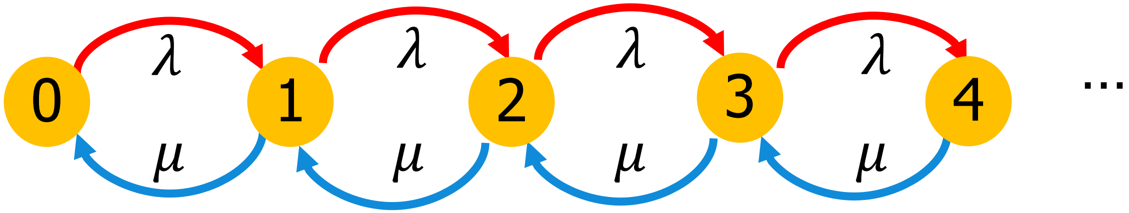
$$P_0 = 1 - \rho$$
$$P_n = \rho^n (1 - \rho)$$

Makes sense!?

The M/M/1 queue

Mathematical analysis

$$P_n = \rho^n (1 - \rho)$$



Goal:

- A closed form expression of the probability of the number of jobs in the queue (P_i) given only λ and μ ✓

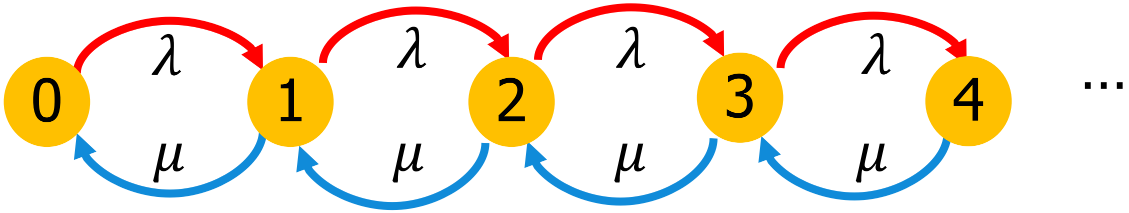
To compute

- #requests in the system: $N = \sum_{n=0}^{\infty} nP_n$
- response time: R = N / λ

The M/M/1 queue

Mathematical analysis

$$\boxed{P_n = \rho^n (1 - \rho)}$$



Compute #requests in the system:

$$N = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n (1-\rho) = \ldots\ldots\ldots\ldots\ldots\ldots$$

$$= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda}$$

Proof by intimidation ☺

take the derivative of the integral!

$$\sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1}$$

geometric series

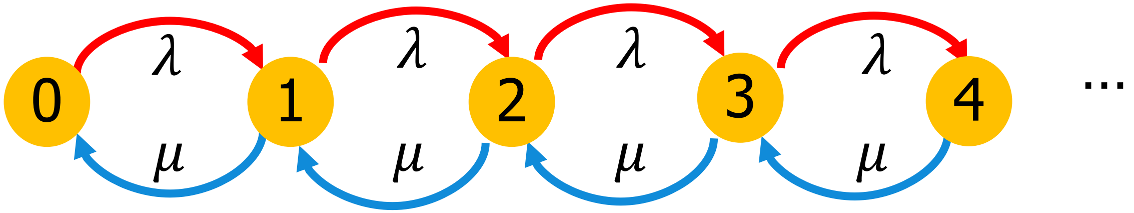
$$(1-\rho)\rho\boxed{\frac{d}{d\rho}\left(\boxed{\sum_{n=0}^{\infty}\rho^n}\right)} = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$(1-\rho)\rho\left(\frac{1}{(1-\rho)^2}\right) = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda}$$

The M/M/1 queue

Mathematical analysis

$$\boxed{P_n = \rho^n (1 - \rho)}$$



Goal:

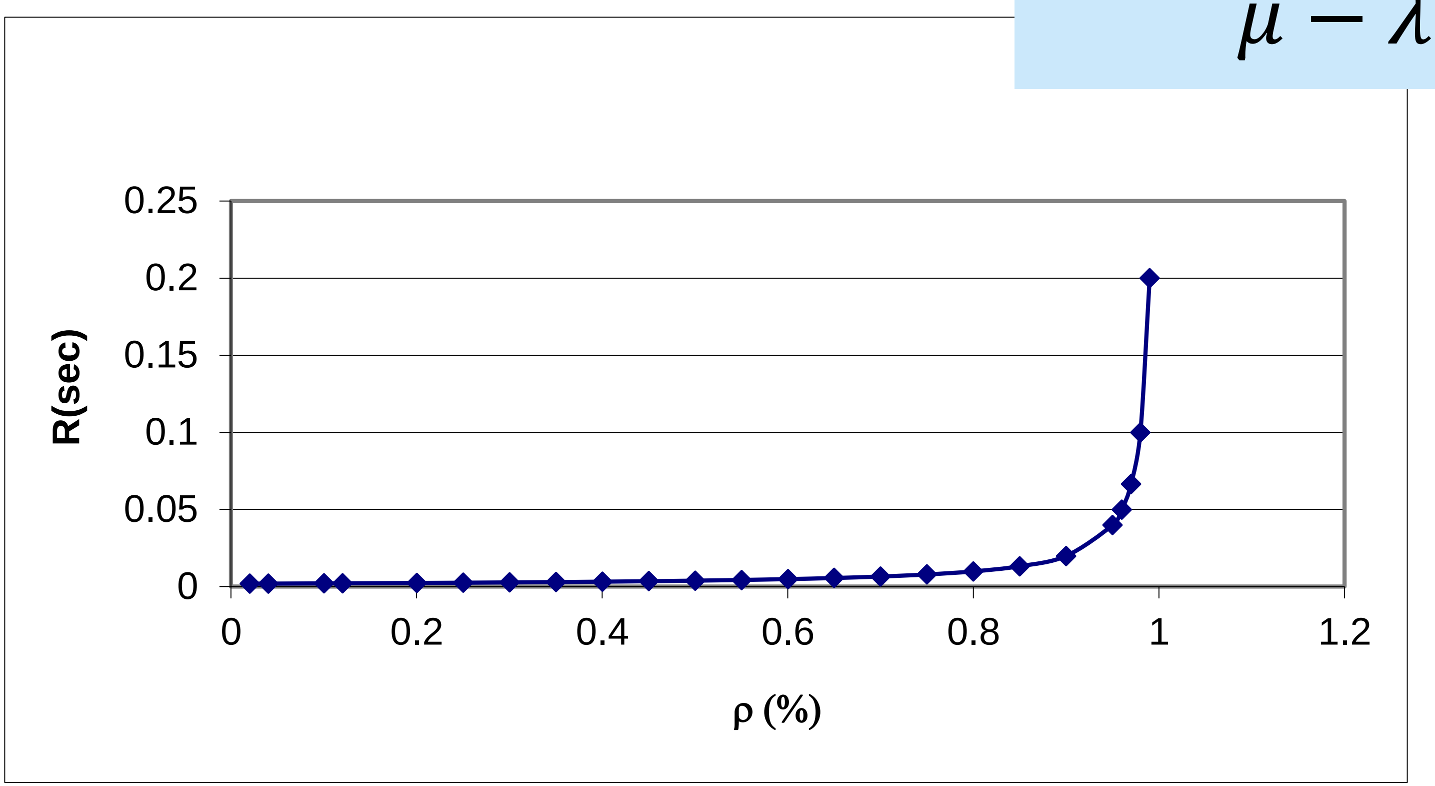
- A closed form expression of the probability of the number of jobs in the queue (P_i) given only λ and μ ✓

To compute

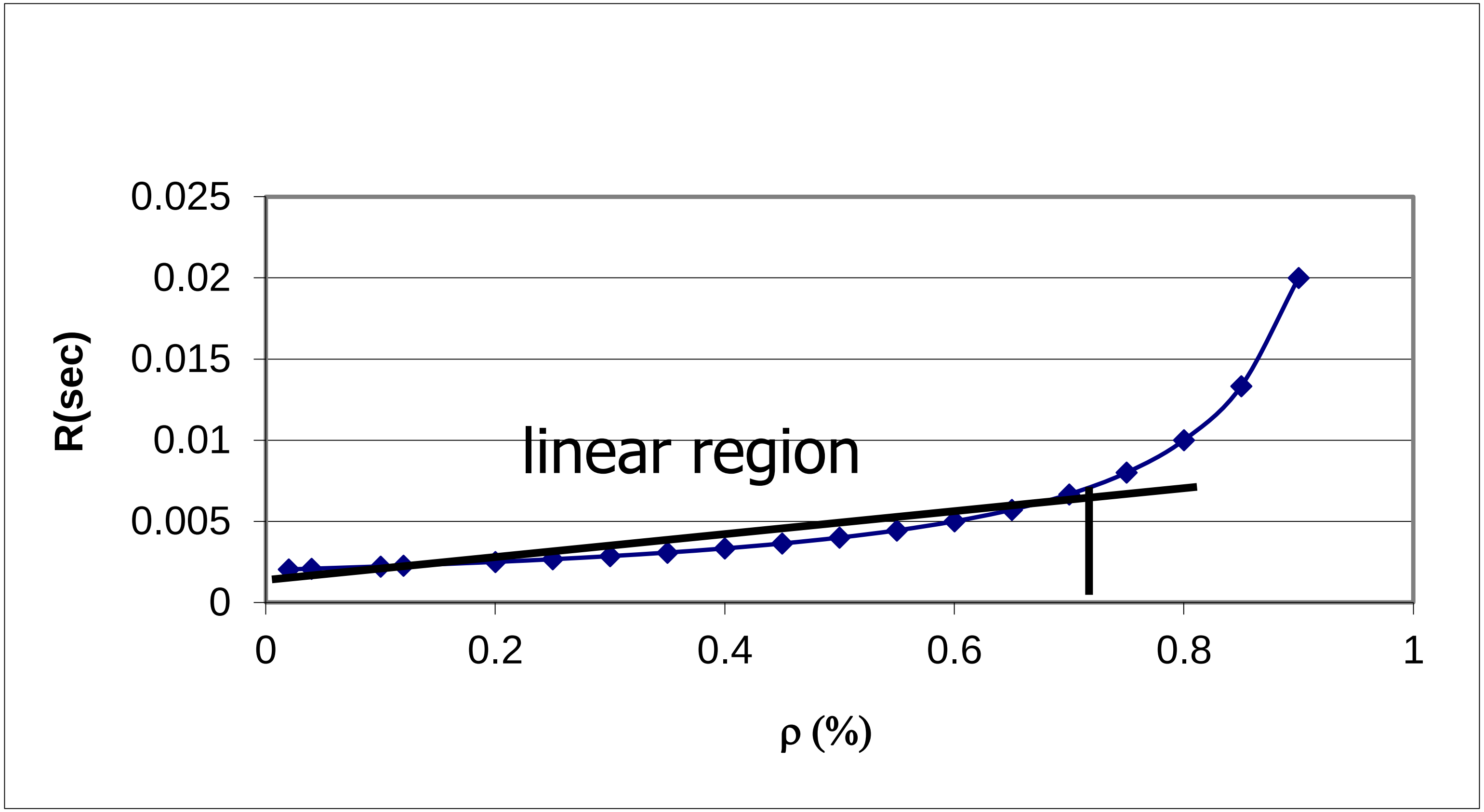
- #requests in the system: $N = \dfrac{\rho}{(1-\rho)} = \dfrac{\lambda}{\mu - \lambda}$ ✓
- response time: $R = 1/(\mu - \lambda)$

Response Time vs. Arrivals

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

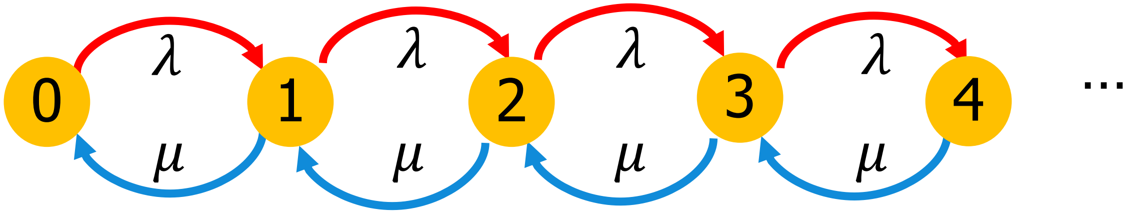


Stable Region



The M/M/1 queue

Main results

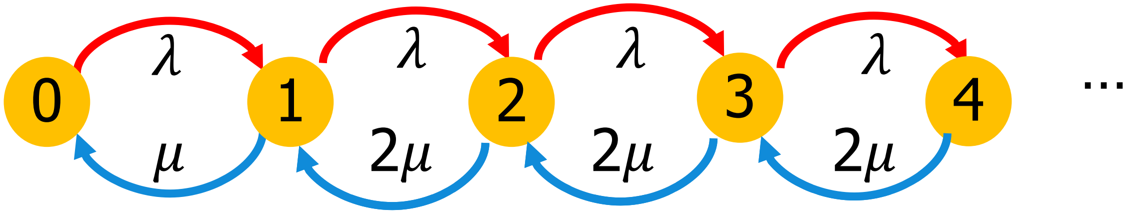


Utilization	$U = X\,S = \lambda/\mu = \rho$
Prob. of n clients in the system	$P_n = \rho^n\,(1 - \rho)$
Mean #clients in the system	$N = \rho\,/\,(1-\rho) = \lambda\,/\,(\mu-\lambda)$
Mean #clients in the queue	$N_Q = N - (1 - P_0) = N - \rho$
Mean response time	$R = N/\lambda = 1/(\mu-\lambda) = S/(1-\rho)$
Mean waiting time	$W = R - S = \rho/(\mu-\lambda)$

The M/M/**2** queue

Mathematical analysis

What needs to be changed?



Steady state (ρ < 1):

- Flows must be in equilibrium

From left to right

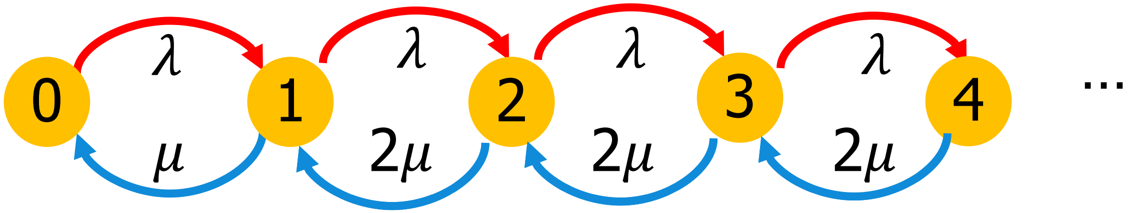
$\rho = \lambda / \mathbf{2}\mu$

$$\left.\begin{aligned} \lambda P_0 &= \mu P_1 \\ \lambda P_1 &= \mathbf{2}\mu P_2 \\ &\vdots \\ \lambda P_{n-1} &= \mathbf{2}\mu P_n \end{aligned}\right\} \quad \begin{aligned} P_1 &= \mathbf{2}\rho P_0 \\ P_2 &= \mathbf{2}\rho^2 P_0 \\ &\vdots \\ P_n &= \mathbf{2}\rho^n P_0 \end{aligned}$$

one too many for n=0 ☹

The M/M/2 queue

Mathematical analysis



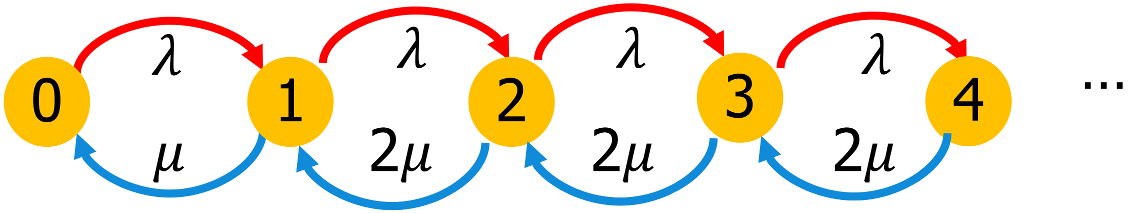
Steady state (ρ < 1):

- Flows must be in equilibrium: $P_n = 2\rho^n P_0$
- Probabilities must sum to one:

$$\sum_{n=0}^{\infty} P_n = 1 \Rightarrow 2P_0 \boxed{\sum_{n=0}^{\infty} \rho^n} - P_0 = 1 \Rightarrow \frac{1+\rho}{1-\rho} P_0 = 1$$

The M/M/2 queue

Mathematical analysis



Steady state (ρ < 1):

- Flows must be in equilibrium
- Probabilities must sum to one

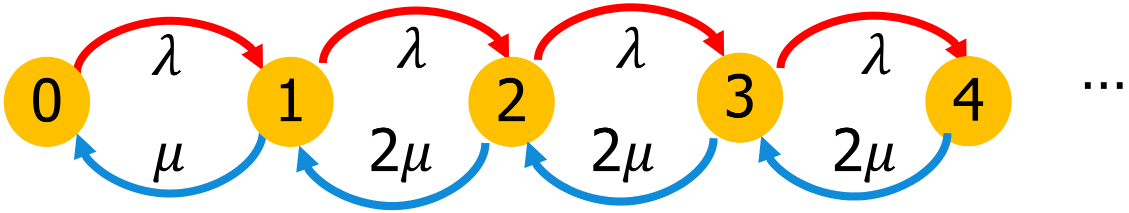
$$P_0 = \frac{1-\rho}{1+\rho}$$

$$P_n = 2\rho^n \frac{1-\rho}{1+\rho}$$

The M/M/2 queue

Mathematical analysis

$$P_n = 2\rho^n \frac{1-\rho}{1+\rho}$$



Compute #requests in the system:

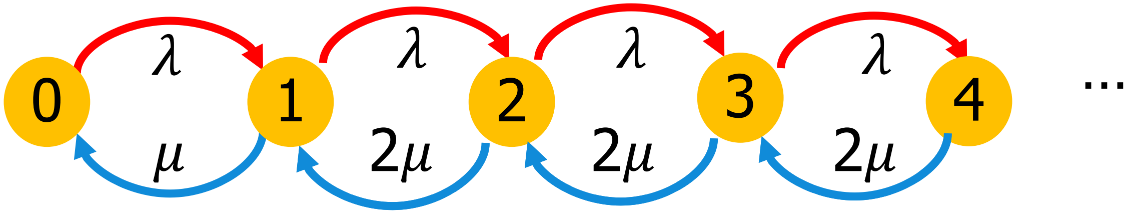
$$N = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} 2n\rho^n \frac{1-\rho}{1+\rho}$$

$$= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots = \frac{2\rho}{(1-\rho^2)}$$

The M/M/2 queue

Main results

$\rho = \lambda/2\mu$



Utilization	$U = 1 - P_0 = 2\rho / (1 + \rho)$
Prob. of n clients in the system	$P_n = 2\rho^n (1 - \rho) / (1 + \rho)$
Mean #clients in the system	$N = 2\rho / (1-\rho^2)$
Mean #clients in the queue	$N_Q = 2\rho^3 / (1-\rho^2)$
Mean response time	$R = N/\lambda = 1 / (\mu (1-\rho^2))$
Mean waiting time	$W = R - 1/\mu = \rho^2 / (\mu (1-\rho^2))$

The inspection paradox

Waiting at a queue

- D/D/1
  - $E[W] = 0$

How come?

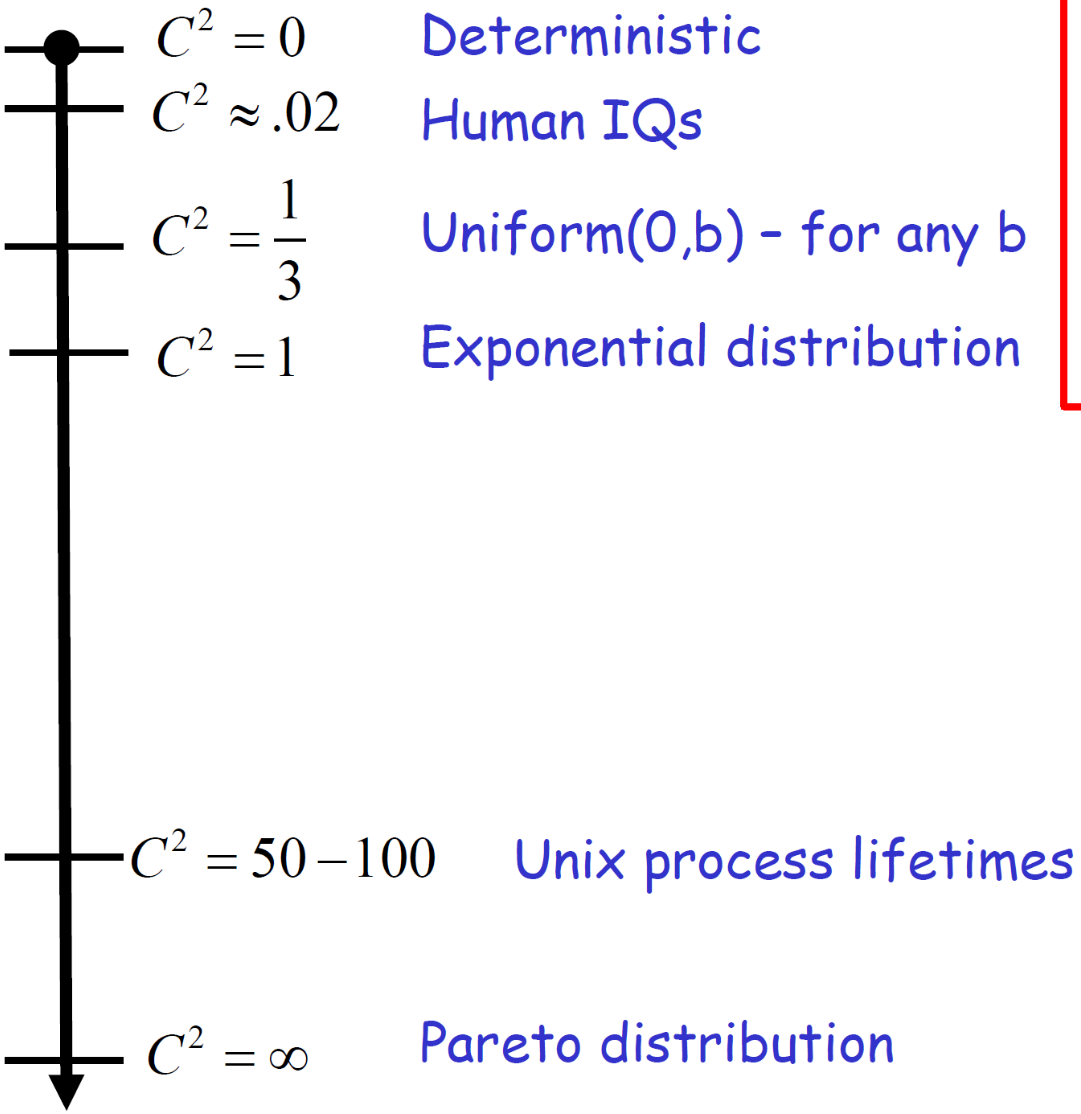
- M/M/1
  - $E[W] = \frac{\rho}{1-\rho} E[S]$

- M/G/1
  - $E[W] = \frac{\rho}{1-\rho} \frac{E[S^2]}{2E[S]} >> \frac{\rho}{1-\rho} E[S]$
  - $E[S^2] = (1 + C_v^2)\, E[S]^2$
  - C_v^2 = squared coefficient of variation

Is this bad?

Waiting for the bus

Variability in Job Sizes



$C^2 = 0$ Deterministic

$C^2 \approx .02$ Human IQs

$C^2 = \frac{1}{3}$ Uniform(0,b) - for any b

$C^2 = 1$ Exponential distribution

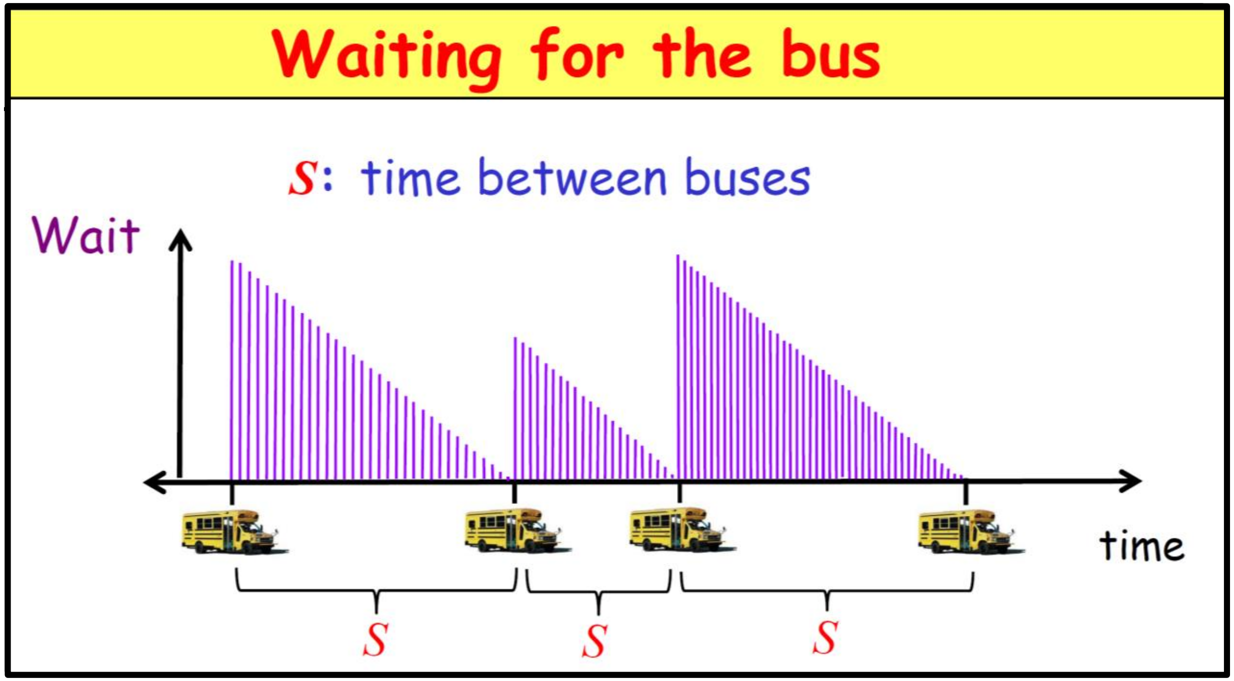
$C^2 = 50 - 100$ Unix process lifetimes

$C^2 = \infty$ Pareto distribution

Squared Coefficient of Variation

$$\mathrm{C}^2 = \frac{Var(S)}{E[S]^2}$$

Back to



- Have a steady stream of students take the bus and average their waiting times

- $E[\text{wait}] = \dfrac{\sum \text{wait}_s}{\#\text{students}} > E[S]/2$

The inspection paradox

Is everywhere

- Examples
  - everybody speeds at the highway (or goes much slower)
  - planes are always filled to the max
  - pubs are noisy
  - …
- M/G/1
  - $$E[W] = \boxed{\frac{\rho}{1-}} \boxed{\frac{E[S^2]}{2E[S]}}$$

?

high load leads to waiting

job size variance leads to waiting

increase server speed

smart scheduling (SRPT)