CS 4215: Quantitative Performance Evaluation for Computing ystems

Lydia Y. Chen
y.chen-10@tudelft.nl

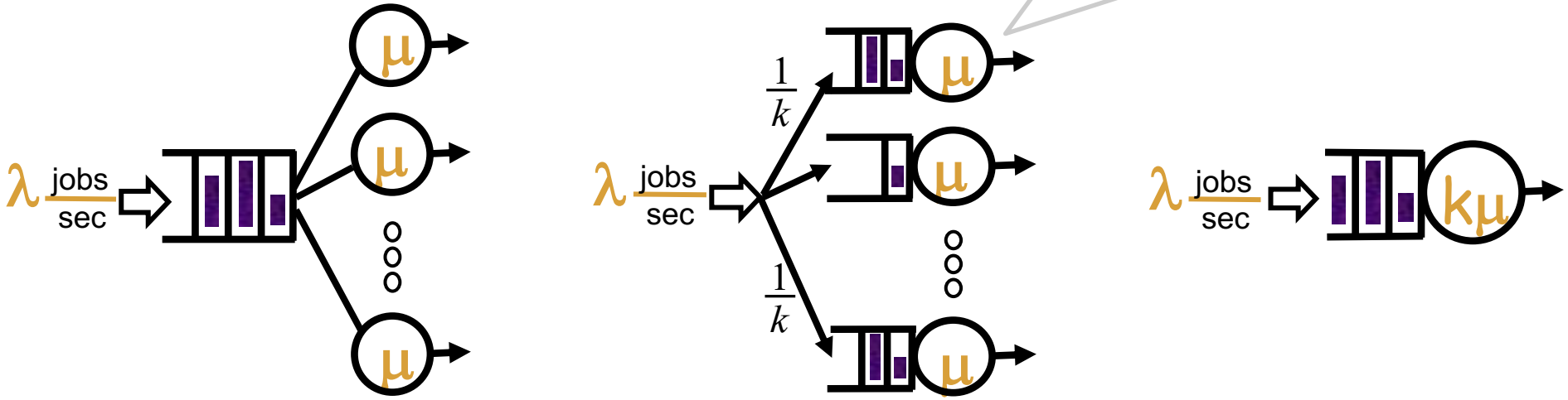
Modern computing systems

- Server
- Number of servers
- Number of cores
- Memory
- Cache
- Communication

- **Workload**
- **Arrival rate**
- **Applications**
- **Number of threads**

You will learn …

Which system design has the lowest average latency?



Last week

Operational Laws

Littles Law

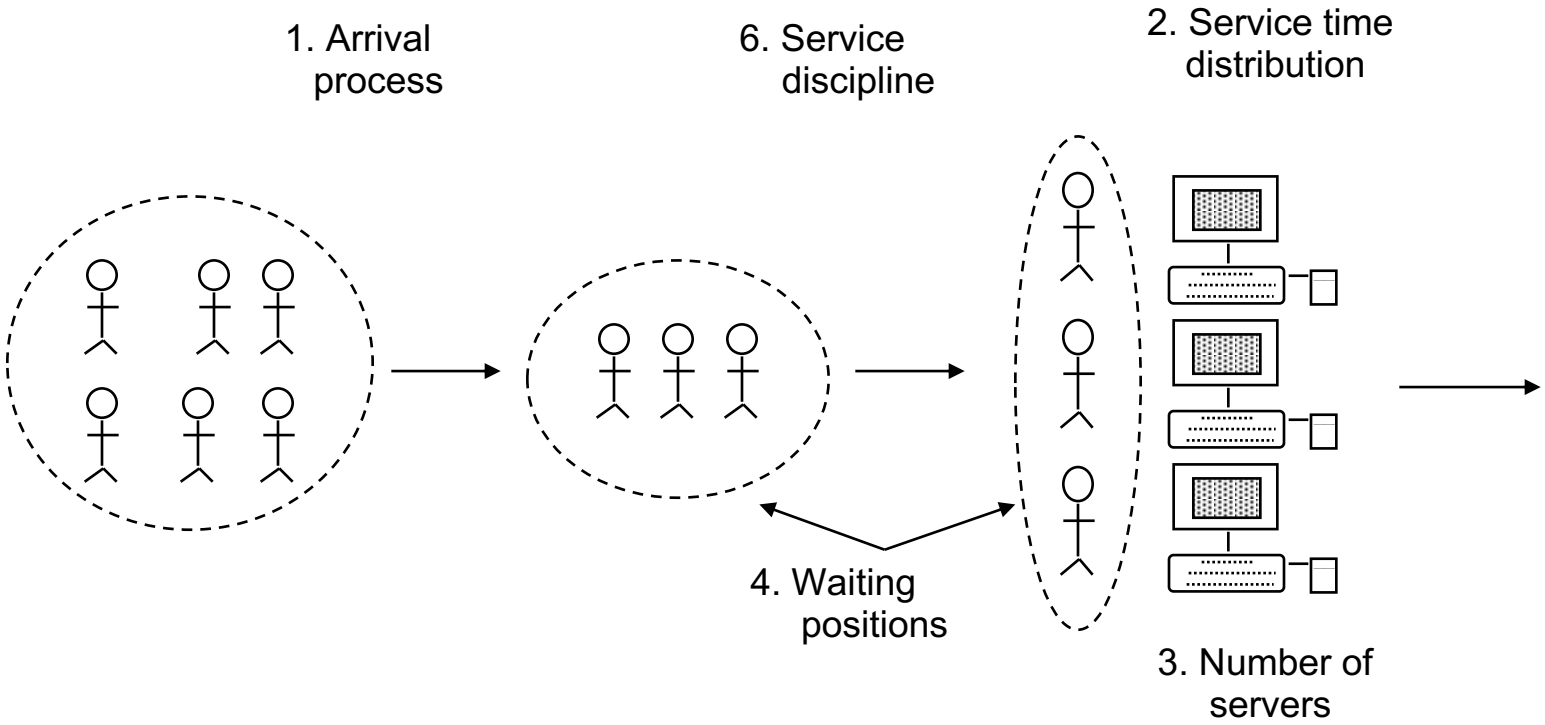
Forced Law

Utilization Law

Asymptotic Bounds

Terminology

Basic Components of a Queue



Kendall Notation *A/S/m/B/K/SD*

- *A*: Arrival process
- *S*: Service time distribution
- *m*: Number of servers
- *B*: Number of buffers (system capacity)
- *K*: Population size, and
- *SD*: Service discipline

Note. Defaults:
Infinite buffer capacity
Infinite population size
FCFS service discipline.
G/G/1 = G/G/1/1/1/FCFS

M: Exponential
E_k: Erlang *M*: Exponential
E_k: Erlang with parameter *k*
H_k: Hyper-exponential with parameter *k*
D: Deterministic ? constant
G: General All
G: General ? All

Arrival Process

- Arrival times: $t_1, t_2, \ldots, t_j$
- Interarrival times: $\tau_j = t_j - t_{j-1}$
- τ_j form a sequence of Independent and Identically Distributed (IID) random variables
- Exponential + IID ? Poisson
- Notation:
  - M = Memoryless = Poisson
  - E = Erlang
  - H = Hyper-exponential
  - G = General $\Rightarrow$ Results valid for all distributions

Service Time

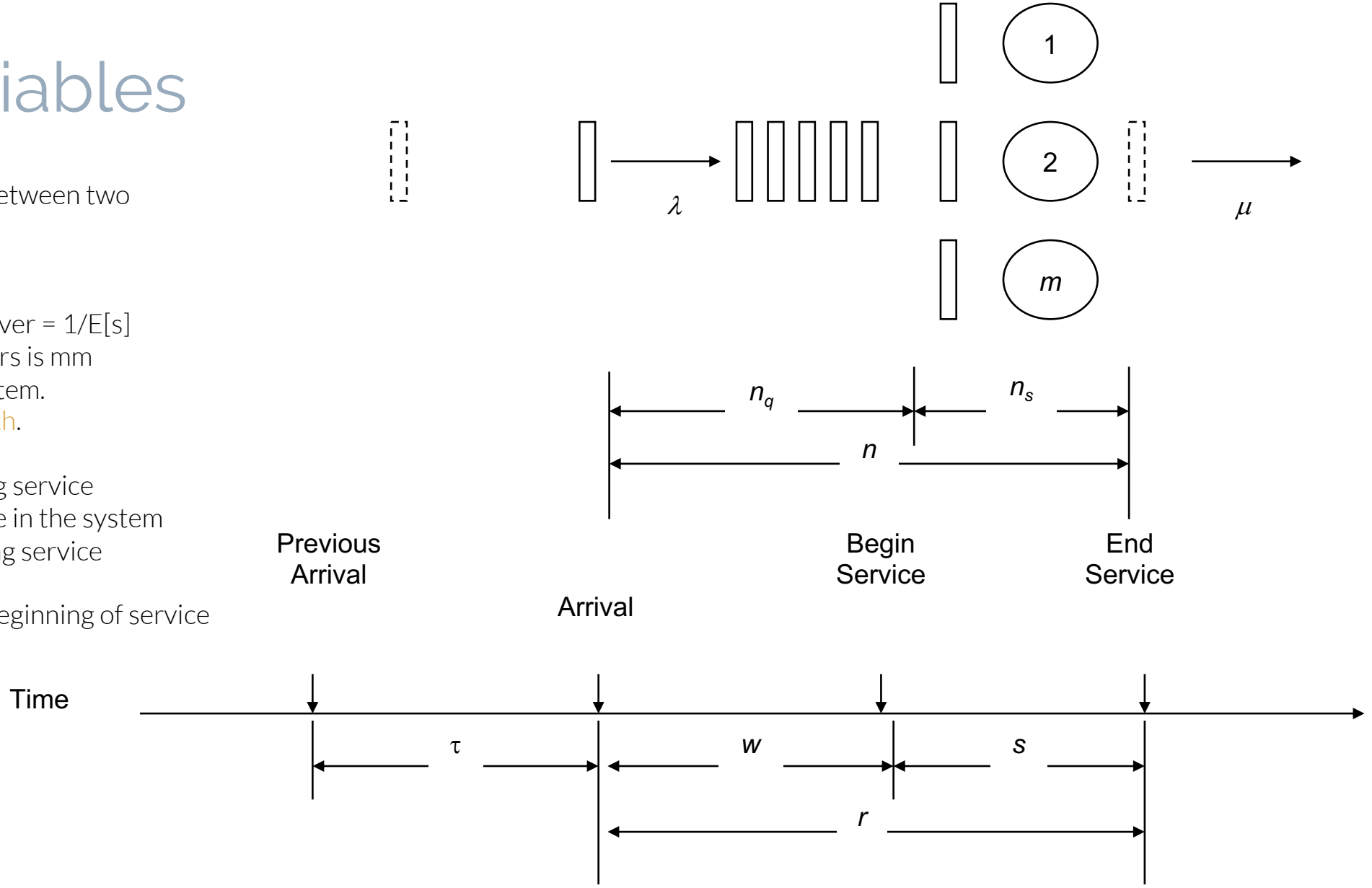
- Distribution
  - Time each student spends at the terminal.
  - Service times are IID.
  - Distribution: M (exponential), E (Erlang), H (hyper exponential), or G (general)
  - Device = Service center = Queue
  - Buffer = Waiting positions
- Discipline
  - First-Come-First-Served (FCFS)
  - Last-Come-First-Served (LCFS)
  - Last-Come-First-Served with Preempt and Resume (LCFS-PR)
  - Round-Robin (RR) with a fixed quantum.
  - Small Quantum $\Rightarrow$ Processor Sharing (PS)
  - Infinite Server: (IS) = fixed delay
  - Shortest Processing Time first (SPT)
  - Shortest Remaining Processing Time first (SRPT)

Example *M/M/3/20/1500/FCFS*

- Time between successive arrivals is exponentially distributed.
- Service times are exponentially distributed.
- Three servers
- *20* Buffers = *3* service + *17* waiting. After *20*, all arriving jobs are lost
- Total of *1500* jobs that can be serviced.
- Service discipline is first-come-first-served.

Key Variables

- t = Inter-arrival time = time between two successive arrivals.
- λ = Mean arrival rate = 1/E[t]
- s = Service time per job.
- μ = Mean service rate per server = 1/E[s]
- Total service rate for m servers is mm
- n = Number of jobs in the system.
  This is also called queue length.
- n_q = Number of jobs waiting
- n_s = Number of jobs receiving service
- r = Response time or the time in the system
  = time waiting + time receiving service
- w = Waiting time
  = Time between arrival and beginning of service



Rules for All Queues

What system is always stable?

1. Stability Condition

$$\lambda \leq m\mu_i$$

2. Little's Law holds

$$\mathrm{E}[N] = \lambda E[T]$$

3. Number in System v.s. Number in Queue

$$E[N] = E[N_Q] + E[N_s]$$
$$Var[N] = Var[N_Q] + Var[N_s]$$

4. Response time in system v.s. in queue

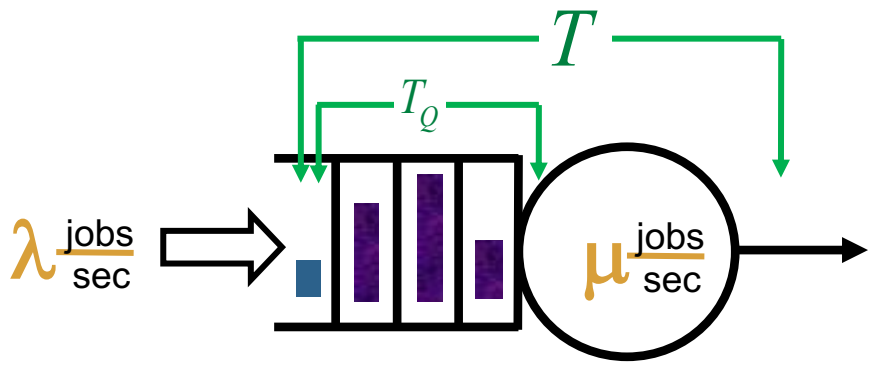
What is necessary condition?

$$E[T] = E[T_Q] + E[T_s]$$
$$Var[T] = Var[T_Q] + Var[T_s]$$

Single queue

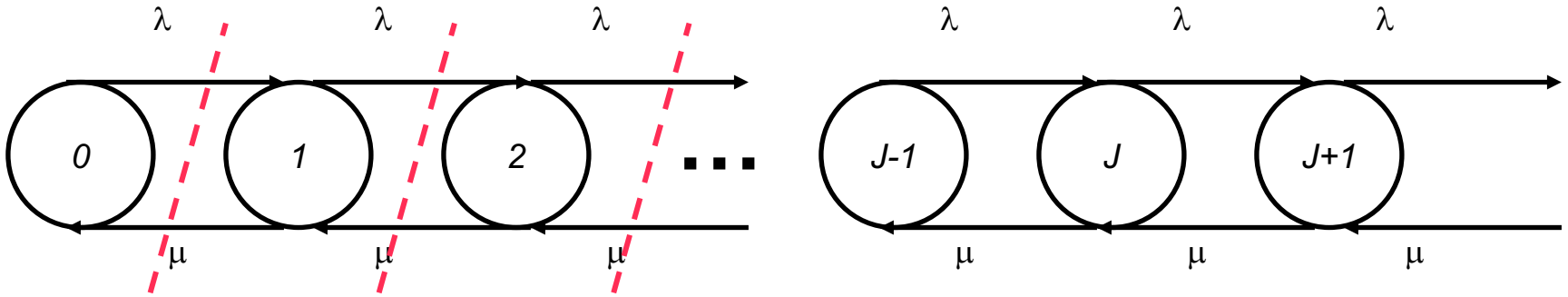
M/M/1 Queue

- The most commonly used type of queue
- Assumes that the interarrival times and the service times are exponentially distributed and there is only one server.
- No buffer or population size limitations and the service discipline is FCFS
- Need to know only the mean arrival rate λ and the mean service rate μ.



Modeling M/M/1 Queue

What is the state?



- Balance equation

$$\lambda P_o = \mu P_1$$
$$\lambda P_1 = \mu P_2$$

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad n = 1, 2, \ldots, \infty$$

$$\lambda P_n = \mu P_n$$

- Probability of *o* jobs in the system, P_o

$$\sum_{i=1}^{\infty} P_n = 1$$

$$P_o =$$

Results for M/M/1 Queue

- The quantity *l/m* is called traffic intensity and is usually denoted by symbol *r*.

$$p_n = \rho^n p_0$$

$$p_0 = \frac{1}{1 + \rho + \rho^2 + \cdots + \rho^\infty} = 1 - \rho$$

$$p_n = (1 - \rho)\rho^n \quad n = 0, 1, 2, \ldots, \infty$$

$$U = 1 - p_0 = \rho$$

Results for M/M/1 Queue(Cont)

- Mean number of jobs in the system:

$$E[n] \quad \text{How?} \quad = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$$

- Variance of the number of jobs in the system:

$$\begin{aligned} \mathrm{Var}[n] &= \text{How?} \\ &= \left(\sum_{n=1}^{\infty} n^2(1-\rho)\rho^n\right) - (E[n])^2 = \frac{\rho}{(1-\rho)^2} \end{aligned}$$

Results for M/M/1 Queue(Cont)

- Probability of n or more jobs in the system:

$$P(\geq n \text{ jobs in system}) \quad \text{How?} \quad = \rho^n$$

- Mean response time

What theorem helps?

How?

$$E[r] = \frac{E[n]}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{1/\mu}{1-\rho}$$

Results for M/M/1 Queue(Cont)

- Cumulative distribution function of the response time:

$$F(r) = 1 - e^{-r\mu(1-\rho)}$$

- The response time is exponentially distributed.
  $\Rightarrow q$-percentile of the response time

Response time distribution is a HARD problem!!!!!

$$1 - e^{-r_q\mu(1-\rho)} = \frac{q}{100}$$

$$r_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100}{100-q}\right)$$

Results for M/M/1 Queue (Advanced)

- Cumulative distribution function of the waiting time:

$$F(w) = 1 - \rho e^{-w\mu(1-\rho)}$$

- This is a truncated exponential distribution. Its q-percentile is given by:

$$w_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100\rho}{100-q}\right)$$

- The above formula applies only if *q* is greater than *100(1-ρ)*. All lower percentiles are zero.

$$w_q = max\left\{0, \frac{E[w]}{\rho} \ln\left(\frac{100\rho}{100-q}\right)\right\}$$

Example of M/</1

On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about two milliseconds to forward them. Using an M/M/1 model, analyze the gateway. What is the probability of buffer overflow if the gateway had only 13 buffers? How many buffers do we need to keep packet loss below one packet per million?

- Probability of buffer overflow

$$
\begin{aligned}
&= \quad \mathrm{P}(\text{more than } 13 \text{ packets in the gateway}) \\
&= \quad \rho^{13} = 0.25^{13} = 1.49 \times 10^{-8} \\
&\approx \quad 15 \text{ packets per billion packets.}
\end{aligned}
$$

Arrival rate l = 125 pps
Service rate m = 1/.002 = 500 pps
Gateway utilization U = *0.25*

- To limit the probability of loss to less than 10^{-6}

$$\rho^n \leq 10^{-6}$$

$$n > \log(10^{-6}) / \log(0.25) = 9.96$$

Other queues

- M/G/1

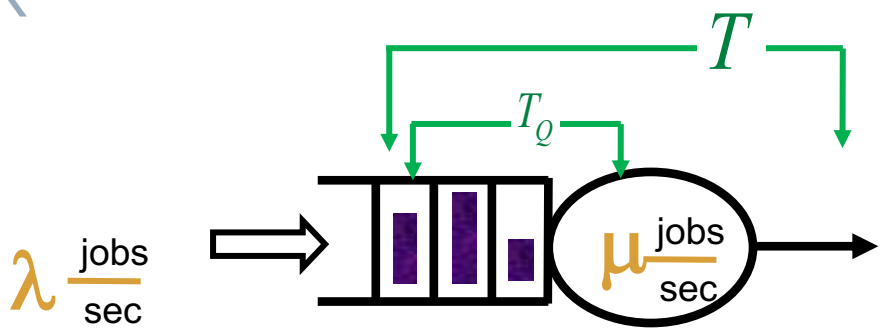
$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

What does that mean?

- G/M/1

- G/G/1

Closed look



Which one has the highest queue?

D/D/1

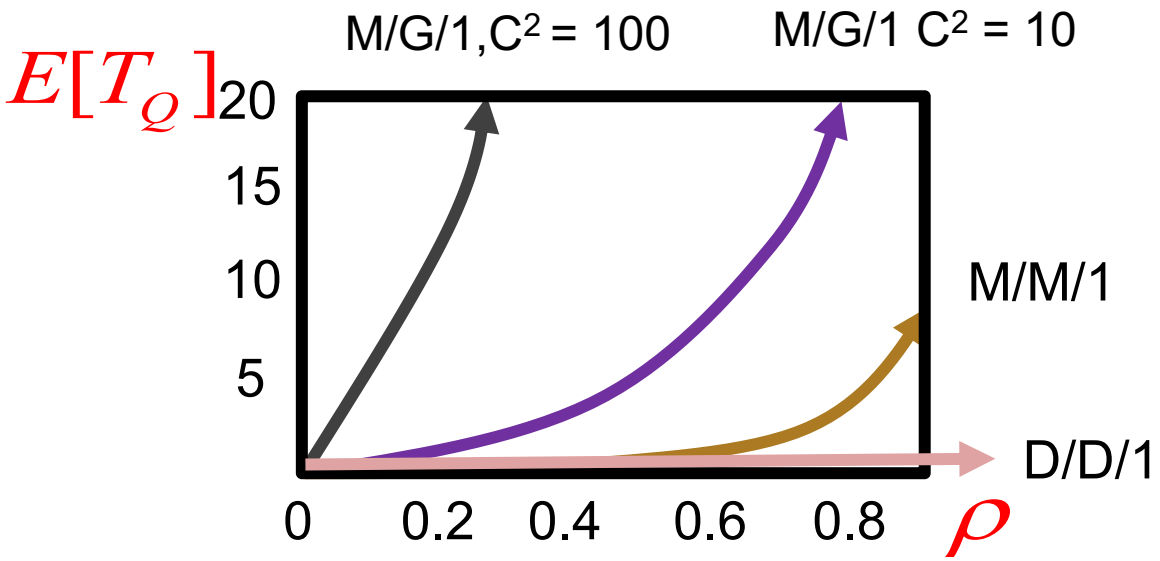
$$E[T_Q] = 0$$

M/M/1

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot E[S]$$

M/G/1

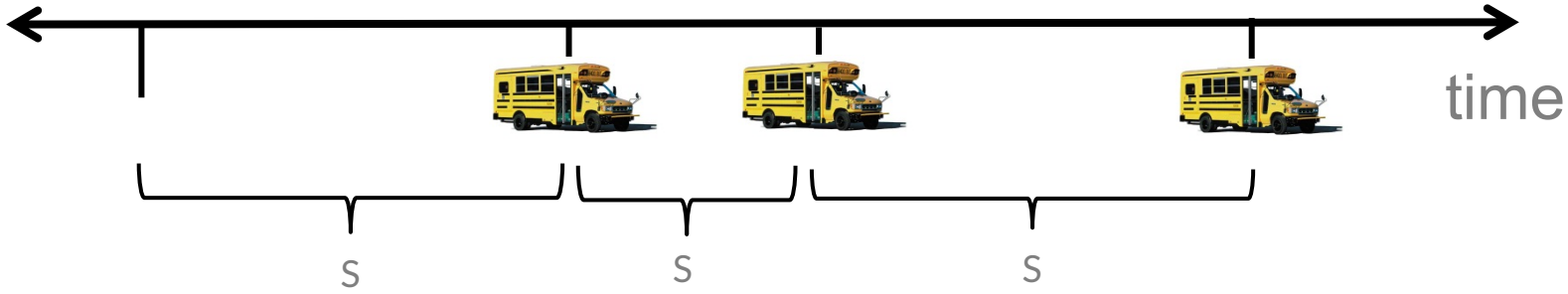
$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$



Variability matters a lot! From where?

Inspection paradox

Average waiting time (S) between buses is 10 minutes



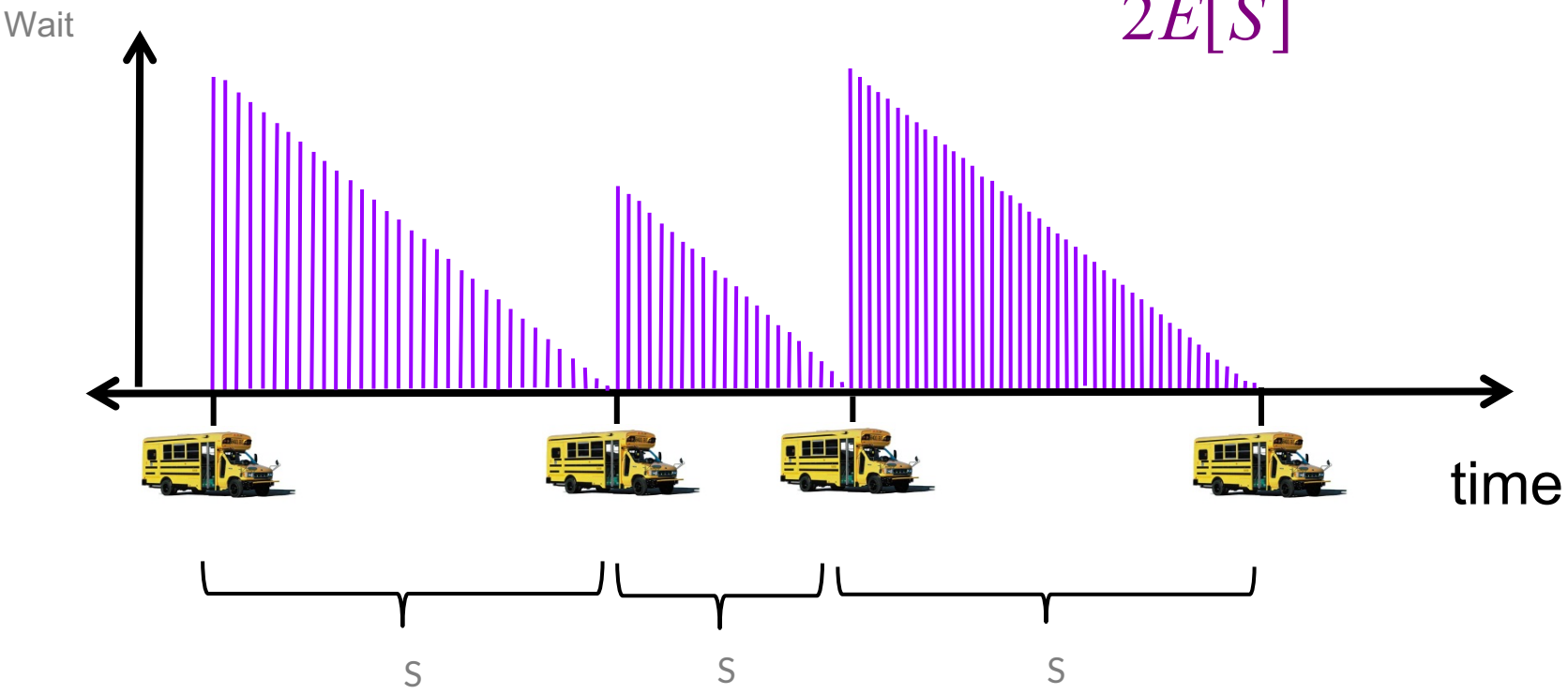
QUESTION:

On average, how long do I have to wait for a bus?

(a) < 5 min
(b) 5 min
(c) 10 min
(d) >10 min

Inspection paradox

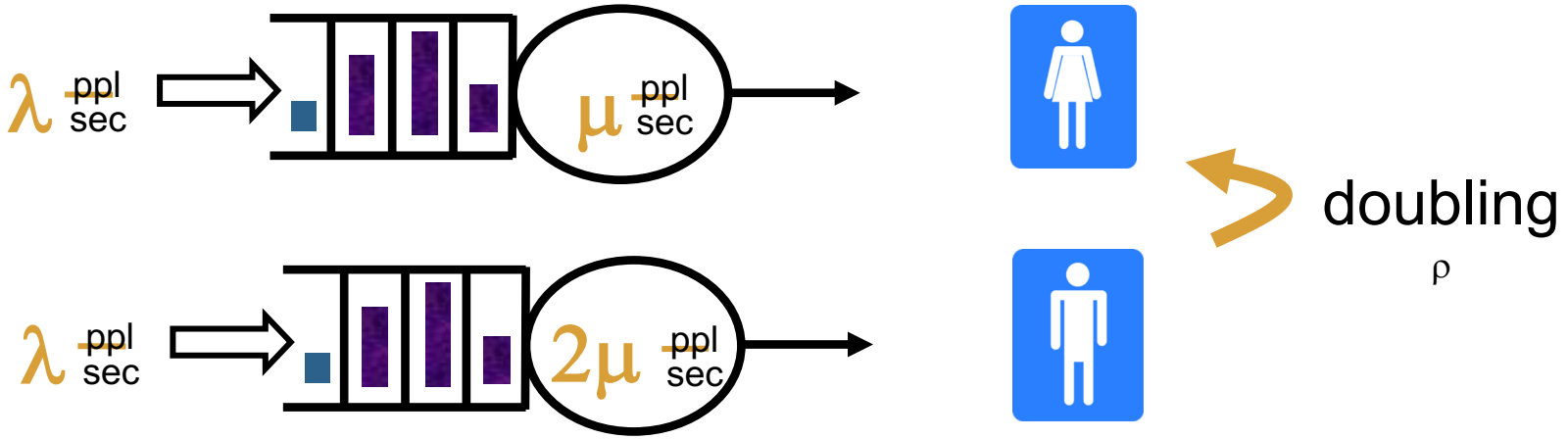
$$E[\text{Wait}] = \frac{E[S^2]}{2E[S]} >> E[S]$$



On average, women spend 88 sec in loo and men spend 40 sec in loo

How about their wait time? ?

Another perspective of variability

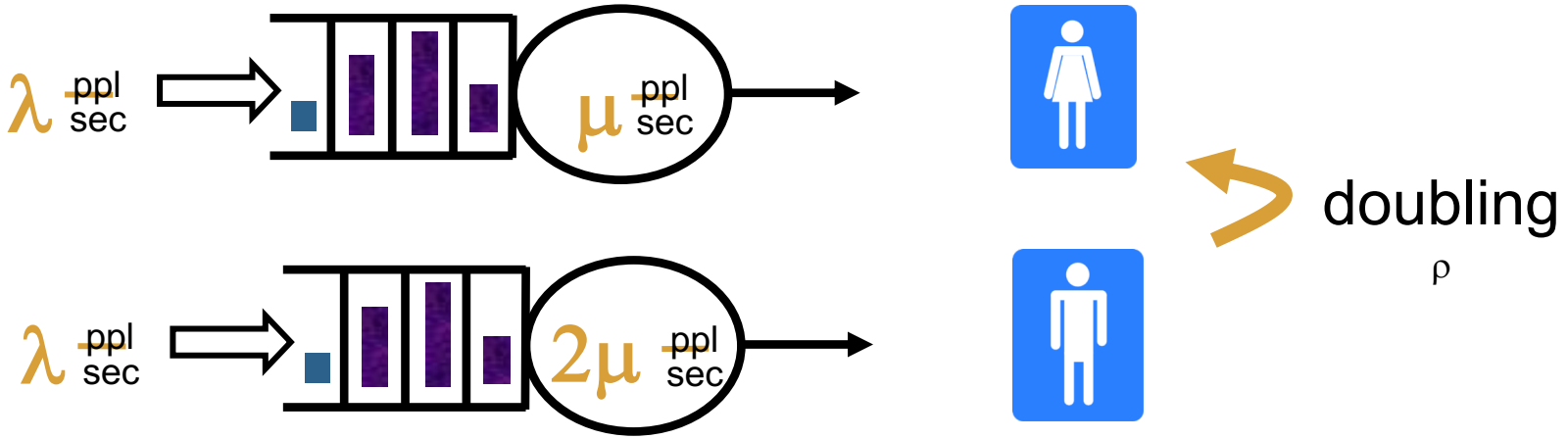


QUESTION:

Women take 2X as long. What's the difference in their wait?

(a) factor < 2
(b) factor 2
(c) factor 4
(d) factor > 4

Another perspective of variability

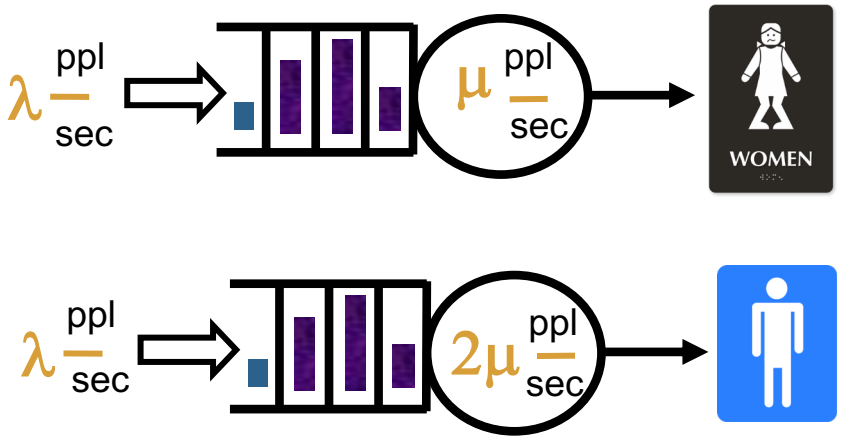


M/M/1 $$E[T_Q] = \frac{\rho}{1-\rho} \cdot E[S]$$

M/G/1 $$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

Doubling r can increase $E[T_Q]$ by factor of 4 to ∞

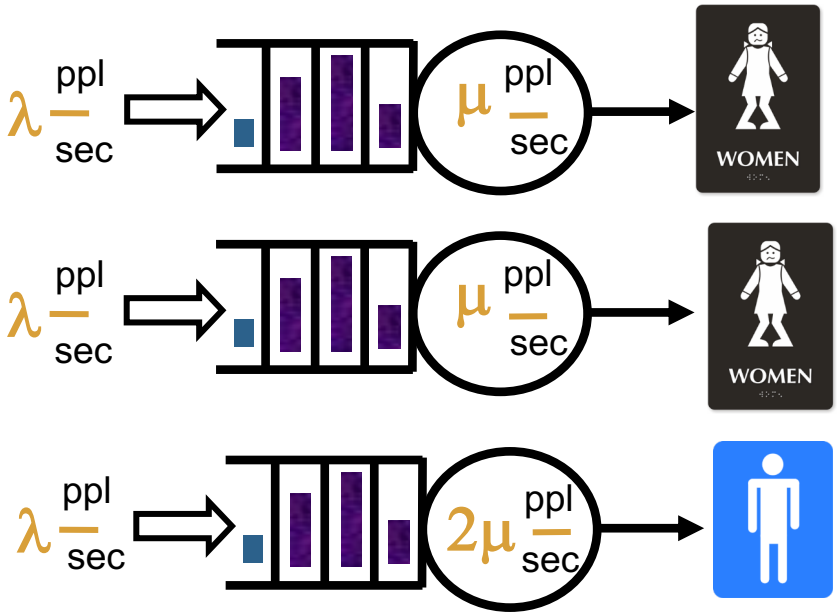
Another perspective of variability



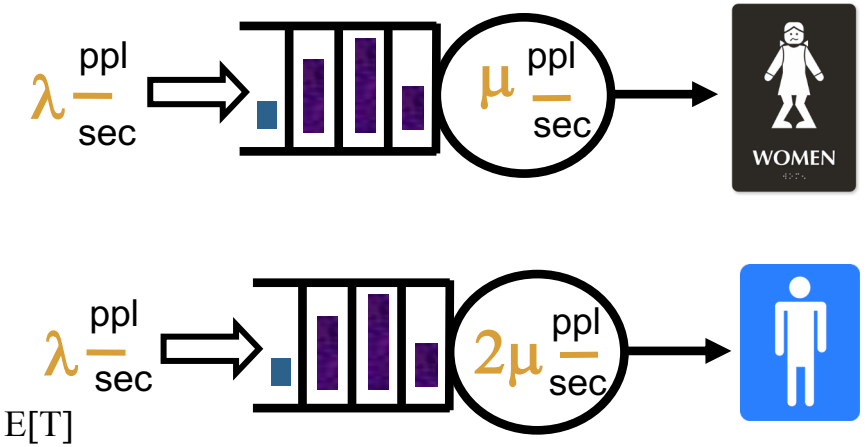
Solutions to reduce the waiting time?

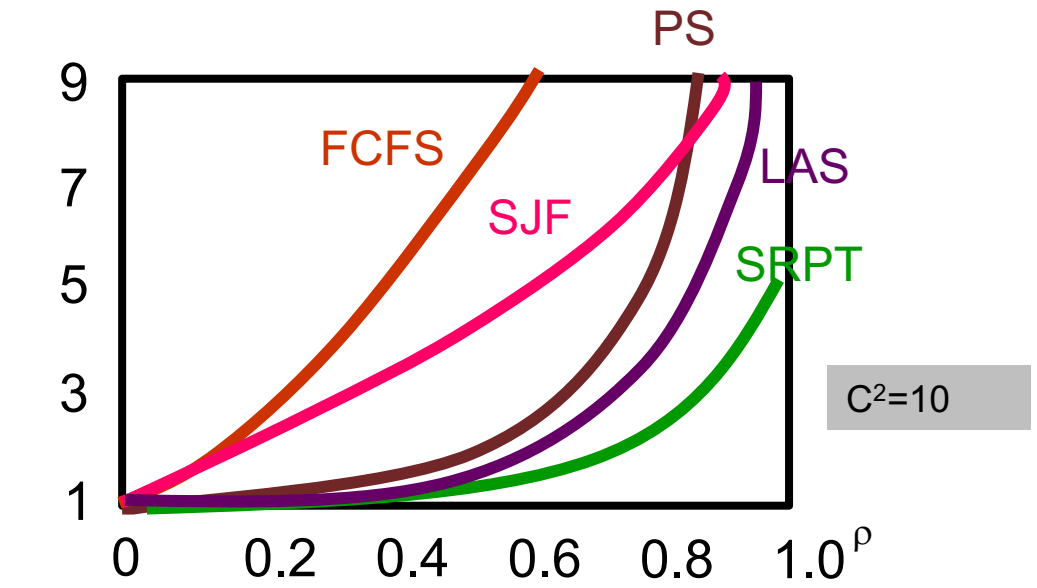
Insufficient!
Waiting time for women is still factor of 2 higher!!!
What else?

Solution 1



Another perspective of variability





Solutions to reduce the waiting time?

Solution 2

First, reduce the load
Secondly, ***smarter scheduling***

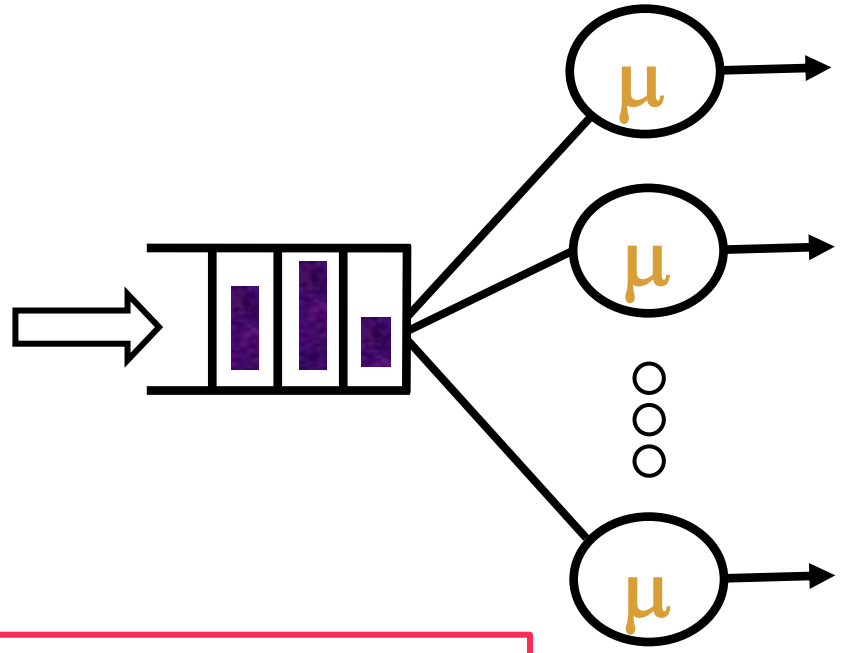
- FCFS (First-Come-First-Served, non-preemptive)
- PS (Processor-Sharing, preemptive)
- SJF (Shortest-Job-First, non-preemptive)
- SRPT (Shortest-Remaining-Processing-Time, preemptive)

Multiple servers

M/M/k Queue

- The mean arrival rate λ and the mean service rate per server is μ

Poisson process with rate λ $\frac{\text{jobs}}{\text{sec}}$

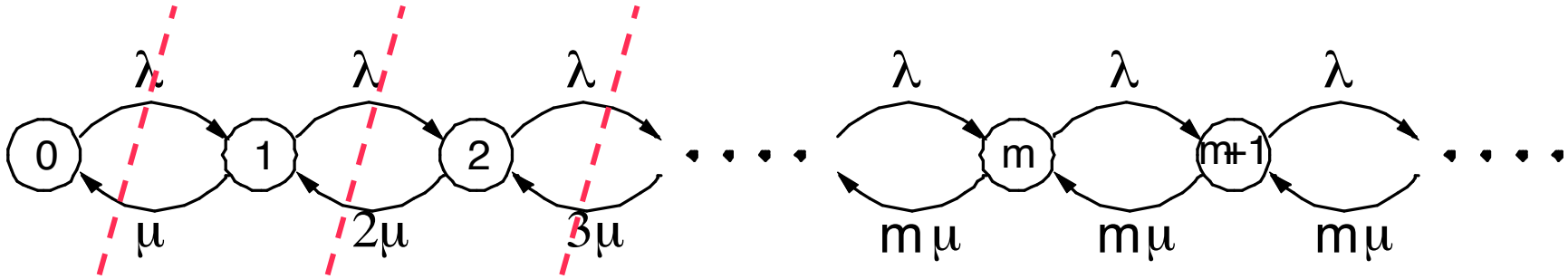


K servers

$$\rho \equiv \text{System Load} \equiv \frac{\lambda}{k\mu}$$

Modeling M/M/k Queue

What is the state?



- Balance equation

$$\lambda P_o = \mu P_1$$
$$\lambda P_1 = \mu P_2$$

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad n = 1, 2, \ldots, \infty$$

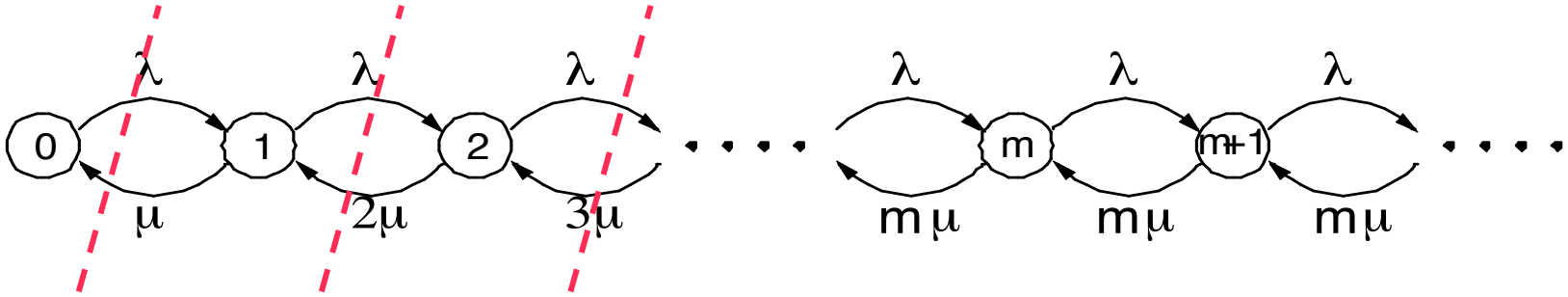
$$\lambda P_n = \mu P_n$$

- Probability of *o* jobs in the system, P_o

$$\sum_{i=1}^{\infty} P_n = 1$$

$P_o =$

Results for M/M/k



- Balance equation

How?

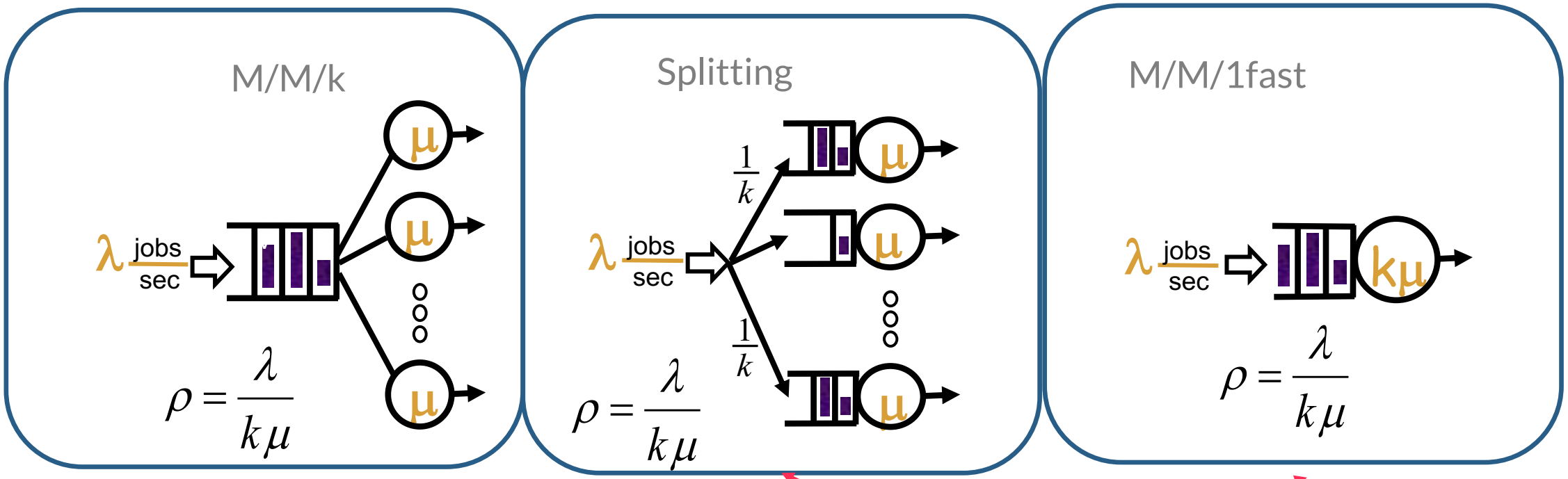
$$P_i = \begin{cases} P_0 \dfrac{(mp)^i}{i!}, & \text{for} \quad i \le m \\ P_0 \dfrac{m^m \rho^i}{m!}, & \text{for} \quad i > m \end{cases}$$

How?

$$P_0 = \left[\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)} \right]^{-1}$$

Comparing systems

Q: Which is best for minimizing E[T]?

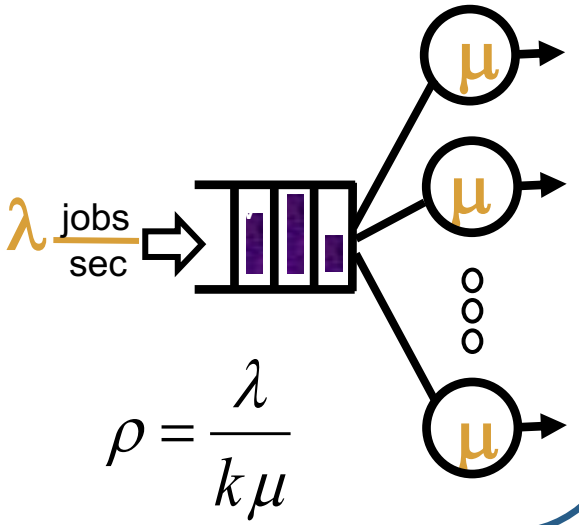


$$E[T_Q]^{M/M/1} = \frac{\rho}{1-\rho} \cdot E[S]$$

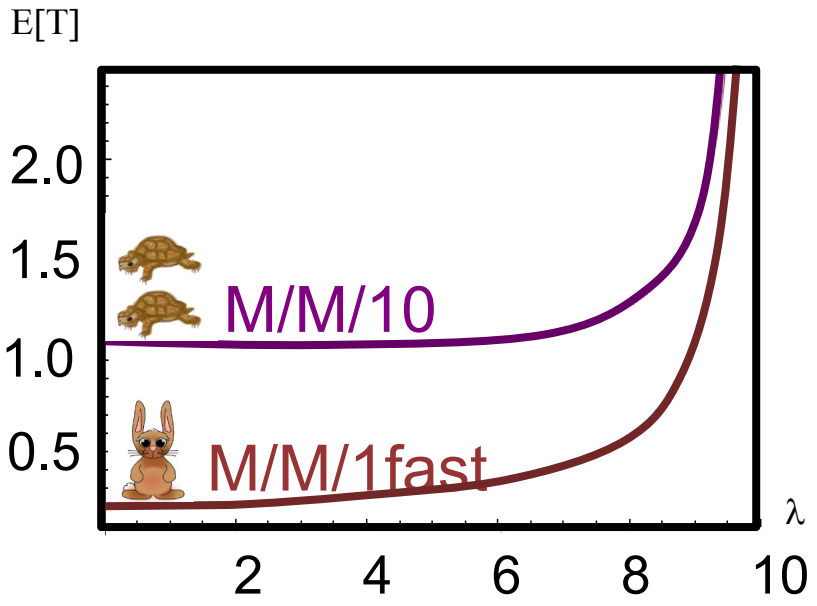
Many slow or 1 fast

Q: Which is best for minimizing E[T]?

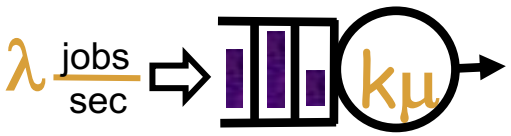
M/M/k



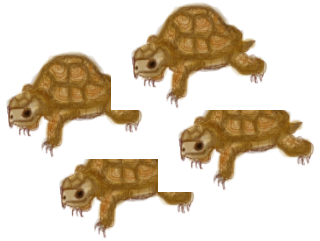
$$\rho = \frac{\lambda}{k\mu}$$



M/M/1fast



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

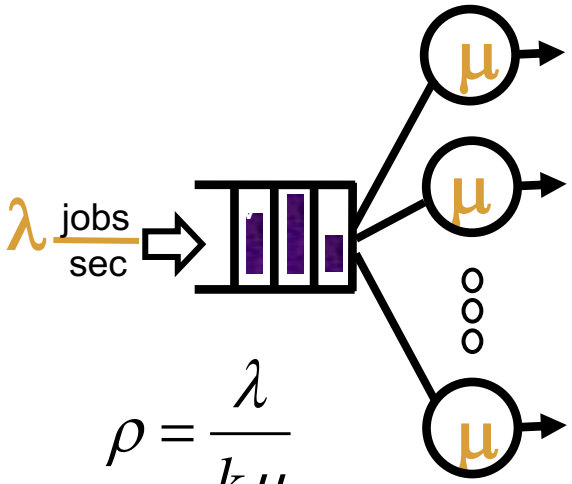


VS.

Many slow or 1 fast: variability

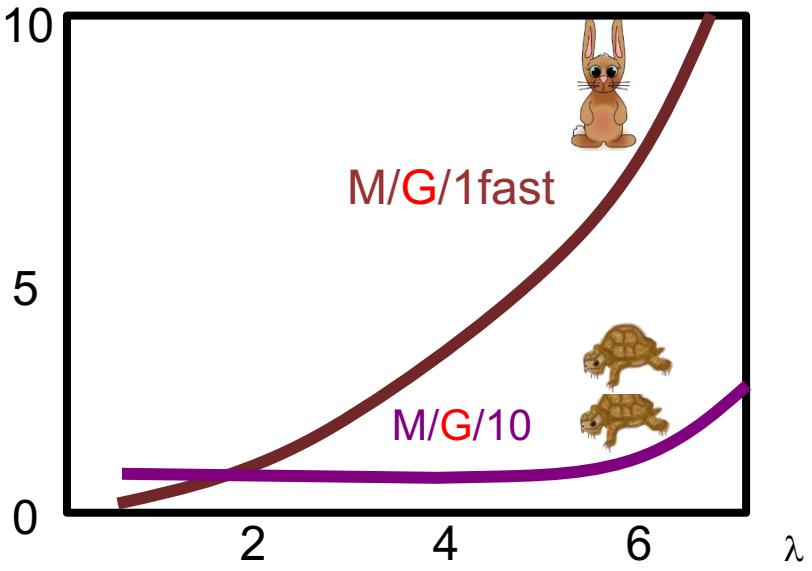
Q: What if $C_G^2 = 100$

M/**G**/k

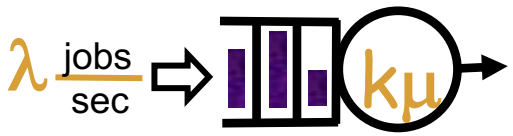


$$\rho = \frac{\lambda}{k\mu}$$

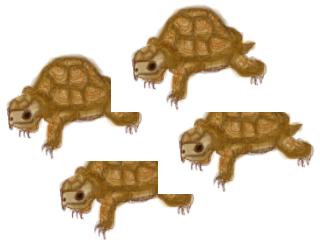
Why??



M/**G**/1fast

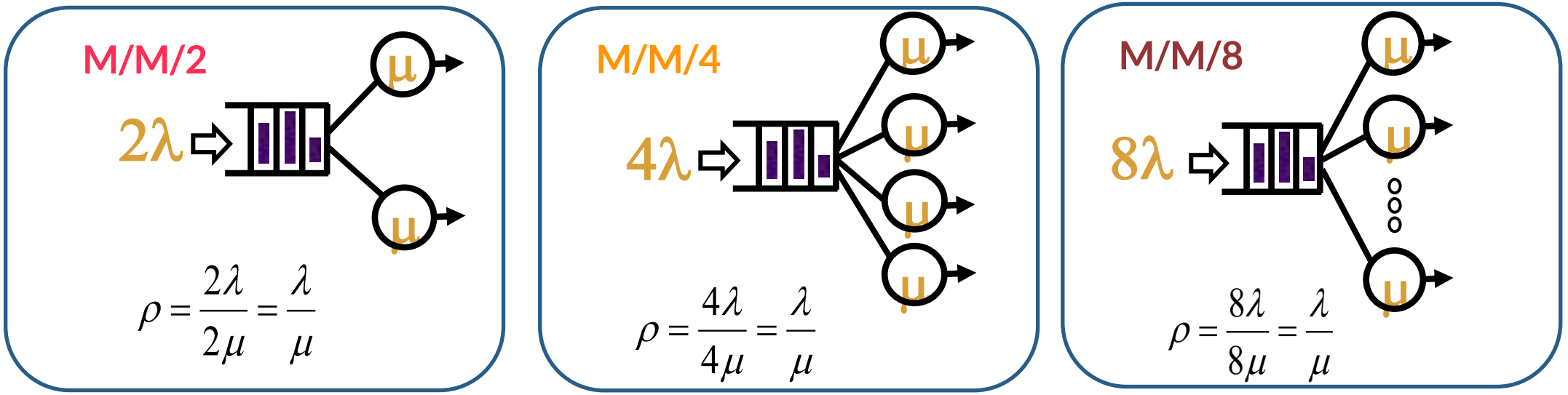


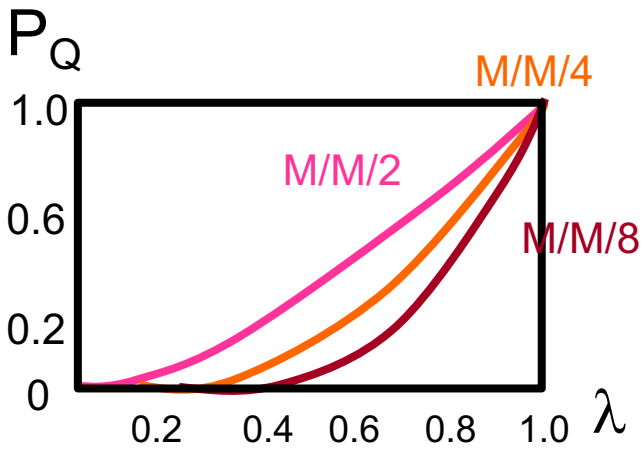
$$\rho = \frac{\lambda}{k\mu}$$

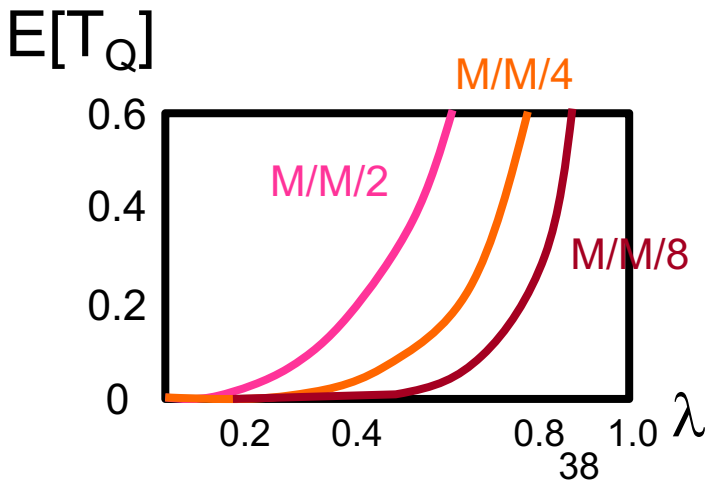


Check: P_Q, probability an arrival has to queue

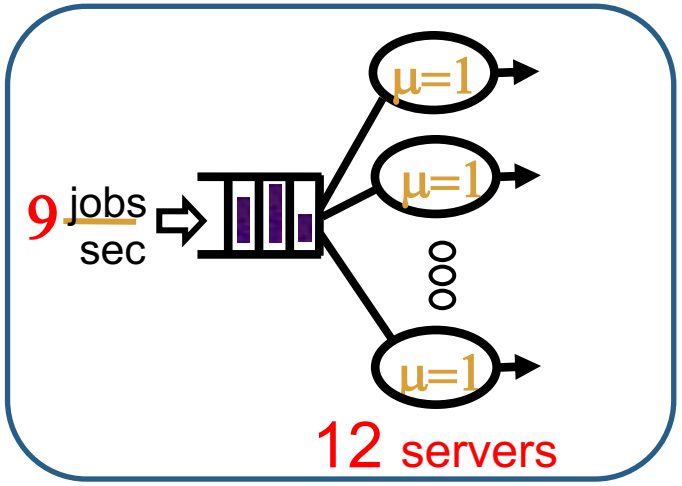
Proportional scaling is over killed





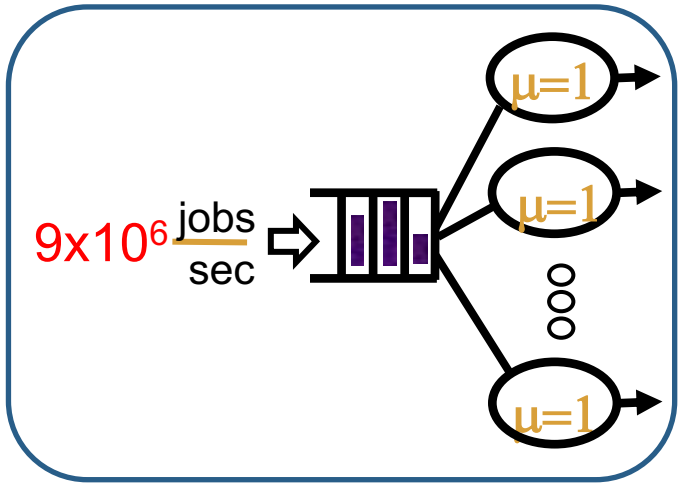


Many slow or 1 fast: variability



QUESTION: If arrival rate becomes 10^6 times higher, how many servers do we need to keep P_Q the same:

(a) 9.1×10^6 (b) 10×10^6 (c) 11×10^6 (d) 12×10^6 (e) 13×10^6



"Square root staffing"
[Halfin, Whitt OR 1981]

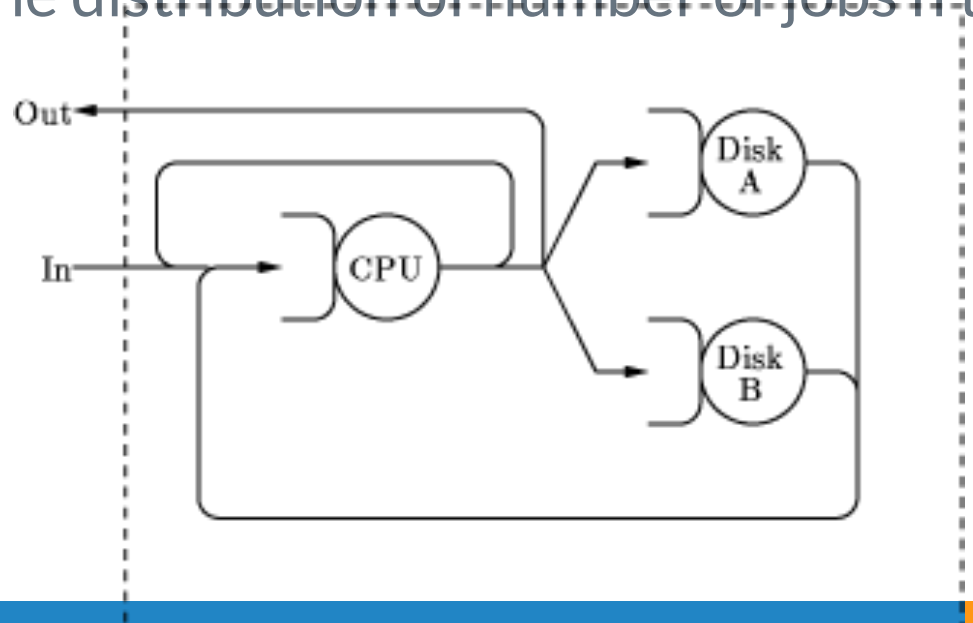
Let R be the minimum #servers for stability.

Then $R + \sqrt{R}$ servers yields $P_Q = 20\%$.

Queueing network

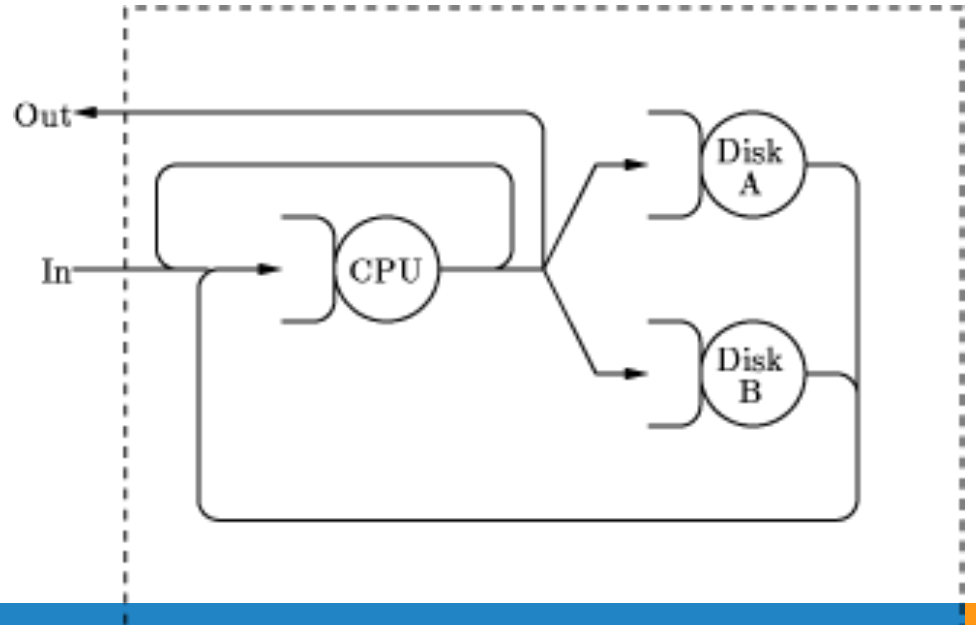
Open Queueing Networks

- **Queueing Network**: model in which jobs departing from one queue arrive at another queue (or possibly the same queue)
- **Open queueing network**: external arrivals and departures
  - Number of jobs in the system varies with time.
  - Throughput = arrival rate
  - Goal: To characterize the distribution of number of jobs n the system.



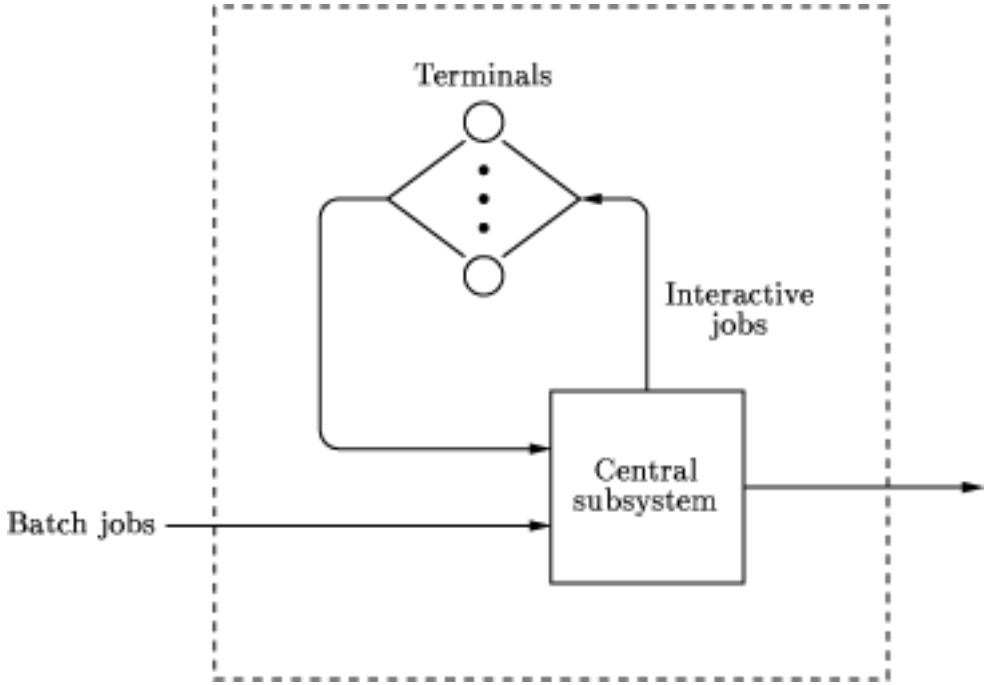
Closed Queueing Networks

- Closed queueing network: No external arrivals or departures
  - Total number of jobs in the system is constant
  - `OUT' is connected back to `IN.'
  - Throughput = flow of jobs in the OUT-to-IN link
  - Number of jobs is given, determine the throughput

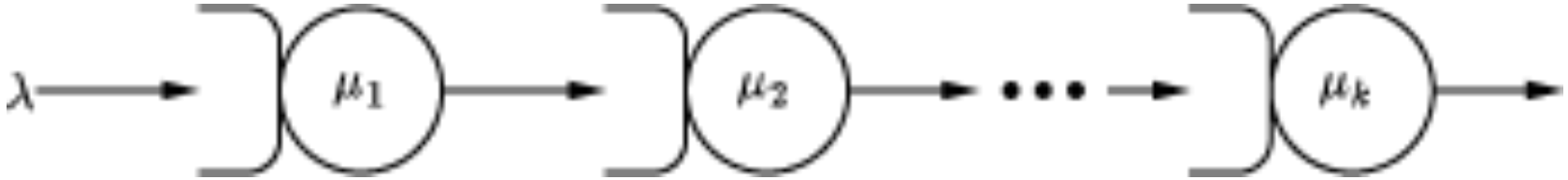


Mixed Queueing Networks

- **Mixed queueing networks**: Open for some workloads and closed for others $\Rightarrow$ Two classes of jobs. **Class** = types of jobs.
  All jobs of a single class have the same service demands and transition probabilities. Within each class, the jobs are indistinguishable.



Series Networks



- *k M/M/1* queues in series
- Each individual queue can be analyzed independently of other queues
- Arrival rate= λ. If μ_i is the service rate for *i*th server:

$$\text{Probability of } n_i \text{ jobs in the } i^{\text{th}} \text{ queue } = (1-\rho_i)\rho_i^{n_i}$$

$$\text{Utilization of } i^{th} \text{ server} \rho_i = \lambda/\mu_i$$

- Joint probability of queue lengths:

$$\begin{aligned} &P(n_1, n_2, n_3, \ldots, n_M) \\ &= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}(1-\rho_3)\rho_3^{n_3}\cdots(1-\rho_M)\rho_M^{n_M} \\ &= p_1(n_1)p_2(n_2)p_3(n_3)\cdots p_M(n_M) \end{aligned}$$

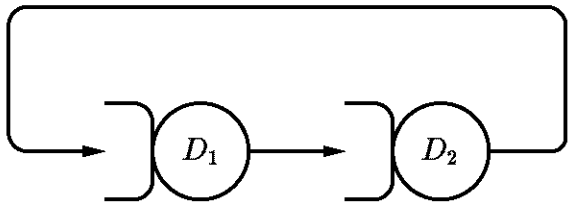
Product-Form Network

- Any queueing network in which:

$$P(n_1, n_2, \ldots, n_M) = \frac{1}{G(N)} \prod_{i=1}^{M} f_i(n_i)$$

- When $f_i(n_i)$ is some function of the number of jobs at the ith facility, $G(N)$ is a normalizing constant and is a function of the total number of jobs in the system.

Example



- Consider a closed system with two queues and N jobs circulating among the queues:
- Both servers have an exponentially distributed service time. The mean service times are 2 and 3, respectively. The probability of having n_1 jobs in the first queue and n_2=N-n_1 jobs in the second queue can be shown to be:

$$P(n_1, n_2) = \frac{1}{3^{N+1} - 2^{N+1}} (2^{n_1} \times 3^{n_2})$$

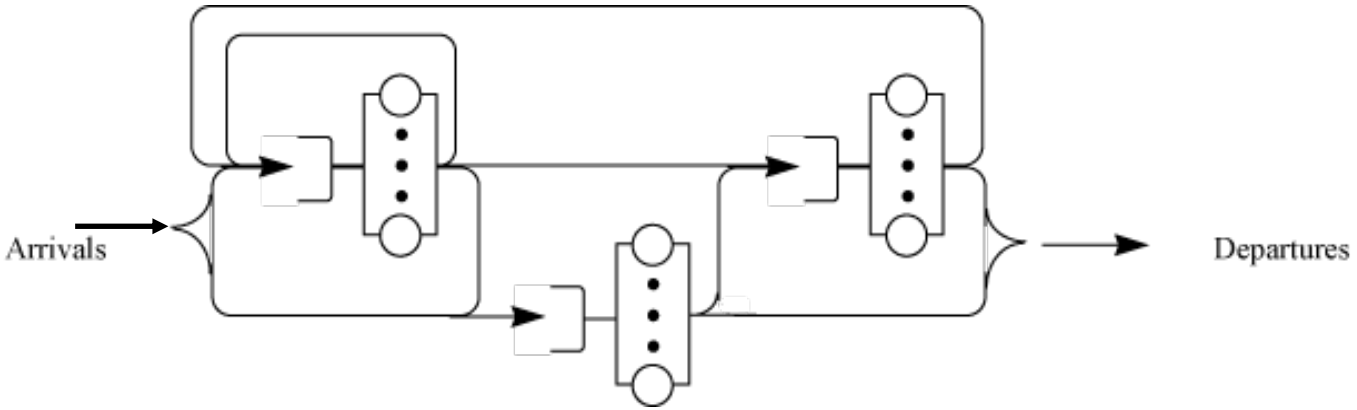
- In this case, the normalizing constant *G(N)* is 3^{N+1}-2^{N+1}.
- The state probabilities are products of functions of the number of jobs in the queues. Thus, this is a ***product form network***.

General Open Network of Queues (Jackson)

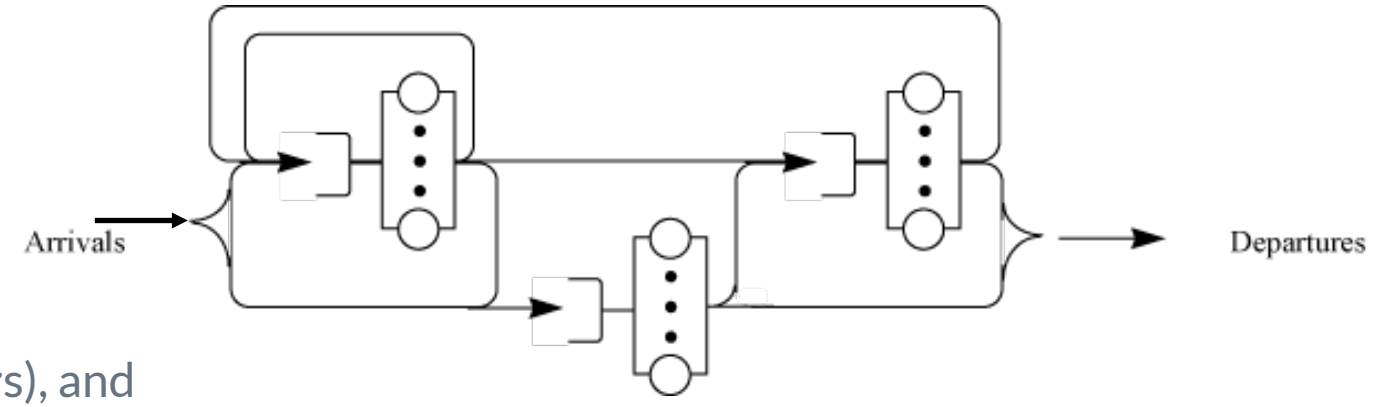
- If all queues are single-server queues, the queue length distribution is:

$$
\begin{aligned}
&P(n_1, n_2, n_3, \ldots, n_M) \\
&= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}(1-\rho_3)\rho_3^{n_3}\cdots(1-\rho_M)\rho_M^{n_M} \\
&= p_1(n_1)p_2(n_2)p_3(n_3)\cdots p_M(n_M)
\end{aligned}
$$

- Note: Queues are not independent *M/M/1* queues with a Poisson arrival process.
- In general, the internal flow in such networks is not Poisson. Particularly, if there is any feedback in the network, so that jobs can return to previously visited service centers, the internal flows are not Poisson.



Closed Product-Form Networks (BCMP)



1. **Service Disciplines**:
   - First-come-first-served (FCFS),
   - Processor sharing (PS),
   - Infinite servers (IS or delay centers), and
   - Last-come-first-served-preemptive-resume (LCFS-PR).
2. **Job Classes**: The jobs belong to a single class while awaiting or receiving service at a service center, but may change classes and service centers according to fixed probabilities at the completion of a service request.
3. **Service Time Distributions**:
   - At FCFS service centers, the service time distributions must be identical and exponential for all classes of jobs.
   - At other service centers, where the service times should have probability distributions with rational Laplace transforms;
   - Different classes of jobs may have different distributions.

BCMP Networks(Cont)

4. State Dependent Service:

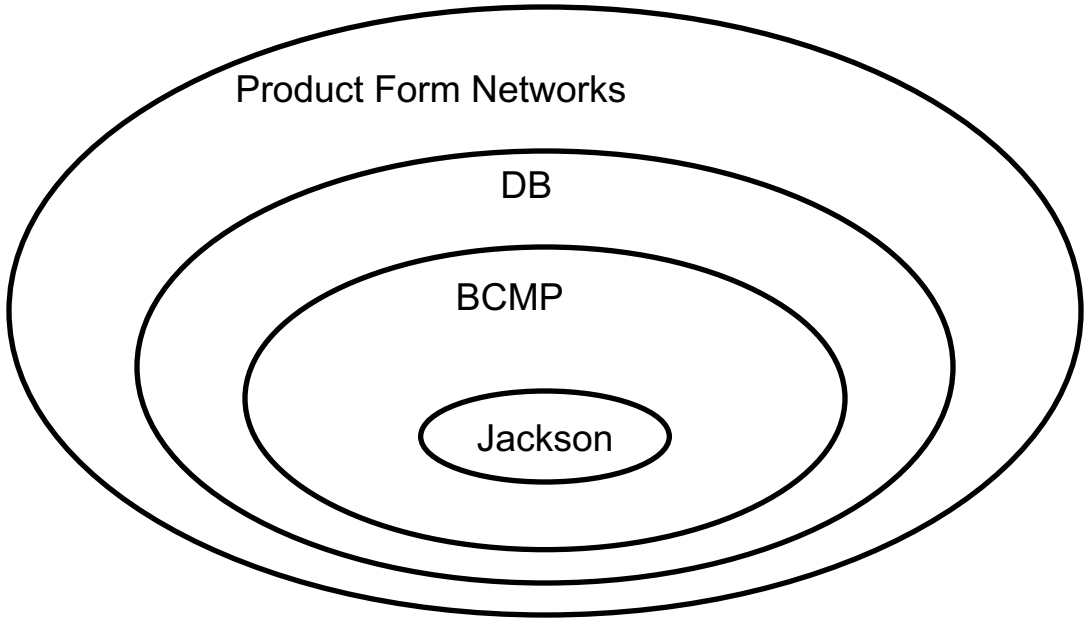
- The service time at a FCFS service center can depend only on the total queue length of the center.
- The service time for a class at PS, LCFS-PR, and IS center can also depend on the queue length for that class, but not on the queue length of other classes.
- Moreover, the overall service rate of a subnetwork can depend on the total number of jobs in the subnetwork

5. Arrival Processes:

- In open networks, the time between successive arrivals of a class should be exponentially distributed.
- No bulk arrivals are permitted.
- The arrival rates may be state dependent.
- A network may be open with respect to some classes of jobs and closed with respect to other classes of jobs.

Example of QN

Summary



- Product form networks: Any network in which the system state probability is a product of device state probabilities
- Jackson: Network of M/M/m queues
- BCMP: More general conditions
- Denning and Buzen: Even more general conditions

Bonus quizs

Next lecture:

Thanks!

Any questions?

y.chen-10@tudelf.nl
lydiaychen@ieee.org

Results for M/M/1 Queue(Cont)

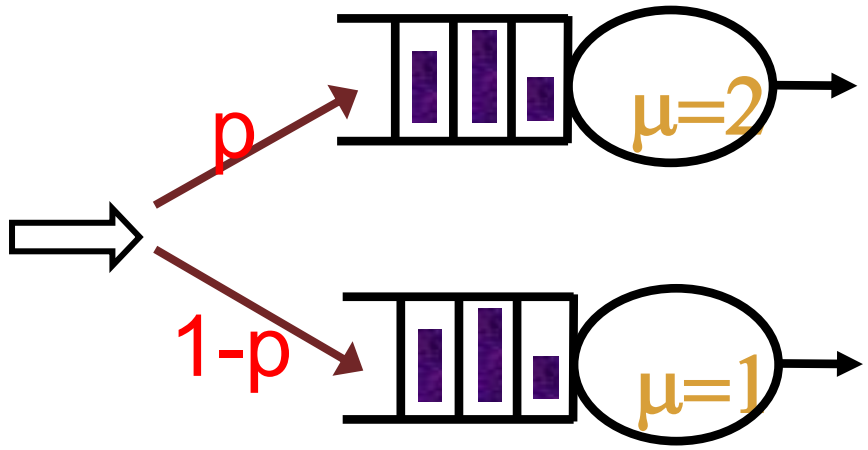
- Mean number of jobs in the queue:

$$E[n_q] = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty}(n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$$

- Idle $\Rightarrow$ there are no jobs in the system
- The time interval between two successive idle intervals

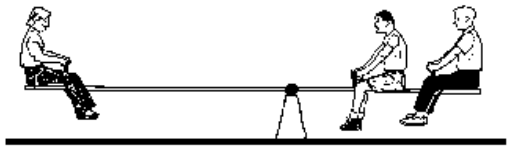
Load balancing

Poisson process with rate λ jobs/sec



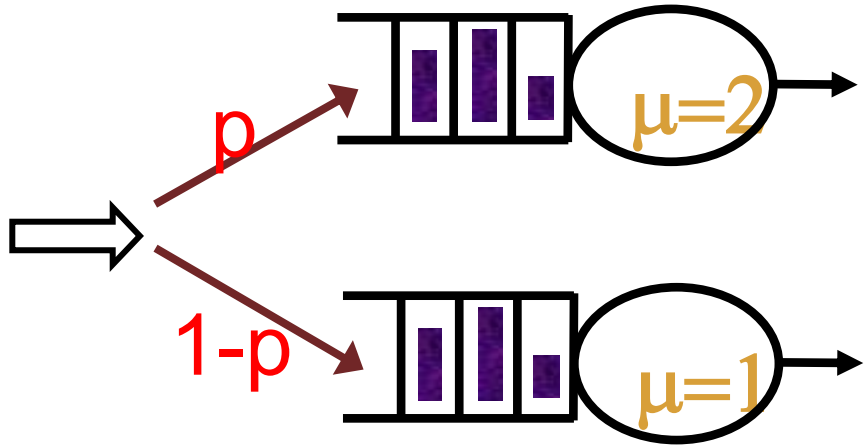
QUESTION: What is the optimal p to minimize **E[T]**?

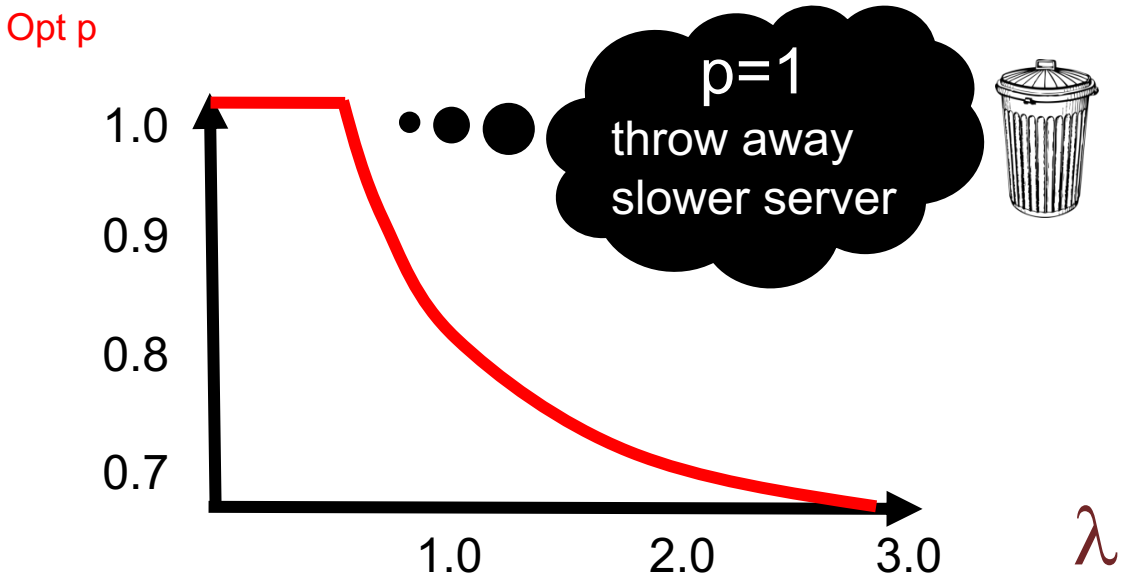
(a) $p = 2/3$ (b) $p > 2/3$ (c) $p < 2/3$



Load balancing

Poisson process with rate λ jobs/sec





Presentation design

This presentations uses the following typographies and colors:

- Titles: **Raleway**
- Body copy: **Lato**

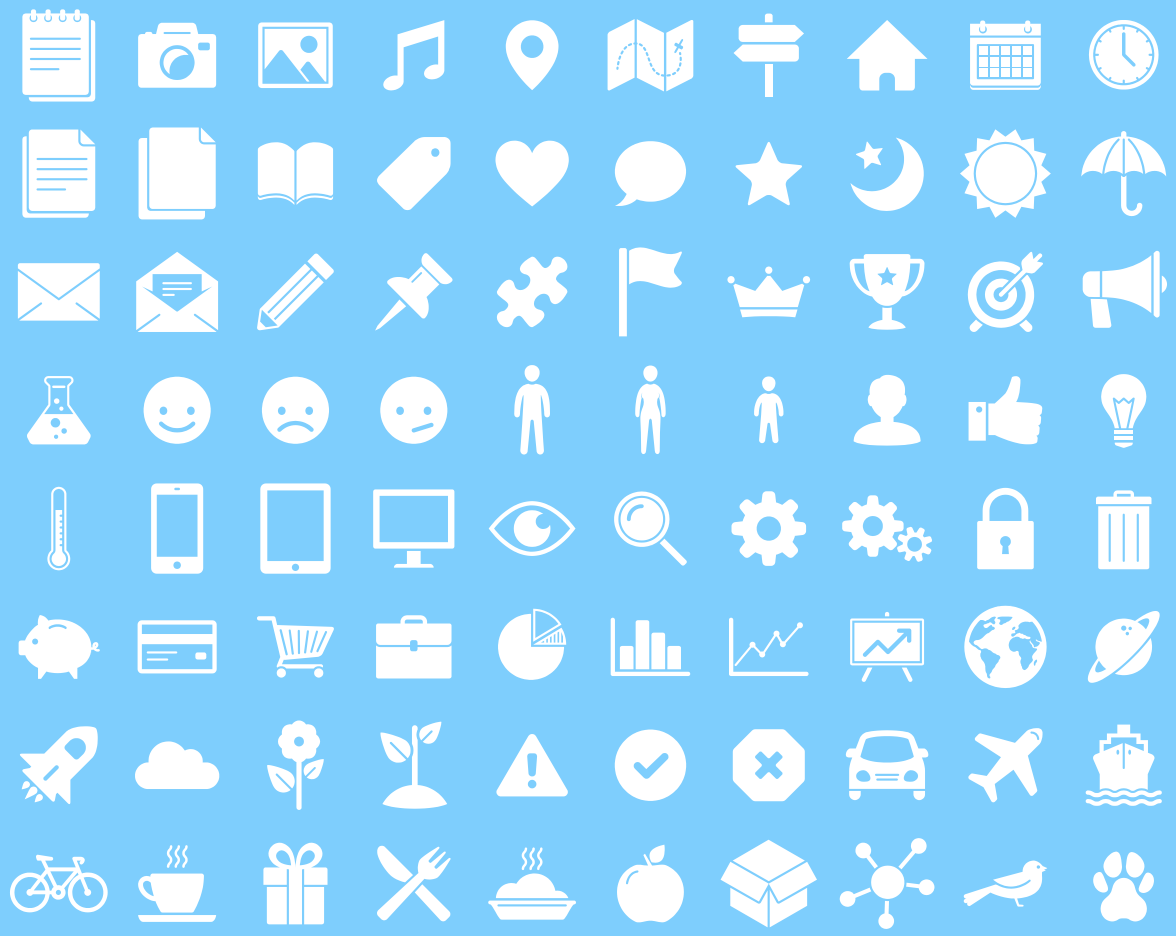
You can download the fonts on these pages:

https://www.fontsquirrel.com/fonts/raleway

https://www.fontsquirrel.com/fonts/lato

- Dark blue **#2185c5**
- Light blue **#7ecefd**
- Yellow **#ff9715**
- Magenta **#f20253**
- Dark gray **#677480**
- Light gray **#97abbc**

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®



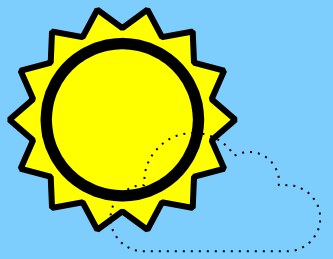
SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change fill color and opacity.
- Change line color, width and style.

Isn't that nice? :)

Examples:



Now you can use any emoji as an icon!
And of course it resizes without losing quality and you can change the color.

How? Follow Google instructions
https://twitter.com/googledocs/status/730087240156643328

and many more...