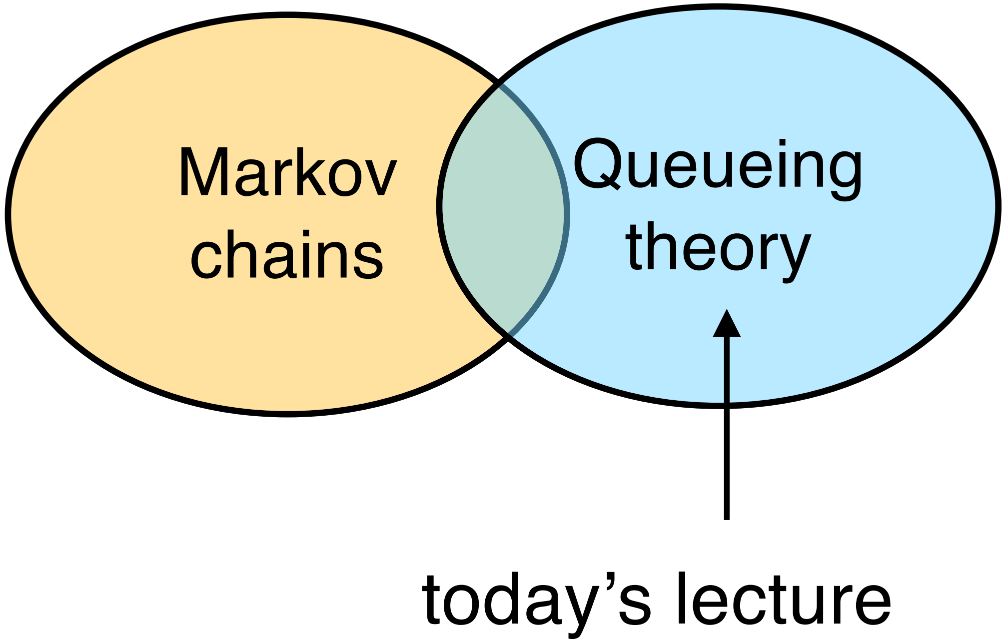
Performance Evaluation of Networks

Sara Alouf

Ch 5 – The General Service Time Queue



- $M / G / 1$ FIFO queue
- $M / G / 1$ FIFO queue with vacations

$M\,/\,G\,/\,1$ FIFO Queue

- Arrivals Poisson process rate λ
- Service time ➜ General distribution and independence
  - Service times are independent identically distributed
  - σ generic service time

$$G(x) = P(\sigma \leq x), \quad x \geq 0$$

$$E[\sigma] = \int_0^\infty (1 - G(x))\, dx = \frac{1}{\mu}$$

$$E[\sigma^2] = \int_0^\infty x^2\, dG(x)$$

$$\mathrm{Var}[\sigma] = E[\sigma^2] - (E[\sigma])^2$$

- First-in-first-out service discipline

$M / G / 1$ FIFO Queue

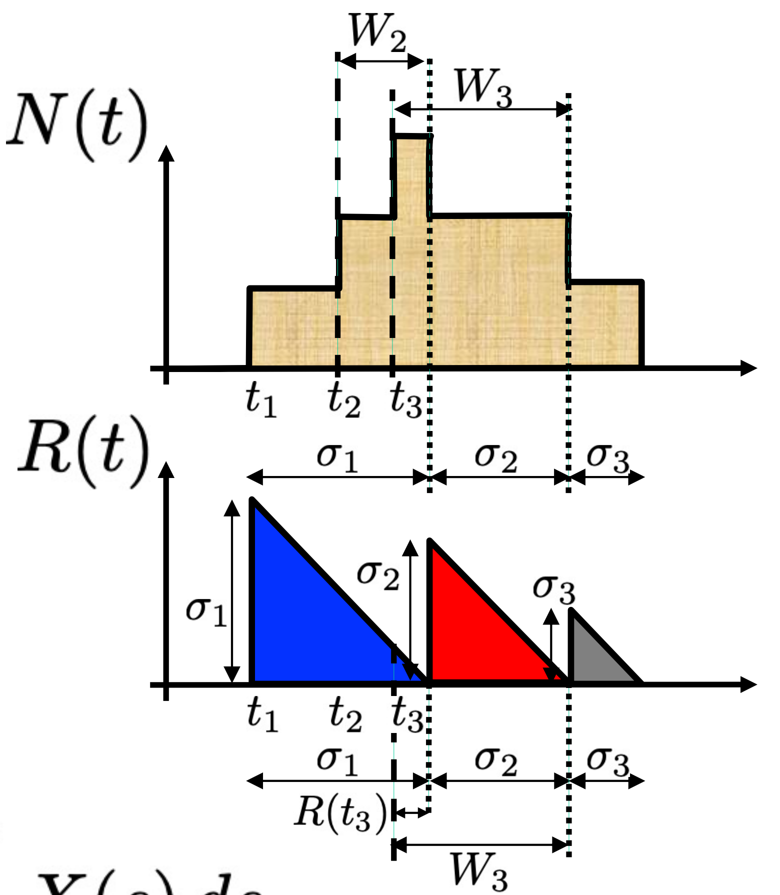
- Load $\rho = \dfrac{\lambda}{\mu}$
- Queue size $N(t)$ ➜ not Markov chain

  sojourn time in a state is not Exp()
- Expected waiting time in steady-state $\overline{W}$
- Pollaczek-Khinchin formula

$$\boxed{\overline{W} = \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)}} = \frac{\rho}{2(1-\rho)} \cdot \frac{\mathrm{Var}(\sigma) + E[\sigma]^2}{E[\sigma]}$$

- Higher service time variability ➜ longer waiting times

Proof of Pollaczek-Khinchin Formula



- For customer i
  - Arrival time t_i
  - Service time σ_i
  - Waiting time W_i
- Number of customers waiting
  - at time t ➜ $X(t)$
  - at time t_i ➜ $X(t_i) = X(t_i^-)$
  - expectation $\overline{X} = \lim_{t\to\infty} \frac{1}{t} \int_0^t X(s)\,ds$
- Residual service time
  - at time t ➜ $R(t)$
  - expectation $\overline{R} = \lim_{t\to\infty} \frac{1}{t} \int_0^t R(s)\,ds$

Proof of Pollaczek-Khinchin Formula

- Customer i sees $X(t_i)$ customers waiting

$$W_i = R(t_i) + \sigma_{i-1} + \sigma_{i-2} + \ldots + \sigma_{i-X(t_i)}$$

$$E[W_i] = E[R(t_i)] + E\left[\sum_{j=1}^{X(t_i)} \sigma_{i-j}\right]$$

- $X(t_i)$ consists of customers $i-1$, …, $i-X(t_i)$

  they have not been served yet

  ➜ $X(t_i)$ independent of their service time

- Use Wald's formula

$$\underbrace{\boxed{E[W_i]}}_{\overline{W}} = \underbrace{\boxed{E[R(t_i)]}}_{??} + \underbrace{\boxed{E[X(t_i)]}}_{??} E[\sigma]$$

Proof of Pollaczek-Khinchin Formula

- Take limit $i \to \infty$ and use PASTA property
- We have

$$\lim_{i\to\infty} E[R(t_i)] = \lim_{t\to\infty} \frac{1}{t}\int_0^t R(s)\,ds = \overline{R}$$

expected residual service time at arrival epochs in steady state

= time average of residual service time

- Similarly

$$\lim_{i\to\infty} E[X(t_i)] = \lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)\,ds = \overline{X}$$

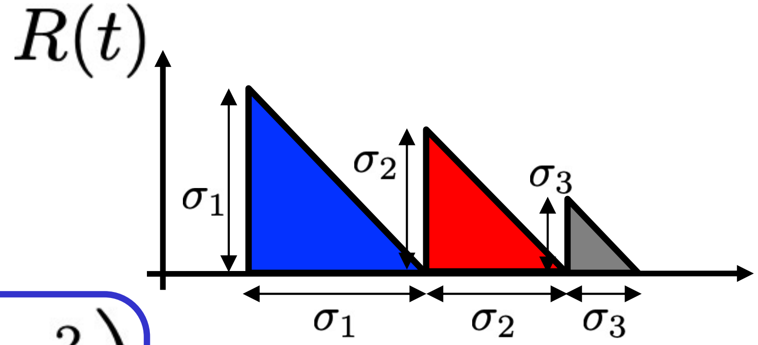
- Therefore

$$E[W_i] = E[R(t_i)] + E[X(t_i)]E[\sigma] \quad \rightarrow \quad \overline{W} = \overline{R} + \frac{\overline{X}}{\mu}$$

Proof of Pollaczek-Khinchin Formula

- If $\rho < 1$ (stability condition of $M / G / 1$)

  ➜ queue empties infinitely often
- Let 0 and C be two times when system is empty
- Let k be number of customers served in (0, C)
- Expected residual service time



$$\overline{R} = \lim_{C\to\infty} \frac{1}{C} \sum_{i=1}^{k} \frac{\sigma_i^2}{2}$$

$$= \boxed{\lim_{\substack{C\to\infty \\ k\to\infty}} \left(\frac{k}{C}\right)} \boxed{\lim_{\substack{C\to\infty \\ k\to\infty}} \left(\frac{1}{k} \sum_{i=1}^{k} \frac{\sigma_i^2}{2}\right)}$$

$$= \lambda \frac{E[\sigma^2]}{2}$$

Proof of Pollaczek-Khinchin Formula

- Apply Little's formula on waiting room

$$\overline{X} = \lambda \overline{W}$$

- Recall $\overline{W} = \overline{R} + \frac{\overline{X}}{\mu} = \overline{R} + \frac{\lambda}{\mu}\overline{W}$

$$\Rightarrow \quad \overline{W}(1-\rho) = \overline{R}$$

$$\Leftrightarrow \quad \overline{W} = \frac{\overline{R}}{1-\rho}$$

$$\Leftrightarrow \quad \boxed{\overline{W} = \frac{\lambda E[\sigma^2]}{2(1-\rho)}}$$

$M\,/\,G\,/\,1$ FIFO Queue

- Expected waiting time

$$\overline{W} = \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)}$$

- Expected sojourn time

$$\overline{T} = \overline{W} + \frac{1}{\mu} = \frac{1}{\mu} + \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)}$$

- Expected number of customers waiting

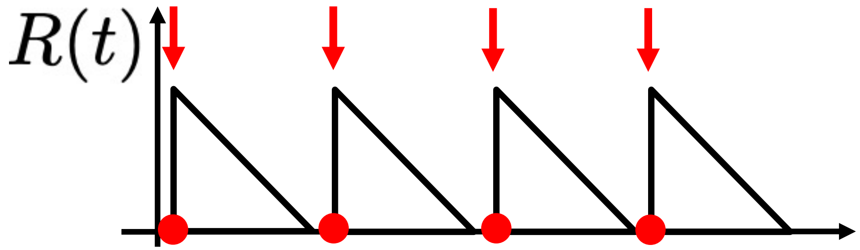
$$\overline{X} = \lambda\,\overline{W} = \frac{\lambda^2\, E[\sigma^2]}{2\,(1-\rho)}$$

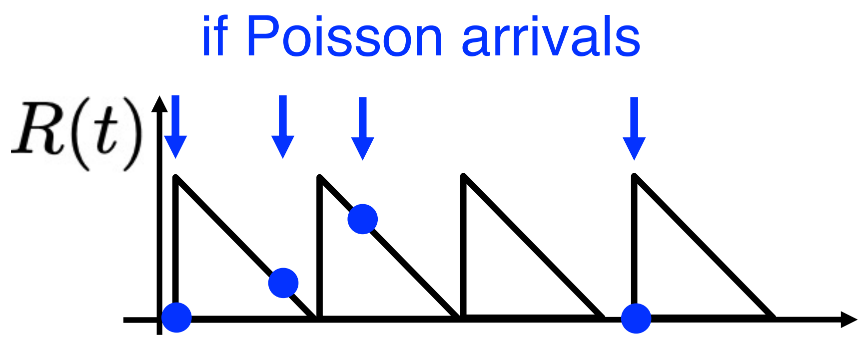
- Expected queue size

$$\overline{N} = \overline{X} + \rho = \lambda \overline{T} = \rho + \frac{\lambda^2\, E[\sigma^2]}{2\,(1-\rho)}$$

Example When PASTA Not True

- Consider $D / D / 1$ FIFO queue
- Arrivals every second ➜ $\lambda = 1$ s^{-1}
- Service time 0.9 second ➜ $\mu = 1/0.9$ s^{-1}
- Load is very high ➜ $\rho = 0.9$





- Time average

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t R(s)\,ds = \lambda\frac{E[\sigma^2]}{2} = \frac{(0.9)^2}{2} = 0.405$$

- Average at arrival epochs $\lim_{i\to\infty} E[R(t_i)] = 0$

$M/M/1$ Versus $M/D/1$ FIFO

- Consider two queues
  - Poisson arrival rate λ
  - One server
  - Infinite waiting room
  - FIFO service discipline
- Different service time distribution but same average $1/\mu$
  - $M/M/1$ queue : Exp(μ) $\quad E[\sigma^2] = \frac{2}{\mu^2}$
  - $M/D/1$ queue : $\sigma = 1/\mu$ $\quad E[\sigma^2] = \frac{1}{\mu^2}$
- Expected waiting time

$$\overline{W}_{M/M/1} = \frac{\lambda E[\sigma^2]}{2(1-\rho)} = \frac{\lambda}{\mu^2(1-\rho)} \qquad \boxed{\overline{W}_{M/D/1} = \frac{\overline{W}_{M/M/1}}{2}}$$

$M\,/\,G\,/\,1$ FIFO Queue With Vacations

- $M\,/\,G\,/\,1$ FIFO but if queue empty: server ➜ vacation
  - Maintenance, background task, sleep mode, power off
- First-in-first-out service discipline
- Arrivals Poisson process rate λ
- Service time ➜ General distribution and independence
  - Service times are independent identically distributed
  - σ generic service time

$$G(x) = P(\sigma \leq x), \quad x \geq 0$$

$$E[\sigma] = \int_0^\infty (1 - G(x))\,dx = \frac{1}{\mu}$$

$$E[\sigma^2] = \int_0^\infty x^2\,dG(x)$$

$M / G / 1$ FIFO Queue With Vacations

- Vacation time ➜ General distribution and independence
  - vacation durations are independent identically distributed
  - V generic vacation duration

$$F(x) = P(V \leq x), \quad x \geq 0$$

$$E[V] = \int_0^\infty (1 - F(x))\, dx$$

$$E[V^2] = \int_0^\infty x^2\, dF(x)$$

- Load $\rho = \frac{\lambda}{\mu}$
- Queue size not Markov chain (sojourn time not Exp())

$M / G / 1$ FIFO Queue With Vacations

- If $\rho < 1$ (stability condition) expected waiting time steady-state

$$\boxed{\overline{W} = \frac{\lambda E[\sigma^2]}{2(1-\rho)} + \frac{E[V^2]}{2E[V]}}$$

$$= \overline{W}_{M/G/1} + \boxed{\frac{E[V^2]}{2E[V]}}$$

- Higher vacations variability ➜ longer waiting times

$$\frac{E[V^2]}{2E[V]} = \frac{\mathrm{Var}[V]}{2E[V]} + \frac{E[V]}{2}$$

- To lessen impact of vacations ➜ V deterministic, small
- If cost to go on vacation ➜ tradeoff to be found

15 minutes break

$M\,/\,G\,/\,1$ FIFO Queue With Vacations

- Expected waiting time in steady-state $\overline{W}$

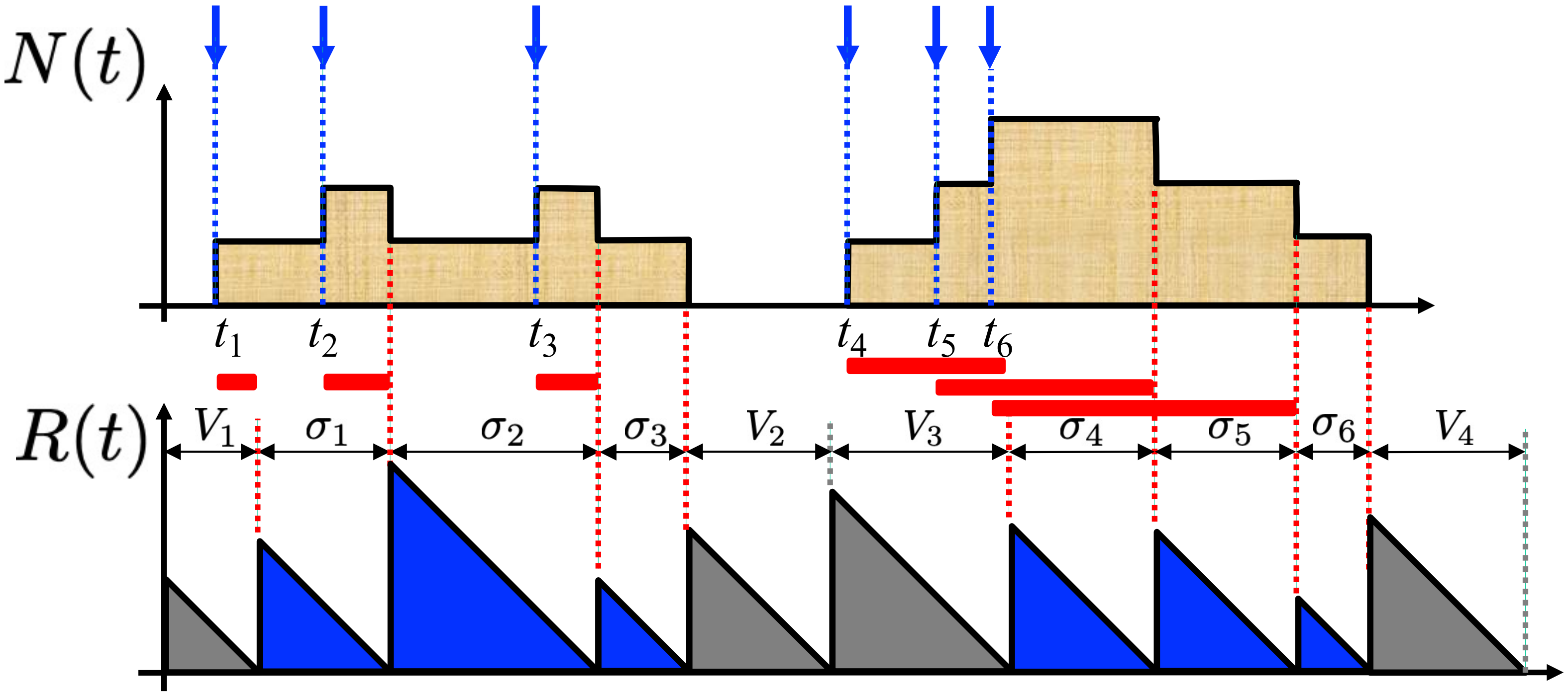
$$\boxed{\overline{W} = \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)} + \frac{E[V^2]}{2E[V]}}$$

Proof

- For customer i
  - Arrival time t_i
  - Service time σ_i
  - Waiting time W_i
- Number of customers waiting
  - at time t ➜ $X(t)$
  - at time t_i ➜ $X(t_i) = X(t_i^-)$
  - expectation $\overline{X} = \lim_{t \to \infty} \frac{1}{t} \int_0^t X(s)\, ds$
- kth server vacation time V_k

Proof

- Residual time at server at time t ➜ $R(t)$
  - if server busy ➜ residual service time
  - if server in vacation ➜ residual vacation time



$$W_6 = R(t_6) + \sigma_4 + \sigma_5$$

Proof

- Customer i sees $X(t_i)$ customers waiting

$$W_i = R(t_i) + \sigma_{i-1} + \sigma_{i-2} + \ldots + \sigma_{i-X(t_i)}$$

$$E[W_i] = E[R(t_i)] + E\left[\sum_{j=1}^{X(t_i)} \sigma_{i-j}\right]$$

- Use Wald's formula ($X(t_i)$ independent of all σ_{i-j})

$$\boxed{E[W_i]} = \boxed{E[R(t_i)]} + \boxed{E[X(t_i)]} E[\sigma]$$

- Take limit $i \to \infty$ and use PASTA property

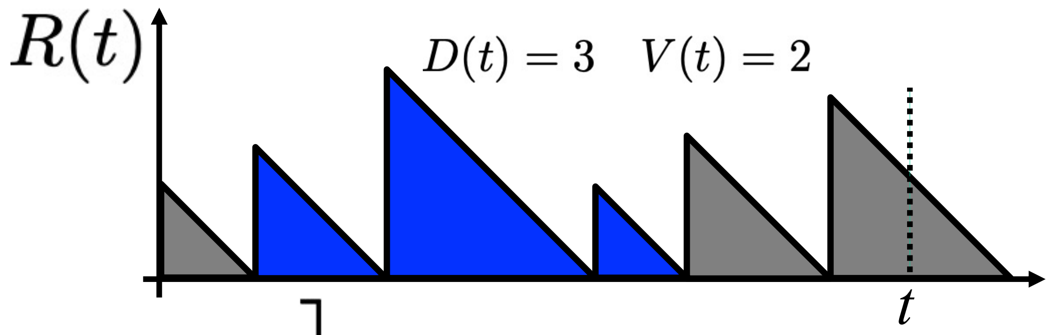
$$\left.\begin{aligned} \overline{W} &= \overline{R} + \frac{\overline{X}}{\mu} \\ \overline{X} &= \lambda \overline{W} \end{aligned}\right\} \Rightarrow \boxed{\overline{W} = \frac{\overline{R}}{1-\rho}}$$

- By Little's formula $\overline{X} = \lambda \overline{W}$

same expression
different
$\overline{R}$

Proof: Find $\overline{R}$

- $\rho < 1$ ➜ queue will empty infinitely often
- $D(t)$ ➜ number of customers fully served in $(0, t)$
- $V(t)$ ➜ number of complete vacations in $(0, t)$
- Expected residual time



$$\overline{R} = \lim_{t\to\infty} \frac{1}{t} \int_0^t R(u)du$$

$$= \lim_{t\to\infty} \frac{1}{t} \left[\sum_{i=1}^{D(t)} \frac{\sigma_i^2}{2} + \sum_{k=1}^{V(t)} \frac{V_k^2}{2} + \text{trapezoid} \right]$$

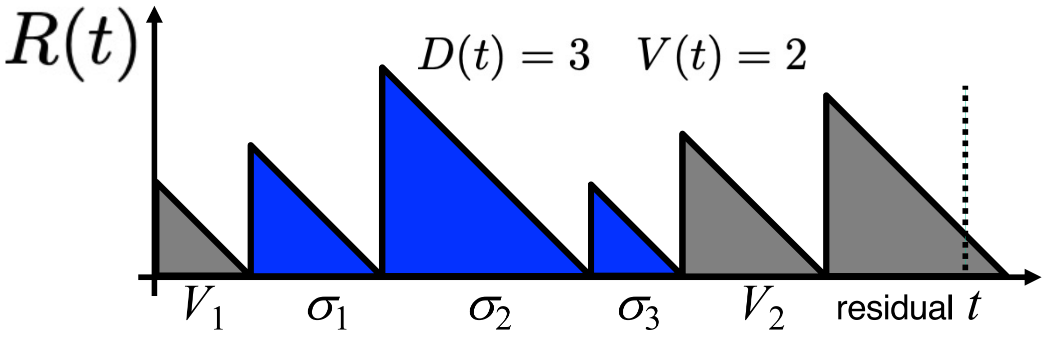
$$= \lim_{t\to\infty} \left[\boxed{\frac{D(t)}{t}} \boxed{\frac{1}{D(t)} \sum_{i=1}^{D(t)} \frac{\sigma_i^2}{2}} + \frac{V(t)}{t} \boxed{\frac{1}{V(t)} \sum_{k=1}^{V(t)} \frac{V_k^2}{2}} + \cancel{\frac{\text{trapezoid}}{t}} \right]$$

$$= \lambda \frac{E[\sigma^2]}{2} + \boxed{\left(\lim_{t\to\infty} \frac{V(t)}{t} \right)} \frac{E[V^2]}{2}$$

?

Proof: Find $\lim_{t\to\infty} \frac{V(t)}{t}$

- We have



$$t = \sum_{i=1}^{D(t)} \sigma_i + \sum_{k=1}^{V(t)} V_k + \text{residual}$$

$$1 = \boxed{\frac{D(t)}{t}} \boxed{\frac{1}{D(t)} \sum_{i=1}^{D(t)} \sigma_i} + \frac{V(t)}{t} \boxed{\frac{1}{V(t)} \sum_{k=1}^{V(t)} V_k} + \cancel{\frac{\text{residual}}{t}}$$

take
limit
$t \to \infty$

$$1 = \lambda E[\sigma] + \boxed{\left(\lim_{t\to\infty} \frac{V(t)}{t} \right)} E[V]$$

$$\Rightarrow \quad \lim_{t\to\infty} \frac{V(t)}{t} = \frac{1-\rho}{E[V]}$$

Proof: Recap

- $$\lim_{t\to\infty} \frac{V(t)}{t} = \frac{1-\rho}{E[V]}$$

- $$\overline{R} = \lambda \frac{E[\sigma^2]}{2} + \left(\lim_{t\to\infty} \frac{V(t)}{t} \right) \frac{E[V^2]}{2}$$
  $$= \lambda \frac{E[\sigma^2]}{2} + \left(\frac{1-\rho}{E[V]} \right) \frac{E[V^2]}{2}$$

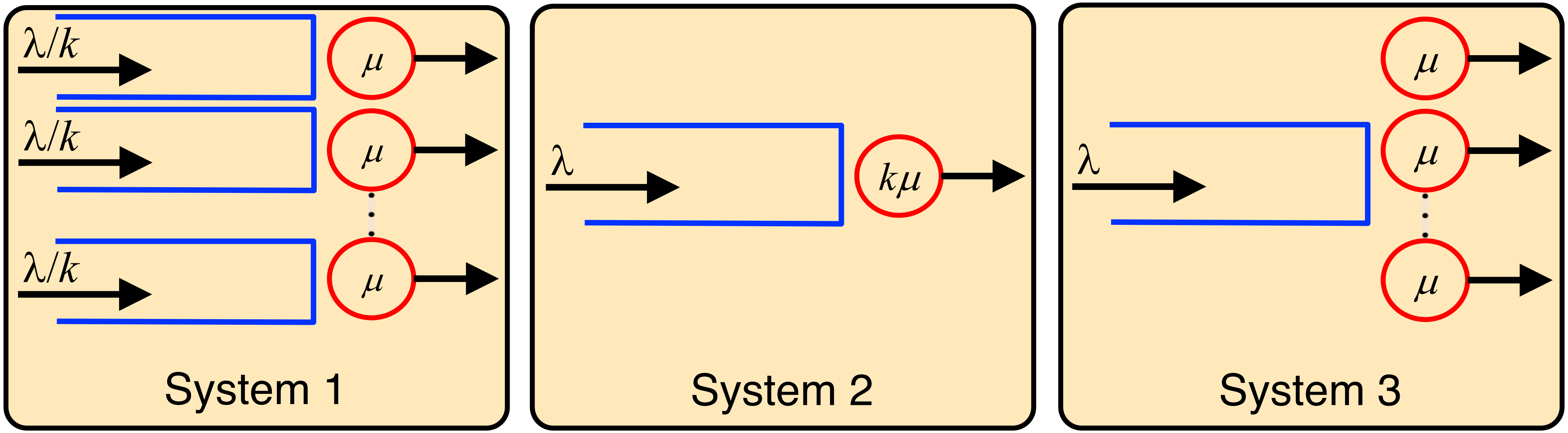
- $$\overline{W} = \frac{\overline{R}}{1-\rho}$$
  $$= \frac{\lambda E[\sigma^2]}{2(1-\rho)} + \frac{E[V^2]}{2E[V]} \quad ✔$$
  $= \overline{W}_{M/G/1} +$ effect of vacations

Exercize: Compare Different Organizations

- Global traffic Poisson rate λ
- Total service rate $k\mu$

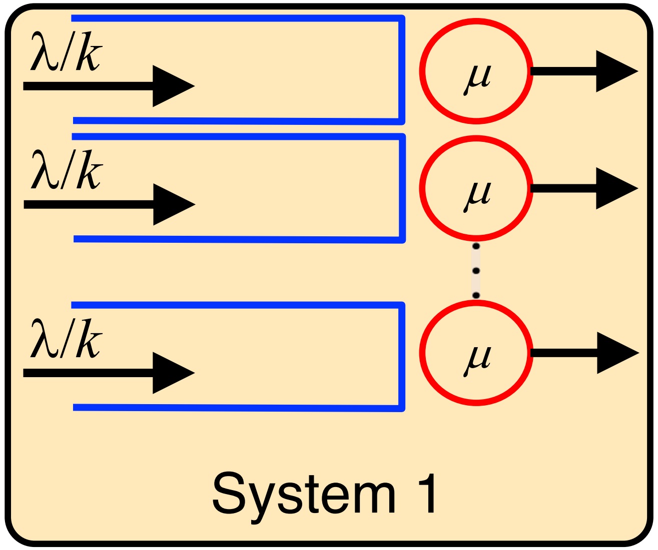
$$\rho = \frac{\lambda}{k\mu}$$

- Service time is exponentially distributed
- Objective: compare three systems organizations



- Order systems according to expected sojourn time

System 1

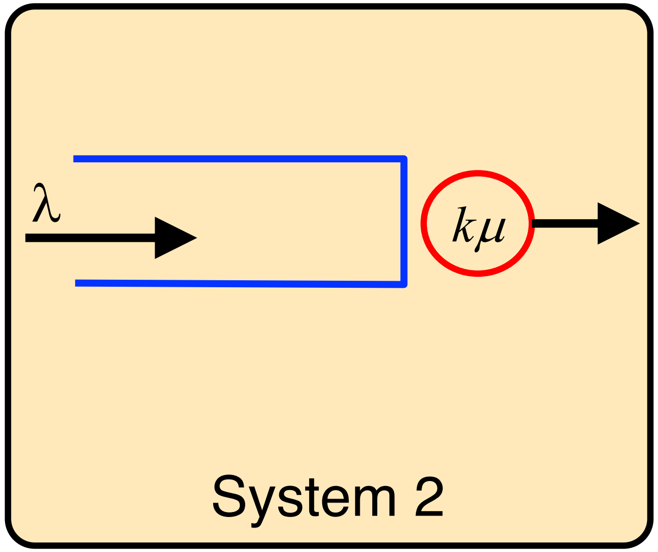


- Each queue is $M / M / 1$ queue
- Arrival rate λ/k
- Service rate μ
- Infinite queue: stability condition $\lambda/k < \mu$ $(\rho < 1)$
- Expected queue size in one queue $\dfrac{\lambda/k}{\mu - \lambda/k}$
- Expected number of customers in System 1

$$\overline{N}_1 = \frac{\lambda}{\mu - \lambda/k} = \frac{k\rho}{1-\rho}$$

- By Little's fomula $$\overline{T}_1 = \frac{\overline{N}_1}{\lambda} = \frac{1}{\mu - \lambda/k} = \frac{k}{k\mu - \lambda}$$

System 2



- Queue is $M / M / 1$ queue
- Arrival rate λ
- Service rate $k\mu$
- Infinite queue: stability condition $\lambda < k\mu$ $(\rho < 1)$

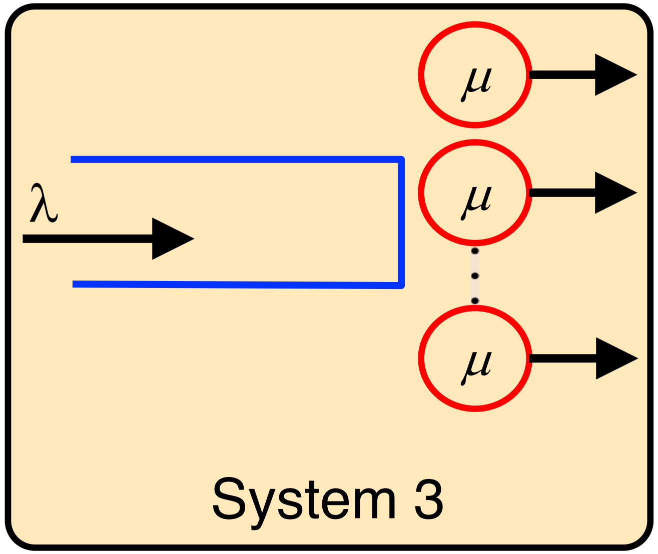
- Expected number of customers in System 2

$$\overline{N}_2 = \frac{\lambda}{k\mu - \lambda} = \frac{\rho}{1 - \rho}$$

- By Little's fomula

$$\overline{T}_2 = \frac{\overline{N}_2}{\lambda} = \frac{1}{k\mu - \lambda}$$

System 3



- Queue is $M/M/k$ queue
- Arrival rate λ
- Service rate μ
- Infinite queue: stability condition $\lambda < k\mu \ \ (\rho < 1)$
- Expected number of customers in System 3

$$\overline{N}_3 = \overline{N}_{\text{wait}} + \overline{N}_{\text{servers}}$$

- By Little's formula on all servers

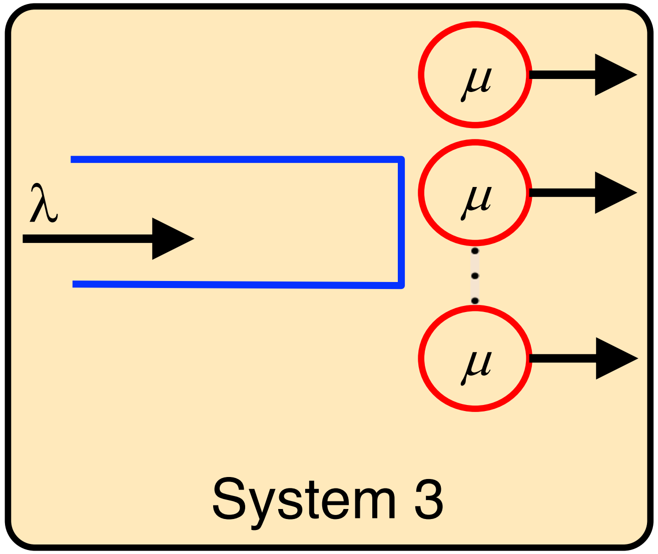
$$\overline{N}_{\text{servers}} = \lambda \frac{1}{\mu}$$

- Let X_{wait} number of customers waiting in queue

$$\begin{aligned} \overline{N}_{\text{wait}} &= E[X_{\text{wait}}] \\ &= E[X_{\text{wait}}|\text{wait}]P_{\text{wait}} + \cancel{E[X_{\text{wait}}|\text{no wait}]}(1 - P_{\text{wait}}) \end{aligned}$$

law of total probabilities

System 3



Conditioning on fact all servers busy
X_{wait} same as queue size in $M / M / 1$
queue with arrival rate λ service rate $k\mu$

$$E[X_{\text{wait}}|\text{wait}] = \overline{N}_2$$

- All servers are busy with probability (see lecture 4)

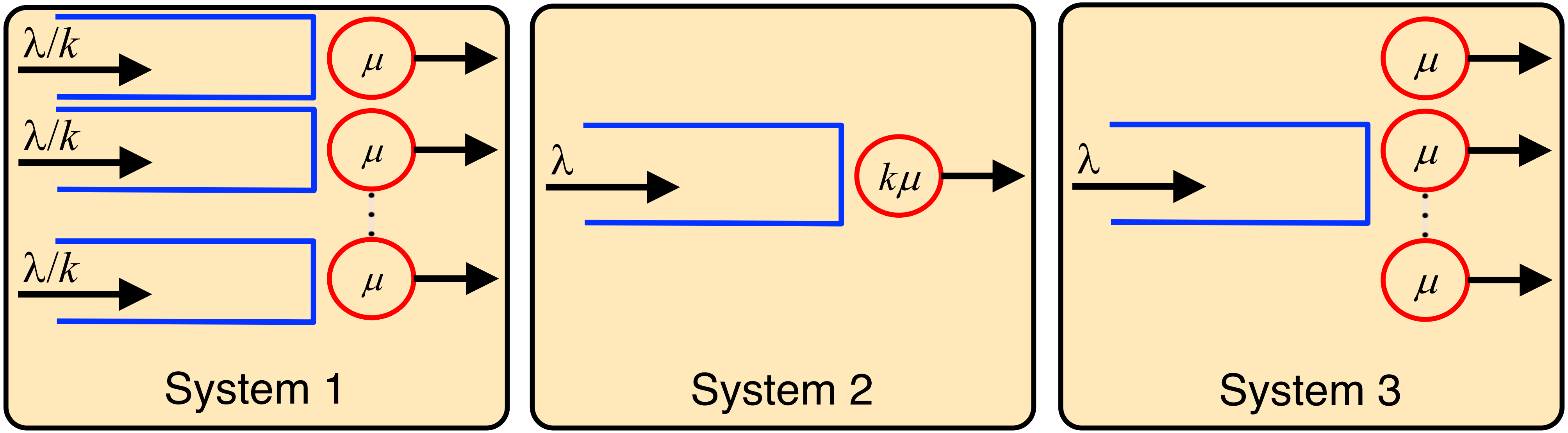
$$P_{\text{wait}} = \frac{\pi_0 (k\rho)^k}{k!(1-\rho)} = \frac{\frac{(k\rho)^k}{k!(1-\rho)}}{\sum_{i=0}^{k-1} \frac{(k\rho)^i}{i!} + \frac{(k\rho)^k}{k!(1-\rho)}}$$

- Expected number of customers in System 3

$$\overline{N}_3 = \frac{\lambda}{k\mu - \lambda} P_{\text{wait}} + \frac{\lambda}{\mu}$$

- By Little's fomula $\overline{T}_3 = \frac{1}{k\mu - \lambda} P_{\text{wait}} + \frac{1}{\mu}$

Exercize: Compare Different Organizations

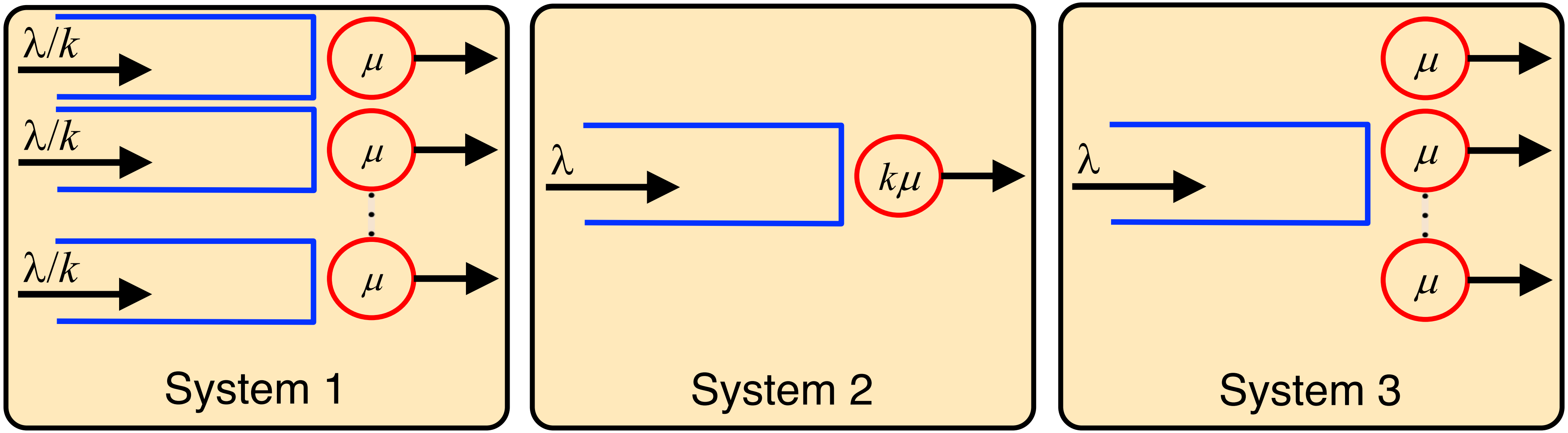


$$\overline{N}_1 = \frac{k\rho}{1-\rho} \qquad \overline{N}_2 = \frac{\rho}{1-\rho} \qquad \overline{N}_3 = \overline{N}_2 P_{\text{wait}} + \frac{\lambda}{\mu}$$

$$\overline{T}_1 = \frac{k}{k\mu - \lambda} \qquad \overline{T}_2 = \frac{1}{k\mu - \lambda} \qquad \overline{T}_3 = \overline{T}_2 P_{\text{wait}} + \frac{1}{\mu}$$

- System 2 is k times better than System 1
- What about System 3? $\dfrac{\overline{T}_3}{\overline{T}_2} = P_{\text{wait}} + k(1-\rho) > 1$

Exercize: Compare Different Organizations



- What about System 3? $$\frac{\overline{T}_3}{\overline{T}_2} = P_{\text{wait}} + k(1-\rho) > 1$$
- If very low utilization ➜ ratio close to k

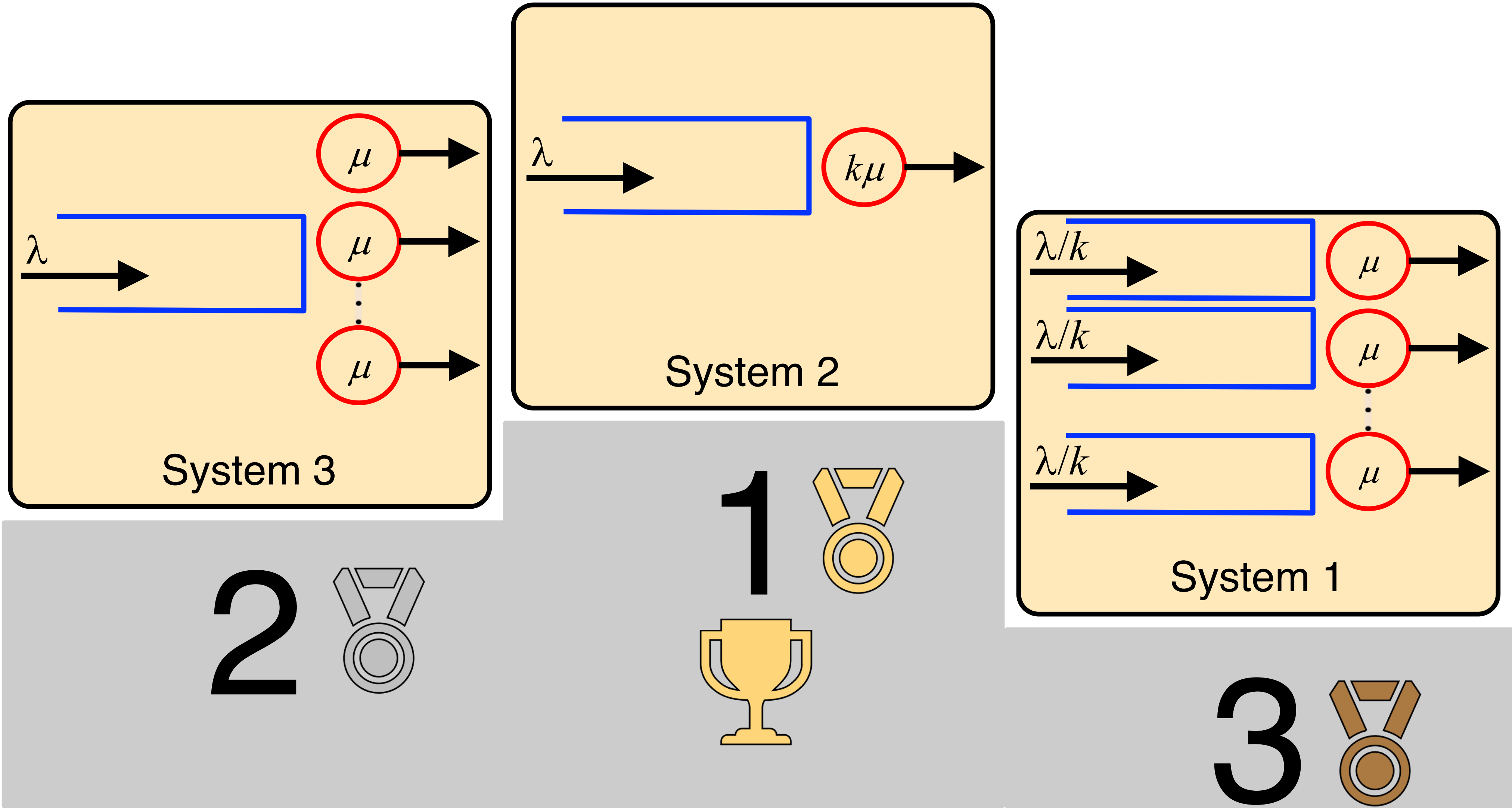
  System 3 almost k times worse than System 2

  ➜ System 3 slightly better than System 1
- If very high utilization ➜ ratio close to 1

  System 3 almost same (slightly worse) as System 2

Conclusion $\overline{T}_2 < \overline{T}_3 < \overline{T}_1$

- … and the winner is: System 2!



For next week

- Lesson 5 to revise

- Homework 5 to return on Tuesday 15 October before 9 am

- Lesson 6 to read before Lecture 6