Lecture 5

Intuitive Solutions for Simple Models

M/M/1 Queue

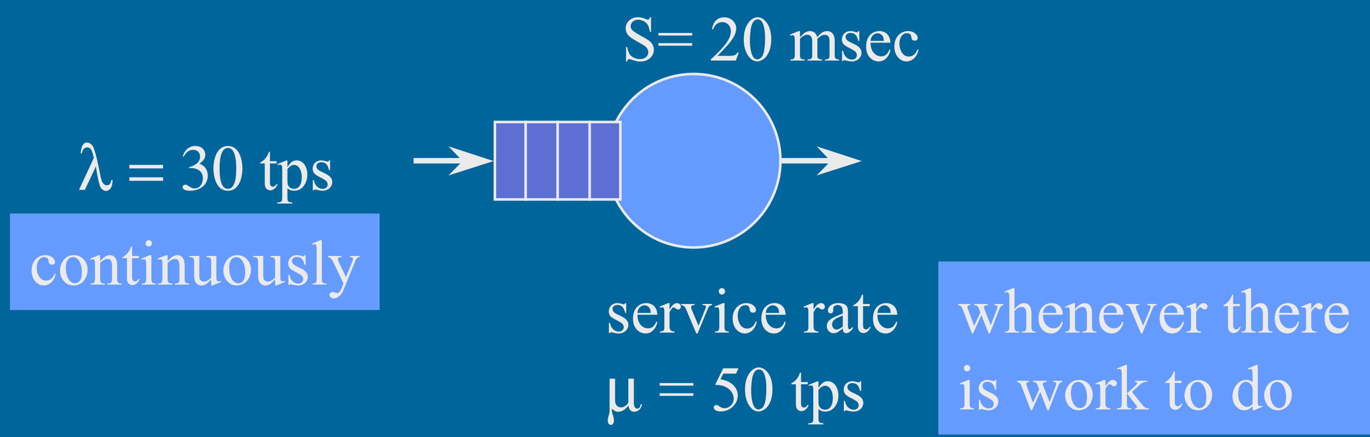
Markov Chains

Little's Law

Solution Methods in Overall Picture

- Baseline model
- Prediction model
- Fig. 4.1

Single Server System

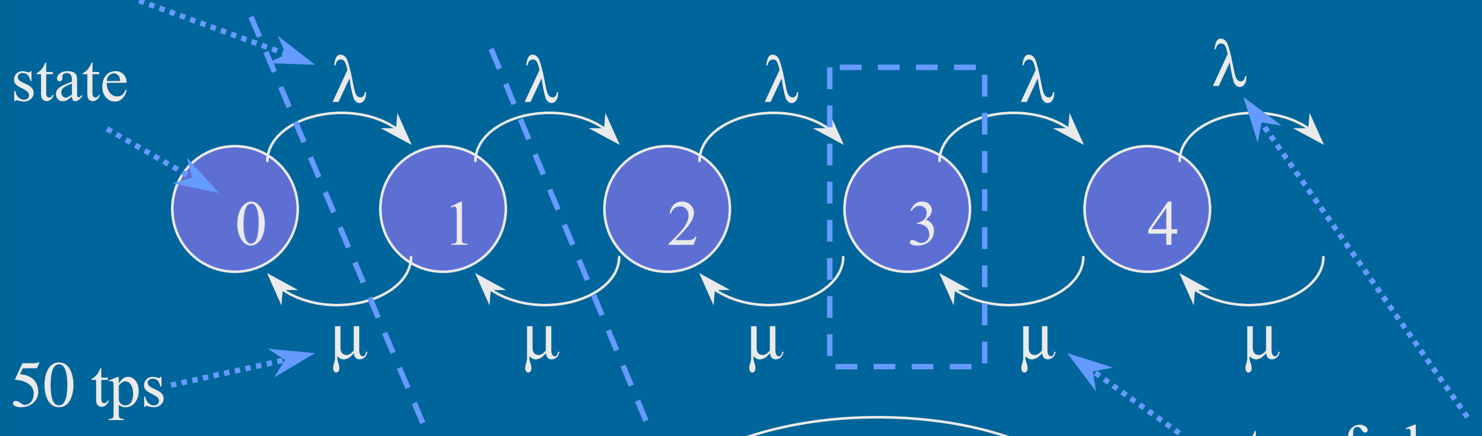


- Server utilization U?
- Server queue length Q?
- Server response time R?
- Server & system throughput X?

Birth-Death System

- Markov chain (history is irrelevant)
- Birth-death (change to adjacent states only)
- Stochastic process (randomness)
- Fig. 4.3
- System states, nr of jobs (k) in system
  $k = 0,1,2, \ldots$
- Need: System statistics
  - average U, R, X, Q
- Can compute them from
  state probabilities P_k $\forall k=0,1,2, \ldots$

Flow Balance



$$\mu P_1 = \lambda P_0$$
$$\mu P_2 = \lambda P_1$$
$$\ldots$$
$$\mu P_k = \lambda P_{k-1}$$
$$\ldots$$

$$P_1 = \lambda/\mu \, P_0$$
$$P_2 = \lambda/\mu \, P_1$$
$$\ldots$$
$$P_k = \lambda/\mu \, P_{k-1}$$
$$\ldots$$

normalizing constant

$$P_0+P_1+P_2+\ldots = 1.0$$

Single Server Solution (2)

$P_1 = \lambda/\mu \; P_0$
$P_2 = \lambda/\mu \; P_1$
...
$P_k = \lambda/\mu \; P_{k-1}$
...

$P_0 + P_1 + P_2 + \ldots = 1.0$

$$\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i P_0 = P_0 \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i = 1$$

$$P_0 = \frac{1}{\sum_0^{\infty} \left(\frac{\lambda}{\mu}\right)^i} = \frac{1}{\frac{1}{1-\frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu}$$

$$P_k = \left(\frac{\lambda}{\mu}\right)^k P_0 = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right)$$

Single Server Solution (contd) (1)

$$P_0 = \frac{1}{\sum_0^{\infty} \left(\frac{\lambda}{\mu}\right)^i} = \frac{1}{\frac{1}{1-\frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu}$$

$$P_k = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right)$$

λ = 30, μ = 50

$$P_0 = 1 - \frac{30}{50} = 0.4 = \boxed{40\%}$$

$$U = 1 - P_0 = \boxed{60\%}$$

(= λ/μ)

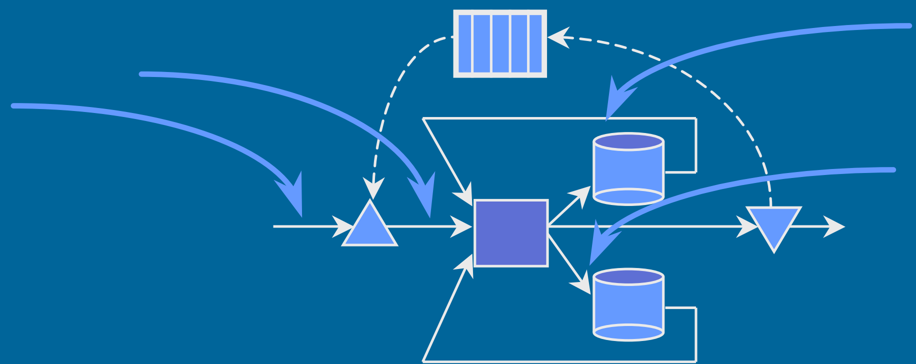
$$X = \sum_{i=0}^{\infty} \mu_i P_i = \sum_{i=1}^{\infty} \mu_i P_i = \mu \sum_{i=1}^{\infty} P_i$$

$$= \mu U = \mu \frac{\lambda}{\mu} = \lambda = \boxed{30 tps}$$

aver. popul?
R?

Arrival Theorem

- Arriving customer to some node (I.e., a customer <u>in transition</u>) sees the system as steady state system with itself removed
  - infinite population:
    overall steady state
  - finite population:
    steady state for system with one job less



Single Server Solution (contd)

number of customers in system (aver. population)

$$\overline{N} = \sum_{0}^{\infty} iP_i = \sum_{1}^{\infty} iP_i = \sum_{1}^{\infty} i\left(\frac{\lambda}{\mu}\right)^i \left(1-\frac{\lambda}{\mu}\right)$$

$$= \left(1-\frac{\lambda}{\mu}\right)\frac{\frac{\lambda}{\mu}}{\left(1-\frac{\lambda}{\mu}\right)^2} = \frac{\lambda}{\mu-\lambda} = \frac{30}{50-30} = 1.5$$

response time = time in queue + time in service

$$R = \boxed{\overline{N}\frac{1}{\mu}} + \boxed{\frac{1}{\mu}} = \frac{\lambda}{\mu-\lambda}\frac{1}{\mu} + \frac{1}{\mu} = \frac{1}{\mu-\lambda} = \frac{1}{20} = 0.05\,\text{sec}$$

arrival theorem → nr of jobs in front of ”me”

Single Server Solution (contd)

aver nr of jobs in queue (not in service)
= nr of jobs in system - utilization

$$\overline{N} - U = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = 1.5 - 0.6 = 0.9$$

Use Solution for Predictions

(IAT=22 ms, S=20 ms)
$\lambda = 45, \mu = 50$

$$U = \frac{\lambda}{\mu} = 90\%$$

$$X = \lambda = 45 tps$$

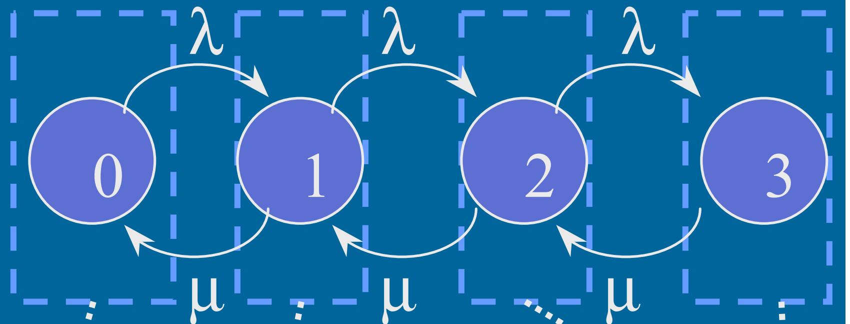
$$\overline{N} = \frac{\lambda}{\mu - \lambda} = 9$$

$$R = \frac{1}{\mu - \lambda} = \frac{1}{5} = 0.2 \sec$$

(IAT=22 ms, S=13 ms)
$\lambda = 45, \mu = 75$

(λ, μ)	(30,50)	(45,50)	(45,75)
U	60%	90%	60%
X	30	45	45 tps
$\overline{N}$	1.5	9	1.5
R	0.05	0.2	0.033 sec

Limited buffer example (1)



- Limited buffer size: 3
  - if packet (job,transaction) arrives and buffer full, then that packet is lost
  - flow balance:

$$
\begin{aligned}
&\mu P_1 = \lambda P_0 \\
&\lambda P_0 + \mu P_2 = \lambda P_1 + \mu P_1 \\
&\lambda P_1 + \mu P_3 = \lambda P_2 + \mu P_2 \\
&\mu P_2 = \lambda P_3
\end{aligned}
$$

Limited buffer (contd) (10)

$$\mu P_1 = \lambda P_0$$
$$\lambda P_0 + \mu P_2 = \lambda P_1 + \mu P_1$$
$$\lambda P_1 + \mu P_3 = \lambda P_2 + \mu P_2$$
$$\mu P_2 = \lambda P_3$$

$\lambda = 30$ tps, $\mu = 50$ tps
$\lambda/\mu = 0.6$

$$P_0 + P_1 + P_2 + P_3 = 1.0$$

$$\cancel{\lambda P_0} + \mu P_2 = \lambda (\lambda/\mu\ P_0) + \cancel{\cancel{\mu} (\lambda/\cancel{\mu})\ P_0}$$

$$P_1 = \lambda/\mu\ P_0$$
$$P_2 = (\lambda/\mu)^2\ P_0$$
$$P_3 = (\lambda/\mu)^3\ P_0$$

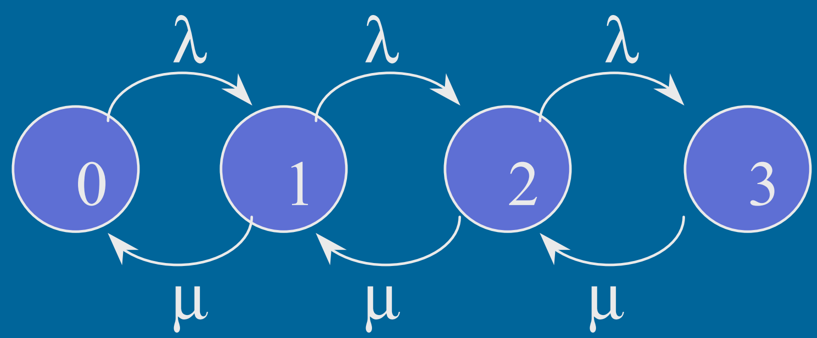
$$P_0 + 0.6\ P_0 + 0.6^2\ P_0 + 0.6^3\ P_0 = 1.0$$

$$(1 + 0.6 + 0.36 + 0.216)\ P_0 = 1.0$$

$$2.176\ P_0 = 1.0$$

$$P_0 = 0.46,\ P_1 = 0.275\ ,\ P_2 = 0.165,\ P_3 = 0.10$$

Limited buffer (contd) (6)



$P_0 = 0.46$, $P_1 = 0.275$, $P_2 = 0.165$, $P_3 = 0.10$

utilization $= 1 - P_0 = 0.54 = U$

throughput $= 0 * P_0 + \mu P_1 + \mu P_2 + \mu P_3$
$= \mu (P_1 + P_2 + P_3) = 50 * 0.54 = 27$ tps $= X$

average population $= 0 * P_0 + 1 P_1 + 2 P_2 + 3 P_3$
$= 0.275 + 2 * 0.165 + 3 * 0.10 = 0.91 = \bar{N}$

average queue length = aver pop – util = 0.91 – 0.54 = 0.35

response time $= (1/\mu)\bar{N} + (1/\mu) = 0.02 * 0.91 + 0.02 = 0.4 = R$

loss rate $= \lambda P_3 = 30 * 0.10 = 3$ tps

Prob(loss) $= P_3 = 0.10$

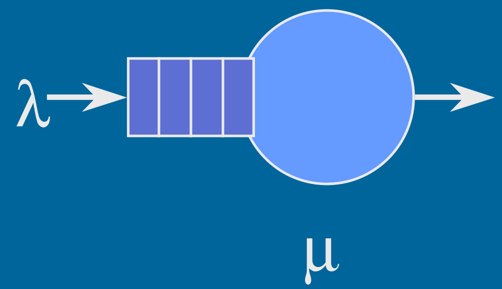
Queue length vs. population

- Two terminologies
- ”Queue length” = population

  Menasce
  - queue includes those in service
- Population N = queue + those in service
  - queue includes only those <u>not</u> in service
  - N = queue length + U

  LZGS
- No big problem – watch out.

M/M/1 Queue



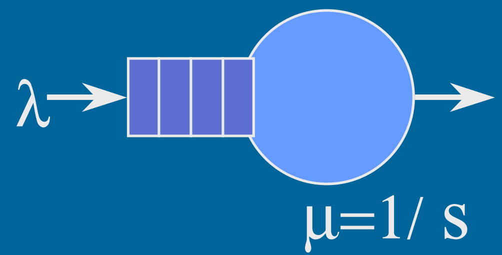
one server

Exponential death process: $\mu_i = \mu$

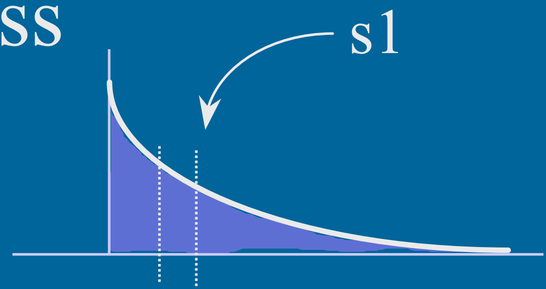
Exponential birth process: $\lambda_i = \lambda / \mu$

- Single server example = M/M/1
- M/M/2 M/M/m M/M/∞
- M/M/1/B M/M/m/B (finite buffer B)
- M/M/1/B/K (finite customer population K)
- M/G/1 G/G/1 M/G/m/B/K

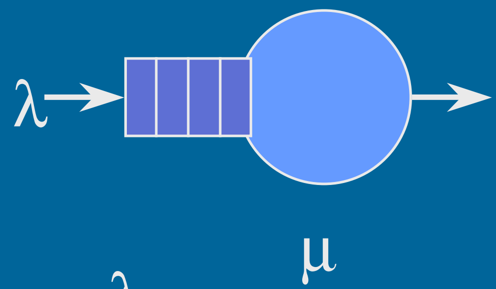
M/M/1 Queue

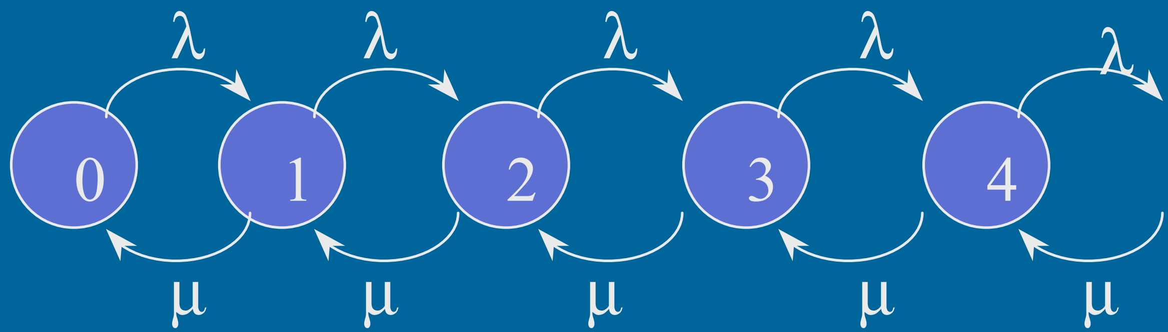


- Markovian birth-death process
- Exponential distributions
  - birth & death
  - memoryless: $\Pr(S \geq s) = \Pr(S^{rem} \geq s \mid S^{serv}{=}s1)$
- Infinite queue (line, buffer) space
- First-in-first-out queueing discipline
- Prob(empty) = $P_0 > 0$ (true when $\mu > \lambda$)



M/M/1 Queue



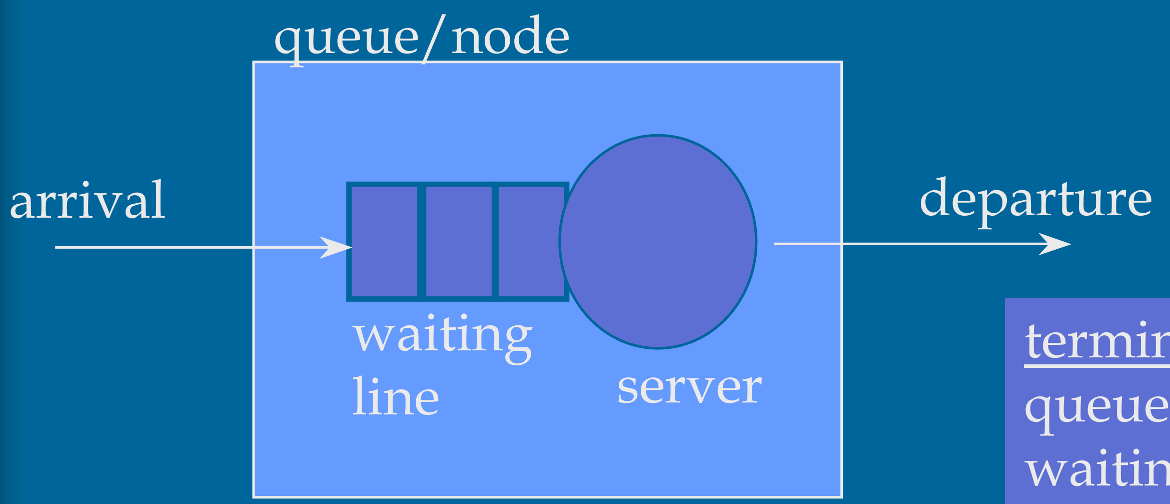


- Transition probability matrix **Q** $\mathbf{Q}[1,2] = \lambda / (\lambda + \mu)$
- Ergodicity: recurrent, aperiodic states if $\lambda < \mu$
- Probability vector of being in state at time t: **p**(t)
- Stable limit probability $\lim_{t \to \infty} \mathbf{p}(t) = \Pi$ $\Pi_i = P_i$
- Steady state solution $\Pi = \Pi\mathbf{Q}$
- Rephrase question: How to find such Π that $\Pi = \Pi\mathbf{Q}$?

Box 31.1 [Jain 91]

Box 31.4 [Jain 91]

Little's Law: N = XR (1)



<u>terminology</u>
queue = node = server?
waiting line = queue?

Mean Number of Customers

= Mean Throughput * Mean Time in Queue

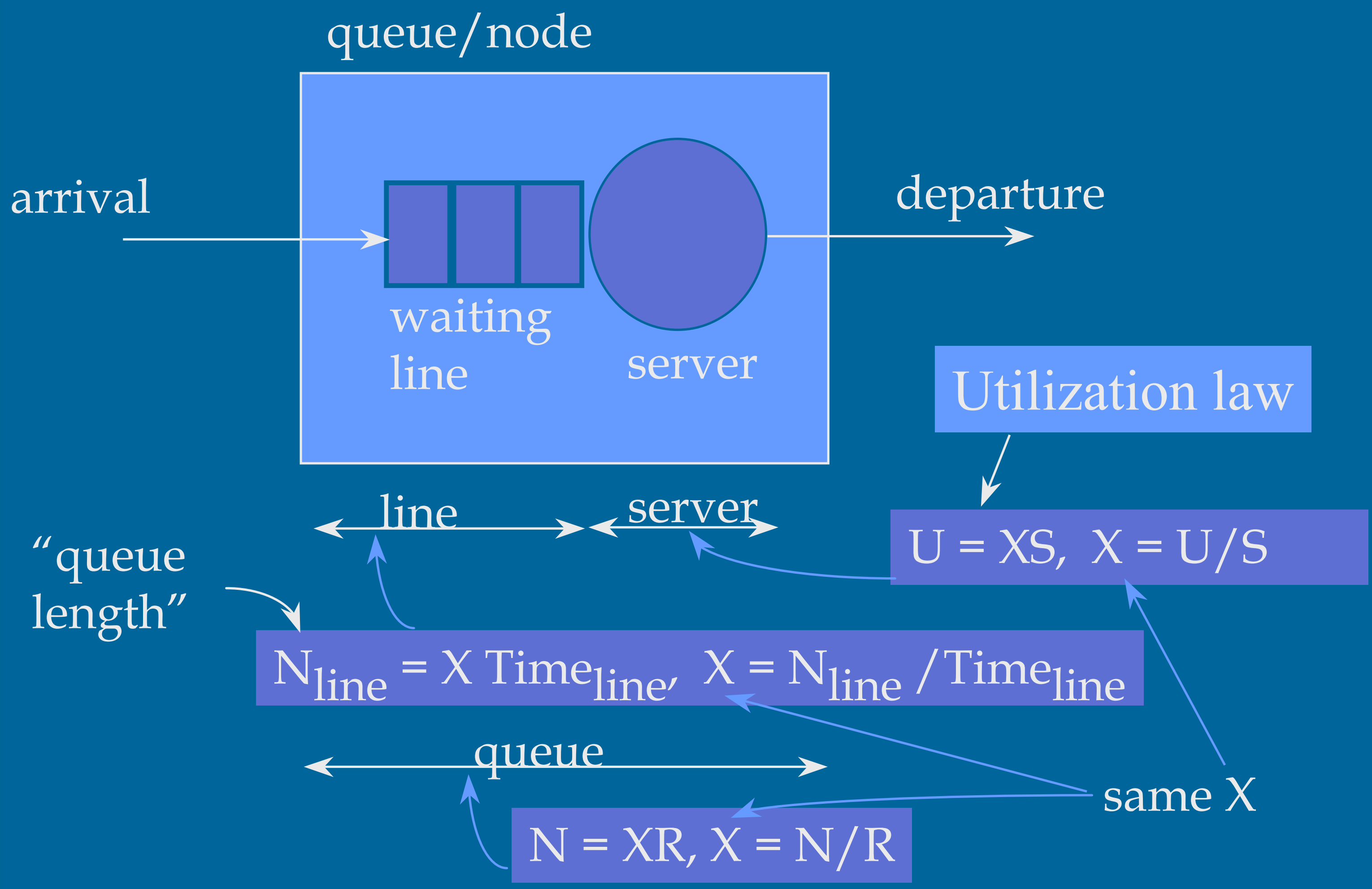
in waiting line *or*
in service

M/M/1: $\overline{N} = \dfrac{\lambda}{\mu - \lambda} = 1.5$

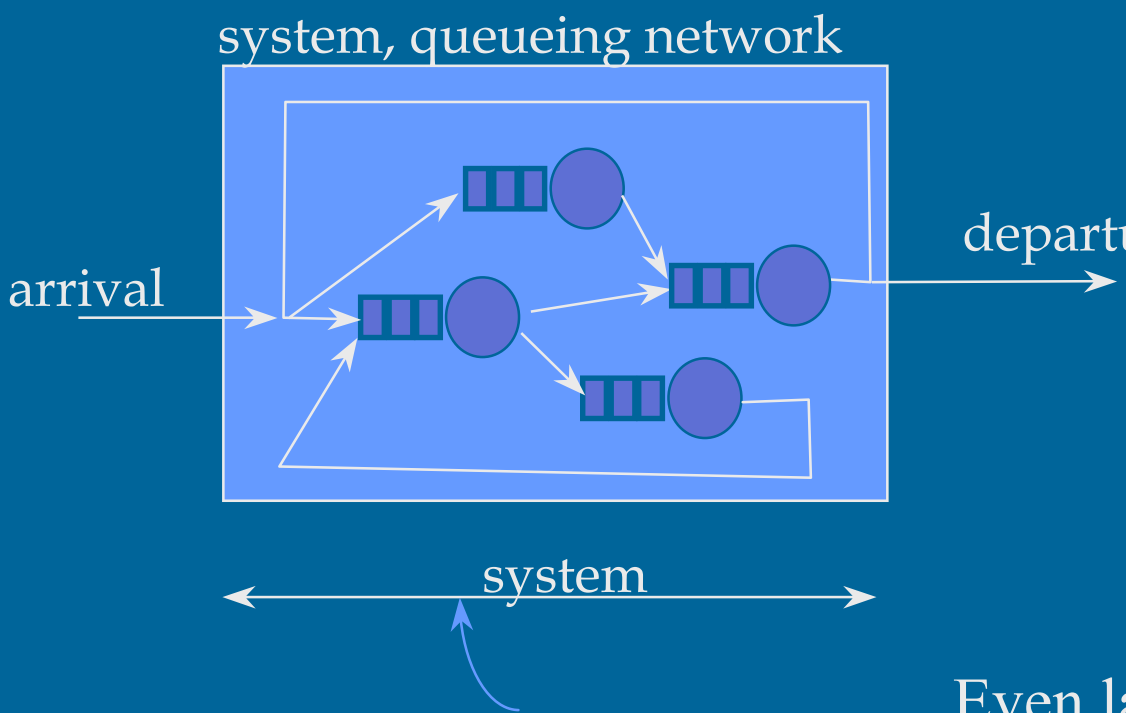
$X = \lambda = 30\, tps$

$R = \dfrac{1}{\mu - \lambda} = 0.05\,\text{sec}$

Little's Law: N = XR
Apply to Queue, Server, or Line



Little's Law: $N = XR$
Apply to Whole System (1)



$N = N_{system} = X_{system}R_{system} = X_0R$
$X_0 = N/R$

Even larger system?
Include human users in "system".