More Queueing Theory

Carey Williamson
Department of Computer Science
University of Calgary

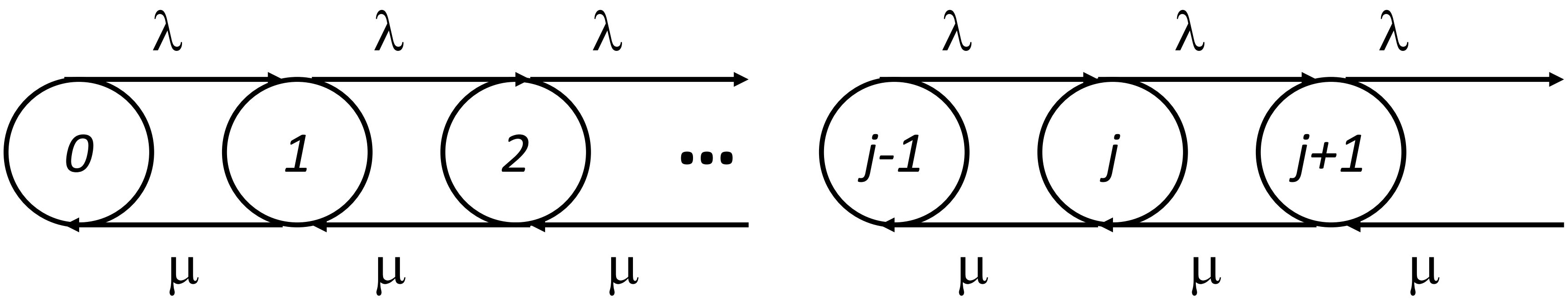
Motivating Quote for Queueing Models

"Good things come to those who wait"

- poet/writer Violet Fane, 1892
- song lyrics by Nayobe, 1984
- motto for Heinz Ketchup, USA, 1980's
- slogan for Guinness stout, UK, 1990's

M/M/1 Queue

- M/M/1 queue is the most commonly used type of queueing model
- Used to model single processor systems or to model individual devices in a computer system
- Need to know only the mean arrival rate λ and the mean service rate μ
- State = number of jobs in the system



Results for M/M/1 Queue (cont'd)

- Mean number of jobs in the system:

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

- Mean number of jobs in the queue:

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

Results for M/M/1 Queue (cont'd)

- Probability of n or more jobs in the system:

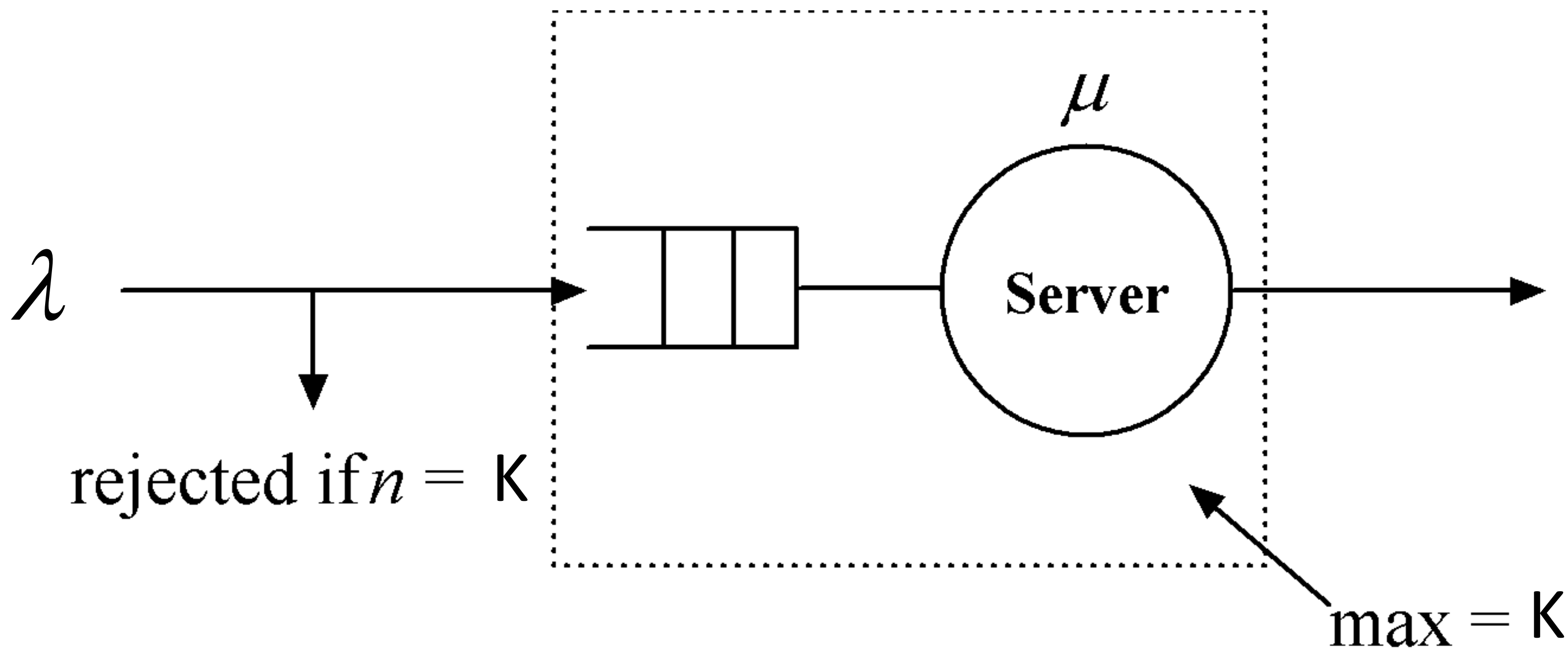
$$P(n \geq k) = \sum_{n=k}^{\infty} p_n = \sum_{n=k}^{\infty} (1-\rho)\rho^n = \rho^k$$

- Mean response time (using Little's Law):
  - Mean number in the system
    = Arrival rate $\times$ Mean response time
  - That is:

$$E[n] = \lambda E[r]$$

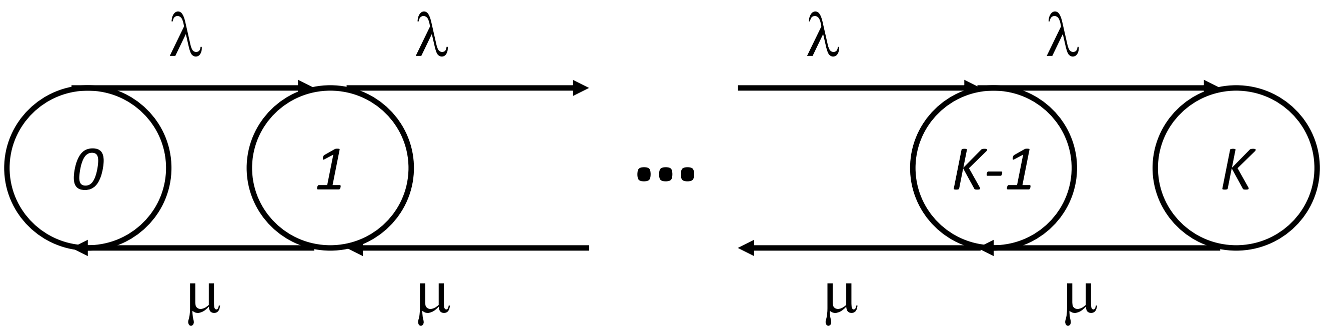
$$E[r] = \frac{E[n]}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{1/\mu}{1-\rho}$$

M/M/1/K – Single Server, Finite Queuing Space



Analytic Results

- State-transition diagram:

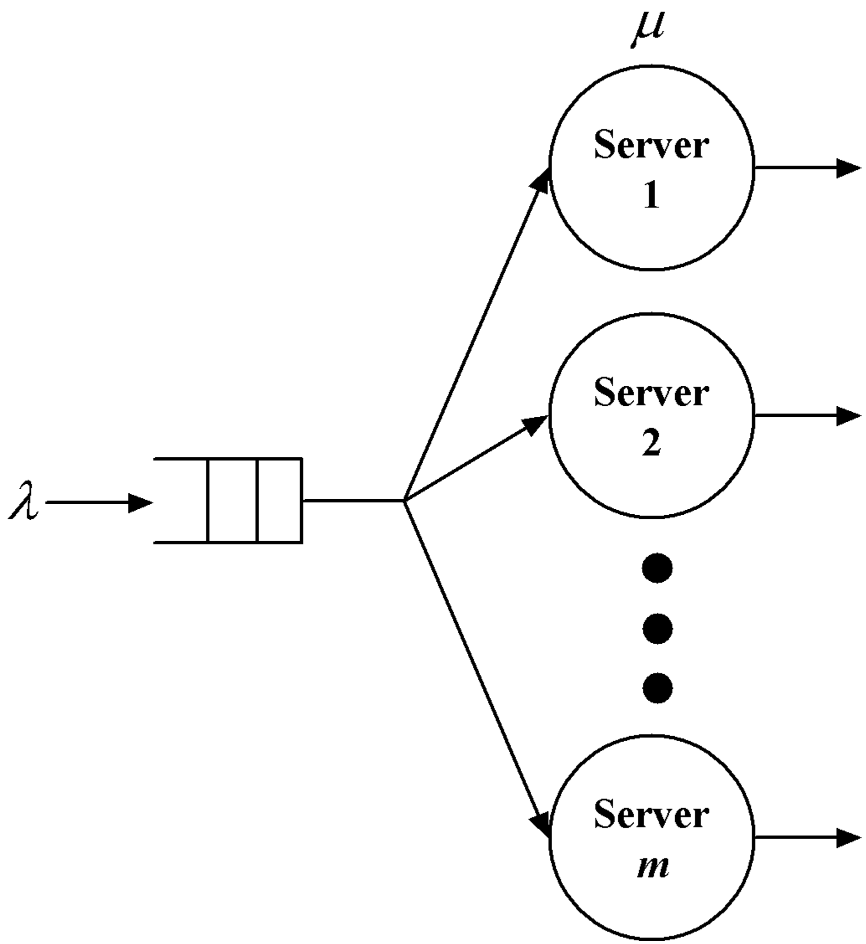


- Solution

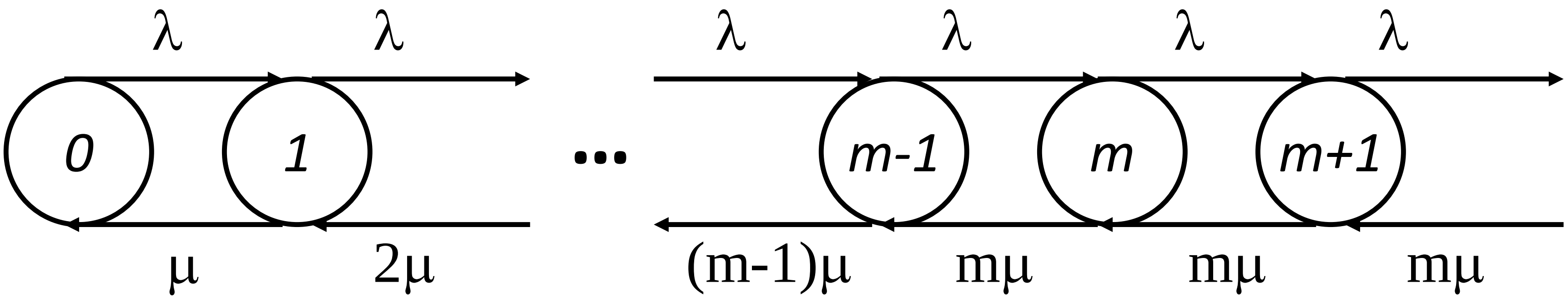
$$p_n = p_0 \rho^n, \qquad \text{where } \rho = \frac{\lambda}{\mu}$$

$$p_0 = \left[\sum_{n=0}^{K} \rho^n\right]^{-1} = \begin{cases} \dfrac{1-\rho}{1-\rho^{K+1}} & \rho \neq 1 \\ \dfrac{1}{K+1} & \rho = 1 \end{cases}$$

M/M/m - Multiple Servers



- State-transition diagram:

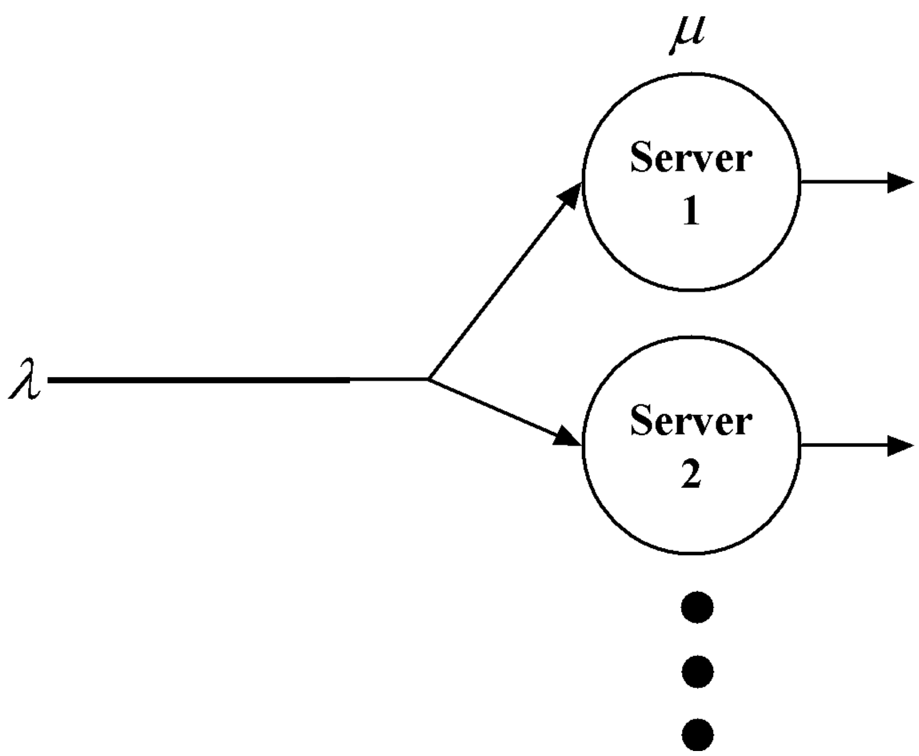


- Solution

$$p_n = p_0 \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}} = \begin{cases} p_0 \rho^n \dfrac{1}{n!} & n \le m \\ p_0 \rho^n \dfrac{1}{m! m^{n-m}} & n > m \end{cases}$$

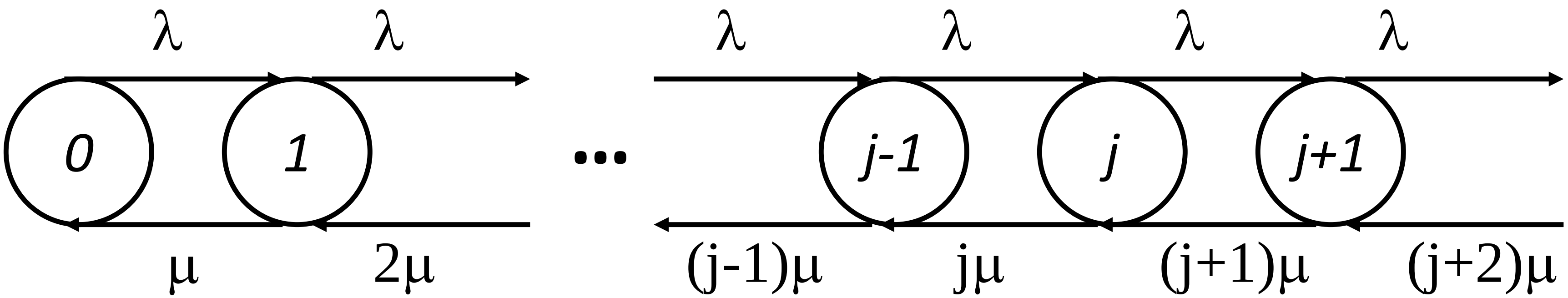
M/M/∞ - Infinite Servers

- Infinite number of servers - no queueing



Analytic Results

- State-transition diagram:



- Solution

$$p_n = p_0 \rho^n \frac{1}{n!}$$

$$p_0 = \left[\sum_{n=0}^{\infty} \rho^n \frac{1}{n!} \right]^{-1} = e^{-\rho}$$

- Thus the number of customers in the system follows a Poisson distribution with rate ρ

M/G/1 Queue

- Single-server queue with Poisson arrivals, general service time distribution, and unlimited capacity
- Suppose service times have mean $\frac{1}{\mu}$ and variance σ^2
- For $\rho < 1$, the steady-state results for $M/G/1$ are:

$$\rho = \lambda / \mu, \quad p_0 = 1 - \rho$$

$$E[n] = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}, \quad E[n_q] = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

$$E[r] = \frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}, \quad E[w] = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$$

M/G/1 Queue

- No simple expression for the steady-state probabilities
- Mean number of customers in service: $\rho = E[n] - E[n_q]$
- Mean number of customers in queue, $E[n_q]$, can be rewritten as:

$$E[n_q] = \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2\sigma^2}{2(1-\rho)}$$

  - If λ and μ are held constant, $E[n_q]$ depends on the variability, σ^2, of the service times.

Effect of Utilization and Service Variability

- For almost all queues, if lines are too long, they can be reduced by decreasing server utilization (ρ) or by decreasing the service time variability (σ^2)

- Coefficient of Variation: a measure of the variability of a distribution

$$CV = \frac{\sqrt{Var(X)}}{E[X]}$$

  — The larger CV is, the more variable is the distribution relative to its expected value.

- Pollaczek-Khinchin (PK) mean value formula:

$$E[n] = \rho + \frac{\rho^2(1+(CV)^2)}{2(1-\rho)}$$

Effect of Utilization and Service Variability

- Consider $E[n_q]$ for *M/G/1* queue:

$$E[n_q] = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

$$= \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{1+(CV)^2}{2}\right)$$

Same as for M/M/1 queue

Adjusts the M/M/1 formula to account for a non-exponential service time distribution