M/G/1 and Priority Queueing

Richard T. B. Ma
School of Computing
National University of Singapore

CS 5229: Advanced Compute Networks
Outline

- PASTA
- $M/G/1$ Workload and FIFO Delay
- Pollaczek-Khinchine Formula
- $M/G/1$ Priority Queueing Delay
- $M/G/1$ Conservation Law
Steady State: A Revisit

- Steady state distribution \( \pi = (\pi_0, \pi_1 \ldots) \)

- \( \pi_i \) is the limiting probability that the steady-state \( X \stackrel{\text{def}}{=} \lim_{t \to \infty} X(t) \) is in state \( i \):
  \[
  \pi_i = P\{X = i\} = \lim_{t \to \infty} P\{X(t) = i\}
  \]

- \( \pi_i \) is also the limiting proportion of time that the system is in state \( i \):
  \[
  \pi_i = \lim_{t \to \infty} \frac{1}{t} \int_0^t 1_{\{X(s) = i\}} ds
  \]
System State Seen by Arrivals

- A point process: \( t = \{t_n: n \geq 0\} \), where \( t_i \) is the arrival time of the \( i^{th} \) observer

- \( \{X(t_n^-): n \geq 0\} \) defines the system states seen by the observers upon their arrivals

- Denote \( \pi_i' = \lim_{n \to \infty} P\{X(t_n^-) = i\} \) or equivalently \( \pi_i' = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n} 1_{\{X(t_j^-) = i\}} \)

- Does \( \pi_i' = \pi_i \) hold?
$X = \{L(t), t \geq 0\}$ under $M/M/1$
Workload Curve

\[ V(t) \]

Departure curve \( N^d(t) \)

\[ W_1 \]

\[ W_2 \]

\[ L(t) \]

Time \( t \)

\[ t = \{ t_n : n \geq 0 \} : \quad t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6 \]
Does $\pi'_i = \pi_i$ hold?

- In a D/D/1 system with $T = 4$, and $S = 3$:

$$\pi_0 = \frac{1}{4} \text{ and } \pi'_0 = 1$$
PASTA

- Poisson Arrivals See Time Averages

Explanation:
- If arrivals are Poisson, then the proportion of time a queueing system spends in a given state ($\pi_i$) is equal to the proportion ($\pi'_i$) of arrivals who find the system in that state.

Notation
- State process: $X = \{X(t): t \geq 0\}$
- Poisson point process: $t = \{t_n: n \geq 0\}$ at rate $\lambda$ with counting process $\{N(t): t \geq 0\}$
Assumption: Lack of Anticipation (LAA)

For any $t \geq 0$, the future (Poisson) increments
\[\{N(t + dt) - N(t): dt \geq 0\}\] are independent of the joint past \[\{(X(u): u \leq t), (N(u): u \leq t)\}\].

Theorem: Any stochastic process \(\{X(t): t \geq 0\}\) (with càdlàg sample paths) jointly with a Poisson point process \(\{t_n: n \geq 0\}\) that satisfies LAA, and let $f$ be any real bounded function. Then with probability 1,
\[
\lim_{t \to \infty} \frac{1}{t} \int_0^t f(X(s))ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n} f(X(t_j^-))
\]
Intuition/Informal “Proof”

Let $U$ be a uniform r.v. in $[0, T]$. $E[X(U)] = \, ?$

$$E[X(U)] = \int_0^T X(t)f_U(t)\,dt = \int_0^T X(t)\frac{1}{T}\,dt = \bar{X}(T)$$
Conditional Arrival Times

- The conditional probability for Poisson:
  \[ P\{t_1 < s \mid N(t) = 1\} = \frac{s}{t} \]
  \[ \Rightarrow E[X(t^-_1) \mid N(T) = 1] = E[X(U)] = \bar{X}(T) \]

- Given \( N(t) = n \), arrival times \( t_1, \ldots, t_n \) are distributed as the order statistics of \( n \) i.i.d. uniform r.v.s \( U_i = \text{uniform}[0, t] \).
  \[ \Rightarrow E\left[ \frac{1}{N(T)} \left( X(t^-_1) + \cdots X(t^-_n) \right) \mid N(T) = n \right] \]
  \[ = E\left[ \frac{1}{n} (X(U_1) + \cdots X(U_n)) \right] = \bar{X}(T) \]
Conditional Expectation

\[ \mathbb{E} \left[ \frac{1}{N(T)} [X(t_1^-) + \cdots + X(t_{N(T)}^-)] \mid N(T) > 0 \right] \]

\[ = \sum_{n=1}^{\infty} \mathbb{E} \left[ \frac{1}{N(T)} (X(t_1^-) + \cdots + X(t_{n}^-)) \mid N(T) = n \right] \mathbb{P}\{N(T) = n\} \]

\[ = \sum_{n=1}^{\infty} \bar{X}(T) \mathbb{P}\{N(T) = n\} = \bar{X}(T) \sum_{n=1}^{\infty} \mathbb{P}\{N(T) = n\} \]

\[ = \bar{X}(T) \mathbb{P}\{N(T) > 0\} \]
Conditional Expectation

\[
\lim_{t \to \infty} E \left[ \frac{1}{N(t)} \left[ X(t_1^-) + \cdots + X(t_{N(t)}^-) \right] \mid N(t) > 0 \right]
\]

\[
= \lim_{t \to \infty} \bar{X}(t) P\{N(t) > 0\} = \lim_{t \to \infty} \bar{X}(t) = \lim_{t \to \infty} \frac{1}{t} \int_0^t X(s) ds
\]

- A special case for \( f(x) = x \) for PASTA

\[
\lim_{t \to \infty} \frac{1}{t} \int_0^t f(X(s)) ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^n f(X(t_j^-))
\]
\[ \lim_{t \to \infty} \frac{1}{t} \int_0^t f(X(s)) \, ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^n f(X(t_j^-)) \]

- Take \( X(t) = V(t) \) and check LAA
- Apply PASTA by using \( f(y) = 1_{\{y \leq x\}} \) for some constant \( x \):

\[ \lim_{t \to \infty} \frac{1}{t} \int_0^t 1_{\{V(s) \leq x\}} \, ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^n 1_{\{V(t_j^-) \leq x\}} \]
\[ \lim_{t \to \infty} \frac{1}{t} \int_0^t P(V(s) \leq x) \, ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^n P(D(t_j^-) \leq x) \]
M/G/1 Workload & FIFO Delay

\[ \lim_{t \to \infty} \frac{1}{t} \int_0^t P(V(s) \leq x) ds = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n} P(D(t_j^-) \leq x) \]

\[ \Rightarrow P\{V \leq x\} = P\{D \leq x\} \]

\[ \Rightarrow V \cong D \Rightarrow E[V] = E[D] \text{ (FIFO delay)} \]

- Intuition: the workload seen by an arrival equals the waiting time for that arrival

- How can we calculate \( E[V] \)?
Brumelle's Formula: A Revisit

- With FIFO service discipline

\[ E[V] = \lambda E[SD] + \frac{\lambda}{2} E[S^2] \]

- Let \( \rho \overset{\text{def}}{=} \lambda E[S] \), with \( G/G/1 \) mean workload
  - Service time is independent of arrivals

\[ E[V] = \lambda E[S] E[D] + \frac{\lambda}{2} E[S^2] = \rho E[D] + \rho E[R_s] \]

- \( M/G/1 \) mean workload and FIFO delay

\[ E[V] = \rho E[D] + \rho E[R_s] = E[D] \, (\text{by PASTA}) \]
M/G/1 Workload & FIFO Delay

- M/G/1 mean FIFO delay has a closed-form

\[ E[D] = \rho E[D] + \rho E[R_s] \Rightarrow E[D] = \frac{\rho}{1 - \rho} E[R_s] = \frac{\lambda E[S^2]}{2(1 - \rho)} \]

- So as the workload under any work-conserving service discipline

\[ E[V] = E[D] = \frac{\rho E[R_s]}{1 - \rho} = \frac{\lambda E[S^2]}{2(1 - \rho)} \]

- Sanity check for M/M/1, \( E[V] = \frac{\rho}{\mu(1-\rho)}. \)
Pollaczek-Khinchine Mean Formula

- For an $M/G/1$ system with FIFO:
  - Mean Queueing Delay (P-K mean formula)
    
    $$E[D] = \frac{\rho E[R_s]}{1 - \rho} = \frac{\lambda E[S^2]}{2(1 - \rho)}$$

  - Mean Customers in Queue: $E[Q] = \lambda E[D]$
  - Mean Customers in System:
    
    $$E[L] = \lambda E[W] = \lambda (E[D] + E[S]) = E[Q] + \rho$$
Example

- FIFO M/G/1 system, service rate $\mu$ Mbps. Packet arrival rate is $\lambda$ packets per minute, with size distribution $P\{L = 10K \text{ bits}\} = 0.3$ and $P\{L = 1K \text{ bits}\} = 0.7$.

- What is the expected number of packet in the system?
Pollaczek-Khinchine Formula (*)

\[ D \equiv \sum_{i=1}^{G} R_i \]

- \( \{R_i\} \) are i.i.d. distributed as \( R_s \)
- \( G \) is geometric, independent of \( \{R_i\} \), with
  \[ P\{G = i\} = (1 - \rho)\rho^i \]
- If \( G = 0 \), then \( D = 0 \), which shows
  \[ P\{D = 0\} = P\{G = 0\} = 1 - \rho \]
- \( E[D] = E[\sum_{i=1}^{G} R_i] = E[G]E[R_s] = \frac{\rho}{1-\rho} E[R_s] \)
- Can be derived by RCL & Laplace transform
How about $E[D]$ under LIFO?

- Let $\rho \overset{\text{def}}{=} \lambda E[S]$, with $G/G/1$ mean workload
  - Service time is independent of arrivals
  \[
  E[V] = \lambda E[S]E[D] + \frac{\lambda}{2} E[S^2] = \rho E[D] + \rho E[R_s]
  \]
  - Avg. work in queue + avg. work in service
    \[
    E[V] = \rho E[D] + \rho E[R_s]
    \]
    \[
    \Rightarrow E[D] = E[V]/\rho - E[R_s]
    \]
- Problem: We derived Brummelle’s formula by assuming $D$ to be FIFO delay
How about $E[D]$ under LIFO?

\[
E[V] = \lambda E[S] E[D] + \frac{\lambda}{2} E[S^2] = \rho E[D] + \rho E[R_s]
\]

- Avg. work in queue + avg. work in service
  
  \[
  E[V] = E[V_q + V_s] = E[V_q] + E[V_s]
  \]

- Average work in service:
  
  \[
  E[V_s] = (1 - \rho)0 + \rho E[R_s] = \rho E[R_s]
  \]

- To show average work in queue:
  
  \[
  E[V_q] = \rho E[D]
  \]
Average workload in queue $E[V_q]$

$$E[V_q] = E \left[ \sum_{i=1}^{Q} S_i \right] = E[Q]E[S] = \lambda E[D]E[S] = \rho E[D]$$

- by Wald’s equation
- $E[Q] = \lambda E[D]$ (by Little’s law)
- $E[D]$ denotes the mean delay under any non-working conserving policy, that does not depend on the service times.
Wald’s equation (Special case)

\[
E \left[ \sum_{i=1}^{Q} S_i \right] = \sum_{q=1}^{\infty} E \left[ \sum_{i=1}^{Q} S_i \mid Q = q \right] P\{Q = q\}
\]

- Independence of \( Q \) and \( S_i \) implies:

\[
= \sum_{q=1}^{\infty} E \left[ \sum_{i=1}^{q} S_i \right] P\{Q = q\} = \sum_{q=1}^{\infty} q \ E[S] P\{Q = q\}
\]

\[
= E[S] \sum_{q=1}^{\infty} q P\{Q = q\} = E[S] E[Q]
\]
How about $E[D]$ under LIFO?

- An intuitive argument: the following two sample paths occur with the same chance.

- **FIFO:**
  - 1
  - 2
  - 3

- **LIFO:**
  - 1
  - 2
  - 3

Time $t$
M/G/1 Priority Queueing

- **Settings:**
  - $I$ service classes, indexed by $i = 1, \ldots, I$
  - Lower index $\Rightarrow$ Higher priority
  - Define $\lambda_i$ as the Poisson arrival rate, $S_i$ as the service time of a class $i$ job

- **Assumptions:**
  - Work-serving
  - Non-preemptive or preemptive
  - FIFO within each service class
Priority Queueing: Illustration

Non-preemptive:

Arrive at $t = 1$

Arrive at $t = 3$

Arrive at $t = 5$

Preempt-resume:

$S = 5$

$S = 3$

$S = 4$
Definitions of Some Entities

- Define $i$th service class’s utilization as
  \[ \rho_i \overset{\text{def}}{=} \lambda_i E[S_i] \]
  - Utilization contributed by jobs from class $i$
  - Mean arrival rate of workload from class $i$

- System arrival rate and mean service time
  \[ \lambda \overset{\text{def}}{=} \sum_{i=1}^{I} \lambda_i, \quad \bar{S} \overset{\text{def}}{=} \sum_{i=1}^{I} (\lambda_i/\lambda)E[S_i]. \]

- System utilization
  \[ \rho \overset{\text{def}}{=} \lambda \bar{S} = \sum_{i=1}^{I} \rho_i \Rightarrow P[\text{server idle}] = 1 - \rho \]
Nonpreemptive Queueing Delay

- $E[D_i] \overset{\text{def}}{=} E[\text{waiting time for jobs from class } i]$
- $E[W_i] \overset{\text{def}}{=} E[\text{class } i \text{ sojourn time}] = E[D_i] + E[S_i]$

\[
D_i = V_R + D_i^1 + D_i^2 = V_R + \sum_{j=1}^{i} V_{qj} + \sum_{j=1}^{i-1} \sum_{k=1}^{N_j(D_i)} S_{jk}
\]

- $V_R$: residual workload in the server
- $D_i^1$: delay due to jobs in the queue upon arrival
- $D_i^2$: delay due to jobs who arrive afterwards

\[
E[D_i] = \overline{V_R} + \overline{D_i^1} + \overline{D_i^2}
\]
Mean Residual Workload $\overline{V_R}$

- Apply PASTA again

$$\overline{V_R} = \sum_{j=1}^{l} \rho_j E[R_j] = \sum_{j=1}^{l} \rho_j \frac{E[S_j^2]}{2E[S_j]} = \sum_{j=1}^{l} \frac{\lambda_j E[S_j^2]}{2}$$

where, $E[R_j] = \frac{E[S_j^2]}{2E[S_j]}$ is the mean residual time

- Under non-preemptive system, $\overline{V_R}$ is independent of priority classes.
$D_i^1$: average workload in queue

- Interpret $D_i^1$ as
  - the average workload in queue upon arrival of a class $i$ customer
  - that needs to be finished before serving $i$

$$
D_i^1 = \sum_{j=1}^{i} \mathbb{E}[V_{qj}] = \sum_{j=1}^{i} \mathbb{E} \left[ \sum_{k=1}^{Q_j} S_{jk} \right] = \sum_{j=1}^{i} \mathbb{E}[Q_j] \mathbb{E}[S_j]
$$

$$
= \sum_{j=1}^{i} \lambda_j \mathbb{E}[D_j] \mathbb{E}[S_j] = \sum_{j=1}^{i} \rho_j \mathbb{E}[D_j]
$$
$D_i^2$: avg. workload arrived later

\[
D_i^2 = E \left[ \sum_{j=1}^{i-1} \sum_{k=1}^{N_j(D_i)} S_{jk} \right] = \sum_{j=1}^{i-1} E \left[ \sum_{k=1}^{N_j(D_i)} S_{jk} \right]
\]

\[
= \sum_{j=1}^{i-1} E[N_j(D_i)]E[S_j] = \sum_{j=1}^{i-1} \lambda_j E[D_i]E[S_j] = \sum_{j=1}^{i-1} \rho_j E[D_i]
\]

\[\textbf{because}\ E[N_j(D_i)] = \int_{x=0}^{\infty} E[N_j(D_i)|D_i = x] f_{D_i}(x)\ dx
\]

\[= \int_{x=0}^{\infty} E[N_j(x)] f_{D_i}(x)\ dx = \int_{x=0}^{\infty} \lambda_j x f_{D_i}(x)\ dx = \lambda_j E[D_i]
\]
Triangular Equations

\[ E[D_i] = V_R + \sum_{j=1}^{i} \rho_j E[D_j] + \sum_{j=1}^{i-1} \rho_j E[D_i] \]

\[ = \frac{V_R + \sum_{j=1}^{i} \rho_j E[D_j]}{1 - \sum_{j=1}^{i-1} \rho_j} \]

- Compute starting from \( i = 1 \)

\[ E[D_i] = \frac{V_R}{(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)} \]
Priority Class $i$

- **Mean queueing Delay**
  \[
  \mathbb{E}[D_i] = \frac{\overline{V}_R}{(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)}
  \]

- **Mean sojourn time**
  \[
  \mathbb{E}[W_i] = \mathbb{E}[D_i] + \mathbb{E}[S_i]
  \]

- **Mean number of customers in queue**
  \[
  \mathbb{E}[Q_i] = \lambda_i \mathbb{E}[D_i]
  \]

- **Mean number of customers in system**
  \[
  \mathbb{E}[L_i] = \mathbb{E}[Q_i] + \rho_i
  \]
How about Preemptive-Resume?

Without preemption:

\[ E[D_i] = \frac{V_R}{(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)} \]

\[ V_R = \sum_{j=1}^{I} \rho_j E[R_j] \]

\[ E[W_i] = E[D_i] + E[S_i] \]

With preemption:

\[ E[\tilde{D}_i] = \frac{\sum_{j=1}^{i} \rho_j E[R_j]}{(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)} \]

\[ E[\tilde{W}_i] = E[\tilde{D}_i] + E[\tilde{S}_i] = E[\tilde{D}_i] + \frac{E[S_i]}{1 - \sum_{j=1}^{i-1} \rho_j} \]
Service Time with Preemption

- Consider a customer of class 2, starting service at time $t_0$, during service time, it might be preempted by class 1 customers.

- Time until it finishes is $\tilde{S}_2 \triangleq \sum_{j=1}^{\infty} S_2^{(j)}$

![Diagram showing service time with preemption]

- $S_2^{(1)}$ = service time before preemption
- $S_2^{(2)}$ = service time after first preemption
- $S_2^{(3)}$ = service time after second preemption

$t_0$ and $t_0 + \tilde{S}_2$ mark the start and end of the total service time with preemptions.
Deriving $E \left[ S^{(j)}_i \right]$

- For any class $i$ customer, $\tilde{S}_i \overset{\text{def}}{=} \sum_{j=1}^{\infty} S_i^{(j)}$
- Condition on the length of the previous cycle

$$E \left[ S_i^{(j+1)} \right] = \int_{x=0}^{\infty} E \left[ S_i^{(j+1)} \big| S_i^{(j)} = x \right] f_j(x) \, dx$$

$$= \int_{x=0}^{\infty} (\lambda_1 x E[S_1] + \cdots + \lambda_{i-1} x E[S_{i-1}]) f_j(x) \, dx$$

$$= \int_{x=0}^{\infty} (\rho_1 + \cdots + \rho_{i-1}) x f_j(x) \, dx$$

$$= (\rho_1 + \cdots + \rho_{i-1}) E \left[ S_i^{(j)} \right]$$
Deriving $E\left[\widetilde{S}_i\right]$

- $E\left[S_{i}^{(j+1)}\right] = (\rho_1 + \cdots + \rho_{i-1})^j E\left[S_i^{(1)}\right]$

- Let $\gamma \overset{\text{def}}{=} \rho_1 + \cdots + \rho_{i-1}$

$$E[\widetilde{S}_i] = \sum_{j=1}^{\infty} E\left[S_i^{(j)}\right] = (1 + \gamma + \gamma^2 + \cdots)E\left[S_i^{(1)}\right]$$

$$= (1 + \gamma + \gamma^2 + \cdots)E[S_i] = \frac{1}{1-\gamma} E[S_i]$$

$$\Rightarrow E[\widetilde{S}_i] = \frac{E[S_i]}{1 - \sum_{j=1}^{i-1} \rho_j}$$
Why $E[\tilde{S}_i] = E[S_i] / (1 - \sum_{j=1}^{i-1} \rho_j)$?

\[
\tilde{S}_i = S_i + \sum_{j=1}^{i-1} \sum_{k=1}^{N_j(\tilde{S}_i)} S_{jk}
\]

\[
E[\tilde{S}_i] - E[S] = \sum_{j=1}^{i-1} E \left[ \sum_{k=1}^{N_j(\tilde{S}_i)} S_{jk} \right] = \sum_{j=1}^{i-1} E[N_j(\tilde{S}_i)] E[S_j]
\]

\[
= \sum_{j=1}^{i-1} \lambda_j E[\tilde{S}_i] E[S_j] = E[\tilde{S}_i] \sum_{j=1}^{i-1} \lambda_j E[S_j] = E[\tilde{S}_i] \sum_{j=1}^{i-1} \rho_j
\]
(Delay) Conservation Law

- “You don’t get something for nothing.”
  - Shorter delay for higher class ➔ longer delay for lower class

- What is unchanged/conserved? workload
  - No idle time when facing non-empty queue
  - In busy periods, independent of order of service
  - Require no work is created/destroyed in system
  - No user leaves the system before finishing
    - Only consider work-conserving service disciplines
M/G/1 Conservation Law

- Under M/G/1, if the service discipline is
  1. non-preemptive and work conserving
  2. independent of the service times

  then the following must hold:

  \[ E[V_q] = \sum_{j=1}^{l} \rho_j E[D_j] = \begin{cases} \frac{\rho}{1 - \rho} \bar{V}_R & \rho < 1 \\ +\infty & \rho \geq 1 \end{cases} \]

- Weighted sum of the queueing delay \( E[D_j] \) can NEVER CHANGE regardless how sophisticated the service discipline is.
Conservation Law: Derivation

- Avg. workload in queue (or $\overline{D_1^I}$) is conserved

$$E[V] - \overline{V_R} = \sum_{j=1}^{l} E[Q_j]E[S_j] = \sum_{j=1}^{l} \lambda_j E[D_j]E[S_j] = \sum_{j=1}^{l} \rho_j E[D_j]$$

$\rho_j E[D_j]$ is the mean workload from class $j$ in queue.

- $E[V]$ is the same under $M/G/1$

$$E[V] = \frac{\lambda E[S^2]}{2(1-\rho)} \Rightarrow E[V] = \frac{\overline{V_R}}{1-\rho} \quad \text{(P-K mean formula)}$$

because $E[S^2] = \sum_{j=1}^{l} \frac{\lambda_j}{\lambda} E[S_j^2] = \sum_{j=1}^{l} \frac{2 \lambda_j E[S_j^2]}{2\lambda} = \frac{2}{\lambda} \overline{V_R}$
Class-homogenous Service Time

- For the special case of \( E[S_j] = E[S] \),

\[
\sum_{j=1}^{I} \rho_j E[D_j] = \frac{\rho}{1 - \rho} \bar{V}_R \Rightarrow \sum_{j=1}^{I} \lambda_j E[D_j] = \frac{\lambda}{1 - \rho} \bar{V}_R
\]

- Avg. # of customers in queue is a constant.

- By Little’s Law, the average queueing delay (“averaged” over all classes) is a constant:

\[
\sum_{j=1}^{I} (\lambda_j / \lambda) E[D_j] = \frac{1}{1 - \rho} \bar{V}_R
\]
Workload in Queue Conserved

Workload in queue $V_q$:

$M/G/1$ FIFO without priority
Workload in Queue Conserved

\[ S = 3 \]
\[ S = 4 \]

Workload in queue \( V_q \):

\[ M/G/1 \text{ FIFO without priority} \]
Workload in Queue Conserved

Workload in queue $V_q$:

$S = 3$, $S = 4$

Preemptive-Resume
G/G/1 Conservation Law (*)

- If select customers in a way that is independent of their service time, then
  - distribution of the number of customers in the system will be invariant of the order of service.
  - Avg. queueing delay is also invariant. (by LL)

- In particular, for non-preemptive systems:

\[
E[V_q] = E[V] - \overline{V_R} = \sum_{j=1}^{I} \rho_j E[D_j] = E[V] - \sum_{j=1}^{I} \frac{\lambda_j E[S_j^2]}{2}
\]
Service-Time Dependency (*)

- The mean delay with FIFO is a tight lower bound for work conserving and service time independent service disciplines.

- Service time dependent scheduling:
  - SPT: shortest processing time first
  - SRPT: shortest remaining processing time first
    \[ E[D]_{FIFO} \geq E[D]_{SPT} \geq E[D]_{SRPT} \]
  - However, uncommon in packet switching because the packet ordering will be modified and delay for large packets increases.
References
