PS queue (Processor Sharing)

- In a single server PS queue the capacity $C$ of the server is equally shared between the customers in system,
  - if there are $n$ customers in system each receives service at the rate $C/n$
  - customers don’t have to wait at all; the service starts immediately upon arrival

- PS queue has become an important tool, e.g., in the flow level modelling of the Internet
  - roughly speaking, TCP shares the resources of the network equally between the flows in progress (that is, transfers of web pages or other documents)

- PS queue is an idealized model, since in general the capacity of the server cannot be divided in continuous (real valued) parts, if at all. PS is, however, a good approximation
  - for the round robin (RR) discipline where the customers are served in turn, each for a small time slice (for instance, time sharing operating systems)
  - for document or file transfer, when these are divided in small packets served in turn or whose transmission rates from the sources have been equalized

- PS queue is theoretically interesting because its average properties are insensitive to the distribution of the service demands of the customers (unlike those of a FIFO queue).
**M/M/1-PS queue**

- The arrival process is Poisson with intensity $\lambda$ and the distribution of the service demand (job size) is exponential such that if the customer got all the capacity of the server the service time would be distributed as $\text{Exp}(\mu)$.

- Then the number of customers in system $N$ obeys the same birth-death process as in the familiar M/M/1-FIFO queue:
  - the probability per time unit for an arrival of a new customer is $\lambda$
  - with $n$ customers in system the finishing intensity of each of them is $\mu/n$; thus the overall probability per time unit that the service of some customer ends is $\mu$

- The queue length distribution of the PS queue is the same as for the ordinary M/M/1-FIFO queue,
  \[ \pi_n = (1 - \rho)\rho^n, \quad \rho = \frac{\lambda}{\mu} \]

- Accordingly, the expected number of customers in system $E[N]$ and, by Little’s theorem, the expected delay in the system $E[T]$ are
  \[ E[N] = \frac{\rho}{1 - \rho}, \quad E[T] = \frac{1/\mu}{1 - \rho} \]
M/G/1-PS queue

- First we derive an auxiliary result valid for a general G/G/1 queue. Denote

\[
\begin{align*}
  x &= \text{the amount of service (work) received by a customer} \\
  F_X(x) &= \text{cdf of the service demand (work) of a customer} \\
  N(x) &= \text{av. number of customers in system that have received service } \leq x \\
  T(x) &= \text{av. time spent in system by customers that have received service } x \\
  n(x) &= \frac{dN(x)}{dx} = \text{av. density of customers (wrt received service)}
\end{align*}
\]

- Now, apply Little’s theorem to a black box defined as follows:

  - customer arrives at the box when the amount of service received passes \( x \); the customer has then spent in system a time \( T(x) \) on average
  - customer departs from the box when the amount of service received passes \( x + \Delta x \); the customer has then spent in system a time \( T(x + \Delta x) \) on average

- Think of the service demand to be discrete so that the amount of required work is always a multiple of \( \Delta x \)

  - then no customer exits the box because of the completion of the job
  - finally, in the limit \( \Delta x \to 0 \) the discreteness becomes immaterial
M/G/1-PS queue (continued)

- As customers arrive at the system at rate $\lambda$ and the fraction $1 - F_X(x)$ of them reach the “service age” $x$, the arrival rate at the box is $\lambda(1 - F_X(x))$.
- The mean delay of a customer in the box is $T(x + \Delta x) - T(x)$.

\[
\begin{array}{c}
\lambda(1-F_X(x)) \\
\downarrow \text{T(x)} \\
\begin{array}{c}
N(x+\Delta x)-N(x) \\
\downarrow \text{T(x+\Delta x)} \\
x \\
x+\Delta x
\end{array}
\end{array}
\]

- Little’s result gives

\[
N(x + \Delta x) - N(x) = \lambda(1 - F_X(x))(T(x + \Delta x) - T(x))
\]

- By dividing by $\Delta x$ we obtain in the limit $\Delta x \to 0$ the desired auxiliary result

\[
n(x) = \lambda \left(1 - F_X(x)\right) \frac{dT(x)}{dx}
\]
M/G/1-PS queue (continued)

• On the other hand, we can directly deduce that

\[
 n(x) = n(0) \cdot (1 - F_X(x))
\]

• This is because all the customers in a PS queue are served at the same rate
  
  – at every instant of time, the “service age” of all the customers increases at the same rate
  
  – the difference in the customer density wrt to the service age arises only due to departures of customers upon completion of their service
  
  – by age \( x \) the fraction \( F_X(x) \) of the customers have departed and the fraction \( 1 - F_X(x) \) of them remains in the system

• By equating the expressions for \( n(0) \) in the framed equations, we obtain

\[
 \frac{dT(x)}{dx} = \frac{n(0)}{\lambda} \quad \text{or} \quad T(x) = \frac{n(0)}{\lambda} x
\]

• \( T(x) \) is besides the average time spent in system by customer with age \( x \), also the total mean delay of those customers whose service demand is \( x \), i.e. the mean delay conditioned on the service requirement.
M/G/1-PS queue (continued)

- Further, one can deduce that
  \[ \lim_{x \to \infty} T(x) = \frac{x}{C(1 - \rho)} \]

- Arrival of a very big job is rare sole event. The job stays in the system for a very long time. Meanwhile, all the other (small) jobs arriving in the system pass by; the big job sees effectively the service rate remaining from the other jobs, \( C(1 - \rho) \).

- Thus the coefficient of proportionality \( n(0)/\lambda \) in the equation \( T(x) = (n(0)/\lambda) x \) is \( 1/C(1 - \rho) \),

\[
T(x) = \frac{x}{C(1 - \rho)}
\]

- By averaging this formula for the conditional delay with respect to the distribution of the job size, and then applying Little’s result, one obtains again the mean formulae

\[
E[T] = \frac{1/\mu}{1 - \rho} \quad 1/\mu = E[X]/C, \quad E[N] = \frac{\rho}{1 - \rho}
\]
M/G/1-PS queue (continued)

- The important thing in these re-derived formulae is that we didn’t make any assumption on the distribution of the job size. The mean formulae for the PS queue are insensitive.

- The equation $T(x) = x/C(1 - \rho)$ tells that the average delay of a customer in system is proportional to the job size
  
  - the mean delay in system of each customer is its service time $x/C$, had it all the capacity of the server, multiplied by the “stretching” factor $1/(1 - \rho)$
  
  - on the average, each customer sees the same effective service capacity $C(1 - \rho)$.

- Because of these properties the PS queue can be considered the most equalitarian queueing discipline.
M/G/1-PS queue (continued)

- According to Pollaczek-Khinchin results the mean queue length and mean delay in an M/G/1-FIFO queue are greater (smaller) than in a corresponding M/M/1-FIFO queue, and thence in an M/G/1-PS queue, if the squared coefficient of variation $C_v^2$ of the service demand is greater (smaller) than 1.

- The superiority of the PS discipline in the case of a large squared coefficient of variation is easy to understand as
  - in FIFO, a large number of small jobs have to wait the completion of a long job
  - whereas, in a PS queue, they can pass by
  - a large number of customers experience better service in the PS system

- In the case of a small squared coefficient of variation (regular traffic), the more disciplined FIFO scheduling is better.

- With the M/M/1 assumptions, whence the means are equal, the variance of the delay distribution in the PS-queue is greater than that in the FIFO queue. One can derive the results:
  \[
  V[T]_{\text{FIFO}} = \frac{1}{\mu^2(1 - \rho)^2} \quad V[T]_{\text{PS}} = \frac{1}{\mu^2(1 - \rho)^2} \frac{2 + \rho}{2 - \rho}
  \]
  the latter factor is in the range 1 \ldots 3
Example: downlink data traffic in a cellular system

- The HSPDA protocol (High speed downlink packet access) of 3G cellular systems uses a time-division type multiplexing:
  - the base station (BS) transmits at full power to only one user in each time slot.
- If slots are assigned in a round robin fashion to the active users, then the BS station realizes a PS queue for the downlink traffic.
- Link adaptation: the bit rate is adapted to the radio channel conditions.
- The rate goes down with the distance $r$ from the BS as signal becomes weaker, e.g.,

$$C(r) = \begin{cases} 
C_0 & r \leq r_0 \\
C_0 \left(\frac{r_0}{r}\right)^\alpha & r > r_0
\end{cases}$$

where $r_0$ is some threshold range within which the maximal rate $C_0$ is obtained; the exponent $\alpha$ is typically in the range 2 . . . 4.
- $C(r)$ is the maximum bit rate for a user at distance $r$
  - when there are $n$ users active in the cell, the rate is $C(r)/n$. 
Example (continued)

Assumptions:

- The cell is approximated by a circular disk with radius $R$.
- Flows arrive at the base station at total rate $\lambda$ (Poisson process).
- Each flow has a size $X$ independently drawn from some distribution with mean $\bar{X}$.
- The location of the destination point of each flow is independently drawn from a uniform distribution in the disk.

The service time $S$ (at full rate, without sharing)

$$ S = \frac{X}{C(r)} $$

is a random variable because both $X$ and the distance $r$ are random variables. $X$ and $r$ are, however, independent and we have

$$ \bar{S} = \frac{\bar{X}}{C} \quad \text{where} \quad \frac{1}{C} = \left( \frac{1}{C(r)} \right) = \frac{1}{\pi R^2} \int_0^R \frac{2\pi r}{C(r)} \, dr $$
Example (continued)

- The queue at BS is a PS queue with load $\rho = \lambda \bar{S} = \lambda \bar{X} / \bar{C}$.
- For a stable queue we must have $\rho < 1$. Thus the greatest sustainable traffic load, $\lambda \bar{X}$, is $\bar{C}$ and we have the cell capacity [kbits/s],

$$\bar{C} = \left( \frac{1}{\pi R^2} \int_0^R \frac{2\pi r}{C(r)} dr \right)^{-1}$$

- Because the system is a PS queue we have the effective service rate (also called flow throughput) at distance $r$

$$C_{\text{eff}} = C(r)(1 - \rho)$$

and average sending time of a flow of size $X$ for a node at distance $r$,

$$\bar{T}(r, X) = \frac{X}{C(r)(1 - \rho)}$$