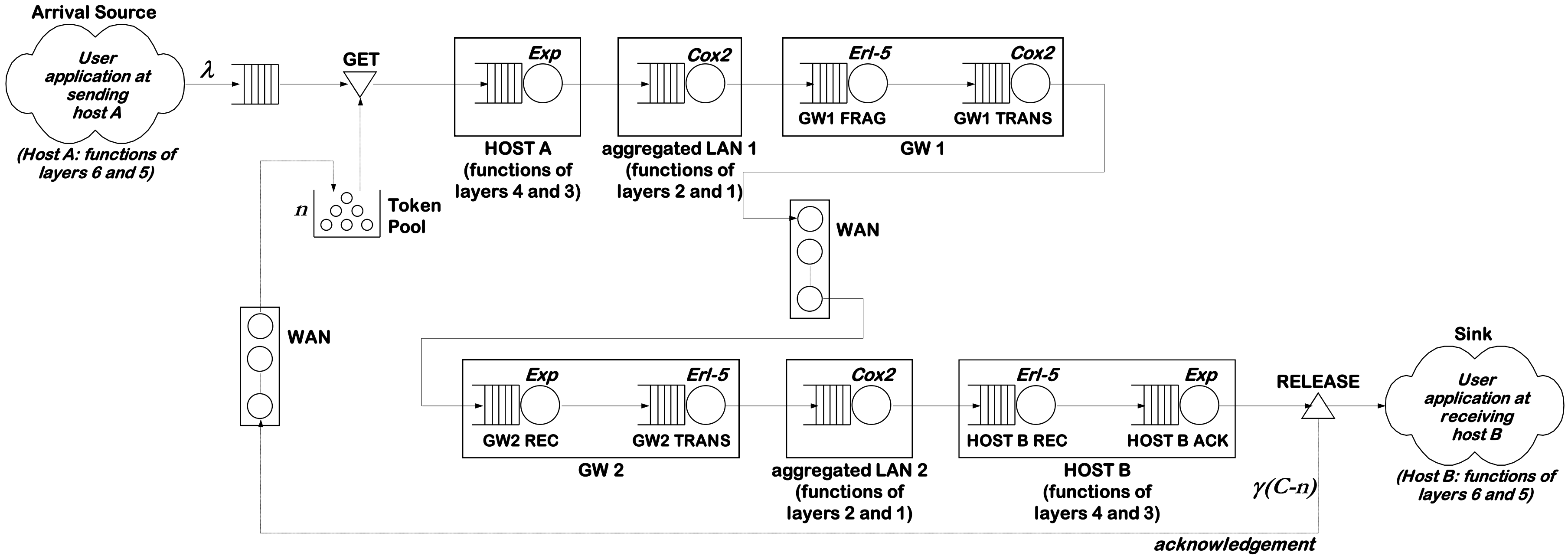
The single center platform

Arrivals, Services , Discipline Laws

From previous lecture: your Internet platform model



Model aggregation

The platform is a mult-center system with

- 11 service centers

  that separate the

- packet arrival Source

  from the

- packet Sink

We'll see further-on how to mathematically deal with multi-center systems.

Model aggregation

For the time being, let us assume we know (by measurements) the time t_s the platform takes to service one packet (or “job” i.e. for the “job” to move from the Source to the Sink) and also (by measurements) the probability law $f(t_s)$ of that service time (e.g.exponential, hyperexponential etc) with its mean $E(t_s)$.

Model aggregation

- In other words let us assume we know the *aggregate* time-behaviour of the 11 centers.
- In this case we can study the entire system as composed of a *single center with*
  - *a*rrival rate λ
  - service-time density $f(t_s)$,
  - mean service time $E(t_s)$
  - (i.e. service rate $\mu=1/E(t_s)$)

with a series of random variables as below:

Definition of center random variables

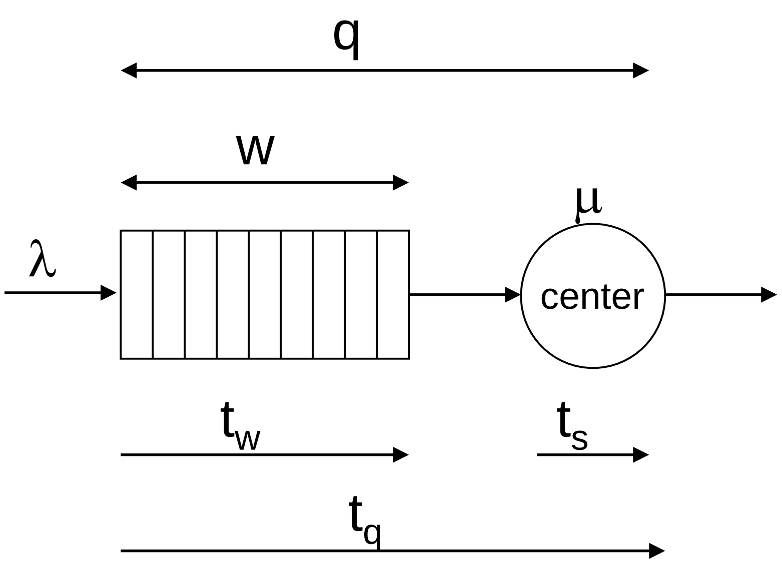


Fig.1: description of a service center

q	number of users at a given instant in the center
w	number of users at a given instant in the queue
t_s	center service time (time to serve one job)
t_w	waiting time at the center (spent in the queue)
t_q	residence time in the center, time between queue entrance and center exit.
n_t	number of users who exit the center in interval t

Definition of random variable averages

E(q) *mean* number of users in the center

E(w) *mean* number of users in the queue

$E(t_w)$ *mean* waiting time in queue

$E(t_q)$ *mean* response time of the center

$E(n)_t$ *mean* number of jobs which exit in interval t

$\lambda = 1/ E(t_a)$ *mean* arrival-rate of jobs (jobs/sec)

$\mu = 1/ E(t_s)$ *mean* service-rate (jobs/sec) or *center work capacity*

The average variable Throughput

Throughput of the center is the mean number of jobs that exit the center in a time unit

==> $E(n)_t$ for $t=1$

Throughput property

The center *does not create nor destroy jobs*, so if λ is the jobs arrival rate to the center then:

The mean number of jobs $E(n)_1$ which exits from the center per time unit will coincide with the mean number that, on average, enters from the source

$$\implies E(n)_1 = \lambda$$

The center utilization factor ρ

$$\rho = \frac{mean\,arrival\,rate\,to\,the\,center}{work\;capacity\,of\;the\,center} = \frac{\lambda}{\mu} = \lambda E(ts)$$

Exercise:

- Draw the throughput trend versus λ for $\lambda < \mu$, $\lambda \geq \mu$

Exercise hint

The condition $\rho \geq 1$ (i.e. $\lambda \geq \mu$) produces endless growth of the queue and the center is said to be *not in stochastic equilibrium*.

On the other hand even if λ grows larger than μ the center capacity remains μ and then:

$$E(n)_1 = \mu \quad \text{for } \lambda \geq \mu$$

In other words, in such a condition, the center throughput $E(n)_1$ remains the work capacity μ of the center (and excess arrivals crowd indefinitely on the queue)

Service center utilization in case of multiple channel center (not this lecture)

By considering a center made up of m identical processors (or channels) each with capacity $\mu=1/E(t_s)$ ==> the total work capacity will be **mμ**

The center utilization will therefore be expressed by

$$\rho = \frac{\lambda}{\mu}$$ If m=1

$$\rho = \frac{\lambda}{m\mu}$$ If m>1

Relationships between instant variables

The following relationship holds between the time variables of the center in Fig.1:

- $t_q = t_w + t_s$

While for the population-variables, the following holds:

- $q = 0$, or 1 if $m = 1$ and $w = 0$
- $q = w+1$ if $m = 1$ and $w \neq 0$
- $q = 0, 1, \ldots, m$ if $m > 1$ and $w = 0$
- $q = w + m$ if $m > 1$ and $w \neq 0$

Where m is the *number of parallel servers* (also called *processors or channels*) with identical service distribution and identical mean service time (m=1 in this lecture).

Relationships between average variables

In addition, the following equations among the mean values hold:

- $E(q) = E(w) + \rho$ if $m = 1$
- $E(q) = E(w) + m\rho$ if $m > 1$
- $E(t_q) = E(t_w) + E(t_s)$

Little theorem

The Little theorem is a *fundamental* relation between the quantities which describe a service center

In particular, Little law links *mean populations and mean times,* where the following is valid

$$E(t_w) = \frac{E(w)}{\lambda}$$

Mean waiting time in the queue

$$E(t_q) = \frac{E(q)}{\lambda}$$

Mean response time

Single center solution

for center in stochastic equilibrium

To synthetically describe a single center model, we use the notation: A/B/m/Z with the following meaning:

- A : distribution of inter-arrival times t_a
- B : distribution of service times t_s
- m : number of parallel processors in the center (for single processor m=1)
- Z : service discipline

The notation is simplified into A/B/m when we presume that the service discipline is independent from the service time (for example FIFO)

The M/G/1 center solution

The notation M/G/1 denotes a *single processor* center having

Exponential distribution (M) inter-arrival times

and

General distribution (G) of type service times

The M/G/1 center solution

The M/G/1 case was resolved by

Khinchin and Pollaczek

by providing the following equation for the mean length of the queue

$$E(w)_{KP} = \frac{\rho^2}{2(1-\rho)}\left[1+\frac{\sigma^2(t_s)}{E^2(t_s)}\right]$$

with $\rho = \lambda/\mu = \lambda E(t_s)$

The M/G/1 center solution

At this point, it is simple to obtain the value of $E(q)$ by applying the relations seen previously

$$E(q) = E(w)_{KP} + \rho$$

From which we can obtain $E(t_w)$ and $E(t_q)$ by the Little theorem

The M/G/1 center solution

The KP solution can be used for centers with

- any general distribution (G) of the service times
- any service discipline that doesn't depend on the service time (thus FIFO, LIFO, RAND, etc.)
- but only exponential interarrival times (M) or, equivalently, Poisson law of arrivals

The M/G/1 center solution

The term $\frac{1}{2}\left[1+\frac{\sigma^2(t_s)}{E^2(t_s)}\right]$ demonstrates that

the center congestion (in terms of mean population in the queue) is directly proportional to the dispersion (variance) of service times requested by the jobs that cross the center

Center solution for various distributions (and variance) of service times

By varying the distributions of the service times, $E(w)_{KP}$ takes on the following forms

- Constant distribution (D) of t_s
  - being $\sigma^2(t_s) = 0$, there results:
  - $E(w) = \dfrac{\rho^2}{2(1-p)}$

Center solution for various distributions (and variance) of service times

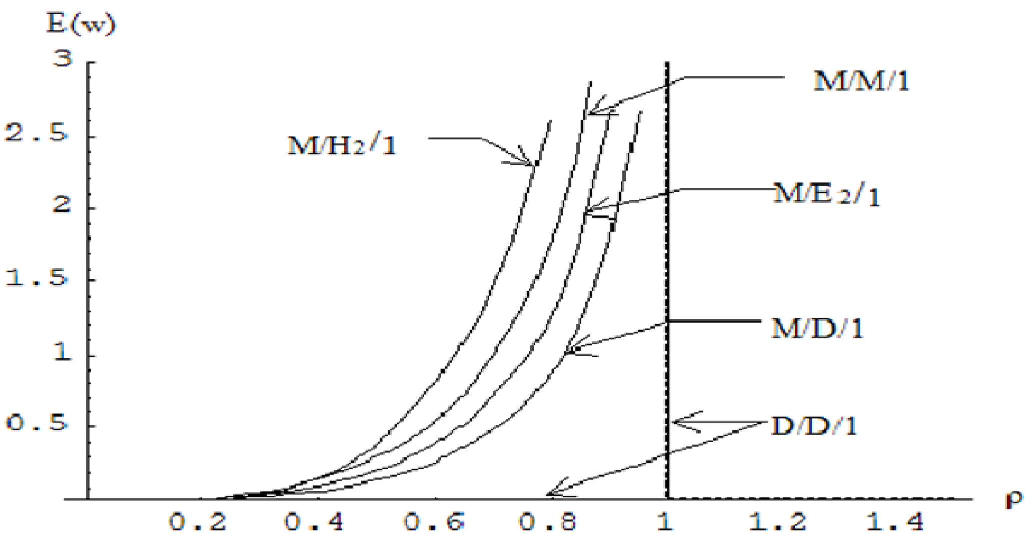
- Exponential distribution (M) of t_s
  - being $\sigma^2(t_s) = E^2(t_s)$, there results:
  - E(w) = $$\frac{\rho^2}{(1-p)}$$
- k-Erlang distribution (E_k) of t_s
  - being $\sigma^2(t_s) = E^2(t_s) / k$, there results
  - E(w) = $$\frac{\rho^2}{2(1-p)}(1+\frac{1}{k})$$

Center solution for various distributions (and variance) of service times

- Hyper-exponential distribution (H_2) of t_s

  - being $\sigma^2(t_s) = \alpha E^2(t_s)$, there results:
  - $E(w) = \dfrac{\rho^2}{2(1-p)}(1+\alpha)$

Qualitative behavior of mean queue length



The figure shows the trend of the mean queue length in function of center utilization and service time distribution

Sensitivity to the factors and disciplines

The Fig. shows that for some service time distributions (exponential and hyperexponential) , the average queue length grows with notable slope for values of ρ (center utilization 80%) from 0.8 and on.

Sensitivity to the factors and disciplines

This growth implies that, when designing the center and presuming 80%utilization we also need to forecast that small increases of λ can cause serious degradations in the behavior of the center

Sensitivity to the factors and disciplines

There is also a sensitivity to the service disciplines which can alter the standard deviation of some of the output variables of the center, like w, q, t_q, t_w

Sensitivity to the factors and disciplines

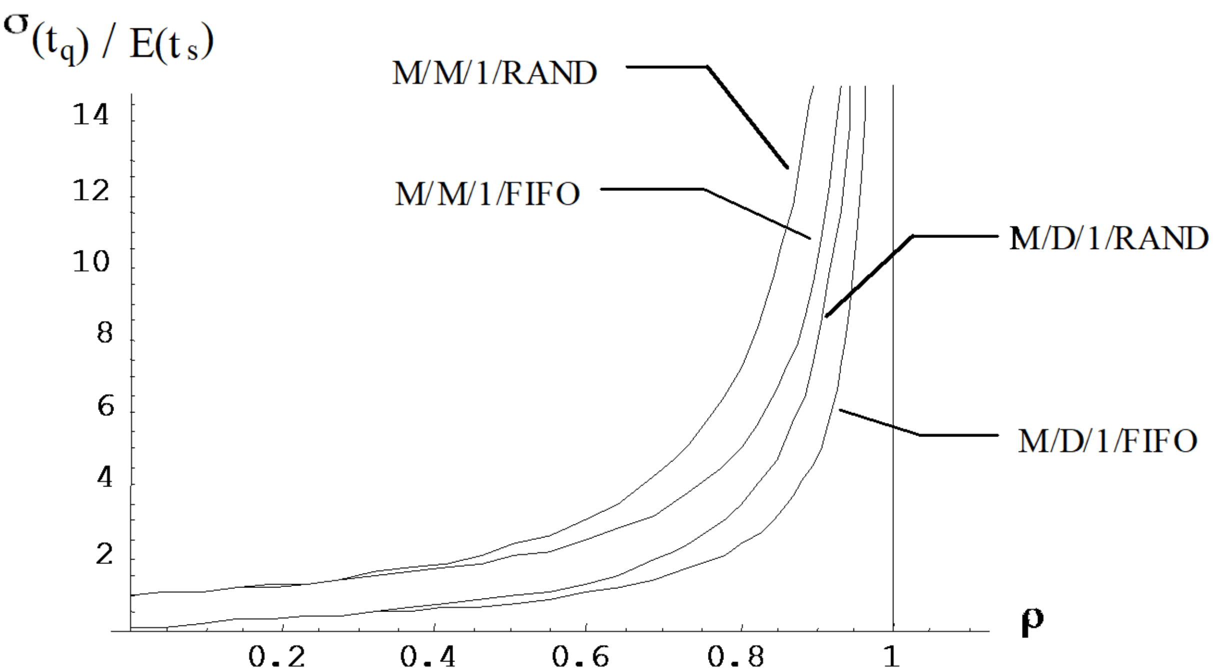
This is important because the variance of the output variables influence the quality of service (QoS), and in particular:

low variance is an index of good quality

high variance no good quality

In the meantime, it is important to know the effect of the disciplines not only on the means but also on the variances

Sensitivity to the factors and disciplines



As we see from the trend of the standard deviation of the response time (per service time unit) the FIFO discipline is the one which offers less variance of response time, while the RAND the highest. The differences between the variances diverge rapidly as ρ tends to 1.

Type D service times also offer less than type M.

Waiting times and seen remaining time

A user who arrives in the queue waits with a time that depends on two factors, called waiting components

1. A factor proportional to the queue length that he finds ahead

2. A factor proportional to the remaining service time the user found at arrival must still spend in the server, indicated with t_{srem} (seen remaining service time)

Waiting time and seen remaining time

We wish to express a relation between the mean waiting time $E(t_w)$ and mean *seen remaining time* $E(t_{srem})$

First remember that one can prove that

$$E(t_{srem}) = \frac{\lambda}{2} E(t_s^{\,2})$$

Waiting time and seen remaining time

Such relation is valid for general (G) service times. For exponential (M) services, the relation becomes

$$E(t_{srem}) = \rho\, E(t_s), \quad \text{since} \quad E(t_s^2) = 2E^2(t_s)$$

Waiting time and seen remaining time

In force of relationship $E(t_s^2) = 2E^2(t_s)$

the exponential mean queue length $E(w)_{KP}$ can be rewritten in function of the *mean seen remaining service time*, as follows

$$E(w)_{kp} = \frac{\lambda^2 E^2(t_s)}{(1-\rho)} = \frac{\lambda E(t_{rimv})}{1-\rho}$$

From which, by the Little theorem:

$$E(t_w)_{kp} = \frac{E(t_{srimv})}{1-\rho}$$

Waiting time and seen remaining time

That proves that the mean waiting time depends on two waiting components

1. A factor proportional to the queue length that he finds ahead, of value $1/(1-\rho)$

2. A factor proportional to the mean remaining service time $E\ (t_{srem})$

Such two factors do not add each other , they instead multiply each other, meaning that if either of two is zero, the resulting waiting time is also zero.