Performance Modeling of Computer Systems and Networks

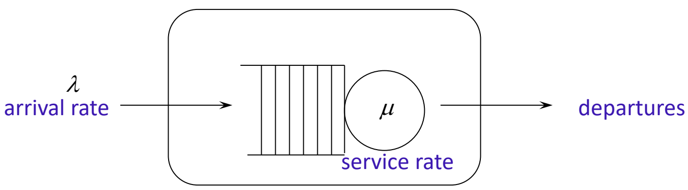
Prof. Vittoria de Nitto Personè

Analytical models
(single resource)

Università degli studi di Roma Tor Vergata
Department of Civil Engineering and Computer Science Engineering

1

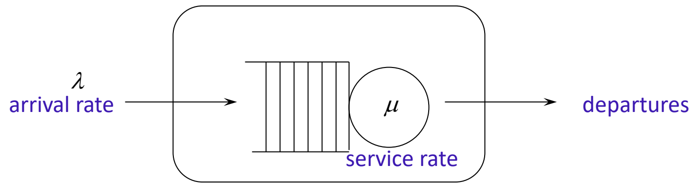
Single server center



Terminology
- service time S $\quad\quad S=1/\mu$
- time in the queue T_Q waiting time
- time in the system T_S residence/response time
- number in the system N_S
- number in the queue N_Q
- number in the service U, ρ

2

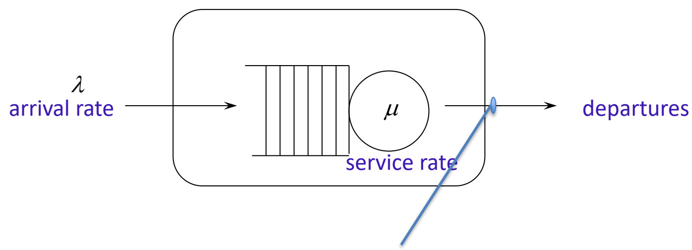
Single server center



E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

1. As λ, the mean arrival rate, increases, all the performance metrics mentioned above increase.
2. As μ, the mean service rate, increases, all the performance metrics mentioned above decrease.

Single server center

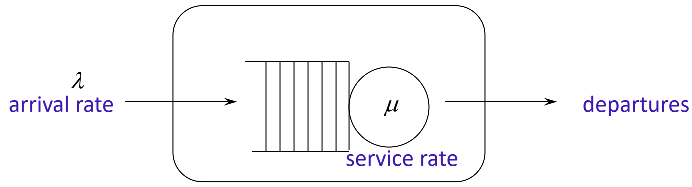


E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

Def. throughput

t=1, E(n)$_1$ n° of completions (departures) in the time unit

Single server center



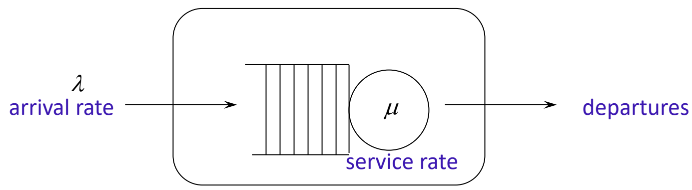
Def. utilization

How can we “mathematically” define the utilization?

$$\rho = \lambda / \mu$$

5

Single server center

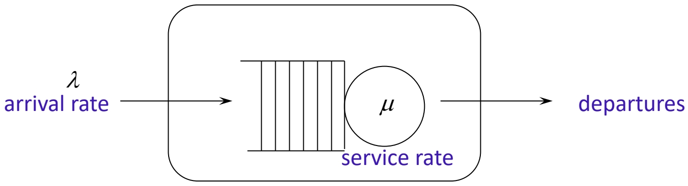


E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

$$E(T_S) = E(T_Q) + E(S)$$

6

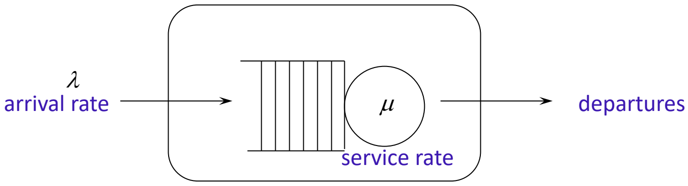
Single server center



E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

$$E(N_S) = E(N_Q) + E(number\ in\ service)$$

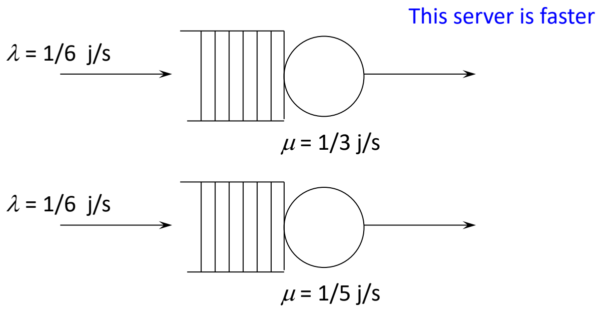
Single server center



E(T_Q), E(T_S), E(N_S), E(N_Q), Prob{ $T_S > t$ }, E(n)$_t$

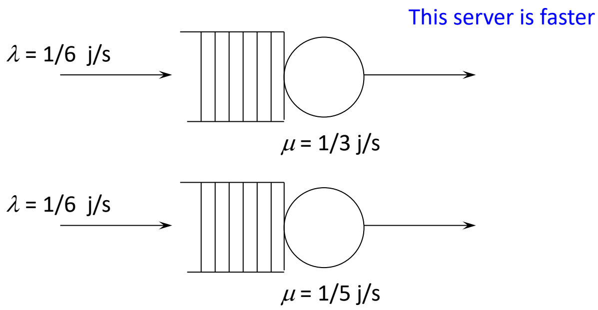
$$E(N_S) = E(N_Q) + \rho$$

Single server center



Which system has greater throughput?

Single server center



By assuming *job flow balance*, the throughput is the same !!
For both systems $X = \lambda$ = 1/6 j/s

BUT the faster server shows the shorter queue and so shorter mean response time

In other words, improving the mean response time does not necessarily improve the throughput

Single server center

If the center is in stochastic equilibrium (stationary condition),

$$\lambda < \mu, \quad \rho = \lambda / \mu < 1$$

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

Throughput is independent of the service rate μ

If the center is NOT in stochastic equilibrium,

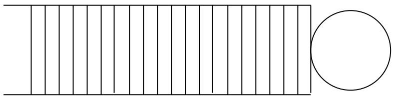
$$\lambda > \mu,$$

$E(n)_1 = X = \mu$

the center cannot work off the arrival rate, the queue grows unlimited

Single server center

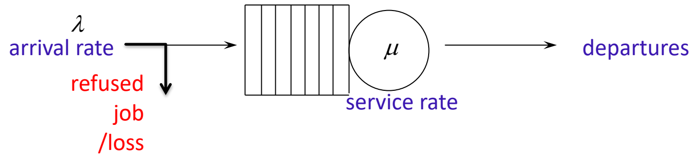
What's up if $\lambda > \mu$?



the center cannot work off the arrival rate, the queue grows unlimited

$$E(N_Q \text{ in } T) \geq \lambda T - \mu T = T(\lambda - \mu) \rightarrow \infty \text{ as } T \rightarrow \infty$$

Single server center with finite buffer



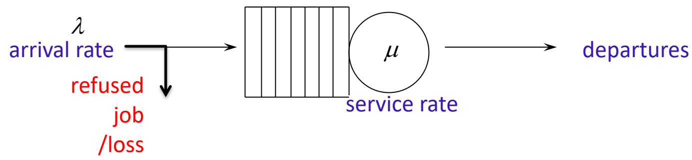
Each arrival when the queue is *full* will be lost

Which is the throughput?

13

Analytical models
basic laws

Single server center with finite buffer



Each arrival when the queue is *full* will be lost

Which is the throughput?

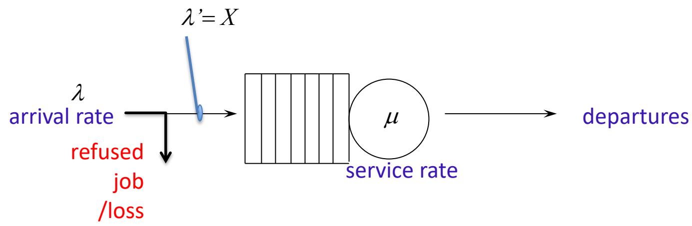
$X \cancel{=} \lambda$ $\cancel{\rho = \lambda / \mu}$

No!
On the contrary

$X < \lambda$

14

Single server center with finite buffer



Each arrival when the queue is *full* will be lost

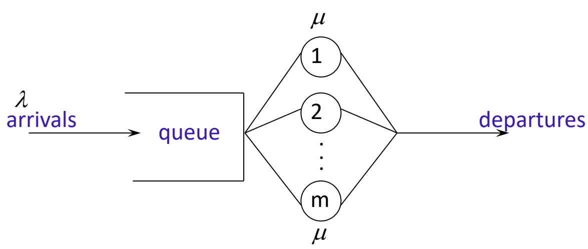
Which is the throughput?

$X \neq \lambda$ $\quad$ $\cancel{\rho = \lambda / \mu}$

No!
On the contrary

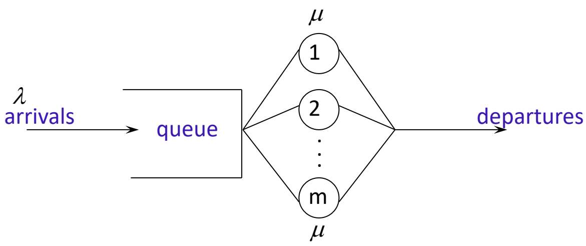
$X < \lambda$

Multi Server Queue



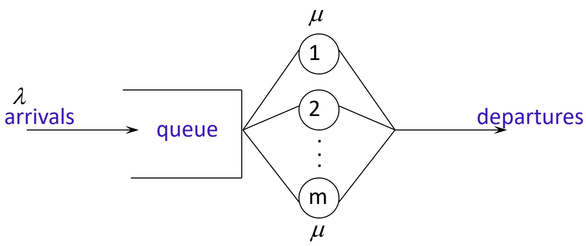
$E(S_i)$ $\qquad$ $E(S)$

Multi Server Queue



$$E(S_i) = 1/\mu \qquad E(S) = 1/m\mu = E(S_i)/m$$

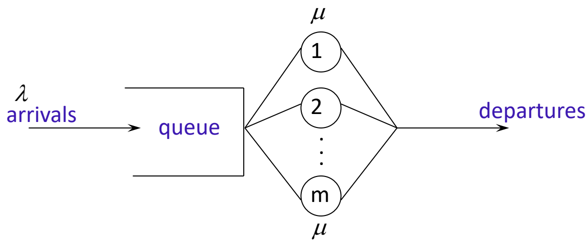
Multi Server Queue



$$E(N_S) = \begin{cases} E(N_Q) + \rho & if\ m = 1 \\ E(N_Q) + m\rho & f\ m > 1 \end{cases}$$

But how is the utilization defined for the multiserver case?

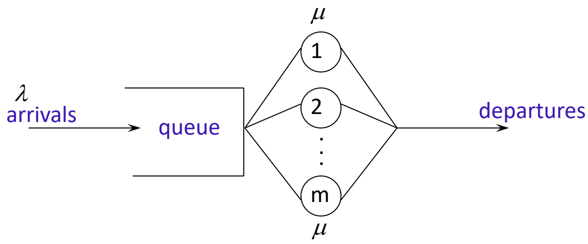
Multi Server Queue



$$\rho_i = \frac{\lambda_i}{\mu} = \frac{\lambda}{m\mu} \qquad \rho_{glob} = \frac{\lambda}{\mu_{glob}} = \frac{\lambda}{m\mu}$$

19

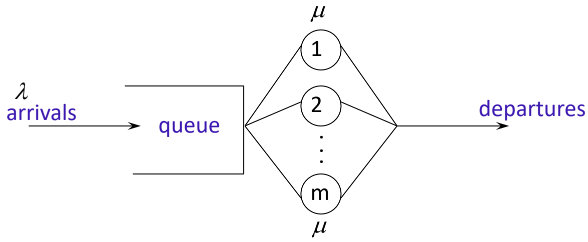
Multi Server Queue



$$\rho = \begin{cases} \frac{\lambda}{\mu} = \lambda E(S_i) & \text{if } m = 1 \\ \frac{\lambda}{m\mu} = \frac{\lambda E(S_i)}{m} & \text{if } m > 1 \end{cases}$$

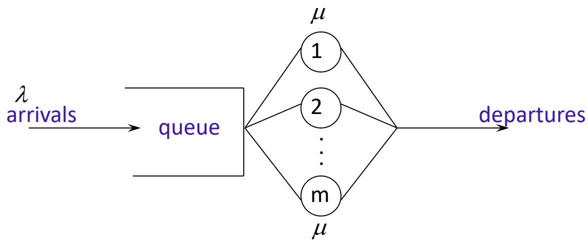
20

Multi Server Queue



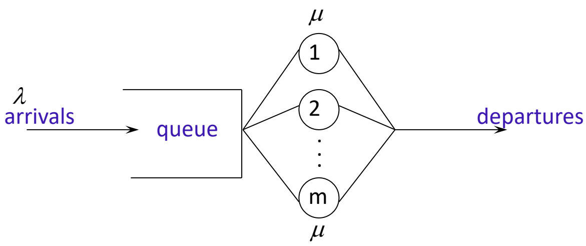
$$\rho_i = \rho_{glob} = \frac{\lambda}{m\mu}$$

Multi Server Queue



$$E(T_S) = E(T_Q) + E(S_i)$$

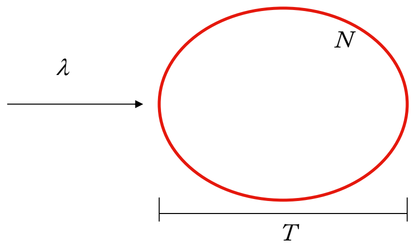
Multi Server Queue



$$E(T_S) = E(T_Q) + E(S_i)$$

$$E(T_Q) = f(\lambda, \mathrm{m}, E(\mathrm{S_i}))$$

Little's law is very important for its broad applicability.
In general, we can see Little's law as applied at a black box:
it states relations between mean values



Little's Law (1961)

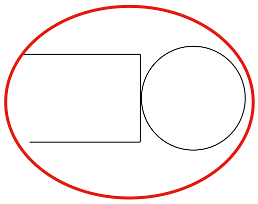
(a) queue discipline is FIFO,

(b) service node capacity is infinite,

(c) flow balance

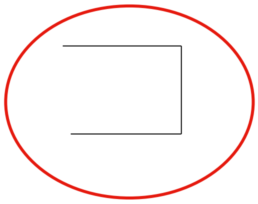
$$N = \lambda T$$

If λ is the mean arrival rate, T is the mean residence time in the black box, N is the mean population in the black box, the theorem asserts that, if the system is "stable", the mean population is given by the "mean arrival flow" multiplied the mean time the jobs spend in the black box



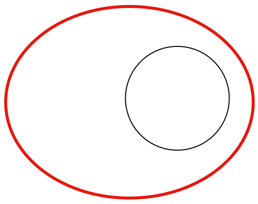
If the black box is the whole center, the theorem is applied to the center mean population:

$$E(N_S) = \lambda E(T_S)$$



If the black box is just the queue, the theorem is applied to the queue mean population:

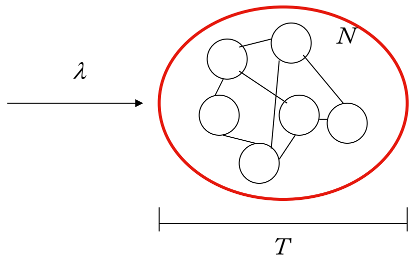
$$E(N_Q) = \lambda E(T_Q)$$



If the black box is just the server, the theorem is applied to the server "mean population", in other words to the utilization:

$$\rho = \lambda E(S)$$

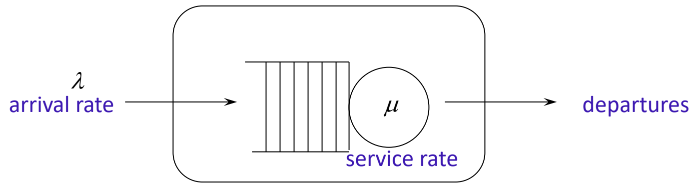
But if the black box is a network of centers, anyway interconnected,



$$N = \lambda T$$

The theorem is applied to the entire network!!

Single server center



$$E(T_S) = E(T_Q) + E(S)$$

$$E(N_S) = E(N_Q) + \rho$$

Little's law

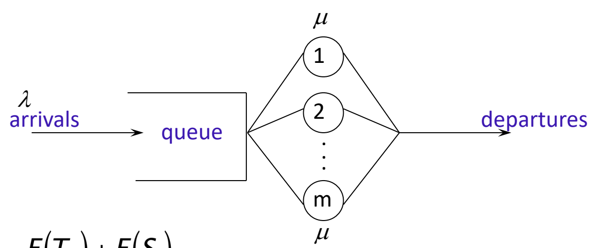
$$E(N_S) = \lambda E(T_S)$$

$$E(N_Q) = \lambda E(T_Q)$$

$$E(T_S) = \frac{E(N_S)}{\lambda}$$

$$E(T_Q) = \frac{E(N_Q)}{\lambda}$$

Multi Server Queue



$$E(T_S) = E(T_Q) + E(S_i)$$

Little's law

$$E(N_S) = \lambda E(T_S)$$

$$E(N_Q) = \lambda E(T_Q)$$

$$E(T_S) = \frac{E(N_S)}{\lambda}$$

$$E(T_Q) = \frac{E(N_Q)}{\lambda}$$

Consider a web server with a mean processing rate of 1.2 job/s.
If the server receives requests with a rate of 0.45 job/s and it has 0.225 enqueued jobs on average, determine:

a) the average utilization
b) the average response time.

During rush hours the arrival rate grows of 20% and the average number of enqueued jobs becomes 0.3681818.
Determine:
c) the performance metrics a) and b)
d) which further increasing in arrival rate makes the server collapsing
e) the performance metrics a) and b) for the limiting case d).

Let us consider a server that processes jobs with rate 0.8 jobs/s.
By assuming that the server receives jobs with a rate depending on the time slot as follows:

- 8.00 a.m. – 12.00 a.m. average arrival rate 1.5 jobs/s
- 12.00 a.m. – 2.00 p.m. average arrival rate 0.5 jobs/s
- 2.00 p.m. – 7.00 p.m. average arrival rate 1.5 jobs/s
- 7.00 p.m. – 9.00 p.m. average arrival rate 0.5 jobs/s
- 9.00 p.m. – 8.00 a.m. average arrival rate 0.05 jobs/s

Determine:

a) average arrival rate per day (24 hours)
b) average utilization per day
c) average throughput per day
d) average throughput for each time slot

Please, justify and comment the results by indicating the used laws.