Performance Modeling
of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Performance Sensitivity to the Service time distribution

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

Copyright © Vittoria de Nitto Personè, 2021
https://creativecommons.org/licenses/by-nc-nd/4.0/ (CC BY-NC-ND 4.0)

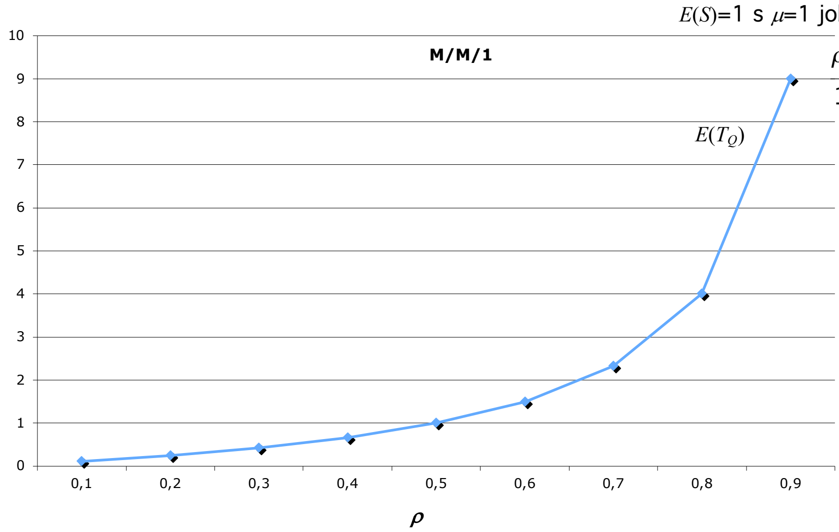
1

Analytical models
M/G/1

Expected waiting time as load grows

$E(S)$=1 s μ=1 job/s

$$\frac{\rho E(S)}{1-\rho}$$



What's happen if the service time distribution changes?

2

The Khinchin Pollaczek equation (KP)

$$E(N_Q) = \frac{\rho^2}{2(1-\rho)}\left[1 + C^2\right], \qquad E(T_Q) = \frac{\rho}{1-\rho}\frac{C^2+1}{2}E(S)$$

$$C^2(S) = \frac{\sigma^2(S)}{E^2(S)}$$

Expected waiting time in an M/G/1 queue can be huge, even under very low utilization ρ, if C^2 is huge.

D $\longrightarrow$ $C^2=0$

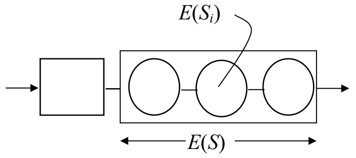
M $\longrightarrow$ $C^2=1$

$E_k \longrightarrow C^2 = \frac{1}{k}$

$H_2 \longrightarrow C^2 = g(p) = \frac{1}{2p(1-p)} - 1$

Expected waiting time as load grows: Erlang case

$E(S)$=0.5 s μ=2 job/s



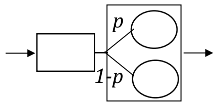
$$E(S_i) = \frac{0.5}{3} = 0.166666666 \quad \text{s}$$

$$\sigma^2(S) = \frac{1}{k}\left(\frac{1}{\mu}\right)^2 = 0.0833333$$

Expected waiting time as load grows: Hyperexponential case

$E(S)$=0.5 s μ=2 job/s

$2p\mu$=0.8 job/s



p=0.2

$2(1-p)\mu$=3.2 job/s

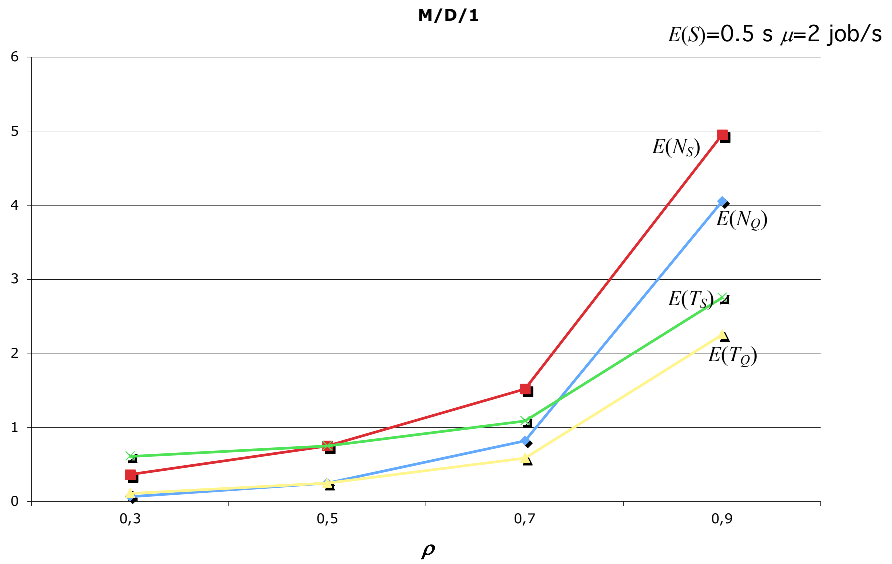
$$\sigma^2(S) = g(p)\left(\frac{1}{\mu}\right)^2 = 0.53125 \qquad g(p) = \frac{1}{2p(1-p)} - 1 = 2.125$$

The Khinchin Pollaczek equation (KP)

$$g(p) = \frac{1}{2p(1-p)} - 1 \qquad E(N_Q) = \frac{\rho^2}{2(1-\rho)}\left[1 + C^2\right], \quad E(T_Q) = \frac{\rho}{1-\rho}\,\frac{C^2+1}{2}\,E(S)$$

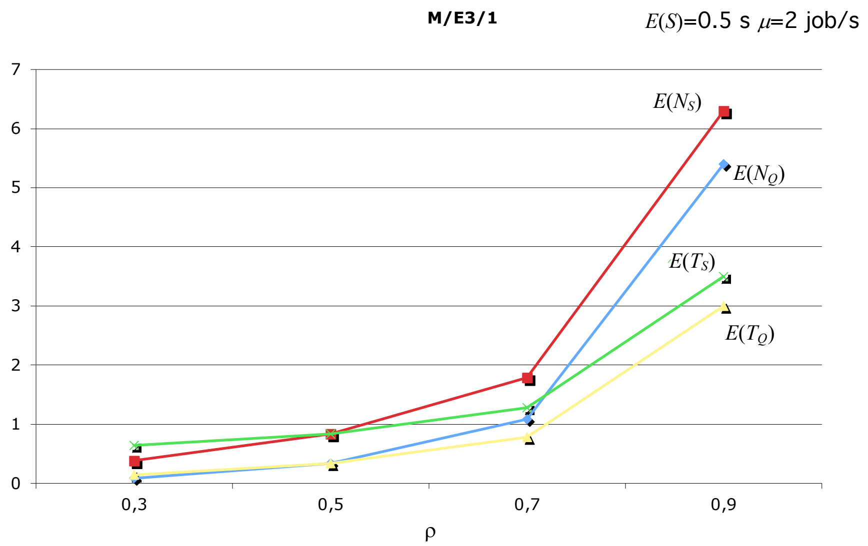
Service time	$E(N_Q)$	$E(T_Q)$
Determinisctic, M/D/1	$\frac{\rho^2}{2(1-\rho)}$	$\frac{\rho E(S)}{2(1-\rho)}$
Markovian, M/M/1	$\frac{\rho^2}{1-\rho}$	$\frac{\rho E(S)}{1-\rho}$
K-Erlang, M/E$_k$/1 $\sigma^2(S) = \frac{E(S)^2}{k}$	$\frac{\rho^2}{2(1-\rho)}\left(1+\frac{1}{k}\right)$	$\frac{\rho E(S)}{2(1-\rho)}\left(1+\frac{1}{k}\right)$
Hyperexpo, M/H$_2$/1 $\sigma^2(S) = E(S)^2 g(p)$	$\frac{\rho^2}{2(1-\rho)}\left(1+g(p)\right)$	$\frac{\rho E(S)}{2(1-\rho)}\left(1+g(p)\right)$

Expected waiting time as load grows: Deterministic case



7

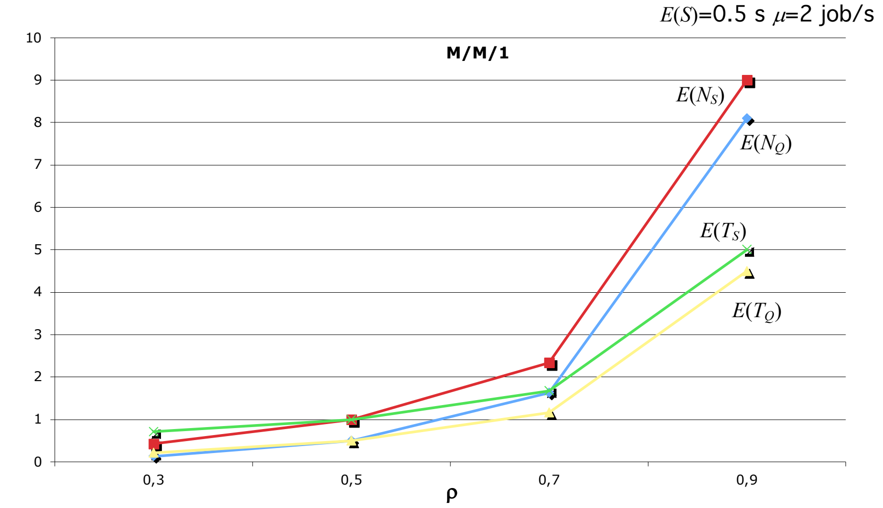
Expected waiting time as load grows: Erlang case



8

Expected waiting time as load grows: Exponential case

$E(S)$=0.5 s μ=2 job/s

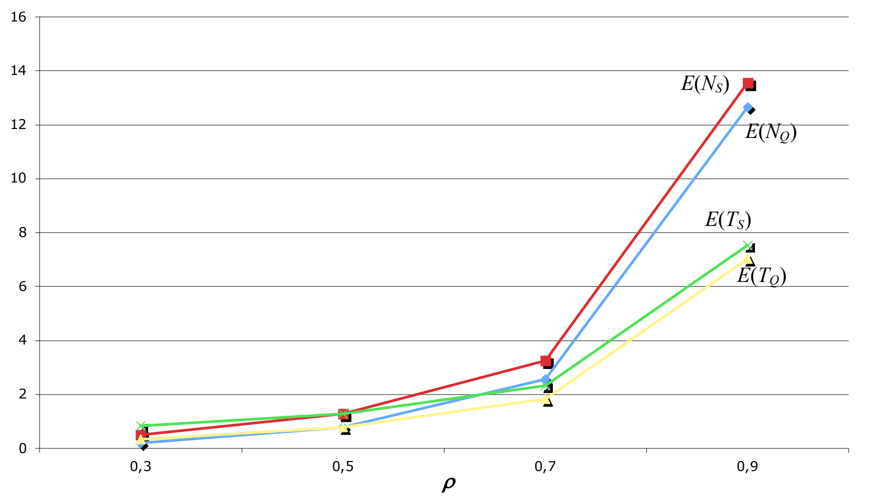


9

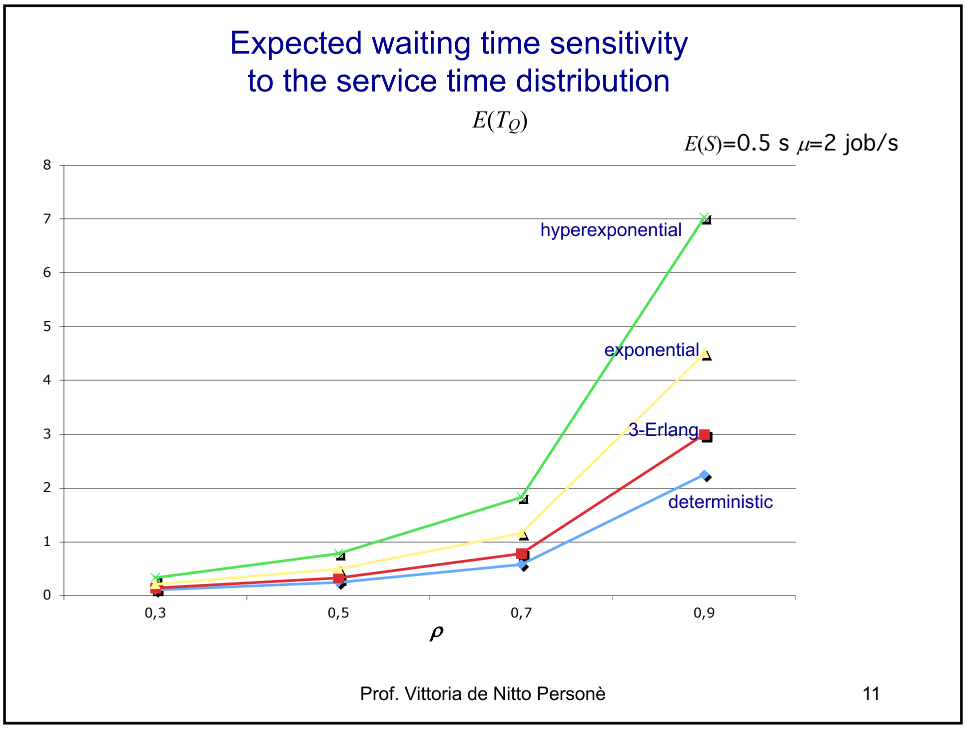
Expected waiting time as load grows: Hyperexponential case

M/H2/1

$E(S)$=0.5 s μ=2 job/s

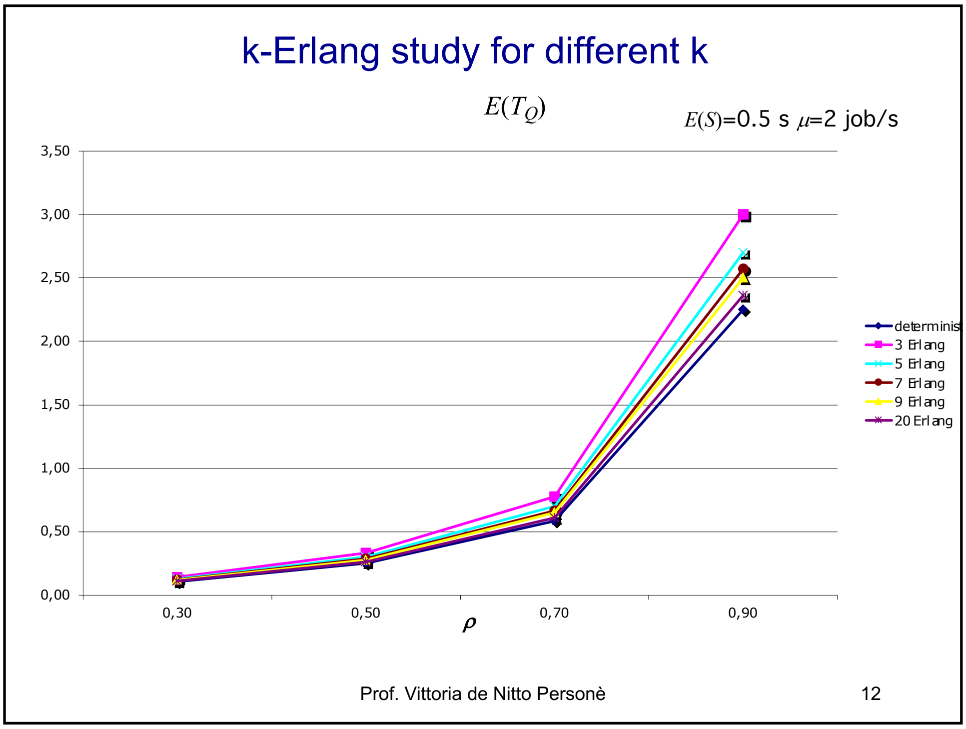


10



Expected waiting time sensitivity
to the service time distribution
$E(T_Q)$
$E(S)$=0.5 s μ=2 job/s
8
7
6
5
4
3
2
1
0
hyperexponential
exponential
3-Erlang
deterministic
0,3
0,5
0,7
0,9
ρ
Prof. Vittoria de Nitto Personè
11

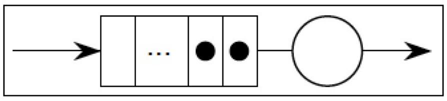
11



k-Erlang study for different k
$E(T_Q)$
$E(S)$=0.5 s μ=2 job/s
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
0,30
0,50
0,70
0,90
ρ
determinis
3 Erlang
5 Erlang
7 Erlang
9 Erlang
20 Erlang
Prof. Vittoria de Nitto Personè
12

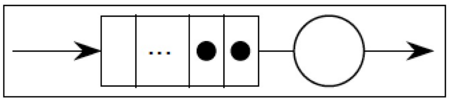
12

A TP system accepts and processes a stream of transactions, mediated through a (large) buffer:



- Transactions arrive "randomly" at some specified rate
- The TP server is capable of servicing transactions at a given service *rate*

Q: If both the arrival rate and service rate are doubled, what happens to the mea response time?



- The arrival rate is 15tps
- The mean service time per transaction is 58.37ms

Q: What happens to the mean response time if the arrival rate increases by 10%?

$$E(T_Q) = \frac{\rho}{1-\rho}\frac{C^2+1}{2}E(S)$$

$$E(T_{Q'}) = \frac{\rho'}{1-\rho'}\frac{C^2+1}{2}E(S)$$

$$\frac{E(T_Q)}{E(T_{Q'})} \cong 0,27 \cong \frac{1}{3,7}$$

Heavy tail distributions properties

esponenziale $\longrightarrow$ memoryless
failure rate costante

Heavy tail $\longrightarrow$ failure rate decrescente

(Pareto: $r(x) = \alpha / x$, $x > 1$)

Where they are

Jobs Unix

Sizes files websites $\alpha \approx 1.1$

Internet topology

Packet n° IP flows 1% → 50%

Natural phenomena

Pareto

$$f(x) = \alpha k^{\alpha} x^{-\alpha-1} \quad k \le x < \infty$$

α , parametro di forma

$$E[X] = \frac{\alpha k}{\alpha - 1} \quad \alpha > 1$$

$$\sigma^2[X] = \frac{\alpha k^2}{(\alpha-1)^2(\alpha-2)} \quad \alpha > 2$$

Bounded Pareto

$$f(x) = \alpha x^{-\alpha-1} \frac{k^{\alpha}}{1-\left(\frac{k}{p}\right)^{\alpha}} \quad k \le x \le p, 0 < \alpha < 2$$

(Vilfredo Pareto, 15 July 1848 – 19 August 1923, economista e sociologo)

Pareto

$$E(T_Q) = \frac{\rho}{1-\rho}\frac{C^2+1}{2}E(S)$$

$$C^2(S) = \frac{\sigma^2(S)}{E^2(S)}$$

$$E[T_Q] = \frac{\rho E[S]}{1-\rho}\frac{1+\alpha(\alpha-2)}{2\alpha(\alpha-2)}$$

$\alpha > 2$

Pareto study as load grows

$E(S)$=0.5 s μ=2 job/s

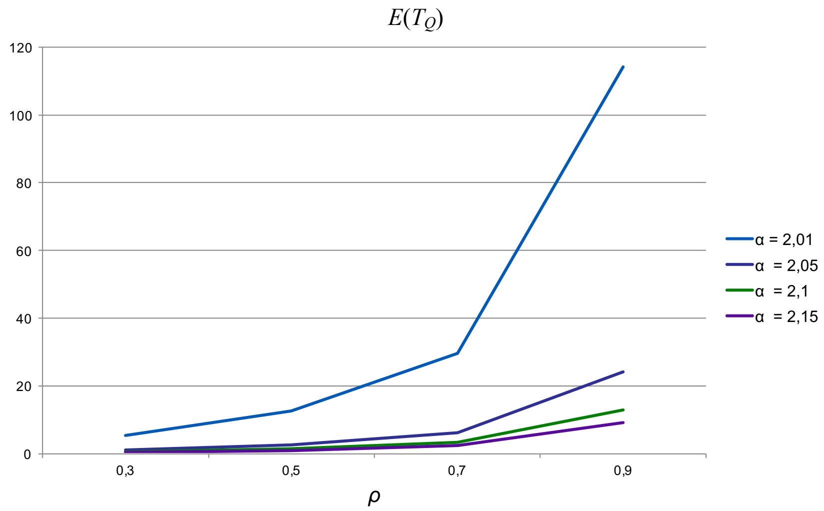
$E(T_Q)$

ro list	α = 2,01	α = 2,05	α = 2,1	α = 2,15	determ	3-Erlang	expo	hyper
0,3	5,437633262	1,152439024	0,617346939	0,439368771	0,107	0,142	0,213	0,333
0,5	12,68781095	2,68902439	1,44047619	1,025193798	0,25	0,333	0,5	0,781
0,7	30	6,274390244	3,361111111	2,392118863	0,583	0,778	1,167	1,823
0,9	114,1902985	24,20121951	12,96428571	9,226744186	2,25	3	4,5	7,031

k=0.2512 k=0.2619

$$E[S] = \frac{\alpha k}{\alpha-1} \Rightarrow k = \frac{\alpha-1}{\alpha}E[S]$$

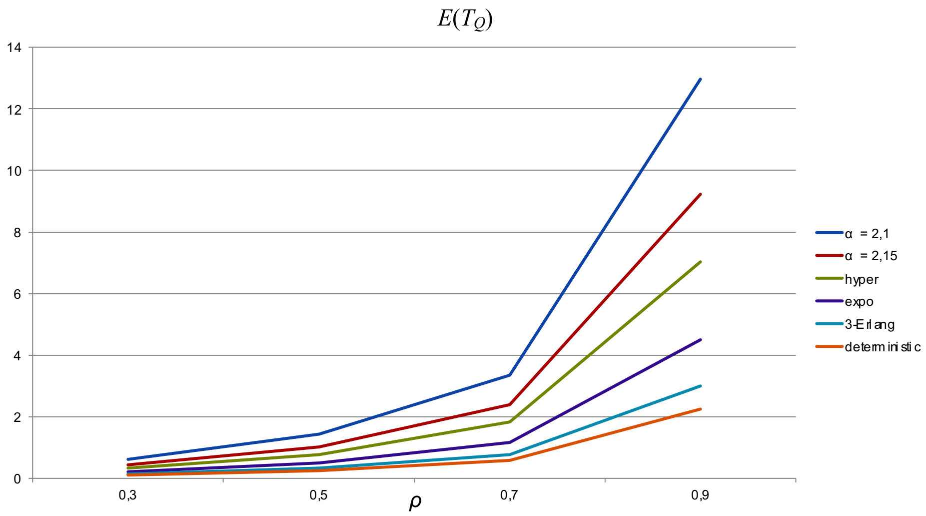
Pareto study as load grows



Prof. Vittoria de Nitto Personè 21

21

Pareto study as load grows



Prof. Vittoria de Nitto Personè 22

22