On overloaded queues

Alain Jean-Marie
INRIA et LIRMM, University of Montpellier 2
161 Rue Ada, 34392 Montpellier Cedex 5, France
ajm@lirmm.fr

Lunteren Conference
January 2005

Based on results obtained with Philippe Robert

Plan of the talk

The Single-Server Queue

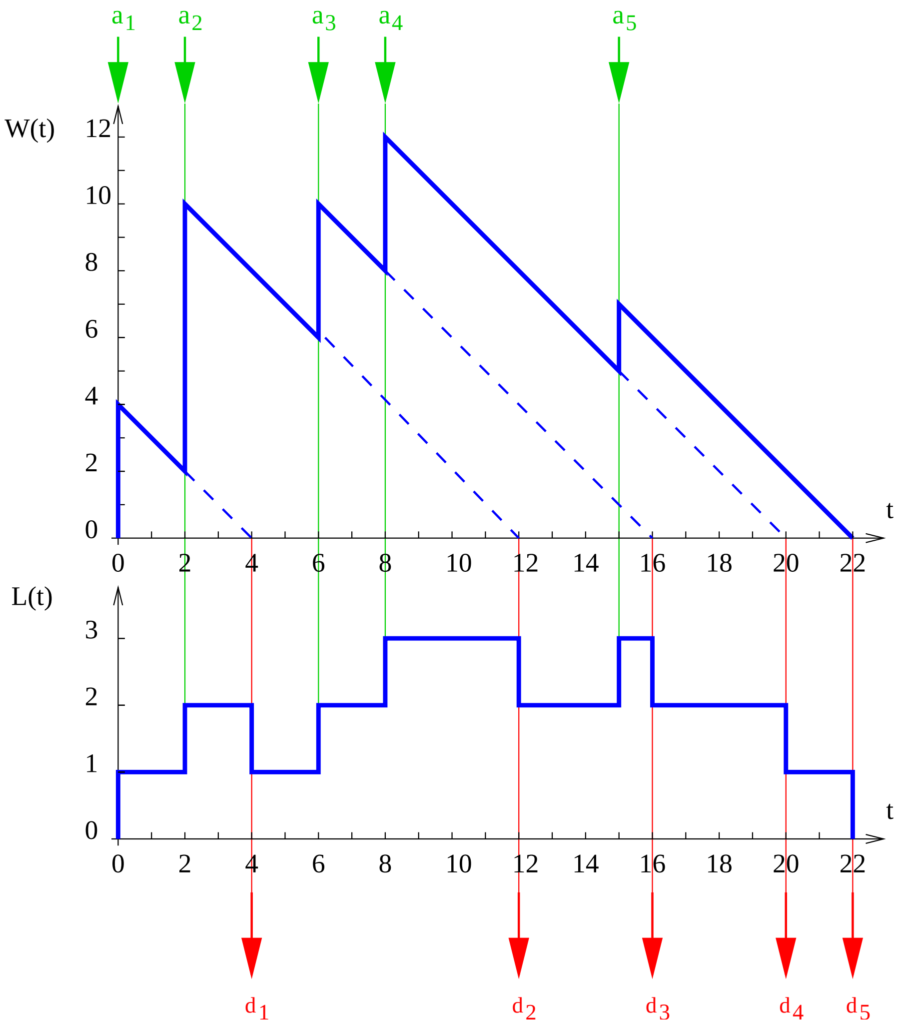
The basic $G/G/1/+\infty/D$ queue

- Arrival of customers, according to a stationary process with inter-arrival times $\tau_1, \tau_2, \ldots, \tau_n, \ldots$
- Service requirements $\sigma_1, \sigma_2, \ldots, \sigma_n, \ldots$
- One single server
- Infinite waiting room for customers
- Service discipline (scheduling) D.

State description of a queue

The state of the queue can be described by several evolving quantities:

- $A(t)$: number of arrivals up to time t
- $D(t)$: number of departures up to time t
- $L(t)$: number of customers at time t ("length" of the queue)
- $W(t)$: workload at time t, number of units of work left to do for the server
- $R(t) = (r_1(t), r_2(t), \ldots, r_{L(t)}(t))$, vector of the residual service times of customers that are in the queue.



a_1 a_2 a_3 a_4 a_5
W(t)
12
10
8
6
4
2
0
t
0 2 4 6 8 10 12 14 16 18 20 22
L(t)
3
2
1
0
t
0 2 4 6 8 10 12 14 16 18 20 22
d_1 d_2 d_3 d_4 d_5

Conservation laws

Conservation of the number of customers

$$L(t) \;=\; L(0) \;+\; A(t) \;-\; D(t)$$

Repartition of the workload

$$W(t) \;=\; \sum_{i=1}^{L(t)} r_i(t)$$

Conservation of work: for a work-conserving service discipline,

$$W(t) \;=\; \sum_{n=1}^{A(t)} \sigma_n \;-\; \int_0^t \mathbf{1}_{\{W(s)>0\}} \mathrm{d}s \;.$$

Stability

The case usually considered in Queueing Theory is when

$$L(t) \;\rightarrow\; L(\infty)\;, \qquad W(t) \;\rightarrow\; W(\infty)$$

in distribution as $t \rightarrow \infty$. Such a queue is called stable.

When stability occurs:

- the response times T_n of customers also have a stationary distribution,
- the queue empties ($\Longleftrightarrow$ the server becomes idle) infinitely often.

Stability condition

When does this happen? If inter-arrival times τ_n and service times σ_n are stationary sequences, there is stability if and only if

$$\boxed{\mathbb{E}(\sigma_0) \quad < \quad \mathbb{E}(\tau_0)\ .}$$

Equivalently,

$$\underset{\text{input rate}}{\lambda} \quad := \quad \frac{1}{\mathbb{E}(\tau_0)} \quad = \quad \lim_{t\to\infty} \frac{A(t)}{t} \quad < \quad \frac{1}{\mathbb{E}(\sigma_0)} \quad =: \quad \underset{\text{service capacity}}{\mu}$$

Q: What about the output rate of customers:

$$\theta := \lim_{t\to\infty} \frac{D(t)}{t} \qquad ?$$

A: it is equal to the input rate:

$$\begin{aligned}
\theta &= \lim_{t\to\infty} \frac{D(t)}{t} \\
&= \lim_{t\to\infty} \frac{A(t)}{t} - \frac{L(t)}{t} \\
&= \lim_{t\to\infty} \frac{A(t)}{t} \\
&= \lambda \, .
\end{aligned}$$

Plan of the talk

Unstable queues

What happens when $\lambda > \mu$?

The queue is overloaded: too much work arrives. Also called unstable, transient, . . .

The number of customers waiting grows, the queue "explodes".

The waiting time of customers tends to grow with time.

...

Does it really?

How fast does it grow? How bad is it?

The answers turn out to depend (only) on the service discipline

General properties

Some general properties can be stated for an overloaded queue:

Properties

- the workload $W(t)$ goes to infinity almost surely,
- its growth rate is

$$\lim_{t\to\infty} \frac{W(t)}{t} = \frac{\lambda - \mu}{\mu} = \frac{\lambda}{\mu} - 1 \, ,$$

- there exists almost surely a time t_0 such that the server is always busy after t_0:

$$W(t) > 0 \qquad \forall t > t_0 \, .$$

Unstable queues (cdt)

The situation for the number of customers $L(t)$ is not so clear:

- does $L(t) \to \infty$?
- if it does, is there a growth rate

$$\alpha \ := \ \lim_{t\to\infty} \frac{L(t)}{t} \ ?$$

- what is the output rate $\theta = \lim_t D(t)/t$? According to the conservation law of customers:

$$\lambda \ = \ \alpha \ + \ \theta \ .$$

Plan of the talk

The FIFO case

For a FIFO queue, we have:

Properties

- the growth rate of $L(t)$ is $\alpha = \lambda - \mu$
- the output rate θ is equal to μ
- the response time of customers grows linearly with time:

$$\lim_{n \to \infty} \frac{T_n}{n} \; = \; \lambda - \mu \; .$$

The FIFO case (ctd)

Key facts:

- All but one customer in the queue have their residual service times equal to their service times:

$$r_i(t) \;=_d\; \sigma_i \;, \qquad i = 2, \ldots, L(t) \;.$$

- The response time of a customer is related to the workload seen at its arrival epoch:

$$T_n \;=\; W(a_n) \;+\; \sigma_n \;.$$

Therefore, the workload is almost equally distributed among customers:

$$
\begin{array}{ccccccc}
W(t) & = & r_1(t) & + & \displaystyle\sum_{i=2}^{L(t)} \sigma_i & & \\
\dfrac{W(t)}{t} & = & \dfrac{r_1(t)}{t} & + & \dfrac{L(t)}{t} \dfrac{1}{L(t)} \displaystyle\sum_{i=2}^{L(t)} \sigma_i & & \\
\downarrow & & \downarrow & & \downarrow & & \\
\dfrac{\lambda}{\mu} - 1 & = & 0 & + & \alpha \mathbb{E}(\sigma_0) & = & \dfrac{\theta}{\mu}
\end{array}
$$

For response times:

$$
\frac{R_n}{n} = \frac{a_n}{n} \frac{W(a_n)}{a_n} \to \lambda \frac{\lambda - \mu}{\mu} .
$$

Extensions of the FIFO case

The key property: "all customers but one have a residual service time equal to σ in distribution" holds for other service disciplines:

- ROS: random order of service
- LIFO non-preemptive
- ...?

Plan of the talk

The overloaded Processor Sharing queue

Under the Processor Sharing discipline, the server serves each of the $L(t)$ customers at rate $1/L(t)$.

Recall: vector of residual service times

$$R(t) \;=\; (r_1(t), r_2(t), \ldots, r_{L(t)}(t)) \;.$$

As long as no arrival occurs and all $r_i(t)$ remain positive:

$$\frac{\mathrm{d}r_i(t)}{\mathrm{d}t} \;=\; -\;\frac{1}{L(t)} \;.$$

Growth rates

Properties

- the growth rate of $L(t)$ is α, unique positive solution of:

$$x \; = \; \lambda \left(1 - E(e^{-x\sigma_0})\right) \; ,$$

- the response time of the n-th customer, grows linearly with n: given its service time,

$$\frac{T_n}{n} \quad \xrightarrow{n \to +\infty} \quad \frac{(e^{\alpha\sigma_0} - 1)}{\lambda} \; ,$$

- the output rate θ is solution of:

$$y \; = \; \lambda \; E(e^{-(\lambda - y)\sigma_0}) \; .$$

A "proof"

Idea of the proof: consider a customer with service time σ_n arriving at time a_n. Its response time T_n is such that:

$$
\begin{aligned}
\sigma_0 &= \int_{a_n}^{a_n+T_n} \frac{1}{L(u)} \mathrm{d}u \\
&\simeq \int_{a_n}^{a_n+T_n} \frac{1}{\alpha u} \mathrm{d}u \\
&= \frac{1}{\alpha} \log\left(\frac{a_n + T_n}{a_n}\right) \\
\Longrightarrow \quad T_n &\simeq a_n \left(e^{\alpha\sigma_0} - 1\right) .
\end{aligned}
$$

Consider now the number of customers still present at time t.

Customer n with $a_n \le t$ and service time σ, is still there if

$$\begin{aligned} a_n + T_n \ \ge \ t \quad & \overset{\sim}{\Longleftrightarrow} \quad a_n \, e^{\alpha\sigma} \ \ge \ t \\ & \Longleftrightarrow \quad a_n \ \ge \ t \, e^{-\alpha\sigma} \ . \end{aligned}$$

Therefore, since $a_n \cong \lambda n$, there are approximately

$$\lambda t \ - \ \lambda t \, e^{-\alpha\sigma} \ = \ \lambda t \ \left(1 \ - \ e^{-\alpha\sigma}\right) \text{ of these.}$$

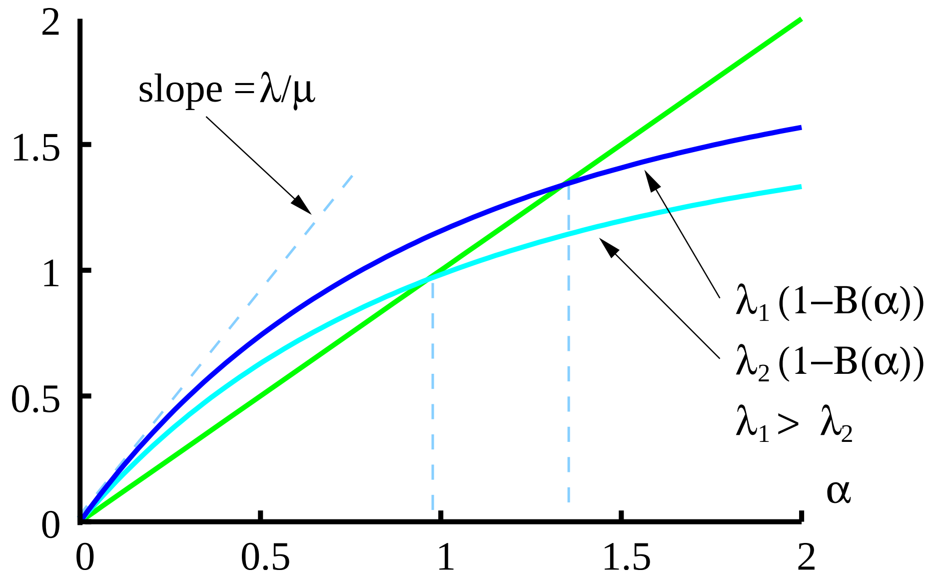
De-conditioning on σ, we get:

$$\begin{aligned} L(t) \quad & \cong \quad \lambda t \, \mathbb{E}\left(1 \ - \ e^{-\alpha\sigma}\right) \\ \Longrightarrow \qquad \alpha \quad & = \quad \lambda \ \left(1 \ - \ \mathbb{E}(e^{-\alpha\sigma})\right) \ . \end{aligned}$$

Input/Output rate relations

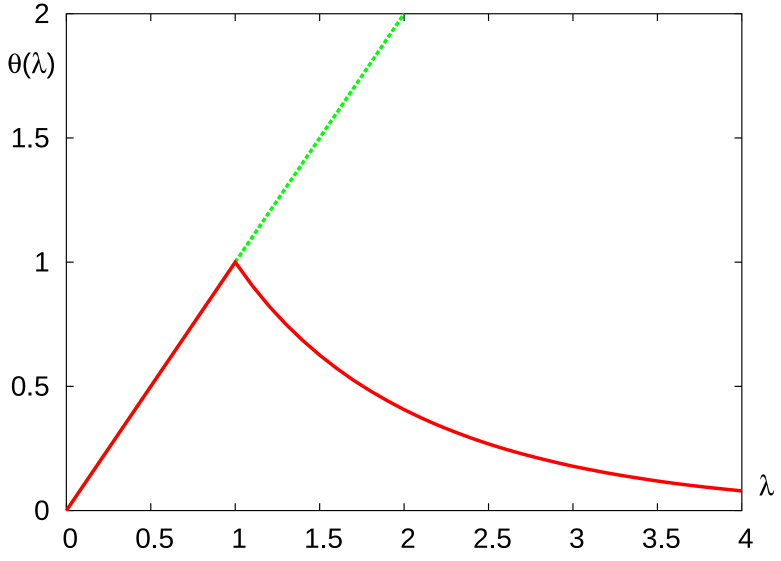
How does the output rate θ vary with λ?

The growth rate α of the queue is increasing with respect to λ:

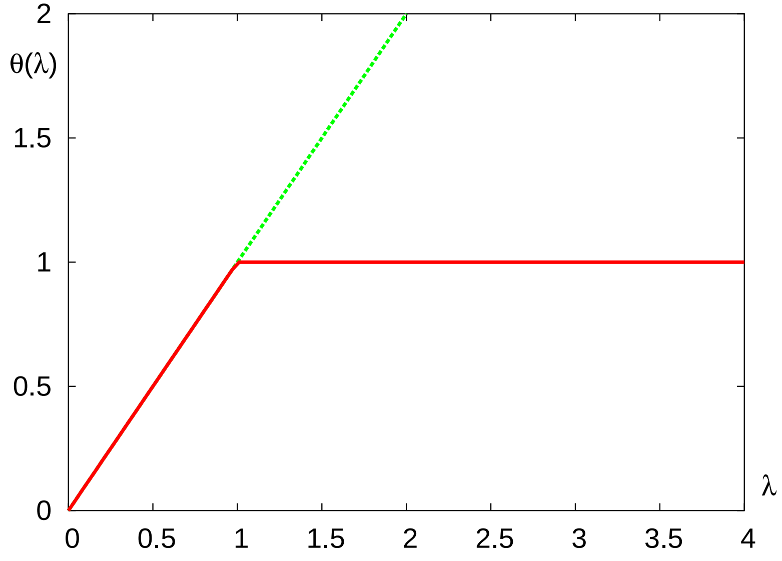


Input/Output rate relations (ctd)

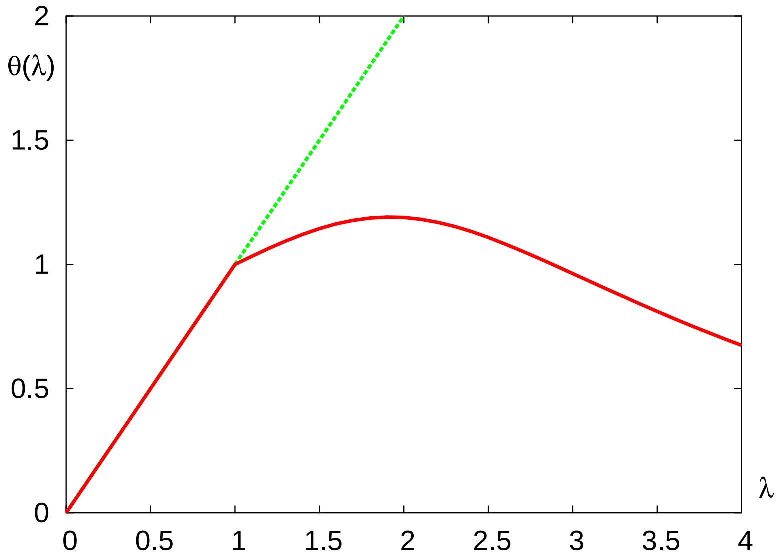
It is also true that $\alpha(\lambda)/\lambda$ is increasing, and $\theta(\lambda)/\lambda$ decreasing.
However: the output rate $\theta(\lambda)$ is not monotone, nor convex.

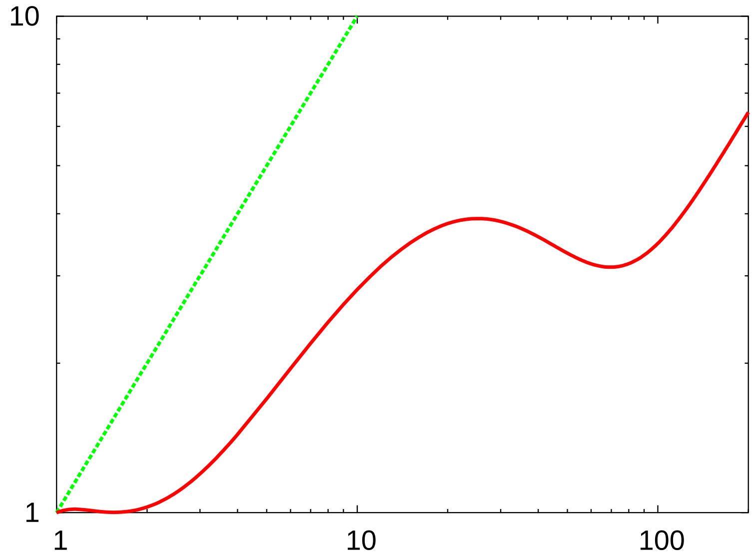


$./D/1/\infty/PS$



$./M/1/\infty/PS$





$$\sigma = \begin{cases} 1/2 & \text{wp } 8/9 \\ 5 & \text{wp } 1/9 \end{cases}$$

$$\sigma = \begin{cases} 0 & \text{wp } 4/125 \\ 18 & \text{wp } 1/250 \\ 3/2 & \text{wp } 4399/7250 \\ 1/20 & \text{wp } 259/725 \end{cases}$$

On Elephants and Mice

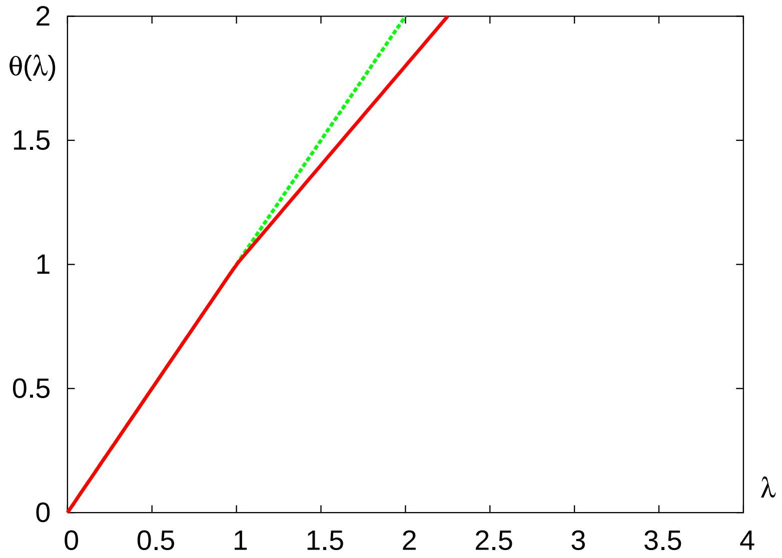
Consider the service time distribution:

$$\sigma \;=\; \begin{cases} 0 & \text{wp } 1-\varepsilon & \text{lots of mice} \\ Exp(\varepsilon\mu) & \text{wp } \varepsilon & \text{few elephants} \end{cases}$$

It has mean $\dfrac{1}{\mu}$ and variance $\left(\dfrac{2}{\varepsilon}-1\right)\dfrac{1}{\mu^2}$.

In this case:

$$\theta(\lambda) \;=\; \lambda - \varepsilon(\lambda-\mu) \qquad \alpha(\lambda) \;=\; \varepsilon(\lambda-\mu)\;.$$



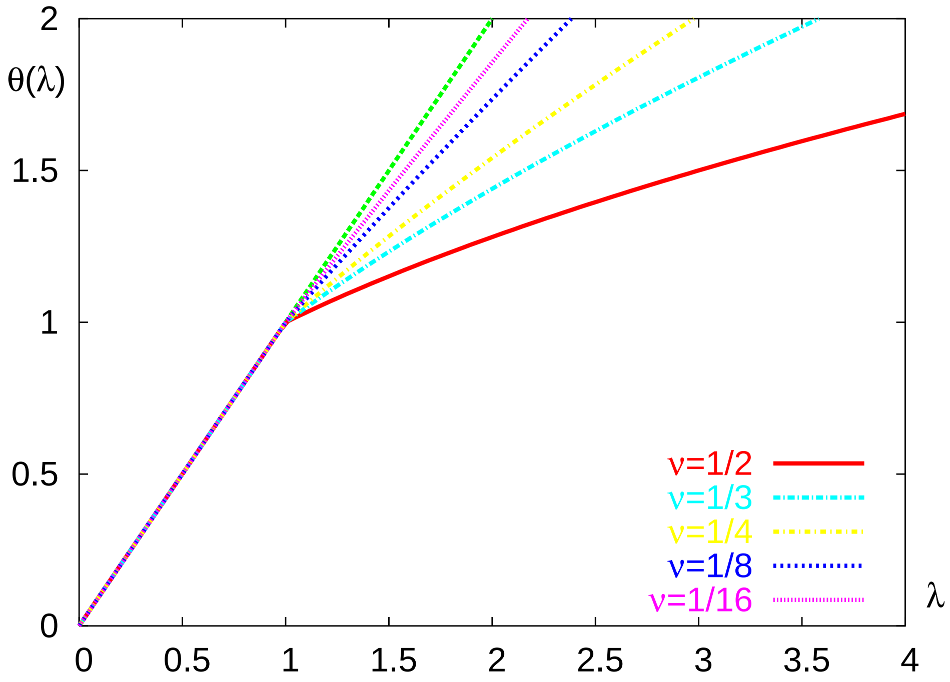
When $\varepsilon \longrightarrow 0$, the output rate remains close to λ. The customers accumulate, but at a very small rate!

$\Longrightarrow$ application to the TCP protocol, see Bonald and Roberts (2003).

Asymptotic Behavior

The assumption that $\sigma = 0$ is not essential. What is relevant is the density close to 0 of the service time distribution.

$\mathrm{d}P(\sigma \leq x)/\mathrm{d}x$ $x \to 0$	$B^*(s)$ $s \to \infty$	$\theta(\lambda)$ $\lambda \to \infty$
$o(x)$	$o(s^{-1})$	0
Ax	$\dfrac{A}{s}$	A
$> O(x)$	$< o(s^{-1})$	$+\infty$



$$\sigma \sim \mathsf{Gamma}(\mu;\nu) \qquad\qquad \mathbb{E}(e^{-\sigma s}) = \left(\frac{\nu\mu}{\nu\mu+s}\right)^{\nu}.$$

A stronger result on Residual Service Times

Residual service times also converge:

Property For any continuous function $f : \mathbb{R}_+ \longrightarrow \mathbb{R}$, almost surely

$$\begin{aligned}\lim_{t\to+\infty} \frac{1}{t}\sum_{i=1}^{L(t)} f(r_i(t)) &= \lambda\mathbb{E}\left(\int_0^{\sigma_0} f(x)\alpha e^{-\alpha(\sigma_0-x)}\mathrm{d}x\right)\\ &= \lambda\int_0^{\infty}\int_0^{u} f(x)\alpha e^{-\alpha(u-x)}\mathrm{d}x\mathrm{d}\mathbb{P}\{\sigma_0\le u\}\ ,\end{aligned}$$

provided that $\mathbb{E}\left(\sup_{x\le\sigma_0}|f(x)|\right) < +\infty$. Moreover the result is valid for all the indicator functions of intervals.

Other Oddities: Response Times

We have seen that the distribution of T_n behaves as:

$$\frac{T_n}{n} \quad \xrightarrow{n \to +\infty} \quad \frac{(e^{\alpha\sigma_0} - 1)}{\lambda} \ ,$$

In expectation,

$$\mathbb{E}(T_n) \ \simeq \ \frac{n}{\lambda} \left(\mathbb{E}(e^{\alpha\sigma_n}) - 1 \right) \ .$$

For instance, for service times $\sigma_n \sim Exp(\mu)$: $\alpha = \lambda - \mu$ and

$$\mathbb{E}(T_n) \ \simeq \ \frac{n}{\lambda} \frac{\lambda - \mu}{2\mu - \lambda} \ .$$

This is infinite if $\lambda \geq 2\mu$!! A consequence of results by Coffman, Muntz and Trotter (1970).

Plan of the talk

Introduction

General properties of overloaded queues

The FIFO Case

The Overloaded Processor Sharing Queue

Other Service Disciplines

Final word

The preemptive LIFO case

Properties in the preemptive LIFO case:

- the growth rate of $L(t)$ is α_2 such that:

$$\alpha_2 = \lambda \mathbb{P}\{W_0^{\mathcal{F}} \;=\; 0\}$$

where $W_n^{\mathcal{F}}$ are the waiting times of customers in the "dual" stable FIFO queue with:

$$\tau_n^{\mathcal{F}} \;=\; \sigma_{-n} \qquad \sigma_n^{\mathcal{F}} \;=\; \tau_{-n} \;,$$

- the output rate θ is equal to $\lambda - \alpha_2$,

- the other customers remain forever in the queue!

Priority queues case

Assume fixed priorities $1 \succ 2 \succ 3 \succ \ldots$, arrival rates λ_k and per-class workload ρ_k.

Properties There exists a priority level k such that

- the queue of customers of class $1, 2, \ldots, k-1$ is stable,
- the queue of customers of class k grows at rate $\alpha_k = \lambda_k \left(1 - \sum_{j=1}^{k-1} \rho_j\right)$
- customers of class $k+1, \ldots$ never enter service and accumulate at rate λ_j.

Similar results hold for the SPT and SRPT disciplines, see Bansal and Harchol-Balter (2001). Preemption does not make a difference.

Plan of the talk

Open questions

Various open issues:

- what about networks of queues?
- what about weighted processor sharing and variants (head-of-the line PS, Fair Queuing, . . .)?
- what about threshold-based disciplines, Foreground-Background, Earliest-Deadline-First, . . . ?
- . . .

Bibliography

JEAN-MARIE, A. AND ROBERT, PH. On the transient behavior of the processor sharing queue. *QUESTA*, 17 (1994), 129–136.

COFFMAN, E. G. JR, MUNTZ, R. AND TROTTER, H. Waiting time distributions for processor-sharing systems. *JACM*, 17 (1970), 123–130.

YASHKOV, S. On a heavy traffic limit theorem for the M/G/1 processor sharing queue. *Commun. Statist. - Stochastic Models 9*, 3 (1993), 467–471.

BONALD, T. AND ROBERTS, J.W. Congestion at flow level and the impact of user behaviour. *Computer Networks*, 42 (2003), 521–536.

BANSAL, N. AND HARCHOL-BALTER, M. Scheduling Solutions for Coping with Transient Overload. Technical Report #CMU-CS-01-134, School of Computer Science, Carnegie Mellon University, May 2001.