Scheduling and Queueing: Optimality under rare events and heavy loads

Bert Zwart
CWI

June 21, 2011

MAPSP

Consider a queue with

- Poisson λ arrivals
- Exponential μ service times, $\mu > \lambda$.
- A single server working according to FCFS discipline
- Let $\rho = \lambda/\mu$

For the steady-state waiting time W we know that

$$E[W] = \frac{\rho}{(1-\rho)\mu}$$

$$P(W > x) = \rho e^{-\mu(1-\rho)x}$$

Key questions

If we consider more general inter-arrival times and service times, it is impossible to compute $E[W]$ and $P(W > x)$ analytically. However, it still can be shown that, under some regularity conditions:

$$E[W] = \Theta\left(\left(\frac{1}{1-\rho}\right)^{\beta}\right), \qquad \rho \uparrow 1,$$

and for fixed ρ and $x \to \infty$,

$$P(W > x) = e^{-\gamma x(1+o(1))} \qquad \text{or} \qquad P(W > x) = \Theta(x^{-\alpha}).$$

How do α, β, γ depend on the scheduling discipline?

How do we choose a scheduling discipline that mitigates the effect of critical loading and the occurrence of long delays?

Overview

- Tail estimates for specific scheduling disciplines (FIFO, LIFO, PS, SRPT)

- Optimizing tail behavior when distribution is not known

- Scheduling under critical loading

The GI/GI/1 FIFO queue

Consider a GI/GI/1 FIFO queue with i.i.d. inter-arrival times (A_i), i.i.d. service times (B_i), working at speed 1. $\rho = E[B]/E[A] < 1$.

Let W be the steady-state waiting time. Well-known is:

$$W \stackrel{d}{=} \sup_{n \geq 0} S_n,$$

with $S_n = \sum_{i=1}^{n} X_i$ and $X_i = B_i - A_i$.

Main question: what is the behavior of

$$P(W > x) = P(\sup_{n \geq 0} S_n > x)$$

as $x \to \infty$?

Simple estimates

The following crude bounds turn out to be sharp enough!

$$P(S_n > x) \le P(\sup_n S_n > x) \le \sum_{n=0}^{\infty} P(S_n > x).$$

Upper bound: Let $u > 0$ be such that $E[e^{uX}] < 1$, and observe that

$$\sum_{n=0}^{\infty} P(S_n > x) \le \sum_{n=0}^{\infty} E[e^{uS_n}]e^{-ux} = \frac{1}{1 - E[e^{uX}]} e^{-ux}.$$

Define $\gamma_F = \sup\{u : E[e^{uX}] \le 1\}$.
Since the above bound is valid for all $u < \gamma_F$, we see that

$$\limsup_{x \to \infty} \frac{1}{x} \log P(W > x) \le -\gamma_F.$$

Lower bound: pick $n = xb$, with b cleverly chosen, and apply "Cramér".

Comments

- The limit

$$\lim_{x\to\infty} \frac{-\log P(W > x)}{x} = \gamma_F = \sup\{u : E[e^{uX}] \leq 1\}$$

  always holds, but could equal 0.

- Important interpretation from proof of "Cramér": rare events under light tails typically occur by a temporary change of the underlying distribution, from F to some exponentially tilted $\tilde{F}$.

- In a queueing context, this causes the drift to change from negative to positive.

- Choosing $\tilde{F}$ typically relates to a minimization problem. In GI/GI/1: trade off between the slope of the new drift, and the duration of the change.

- bx can be interpreted as the most likely time it takes to create a workload of level x.

The results obtained so far are not very meaningful if

$$E[e^{\epsilon X}] = \infty$$

for all $\epsilon > 0$.

In this case, we say that X has a heavy (right) tail.

Examples of heavy tails:

- Lognormal: $P(X > x) \sim e^{-(\log x)^2}$
- Weibull: $P(X > x) \sim e^{-x^\alpha}$, $\alpha \in (0, 1)$.
- Pareto: $P(X > x) \sim Cx^{-\alpha}$
- Regular variation: $P(X > x) = L(x)x^{-\alpha}$. $L(ax)/L(x) \to 1$ (example: $L(x) = \log x$).

Properties

If $P(X > x) = L(x)x^{-\alpha}$, then

$$P(X > x + y \mid X > x) \to 1.$$

for fixed $y > 0$ as $x \to \infty$.
"If things go wrong, they go totally wrong."

If X' is an i.i.d. copy of X, then

$$P(X + X' > x) \sim P(\max\{X, X'\} > x) \sim 2P(X > x).$$

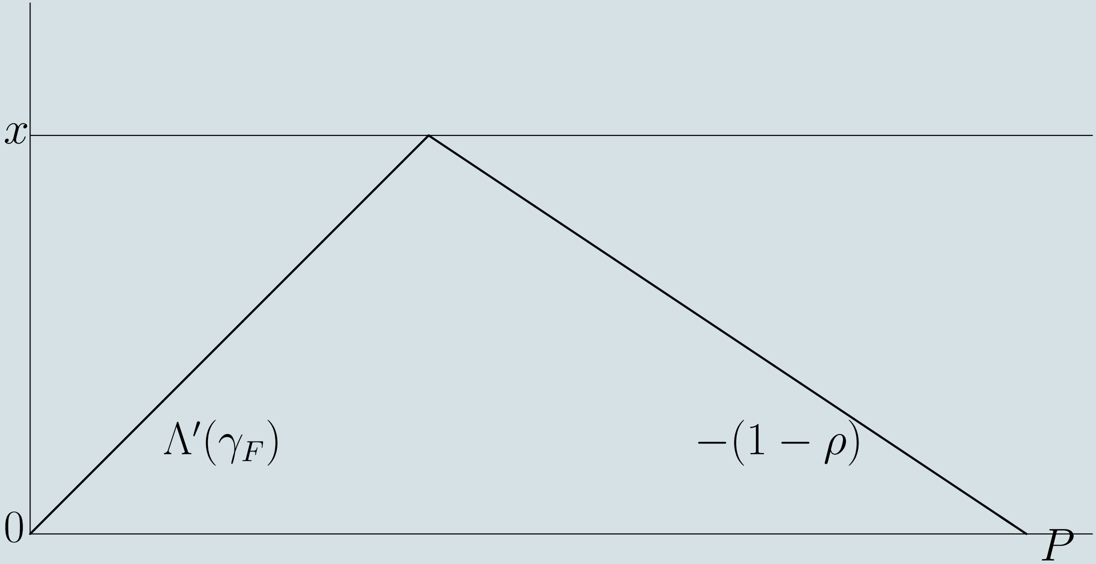
"Maximum dominates the sum."

The principle of a single big jump

- Remember $W \stackrel{d}{=} \sup_n S_n$, $X_i = B_i - A_i$.
  Suppose $P(B_1 > x) = L(x)x^{-\alpha}$.
- At some time n, the random walk S_n has the typical value $-an$, $a = -E[X]$.
- $X_{n+1} = B_{n+1} - A_{n+1}$ is so large that $S_{n+1} > x$. For this to happen, we need $X_n > an + x$.
- This can happen at any time n.

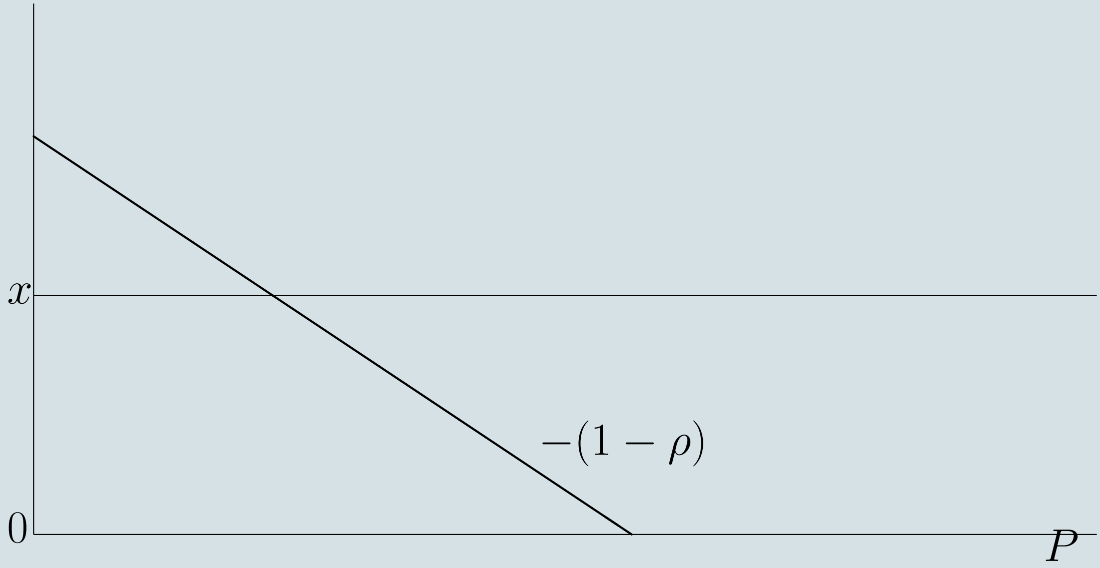
$$
\begin{aligned}
P(W > x) &\approx P(\cup_{n=1}^{\infty}\{S_n \approx -an; X_{n+1} > an + x\}) \\
&\approx \sum_{n=0}^{\infty} P(X_{n+1} > an + x) \\
&\sim \frac{1}{a}\int_x^{\infty} \bar{P}(B > u)du \\
&\sim \frac{\rho}{1-\rho}\frac{1}{E[B](\alpha - 1)}L(x)x^{1-\alpha}.
\end{aligned}
$$

Summary: The light-tailed case



- In beginning of busy period: Sample from exponentially(γ_F) tilted distribution until level x is crossed.
- Maximum in busy cycle: $x + O(1)$

Summary: The heavy-tailed case



- In beginning of busy period (after $O(1)$ time): Huge job arrives
- Maximum in busy cycle: $x + O(x)$.

Preemptive LIFO

Consider a GI/GI/1 queue with i.i.d. inter-arrival times (A_i), i.i.d. service times (B_i), working at speed 1. $\rho = E[A]/E[B] < 1$.

Assume the service discipline is Preemptive LIFO.

Observation: sojourn time has same distribution as GI/GI/1 busy period P (you enter first and leave last).

We will review the behavior as $\mathbf{P}[P > x]$ as $x \to \infty$, both for light tails and heavy tails.

In both case, assume a job of size B enters an empty system at time 0.

Let $A(x) = \sum_{n=1}^{N(x)} B_i$ be the amount of work arriving to the system $(0, x]$.

$N(x) = \max\{n : A_1 + \ldots + A_n \leq x\}$.

Upper bound:

$$\begin{aligned} \mathbf{P}[P > x] &\leq \mathbf{P}[B + A(x) > x] \\ &\leq E[e^{sB}]E[e^{sA(x)}]e^{-sx}. \end{aligned}$$

Mandjes & Zwart (2004), Glynn & Whitt (1991):

$$\lim_{x\to\infty} \frac{1}{x} \log E[e^{sA(x)}] = \Psi(s) := -\Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(s)}\right).$$

$$\Phi_A(s) = E[e^{sA}], \qquad \Phi_B(s) = E[e^{sB}].$$

Upper bound (2)

Thus,

$$\frac{1}{x}\log \mathbf{P}[P > x] \leq \frac{\log E[e^{sB}]}{x} + \Psi(s)(1+o(1)) - s.$$

optimizing over s, we obtain

$$\limsup_{x\to\infty} \frac{1}{x}\log \mathbf{P}[P > x] \leq -\gamma_L,$$

with

$$\gamma_L = \sup_{s\geq 0}[s - \Psi(s)].$$

This upper bound is sharp.
Intuition: large busy period happens as a consequence of the fact that system behaves as if $\rho = 1$ for x units of time.

Comparison with FIFO

Observe

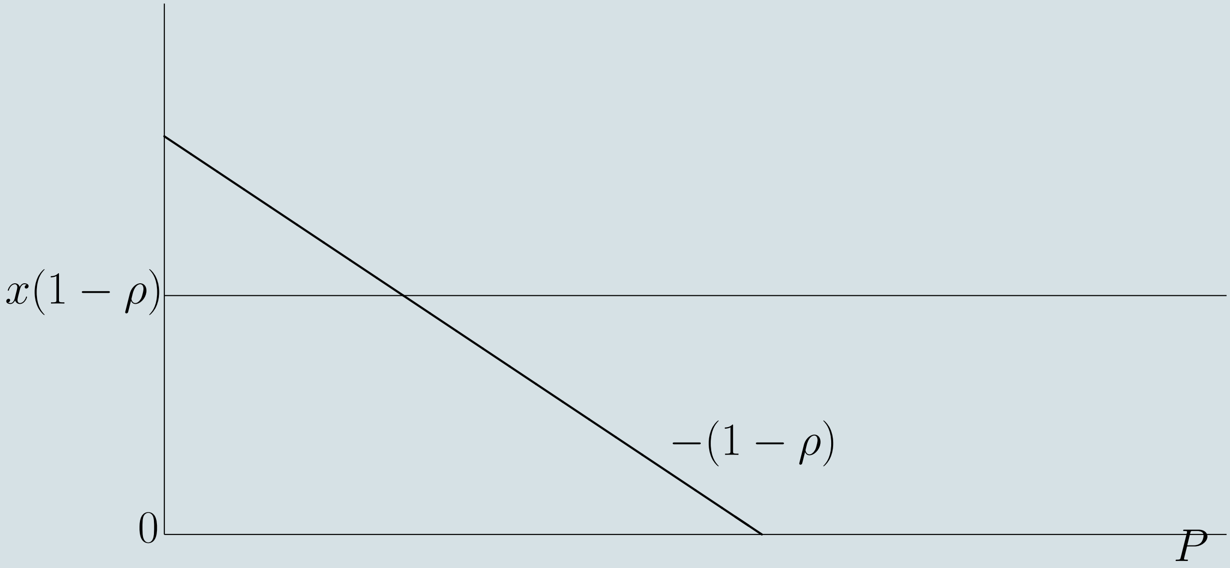
$$\begin{aligned}\gamma_F &= \sup\{s : \Phi_A(-s)\Phi_B(s) \le 1\} \\ &= \sup\{s : -s \le \Phi_A^{\leftarrow}(1/\Phi_B(s))\} \\ &= \sup\{s : s - \Psi(s) \ge 0\}.\end{aligned}$$

Since $\Psi'(0) = \rho$, and using strict convexity, it follows that

$$\gamma_L < (1-\rho)\gamma_F.$$

Conclusion: LIFO is not optimal in the light-tailed case.

Heavy tails:intuition



- In beginning of busy period (after $O(1)$ time): Huge job arrives with size $x(1-\rho)$
- Workload process drifts down at rate $1-\rho$.

Idea of proof

Based on picture:

$$\begin{aligned}\mathbf{P}[P > x] &\approx \mathbf{P}[B_{max} > x - A(x)] \\ &\approx \mathbf{P}[B_{max} > (1-\rho)x].\end{aligned}$$

Made rigorous for regularly varying service times in Zwart (2001), extended to lognormal and some Weibullian tails by Jelenkovic & Momcilovic (2004).

Boxma (1979)/Asmussen (1999): $\mathbf{P}[B_{max} > x] \sim \mathbf{E}[N]\mathbf{P}[B > x]$.

Conclusion:

$$\mathbf{P}[P > x] \sim \mathbf{E}[N]\mathbf{P}[B > x(1-\rho)].$$

If $\mathbf{P}[B > x] \sim L(x)x^{-\alpha}$, then

$$\mathbf{P}[P > x] \sim \mathbf{E}[N](1-\rho)^{-\alpha}\mathbf{P}[B > x].$$

Thus, the sojourn time under LIFO has the same tail as the service time, up to a constant!

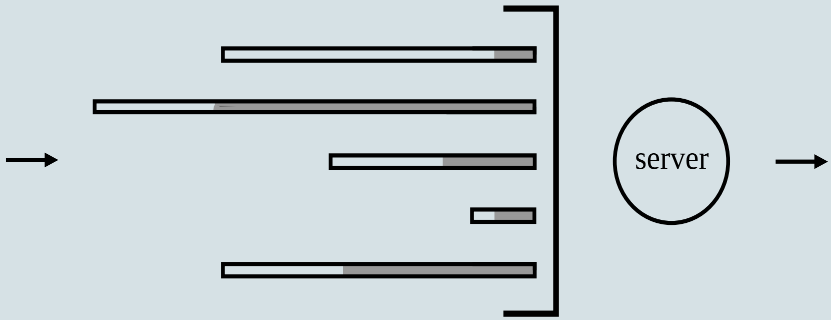
Thus, it is optimal (up to a constant).

Conclusion:

- FIFO outperforms LIFO for light tails
- LIFO outperforms FIFO for regularly varying tails.

Processor Sharing

- Processor Sharing is a service discipline where each job in the system receives the same service rate.
- Old application: time-sharing in computer systems.
- New application: TCP-like bandwidth allocation mechanisms.



How does a large response time occur?

1. Huge amount of work/number of jobs upon arrival
2. Increased amount of work/arrivals during sojourn
3. Unusually large service time

- FIFO: Always case 1.
- LIFO with light tails: case 2
- LIFO with heavy tails: case 2 or 3.
- PS ??

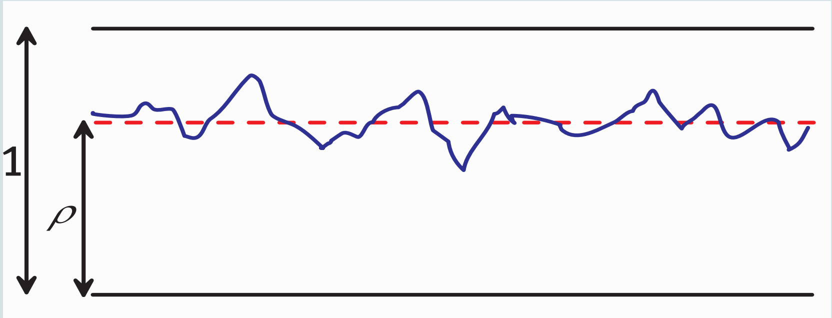
Heavy tails

One way to achieve sojourn time of length x is that your own service time is $(1-\rho)x$.

All other jobs will regard the big job as permanent (separation of timescales).

PS with one permanent customer is stable, so throughput must be ρ. Thus, service rate of $1-\rho$ is allocated to large customer, leading to sojourn of x

$$\mathbf{P}[V > x] \sim \mathbf{P}[B > x(1-\rho)]$$



Comments

$$\mathbf{P}[V > x] \sim \mathbf{P}[B > x(1-\rho)]$$

- Called a reduced service rate approximation or reduced load approximation.
- Sojourn time is primarily large because of a large service time.
- "If you stay in the system for a long time, its your own fault".

Light-tailed case

Let P^* be the time to empty the system starting from equilibrium.

Upper bound

$$\mathbf{P}[V > x] \;\leq\; \mathbf{P}[P^* > x]$$

Using similar arguments as before, we obtain

$$\limsup_{x\to\infty} \frac{\log \mathbf{P}[V > x]}{x} \leq -\gamma_L.$$

This bound is sharp if B can take arbitrary large values.

Conclusion: PS outperforms FIFO for heavy tails, but is as bad as LIFO for light tails.

- Heavy-tailed case like PS:

$$\mathbf{P}[V > x] \sim \mathbf{P}[B > x(1-\rho)]$$

with similar intuition.

- Light tails like LIFO:

$$\mathbf{P}[V > x] \geq \mathbf{P}[V > x; B > x_0]$$

This can be lower bounded by a busy period of jobs smaller than x_0, which has decay rate $\gamma_{L,\leq x_0}$. Then take $x_0 \to \infty$.

- Does not work if B has bounded support with mass at right end point x_B. In that case, there is a connection with a priority queue, and the decay rate is in the interval $(\gamma_L, \gamma_F]$.

Other disciplines

- Extension of SRPT to wider family of size-based scheduling disciplines, so called "SMART" disciplines (Wierman et al): results stay qualitatively the same

- Same story for FB (LAS).

- What makes a scheduling discipline optimal for light tails, and what makes it optimal for heavy tails?

- More general framework is needed.

The setup

- Scheduling discipline π with following properties:
  - work-conserving,
  - non-anticipative,
  - non-learning (scheduling policy is independent of events before last regeneration epoch).

- Let $V_{\pi,i}$ be sojourn time of ith arriving customer and let N be the number of customers served during a busy period. Then, if $\rho < 1$, $V_{\pi,i} \xrightarrow{d} V_\pi$ with

$$P(V_\pi > x) = \frac{1}{E[N]} E\left[\sum_{i=1}^{N} I(V_{\pi,i} > x)\right].$$

Tail optimal scheduling

- We call a scheduling discipline π_0 optimal under P if

$$\limsup_{x\to\infty} \frac{P(V_{\pi_0} > x)}{P(V_\pi > x)} < \infty$$

for any scheduling discipline π. If the limsup is ≤ 1 we call π_0 strongly optimal.

- π_0 is weakly optimal if

$$\limsup_{x\to\infty} \frac{P(V_{\pi_0} > x)^{1+\epsilon}}{P(V_\pi > x)} < \infty$$

for every scheduling discipline π and any $\epsilon > 0$.

- Challenge: what if we are allowed to vary $P(\cdot)$ as well?

How to verify optimality

Lower bounds for any service discipline:

$$
\begin{aligned}
P(V_\pi > x) &\geq P(B > x) \\
P(V_\pi > x) &= \frac{1}{E[N]} E\left[\sum_{i=1}^{N} I(V_{\pi,i} > x)\right] \\
&\geq \frac{1}{E[N]} E\left[\sum_{i=1}^{N} I(V_{\pi,i} > x) I(C_{max} > x)\right] \\
&\geq \frac{1}{E[N]} P(C_{max} > x).
\end{aligned}
$$

C_{max} is the maximal amount of work in system during a busy period.

Upper bound: time it takes to empty entire system from stationary just after an arrival (residual busy period).

Optimality

- Recall that C_{max} is the maximal amount of work in system during a busy period.

- It can be shown that $\gamma_{C_{max}} = \gamma_F$, so FIFO is weakly optimal for light tails. This is shown before in a different setting by Ramanan & Stolyar (2001).

- For heavy tails, PS,LIFO and SRPT are optimal.

- Main question: Can we construct a work-conserving non-anticipative non-learning scheduling algorithm that will be weakly optimal for $P \in \mathcal{P}$ with $\mathcal{P}$ containing both light tails and heavy tailed service times?

NO!

Some intuition:

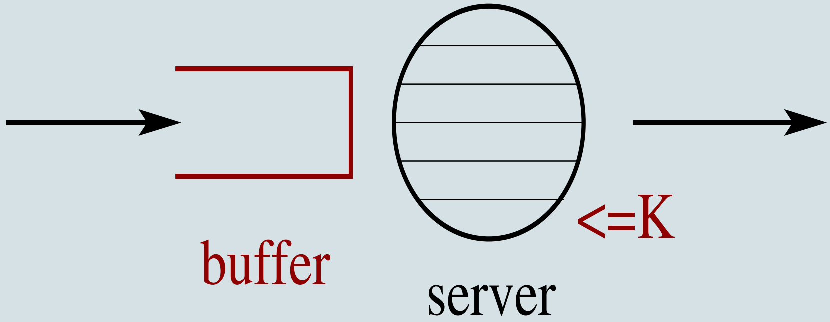
- Non-preemptive scheduling disciplines are not optimal, since $O(x)$ big jobs get stuck after a single big job of size $\geq x$ arrives. This is bad in case of heavy tails.

- PS, LIFO and SRPT all have the appealing property that system stays stable if an infinite-size job is added. This seems a necessary condition to be optimal for heavy tails.

- Suppose that a scheduling discipline retains stability after adding an infinite-size job. If you are a large job, you will likely have to wait for a busy period of small jobs to pass you, leading to busy-period type behavior, which is bad in case of light tails.

- Proof is actually based on this intuition and shows that disciplines that are optimal in one case are worst case in the other case, and vice versa.

Limited Processor Sharing



- At most K jobs can be served simultaneously, according to PS
- Additional jobs wait in FIFO buffer.
- Idea: clever choice of K, for example as function of ρ (assuming we know the load).

Results for LPS

- If $\mathbf{P}[B > x] \sim L(x)x^{-\alpha}$, then

$$-\log \mathbf{P}[V > x] \sim \min\{\alpha, (\alpha - 1)k\} \log x,$$

  with $k = \inf\{n : \rho > (1 - n/K)\}$ the number of big jobs necessary to saturate the system.

- If B has decay rate $\gamma_B > 0$, then

$$\gamma_{LPS-K} = \inf_{a \in [0,1]} \{(1 - a)\gamma_F + a\gamma_B/K + \sup_{s \geq 0}[sa(1 - 1/K) - \Psi(s)]\}$$

- $K = \lceil \frac{1}{1-\rho} \rceil$ seems a robust choice, leading to better than worst case behavior for large classes of light-tailed and heavy-tailed distributions.

- Knowing the load helps!

Critical loading

For most service disciplines

$$E[V_\pi] = \Theta\left(\frac{1}{1-\rho}\right)$$

Nikhil Bansal (2004) found a counterexample: for M/M/1 SRPT, he found that:

$$E[V_\pi] = \Theta\left(\frac{1}{(1-\rho)\log(1/(1-\rho))}\right) = o\left(\frac{1}{1-\rho}\right)$$

Proof is based on an "explicit" (triple integral) formula for $E[V_\pi]$ and many laborious manipulations.

Critical loading (2)

Lin/Wierman/Z (2011): be even more laborious manipulations, we found for generally distributed service times that:

- If job sizes have a Pareto law with infinite variance, then

$$E[V_\pi] = \Theta\left(\log(1/(1-\rho))\right).$$

- If job sizes have finite variance, then

$$E[V_\pi] = \Theta\left(\frac{1}{(1-\rho)G^{-1}(\rho)}\right)$$

  with $G(x) = E[B; B < x]/E[B]$.

- The heavier the tail the slower the growth
- Proofs are not probabilistic so no intuition yet...

Concluding remarks

- Challenge 1: get better understanding of SRPT
- Challenge 2: combine techniques from queueing and scheduling. Example: Suppose one needs to schedule n items and the goal is to minimize mean response time. Optimal blind scheduling policy has a competitive ratio of $O(\log n)$ for n large. In the queueing world, a busy period has roughly the length $1/(1-\rho)$, so one would expect that any blind policy would be $O(\log(1/(1-\rho))$ worse than SRPT, which is consistent with Bansal's result for $M/M/1$.

  Difficult to make this precise.