CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Optimal scheduling discipline in a single-server queue with Pareto type service times

Samuli Aalto (Helsinki University of Technology)
Urtzi Ayesta (LAAS-CNRS)

ValueTools 2008
23 october, 2008

Scheduling in an M/G/1 Queue

- Poisson arrivals with rate λ.
  Service requirements are i.i.d. with distribution $F(x)=P[X \leq x]$.
  Complementary cumulative distribution denoted by $\overline{F}(x)=1-F(x)$

- Attained service is known (total service requirement unknown)

- Optimality criterion: Mean number of jobs in the system

Scheduling disciplines

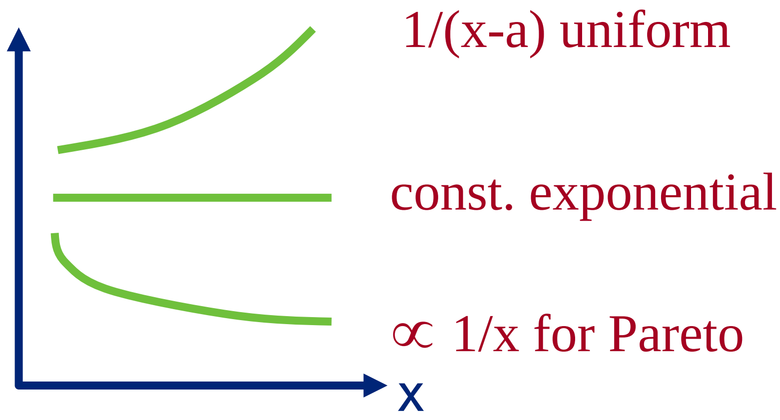
Two important set of disciplines depending on whether the size of jobs is known.

- The size is known: Shortest-Remaining-Processing-Time (SRPT) is optimal with respect to the average response time of the system.
- The size is not known, but we know the *attained service* of jobs. The most appropriate scheduling discipline depends on the service time distribution characteristics

Monotone Hazard Rate

Hazard rate of a distribution function: $h(x)dx = P[x < X \leq x+dx \mid X > x]$

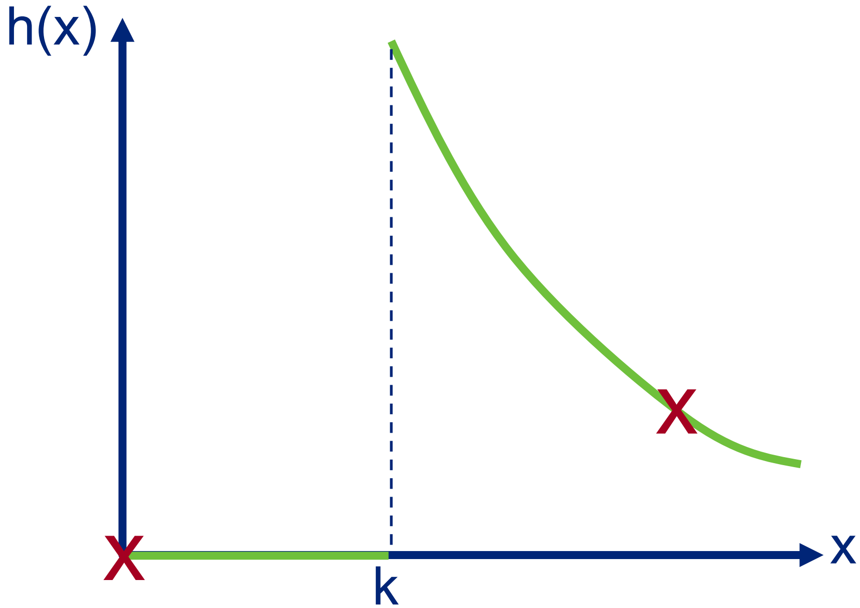
$$h(x) = \frac{f(x)}{1 - F(x)}$$



- Increasing Hazard Rate (IHR):
  Non-preemptive discipline (FCFS etc.) is optimal
- Constant hazard rate, i.e. exponential distribution:
  Mean number of jobs is policy independent
- Decreasing Hazard Rate (DHR):
  Least Attained Service (LAS) is optimal. Job(s) with the least attained service is served.

Scheduling for non-monotone hazard rate?

- What if the distribution is defined on a interval [k,∞).
  For example a Pareto-type distribution



$$\bar{F}(x) = \begin{cases} 1, & 0 < x \le k \\ \left(\dfrac{k}{x}\right)^{\alpha}, & x > k \end{cases}$$

- What if the support is bounded, that is, if $\overline{F}(x) = 0$ for all p>x ?

Optimality of Gittins policy

Theorem [Gittins89]:

Gittins index policy minimizes the mean number of jobs in the system among all non-anticipating scheduling policies

Introduced by Sevcik [1974] for static scheduling (Smallest-Rank policy)
Optimality in an M/G/1 queue by Gittins [1989].

Gittins index

Job with attained service a has the Gittins index $G(a) = \sup_{\Delta \geq 0} J(a, \Delta)$

with $$J(a,\Delta) = \frac{\int_a^{a+\Delta} f(y)\,\mathrm{d}y}{\int_a^{a+\Delta} \overline{F}(y)\,\mathrm{d}y} = \frac{\text{reward}}{\text{investment}}$$

- reward: P[a $\leq$ X $\leq$ a+Δ |X>a]
- investment: E[min(X-a, Δ) |X>a]

– In particular: *J(a,0)=h(a)* and *J(a,∞)=1/E[X-a|X>a]*

Gittins index policy

Serve at every instant of time the job with highest value G(a).

Relation between Gittins and SR

Gittins index policy

Serve at every instant of time the job with highest value G(a).

Sevcik's Smallest-Rank policy index policy:

Pick the job with highest index value G(a) and assign him a service quota $\Delta^*(a) = \inf\{\Delta \geq 0 \mid G(a) = J(a,\Delta)\}$

This job will be served until:

- It receives $\Delta^*(a)$ units of service
- It departs from the system
- A new job with higher Gittins index arrives to the queue

Gittins and SR (cont.)

Lemma: For all $a \leq x \leq a+\Delta^*(a)$,

- $G(x) \geq G(a)$
- $x+\Delta^*(x) \leq a+\Delta^*(a)$

Proposition: The Gittins discipline and SR are equivalent sample-path wise.

– Not surprising result:
  - Optimality of $c\mu$-rule (without arrivals Smith'56, with arrivals Fife'65)
  - Multi-class single server queue with feedback and non-preemptive policy: The optimal policy without arrivals is also optimal with Poisson arrivals [MW76]

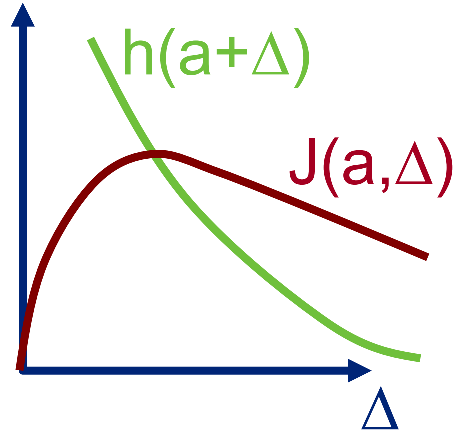
Gittins index policy

Theorem: For any attained service a≥0,

$$G(a) = h\left(a + \Delta^*(a)\right)$$

Sketch of the proof:

$$\frac{\partial}{\partial \Delta} J(a,\Delta) = 0 \Rightarrow J\left(a, \Delta^*(a)\right) = h\left(a + \Delta^*(a)\right)$$



Proposition: G(a) is decreasing for all a if and only if the service time distribution is DHR

Sketch of the proof:

$\leftarrow$ For any fixed a, J(a,Δ) is decreasing with respect to Δ.

- Then for all a, G(a)=J(a,0)=h(a), and note that h(a) is decreasing

$\rightarrow$ It can be shown that G(a)=h(a)

Theorem: LAS minimizes stochastically the number of jobs if and only if the service time distribution is DHR

Sketch of the proof:

$\leftarrow$ By [RS,1989]

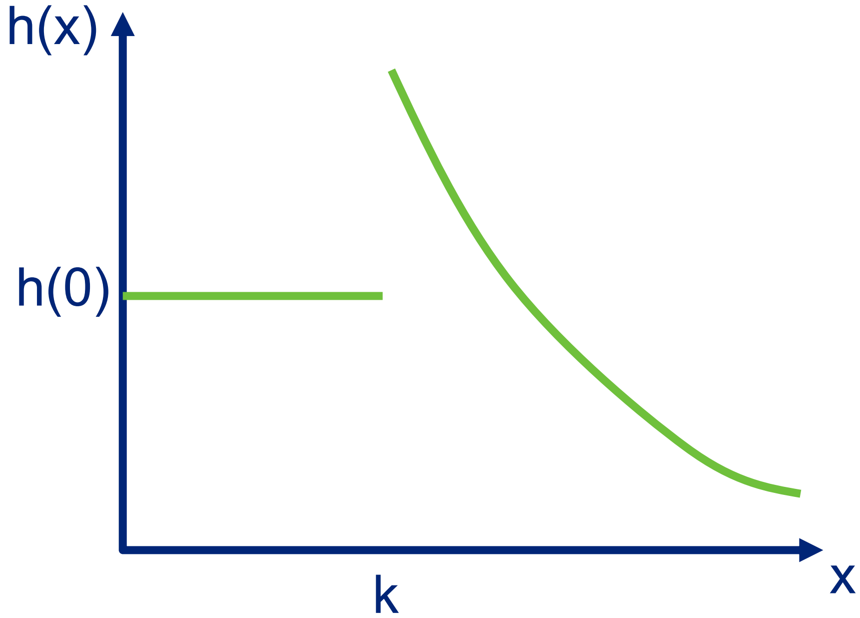
$\rightarrow$ in particular LAS minimizes the mean, thus G(a) is decreasing, and hence distribution must be DHR.

Equivalent result for FCFS and NBUE distributions.

CDHR(k) or Pareto-type distributions

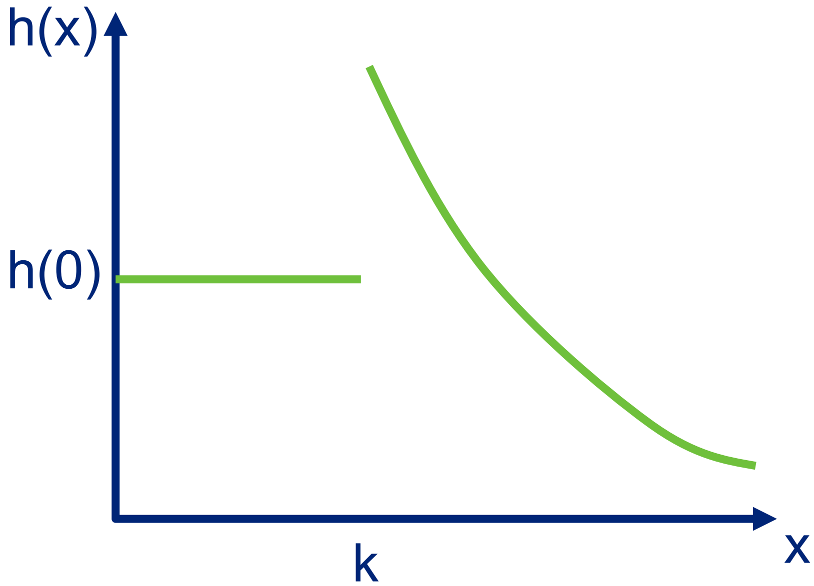
Definition of CDHR(k) distribution:

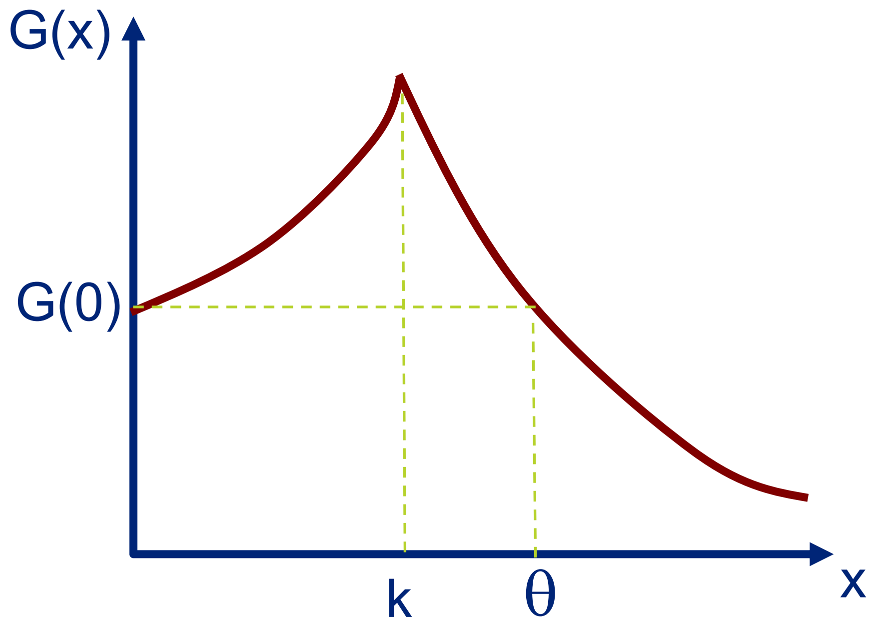
- A1: $h(x)$ is constant for all $x < k$,
- A2: $h(x)$ is decreasing for all $x \geq k$.
- A3: $h(0) < h(k)$.



Proposition: Assume the service time distribution belongs to the class CDHR(k). Then there exists a θ >k s.t,

- G(x) ≥G(0) for all x < θ,
- G(θ) ≤ G(0), and
- G(x) is decreasing for all x ≥ θ.





Theorem: Assume a CDHR(k) service time distribution:

(i) If A3 is not satisfied, then G(x) is decreasing for all x, hence LAS is optimal.

(ii) If A3 is satisfied, then there is $\theta > k$ such that FCFS+ LAS(θ) is optimal.

The precise value of θ depends only on the parameters of the service time distribution.

FCFS+LAS(θ):

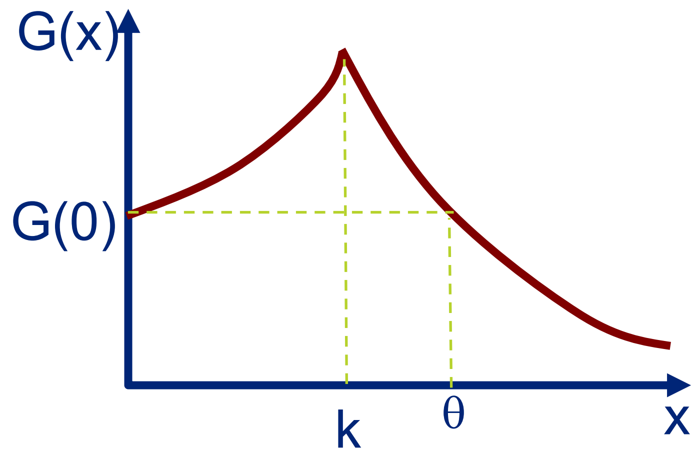
Classify jobs into two classes:

High Priority:

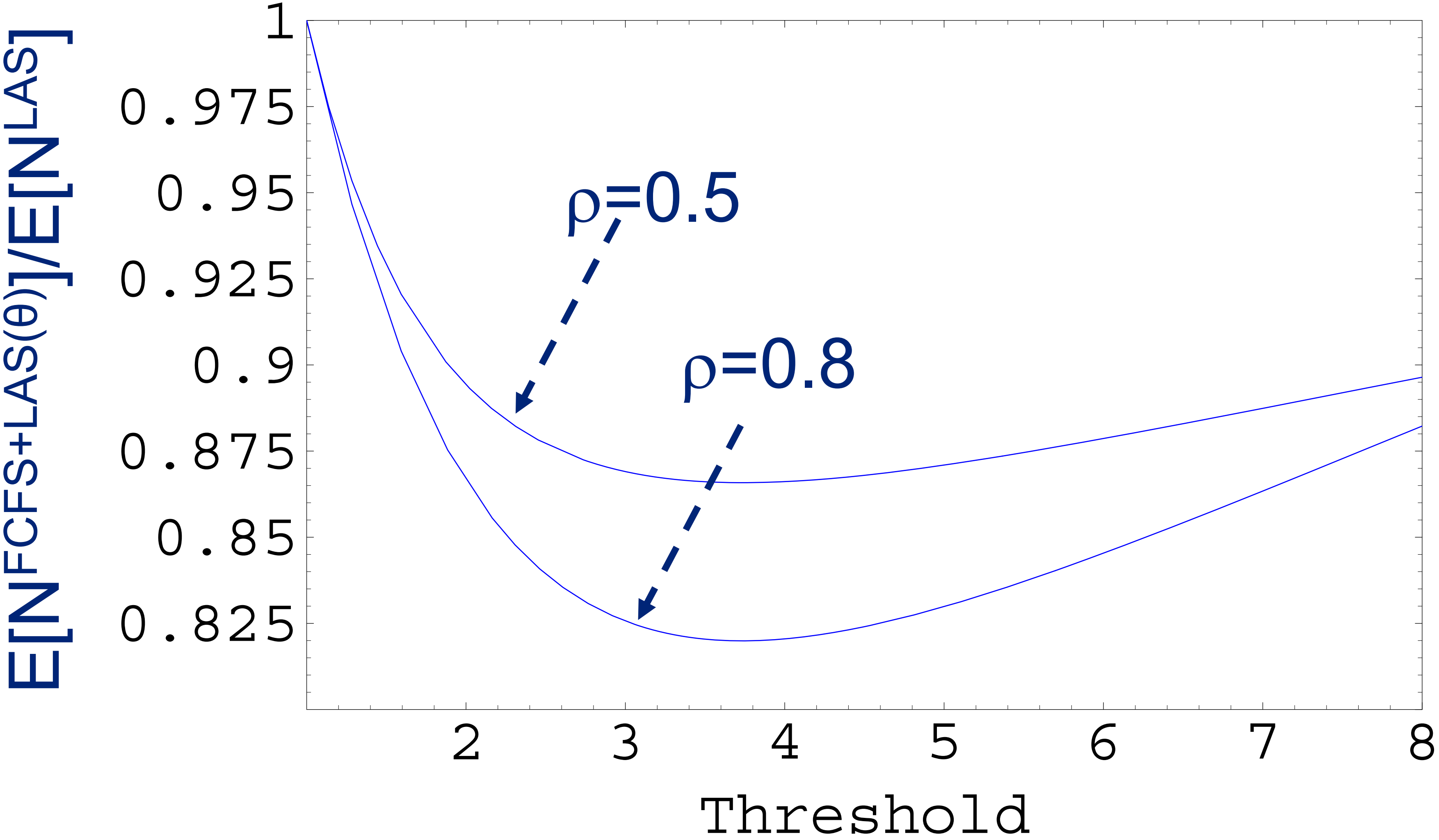
- Jobs with attained service less than θ
- Serve within this class according to FCFS

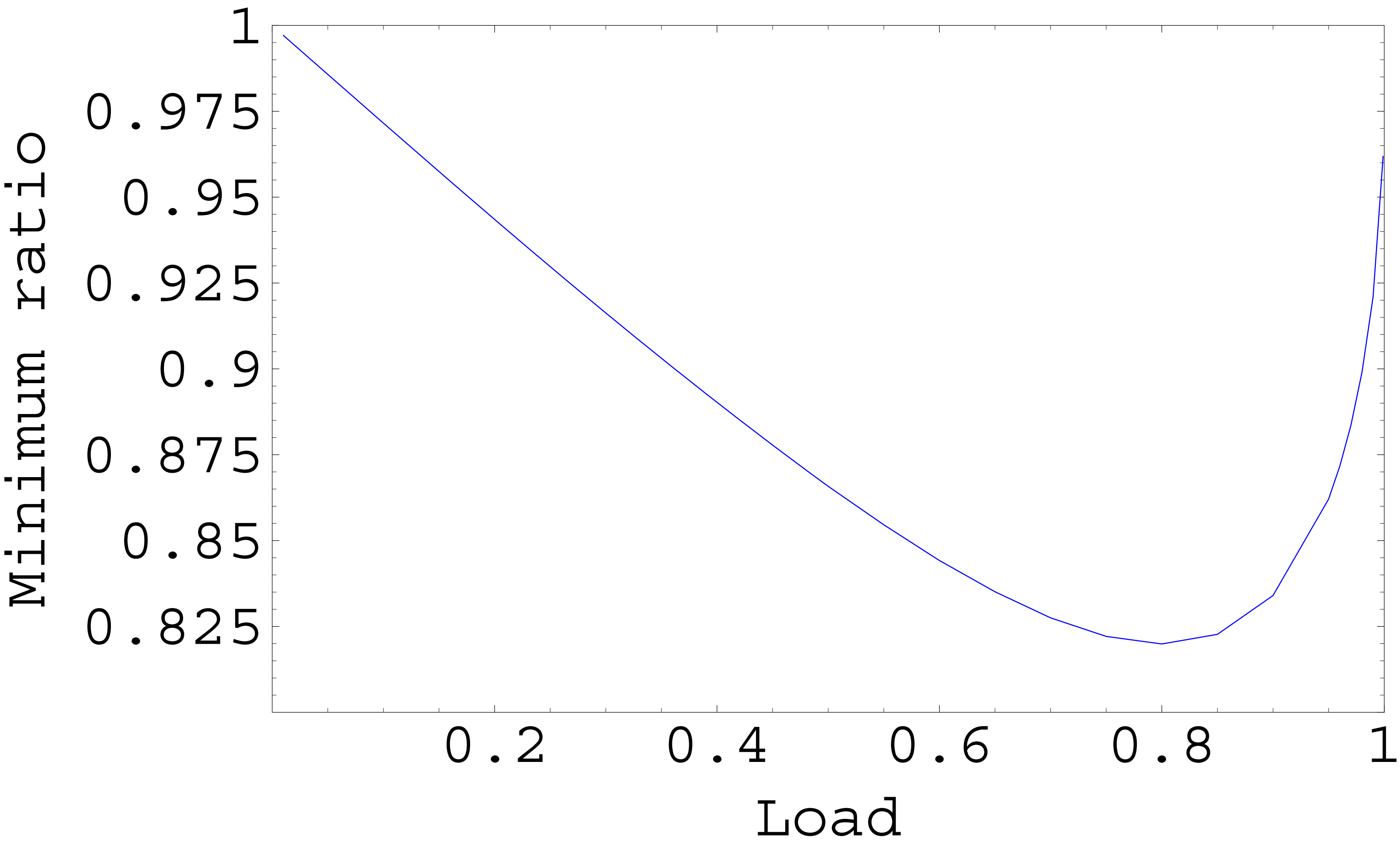
Low Priority:

- Jobs with attained service more than θ
- Serve within this class according to LAS



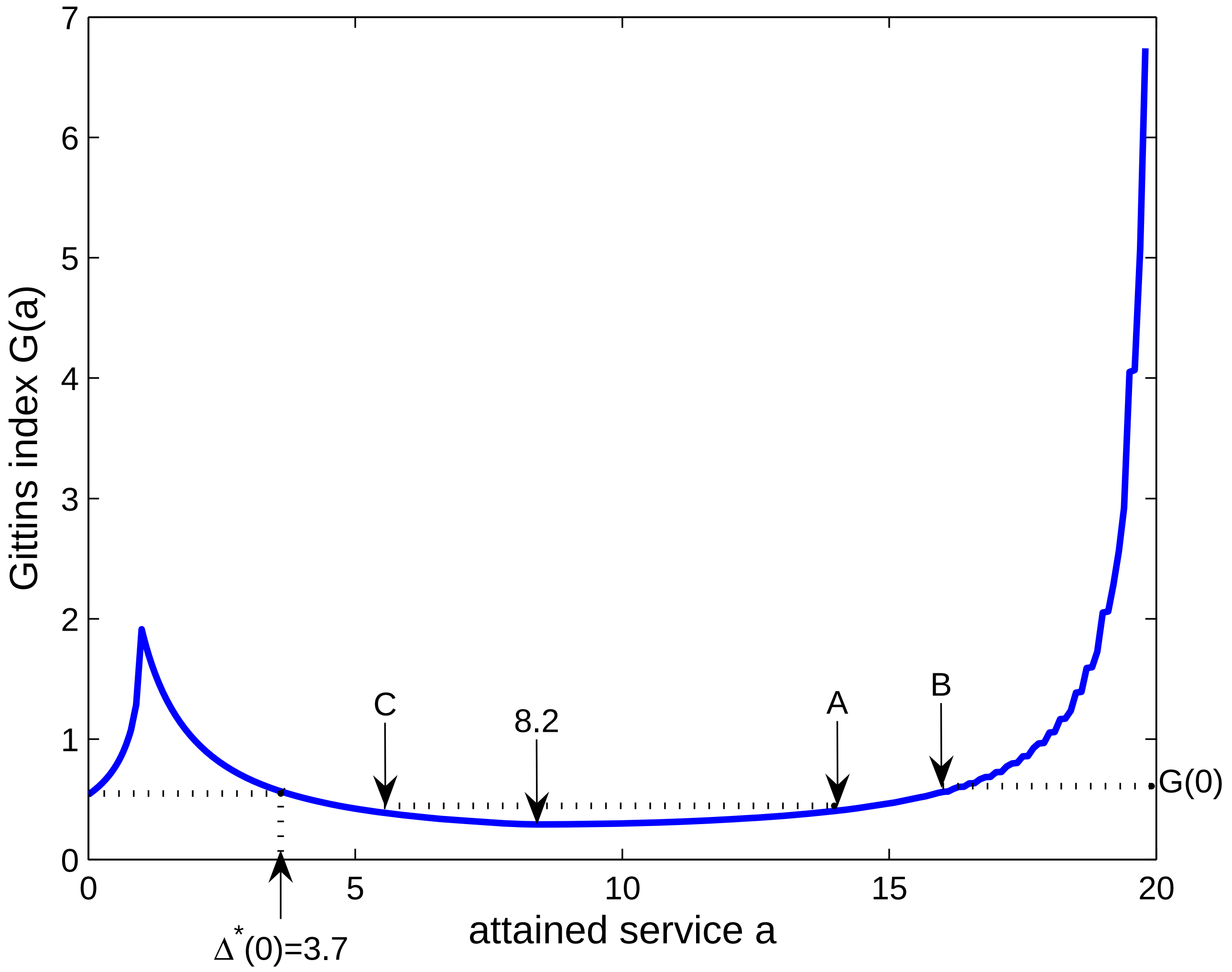
Numerical example: Pareto distribution with k=1 and α =2





1
0.975
0.95
0.925
0.9
0.875
0.85
0.825
Minimum ratio
0.2
0.4
0.6
0.8
1
Load

Impact of an upper bound bounded distribution: Bounded Pareto



Conclusion and future research

- Application of index policy for non work-conserving systems:
  - Multi-server systems
  - Time-varying capacity like in wireless systems

- Scheduling in a G/G/1 queue. LAS and FCFS are optimal with DHR and IHR respectively.
  - What if hazard-rate is not monotone?

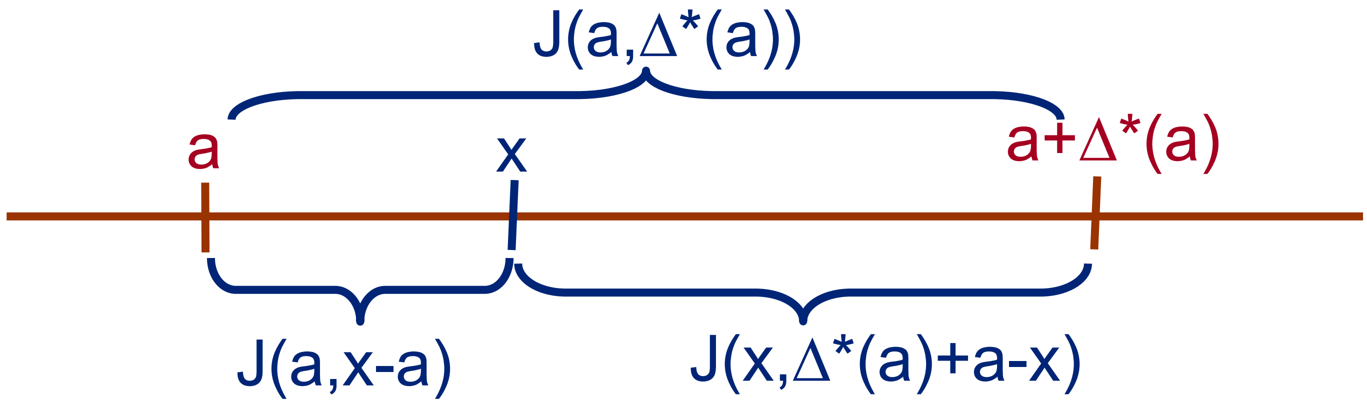
- Calculate performance metrics for a given function G(a)?

- Relation between optimal scheduling in static and stochastic scenarios.

- Application of Gittins for multi-class queues
  - Optimal policy for cases that $c\mu$-rule does not cover

Sketch of the proof: For all $a \leq x \leq a+\Delta^*(a)$, there exists a function $p(x) \leq 1$ such that

$$J(a,\Delta^*(a))=p(x)\, J(a,x-a) + (1-p(x))\, J(x,\Delta^*(a)+a-x).$$



But $J(a,\Delta^*(a)) \geq J(a,x-a)$, thus $J(x,\Delta^*(a)+a-x) \geq J(a,\Delta^*(a))$.

Now it follows that

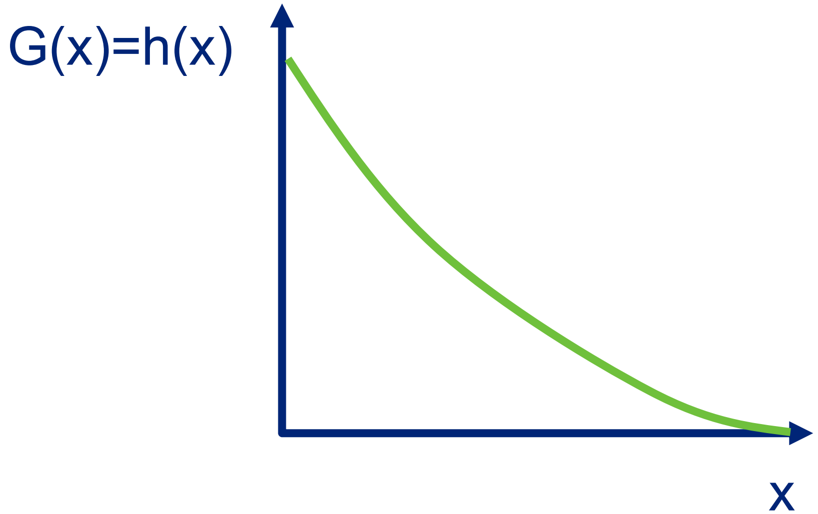
$$G(x) \geq J(x,\Delta^*(a)+a-x) \geq J(a,\Delta^*(a)) = G(a).$$

Theorem: If the distribution is of type DHR, LAS (Least-Attained Service) minimizes the mean number of jobs in the system

$$J(a,\Delta)=\frac{\int_a^{a+\Delta} f(y)\mathrm{d}y}{\int_a^{a+\Delta} \overline{F}(y)\mathrm{d}y}$$

Sketch of the proof:

- For any fixed a, $J(a,\Delta)$ is decreasing with respect to Δ.
- Then for all a, $G(a)=J(a,0)=h(a)$, and note that $h(a)$ is decreasing



Similar result for IHR, then $G(a) \geq G(0)$, for all a

Hence any non-preemptive policy is optimal