Mean delay comparison among MLPS scheduling disciplines

Samuli Aalto (TKK)
Urtzi Ayesta (CWI)

Contents

Teletraffic application: scheduling elastic flows

- Consider a bottleneck link in an IP network
  - loaded with elastic flows, such as file transfers using TCP
  - if RTTs are of the same magnitude, then approximately fair bandwidth sharing among the flows
- Internet measurements propose that
  - a small number of large TCP flows responsible for the largest amount of data transferred (elephants)
  - most of the TCP flows made of few packets (mice)
- Intuition says that
  - favouring short flows reduces the total number of flows, and, thus, also the mean file transfer time
- How to schedule flows and how to analyse?
  - Guo and Matta (2002), Feng and Misra (2003), Avrachenkov et al. (2004), Aalto et al. (2004a,2004b)

Own references

- K. Avrachenkov, U. Ayesta, P. Brown and E. Nyberg:
  - “Differentiation between Short and Long TCP Flows: Predictability of the Response Time”
  - IEEE INFOCOM 2004
- S. Aalto, U. Ayesta and E. Nyberg-Oksanen:
  - “Two-level processor-sharing scheduling disciplines: mean delay analysis”
  - ACM SIGMETRICS – PERFORMANCE 2004
- S. Aalto, U. Ayesta and E. Nyberg-Oksanen:
  - “M/G/1/MLPS compared to M/G/1/PS”
  - INRIA Research Report RR-5219, June 2004
  - To appear in Operations Research Letters

Queueing model

- Assume that
  - flows arrive according to a Poisson process with rate λ
  - each flow has a random service requirement with distribution function $F(x)$, density function $f(x)$ and hazard rate $h(x)$
  - service time distribution is of type **DHR** (decreasing hazard rate) such as hyperexponential or Pareto
- So, we have an M/G/1 queue at the flow level
  - customers in this queue are flows (and not packets)
  - service time = file size = the total number of packets to be sent
  - attained service time = the number of packets sent
  - remaining service time = the number of packets left
- Reference model: M/G/1/PS

Scheduling disciplines at flow level

- PS = Processor Sharing
  - Without any specific scheduling policy at packet level, the elastic flows are assumed to divide the bottleneck link bandwidth evenly

- SRPT = Shortest Remaining Processing Time
  - Choose a packet from the flow with least packets **left**

- FB = Foreground-Background = **LAS** = Least Attained Service
  - Choose a packet from the flow with least packets **sent**

- MLPS = Multilevel Processor Sharing
  - Choose a packet of a flow with less packets **sent** than a given threshold

MLPS scheduling disciplines

- **Definition**: MLPS scheduling discipline
  - introduced in Kleinrock (1976)
  - based on the attained service times
  - N+1 levels defined by N thresholds $0 < a_1 < \ldots < a_N < \infty$
  - between the levels, a strict priority is applied
  - within a level, FB, PS, or FCFS is applied

- **Examples**: Two levels with threshold a
  - FB+FB = FB = LAS
  - FB+PS = FLIPS
    - Feng and Misra (2003)
  - PS+PS = ML-PRIO
    - Guo and Matta (2002), Avrachenkov et al. (2004)

Optimality results for M/G/1

- Schrage (1968)
  - If the remaining service time is known, then SRPT optimal minimizing the mean delay $E[T]$
- Yashkov (1978, 1987)
  - If only the attained service time is known, then DHR implies that FB optimal minimizing the mean delay $E[T]$
- **Remark**: in this study we consider work-conserving (WC) and non-anticipating (NA) service disciplines such as FB, MLPS and PS

Contents

- Introduction
- Earlier results
- New results

Earlier results: comparison to PS

- Aalto et al. (2004a):
  - Two levels with FB and PS allowed as internal disciplines

$$\text{DHR} \quad \Rightarrow \quad E[T^{\text{FB}}] \leq E[T^{\text{FB+PS}}] \leq E[T^{\text{PS+PS}}] \leq E[T^{\text{PS}}]$$

- Aalto et al. (2004b):
  - Any number of levels with FB and PS allowed as internal disciplines

$$\text{DHR} \quad \Rightarrow \quad E[T^{\text{FB}}] \leq E[T^{\text{MLPS}}] \leq E[T^{\text{PS}}]$$

Idea of the proof

- Key variable: U_x = unfinished truncated work with threshold x
  - sum of remaining truncated service times $\min\{S, x\}$ of those customers who have attained service less than x

- Steps in the proof:
  - **First step**: prove that for any π and π'

$$\text{DHR} \ \& \ \text{WC} \ \& \ \text{NA} \ \& \ E[U_x^{\pi}] \leq E[U_x^{\pi'}] \ \forall x \quad \Rightarrow \quad E[T^{\pi}] \leq E[T^{\pi'}]$$

  - **Second step**: prove that for any x (and t)

$$U_x^{\text{FB}}(t) \leq U_x^{\text{FB+PS}}(t) \leq U_x^{\text{PS+PS}}(t) \quad \& \quad E[U_x^{\text{PS+PS}}] \leq E[U_x^{\text{PS}}]$$

  - **Third step**: prove that for any x (and t)

$$U_x^{\text{FB}}(t) \leq U_x^{\text{MLPS}}(t) \quad \& \quad E[U_x^{\text{MLPS}}] \leq E[U_x^{\text{PS}}]$$

First step

- WC & NA implies that

$$E[T^{\pi}] = \frac{1}{\lambda}\int_0^{\infty}(E[U_x^{\pi}])'h(x)dx$$

- DHR & WC & NA implies that

$$E[U_x^{\pi}] \le E[U_x^{\pi'}] \ \forall x \quad \Rightarrow \quad E[T^{\pi}] \le E[T^{\pi'}]$$

- Proof:
  - Follows from above
  - In particular, if the hazard rate is differentiable, then by partial integration (note: $U_0 = 0$ and $E[U_\infty{}^{\pi}]$ independent of π)

$$E[T^{\pi}] - E[T^{\pi'}] = -\frac{1}{\lambda}\int_0^{\infty}(E[U_x^{\pi}] - E[U_x^{\pi'}])h'(x)dx$$

Second step (1)

- Key problem: splitting the PS level
  - π = PS+PS(a)
  - π' = PS
- Solution steps:
  - By Prop. 10 in Aalto & al. (2004b), for any $x \le a$ (and t)

$$U_x^{\pi}(t) \le U_x^{\pi'}(t)$$

  - Known result for WC disciplines:

$$E[U_\infty^{\pi}] = E[U_\infty^{\pi'}]$$

  - Based on the known integral equation for the derivative of the conditional mean delay, for any $x > a$

$$\frac{d}{dx}E[T^{\pi}(x)] \ge \frac{d}{dx}E[T^{\pi'}(x)] = \frac{1}{1-\rho}$$

Second step (2)

- Known integral equation

$$\alpha'(x) = \frac{\lambda}{1-\rho_a} \int_0^x \alpha'(y)(1-F(a+x-y))dy$$
$$+ \frac{\lambda}{1-\rho_a} \int_0^\infty \alpha'(y)(1-F(a+x+y))dy + c(x) + 1$$

- Lemma needed:

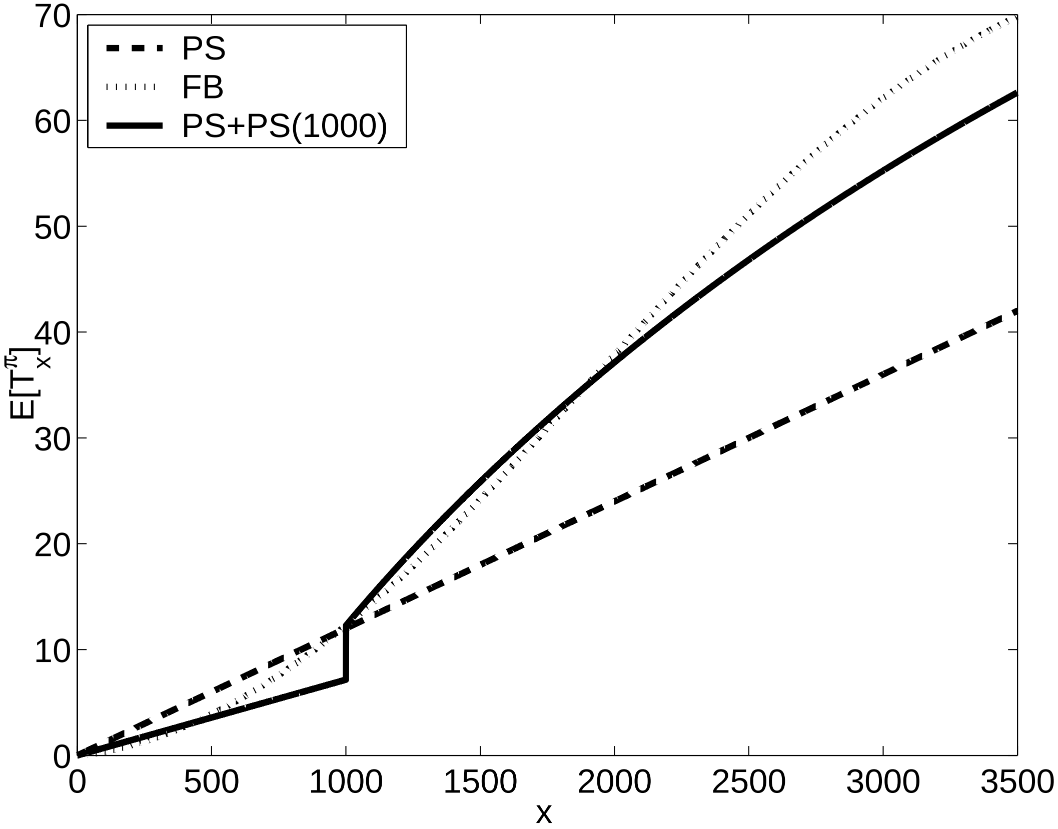
$$E[U_a^{\pi}] \le E[U_a^{\pi'}] \quad \& \quad E[U_b^{\pi}] \le E[U_b^{\pi'}] \quad \&$$
$$\frac{d}{dx}E[T^{\pi}(x)] \ge \frac{d}{dx}E[T^{\pi'}(x)] \quad \forall x \in (a,b)$$
$$\Rightarrow \quad E[U_x^{\pi}] \le E[U_x^{\pi'}] \quad \forall x \in [a,b]$$

  - Based on the known expression:

$$E[U_x^{\pi}] = E[U_a^{\pi}] + \lambda \int_a^x E[T^{\pi}(y)](1-F(y))dy$$

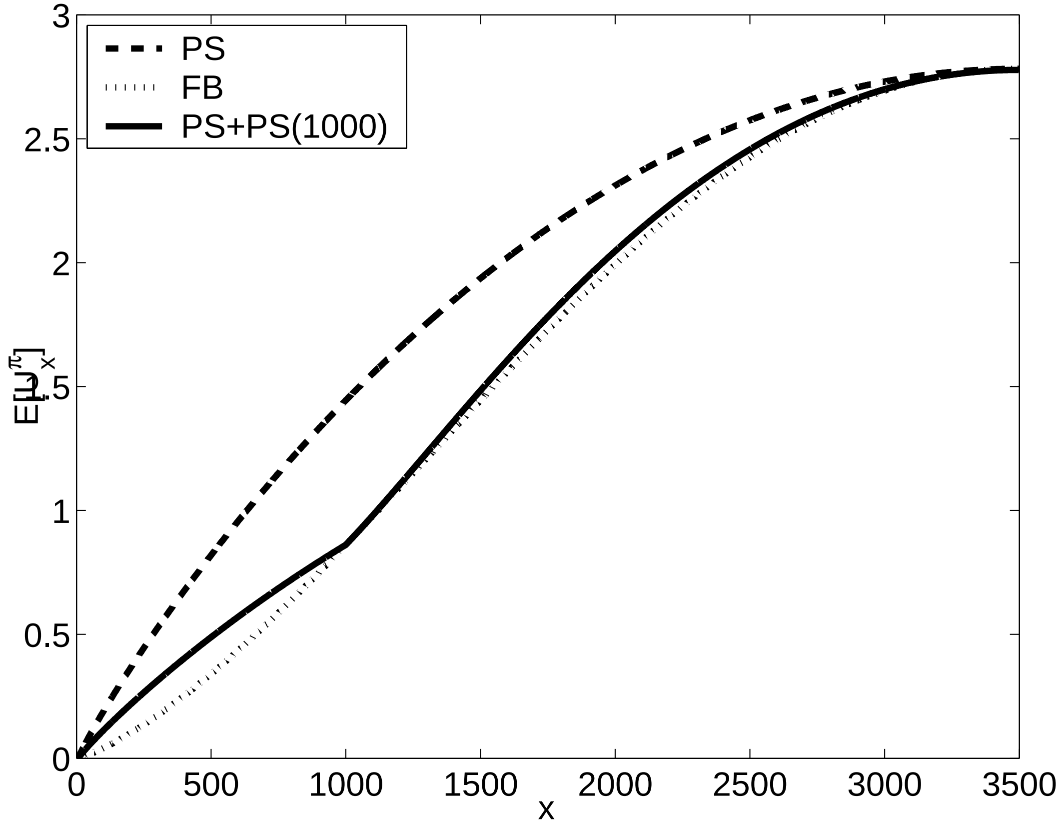
Conditional mean delay E[T(x)]

bounded Pareto file size distribution



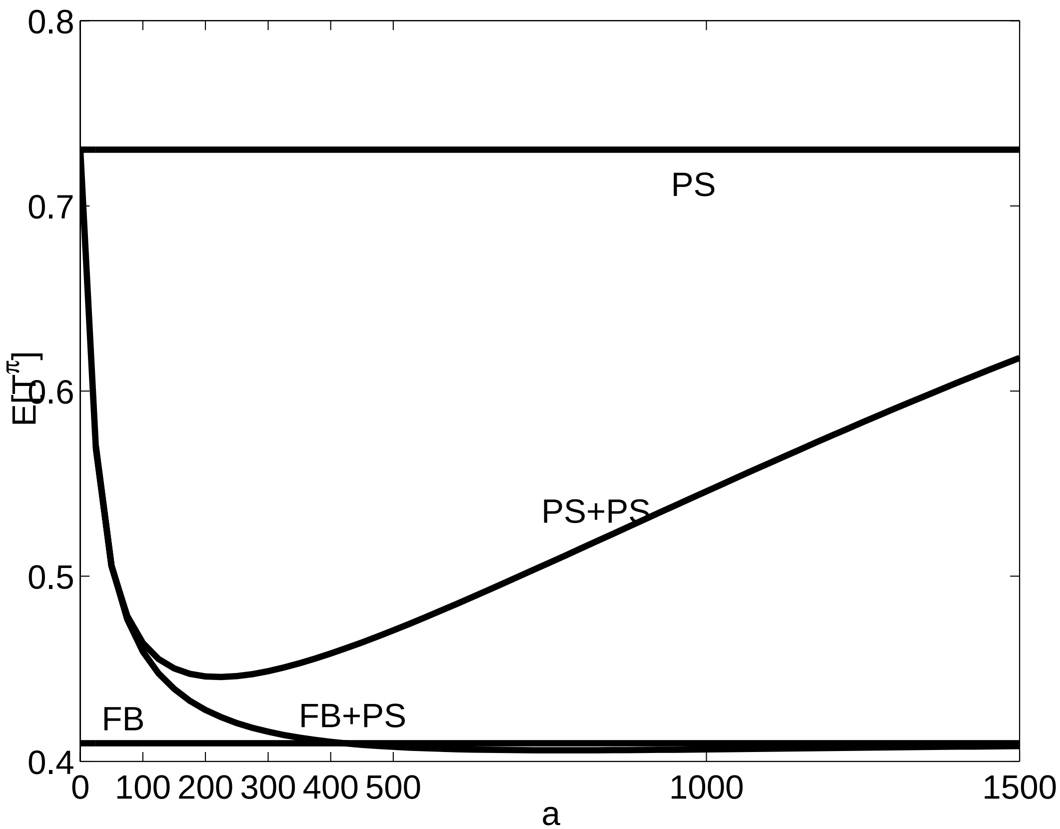
Mean unfinished truncated work $E[U_x]$

bounded Pareto file size distribution



Mean delay E[T]

bounded Pareto service time distribution



Contents

- Introduction
- Earlier results
- New results

New results: comparison among MLPS disciplines

- **Theorem 1** (not really a new one):
  - Any number of levels with all original internal disciplines allowed

$$\text{DHR} \quad \Rightarrow \quad E[T^{\text{FB}}] \leq E[T^{\text{MLPS}}] \leq E[T^{\text{FCFS}}]$$

- **Theorem 2**:
  - Any number of levels with all original internal disciplines allowed
  - MLPS is derived from MLPS' by splitting a level and copying the internal discipline

$$\text{DHR} \quad \Rightarrow \quad E[T^{\text{MLPS}}] \leq E[T^{\text{MLPS'}}]$$

- **Theorem 3**:
  - Any number of levels with all original internal disciplines allowed
  - MLPS is derived from MLPS' by changing an internal discipline from PS to FB (or from FCFS to PS)

$$\text{DHR} \quad \Rightarrow \quad E[T^{\text{MLPS}}] \leq E[T^{\text{MLPS'}}]$$

Theorem 2: Splitting a PS level (1)

- Proof based both on sample path and mean value arguments
- Prove that for all x,

$$E[U_x^{\text{MLPS}}] \leq E[U_x^{\text{MLPS}'}]$$

  - The tough nut!

Theorem 2: Splitting a PS level (2)

- Key problem: splitting the highest level. For example,
  - MLPS = PS+PS+PS(a_1, a_2)
  - MLPS' = PS+PS(a_1)
- Solution steps:
  - By Prop. 6 in Aalto & al. (2004b), for any $x \le a_2$ and t

$$U_x^{\mathrm{MLPS}}(t) \le U_x^{\mathrm{MLPS'}}(t)$$

  - Known result for WC disciplines:

$$E[U_\infty^{\mathrm{MLPS}}] = E[U_\infty^{\mathrm{MLPS'}}]$$

  - Tough new result based on the known integral equation for the derivative of the conditional mean delay: for any $x > a_2$

$$\frac{d}{dx}E[T^{\mathrm{MLPS}}(x)] \ge \frac{d}{dx}E[T^{\mathrm{MLPS'}}(x)]$$

  - Apply Lemma of Slide 14

Theorem 2: Splitting a PS level (3)

- Additional problem: splitting another level. For example,
  - MLPS = PS+PS+PS(a_1, a_2)
  - MLPS' = PS+PS(a_2)
- Solution:
  - Easily, for any $x \geq a_2$ and t

$$U_x^{\text{MLPS}}(t) = U_x^{\text{MLPS'}}(t)$$

  - Truncate service times and prove by the "splitting the highest level" result that for any $x < a_2$ (and t)

$$E[U_x^{\text{MLPS}}(S \wedge a_2)] \leq E[U_x^{\text{MLPS'}}(S \wedge a_2)]$$

Theorem 2: Splitting an FCFS level (1)

- Proof based both on sample path and mean value arguments
- Prove that for all x,

$$E[U_x^{\mathrm{MLPS}}] \le E[U_x^{\mathrm{MLPS'}}]$$

  - An easy exercise

Theorem 2: Splitting an FCFS level (2)

- Problem: splitting **any** level. For example,
  - MLPS = FCFS+FCFS+FCFS(a_1, a_2)
  - MLPS' = FCFS+FCFS(a_2)
- Solution steps:
  - By definition, for any t

$$U_0^{\text{MLPS}}(t) = U_0^{\text{MLPS'}}(t) = 0$$

  - Easily, for any $x \geq a_2$ and t

$$U_x^{\text{MLPS}}(t) = U_x^{\text{MLPS'}}(t)$$

  - Easy result based on the known expression for the conditional mean delay: for any $x < a_2$

$$\frac{d}{dx} E[T^{\text{MLPS}}(x)] \geq \frac{d}{dx} E[T^{\text{MLPS'}}(x)]$$

  - Apply Lemma of Slide 14

Theorem 3: Changing an internal discipline

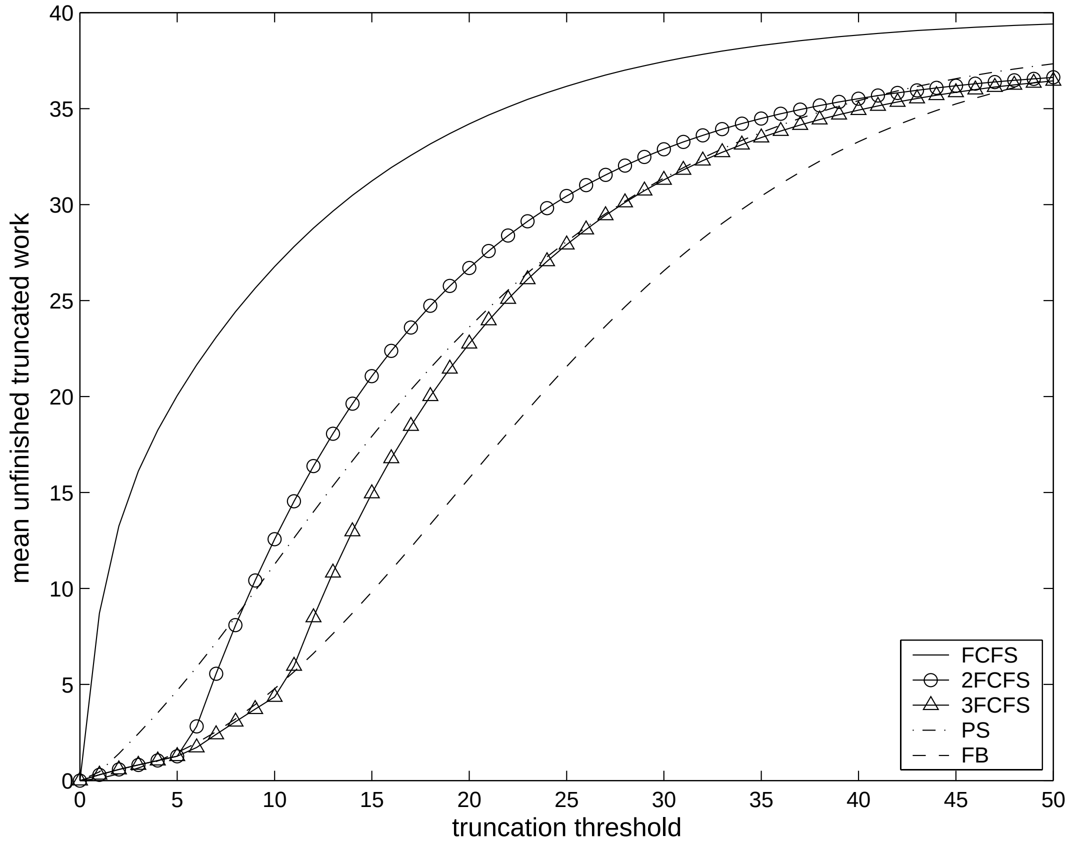
- Proof based on sample path arguments
- Prove that for all x and t,

$$U_x^{\text{MLPS}}(t) \le U_x^{\text{MLPS'}}(t)$$

  - Tedious but straightforward

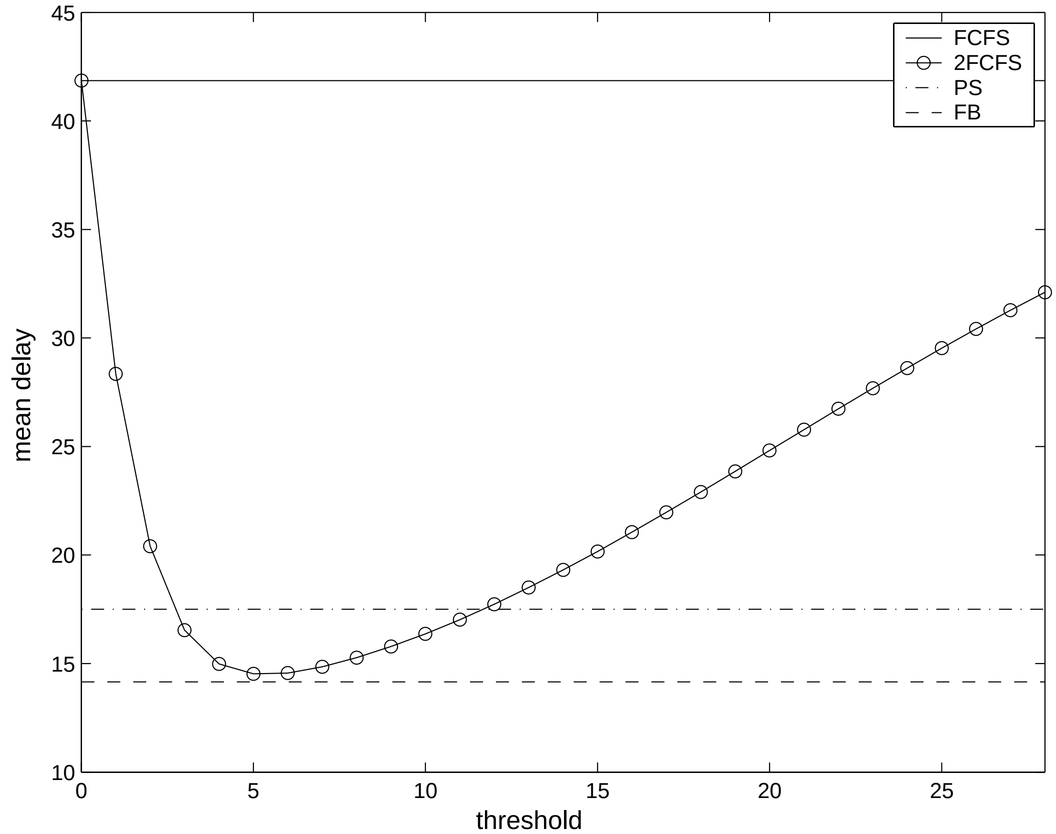
Mean unfinished truncated work $E[U_x]$

hyperexponential file size distribution



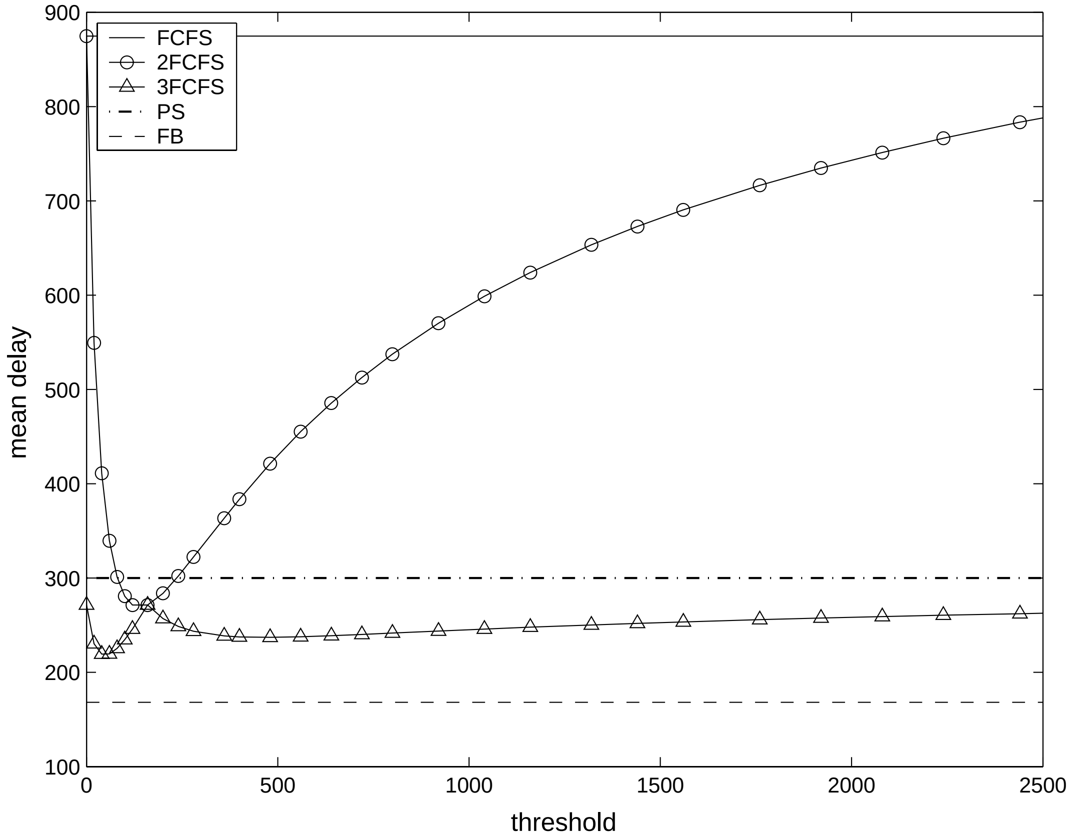
Mean delay E[T]

hyperexponential file size distribution



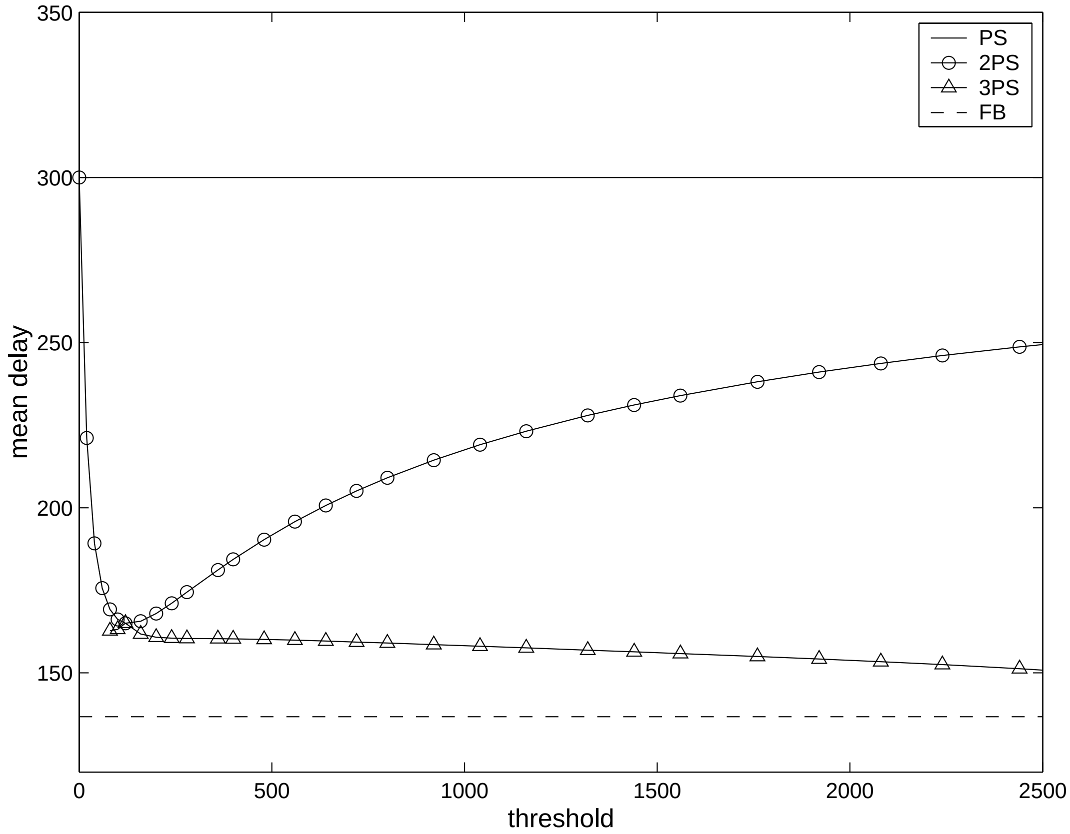
Mean delay E[T]

Pareto file size distribution



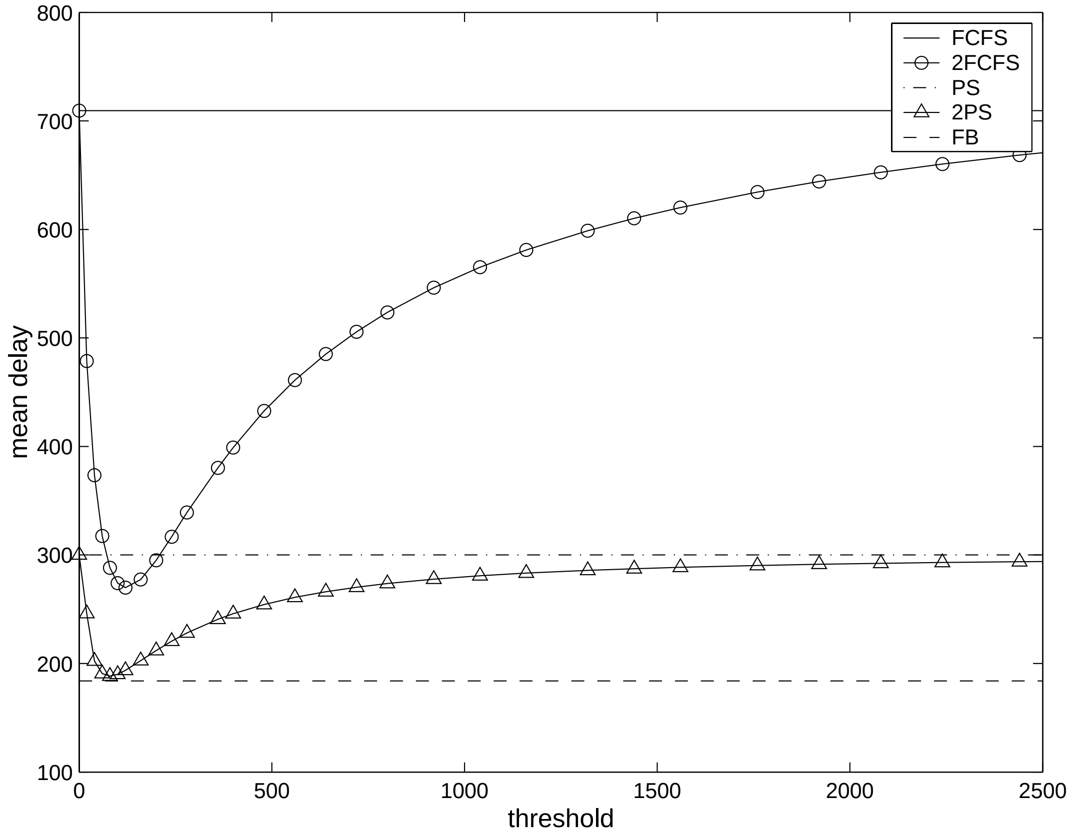
Mean delay E[T]

Pareto file size distribution



Mean delay E[T]

Pareto file size distribution



The End