Multilevel Processor Sharing Scheduling Disciplines: Mean Delay Analysis

Samuli Aalto
Networking Laboratory
Helsinki University of Technology

Joint work with
Urtzi Ayesta (INRIA) and Eeva Nyberg (HUT)

Background

- File transfers in the Internet use **TCP**
  - a file is splitted into packets which are sent (in a controlled way) from the source node to the destination node
  - flow = packets related to a file
  - due to the congestion control part of TCP, the network resources are shared fairly (in the ideal case)
- Internet measurements show 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
  - favoring short flows reduces the total number of flows, and thus, by Little's law, also the mean "file transfer" time

Mathematical model

- Consider a bottleneck link loaded with **elastic flows**
  - such as file transfers using TCP
- Assume that
  - the flows arrive according to a Poisson process with rate λ
  - each flow has a random service requirement (= file size) with a general distribution with mean L
    - cumulative distribution function $F(x)$, tail distribution function $G(x) = 1 - F(x)$, density $f(x)$, hazard rate $h(x) = f(x)$ / $G(x)$
    - typically heavy-tailed such as **Pareto** $\Rightarrow$ decreasing hazard rate
- So, at the flow level, we have a **M/G/1** queueing system
  - customers = flows = file transfers (not individual packets!)
  - delay = file transfer time
  - service time = file size / link capacity C
  - service rate = μ = link capacity C / mean file size L
  - load = $\rho = \lambda / \mu$

Scheduling disciplines

- **PS** = Processor Sharing
  - Without any specific scheduling policy, the flows are assumed to divide the bottleneck link capacity evenly (= fairness in the ideal case)
- **SRPT** = Shortest Remaining Processing Time
  - Choose a packet of the flow with least packets left
- **LAS** = Least Attained Service
  - Choose a packet of the flow with least packets sent
  - Also called: **FB** = Foreground-Background
- **MLPS** = Multilevel PS (cf. Kleinrock (1976))
  - Choose a packet of the flow with less packets sent than a given threshold

- Notes:
  - All of them are **work-conserving** disciplines
  - Only SRPT uses “future” information

Optimality results for M/G/1

- If the **remaining service times** (= number of packets left) are **known** for each customer (= flow), then
  - Schrage (1968):
    SRPT optimal minimizing the mean delay (= file transfer time)

- If only the **attained service times** (= number of packets sent) are **known** for each customer (= flow), then
  - Yashkov (1978):
    Decreasing hazard rate $\Rightarrow$
    FB optimal among work-conserving scheduling disciplines
  - Feng and Misra (2003):
    the same result as above proved (?) in another way
  - Wierman et al. (2002):
    Decreasing hazard rate $\Rightarrow$ **FB better than PS**

MLPS scheduling disciplines

- **Definition**:
  - Based on the attained service times
  - Thresholds $0 = a_0 < a_1 < \ldots < a_N < a_{N+1} = \infty$ define $N{+}1$ levels, with a strict priority between the levels
  - Within a level, either FB or PS is applied

- **Example**: Two levels with threshold a
  - FB+FB = FB = LAS
  - FB+PS = FLIPS (Feng and Misra (2003))
  - PS+PS = ML-PRIO (Guo and Matta (2002))

Conditional mean delay formulas for M/G/1

- **Notation**: $T(s)$ = delay of a customer with service time s
- PS:

$$E[T(s)] = \frac{s}{1-\rho}$$

- FB:

$$E[T(s)] = \frac{E[W_s]+s}{1-\rho_s}$$

- PS+PS(a):

$$E[T(s)] = \begin{cases} \dfrac{s}{1-\rho_a}, & s \le a \\ \dfrac{E[W_a]+a}{1-\rho_a} + \dfrac{\alpha(s-a)}{1-\rho_a}, & s > a \end{cases}$$

Related queueing systems

- M/G/1 with truncated service times $\min\{S,x\}$:

$$\rho_x = \lambda E[\min\{S,x\}]$$

$$E[W_x] = \frac{\lambda E[(\min\{S,x\})^2]}{2(1-\rho_x)}$$

- M^X/G/1-PS with modified service times $\tilde{S}$:

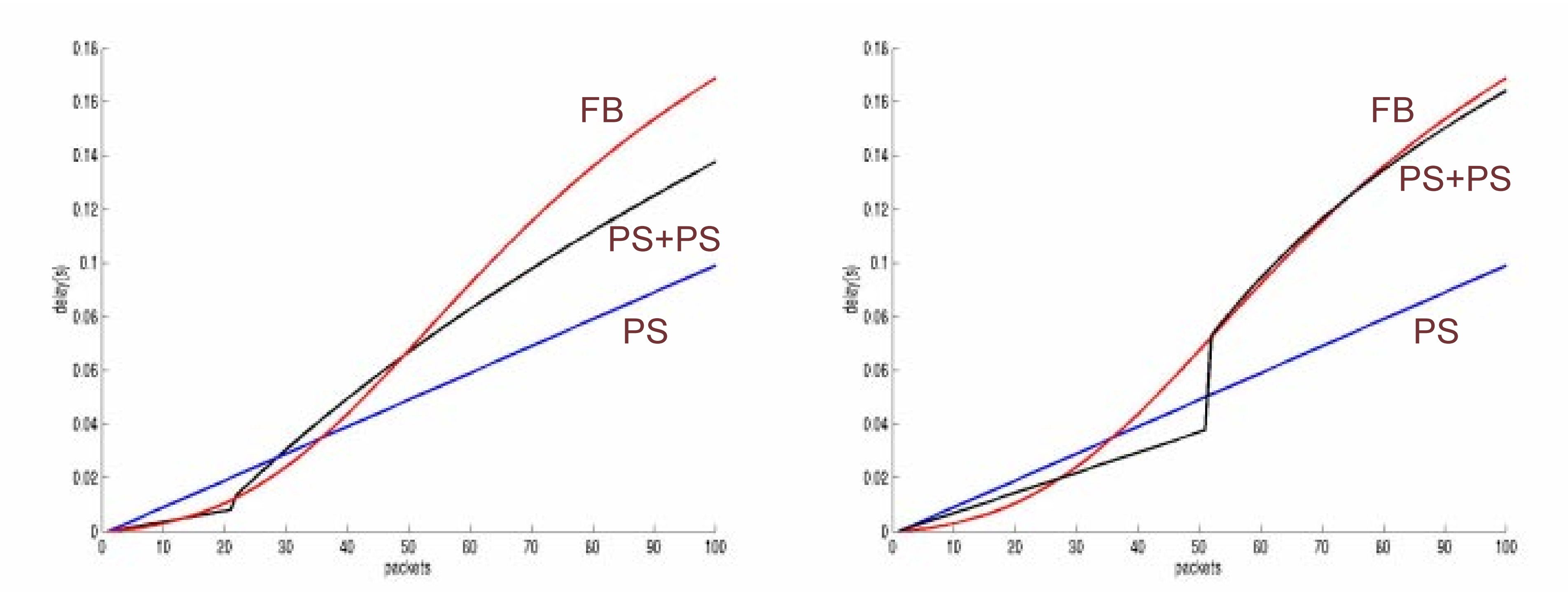
$$P\{\tilde{S} \le x\} = P\{S \le a + x \mid S > a\}$$

$$\alpha(x) = E[\tilde{T}(x)] \quad \text{satisfying}$$

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

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

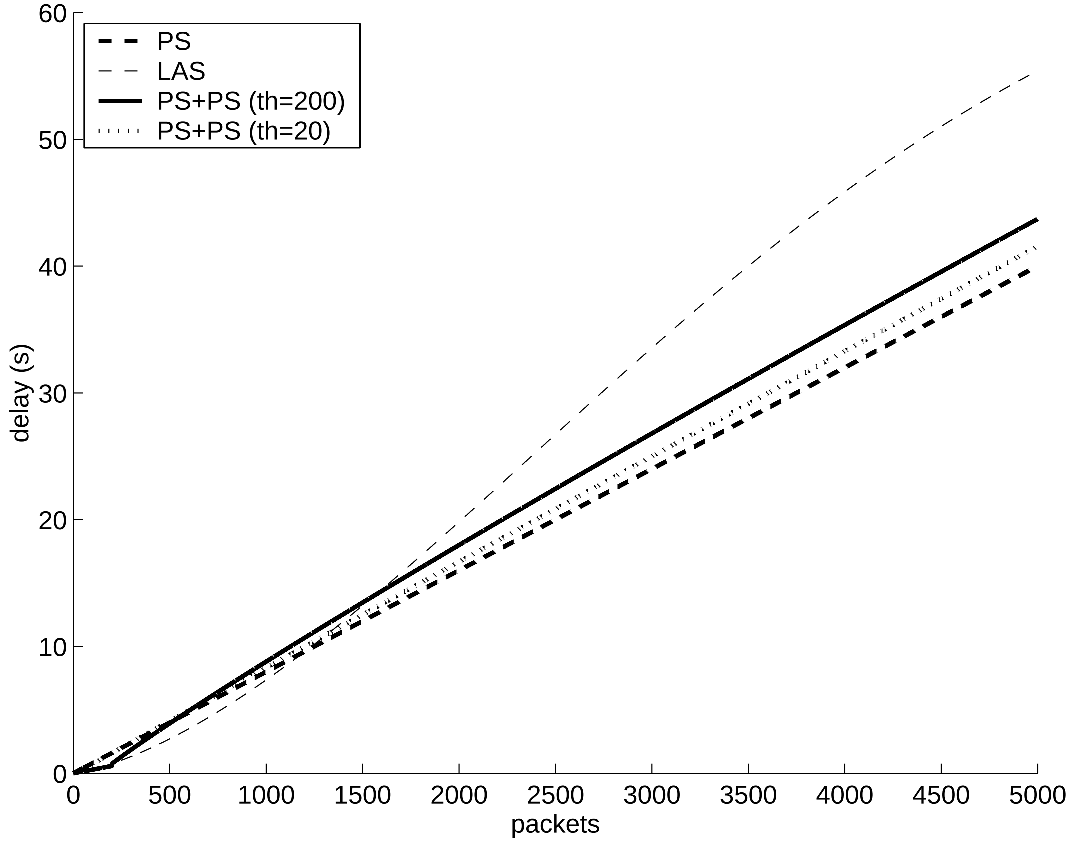
Conditional mean delay E[T(s)]



Note: exponential service time distribution

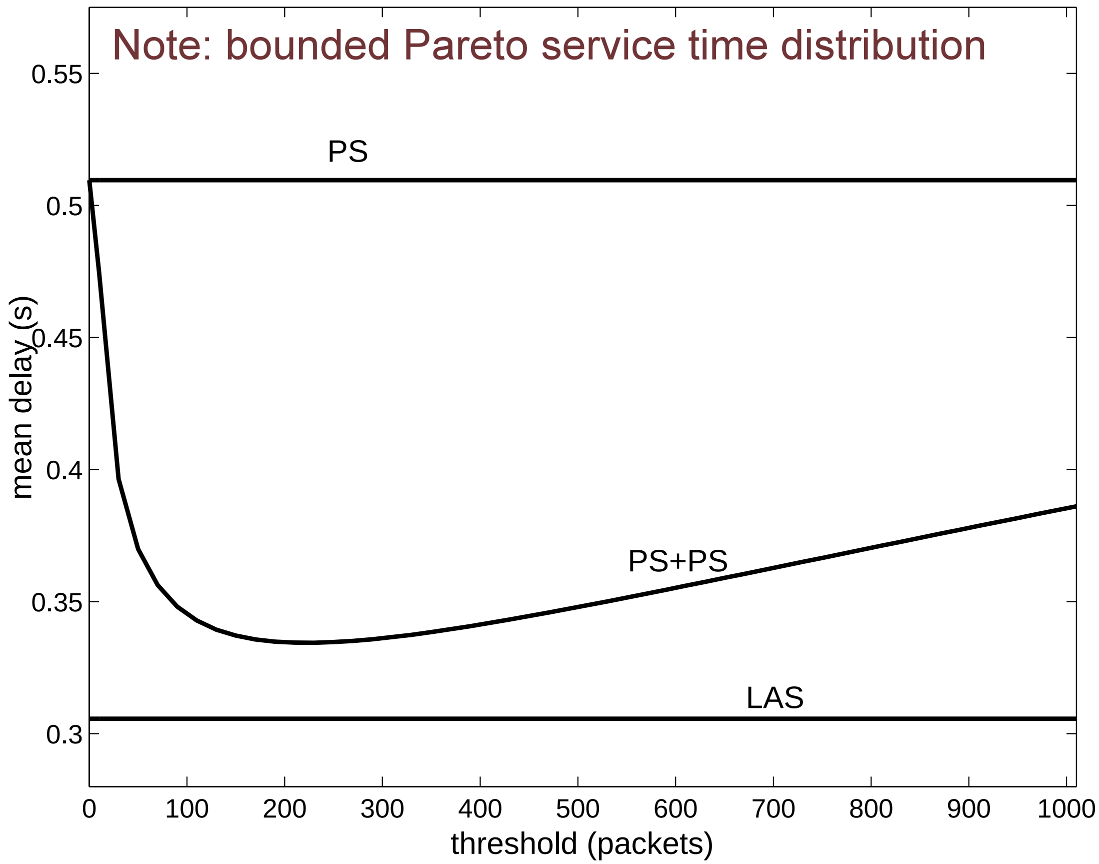
Asymptotic properties of the conditional mean delay E[T(s)]

Note: bounded Pareto service time distribution



- Conclusion: PS+PS seems to be better than FB in the asymptotic region (when hazard rate decreasing)

Mean delay E[T]



- Conclusion: PS+PS seems to be better than PS in the mean delay sense (when hazard rate decreasing)

Problem

- **Theorem**: With decreasing hazard rate,

$$E[T^{\text{FB}}] \le E[T^{\text{FB+PS}}] \le E[T^{\text{PS+PS}}] \le E[T^{\text{PS}}]$$

- Steps in the proof:
  - **First**: prove that for any work-conserving disciplines D_1 and D_2

$$E[U_x^{D_1}] \le E[U_x^{D_2}] \;\; \forall x \quad \Rightarrow \quad E[T^{D_1}] \le E[T^{D_2}]$$

    - T = delay
    - U_x = unfinished truncated work = sum of remaining **truncated** service times $\min\{S,x\}$ of those customers who have attained service at most x time units
  - **Second**: prove that for any x

$$E[U_x^{\text{FB}}] \le E[U_x^{\text{FB+PS}}] \le E[U_x^{\text{PS+PS}}] \le E[U_x^{\text{PS}}]$$

Solution: mean value arguments (1)

- **Proposition 1**: If no "future" information used, then

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

- Proof:
  - Kleinrock (1976) by Little's formula:

$$dE[N(y)] = \lambda G(y) dE[T(y)]$$

    - $N(y)$ = #customers with attained service time at most y
    - $T(y)$ = delay of a customer with service time y
  - Easy to see:

$$E[R_x(y)] = \frac{1}{G(y)} \int_y^x G(t) dt$$

    - $R_x(y) = \min\{S(y),x\} - \min\{y,x\}$ = remaining truncated service time of a customer with attained service time y
    - $S(y)$ = service time of a customer with attained service time y

Solution: mean value arguments (2)

– No "future" information used:

$$E[U_x] = \int_{0^-}^{x^-} E[R_x(y)]dE[N(y)]$$

- U_x = unfinished truncated work:

$$U_x = \sum_i (\min\{S_i, x\} - \min\{X_i, x\})$$

- S_i = service time of customer i
- X_i = attained service time of customer i

– By combining the results above, we finally get

$$(E[U_x])' = \lambda G(x) E[T(x)]$$

implying that

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

Solution: mean value arguments (3)

- **Proposition 2**: With decreasing hazard rate,

$$E[U_x^{D_1}] \leq E[U_x^{D_2}] \;\forall x \quad \Rightarrow \quad E[T^{D_1}] \leq E[T^{D_2}]$$

- Proof:
  - Follows directly from Proposition 1.
  - If the hazard rate differentiable, then simply by partial integration:

$$E[T^{D_1}] - E[T^{D_2}] = \frac{1}{\lambda} \int_0^\infty (E[U_x^{D_1}] - E[U_x^{D_2}])' h(x) dx$$

$$= -\frac{1}{\lambda} \int_0^\infty (E[U_x^{D_1}] - E[U_x^{D_2}]) h'(x) dx$$

Solution: mean value arguments (4)

- **Proposition 3**: For any a and x,

$$E[U_x^{\mathrm{PS+PS}(a)}] \le E[U_x^{\mathrm{PS}}]$$

- Proof:
  - From slide 7:

$$E[T^{\mathrm{PS+PS}}(s)] = \begin{cases} \frac{s}{1-\rho_a} \le \frac{s}{1-\rho} = E[T^{\mathrm{PS}}(s)], & s \le a \\ E[T^{\mathrm{FB}}(a)] + \frac{\alpha(s-a)}{1-\rho_a}, & s > a \end{cases}$$

  - Notation:

$$\alpha^* = \inf_{x>0} \alpha'(x)$$

  - From slide 8:

$$\inf_{s>a}(T^{\mathrm{PS+PS}}(s))' = \frac{\alpha^*}{1-\rho_a} \ge \frac{1}{1-\rho} = (T^{\mathrm{PS}}(s))'$$

Solution: mean value arguments (5)

- Notation:

$$x^* = \inf\{s \geq a \mid E[T^{\mathrm{PS+PS}}(s)] \geq E[T^{\mathrm{PS}}(s)]\}$$

- For all $x \leq x^*$,

$$E[U_x^{\mathrm{PS+PS}}] = \int_0^x \lambda G(s) E[T^{\mathrm{PS+PS}}(s)] ds$$

$$\leq \int_0^x \lambda G(s) E[T^{\mathrm{PS}}(s)] ds = E[U_x^{\mathrm{PS}}]$$

- For all $x > x^*$,

$$(E[U_x^{\mathrm{PS+PS}}])' = \lambda G(x) E[T^{\mathrm{PS+PS}}(x)]$$

$$\geq \lambda G(x) E[T^{\mathrm{PS}}(x)] = (E[U_x^{\mathrm{PS}}])'$$

- Finally, since both PS and PS+PS are work-conserving, we have

$$E[U_\infty^{\mathrm{PS+PS}}] = E[U_\infty^{\mathrm{PS}}]$$

Solution: sample path arguments (1)

- **Notation**: unfinished truncated work for discipline D at time t:

$$U_x^D(t) = \sum_{i=1}^{A(t)} (\min\{S_i, x\} - \min\{X_i(t), x\})$$
$$= \sum_{i=1}^{A(t)} \min\{S_i, x\} - \int_0^t \sigma_x^D(u)du$$

  - $A(t)$ = #arrivals up to time t
  - X_i = service time of customer i
  - $X_i(t)$ = attained service time of customer i at time t
  - $\sigma_x^D(t)$ = service rate of customers with attained service less than x at time t

- For any scheduling discipline D,

$$\sigma_x^D(t) = 0, \quad \text{if } N_x^D(t) = 0$$
$$\sigma_x^D(t) \leq 1, \quad \text{if } N_x^D(t) > 0$$

  - $N_x^D(t)$ = #customers with attained service less than x at time t

Solution: sample path arguments (2)

- **Definition**: set $D_x{}^*$ of scheduling disciplines:

$$D \in D_x^* \quad \Leftrightarrow \quad \sigma_x^D(t) = 1, \text{ if } N_x^D(t) > 0$$

- By definition, for any D^* in $D_x{}^*$, x, t,

$$U_x^{D^*}(t) = \min_D U_x^D(t)$$

- **Proposition 4**: For any a, x, t,

$$U_x^{\mathrm{FB}}(t) \le U_x^{\mathrm{FB+PS}(a)}(t) \le U_x^{\mathrm{PS+PS}(a)}(t)$$

- Proof:
  - Clearly, for all x and $a \ge x$,

$$\mathrm{FB}, \mathrm{FB} + \mathrm{PS}(a) \in D_x^*$$

  - On the other hand, for all $a \le x$,

$$\sigma_x^{\mathrm{FB+PS}(a)}(t) \equiv \sigma_x^{\mathrm{PS+PS}(a)}(t)$$

Solution: sample path arguments (3)

- Give an example of x and t such that

$$U_x^{\mathrm{PS+PS}}(t) > U_x^{\mathrm{PS}}(t)$$

- Not so easy. But it is another story …