Multilevel Processor Sharing Scheduling Disciplines: Mean Delay Analysis

Samuli Aalto, Eeva Nyberg
HUT/Networking Laboratory
Urtzi Ayesta
INRIA/MISTRAL

Background

- 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, also the mean "file transfer" time
- How to schedule flows and how to analyse?

References

- Earlier work:
  - K. Avrachenkov (INRIA), U. Ayesta (INRIA/FT), P. Brown (FT) and E. Nyberg (HUT):
  - “Differentiation between Short and Long TCP Flows: Predictability of the Response Time”
  - To be presented in IEEE INFOCOM 2004, Hong Kong, March 2004
- Present work:
  - S. Aalto (HUT), U. Ayesta (INRIA/FT), E. Nyberg (HUT):
  - “Multilevel Processor Sharing Scheduling Disciplines: Mean Delay Analysis”
  - Accepted to ACM SIGMETRICS – PERFORMANCE 2004, New York, June 2004

Mathematical model

- Consider a bottleneck link loaded with elastic flows
  - such as file transfers using TCP
- Assume that
  - flows arrive according to a Poisson process
  - each flow has a random service requirement (= file size) with a general distribution
  - Note: file sizes typically heavy-tailed such as Pareto $\Rightarrow$ **decreasing hazard rate**
- So, we have a ***M/G/1*** queue on the flow level
  - Note: customers in this queue are flows (and not packets)

Scheduling disciplines

- **PS** = Processor Sharing
  - Without any specific scheduling policy, the elastic 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**
- **FB** = Foreground-Background
  - Choose a packet of the flow with least packets **sent**
- **MLPS** = Multilevel Processor Sharing
  - Choose a packet of a flow with less packets sent than a given threshold

Known optimality results for M/G/1

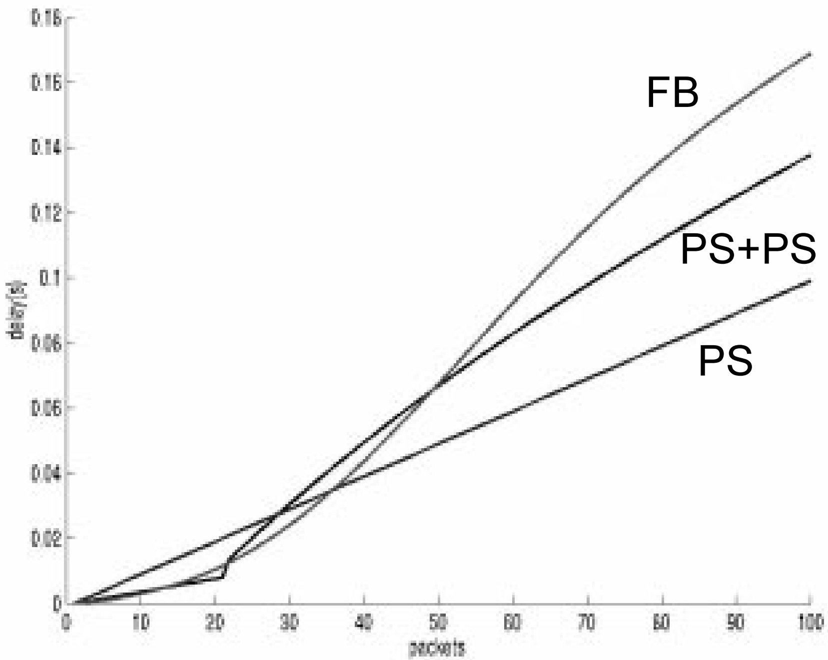
- If the number of packets **left** known, then
  - **SRPT optimal** minimizing the mean file transfer time
- If only the number of packets **sent** known, then
  - **decreasing hazard rate** implies that
    FB optimal among work-conserving scheduling disciplines

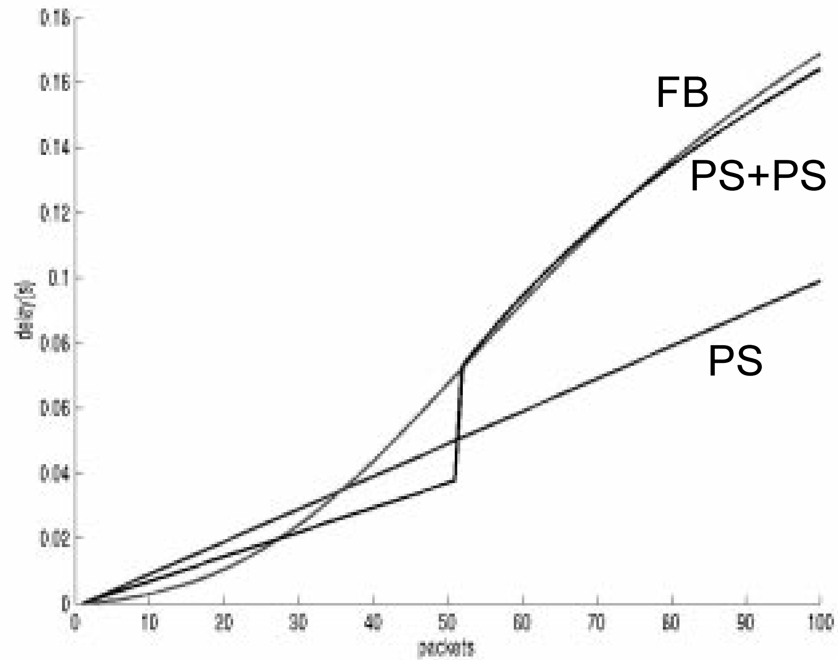
MLPS scheduling disciplines

- **Definition**: MLPS scheduling discipline
  - based on the attained service times (= #packets sent)
  - 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), Avrachenkov et al. (2004)

Conditional mean delay E[T(s)]

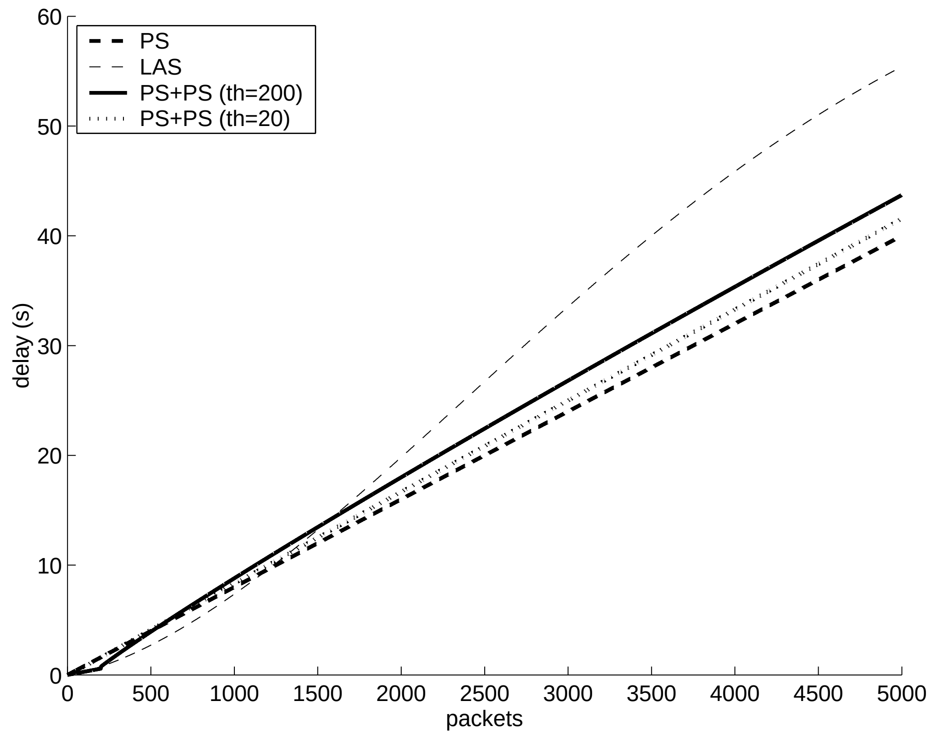
exponential file size distribution





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

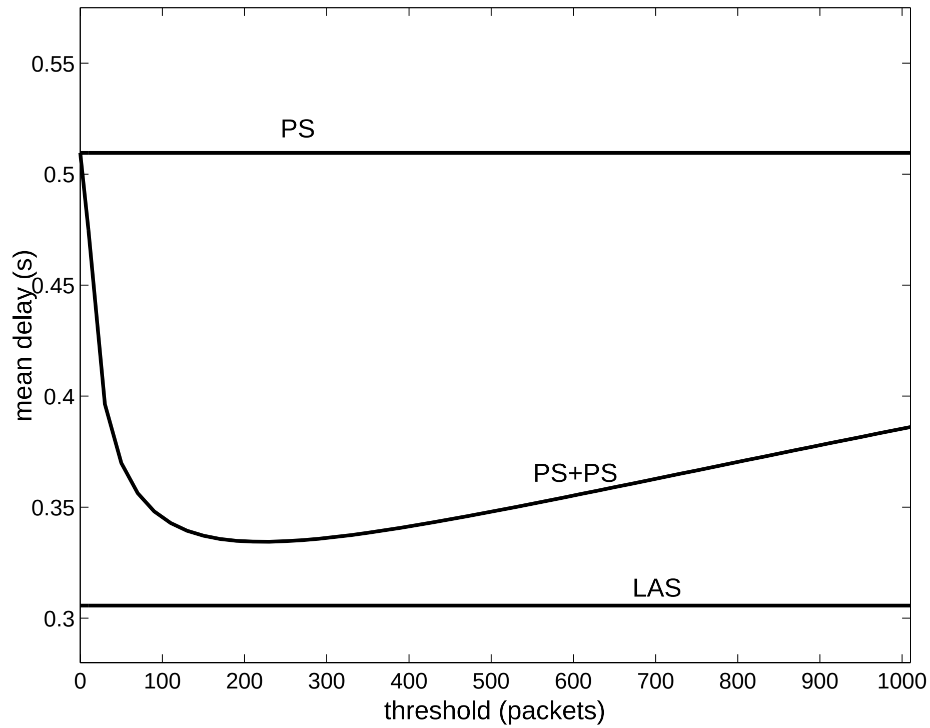
bounded Pareto file size distribution



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

Mean delay E[T]

bounded Pareto file size distribution



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

Problem that we solved

- **Theorem**:
  - With decreasing hazard rate, the order of the mean delays is as follows:

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

Solution: general comments

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

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

  - **Second**: prove that for any x

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

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

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:
  - Start with a known result from Kleinrock (1976)
  - Then, proceed along the lines of Feng and Misra (2003) but correcting their slight mistake

Solution: mean value arguments (2)

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

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

- Proof:
  - Follows directly from Proposition 1
  - If hazard rate differentiable, then 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 (3)

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

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

- Proof:
  - Based on a known analytical result concerning the conditional mean delays by Kleinrock (1976):

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

Solution: sample path arguments (1)

- **Definition**:
  - Unfinished truncated work for discipline D at time t:

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

  - $\sigma_x{}^D(t)$ = service rate of customers with attained service time less than x at time t

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

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

  - $N_x{}^D(t)$ = number of customers with attained service time less than x at time t

Solution: sample path arguments (2)

- **Definition**:
  - Set D_x* of scheduling disciplines:

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

- Observation:
  - By definition, for any D* in D_x* and any x, t,

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

Solution: sample path arguments (3)

- **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 $a \le x$,

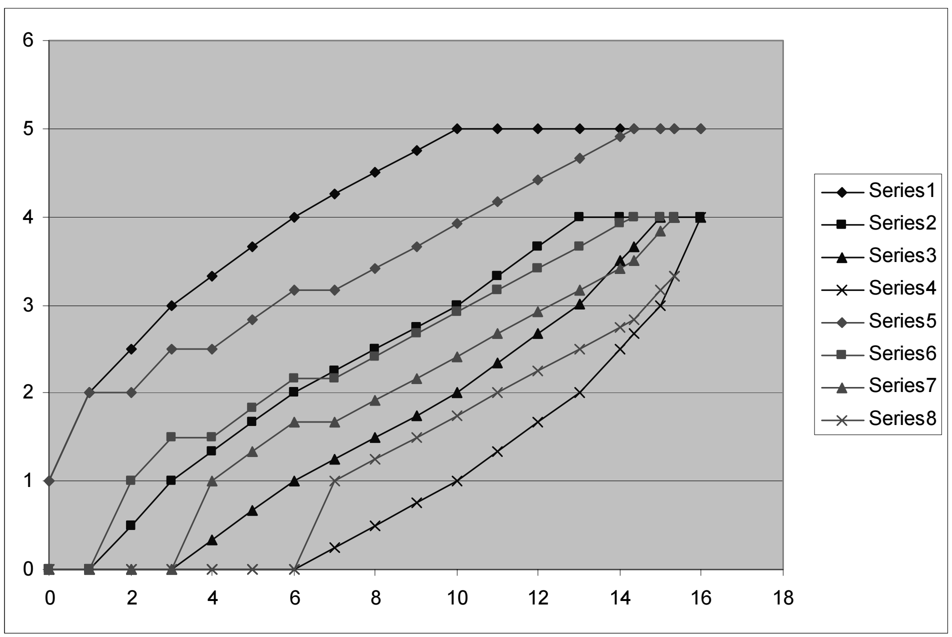
$$\sigma_x^{\mathrm{FB+PS}(a)}(t) \equiv \sigma_x^{\mathrm{PS+PS}(a)}(t)$$

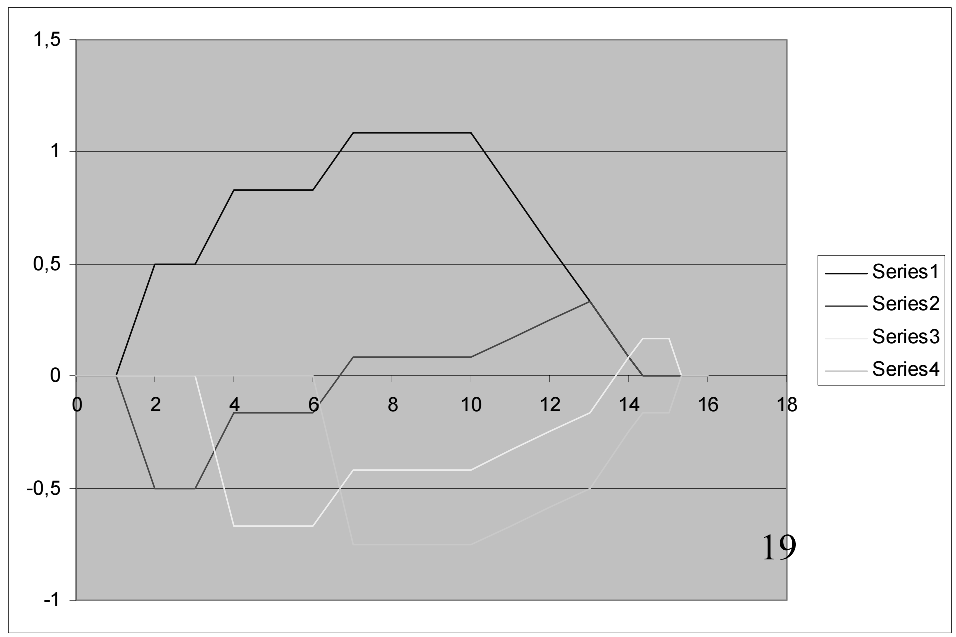
Solution: sample path arguments (4)

- 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 …





THE END