

The strict priority scheduler

- The strict priority scheduler is very simple and efficient
- In the following, we will carry out the analysis of this scheduler to determine the formulas for performance analysis, resource allocation and admission control
- The service envelope of service priority i , in the strict priority scheduler, is

$$S_i(t) = \max \left(0, Ct - \sum_{j=1}^{i-1} X_j(t) \right)$$

- Where C is the line capacity
- In fact, service class i is interfered only by service classes with a better service priority, that is, classes from 1 to $i-1$

The strict priority scheduler

- Thus

$$\alpha_i(t) = - \frac{E \left(X_i(t) - \max \left(0, C(t + d_i) - \sum_{j=1}^{i-1} X_j(t + d_i) \right) \right)}{\sqrt{\text{var} \left(X_i(t) - \max \left(0, C(t + d_i) - \sum_{j=1}^{i-1} X_j(t + d_i) \right) \right)}}$$

- The specific form of $\alpha_i(t)$ depends on the average values and variances of traffic flows from 1 to $i-1$

The strict priority scheduler

- When all traffic flows admit a linear variance envelope, a closed-form analysis is possible
- In this case, the average value of $S_i(t)$ is approximated as

$$E(S_i(t)) = Ct - \sum_{j=1}^{i-1} N_j r_j t$$

- And the variance of $S_i(t)$ is approximated as

$$\text{var}(S_i(t)) = \sum_{j=1}^{i-1} N_j r_j b_j t$$

- Thus,

$$\alpha_i(t) = \frac{-N_i r_i t + C(t + d_i) - \sum_{j=1}^{i-1} N_j r_j (t + d_i)}{\sqrt{N_i r_i b_i t + \sum_{j=1}^{i-1} N_j r_j b_j (t + d_i)}}$$

The strict priority scheduler

- The absolute minimum of $\alpha_i(t)$ is

$$\alpha_{i,\min} = 2\sqrt{\frac{(C - A_i) \left((C - A_{i-1}) B_i - (C - A_i) B_{i-1} \right)}{B_i^2}} d_i$$

- where

$$A_i = \sum_{j=1}^i N_j r_j \quad B_i = \sum_{j=1}^i N_j r_j b_j$$

- From which we can calculate the closed-form formulas for
 - ◆ Delay distribution
 - ◆ Capacity planning
 - ◆ Admission control

The strict priority scheduler: delay distribution with linear-bounded variance traffic

- Delay distribution

$$\Pr(D_i > d_i) \approx \exp\left(-2 \frac{C - A_i}{B_i^2} \left((C - A_{i-1}) B_i - (C - A_i) B_{i-1} \right) d_i \right)$$

- Average delay

$$E(D_i) \approx \frac{B_i^2}{2(C - A_i) \left((C - A_{i-1}) B_i - (C - A_i) B_{i-1} \right)}$$

The strict priority scheduler: capacity planning with linear-bounded variance traffic

- Given a strict priority scheduler with n service classes, where the traffic of class i , $X_i(t)$ has average value $E(X_i(t))=r_i t$ and variance upper bounded by $\text{var}(X_i(t)) \leq r_i b_i t$
- Given the statistical delay SLA of class i : (d_i, p_i)
- The line capacity C needed to satisfy concurrently all SLAs is obtained by inverting the delay distribution formula

$$C_i \geq \frac{(2A_i B_{i-1} - A_{i-1} B_i - A_i B_i) d_i - \sqrt{A_{i-1}^2 B_i^2 d_i^2 - 2A_{i-1} A_i B_i^2 d_i^2 + A_i^2 B_i^2 d_i^2 + 2B_{i-1} B_i^2 d_i \ln(p_i) - 2B_i^3 d_i \ln(p_i)}}{2(B_{i-1} - B_i) d_i}$$

- Then:

$$C = \sup_{1 \leq i \leq n} (C_i)$$

The strict priority scheduler: capacity planning with linear-bounded variance traffic

- The required capacity depends on the priority level assigned to each service class
- Given n service classes, there are $n!$ possible permutations for this assignment
- The problem of finding the optimal permutation is NP-complete
- Frequently, the best permutation is that assigning the traffic with the largest average value to priority 1, the second-largest average value to priority 2 and so on
- However, this is not necessarily true in the general case, as also variance has a significant effect

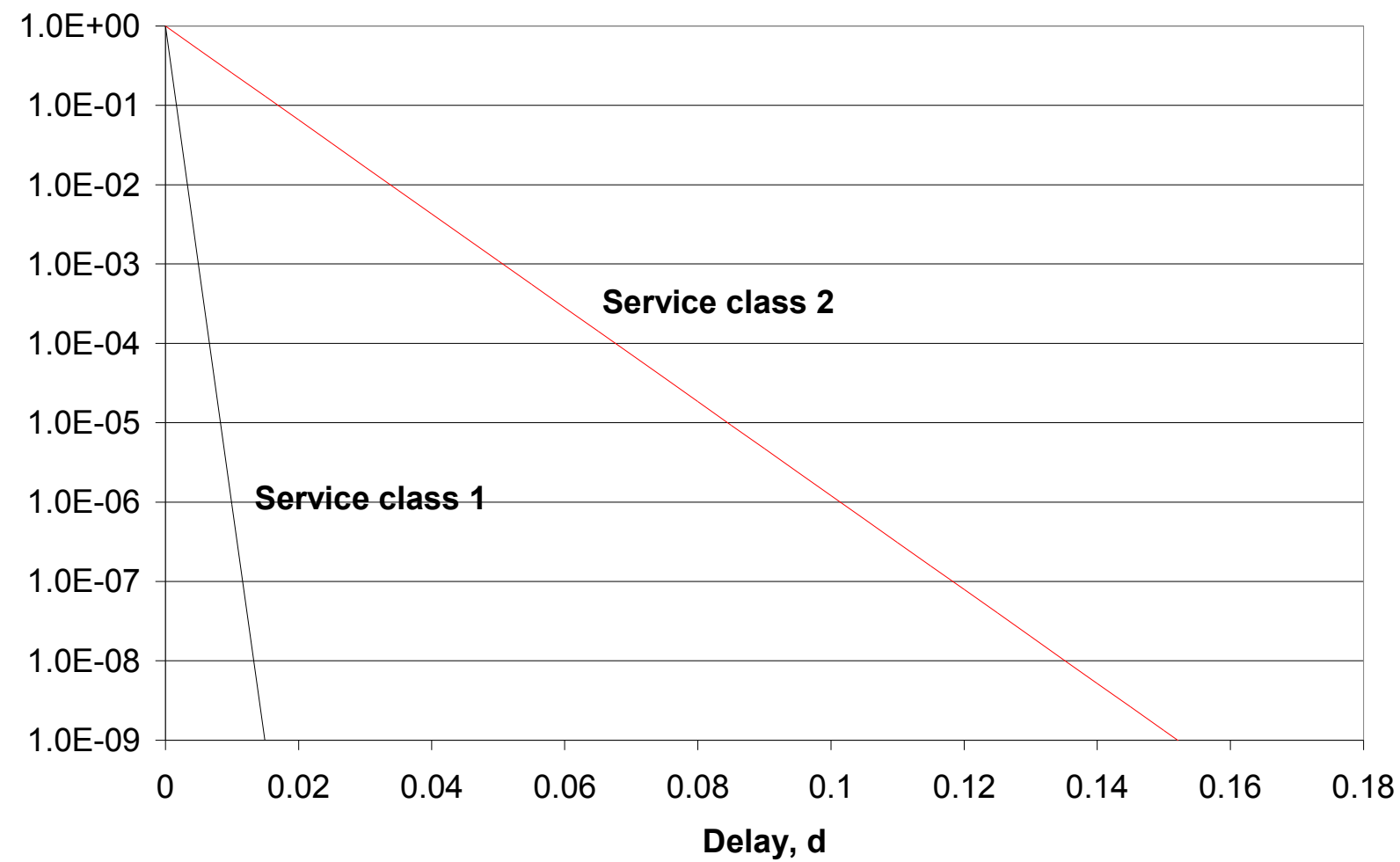
The strict priority scheduler: admission region with linear-bounded variance traffic

- Given a strict priority scheduler with n service classes, where the traffic of class i , $X_i(t)$ has average value $E(X_i(t))=r_i t$ and variance upper bounded by $\text{var}(X_i(t))\leq r_i b_i t$
- Given the statistical delay SLA of class i : (d_i, p_i)
- Given the line capacity C
- The admission region is the set of tuples (N_1, N_2, \dots, N_n) for which the SLAs are concurrently fulfilled
- The admission region is calculated by inverting the delay curve
- Algebraic calculations are very complex, but it is possible to find a closed-form expression of the admission region

The strict priority scheduler: an example of calculation of delay curves

- For example, let us consider the case of two service classes, with the following parameters
 - ◆ $r_1 = 200$ kbit/s, $b_1 = 9,600$ bit
 - ◆ $r_2 = 200$ kbit/s, $b_1 = 4,600$ bit
 - ◆ $N_1 = 30$ sources
 - ◆ $N_2 = 10$ sources
 - ◆ $C = 1 \times 10^7$ bit/s
- The delay curves in this case are calculated with the proposed formula and they are plotted in the next slide

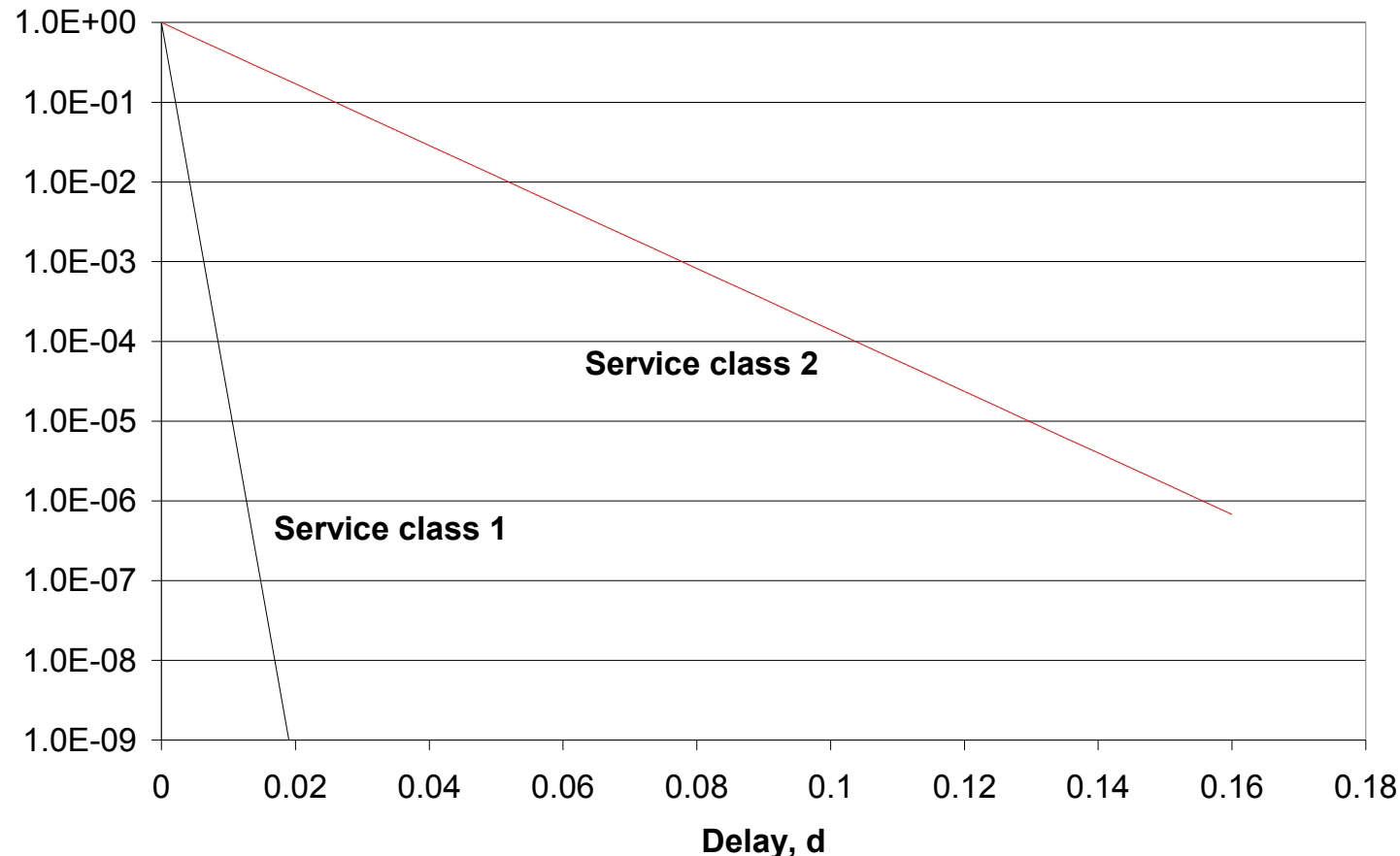
The strict priority scheduler : an example of calculation of delay curves



The strict priority scheduler : an example of calculation of capacity planning

- Continuing with the previous example, let us assume that capacity is unknown and let us add statistical delay SLAs
 - ◆ $r_1 = 200$ kbit/s, $b_1 = 9,600$ bit
 - ◆ $r_2 = 200$ kbit/s, $b_1 = 4,600$ bit
 - ◆ $d_1 = 50$ ms, $p_1 = 1 \times 10^{-3}$
 - ◆ $d_2 = 70$ ms, $p_2 = 2 \times 10^{-3}$
 - ◆ $N_1 = 30$ sources
 - ◆ $N_2 = 10$ sources
- The resulting capacity requested to fulfill concurrently all SLAs is:
 - ◆ $C_1 = 6.6 \times 10^6$ bit/s
 - ◆ $C_2 = 9.35 \times 10^6$ bit/s
- Thus
 - ◆ $C = \max(6.6 \times 10^6 \text{ bit/s}, 9.35 \times 10^6 \text{ bit/s}) = 9.35 \times 10^6 \text{ bit/s}$

The strict priority scheduler : an example of calculation of capacity planning



- In the scenario of the previous example, the delay curve is that plotted in the figure; the SLA of service class 2 is provided exactly, while the dimensioning for class 1 is large

The strict priority scheduler: admission region with linear-bounded variance traffic

$$\left\{ \begin{array}{l} N_i \leq \frac{\left(\left(\frac{B_{i-1}}{b_i} \right)^2 - \left(2C - 2A_{i-1} + \frac{B_{i-1}}{b_i} \right)^2 - 4B_{i-1} \frac{\ln p_i}{d_i} - 4 \left(C - A_{i-1} + \frac{B_{i-1}}{b_i} \right) \sqrt{\left(C - A_{i-1} \right)^2 + 2B_{i-1} \frac{\ln p_i}{d_i}} \right)}{-4r_i \left(2 \left(C - A_{i-1} + \frac{B_{i-1}}{b_i} \right) - \frac{b_i \ln p_i}{d_i} \right)} \quad \text{for } 0 \leq N_{i-k} \leq \min(N_{i-k}^*, N_{i-k, \max}), \text{ and } i > k = 1, 2, \dots, n \\ N_i = 0 \quad \text{for } N_{i-k}^* \leq N_{i-k} \leq N_{i-k, \max}, \text{ and } i > k = 1, 2, \dots, n \end{array} \right.$$

• where

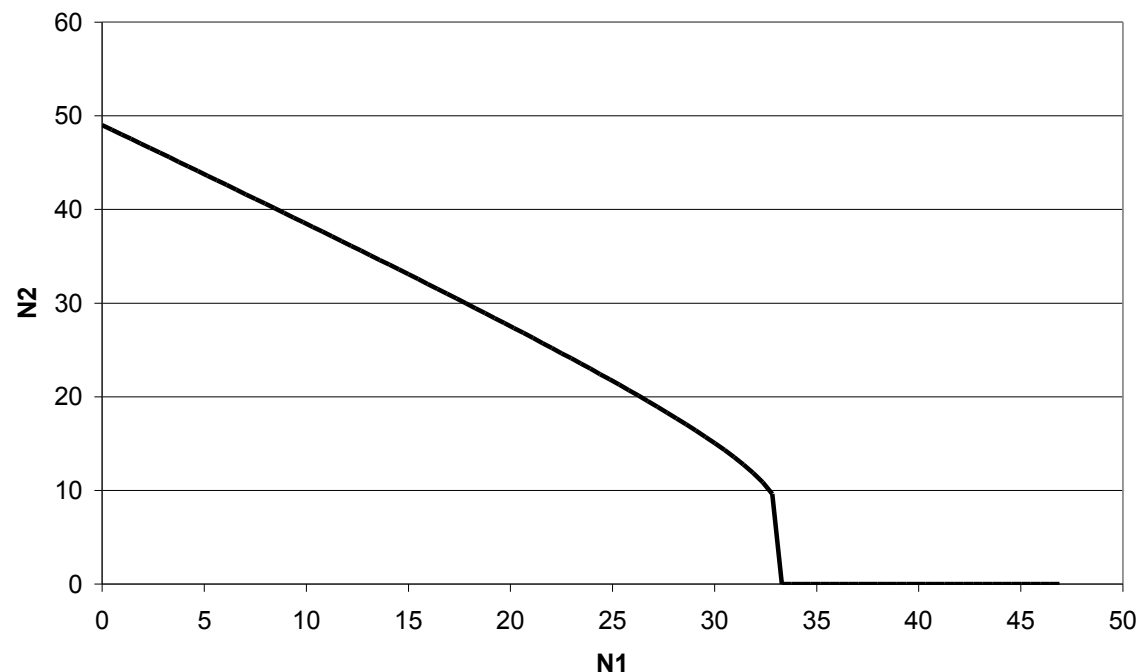
$$N_{i-k}^* = \frac{(Cd_i - b_{i-k} \ln p_i - d_i A_{i-1-k}) - \sqrt{-\ln p_i (-b_{i-k}^2 \ln p_i + 2d_i b_{i-k} (C - A_{i-1-k}) + 2d_i B_{i-1-k})}}{r_{i-k} d_i}$$

$$N_{i-k, \max} = \frac{2d_{i-k} C^2}{2d_{i-k} r_{i-k} C - r_{i-k} b_{i-k} \ln p_{i-k}}$$

Admission control curves for the SP scheduler

- For example, with the selected parameters, the admission control region is plotted in the figure
- It is interesting to note that for more than 34 sources of class 1, no sources of class 2 can be admitted
- This depends on the specific values of the TCA and SLA parameters
- With other parameters, this phenomenon does not necessarily occur

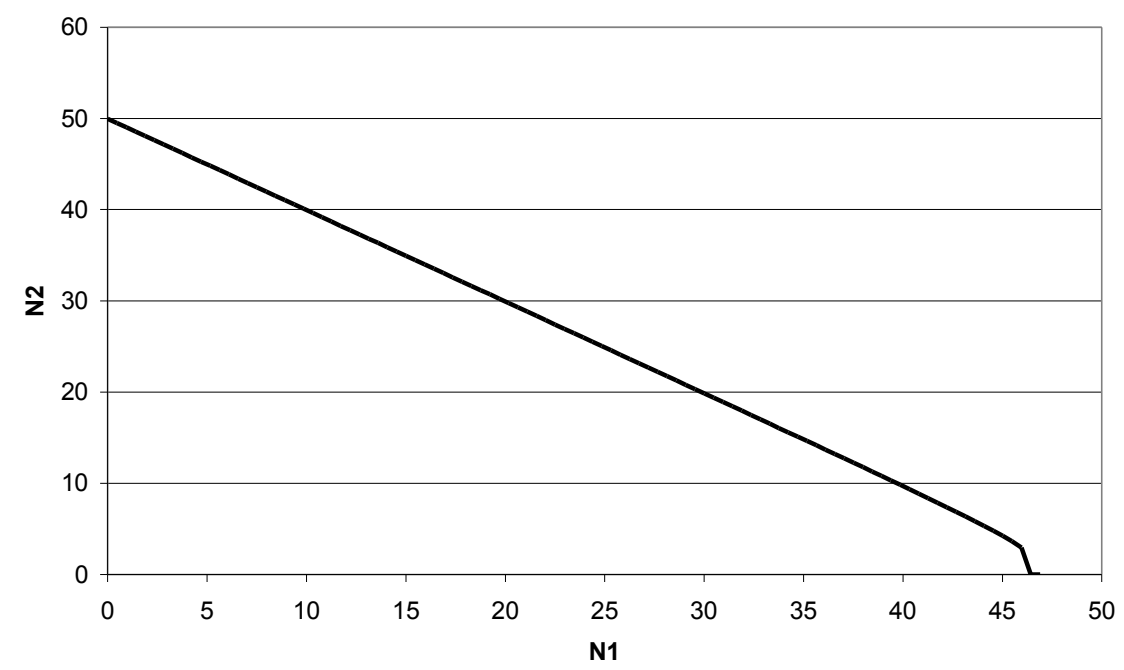
class 1		class 2	
r1	2.00E+05 bit/s	r2	2.00E+05 bit/s
b1	9.60E+03 bit	b2	4.60E+03 bit/s
r1b1	1.92E+09 bit ² /s	r2b2	9.20E+08 bit ² /s
d1	0.05	d2	0.07
p1	1.00E-03	p2	2.00E-03
C		1.00E+07	



Admission control curves for the SP scheduler

- This is another example, providing a qualitatively different admission region

class 1			class 2		
r1	2.00E+05	bit/s	r2	2.00E+05	bit/s
b1	9.60E+03	bit	b2	4.60E+03	bit/s
r1b1	1.92E+09	bit ² /s	r2b2	9.20E+08	bit ² /s
d1	0.05		d2	0.8	
p1	1.00E-03		p2	1.00E-01	
C			1.00E+07		



The GPS scheduler

- In a GPS with n service classes the statistical service of the class- i traffic is given by

$$S_i(t) = w_i C t + \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} \max(0, w_j C t - X_j(t)), \quad \begin{cases} j = 1, \dots, n \\ k = 1, \dots, n \text{ backlogged classes} \end{cases}$$

- Where the j index represents non-backlogged classes while the k index refers to backlogged classes
- Therefore, the average value of the statistical service envelope available for service class i can be evaluated as

$$\alpha_i(t) = \frac{E \left[X_i(t) - \left(w_i C(t+d_i) + \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} \max(0, w_j C(t+d_i) - X_j(t+d_i)) \right) \right]}{\sqrt{\text{var} \left[X_i(t) - \left(w_i C(t+d_i) + \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} \max(0, w_j C(t+d_i) - X_j(t+d_i)) \right) \right]}}$$

The GPS scheduler

- With linear-variance bounded traffic,

$$E(S_i(t)) \approx w_i C t + \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} (w_j C t - N_j r_j t)$$

$$\text{var}(S_i(t)) \approx \sum_{j \neq i} \left(\frac{w_i}{\sum_{k \neq j} w_k} \right)^2 N_j r_j b_j t$$

- thus

$$\alpha_i(t) = \frac{-N_i r_i t + \left(w_i C + \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} (w_j C - N_j r_j) \right) (t + d_i)}{\sqrt{N_i r_i b_i t + \sum_{j \neq i} \left(\frac{w_i}{\sum_{k \neq j} w_k} \right)^2 N_j r_j b_j (t + d_i)}}$$

The GPS scheduler

- The absolute minimum of $ai(t)$ is

$$\alpha_{i,\min} = \sqrt{\frac{4N_i r_i d_i (w_i b_i C + b_i A_i + B_i) \times (w_i C + A_i - N_i r_i)}{(N_i r_i b_i + B_i)^2}}$$

- where

$$A_i = \sum_{j \neq i} \frac{w_i}{\sum_{k \neq j} w_k} \times \max(0, w_j C - N_j r_j); \quad B_i = \sum_{j \neq i} \left(\frac{w_i}{\sum_{k \neq j} w_k} \right)^2 N_j r_j b_j.$$

The GPS scheduler: probability of violation of delay bound

• Thus,

$$\Pr(D_i > d_i) \approx \exp\left(-\frac{2N_i r_i (w_i b_i C + b_i A_i + B_i) (w_i C + A_i - N_i r_i)}{(N_i r_i b_i + B_i)^2} d_i\right)$$

The GPS scheduler: admission control rule

- For the admission control rule, we define

$$N_{i,\min} = \frac{2d_i(w_i C)^2}{r_i(2d_i w_i C - b_i \ln p_i)} \quad N_{i,\max} = \frac{d_i(w_i C + A_i)(w_i b_i C + b_i A_i + B_i) + b_i B_i \ln p_i + \sqrt{d_i(w_i b_i C + b_i A_i + B_i)^2 (d_i(w_i C + A_i)^2 + 2B_i \ln p_i)}}{r_i(2d_i(w_i b_i C + b_i A_i + B_i) - b_i^2 \ln p_i)}$$

- Then the following condition is checked

$$\text{If } \left(d_i(w_i C + A_i)^2 + 2B_i \ln p_i \right) > 0 \quad \text{then } N_i \leq \max(N_{i,\min}, N_{i,\max})$$
$$\text{else } N_i \leq N_{i,\min}$$

SP versus GPS: an example

- The figure shows a comparison between the admission curves of a Strict Priority and a GPS scheduler
- The GPS scheduler guarantees a minimum bandwidth for all classes
- Thus, service class 2 is not truncated as in the case of the SP scheduler

r1	600000
b1	40000
r2	800000
b2	100000
d1	0.01
d2	0.05
p1	0.000001
p2	0.000001
C	45000000
w1	0.8
w2	0.2

