

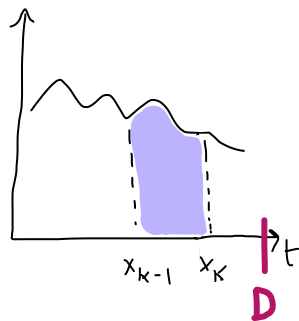
Se riprendiamo, per il caso astratto, la formula del tempo medio in coda, abbiamo:  $E[T_q] = \frac{E[S_{sum}]}{\left(1 - \sum_{i=1}^k p_i\right) \left(1 - \sum_{i=1}^{k-1} p_i\right)}$

Nel caso size-based, l' "unico" cambiamento lo si ha in come calcoliamo  $p_i = \lambda_i E[S_i]$

Nel caso astratto  $E[S_i]$  è uguale per ogni classe, qui varia.

Nel caso astratto il tasso di arrivo della singola classe dipende da fattori esterni alla size, ora dipende dalla size.

Possiamo dunque dire che nel size-based entrambi questi fattori dipendono dalla distribuzione del tempo di servizio.



la sottoarea viola mi dice quanti job sono di classe k. La funzione in figura è la densità del tempo di servizio  $f(t)$ , che è continua. Ciò che si fa è fissare un upper bound "D", grande abbastanza per far sì che la probabilità di servizi più grandi di D tenda a 0.

Un utente chiede servizio "s", a quale classe appartiene? Appartiene alla classe "k" se  $s \in (X_{k-1}, X_k]$

Poichè si tratta di un'area, dobbiamo integrare.

L'area viola è data da:  $F(X_k) - F(X_{k-1})$ , dove F è la cumulativa.

Se il tasso è  $\lambda$  (generale), a me arriva una sottoporzione data da:

$$\lambda_k = \lambda (F_{(X_k)} - F_{(X_{k-1})}) \quad \text{Quindi quell'area è come se fosse la "Probabilità di cadere lì"}$$

Passiamo adesso al tempo di servizio medio  $E[S_k] = \int_{x_{k-1}}^{x_k} t \cdot f(t) dt$

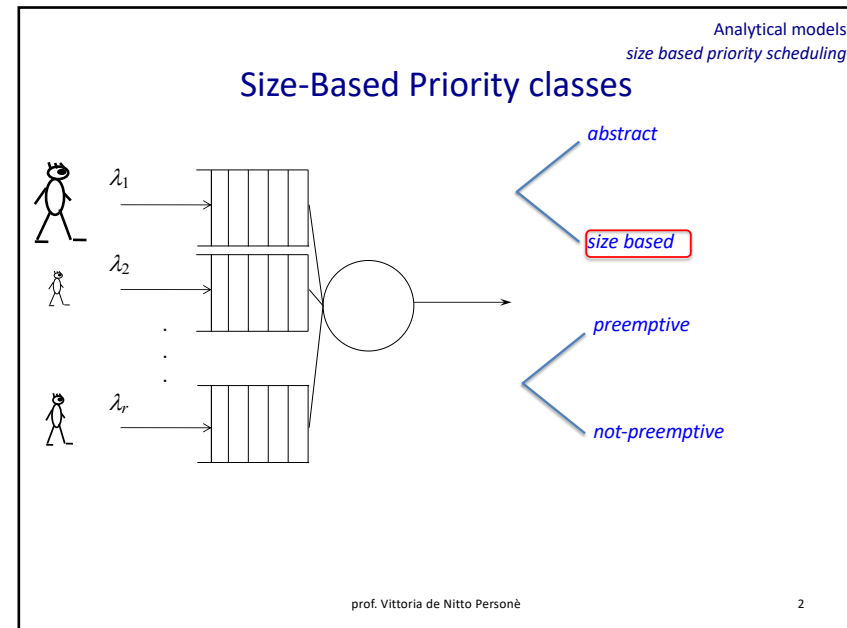
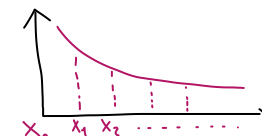
dove  $f(t)$  viene "Normalizzata", ovvero la densità viene normalizzata per poi avere somma 1.

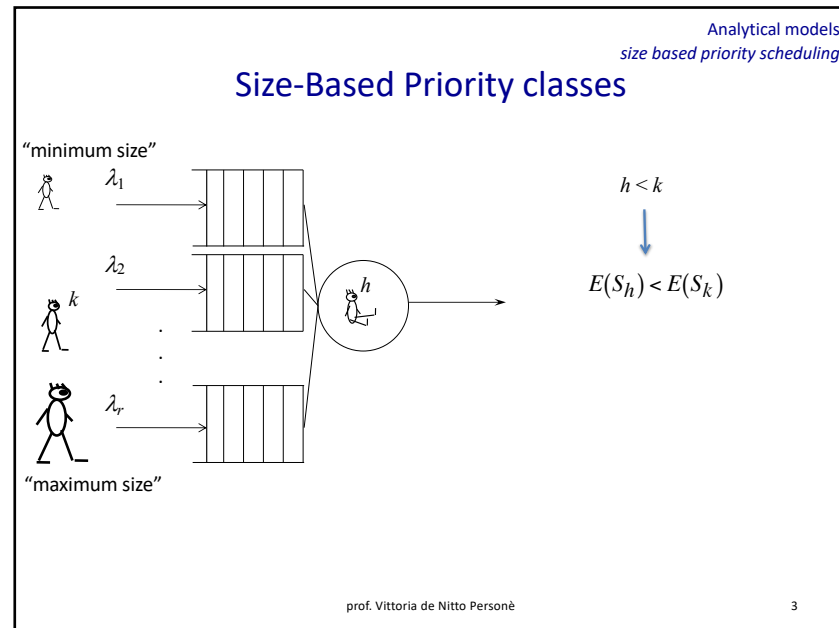
Nell'integrale abbiamo un'impostazione di tipo "peso" \* "valore assunto" (per quanto tempo assumo un comportamento \* il comportamento). Possiamo notare che, nel calcolo di  $p_k = \lambda_k \cdot E[S_k]$ ,  $F_{(X_k)} - F_{(X_{k-1})}$  si semplifica (appare sia in lambda che nella normalizzazione dentro  $E[S_k]$ ).

Osservazione: prendiamo un'esponenziale (ma vale anche per hyper, etc con forma simile).

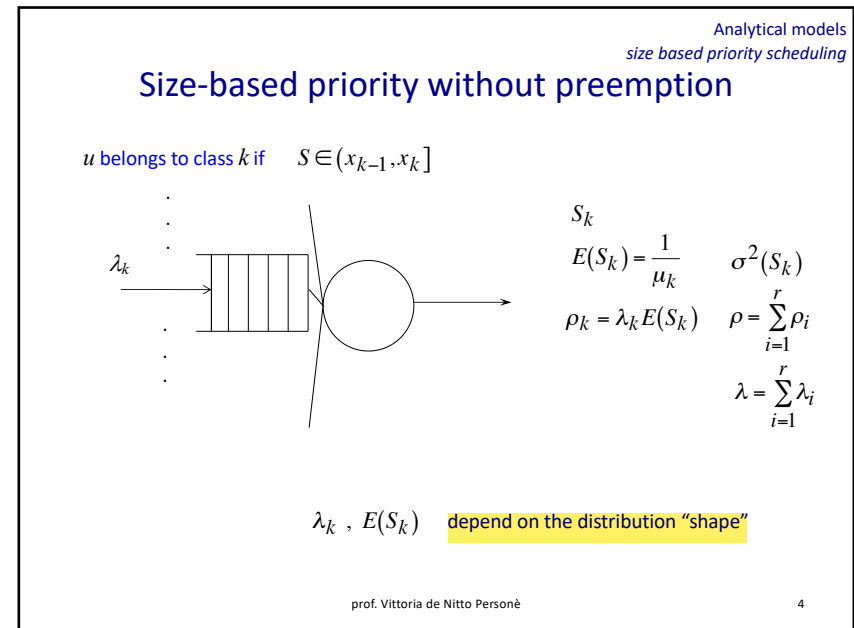
Si osserva che  $\lambda_1 > \lambda_2 \geq \lambda_3$ , ovvero l'area è monotona decrescente. Se pensiamo ai tempi di servizio associati, possiamo invece dire:  $E[S_1] < E[S_2] < \dots < E[S_n]$  ovvero è monotona crescente. Esiste una proprietà che ci dice che: la somma di una serie crescente ed una serie decrescente forma una somma minima

Per questo la size based di classe k è migliore di una astratta di classe k, perchè è sempre una somma minima, sotto non scendo!





3



4

Analytical models  
size based priority scheduling

$$E(S_k) = \int_{x_{k-1}}^{x_k} t f^n(t) dt$$

$$f^n(t) = \frac{f(t)}{F(x_k) - F(x_{k-1})}$$

densità normalizzata, così le medie sono 'corrette' in modo che la somma sia 1.

$$\lambda_k = \lambda(F(x_k) - F(x_{k-1})), \quad p_k = \frac{\lambda_k}{\lambda} = F(x_k) - F(x_{k-1}) = \text{probabilità di essere di classe } k$$

$$\rho_k = \lambda_k E(S_k) = \lambda(F(x_k) - F(x_{k-1})) \int_{x_{k-1}}^{x_k} t f^n(t) dt$$

$$= \lambda(F(x_k) - F(x_{k-1})) \int_{x_{k-1}}^{x_k} t \frac{f(t)}{F(x_k) - F(x_{k-1})} dt$$

$$= \lambda \int_{x_{k-1}}^{x_k} t f(t) dt$$

comunque lambda è sempre dipendente dalla classe k, anche se qui si semplifica.

prof. Vittoria de Nitto Personè

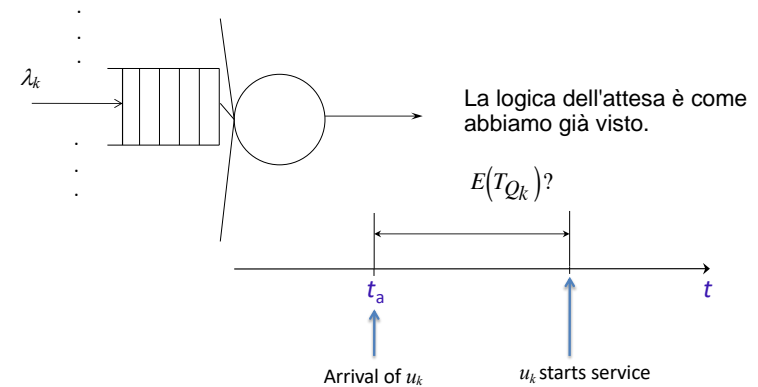
5

5

Analytical models  
size based priority scheduling

## Size-based priority without preemption

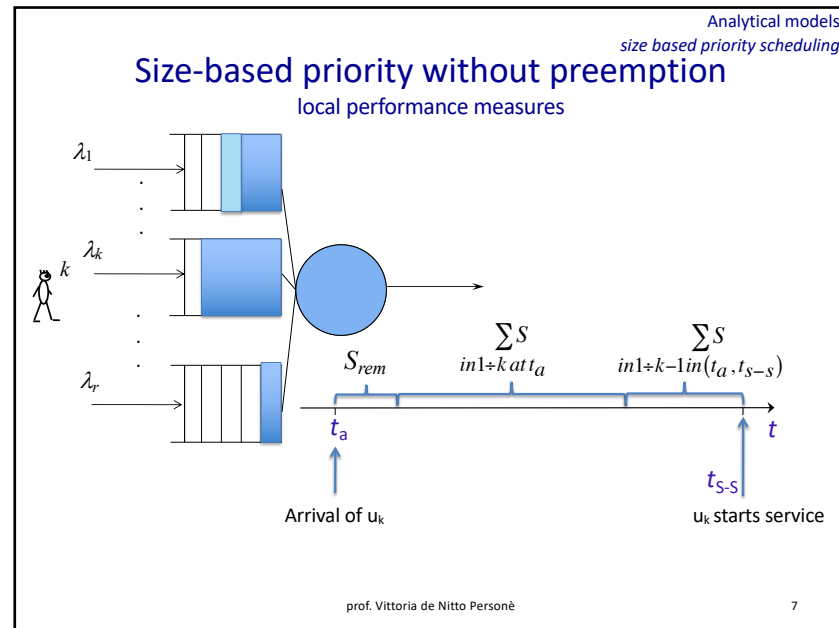
local performance measures



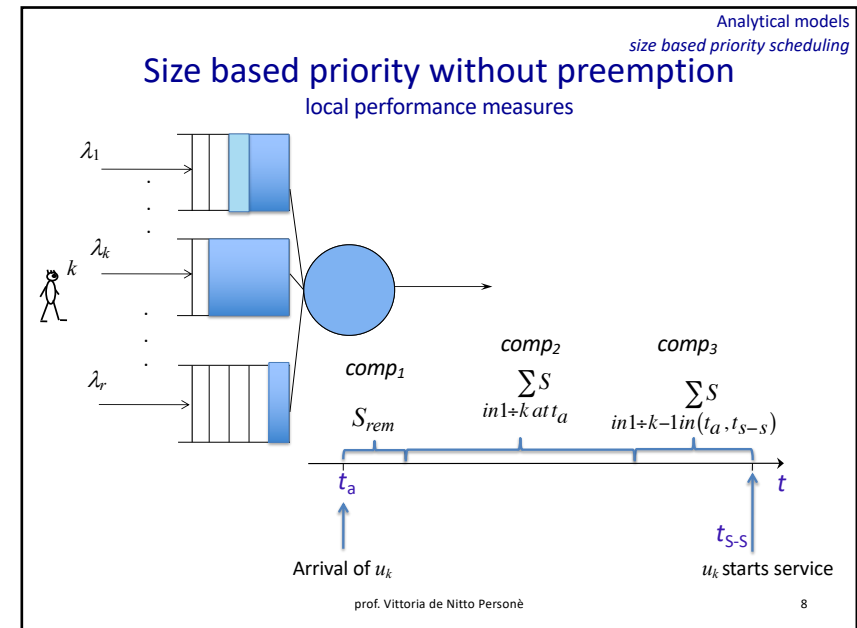
prof. Vittoria de Nitto Personè

6

6

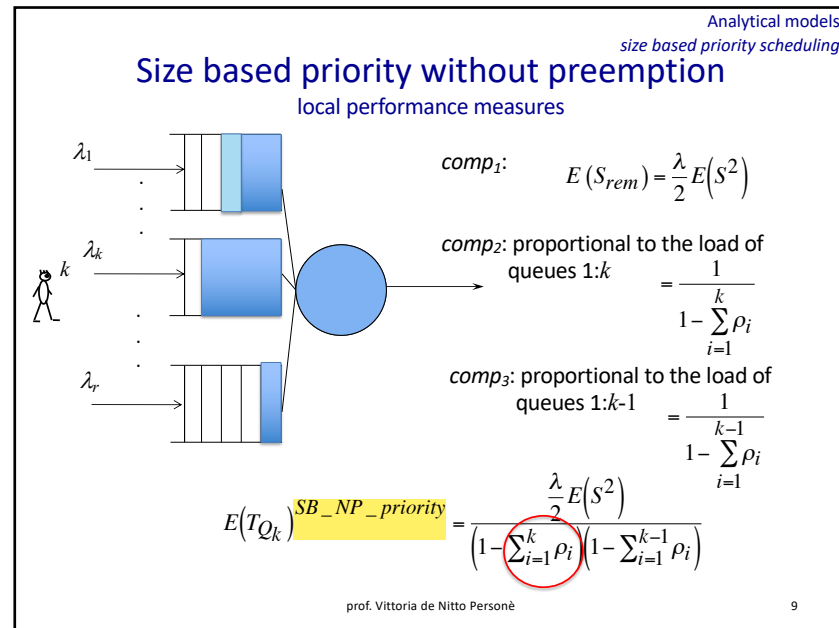


7



8

ciò che cambia è come scriviamo "rho", le formule sono uguali.



Analytical models  
size based priority scheduling

$$\rho_k = \lambda \int_{x_{k-1}}^{x_k} tf(t) dt$$

$$\sum_{i=1}^k \rho_i = \sum_{i=1}^k \lambda \int_{x_{i-1}}^{x_i} tf(t) dt$$

$$= \lambda \int_0^{x_k} tf(t) dt$$

$$E(T_{Q_k})^{SB\_NP\_priority} = \frac{\frac{\lambda}{2} E(S^2)}{\left(1 - \lambda \int_0^{x_k} tf(t) dt\right) \left(1 - \lambda \int_0^{x_{k-1}} tf(t) dt\right)}$$

prof. Vittoria de Nitto Personè 10

Analytical models  
size based priority scheduling

## Size based priority without preemption

global performance measures

And the "global" performance?

somma pesata delle attese delle varie classi.

$$E(T_Q)^{SB-NP-priority} = E(E(T_{Q_k})) = \sum_{k=1}^r p_k E(T_{Q_k})$$

$$p_k = \frac{\lambda_k}{\lambda} = \frac{\lambda(F(x_k) - F(x_{k-1}))}{\lambda} = F(x_k) - F(x_{k-1})$$

probabilità per ogni "k"

$$E(T_Q)^{SB-NP} = \frac{\lambda}{2} E(S^2) \sum_{k=1}^r \frac{F(x_k) - F(x_{k-1})}{\left(1 - \lambda \int_0^{x_k} tf(t)dt\right) \left(1 - \lambda \int_0^{x_{k-1}} tf(t)dt\right)}$$

prof. Vittoria de Nitto Personè

11

Analytical models  
priority scheduling

## Size-based vs abstract priority

local performance measures

$$E(T_{Q_k})^{NP-priority} = \frac{\frac{\lambda}{2} E(S^2)}{\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

$$E(T_{Q_k})^{SB-NP} \leq E(T_{Q_k})^{abstract-NP}$$

$$\left[\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)\right]^{SB-NP} \geq \left[\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)\right]^{abstract-NP}$$

$$\left[\sum_{i=1}^h \rho_i\right]^{SB-NP} \leq \left[\sum_{i=1}^h \rho_i\right]^{abstract-NP} \quad \text{for each } h$$

$$E(T_{S_k})^{SB-NP} \not\geq E(T_{S_k})^{abstract-NP}$$

prof. Vittoria de Nitto Personè

12

Sulla risposta classe per classe non posso dire nulla, il tempo di servizio SIZEBASED sulle classi meno importanti, possono eccedere la media del tempo. Non posso dirlo a priori.

## Size-based vs abstract priority

global performance measures

(come indice globale,  
non in funzione della  
classe)

$$E(T_Q)^{SB-NP} \leq E(T_Q)^{abstract-NP}$$

$$E(T_S)^{SB-NP} \leq E(T_S)^{abstract-NP}$$

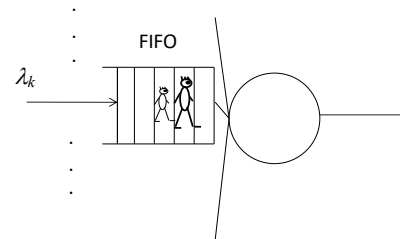
$$E(T_S)^{x-NP} = E(T_Q)^{x-NP} + E(S)^{x-NP}$$

(sia size based che astratto)

$$E(S)^{SB-NP} = E(S)^{abstract-NP} = E(S)$$

Tra abstract e SB cambia l' "ordinamento delle classi", ma i valori dei tempi di servizio sono uguali, e quindi la media è sempre quella.

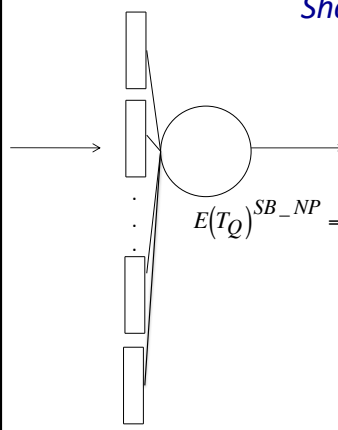
## Size based priority without preemption



Nelle singole classi, abbiamo ordine FIFO. Senza prelazione, nella coda posso avere disordine in base alla size. Per ordinare, perchè non ordino tra le size presenti?  
cioè invece di raggruppare per classi, raggruppo per singolo tempo di servizio richiesto.

Analytical models  
size based priority scheduling

### Size based priority without preemption Shortest Job First



$r \rightarrow \infty$

faccio il limite di  $E[T_q]$  SB\_NP

$$E(T_Q)^{SB\_NP} = \frac{\lambda}{2} E(S^2) \sum_{k=1}^r \frac{F(x_k) - F(x_{k-1})}{\left(1 - \lambda \int_0^{x_k} t f(t) dt\right) \left(1 - \lambda \int_0^{x_{k-1}} t f(t) dt\right)}$$

$$E(T_Q)^{SJF} = \frac{\lambda}{2} E(S^2) \int_0^{\infty} \frac{dF(x)}{\left(1 - \lambda \int_0^x t f(t) dt\right)^2}$$

prof. Vittoria de Nitto Personè

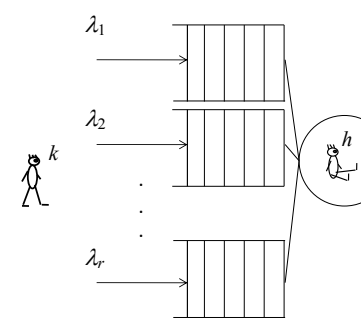
15

15

eseguo sempre job con size più piccola (ovviamente sapendo la size).  
Per  $r \rightarrow \infty$ , non calcolo più prestazioni singola classe (questo "concetto" sparisce)  
ma solo prestazioni globali, in quanto "ordino il singolo".

Analytical models  
priority scheduling

### SB preemptive priority



se arriva job  $k < h$   
più importante  
 $E(S_k) < E(S_h)$

ma se il job h stava proprio per finire, mi conviene cambiarli?

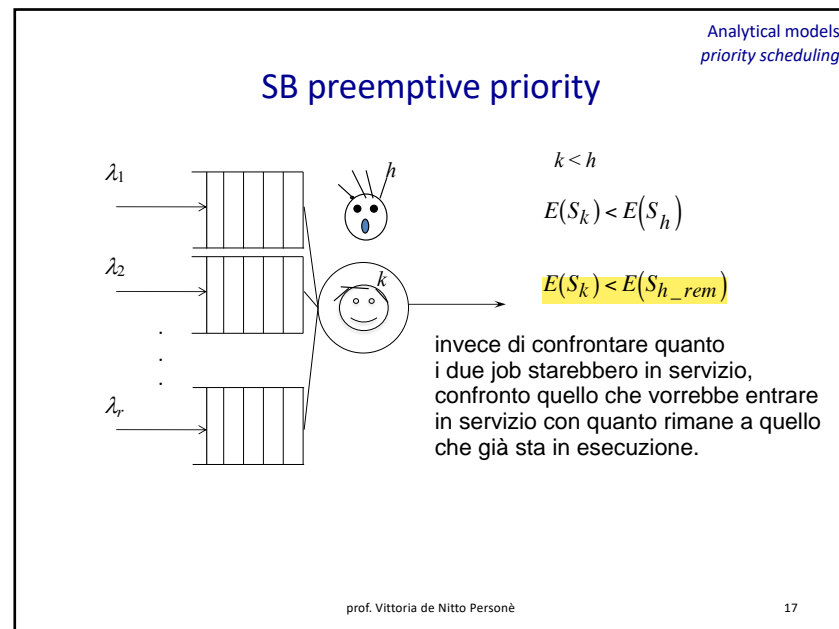
posso fare prelazione solo se:  
*Preemption without loss:*  
the service of the interrupted job  
will be resumed from the  
interruption point

prof. Vittoria de Nitto Personè

16

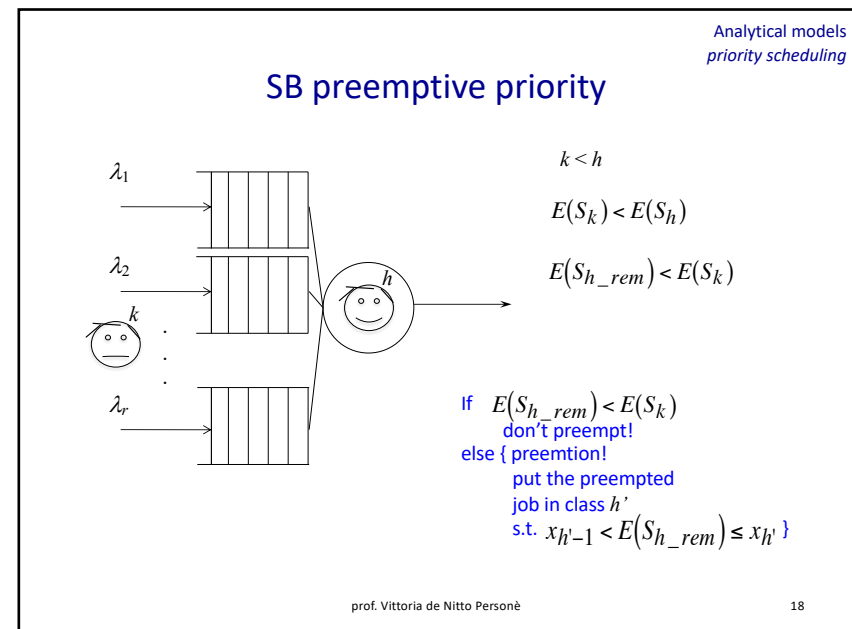
16





17

Se invece faccio lo scambio, il job tolto lo metto non nella sua classe di origine, bensì nella classe in funzione a quanto tempo di servizio gli rimane!  
(es: classe 0: esecuzione 1 min, classe 1: esecuzione 2 min)  
Se job classe 1 esegue, e gli rimangono 58 secondi, ma viene sostituito, viene messo in classe 0, non in classe 1.



18

Analytical models  
priority scheduling

### SB preemptive priority

local performance measures

$$E(T_{Qk})^{SB-P} = \frac{??}{\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

the remaining service time of the non-preemptible jobs + the remaining service time of the preemptible jobs that immediately restart service

(non interrompibili)

(interrompibili, ma nel valutarlo vedo che tempo rimanente è < k, e quindi riprendono servizio subito)

prof. Vittoria de Nitto Personè 19

19

Analytical models  
priority scheduling

### SB preemptive priority

local performance measures

the remaining service time of the non-preemptible jobs

abstract:  $\frac{1}{2} \sum_{i=1}^k \lambda_i E(S^2)$  tempo servizio rimanente con prelazione

size based:  $E(S_k^2) = \int_{x_{k-1}}^{x_k} t^2 f^n(t) dt$

$\frac{\lambda}{2} \int_0^{x_k} t^2 dF(t)$

prof. Vittoria de Nitto Personè 20

20

Analytical models  
priority scheduling

### SB preemptive priority

local performance measures

the remaining  
service time of the  
non-preemptible jobs

$$\frac{\lambda}{2} \int_0^{x_k} t^2 dF(t)$$

the remaining  
service time of the  
preemptible jobs that  
immediately restart  
service

$$\frac{\lambda}{2} x_k^2$$

Prob {preemptible job} = Prob {S > x<sub>k</sub>} = 1 - Prob {S ≤ x<sub>k</sub>}

prof. Vittoria de Nitto Personè

21

21

Analytical models  
priority scheduling

### SB preemptive priority

local performance measures

$$E(T_{Q_k})^{SB-P} = \frac{\frac{\lambda}{2} \left[ \int_0^{x_k} t^2 dF(t) + (1 - F(x_k)) x_k^2 \right]}{\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

$$E(T_{Q_k}) \leq E(T_{Q_{k+1}})$$

$$E(T_{Q_k})^{SB-P} \leq E(T_{Q_k})^{SB-NP}$$

$$E(T_{S_k}) = E(T_{Q_k}) + E(S_{virt-k})$$

dove conto tutte le  
"interruzioni"

$$E(S_{virt-k}) = \frac{E(S_k)}{1 - \sum_{i=1}^{k-1} \rho_i}$$

allungamento tempo di servizio  
proporzionale al carico delle classi  
che hanno diritto di interrompere.

prof. Vittoria de Nitto Personè

22

22

## SB preemptive priority

global performance measures

Analytical models  
priority scheduling

$$E(T_Q)^{SB-P} = E\left(E(T_{Q_k})^{SB-P}\right) = \sum_{k=1}^r p_k E(T_{Q_k})^{SB-P}$$

$$p_k = \frac{\lambda_k}{\lambda} = F(x_k) - F(x_{k-1})$$

$$E(T_Q)^{SB-P} = \frac{\lambda}{2} \sum_{k=1}^r \frac{[F(x_k) - F(x_{k-1})] \left[ \int_0^{x_k} t^2 dF(t) + (1 - F(x_k)) x_k^2 \right]}{\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

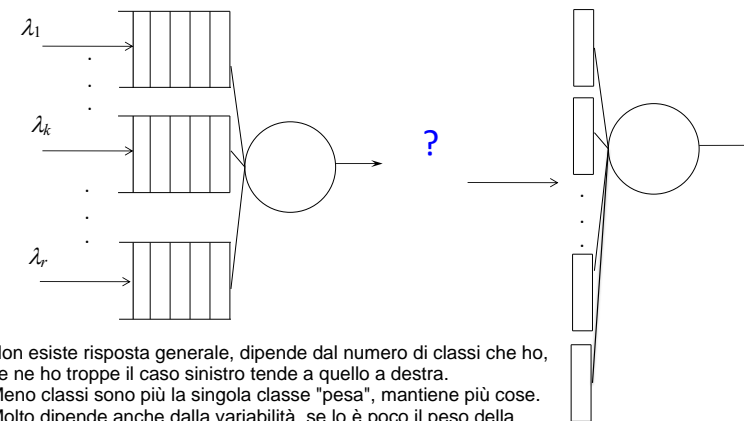
prof. Vittoria de Nitto Personè

23

23

## SB\_P vs SJF

Analytical models  
priority scheduling



Non esiste risposta generale, dipende dal numero di classi che ho, se ne ho troppe il caso sinistro tende a quello a destra. Meno classi sono più la singola classe "pesa", mantiene più cose. Molto dipende anche dalla variabilità, se lo è poco il peso della singola classe è poco, e conviene prendere i vantaggi della prelazione. Perché non mettere la prelazione nel caso SJF? E infatti...

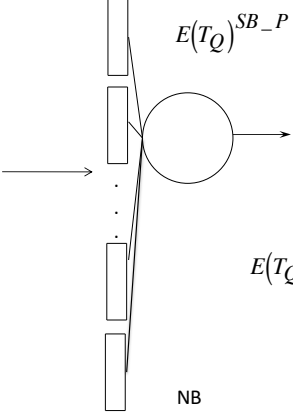
prof. Vittoria de Nitto Personè

24

24

Analytical models  
priority scheduling

### Shortest Remaining Processing Time



$$E(T_Q)^{SB-P} = \frac{\lambda}{2} \sum_{k=1}^r \frac{[F(x_k) - F(x_{k-1})] \left[ \int_0^{x_k} t^2 dF(t) + (1 - F(x_k)) x_k^2 \right]}{\left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}$$

$r \rightarrow \infty$

$$E(T_Q)^{SRJF} = \frac{\lambda}{2} \int_0^\infty \frac{\left[ \int_0^x t^2 dF(t) + (1 - F(x)) x^2 \right]}{\left(1 - \lambda \int_0^x t f(t) dt\right)^2} dF(x)$$

NB  
SRJF=SRPT

web servers under overload

prelazione valutata su tempo di servizio rimanente, sia per sostituzione che per posizionamento.

25

25

Analytical models  
priority scheduling

### Shortest Remaining Processing Time

$$E(T_Q(x)) = \frac{\frac{\lambda}{2} \int_0^x t^2 f(t) dt + \frac{\lambda}{2} x^2 (1 - F(x))}{(1 - \rho_x)^2}$$

non interrompibili      interrompibili (sarebbe  $F > x$ )  
ma che riprendono servizio subito

In questa prima formula, se volessi sapere tempo attesa di tutti i job che hanno una certa size "x", non ho il problema dell'integrale doppio, è un caso "più semplice".

$$E(T_s(x)) = E(T_Q(x)) + \underbrace{\int_0^x \frac{dt}{1 - \rho_t}}_{\text{tempo di servizio virtuale, varia nel tempo. Il job diventa più piccolo quando tempo avanza, e sale di priorità. (anche rho cambia, è in funzione di t, perchè sono sempre meno i job che possono interromperlo).}$$

$$\rho_x = \lambda \int_0^x t f(t) dt$$

Queste formule avanzate, se ci fossero, verranno fornite all'esame.  
Quelle base no.

26

26