

SITA ~~A~~ (Grover 98)

Dispatcher

S1

----- α_0 (cutoff)

S2

if $t \leq \alpha_0 \Rightarrow$ assign to S1
 \Rightarrow otherwise assign
to S2

Objective: Optimize Slow Down (SD)

Given a task 't',

$$SD(t) = \frac{\text{Waiting Time}(t)}{\text{Task Size}(t)}$$

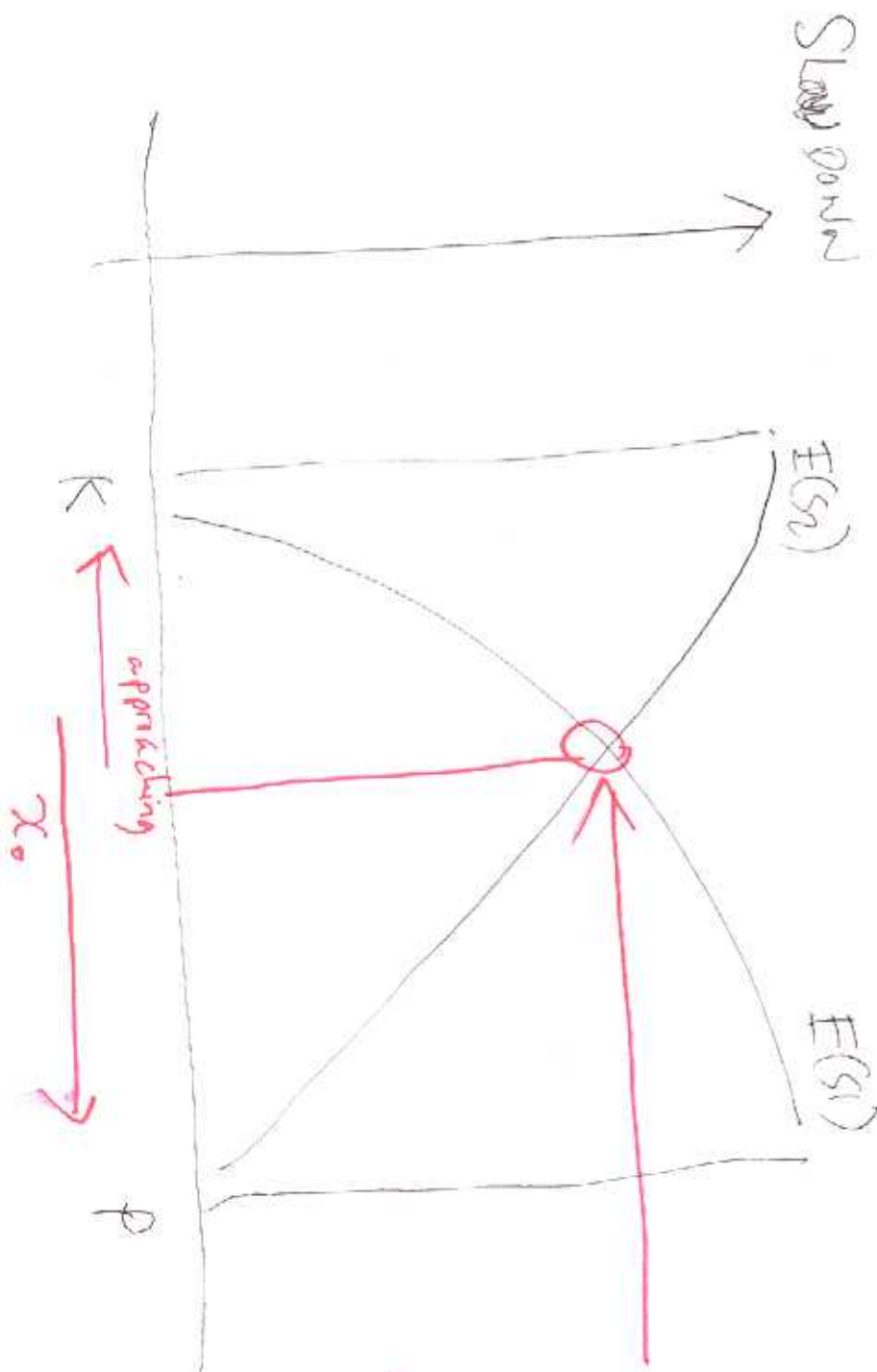
48-1

$$E(S) = P_1 E(S_1) + P_2 E(S_2)$$

where P_1 : fraction of tasks on S_1

P_2 : ——— on S_2

$E(S_i)$: expected slowdown at S_i



BEST VALUE

FOR x_0

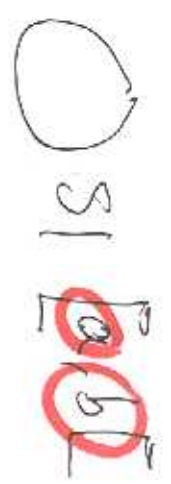
because you have the same slowdown in S_1 and S_2

task size

18-2

SITA-E (Harshol 1999)

Size range
↓



Objective: to compute a, b, c, d, \dots
so the waiting time
is optimized.



⋮

(L8-3)

~~Objective~~

$$E(W_{s_1}) \approx E(W_{s_2}) \approx \dots \approx E(W_{s_{10}})$$

$$\int_a^b x f(x) dx \approx \int_b^c x f(x) dx \approx \dots$$

Level

Basics

Metrics

Load index

Task assignment

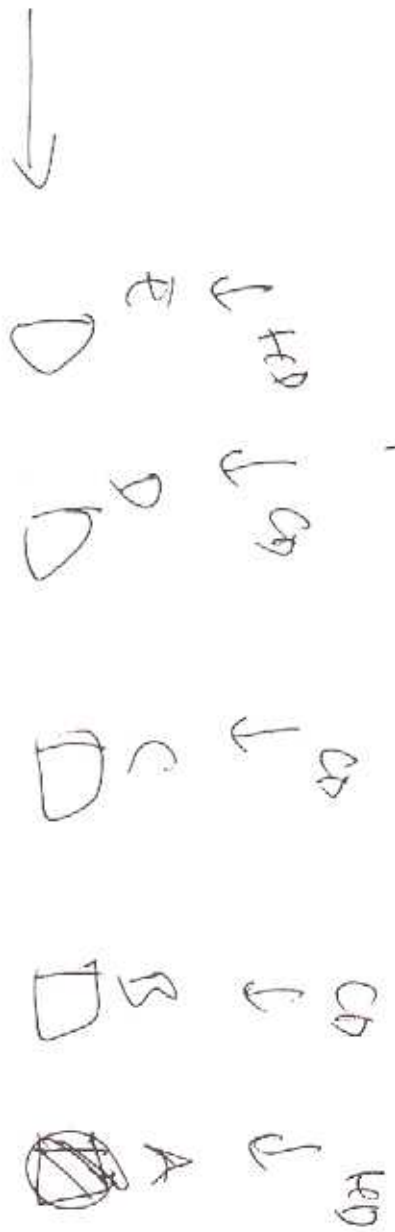
Task size is known

(SITA-V, SITA-E)

Task size is unknown

(TAGS, TARTF)

RRR



Request

(S1) A

(S2) B E

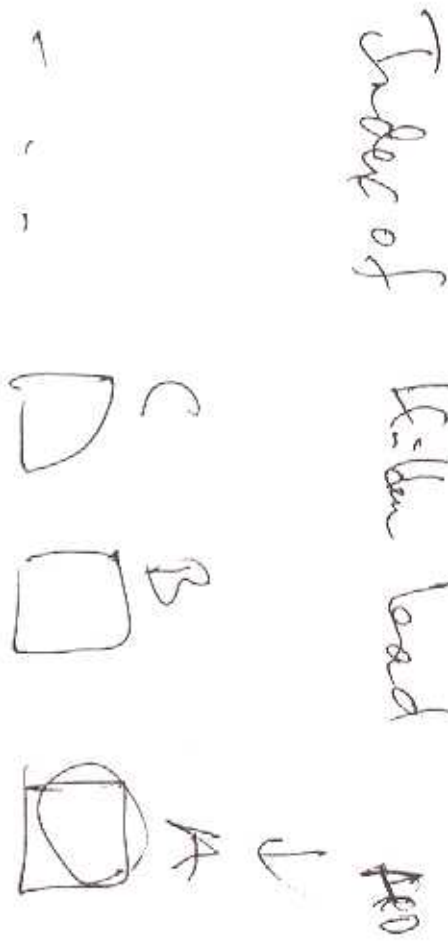
(S3) C

(S4) D

(S5)

(L8-5)

Index of Wilson load



load index ↓

(S1) A, HLD_A

(S2) B, HLD_B

(S3) C, ...

(48-6)