

# NOTE PMCSN

$$E[T_a] = \frac{\frac{1}{2} E[S^2]}{1-p}$$

GENERALI

SOLO per M/M/1 ho  $E[T_a] = \frac{p E[S]}{1-p}$

al numeratore ho sempre  $E[S_{\text{REM}}]$

Per avere  $E[T_s]$  sommo  $E[T_a] + E[S]$ ,

nel caso esponenziale  $E[T_s] = \frac{1}{\mu - \lambda}$ , questo

lo si trova spesso nei problemi

in cui viene chiesto  $E[T_s]$ .

Per Little  $E[N_q] = \lambda \cdot E[T_a]$ , applicabile

in coda, nel centro o nel servente.

$$(E[N_s] = \lambda E[T_s]) \quad (p = \lambda E[S])$$

CASO PARTICOLARE e' L'HYPEREXP:

$$\frac{p E[S]}{2(1-p)} \left( 1 + g(p) \right) \quad \text{con } g(p) = \frac{1}{2p(1-p)} - 1$$

Attenzione all' Hyperexponential, dove ho un SERVENTE che lavora a  $2\mu$  con probabilità "p" e  $(1-p)\mu$  con probabilità " $1-p$ ".

Concepto spesso chiesto è lo **SLOWDOWN**, generalmente espresso come  $E[Sd(x)] = \frac{E[T_q]}{x} + 1$ , se ho  $\geq 2$  code devo scegliere la coda in cui "x" code! Spesso viene confrontato con il Processor Sharing, che è FAVO ~

$$E[Sd(x)] = \frac{1}{1-p} \quad \forall x$$

$$\text{Per il PS si ha } E[T_s] = \frac{E[S]}{1-p}$$

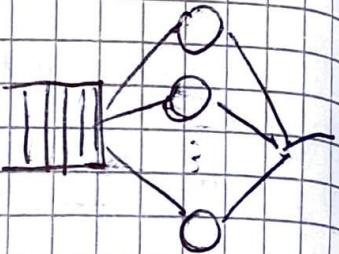
Nei casi M/M/1 (1 servente) le formule usate provengono dalla KP.

Nei casi M/M/m (m serventi) le formule usate provengono dalla ERLANG-C

# ERLANG C

## Potremo al Multiserv com CODA M/M/m

$$\bullet P(0) = \left[ \sum_{i=0}^{m-1} \frac{(mp)^i}{i!} + \frac{(mp)^m}{m!(1-p)} \right]^{-1}$$



usata in  $P_{02} = \frac{(mp)^m}{m!(1-p)} P(0) = \frac{\left(\frac{\lambda}{m}\right)^m}{m!(1-p)} P(0)$

Inoltre  $E[S_i] = \frac{1}{\mu}$  del server "i"

$$E[S] = \frac{E[S_i]}{m}; p_i = \left(\frac{\lambda}{m}\right) = \frac{\lambda}{m\mu} e$$

$p = \frac{\lambda}{(m\mu)}$  sono uguali anche se ci danno due "visioni diverse"

"visioni diverse"

$$E[T_a] = \frac{P_a \cdot E[S]}{1 - P}$$

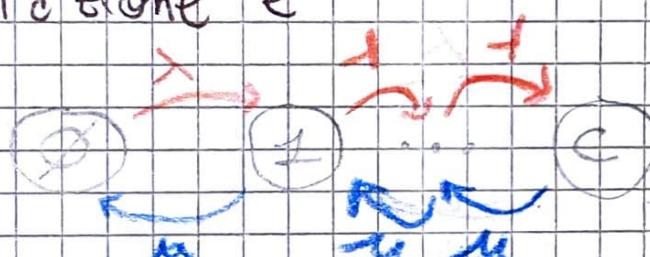
si libera un server qualsiasi, al job non interessa entrare in un server specifico, aspetta che si liberi il primo.

$$E[T_g] = E[T_a] + E[S_i]$$

↳ "scelto un server" deve completare lì dentro il servizio.

Fino ad ora abbiamo sempre analizzato  
con una coda infinita, quindi al 100%  
un job riceverà il servizio, sempre.

Potiamo introdurre il **buffer finito**  
di capacità  $C$  ( $C = 1$  arrivante +  $C-1$  posti  
in coda). La rappresentazione è  
fatta con Markov:

M/M/1/C  $\xrightarrow{\lambda}$  

Sia  $\pi_i$  = prob di stare nello stato " $i$ "  
devo scrivere le **equazioni di flusso**:

$$\begin{cases} \pi_0 \lambda = \pi_1 \mu \\ \pi_1 (\lambda + \mu) = \pi_0 \lambda + \pi_2 \mu \\ \vdots \\ \sum \pi_i = 1 \end{cases}$$

cioè che esce = ciò che entra,  
lo scrivo in funzione di  $\pi_0$ ,  
si troverà sempre:

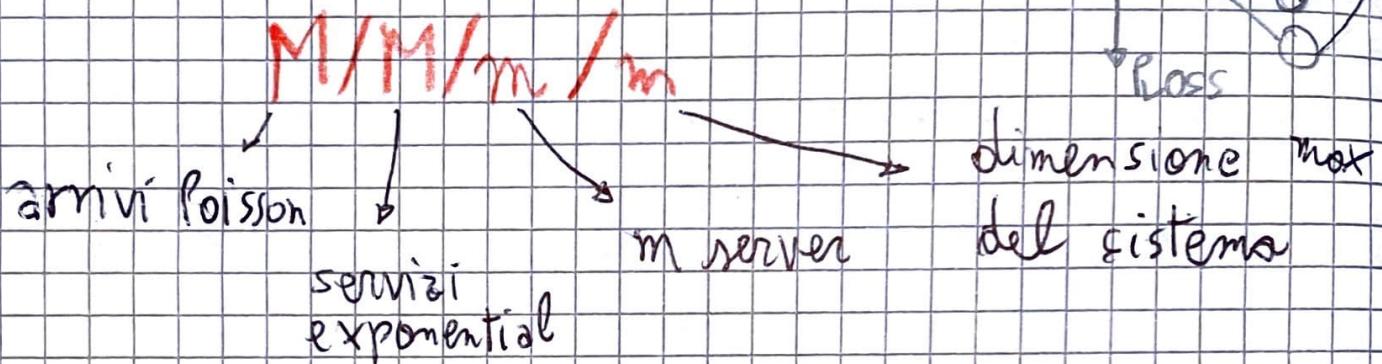
$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

$\pi_0$  lo trovo grazie alla **NORMALIZZAZIONE** =  $\frac{1}{\sum_{i=0}^C \left(\frac{\lambda}{\mu}\right)^i}$

A me interessa  $\pi_C = \text{Ploss} = \left(\frac{\lambda}{\mu}\right)^C \cdot \pi_0 \sim \lambda^C = \lambda(1 - \text{Ploss})$

Con capacità finita il sistema è sempre stazionario.

## MULTISERVER SENZA CODA $\lambda$



se dim. max del sistema =  $m$  server allora

non ho coda. Come combia da  $M/M/1/\epsilon$ ?

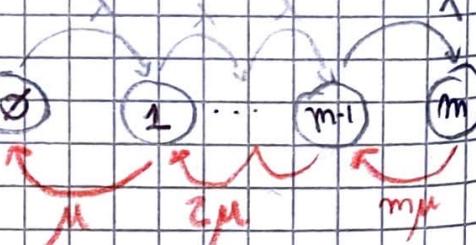
Qui ogni nuovo job va direttamente nel server

Libero (re de). La catena è:

cioè che combia è la parte del servizio, dove c'è  $\mu, 2\mu, 3\mu, \dots m\mu$  in quanto ogni job che esce libera un server.

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \frac{\pi_0}{i!} = \frac{\left(\frac{\lambda}{\mu}\right)^i / i!}{\sum_{j=0}^m \left(\frac{\lambda}{\mu}\right)^j / j!}$$

ERLANG - B



quindi cambiano "solo" i! e j!

~~erlang~~ B: n° chiamate perse, senza coda.

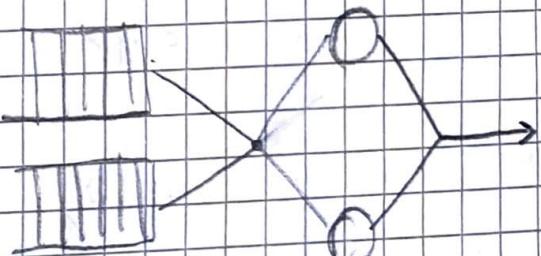
~~erlang~~ E: n° chiamate in attesa, cioè in coda.



## MULTISERVER MULTICODA

Visto nel caso di 2 code

e 2 server. La coda 1 è



prioritaria con  $P_1$ , poi c'è  $P_2$ .

CON PRELAZIONE

$$E[T_q] = P_1 \cdot \frac{P_2(P_2) \cdot E[S]}{1 - P_1} + P_2 \cdot \frac{P_1(P_1) E[S]}{(1 - P)(1 - P_1)}$$

per avere  $E[T_s]$  nommo  $E[S_{\text{virt\_k}}] = \frac{E[S]}{\sum_{i=1}^k p_i}$

SENZA PRELAZIONE

$$E[T_q] = P_1 \cdot \frac{P_2 \cdot E[S]}{1 - P_1} + P_2 \cdot \frac{P_1 E[S]}{(1 - P)(1 - P_1)}$$

per avere  $E[T_s]$  nommo  $E[S]$

# PRIORITÀ

Tutto ciò che abbiamo visto prevedeva sempre al massimo una coda.

L'idea è quello di dividere le code per avere livelli di priorità diverse.

Dobbiamo SEMPRE chiederci:

- conosco la Taglia dei job?
- un job "importante" può prendere il posto di uno "meno importante" che nel mentre era servito?

La prima domanda è associata alla proprietà **SIZE BASED / ABSTRACT PRIORITY**.

La seconda è **PREEMPTIVE / NOT PREEMPTIVE**

Vediamo "caso per coda":

## ABSTRACT + NON PREEMPTIVE

Se sono un job più importante, "sono avanti" nello scheduling viene servito. Se un job poco importante viene servito, devo aspettare.

$$E[T_{a_k}] = \frac{\frac{1}{2} E[S^2]}{(1 - \underbrace{\sum_{i=1}^k p_i}_{\text{Job della mia classe in Fila prima di me}}) \left(1 - \underbrace{\sum_{i=1}^{k-1} p_i}_{\text{Job di classi più importanti}}\right)}$$

$E[S_{\text{REM}}]$   
("non preemptive")

Per avere  $E[T_{S_k}]$  valore nommo  $E[S]$  ("abstract")

$E[T_{a_1}] \leq E[T_{a_2}] \leq \dots$  ugualmente per  $E[T_{S_1}] \leq E[T_{S_2}] \dots$

Per  $E[T_a]$  o  $E[T_S]$  GLOBALI devo pesare i tempi per un peso "p<sub>k</sub>" che spesso è da trovare per avere un certo QoS.

- Se  $\text{EXP } \frac{1}{2} E[S^2] = P E[S]$ .

- generalmente, per ABSTRACT,  $P_k = \tilde{P} \cdot P \sim$  peso classe k

# Miglioro i tempi?

- LOCALMENTE, miglioro clam prioritate, peggioro le meno importanti.
- GENERALMENTE, NON miglioro rispetto allo KP, poiché sono 2 scheduling ASINTOTICI (UGUALI).

## ABSTRACT + PREEMPTIVE

Cio' che cambia e' che un job importante puo' scavalcare uno meno importante anche se era stato venendo servito. Quindi cio' che mi aspetto che combini solo il numeratore, poiche' non aspetto piu' tutti, ma solo job piu' importanti di me!

$$E[T_{0K}] = \frac{\frac{1}{2} E[S^2] \left( \sum_{i=1}^K \lambda_i \right)}{\left( 1 - \sum_{i=1}^K P_i \right) \left( 1 - \sum_{i=2}^{K-1} P_i \right)}$$

ovvero mi aspetto  $\lambda$  in funzione di ' $K$ ', prima mi aspettavo sempre TUTTI.

Va da sé che ogni classe viene migliorata,

prima avevo  $\frac{1}{2} E[S^2] \cdot \lambda$ , ora  $\frac{1}{2} E[S^2] \sum \lambda_k$  che solo nell'ultima classe coincidono!

( $\sum \lambda_k = \lambda$  se sommo tutte le classi!)

Per  $E[TS_K]$ ? ogni job deve considerare anche il tempo "rubato" da job più importanti.

Allora  $E[TS_K] = E[T_{ak}] + \frac{E[S]}{1 - \sum_{i=1}^{K-1} p_i}$

SERVIZIO VIRTUALE

SULLE PRESTAZIONI?

LOCALMENTE:  $E[T_{ak}] \leq E[T_{ak}]$

GENERALMENTE  $E[T_a]_{\text{P priority}} \leq E[T_a]_{\text{NP priority}} = E[T_a]$

NON POSSO CONFRONTARE  $E[TS_K]$  vs  $E[TS_K]$

(perché  $E[S] \leq E[S_{virt}]$ , ma  $E[T_{ak}]^P \leq E[T_{ak}]^{\text{NP}}$ )

• SE EXP perdo ogni vantaggio, e' come KP, colpa della Memoylen.

• Cio' che ho scritto vale per **ARRIVI RANDOM** (**DISTR POISSON**)

## SIZE BASED + NO PRELAZIONE

Mentre nell'abstract NON C'ERA un criterio specifico per ordinare i job, qui li ordiniamo in base alla loro SIZE.

Essendo SENZA PRELAZIONE, un job di qualsiasi classe aspetta la fine del servizio dei job precedenti.

Nella size based si usano gli INTEGRALI.

La divisione in classi avviene in base alle dimensioni dei job, metto una due classi, con Tempi servizio  $[0, E[S]]$  e  $(E[S], +\infty)$ .

Le formule usate sono:

$$E[S_k] = \frac{1}{F(X_k) - F(X_{k-1})} \int_{X_{k-1}}^{X_k} t \cdot f(t) dt$$

$P_k$  — cedo in questo "risparmio".

Poi trovo  $\lambda_k = \lambda \cdot P_k$

MIGLIORAMENTE  $P_k = \lambda_k E[S_k]$ .

Non posso FARE  $P_k = P \cdot p_k$ , questo è OK solo nell'ABSTRACT !!

Le formule sono come prima !

$$E[T_{0:k}] = \frac{\frac{1}{2} E[S^2] }{ \left( 1 - \sum_{i=1}^{k-1} p_i \right) \left( 1 - \sum_{i=1}^k p_i \right)} \text{ se exp}$$

$$\lambda \int_0^{x_{k-1}} t \cdot f(t) dt \quad \lambda \int_0^{x_k} t \cdot f(t) dt$$

**CAMBIA SOLO  
COME TROVO  $P_k$**

GLOBALMENTE, moltiplica per i tempi in coda per la probabilità  $p_k$ . Per  $E[T_{S_k}]$  devo sommare a  $E[T_{0:k}]$  il rispettivo  $E[S_k]$  !

LOCALMENTE  $E[T_{0:k}]^{SB-NP} \leq E[T_{0:k}]^{abstract-NP}$

(combi come calcolo i  $p_k$ ), mentre confrontando  $E[T_{S_k}]^{SB-NP}$  vs  $abstract-NP$  NON POSSO DIRE

NULLA (posso avere  $E[S_k]^{SB-NP} \Rightarrow E[S_k]^{ab-NP}$ , magari nelle ultime cloni, perché in 'ab' non ordino)

GLOBALMENTE  $E[T_s]^{SB-NP} \leq E[T_s]^{ab}$  perchè

$$E[T_a]^{SB-NP} \leq E[T_a]^{ab-NP} \quad (\text{detto min})$$

$$E[S]^{SB-NP} = E[S]^{ab-NP}, \text{ i job quello sono!}$$



### SIZE BASED + PREFERENZE

Caso più complesso, dove non basta che il Job che vuole entrare abbia priorità superiore a quello in servizio; bensì il suo tempo di servizio deve essere MINORE del tempo di servizio RIMANENTE del job che vuole sostituirlo

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

$$E[S_{vir-K}] = \frac{E[S_k]}{1 - \sum_{i=1}^{k-1} p_i}$$

passando alle globali devo sempre pensare per  
 $P_k = F(X_k) - F(X_{k-1})$

Spero lo mi confronta con S. Job. First No Preemp.,  
ma sono incomparabili, allora le unisco in  
Shortest Remaining Processing Time.

Completo, infatti al massimo lo mi calcolo  
per una specifica size.

Normalmente queste formule vengono FORNITE.

## TEORIA delle V.A.

la sequenza di arrivi  $A_i, A_{i+1}, A_{i+2}, \dots$  è visibile come  $A_i = A_{i-1} + R_i$ , dove  $R_i$  è il tempo passante (di interarrivo) tra  $A_{i-1}$  ed  $A_i$ . Se questi  $R_i$  sono INDIP. e oI IDENT. DISTRIBUITI con  $E[R_i] = \frac{1}{\lambda}$  ALLORA

La sequenza di arrivi  $A_1, A_2, \dots$  è un processo di arrivi STAZIONARIO con rate  $\lambda$ .

Se  $\lambda$  varia? NON STAZIONARIO.

Gli interarrivi sono sequenza i.i.d di  $\text{Exp}(\frac{1}{\lambda})$ , da cui la sequenza di arrivi di Poisson( $\lambda$ ).  
Il singolo arrivo  $A_i$  è Erlang( $i, \frac{1}{\lambda}$ )

# RETI CODE e MARKOV (Routing)

Abbiamo ora che fare con come M/M/1 legate tra loro, quindi il flusso può andare da un centro all'altro, disperdersi etc..

Abbiamo CASO CHIUSO e CASO APERTO

CASO CHIUSO (non c'è flusso da/verso l'esterno).

•  $\forall$  centro  $i$  ricevo eq. traffico :  $= \text{FLOW IN}$

Sono linearmente indipendenti (throughput relativi), ne fanno uno e trovano gli altri.

• Calcolo le visite  $V_{i,j} = \frac{Y_i}{Y_j} \rightarrow$  "fisso il riferimento"

Dati  $N$  job,  $M$  centri, sistema chiuso, si procede con Mean Value Analysis, algoritmo ricorsivo, si parte da  $N=0$  job fino ad  $N$ .

Per definizione,  $\forall$  centro  $i$ , con  $N=0$  job si ha  $E[n_i | N=0] = 0$

MVA per reti chiuse! (Product Form)

Poi si pone a  $N=1$  solo, e si calcola

$$E[t_i(N)] = E[S_i] \left( 1 + E[n_i(N-1)] \right)$$

"Tempo medio risposta entro"

(tempo di  
un job nel  
centro "i" con  
 $N$  job nel sist.)

Colliodiamo  $\forall$  centro le entrate  $\lambda_i (N=1)$

$$\lambda_i(N) = \frac{N}{\sum v_{j,i} E[t_j(N)]}$$

ed infine throughput  
entro "i"

$$E[n_i(N)] = \lambda_i(N) \cdot E[t_i(N)] \quad \text{da Little}$$

Devo arrivare al  $\text{n}^{\circ}$  di job  $N$  clienti.

Il tempo di risposta di un centro "i" è dato da:

$$E(t_{i,i}(N)) = \sum_{j \neq i} v_{j,i} E[t_j(N)] \quad \text{cioè visite e tempi negli ALTRI centri.}$$

Il tempo di ciclo rispetto "i" è come  
prima, ma considera anche il centro "i".

$$\text{Vale } R_i(N) = V_i E(t_{i,i}(N)) = \underbrace{E[S_i]}_{D_i} \cdot N_i \cdot (1 + E[n_i(N-1)])$$

## CASO APERTO

• anche qui, per ogni centro, scrivo  $\lambda_i = \text{FLOW IN}$ , procedendo SEMPRE per sostituzione, mai per adegnozione di una variabile. Normalmente sono esercizi "reti di Jackson".

• Per centro calcolo  $E[T_{S,i}] = \frac{1}{\mu_i - \lambda_i}$  (se  $M/M/1$ ) e calcolo le visite  $V_{i,j} = \frac{\lambda_i}{\lambda_j}$  (normalmente  $j$  è il flusso esterno)

Infine Per centro calcolo  $E[t_{r,i}] = \sum V_{j,i} \cdot E(t_j)$

Se servire,  $E[m_i] = \lambda_i \cdot E(t_i) \rightarrow E[T_{S,i}]$   
è anche il throughput (STAZ)

Nel sistema aperto i tempi di risposta sono del tipo  $V_i \cdot E(t_i)$ , indipendenti da  $N$ .

Nel chiuso (vedi formula a minuti) uno  $P_i \cdot (1 + E(m_i(N-1)))$ , dipendente da  $N$ .

In questi esercizi con MIMIZ e arrivo al Poisson Vale Burke: ciò che esce da un centro va in quello successivo a seconda di come è definito (non ho "perdite" inutile)

Burke vale se NON C'È FEEDBACK, perché solo grazie all'indipendenza le probabilità dello stato congiunto è calcolabile come prodotto delle probabilità.

Con feedback + sist aperto risolve come nel caso aperto.

# ANALISI OPERAZIONALE

Si avranno dati misurati come parametri

È importante capire i dati forniti.

$T$  = tempo osservazione

$B$  = tempo sist. occupato

$A$  = arrivi in  $T$

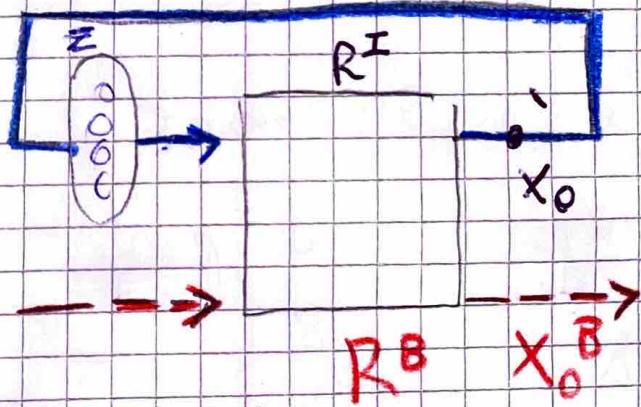
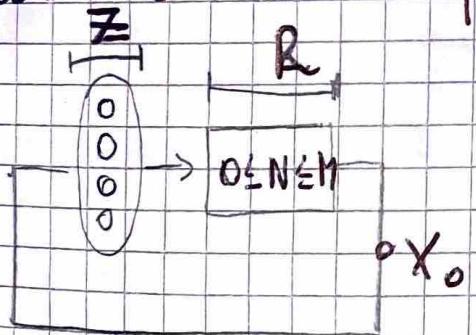
$C$  = completamenti

$$\lambda = \frac{A}{T}, \quad \frac{C}{T} = \text{freq. usata } X, \quad U = \frac{B}{T}, \quad S = \frac{B}{C}$$

( $\approx$  throughput)

$$U = \frac{C}{T} \cdot \frac{B}{C} = X \cdot S \quad (\text{legge utilizzazione})$$

Troviamo due tipi di sistemi:



Vediamo le leggi governanti tali sistemi

- Legge flusso forzato  $X_i = V_i X_0$

$\underbrace{V_i}_{\substack{\text{SINGOLO} \\ \text{CENTRO}}}$        $\underbrace{X_0}_{\substack{\text{TOT} \\ \text{SISTEMA}}}$   
 $\underbrace{X_i}_{\text{VISITE}}$

- Per sistemi INTERATTIVI,

il response time è  $R = \frac{M}{X_0} - Z$  dove  
INTERATTIVO

M è n° terminali, Z è think time.

- Il numero di thinker è  $\# \text{think} = Z \cdot X_0$

- Legge utilizzazione è  $U_i = X_i \cdot S_i$   
 (vale per i singoli centri, non per  $X_0$ )

- Se il sistema è MISTO dev'essere considerate  
 separate le parti batch ed interattive, cioè:

$$X_{\text{DISK}} = X_{\text{DISK}}^B + X_{\text{DISK}}^I$$

, oltre mare i dati  
 in modo corretto!

N.B.:  $X_0 = \frac{X_i \cdot \text{UTILIZZ}}{V_i}$   $\left( \frac{U_i}{S_i} \right) \cdot \frac{1}{V_i} = \frac{U_i}{D_i}$

FLUSSO      UTILIZZ      DOMANDA  
FORZAT.       $V_i$        $D_i$

$$N_i = X_i \cdot R_i \text{ (Little)}$$

$$X^I_{\text{disk}} = X_{\text{disk}} - X^B_{\text{disk}} \text{ e' minimo se}$$

$X_{\text{disk}}$  realizza  $P \rightarrow 1$  cioè  $X_{\text{disk}} = \frac{1}{S_{\text{disk}}}$   
(completa più job ne il disco lo uso al minimo,  
cioè ne opera in ETS<sub>disk</sub>)

- Se non si parla esplicitamente di sistemi interattivi, il tempo di risposta è :  $R = \sum_i v_i R_i$
- Se abbiamo  $T, C_0$  allora  $X_0 = \frac{C_0}{T}$ , devo vedere i dati che ho!

# BOUND & Bottleneck

$$D_i = V_i \cdot S_i, \quad V_i = \frac{X_i}{X_0}, \quad D_i = \frac{X_i}{X_0} S_i = \frac{(C_i/T)}{\frac{1}{C_0} C_i} B_i = \frac{B_i}{T/C_0}$$

$$U_i = X_i S_i = (X_0 V_i) S_i = X_0 D_i$$

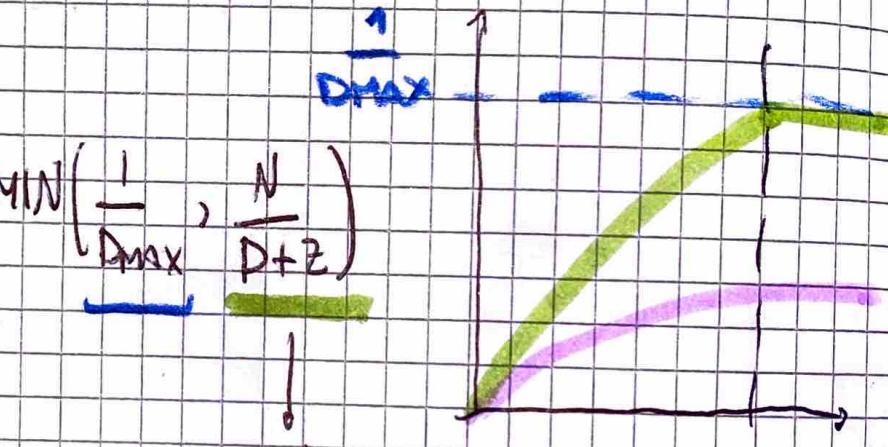
$D_{MAX}$  = max tra le "D" domande.

$$D_{TOT} = D = \sum_i D_i \quad (i \neq \text{centro})$$

## THROUGHPUT

$$\frac{N}{ND+Z} \leq X(N) \leq \min\left(\frac{1}{D_{MAX}}, \frac{N}{D+Z}\right)$$

Sono l'ultimo  
in coda sempre



non ha mai nemmeno  
avanti.

$N^2$

dipende dai movimenti nel centro,  
e' vicini ne tutti reggono stessa persona, VERE altimenti

La soluzione la ho per  $N^2 = D+Z$ , cioè

$$\text{impongo } \frac{1}{D_{MAX}} = \frac{N}{D+Z}$$

Per il tempo di Risposta R?

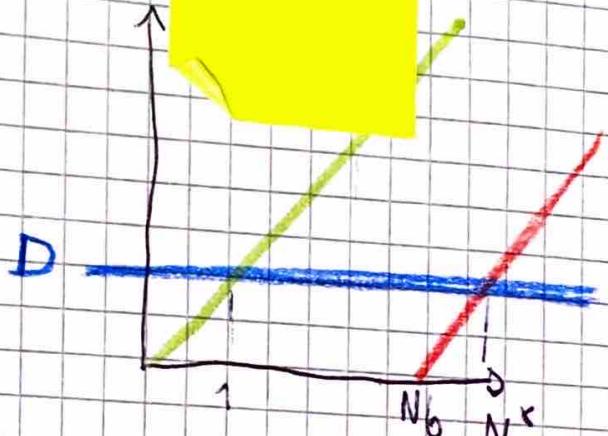
caso peggiore: aspetto tutti: ND

caso migliori:

aspetto solo me: D

•  $R = 0 \rightsquigarrow X_0 = \frac{1}{D_{MAX}} \rightsquigarrow \underline{ND_{MAX} - 2}$   
(non ci sono job)

$$\underline{ND} \geq R \geq \max \{ D, \underline{ND_{MAX} - 2} \}$$



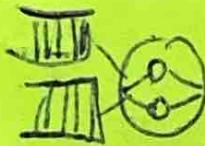
# Random Notes

servizi exp,  
distribuzione ignota  
(exp + k-exp o...)  
 $E[T_a] = \frac{P[E[S]]}{1-P} \cdot \left( \frac{1+C^2}{2} \right)$

In Markov,  
 $E[N_S] = \sum_{i=0}^c \pi_i \cdot i =$

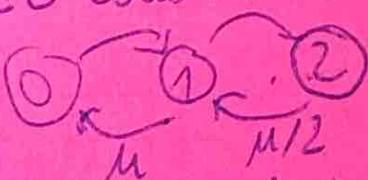
$$\pi_0 \cdot 0 + \pi_L \cdot 1 + \dots$$

MULTISER  
MULTICODA  
PREEMPTIVE,  
per  $E[T_{S1}]$  sommo  $E[S]$   
(mai interrotto), per  
 $E[T_{S2}]$  uno  $E[S_{virt-2}]$



SIZE BASED NP  
 con 1 clone, allora  
 non posso avere  $E[S_K]$ ,  
 $\lambda_K$  etc... ma  $\rho = \lambda E[S]$   
 è come uno KP

$$F_{\text{unit}} = \frac{x-a}{b-a}$$

Se Coda FINITA + PS  
  
 perché divido!

"Z" si misura in "op"

se devo con  
slowdown PS e  
 $PS = \frac{1}{1-P}$ , FIFO :  
 val( $E[T_a]/x$ ) +

$$\text{medio Req} = 0 \cdot \pi_0 + 1 \cdot \pi_1 + \dots$$

$$C^2(\text{Req}) = E[\text{Req}^2] - \text{medio}^2$$

(1 ·  $\pi_1 + 2^2 \cdot \pi_2$ )

In un sistema interattivo

$$R = \frac{M}{X_0} - Z$$

$R \rightarrow$  se  $X_0 \nearrow$ ,  $X_0 \text{ MAX}$

se tutto ciò che arriva serve  
cioè  $P=2 \rightarrow X_0 = \frac{1}{\text{tempo servizio}}$

Se coda chiusa, e  
ha " $D_i \rightarrow MVA$  ridotto",  
per tempo di risposta

$$V_i E(t_i(N)) < E(S_i) V_i / (1 + E(N-1))$$

$$R_i(N) = D_i (1 + E_i(N-1))$$

$$X_0(N) = \frac{N}{\sum R_i}$$

e  $\sum R_i = \frac{\text{tempo risp. totale}}{\text{tempo servizio}}$