Reti di code per l'analisi dei sistemi di calcolo

Nello Scarabottolo – dicembre 2003 – versione 1.2

Obiettivo dell'analisi di un sistema di calcolo è la previsione delle prestazioni che è possibile ottenere durante diverse fasi del suo ciclo di vita. Ci sono diversi momenti nei quali è importante effettuare un'analisi di questo tipo. Innanzi tutto durante la fase di progetto e di realizzazione di un sistema informativo, quando devono essere decisi i meccanismi di accesso al data-base aziendale, quelli di sincronizzazione dei processi, e così via. Oppure quando ci si trova a dover dimensionare o acquisire un nuovo sistema di calcolo, o ancora quando ci si trova di fronte ad una evoluzione del sistema e/o dei carichi. In tutte queste occasioni è necessario disporre di uno strumento che consenta di valutare le prestazioni offerte dalle diverse alternative che si hanno a disposizione, in modo da selezionare la migliore.

Un primo metodo che può essere utilizzato per condurre quest'analisi è quello che si basa sull'intuizione e sull'estrapolazione delle tendenze. Si tratta di un metodo fortemente basato sull'esperienza, rapido e flessibile, ma poco affidabile e poco accurato. All'altro opposto si pone la metodologia di analisi basata sulla valutazione sperimentale delle alternative, che dà risultati più accurati e più precisi, ma che richiede grossi investimenti sia in termini di tempo che di denaro.

Una soluzione intermedia tra l'intuizione e la valutazione sperimentale è quella basata sulla **modellizzazione** del sistema di calcolo. Si tratta di un compromesso tra i due metodi precedenti che risulta essere da un lato più flessibile e meno dispendioso della valutazione sperimentale (richiede meno lavoro e minori spese), dall'altro più attendibile ed accurato del metodo intuitivo (consente un'analisi metodica).

Nel seguito si discutono gli aspetti fondamentali di una delle più diffuse metodologie di analisi dei sistemi di calcolo, basata appunto sul metodo della modellizzazione del sistema stesso. In particolare, per modello si intende l'astrazione di un sistema che ne mette in evidenza solo quegli aspetti che sono essenziali per rappresentarne il comportamento. Modellizzare significa quindi organizzare uno schema per raccogliere, ordinare, capire e valutare le informazioni relative a un sistema (di calcolo).

Il ciclo di modellizzazione si compone di tre fasi principali. Dapprima viene validato il modello, ovvero viene analizzato il sistema esistente, in modo da poter definire un modello base e verificarne l'affidabilità. Il secondo passo consiste nella proiezione del modello: le modifiche previste vengono apportate al modello e vengono studiati gli effetti. Infine, dopo la realizzazione delle modifiche, viene effettuata una verifica dei risultati, confrontando i risultati della proiezione con le misure reali.

Argomento specifico di questo testo sono i modelli a reti di code dei sistemi di calcolo. I modelli a rete di code rappresentano un sottoinsieme della teoria delle code che, per lo più, è orientata alla rappresentazione di sistemi complessi tramite singoli centri di servizio, caratterizzati a loro volta in modo complesso. Per l'analisi di questi centri vengono poi utilizzate tecniche matematiche sofisticate, che consentono di ottenere indici delle prestazioni del sistema sufficientemente precisi. I modelli a reti di code, invece di utilizzare un singolo centro molto complesso, rappresentano il sistema come una rete di centri di servizio semplici. Questo approccio è vantaggioso quando viene applicato a sistemi di calcolo, infatti in questo caso può essere utilizzato un opportuno sottoinsieme delle reti di code: i modelli così costruiti risultano abbastanza semplici e intuitivi, e gli algoritmi di valutazione delle prestazioni possono essere implementati in modo che siano al tempo stesso precisi ed efficienti.

I modelli a reti di code

Per modello a reti di code di un sistema di calcolo si intende una sua rappresentazione sotto forma di una rete di code che possa essere valutata in maniera analitica. Una rete di code è un insieme di centri di servizio (che rappresentano le risorse del sistema) e di clienti (che rappresentano gli utenti o le transazioni). In genere, una volta definita una rete di code si usano appositi programmi per risolvere in maniera efficiente l'insieme di equazioni "indotto" dalla rete e dai suoi parametri.

Un esempio di centro di servizio singolo è riportato in figura 1.1. I clienti affluiscono al centro, eventualmente attendono in coda, ricevono quindi un servizio ed escono dal centro.

Figura 1.1 – Centro di servizio singolo

Anche intuitivamente si può comprendere come per valutare correttamente il comportamento di un centro di questo tipo è necessario specificare due parametri:

1. l'**intensità del carico** (frequenza di arrivo dei clienti) e
2. la **domanda di servizio** (richiesta media di servizio da parte dei clienti).

Le prestazioni del centro saranno quindi misurate valutando alcuni tra i seguenti valori:

a) **utilizzo** (percentuale di tempo in cui server è occupato);

b) **tempo di residenza** (tempo medio passato da un cliente al centro di servizio, sia in coda che in fase di ottenimento del servizio);

c) **lunghezza della coda** (numero medio dei clienti presenti al centro si servizio, sia in coda che ricevendo servizio);

d) **traffico in uscita** (frequenza di attraversamento del centro di servizio).

L'andamento tipico di questi valori in funzione del carico a cui è sottoposto il centro è riportato nella figura 1.2.

Per definire il modello di un sistema complesso è necessario ricorrere a più centri di servizio collegati tra loro a formare una vera e propria rete, come illustrato per esempio nella figura 1.3 (in particolare questa rete può essere utilizzata per rappresentare un sistema con più terminali e con carichi di tipo interattivo).

In pratica a ogni risorsa del sistema di calcolo viene associato un centro di servizio per cui vengono specificati i parametri indicati prima, ovvero l'intensità del carico e la domanda di servizio. Nel modello i clienti rappresentano le transazioni che vengono effettuate nel sistema di calcolo, quindi l'intensità di carico corrisponde alla frequenza con cui gli utenti mandano transazioni al sistema, mentre la domanda di servizio a un centro è uguale alla richiesta complessiva di servizio da parte di una transazione nei confronti di una risorsa.

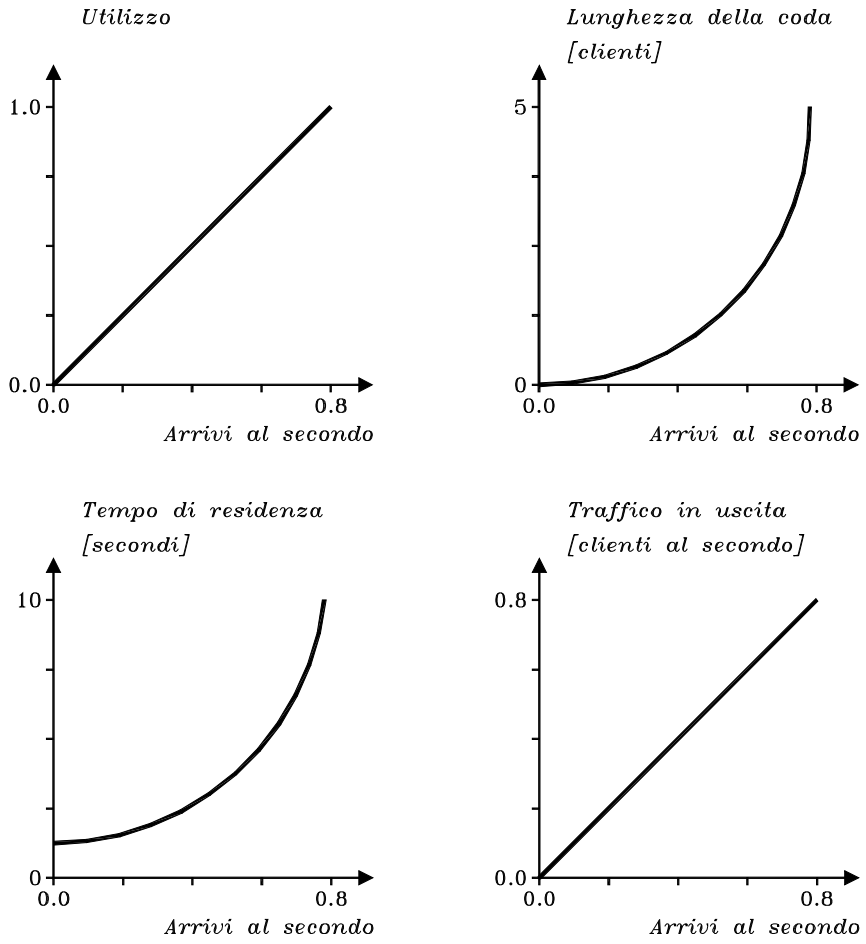

Figura 1.2 – Andamento degli indici di prestazione di un centro di servizio singolo

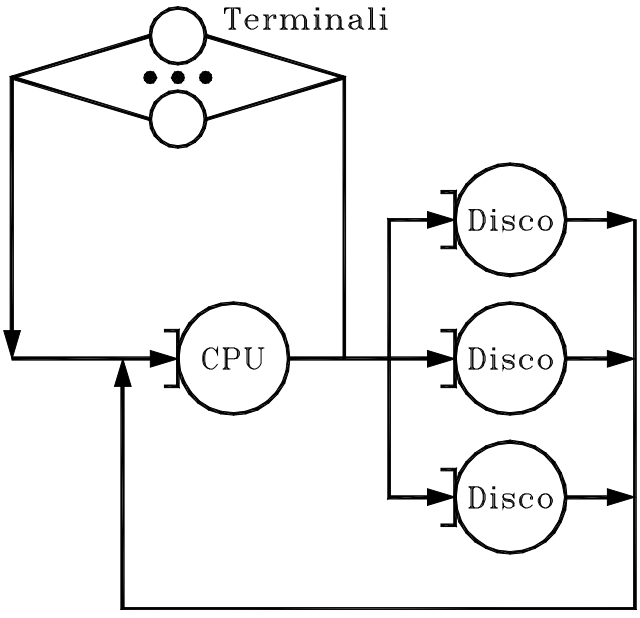

Figura 1.3 – Esempio di rete di code con più centri di servizio

Sviluppo di un modello a reti di code

Durante lo sviluppo di un modello a reti di code è preferibile utilizzare una metodologia top-down in modo da identificare quali sono i componenti principali e come interagiscono tra di loro, e poter poi introdurre tutti e soli i dettagli significativi. L'abilità dell'analista entra in gioco quando vengono adottate delle ipotesi di semplificazione (ovvero quando si introducono delle approssimazioni). Per ogni ipotesi fatta è importante indicare le motivazioni e le argomentazioni che la rendono plausibile.

Leggi fondamentali

Grandezze base

In un qualsiasi sistema di calcolo è possibile definire le seguenti grandezze fondamentali:

T tempo di osservazione del sistema [*sec.*]

A numero di richieste arrivate [*clienti*]

C numero di richieste completate [*clienti*]

In base a queste grandezze è possibile definire:

$\delta = \frac{A}{T}$ frequenza di arrivo [*clienti/sec.*]

$X = \frac{C}{T}$ traffico in uscita [*clienti/sec.*]

Nel caso di sistemi composti da una sola risorsa si può definire anche

B tempo in cui la risorsa è rimasta occupata [*sec.*]

e in base a questo valore si possono definire

$U = \frac{B}{T}$ utilizzo

$S = \frac{B}{C}$ tempo medio di servizio [*sec.*]

Prima della presentazione delle leggi che consentono di mettere in relazione le grandezze fondamentali definite sopra è necessario introdurre un'ipotesi fondamentale per lo sviluppo delle stesse leggi: all'interno dell'intervallo di osservazione T il sistema deve essere in *equilibrio operazionale*, ovvero il flusso deve essere bilanciato e quindi il numero degli arrivi deve essere uguale a quello dei completamenti ($A = C$). Questa ipotesi implica anche che la frequenza di arrivo eguagli il traffico in uscita ($\delta = X$). Utilizzando un intervallo di osservazione T, l'errore che si commette assumendo che sia valida l'ipotesi di equilibrio operazionale può essere calcolato come:

$$\varepsilon \leq \left| \frac{A-C}{C} \right|$$

Legge dell'utilizzo (sistemi composti da 1 risorsa)

Dalle relazioni fondamentali si può ricavare la seguente legge:

$$U = \frac{B}{T} = \frac{C}{T}\frac{B}{C} = X\,S$$

ovvero l'utilizzo è uguale al traffico moltiplicato per il tempo medio di servizio.

Si consideri per esempio un disco che debba servire 40 richieste al secondo, ciascuna delle quali richieda 22.5 *msec.* di servizio. L'utilizzo del disco risulta essere:

$$U = X\,S = 40 \times 22.5 \times 10^{-3} = 0.9 = 90\%$$

Legge di Little (sistemi con più risorse)

Si definisce come **tempo di accumulo** (*W*) del sistema la somma del tempo di residenza dei clienti che entrano nel sistema durante il tempo di osservazione. Si consideri per esempio il grafico di figura 1.4, dove sono riportate le curve degli arrivi e dei relativi completamenti. Il tempo di accumulo è rappresentato dall'area compresa tra le due curve. I rettangoli tratteggiati che formano quest'area indicano appunto il tempo di residenza nel sistema di ogni singolo cliente. Per esempio il cliente 1 rimane 1 secondo nel sistema, mentre il cliente 16 vi rimane 2 secondi. Il tempo di accumulo sarà perciò $W = 1+2+2+2+2+1+2+2+1+1+2+1+2+1+1+2 = 25$.

Si può quindi definire come tempo medio di residenza nel sistema (*R*) il rapporto tra il tempo di accumulo e il numero dei completamenti:

$$R = \frac{W}{C}$$

Un altro modo di rappresentare il tempo di accumulo è quello illustrato in figura 1.5, dove viene riportato il grafico della curva dei clienti presenti nel sistema nei diversi istanti di tempo (ottenuta dalla differenza tra la curva degli arrivi e quella dei completamenti riportate in figura 1.4). Il tempo di accumulo corrisponde quindi all'area sottesa a questa curva e può essere considerato come la somma dei clienti presenti nei diversi istanti di tempo. Per esempio nel primo istante il sistema contiene 3 clienti, nel secondo istante ne contiene 4, nel terzo 5, e così via. Si ottiene $W = 3+4+5+2+0+3+3+4+1+0 = 25$.

Si può quindi definire come numero medio di clienti presenti nel sistema (*N*) il rapporto tra il tempo di accumulo e il tempo di osservazione:

$$N = \frac{W}{T}$$

Quest'ultima relazione può anche essere riscritta nel seguente modo:

$$N = \frac{W}{T} = \frac{C}{T}\frac{W}{C} = X\,R$$

La legge di Little può quindi essere espressa come $N = X\,R$, ovvero il numero medio dei clienti presenti in un sistema è uguale al traffico moltiplicato per il tempo medio di residenza.

Se si considera ancora l'esempio riportato nelle figure 1.4 e 1.5, è possibile ottenere i seguenti risultati:

$$\delta = \frac{A}{T} = X = \frac{C}{T} = \frac{16}{10} = 1.6 \textit{ clienti/sec.}$$

$$U = \frac{B}{T} = \frac{8}{10} = 80\%$$

$$S = \frac{B}{C} = \frac{8}{16} = 0.5 \textit{ sec.}$$

$$N = \frac{W}{T} = \frac{25}{10} = 2.5 \textit{ clienti}$$

$$R = \frac{W}{C} = \frac{25}{16} = 1.56 \textit{ sec.}$$

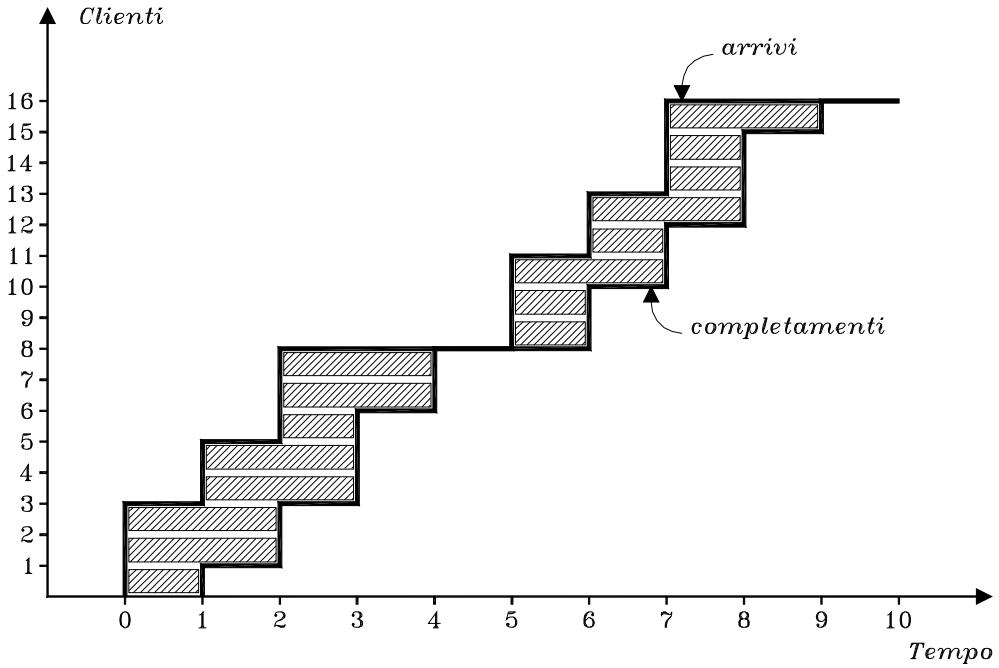

Figura 1.4 – Relazione fra tempo di accumulo e curve degli arrivi e dei completamenti

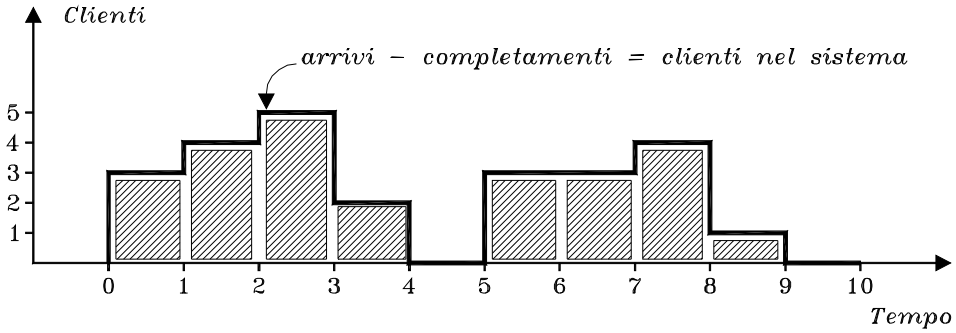

Figura 1.5 – Relazione fra tempo di accumulo e curva dei clienti presenti nel sistema

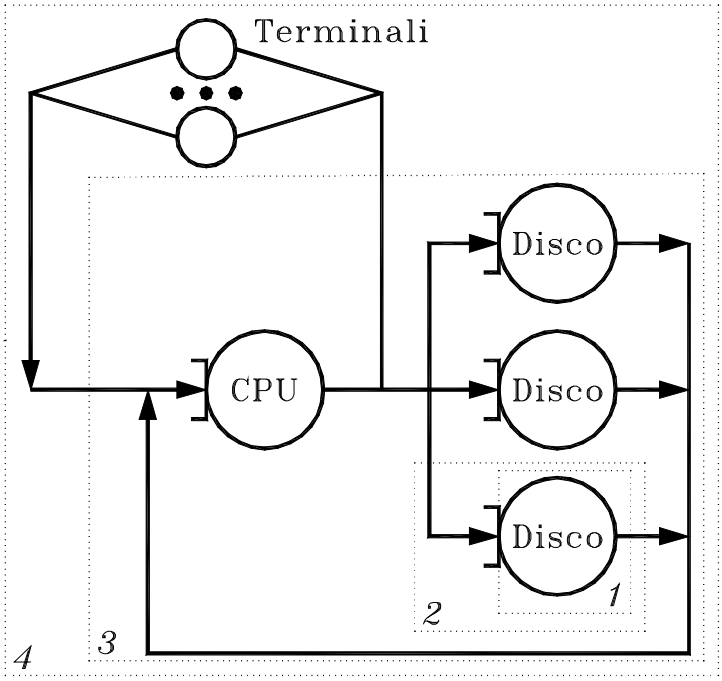

Figura 1.6 – Esempio di sistema per l'applicazione della legge di Little a diversi livelli

Applicazione della legge di Little a diversi livelli di un sistema

Viene ora presentato un esempio di applicazione della legge di Little a diversi livelli di un sistema di calcolo. In particolare viene preso in considerazione il sistema presentato in figura 1.6. I livelli sono identificati dalle linee tratteggiate e dai relativi numeri.

1. **Solo disco (senza coda)**

   In ogni istante di tempo, nel centro può essere presente 1 cliente oppure nessuno e la risorsa è utilizzata solo quando è presente un cliente. Per come sono stati definiti, S è il tempo medio di servizio per richiesta, mentre R è il tempo medio di residenza nel sistema. Siccome il sistema sotto esame è costituito da un centro senza coda, R deve essere uguale a S, quindi le due leggi dell'utilizzo ($U = X\,S$) e di Little ($N = X\,R$) coincidono. In questo modo si ottiene che $N = U$: la popolazione media N del centro di servizio viene a coincidere con l'utilizzo U delle risorse presenti nello stesso centro.

2. **Disco (compresa la relativa coda)**

   In questo caso bisogna considerare anche i clienti presenti in coda e il tempo trascorso in attesa che il centro si liberi. Perciò $N \neq U$ e $R \neq S$, valgono infatti le seguenti relazioni:

   N popolazione = clienti in coda (N–U) + clienti in servizio (U);

   R tempo di residenza = tempo di attesa in coda (R–S) + tempo di servizio (S);

   X velocità alla quale vengono soddisfatte le richieste.

3. **Sistema senza terminali**

   A questo livello le grandezze messe in relazione dalla legge di Little si riferiscono al vero e proprio sistema di calcolo:

   N clienti presenti nel sottoinsieme centrale, cioè processi che non sono sospesi in attesa della risposta dell'utente;

X frequenza alla quale fluiscono le transazioni tra sottosistema centrale e terminali;

R tempo in cui le transazioni lanciate dal terminale restano nel sottosistema centrale (si noti come questo sia il concetto tradizionale di tempo di risposta di un sistema di calcolo, perciò in questo caso il tempo di residenza coincide con il tempo di risposta).

4. Intero sistema (compresi i terminali)

In questo caso bisogna considerare anche i processi che sono momentaneamente sospesi in attesa che l'utente dia una risposta (riflessione).

N numero totale degli utenti interattivi;

X frequenza alla quale fluiscono le transazioni tra sottosistema centrale e terminali;

R_d tempo di residenza = tempo di risposta (R) + tempo di riflessione (Z).

Si noti che il tempo di risposta coincide con il tempo di residenza del sistema al livello precedente, quando non erano stati considerati i terminali.

Riscrivendo la legge di Little si ottiene $N = X R_d = X (R + Z)$ da cui è possibile ricavare la cosiddetta **legge del tempo di risposta**:

$$R = \frac{N}{X} - Z$$

Legge del flusso forzato

Questa legge consente di stabilire quale sia il rapporto di proporzionalità che lega tra loro i traffici all'interno di un sistema. Si supponga di rilevare non solo i completamenti C di sistema, ma anche i completamenti C_k di ogni singolo centro di servizio.

Le **visite** effettuate da un singolo cliente al k-esimo centro possono essere definite come

$$V_k = \frac{C_k}{C}$$

Se per esempio nel periodo di osservazione vengono rilevati $C = 50$ completamenti del sistema e $C_{disco} = 1000$ completamenti da parte di un disco, si può dedurre che ogni cliente servito dal sistema ha visitato in media il disco

$$V_{disco} = \frac{C_{disco}}{C} = \frac{1000}{50} = 20 \text{ volte.}$$

La definizione precedente può essere riscritta come

$$C_k = V_k C$$

Dividendo quindi per il tempo di osservazione T si ottiene

$$\frac{C_k}{T} = \frac{V_k C}{T}$$

da cui si ricava la legge del flusso forzato

$$X_k = V_k X$$

Se per esempio in un sistema ciascun cliente richiede 10 accessi a un disco ($V_{disco1} = 10$) e lo stesso disco soddisfa in media 20 richieste al secondo ($X_{disco1} = 20$), allora il traffico complessivo del sistema sarà

$$X = \frac{X_{disco1}}{V_{disco1}} = \frac{20}{10} = 2 \textit{ clienti/sec.}$$

Si noti come, grazie alla legge del flusso forzato, è sufficiente osservare il comportamento del disco per ottenere il traffico in uscita dal sistema complessivo.

Si riprenda in esame l'esempio presentato sopra; se inoltre un altro disco soddisfa X_{disco2} richieste al secondo è possibile calcolare il numero delle visite che ogni cliente effettua al secondo disco

$$V_{disco2} = \frac{X_{disco2}}{X} = \frac{30}{2} = 15$$

richieste per cliente.

In tabella 1.1 sono riassunte le definizioni dei parametri utilizzati nei modelli a reti di code e le principali relazioni intercorrenti fra di esse.

CONVENZIONI

T	= tempo di osservazione [*sec.*]
A_k	= arrivi [*clienti*]
C_k	= completamenti [*clienti*]
δ	= frequenza di arrivo [*clienti/sec.*]
X_k	= traffico [*clienti/sec.*]
B_k	= tempo di occupazione [*sec.*]
U_k	= utilizzo
S_k	= tempo di servizio per visita [*sec.*]
N	= popolazione clienti
R_k	= tempo di residenza [*sec.*]
Z	= tempo di riflessione [*sec.*]
V_k	= numero di visite
D_k	= domanda di servizio [*sec./cliente*]

LEGGI FONDAMENTALI

Legge dell'utilizzo:	$U_k = X_k S_k = X D_k$
Legge di Little:	$N = XR$
Legge del tempo di risposta:	$R = N/X - Z$
Legge del flusso forzato:	$X_k = V_k X$

ALTRE RELAZIONI

$\delta = A_k/T$
$X_k = C_k/T$
$U_k = B_k/T$
$S_k = B_k/C_k = (U_k T)/C_k$
$V_k = C_k/C$
$D_k = V_k S_k = B_k/C = (U_k T)/C$

Tabella 1.1 – Parametri e relazioni usati nei modelli a reti di code

Ingressi e uscite di un modello a reti di code

Vengono considerate solo le **reti di code separabili**, sottoinsieme delle reti di code, perché, sono più semplici e richiedono un volume minore di calcoli. Da un lato, le richieste per l'applicabilità di questi modelli non sono sempre soddisfatte; dall'altro, le imprecisioni che risultano dall'applicazione di questi modelli a sistemi reali sono di solito trascurabili rispetto a quelle introdotte in altre fasi del processo di modellizzazione. Un'analisi più dettagliata delle reti di code generiche e delle ipotesi che debbono essere soddisfatte affinché, un sistema possa essere rappresentato da una rete di code separabile può essere trovata in:

> Lazowska, Zahor, Scott Graham, Sevcik, *Quantitative System Performance – Computer System Analysis Using Queueing Network Models*, Prentice-Hall, 1984.

Ingressi del modello separabile con una sola classe di clienti

Per consentire l'analisi di un sistema con una sola classe di clienti è necessario descrivere il cliente medio che usufruisce dei servizi forniti dal sistema, specificare i centri che compongono il sistema e indicare il servizio che i clienti richiedono ai diversi centri.

Descrizione cliente

La descrizione del cliente medio permette di identificare l'intensità del carico che viene applicato al sistema in esame. Si distinguono tre diversi casi:

1. Lavori **transazionali**. Il carico viene espresso in termini di frequenza di arrivo (δ). In questo caso, quando un cliente ha completato il servizio, lascia il sistema. La popolazione (numero di clienti presenti nel sistema) varia molto nel tempo.
2. Lavori **batch**. Il carico viene indicato dalla popolazione (N), che in questo caso è fissa. Ogni cliente che ha completato il servizio lascia il sistema, ma viene subito sostituito da un altro (è come se l'uscita fosse cortocircuitata con l'ingresso).
3. Lavori **interattivi**. Il carico viene specificato tramite la popolazione (N), che in questo caso corrisponde al numero dei terminali attivi, e il tempo di riflessione (Z).

Si definiscono **modelli aperti** quei modelli che si riferiscono a sistemi in cui il numero di richieste di elaborazione (clienti) presenti contemporaneamente nel sistema non è limitato (carico *transazionale*). Si chiamano **modelli chiusi** quelli che invece si riferiscono a sistemi in cui il numero di richieste presenti è costante (carico di tipo *batch*) o comunque limitato superiormente (carico di tipo *interattivo*).

Si noti come un carico dovuto a lavori interattivi assomiglia ad un carico dovuto a lavori batch perché, la popolazione (N) è costante in entrambi i casi mentre il tempo di riflessione (Z) è nullo per i lavori batch. Però nel caso di carico dovuto a lavori interattivi, la popolazione all'interno del sottosistema centrale è variabile, cioè simile al caso transazionale. Quindi nel sottosistema centrale N è:

- variabile senza limite superiore per lavori transazionali;
- variabile fino al numero di terminali per lavori interattivi;
- fisso per lavori batch.

Descrizione dei centri

Prima di tutto è necessario specificare il numero complessivo dei centri (K), poi per ogni centro bisogna indicare se si tratta di un centro **ad accodamento** (il tempo di attraversamento comprende il tempo di servizio e il tempo di attesa in coda) oppure di un centro **di ritardo** (ogni cliente ha un proprio punto di servizio e quindi il tempo di residenza equivale al tempo di servizio, infatti non c'è competizione per il servizio e quindi non c'è attesa in coda).

Richiesta di servizio

Il servizio che un cliente richiede a un centro del sistema può essere espressa in due modi:

1. specificando la richiesta di servizio S_k per ogni visita al centro, e il numero V_k degli accessi (visite) che ogni cliente richiede al centro;
2. indicando la domanda di servizio complessiva D_k, che può essere calcolata come il prodotto della richiesta di servizio e il numero di visite:

   $$D_k = V_k S_k$$

   oppure, considerando le definizioni:

   $$S_k = \frac{B_k}{C_k} \text{ e } V_k = \frac{C_k}{C}$$

   può essere espressa come:

$$D_k = \frac{B_k}{C_k}\frac{C_k}{C} = \frac{B_k}{C}$$

Si definisce anche la domanda di servizio totale richiesto da un cliente a tutte le risorse come:

$$D = \sum_{k=1}^{K} D_k$$

Uscite del modello separabile con una sola classe di clienti

Valori per i singoli centri

U_k utilizzo del centro k (percentuale di tempo in cui il centro è occupato oppure numero medio di clienti riceventi un servizio da quel centro);

R_k tempo di residenza complessivo al centro k, sia in servizio (S_k) che in coda ($R_k - S_k$); questo valore si riferisce a tutte le visite; il tempo di residenza per ogni singola visita è R_k / V_k

X_k traffico in uscita dal centro k;

Q_k clienti presenti al centro k, sia in servizio (U_k) che in coda ($Q_k - U_k$).

Valori per il sistema

R tempo di risposta del sistema:

$$R = \sum_{k=1}^{K} R_k$$

X traffico in uscita dal sistema. Se il sistema è stato parametrizzato in termini di D_k si riesce a calcolare X, ma non a suddividerlo nei vari X_k (la legge del flusso forzato è $X_k = V_k X$); si nota quindi come la parametrizzazione in termini di S_k e V_k dia un maggiore livello di dettaglio;

Q numero medio di clienti presenti nel sistema, che può essere calcolato sia come somma dei clienti presenti ai diversi centri

$$Q = \sum_{k=1}^{K} Q_k$$

sia applicando la legge di Little, da cui si ricava

$Q = N$ per clienti batch

$Q = X R$ per clienti transazionali

$Q = N - X Z$ per clienti interattivi

Ingressi per modelli con più classi di clienti

Si considerano ora diverse classi di clienti e con C si indica il numero di queste classi. Ogni parametro di ingresso deve essere riferito a una classe di clienti e viene perciò affiancato da un pedice.

Si distinguono:

1. modelli **aperti**, se tutte le classi sono di tipo transazionale;
2. modelli **chiusi**, se tutti i lavori sono di tipo interattivo o batch;

3. modelli **misti**, se ci sono classi di tipo interattivo o batch e classi di lavori transazionali.

Descrizione cliente

Per ogni classe l'intensità di carico è data specificando uno dei seguenti valori

δ_c frequenza di arrivo (transazionali);

N_c popolazione (batch)

N_c, Z_c popolazione e tempo di riflessione (interattivi)

Descrizione dei centri di servizio

Per ogni centro bisogna specificare se si tratta di centro ad accodamento o di ritardo.

Richieste di servizio

Per ogni classe c e centro k si specifica la richiesta di servizio indicando $D_{c,k}$ oppure $V_{c,k}$ e $S_{c,k}$.

Uscite per modelli con più classi di clienti

Valori per il singolo centro di servizio

I valori d'uscita possono essere espressi suddivisi per classi oppure aggregati. Si considerino dapprima i valori suddivisi per classe:

$U_{c,k}$ utilizzo del centro k dovuto a clienti della classe c;

$R_{c,k}$ tempo di residenza dei clienti di classe c al centro k;

$X_{c,k}$ traffico dei clienti di classe c in uscita dal centro k;

$Q_{c,k}$ clienti di classe c presenti al centro k;

I valori aggregati saranno invece:

U_k utilizzo del centro k:

$$U_k = \sum_{c=1}^{C} U_{c,k}$$

R_k tempo di residenza complessivo al centro k:

$$R_k = \frac{\sum_{c=1}^{C} R_{c,k} X_{c,k}}{X_k}$$

X_k traffico in uscita dal centro k:

$$X_k = \sum_{c=1}^{C} X_{c,k}$$

Q_k clienti presenti al centro k:

$$Q_k = \sum_{c=1}^{C} Q_{c,k}$$

Valori di sistema

Anche i valori di uscita relativi al sistema complessivo possono essere suddivisi per classe oppure aggregati. Si presentano dapprima quelli ripartiti per classe:

R_c tempo di risposta per tutti i clienti di classe c;

X_c traffico dei clienti di classe c in uscita dal sistema;

Q_c numero medio di clienti di classe c presenti nel sistema;

I valori aggregati sono invece:

R tempo di residenza nel sistema:

$$R = \frac{\sum_{c=1}^{C} R_c X_c}{X}$$

X traffico in uscita dal sistema:

$$X = \sum_{c=1}^{C} X_c$$

Q numero medio di clienti presenti nel sistema:

$$Q = \sum_{c=1}^{C} Q_c$$

Limiti sulle prestazioni

La valutazione di un sistema può essere effettuata anche attraverso l'analisi dei limiti superiore e inferiore delle prestazioni che questo sistema può fornire. Questo approccio è caratterizzato dalla semplicità di calcolo, che consente di ottenere rapidamente una prima approssimazione del comportamento del sistema. Nel seguito la trattazione sarà comunque limitata a sistemi con una sola classe di clienti. L'analisi dei limiti sulle prestazioni di sistemi con più classi di clienti è possibile, ma vengono a mancare le caratteristiche di semplicità e rapidità che rendono conveniente l'applicazione di questa tecnica.

Due sono i tipi di limiti che verranno presentati nel seguito: quelli asintotici e quelli di sistema bilanciato. I primi sono più semplici da calcolare, mentre i secondi forniscono risultati più precisi. In entrambi i casi il sistema viene definito indicando:

- il tipo di clienti (transazionali, interattivi o batch);
- il numero dei centri di servizio K;
- la domanda di servizio massima tra quelle relative ai diversi centri di servizio:

$$D_{max} = \max \{D_k\}$$

- la domanda complessiva di servizio:

$$D = \sum_{k=1}^{K} D_k$$

- in caso di clienti interattivi, il tempo medio di riflessione Z.

Limiti asintotici

L'ipotesi di base è che la domanda di servizio di un cliente nei confronti di un centro non dipenda né dal numero dei clienti presenti nel sistema, né dalla loro distribuzione nei diversi centri che compongono il sistema.

Modelli aperti (carico transazionale)

Al crescere della frequenza di arrivo dei clienti, il sistema raggiunge un limite oltre il quale non è più capace di soddisfare le richieste dei nuovi clienti (saturazione). Il valore di saturazione è quindi il valore massimo di frequenza di arrivo sopportabile dal sistema. Si consideri la legge dell'utilizzo applicata a un qualsiasi centro di servizio:

$$U_k = X_k S_k$$

Dalla legge del flusso forzato $X_k = X V_k$ e dall'ipotesi del bilancio di flusso $\delta = X$ si ricava

$$X_k = \delta V_k$$

Sostituendo quest'ultima espressione nella legge dell'utilizzo presentata prima si ottiene:

$$U_k = \delta V_k S_k = \delta D_k$$

Per tutti i centri di servizio l'utilizzo deve comunque essere inferiore (o al massimo uguale) al 100% perciò per ogni centro k deve valere:

$$U_k = \delta D_k \leq 1$$

da cui si ricava

$$\delta \leq \frac{1}{D_k}$$

ovvero

$$\delta \leq \delta_{sat} = \frac{1}{D_{max}}$$

Si consideri ora il tempo di risposta del sistema. Nel caso migliore (limite inferiore) ogni cliente non interferisce con nessun altro cliente e quindi non ci sono ritardi dovuti all'attesa in coda. In questo caso il tempo di risposta equivale alla domanda di servizio complessiva ($R = D$). Per valutare il caso peggiore (limite superiore) bisogna invece considerare la possibilità che i clienti possano arrivare in gruppi di n ogni n/δ unità di tempo. La frequenza di arrivo è ancora $n / {}^{n}/_{\delta} = \delta$. I clienti che si trovano in fondo al gruppo dovranno accodarsi agli altri e aspettare che questi siano serviti prima di poter ricevere il servizio richiesto al sistema. Al crescere di n aumenta anche il tempo di attesa, perciò, se non è nota la distibuzione degli arrivi, ma si conosce solo la fequenza media δ, non è possibile stabilire un limite superiore al tempo di risposta.

Modelli chiusi (carico batch o interattivo)

Per la definizione dei limiti nel caso dei modelli chiusi vengono considerati carichi di tipo interattivo. I limiti nel caso di carico batch possono essere ricavati tramite lo stesso procedimento, semplicemente eliminando il tempo di riflessione (come già detto prima i carichi batch sono equivalenti a quelli interattivi a patto di porre $Z = 0$).

In maniera analoga a quanto fatto nel caso di lavori transazionali, e notando che valgono le seguenti relazioni:

$$\delta \leq \frac{1}{D_k}$$

$$U_k = X_k S_k$$

$$X_k = X V_k$$

$$D_k = V_k S_k$$

da cui

$$U_k = X V_k S_k = X D_k$$

è possibile riscrivere la legge dell'utilizzo come:

$$U_k(N) = X(N) D_k$$

Siccome l'utilizzo di ogni centro deve comunque essere inferiore a 1 si ricava che per ogni centro k deve valere:

$$X(N) \leq \frac{1}{D_k}$$

e quindi si ottiene un primo vincolo sul traffico del sistema:

$$X(N) \leq \frac{1}{D_{max}}$$

Si consideri ora la situazione che si viene a creare quando nel sistema circola un solo cliente; il traffico è:

$$X(1) = \frac{1}{D+Z}$$

infatti si potrà osservare un completamento ogni $D+Z$ secondi. Quando si aggiungono gli altri N-1 clienti, nel caso migliore non si verifica nessuna interferenza e si ottiene:

$$X(\mathrm{N}) = \frac{N}{D+Z}$$

cioè si osservano N completamenti ogni $D+Z>$ secondi. Nel caso peggiore, invece, ogni cliente deve aspettare in coda che tutti gli altri siano serviti, e perciò deve spendere (N-1) secondi in coda, oltre a D secondi in servizio e Z secondi in riflessione. Si ottiene quindi:

$$X(\mathrm{N}) = \frac{N}{ND+Z}$$

Riassumendo, nel caso di carico interattivo, si può scrivere:

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

mentre per i carichi di tipo batch si ottiene:

$$\frac{1}{D} \leq X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{D}\right\}$$

L'andamento di questi limiti è riportato in figura 1.7.

Dalla legge del tempo di risposta si ricava che

$$X = \frac{N}{R+Z}$$

e sostituendo questa espressione nelle disequazioni presentate sopra si ottengono i limiti asintotici del tempo di risposta per sistemi interattivi:

$$\max\left\{D, ND_{max} - Z\right\} \leq R(N) \leq ND$$

e per sistemi batch:

$$\max \{D, ND_{max}\} \leq R(N) \leq ND$$

riportati entrambi in figura 1.8.

Il punto di equivalenza dei due limiti superiori del traffico e dei due limiti inferiori del tempo di risposta è:

$$N^* = \frac{D+Z}{D_{max}}$$

Si definisce carico **leggero** un carico dovuto a un numero di clienti inferiore a N^*, quando cioè il limite superiore al traffico è stabilito dalla retta

$$X = \frac{N}{D+Z}$$

mentre il limite inferiore del tempo di risposta è indicato dalla retta

$$R = D$$

Se invece il numero dei clienti presenti nel sistema è maggiore di N^* si dice che il carico è **pesante** ed il traffico è limitato superiormente dalla retta

$$X = \frac{1}{D_{max}}$$

mentre il tempo di risposta ha come limite inferiore la retta

$$R = N D_{max} - Z$$

Nel caso di carichi leggeri, traffico e tempo di risposta sono vincolati dal valore della domanda complessiva, quindi è possibile migliorare le prestazioni diminuendone il valore. Quando invece ci si trova in situazione di carico pesante il vincolo è la domanda massima, quella cioè che identifica quale dei centri di servizio è il collo di bottiglia del sistema. Miglioramenti delle prestazioni sono possibili solo se viene rimosso questo collo di bottiglia.

In tabella 1.2 sono riassunte le formule di calcolo dei limiti asintotici.

Tabella 1.2 – Schema riassuntivo per il calcolo dei limiti asintotici

Sistemi	Traffico	Tempo di risposta
Transazionali	$X(\delta) \leq \delta_{sat} = \frac{1}{D_{max}}$	$R(\delta) \geq D$
Batch	$\frac{1}{D} \leq X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{D}\right\}$	$\max\{D, ND_{max}\} \leq R(N) \leq ND$
Interattivi	$\frac{N}{ND+Z} \leq X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{D+Z}\right\}$	$\max\{D, ND_{max} - Z\} \leq R(N) \leq ND$

Limiti di sistemi bilanciati

I limiti di sistemi bilanciati consentono di ottenere risultati più precisi di quelli dati dai limiti asintotici. Richiedono però dei calcoli più complessi.

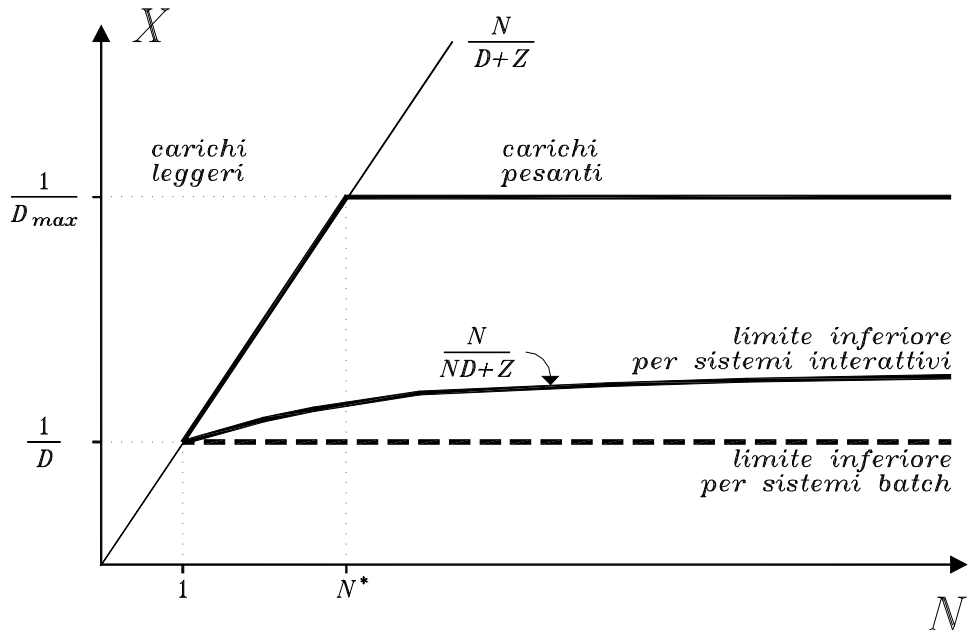

Figura 1.7 – Andamento dei limiti asintotici per il traffico

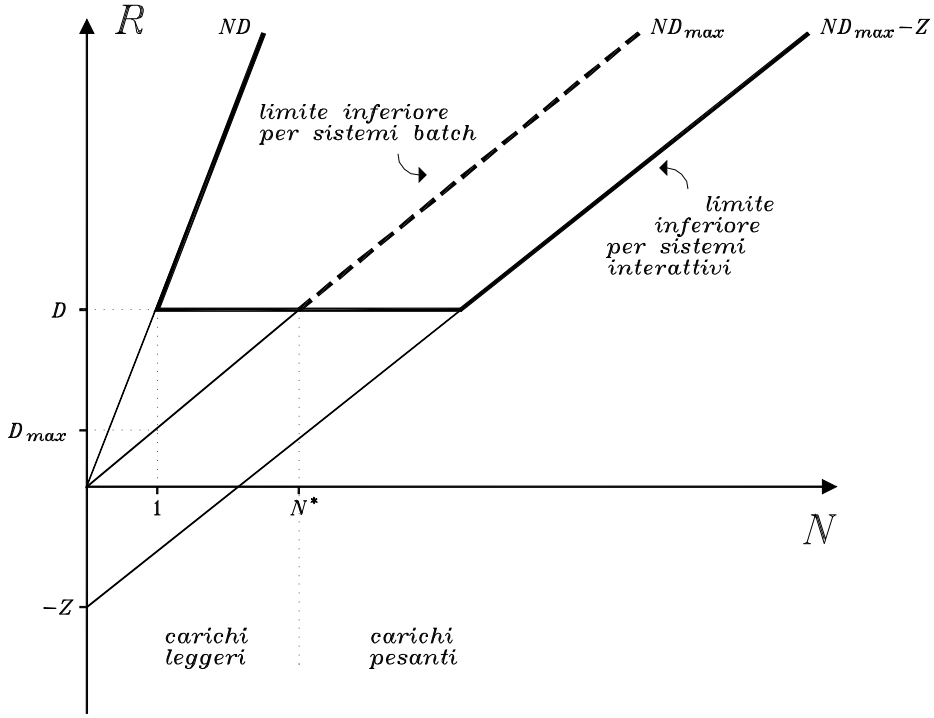

Figura 1.8 – Andamento dei limiti asintotici per il tempo di risposta

Per sistema bilanciato si intende un sistema in cui tutte le domande ai diversi centri di servizio sono uguali ($D_1 = D_2 = ... = D_K = D/K = D_b$). Nel seguito, per semplicità, vengono presentati i limiti validi per i sistemi batch. I valori relativi ai casi transazionale e interattivo sono riportati soltanto

nella tabella 1.3 riassuntiva; comunque possono essere ottenuti in maniera analoga a quanto riportato per i carichi batch.

Il tempo di residenza al centro k è:

$$R_k(N) = D_k \left[1 + A_k(N)\right]$$

dove $A_k(N)$ è la coda che un cliente (l'N-esimo) vede quando arriva al centro k. Si presti attenzione a non confondere questo valore con quello indicato da $Q_k(N)$ che rappresenta invece il numero dei clienti (compreso l'N-esimo) presenti in media al centro k. Si noti inoltre come per i centri di ritardo risulti che $A_k(N) = 0$.

Il tempo di risposta del sistema si ottiene come somma dei tempi di residenza ai k centri:

$$R(N) = \sum_{k=1}^{K} R_k(N) = \sum_{k=1}^{K} D_k \left[1 + A_k(N)\right]$$

Nel caso dei sistemi bilanciati, la domanda è uguale per tutti i centri ($\forall k$: $D_k = D_b$), si ottiene quindi:

$$R(N) = D_b \sum_{k=1}^{K} \left[1 + A_k(N)\right] = D_b \left[\sum_{k=1}^{K} 1 + \sum_{k=1}^{K} A_k(N) \right] = D_b \left[K + \sum_{k=1}^{K} A_k(N) \right]$$

Per i carichi di tipo batch si trova che:

$$\sum_{k=1}^{K} A_k(N) = N - 1$$

in pratica il numero complessivo dei clienti visti in coda dall'N-esimo cliente è pari a tutti i clienti presenti nel sistema tranne appunto quest'ultimo cliente. A questo punto è possibile esprimere il tempo di risposta come:

$$R(N) = D_b \left[K + N - 1\right]$$

Dal sistema sotto esame è possibile ricavare

$$D_{max} = \max\{D_k\} \qquad \text{e} \qquad D_{min} = \min\{D_k\}$$

Tra tutti i sistemi con K centri, N clienti, domanda massima non superiore a D_{max} e domanda minima non inferiore a D_{min}, quello che ha prestazioni peggiori è ovviamente il sistema bilanciato in cui tutte le domande sono pari a D_{max}, mentre quello che ha prestazioni migliori è altrettanto ovviamente il sistema bilanciato con tutte le domande pari a D_{min}. Dall'esame di questi casi limite si può ricavare:

$$(N + K - 1)\, D_{min} \le R(N) \le (N + K - 1)\, D_{max}$$

Dalla legge di Little $R = N / X$ è possibile ottenere i limiti del traffico:

$$\frac{N}{N + K - 1} \frac{1}{D_{max}} \le X(N) \le \frac{N}{N + K - 1} \frac{1}{D_{min}}$$

Si può restringere ulteriormente questo limite considerando soltanto i sistemi con domanda complessiva pari a:

$$D = \sum_{k=1}^{K} D_k$$

Tra tutti questi sistemi, quello con prestazioni migliori è quello bilanciato, in cui la domanda di servizio ai diversi centri è uguale a $D_{med} = D / K$. Si possono quindi ottenere i seguenti limiti ottimistici:

$$R(N) \geq (N + K - 1)\, D_{med} = K\, D_{med} + (N - 1)\, D_{med} = D + (N - 1)\, D_{med}$$

$$X(N) \leq \frac{N}{N + K - 1}\ \frac{1}{D_{med}} = \frac{N}{(N - 1)\, D_{med} + K\, D_{med}} = \frac{N}{(N - 1)\, D_{med} + D}$$

Il caso peggiore, invece, si ha quando solo D / D_{max} centri hanno domanda pari a D_{max}, mentre la domanda degli altri è pari a 0. Questo sistema corrisponde quindi a un sistema bilanciato con $K' = D / D_{max}$ centri di servizio. Si ottengono perciò i seguenti limiti pessimistici:

$$R(N) \leq \left(N + \frac{D}{D_{max}} - 1 \right) D_{max} = D + (N - 1)\, D_{max}$$

$$X(N) \geq \frac{N}{N + \dfrac{D}{D_{max}} - 1}\ \frac{1}{D_{max}} = \frac{N}{D + (N - 1)\, D_{max}}$$

In tabella 1.3 sono riassunti i limiti di sistemi bilanciati anche per i casi transazionali e interattivi.

Tabella 1.3 – Schema riassuntivo per il calcolo dei limiti di sistema bilanciato

Sistemi	**Traffico**
Transazionali	$X(\delta) \leq \delta_{sat} = \dfrac{1}{D_{max}}$
Batch	$\dfrac{N}{D + (N-1)\, D_{max}} \leq X(N) \leq \min\left\{ \dfrac{1}{D_{max}}, \dfrac{N}{(N-1)\, D_{med} + D} \right\}$
Interattivi	
	Tempo di risposta
Transazionali	$\dfrac{D}{1 - \delta D_{med}} \leq R(\delta) \leq \dfrac{D}{1 - \delta D_{max}}$
Batch	$\max\{N D_{max},\ D + (N-1)\, D_{med}\} \leq R(N) \leq D + (N-1)\, D_{max}$
Interattivi	

Tecniche di soluzione analitiche per modelli con una sola classe di clienti

La procedura di soluzione analitica dei modelli con una sola classe di clienti (*One Job Class – OJC*) è diversa a seconda che si riferisca a modelli aperti oppure a modelli chiusi.

Tecnica di soluzione per modelli aperti

Come già visto in precedenza, il massimo carico tollerabile da un sistema interattivo è pari a $\delta_{sat} = 1/D_{max}$. Oltre questo valore il sistema satura. Nel seguito si considera sempre un sistema non saturo, ovvero con un carico $\delta \leq \delta_{sat}$.

Dall'ipotesi di bilancio del flusso si ottiene che $X(\delta) = \delta$, cioè il traffico in uscita dal sistema è uguale al carico in ingresso. La legge del flusso forzato consente di ottenere il valore del traffico in uscita da ogni singolo centro come $X_k(\delta) = \delta V_k$.

L'utilizzo dei diversi centri di servizio può essere ricavato dalla legge dell'utilizzo che dà:

$$U_k(\delta) = X_k(\delta)S_k = \delta V_k S_k = \delta D_k$$

Per quello che riguarda il tempo di residenza bisogna distinguere tra centri di ritardo e centri ad accodamento. Nel primo caso il tempo di residenza è pari alla domanda di servizio, perché non si spende tempo in coda e perciò si ottiene $R_k(\delta) = V_k S_k = D_k$. Se invece si considerano i centri ad accodamento, nel tempo di residenza deve essere compreso il tempo di servizio ($V_k S_k$) e il tempo di attesa in coda ($V_k S_k A_k(\delta)$), dove con $A_k(\delta)$ viene indicata la coda che un cliente trova quando arriva al centro k. Si ottiene quindi $R_k(\delta) = D_k \left[1 + A_k(\delta)\right]$.

Nel caso di carichi di tipo transazionale, la lunghezza della coda che un nuovo cliente vede quando arriva al centro k (espressa da $A_k(\delta)$) è uguale al numero medio dei clienti presenti allo stesso centro (indicato da $Q_k(\delta)$).

Applicando la legge di Little, $N = XR$, al centro k si ottiene $Q_k(\delta) = X_k(\delta)\ \overline{R_k(\delta)}$, dove con $\overline{R_k(\delta)}$ si indica il tempo di residenza al centro k per ogni singola visita ($\overline{R_k(\delta)} = R_k(\delta) / V_k$). Applicando la legge del flusso forzato, come già mostrato prima, si ricava $Q_k(\delta) = \delta V_k \overline{R_k(\delta)}$, e quindi $Q_k = \delta R_k(\delta)$. Sostituendo questa espressione in quella che in precedenza dava il valore del tempo di risposta, si ottiene ora:

$$R_k(\delta) = D_k \left[1 + \delta R_k(\delta)\right] = D_k + \delta D_k R_k(\delta) = D_k + U_k(\delta)R_k(\delta)$$

da cui è immediato ricavare la seguente espressione del tempo di risposta:

$$R_k(\delta) = \frac{D_k}{1 - U_k(\delta)}$$

Si noti come questa espressione soddisfi la nozione intuitiva del tempo di risposta secondo cui valgono i seguenti comportamenti asintotici:

$$\lim_{U_k \to 0} R_k(\delta) = D_k$$

$$\lim_{U_k \to 1} R_k(\delta) = \infty$$

Il tempo di risposta complessivo del sistema si ottiene sommando i tempi di residenza ai diversi centri:

$$R(\delta) = \sum_{k=1}^{K} R_k(\delta)$$

La lunghezza della coda al centro k (ovvero il numero medio di clienti presenti in un centro) risulta essere $Q_k(\delta) = \delta R_k(\delta)$. Quando si considerano i centri di ritardo, si ottiene:

$$Q_k(\delta) = \delta D_k = U_k(\delta)$$

Nel caso invece dei centri ad accodamento, si trova:

$$Q_k(\delta) = \frac{U_k(\delta)}{1 - U_k(\delta)}$$

Il numero medio dei clienti presenti complessivamente nel sistema si ottiene sommando le code che si trovano ai diversi centri:

$$Q(\delta) = \sum_{k=1}^{K} Q_k(\delta) = \delta R(\delta)$$

Tutte queste formule sono sufficienti a risolvere in maniera analitica un modello con un carico transazionale.

Tecnica di soluzione per modelli chiusi

La tecnica che permette la soluzione nel caso di modelli chiusi va sotto il nome di analisi del valor medio (*Mean Value Analysis – MVA*) e si basa sull'applicazione ripetuta delle seguenti formule:

1. Tempo di residenza per centri di ritardo:

   $R_k(N) = D_k$

   e per centri ad accodamento:

   $R_k(N) = D_k[1+A_k(N)]$

2. Legge di Little applicata al sistema complessivo:

   $$X(N) = \frac{N}{R(N)+Z}$$

3. Legge di Little applicata a ogni centro di servizio:

   $Q_k(N) = X(N)\, R_k(N)$

Con $A_k(N)$ si indica la lunghezza della coda vista dall'ultimo cliente (l'N-esimo) quando arriva al centro k. Questo valore equivale al numero dei clienti presenti in media al centro k quando i clienti complessivi non sono N, ma N–1. Perciò si può scrivere $A_k(N) = Q_k(N–1)$. Per ottenere la soluzione del modello con N clienti si deve utilizzare un procedimento iterativo: con un solo cliente le code sono vuote e quindi si può scrivere $A_k(1) = Q_k(0) = 0$, da cui è possibile ricavare il tempo di residenza ai singoli centri $R_k(1)$, il traffico $X(1)$ e le code ai diversi centri $Q_k(1)$ utilizzando le formule presentate prima. A questo punto si conoscono anche i valori di $A_k(2) = Q_k(1)$, e quindi le stesse formule possono essere applicate di nuovo per risolvere il sistema con due clienti. Procedendo in questo modo si giunge fino al numero di clienti desiderato N.

Si noti che insieme alla soluzione del modello con N clienti si ottiene anche la soluzione dello stesso modello con N–1, N–2, ... 2, 1 clienti. Il calcolo può perciò essere pesante quando N assume valori elevati. In questi casi ci si può accontentare di una soluzione approssimata che richieda una quantità inferiore di calcoli. L'espressione che dà la lunghezza della coda può essere approssimata nel seguente modo:

$$A_k(N) = Q_k(N–1) \approx \frac{N–1}{N} Q_k(N)$$

Questa ipotesi è asintoticamente corretta, infatti

$$\lim_{N\to\infty} \frac{Q_k(N)}{N} = \lim_{N\to\infty} \frac{Q_k(N–1)}{N–1}$$

Al primo passo si sceglie un valore qualsiasi per le code (in genere si assume $Q_k(N) = N/K$ per rendere più rapida la convergenza dell'algoritmo), si calcolano poi i relativi valori del tempo di residenza, del traffico e anche i nuovi valori delle code. Questi nuovi valori vengono confrontati con i precedenti: se lo scostamento è superiore a una certa soglia si ripete il procedimento utilizzando questi nuovi valori. Se invece lo scostamento è inferiore alla soglia i valori ottenuti danno la soluzione del modello.

In tabella 1.4 sono riportati schematicamente gli algoritmi per la valutazione analitica dei modelli a una classe di clienti (*One Job Class – OJC*).

Tabella 1.4 – Algoritmi di soluzione analitica per modelli OJC

Algoritmo OJC esatto	Algoritmo OJC approssimato

for (k = 1;$k \le K$; k++)
  Q_k = 0;
for (i = 1;$i \le N$; i++) {
  R = 0;
  for (k = 1;$k \le K$; k++) {
  $A_k = Q_k$;
  if (ritardo) $R_k = D_k$;
  if (accodamento) $R_k = D_k (1 + A_k)$;
  $R = R + R_k$;
  }
  $X = i / (R + Z)$;
  for (k = 1;$k \le K$; k++)
  $Q_k = X R_k$;
  }
	for (k = 1;$k \le K$; k++)
	  $Q_k = N / K$;
	do {
	  R = 0;
	  for (k = 1;$k \le K$; k++) {
	  $A_k = Q_k (N–1) / N$;
	  if (ritardo) $R_k = D_k$;
	  if (accodamento) $R_k = D_k (1 + A_k)$;
	  $R = R + R_k$;
	  }
	  $X = N / (R + Z)$;
	  $diff$ = 0;
	  for (k = 1;$k \le K$; k++) {
	  if ($Q_k – XR_k/Q_k > diff$) $diff = Q_k – XR_k/Q_k$;
	  $Q_k = X R_k$;
	  }
	  } while ($diff > soglia$);

Tecniche di soluzione analitiche per modelli con più di una classe di clienti

I sistemi con più classi di clienti (*Multiple Job Class – MJC*) possono essere classificati a seconda che i clienti siano solo transazionali (modelli aperti), solo interattivi o batch (modelli chiusi) oppure sia transazionali che interattivi o batch (modelli misti).

In ogni caso è necessario indicare, per tutte le classi di clienti c e per tutti i centri k, la domanda di servizio $D_{c,k}$.

Tecnica di soluzione per modelli aperti

Nel caso dei modelli aperti l'intensità del carico viene espressa dal vettore $\vec{\delta} = [\delta_1, \delta_2, ..., \delta_C]$. La capacità di elaborazione è limitata dalla seguente relazione:

$$\forall k: \sum_{c=1}^{C} \delta_c D_{c,k} \le 1$$

Il traffico in uscita dal sistema segue la legge del flusso forzato e perciò può essere espresso come:

$$X_{c,k}(\vec{\delta}) = \delta_c V_{c,k}$$

L'utilizzo di ogni singolo centro viene ricavato applicando l'omonima legge:

$$U_{c,k}(\vec{\delta}) = X_{c,k}(\vec{\delta}) S_{c,k} = \delta_c D_{c,k}$$

L'utilizzo complessivo del centro k si ottiene dalla somma degli utilizzi relativi alle diverse classi:

$$U_k = \sum_{c=1}^{C} U_{c,k}$$

Il tempo di residenza di un cliente di classe *c* al centro *k* può essere espresso come:

$$R_{c,k}(\vec{\delta}) = D_{c,k}$$

nel caso di centri di ritardo e come:

$$R_{c,k}(\vec{\delta}) = \frac{D_{c,k}}{1 - \sum_{j=1}^{C} U_{j,k}(\vec{\delta})}$$

per i centri ad accodamento (la dimostrazione di questa relazione non viene riportata per ragioni di semplicità; può comunque essere ottenuta in maniera analoga a quanto fatto nel caso di sistemi con una sola classe di clienti).

Il tempo di risposta complessivo del sistema si ottiene sommando i tempi di residenza ai diversi centri:

$$R_c(\vec{\delta}) = \sum_{k=1}^{K} U_{c,k}(\vec{\delta})$$

Il numero di clienti di classe *c* presenti al centro *k* (normalmente definito come lunghezza della coda) è:

$$Q_{c,k}(\vec{\delta}) = \delta_c R_{c,k}(\vec{\delta})$$

Si ottiene perciò:

$$Q_{c,k}(\vec{\delta}) = U_{c,k}(\vec{\delta})$$

per i centri di ritardo e

$$Q_{c,k}(\vec{\delta}) = \frac{U_{c,k}(\vec{\delta})}{1 - U_{c,k}(\vec{\delta})}$$

Il numero dei clienti presenti in media nel sistema si ottiene sommando il numero di clienti in coda ai diversi centri:

$$Q_c(\vec{\delta}) = \delta_c R_c(\vec{\delta}) = \sum_{k=1}^{K} Q_{c,k}(\vec{\delta})$$

Tecnica di soluzione per modelli chiusi

In questo caso il carico viene espresso dai vettori $\vec{N} = [N_1, N_2, ..., N_C]$ e $\vec{Z} = [Z_1, Z_2, ..., Z_C]$ con $Z_c = 0$ per le classi di tipo batch. La tecnica che permette la soluzione nel caso di modelli chiusi si basa sull'applicazione ripetuta delle seguenti formule:

1. Tempo di residenza per centri di ritardo:

$R_{c,k}(\vec{N}) = D_{c,k}$

e per centri ad accodamento:

$R_{c,k}(\vec{N}) = D_{c,k}[1 + A_{c,k}(\vec{N})]$

2. Legge di Little applicata al sistema complessivo, classe per classe:

$$X_c(\vec{N}) = \frac{N_c}{Z_c + \sum_{k=1}^{K} R_{c,k}(\vec{N})}$$

3. Legge di Little applicata a ogni centro di servizio, classe per classe:

$Q_{c,k}(\vec{N}) = X_c(\vec{N})\, R_{c,k}(\vec{N})$

Da questi valori è possibile ottenere la coda totale al centro *k*:

$$Q_k(\vec{N}) = \sum_{c=1}^{C} Q_{c,k}(\vec{N})$$

Con $A_{c,k}(\vec{N})$ si indica la lunghezza della coda vista dall'ultimo cliente di classe *c* quando arriva al centro *k*. Questo valore equivale al numero dei clienti presenti in media al centro *k* quando i clienti complessivi non sono $\vec{N}$, ma $\vec{N}-\vec{1}_c$ con cui si indica la popolazione originaria a cui è stato rimosso un cliente di classe *c*. Perciò si può scrivere:

$A_{c,k}(\vec{N}) = Q_{c,k}(\vec{N}-\vec{1}_c)$

Per ottenere la soluzione del modello con $\vec{N}$ clienti si deve utilizzare un procedimento iterativo: la soluzione per una popolazione nulla può essere ricavata semplicemente perché, si verifica che

$\forall k: \; Q_k(\vec{0}) = 0$

Si possono quindi utilizzare le formule presentate prima per costruire la soluzione per popolazioni via via crescenti fino ad arrivare a quella desiderata. In figura 1.9 è riportato un diagramma delle precedenze di calcolo relativo a una popolazione composta da 3 clienti di classe A e 2 di classe B.

Si noti che insieme alla soluzione del modello con $\vec{N}$ clienti si ottiene anche la soluzione dello stesso modello con un numero di clienti di classe *c* compreso tra 0 e N_c. Il calcolo può perciò essere pesante quando i diversi N_c assumono valori elevati. In questi casi, come per i sistemi con una sola classe di clienti, ci si può accontentare di una soluzione approssimata che richieda una minor mole di calcoli. Il punto cruciale è l'espressione della lunghezza della coda che può essere approssimata come

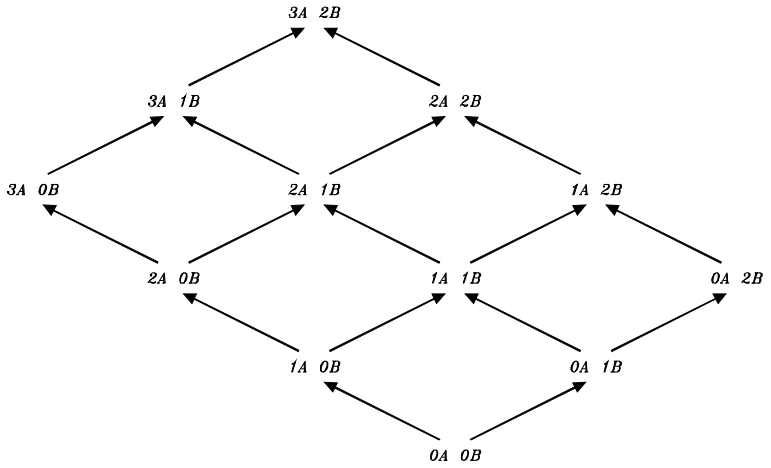

Figura 1.7 – Precedenze di calcolo in un sistema con due classi chiuse di clienti

$$A_{c,k}(\vec{N}) = Q_{c,k}(\vec{N-1_c}) \approx \frac{N_c - 1}{N_c} \; Q_{c,k}(\vec{N}) + \sum_{\substack{j=1 \\ j \neq c}}^{C} Q_{j,k}(\vec{N})$$

Al primo passo si sceglie un valore qualsiasi per le code (in genere si assume $Q_{c,k}(\vec{N}) = N_c / K$ per rendere più rapida la convergenza dell'algoritmo), si calcolano poi i relativi valori del tempo di residenza, del traffico e anche i nuovi valori delle code. Questi nuovi valori vengono confrontati con i precedenti: se lo scostamento è superiore a una certa soglia si ripete il procedimento utilizzando questi nuovi valori, se invece lo scostamento è inferiore alla soglia i valori ottenuti danno la soluzione del modello.

Una ulteriore possibilità per una valutazione approssimativa di modelli chiusi a più classi di clienti si basa su una modifica dell'algoritmo OJC esatto, presentato in precedenza per i sistemi con una sola classe di clienti. Si consideri per esempio il caso di un sistema a due classi di clienti, nel quale i clienti delle due classi interferiscano tra di loro in un solo centro. Una soluzione approssimata del modello può essere ottenuta applicando l'algoritmo OJC esatto per entrambe le classi di clienti e sommando, ad ogni passaggio, le code del centro condiviso, in modo da tener conto dell'interferenza tra i clienti delle due classi. Nel caso le domande delle due classi al centro condiviso siano uguali, è sufficiente applicare l'algoritmo OJC esatto per una sola classe di clienti, raddoppiando la code al centro condiviso. Per estensione, la stessa tecnica si può applicare a sistemi con più classi di clienti, che condividano più centri di servizio.

I risultati che si ottengono dall'applicazione di questo algoritmo OJC modificato sono in genere sovrastimati rispetto a quelli ottenuti dagli algoritmi MJC: infatti, sommando le code, ad ogni iterazione viene considerato il carico calcolato al passo precedente, quindi nel risultato finale non viene considerato l'ultimo cliente delle classi diverse da quella cui si applica l'algoritmo modificato. Perciò il risultato che si ottiene si riferisce a un numero di clienti inferiore rispetto a quello effettivo del sistema.

Tecnica di soluzione per modelli misti

In questi casi l'intensità di carico viene descritta da un vettore:

$$\vec{I} = [N_1 \, o \, \delta_1, N_2 \, o \, \delta_2, ..., N_C \, o \, \delta_C]$$

La tecnica di soluzione utilizza passi che appartengono alle tecniche di soluzione presentate prima per i modelli aperti e per quelli chiusi. Nelle espressioni che vengono presentate nel seguito con A si indica l'insieme delle classi aperte e con B quello delle classi chiuse.

Il primo passo consiste nel calcolo, per tutti i K centri, dell'utilizzo dovuto alle classi aperte:

$$\forall k \in K, \forall c \in A: \; U_{c,k}(\vec{I}) = \delta_c D_{c,k}$$

$$\forall k \in K: \; U_{A,k}(\vec{I}) = \sum_{c \in A} \delta_c D_{c,k}$$

Il valore dell'utilizzo viene quindi usato come fattore correttivo delle domande di servizio relative alle classi chiuse:

$$\forall k \in K, \forall c \in B: \; D^{*}_{c,k} = \frac{D_{c,k}}{1 - U_{A,k}(\vec{I})}$$

Uno degli algoritmi presentati prima per le classi chiuse può quindi essere utilizzato per la soluzione delle classi $c \in B$, considerando le domande $D^{*}_{c,k}$, ricavando il traffico $X_{c,k}$, la lunghezza delle code $Q_{c,k}$ e il tempo di residenza $R_{c,k}$. A questo punto è possibile calcolare l'utilizzo dei centri come

$$U_{c,k} = X_{c,k} S_{c,k} = X_{c,k} D_{c,k}$$

e il numero medio dei clienti delle classi chiuse presenti ai diversi centri come

$$Q_{B,k}(\vec{I}) = \sum_{c \in B} Q_{c,k}(\vec{I})$$

Rimangono infine da calcolare il tempo di residenza e la lunghezza delle code per le classi aperte:

$$\forall k \in K, \forall c \in A: \; R_{c,k}(\vec{I}) = \frac{D_{c,k}[1 + Q_{B,k}(\vec{I})]}{1 - U_{A,k}(\vec{I})}$$

e

$$Q_{c,k}(\vec{I}) = \delta_c R_{c,k}(\vec{I})$$