La teoria delle code

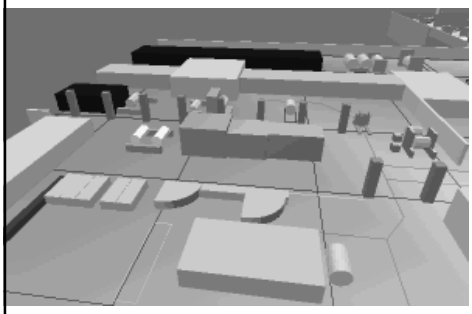
Il problema:
trovare la configurazione ottima di un **sistema produttivo** (quali e quante risorse acquisire) dato un problema produttivo (caratteristiche tecnologiche e volumi richiesti per ciascun tipo di parte da produrre).

Le prestazioni di un sistema produttivo vengono determinate in massima parte in fase di progettazione: le scelte effettuate in fase operativa, di natura prettamente gestionale, possono determinare una efficienza superiore a quella definita nella fase di progetto per non più del 20%...

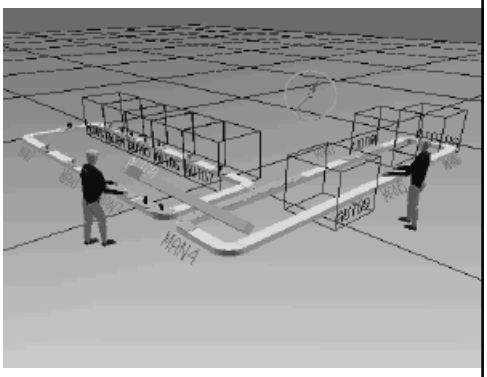
La teoria delle code

Alcuni esempi di configurazione del sistema

Quanti carrelli trasportatori…?



Quanti operatori di assemblaggio… ?



Introduzione

Il problema si articola in due fasi:
❶ generazione della configurazione del sistema
❷ valutazione della bontà della configurazione generata.

L'esigenza di ridurre al minimo il time-to-market (tempo tra l'ideazione del prodotto e l'entrata dello stesso nel mercato) ha portato allo sviluppo di strumenti capaci di supportare il progettista in ciascuna fase del processo di configurazione in maniera efficiente dal punto di vista sia della qualità sia del tempo.

Introduzione

Esistono, pertanto, due macro-categorie di metodi a supporto dell'attività di configurazione dei sistemi produttivi:

- ✓ **metodi di generazione**
- ✓ **metodi di valutazione**

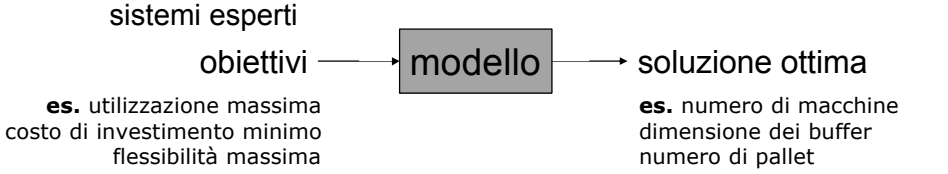
che fanno capo a tre tecniche generali:

- modellazione matematica
- tecniche di simulazione
- intelligenza artificiale

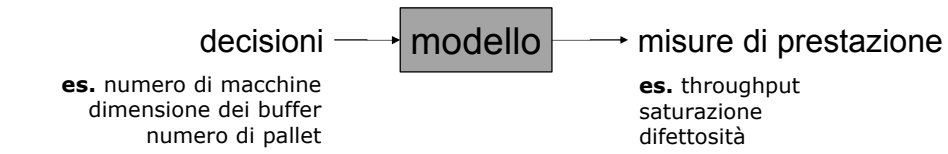
Queste tecniche sono fra di loro complementari e non esclusive.

Introduzione

Modelli di generazione:
es. ottimizzazione matematica, programmazione stocastica, sistemi esperti

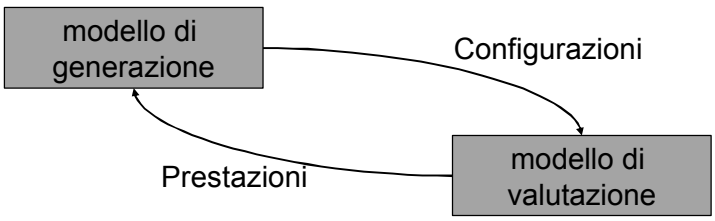


Modelli di valutazione:
es. simulazione, teoria delle code, reti di Petri



Introduzione

Le due famiglie di modelli devono essere combinate ed integrate in una *struttura ad anello chiuso*, in cui i modelli di generazione forniscono ai modelli di valutazione le configurazioni da valutare e usano il feedback che ne deriva per prendere le decisioni o modificare le decisioni prese.

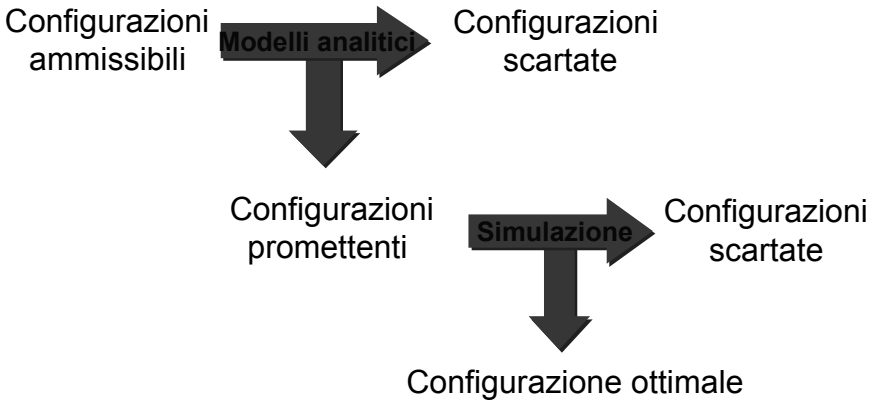


Introduzione

Tra i modelli di valutazione, i modelli analitici permettono di determinare velocemente e con un margine di errore dell'ordine del 10%, le prestazioni di un sistema produttivo modellato in maniera "semplificata", attraverso relazioni matematiche tra grandezze rappresentative del flusso di materiali all'interno del sistema stesso.

Se ci sono numerose configurazioni alternative, si usano i modelli analitici per eliminare rapidamente quelle meno promettenti. Successivamente, metodi di valutazione più sofisticati e precisi, saranno usati per scegliere la migliore tra le configurazioni rimaste.

Introduzione



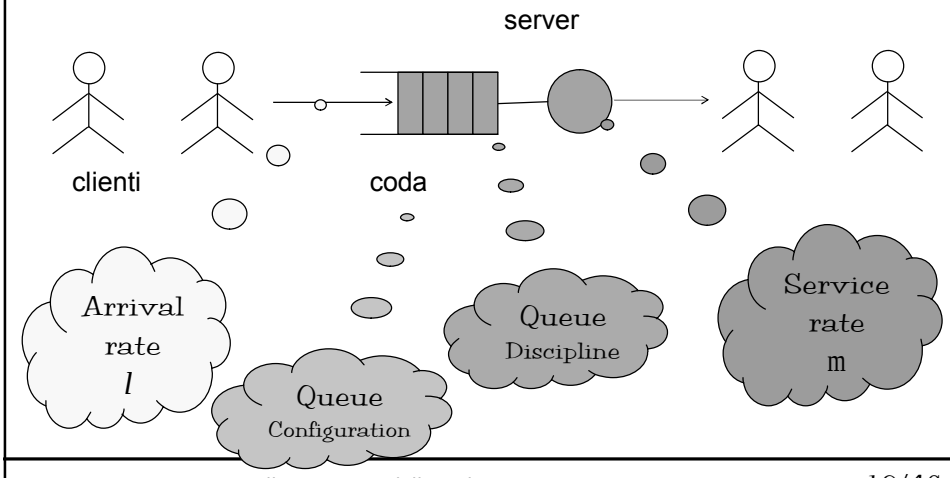
Sistemi a coda

Tra i metodi analitici, ci sono i metodi basati sulla Teoria delle code, nei quali il sistema produttivo viene assimilato ad un sistema a coda

Un flusso di *clienti* in arrivo nel sistema chiede di essere servito da una risorsa (*servitore*) per un determinato tempo di servizio. Poiché i servitori sono generalmente in numero limitato, l'istante di arrivo dei clienti è casuale e, inoltre, essi richiedono un servizio la cui durata può non essere costante, una parte dei clienti dovrà attendere in *coda.*

Sistemi a coda

I sistemi a coda possono essere costituiti da una coda isolata:



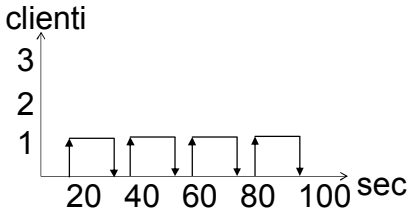
Sistemi a coda

Siano:
λ tasso di arrivo medio dei clienti nel sistema
μ tasso di servizio medio per cliente
deterministici e sia $\mu > \lambda$.

➲Il cliente trova sempre il sistema vuoto e il server libero

Es.

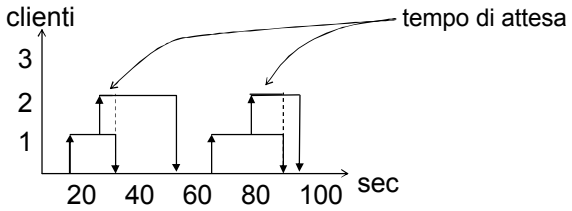
λ=3 clienti/min
μ=1/15 cliente/sec



Sistemi a coda

Siano:
λ tasso di arrivo medio dei clienti nel sistema
μ tasso medio di servizio per cliente
e sia $\mu > \lambda$.

Numero di clienti in arrivo e tempo di servizio sono **stocastici** ...
➲Il cliente *può* trovare il server occupato e in questo caso deve attendere in coda finché esso non si libera.



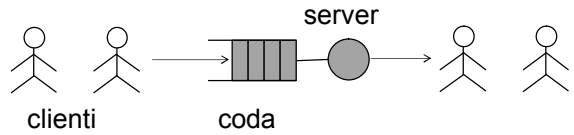
Sistemi a coda

Nel caso $\mu < \lambda$, i clienti in arrivo trovano sempre il server occupato (il tempo di servizio, deterministico o stocastico, è maggiore del tempo tra due arrivi consecutivi) e la coda cresce infinitamente (il sistema non si svuota mai)...

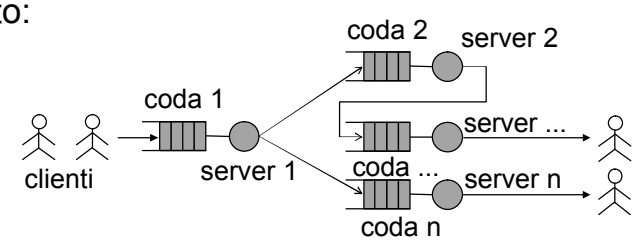
➲Il sistema non è stabile.

Sistemi a coda

I sistemi a coda possono essere costituiti da una coda isolata:



oppure possono avere una struttura a rete, in cui più sistemi a coda isolata concorrono all'esecuzione del servizio completo:



Sistemi a coda

Se il sistema a coda è un ufficio postale o bancario...

Un flusso di *clienti* in arrivo nel sistema chiede di essere servito da una risorsa (*servitore*) per un determinato tempo di servizio. Poiché i servitori sono generalmente in numero limitato, l'istante di arrivo dei clienti è casuale e, inoltre, essi richiedono un servizio la cui durata può non essere costante, una parte dei clienti dovrà attendere in *coda*

Tempo necessario ad eseguire le operazioni richieste

Persone

File allo sportello

Impiegati

Sistemi a coda

Se il sistema a coda è una rete di calcolo...

Un flusso di *clienti* in arrivo nel sistema chiede di essere servito da una risorsa (*servitore*) per un determinato tempo di servizio. Poiché i servitori sono generalmente in numero limitato, l'istante di arrivo dei clienti è casuale e, inoltre, essi richiedono un servizio la cui durata può non essere costante, una parte dei clienti dovrà attendere in *coda*

Tempo necessario ad eseguire le operazioni richieste

Task

memoria

CPU

Sistemi a coda

Se il sistema a coda è un sistema produttivo...

Un flusso di *clienti* in arrivo nel sistema chiede di essere servito da una risorsa (*servitore*) per un determinato tempo di servizio. Poiché i servitori sono generalmente in numero limitato, l'istante di arrivo dei clienti è casuale e, inoltre, essi richiedono un servizio la cui durata può non essere costante, una parte dei clienti dovrà attendere in *coda*

Tempo di lavorazione, di attrezzaggio, di trasporto

Parti da lavorare

Buffer interoperazionali

Centri di lavoro, stazioni di carico/scarico, trasportatori

Modelli analitici e sistemi a coda

Le grandezze rappresentative del flusso di materiali all'interno del sistema produttivo che, attraverso relazioni analitiche, modellano il sistema stesso, sono quindi:

- ☑ il numero di server (MC, L/U, trasportatori) per ciascun tipo di risorsa
- ☑ l'ampiezza dei buffer di sistema e locali
- ☑ il tempo di interarrivo delle parti da lavorare per ciascun tipo di parte
- ☑ il tempo di servizio richiesto da ciascun tipo di parte su ciascun tipo di risorsa
- ☑ la regola di priorità secondo cui le parti in coda vengono servite

Teoria delle code

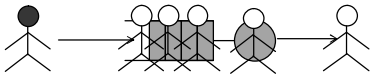
Studia i sistemi a coda e si propone di determinarne le prestazioni, ad esempio l'entità del flusso di clienti in uscita dal sistema, l'entità dei ritardi e delle code, in funzione delle grandezze caratteristiche del sistema a coda considerato.

La strutturazione del modello si fonda fortemente sulle assunzioni fatte circa le distribuzioni di probabilità associate alle variabili che modellano il tempo di arrivo dei clienti e il tempo di servizio.

Teoria delle code e code isolate

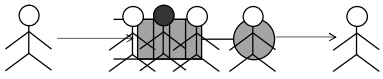
La dinamica del processo:

❶ Il cliente $\mathbf{C_n}$ entra nel sistema all'istante $\mathbf{i_n}$



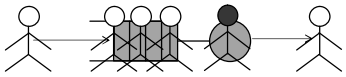
N-1 clienti sono entrati nel sistema. Di questi, alcuni sono stati serviti e hanno abbandonato il sistema, altri sono in servizio, i rimanenti in attesa

❷ $\mathbf{C_n}$ subisce un tempo di attesa $\mathbf{w_n}$ dal momento in cui entra nel sistema al momento in cui viene servito: durante questo periodo, altri clienti vengono serviti e altri ne arrivano

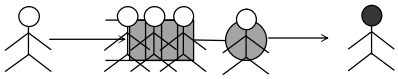


Teoria delle code e code isolate

❸ $\mathbf{C_n}$ viene prelevato dalla coda secondo la regola di priorità in vigore e trascorre un tempo $\mathbf{t_n}$ in servizio



❹ $\mathbf{C_n}$ esce dal sistema dopo un tempo $s_n = w_n + t_n$ detto tempo di attraversamento del sistema:



Distribuzione degli arrivi

Tra due clienti successivi $\mathbf{C_{n-1}}$ e $\mathbf{C_n}$ trascorre un intervallo di tempo $t_n = i_n - i_{n-1}$, che la teoria delle code modella come una variabile aleatoria con distribuzione nota.

Le distribuzioni più comunemente impiegate per modellare t_n sono:

❍ Distribuzione costante

$f(\tau)=\delta(\tau - \tau_0)$ dove δ è la funzione di Dirac

- Distribuzione esponenziale:

  $f(\tau)=\lambda e^{-\lambda\tau}$ dove:

  - $E(\tau)= \lambda^{-1}$ e $\sigma(\tau)=\lambda^{-1}$
  - λ è il numero medio di arrivi nell'unità di tempo
  - $F(\tau)=1-e^{-\lambda\tau}$ è la probabilità cumulata

Distribuzione degli arrivi

La distribuzione esponenziale per la modellazione dell'intervallo di tempo tra arrivi successivi riveste particolare importanza ed è spesso impiegata perché:

① c'è una buon accordo con le osservazioni sperimentali

② gode di un certo numero di proprietà per cui è possibile determinare per via analitica senza approssimazione le principali caratteristiche di un sistema a coda

La distribuzione esponenziale è l'unica distribuzione a godere della proprietà secondo cui la distribuzione del tempo residuo è indipendente dal tempo già maturato!

Distribuzione del tempo di servizio

Anche t_n è modellato come una variablile aleatoria e le distribuzioni più comunemente impiegate sono:

- Distribuzione costante:
  $f(t)=\delta(t-t_0)$ dove δ è la funzione di Dirac

- Distribuzione esponenziale:
  $f(t)=\mu e^{-\mu t}$ dove:

  - $E(t)= \mu^{-1}$
  - $\sigma(t)= \mu^{-1}$

- altre (Erlang, iper-esponenziale)

Discipline di priorità

❶ Discipline basate sull'ordine di arrivo:

⊙ FCFS (First Come, First Served): i clienti vengono serviti in stretto ordine di arrivo
- ⊙ (es. coda su stazione di sosta di un nastro trasportatore)

⊙ LCFS (Last Come, First Served): l'ultimo cliente arrivato, viene servito
- ⊙ (es. se i pezzi in uscita da uno stadio produttivo vengono impilati)

⊙ Serve in random order: i clienti vengono serviti in ordine casuale
- ⊙ (es. se i pezzi sono alla rinfusa in un cesto)

Discipline di priorità

❷Discipline basate sulla classe dei clienti:

HOL (Head Of Line): i clienti vengono assegnati ad una classe e a ciascuna di queste viene associato un livello di priorità: nell'ambito di una stessa classe la disciplina di servizio è FCFS o LCFS, ma i clienti di classi a priorità inferiore vengono serviti solo se non ci sono clienti di classi a priorità maggiore.

❸ Discipline basate sulla durata del servizio richiesto:

SPT (Shortest Processing Time): viene servito il cliente che chiede il servizio di durata minima

LPT (Longest Processing Time): viene servito il cliente che chiede il servizio di durata massima

Classificazione delle code isolate

Ciascuna coda viene caratterizzata con la notazione:

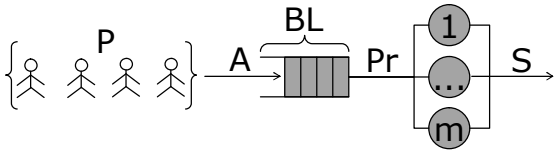
A/S/m/P/BL/Pr

dove:

- A indica la distribuzione degli arrivi
- S indica la distribuzione dei tempi di servizio

{ D: costante
M: esponenziale
G: generale }

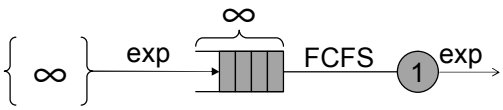
- m è il numero di servitori
- P è la numerosità della popolazione di clienti
- BL è la capacità massima della coda
- Pr indica disciplina di priorità



Esempi notevoli

Coda M/M/1

- distribuzione degli arrivi: esponenziale
- distribuzione dei tempi di servizio: esponenziale
- coda SS (single server)
- popolazione di clienti infinita (omessa)
- spazio di attesa infinito (omesso)
- disciplina di priorità: FCFS (omessa)



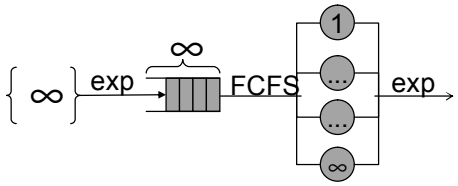
Se ci sono n>1 clienti nel sistema, n-1 sono in coda!

Le code SS possono modellare le stazioni di lavoro!

Esempi notevoli

Coda M/M/¥

- distribuzione degli arrivi: esponenziale
- distribuzione dei tempi di servizio: esponenziale
- coda IS (infinite server)
- popolazione di clienti infinita (omessa)
- spazio di attesa infinito (omesso)
- disciplina di priorità: FCFS (omessa)



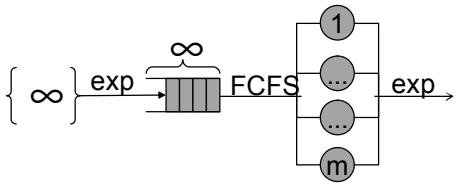
Le code IS possono modellare segmenti di rulliera!

Non ci sono mai clienti in coda, perché vi è sempre un server disponibile (la coda è solo un elemento di ritardo).

Esempi notevoli

Coda M/M/m

- distribuzione degli arrivi: esponenziale
- distribuzione dei tempi di servizio: esponenziale
- coda MS (multiple server)
- popolazione di clienti infinita (omessa)
- spazio di attesa infinito (omesso)
- disciplina di priorità: FCFS (omessa)



Le code MS possono modellare stazioni di lavoro, L/U, trasportatori,...

Se ci sono n>m clienti nel sistema, n-m devono attendere in coda.

Prestazioni delle code isolate

Intensità di traffico

quantità di servizio richiesto nell'unità di tempo (**UDT***)*

arrivi nell'unità di tempo * tempo medio di servizio

Capacità del sistema

quantità di servizio che il sistema può produrre per UDT

numero di servitori

Prestazioni delle code isolate

Fattore di utilizzo

rapporto tra intensità di traffico e capacità del sistema

Il fattore di utilizzo si può interpretare anche come percentuale di servitori mediamente occupata

Prestazioni delle code isolate

Es.

l=1.5 ➡ in media viene richiesto un servizio di 1.5 UDT per UDT

se C=m=2, ρ=0.75 ➡ in media ciascun servitore viene impiegato per il 75% del tempo

Prestazioni delle code isolate

Flusso uscente (throughput)

quantità di clienti che escono dal sistema per UDT

Nei casi più semplici (popolazione infinita e spazio d'attesa infinito)

= min (flusso entrante, capacità massima del sistema)

Ciascun servitore può smaltire, infatti, al più μ clienti per UDT

Prestazioni delle code isolate

Dal punto di vista del cliente...

Le misure di prestazione che interessano sono i valori medi e le deviazioni standard delle variabili:

- tempo di attraversamento del sistema s_n
- tempo di attesa in coda w_n
- tempo di servizio t_n

$$s_n = w_n + t_n$$

Prestazioni delle code isolate

Da un punto di vista aggregato...

Le grandezze di interesse sono i valori medi e le deviazioni standard di:

- il numero di clienti nel sistema, n(t)
- il numero di clienti in coda, q(t)
- il numero di clienti in servizio, $n_s(t)$

Se il sistema ha capacità m, deve essere:

$$q(t) = \max(0, n(t) - m)$$

Il teorema di Little

Stabilisce una relazione tra flusso in entrata, tempo di attraversamento e numero di clienti compresi tra due confini specificati di un sistema a coda.

$\lambda E(s)=E(n)$

$\lambda E(w)=E(q)$

$\lambda E(t)=E(n_s)$

Sebbene i risultati di questo teorema siano stati acquisiti ed impiegati da lungo tempo, esso è stato dimostrato solo nel 1961.
La dimostrazione verrà omessa

Il risultato non dipende dalla distribuzione degli arrivi e del tempo di servizio e rimane valido qualunque sia la disciplina di servizio, purché conservativa.

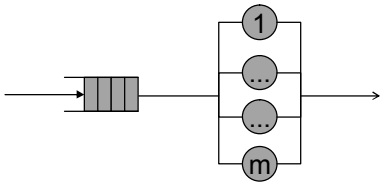
Il teorema di Little

Esempio.

La capacità produttiva di uno skilift è di 12 sciatori al minuto.
La risalita richiede 15 minuti.
Di quanti ganci dispone l'impianto?

La discesa richiede 10 minuti. Sapendo che il numero totale di sciatori è 500, quanto si attende in media alla base dello skilift?

Il teorema di Little



server: ganci
flusso in arrivo: corrisponde al flusso degli sciatori che scendono dalla pista
coda: alla base dello skilift attendono gli sciatori che non stanno scendendo dalla pista e che non stanno salendo con lo skilift
tempo di servizio: tempo di risalita
throughput: sciatori sganciati al minuto

Il teorema di Little

Hp. L'impianto è sempre saturo

Applico il teorema di Little: $\overline{n_s} = ?\bar{t} = 12 \times 15 = 180$

Se ci sono 180 clienti in salita, avremo in totale 360 ganci.

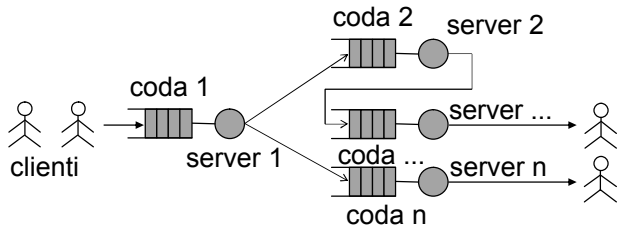
Applico il teorema di Little: $\overline{n_s} = ?\bar{t} = 12 \times 10 = 120$

Se ci sono 120 clienti in discesa, avremo in totale 300 clienti in servizio e i rimanenti 500-300=200 in coda.

Applico ancora il teorema di Little: $\overline{w} = \frac{\bar{q}}{?} = \frac{200}{12} = 16 \text{ min}$

Reti di code

Si tratta di sistemi a coda con struttura a rete, in cui più sistemi a coda isolata concorrono all'esecuzione del servizio completo:



Si prestano a modellare contesti applicativi che prevedono insiemi di moduli o elementi tra loro interconnessi in modo che l'input di ciascuno è una combinazione degli output degli altri.

Classificazione delle reti di code

Assi di classificazione:

- sistema chiuso vs. aperto (il numero di clienti che circolano nel sistema è fisso vs. è una variabile random)
- blocking vs. non-blocking (i buffer sono finiti e, se il buffer a valle di un server è pieno, il server a monte si blocca o i buffer sono infiniti e nessun server può mai bloccarsi)
- distribuzione degli arrivi al sistema
- distribuzione dei tempi di servizio
- disciplina di servizio

Classificazione delle reti di code

- reliable vs. unreliable server (affidabilità dei server inferiore a 1 o pari a 1)
- single class vs. multiple class (aggregazione dei parametri caratteristici dei diversi tipi di cliente in un cliente rappresentativo o modellazione esplicita dei diversi tipi di cliente del sistema)
- routing dipendente vs. indipendente dallo stato del sistema (i clienti visitano i server seguendo un percorso che dipende dallo stato del sistema o meno)

Reti di code

L'estensione della teoria delle code alle reti di code ha portato ai seguenti risultati:

☑ teoremi che consentono il calcolo esatto delle prestazioni delle reti caratterizzate da distribuzione esponenziale dei tempi di servizio e interarrivo

☑ metodi approssimati che consentono di valutare con discreta precisione le prestazioni delle reti con caratteristiche generali

Prestazioni delle reti di code

- probabilità marginale: è la probabilità di trovare n_i clienti nella coda i-esima
- flusso uscente dalla coda i-esima e dal sistema
- utilizzo delle stazioni della rete
- dimensione della coda i-esima, da intendersi come numero medio di clienti in attesa o in servizio alla coda i-esima
- tempo di attraversamento della coda i-esima e della rete

Reti di code

I risultati ottenuti estendendo la teoria delle code alle reti di code non sono in forma esplicita come nel caso delle code isolate.
Rispetto alle code isolate, occorre infatti modellare le transizioni dei clienti da una coda all'altra e, in generale, non è possibile analizzare ciascuna coda indipendentemente dalle altre.
Occorre pertanto implementare, avvalendosi di un calcolatore, gli algoritmi utili alla valutazione di tali parametri di prestazione.