Appunti di Analisi dei Sistemi ad Eventi

Giacomo Sturm

AA: 2023/2024 - Ing. Informatica

Sorgente del file LaTeX disponibile su
https://github.com/00Darxk/Analisi-dei-Sistemi-ad-Eventi

Indice

1 Introduzione

Verrano forniti due modelli di sistemi, reali o astratti, un modello matematico, reti di Code, ed un modello logico, reti di Petri. Quest'ultimo è un modello grafico, simile ad un diagramma di flusso. La rete di Petri analizza le interazioni tra gli elementi del sistema, mentre la rete di Code analizza nel tempo queste interazioni ingresso-uscita. In questi modelli si analizza l'evoluzione di una variabile di stato, da individuare nel sistema analizzato, per studiare la funzione obiettivo. La variazione della variabile di stato si studia tramite derivata continua o discreta, oppure si campiona il suo valore ad intervalli fissi. L'analisi ad eventi consiste nel misurare solamente se succede qualcosa al sistema, se avviene un evento, ovvero non c'è spreco di memoria campionando lo stesso valore. Per determinare un evento si controlla se la variabile di stato considerata è cambiata, questa variabile può essere sia deterministica oppure aleatoria. In caso sia aleatoria, conoscere la sua distribuzione di probabilità non è sufficiente per determinarne l'evouzione, sono necessaria la media, il valore centrale della distribuzione, e la varianza, la distanza dal valore centrale nella distribuzione.

Si usano sistemi manufatturieri come esempi, poiché sono comuni e seplici da studiare. Viene definita una coda un luogo dove i clienti o utenti aspettano il servizio. Quando un cliente entra nel sistema, se è disponibile un servente, viene servito, se non è disponibile si mette in coda. Viene definito tempo di processamento il tempo necessario per un cliente affinché sia servito. Si considerano i clienti usciti dal sistema dopo essere stati serviti. Si considera per ipotesi la coda ordinate in FIFO (First In First Out), ovvero si considera il primo cliente entrato in coda, il primo servito, se sono disponibili serventi. La coda del modello può essere illimitata oppure limitata con un massimo numero clienti k. Si indica il numero dei serventi con s. Si definisce la variabile di stato di questo sistema il valore intero n che rappresenta il numero di clienti all'interno del sistema, il suo valore massimo corrisponde alla massima capienza dei servneti e della coda: $n \in [0, s + k]$. Questo valore si incremente o decrementa di uno ogni volta che un cliente entra o esce dal sistema. In uno stesso istante non può avvenire più di un evento, ovvero la variabile può variare di uno in ogni istante. Si chiamano questi eventi di incremento e decremento processi di nascita e morte. Questo sistema è descritto da una legge di transizione:

$$\begin{cases} n = n + 1 \\ n = n - 1 \end{cases}$$

Quest'equazione rappresenta la legge di evoluzione del sistema. Un evento rappresenta l'arrivo o la partenza di un cliente dal sistema. In questo modello la variazione è slegata dal tempo, noto solo il cambiamento della variabile di stato ad ogni evento, per cui rappresenta un modello logico. Se ad ogni evento viene assegnato una durata di tempo il modello diventa temporizzato, in maniera asincrona, ovvero ogni evento corrisponde ad intervalli di tempo diversi. L'obiettivo del modello è determinare l'evoluzione del sistema, questo può comprendere il numero di clienti, il tempo di servizio, differenza tra il tempo di entrata ed il tempo di uscita di un cliente, il tempo di attesa. Conoscendo il tempo di processamento si può deterinare se il sistemaè sotto o sovra-utilizzato.

2 Reti di Petri

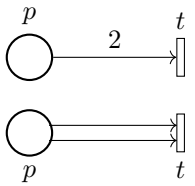
La rete di Petri è un modello logico per rappresentare sistemi ad eventi deterministici (DES), può rappresentare comportamenti complessi come la sincronizzazione, il succedersi asincrono di eventi che avvengono in intervalli di tempo diversi, operazioni concorrenti che avvengono totalemente indipendentemente tra di loro, conflitti ed altre caratteristiche di sistemi ad eventi.

La rete di petri è una rappresentazione grafica con una struttura matematica, è modulare e limitata, potendo rappresentare un ciclo continuo, è possibile gestire il ridimensionamento della rete senza perdere le sue proprietà. La rete è un grafo bipartito, formato da due tipi di nodi, i posti p e le transizioni t. Si possono unire solamente posti-transizioni tramite archi orientati.

Viene definito pre-set di un nodo x l'insieme dei nodi immediatamente a monte di x: $\bullet x$, viene invece definito post-set di un nodo x l'insieme dei nodi immediatamente a valle di x: $x\bullet$. Lo stato del sistema viene definito dalla marcatura x un vettore colonna di dimensione pari al numero di posti $|P|$, dove P indica l'insieme dei posti, ed avente ogni componente di valore uguale al numero di gettoni presenti nel posto associato:

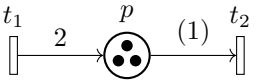
$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_{|P|} \end{pmatrix}$$

Viene definita marcatura iniziale x_0 lo stato assunto dal sistema all'inizio della sua analisi. I nodi sono collegati da archi pesati, il peso di un arco esprime il numero di gettoni generati, in caso sia in entrata ad un posto, oppure consumati, in caso sia in entrata ad una transizione. Il peso di un arco può essere indicato come un numero espresso sopra l'arco, per convenzione se è omesso il peso si considera di peso unitario, oppure si possono rappresentare come un numero di archi pari al peso dell'arco:



2.1 Evoluzione

Una transizione è abilitata se i posti a monte della transizione contengono almeno abbastanza gettoni da poter essere tutti consumati dai rispettivi archi.



In questo esempio la transizione t_2 è abilitata, poiché l'arco consuma tre gettoni e nel posto immediatamente a monte della transizione sono presenti tre gettoni. Ad ogni transizione può essere associato un tempo di processamento, in modo da temmporizzare il sistema. Se il pre-set di una transizione è vuoto, allora quella transizione è sempre abilitata.

Il numero di stati possibili in una qualsiasi configurazione corrisponde al numero di transizioni abilitate in quella data configurazione. Questi stati possibili possono essere rappresentati con un grafo di stato, in base alla rete e alla marcatura iniziale considerata x_0.

Si definiscono posti con un pre-set nullo appesi ed il numero di gettoni al loro interno può o rimanere costante o diminuire. Per cui se il sistema si basa solamente su posti appesi, allora sicuramente si bloccherà, incontra un "deadlock".

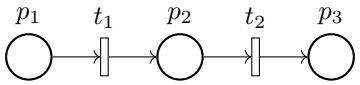
L'evoluzione di un sistema viene determinata dall'accadimento di eventi abilitati, ognuno con una sua abilità di accadere. La possibilità che un evento accada dipende dall' abilitazione di una transizione, l'effetto del suo accadimento corrisponde allo scatto di una transizione. L'abilitazione di una transizione dipende solamente dal peso dell' arco in entrata e dai gettoni nel pre-set, è abilitata se il numero dei gettoni nel pre-set è almeno uguale al peso dei rispettivi archi. Lo scatto di una transizione provoca un "flusso" di gettoni, questo flusso non è continuo, poiché i gettoni in entrata alla transizione vengono consumati e ne vengono creati di nuovi sulla base del peso dell'arco in uscita, numero indipendente dal numero dei gettoni consumati.

L'evoluzione comprende quattro passaggi ciclici: data una marcatura corrente si individua l'insieme delle transizioni abilitate, si sceglie casualmente, se non è specificato, una sola di queste transizioni, si provoca lo scatto di questa transizione che cambia la marcatura, si considera la nuova marcatura corrente e si ripetono questi passaggi.

Una sequenza di transizioni, abilitate, $(t_1, \cdots, t_n)$ si esprime con il simbolo S. Questa sequenza rappresenta l'ordine con cui le transizioni scattano, affinché rappresenti una sequenza valida, le transizioni considerate devono essere abilitate quando è il loro turno di scattare. Diverse sequenze possono arrivare alla stessa marcatura. Un singolo posto può abilitare più di una transizione, ma dopo lo scatto di una delle transizioni abilitate, potrebbe non avere gettoni rimanenti per abilitare le altre.

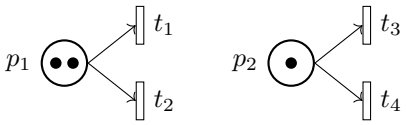
2.2 Strutture Fondamentali

Due transizioni si dicono in sequenza se sono collegate da un singolo posto:



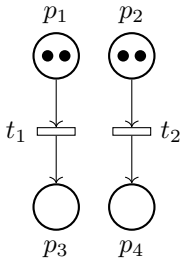
Le transizioni t_1 e t_2 si dicono in sequenza.

Un posto d'ingresso a due o più transizioni rappresenta un conflitto strutturale. Questo conflitto può essere effettivo se data una marcatura M, lo scatto di una transizione disabilita le altre transizioni. Il conflitto è potenziale se questo scatto non disabilita le altre transizioni.



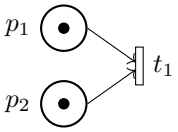
Le transizioni t_1 e t_2 si dicono in conflitto potenziale, le t_3 e t_4 si dicono in conflitto effettivo.

Due, o più, transizioni si dicono concorrenti se la loro evoluzione è indipendente l'una dall'altra:



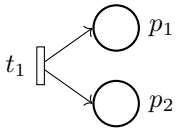
Le transizioni t_1 e t_2 si dicono in concorrenza strutturale, essendo entrambe abilitate si dicono in concorrenza effettiva.

Due o più posti si dicono sincronizzati se come post-set presentano la stessa transizione:



I posti p_1 e p_2 si dicono sincronizzati tra di loro, la transizione t si identifica come transizione di sincronizzazione

Due o più posti si dicono concorrenti se presentano in pre-set la stessa transizione, per cui allo scatto di quella transizione vengono generati dei gettoni in entrambi i posti:

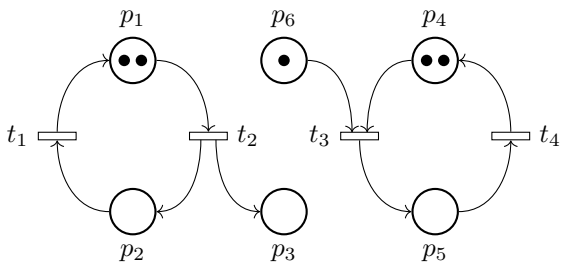


I due posti p_1 e p_2 si dicono concorrenti tra di loro, la transzizione t si identifica come transizione di inizio concorrenza.

Una rete di Petri si dice completa se non presenta nessun posto e transizione appese.

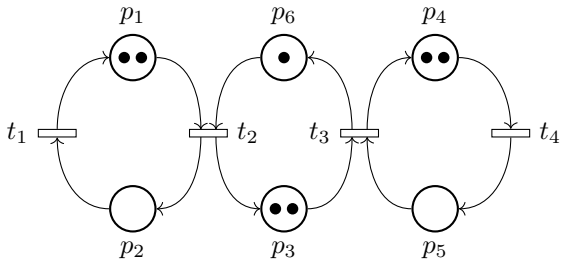
2.3 Esempio: Sistema Produttori/Consumatori

Si considera un sistema semplice formato da uno o più produttori che creano oggetti e li depositano in un buffer condiviso da cui uno o più consumatori possono prelevarli e consumarli. Per rappresentare un consumatore o un produttore si un ciclo che produce uno o più gettoni e lo depositano in un posto esterno al ciclo, oppure prelevano uno o più gettoni per poi consumarli con lo scatto di una transizione del ciclo:



In questo caso ogni volta che la transizione t_2 scatta, viene generato un gettone nel posto p_3, quindi il ciclo rappresenta un ciclo di produttori ed il numero di gettoni nel ciclo, indica il numero di produttori. La transizione t_3 è abilitata solo se è presente almeno un gettone nel posto p_6, per cui questo ciclo consuma un gettone ogni volta che scatta t_6, rappresenta un ciclo di consumatori, ed il numero di gettoni nel ciclo rappresenta il numero di consumatori del sistema. In un qualsiasi ciclo il numero di gettoni rimane sempre costante, se il peso degli archi che generano gettoni è uguale al peso dei gettoni che consumano gettoni, e se le transizioni sono sempre abilitate, altrimenti il ciclo non potrebbe né consumare né generare gettoni.

Per creare un sistema unico produttori-consumatori si considera un posto dove vengono depositati i gettoni generati dal ciclo dei produttori e consumati dal ciclo dei consumatori. Questo deposito può essere sia illimitato, nelle situazioni precedenti, oppure limitato. In questo caso è necessario un controllo nelle transizioni pre-set del buffer per impedire siano generatati gettoni se il deposito non può accomodarli, analogamente è necessario un controllo post-set per segnalare che un numero di gettoni è diminuito e quindi il deposito può accomodare più gettoni. Per indicare questo limite si crea un ciclo composto dalle transizioni generatrici nel cicli dei produttori, il posto buffer, le transizioni consumatrici dei cicli dei consumatori, ed un altro poste. In questo ciclo così definito il numero di gettoni rimane invariato, per cui il massimo numero di gettoni presenti nel deposito non può eccedere un limite imposto a priori:



In questo caso il deposito presenta un limite massimo di tre gettoni, e nel sistema sono presenti due consumatori e due produttori. Generalmente modelli di sistemi di produttori e consumatori presentano sempre dei cicli simili comunicanti tra di loro.

3 Proprietà

Una stessa rete presenta proprietà diverse in base ad una diversa marcatura iniziale M_0.

3.1 Raggiungibilità

Una marcatura M^* si dice raggiugibile se esiste almeno una sequenza S di transizioni abilitate tale che sia possibile, da una marcatura iniziale M, raggiungere la marcatura M^*:

$$M\,[S > M^*$$

Si definisce, data una rete di Petri N marcata con una marcatura M_0, l'insieme di raggiungibilità $R(N, M_0)$, l'insieme più piccolo di marcature tale che la marcatura iniziale appartiene all'insieme, e data una qualsiasi marcatura M^* appartenente all'insieme, ed una qualsiasi transizione t, abilitata, appartenente all'insisme delle transizioni T nella marcatura M^*. La transizione M^{**} ottenuta facendo scattare la transizione t nella marcatura M^* anch'essa appartiene all'insieme di raggiungibilità.

$$M_0 \in R(N, M_0)$$
$$M^* \in R(N, M_0) \wedge t \in T \text{ t.c. } M^*\,[\,t > M^{**} \implies M^{**} \in R(N, M_0)$$

3.2 Limitatezza

Un posto p_i di una rete N si dice $k-$limitato se in tutte le marcature raggiungibili, da una marcatura iniziale M_0, quel posto presenta al massimo k gettoni al suo interno:

$$\forall M \in R(N, M_0) \rightarrow m_i \leq k$$

Una rete N in una marcatura iniziale M_0 si dice $k-$limitata se tutti i suoi posti sono $k-$limitati. Se $k = 1$, la rete si dice binaria, poiché ogni posto può avere o zero o un singolo gettone. Una rete si dice limitata al massimo numero di gettoni che possono esistere in uno dei suoi posti, per cui è sufficiente un singolo posto illimitato affinché l'intera rete sia illimitata.

I cicli, analizzati precedentemente, rappresentano un caso semplice di rete limitata, poiché il numero di gettoni presenti nel ciclo rimane costante.

3.3 Reversibilità

Una rete N si dice reversibile, a partire da una marcatura iniziale M_0, se per ogni marcatura M appartenente all'insieme di raggiungibilià, la marcatura iniziale appartiene all'insieme di raggiungibilità della marcatura M:

$$\forall M \in R(N, M_0) \implies M_0 \in R(N, M)$$

Per cui una rete si dice reversibile se per ogni marcatura M raggiungibile deve esistere una serie di scatti S tali da ritornare alla marcatura originale M_0:

$$\forall M \in R(N, M_0) \implies M\,[\,S > M_0$$

3.4 Conservatitivà

Una rete N con marcatura iniziale M_0 si dice conservativa in riferimento ad un vettore peso W (colonna), maggiore uguale al vettore nullo 0, di dimensione pari alla cardinalità dell'insieme dei posti dim$W = |P|$, se per ogni marcatura M appartenente all'insieme di raggiungibilità R il prodotto matriciale tra la trasposta del vettore peso W ed il vettore marcatura M assume un valore finito e costante:

$$\exists W \geq 0 \text{ t.c. } \forall M \in R(N, M_0) \implies W^T \cdot M = k \in \mathbb{R}^+$$

Il vettore peso W poiché è maggiore uguale al vettore nullo presente al minimo una sola componente non nulla positiva, ed al massimo tutte componenti positive non nulle. Il prodotto tra W^T e M si può esprimere in diversi modi:

$$\forall M \in R(N, M_0) : W^T \cdot M = \begin{pmatrix} w_1 & \cdots & w_{|P|} \end{pmatrix} \cdot \begin{pmatrix} m_1 \\ \vdots \\ m_{|P|} \end{pmatrix} = \sum_{j=1}^{|P|} w_j m_j = k$$

$$M_0 \in R(N, M_0) \implies W^T \cdot M = W^T \cdot M_0 = k$$

$$\sum_{j=1}^{|P|} w_j m_j = \sum_{j=1}^{|P|} w_j m_{j0} \rightarrow \sum_{j=1}^{|P|} w_j (m_j - m_{j0}) = 0$$

Una rete si dice conservativa, se esiste un vettore peso W strettamente maggiore del vettore nullo, per cui il prodotto la trasposta del vettore ed una qualsiasi marcatura appartenente all'insieme di raggiungibilità risulta sempre costante:

$$\exists W > 0 \text{ t.c } \forall M \in R(N, M_0) \implies W^T \cdot M = k \in \mathbb{R}^+$$

Una rete conservativà quindi è $k-$limitata poiché non è necessario azzerare il contributo di un posto, a differenza del caso della conservativà in riferimento ad un vettore dove il vettore W contiene tanti zeri quanti sono i posti illimitati nella rete N, in modo da azzerare i loro contriubti nella somma.

Una rete si dice strettamente conservativa se è conservativa con riferimento al vettore identità, per cui tutte le componenti del vettore peso W assumono valore unitario: $\forall j \in 1, \cdots, |P| \implies w_j = 1$.

Una rete si dice non conservativa se è conservativa con riferimento ad un vettore peso nullo, per cui tutti i suoi posti sono illimitati, per cui il numero di posti rimane costante solo se non si considera nessun posto della rete.

In generale per controllare la conservatività si cerca il sottoinsieme più grande dell'insieme dei posti della rete N dove il numero di gettoni rimane complessivamente costante. Si deduce quindi che un ciclo rappresenta un elemeno conservativo, e se è presente in una rete, sarà sempre conservativà rispetto ad un vettore, se i posti del ciclo presentano almeno un vettore.

3.4.1 Vivezza

Una transizione t, di una rete N con marcatura iniziale M_0, si dice viva, se e solo se per ogni marcatura M appartenente all'insieme di raggiungibilità R esiste una marcaura M^* raggiungibile

da M, tale che la transizione t sia abilitata:

$$t : \text{viva} \iff \forall M \in R(N, M_0),\ \exists M^* \in R(N, M) \text{ t.c. } t : \text{abilitata in } M^*$$

Una rete N, con marcatura iniziale M_0, si dice raggiungibile se e solo se tutte le sue transizioni t_j sono vive:

$$N : \text{viva} \iff \forall t \in T \to t : \text{viva}$$

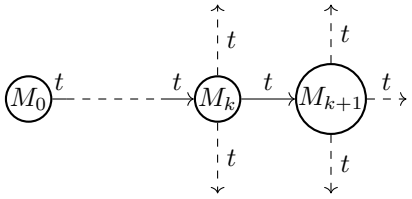
4 Analisi di Una Rete

4.1 Analisi Dinamica

Partendo da una rete è possibile creare un grafo di stato o grafo di raggiungibilità, che racchiude le relazioni tra ogni marcatura appartenente all'insieme di raggiungibilità, data una marcatura iniziale M_0, tramite li scatti di una singola transizione. Questo grafo è sempre limitato, anche se la rete non lo è. Dal grafo è possible inferire sulle proprietà della rete in quella data configurazione. Bisogna tenere conto delle differenza tra le prorpietà strutturali di una rete, che non dipendono dalla marcatura iniziale M_0, e le proprietà dinamiche che dipedono dalla marcatura M_0. Le proprietà individuate da un'analisi strutturale sono più importanti poiché intrinseche alla rete e verrano studiate nelle sezioni successive.

4.1.1 Grafo di Raggiungibilità e di Copertura

Un grafo di raggiungibilità è un grafo con un unico tipo di nodo che corrisponde ad una marcatura M. Sono presenti tanti nodi quante sono le marcature presenti nell'insieme di raggiungibilità R, a partire da una marcatura iniziale M_0. Gli archi del grafo uniscono due marcature collegate dallo scatto di una singola transizione, abilitata. Se è presente un numero finito di nodi, la rete è limitata, se sono presenti solo valori di 0 e 1, allora la rete è binaria. Se da ogni nodo del grafo esiste un percorso che abilita tutte le transizioni allora la rete è viva. Se da ogni nodo esiste un percorso che ritorna allo stesso nodo, la rete è reversibile.
Per costruire un grafo di raggiungibilità si parte dalla marcatura iniziale M_0, segnandola come nodo corrente. Si indica M_k la marcatura associata al nodo corrente; se non ci sono più transizioni attivabili a partire dal nodo corrente, non considerate in precedenza rispetto allo stesso nodo, e se il nodo corrente non corrisponde alla marcatura iniziale $k > 0$ allora si assegna come nodo corrente M_{k-1}, altrimeni l'algoritmo termina. Si considera la prima transizione abilitata, non considerata in precedenza con riferimento allo stesso nodo, e si calcola la marcatura raggiunta dal suo scatto. Se questa marcatura non corrisponde ad una marcatura già analizzata la si chiama M_{k+1}, e si crea un nodo associato ad essa collegato al nodo corrente da un arco, indicando la transizione scattata per arrivarci. Questo nodo diventa il nuovo nodo corrente e si ricomincia l'algoritmo cercando transizioni attivabili a partire da questo nodo corrente.



Al termine di questo algoritmo si ottiene il grafo di raggiungibilità di una rete N con marcatura iniziale M_0. Se tutti i nodi contengono marcatura, la cui somma dei gettoni è costante, allora la rete è strettamente conservativa, mentre se solo la somma di alcune posizioni delle marcature sono costanti allora la rete è conservativa in riferimento ad un vettore. Per controllare se la rete è conservativa rispetto ad un vettore $W > 0$, bisogna controllare che la somma pesata assuma valore

costante. Generalmente è meglio un vettore peso strettamente maggiore al vettore nullo che un vettore maggiore uguale al vettore nullo. Dato un vettore peso W, è possibile identificarne infiniti, combinazioni lineari del vettore W.
Se la rete è illimitata, si considera invece del grafo di raggiungibilità il grafo di copertura, che presenta un numero finito di nodi per poter descrivere la rete illimitata. Per identificare se una rete è illimitata si cerca una sequenza ammissibile di transizioni S da una marcatura M^* ad una marcatura M^{**}: $M^*[\, S > M^{**}$, tale che la marcatura M^{**} sia maggiore uguale alla marcatura M^*: $M^{**} \geq M^*$. Per cui presenta almeno un elemento maggiore della marcatura di partenza, ciò implica che il numero di gettoni complessivo è aumentato durante la sequenza S, ed almeno un posto presenta più gettoni rispetto all'inizio della sequenza. Necessariemente quindi la sequenza S è abilitata nella nuova marcatura M^{**}, e può scattare portando ad un'altra marcatura M^{***} maggiore uguale della precedente: $M^{**}[\, S > M^{***}$ t.c. $M^{***} \geq M^{**}$. Continuando aribtrariamente questo processo è possibile aumentare il numero di gettoni all'interno di almeno un posto della rete, per cui quei posti sono illimitati. La posizione dei posti illimitati o strettamente maggiori si indica con il simbolo ω, per indicare un numero arbitrario di gettoni:

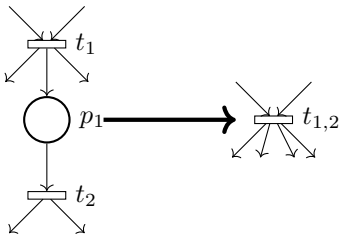
$$M = \begin{pmatrix} \vdots \\ \omega \\ \vdots \end{pmatrix}$$

Nel grafo di copertura, alla prima istanza di questo aumento arbitrario di gettoni si inserisce il termine ω nella posizione corrispondente ai posti illimitati, e si continua la costruzione del grafo seguendo le regole precedentemente definite. In alcuni casi è possibile, dato un determinato valore di ω, svuotare il posto illimitato arrivando ad un nodo con una marcatura senza ω. Oppure è possibile ritornare ad una marcatura finita precedentemente analizzata quindi collegando i due nodi con un'istruzione condizionale $\omega = k$, oltre al nome della transizione scattata. Se una rete è illimitata allora non può essere ciclica, ma può essere conservativa rispetto ad un vettore.

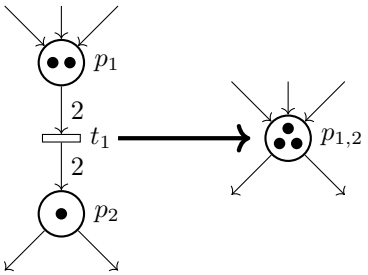
4.1.2 Tecniche di Riduzione

Si può ridurre il numero di posti di una rete, pur mantenendo le stesse proprietà, eccetto la conservatività. Poiché il vettore peso dipende dalla specifica rete considerata.

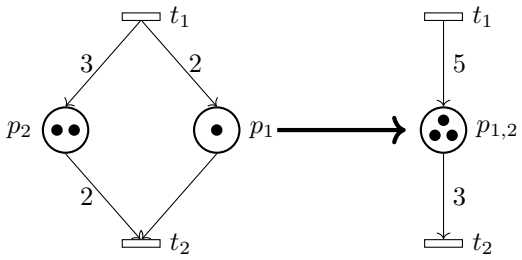
Se due posti o due transizioni sono connessi in serie, avendo in comune una singola transizione o posto (vuoto), possono essere sotituiti da un singolo posto o transizione. In caso di transizioni poste in serie, si possono unire solo se il posto in comune tra di loro è vuoto, altrimenti si perderebbe l'informazione dei suoi gettoni, e se il peso dell'arco entrante al posto equivale al peso dell'arco uscente dal posto:



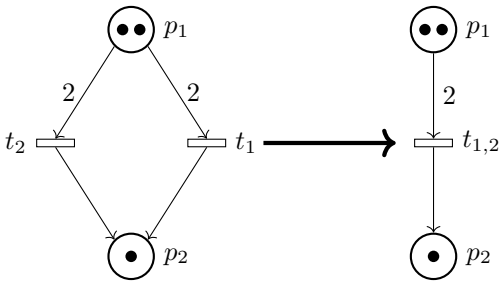
Due posti si possono unire se sono collegati da un'unica transizione (abilitata), ed il peso degli archi entranti equivale il peso degli archi uscenti da essa. Il posto risultante contiene la somma dei gettoni presenti nei due posti:



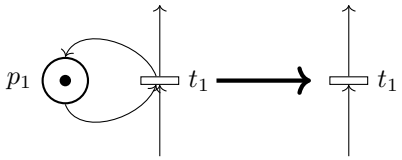
Inoltre è possibile unire insieme transizioni o posti in parallelo, ovvero aventi gli stessi insieme di pre-set e post-set. In caso siano due posti connessi in parallelo, il posto risultante contiene la somma dei gettoni, mentre l'arco entrante al posto ha peso dato dalla somma dei pesi degli archi entranti nei due posti originali, analogamente per l'arco uscente dal posto risultante:



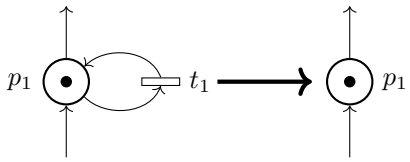
In caso siano presenti due transizioni poste in parallelo, per unirle è necessario che gli archi entranti in entrambe le transizioni abbiano lo stesso peso, analogamente per le transizioni in uscita, in questo modo date due marcatura prima e dopo lo scatto di una delle due transizioni è impossibile distinguere quale delle due sia scattata, per cui si considerano come un'unica transizione:



Se in una rete è presente un autociclo, ovvero un posto o una transizione che presenta in pre-set e post-set lo stesso insieme, si può eliminare poiché non altera il comportamento della rete. In caso sia un posto in autociclo, se gli archi in entrata ed in uscita al posto hanno lo stesso peso, il numero di gettoni al suo interno rimane costante; se il peso dell'arco in entrata alla transizione è maggiore del numero dei gettoni interni al posto, la rete è morta, poiché quella transizione non sarà mai abilitata, quindi eliminandolo bisogna indicare che la rete sia morta, anche se la rete ridotta non lo è. Se il peso dell'arco in uscita è minore del peso dell'arco in entrata si raggiunge la stessa situazione, e la transizione diventa morta.



Analogamente se è presenta una transizione avente in pre-set ed in post-set lo stesso posto, se il posto è $k-$limitato ed il peso del'arco entrante nella transizione è maggiore di k, allora quella transizione è morta. Se il peso degli archi entranti ed uscenti dalla transizione è uguale, non altera il numero dei gettoni nel posto associato.



Gli autociclici possono essere espressi come una doppia freccia, invece di due archi separati, quando entrambi i pesi uguali.

4.2 Rappresentazione Algebrica

Una qualsiasi rete di petri può essere rappresentata in riferimento a 3 matrici:

La matrice di ingresso I (Input) è una matrice avente tante righe quanti sono i posti nella rete e tante colonne quante sono le transizioni. Gli elemnenti della matrice sono interi positivi o nulli: $I \in M(|P|, |T|, \mathbb{N})$. L'elemento i_{ij} della matrice $I := (i_{ij})$ viene definito come il peso dell'arco che

collega il $i-$esimo posto alla $j-$esima transizione. Per cui questa matrice racchiude tutti i pesi degli archi entranti nelle transizioni.

Analogamente la matrice di uscita $O \in M(|P|, |T|, \mathbb{N})$ (Output), contiene tutti i pesi degli archi uscenti dalle transizioni. L'elemento o_{ij} della matrice $O := (o_{ij})$ è definito come il peso dell'arco uscente dalla transizione j in entrata al posto i.

La matrice di incidenza C è definita come la differenza tra la matrice di uscita e la matrice di entrata:

$$C := O - I \in M(|P|, |T|, \mathbb{Z})$$

Se nella rete è presente un ciclo, a gettoni costanti, le componenti della matrice di incidenza relative a quegli archi assumono valori nulli. Se un elemento c_{ij} della matrice di incidenza $C := (c_{ij})$ è negativo, allora la transizione t_j è in ingresso al posto p_i, se è positivo allora la transizione t_j è di uscita dal posto p_i, ed il modulo dell'elemento indica il peso dell'arco corrispondente.

Questa rappresentazione corrisponde ad un'analisi puramente strutturale di una rete, poiché si perdono le informazioni sui gettoni contenuti nei posti. Se ogni elemento della matrice di incidenza è nullo, allora la rete è strettamente, strutturalmente, conservativa, poiché è indipendente dalla marcatura M_0 della rete. Gli elementi della matrice di incidenza possono essere nulli in caso sia presente un ciclo o un autociclo nelle rispettive posizioni. Una rete senza autocicli si definisce pura.

Per determinare se una data transizione t_i è abilitata, si controlla se il vettore marcatura corrente M è maggiore uguale alla colonna $i-$esima della matrice di ingresso I_i, poiché quella colonna indica quanti gettoni devono essere consumati per lo scatto della transizione t_i e se la marcatura M è maggiore uguale a quella colonna, sono presenti necessariamente abbastanza gettoni per scattare la transizione t_i in tutti i posti collegati, rendendola abilitata:

$$t_i : \text{ abilitata } \iff M \geq I_i$$

La marcatura corrente M se viene sommata con la colonna $i-$esima della matrice di uscita risulta maggiore uguale di M prima della somma, poiché gli elementi della matrice O sono strettamente positivi, allora:

$$M + O_i \geq M \geq I_i \rightarrow M + O_i \geq I_i$$
$$M + O_i - I_i = M + C_i \geq 0$$

La marcatura M rappresenta lo stato della rete, mentre l'espressione $M+C_i$ rappresenta l'equazione di stato della rete, ovvero la sua evoluzione dopo lo scatto di una transizione t_i, se abilitata. Rappresenta una nuova marcatura raggiunta dalla rete aggiungendo e togliendo gettoni in base ai pesi degli archi entranti ed uscenti alla transizione t_i. L'evoluzione della rete non dipende dalla marcatura, ma dipende interamente dalla topo logia della stessa. Per ottenere la colonna $i-$esima della matrice di incidenza si considera il versore s_i, vettore di dimensione pari alla cardinalità dell'insieme delle transizioni dim $s = |T|$, avente tutti elementi nulli eccetto per l'elemento nella

posizione $i-$esima:

$$s_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} (1) \\ \vdots \\ (i) \\ \vdots \\ (|T|) \end{matrix}$$

Per cui si può esprimere una colonna corrispondente ad una transizione t_i attraverso il prodotto della matrice di incidenza per il versore associato a quella transizione:

$$C_i = C \cdot s_i$$

Una sequenza di scatti $S = t_{k_1}, \cdots, t_{k_n}$ abilitata in una marcatura iniziale M_0 è una sequenza di transizioni $t_{k_j} \in T,\ \forall j = 1, \cdots, n$ tali che la marcatura raggiunta dallo scatto di t_{k_j} da M_j porta ad una marcatura M_{j+1} dove la transizione $t_{k_{j+1}}$ è abilitata:

$$M_0 [\ t_{k_1} > M_1, \cdots, M_{n-1} [\ t_{k_n} > M_n \implies M_0 [\ t_{k_1}, \cdots, t_{k_n} = S > M_n$$

Una sequenza di transizioni S si dice sequenza di scatti solo se tutte le transizioni sono abilitate quando devono scattare, non necessariamente vero per una sequenza arbitraria. Se questo è verificato la sequenza di transizioni si dice ammissibile.

L'effetto di una sequenza di scatti $S = t_{k_1}, \cdots, t_{k_n}$ da una marcatura M_0 ad una marcatura M^* può essere espresso mediante l'equazione di stato della rete:

$$M^* = M + C_{k_1} + \cdots + C_{k_n} = M + C \cdot (s_{k_1} + \cdots + c_{k_n})$$

Si definisce il vettore colonna s_k, di dimensione pari alla cardinalità dell'insieme delle transizioni, associato alla sequenza di scatti, che indica quante volte ogni singola transizione è scattata, senza fornire informazioni sull'ordine in cui le transizioni scattano. Presenta nella posizione $i-$esima il valore corrispondente al numero di occorrenze della transizione t_i nella sequenza S:

$$s_{k_1} + \cdots + s_{k_n} = s_k$$

Per cui si può esprimere l'evoluzione della rete da una marcatura M_0 dopo una sequenza di scatti S come:

$$M_0 [\ S > M^* : \quad M^* = M_0 + C \cdot s_k$$

Quest'espressione indica una relazione lineare.

4.3 Analisi Strutturale

L'analisi della rappresentazione algebrica di una rete permette di trovare proprietà esclusivamente strutturali, intrinseche alla rete. Quest'analisi si basa solo su informazioni contenute nel grafo di incidenza. Questa analisi punta ad individuare delle strutture specifiche nella rete. Alcune di queste strutture si chiamano invarianti, possono essere di posto, $P-$invarianti, di transizione, $T-$invarianti.

Si dicono invarianti canonici, se il minimo comune multiplo dei loro elementi è pari ad uno, per trovare l'invariante più grande si sommano tutti gli invarianti più piccoli trovati. Si definisce il supporto di un invariante $||x||$, l'insieme di elementi associati alla rete corrispondenti ai componenti del vettore x. Un invariante si dice a supporto minimo se il suo supporto non contiente il supporto di altri invarianti. Generalemente quando si risolvono i sistemi associati agli invarianti, se è possibile scegliere, ai componenti si assegnano i valori nulli altrimenti 1, per trovare quelli a supporto minimo e canonici. Gli invarianti sia a supporto minimo che canonici formano una base dell'insieme degli invarianti della rete.

4.3.1 P-invarianti

Gli invarianti di posto sono delle strutture associate all'insieme dei posti P, dove la somma pesata dei gettoni è costante, si studiano per determinare la conservatività strutturale della rete. Un $P-$invariante di una rete N viene definito come un vettore colonna x di dimensione pari alla cardinalità dell'insieme dei posti $|P|$, tale che il prodotto matriciale tra la sua trasposta ed una qualsiasi marcatura M raggiungibile da M_0 è costante e finito:

$$x^T M = x^T M_0,\ \forall M \in R(N, M_0)$$

Per calcolare i $P-$invarianti si considera una generica marcatura raggiungibile M da una sequenza di scatti S, con associato un vettore s:

$$M = M_0 + Cs$$

Gli ivarianti di posto x sono tali che:

$$x^T M = x^T (M + Cs) = x^T M_0 + Csx^T$$

Per definizione $x^T M = x^T M_0$, allora, un vettore colonna x si dice $P-$invariante se rispetta la seguente equazione:

$$x^T Cs = 0$$

Per definizione un vettore associato ad una sequenza di scatti S non è nullo, per cui non si considera come soluzione $s = 0$. Per cui il l'equazione diventa un sistema di $|P|$ equazioni:

$$x^T C = 0 \rightarrow \begin{pmatrix} x_1 & \cdots & x_{|P|} \end{pmatrix} \cdot \begin{pmatrix} c_{11} & \cdots & c_{1|T|} \\ \vdots & \ddots & \vdots \\ c_{|P|1} & \cdots & c_{|P||T|} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow \begin{cases} x_1 c_{11} + \cdots + x_{|P|} c_{|P|1} = & 0 \\ \vdots & \vdots \\ x_1 c_{1|T|} + \cdots + x_{|P|} c_{|P||T|} = & 0 \end{cases}$$

Si omette la soluzione x uguale al vettore nullo, poiché inutile nell'analisi di una rete di Petri. Le soluzioni accettate sono vettori colonna con componenti intere. Se non sono presenti soluzioni oppure se le componenti dell'invariante sono minori o uguali a 0: $x \leq 0$, la rete non è strutturalmente conservativa, se è presente almeno una soluzione è possibile crearne infinite tramite combinazione lineare e la rete è strutturalmente conservativa. Se il $P-$invariante è tale che $x \geq 0$ la rete è conservativa rispetto a quel vettore peso, se è strettamente positivo, la rete è strutturalmente conservativa, se il vettore corrisponde al vettore identità, la rete si dice strettamente strutturalmente conservativa.

4.3.2 T-invarianti

I $T-$invarianti sono analoghi ai $P-$invarianti associati alle transizioni T. Un vettore colonna y, di dimensione pari alla cardinalità dell'insieme delle transizioni, si definisce $T-$invariante, se dopo lo scatto delle transizioni indicate dal vettore si ritorna alla marcatura iniziale, quindi esprime la reversibilità della rete. Si considera uno sequenza di scatti $S : M_0 [S > M_0$ associata al vettore y:

$$M = M_0 + Cy = M_0 \rightarrow Cy = 0$$

Non si considera $y = 0$ una soluzione, per cui un vettore colonna si definisce $T-$invariante se non è il vettore nullo e ha componenti interi. Per trovare le componenti del vettore y bisogna risolvare un sistema di $|T|$ equazioni:

$$Cy = 0 \rightarrow \begin{pmatrix} c_{11} & \cdots & c_{1|T|} \\ \vdots & \ddots & \vdots \\ c_{|P|1} & \cdots & c_{|P||T|} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_{|T|} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow \begin{cases} y_1 c_{11} + \cdots + y_{|T|} c_{1|T|} = & 0 \\ \vdots & \vdots \\ y_{|T|} c_{|P|1} + \cdots + y_{|T|} c_{|P||T|} = & 0 \end{cases}$$

Se il vettore y è maggiore uguale del vettore nullo, ovvero ha componenti tutte positive, la rete si dice reversibile, altrimenti se presenta almeno una componente negativa implicherebbe che per ritornare alla marcatura iniziale bisognerebbe invertire lo scatto di una transizione, ma ciò è impossibile, quindi in questo caso la rete non è reversibile. Se una ammete un $T-$invariante, ne ammette infiniti. Se non ammette $T-$invarianti sicuramente non è reversibile. Bisogna comunque verificare che la sequenza di scatti indicata dal $T-$invariante sia abilitata, per dimostrare che la rete è reversibile.

5 Modellazione di Un Sistema

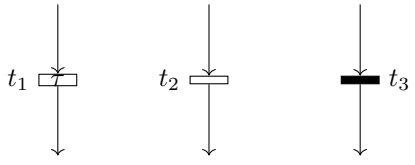
Si modellano sistemi produttivi per la loro semplicità, poiché sono facilmente intuibili. In questi sistemi un materiale grezzo subisce delle lavorazioni fino a diventare un prodotto finito. Per modellare un sistema vengono fornite le informazioni sui passaggi che il grezzo attraversa nel sistema, sulla base di queste informaizoni si modellano le varie parti.

In generale un sistema produttivo è formato da varie risorse, macchine per la lavorazione, dispositivi per lo spostamento dei materiali, e dei depositi dove conservare i materiali. Ogni passaggio effettuato dal grezzo richiede l'uso di almeno una risorsa disponibile, poiché possono essere impegnate da altri compiti. Le azioni associate a questi passaggi vengono espresse come gli eventi del sistema.

Bisogna tenere conto delle condizioni che permettono lo scatto di una transizione, ovvero l'accadimento di un evento nel sistema. Per cui tutte le risorse vengono rappresentate come dei cicli, poiché se una risorsa è limitata, il sistema produttivo si blocca, ed è necessario un modo per modellare una risorsa che non si esaurisce, ciò viene quindi rappresentato come un ciclo in una rete di Petri. L'uso di un ciclo permette anche di modellare la disponibilatà di una risorsa per effettuare una serie di eventi.

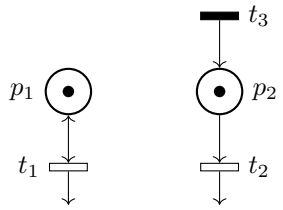
5.1 Elementi Fondamentali

Ad ogni evento viene associata una transizione temporizzata, che rappresenta la possibilità accada un evento, mentre gli stati di una risorsa sono associati a posti collegati alle rispettive transizioni. Gli stati associati ai posti forniscono informazioni sulla disponibilità di un evento e di una o più risorse, possono esprimere che un evento sta accadendo, informano il sistema quando l'attività è conclusa, e restituiscono le risorse usate per quell'evento al termine dell'attività. Ogni attività è rappresentata da un ciclo, per indicare la disponibilità delle risorse utilizzare, e per esprimere i passaggi che compie il grezzo tramite quelle risorse. Quindi per definire una serie di eventi, si segue il percorso di un generico materiale grezzo e si associano tutti gli eventi a rispettive transizioni, con pre-set e post-set determinati dal comportamento del sistema, per ogni possibile percorso che il materiale può attraversare nel sistema, senza saltare passaggi. In questo modo si analizzano algoritmicamente tutte le possibili interazioni tra le risorse presenti nel sistema. Quando un materiale esce dal sistema, non deve essere più considerato nella costruzione del modello. Ogni evento del sistema può richiedere un certo intervallo di tempo per il suo accadimento, per cui deveno essere associati a rispettive transizioni temporizzate. Per cui è necessario distinguere nella rete tra le transizioni istantanee e quelle temporizzate. Si indicano le transizioni temporizzate come dei blocchi vuoti, mentre quelle istantanee sono rappresentate come delle barrette. Se è noto l'intervallo di una transizione temporizzata τ, viene inserito dentro la transizione:

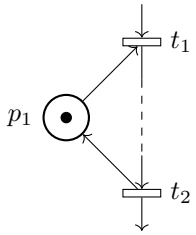


La transizione t_1 è temporizzata, con associato un intervallo di tempo non necessariamente noto, mentre la transizione t_2 è istantanea, ovvero l'evento associato avviene in un itervallo di tempo tendente a zero. Le marcature in un posto vengono consumate da una transizione solamente al termine dell'intervallo di tempo associato ad essa.

In caso alcune risorse sono sempre disponibili, possono essere rappresentate con un transizione istantanea appesa a monte, oppure con un autociclo, rappresentabile come una doppia freccia:

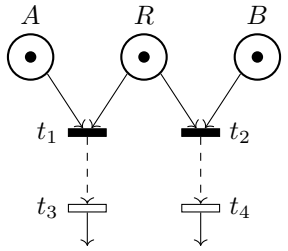


Entrambe queste rappresentazioni esprimono lo stessa risorsa sempre disponibile, ma in caso siano presenti transizioni appese nella rete, si perde la limitatezza, per cui è preferibile usare un autociclo. Altrimenti una risorsa può essere disponibile al passare di un intervallo di tempo finito, in questo caso si modella tramite una transizione temporizzata posta a monte. Se la risorsa non è capacitata ad 1, si può modellare come un autociclo di peso pari ai gettoni necessari marcato sufficientemente, oppure con una transizione a monte, ma ciò rende la rete illimitata. La disponibilità di una risorsa si esprime tramite un posto, che presenta in pre-set la transizione che determina la fine dell'uso di quella risorsa, mentre in post-set la transizione che determina l'inizio dell'uso della risorsa. Per cui lo stato del posto indica se la risorsa è disponibile, oppure è occupata da un'attività:



5.2 Conflitto Strutturale

Quando è presente una risorsa condivisa nel sistema, il posto associato può creare un conflitto strutturale in caso non sia modellato correttamente. Se il posto è in entrata a due transizioni temporizzate, poiché i gettoni vengono consumati solo dopo l'intervallo di tempo della transizione, è possibile che abiliti entrambi le transizioni ed entrambi lavorino usando lo stesso numero di gettoni, ciò non corrisopnde alla realtà. Poiché una volta che una risorsa è usata in processo, non può essere usata contemporaneamente in un altra attività. Per cui per rappresentare una risorsa condivisa, si usano delle transizioni istantane prima delle transizioni temporizzate, per indicare che la risorsa non è più disponibile, ed è impegnata da una delle attività, in questo modo l'altra transizione in entrata non è più abilitata. Questo conflitto diventa quindi solo potenziale:



Se non viene specificato, il conflitto non viene risolto, ovvero non si modella quale delle due o più transizioni ha la precedenza. Generalmente, se è presente una lavorazione in parallelo, con delle risorse condivise, è più efficiente richiedere la disponibilità della risorsa condivisa, all'inizio della lavorazione, in modo da non aspettare la sua disponibilità, e bloccare il resto del processo.

5.3 Esempio: Cella di Assemblatura

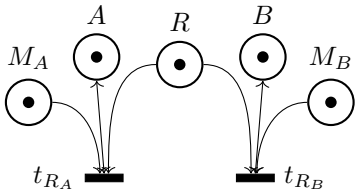
In generale le macchine come tutte le risorse reali sono capacitate, per cui vanno modellate come un ciclo, marcato. All'inizio di ogni modello, bisogna marcare i cicli di produzione e lavorazione, per permettere al sistema di operare.

Si considera una cella di assemblatura, dove due materiali grezzi A e B vengono inseriti nel sistema tramite rulli trasportatori. All'interno della cella è presente un braccio meccanico, capacitato ad uno, che può spostare i grezzi in due diverse macchine di lavorazione, una per ogni grezzo M_A e M_B, anch'esse capacitate ad uno. Quando entrambe le macchine hanno processato un grezzo, possono essere raccolte dal braccio meccanico ed assemblate in un prodotto finito. Prima che il robot posso traspostare altri materiali grezzi, deve spostare il prodotto finito in un buffer di output, capacitatato ad uno, prima che il prodotto possa essere traspostato all'esterno del sistema.

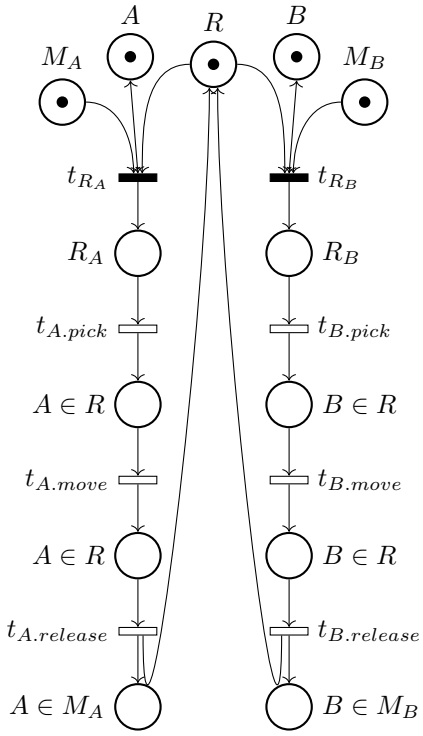
In questo caso ogni risorsa è capacitata ad uno, per cui il robot all'inizio può prendere uno dei due grezzi, per cui è presente un conflitto strutturale all'inizio. Si modella senza risolvere il conflitto. Se il robot prendere un grezzo quando entrambe le macchine hanno processato sono occupate con un grezzo il sistema si blocca, per cui bisogna modellare il sistema in modo da impedire questo "deadlock".

Si considerano i materiali grezzi sempre disponibili nel sistema, per cui si rappresentano come degli autocicli, legati a monte della transizione, istantanea, per attivare una delle due sequenza di lavorazione A e B. Gli eventi che avvengono processando uno dei due grezzi sono gli stessi tra di loro, per cui si può modellare uno solo di questi processi ed includere una sua copia per rappresentare l'altro, fino alla fine della lavorazione del grezzo nella macchina apposita. Quindi si rappresenta una rete, in parte simmetrica.

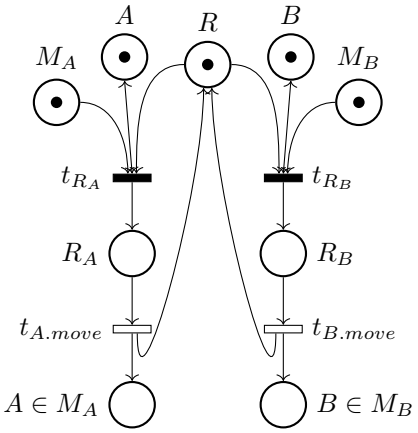
Se il robot è disponibile R, il grezzo è disponibile A/B, e la macchina è disponibile M_A/M_B, allora può cominciare la fase di lavorazione del grezzo. Si rappresenta come 3 posti a monte di una transizione istantanea t_{R_A}/t_{R_B}, poiché è presente un conflitto strutturale con l'altrs sequenza di lavorazione dell'altro grezzo:



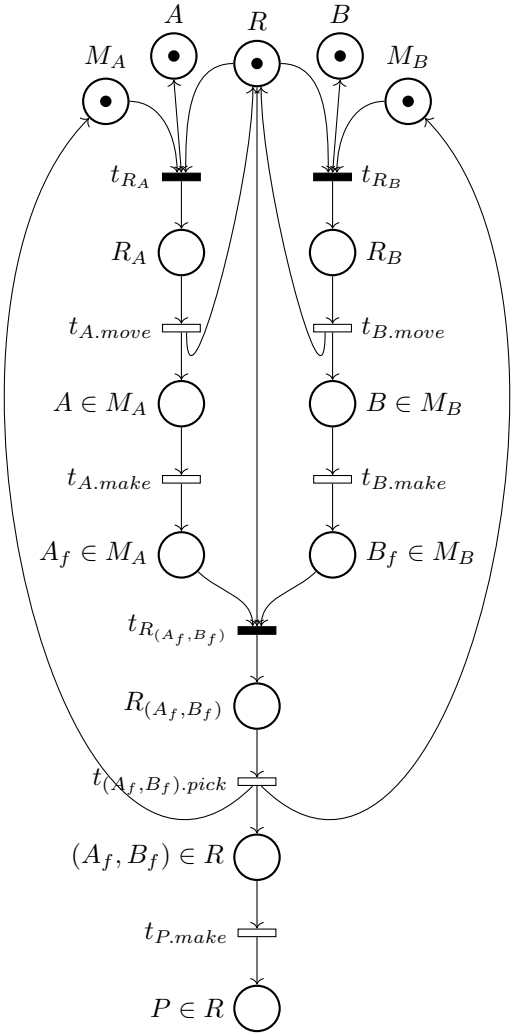
Le transizioni istantanee portano ad un posto R_A/R_B, che esprime lo stato del braccio meccanico, occupato nel carico $A/B.pick$, traporto $A/B.mov$ e scarico $A/B.release$ del grezzo A o B, e rappresentano il primo evento nella sequenza della lavorazione della macchina. Ognuna delle operazioni del braccio viene espressa come una transizione temporizzata, e solo al termine, ovvero allo scarico $t_{A/B.release}$, il braccio torna di nuovo disponibile allo spostamento di un altro grezzo.



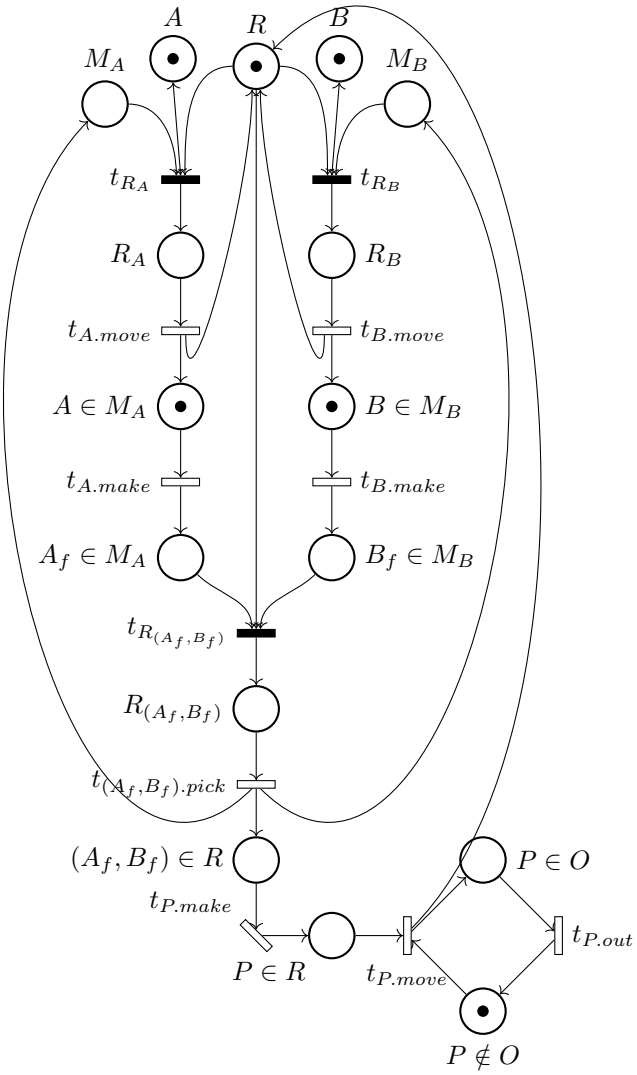
Le transizioni in sequenza $t_{A/B.pick}$, $t_{A/B.move}$, $t_{A/B.release}$ possono essere sostituite da un'unica transizione associata ad un intervallo di tempo pari alla somma dei tre intervalli delle transizioni originali:



Quando le macchine presenta al loro interno un grezzo, lo lavorano fino ad un prodotto finito, quest'operazione viene rappresentata come una transizione temporizzata $t_{A/B.make}$ che passa dallo stato $A/B \in M_{A/B}$ allo stato $A_f/B_f \in M_{A/B}$, entrambi espressi come posti. Quando entrambi i posti che rappresentano lo stato del prodotto finito in una macchina sono marcati, ed il braccio meccanico è disponibile, allora il braccio può raccogliere entrambi i prodotti finiti per cominciare il processo di assemblaggio finale. Quest'operazione si rappresenta mediante una transizione $t_{R_{(A_f,B_f)}}$ istantanea, poiché lo stesso braccio è in entrata ad altre due transizioni, ma il conflitto non avverrà mai con questa transizione, poiché può essere abilitata se e solo se le due transizioni in ingresso alle sequenze di lavorazione dei grezzi sono disabilitate. Questo perché in ogni ciclo di lavorazione di un grezzo è presente un singolo gettone, che abilita solo una transizione. Questa transizione porta ad uno stato $R_{(A_f,B_f)}$, ovvero la sequenza di assemblaggio finale. Il braccio meccanico quindi raccogli i due prodotti finiti da M_A e M_B, rendendoli nuovamente disponibili, e arriva ad uno stato $(A_f, B_f) \in R$, dove entrambi i prodotti finiti sono all'interno del robot munito di braccio meccanico.



A questo punto si rappresenta l'assemblaggio finale come una transizione temporizzata $t_{P.make}$, che porta allo stato $P \in R$, dove P denota il prodotto finale, contenuto ancora nel robot. Il prodotto finito viene riposto nel buffer di uscita O, se è disponibile, rendendo nuovamente disponibile il robot per spostare altri grezzi. Il prodotto finito nel buffer finale a questo punto viene trasportato fuori dal sistema, modellato con un altra transizione temporizzata $t_{P.out}$:

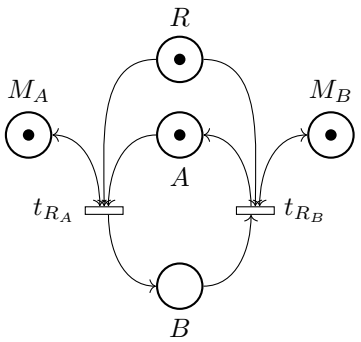


Lo stato iniziale del modello viene scelto arbitrariamente. Bisogna inserire gettoni nei cicli altrimenti non sarebbero abilitati. Si evita di partire da un conflitto effettivo, per cui si inseriscono

i gettoni in $M_{A/B}$. In questo caso tutte le risorse sono capacitate ad uno, per cui è necessario inserire un ungico gettone in un posto per ogni ciclo. Se si parte da un conflitto effettivo, se non viene specifiato, il simulatore sceglie arbitrariamente quale delle transizioni ha priorità sulle altre. Per com'è strutturata la rete, l'assemblaggio non è un conflitto effettivo, poiché la transizione che denota l'inizio della fase di assemblaggio è abilitata solo se le macchine A e B hanno finito il processamento dei grezzi, quindi non sono abilitate le transizioni in entrata a M_A/B.

5.4 Risoluzione di un Conflitto

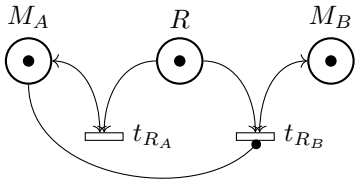
Per risolvere un conflitto effettivo, si può modellare una priorità tra le due transzioni in conflitto. Quando un conflitto viene risolto, non è necessario introdurre delle transizioni istantanee. Si modella la sequenzialità tra le due transizioni inserendo un ciclo controllore, che permette alle transizioni di scattare solamente una dopo l'altra. Questo controllore come ogni ciclo deve essere marcato:



In caso R abiliti solo queste due transizioni, si può omettere. Quando si modella l'alternanza, le transizioni possono scattare solo in questo ordine prestabilito, anche indipendentemente dalla disponibilità di $M_{A/B}$. Per cui questo metodo di risoluzione di un conflitto è inefficiente.

Si definisce un nuovo tipo di connesione tra posto e transizioni, detto arco inibitore, questo arco, uscente da un posto, si collega ad una transizione mediante una punta di freccia circolare. Se la marca presente nel posto è maggiore uguale al peso k dell'arco inibitore, allora la transizione collegata non può scattare, anche se è abilitata.

In questo caso la transizione non può scattare. In questo modo si può risolvere un conflitto tra due transizioni. Una transizione ha priorità sull'altra, quindi si include un arco inibitore, di peso pari al peso dell'arco entrante nella transizione che ha priorità. In questo modo solo se la prima transizione non è abilitata, la seconda è in grado di scattare:

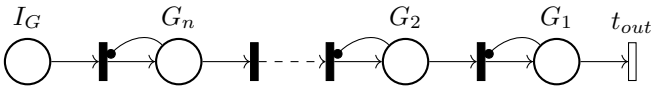


In questo modo non si spreca del tempo ad aspettare che la seconda transizione sia abilitata. Questo metodo di risoluzione è più efficiente rispetto alla priorità.

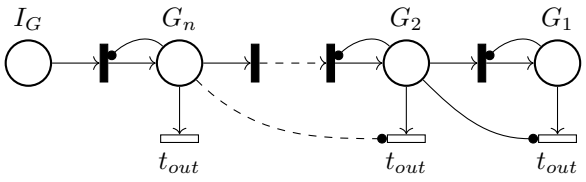
Se non viene esplicitamente richiesto, non si modella la risoluzione del conflitto, se invece si modella la risoluzione si usano direttamente le transizioni temporizzate.

5.5 Buffer FIFO e LIFO

In caso il sistema presenta una pila o una coda, bisogna modellare un magazzino con disciplina FIFO o LIFO. Si modella un buffer nella convenzione First In First Out, tramite archi inibitori, dove la transizione t_{out} determina l'uscita dell'elemento dal buffer. Si considera un buffer di n elementi:



Si modella ora un buffer LIFO, di n elementi:

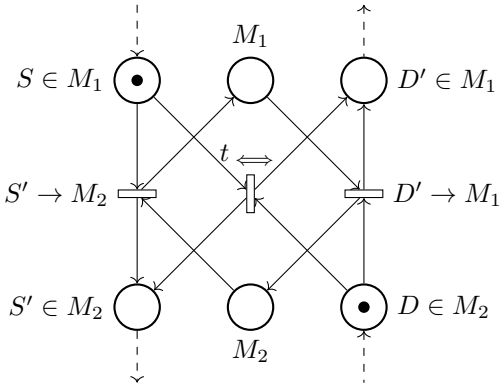


5.6 Diagramma di Gant

Il grafico di Gant è una rappresentazione grafica dell'uso delle risorse in un sistema. Ogni risorsa si indica come una linea parallela. Si scorrono verso destra, rispetto al tempo, e viene associato un evento in un determianto tempo per ogni scatto di una o più transizioni. Si può identificare nel grafico di Gant quando è presente un conflitto, ovvero quando una stessa risorsa è richiesta per due, o più, processi diversi. Il grafico di Gant è completamente indipendente dalla rete associata ad un sistema. Attraverso questa visualizzazione, si possono individuare tutti i conflitti effettivi della rete, quindi si possono modellare adeguatamente le risoluzioni, se sono richieste. La precedenza nel grafico di Gant viene scelta arbitrariamente. Questa rappresentazione dipende dalla marcatura iniziale del sistema, in caso sia presente un deadlock, la rappresentazione non è periodica. Generalmente un grafico di Gant di un sistema produttivo ciclico, è periodico, per cui è sufficiente rappresentare solo un periodo del sistema.

5.7 Scambio

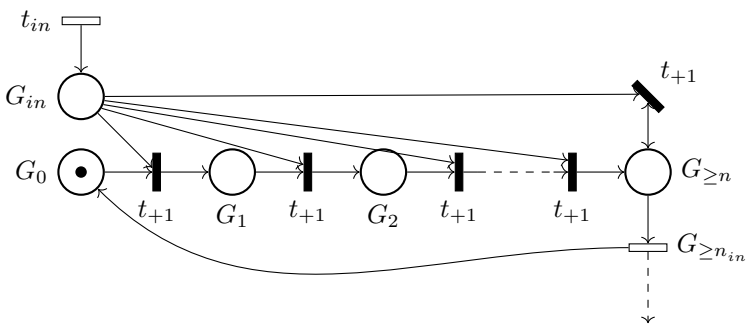
Un sistema può arrivare ad una situazione dove due processi sono bloccati, poiché richiedono alcune risorse usate nell'altro processo. Se il sistema si blocca se incontra questa situazione, è necessario modellare un metodo in modo in cui il sistema posso continuare ad operare, anche se ragiunta questa situazione. Per modellarlo, bisogna dare priorità ad uno dei processi, e si evita di richiedere la disponibilità della risorsa scambiandola con l'altro processo, e riportando quel processo ad uno stato precedente all'uso della risorsa.



In questo esempio è presente una lavorazione in parallelo di due materiali S e D, con delle macchine M_1 e M_2. Nella sequenza di destra, è stata usata la risorsa M_1, e per continuare la lavorazione è necessaria la risorsa M_2, impegnata nella lavorazione di destra. Le risorse M_1 ed M_2, sono entrambe non disponibili per entrambe le sequenza, per cui ci si trova in un deadlock. Si prioritizza la lavorazione di sinistra, per cui si inserisce una transizione $t \iff$ che evita la richiesta delle due risorse entrambe occupate. Dando priorità alla lavorazione di sinistra, il processo di destra, arriva da $D \in M_2$ ad uno stato $D' \in M_1$ senza dover richiedere la risorsa M_1 tramite la transizione $D' \to M_1$. Mentre il processo di sinistra, arriva da $S \in M_2$ ad uno stato $S' \in M_1$ dove ha già richiesto ed ottenuto la risorsa M_1, senza lo scatto della transizione $S' \to M_2$, effettivamente scambiando la risorsa tra i due processi.

5.8 Contatore

In alcuni sistemi, un processo può essere attivato solo dopo avere almeno n elementi disponibili. Per modellarlo è necessario un contatore, che si aggiorni per ogni nuovo elemento, fino alla raggiunta degli n elementi necessari. Per modellare questo contatore, sono necessari $n+1$ posti, che rappresentano il numero di elementi disponibili dal posto G_0, che indica non sono presenti elementi, fino al posto $G_{\geq n}$, che indica sono presenti almeno n elementi. Inoltre è necessario un qualche tipo di generatore, per poter modellare l'entrata degli elementi nel sistema, ogni intervallo di tempo. Poiché se fossero sempre disponibili non sarebbe necessario un contatore.


t_{in}
G_{in}
t_{+1}
G_0
$G_{\geq n}$
t_{+1}
G_1
t_{+1}
G_2
t_{+1}
t_{+1}
$G_{\geq n_{in}}$

6 Reti di Code

6.1 Nozioni di Statistica

Prima di descrivere il modello matematico delle reti di Code, è necessario introdurre concetti della statistica usati nel modello. Gli eventi considerati dai modelli sono associati ad un'alea, poiché non è noto a priori come si verica. Si definisce lo spazio campione Ω l'insieme di tutti gli eventi ω che possono accadere nel sistema. Si definisce la funzione probabilità p di un evento ω appartenente allo spazio campione come il rapporto tra il numero dei casi favorevoli ed il numero dei casi totali:

$$p : \Omega \to [0, 1]$$
$$p(\omega) = \frac{n_\omega}{n_\Omega}$$

Poiché gli eventi sono assocati ad un alea, si descrivono mediante una variabile aleatoria, definita come una funzione arbitraria applicata sullo spazio campione, che restituisce un reale:

$$X : \Omega \to \mathbb{R}$$

Questa variabile può essere sia continua che discreta, la cui analisi è più semplice rispetto ad una varibile continua. Si definisce la distribuzione di una variabile aleatoria, una funzione che indica quanto la variabile è uguale o minore di un certo valore:

$$f_x : X \to [0, 1]$$

Si definisce la densità di probabilità di un evento, la probabilità con cui la varibile aleatoria assume un certo valore, nel discreto corrisponde ad una somma di probabilità:

$$F_x(x) = \sum_{y \leq x} p_x(y)$$
$$F_x(x) = \int_{-\infty}^{x} f_x(y)dy$$

Si definisce la media d'insieme la media di tutti i valori di uno spazio campione:

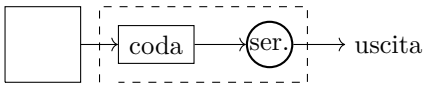
$$n := \frac{\sum_{\omega \in \Omega} \omega}{\dim \Omega}$$

Si definisce invece il valore aspettato di un insieme la media pesata dei valori dell'insieme:

$$N := \sum_{\omega \in \Omega} \omega p_\omega$$

6.2 Teoria delle Code

Si definisce coda di un sistema, lo spazio finito di attesa e l'insieme dei serventi. Un sistema che presenta una coda, oltre alla coda contiene una sorgente, da dove arrivano i clienti, ed un'uscita, non modellata, da dove escono:



Il sistema di code viene gestito tramite disciplina FIFO. Dato un sistema, si definiscono variaibli esogene, variabili esterne al sista, ovvero segnali di input, mentre si definiscono variabili endogene, variaibli interne al sistema, ovvero variabili di stato.

I sistemi di code vengono analizzati dal punto di vista temporale, per determinare se è conveniente usufruire del servizio è necessario individuare il tempo medio di attesa, ed il tempo medio di servizio, poiché il tempo di attesa deve essere proporzionale alla richiesta. Quest'analisi si interessa solamente di individuare le grandezze di interessa, non di svolgere un'analisi su di esse.

In queste analisi il tempo, è una variabile aleatoria continua. Si definisce una serie di grandezze:

- t_a: Tempo d'arrivo, il tempo tra due entrate diverse dentro al sistema;
- t_s: Tempo di servizio, indica il tempo in cui un servente è impegnato da un cliente;
- t_q: Tempo di attesa nella coda di un cliente, prima di essere servito;
- t_w: Tempo speso da un cliente all'interno del sistema;
- s: Numero di serventi nel sistema;
- n: Numero totale di clienti nel sistema;
- Disciplina di serizio (FIFO);
- Dimensione della popolazione di serventi;
- Comportamento del cliente dopo il servizio;
- l: Lunghezza della coda.

Queste variaibli non sono tutte indipendenti, si può notare facilmente come il tempo speso nel sistema, corrisponde alla somma tra il tempo di attesa ed il tempo di servizio:

$$t_w = t_a + t_s \tag{6.2.1}$$

Analogamente la lunahezza della coda, è dato dalla differenza tra il numero dei clienti totali nel sistema meno la popolazione dei serventi, solo quando il numero dei clienti è minore dei serventi totali presenti nel sistema:

$$l = \begin{cases} n - s & n \geq s \\ 0 & n < 0 \end{cases} \tag{6.2.2}$$

Per lavorare con le variabili aleatorie, non si considera l'intera funzione che definisce l'alea, si usano la varianza, la differenza dal valore medio, e la media, il valore medio assunto dalla distribuzione. Si considerano per semplicità distribuzioni esponenziali per il tempo di attesa t_a e per il tempo di servizio t_s, dove la media e la varianza assumono lo stesso valore. Si definiscono quindi la frequenza degli arrivi λ, e la velocità di servizio μ:

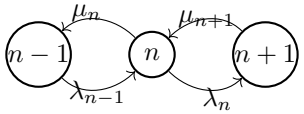
$$\lambda = \frac{1}{E\{t_a\}} \tag{6.2.3}$$

$$\mu = \frac{1}{E\{t_s\}} \tag{6.2.4}$$

Mentre il fattore $E\{\cdot\}$ indica il valore atteso della distribuzione. Lavorando con distribuzioni note, si semplifica il calcolo della probabilità P_n.

6.3 Sistemi di Nascita e Morte

Nell successive analisi si considerano solamente sistemi di nascita e morte, ovvero sistemi dove non avvengon osalti multipli, solo salti tra stati adiacenti tra di loro. La presenza di salti multipli dipende dall'intervallo di osservazione, per cui utilizzando un intervallo di osservazione Δt infinitesimo, si può verificare solamente un entrata o un uscita per ogni istante Δt. Per cui si possono verificare solo i seguenti salti, adiacenti in uno stato n:



Dove μ_n e λ_n indicano la velocità di servizio e la frequenza di arrivo allo stato n. Si vogliono considerare tutti i possibili eventi che possono avvenire in un singolo istante $\Delta t \to 0$, partendo da un tempo t in uno stato n, la variabile aleatoria discreta. L'obiettivo finale di quest'analisi è individuare il valore attesto della distribuzione di probabilità $P_n(t)$. Si considerano tutti i possibili modi di arrivare allo stato n:

Stato in t	Evento in Δt	Probabilità P_n
$n-1$	Nascita	$P_{n-1}(t)\lambda_{n-1}\Delta + o(\Delta t)$
$n+1$	Morte	$P_{n+1}(t)\mu n + 1\Delta t + o(\Delta t)$
?	Salti multipli	$o(\Delta t)$
n	Nessun salto	$P_n\left[1 - \lambda_n\Delta t - \mu_n\Delta t\right] + o(\Delta t)$

Per ricordare che il tempo di osservazione è infinitesimo, si inserisce l'elemento $o(\Delta t)$, infinitesimo di ordine superiore a Δt, che tende a 0 più velocemente di Δt. La probabilità di che un evento rimanga allo stato n, corrisponde alla probabilità che non avvenga nessun tipo di salto da n. Si esprime quindi la probabilità di essere allo stato n, ovvero la somma delle probabilità di una nascita

dallo stato $n-1$, di una morte da $n+1$, di salti multipli e che non avvenga nulla allo stato n:

$$P_n(t+\Delta t) = P_{n-1}(t)\lambda_{n-1}\Delta t + P_{n+1}(t)\mu_{m+1}\Delta t + P_n(t)[1-(\lambda_n+\mu_n)\Delta t] + o(\Delta t)$$
$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{m+1} - P_n(\lambda_n+\mu_n) + \frac{0(\Delta t)}{\Delta t}$$

Si applica il limite $\Delta t \to 0$, poiché non sono ammessi nel sistema salti multipli:

$$\lim_{\Delta\to 0}\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \frac{dP_n(t)}{dt} = P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{m+1} - P_n(\lambda_n+\mu_n) + \cancelto{0}{\frac{0(\Delta t)}{\Delta t}}$$

Si considerano sistemi stazionari, con una risposta a regime permanente, non nulla e costante. Ovvero sistemi asintoticamente stabili, dove il transitorio tende a zero. Per cui per tempi abbastanza grandi, la probabilità assume valori costanti:

$$\lim_{t\to\infty} P_n(t) = P_n$$

Assumendo valori costanti, la sua derivata è nulla, perciò:

$$0 = P_{n-1}\lambda_{n-1} + P_{n+1}\mu_{m+1} - P_n(\lambda_n+\mu_n)$$

Si dimostra la relazione ottenuta da quest'equazione tramite induzione. Si considera il passo induttivo per $n=0$, poiché non puà esistere uno stato $n<0$, non può esistere la probabilità $P_{n<0}$:

$$P_1\mu_1 - P_0\lambda_0 - P_0\mu_0 = 0$$

Analogamente non si può ritornare ad uno stato $n<0$, la velocità di servizio per lo stato $n=0$ è nulla:

$$P_1\mu_1 - P_0\lambda_0 = 0$$
$$P_1 = \frac{\lambda_0}{\mu_1}P_0$$

Per uno stato $n=1$ si ottiene la seguente equazione:

$$P_0\lambda_0 + P_2\mu_2 - P_1\lambda_1 - P_1\mu_1 = 0$$

Considerando la relazione ottenuta precedentemenre si ottiene:

$$P_2\mu_2 - P_1\lambda_1 = 0$$
$$P_2 = \frac{\lambda_1}{\mu_2}P_1 = \frac{\lambda_1\lambda_0}{\mu_2\mu_1}P_0$$

Per cui per uno stato n generico:

$$P_n\mu_n - P_{n-1}\lambda_{n-1} = 0$$
$$P_n = \frac{\lambda_{n-1}}{\mu_n}P_{n-1}$$

Si ottiene quindi la seguente formula generale per ottenere la distribuzione di probabilità per uno stato n:

$$P_n = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} P_0 \tag{6.3.1}$$

Essendo una distribuzione di probabilità, la somma di tutte le probabilità di tutti i possibili stati deve assumere valore unitario:

$$\sum_{n=0}^{\infty} P_n = 1$$

$$P_0 + \sum_{n=1}^{\infty} P_n = 1$$

$$P_0 + \sum_{n=1}^{\infty} \left[\frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} P_0 \right] = 1$$

La probabilità P_0 assume valore costante, per cui è possibile esprimerla come:

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \left[\frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} \right]} \tag{6.3.2}$$

6.4 Sistemi M/M/1

Si considera un sistema dove i tempi di arrivo e di attesa sono distribuiti esponenzialmente, con un singolo servente, ed una coda illimitata. Tramite la notazione di Kendal si esprime come: $M/M/1$. Poiché i tempi sono distribuzioni note, si possono espimrere la frequenza d'arrivo e la velocità di servizio come:

$$\begin{cases} \lambda_n = \lambda & \forall n \in [0, +\infty] \\ \mu_n = \mu & \forall n \in [1, +\infty] \end{cases}$$

La produttoria diventa quindi:

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \left[\frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} \right]} = \frac{1}{1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu} \right)^n}$$

Si considera la serie geometrica di ragione $\lambda/\mu < 1$:

$$\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^n = 1 + \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^n = \frac{1}{1 - \frac{\lambda}{\mu}}$$

La probabilità che il sistema contenga 0 clienti è quindi:

$$P_0 = \frac{1}{\frac{1}{1 - \frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu}$$

Per cui la probabilità che il sistema stia lavorando, ovvero abbia più di 0 clienti si ottiene come il complementare della relazione precedente:

$$1 - P_0 = \frac{\lambda}{\mu} = \rho$$

Questo fattore ρ si chiama fattore di utilizzo del servente.

Data la probabilità 0, si esprimre la probabilità per uno stato generico n:

$$P_n = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} (1 - \rho) = \rho^n (1 - \rho)$$

Si calcola ora il valore atteso N di questa distribuzione P_n, ovvero la media dei clienti presenti nel sistema in un qualunque istante:

$$N = \sum_{n=0}^{\infty} n P_n = \sum_{n=0}^{\infty} n \rho^n P_0$$

Si porta un fattore ρ ed il fattore P_0, costanti, fuori dalla sommatoria, e si ottiene la derivata di ρ^n dentro la sommatoria:

$$N = \rho P_0 \sum_{n=0}^{\infty} n \rho^{n-1} = \rho P_0 \sum_{n=0}^{\infty} \frac{d}{d\rho} \rho^n$$

Poiché la derivata è un'operazione lineare, si può portare fuori dalla sommatoria, una serie geometrica di ragione $\rho < 1$, per cui converge:

$$N = \rho P_0 \sum_{n=0}^{\infty} \frac{d}{d\rho}\rho^n = \rho P_0 \frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) = \rho P_0 \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}$$

$$N = \frac{\lambda}{\mu - \lambda} \tag{6.4.1}$$

Questa relazione corrisponde ad una delle tre leggi di Little. Per determinare le altre due leggi si considera il valore atteso del tempo di attraversamento del sistema W, pari alla somma del valore atteso del tempo di servizio ed il tempo speso in coda:

$$W = W_q + W_s \tag{6.4.2}$$

Il tempo di attesa nella coda dipende dal sistema, mentre il valore atteso del tempo di servizio corrisponde al reciproco della velocità di servizio:

$$W = W_q + \frac{1}{\mu}$$

Il tempo medio in cui un cliente sta nel sistema è direttamente proporzionale al numero di clienti nel sistema. Il valore atteso dei clienti nel sistema corrisponde alla frequenza di arrivo moltiplicata per il valore atteso del tempo di attraversamento del sistema, poiché il numero di clienti in coda è nullo in caso la frequenza di arrivo sia nulla, indipendentemente dal tempo di attraversamento del sistema:

$$N = \lambda W$$

$$W = \frac{N}{\lambda} = W_q + \frac{1}{\mu}$$

$$W_q = \frac{1}{\mu - \lambda} - \frac{1}{\mu} \tag{6.4.3}$$

L'ultima legge di Little lega il valore atteso della lunghezza della coda l del sistema al valore atteso del tempo passato in coda t_q, ovvero mette in relazione diretta il numero dei clienti in coda L con il tempo speso in coda W_q. Questo valore atteso dipende solo dalla velocità degli arrivi in coda, per cui si può esprimere come la frequenza di arrivo moltiplicata per il tempo madio passato in coda:

$$L = \lambda W_q = \lambda\left(\frac{1}{\mu - \lambda} - \frac{1}{\mu}\right) = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu}$$

$$L = N - \rho \tag{6.4.4}$$

6.5 Sistemi M/M/1/K

Si considera un sistema a coda con un massimo di k clienti, per cui la coda ha una lunghezza massima di $k-1$, ovvero il numero massimo di clienti nel sistema, meno il numero di serventi del sistema. Come nell'analisi precedente li tempo di arrivo e di servizio sono distribuiti esponenzialmente con parametro caratteristico γ e μ, ma bisogna considerare quando non può più accettare clienti, quando arriva allo stato $n = k$. La velocità di servizio è indipendente dal numero di clienti nella coda per cui si esprime come, ma è comunque definita per ogni stato accessibile del sistema:

$$\mu_n = \mu \, \forall n = 1, \cdots, k$$

Mentre la frequenza di arrivo si annulla quando all'interno del sistema sono presenti esattamente k clienti:

$$\lambda_n = \begin{cases} \gamma & 0 \leq n < k \\ 0 & n \geq k \end{cases}$$

Quando il sistema può accettare clienti, la frequenza di arrivo è distribuita esponenzialmente con parametro caratteristico γ, mentre quando non può accettare più clienti la frequenza si annulla. Per cui la frequenza γ è una frequenza più grande che descrive la frequenza di arrivo dall'esterno, indipendentemente dallo stato del sistema, mentre la frequenza λ di arrivo effettivo è minore, poiché da un valore $n = k$ si annulla. Dato che la frequenza γ non considera quando può lavorare il servente, consdiera anche i clienti che vengono rifiutati dal sistema.

Si considera l'ipotesi di funzionamento dei sistemi stazionari, ma si include questa distinzione della frequeza di arrivo:

$$P_n = \begin{cases} \dfrac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} = \left(\dfrac{\gamma}{\mu}\right)^n & n \leq k \\ 0 & n > k \end{cases}$$

Per cui la probabilità che il sistema si trovi allo stato $n = 0$ si ottiene tramite la seguente espressione:

$$P_0 = \frac{1}{1 + \sum_{n=0}^{k} \left(\dfrac{\gamma}{\mu}\right)^n} = \frac{1}{\sum_{n=0}^{k} \left(\dfrac{\gamma}{\mu}\right)^n}$$

Poiché la sommatoria si annulla per valori $n > k$, per cui non è necessario dimostrare la convergenza con $\gamma/\mu < 1$ poiché il sistema è limitato:

$$\sum_{n=0}^{\infty} \left(\frac{\gamma}{\mu}\right)^n = 1 + \cdots + \left(\frac{\gamma}{\mu}\right)^k + 0 + \cdots$$

Questa sommatoria si può esprimere mediante la differenza tra due serie geometriche:

$$\sum_{n=0}^{k}\left(\frac{\gamma}{\mu}\right)^n = \sum_{n=0}^{\infty}\left(\frac{\gamma}{\mu}\right)^n - \sum_{n=k+1}^{\infty}\left(\frac{\gamma}{\mu}\right)^n$$

$$l = n - (k+1)$$

$$\frac{1}{1-\frac{\gamma}{\mu}} - \sum_{l=0}^{\infty}\left(\frac{\gamma}{\mu}\right)^{l+k+1} = \frac{1}{1-\frac{\gamma}{\mu}} - \left(\frac{\gamma}{\mu}\right)^{k+1}\frac{1}{1-\frac{\gamma}{\mu}}$$

$$\sum_{n=0}^{k}\left(\frac{\gamma}{\mu}\right)^n = \frac{1}{1-\frac{\gamma}{\mu}}\left(1-\left(\frac{\gamma}{\mu}\right)^{k+1}\right)$$

Per cui la probabilità P_0 corrisponde si può esprimere come:

$$P_0 = \frac{1-\frac{\gamma}{\mu}}{1-\left(\frac{\gamma}{\mu}\right)^{k+1}} \tag{6.5.1}$$

Si vuole determinare ora il valore atteso N del numero di clienti n nel sistema:

$$N = \sum_{n=0}^{k} nP_n = \sum_{n=0}^{k} n\left(\frac{\gamma}{\mu}\right)^n P_0 = P_0\frac{\gamma}{\mu}\sum_{n=0}^{k} n\left(\frac{\gamma}{\mu}\right)^{n-1} = P_0\frac{\gamma}{\mu}\sum_{n=0}^{k}\frac{d}{d\frac{\gamma}{\mu}}\left[\left(\frac{\gamma}{\mu}\right)^n\right]$$

Estraendo la derivata dalla sommatoria si ottiene la stessa sommatoria precedente, sostituendo il risultato precedentemente ottenuto diventa:

$$N = P_0\frac{\gamma}{\mu}\frac{d}{d\frac{\gamma}{\mu}}\sum_{n=0}^{k}\left(\frac{\gamma}{\mu}\right)^n = P_0\frac{\gamma}{\mu}\frac{d}{d\frac{\gamma}{\mu}}\left[\frac{1-\left(\frac{\gamma}{\mu}\right)^{k+1}}{1-\frac{\gamma}{\mu}}\right]$$

$$P_0\frac{\gamma}{\mu}\frac{d}{d\frac{\gamma}{\mu}}\left[\frac{1}{1-\frac{\gamma}{\mu}} - \frac{\left(\frac{\gamma}{\mu}\right)^{k+1}}{1-\frac{\gamma}{\mu}}\right] = P_0\frac{\gamma}{\mu}\left(\frac{1}{\left(1-\frac{\gamma}{\mu}\right)^2} - \frac{(k+1)\left(\frac{\gamma}{\mu}\right)^k - \left(\frac{\gamma}{\mu}\right)^{k+1}}{\left(1-\frac{\gamma}{\mu}\right)^2}\right)$$

$$P_0\frac{\gamma}{\mu}\left(\frac{1}{\left(1-\frac{\gamma}{\mu}\right)^2} - \left(\frac{\gamma}{\mu}\right)^k\frac{(k+1)\left(1-\frac{\gamma}{\mu}\right)}{\left(1-\frac{\gamma}{\mu}\right)^2}\right) = P_0\frac{\gamma}{\mu}\left(\frac{1-\left(\frac{\gamma}{\mu}\right)^{k+1}}{\left(1-\frac{\gamma}{\mu}\right)^2} - (k+1)\frac{\left(\frac{\gamma}{\mu}\right)^k}{\left(1-\frac{\gamma}{\mu}\right)^2}\right)$$

Sostituendo il valore di P_0 ottenuto precedentemente si ottiene:

$$N = \frac{\frac{\gamma}{\mu}}{1 - \frac{\gamma}{\mu}} - (k+1)\frac{\left(\frac{\gamma}{\mu}\right)^{k+1}}{1 - \left(\frac{\gamma}{\mu}\right)^{k+1}} \tag{6.5.2}$$

Si vuole esprimere la frequenza degli arrivi λ rispetto alla frequenza effettiva γ tramite un fattore di perdita $1 - \varepsilon$, che indica quale percentuale degli arrivi sono rifiutati dal sistema.

$$\lambda = \varepsilon\gamma$$

La frequenza degli arrivi corrisponde alla velocità di servizio moltiplicata per il fattore di utilizzo del servente ρ, esprimibile come il complementare di P_0:

$$\lambda = \mu\rho = \mu(1 - P_0) = \mu\left[1 - \frac{1 - \frac{\gamma}{\mu}}{1 - \left(\frac{\gamma}{\mu}\right)^{k+1}}\right] = \mu\left[\frac{\frac{\gamma}{\mu} - \left(\frac{\gamma}{\mu}\right)^{k+1}}{1 - \left(\frac{\gamma}{\mu}\right)^{k+1}}\right]$$

$$\lambda = \mu\frac{\gamma}{\mu}\left[\frac{1 - \left(\frac{\gamma}{\mu}\right)^{k}}{1 - \left(\frac{\gamma}{\mu}\right)^{k+1}}\right] = \gamma\varepsilon$$

Per cui la percentuale di clienti persi per un sistema $M/M/1/k$ si esprime tramite la seguente espressione:

$$\varepsilon = \frac{1 - \left(\frac{\gamma}{\mu}\right)^{k}}{1 - \left(\frac{\gamma}{\mu}\right)^{k+1}} < 1 \tag{6.5.3}$$

Si perdono dei clienti ogni volta che ne arriva uno nuovo nello stato $n = k$. Da questo relazione si può ottenere la probabilità che il sistema si trovi nello stato $n = k$, ovvero il fattore di perdita del sistema:

$$P_k = 1 - \varepsilon \tag{6.5.4}$$

Poiché la percentuale dei clienti persi corrisponde alla probabilità che il sistema si trova in uno stato dove non può accettare clienti $n = k$.

6.6 Sistemi M/M/s

Questo tipo di sistemi presenta una distribuzione esponenziale del tempo di arrivo e di attesa. Il sistema è illimitato, ma in questo caso è presente un numero s di serventi. Questi serventi si

considerano tutti uguali, e possono lavorare in parallelo tra di loro, per cui l velocità di servizio dipende dal numero di serventi occupati, quindi dallo stato n. La velocità effettiva μ_n corrisponde a n volte la velocità μ, quando sono presenti meno di s clienti, altrimenti il sistema lavora ad una velocità massima equivalente al prodotto tra il numero dei serventi e la velocità μ:

$$\mu_n = \begin{cases} n\mu & n < s \\ s\mu & n \geq s \end{cases}$$

La cosa è illimitata, per cui deve essere verificata la condizione di stazionarietà:

$$\lambda < s\mu$$

Si considera la probabilità che non sia presente nessun cliente nel sistema:

$$P_0 = \frac{1}{\displaystyle\sum_{n=0}^{\infty} \left[\frac{\displaystyle\prod_{i=0}^{n-1} \lambda_i}{\displaystyle\prod_{j=1}^{n} \mu_j} \right]}$$

Si considera solamente il denominatore del denominatore, poiché solo la velocità di servizio dipende dallo stato:

$$\sum_{n=0}^{\infty} \prod_{0=1}^{n} \mu_j = \sum_{n=0}^{s-1} \prod_{j=1}^{n} \mu_j + \sum_{n=s}^{\infty} \prod_{j=1}^{n} \mu_j$$

$$0 + 1\mu + \cdots + (s-1)!\mu^{s-1} + s!\mu^s + (s!\mu^s)s\mu + (s!\mu^s)(s\mu)^2 + \cdots$$

$$\sum_{n=0}^{s-1} n!\mu^n + \sum_{n=s}^{\infty} s!s^{n-s}\mu^n$$

$$\sum_{n=0}^{\infty} \left(\frac{\displaystyle\prod_{i=0}^{n-1} \lambda_i}{\displaystyle\prod_{j=1}^{n} \lambda_j} \right) = \sum_{n=0}^{s-1} \frac{1}{s!} \left(\frac{\lambda}{\mu} \right)^n + \sum_{n=s}^{\infty} \frac{1}{s!s^{n-s}} \left(\frac{\lambda}{\mu} \right)^n$$

$$\sum_{n=s}^{\infty} \frac{1}{s!s^{n-s}} \left(\frac{\lambda}{\mu} \right)^n \text{ per } n - s = l \rightarrow \frac{1}{s!} \sum_{l=0}^{\infty} \frac{1}{s^l} \left(\frac{\lambda}{\mu} \right)^{l+s}$$

$$\frac{1}{s!} \left(\frac{\lambda}{\mu} \right)^s \sum_{l=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^l = \frac{1}{s!} \left(\frac{\lambda}{\mu} \right)^s \frac{s\mu}{s\mu - \lambda}$$

Per cui la probabilità si epsimere tramite la esguente espressione:

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu-\lambda}} \tag{6.6.1}$$

Si esprime la probabilità di trovarsi in uno stato n, in base al fatto se tutti i serventi sitano lavorando o solo un numero $n < s$:

$$P_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & n < s \\ \frac{1}{s!}\frac{1}{s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_0 & n \geq s \end{cases}$$

La probabilità di trovarsi allo stato s può essere calcolata analogamente con entrambe le espressioni:

$$\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_0 = \frac{1}{s!}\frac{1}{s^{s-s}}\left(\frac{\lambda}{\mu}\right)^s P_0$$

Si vuole determinare il valore atteso della lunghezza della coda:

$$L = \sum_{l=0}^{\infty} lP_l = \sum_{n=s}^{\infty}(n-s)P_{n-s}$$

La probabilità di trovarsi allo stato P_{n-s} coincide con la probabilità che nel sistema siano presenti n clienti P_n. Per cui si applica un'altra sostituzione $k = n - s$:

$$\sum_{k=0}^{\infty} kP_{k+s} = \sum_{k=0}^{\infty} k\frac{1}{s!}\frac{1}{s^k}\left(\frac{\lambda}{\mu}\right)^{k+s} P_0 = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_0 \sum_{k=0}^{\infty} k\left(\frac{\lambda}{s\mu}\right)^k$$

$$\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} P_0 \sum_{k=0}^{\infty} k\left(\frac{\lambda}{s\mu}\right)^{k-1} = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} P_0 \sum_{k=0}^{\infty} \frac{d}{d\frac{\lambda}{s\mu}}\left(\frac{\lambda}{s\mu}\right)^k = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} P_0 \frac{d}{d\frac{\lambda}{s\mu}} \sum_{k=0}^{\infty}\left(\frac{\lambda}{s\mu}\right)^k$$

$$\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} P_0 \frac{d}{d\frac{\lambda}{s\mu}} \frac{1}{1-\frac{\lambda}{s\mu}} = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_0 \frac{\frac{\lambda}{s\mu}}{\left(1-\frac{\lambda}{s\mu}\right)^2}$$

Per cui il valore atteso è calcolabile come:

$$L = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\frac{\lambda}{s\mu}}{\left(1-\frac{\lambda}{s\mu}\right)^2} \frac{1}{\sum_{n=0}^{s-1} \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu-\lambda}} \tag{6.6.2}$$

6.7 Rete Aperta

Si definisce una rete di code l'interconnesione di un insieme di M stazioni, numerate da uno ad M. Le espressioni per le grandezze ottenute precedentemente possono essere usate solamente per una singola stazione.

Una rete viene chiamata aperta se è presente un esplicito scambio con l'esterno, tutti i clienti che entrano dal sistema, dovranno necessariamente uscire dal sistema. L'entrata dei clienti nel sistema è determinata da una frequenza di arrivo al sistema completo λ, se escono lo stesso numero di clienti dal sistema, allora la velocità di servizio complessiva del sistema μ dovrà essere maggiore della frequenza di arrivo del sistema affinché sia in grado di processare tutti i clienti in arrivo.

La frequenza di arrivo distribuita esponenzialmente con parametro caratteristico λ corrisponde ad una distribuzione poissiana del tempo di interarrivo al sistema.

In un sistema $M/M/S$ l'uscita è distribuita esponenzialmente esattmamente come l'arrivo, con lo stesso parametro caratteristico λ, poiché in una rete aperta il numero di clienti in entrata corrisponde al numero di clienti in uscita al sistema. Inoltre si considera la coda illimitata, per cui tutti i clienti in entrata vengono serviti in un tempo finito per poi uscire dal sistema. Per cui in una serie di stazioni connesse tra di loro l'entrata della stazione a valle corrisponde all'entrata della stazione a monte. Le stazioni sono collegate tra di loro da archi pesati, che indicano con quale percentuale l'uscita dalla stazione a monte arrivi ad una delle stazioni collegate a valle. Dati più archi entranti in una singola stazione, la sua frequenza di arrivo corrisponde alla somma tra tutte le grandezze in entrata. Il peso degli archi rappresenta la probabilità di instradamento P_{ij} dalla stazione i alla stazione j. La somma di tutte le probbabilità queste probabilità, rappresenta l'intero spazio campione, per cui dovrà essere pari ad uno, si considera l'uscita verso l'esterno P_{i0}. Può essere presente un autociclo, per cui si include la probabilità P_{ii}. L'espressione è quindi:

$$\sum_{j=0}^{M} P_{ij} = 1 \tag{6.7.1}$$

Per trattare una rete di code aperta, si considerano cinque ipotesi fondamentali, per definire che l'interconnesione tra due singole stazione rappresenta una rete aperta. Per una singola coda la variabile di stato rappresenta il numero di clienti in coda n, invece per due stazioni collegate tra di loro la variabile di stato è un vettore di dimensione pari al numero di stazioni presenti nel sistema. Per cui per una rete composta da M stazioni, la variabile di stato n è un vettore:

$$n = \begin{pmatrix} n_1 \\ \vdots \\ n_M \end{pmatrix} \tag{6.7.2}$$

Per poter determinare il valore della varibile di stato, è necessario conoscere la sua distribuzione di probabilità P_n.

Una rete si dice aperta se rispetta queste cinque ipotesi:

- Il tempo di arrivo t_a di ogni stazione j è distribuito esponenzialmente (o poisoniano) con parametro caratteristico λ_j per $j = 1, \cdots, M$;

- Il tempo di servizio t_s di ogni stazione j è distribuito esponenzialmente con parametro caratteristico μ_j per $j = 1, \cdots, M$;
- La disciplina di servizio è FIFO;
- Ogni stazione è una coda illimitata, per garantire la distribuzione esponenziale, per cui ogni stazione è un sistema del tipo $M/M/1$ o $M/M/s$;
- La probabilità di instradamento complessiva da una stazione i verso tutte le possibli stazioni j è pari ad uno: $\sum_{j=0}^{M} P_{ij} = 1$.

6.7.1 Teorema di Jackson

Il teorema di Jackson, fornisce, se sono soddisfatte le ipotesi di una rete aperta, un'espressione della distribuzione di probabilità dell'intero sistema e di ogni singola stazione. Inoltre il teorema è applicabile se sono soddisfatte altre due ipotesi. Si considera la frequenza di arrivo effettiva λ'_j ad ogni singola stazione, espressa dalla frequenza di arrivo dall'esterno λ_j e da tutte le altre stazioni connesse i, rappresentato dalla loro frequenza di uscita dal sistema, pari alla frequenza di arrivo, moltiplicata per la probabilità di instradamento alla stazione j:

$$\lambda'_j = \lambda_j + \sum_{i=0}^{M} P_{ij}\lambda'_i \tag{6.7.3}$$

Per ogni stazione j, poiché sono sistemi $M/M/s$ o $M/M/1$, si possono trattare le stazioni come fossero indipendenti, esprimendo la loro distribuzione tramite la seguente espressione, dove le interconnesioni tra le stazioni sono contenute nella frequenza effettiva λ'_j:

$$f_j(n_j) = \begin{cases} \dfrac{1}{n_j!}\left(\dfrac{\lambda'_j}{\mu_j}\right)^{n_j} f_j(0) & n_j < s_j \\ \dfrac{1}{s_j! s_j^{n_j - s_j}}\left(\dfrac{\lambda'_j}{\mu_j}\right)^{n_j} f_j(0) & n_j \geq s_j \end{cases} \tag{6.7.4}$$

Si enunciano quindi le ulteriori ipotesi del teorema di Jackson:

- Il sistema deve essere stazionario, per cui ogni stazione deve essere stazionaria: $\lambda'_j < s_j\mu_j$;
- Per ogni stazione j, la somma delle sue distribuzioni per ogni stato n_j è pari ad uno: $\sum_{n_j=0}^{+\infty} f_j(n_j) = 1$.

Se queste ipotesi, insieme alle ipotesi di una rete aperta, sono soddisfatte, allora la probabilità di trovarsi in uno stato n, è pari alla produttoria di tutte le distribuzioni di ogni stazione j allo stato

associato n_j:

$$P(n) = P(n_1, \cdots, n_M) = \prod_{j=1}^{M} f_j(n_j)$$

Inoltre la distribuzione di stato di ogni stazione j coincide con la sua probabilità di trovarsi allo stato j:

$$P_j(n_j) = f_j(n_j)$$

Allora la distribuzione di probabilità probabilità del è data dalla produttoria su tutte le stazioni, della loro distribuzione probabilità di probabilità:

$$P(n) = \prod_{j=1}^{M} P_j(n_j) \tag{6.7.5}$$

Per cui si possono trattare le stazioni come fossero indipendenti, poiché la dipendenza è racchiusa dalla frequenza d'arrivo effettiva λ'_j e possono essere usate le espressioni ottenute precedentemente per determinare le grandezze di interesse di ogni singola stazione, considerando la frequenza effettiva λ'_j al posto della frequenza di arrivo λ_j.

Il valore atteso del numero dei clienti N può essere banalmente calcolato come la somma del valore atteso del numero dei clienti su ogni stazione j:

$$N = \sum_{j=1}^{M} N_j \tag{6.7.6}$$

Per ottenere il valore atteso del tempo di attraversamento è sufficiente applicare la legge di Little $W = N/\lambda$, dove si calcola la frequenza di arrivo esterna al sistema, come la somma della frequenza di arrivo esterna ad ogni stazione λ_j:

$$\lambda = \sum_{j=1}^{M} \lambda_j$$

Per determinare invece il valore atteso del tempo di attraversamento di una singola stazione j, poiché sono interconnesse tra di loro, bisogna considerare che lo stesso cliente potrebbere passare per la stessa stazione j più di una volta. Per cui solo se il cliente attraversa la stazione j una sola volta si può calcolare il valore atteso del tempo di attraversamento come $W_j = 1/(s_j\mu_j - \lambda'_j)$. Per quantificare questo fenomeno si introduce il "vist count" ν_j, che indica quante volte uno stesso cliente attraversa la stessa stazione j. Per calcolare questa percentuale (o contatore) si considerano gli arrivi effettivi alla singola stazione ed all'intero sistema, per cui si calcola come il rapporto tra la frequenza di arrivo effettiva alla stazione j per la frequenza di arrivo complessiva:

$$\nu_j = \frac{\lambda'_j}{\lambda} \tag{6.7.7}$$

Per cui il valore atteso del tempo di attraversamento di una singola stazione j si ottiene come la grandezza ottenuta considerando la stazione indipendente, moltiplicato per il suo "visit count" ν_j:

$$W_j = \nu_j W\big|_j = \frac{\nu_j}{s_j\mu_j - \lambda'_j} \tag{6.7.8}$$

Si può quindi esprimere il valore atteso del tempo di attraversamento dell'intero sistema come la somma di tutti i valori medi di ogni stazione j:

$$W = \sum_{j=1}^{M} W_j = \sum_{j=1}^{M} \nu_j W\big|_j = \sum_{j=1}^{M} \frac{\nu_j}{\mu_j - \lambda'_j} \tag{6.7.9}$$

6.8 Produttività

Data una rete di code, o una singola stazione, si definisce la produttività reale X_R, la frequenza con cui crea prodotti. Poiché è un sistema stazionario, la stessa quantità di prodotti o clienti che entra, esce dal sistema, per cui la produttività reale corrisponde alla frequenza di arrivo al sistema dall'esterno. Per cui per una rete di code aperta si ottiene la seguente espressione per la produttività reale:

$$X_R = \lambda = \sum_{j=1}^{M} \lambda_j \tag{6.8.1}$$

Per ogni stazione j invece si posono individuare due diverse produttività reali. Si possono consideare solamente i prodotti finiti dalla stazione, ovvero non si considerano i prodotti e clienti che attraversano la stazione più di una volta, per cui questa produttività si esprime come:

$$X_{R_j} = \frac{\lambda'_j}{\nu_j} = \frac{\lambda'_j}{\frac{\lambda'_j}{\lambda_j}} = \lambda_j$$

La produttività reale di una stazione j, considerando solo i prodotti finiti, coincide agli arrivi dall'esterno, questo era facilmente individuabile poiché essendo la stazione stazionaria il numero di clienti in entrata coincide con il numero di clienti in uscita.

Invece se si considerano tutte le uscite dalla stazione nel calcolo della produttività, allora coincide con la frequenza di arrivo effettiva, includendo i clienti che entrano nella stazione più di una volta:

$$X_{R_j} = \lambda'_j$$

La produttività teorica di una stazione rappresenta la massima velocità a cui i prodotti possono uscire dal sistema. Per un sistema $M/M/s$ è quindi:

$$X_{T_j} = s_j \mu_j$$

Analogamente alla produttività reale, questo valore considera anche i clienti che hanno attraversato la stazione più di una volta, per cui per ottenere la produttività che tiene conto solamente dei prodotti finiti da una stazione j, si divide per il suo "visit count":

$$X_{T_j} = \frac{s_j \mu_j}{\nu_j}$$

Questa grandezza considera quindi solo i prodotti che sono entrati ed usciti dal sistema dall'esterno, senza attraversare nuovamete la stazione.

Per cui per ogni stazione si ottengono le seguenti relazioni:

	Finiti	Tutti
X_{R_j}	λ_j	λ'_j
X_{T_j}	$\frac{s_j\mu_j}{\nu_j}$	$s_j\mu_j$

La produttività reale del sistema coincide alla somma delle produttività di ogni singola stazione, mentre la produttività teoretica della rete, corrisponde al minimo valore teoretico di produttività delle stazioni, poiché anche aumentando gli arrivi dall'esterno, il sistema non può processare una frequenza di clienti maggiore della velocità di processamento più piccola. Questa stazione viene chiamata collo di bottiglia, poiché costringe il sistema a rallentare.

$$X_T = \min_{j=1,\cdots,M}\{X_{T_j}\} = \min_{j=1,\cdots,M}\left\{\frac{s_j\mu_j}{\nu_j}\right\} \tag{6.8.2}$$

La produttività reale non potrà mai oltrepassare la prduttività teoretica, per mantenere verificata la condizione di stazionarietà:

$$X_R < X_T$$

La massima produttività reale raggiungibile si esprime quindi come:

$$X_{R_{\max}} = X_T - \varepsilon \tag{6.8.3}$$

Dove ε è una frequenza arbitrarimaente piccola, ma non è consigliato scegliere valori ristretti poiché alla minima fluttuazione degli arrivi esterni si potrebbe perdere la condizione di stazionarietà del sistema.

Per cui per aumentare la produttività reale di un sistema, si aumentano gli arrivi dall'esterno fino a raggiunere un valore minore della produttività teoretica del collo di bottiglia del sistema. Mentre per aumentare la produttività teoretica si aumenta la velocità di servizio del collo di bottiglia fino a che la sua produttività teoretica non assuma un valore pari alla seconda produttività teoretica minore del sistema. Poiché per alterare il "visit count" o il numero dei serventi bisognerebbe alterare la struttura del sistema.