UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

"SISTEMI CODA-INVENTORY"

Relatore: Prof. Roberto Corvaja

Laureando: Luca Vergolani

ANNO ACCADEMICO 2021 – 2022

Data di laurea 21/09/2022

INDICE

INTRODUZIONE

I sistemi coda-inventory sono dei particolari sistemi capaci di modellizzare problemi reali relativi ai sistemi di tipo produttivo, logistico e gestionale, che normalmente vengono semplificati attraverso i modelli della teoria inventory o della teoria delle code indipendentemente. Attraverso lo studio di questi sistemi nasce nei primi anni ’90 un ambito di ricerca che tutt’ora è ancora aperto e ampiamente discusso.

Nel corso dei prossimi capitoli verrà proposta un’analisi dei sistemi coda-inventory allo scopo di mostrarne le caratteristiche fondamentali e gli approcci che sono stati utilizzati per affrontare lo studio di questi sistemi. A questo proposito verranno presentati i principali contributi teorici e gli sviluppi fondamentali della ricerca corrispondente. In particolare, verranno analizzati alcuni degli studi più significativi descrivendone le finalità, l’obiettivo della ricerca e il modello matematico utilizzato.

Nel primo capitolo vengono presentati i sistemi coda-inventory attraverso un’introduzione ai sistemi inventory, un’introduzione ai sistemi a coda e un’analisi dei primi due studi relativi ai sistemi coda-inventory.

Nel secondo capitolo invece vengono presentati gli ambiti di ricerca fondamentali e le principali variazioni che sono state introdotte attraverso l’analisi dei modelli e delle metodologie di alcuni studi caratteristici.

Mentre nel terzo capitolo vengono presentati gli ambiti di ricerca aperti e futuri attraverso l’analisi di uno studio molto recente e attraverso la presentazione di alcune caratteristiche non ancora introdotte nella ricerca.

CAPITOLO 1
INTRODUZIONE AI SISTEMI CODA-INVENTORY

1.1 Introduzione alla teoria inventory

La teoria matematica dell'inventory e della produzione è una teoria matematica che si pone l'obiettivo di modellizzare il comportamento di un sistema produttivo e di immagazzinamento al fine di minimizzarne i costi. La gestione degli inventory, scorte di oggetti destinate alla produzione o alla vendita, è infatti un processo essenziale per tutte le aziende che utilizzano o producono oggetti fisici. Le aziende manufatturiere, ad esempio, necessitano di magazzini per l'accumulo sia di materie prime sia di prodotti finiti che devono essere spediti; i grossisti e i rivenditori invece devono poter assicurare ai propri clienti un numero adeguato di prodotti al fine della vendita, per questo motivo è essenziale la gestione dei magazzini. I costi di gestione e mantenimento hanno però un impatto non indifferente sulle aziende che ne fanno uso, per questo è di fondamentale importanza cercare di limitarli il più possibile riducendo gli spazi inutilizzati e l'accumulo di oggetti non necessari. Nasce così la necessità di studiare più approfonditamente tale problematica che prende il nome di inventory control problem.

Lo studio di questi problemi viene effettuato attraverso l'utilizzo di strumenti matematici e di ricerca operativa. L'approccio utilizzato parte dalla creazione di modelli matematici che vengono poi analizzati per trovare le soluzioni ottimali da applicarle ai sistemi reali. Questi modelli possono essere raggruppati in due macrocategorie: modelli deterministici e modelli stocastici, a seconda del grado di aleatorietà dei parametri. In particolare, ciò che li caratterizza è la variabile che descrive la domanda di un prodotto, ovvero il numero di oggetti che vengono ritirati dall'inventory in un preciso periodo di tempo: se la domanda è approssimabile ad una costante allora si parla di modelli deterministici, altrimenti di modelli stocastici.

Questi modelli presentano una struttura standard costituita dai compratori, i quali richiedono una certa quantità di oggetti dall'inventory, e dai venditori, che servono i compratori. Le richieste dei compratori possono essere prese in carico, procedendo al trasferimento delle corrispondenti quantità dall'inventory al compratore, se le quantità richieste sono fisicamente presenti nell'inventory; oppure possono essere rifiutate in caso contrario. Le richieste rifiutate prendono il nome di shortages, e possono essere gestite in modo diverso a seconda del modello che si sta prendendo in considerazione. Se si sceglie di non tenere traccia delle richieste

rifiutate, si parla di modello con lost sales, in questo caso tutte le richieste non gestite immediatamente vengono perse per sempre; in caso contrario, queste prendono il nome di backordes o backlogs e vengono messe in attesa finché il sistema non torna fisicamente disponibile. Per gestire le richieste pendenti viene generalmente allocato uno spazio che può essere limitato oppure illimitato e, a seconda della scelta, si parla rispettivamente di modello con backlog finito o infinito.

Un'altra caratteristica importante di questi modelli è che il venditore può richiedere una determinata quantità di oggetti extra al fine di rifornire l'inventory. Tale richiesta viene generalmente completata dopo una certa quantità di tempo chiamato lead time. Nello specifico il lead time è il tempo che trascorre dal momento in cui il venditore richiede gli oggetti, al momento in cui li riceve, e l'inventory viene rifornito.

Il modo in cui un inventory viene gestito è determinante rispetto ai costi e al tipo di sistema che si sta modellizzando; in particolare, nel momento in cui viene richiesto un rifornimento di oggetti, è di fondamentale importanza capire quale sia la quantità di oggetti necessaria per rifornire l'inventory in modo che sia la più efficiente possibile in termini di costi. Ci sono diverse politiche di controllo del rifornimento e prendono il nome di policy. Le principali sono:

- (s, S) - policy;
- (S-1, S) - policy;
- (s, Q) - policy;
- Randomized order size.

Nella policy (s, S) S rappresenta il massimo numero di oggetti allocabili nell'inventory, s è invece la soglia minima in cui effettuare una richiesta di rifornimento. In questa politica di controllo la quantità di oggetti richiesta per un rifornimento è quella necessaria a riportare il numero di oggetti pari a livello massimo S.

Nella policy (S-1, S) viene effettuata una richiesta di rifornimento ogni volta che un oggetto viene venduto. Viene chiamata alternativamente anche come one-for-one ordering policy e il suo utilizzo è prevalente in sistemi in cui gli oggetti sono particolarmente costosi o che vengono venduti raramente.

Nella policy (s, Q) il numero di oggetti richiesto in un rifornimento è fissato ed è pari a Q, dove s rappresenta il numero di oggetti minimo in cui effettuare una richiesta di rifornimento e Q rappresenta la quantità fissata. Solitamente Q viene scelto in modo tale che sia molto maggiore rispetto ad s.

Nella politica randomized order size la decisione del numero di oggetti per il rifornimento viene gestito da una funzione di probabilità discreta p sui valori {1, 2, …, S}. In questo caso la quantità di oggetti è una variabile n con probabilità p_n, tale per cui vale che:

$$\sum_{n=1}^{S} p_n = 1$$

I modelli classici della teoria inventory si sviluppano intorno alle seguenti proprietà, e tutti vengono caratterizzati da un'assunzione fondamentale: il tempo di servizio effettuato dai venditori è nullo. In altre parole, la richiesta di oggetti fatta da un compratore viene soddisfatta istantaneamente se l'inventory dispone fisicamente del numero di oggetti richiesto. Questa assunzione fa sì che in sistemi inventory di questo tipo si formi una coda di compratori solo nel momento in cui l'inventory non è fisicamente disponibile, ovvero quando la quantità di oggetti al suo interno è nulla o inferiore alla quantità richiesta. Modificare questa assunzione significa cambiare radicalmente il comportamento di questi sistemi, infatti, se consideriamo un tempo di servizio positivo, ci dobbiamo aspettare la creazione di una coda di compratori anche quando l'inventory è disponibile, questo perché ciascun compratore impiega una certa quantità di tempo per essere servito. Sistemi inventory di questo tipo, con tempo di servizio positivo, prendono il nome di sistemi coda-inventory e per essere studiati necessitano di un approccio completamente diverso dall'usuale, ovvero vanno studiati utilizzando l'approccio della teoria delle code, ma come vedremo, anche la teoria delle code classica non è sufficiente a descrivere in maniera completa questo tipo di sistemi. Nei prossimi paragrafi vedremo come lo studio di sistemi coda-inventory sia diventato un ambito di ricerca molto importante e come si siano sviluppati i primi modelli per descriverne le caratteristiche principali.

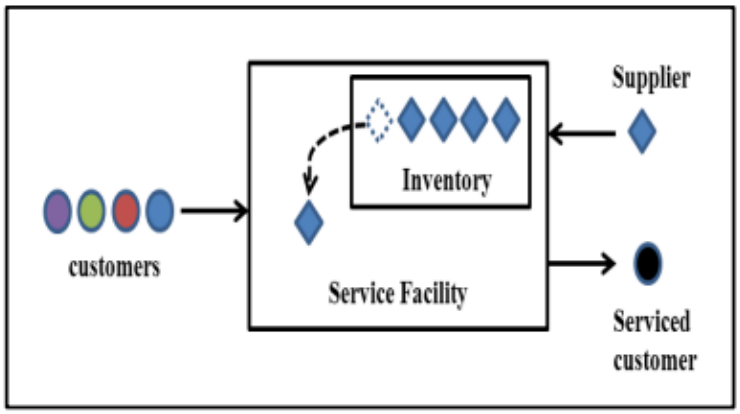

Figura 1: Modello di un sistema inventory.

1.2 Introduzione alla teoria delle code

La teoria delle code è una teoria matematica probabilistica che si pone l'obbiettivo di studiare e analizzare particolari tipi di sistemi chiamati sistemi a coda. Aspettare in coda è una situazione molto frequente, può capitare al supermercato, in autostrada e in qualsiasi altra situazione in cui è presente una struttura adibita al servizio di un qualche tipo di utenti. Questo tipo di situazioni possono verificarsi anche in ambiti meno tangibili e concreti come, ad esempio, quando si fa uso di chat o in generale quando si naviga online: questi servizi, infatti, sono forniti da sistemi di telecomunicazione che gestiscono una grande quantità di utenti e di conseguenza la creazione di code di attesa. In generale ogni volta che un qualche tipo di servizio viene richiesto da molti più utenti di quanti possa servire simultaneamente, si osserva la creazione di una coda, che viene gestita man mano che il servizio ritorna disponibile. Ridurre i tempi di attesa diventa dunque fondamentale, nasce da qui l'esigenza di sviluppare metodi e tecniche che permettano di analizzare queste situazioni.

La teoria delle code trova una soluzione a questa esigenza attraverso lo studio di modelli matematici che permettono di ricavare parametri e metriche di valutazione quali il tempo di attesa medio, il numero medio di utenti in coda in un sistema, il tempo medio in cui il servizio è occupato e tanti altri. Questo fa sì che la teoria delle code possa essere utilizzata in svariati ambiti di applicazione, dai più concreti, come ad esempio stabilire la durata di un semaforo, determinare il numero di commessi in supermercato o di sedie in una stanza d'attesa di uno studio medico, ai più astratti come ad esempio nel campo delle reti di telecomunicazioni, dove la teoria delle code può essere usata per dimensionare vari aspetti del sistema, come il tasso di trasmissione dei dati o la quantità di memoria allocabile da un router.

Come introdotto precedentemente, la teoria delle code analizza e studia particolari tipi di problemi relativi ai cosiddetti sistemi a coda. Un sistema a coda è composto da tre elementi principali: Clienti, servitori e area di accordamento. Nello specifico un sistema a coda è definito come una collezione di eventi ed attività volti a fornire un determinato servizio ad un particolare tipo di clienti. A seconda del tipo di sistema preso in considerazione, un cliente può necessitare più servizi, offerti da un determinato numero di servitori, e conseguentemente può dover aspettare in diverse aree d'attesa prima che possa uscire dal sistema. Un cliente può essere una persona, una macchina o un oggetto e più in generale rappresenta ciò che in un determinato sistema necessita la presenza di un servitore per poter essere servito. Anche un servitore a sua volta può essere una persona, una macchina o un meccanismo particolare ma, a differenza del cliente, rappresenta ciò che soddisfa le richieste dei clienti. L'interfaccia con cui il cliente

interagisce con il sistema prende il nome di struttura di servizio, e può essere progettata in modi diversi a seconda del modello che si vuole utilizzare.

A prescindere dalle caratteristiche specifiche che un sistema a coda può assumere, questo segue un comportamento standard che è possibile descrivere in questo modo: quando un cliente entra in un sistema a coda, questo viene immediatamente indirizzato verso i servitori disponibili, se ce ne sono. Nel caso in cui tutti i servitori sono occupati, ovvero nel caso in cui sono già impegnati nel servizio di altri clienti, i nuovi utenti che arrivano nel sistema entrano in una cosiddetta coda di attesa, o buffer, che rappresenta un'area in cui i clienti attendono temporaneamente che il servizio ritorni disponibile. Ogni servitore può infatti servire un solo cliente per volta e quando il servizio è completato, il cliente servito lascia il sistema rendendo nuovamente disponibile il servizio di un nuovo cliente.

Per poter descrivere adeguatamente un sistema a coda è dunque necessario definire rigorosamente le caratteristiche dei clienti, dei servitori e delle code di attesa; per questo è necessario introdurre quattro elementi base:

- Il processo di arrivo dei clienti nel sistema;
- Il processo di servizio di ogni servitore;
- La struttura dell'area di accodamento;
- La disciplina del servizio.

Per descrivere rigorosamente un modello di sistema a coda è necessario specificare il comportamento dei clienti che arrivano nel sistema, come ad esempio il tasso di arrivo; questo comportamento viene definito attraverso il cosiddetto processo degli arrivi, un processo che generalmente è per l'appunto aleatorio. Lo stesso discorso vale per quanto riguarda il processo di servizio, ovvero il processo che descrive il comportamento di servizio di ciascun servitore. Per quanto riguarda la struttura dell'area di accodamento è essenziale stabilire, invece, il massimo numero di clienti che possono essere messi in attesa, ovvero la capacità massima del sistema a coda. Inoltre, è fondamentale stabilire il modo in cui viene gestita la coda, ovvero stabilire la disciplina del servizio. Per essere più chiari si tratta di stabilire l'ordine di priorità in coda e dunque il modo in cui essa si muove nel sistema. Esempi di discipline sono ad esempio first-in-first-out, in cui il primo che arriva nel sistema è il primo che viene servito, oppure last-in-first-out, in cui l'ultimo che arriva nel sistema è il primo ad essere servito.

Tutte queste caratteristiche servono per descrivere il tipo di modello utilizzato. Un modo standard per specificare il tipo di sistema a coda attraverso queste caratteristiche è la notazione di Kendall, che consiste nello specificare i descrittori corrispondenti alle caratteristiche

fondamentali del modello che si sta utilizzando. Ad esempio: M/M/1 è il modello di sistema a coda in cui gli arrivi e il servizio sono processi markoviani e il sistema presenta 1 solo servitore, M/M/m/k è il modello in cui arrivi e servizio sono processi markoviani ma il sistema presenta m servitori e la coda ha una capacità massima di k utenti e così via.

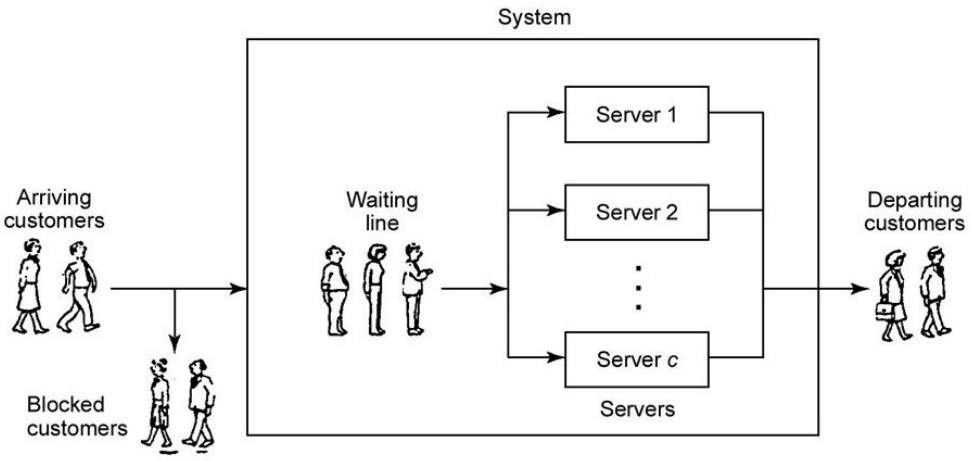

Figura 2: Modello di un sistema a coda.

Alla fine del paragrafo precedente sono stati introdotti i sistemi inventory con tempo di servizio positivo. Questi sistemi, come accennato brevemente, seguono un comportamento che sembra essere tipico dei sistemi a coda. Infatti, dopo essere entrato nel sistema, il compratore viene indirizzato verso il servitore disponibile; se l'inventory presenta fisicamente la quantità di oggetti richiesti, egli viene servito, ovvero gli viene trasferita la quantità richiesta di oggetti dall'inventory. Questo processo però richiede una certa quantità di tempo e dunque in presenza di altri compratori può portare alla creazione di una coda. Osservando in modo più critico il comportamento dei sistemi a coda e confrontandolo con quello dei sistemi inventory con tempo di servizio positivo, ci si può accorgere di una differenza fondamentale. I modelli classici della teoria delle code assumono che se un cliente attende di essere servito e un servitore è disponibile, allora viene subito indirizzato e comincia ad essere servito. Nei sistemi inventory con tempo di servizio positivo questo può non succedere, in particolare nel momento in cui l'inventory non dispone fisicamente degli oggetti richiesti e viene richiesto un rifornimento di oggetti.

Questa ipotesi fa sì che i sistemi inventory con tempo di servizio posivito non possano essere trattati unicamente attraverso la teoria delle code, in quanto non coincidono completamente con i sistemi a coda classici. Lo studio di questi sistemi necessita un approccio diverso, che tenga

conto tanto della gestione dell'inventory quanto dello studio delle code, per questo motivo sistemi inventory di questo tipo, con tempo di servizio positivo, prendono il nome di sistemi coda-inventory. Nasce così un nuovo ambito di ricerca che pone le sue radici negli studi di Melivk e Molchanov [1] e Sigman Simchi – Levi [2], studi che verranno presentati e analizzati nel prossimo paragrafo.

1.3 Presentazione e analisi dei sistemi coda-inventory

Come accennato precedentemente, i sistemi coda-inventory sono dei sistemi complessi caratterizzati dall'unione delle caratteristiche tipiche dei sistemi a coda con quelle tipiche dei sistemi inventory. Più precisamente, si tratta di sistemi caratterizzati da una strutta base tipica dei sistemi a coda, con clienti, servitori e area di accodamento, ma che contemporaneamente tengono in considerazione la gestione di un inventory annesso, che, invece, segue le caratteristiche tipiche della teoria inventory.

I sistemi coda-inventory sono estremamente importanti per modellizzare problemi reali relativi ai sistemi di tipo produttivo, logistico e gestionale, che normalmente vengono semplificati attraverso i modelli della teoria inventory o della teoria delle code indipendentemente. Nasce da questi problemi, che generalmente sono problemi di ricerca operativa e ottimizzazione, un ambito di ricerca che tutt'ora è ancora aperto e ampiamente discusso.

Lo studio di questi sistemi pone le sue radici nei primi anni '90 attraverso gli studi di Melikov e Molchanov [1] e di Sigman e Simchi-Levi [2]. Entrambi questi studi nascono dall'esigenza di risolvere dei classici problemi di ottimizzazione considerando però in modo più completo i tempi di attesa complessivi e i costi di immagazzinamento. Da questo primo obiettivo si sviluppa la ricerca di nuovi metodi risolutivi nel tentativo di trovare delle espressioni matematiche per descrivere le caratteristiche fondamentali dei modelli. Nel corso di questo paragrafo verranno presentati entrambi gli studi citati e verranno analizzati descrivendone le finalità, l'obiettivo della ricerca e il modello matematico utilizzato. Questo tipo di approccio agli studi e agli articolì verrà utilizzato in tutti i successivi capitoli, in modo da fornire uno sguardo completo e rigoroso, ma non completamente esaustivo, sugli argomenti che verranno presentati.

Il modello che verrà utilizzato per affrontare gli studi relativi ai sistemi coda-inventory è costituito generalmente da un'area di accodamento, da un'area di servizio e da un inventory con le seguenti caratteristiche:

- N, con $N \in [1, \infty)$, è il numero di clienti massimo che può attendere in coda;
- $m,\ con\ m < \infty$, è il numero di servitori presenti nel sistema;
- $S,\ con\ S < \infty$, è il numero di oggetti massimo allocabile nell'inventory;
- $s,\ con\ s < \infty$, è la soglia minima oltre alla quale viene richiesto un rifornimento.

Inoltre, vengono definiti i processi degli arrivi, del servizio e del rifornimento dell'inventory tramite delle variabili tipicamente esponenziale di parametri:

- λ per il tasso del processo degli arrivi nel sistema;
- μ per il tasso del processo del servizio;
- γ per tasso di rifornimento del magazzino;

Questo modello verrà esteso e adeguato a tutti gli studi presentati nel corso dei prossimi capitoli e fungerà da base comune per poterne fare una presentazione coerente. Di seguito una rappresentazione del modello appena descritto.

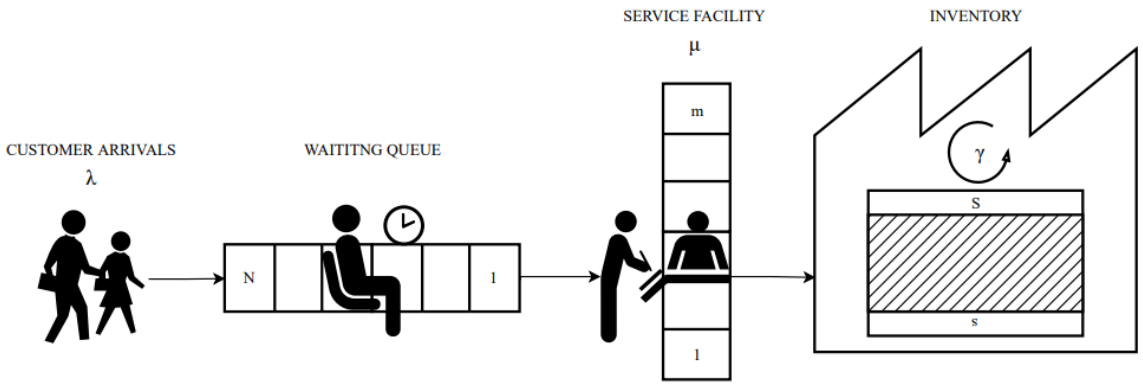

Figura 3: Modello di un sistema coda-inventory.

Stock optimization in transportation/storage systems

Uno dei primi studi riconosciuti nell'ambito dei sistemi coda-inventory è quello di Melikov e Molchanov [1]. Nella loro trattazione "Stock optimization in trasportation/storage system" cercano di affrontare un problema di ottimizzazione, chiamato comunemente "inventory control problem", di un sistema di trasporto e immagazzinamento (transpostation/storage systems, TSS); nello specifico, l'obbiettivo della loro ricerca è quello di valutare il tipo di controllo del

magazzino ottimale per un sistema di rifornimento di prodotti petroliferi. Questo tipo di problema è un elemento importante nell'ottimizzazione di sistemi inventory complessi e rappresenta uno dei problemi classi della teoria inventory.

In molti studi riguardanti i sistemi di controllo dell'inventory, come anticipato precedentemente, viene però assunto che le richieste dei clienti vengano soddisfatte istantaneamente e conseguentemente che il magazzino sia ridotto istantaneamente di uno stock per ogni cliente servito. Ciò significa che non si tiene in considerazione la formazione di una coda nel sistema quando l'inventory è ancora disponibile, bensì solo nel caso in cui l'inventory è vuoto. In questo modo i modelli classici sono ottimizzati ignorando i costi delle richieste dei clienti in coda e i costi legati al numero limitato di posti in coda, entrambe situazioni che si verificano normalmente in sistemi TSS ed in particolare nei sistemi di fornitura di prodotti petroliferi, come ad esempio stazioni di servizio o depositi di petrolio.

Il modello utilizzato all'interno della loro ricerca si basa su un classico modello TSS con le seguenti caratteristiche. Come input del sistema viene scelto un processo degli arrivi, ovvero processo delle richieste dei compratori, Poissoniano di tasso λ. Le risorse vengono servite ai clienti attraverso un solo canale e il tempo di trasferimento, o tempo di servizio, viene modellizzato attraverso una variabile aleatoria esponenziale di parametro μ.

Posto S la capacità massima del magazzino, il numero di oggetti immagazzinato si riduce di una unità solo dopo che il servizio di un cliente è stato completato. Questa unità può essere trasferita dall'inventory al cliente solo finché il numero di oggetti presente nel magazzino non raggiunge il valore $s, 0 \leq s \leq S-1$, valore dopo il quale il magazzino smette di trasferire gli oggetti a prescindere dal numero di richieste in coda presenti nel sistema. Raggiunto il livello di soglia minima s, il sistema può procedere con la richiesta di un numero di oggetti aggiuntivo. A questo proposito viene assunto che il sistema richieda una quantità $i,\ i \in \{1, 2, \dots, S-s\}$, con probabilità $\alpha_i(s,n)$, tale che $\sum_{i=1}^{S-s} \alpha_i(s,n) = 1\ \forall n$, dove n è la lunghezza della coda di richieste dei clienti.

Gli oggetti aggiuntivi richiesti vengono ricevuti dopo una certa quantità di tempo, dovuta al trasporto e allo scarico degli oggetti nel magazzino, durante il quale il sistema non serve alcun utente. Questa quantità di tempo, chiamata lead time, viene modellata attraverso una variabile aleatoria esponenziale con media γ.

Inoltre, i clienti sono assolutamente pazienti, ovvero aspettano in coda finché non vengono serviti e le loro richieste sono indipendenti dal numero di oggetti presenti nel magazzino. Il numero massimo di clienti ammessi nel sistema è N. Ciò significa che se un utente che arriva

nel sistema in cui sono già presenti N utenti in coda fa una richiesta, questa non verrà considerata e verrà persa.

Per formalizzare il problema di ottimizzazione di questo modello, Melichov e Molkanov utilizzano la seguente notazione:

- (m, n) è il vettore di stato bidimensionale del sistema, dove m rappresenta il numero di oggetti presente nel magazzino, n invece rappresenta il numero di utenti in coda;
- $E \coloneqq \{(m, n) : m = s, \dots, S,\ n = 0, \dots, N\}$ è lo spazio delle fasi del sistema, ovvero lo spazio in cui sono rappresentati tutti gli stati del sistema;
- $\pi(m, n)$ rappresenta la probabilità stazionaria dello stato (m, n);
- c_1 è il costo associato al tempo di attesa di una richiesta di un cliente per unità di tempo;
- c_2 è il costo associato alla perdita di una richiesta di un cliente;
- c_3 è il costo di stoccaggio di un oggetto nel magazzino per unità di tempo;
- $\theta(x; y)$ è il tempo di attesa per una transizione $x -> y$ di due stati del sistema.

Posto:

$$G = \sum_{m,n \in E} (nc_1 + c_3 m)\pi(m, n) + \sum_{k>N} c_2 P[n = k]$$

il costo totale per unità di tempo associato ai costi di attesa, di perdita delle richieste e di stoccaggio di oggetti nel magazzino, l'obiettivo di Marchov e Melikov è stato quello di trovare il valore $\alpha_i(s, n), i = 1, \dots, S-1; s, n \in E$, che minimizza G in condizione di stazionarietà.

Per risolvere questo problema Marchov e Melikov utilizzano i metodi classici usati per risolvere le catene di Markov e metodi di programmazione dinamica e lineare. Propongono inoltre una soluzione approssimata da utilizzare nel caso in cui la dimensione dello spazio delle fasi è talmente grande da essere difficile da calcolare computazionalmente. Questa strategia assume innanzitutto di essere in condizione di "high load", secondo la quale il tasso di arrivo è molto maggiore del tasso di servizio e il rifornimento del magazzino avviene molto raramente, e inoltre utilizza metodi di decomposizione e aggregazione dello spazio delle fasi grazie ai quali è possibile progettare degli algoritmi più efficienti.

Light traffic heuristic for an M/G/1 queue with limited inventory

Un altro importante studio che pone le basi allo studio di sistemi coda-inventory è quello di Sigman e Simchi-Levi [2] che, nel loro studio "Light traffic heuristic for an M/G/1 queue with limited inventory", cercano di ricavare delle formule chiuse per esprimere le metriche fondamentali del sistema e nello specifico il tempo di attesa di un sistema a coda M/G/1 con inventory annesso. Questo studio nasce dall'intenzione ultima di risolvere un cosiddetto "facility location problem", FLP, ovvero un problema matematico di ottimizzazione in cui si cerca di determinare la posizione ottimale di una fabbrica o di un magazzino, considerando la geografia delle richieste, i costi di servizio, e le distanze di trasporto. Lo scopo di questa ricerca è quello di minimizzare i costi complessivi e conseguentemente di massimizzare i profitti rispetto alla richiesta dei clienti e alla loro posizione relativa. Lo studio di Sigman e Levi viene dunque affrontato in due fasi principali: la prima consiste nel fornire un metodo matematico che consenta di ricavare una formula chiusa per calcolare il tempo complessivo di attesa del sistema. Nella trattazione classica, infatti, formule chiuse per sistemi M/G/1 con inventory erano ricavate considerando solo il caso di oggetti illimitati, e dunque riconducendosi al caso di sistemi a coda classici; in ogni caso ricavare un'espressione per il calcolo del tempo di attesa è di fondamentale importanza nei problemi di ottimizzazione dove si è interessati al ridurre al minimo i costi di immagazzinamento legati ai tempi di attesa. Nella seconda parte, tramite i risultati ottenuti, viene presentato lo studio e l'analisi del problema di ottimizzazione introdotto precedentemente.

Il modello considerato da Sigman e Levi viene costruito su un sistema M/G/1 con criterio FIFO, tasso di arrivo λ e tasso di servizio μ. A questo modello viene aggiunto un magazzino che contiene un numero finito di oggetti iniziali pari a S. Ogni arrivo corrisponde ad un cliente, il quale può richiedere dall'inventory un solo oggetto che gli viene trasferito al termine del servizio, in questo modo il numero di oggetti del magazzino si riduce di una unità per volta. Il tempo richiesto per servire un cliente e per preparare il sistema al servizio di quello successivo, che eventualmente si trova in coda, è il tempo di servizio e rispetto allo studio precedente corrisponde ad una variabile aleatoria T, che nel corso dello studio varia in esponenziale, uniforme e Erlang. Con l'aumentare del numero di clienti serviti, si riduce progressivamente il numero di oggetti presente nell'inventory e, una volta raggiunto il valore s, $0 \leq s \leq S$, viene richiesto un rifornimento di oggetti pari al numero necessario per riportare il livello del magazzino al valore iniziale S. Questo processo impiega un tempo I rappresentato in questo modello da una variabile aleatoria esponenziale di parametro γ, ovvero $I \sim exp(\gamma)$. Durante il

rifornimento le richieste di eventuali nuovi utenti continuano ad essere assecondate finché il magazzino non esaurisce gli oggetti, da questo momento in poi eventuali altri clienti vengono messi in coda di attesa finché il sistema non torna disponibile.

Al fine di ricavare una formula chiusa per il calcolo del tempo medio di attesa in coda t_w, Sigma e Levi utilizzano l'assunzione di "light traffic", ovvero assumono che i parametri $\lambda, \mu \, e \, \gamma$ soddisfano il seguente comportamento: $\gamma \leq \lambda \leq \mu$; in altre parole, il taso di rifornimento è inferiore al tasso di arrivo che è inferiore al tasso di servizio. Questa assunzione fa sì che con grande probabilità si formi una coda solo ed esclusivamente quando il magazzino è vuoto, finché non viene rifornito nuovamente. Questo permette di cambiare il modo di studiare il sistema ponendo l'attenzione sul cosiddetto "busy period" (B_a), ovvero il periodo di tempo in cui il sistema è impegnato nel servire i clienti: tanto più clienti sono presenti nel sistema, tanto maggiore è lavoro del sistema e di conseguenza il "busy period". Da questa considerazione è possibile approssimare il tempo di attesa complessivo considerando l'"exceptional first service (EFS) busy period" con tempo di servizio pari a $S_a = I + S$, ovvero considerando il numero di clienti che entrano nel sistema vuoto portandolo in stato di "busy", i clienti che entrano nel sistema durante un EFS "busy period" e il lavoro complessivo del sistema in quel periodo di tempo. Questa metodologia permette di ricavare con più facilità i parametri del sistema ed in particolare il tempo medio di attesa nel sistema.

Utilizzando questa strategia e definendo

- n_1 il numero di clienti che iniziano un EFS busy period;
- n_2 il numero di clienti che arrivano durante un EFS busy period;
- V_a il lavoro complessivo del sistema durante un momento casuale dell'EFS busy period;
- $E(V_a)$ il tempo di attesa medio di un cliente durante un EFS busy period;

Sigman e Levi riescono ad esprimere il tempo medio di attesa nel sistema come:

$$t_w = n_1/\gamma + n_2 E(V_a)$$

Relazione che nel corso della trattazione viene espressa esclusivamente in funzione dei parametri base. Considerando $S = 50, \mu = 1,$ un'analisi numerica mostra come varia il tempo medio di attesa in funzione del valore di soglia, del tasso di arrivo e del tasso di rifornimento. In Figura 4 è possibile vedere come varia il tempo di attesa medio calcolato tramite relazione algebrica t_w, in figura chiamato H_1 e tramite una simulazione computazionale, in figura chiamata SIM_1. Si osserva come i risultati del modello seguono l'andamento stimato in modo corretto.

Average delay for exponential service time*.

	u =	0	10	25	49
$\lambda = 0.2,\ \gamma = 0.01$	SIM_1	59.64	39.76	24.17	16.28
	H_1	53.14	34.18	20.07	12.69
$\lambda = 0.1,\ \gamma = 0.01$	SIM_1	23.88	10.76	2.78	1.46
	H_1	24.70	10.32	3.34	1.06
$\lambda = 0.05,\ \gamma = 0.01$	SIM_1	10.42	2.38	0.15	0.07
	H_1	12.74	2.33	0.22	0.014
$\lambda = 0.1,\ \gamma = 0.005$	SIM_1	101.78	86.66	46.97	25.00
	H_1	93.83	60.35	35.44	22.41
$\lambda = 0.05,\ \gamma = 0.005$	SIM_1	41.87	19.87	4.82	3.11
	H_1	46.54	19.44	6.30	2.00

*Simulation was programmed in C and run on IBM/PC.

Figura 4: Variazione del tempo di attesa medio in funzione della soglia u.

I risultati ottenuti vengono poi applicati ad un problema di ottimizzazione che viene formalizzato considerando un grafo orientato $G(N, L)$ con N nodi e L archi. Viene assunto che una unità di servizio è localizzata nel grafo in un punto fissato x. Le richieste di servizio dei clienti provengono esclusivamente dai nodi del grafo, da cui vengono generati processi di richieste Poissoniani indipendenti l'uno dall'altro. Il tasso di richiesta è pari a $\lambda h_i, \forall i \in N, tali\ che \sum_{i \in N} h_i = 1$.

Ogni richiesta proveniente da un nodo richiede che l'unità di servizio, proveniente da x, raggiunga il nodo corrispondente, serva il nodo e infine ritorni in x. Le ulteriori richieste che arrivano mentre il servizio è occupato, attendono in una coda gestita tramite la disciplina FIFO. Viene poi definito il tempo associato al noto i, considerando:

- c_{ab} la distanza minima da un nodo a ad un nodo b appartenenti a G;
- v la velocità di spostamento;
- c_{xi}/v il tempo di spostamento da x a i;
- Il tempo di servizio nel nodo;
- $(\beta - 1)c_{xi}/v$ il tempo di spostamento da i a x, dove $\beta > 1$ è un fattore che modifica la velocità di ritorno;
- Il tempo di servizio nel nodo base.

Questo modello, rispetto alla trattazione classica, considera non solo la formazione di una coda come risposta al tempo di servizio, riducendosi quindi ad un sistema M/G/1 standard, ma anche come conseguenza dello spazio limitato del magazzino. Per questo motivo il tempo di attesa stimato del sistema, $E(W)$, può cambiare a seconda della posizione dell'unità di servizio.

Definendo,

$$E(W) = t_w + \sum_{i \in N} h_i \, c_{xi}/v$$

il problema si riduce dunque al ricavare la posizione ottimale x per l'unità di servizio che minimizza il tempo di attesa del sistema $E(W)$. La metodologia proposta è quella di utilizzare il partizionamento dei segmenti e utilizzare delle regioni di convessità per stabilirne il minimo.

I risultati ottenuti da questi due studi sono fondamentali poiché pongono le basi verso uno studio più approfondito dei sistemi coda-inventory. Nello specifico verrà data sempre più importanza alla ricerca di formule chiuse ed espressioni semplici che descrivano matematicamente e in modo rigoroso le metriche fondamentali di questi sistemi. Un altro fondamentale contributo in questa direzione è quello di Berman e i suoi collaboratori, che con una serie di articoli tentano di studiare più nel dettaglio i sistemi coda-inventory. Sulla scia dei risultati ottenuti da Sigman, Simchi-Levi, Malikov e Molchanov, riescono a dimostrare che in un sistema esponenziale con lead time nullo, una politica ottima di gestione dell'inventory non richiede un rifornimento di oggetti finchè il magazzino non è nullo e una certa quantità di clienti non si trova in coda.

Da questo momento grande attenzione viene posta sullo studio di sistemi coda-inventory e in particolare nella ricerca di una soluzione analitica esplicita nel tipo di "product form". In altre parole, assumendo $N(t)$ e $I(t)$ rispettivamente il numero di clienti nel sistema e il numero di oggetti nell'inventory al tempo t, si dice che il sistema ammette una soluzione del tipo "product form" se la distribuzione asintotica di $\big(N(t), I(t)\big)$ può essere scritta come il prodotto della distribuzione di due variabili aleatorie indipendenti. Uno dei primi contributi in questo ambito è lo studio di Schwarz [3] che nella trattazione "M/M/1 Queuing systems with inventory" analizza per l'appunto un sistema a coda di tipo M/M/1 con diverse tipologie di inventoy e politiche di controllo associate. Questo studio è di fondamentale importanza poiché oltre a ricavare delle espressioni matematiche esplicite delle metriche di performance del sistema, riesce ad esprimere le funzioni di probabilità stazionaria del sistema nel tipo "product form" presentato precedentemente.

Nel corso degli anni successivi, i modelli dei sistemi coda-inventory presi in considerazione estendono sempre di più le ipotesi e le caratteristiche, portando la ricerca ad approfondire sempre più rami ed aspetti peculiari di questi sistemi. Nei prossimi capitoli verranno presentati alcuni di questi aspetti caratteristici attraverso l'analisi degli studi che hanno introdotto tali caratteristiche.

CAPITOLO 2
PRESENTAZIONE E ANALISI DEGLI AMBITI DI RICERCA FONDAMENTALI

L'analisi dei sistemi coda-inventory nasce, come visto nel capitolo precedente, tramite lo studio di due problemi di ottimizzazione legati agli ambiti produttivo e logistico-gestionale. Da allora l'ambito di ricerca relativo ai sistemi coda-inventory si è sviluppato analizzando sempre più nel dettaglio le caratteristiche fondamentali dei modelli e cercando delle proprietà matematiche che ne descrivano il comportamento. In particolare, nasce il desiderio di esplorare le potenzialità e i limiti di questi sistemi; per questo motivo vengono considerati modelli più complessi che tengono in considerazione aspetti sempre più specifici di ciascun sistema. Alcune delle caratteristiche introdotte si basano sull'allentare alcune ipotesi precedentemente prese in considerazione, come ad esempio il fatto che gli oggetti dei magazzini non abbiamo un tempo di scadenza e non siano soggetti al deterioramento; altre invece aggiungono dei comportamenti particolari, come ad esempio il fatto che un servitore possa prendersi una pausa o che i clienti possano riprovare a richiedere un servizio prima di lasciare il sistema. Nei prossimi paragrafi verranno presentati alcuni ambiti della ricerca relativi ai sistemi coda-inventory, ciascuno dei quali presenta delle variazioni caratteristiche sui modelli.

2.1 Sistemi coda-inventory con retrial queue

Una delle caratteristiche fondamentali dei sistemi inventory è quella di specificare il comportamento del sistema quando il magazzino è vuoto. Come visto nei capitoli precedenti, nella teoria classica vengono generalmente considerati due approcci distinti:

- Approccio con "lost sales";
- Approccio con "backlog";

Artalejo in [4], introduce per la prima volta un terzo approccio per gestire la situazione di non disponibilità del sistema. Si tratta di considerare il caso in cui i clienti che vengono rifiutati dal sistema, invece di abbandonarlo per sempre, lasciano temporaneamente l'area di servizio e

riprovano ripetutamente a richiedere un servizio, finché il magazzino non viene rifornito e la loro richiesta presa in carico.

Capita spesso nella vita quotidiana che un cliente che visita un negozio o un magazzino non trovi quello che sta cercando, ad esempio, nel caso di un negozio di scarpe, un cliente può non trovare il paio con la taglia corrispondente oppure può scoprire che il paio di scarpe cercato è attualmente esaurito; in questa situazione tipicamente il cliente o cerca in un altro negozio, oppure prova a ritornare dopo una certa quantità di tempo. Questo ultimo comportamento formalmente prende il nome di "repeated request".

A differenza della teoria inventory, nella teoria delle code è presente una descrizione dettagliata di alcuni modelli chiamati "retrial queue". Un sistema con "retrial queue" è simile ad un classico sistema a coda che però presenta le seguenti caratteristiche. Quando un cliente trova tutti i servitori davanti a sé occupati, entra in un'area infinita, chiamata "orbita", dalla quale può provare ripetutamente ad accedere al servizio. In questo modo non è necessario che si crei una coda effettiva nel sistema. In particolare, un'orbita non descrive un'area di attesa fisica realmente presente, ma al contrario rappresenta un insieme di clienti che attendono di ritornare nel sistema. Questo tipo di code sono molto usante negli ambiti di telecomunicazioni e di computer networks.

Artajelo cerca dunque di integrare i risultati ottenuti dalla teoria delle code nell'ambito delle "retrial queue" con quelli della teoria inventory al fine di ricavarne delle relazioni e delle soluzioni numeriche. Questo studio non entra meramente nel merito dei sistemi coda-inventory per come sono stati presentati, poiché per semplicità non considera il tempo di servizio positivo. Ciononostante, lo studio di Artajelo risulta essere la base di partenza verso lo studio di sistemi coda-inventory con retrial queue. Artajelo considera un modello inventory con politica (s, S) single-item; al tempo $t = 0$ il magazzino contiene il massimo numero di oggetti S e appena raggiunge il livello s, viene richiesto il rifornimento di una quantità pari a $S - s$. Il valore di S viene scelto in modo tale che sia maggiore di $2s$ per evitare richieste di rifornimento perpetue. Inoltre, il lead time viene rappresentato attraverso una variabile aleatoria esponenziale indipendente e identicamente distribuita di parametro $\gamma > 0$.

Le richieste dei clienti vengono invece modellizzate attraverso un processo di Poisson con tasso $\lambda > 0$ e vengono servite finché il magazzino dispone di una quantità di oggetti positiva, indipendentemente dal valore di soglia stabilito per il rifornimento. Quando il livello del magazzino raggiunge il valore nullo, i clienti che trovano il sistema non disponibile lasciano

temporaneamente l'area di servizio e provano ripetutamente a richiedere un servizio finché non viene rifornito l'inventory e conseguentemente accattata la richiesta.

Le richieste ripetute e rifiutate vengono considerate parte di uno stesso gruppo, chiamato "retrial group", in cui si considera che ciascuna richiesta sia indipendente dalle altre. Dato j il numero di richieste ripetute presenti nel "retrial group" al tempo t, la probabilità di avere un tentativo di richiesta durante il tempo $(t, t + dt)$ è $j\alpha + o(dt)$.

In questo modello viene inoltre assunto che i tempi di interarrivo delle richieste, il lead time e il tempo di ripetizione delle richieste rifiutate sono mutualmente indipendenti. In questo modo lo stato del sistema al tempo t può essere descritto attraverso il processo

$$X = \{\big(I(t), R(t)\big); t > 0\}$$

dove $I(t)$ rappresenta il numero di oggetti contenuto nel magazzino, mentre $R(t)$ il numero di richieste rifiutate tra quelle del "retrial group" al tempo t. Sotto le precedenti assunzioni il processo X è trattabile attraverso il metodo delle catene di Markov.

Una delle osservazioni principali relative al comportamento del processo X, è che esso si comporta come se fosse un sistema M/M/m con retrial queue dove le partenze dei clienti avvengono solo se il numero di utenti nell'area di servizio è maggiore di un valore $m', 0 < m' \leq m$, e i clienti vengono serviti in gruppi di capacità m'.

Altri ricercatori riprendono in mano il modello considerato da Artalejo per studiarne gli aspetti caratteristici. Un esempio è quello di Ushakumari [5], che ricava una soluzione esatta per l'analisi della distribuzione di probabilitià del sistema in considerazione. Successivamente questo modello viene esteso ai casi di sistemi coda-inventory, come ad esempio nello studio di Krishnamoorty e Jose [6], che analizza il comportamento di un sistema M/M/1 caratterizzato da una retrial queue con annesso un inventory e con tempo di servizio positivo.

2.2 Sistemi coda-inventory con multiserver e retrial queue.

Tutti i modelli considerati fino ad ora vengono accomunati dalla presenza di un solo servitore che gestisce le richieste dei clienti. Ma esiste un altro ambito di ricerca legato ai sistemi coda-inventory relativo ai sistemi multiserver, ovvero ai sistemi con un numero maggiore di servitori. Tramite lo studio di Artalejo et al. [4] è stato possibile osservare come i sistemi inventory con retrial queue si comportano come se fossero sistemi M/M/m, dove m è il numero di servitori.

Proprio da questa osservazione cominciano ad essere studiati più approfonditamente i sistemi coda-inventory con multiserver. Uno studio particolarmente interessante è quello di Nair et al. [7] che oltre a proporre un modello coda-inventory multiserver con retrial queue, introduce un approccio completamente diverso nella formalizzazione del modello.

La maggior parte dei modelli relativi ai sistemi coda-inventory si basa sull'idea che un magazzino viene servito ai clienti tramite un'area di servizio. Ci sono alcuni casi però in cui un cliente che entra in un sistema si serve autonomamente. Questo è il caso di sistemi inventory "self-serviceable". Sistemi self-service sono molto utilizzati al giorno d'oggi in particolare negli store online dove ciascun cliente procede autonomamente all'acquisto del prodotto desiderato. Questo tipo di sistema viene modellizzato da Nair assumendo che l'inventory stesso sia l'area di servizio. Nello specifico, quando un cliente lascia il sistema dopo il servizio, anche il suo servitore, ovvero l'oggetto del servizio, lascia il sistema, comportando la riduzione del numero di servitori di una unità e conseguentemente la diminuzione degli oggetti del magazzino. Questo approccio risulta essere profondamente diverso sia rispetto ai sistemi inventory classici sia a quelli multiserver della teoria delle code in quanto nel modello considerato il numero di servitori è variabile.

La formalizzazione del modello utilizzato da Nair comincia dunque considerando un inventory gestito con politica (s, S), senza lead time e in cui l'inventory si comporta come un insieme di servitori. L'arrivo dei clienti nel sistema viene modellizzato tramite una variabile aleatoria Poissoniana con tasso di arrivo λ, tali clienti vengono serviti ogni volta che è presente un servitore libero. Il tempo di servizio delle richieste dei clienti è invece esponenzialmente distribuito con tasso di servizio μ. Al termine di ciascun servizio il cliente e il servitore lasciano il sistema in modo tale che il numero totale di servitori, ovvero di oggetti nel magazzino, vengano ridotti di una unità. Appena il livello dell'inventory raggiunge il valore s, il magazzino viene riempito immediatamente fino al valore massimo S.

Un cliente, che arriva nel sistema e trova tutti i servitori occupati, entra in un buffer chiamato orbita, nel rispetto della struttura con "retrial queue" presentata precedentemente. Dall'orbita un cliente prova ripetutamente a richiedere un servizio; ciascun tentativo viene effettuato dopo un intervallo di tempo esponenzialmente distribuito di parametro λ_r. Nel momento in cui un cliente presente nell'orbita trova un servitore disponibile richiede il servizio, al completamento del quale lascia il sistema.

Per descrivere matematicamente il modello vengono considerati $N(t)$ il numero di clienti presenti nell'orbita e $A(t)$ il numero totale di servitori nel sistema, liberi o occupati, al momento

t. Inoltre viene definito $B(t)$ il numero di servitori occupati nel sistema. Si osserva che $B(t)$ è limitato superiormente da $A(t)$. In questo modo il modello viene descritto tramite un processo tridimensionale,

$$\Omega = \{N(t), A(t), B(t)\}.$$

Il sistema, dunque, viene descritto attraverso un processo stocastico continuo avente uno spazio

$$E = \{(n, a, b); n \geq 0, a = s+1, s+2, \dots, S \; e \; b \leq a\}.$$

Durante la trattazione, Nair procede con la definizione delle condizioni di stabilità ed ergodicità del sistema e propone un problema di ottimizzazione allo scopo di ridurre il costo totale indotto dai tempi di attesa, gestione e di rifornimento.

Dati:

- E_0 il numero medio di utenti nell'orbita;
- E_a il numero medio di servitori nel sistema, liberi e occupati;
- C_w il costo di attesa di un utente nell'orbita per unità di tempo;
- C_h il costo di stoccaggio di un oggetto, servitore, per unità di tempo;
- C_0 il costo di un rifornimento;
- $E(\tau)$ il tempo atteso di un "recycle time", ovvero il tempo che trascorre tra due rifornimenti consecutivi;

Il costo complessivo del sistema può essere descritto come:

$$TC = E_0 C_w + E_a C_h + \frac{C_0}{E(\tau)}$$

Il problema di ottimizzazione si concretizza nel trovare il valore minimo s ottimo tale da minimizzare TC. Posto:

- $\lambda = 2$;
- $\mu = 3$;
- $\alpha = 2$;
- $C_w = 10$;
- $C_h = 2$;
- $C_0 = 0$;
- $S = 4$;

Si osserva, Figura 5, come all'aumentare della soglia di rifornimento dell'inventory aumenta il numero di servitori/oggetti presenti nel magazzino e conseguentemente si riduce molto velocemente il numero di utenti nell'orbita. Il numero medio di servitori occupati E_b cresce inizialmente finchè $s = 1$, per poi portarsi ad un valore costante inferiore.

Inoltre, viene stimato che il valore di soglia ottimale che minimizza il costo totale del sistema è $s = 2$, Figura 6.

s	E_o	E_a	E_b
0	0.8542	2.6634	1.1987
1	0.1371	3.4787	1.4966
2	0.0284	3.5005	0.6667
3	0.0064	4	0.6667

Figura 5: Metriche di prestazione del sistema.

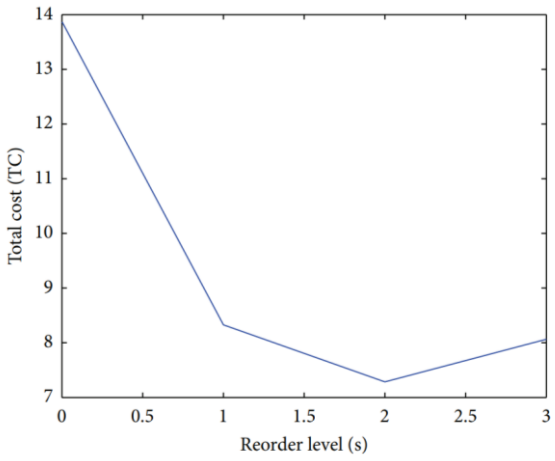

Figura 6: Andamento del costo totale del sistema al variare della soglia di rifornimento.

2.3 Sistemi coda-inventoy con perishable items and server vacation

Generalmente nei modelli classici dei sistemi coda-inventory viene ignorato il contenuto specifico degli oggetti del magazzino e si assume che questi abbiano una durabilità infinita e dunque che non si deteriorino con il passare del tempo. Esiste però una particolare classe di sistemi coda-inventory che invece tiene conto anche del "tempo di vita" degli oggetti immagazzinati. Questi sistemi prendono il nome di "perishable inventory systems".

Nei sistemi reali è fondamentale tenere in considerazione il tempo di vita degli oggetti nella gestione dei magazzini; il deterioramento della merce, infatti, può avere un grande impatto sui costi complessivi e dunque sul tipo di gestione da applicare: da un lato immagazzinare una grande quantità di oggetti può portare a grosse perdine nel caso di scadenza e deterioramento, dall'altro, l'immagazzinamento di una quantità di oggetti limitata può essere controproducente nel caso in cui tale quantità non sia sufficiente a soddisfare le richieste dei clienti. Un esempio emblematico di sistemi reali in cui è necessario tenere in considerazione anche della durata degli oggetti è quello della gestione delle banche del sangue dove il corretto mantenimento delle donazioni è fondamentale ed estremamente importante. Ma esempi di questo tipo si possono ritrovare anche in situazioni più comuni come, ad esempio, nei sistemi di gestione e trasporto di prodotti alimentari.

Nei modelli della teoria inventory viene fatta una distinzione in base a come il sistema considera il deterioramento degli oggetti. Generalmente si considerano due casi, il primo consiste nello stabilire che gli oggetti si deteriorano uno alla volta, come ad esempio certi tipi di alimenti; il secondo, invece, assume che tutti gli oggetti di un unico gruppo, come ad esempio un lotto di medicine con stessa scadenza, si deteriorano immediatamente dopo una certa quantità di tempo.

Uno dei primi studi ad introdurre la caratteristica di deterioramento degli oggetti, tipica dei sistemi inventory, nell'analisi di un sistema a coda è quello di Melikov et al. [8]. Questo studio prende in esame un modello inventory con tempo di servizio nullo a cui viene annessa una "retrial queue".

Tra i modelli coda-inventory con "perishabel items" una classe particolarmente importante è però quella che assume contemporaneamente che il servitore possa prendere una pausa a seconda del livello dell'inventory e/o del numero di richieste nel sistema. Questa caratteristica prende il nome di "server vacation".

Uno degli studi più caratteristici riguardante sistemi coda-inventory con perishable items e server vacation è quello di Koroliuk e Melikov in [9]. In particolare, la trattazione presa in esame analizza il sistema attraverso due modelli, uno in cui la coda dei clienti è limitata e uno in cui questa è illimitata. Inoltre, il comportamento del servitore segue la seguente logica: se concluso l'ultimo servizio non ci sono utenti in coda, il servitore può prendere una pausa, al termine della quale ritorna disponibile solo se il numero di clienti nel sistema è superiore ad una soglia stabilita.

Come nei modelli presentati precedentemente, il sistema preso in considerazione in [9] è caratterizzato pima di tutto da un inventory limitato di capacità S gestito da un unico servitore.

Il flusso di arrivo dei clienti viene modellizzato tramite una variabile aleatoria di Poisson di parametro λ e, a prescindere dello stato del servitore e del livello del magazzino, i clienti in arrivo vengono indirizzati verso una coda gestita con disciplina FIFO. Per quanto riguarda l'area di accodamento, vengono considerati due modelli: uno con una coda finita, in cui se un cliente che entra nel sistema trova davanti a sé N clienti, $1 < N < \infty$, questo non viene considerato, e uno con coda illimitata. Il tempo di servizio è rappresentato da una variabile aleatoria esponenziale indipendente e identicamente distribuita con parametro $\mu < \infty$. Inoltre, ogni cliente può richiedere un solo oggetto dall'inventory che si riduce di una unità al termine di ciascun servizio.

Oltre che per il servizio, il numero di oggetti presenti nel magazzino si riduce conseguentemente al deterioramento, comportamento che viene considerato indipendente da ciascun oggetto e che avviene dopo una certa quantità di tempo casuale descritta da una variabile aleatoria esponenziale di parametro $v > 0$. Viene assunto inoltre che un oggetto può deteriorarsi durante il servizio.

La politica di rifornimento dell'inventory è la (s, S), dove $s < S$ è la soglia sotto la quale il sistema richiede un rifornimento di una quantità pari a $S - s$. Per evitare rifornimenti ripetuti viene scelta la soglia minima di rifornimento s in modo tale che $s < S/2$. Inoltre, il lead time viene descritto tramite una variabile aleatoria esponenziale che dipende dallo stato del servizio: se il servitore si trova in pausa il parametro della variabile aleatoria è γ_0, altrimenti è γ_1.

Per quanto riguarda il comportamento del servizio, viene assunto che se alla fine dell'ultimo servizio è presente almeno un cliente nel sistema e il magazzino è disponibile, il servitore sceglie istantaneamente il prossimo cliente e comincia a servirlo. Altrimenti, se non sono presenti clienti in coda allora il servitore può entrare in pausa a prescindere dal livello del magazzino. Il tempo della pausa del servitore viene descritto attraverso una variabile aleatoria esponenziale di parametro $\beta < \infty$. Al termine della pausa il servitore torna disponibile se nel sistema sono presenti almeno $r \geq 1$ clienti e se il numero di oggetti dell'inventory è positivo, altrimenti torna in pausa.

Un'altra caratteristica presa in considerazione è che i clienti sono impazienti, ovvero possono decidere di abbandonare il sistema mentre si trovano in coda di attesa. Il grado di impazienza dei clienti dipende dallo stato del servitore; nello specifico, il tempo di attesa minimo accettabile dai clienti è modellato attraverso un variabile aleatoria esponenziale che, in caso di servitore in pausa ha un parametro α_0, altrimenti α_1, $\alpha_0 \neq \alpha_1$.

Il problema proposto in [9] è quello di trovare una distribuzione congiunta che tenga in considerazione il livello dell'inventory, il numero di clienti e lo stato del servizio. Questo permette di calcolare il livello medio dell'inventory S_{aw}, il tasso di deterioramento dell'inventory Γ_{av}, ovvero il numero medio di oggetti che si deteriorano per unità di tempo, il tasso di rifornimento RR, ovvero il numero medio di richieste di rifornimento per unità di tempo, la probabilità del servitore di entrare in pausa P_{vac}, la probabilità di perdita dei clienti PL, e la probabilità di perdita dei clienti in coda a causa dell'impazienza PL_{av}.

Dalla formalizzazione del modello viene determinato che il processo del sistema è una catena di Markov tridimensionale. Infatti, è possibile descrivere lo stato del sistema in un istante di tempo casuale attraverso un processo stocastico rappresentabile tramite un vettore tridimensionale $n = (n_1, n_2, \theta)$, dove n_1 e n_2 rappresentano rispettivamente il livello dell'inventory e il numero di clienti nel sistema; mentre θ rappresenta lo stato del servitore.

Nello specifico:

$$\theta = \begin{cases} 0 & nel\ caso\ di\ servitore\ in\ pausa, \\ 1 & nel\ caso\ di\ servitore\ disponibile. \end{cases}$$

A questo punto lo spazio degli stati del sistema può essere descritto in questo modo:

$$E = E_0 \cup E_1, E_0 \cap E_1 = \emptyset$$

Dove:

$$E_0 = \{n: n_1 = 0, 1, \dots, S; n_2 = 0, 1, \dots, N; \theta = 0\}$$

$$E_1 = \{n: n_1 = 0, 1, \dots, S; n_2 = 0, 1, \dots, N; \theta = 1\}$$

Nel modello con coda limitata viene posto $N < \infty$, mentre nel modello con coda illimitata $N = \infty$.

CAPITOLO 3
AMBITI DI RICERCA APERTI

3.1 Ambiti di ricerca attuali

Gli studi sui sistemi coda-inventory, come visto nei capitoli precedenti, si sviluppano attraverso la formalizzazione dei relativi modelli matematici che successivamente vengono analizzati per cercarne le condizioni di stabilità e i parametri fondamentali. Uno dei parametri fondamentali è il tempo di attesa, questo infatti è uno dei parametri da tenere in considerazione al fine di ridurre i costi complessivi di qualunque sistema che si prende in considerazione. Per questo motivo la ricerca sui sistemi coda-inventory ha posto grande attenzione sulla riduzione del tempo di attesa totale dei sistemi.

Nel merito delle soluzioni e dei problemi analizzati si possono però osservare alcune lacune. Innanzitutto, lo studio dei sistemi multiserver permette di poter valutare teoricamente se l'aumento del numero di servitori in un sistema migliora le prestazioni del servizio con conseguente riduzione del tempo di attesa. Da un punto ti vista pratico però, si può osservare che la maggior parte dei sistemi coda-inventory reali utilizzano un unico servitore, ad esempio basti pensare alle gelaterie, dove generalmente i clienti vengono serviti uno alla volta in ordine di arrivo. Questi sistemi raramente vengono estesi a sistemi multiserver, sia per mancanza di spazio effettivo, sia per una questione prettamente economica.

Inoltre, nella maggior parte dei modelli relativi ai sistemi coda-inventory viene assunto che il tempo di servizio sia un processo omogeneo, spesso esponenziale. Nei sistemi reali il tempo di servizio omogeneo è una assunzione molto azzardata, ma utile per approssimarne il comportamento. Nella maggior parte di questi sistemi però, il tempo di servizio è molto variabile e dipende sia dal tipo di richieste dei clienti e dal loro comportamento, sia dalla predisposizione del servitore in quell'istante.

Nell'ambito dell'economia e del marketing, sempre più importanza viene data alla figura del cliente, che da una posizione prettamente ricettiva, ora acquisisce un ruolo predominante. Infatti, il cliente di una qualsiasi azienda al giorno d'oggi ha la possibilità di scegliere sia tra diverse tipologie di prodotto, sia tra diverse aziende che offrono lo stesso prodotto. Fidelizzare i clienti diventa dunque un obiettivo importante per qualunque azienda che voglia restare

competitiva nel mercato e che voglia distinguersi tra i suoi competitors. A questo scopo l'attenzione dell'azienda si sposta sul soddisfacimento dei bisogni dei propri clienti. Uno degli aspetti caratteristici nella soddisfazione del cliente, oltre alla qualità dei prodotti, è la qualità del servizio offerto, che, tra i vari parametri, si concretizza anche nel tempo che un cliente deve attendere prima di essere servito. Il tempo di attesa, dunque, non è più semplicemente un parametro da ottimizzare al fine di aumentare il profitto o di minimizzare le spese, ma diventa un elemento caratterizzante dell'azienda stessa. Per questi motivi, se l'intenzione è quella di analizzare nel dettaglio l'andamento del sistema e di ridurre il tempo di attesa corrispondente, è necessario tenere in considerazione anche il comportamento dei clienti nel sistema e dunque assumere un tempo di servizio omogeneo non è più sufficiente.

In una situazione reale, si possono osservare alcuni comportamenti standard che i clienti manifestano quando decidono di acquistare un qualche tipo di oggetto. Alcuni clienti per evitare di trovarsi in coda preferiscono chiedere preventivamente un appuntamento o prenotare ciò di cui hanno bisogno. Altri preferiscono invece attendere in coda ma alcuni sono indifferenti al tempo che questo può comportare, mentre altri dopo una certa quantità di tempo diventano impazienti e lasciano il sistema. Tra i clienti che preferiscono attendere in coda vengono formalmente modellizzati tre tipologie di cliente:

- Balking, sono i clienti che decidono di non aspettare in coda se la coda è troppo lunga;
- Reneging, sono i clienti che lasciano la coda se hanno atteso troppo tempo in coda;
- Jockeying, sono i clienti che cambiano la coda se pensano che potrebbero essere serviti più velocemente.

Oltre a ridurre il tempo di attesa è dunque fondamentale cercare di ridurre al minimo le perdite dei clienti che approcciano il sistema. Una delle strategie molto utilizzate nei modelli e ampiamente studiata è, come visto precedentemente, quella dell'utilizzo di "retrial queue" che riduce il numero di clienti "balking"; concretamente questa strategia descrive semplicemente il fatto che un cliente possa ritornare nel sistema dopo una certa quantità di tempo. Ma una strategia ancora più efficace è quella di considerare il tempo di servizio dipendente dalla lunghezza della coda.

L'unico modo per ridurre i tempi di attesa considerando l'impazienza dei clienti e il loro comportamento di fronte ad una coda è, infatti, quello di considerare un sistema con un tempo di servizio che dipende dalla coda stessa. Questo da un lato limita la perdita di clienti (balking e reneging), dall'altro aumenta il numero totale di clienti nel sistema. Se infatti il tempo di attesa si riduce, in un sistema reale si osserva un aumento nell'arrivo di nuovi clienti; questo aspetto

però non viene considerato nei modelli in quanto generalmente l'arrivo dei clienti viene modellizzato attraverso una variabile aleatoria il cui parametro non dipende dal tempo di attesa.

Modelli che introducono un servizio dipendente dalla coda non sono stati ancora studiati ampiamente e rappresentano uno degli ambiti di ricerca attuali. Uno degli ultimi studi che analizza un modello con tempo di servizio non omogeneo è quello di Jaganathan et al. [10] che descrive un sistema coda-inventory single server e con tasso di servizio dipendente dalla coda.

Il modello scelto per studiare questo sistema consiste in sistema coda-inventory con un servitore unico, una coda finita di capacità N, un inventory con capacità massima di S oggetti.

Viene assunto che il servizio sia sempre occupato se il numero di oggetti presenti nel magazzino e il numero di clienti in coda sono valori positivi. Altrimenti il servizio diventa disponibile.

Ogni volta che il servizio è occupato, un cliente che entra nel sistema deve attendere in coda. I clienti che entrano nel sistema per la prima volta vengono chiamati clienti prioritari e vengono modellizzati attraverso un processo degli arrivi Poissoniano di parametro λ_p.

Per evitare la perdita di clienti, viene considerata anche la presenza di un'orbita di capacità infinita. Ogni volta che un cliente trova la coda completamente piena, può entrare nell'orbita con probabilità p, oppure può abbandonare il sistema con probabilità complementare $1 - p$. I clienti nell'orbita che trovano la coda con un numero di clienti inferiore a N entrano immediatamente. Altrimenti riprovano ripetutamente ad accedere al servizio tramite un processo esponenziale con tasso di arrivo pari a λ_r seguendo il comportamento di una classica "retrial queue".

Il servizio ha una velocità variabile che cambia in relazione alla lunghezza della coda in modo da ridurre il tempo di attesa. Il processo di servizio viene quindi rappresentato tramite una variabile esponenziale di parametro μ_w, dove $1 \leq w \leq N$, e w è il numero di clienti in coda. I clienti ricevono l'oggetto richiesto dall'inventory solo al termine del servizio corrispondente. Poiché il processo del servizio dipende dalla lunghezza della coda viene considerato un processo eterogeneo.

Il sistema tiene in considerazione altri due aspetti fondamentali una politica di gestione del magazzino e il fatto che gli oggetti possono danneggiarsi e degradarsi col passare del tempo.

Per quanto riguarda la politica di gestione dell'inventory, viene richiesta un rifornimento del magazzino di $Q = S - s$ oggetti ogni volta che il numero di oggetti immagazzinati raggiunge il valore soglia s. Chiamata anche politica (s, Q). Il processo di rifornimento segue un comportamento esponenziale di parametro γ.

Per quanto riguarda invece il deperimento degli oggetti viene scelto un processo esponenziale di parametro v che rappresenta i possibili danneggiamenti degli oggetti dovuti alla loro gestione, le scadenze dei prodotti e i possibili difetti di fabbricazione.

Per l'analisi di questo sistema viene inoltre scelto $\mu_w = \mu w^{\alpha}$, dove $0 \leq \alpha \leq 1$. Se $\alpha = 0$ il modello segue un processo di servizio omogeneo, se $\alpha = 1$ il processo è lineare rispetto a w, altrimenti è un processo non omogeneo.

Posto inoltre $P_1(t)$ il numero di clienti nell'orbita, $P_2(t)$ il numero di oggetti presenti nel magazzino e $W(t)$ la lunghezza della coda al tempo t, è possibile descrivere il processo del sistema al tempo t come:

$$\{P(t), t \geq 0\} = \{\big(P_1(t), P_2(t), W(t)\big), t \geq 0\},$$

Il cui spazio delle soluzioni D è tale che:

$$D = \{(p_1, p_2, w) : p_1 = 0, 1, \dots, \infty; p_2 = 0, 1, \dots, S; w = 0, 1, \dots, M\}.$$

3.2 Ambiti di ricerca futuri

Prenotazioni e Overbooking

Si è visto nel paragrafo precedente come l'attenzione della ricerca si è spostata sulla figura del cliente e sul modo in cui considerare il suo comportamento all'interno del sistema. Un'altra caratteristica importante che non è stata ancora introdotta nei modelli coda-inventory è quella delle prenotazioni. Infatti, nella maggior parte dei modelli studiati un cliente che entra nel sistema ordina un certo oggetto esclusivamente nel momento in cui viene servito. L'introduzione delle prenotazioni permette di considerare sistemi in cui i clienti possono prenotare gli oggetti richiesti prima ancora di essere serviti. Questo è infatti una possibilità molto frequente che si ritrova nella maggior parte degli store, sia fisici che online.

In merito a ciò un altro elemento che potrebbe essere introdotto è quello dell'overbooking. Per overbooking si intende la possibilità di prenotare un qualche tipo di oggetto nonostante non sia fisicamente disponibile nel momento della prenotazione. Questo avviene solitamente per gestire le cancellazioni eventuali dei clienti che hanno precedentemente prenotato lo stesso oggetto. L'overbooking avviene frequentemente nelle prenotazioni di voli o di hotel e in certi casi permette di evitare le perdite relative alle cancellazioni all'ultimo momento. Modellare sistemi

coda-inventory con prenotazioni, cancellazioni e overbooking permette di analizzare ancora più nel dettaglio sistemi reali e di ottimizzarne le prestazioni.

Oggetti multipli per il servizio

Nei modelli classici dei sistemi coda-inventory viene solitamente considerato che il magazzino contenga sempre la stessa tipologia di oggetti e dunque che i clienti del sistema corrispondente vengano serviti tutti con gli stessi oggetti. Uno degli ambiti di ricerca aperti riguarda per l'appunto lo studio di sistemi che dispongano di oggetti diversi da servire ai rispettivi clienti che li richiedono. Sempre relativamente a questo ambito ci sono dei problemi ancora aperti sullo studio di sistemi produttivi che per completare un servizio richiedono oggetti diversi. Ad esempio, si consideri l'esempio di un sistema produttivo che richiede due tipi di materie prime diversi R_1 e R_2 da processare ed entrambi necessari per il servizio.

Discrete Systems

A causa della complessità che i modelli dei sistemi coda-inventory possono avere, non è ancora stata posta grande attenzione ai sistemi a tempo discreto, che risultano ancora più complessi di quelli a tempo continuo. Pertanto, non sono ancora presenti risultati particolarmente rilevanti. Risulta quindi ancora un ambito di ricerca aperto che probabilmente verrà studiato e approfondito ulteriormente nei prossimi anni.

CONCLUSIONE

I sistemi coda-inventory nascono da un modello abbastanza semplice caratterizzato da un'area di accodamento, una di servizio e da un magazzino. Ciononostante, il loro studio è molto meno semplice di quanto sembra, in quanto tiene in considerazione contemporaneamente vari aspetti che generalmente vengono trattati indipendentemente. Inoltre, poichè sono richiesti metodi ed approcci che necessitano di competenze molto elevate, è estremamente complesso ottenere delle formule esatte e delle metodologie standard al fine di ricavarne una stima delle metriche fondamentali e del loro andamento.

Ciò che risulta evidente però, è come allo stesso tempo siano sistemi estremamente versatili: possono essere estesi attraverso l'introduzione di caratteristiche e comportamenti particolari e possono essere utilizzati al fine di risolvere problemi di ottimizzazione di natura diversa. Proprio in questo risiede la potenzialità dei sistemi coda-inventory e per questo il loro studio ha acquisito un tale interesse da essere tutt'ora un ambito di ricerca molto attuale.

BIBLIOGRAFIA

[1] A. Z. Melikov e A. A. Molchanov, «Stock Optimization in Transportation/Storage Systems,» *Kibernetika i Sistemnyi Analiz,* pp. 179-182, 1992.

[2] K. Sigman e D. Simchi-Levi, «Light Traffic Heuristic for an M/G/1 Queue with Limited Inventory,» *Annals of Operation Research,* pp. 371-380, 1992.

[3] M. Schwarz, C. Sauer, H. Daduna, R. Kulik e R. Szekli, «M/M/1 Queuing systems with inventory,» *Queuing Systems,* pp. 55-78, 2006.

[4] J. R. Artalejo, A. Krishnamoorty e M. J. Lopez-Herrero, «Numerical analysis of (s, S) inventory systems with repeated attempts,» *Annals of Operations Research,* pp. 67-83, 2006.

[5] P. Ushakumari, «On (s, S) inventory system with random lead time and repeated demands,» *Journal of Applied Mathematics and Stochastic Analysis,* pp. 1-22, 2006.

[6] A. Krishnamoorty e K. P. Jose, «An (s, S) Inventory system with positive lead time, loss and retrial of customers,» *Stochastic Modelling and Applications,* p. 5671, 2005.

[7] A. N. Nair e M. J. Jacob, «(s, S) Inventory with Positive Service Time and Retrial of Demands: An Approach through Multiserver Queues,» *ISRN Operations Research,* pp. 1-6, 2014.

[8] A. Z. Melikov, L. A. Ponomarenko e M. O. Shahmliyev, «Models of Perishable Queuing-Inventory System with Repeated Customers,» *Journal of Automation and Information Sciences,* pp. 22-37, 2016.

[9] V. S. Koroliuk, A. Z. Melikov, L. A. Ponomarenko e A. M. Rustamov, «Models of Perishable Queuing-Inventory Systems with Service Vacation,» *Cybernetics and Systems Analysis,* pp. 31-44, 2018.

[10] K. Jeganathan, S. Selvakumar, N. Anbazhagan, S. Amutha e P. Hammachukiattikul, «Stochastic modeling on M/M/1/N inventory system with queue-dependent service rate and retrial facility,» *AIMS Mathematics,* pp. 7386-7420, 2021.

[11] J. A. White, J. W. Schmidt e G. K. Bennett, Analysis of Queuing Systems, Academic Press, 1975.

[12] E. L. Porteus, «Stochastic Inventory Theory,» *Handbooks in Operation Research and Management Science,* vol. 2, n. 12, pp. 605-652, 1990.

[13] A. Logeswari e M. Kavitha, «M/M/1 queuing model by using inventory theory,» *Malaya Journal of Matematik,* pp. 3803-3806, 2020.

[14] A. Krishnamoorty, D. Shajin e V. C. Narayanan, «Inventory with Positive Service Time: a Survey,» *Queuing Theory 2: Advanced Trends,* pp. 202-237, 2020.

[15] L. Kleinrock, Queuing Systems, vol. I: Theory, John Wiley & Sons, 1975.

[16] N. Benvenuto e M. Zorzi, Principles of Communications Networks and Systems, Wiley, 2011.