

Business Analytics – 2024/25

Sistemi MRP/ERP e approccio JIT

Prof. Paolo Brandimarte

Dip. di Scienze Matematiche – Politecnico di Torino

e-mail: paolo.brandimarte@polito.it

URL: staff.polito.it/paolo.brandimarte

Versione (provvisoria): 11 aprile 2025

NOTA: Per uso didattico interno, laurea magistrale in ingegneria matematica. Non postare e non redistribuire.

Riferimenti

- P. Brandimarte, A. Villa. *Gestione della produzione industriale*. UTET Libreria, 1995.
- W.J. Hopp, M.L. Spearman. *Factory physics* (3rd ed). Waveland Press, 2011.
<https://factoryphysics.com/>
- J.F. Proud, E. Deutsch. *Master Planning and Scheduling: An Essential Guide to Competitive Manufacturing* (4th ed). Wiley, 2022.

Problema: gestione delle scorte

Approcci classici già visti: I. PERIODIC REVIEW: tempo o discreto \rightarrow piccola e grande S

II. CONTINUOUS REVIEW: tempo continuo \rightarrow rifornimento di quantità fissa quando arrivo al punto di rifornimento

Limiti: i prodotti NON sono accoppiati così

Limiti degli approcci classici della gestione delle scorte

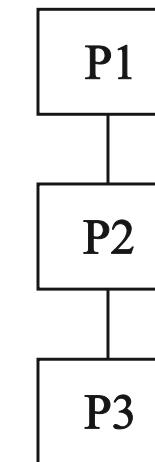
Assumiamo domanda costante nel tempo

I classici approcci di controllo delle scorte presentano limiti severi:

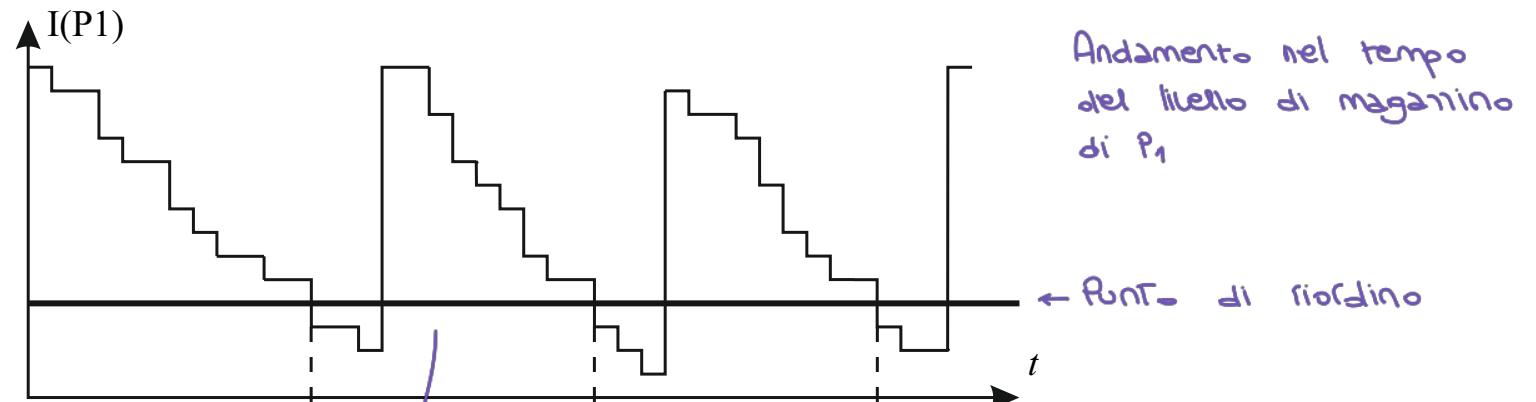
- in ambienti non make-to-stock, come make-to-order e assemble-to-order (ignorano variabilità prevedibile);
- in presenza di vincoli di capacità produttiva (si prestano a problemi retail);
- in presenza di strutture di prodotto complesse (rappresentate mediante una distinta base).
↳ inteso come codice di un elemento che insieme ad altri formerà il prodotto finito

Consideriamo una distinta base banale, con un finito P_1 che richiede un modulo P_2 , che a sua volta richiede un componente P_3 .

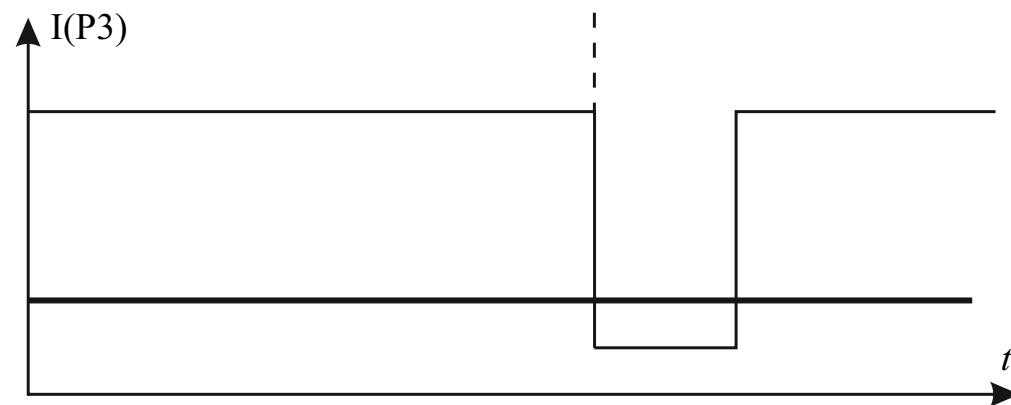
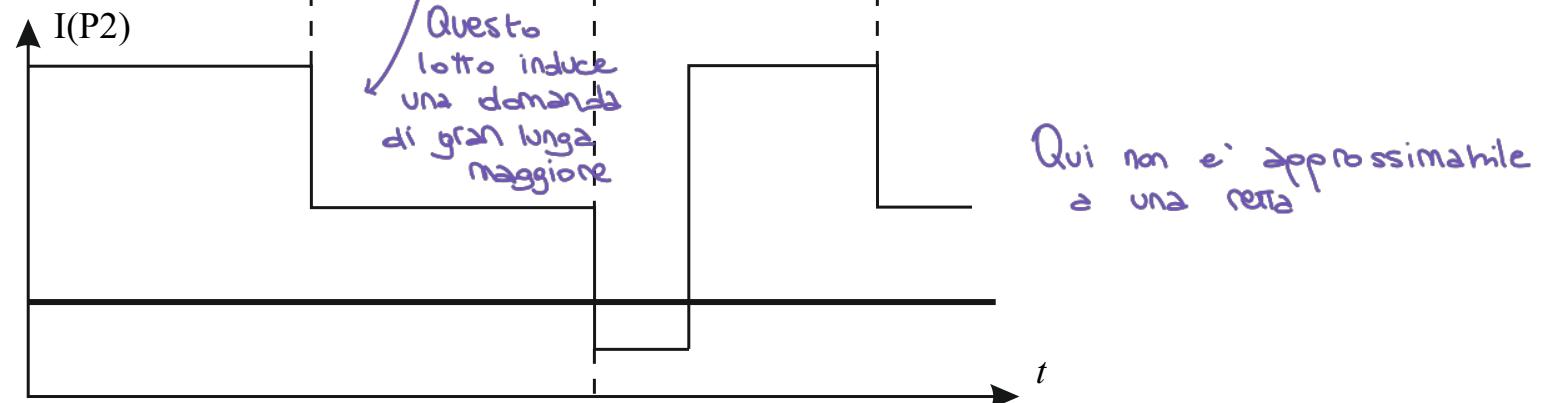
Anche se la domanda per il finito è regolare nel tempo, la propagazione dei fabbisogni lungo la distinta base può indurre un amplificazione della variabilità.



Domanda indipendente:
vicine dal mercato



Domanda dipendente



Sembra un problema banale, ma risolverlo computazionalmente e' molto difficile

Evoluzione: dalla logica MRP ai sistemi ERP

In linea di principio possiamo costruire un modello MILP per il lot-sizing multilivello, in modo da collegare domanda indipendente e dipendente.

Difficile da risolvere, impossibile decenni fa. Negli anni 70 sono nati i sistemi MRP (Material Requirements Planning).

Si cerca di disaccoppiare
i codici, non le domande

Logica a capacità infinita: rilassare il vincolo di capacità implica un disaccoppiamento parziale tra item (non tra domanda dipendente e indipendente). L'interazione viene anticipata (surrogata) da un lead time (fissato a priori).

Evoluzione successiva: sistemi Manufacturing Resource Planning (MRPII). Strumenti RCCP (Rough Cut Capacity Planning) e CRP (Capacity Requirement Planning) per verificare il soddisfacimento dei vincoli di capacità.

Successiva evoluzione a sistemi ERP (Enterprise Resource Planning). Integrazione con parte commerciale e finanziaria.

La logica è a capacità infinita: il vincolo di capacità produttiva non viene esplicitamente considerato, ma viene "surrogato" attraverso il suo effetto principale, cioè la creazione di ritardi, che impongono di lanciare gli ordini di produzione o di acquisto con sufficientemente anticipo rispetto alla domanda.

L'anticipo è rappresentato da un lead time fissato a priori. La tabella seguente illustra il *lead time offsetting* per un lead time di due periodi.

periodo	1	2	3	4	5	6	7	8
fabbisogni					50		60	
ordini pianificati			50		60			

Record MRP:

Se mi serve qds, quando devo emettere l'ordine di produzione?

periodo	1	2	3	4	5	6	7	8
fabbisogni lordi	→ Quello che richiede il cliente							
consegne attese	→ "On order"							
magazzino disponibile	→ "On hand"							
fabbisogni netti	→ E' ciò che poi decide gli ordini pianificati							
ordini pianificati								

Il fabbisogno lordo di P definisce dalla domanda, da P si avanza il fabbisogno lordo di P, etc.

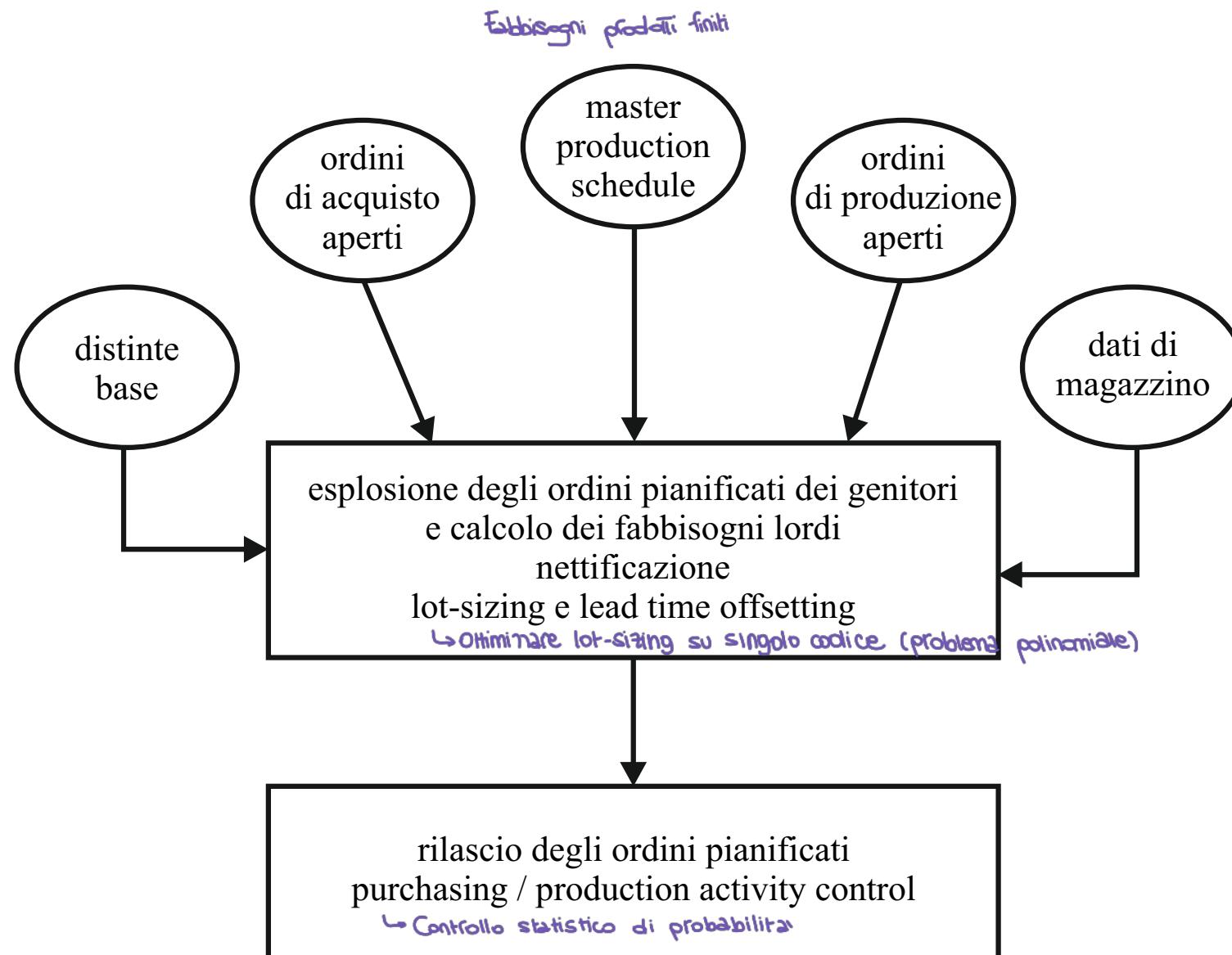
Prima di procedere al lead time offsetting, occorre però calcolare i fabbisogni netti di ciascun codice, tenendo conto di materiali on-hand e on-order.

I fabbisogni netti vengono ottenuti “nettificando” i fabbisogni lordi.

I fabbisogni lordi di un codice dipendono dagli ordini di produzione di tutti i predecessori nelle distinte base.

Un prodotto finale (end item) corrisponde alla radice di una distinta base, e non ha ovviamente dei genitori. I fabbisogni lordi sono rappresentati in questo caso da un Master Production Schedule (MPS). → *Con ordine dei clienti, scheduling, etc
figurda i prodotti finiti*

La logica MRP procede ricorsivamente a partire dagli end item, scendendo lungo le varie distinte base ed “esplodendo” i fabbisogni di ciascun codice, fornendo i tempi di lancio degli ordini a tutti i livelli. Gli ordini pianificati per ciascun codice costituiscono i fabbisogni lordi per i suoi componenti.





Esempio di applicazione della logica MRP

Un prodotto finito P1 richiede un semilavorato P2; la produzione di un pezzo P2 richiede a sua volta due pezzi P3.

In magazzino sono disponibili 10 unità P1 e 20 unità P2; inoltre è attesa la consegna di 20 P2 all'inizio della terza settimana.

Il lead time è di una settimana per P1, due per P2 e tre per P3.

Il dimensionamento dei lotti è banale per P1 e P2: i fabbisogni netti (questa regola di lot sizing è nota come lot-for-lot). Nel caso di P3, gli ordini hanno un vincolo: occorre ordinare quantità multiple di 50.

Fai quanto hai effettivamente bisogno

CODICE P1

periodo	1	2	3	4	5	6	7	8	
fabbisogni lordi					50	60			MPS
consegne attese									
magazzino disponibile	10	10	10	10	10	0	0	0	
fabbisogni netti					40	60			Quello che devi procurare
ordini pianificati					40	60			

CODICE P2

periodo	1	2	3	4	5	6	7	8
fabbisogni lordi				40		60		
consegne attese → Ordine che sta arrivando			20					
magazzino disponibile	20	20	20	40	0	0	0	0
fabbisogni netti					60			
ordini pianificati → Quando ordinare			60					

CODICE P3

periodo	1	2	3	4	5	6	7	8
fabbisogni lordi				120				
consegne attese								
magazzino disponibile	0	0	0	0	30	30	30	30
fabbisogni netti				120				
ordini pianificati	150							

x2 per necessita' di distinta base

E' un ordine esecutivo di acquisto perche' e' nel primo periodo

È importante comprendere la *dinamica* di un sistema MRP. Gli ordini pianificati non sono ordini esecutivi. Gli ordini pianificati dell'action bucket, una volta rilasciati, vengono trasformati in consegne attese.

CODICE P3

periodo	1	2	3	4	5	6	7	8
fabbisogni lordi				120				
consegne attese					150			
magazzino disponibile	0	0	0	0	30	30	30	30
fabbisogni netti								
ordini pianificati								

Può capitare che andando avanti nel processo alla fine si arrivi a ordini che avrei dovuto fare prima rispetto al periodo in cui sono ora

Un'ulteriore differenza tra ordini pianificati e operativi è che quando un ordine di produzione viene rilasciato vengono modificati i record relativi alle giacenze di magazzino dei componenti; una parte di giacenza viene allocata per l'ordine, ed è da considerarsi non disponibile nel calcolo dei fabbisogni netti.

→ Anche se sono in magazzino NON e' detto che tu possa usarli → bisogna "prenotarli"

I pacchetti MRP permettono di specificare un orizzonte temporale di release; solo gli ordini che cadono all'interno di questo orizzonte andrebbero rilasciati, in quanto gli altri sono soggetti a incertezze eccessive.

→ Non partì troppo in anticipo perché sei troppo vulnerabile alle incertezze

Nel calcolo dei fabbisogni è possibile tenere conto di scorte di sicurezza; in questo caso, i fabbisogni netti vengono generati non quando il magazzino disponibile va sotto zero, ma quando va sotto un certo livello di soglia.

Esiste una gamma di regole di lot-sizing, a quantità fissa o variabile, oppure euristiche per la minimizzazione dei costi totali (giacenza e ordinazione).

Il problema del nervosismo

Il dimensionamento dei lotti a partire dai livelli alti della distinta base può avere effetti sgradevoli e non intuitivi nel caso di variazioni dell'MPS.

Le regole a quantità variabile sono soggette ad un fenomeno detto nervosismo (*nervousness*), per cui piccole variazioni nell'MPS possono avere effetti rilevanti sul processo di calcolo dei fabbisogni.

Consideriamo una semplice distinta base a due livelli composta da un prodotto finale P1, il cui lead time è di 2 periodi, e da una singola materia prima P2, il cui lead time è di 4 periodi. Assumiamo che la regola di lot sizing scelta sia un periodo di ricopertura pari a 5 periodi. → *Regola semplice a tasso variabile basata sul lot sizing classico*

CODICE P1

periodo	1	2	3	4	5	6	7	8
fabbisogno lordo	2	24	3	5	1	3	4	50
consegne attese								
magazzino disponibile	28	26	2	13	8	7	4	0
ordini pianificati	14	→ Copre 5 periodi	Non e' detto che sia veramente la scelta migliore	50				

CODICE P2 → Lead time 4 periodi

periodo	1	2	3	4	5	6	7	8
fabbisogno lordo	14				50			
consegne attese	14							
magazzino disponibile	2	2	2	2	2	0	0	0
ordini pianificati	48							

Cambiamo le regole, da 24 a 23 quinoli ordino un po' in meno

Supponiamo che il fabbisogno lordo di P1 durante il secondo periodo scenda da 24 a 23 unità.

CODICE P1

periodo	1	2	3	4	5	6	7	8
fabbisogno lordo	2	23	3	5	1	3	4	50
consegne attese								
magazzino disponibile	28	26	3	0	58	57	54	50
ordini pianificati			63					0

Ora scatta il problema!

CODICE P2

periodo	1	2	3	4	5	6	7	8
fabbisogno lordo			63					
consegne attese		14						
magazzino disponibile	2	16	-47	—> Questi li avrei dovuti ordinare 2 settimane fa: problema!				
ordini pianificati	47							

Il risultato ottenuto è anti-intuitivo: diminuendo i fabbisogni, abbiamo generato un ordine urgente, in quanto l'ordine per P2 è in ritardo di due settimane; paradossalmente, non avremmo avuto alcun problema se il fabbisogno lordo di P1 fosse stato incrementato, anziché diminuito, di una unità.

Un altro punto da considerare è l'effetto di bordo dovuto alla ripianificazione rolling horizon, in cui si aggiunge un time bucket all'orizzonte di pianificazione.

L'aggiunta del fabbisogno relativo a tale time bucket può alterare l'accorpamento dei fabbisogni, con esiti imprevedibili.

Esistono diversi modi per evitare il fenomeno del nervosismo. L'adozione di regole a quantità fissa permette di filtrare naturalmente variazioni di piccola entità, al costo di un aumento nel livello delle scorte.

È anche possibile adottare strategie di gestione differenziata dell'orizzonte temporale di pianificazione, dette strategie di **time fencing**. → Si mette una "siepe" nel MPS all'inizio, cioè un portone di ordine congelata

Nell'immediato non è possibile cambiare nulla nell'MPS; in un periodo intermedio è possibile alterare l'MPS dopo specifica analisi e autorizzazione; è invece possibile cambiare l'MPS a piacere nei periodi più lontani.

Infine, uno strumento utile per evitare il nervosismo, e per risolvere situazioni anomale, è l'uso dei **firm planned orders**; si tratta di ordini che non possono essere modificati dall'MRP quando le condizioni cambiano, ma soltanto dietro istruzione del pianificatore.

Master Production Scheduling

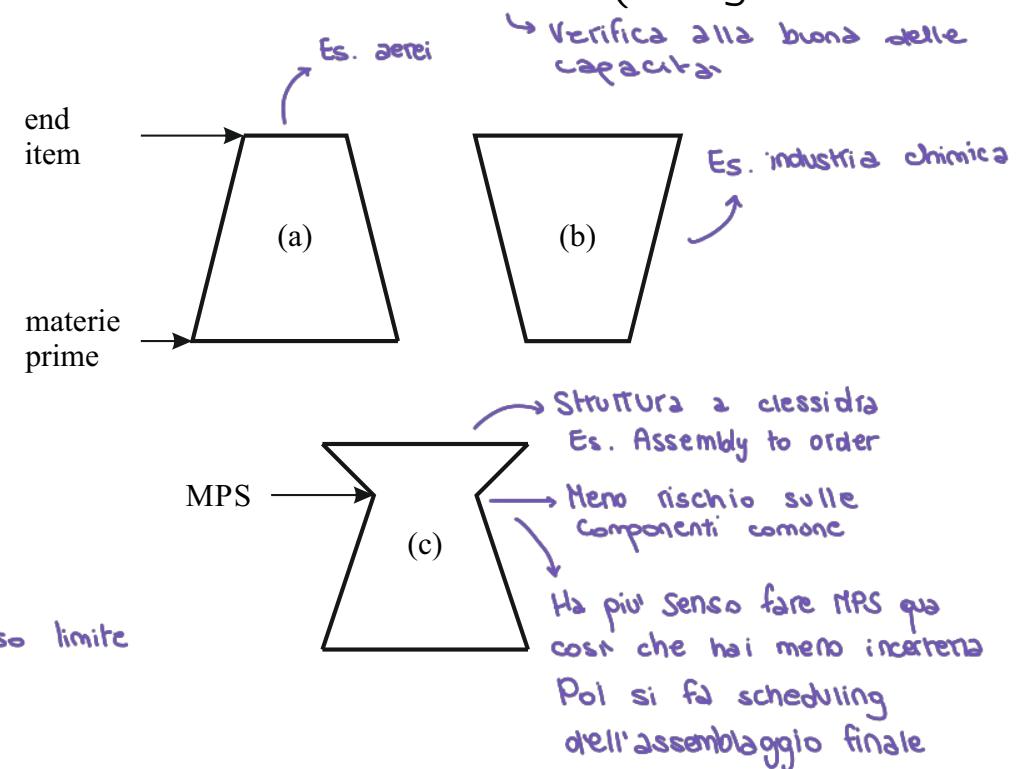
Che tipo di distinta base hai?

L'MPS è l'input primario per la logica MRP, e si basa in parte su ordini cliente, in parte su forecasting (demand planning).

Nella vecchia logica MRP, esso può essere validato da moduli RCCP (Rough Cut Capacity Planning).

Non è detto che l'MPS venga costruito per i codici alla radice delle distinte base. Può esserci domanda indipendente per parti di ricambio, e in ambito ATO può essere preferibile avere due livelli (MPS a livello di moduli e Final Assembly Scheduling a livello alto).

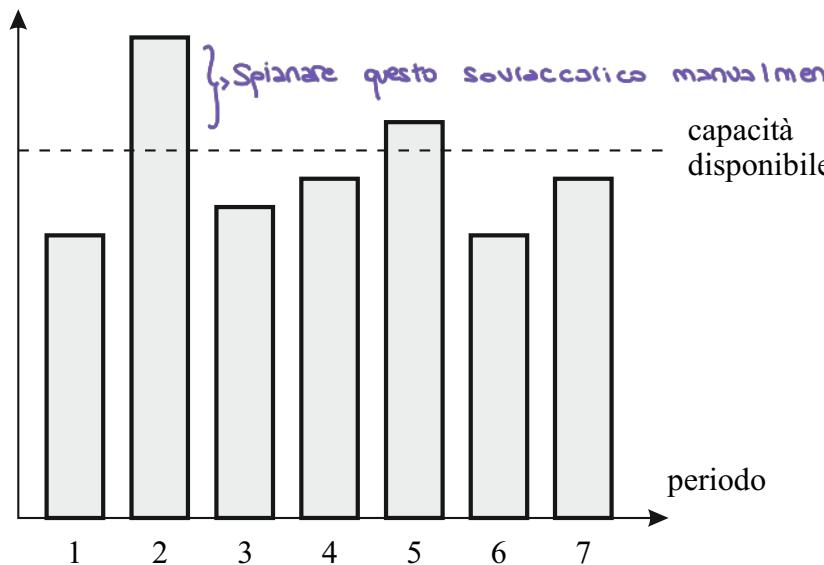
Ma questa logica c'è a capacità infinita → che è un grosso limite



Capacity Requirements Planning (CRP)

La logica MRP è a capacità infinita. È possibile fare una verifica a posteriori del carico di lavoro rispetto alla capacità effettivamente disponibile, mediante moduli CRP.

Idea: fai a posteriori un check carico sulla capacità disponibile (ore macchina)



Non è facile risolvere (manualmente) eventuali non ammissibilità (a meno di usare modelli di ottimizzazione o comunque procedure euristiche a capacità finita).

Inoltre, per evitare ritardi si tende a gonfiare il lead time presunto, creando work in process, che a sua volta allunga i lead time, creando un potenziale circolo vizioso.

↳ c'è l'anticipazione del flow time "verso"

↳ Dipende - tempi di attesa in coda
- tempi di lavorazione

Factory physics: la legge di Little

A livello shop floor le misure di prestazione essenziali sono:

1. Throughput (e.g., pezzi per unità di tempo). → Quanti pezzi che attraversano il sistema
2. Flow time (legato al lead time; comprende tempi di attesa in coda, lavorazione, movimentazione). → Cioè il waiting time
3. Work in process (WIP; legato ai materiali in coda). → Lunghezza della coda

Idealmente, vorremmo throughput alto, con flow time e WIP bassi. È possibile? Quale è l'impatto della variabilità (prevedibile e imprevedibile)?

Un risultato fondamentale nella teoria delle code è la legge di Little:

$$\text{WIP} = \text{throughput} \times \text{flow time}$$

→ Vale in generale e anche per reti di code

La legge esprime la relazione tra WIP e flow time.

Il ruolo della variabilità nel flusso dei materiali

Consideriamo una singola macchina, con un buffer per il WIP:

- W_q waiting time in coda; → E' il valore medio
- t_s tempo medio di lavorazione (servizio);
- μ tasso medio dei servizi (pezzi per unità di tempo: $\mu = 1/t_s$)
- λ tasso medio degli arrivi, che in equilibrio è il throughput;
- L lunghezza media della coda (WIP).

La legge di Little può essere espressa come $L = \lambda \times (W_q + t_s)$.

Misuriamo l'utilizzazione del sistema come $u = \lambda/\mu$ (limitata tra 0 e 1).

$$\hookrightarrow \lambda \ll \mu$$

Una coda G/G/1 non è trattabile analiticamente, ma una formula approssimata (che è esatta nel caso M/M/1) fornisce: Effetto della variabilità, Fattore di congestione $u \rightarrow 1$ ($\frac{u}{1-u} \rightarrow \infty$)

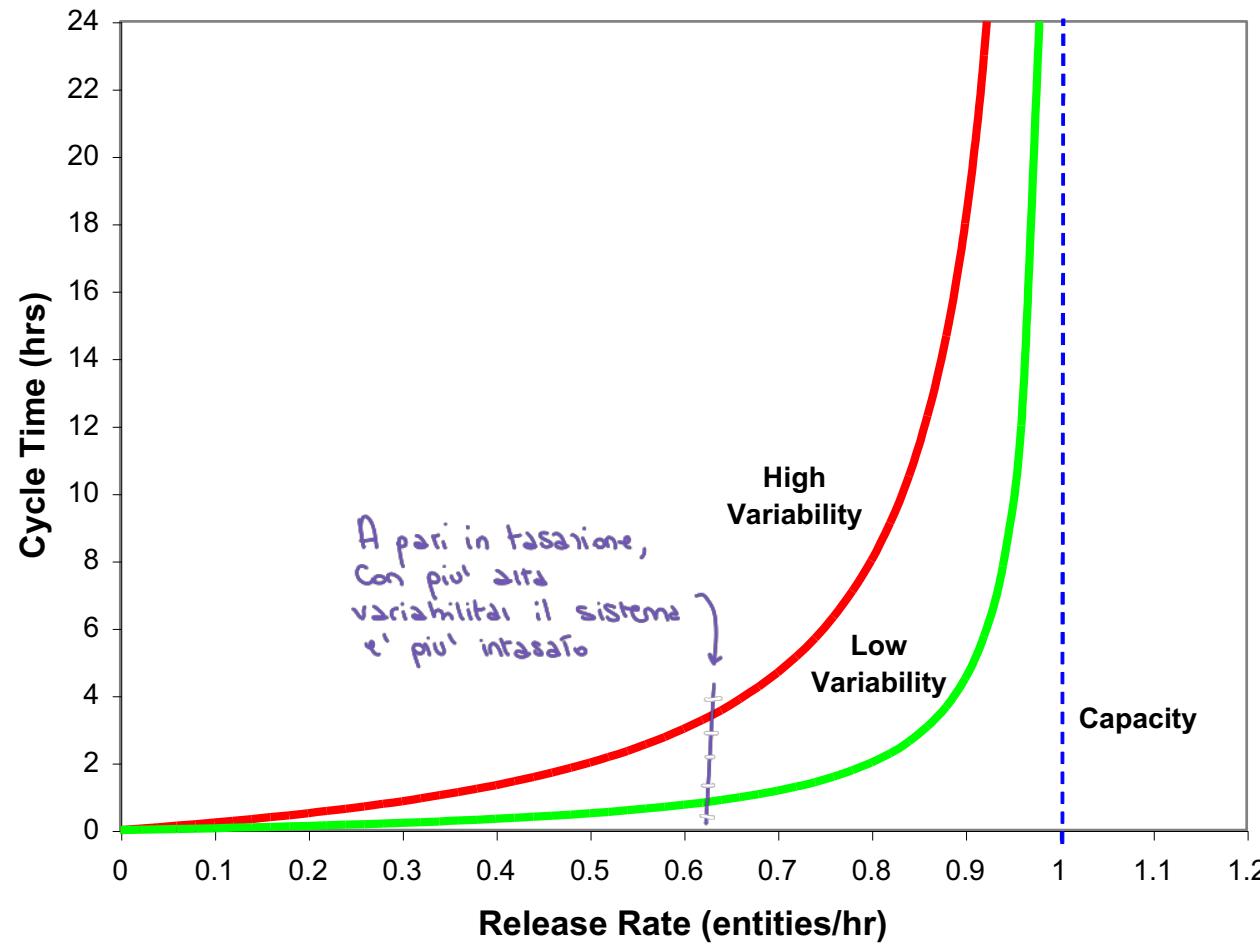
Per M/M/1 (1) è una formula esatta
 $\frac{C_a^2 + C_s^2}{2} = 1$

$$W_q \approx \left(\frac{C_a^2 + C_s^2}{2} \right) \times \left(\frac{u}{1-u} \right) \times t_s \quad \begin{array}{l} \text{Tempo di servizio} \\ \rightarrow \text{E' un'approssimazione e si capisce come ogni fattore impatta} \end{array} \quad (1)$$

dove C_a e C_s sono i coefficienti di variazione dei tempi di interarrivo e servizio, rispettivamente.

Più hai varianza peggio è: non hai mai coda in un sistema deterministico

Il grafico permette di cogliere le implicazioni pratiche del risultato.



Anvicinandoci a 1.2 capacity,
l'attesa in coda va a +
cioe' si crea un arretrato

Buffering Law: Per ovviare alla variabilità occorre introdurre dei buffer come combinazione di:

1. magazzino/WIP → Se ho variabilità della domanda, che sempre bisogno di magazzino
↳ Finiti
↳ Materie prime
Sai colli di bottiglia devi sempre avere wip
2. capacità (in eccesso) → Vincolo di capacità. In generale non ho capacità sovra dimensionata
3. tempo (lead time gonfiato) → Rispetto al tempo medio di servizio i.e. tempi di lavorazione, set-up

Se non si riduce la variabilità, se ne paga il prezzo in termini di WIP alto, capacità sottoutilizzata, riduzione nel livello di servizio al cliente (vendite perse, lead time lunghi, e/o consegne in ritardo).

In effetti, uno dei fondamenti del (vecchio) approccio Toyota è la produzione livellata (production smoothing). → Avere dei flussi il più possibile regolari di produzione

Quali sono le ragioni di variabilità?

La variabilità non è solo legata ad eventi casuali (compresi i guasti alle macchine):

- batching a causa di tempi di setup; → Quando fai il setup devi fare lotti grandi di produzione
- batching nella movimentazione (wait to ^{move} batch); → Wait to batch significa che aspetti fino a quando non hai un batch completo, poi esegui la determinata mansione
- scarso coordinamento nell'assemblaggio (wait to match);
- variabilità della domanda (MPS).

Approccio Just in Time (Toyota)

Idea: ridurre la variabilità mediante la ripetizione di un mix di produzione ripetitivo *mixed-model* (production smoothing). → Versione facile di production smoothing

Facciamo ruotare i prodotti
una volta al mese

codice prodotto	fabbisogni mensili	produzione giornaliera	produzione oraria
A	960	48	6
B	320	16	2
C	1280	64	8
D	480	24	3

Assunzione: un mese di venti giorni lavorativi di otto ore.

↳ Se ripero il mix di produzione ogni ora ho una produzione livellata
= Flusso di produzione regolare

Presupposto: riduzione/annullamento dei tempi di setup. → Per effettivamente avere produzioni livellate

Idea: controllare il WIP mediante un approccio pull (basato sul prelievo fisico di materiali) invece che push (basato su rilascio di ordini pianificati sulla base di previsioni di fabbisogno).

↳ "Spingo" i materiali i.e. anticipo la produzione ideando il fabbisogno
↳ Idea contraria di push

La chiave del controllo pull è il sistema kanban (cartellino; alternativa è il sistema CONWIP).

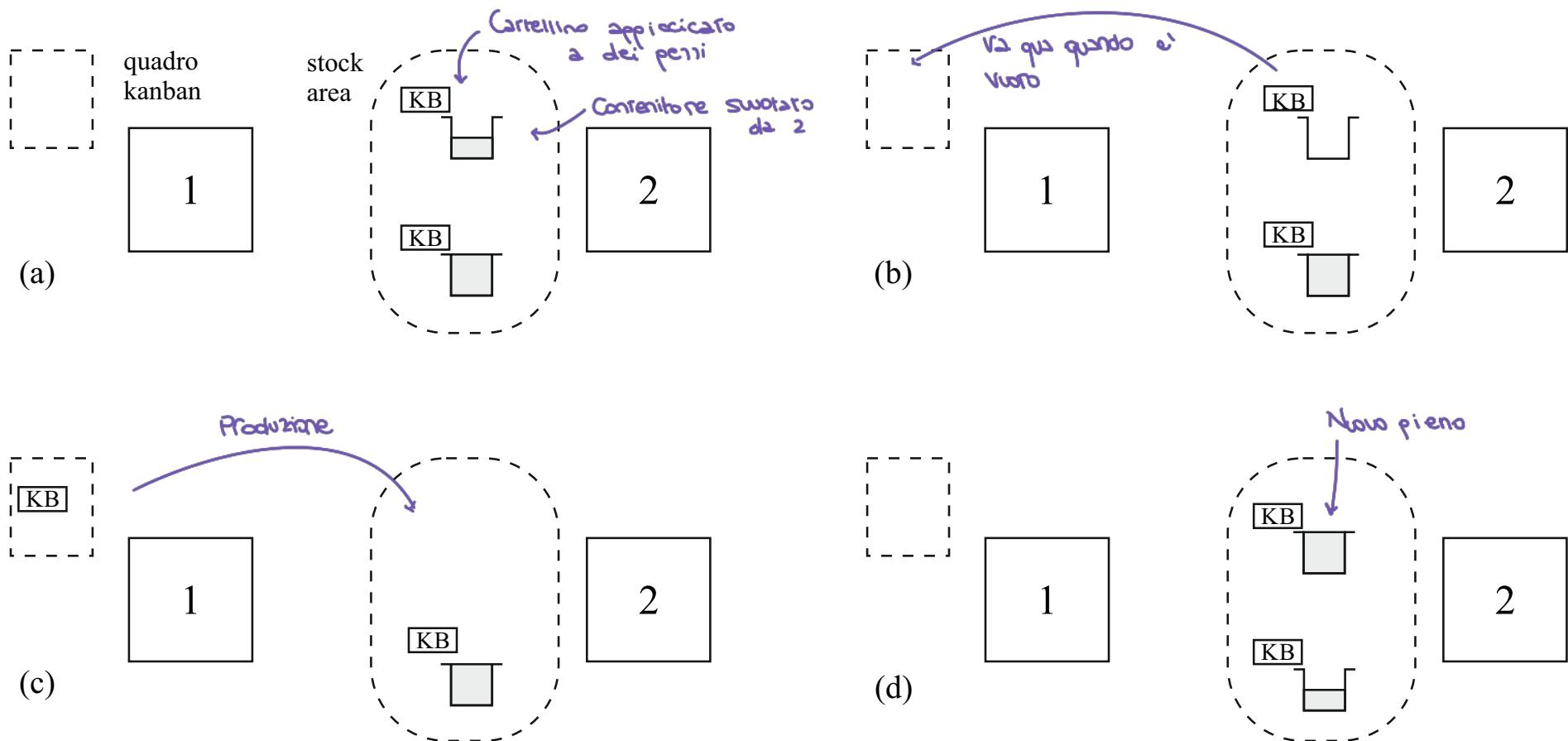
Regola il flusso di clienti, quindi metti dei "certini" così
che il cliente può entrare solo se ha il certino

Ricorda biblioteca a berlino est

Problema: come si ottimizza il numero giusto di Kanban per non avere troppo work in progress, ma neanche stop?

Controllo Pull: Single Kanban

→ 1 Kanban circolano tra 2 stazioni consecutive



Come mettiamo tutto assieme e governiamo con un modello di ottimizzazione?

Toyota Goal Chasing

Supponendo di dover sequenziare le attività di una linea di assemblaggio, abbiamo diverse sequenze che realizzano lo stesso mix.

Possiamo per esempio considerare una sequenza del tipo

AAAAAAABBCCCCCCCCDDDEEEFFF,

→ Assemblare ripetutamente nell'ottica del production smoothing

oppure una del tipo

Vorresti che i componenti specifici avessero un fabbisogno costante nel tempo ACAECFBCACDAEFCBCADCFEAC.

Dato questo, posso fare tutte le permutazioni. Esiste un criterio di scelta per ottenere la permutazione migliore?

Entrambe le sequenze, ripetute ciclicamente, realizzano lo stesso mix. È ragionevole pensare che esista un criterio di scelta in base al quale scegliere una sequenza?

Consideriamo le linee che alimentano di componenti la linea principale di assemblaggio. L'idea è fare in modo da rendere il più possibile costante il fabbisogno di componenti. Vogliamo controllare le linee laterali mediante un sistema pull, che richiede flussi regolari.

formalizzazione del metodo in Toyota che risolvevano (negli anni '70) con esistenze



Formalizzazione:

$\nearrow N$ modelli di automobili diverse

- La linea di assemblaggio realizza N prodotti finiti sulla base di M moduli prodotti su linee laterali.
- La distinta base è piatta: indichiamo con b_{ij} il numero di componenti di tipo j richiesti per assemblare una unità del finito i .
- Il mix ripetitivo prevede l'assemblaggio di Q_i finiti di tipo i . \rightarrow Per ogni modello, quanti ne devi fare

Il fabbisogno per ciclo di unità di tipo j è

$$N_j = \sum_{i=1}^N b_{ij} Q_i$$

*Quanti componenti servono
Finiti di tipo i*

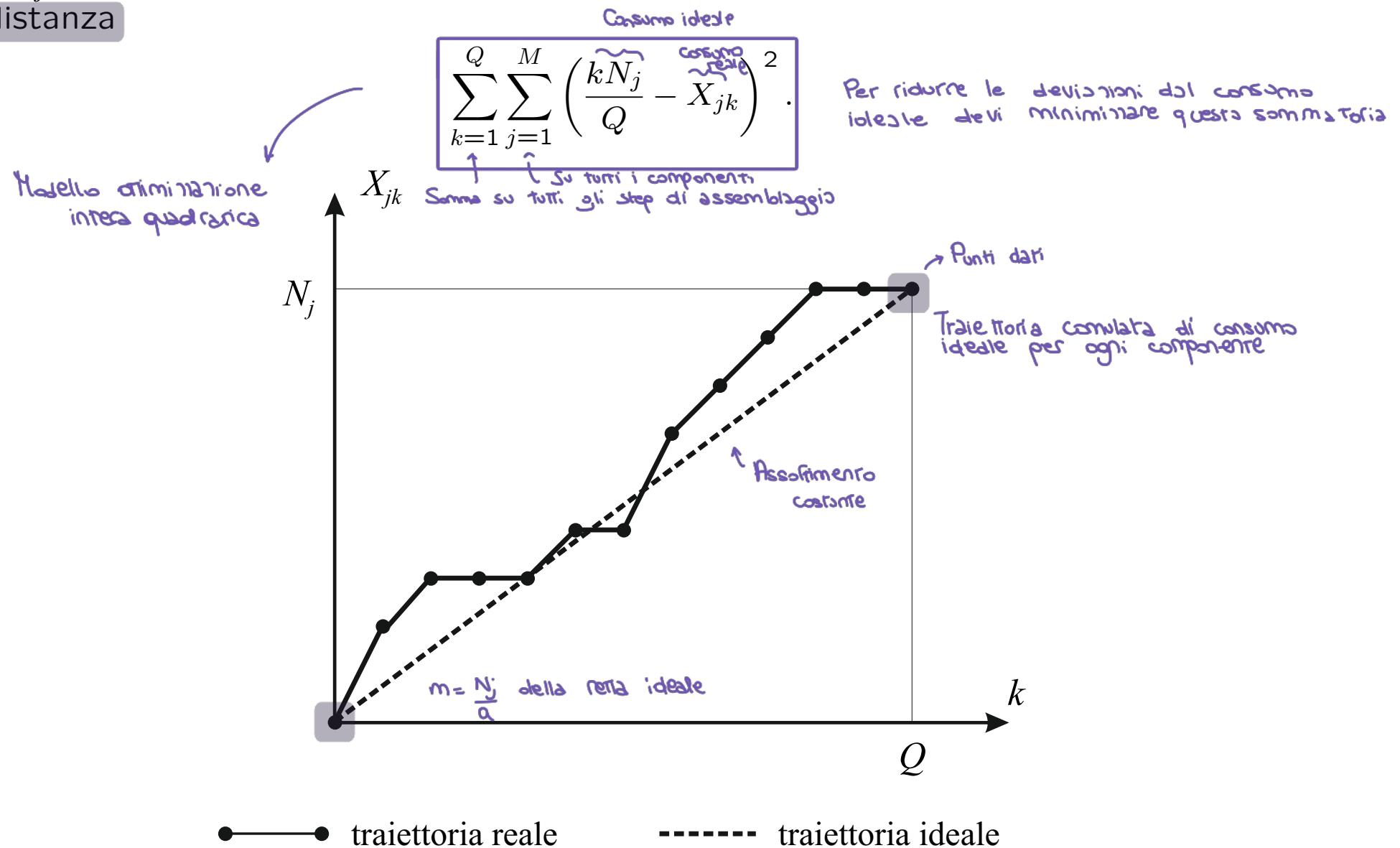
Sia $Q = \sum_{i=1}^N Q_i$ il numero totale di assemblaggi per ciclo. \rightarrow Modelli totali di pezzi finiti i.e. automobili

Il numero di componenti j consumati passa da 0 a N_j dopo Q passi.

Idealmente, per rendere costante l'assorbimento, al passo k il consumo accumulato di componenti j dovrebbe essere

$$\frac{kN_j}{Q}.$$

Sia X_{jk} il consumo cumulato di componenti j al passo k . L'obiettivo è minimizzare la distanza



Tempi di setup e livelli di magazzino

Nel JIT si pone forte enfasi sulla riduzione dei tempi di setup, con buone ragioni. Tuttavia, occorre sempre analizzare l'impatto di tutti i fattori.

Storia di vita vissuta:

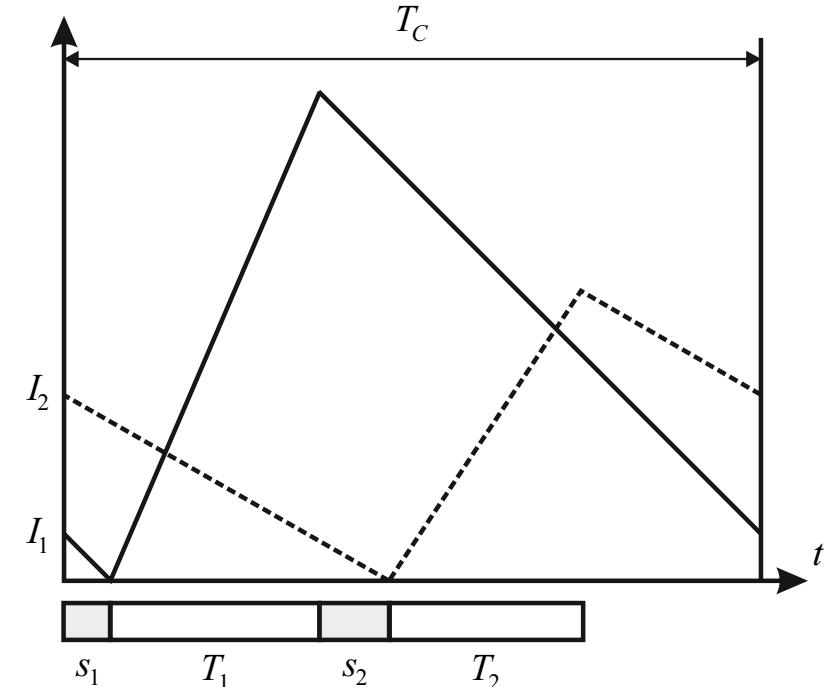
- Un'azienda che produceva componenti in plastica per applicazioni biomedicali (es., kit per la dialisi), aveva un tempo di setup di tre ore sulle presse a iniezione, e aveva problemi di eccesso di capitale immobilizzato a magazzino.
- Un consulente (in giacca e cravatta) rimproverò duramente i dipendenti: “Se ci mettete tre ore, si vede che non sapete fare il vostro lavoro!”.
- Obiezione: un concorrente, che aveva livelli di magazzino più bassi, aveva tempi di setup maggiori.

Dove sta la spiegazione della (apparente) contraddizione?

Consideriamo una linea su cui ruotano ciclicamente N prodotti e, per costruire un modello matematico semplice, introduciamo:

- p_i , il tasso di produzione per il prodotto i (pezzi per unità di tempo);
- $d_i < p_i$ il tasso di domanda (assunto costante);
- s_i il tempo di setup (assunto indipendente dalla sequenza).

Il periodo di rotazione T_c deve essere ridotto, per ridurre il livello medio di giacenza. Qual è il suo limite inferiore?



Indichiamo con T_i la durata del lotto di i per ogni rotazione (escludiamo lotti spezzati). Esso dipende dal tempo complessivo di rotazione, in una situazione di equilibrio:

$$p_i T_i = d_i T_c \quad \Rightarrow \quad T_i = \frac{d_i}{p_i} T_c.$$

Inoltre, deve valere la condizione:

$$T_c \geq \sum_{i=1}^N s_i + \sum_{i=1}^N T_i. \tag{2}$$

L'eventuale slack è utile per assorbire guasti, ritardi, effettuare manutenzioni.

Sostituendo T_i in (2) e manipolando la diseguaglianza, otteniamo il limite inferiore:

$$T_c \geq \frac{\sum_{i=1}^N s_i}{1 - \sum_{i=1}^N \frac{d_i}{p_i}}.$$

Come ci si poteva aspettare, c'è un impatto legato ai tempi di setup, ma c'e' un altro fattore già visto per la coda G/G/1 [vedere Eq. (1)].

Cosa rappresenta? Che impatto ha il tasso di produzione? Che legame vi può essere tra tempo di setup e tasso di produzione?