| | |
|---|---|
| **Titre:** Title: | Contrôle du paramètre Decoupled Lead Time (DLT) et des temps de réponse dans un système piloté en DDMRP |
| **Auteur:** Author: | Guillaume Dessevre |
| **Date:** | 2021 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Dessevre, G. (2021). Contrôle du paramètre Decoupled Lead Time (DLT) et des temps de réponse dans un système piloté en DDMRP [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9929/ |

**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

# Contrôle du paramètre *Decoupled Lead Time* (DLT) et des temps de réponse dans un système piloté en DDMRP

**GUILLAUME DESSEVRE**

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie industriel

Décembre 2021

**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

Cette thèse intitulée:

# Contrôle du paramètre *Decoupled Lead Time* (DLT) et des temps de réponse dans un système piloté en DDMRP

présentée par **Guillaume DESSEVRE**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

**Jean-Marc FRAYRET**, président
**Pierre BAPTISTE**, membre et directeur de recherche
**Jacques LAMOTHE**, membre et codirecteur de recherche
**Robert PELLERIN**, membre et codirecteur de recherche
**Bruno AGARD**, membre interne
**Evren SAHIN**, membre externe

# DÉDICACE

*Ce manuscrit de thèse est à la portée de tous. Bien qu'il soit spécifique à un domaine particulier (la gestion d'un atelier piloté en DDMRP), de nombreux exemples et analogies issus de la vie de tous les jours sont éparpillés entre les réflexions scientifiques, permettant au lecteur de mieux comprendre les enjeux de cette thèse et les évènements qui nous entourent.*

*Je dédicace donc ce manuscrit à tous ceux qui seront assez curieux pour prendre le temps de le lire : famille, ami(e)s, chercheur(e)s, etc.*

# REMERCIEMENTS

Je tiens à remercier en premier mon directeur de recherche, Pierre Baptiste, pour m'avoir fait confiance il y a 4 ans, pour m'avoir aidé tout au long du doctorat lorsque tu avais du temps libre, et pour m'avoir appris le sens du mot « autonomie » pour le reste !

Ensuite, merci à Jacques Lamothe, co-directeur de recherche à l'IMT Mines Albi, pour m'avoir notamment permis de travailler sur un cas industriel pendant ma thèse. Ce fut une excellente expérience, bien qu'elle ne soit pas intégrée dans ce manuscrit. J'en profite donc pour remercier Vincent Pomponne, de Pierre Fabre Dermo-Cosmétique, pour toutes les informations reçues lors de nos nombreuses réunions de travail.

Merci à Robert Pellerin, co-directeur de cette thèse, en particulier pour ta rigueur de travail et pour tous tes commentaires lors de l'écriture des publications.

Pour avoir accepté d'être membres du jury et de devoir lire ce manuscrit en détail, je remercie Jean-Marc Frayret, Bruno Agard et Evren Sahin, respectivement professeurs à Polytechnique Montréal et professeure à CentraleSupélec.

Je tiens également à remercier Maha Ben Ali, professeure à Polytechnique Montréal, pour ton expertise, mais surtout pour m'avoir donné la chance de connaître le milieu de l'enseignement à travers la préparation et l'animation des TD et toutes les heures de correction d'examens !

Un très grand merci à Paul Brunelle, technicien au département de MAGI, pour toutes ces heures « perdues » à installer, désinstaller, réinstaller le logiciel Arena presque tous les mois durant plus d'un an ! J'espère que la bouteille de vin était bonne et saura me faire pardonner.

Merci à Guillaume, Ahlam, et JB, avec qui j'ai travaillé durant mon doctorat, pour avoir installé une bonne ambiance de travail. Et merci à Amandine et Dollon pour l'ambiance hors-travail !

Un grand merci à Emilie, notamment pour le jeu du « Je cherche un mot qui veut dire… » ou encore « T'as pas un synonyme de… » pendant les phases de rédaction !

Enfin, merci à celles et ceux que j'ai oublié sur cette page, mais qui ont leur place ici, ils sauront se reconnaître.

# RÉSUMÉ

La méthode *Demand Driven Material Requirements Planning* (DDMRP) est une méthode récente de gestion de la production, intégrant différents principes du génie industriel tels que le flux poussé, le flux tiré, l'utilisation de stocks tampons, ou encore le mécanisme des méthodes à point de commande. En combinant ces éléments, les auteurs du DDMRP affirment que leur méthode est la solution aux limites des méthodes traditionnelles de gestion de la production.

Un des paramètres les plus importants du DDMRP est le *Decoupled Lead Time* (DLT). Il est défini comme le temps que l'on alloue aux ordres de fabrication pour être complétés. Le temps réellement mis pour fabriquer un ordre est appelé temps de réponse. Ainsi, si le temps de réponse d'un ordre est plus grand que le DLT, l'ordre est alors terminé en retard et l'on risque des ruptures de stocks. De plus, ce paramètre sert à dimensionner les stocks tampons qui génèrent les ordres de fabrication. Par conséquent, le DLT est également relié aux niveaux de stocks de l'entreprise. Il est donc intéressant de réduire au maximum ce paramètre, pour éviter des coûts de stockage inutiles.

C'est ainsi qu'apparaît le compromis autour du dimensionnement du DLT, enjeu principal de cette thèse : d'un côté on cherche à dimensionner le DLT assez grand pour que les temps de réponse y soient le plus souvent possible inférieurs ou égaux, et de l'autre on cherche à diminuer au maximum ce paramètre pour réduire les coûts liés aux stocks. Le DLT est alors relié à deux indicateurs de performances clés en gestion de la production : le taux de service client et le niveau des stocks. Notre objectif général est donc d'améliorer les performances des ateliers de production pilotés par la méthode DDMRP, en proposant des outils de maîtrise des temps de réponse ou d'ajustement du paramètre DLT.

Après avoir fait le tour des publications sur les différents sujets abordés et réalisé une revue de la littérature permettant de définir trois objectifs spécifiques de recherche différents, nous utilisons un protocole expérimental basé sur la simulation à évènements discrets pour atteindre ces objectifs, car elle permet de modéliser des environnements complexes soumis à de fortes variabilités.

Nous proposons dans un premier temps une boucle de régulation permettant d'ajuster dynamiquement le paramètre DLT en fonction des temps de réponse observés dans l'atelier. Nous utilisons par la suite la simulation pour créer des abaques d'aide à la décision. Ces outils permettent de choisir la meilleure option possible de capacité pour une demande prévisionnelle donnée. L'idée étant de maîtriser les temps de réponse en contrôlant le taux de charge des ressources goulots.

Enfin, nous couplons le DDMRP à la méthode *Constant Work-In-Process* (ConWIP) pour créer un modèle de génération des ordres de fabrication plus robuste que la méthode DDMRP classique. Notre modèle prend en compte la variabilité de la demande et celle de la production pour adapter les tailles de lot de fabrication, et ainsi lisser la charge et fluidifier l'atelier.

Les principaux résultats de nos travaux de recherche sont la démonstration que lorsque des ordres de fabrication utilisent la même ressource goulot, leurs temps de réponse sont sensiblement les mêmes. Par conséquent, le paramètre DLT des stocks tampons de produits doit être le même pour tous. De plus, l'ajustement dynamique de ce paramètre permet de réduire les niveaux de stocks, notamment lors d'un mauvais paramétrage initial. Ensuite, il paraît plus judicieux de chercher à maîtriser les temps de réponse plutôt que d'ajuster le DLT. Cette maîtrise passe par le contrôle du taux de charge des ressources goulots. C'est pourquoi nous proposons des outils visuels d'aide à la décision permettant d'identifier les meilleures options d'ajustement de la capacité. Ces outils permettent de corréler le taux de charge des ressources, la distribution des temps de réponse, le paramètre DLT, et les taux de service atelier et client. Enfin, le couplage des méthodes DDMRP et ConWIP permet d'ajuster automatiquement la taille des lots de fabrication en fonction de l'état de l'atelier, fluidifiant ce dernier. Ce couplage rend plus robuste le système et offre de meilleures performances que la méthode DDMRP classique, en termes de volume d'en-cours, de distribution des temps de réponse, et de taux de service.

# ABSTRACT

The Demand Driven Material Requirements Planning (DDMRP) method is a recent production management method that integrates various industrial engineering principles such as push flow, pull flow, the use of inventory buffers, and the mechanism of reorder point methods. By combining these elements, the authors of DDMRP claim that their method is the solution to the limitations of traditional production management methods.

One of the most important parameters of DDMRP is the Decoupled Lead Time (DLT). It is defined as the time that production orders are allowed to be completed. The time actually taken to produce an order is called flow time. Thus, if the flow time of an order is greater than the DLT, the order is completed late and there is a risk of stock-outs. Moreover, this parameter is used to size the stock buffers that generate the production orders. Consequently, the DLT is also linked to the company's inventory levels. It is therefore interesting to reduce this parameter as much as possible, to avoid unnecessary storage costs.

This is how the compromise around the DLT dimensioning appears, which is the main issue of this thesis: on the one hand, we try to dimension the DLT large enough so that the flow times are as often as possible lower than or equal to the DLT, and on the other hand, we try to reduce as much as possible this parameter to reduce the costs related to the stocks. Our objective is therefore to propose methods allowing either to control the flow times so that they are lower than the DLT, or to adjust the DLT accordingly.

After having surveyed the literature on the different topics and performed a critical review of the literature to define three different research objectives, we use an experimental protocol based on discrete event simulation to achieve these objectives, as it allows to model complex environments subject to high variability.

First, we propose a control loop that dynamically adjusts the DLT parameter according to the real times observed in the workshop. We then use simulation to create decision support charts. These tools allow to choose the best possible capacity option for a given forecasted demand. The idea is to control the flow times by controlling the load rate of bottleneck resources. Finally, we couple DDMRP to the Constant Work-In-Process (ConWIP) method to create a more robust model for generating production orders than the classical DDMRP method. Our model takes into account the

variability of demand and production to adapt the manufacturing batch sizes, thus smoothing the load and making the shop floor more fluid.

The main results of our research are the demonstration that when production orders use the same bottleneck resource, their flow times are approximately the same. Therefore, the DLT parameter of the product buffers must be the same for all. Moreover, the dynamic adjustment of this parameter makes it possible to reduce the stocks, in particular at the time of a bad initial parameter setting. Secondly, it seems more sensible to try to control the flow times rather than adjusting the DLT. This control is achieved by controlling the load rate of bottleneck resources. This is why we propose visual decision support tools to identify the best capacity adjustment options. These tools allow to correlate the resource load rate, the flow time distribution, the DLT parameter, and the workshop and customer service rates. Finally, the coupling of DDMRP and ConWIP methods allows for the automatic adjustment of batch sizes according to the state of the shop floor, making the shop floor more fluid. This coupling makes the system more robust and offers better performance than the traditional DDMRP method, in terms of WIP volume, flow time distribution, and service rate.

# TABLE DES MATIÈRES

# LISTE DES TABLEAUX

# LISTE DES FIGURES

# LISTE DES SIGLES ET ABRÉVIATIONS

| | |
|---|---|
| ADU | *Average Daily Usage* (consommation moyenne journalière) |
| CMJ | Consommation Moyenne Journalière |
| ConWIP | *Constant Work-In-Process* |
| DDAE | *Demand Driven Adaptive Enterprise* |
| DDMRP | *Demand Driven* MRP (MRP pilotée par la demande) |
| DDOM | *Demand Driven Operating Model* |
| DES | *Discrete Event Simulation* (simulation à évènements discrets) |
| DLT | *Decoupled Lead Time* |
| FD | Facteur de Délai |
| FP | *Final Product* (produit fini) |
| FT | *Flow Time* (temps de réponse) |
| FV | Facteur de Variabilité |
| LTF | *Lead Time Factor* (facteur de délai) |
| MP | Matière Première |
| MRP | *Material Requirements Planning* |
| OF | Ordre de Fabrication |
| PF | Produit Fini |
| PSF | Produit Semi-Fini |
| RM | *Raw Materials* (matières premières) |
| SFP | *Semi-Finished Product* (produit semi-fini) |
| TdJ | Top du Jaune |
| TdR | Top du Rouge |
| TdV | Top du Vert |

| | |
|---|---|
| ToG | *Top of Green* (Top du vert) |
| ToR | *Top of Red* (Top du rouge) |
| ToY | *Top of Yellow* (Top du jaune) |
| VF | *Variability Factor* (facteur de variabilité) |
| WIP | *Work-In-Process* |

# LISTE DES ANNEXES

# CHAPITRE 1 INTRODUCTION

> Déjà, une histoire, il faut bien la commencer. Aristote dit qu'un tout est ce qui est constitué d'un début, d'un milieu et d'une fin. C'est pour ça qu'il y a l'histoire des trois actes. Donc, il faut avoir un bon début : ce matin en réunion, première erreur, dès le début on ne pigeait rien. (Alexandre Astier, 2006, Kaamelott Livre III, La poétique)

*Vous prévoyez de faire un barbecue entre amis en fin de semaine, et vous planifiez donc les tâches suivantes : jeudi vous achèterez la viande et la sauce (vous avez besoin d'épices, mais c'est bon, il vous en reste d'un barbecue précédent), vendredi vous préparerez les steaks, et samedi vous les ferez griller*. Dans le jargon industriel, on dit que vous raisonnez en flux poussé : on se base sur des prévisions (barbecue samedi), on fait l'inventaire de ce dont on aura besoin (viande, épices, sauce), on retranche ce qu'on a déjà (épices), et on planifie les activités (achats, préparation, cuisson).

Aujourd'hui, les entreprises utilisent majoritairement des méthodes à flux poussé pour planifier et exécuter les ordres de fabrication. La méthode la plus connue et utilisée est le *Material Requirements Planning* (MRP), créé dans les années 1960 et développée par Orlicky (1975), qui reprend les étapes énoncées précédemment. Elle a ensuite été améliorée, devenant le MRP II (Plossl and Orlicky, 1994), pour prendre en compte notamment les ressources nécessaires pour accomplir les différentes tâches, et pour réaliser des calculs de coûts. Le MRP II est très pratique puisqu'il permet de synchroniser la production de systèmes complexes en créant une dépendance entre les éléments d'une nomenclature (la « recette » pour fabriquer votre produit).

*Enfin, pour votre barbecue de samedi, vous commandez à votre ami brasseur une douzaine de bières (des IPA, car vous avez du goût). Dans sa cave, il lui en reste une cinquantaine, mais vu qu'il lui faut un mois pour en faire, il décide de lancer un lot de fabrication pour remonter son stock*. C'est le principe du flux tiré : on ne se base plus sur des prévisions (de ventes), mais sur la consommation d'un stock ou sur une commande réelle qui vient de tomber.

*Et puis si l'on n'a pas de chance, les invités sont en retard et la moitié n'a pas pu venir, il n'y a plus de viande chez le boucher, et le barbecue tombe en panne… de toute façon vous décidez de manger à l'intérieur, car il pleut dehors !*

Ça, c'est la variabilité des systèmes, et les évènements imprévus. Il en existe partout, que l'on fabrique des produits ou que l'on prépare un barbecue. D'après Ptak and Smith (2011), il existe quatre sources de variabilité : deux externes (demande et approvisionnement) et deux internes (production et organisation). Les causes de variabilité, illustrées en Figure 1.1, sont :

1) La demande client. Elle fluctue sans cesse et peut être soumise à des saisonnalités (on achète de la crème solaire surtout en été dans l'hémisphère nord par exemple), à des modes, des règlementations, etc. *Qu'est-ce que vous allez faire de vos grillades en trop, car la moitié des invités n'a pas pu venir ?* ;

2) L'approvisionnement. Les matières premières de votre entreprise sont les produits finis de votre fournisseur : si vous avez des problèmes à fournir vos clients, lui aussi peut en avoir. De plus, le manque d'un simple petit composant peut bloquer toute la fabrication d'un produit fini. *Comme par hasard, aujourd'hui il n'y avait plus de viande chez le boucher…*;

3) La production. On ne met jamais exactement le même temps pour fabriquer deux produits, tout comme on ne met jamais deux fois le même temps pour faire le même trajet en voiture. La production est truffée de variabilité : les temps de fabrication, les changements de série, les pannes des machines, les employés malades, etc. *C'est rare qu'il tombe en panne votre barbecue, mais ce jour-là, c'est arrivé !* ; et

4) L'organisation et les prises de décision. C'est la variabilité « humaine », résultat des décisions prises dans l'entreprise. *Tout le monde à l'intérieur, il pleut dehors ! Bon ça va prendre plus de temps que prévu, car il faut préchauffer le four….*

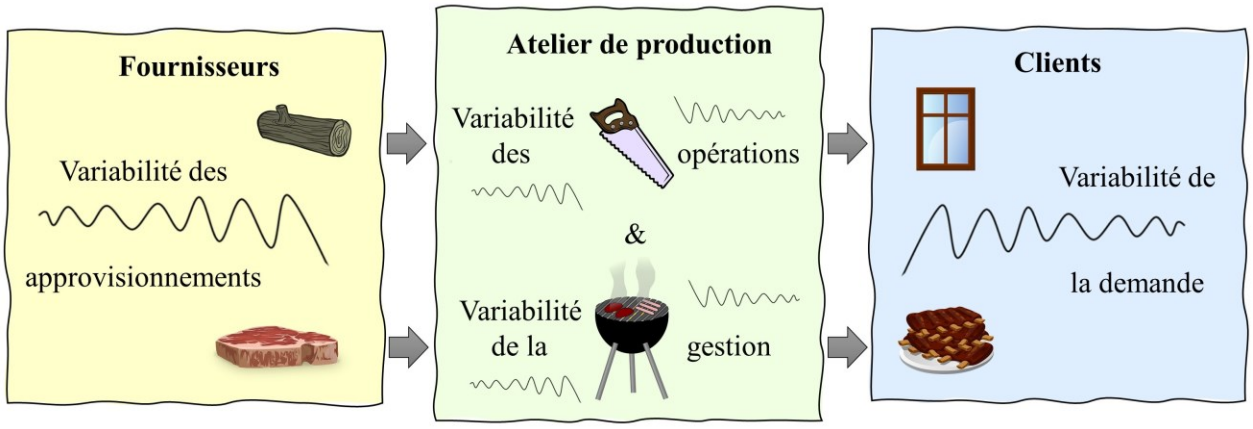


Figure 1.1 : Illustration des quatre formes de variabilité de Ptak and Smith (2011)

Toutes ces causes de variabilité vont changer l'exécution de ce qui a été planifié, et peuvent causer des dégâts : un simple changement dans la demande peut créer des variations intenses dans le stock de matières premières, et vice versa. On parle alors d'effet « coup de fouet » (Lee and Billington, 1992). Imaginez un fouet, ou plutôt une corde posée à terre, qui subit un bref mouvement à une extrémité et qui propage l'onde tout le long, en l'amplifiant jusqu'à l'autre extrémité.

Comment arrêter cette onde ? En posant son pied au milieu de la corde pour empêcher sa propagation : c'est l'idée des points de découplage de la méthode DDMRP. Le but est de disposer des stocks « tampons » sur certains éléments stratégiques de la nomenclature (matières premières, composants ou produits finis) de façon à les protéger des variations et éviter un coup de fouet. Ensuite, après avoir dimensionné ces stocks tampons, on lance des ordres de réapprovisionnement pour les recompléter lorsque ces derniers tombent en-dessous d'un certain seuil (flux tiré). Et pour les éléments non protégés par un stock tampon, on réalise le calcul classique du MRP (flux poussé). *Et peut-être que si le boucher avait mis un stock tampon sur sa viande il en aurait eu ce jeudi là…*

La méthode DDMRP est aujourd'hui d'un intérêt grandissant (Azzamouri et al., 2021), notamment dans le milieu industriel, où de plus en plus d'entreprises se tournent vers le déploiement de cette méthode qui permet, d'après les retours d'expérience, d'améliorer le taux de service client tout en réduisant les niveaux de stocks (Bahu, Bironneau, and Hovelaque, 2019). Dans le milieu de la recherche, le sujet est encore assez controversé, bien que le nombre de publications annuelles sur le sujet n'ait de cesse d'augmenter depuis 2017. Parmi ces publications, de nombreux chercheurs proposent des ouvertures et des pistes d'améliorations possibles sur la méthode (Baptiste, 2018; Dessevre et al., 2020; Orue, Lizarralde, and Kortabarria, 2020).

Parmi les différents paramètres de la méthode DDMRP, nous avons choisi de nous focaliser sur le *Decoupled Lead Time* (DLT). Le DLT est défini par Ptak and Smith (2016) comme le délai le plus long non protégé par un stock tampon dans une nomenclature. Autrement dit, il représente le temps que l'on alloue pour produire un ordre de fabrication : lorsqu'un besoin de recomplètement du stock tampon de produits finis est émis, l'ordre relié à ce besoin doit être complété dans un délai inférieur ou égal au DLT, sous peine de subir des pénuries dégradant le taux de service client (défini comme le nombre de commandes client honorées dans les temps divisé par le nombre total de commandes client). De plus, il sert à dimensionner les stocks tampons DDMRP : plus le DLT est grand, plus les niveaux de stocks le seront également, ce que l'on veut éviter pour des raisons

financières (coût de stockage notamment). Ainsi, le DLT est relié à deux indicateurs de performances importants et souvent utilisés dans l'étude des méthodes de gestion de la production : le taux de service client et le niveau de stock.

On appelle « temps de réponse » le temps que met un ordre de fabrication pour être complété, et on définit le taux de service atelier comme le ratio entre le nombre d'ordres de fabrication dont le temps de réponse est inférieur ou égal au DLT et le nombre total d'ordres de fabrication (cette définition est adaptée de celle présentée par Hopp and Spearman (1996)). Ainsi, plus ce taux est élevé, plus les ordres respectent le temps qu'on leur alloue. On cherche donc un taux de service atelier le plus élevé possible. Par conséquent, il est nécessaire de bien dimensionner le DLT, compromis entre taux de service élevé et faibles niveaux de stocks. Ce compromis, enjeu principal de cette thèse, est représenté en Figure 1.2.

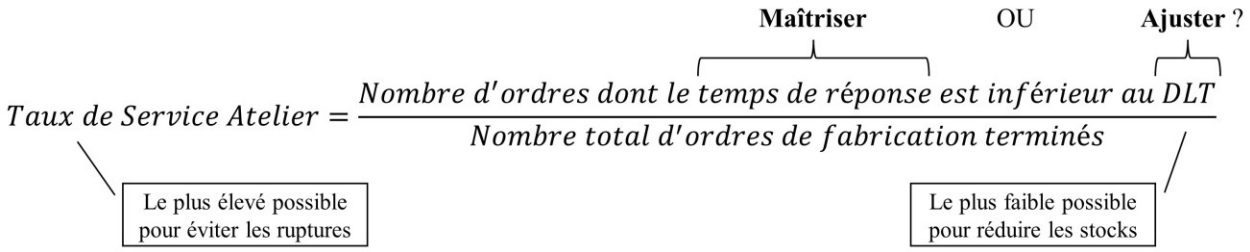


Figure 1.2 : Illustration du compromis sur le dimensionnement du DLT

Les solutions pour obtenir un taux de service atelier élevé sont donc les suivantes : soit on cherche à maîtriser les temps de réponse pour faire en sorte que leur distribution (moyenne et/ou écart-type) soit inférieure ou égale au DLT fixé, soit on ajuste le DLT en conséquence. Nous verrons par ailleurs que la capacité des ressources (machines ou opérateurs par exemple) joue un rôle crucial dans la détermination des temps de réponse. Car si la capacité ne permet pas d'absorber la charge, les ordres de fabrication s'empilent devant les machines, rallongeant les temps de réponse. C'est d'ailleurs en partie ainsi que se créent les files d'attente (devant un guichet ou une caisse, ou sur l'autoroute). *Pensez à votre barbecue : si vous ne prévoyez pas le nombre d'invités à l'avance, et que la grille ne peut contenir que 4 steaks de viande en même temps alors que vous êtes 8, qui va devoir attendre pour manger ? Pourquoi ne pas avoir prévu un deuxième barbecue ?*

Nous allons voir dans le chapitre suivant que capacité, taux de charge, temps de réponse et taux de service atelier sont liés entre eux : pour avoir un taux de service atelier correct, il faut contrôler ses temps de réponse en maîtrisant le ratio charge/capacité. Dans cette thèse, on propose alors de

répondre à la problématique suivante : **Faut-il maîtriser les temps de réponses ou ajuster le DLT lors d'une variation anticipée ou observée des temps de réponse dans un atelier piloté en DDMRP ?**

Pour répondre à cette problématique et aux objectifs spécifiques de recherche présentés dans le chapitre 3, nous avons décidé d'utiliser la simulation à évènements discrets. La modélisation et la simulation des systèmes de production sont de plus en plus utilisées aujourd'hui dans la recherche, car elles permettent de modéliser et d'analyser des environnements complexes ayant plusieurs sources de variabilité (Mourtzis, 2020). L'idée est de reproduire des processus automatiques (fabriquer un produit, transporter des marchandises) et/ou des comportements humains (vérifier un état de stock, passer une commande) pour avoir un modèle proche de la réalité. On peut y introduire de la variabilité, en la mesurant sur des données ou des machines par exemple, ou en utilisant directement des données réelles dans le modèle. Une fois validé, on utilise ce modèle pour tester différents scénarios et ainsi valider, ou non, nos hypothèses de recherche.

Ce travail de recherche a produit trois articles de journaux, dont deux sont inclus dans ce manuscrit[1], précédés d'un article de conférence. Les trois articles présentés dans ce document permettent chacun de répondre à un objectif en lien avec notre problématique de recherche. Quatre autres articles de conférence ont également été publiés: un article sur l'impact du paramétrage du DLT, en collaboration avec Guillaume Martin, présenté à CIGI QUALITA 2019 à Montréal (Dessevre, Martin, Baptiste, Lamothe, and Lauras, 2019); un article sur les problématiques industrielles soulevées par un déploiement de la méthode DDMRP dans une industrie dermo-cosmétique à ILS 2020, initialement prévue à Austin, Texas (Dessevre et al., 2020); et deux articles présentés en ligne à MOSIM 2020 à Agadir, un premier sur des abaques capacitaires, prélude du premier article de ce manuscrit (Dessevre, Baptiste, and Lamothe, 2020), et un deuxième sur un module d'ajustement de la capacité, co-écrit avec Maha Ben Ali (Dessevre and Ben Ali, 2020).

Ce manuscrit est organisé comme suit. Le chapitre 2 présente une revue de la littérature sur tous les sujets abordés dans cette thèse. Dans le chapitre 3, on définit les éléments principaux de la

[1] Le troisième article de journal produit, non présent ici, porte sur l'évaluation de politiques de pilotage d'un processus divergent dans un contexte DDMRP, avec une application sur un cas industriel.

méthodologie employée pour répondre scientifiquement à la problématique. On y retrouve également l'organisation générale des articles et la cohérence de ces derniers par rapport aux objectifs de recherche. Cette thèse étant une thèse par articles incorporés dans le corps du travail, les chapitres 4 à 6 présentent les trois articles de conférence et de journaux. Le chapitre suivant est une discussion générale en regard des aspects méthodologiques et des résultats en lien avec la revue de la littérature. Enfin, le dernier chapitre présente les conclusions et les recommandations de cette thèse.

Bonne lecture !

# CHAPITRE 2 REVUE DE LA LITTÉRATURE

> - Bon, alors maintenant il s'agit de pas faire n'importe quoi.
> - De l'OR-GA-NI-SA-TION.
> - Déjà, on devrait commencer par trier les fioles.
> - Selon quoi… les couleurs ?
> - Ah non, non, non, non, non, non. Les couleurs ça peut être trompeur ! Regardez les haricots : les rouges, ils sont plus jolis que les blancs, mais ils bousillent les boyaux. (Alexandre Astier, 2005, Kaamelott Livre II, Les alchimistes)

La revue de la littérature est un état de l'art des publications qui portent sur tous les sujets liés à la compréhension de la problématique. Elle a pour but de présenter ce qui a été produit par les chercheurs, de définir les limites de ces études, et enfin de présenter la problématique de recherche de cette thèse. Pour cela, on explore les différents sujets de recherche, les problématiques et les solutions trouvées, en s'appuyant sur les publications scientifiques. Si d'autres chercheurs ont déjà publié des solutions, on se base sur leurs travaux pour répondre à nos questions ou pour aller plus loin, sinon on constate une lacune dans la littérature, qui doit être comblée.

Notre principale partie de la revue porte sur le fonctionnement et les limites de la méthode DDMRP. Nous verrons que la génération des ordres d'approvisionnement est similaire aux méthodes dites « à point de commande ». Il sera donc intéressant de comprendre ce qu'est un point de commande pour de futures comparaisons. De plus, un des paramètres importants de la méthode DDMRP est le DLT, délai que l'on se donne pour compléter les ordres de fabrication. Pour maîtriser les temps de réponse, temps réellement observés, il existe une méthode appelée ConWIP. En réalité, cette méthode permet de contrôler les temps de cycle, et c'est pourquoi il est important de faire un point sur les différentes définitions des temps dans la littérature : il s'agit donc de la première partie de la revue. Enfin, la théorie des files d'attente nous permet d'expliquer mathématiquement d'où viennent les temps d'attente, qui représentent en général 80% des temps de réponse (Hopp and Spearman, 2011).

Ainsi, nos cinq parties de la revue de la littérature ont pour objectif de :

1. Définir ce qu'est un temps de réponse, un délai de production, un temps de cycle, et un DLT ;
2. Présenter le fonctionnement de la méthode DDMRP ainsi que ses limites ;
3. Identifier les caractéristiques d'une méthode à point de commande ;

4. Décrire le fonctionnement de la méthode ConWIP ; et
5. Assimiler les principes de la théorie des files d'attente.

La Figure 2.1 reprend le chemin de pensée précédent en illustrant les liens entre ces différents sujets traités dans cette revue, et *a fortiori* dans le manuscrit : les définitions des temps (chapitre 2.1), le DDMRP (2.2), les méthodes à point de commande (2.3), le ConWIP (2.4), et enfin la théorie des files d'attente (2.5).

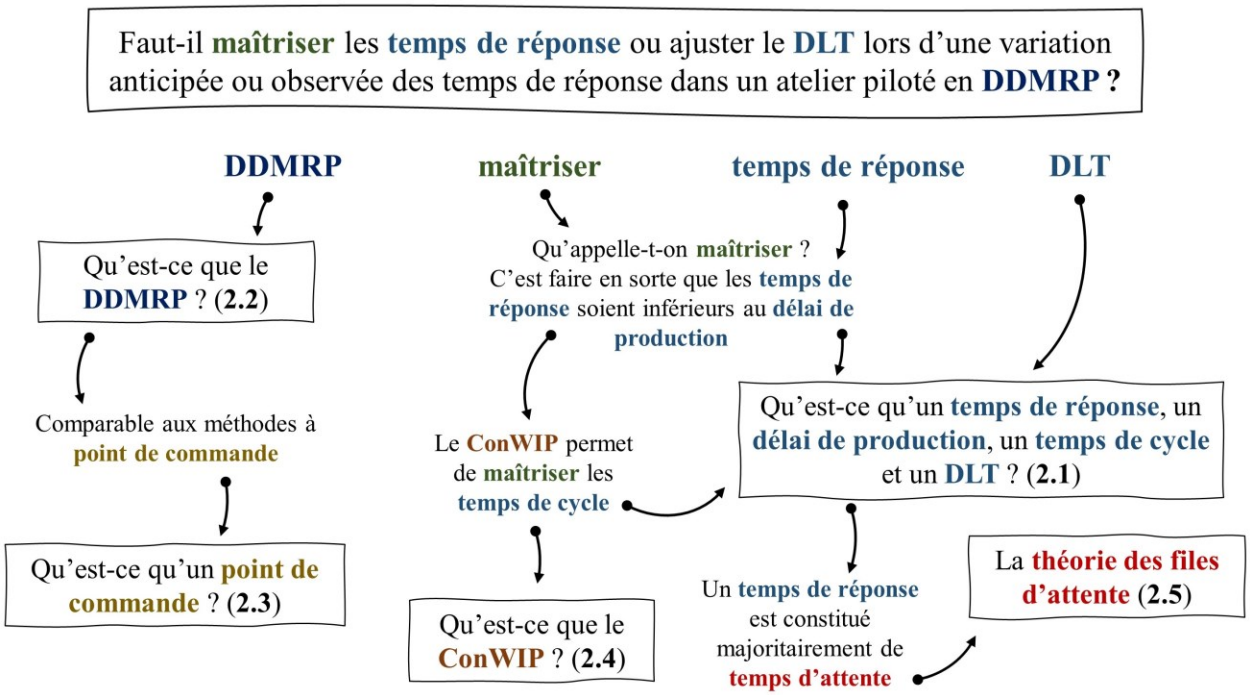


Figure 2.1 : Feuille de route de la revue de littérature

## 2.1 Définitions du temps de réponse, délai de production, temps de cycle et DLT

Dans la littérature anglophone, les termes *lead time* et *flow time* sont les expressions les plus utilisées pour définir le « temps de parcours » d'un ordre de fabrication. Ces deux termes sont différents, car ils ne définissent pas la même chose, mais sont pourtant souvent confondus et il devient difficile de ne pas les mélanger et de ne pas comprendre de quoi il est question dans un article. Dans le dictionnaire APICS par exemple, Blackstone (2013) donne deux définitions différentes du *lead time*, qui se rapprochent plus de ce qu'on peut appeler un *flow time*. Dans le pire des cas, le terme *lead time* est utilisé dans les articles de recherche, mais non défini, et on ne

sait pas si l'article parle de temps observé ou de temps alloué. De plus, ils sont généralement traduits par l'expression « temps de défilement » en français, ce qui rend la distinction encore plus compliquée.

Dans le livre *Factory Physics*, Hopp and Spearman (2011) distinguent ces termes et définissent le *lead time* ainsi : « *The lead time of a given routing or line is the time allotted for production of a part on that routing or line. As such, it is a management constant.* ». Le *lead time* est donc un paramètre de gestion, c'est un choix, c'est le temps alloué pour compléter un ordre de fabrication. À l'inverse, on va définir le *flow time* comme une variable observée, ce n'est pas un choix, c'est le temps exact que met un ordre de fabrication spécifique pour être complété. C'est pourquoi j'ai décidé de traduire ces termes en français par deux termes différents, en expliquant leur définition ci-après. En conséquence, dans ce manuscrit nous utilisons les termes et définitions suivants :

- **Délai de production** : (proche du *lead time*) temps alloué pour la fabrication d'un ordre, depuis l'identification de son besoin jusqu'à la mise en stock des produits attachés à cet ordre. Le délai de production est un paramètre de gestion; et
- **Temps de réponse** : (proche du *flow time*) temps s'écoulant entre l'identification d'un besoin de fabrication d'un produit et la mise en stock des produits répondants à ce besoin. Il s'agit donc du temps que met le système à répondre à un besoin. Les temps de réponse sont variables d'un ordre de fabrication à un autre. Dans un contexte DDMRP, le temps de réponse d'un ordre commence lorsque la position de stock atteint le Top du Jaune (concept expliqué plus tard, partie 2.2).

Un exemple est illustré en Figure 2.2, où l'on mesure les temps de réponse de trois Ordres de Fabrication (OF) différents. Le temps de réponse commence dès qu'un besoin est émis (à $t = 0$ pour les trois ordres ici), et se termine lors de la mise en stock des produits. Le délai de production lui, est fixe et vaut 8 jours.

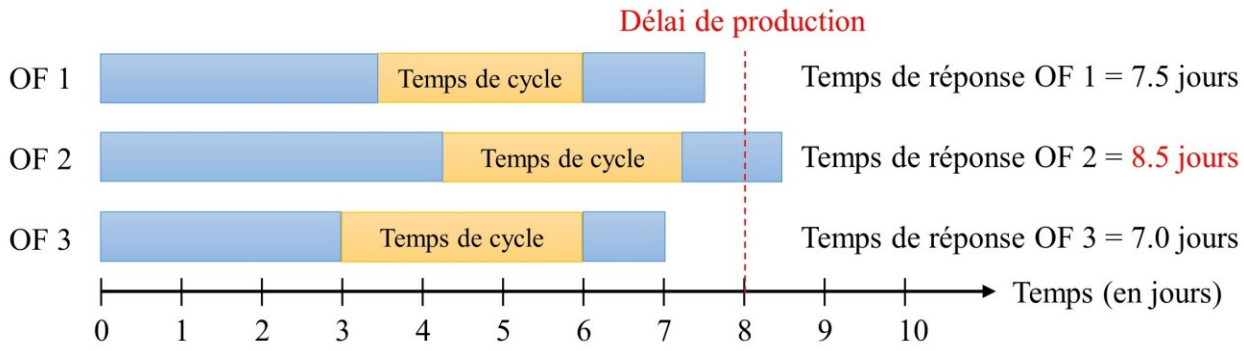


Figure 2.2 : Illustration de trois Ordres de Fabrication (OF) ayant des temps de réponse différents pour un même délai de production

En s'inspirant de Hopp and Spearman (2011), on peut ainsi définir le taux de service d'un atelier comme le ratio entre le nombre d'ordres de fabrication dont le temps de réponse est inférieur ou égal au délai de production et le nombre total d'ordres terminés (Équation 1) :

$$Taux\ de\ Service\ Atelier = \frac{Nombre\ d'ordres\ dont\ le\ temps\ de\ réponse\ est\ inférieur\ au\ délai}{Nombre\ total\ d'ordres\ de\ fabrication\ terminés} \quad (1)$$

En d'autres termes, ce taux de service indique le nombre d'ordres dont le temps de réponse respecte le temps alloué sur le nombre d'ordres total. Dans l'exemple de la Figure 2.2, deux des trois ordres respectent le temps alloué, le taux de service atelier est donc de 66%. On remarque que pour améliorer ce taux de service, il faut soit maîtriser les temps de réponse pour faire en sorte qu'ils finissent dans les temps, soit adapter le délai de production en conséquence.

*Prenons un exemple commun pour mieux comprendre : vous comptez faire le trajet « Montréal – Saint-Profond-du-Nowhere » cette fin de semaine, car vous êtes invités chez des amis pour déjeuner à midi (et vous comptez être à l'heure évidemment). À quelle heure partez-vous ?*

*Les dernières fois où vous y êtes allés, vous avez mis 3h, 2h30 et 4h15. Ce sont vos temps de réponse. Ils sont variables, ils dépendent de plusieurs paramètres (internes, comme votre limite personnelle de vitesse, et externes, comme l'état de la route ou la météo) et sont tous différents. En moyenne vous mettez donc 3h15, mais si vous voulez être sûr d'arriver pour l'apéritif, vous prévoyez 4h. Votre délai (de production) est donc de 4h : ce n'est pas un temps observé, mais le temps que l'on se donne pour finir une tâche (ici, un trajet en voiture). Ce jour-là, vous avez mis 3h50…ouf !*

Sur la Figure 2.2, on a également représenté le temps de cycle. Il est important de le différencier du temps de réponse. Dans ce manuscrit, on définit le temps de cycle comme le temps de passage d'un ordre dans une boucle de production (plus d'explications dans la partie 2.4) :

- **Temps de cycle** : temps s'écoulant entre l'entrée d'un ordre de fabrication dans une boucle de production et sa sortie de la boucle. Autrement dit, temps s'écoulant entre la prise d'un ticket ConWIP et sa libération lorsque la boucle est contrôlée par un ConWIP.

Le dernier terme à définir est le *Decoupled Lead Time* (DLT), terme venant du DDMRP et servant à dimensionner les stocks tampons. Le DLT est défini comme le délai le plus long non protégé par un stock tampon dans une nomenclature (Ptak and Smith, 2016). Ce délai représente un chemin critique. Pour déterminer le DLT, on somme les délais de production des différents composants du chemin critique.

La Figure 2.3 représente la nomenclature d'un produit fini. Le produit fini et ses composants ont chacun un délai de production ou d'approvisionnement représenté dans les cercles. En pointillés se trouve le chemin critique : à gauche, où la production serait pilotée par un MRP, le délai de production cumulé du produit se trouve sur le chemin critique et vaut 2 + 5 + 8 + 30 = 45 jours; à droite, la production est pilotée par un DDMRP avec des stocks tampons placés (stratégiquement) à certains endroits de la nomenclature, et le DLT du produit fini est de 2 + 5 + 8 = 15 jours (on ne compte pas le délai d'approvisionnement du composant 32 puisque celui-ci est protégé par un stock tampon).

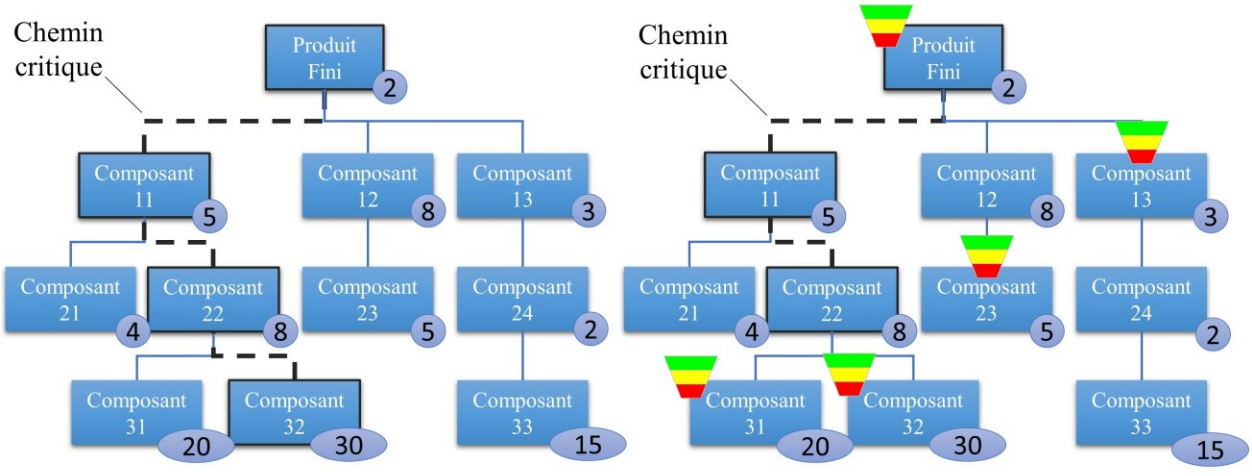


Figure 2.3: Exemple d'une nomenclature et des chemins critiques

*Reprenons l'exemple du trajet en voiture, où Montréal est votre stock tampon initial, Saint-Profond-du-Nowhere est une étape (un composant non protégé par un stock tampon) et votre destination finale (stock tampon de fin) est maintenant la ville de Québec. Vous savez qu'en moyenne vous mettez 2h pour faire « Saint-Profond-du-Nowhere – Québec » (c'est la moyenne de vos temps de réponse), vous décidez donc de vous donner 2h30 pour le faire sans être en retard (c'est votre délai de production). Votre DLT, somme des délais entre deux stocks tampons, est donc de 4h + 2h30 = 6h30 : pour faire « Montréal – Québec » en passant par Saint-Profond-du-Nowhere, vous planifiez un DLT de 6h30.*

Le DLT est donc un paramètre important, car il permet à la fois de définir la taille des zones de chaque stock tampon, mais également de servir de paramètre de contrôle des temps de réponse. Un exemple est illustré en Figure 2.4 où le DLT vaut quatre jours. L'en-cours est représenté en bleu clair, la position de stock est en pointillée, et le stock de produits finis en bleu foncé. Le troisième jour, un ordre de fabrication est lancé et il se termine deux jours plus tard : il respecte le temps alloué et le stock ne descend donc pas dans la zone rouge (qui correspond au stock de sécurité, les différentes zones – verte, jaune et rouge – sont expliquées dans la partie suivante 2.2). Le septième jour, un autre ordre est lancé, mais il met cinq jours pour être complété, au lieu de quatre maximum : le stock (trait plein noir) atteint donc la zone rouge, indiquant que l'on consomme le stock de sécurité. Dans notre exemple, l'ordre se termine juste à temps pour éviter une pénurie qui ferait chuter le taux de service client (le taux de service client lui est dégradé). Un troisième ordre (trait bleu clair en tirets) est lancé au jours 10, augmentant l'en-cours, et se termine au jour 14.

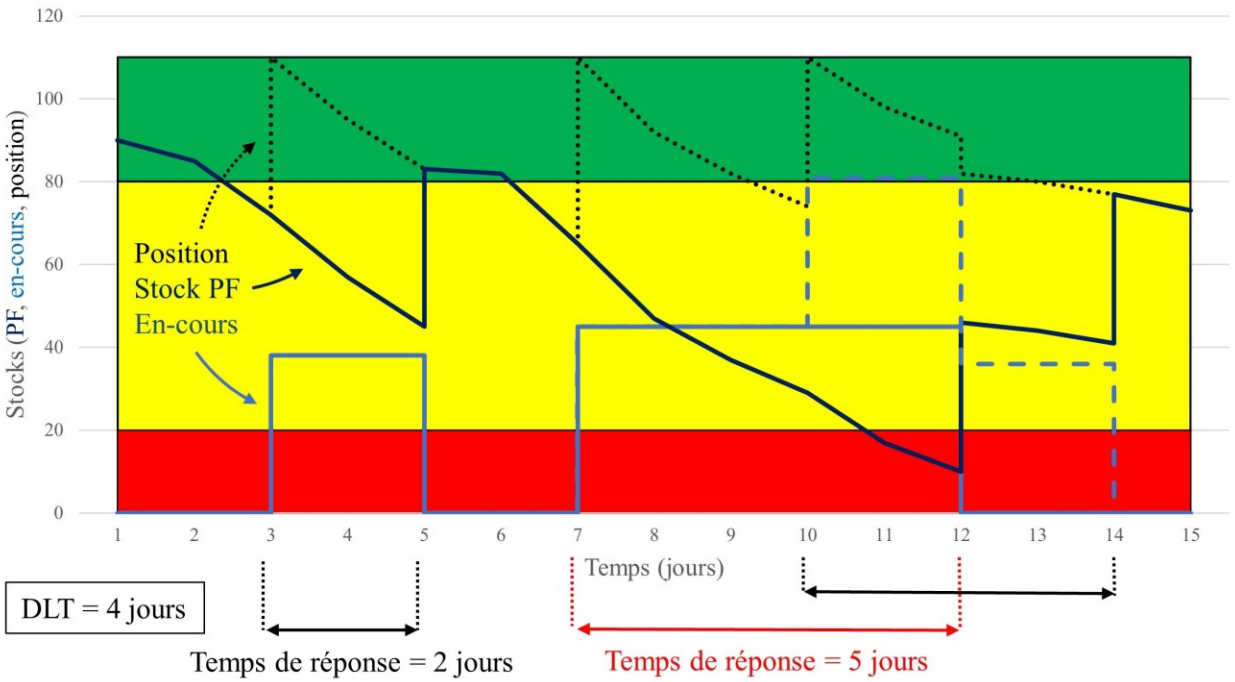


Figure 2.4 : Conséquence des temps de réponse sur l'évolution des stocks dans le temps

Un ordre de fabrication qui ne respecte pas le temps alloué, *c.-à-d.* le DLT ici, risque de détériorer le taux de service client, ou bien de mettre en péril toute la chaîne logistique s'il s'agit d'un composant par exemple. On pourrait donc se dire qu'il faut augmenter le DLT pour éviter les pénuries, mais cela augmenterait les zones des stock tampons, et donc l'en-cours et le stock de produits finis. Ainsi, bien qu'un DLT suffisamment grand soit une solution pour éviter les ruptures, la hausse des stocks et les coûts associés doivent également être pris en compte. Le choix d'un bon DLT est donc primordial lors du paramétrage de la méthode DDMRP.

Enfin, on remarque que si les temps de réponse sont toujours égaux au DLT, ce qui n'arrive jamais en réalité, l'en-cours de produits est égal à la zone jaune, et le stock physique à la zone rouge plus la moitié de la zone verte. C'est le cas dans la Figure 2.5, après la mise en régime permanent à partir du cinquième jour. Ce sont les formules données par Ptak and Smith (2016) pour calculer en-cours et stocks physiques moyens, mais comme les temps de réponse ne sont jamais égaux, surtout au DLT, ces formules sont toujours fausses.

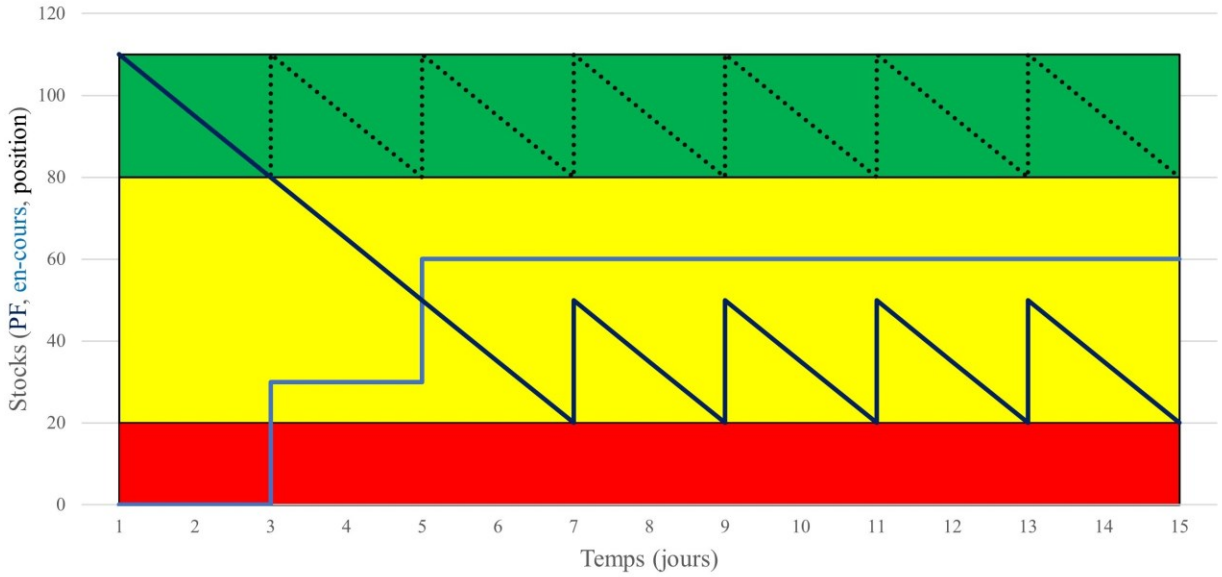


Figure 2.5 : Cas utopique sans variabilité où temps de réponse = délai de production

En conclusion, les définitions des différents temps sont assez vagues dans la littérature. Nous avons donc défini nos propres termes ici, où nous faisons la différence entre temps de réponse, délai de production, et temps de cycle. De plus, le DLT, introduit par la méthode DDMRP est l'un des paramètres les plus importants pour le dimensionnement des stocks tampons, et maîtriser les temps de réponse c'est faire en sorte qu'ils soient inférieurs ou égaux au DLT.

## 2.2 Fonctionnement et publications sur la méthode DDMRP

La méthode DDMRP est une méthode de planification et d'exécution de la production, créée par Carol Ptak et Chad Smith. Publiée dans la troisième édition du *Orlicky's MRP* (Ptak and Smith, 2011), la méthode a ensuite été détaillée dans le livre éponyme (Ptak and Smith, 2016). Elle continue aujourd'hui d'être améliorée et documentée par les auteurs, testée et optimisée par les chercheurs, et déployée dans l'industrie par les consultants (Azzamouri et al., 2021; Bahu, Bironneau, and Hovelaque, 2019). C'est une méthode hybride entre flux poussé et flux tiré, de plus en plus déployée dans le milieu industriel et étudiée dans le monde académique. Pour mieux comprendre les problématiques soulevées par la méthode DDMRP, il faut dans un premier temps comprendre d'où elle vient et comment elle fonctionne.

### 2.2.1 Fonctionnement de la méthode DDMRP

Comme dit précédemment, la méthode DDMRP est une méthode hybride entre flux poussé et flux tiré. Mais de quoi parle-t-on au juste ?

On parle de flux poussé lorsque la gestion de la production est basée sur des prévisions de ventes. Ces prévisions servent dans un premier temps à définir combien de produits finis il faut fabriquer, puis en remontant la nomenclature on va déterminer combien de composants il faut fabriquer et pour quand, et enfin combien et pour quand on a besoin de matières premières. On utilise pour cela le Calcul des Besoins Nets, appelé *Materials Requirements Planning* (d'où vient le MRP de DDMRP) de Orlicky (1975). Il s'agit de comparer ce que l'on compte produire avec ce que l'on a déjà en stock, et ce pour chaque élément de la nomenclature. Il en sort alors une planification de la production. Le MRP, apparu dans les années 1960, est par la suite devenu le MPR II, qui prend en compte notamment les ressources nécessaires et les coûts engendrés (Plossl and Orlicky, 1994). Le MRP II est actuellement la méthode de planification la plus utilisée dans le monde. Ayant plusieurs avantages, notamment la facilité de planification et de synchronisation depuis l'arrivée de l'informatique et des progiciels de gestion intégré, elle possède deux inconvénients majeurs soulevés notamment par Ptak and Smith (2011) :

1) Elle se base sur des prévisions de vente, et les prévisions ne sont jamais fiables. On risque donc de ne pas produire ce qu'il faut, et de produire ce qu'il ne faut pas; et
2) Elle synchronise les éléments d'une nomenclature (matières premières, composants, et produits finis) lors du calcul des besoins, les rendant dépendants. La fabrication d'un produit fini va engendrer l'approvisionnement de matières premières 3 mois à l'avance par exemple, alors qu'on n'est même pas sûr de vendre ce produit (car prévisions peu fiables).

C'est pourquoi sont apparues les méthodes dites à flux tiré. Flux tiré par quoi ? Par la demande, que l'on traduit en anglais par *Demand Driven*, d'où vient le DD de DDMRP. Dans les méthodes à flux tiré, la production est entraînée par une commande client, et cette information va remonter la chaîne logistique pour fabriquer les produits finis, les composants et acheter les matières premières en fonction du besoin réel. Ces méthodes permettent de limiter les stocks, puisque l'on fabrique pour honorer des commandes précises. Les produits finis peuvent être déjà fabriqués lorsqu'on travaille en *Make-To-Stock* (MTS), où l'on va recompléter les stocks, ou bien fabriqués sur demande en *Make-To-Order* (MTO).

Parmi les méthodes à flux tiré, la plus connue est le système Kanban, développé par Toyota (Sugimori et al., 1977). Pour gérer la production dans un système Kanban, on utilise des étiquettes attachées à des conteneurs de pièces et l'on met en place une « boucle » de production : dans une boucle Kanban, le nombre d'étiquettes est fixe, ce qui permet de limiter l'en-cours (*c.-à-d.* le nombre de pièces en cours de fabrication). Lorsqu'un conteneur de pièces finies (des composants par exemple) est utilisé pour la prochaine étape de fabrication (les produits finis), l'étiquette attachée à ce conteneur revient à son point de départ : un tableau Kanban. Sur ce tableau, l'opérateur en charge de la production sur sa machine voit combien d'étiquettes sont présentes pour chaque produit : plus il y en a sur le tableau, moins il y a de composants fabriqués, et donc plus il faut en lancer en production. Une boucle Kanban permet donc de gérer la production en fonction de ce qui a été consommé dans le stock de pièces finies tout en limitant l'en-cours dans la boucle. Il existe aujourd'hui de nombreuses méthodes à flux tiré, notamment celles basées sur des jetons similaires aux étiquettes Kanban (González-r, Framinan, and Pierreval, 2012), comme le Kanban généralisé (Frein, Di Mascolo, and Dallery, 1995), le Kanban étendu (Dallery and Liberopoulos, 2000), le ConWIP (Spearman, Woodruff, and Hopp, 1990) et tous ses dérivés (Prakash and Chin, 2015), le Base-Stock (Duenyas and Patana-Anake, 1997), etc.

Ces deux grandes familles de gestion de la production sont donc diamétralement opposées : le flux poussé planifie les ordres de production en fonction des informations externes et prône la synchronisation de tous les éléments d'une nomenclature, rendant la fabrication d'un produit dépendant de ses composants; tandis que le flux tiré autorise les ordres en se basant sur des informations internes et crée une indépendance entre les éléments de la nomenclature, rendant la synchronisation plus délicate lors de fluctuation de la demande par exemple.

C'est ainsi qu'est apparu le DDMRP, qui mélange flux poussé et flux tiré, dans le but de tirer profit des avantages de ces deux philosophies de gestion. Le principe est simple : créer des points de découplage, appelés stocks tampons, qui vont servir à absorber les différentes variations dans le système. Chaque produit ou composant protégé par un stock tampon est piloté en flux tiré, et chaque produit ou composant non protégé par un stock tampon est piloté en flux poussé.

Il y a 5 étapes principales pour implémenter la méthode DDMRP (Ptak and Smith, 2016) :

1) Le positionnement stratégique des stock tampons : il s'agit de placer des stocks tampons à des endroits stratégiques dans la nomenclature des produits. Ces stocks tampons vont servir

de point de découplage, pour notamment absorber de la variabilité et éviter ce qu'on appelle l'effet « coup de fouet » qui propage cette variabilité tout le long de la chaîne logistique (Lee and Billington, 1992).

2) Le dimensionnement des stocks tampons : une fois positionnés, il faut dimensionner les stocks tampons. Pour cela, le DDMRP utilise quatre paramètres (Ptak and Smith, 2016) :

    a. La Consommation Moyenne Journalière (CMJ) : on calcule cette consommation en se basant sur un historique de ventes, des prévisions, ou un mélange des deux;

    b. Le *Decoupled Lead Time* (DLT) : défini comme le délai le plus long non protégé par un stock tampon, il s'agit du temps que l'on se donne pour recompléter un stock tampon (de produits finis, de composants, ou encore de matières premières, plus de détails plus loin);

    c. Le Facteur de Délai (FD) : c'est un paramètre intrinsèque à la méthode, qui permet de définir la taille de lot minimale et le nombre d'ordre de réapprovisionnement en cours, puisqu'il intervient dans le calcul de la taille de lot minimale (Figure 2.6). Les auteurs préconisent de former des catégories de stocks tampons en fonction des délais de réapprovisionnement, et ainsi de choisir un FD entre 20% et 100% en respectant la règle suivante : plus le DLT est long, moins le FD doit être grand (et inversement). Ceci favorise des tailles de lot plus petites pour les produits ayant des délais importants; et

    d. Le Facteur de Variabilité (FV) : autre paramètre intrinsèque, il est lié à la variabilité de la demande et des fournisseurs, et permet de dimensionner le stock de sécurité des stocks tampons. Ce paramètre varie entre 0% et 100%, en fonction de la variabilité.

Ces quatre paramètres permettent de dimensionner les trois zones du stock tampon (Figure 2.6). La zone rouge représente ainsi le stock de sécurité du stock tampon, qu'il faut éviter de franchir. La zone jaune est le stock d'en-cours, c'est pourquoi elle est dimensionnée en multipliant la demande moyenne par le DLT : ainsi, lorsqu'un ordre est lancé, le stock physique parcours la zone jaune puis remonte lors de la mise en stock, qui devrait se faire avant d'entamer la zone rouge si les temps de réponse sont respectés. Et enfin la zone verte représente la taille de lot minimale à produire.

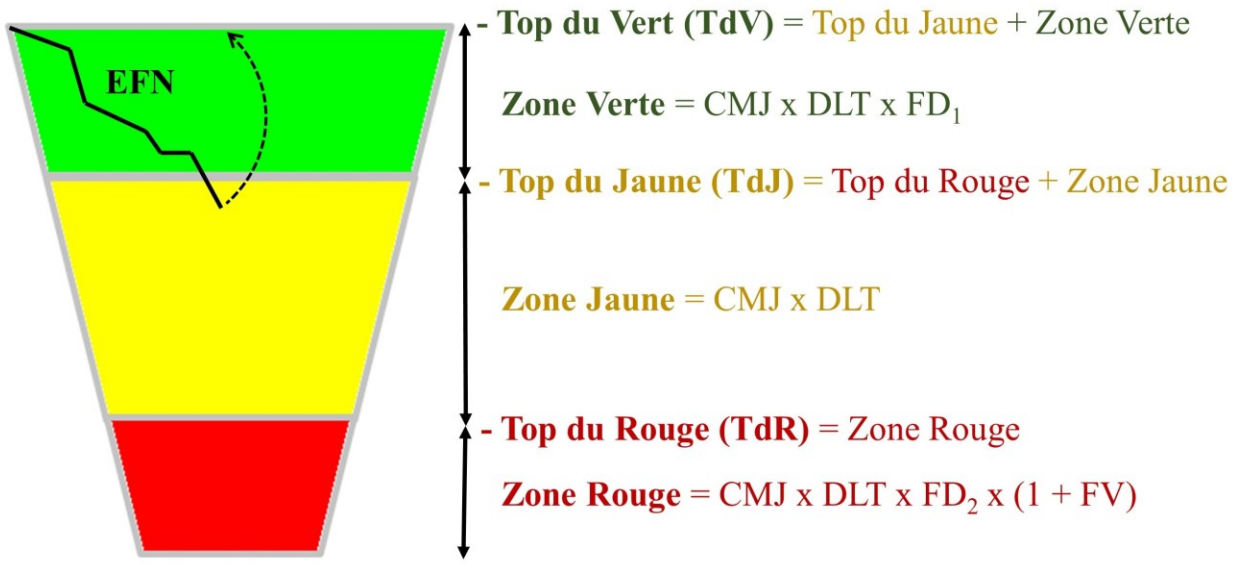


Figure 2.6 : Dimensionnement des zones d'un stock tampon de stock DDMRP[2]

3) L'ajustement dynamique : étape qui consiste à ajuster les zones précédemment définies en fonction de l'estimation de la demande (accroissement de la demande, saisonnalité, fin de vie d'un produit, etc.). Ainsi, le paramètre CMJ joue un rôle de prévision de la demande et est amené à évoluer dans le temps.

4) La planification pilotée par la demande : il s'agit ici de générer des ordres d'approvisionnement (pour les matières premières pilotées par un stock tampon) ou de fabrication (pour les composants et les produits finis pilotés par un stock tampon). Pour cela, on se base sur l'Équation de Flux Net (EFN) pour déterminer la position de stock, dont la formule est la suivante (Équation 2) :

$$Position\ de\ stock = Stock\ physique + En\text{-}cours - Demande\ qualifiée \quad (2)$$

Où la demande qualifiée est la somme de la demande du jour, des pics de commande détectés et des demandes en retard. Un pic de commande correspond à une somme de commandes journalières dépassant un seuil fixé sur un horizon donné. Pour cela, il faut

[2] Il est à noter que le Facteur de Délai sert à dimensionner la zone verte et la zone rouge. Il peut cela dit avoir une valeur différente (mais proche selon les auteurs de la méthode DDMRP) pour les deux zones, c'est pourquoi il est noté $FD_1$ et $FD_2$ sur la Figure 2.6.

définir deux autres paramètres : le seuil de détection des pics, qui va permettre de définir si la somme des commandes d'un jour est un pic ou non, et l'horizon de détection, qui permet de fixer l'horizon sur lequel on considère les pics de commande.

Dans la Figure 2.7, par exemple, le seuil est fixé à 100 pièces et l'horizon est de six jours. Dans le calcul de la demande qualifiée, on prend en compte les 50 pièces du jour 1, ainsi que les 120 pièces du jour 5. S'il n'y a pas de demande en retard à livrer également, la demande qualifiée d'aujourd'hui est de 170 pièces.

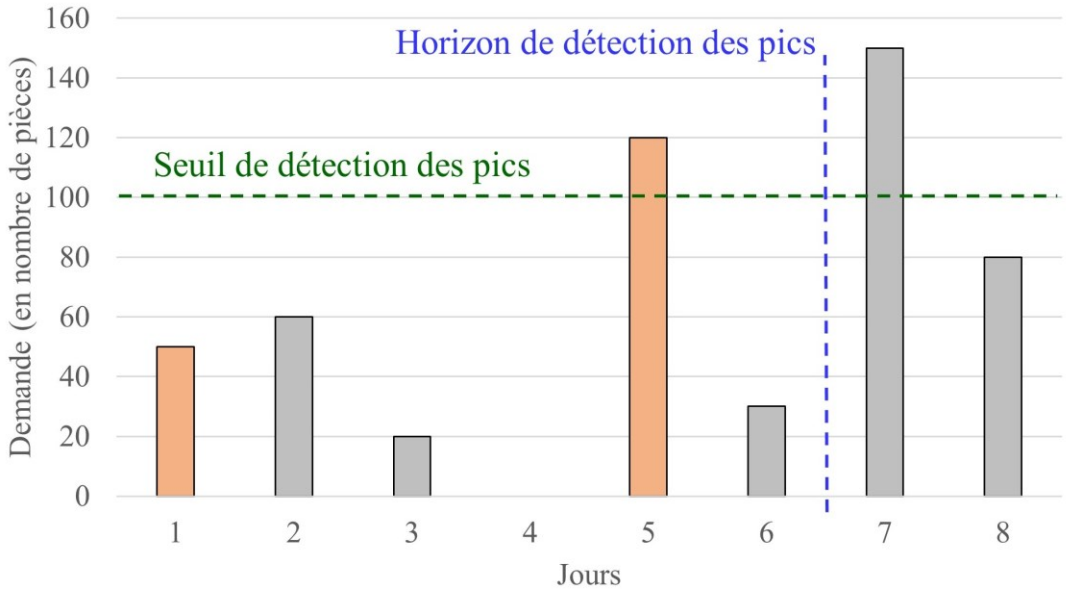


Figure 2.7 : Illustration d'un calcul de la demande qualifiée

Lorsque la position de stock donnée par l'EFN tombe en-dessous du Top du Jaune du stock tampon, un ordre de réapprovisionnement est créé pour recompléter le stock tampon à son Top du Vert, comme illustré sur la Figure 2.6. Ainsi, bien qu'un produit piloté par un stock tampon soit généralement fabriqué en MTS, une commande importante peut déclencher la fabrication plus tôt que prévu, se rapprochant d'un fonctionnement en MTO. De plus, Ptak and Smith (2016) définissent la priorité des ordres en fonction du ratio entre la position de stock calculé et le seuil de réapprovisionnement Top du Jaune : plus ce ratio est faible, plus le stock tampon a été consommé, plus il est prioritaire.

5) L'exécution visuelle et collaborative : étape servant à surveiller les ordres générés à l'étape précédente, et à lancer des alertes si besoin.

Ce qu'on appelle « DDMRP » est en réalité une partie de la méthode complète, qui se nomme *Demand Driven Adaptive Enterprise* (DDAE). Le DDAE, développé plus tardivement par Ptak and Smith (2018), est composé de trois blocs principaux, chacun ayant une vision et un but précis (Figure 2.8).

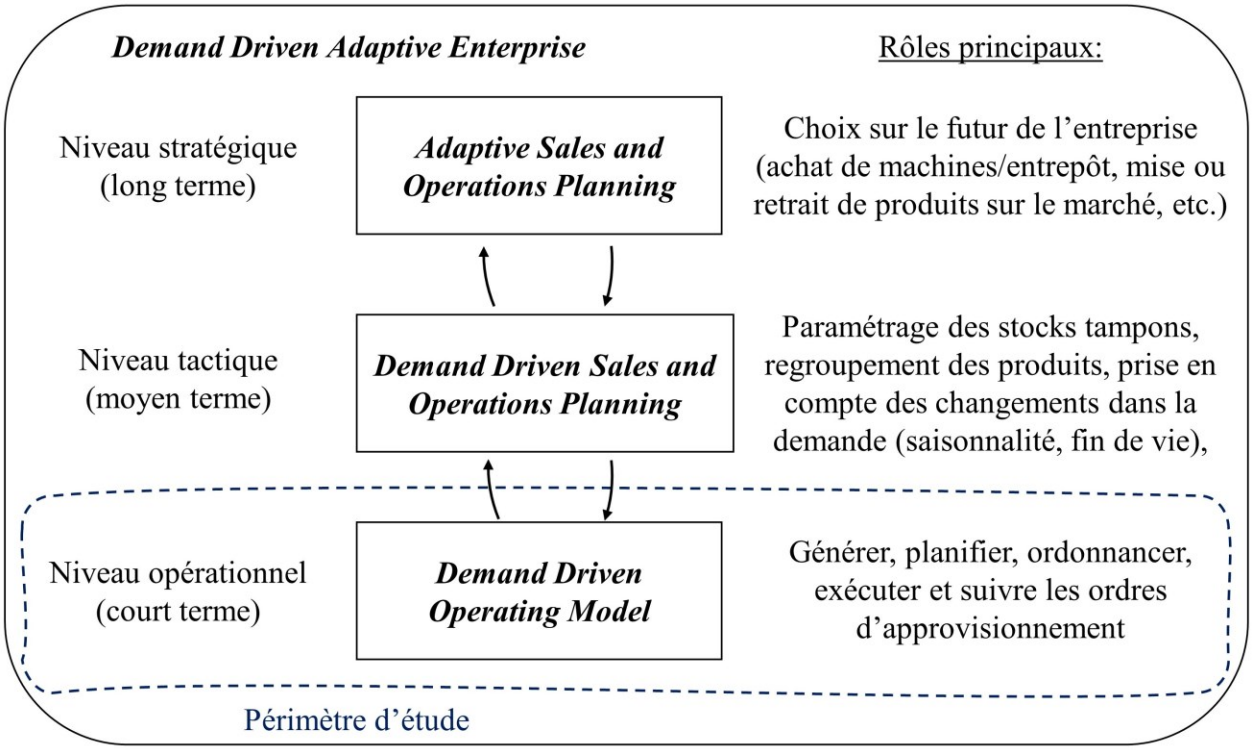


Figure 2.8 : Les différentes parties du *Demand Driven Adaptive Enterprise* de Ptak and Smith (2018)

Les prises de décisions stratégiques se font dans l'*Adaptive Sales & Operations Planning* (AS&OP), où la vision est à long terme pour décider du futur lointain (*c.-à-d.* les prochaines années) de l'entreprise (achat de nouvelles machines, mise sur le marché de nouveaux produits, changement de stratégie de protection des produits par pilotage sur stock tampon, etc.).

Les décisions tactiques, comme le paramétrage ou la catégorie d'affectation des stocks tampons, se font par le *Demand Driven Sales & Operations Planning* (DDS&OP). L'horizon de prise de décision est généralement au mois, pour permettre une adaptation du modèle de production aux variations de l'environnement à moyen terme.

Enfin, le niveau opérationnel est géré par le *Demand Driven Operating Model* (DDOM). Il s'agit ici de gérer les opérations de production, à court terme (semaine, jour, ou heure). Il est composé de trois parties : le DDMRP, qui génère les ordres d'approvisionnement et de production; l'ordonnancement, qui sert à organiser l'enchainement des ordres, si besoin; et l'exécution, pour surveiller le déroulement des étapes de production ou distribution.

Ces trois blocs s'échangent des informations en continu puisque le DDOM va remonter des informations du terrain comme variables d'entrée du DDS&OP, qui va prendre des décisions impactant le DDOM, tout en subissant celles de l'AS&OP, qui utilise les informations du DDS&OP pour se projeter dans le futur.

Dans nos travaux de recherche, on s'intéresse surtout à la génération des ordres de fabrication et à leur suivi. Notre périmètre d'étude porte alors en partie sur le DDOM. On verra qu'en réalité, le deuxième article (chapitre 5), peut s'appliquer à n'importe quel niveau de prise de décision, en fonction des leviers de capacité.

Ce qu'il est important de retenir est la mécanique triviale de génération des ordres de réapprovisionnement du DDMRP : chaque fois que la position de stock (donnée par l'EFN) tombe en dessous du seuil appelé Top du Jaune, on génère un ordre pour recompléter cette position jusqu'au seuil appelé Top du Vert.

### 2.2.2 Publications sur le DDMRP

Dans la littérature, on distingue deux grands courants de recherche sur le DDMRP : les études le comparant avec d'autres méthodes de gestion, et les études portant sur la méthode elle-même, dans le but de l'améliorer.

Les publications originelles sur le DDMRP correspondent au premier courant de recherche : la comparaison. Champ de recherche inévitable dans le génie industriel, on cherche toujours à savoir quelle méthode est la plus performante dans différents environnements. C'est ainsi que Ihme and Stratton (2015) ont comparé le DDMRP au traditionnel MRP, en simulant les modèles sur Excel et SQL. Il s'ensuivit plusieurs travaux comparant, par simulation, un atelier géré en DDMRP, en MRP, et en système Kanban par (Miclo et al., 2015, 2016a; Miclo et al., 2018). Shofa and Widyarto (2017) comparent MRP et DDMRP par la simulation d'une entreprise automobile en Indonésie, tout comme Kortabarria et al. (2018) le font par collecte de données réelles, avant et après

déploiement de la méthode dans une entreprise de serrurerie. Ces comparaisons sont toujours d'actualité, puisque dernièrement Thürer, Fernandes, and Stevenson (2020) comparent le DDMRP, MRP, Kanban et le *Drum-Buffer-Rope* (de la théorie des contraintes) sur des processus d'assemblage en multiproduits, où le MRP offre les pires résultats tandis que le DDMRP obtient les meilleures performances.

Toutes ces comparaisons convergent vers les mêmes résultats : la méthode DDMRP offre de meilleures performances que les autres en général, en termes de compromis entre faible niveaux de stocks (d'en-cours et de produits finis) et taux de service client élevé. La pertinence et les forces de la méthode viennent notamment de l'anticipation des pics de demande dans le calcul de la position de stock, de l'ajustement dynamique des stocks tampons, et de son aptitude à fonctionner dans des environnements à forte variabilité grâce à l'utilisation des stocks tampons.

L'autre branche de recherche sur le DDMRP porte sur l'amélioration directe de la méthode, ou des propositions d'amélioration. Les chercheurs abordent les trois différents niveaux de prise de décision, avec par exemple Vidal et al. (2020) qui s'attaquent au niveau stratégique en proposant un modèle AS&OP par agrégation de données; Martin et al. (2019) qui proposent un arbre de décision pour aider au paramétrage des stocks tampons lors du DDS&OP; et Dessevre, Martin, Baptiste, Lamothe, Pellerin, et al. (2019) qui cherchent à mettre sous contrôle le DLT, entre les niveaux tactique et opérationnel, où la mise à jour du paramètre se fait par retour d'informations venant de l'atelier. C'est un paramètre important puisqu'il est à la fois relié aux niveaux de stocks (il sert à dimensionner les zones des stocks tampons), et également lié au taux de service puisqu'un ordre ne respectant pas ce DLT arrivera en retard, causant des pénuries pour les clients. Lee and Rim (2019) proposent une méthode alternative à celle de Ptak and Smith (2016) pour calculer la zone rouge des stocks tampons (*c.-à-d.* le stock de sécurité), permettant de réduire les niveaux de stocks tout en maintenant un taux de pénurie faible. Velasco Acosta, Mascle, and Baptiste (2019) étudient l'impact du positionnement des stocks tampons lors de nomenclatures à quatre étages sur les performances de l'entreprise, et posent notamment la question du choix de positionnement des stocks tampons lors de nomenclatures encore plus complexes. Enfin, Achergui, Allaoui, and Hsu (2020) développent un algorithme pour résoudre le problème d'optimisation du positionnement stratégique des stocks tampons DDMRP.

Les propositions d'amélioration de la méthode font aussi l'objet de plusieurs publications. Ainsi, Baptiste (2018) expose les opportunités d'ordonnancement de la production, lors de nomenclature à plusieurs étages offrant une visibilité à court ou moyen terme; Dessevre et al. (2020) répertorient les neufs principaux obstacles à la mise en œuvre de la méthode DDMRP dans une industrie dermo-cosmétique, comme les problèmes de *double sourcing* en approvisionnement, le pilotage d'un processus divergent en production, et la gestion capacitaire en distribution; et Orue, Lizarralde, and Kortabarria (2020) déplorent le manque de guide pour déployer la méthode efficacement et invitent les chercheurs à déterminer un processus standardisé de mise en œuvre de la méthode DDMRP.

Finalement, Bagni et al. (2021) montrent que le DDMRP est la seule méthode, parmi la douzaine de méthodes émergentes ces vingt dernières années, à avoir été étudiée et développée à la fois en théorie et en pratique. Vous trouverez plus de détails sur toutes ces publications dans la revue systématique de Azzamouri et al. (2021), qui proposent également des perspectives de recherche.

En conclusion, la méthode DDMRP est de plus en plus déployée dans l'industrie et étudiée dans le monde de la recherche, les comparaisons aux méthodes traditionnelles ont prouvé son efficacité, et les recherches sont très variées et touchent tous les niveaux de prise de décision. De plus, le paramètre DLT est un élément important du paramétrage de la méthode, lié aux niveaux de stocks et au taux de service. Enfin, de nombreuses pistes de recherche sont encore à creuser, notamment les questionnements industriels tels que la gestion de la capacité et le pilotage d'un processus divergent en DDMRP.

Pour comprendre le reste de la thèse, qui ne porte pas uniquement sur le DDMRP, il est important de connaître d'autres concepts sur les méthodes de production. Le chemin de pensée, illustré précédemment en Figure 2.1, est le suivant : le système de réapprovisionnement du DDMRP est similaire à celui d'un point de commande (2.3); le contrôle des temps de cycle peut se faire par une boucle ConWIP (2.4); qui aide à maîtriser une partie des temps de réponse pour qu'ils soient inférieurs au délai de production (ici le DLT); et enfin, un temps de réponse est principalement constitué de temps d'attente (2.5).

## 2.3 Les méthodes à point de commande

Le modèle de réapprovisionnement de la méthode DDMRP est similaire aux méthodes à point de commande, c'est d'ailleurs pourquoi dans le chapitre 6, nous comparons le modèle classique DDMRP, notre modèle DDMRP-ConWIP Intégré, et une méthode à point de commande.

Une méthode à point de commande est une technique de réapprovisionnement en flux tiré qui se base sur un point de commande qui, lorsqu'il est atteint par la position de stock (semblable au calcul de l'EFN vue précédemment), déclenche un ordre d'approvisionnement. On appelle $R$ le seuil de Réapprovisionnement, $Q$ la Quantité à approvisionner et $S$ le Stock maximum. On distingue deux grandes familles dans les méthodes à point de commande (Chaudhary, Kulshrestha, and Routroy, 2018) : les méthodes de recomplètement à quantité fixe, nommées ($R$, $Q$), et les méthodes de recomplètement à quantité variable, ($R$, $S$).

Le modèle ($R$, $Q$) est illustré en Figure 2.5 dans un exemple où $R$ = 70 pièces et $Q$ = 30 pièces : à chaque fois que la position de stock calculée une fois par jour (en pointillée) tombe sous le seuil $R$, on lance un ordre de réapprovisionnement d'une taille de lot fixe $Q$. L'en-cours est représenté en bleu clair, et le stock de produits finis en bleu foncé. Lorsque l'ordre est terminé, le stock remonte donc de $Q$ pièces.

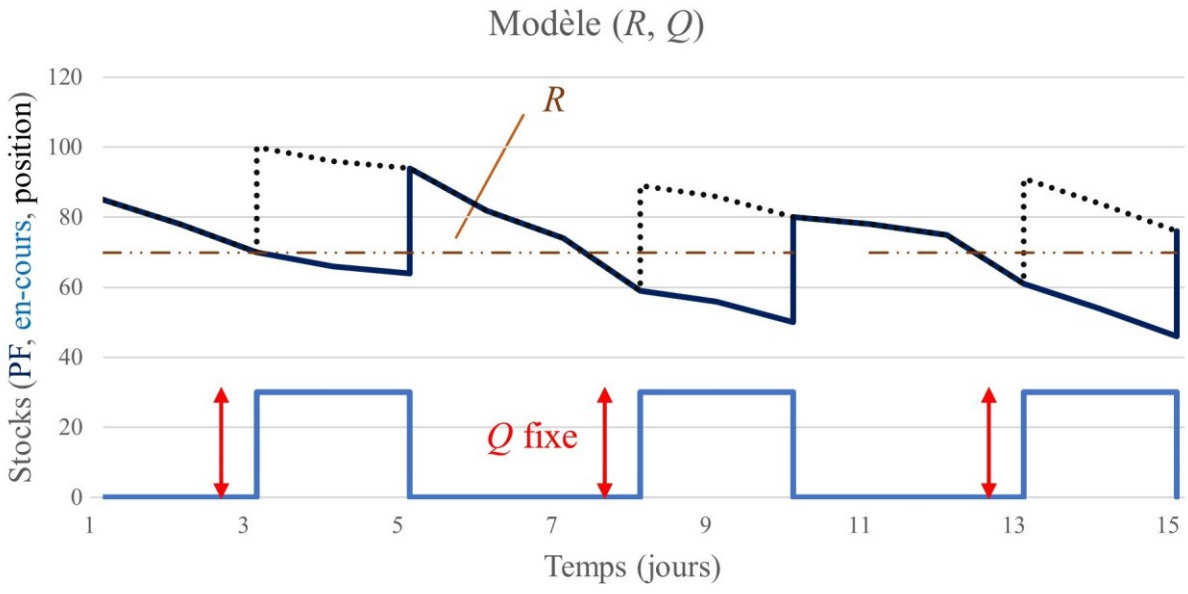


Figure 2.9 : Illustration du modèle ($R$, $Q$)

Ce modèle de point de commande existe depuis plus d'un demi-siècle, l'une des premières publications étant de Hadley and Whitin (1963) qui proposaient un algorithme pour trouver les paramètres optimaux $R$ et $Q$ permettant de minimiser les coûts totaux.

Depuis, de nombreux chercheurs se sont penchés sur ce modèle, le but étant toujours de trouver les meilleurs paramètres $R$ et $Q$ dans différentes situations : Nahmias and Wang (1979) se sont focalisés sur les stocks de produits qui se détériorent dans le temps; De Bodt and Graves (1985) se sont attaqués à un modèle multi-étages avec une demande stochastique; et Chang, Yao, and Lee (1998) et Kao and Hsu (2002) ont étudié les problèmes de demande « floue ».

Initialement fixe, la quantité $Q$ à réapprovisionner a fait l'objet d'études où elle devient variable. Dans les travaux de Axsäter (2005) et Jodlbauer and Dehmer (2020) par exemple, la quantité à réapprovisionner est un multiple $n$ de la taille de lot fixe Q. Ce modèle ($R$, $nQ$) permet entre autres des tailles de lot optimales plus petites que le modèle ($R$, $Q$), la réduction du coût total d'inventaire bien que le seuil $R$ soit plus élevé, et enfin la réduction des coûts de mise en course lorsqu'il faut produire plusieurs lots de fabrication.

L'autre méthode à point de commande populaire est le modèle ($R$, $S$) où $R$ correspond toujours au seuil de Recomplètement, et $S$ correspond au niveau de Stock maximum. Ainsi, lorsque la position de stock atteint le seuil $R$, on lance un ordre de réapprovisionnement avec une taille de lot variable $Q = S - R$, pour remonter la position au niveau maximum. Le modèle ($R$, $S$) est illustré en Figure 2.6 dans le même exemple que précédemment, avec $R = 70$ pièces et $S = 100$ pièces. La différence entre les deux méthodes réside dans la quantité $Q$ à approvisionner, fixe ou variable.

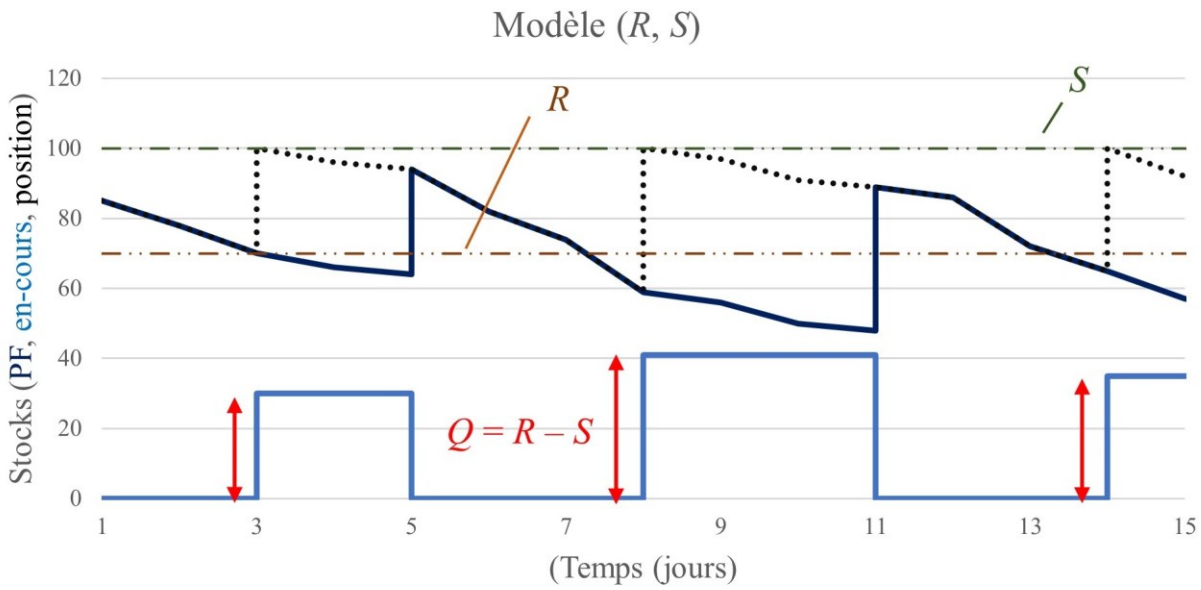


Figure 2.10 : Illustration du modèle (*R*, *S*)

Dans la littérature, le but est toujours de trouver les paramètres optimaux selon l'environnement d'étude : pour cela, Veinott Jr and Wagner (1965) font partie des premiers à utiliser des approches informatiques; Liu (1990) se penche sur le cas où la demande suit une loi de Poisson et la vie des produits suit une loi exponentielle; Liu and Lian (1999) s'attaquent aux modèles de produits périssables avec réapprovisionnement instantané, etc. Plus de détails sont disponibles dans la revue de littérature de Khanlarzade et al. (2014) sur les modèles dont les produits se détériorent dans le temps, et l'état de l'art de Chaudhary, Kulshrestha, and Routroy (2018) sur les produits périssables.

Ainsi, le modèle (*R*, *Q*) permet de connaître à l'avance la taille de lot à fabriquer, donnant une prévision du taux de charge et un approvisionnement pour les matières premières plus précis, tandis que le modèle (*R*, *S*) permet de mieux adapter la production aux variations de la demande. Dans la littérature, la plupart des études considèrent que les délais d'approvisionnement sont variables, mais la cause de cette variabilité est externe au système : il n'y a pas de relation entre la quantité commandée et le délai d'approvisionnement. Ce constat peut être valide pour des ordres d'approvisionnement de matières premières par exemple (car la capacité est externe), mais semble moins valide pour des ordres de production où la taille de lot est liée à la charge des machines, qui est elle-même liée aux temps de réponse (comme nous verrons dans la théorie des files d'attente). Kim and Benton (1995) proposent un modèle mathématique reliant la quantité *Q* au temps de réponse, mais nous verrons que leur raisonnement peut être remis en cause.

Il existe également des méthodes à période fixe, c'est le cas dans la production cyclique ou dans les roues de production (Koenigsberg, 1959; King and King, 2013). Ces méthodes s'éloignent du fonctionnement du DDMRP (bien qu'il possède une certaine périodicité lui aussi, si la demande est stable), c'est pourquoi elles ne sont pas étudiées.

En conclusion, il existe deux grandes familles de méthodes à point de commande : les modèles ($R$, $Q$) et les modèles ($R$, $S$). Les deux méthodes sont étudiées depuis des dizaines d'années, et bien que les chercheurs proposent des solutions pour trouver les meilleurs paramètres ($R$, $Q$ ou $S$) en fonction des caractéristiques des problèmes (type de demande, de produit, distribution des délais d'approvisionnement, etc.), ils ne considèrent pas de relation entre la quantité à réapprovisionner et le délai de réapprovisionnement. De plus, la mécanique de lancement des ordres d'approvisionnement de la méthode DDMRP peut être considérée comme une méthode à point de commande ($R$, $S$) où $R$ est le « Top du Jaune » et $S$ le « Top du Vert » : il s'agit d'un modèle (Top du Jaune, Top du Vert). Les nuances apportées par le DDMRP sont que : (i) les paramètres ($R$, $S$) évoluent au cours du temps du fait de l'adaptation de la prévision de demande (paramètre CMJ); et (ii) le DDMRP intègre la notion de pics dans le calcul de sa position de stock, permettant d'anticiper les commandes importantes pouvant engorger le système.

## 2.4 Le ConWIP pour contrôler les temps de cycle

Parmi les méthodes et outils de gestion de la production, la méthode Constant Work-In-Process (ConWIP) se démarque du lot pour quatre raisons principales (Jaegler et al., 2018) : elle permet de contrôler les temps de cycle, elle donne de très bons résultats pour gérer les ateliers de type ligne de production, elle permet de gérer une très grande variété de produits facilement (alors que d'autres méthodes comme le Kanban en sont incapables), et enfin elle s'intègre très facilement à n'importe quelle méthode de génération d'ordres (MRP, DDMRP, point de commande, etc.). C'est pourquoi elle nous intéresse dans notre quête de maîtrise des temps de réponse.

La méthode ConWIP se base sur la loi de Little (1961) :

$$En\text{-}cours = Temps\ de\ cycle \times Cadence \tag{3}$$

Où l'en-cours (WIP en anglais) est le volume de pièces en cours de production, le temps de cycle (défini dans la partie 2.1) est le temps de parcours dans la boucle (la « boucle » est expliquée juste après), et la cadence correspond au débit de production. L'idée est qu'en maîtrisant l'en-cours, on

maîtrise les temps de cycle, pour une cadence donnée. Pour cela, on délimite le périmètre d'action du ConWIP, que l'on appelle boucle, et on utilise des tickets que l'on attache aux ordres de fabrication. Lorsqu'un ordre de fabrication, issu de n'importe quelle méthode de génération d'ordres, arrive devant la boucle ConWIP, on lui associe un ticket qu'il va garder jusqu'à sa sortie de la boucle (correspondant à l'entrée au stock de produits finis en général). Si aucun ticket ConWIP n'est disponible, l'ordre attend qu'un ticket soit relâché par un des ordres de fabrication dans la boucle lorsqu'il sera terminé.

Ainsi, en maîtrisant l'en-cours, on maîtrise les temps de parcours dans la boucle, ce qu'on appelle temps de cycle. Le temps total lui, peut-être plus long. C'est pourquoi dans la Figure 2.2 (partie 2.1) le temps de cycle ne représente qu'une partie du temps de réponse : il peut y avoir du temps avant la boucle (temps passé sur d'autres machines par exemple, ou temps d'attente d'un ticket), et du temps après (transfert vers un autre atelier, en-dehors de la boucle).

La Figure 2.11 représente un atelier composé de trois machines consécutives (une ligne de production à trois étages) avec une boucle ConWIP autour des machines. Le nombre de tickets est fixé à six. Dans le premier cas, il reste trois tickets disponibles. Si un nouvel ordre de fabrication arrive, il va s'emparer d'un des tickets et entrer directement dans l'atelier. Dans le deuxième cas, les six tickets sont utilisés, c'est pourquoi il y a deux ordres en attente d'entrer dans la boucle. Lorsque l'ordre de fabrication sur la troisième machine sera fini, les produits finis seront stockés et il relâchera un ticket. Un nouvel ordre pourra alors entrer dans la boucle. Le ticket ConWIP, à la différence d'une étiquette Kanban, peut être relié à n'importe quel ordre de fabrication, quel que soit le type de produit. Il représente plus une charge de travail qu'une référence de produit (Spearman, Woodruff, and Hopp, 1990), c'est d'ailleurs pourquoi il est facilement utilisable même si l'on doit gérer une grande variété de produits différents.

Ainsi, une boucle ConWIP permet de maîtriser les temps de cycle (qui seront variables, mais assez proches les uns des autres), mais pas forcément les temps de réponse. Avec une boucle ConWIP, on déplace les temps d'attente initialement à l'intérieur de la boucle vers l'avant de cette boucle : au lieu d'attendre 4h devant une machine, l'ordre va attendre 4h devant la boucle, son temps de cycle va donc être diminué de 4h, mais pas forcément son temps de réponse. Pour réduire le temps de réponse, il faut utiliser la boucle et les temps d'attente à bon escient.

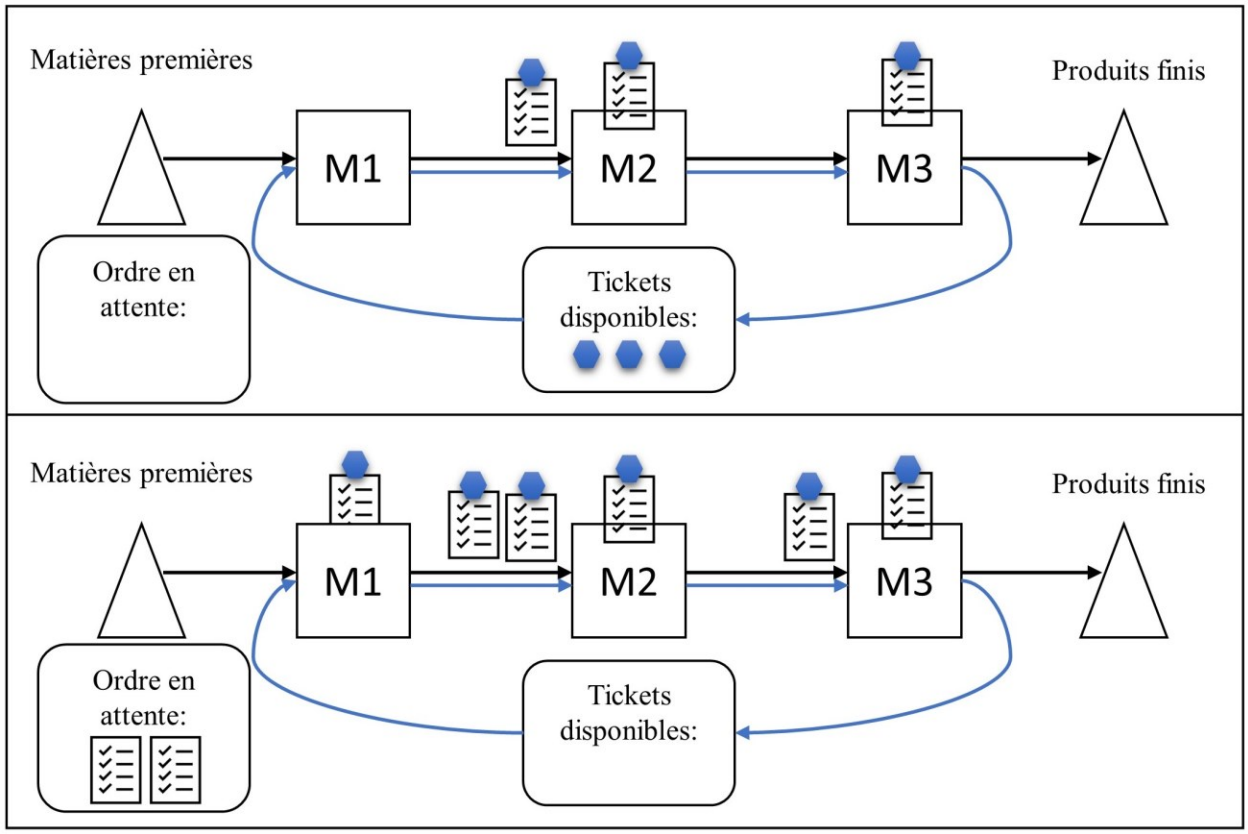


Figure 2.11 : Fonctionnement d'une boucle ConWIP

*Une analogie très simple à faire aujourd'hui pour comprendre ce fonctionnement est le nombre maximal de personnes autorisées en même temps dans une épicerie pendant une pandémie. Si par exemple, pour des raisons de distanciation physique, il ne peut y avoir plus de 20 clients en même temps : pour qu'un nouveau client entre dans l'épicerie, il faut qu'un autre en sorte. En limitant le nombre de chariots à 20 et en obligeant les clients à prendre un chariot en entrant, on reproduit le fonctionnement d'un ConWIP. Par conséquent, les clients passent moins de temps dans les files d'attente des caisses, donc moins de temps dans la « boucle ConWIP » de l'épicerie. Les temps d'attente sont déplacés avant la boucle, c.-à-d. devant l'épicerie.*

Dans la littérature, selon la récente revue de Jaegler et al. (2018), il existe quatre grands champs de recherche sur la méthode ConWIP.

Le premier est le dimensionnement des paramètres du ConWIP. Il y a deux paramètres : le nombre de tickets ConWIP et la taille de lot des ordres de fabrication (Spearman, Woodruff, and Hopp, 1990; Hopp and Roof, 1998). L'en-cours moyen dans la boucle est déterminé par la taille de lot (moyenne) multipliée par le nombre de tickets (Équation 4).

$$En\text{-}cours = taille\ de\ lot \times nombre\ de\ tickets \quad (4)$$

Si on a 10 tickets avec une taille de lot moyenne de 20 pièces, alors il y a 200 pièces dans la boucle. Et si la cadence est de 50 pièces à l'heure, alors le temps de cycle moyen est de quatre heures. En réduisant l'en-cours de moitié, le temps de cycle l'est aussi, tant que la cadence reste inchangée. Le risque est qu'en diminuant trop l'en-cours, la ressource goulot va se désamorcer (elle n'aura rien à produire alors qu'elle le devrait), et l'atelier va perdre en cadence. C'est pourquoi le choix du nombre de tickets et de la taille de lot est un compromis entre faible en-cours, faible temps de cycle et forte cadence.

Deux approches sont principalement utilisées dans la littérature pour déterminer ces paramètres : le dimensionnement statique du nombre de tickets, comme le font Hopp and Spearman (1996) par une formule, Marek, Elkins, and Smith (2001) par une méthode heuristique, et Pergher and Vaccaro (2014) par une méthode de tri multicritères; et le dimensionnement dynamique, permettant d'ajuster le nombre de tickets dans la boucle (Hopp and Roof, 1998; Tardif and Maaseidvaag, 2001; Belisário and Pierreval, 2015).

Le deuxième champ de recherche est l'environnement dans lequel est intégrée la méthode ConWIP. Comme l'explique Stevenson, Hendry, and Kingsman (2005), le choix d'un système de contrôle de la production, comme le ConWIP, dépend de la configuration de l'atelier et des caractéristiques de la demande (sur stock ou sur commande). Dans la littérature, la majorité des publications sur le ConWIP étudie des lignes de production en MTS.

Le troisième axe de recherche porte sur la comparaison du ConWIP aux autres méthodes de contrôle de la production. Ainsi, le ConWIP a été comparé aux méthodes à flux poussé (Bahaji and Kuhl, 2008), au Kanban (Marek, Elkins, and Smith, 2001), au ConWIP modifié (Takahashi and Nakamura, 2002; Prakash and Chin, 2015), au Base-Stock (Khojasteh, 2015), à la méthode POLCA (Harrod and Kanet, 2013), ou encore au COBACABANA (Land, 2009). En conclusion de ces articles, la méthode ConWIP n'est pas toujours la plus efficace, notamment dans des environnements complexes, mais elle reste la méthode la plus facile à mettre en place et à maintenir. Il existe également des méthodes couplant le ConWIP avec d'autres méthodes de gestion de la production, telles que le système Hybrid Kanban ConWIP, le Base Stock Kanban ConWIP ou encore le Base Stock ConWIP (Onyeocha et al., 2015; Al-Hawari, Qasem, and Smadi, 2018). Toutes ces méthodes diffèrent sur la remontée de la demande client dans le système : quand et où

remonter l'information ? À la station précédente ou à toutes les stations en même temps, quand un ticket est relâché ou quand il est pris, etc.

Le dernier champ de recherche repose sur les méthodologies employées pour étudier le ConWIP, où la simulation est largement plus utilisée que les modèles mathématiques. Plus de détails sont disponibles dans les revues de littérature de Framinan, González, and Ruiz-Usano (2003) et de Jaegler et al. (2018).

En conclusion, la méthode ConWIP semble être très efficace et simple à mettre en place, et peut s'intégrer à n'importe quelle méthode générant des ordres de fabrication. Aucun couplage DDMRP-ConWIP n'a encore été étudié dans la littérature. Ce couplage pourrait permettre de maîtriser les temps de cycle dans la boucle, mais peut-être pas les temps de réponse (la différence est expliquée dans partie 2.1). Enfin, puisque le DDMRP détermine les tailles de lot, le seul paramètre à dimensionner est le nombre de tickets dans la boucle ConWIP.

## 2.5 La théorie des files d'attente et le ratio charge/capacité

Les files d'attente, il y en a partout : des ordres de fabrication dans un atelier aux embouteillages sur l'autoroute, en passant par les guichets de poste. Mais qu'est-ce qui fait qu'un jour il n'y a personne à la caisse d'un supermarché, et que le lendemain il faut faire la queue pendant plus d'une heure ? Réponse : la variabilité et le ratio charge/capacité.

En réalité, il y a trois facteurs principaux qui expliquent les temps d'attente dans une file (Kingman, 1962) :

1) Le temps de service moyen (temps que prend un ordre de fabrication pour être traité sur une machine, une voiture pour payer au péage, ou encore temps qu'il faut pour faire passer les articles à la caisse et payer);
2) Le taux de charge (noté $\rho$) de la ressource utilisée (machine, station de péage, caissier ou caissière), appelé aussi ratio charge/capacité; et
3) La variabilité des temps de service et des temps d'arrivée.

Le formule la plus connue pour estimer le temps moyen d'une file d'attente devant une ressource est celle de Kingman (1962), surnommée « l'équation VUT » (Équation 5) :

$$T_{attente} = \left(\frac{cv^2_{arrivée} + cv^2_{service}}{2}\right) \times \left(\frac{\rho}{1-\rho}\right) \times T_{service} \quad (5)$$

On retrouve dans cette formule nos trois facteurs : le temps de service moyen $T_{service}$ (moyenne de la somme des temps de mise en course et de production de chaque ordre de fabrication pour une machine par exemple), le taux de charge $\rho$, et les coefficients de variabilité des temps de service $cv^2_{service}$ et d'arrivée $cv^2_{arrivée}$, définis comme le ratio entre l'écart-type et la moyenne des temps de service ou d'arrivée. Ces derniers permettent de mesurer la variabilité d'un modèle : plus le coefficient est faible, moins le modèle est variable (si les $cv$ sont nuls dans l'équation, alors il n'y a pas de temps d'attente !).

Dans les faits, cette formule fonctionne très bien pour les modèles ayant une seule ressource et dont les temps de service et d'arrivée suivent une loi exponentielle (les coefficients valent alors 1), mais elle n'est pas précise pour les autres situations. D'autres chercheurs ont proposé des formules plus précises, comme Marchal (1976) et Krämer and Langenbach-Belz (1976). Bien que les formules datent du siècle dernier, la recherche d'approximation des temps d'attente est toujours d'actualité, comme le montrent Wu, Srivathsan, and Shen (2018). Le plus important à retenir de ces équations est que le temps d'attente moyen dépend surtout de la variabilité du modèle et du taux de charge $\rho$ de la ressource. La variabilité est un paramètre difficilement contrôlable, qui est en général subi, alors que le taux de charge lui est plus maniable (on ne choisit pas le nombre de clients qui arrivent à un bureau de poste, mais on peut choisir le nombre de guichets à ouvrir).

De plus, l'expression sur le taux de charge dans la formule de Kingman est la fonction $f(\rho) = \left(\frac{\rho}{1-\rho}\right)$, présentée graphiquement dans la Figure 2.12 pour un taux de charge allant de 50% à 100%.

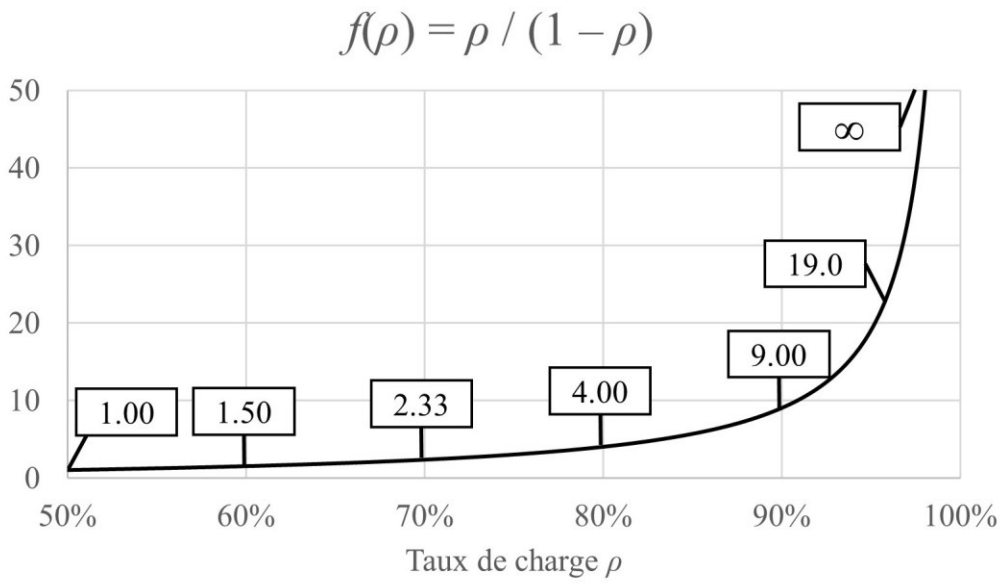


Figure 2.12 : Fonction $f(\rho) = \rho / (1 - \rho)$

La courbe de la fonction a une croissance infinitésimale et possède une asymptote lorsque $\rho$ s'approche de 100%. Par exemple, pour un taux de charge $\rho$ de 50% on a $f(0.5) = 1$, pour $\rho = 75\%$ on a $f(0.75) = 3$ et pour $\rho = 95\%$ on a $f(0.95) = 19$. Ainsi, plus le taux de charge augmente, plus le temps d'attente augmente, et ce drastiquement. Dans des environnements variables où la somme des coefficients $cv^2_{service}$ et $cv^2_{arrivée}$ vaut 1, le taux de charge représente le pourcentage du temps d'attente sur le temps total.

*Dans un supermarché où règne de la variabilité, disons que les coefficients de variabilité valent 1 par exemple, si la caisse est utilisée à 80%, alors le temps d'attente sera en moyenne quatre fois plus long que le temps de service (et il représente donc 80% du temps total).*

C'est pourquoi il faut éviter de (sur)charger une ressource au-delà des 85-90%. Le choix de la charge sur une ressource goulot est donc un compromis entre maximiser la charge pour produire le maximum possible (ce qui revient à augmenter la cadence), tout en minimisant les temps d'attente en évitant de la surcharger.

Pour essayer de réduire le taux de charge, il faut d'abord comprendre comment le calculer. Dans la littérature, le taux de charge est généralement exprimé comme le ratio entre le temps moyen de service $T_{service}$ et le temps moyen entre deux arrivées $T_{arrivée}$ :

$$\rho = \frac{T_{service}}{T_{arrivée}} \tag{6}$$

Par exemple, si un client arrive toutes les cinq minutes à un guichet et qu'il faut en moyenne quatre minutes pour traiter sa demande, le taux de charge du guichet est de 4/5 = 80%.

On peut également exprimer le taux de charge d'une ressource par le ratio entre le temps total d'utilisation de cette ressource et son temps d'ouverture. Dans le milieu industriel, cela revient à sommer les temps de mise en course $T_{mise-en-course}$ d'une machine et les temps de fabrication $T_{fabrication}$, et à les diviser par le temps de disponibilité $T_{ouverture}$ de la machine :

$$\rho = \frac{\sum(T_{mise-en-course} \times N) + \sum(T_{fabrication} \times Q)}{T_{ouverture}} \tag{7}$$

Où *N* correspond au nombre de changements de série à faire et *Q* au nombre total de pièces à produire, durant le temps d'ouverture.

Si dans une semaine de 40h, il faut produire 1 000 pièces par lot de 100 (*c.-à-d. N* = 10) sur une machine, qu'il faut en moyenne une heure pour changer de lot, et que la cadence de production est d'une pièce à la minute, alors le taux de charge $\rho$ de la machine est de :

$$\rho = \frac{(1 \times 10) + \left(\frac{1}{60} \times 1000\right)}{40} \approx \frac{26.67}{40} = 66\% \tag{8}$$

On comprend alors comment on peut réduire le taux de charge, il faut soit diminuer le numérateur, soit augmenter le dénominateur.

La première solution consiste donc à réduire :

- Les temps unitaires de mise en course ou les temps de production. Ce qui est possible jusqu'à un certain point, puis de plus en plus difficile, voire impossible;
- Le nombre de pièces à produire. Ce qui correspond en général à la demande. Si on réduit ce nombre, on ne peut plus satisfaire toutes les commandes. Il faut sinon sous-traiter une partie de la production; et
- Le nombre de lots à produire, pour réduire le temps total de mise en course. Pour cela, il faut augmenter judicieusement les tailles de lots. Dans un contexte DDMRP, on augmente

les tailles de lots en augmentant la Zone Verte, c'est-à-dire en augmentant les paramètres DLT ou FD (Figure 2.6).

Enfin, l'autre solution revient à augmenter le temps d'ouverture. Cette décision peut se faire à n'importe quel niveau, en fonction de la capacité souhaitée, des leviers possibles (embaucher plus d'opérateurs, faire des heures supplémentaires, acheter une nouvelle machine, etc.) et des coûts associés (Taal and Wortmann, 1997; Jodlbauer and Reitner, 2012).

*Dans un supermarché par exemple, l'ouverture d'une caisse supplémentaire permet de désengorger celles déjà bien chargées. Vous pouvez également choisir de faire vos courses une fois par mois au lieu de les faire une fois par semaine, si possible, pour réduire vos temps d'attente.*

C'est pourquoi il est difficile d'établir un modèle mathématique estimant le temps de réponse d'un ordre, car il y a énormément de paramètres à prendre en compte : le temps de service $T_{service}$, qui dépend des temps de mise en course $T_{mise-en-course}$ et du nombre de changement de série $N$, des temps de fabrication $T_{fabrication}$ et de la taille de lot $Q$, du taux de charge $\rho$ qui lui aussi dépend des temps de production, mais également de la capacité des ressources, des coefficients de variabilité des temps de service et d'arrivée $cv_{service}$ et $cv_{arrivée}$, des règles de gestion de l'atelier, d'ordonnancement, etc.

Kim and Benton (1995), étudiant la relation entre taille de lot $Q$ et temps de réponse dans une méthode à point de commande ($R$, $Q$), établissent un modèle mathématique du temps de réponse (qu'ils appellent délai de production (*lead time* en anglais) au passage). En s'inspirant des résultats de Karmarkar (1987), ils estiment le temps de réponse ainsi :

$$Temps\ de\ réponse = \left(T_{mise-en-course} \times N + T_{fabrication} \times Q\right) \times \delta \qquad (9)$$

Où $\delta$ représente le « facteur de file d'attente », terme considérant les temps d'attente dans le système. Ainsi, si $\delta$ vaut 1, il n'y a pas de temps d'attente, s'il vaut 10 alors le temps d'attente représente 90% du temps de réponse. Ils ne donnent pas d'expression mathématique au facteur $\delta$. Le raisonnement est alors le suivant : dans un système où $\delta$ est grand, une hausse de la quantité $Q$ à produire peut entraîner une forte augmentation du temps de réponse. En réalité on vient de voir que c'est plus compliqué, puisqu'en combinant leur formule avec celle de Kingman, on peut estimer mathématiquement le facteur $\delta$ dans un modèle à une ressource:

$$\delta = \left(\frac{cv_{arrivée}^2 + cv_{service}^2}{2}\right) \times \left(\frac{\rho}{1-\rho}\right) \quad (10)$$

Par conséquent, $\delta$ est relié au taux de charge $\rho$, qui lui-même est relié à la quantité à produire $Q$. En produisant des ordres de fabrication plus gros, mais moins souvent, on réduit le taux de charge $\rho$, et donc le facteur $\delta$. Que vous achetiez un article ou que vous fassiez vos courses pour le mois, vous passerez autant de temps dans la file d'attente, tout comme vous passerez autant de temps dans un embouteillage, que vous conduisiez une voiture de sport ou une Renault 21 datant de 1989. C'est pourquoi il est très difficilement d'estimer avec précision les temps de réponse.

Ce qu'il faut retenir de cette sous-section est que : (i) le taux de charge d'une ressource goulot joue un rôle important dans la distribution des temps de réponse; (ii) il existe plusieurs formules permettant de déterminer le temps d'attente des « entités » (ordre de fabrication, véhicule, personne, etc.) dans les files dont l'élément commun est le ratio $\left(\frac{\rho}{1-\rho}\right)$ où $\rho$ représente le taux de charge de la ressource; (iii) plus ce taux de charge augmente en s'approchant des 100%, plus les temps d'attente augmentent drastiquement; et (iv) la taille des lots ou les leviers capacitaires sont des solutions pour diminuer ce taux de charge.

## 2.6 Conclusion

En conclusion générale de cette revue littéraire, on a répondu en partie à nos objectifs de revue et mis l'accent sur les limites ou lacunes à étudier dans la littérature. Ce qui en ressort le plus est que : (i) la méthode DDMRP, récente et donc peu étudiée, peut encore être améliorée; que (ii) le DLT est l'un des paramètres les plus importants de la méthode DDMRP mais qu'aucune publication scientifique n'a été publiée sur l'étude du DLT, que (iii) il est crucial que les temps de réponse soient maitrisés pour maintenir des taux de service atelier et client élevés; que (iv) cette maîtrise passe par le contrôle du taux de charge des ressources; que (v) la méthode ConWIP permet de maîtriser les temps de cycle dans une boucle où l'en-cours est limité par un nombre de tickets; et que (vi) aucune publication ne porte sur le couplage des méthodes DDMRP et ConWIP.

D'où notre problématique suivante : **Faut-il maîtriser les temps de réponses ou ajuster le DLT lors d'une variation anticipée ou observée des temps de réponse dans un atelier piloté en DDMRP ?**

La résolution de cette problématique a pour but d'aider les entreprises ayant déployé (ou voulant déployer) la méthode DDMRP, à mieux paramétrer le DLT ou contrôler les temps de réponse : en contrôlant les temps de réponse (leur moyenne et leur dispersion) ou en adaptant le DLT, on s'assure que les ordres de fabrication soient finis dans les temps (ce qui permet d'avoir des taux de service atelier et client élevés), et en diminuant le DLT, on diminue les stocks (et donc les coûts).

# CHAPITRE 3 MÉTHODOLOGIE

*Nove sed*, *non nova*. La manière est nouvelle, mais non la matière. Citation que j'ai jamais pu replacer correctement dans une conversation. (Alexandre Astier, 2007, Kaamelott Livre V, La roche et le fer)

## 3.1 Objectif principal et méthodologie globale

L'objectif principal de ce travail de recherche est d'améliorer les performances des ateliers de production pilotés par la méthode DDMRP, en proposant des outils de maîtrise des temps de réponse ou d'ajustement du paramètre DLT. Par performance, on entend taux de service client élevé et niveaux de stocks faibles. La méthodologie globale employée pour répondre à notre objectif de recherche est présentée en Figure 3.1. Chaque article (chapitres 3 à 6) suit globalement cette méthodologie (avec plus de détails dans les chapitres suivants), tout comme ce manuscrit, dont les premières étapes ont déjà été réalisées dans les deux premiers chapitres.

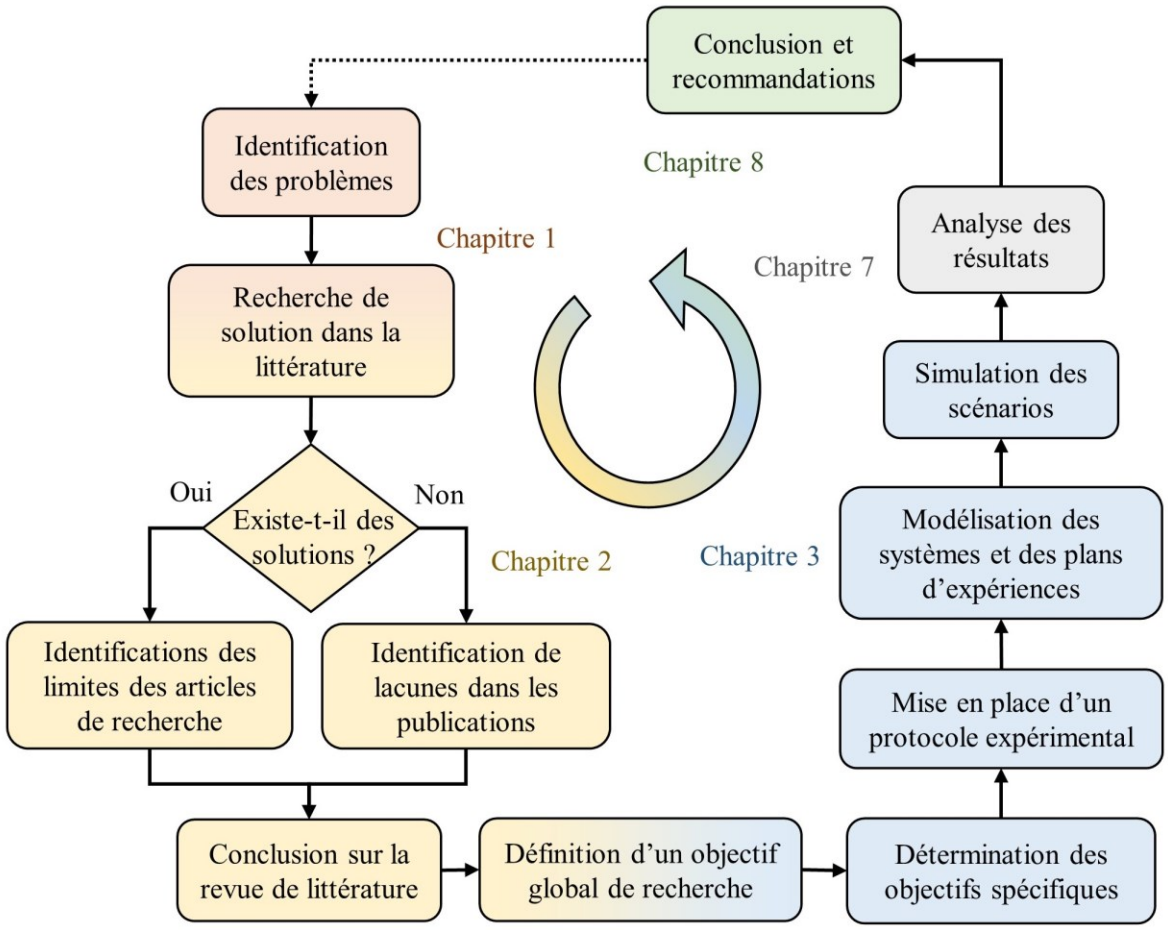


Figure 3.1 : Schéma de la méthodologie globale

Tout commence par l'identification de certains problèmes et de questionnements, comme le paramétrage du DLT dans un atelier géré en DDMRP par exemple : Faut-il maîtriser les temps de réponses ou ajuster le DLT lors d'une variation anticipée ou observée des temps de réponse dans un atelier piloté en DDMRP ?

On cherche par la suite des solutions dans la littérature en s'appuyant sur les publications scientifiques disponibles. Le but étant de ne pas se lancer dans les objectifs de recherche, la modélisation et l'analyse s'il existe déjà des solutions ou des recherches sur les sujets abordés. Il s'agit également d'identifier les limites dans les articles de recherche, les recommandations, ou encore les sujets non traités. Par exemple, dans notre recherche de maîtrise des temps de réponse, nous avons vu qu'il existe la méthode ConWIP. Bien qu'elle soit prometteuse, nous avons vu qu'elle permet de maîtriser les temps de cycle et non les temps de réponse, et nous n'avons trouvé aucune publication sur le couplage de la méthode ConWIP avec la méthode DDMRP, encore moins sur l'intégration du ConWIP dans la génération des ordres de fabrication. On peut alors réaliser une conclusion sur la revue de la littérature, pour déterminer les objectifs de recherche.

Toutes ces étapes ont été présentées dans les deux premiers chapitres de ce manuscrit, et nous avons identifié trois objectifs spécifiques à accomplir, présentés dans la partie suivante.

Ensuite, on met en place un protocole expérimental nous permettant de répondre à nos objectifs de recherche. Il s'agit de déterminer le périmètre d'étude ainsi que la manière utilisée pour répondre à notre problématique. Dans notre cas, nous nous intéressons aux lignes de production constituées de plusieurs machines en série, pilotées par la méthode DDMRP. Ce type d'atelier a été choisi puisqu'il s'agit de l'atelier le plus commun dans l'industrie, en particulier entre deux stocks tampons DDMRP. Ainsi, même si une nomenclature présente des stocks tampons intermédiaires (pour les composants par exemple), on risque fortement de retrouver des lignes de production sur le chemin critique entre deux stocks tampons. Dans ce cas, le DLT correspond donc au délai alloué pour traverser l'atelier. Pour évaluer nos propositions, nous avons choisi la simulation à évènements discrets, puisqu'elle permet facilement de modéliser des environnements complexes ayant plusieurs sources de variabilité (Mourtzis, 2020).

Les étapes suivantes sont donc la modélisation des différents systèmes étudiés. Dans le chapitre 4 par exemple, on étudie et modélise une ligne de production constituée de huit machines en série fabriquant neuf produits finis différents, dans le chapitre 5 il s'agit d'une ligne de production

constituée de six machines en série, ainsi qu'un cas industriel où les produits semi-finis sont pilotés par un stock tampons, et enfin dans le chapitre 6 on modélise une ligne de production composée de trois machines en série. Les éléments principaux à modéliser sont détaillés dans la partie 3.3, suivis des indicateurs généralement utilisés pour analyser les résultats (il est important de ne pas en oublier pour éviter de tout devoir recommencer…). Dans cette étape, on définit également un plan d'expérience qui va servir à valider ou non nos hypothèses de départ. Dans l'article du chapitre 6 par exemple, nous faisons l'hypothèse que notre modèle DDMRP-ConWIP Intégré est plus robuste que le modèle DDMRP classique ou le modèle (*R*, *Q*), et ce pour n'importe quelle variabilité de la demande et de la production. Nous testons donc les trois méthodes de réapprovisionnement, subissant trois signaux de demande différents, ainsi que trois niveaux de panne machine différents dans l'atelier. On obtient alors un plan d'expérience composé de 3x3x3 = 27 scénarios, illustré en Figure 3.2.

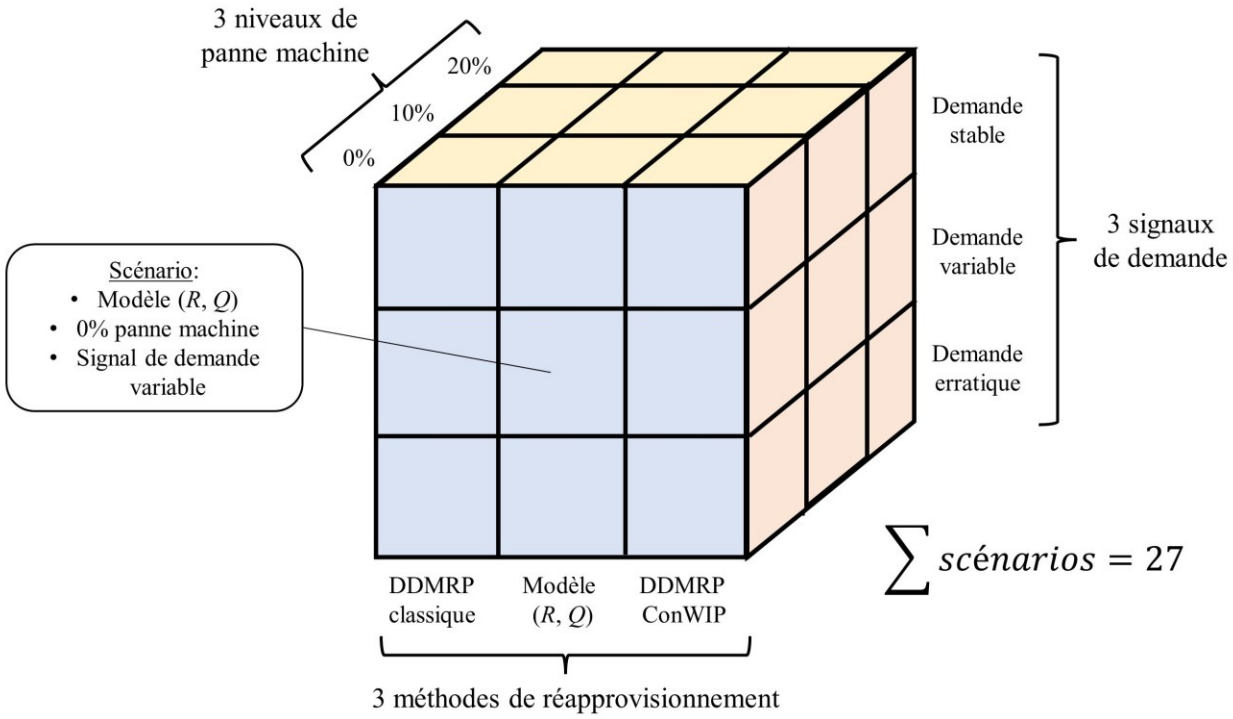


Figure 3.2 : Exemple d'un plan d'expérience à trois dimensions pour un total de 27 scénarios

Il faut enfin déterminer la longueur des simulations et le nombre de réplications. Le modèle du chapitre 6 dure deux ans par exemple, et en général une centaine de réplications est réalisée pour s'assurer d'avoir des résultats représentatifs (on veut éviter de tomber dans un cas particulier où une méthode est moins bonne qu'une autre alors qu'en général c'est l'inverse). Les différences

entre les réplications d'un même scénario reposent sur la génération des nombres aléatoires venant de la variabilité dans les modèles. On définit également une durée de « mise en régime permanent » durant laquelle le modèle n'enregistre pas les données pour éviter de fausser ces dernières puisque l'atelier de production est généralement vide au début d'une simulation. On s'assure ainsi d'avoir des résultats correspondant à un régime d'atelier classique.

L'étape suivante est l'analyse des résultats. En fonction des valeurs des indicateurs de performance, on peut valider ou non nos hypothèses, et déterminer quelle méthode ou politique est plus efficace qu'une autre dans telle situation. On peut également réaliser une discussion pour nuancer les résultats et mettre en évidence les limites de l'étude, ce qui correspond au chapitre 7 de ce manuscrit.

Enfin, la dernière étape consiste à conclure sur le travail réalisé et préconiser des pistes de recherche pour de futurs travaux. Tout le travail est récapitulé brièvement dans le chapitre 8, en valorisant les informations importantes, notamment le résultat global. Cette étape permet de boucler avec la première puisqu'en soulevant des zones d'ombre ou des améliorations possibles, on identifie de nouveaux problèmes à étudier.

## 3.2 Les trois objectifs spécifiques de recherche

Nous décomposons notre problématique et objectif principal en trois objectifs spécifiques de recherche distincts. Le but étant d'une part de chercher à ajuster le DLT en fonction des variations observées des temps de réponses, et d'une autre part de proposer des méthodes pour maîtriser les temps de réponse par le contrôle du taux de charge. Le contrôle du taux de charge se fait soit en ajustant la capacité (dénominateur du calcul du taux de charge, équation 7), soit en diminuant l'utilisation des ressources (numérateur, équation 7), notamment en réduisant le nombre de changement de série.  Ainsi, nous proposons les trois objectifs spécifiques de recherche suivants.

**Premier objectif spécifique de recherche :** Ajuster dynamiquement le DLT en fonction des temps de réponses observés pour l'adapter aux variations de charge. Une baisse de charge réduira le DLT et les stocks, tandis qu'une hausse augmentera le DLT pour aboutir à un taux de service élevé. La méthode proposée adaptera automatiquement le DLT.

**Deuxième objectif spécifique de recherche :** Vérifier le phénomène de la théorie des files d'attente où les temps de réponses augmentent drastiquement lorsque le taux de charge approche

les 100%, et proposer un outil visuel de gestion des temps de réponses par des leviers de capacité, destiné au responsable de production d'atelier géré en DDMRP. Cet outil servira à maîtriser les temps de réponses ou ajuster le DLT « manuellement ».

**Troisième objectif spécifique de recherche :** Créer une méthode capable d'ajuster les tailles de lots en fonction de l'état de l'atelier de production, en s'inspirant du fonctionnement de la méthode ConWIP et du système de génération des ordres du DDMRP. La méthode proposée lissera automatiquement la charge.

Le lien entre la problématique de recherche, l'objectif principal et les trois objectifs spécifiques est illustré en Figure 3.3.

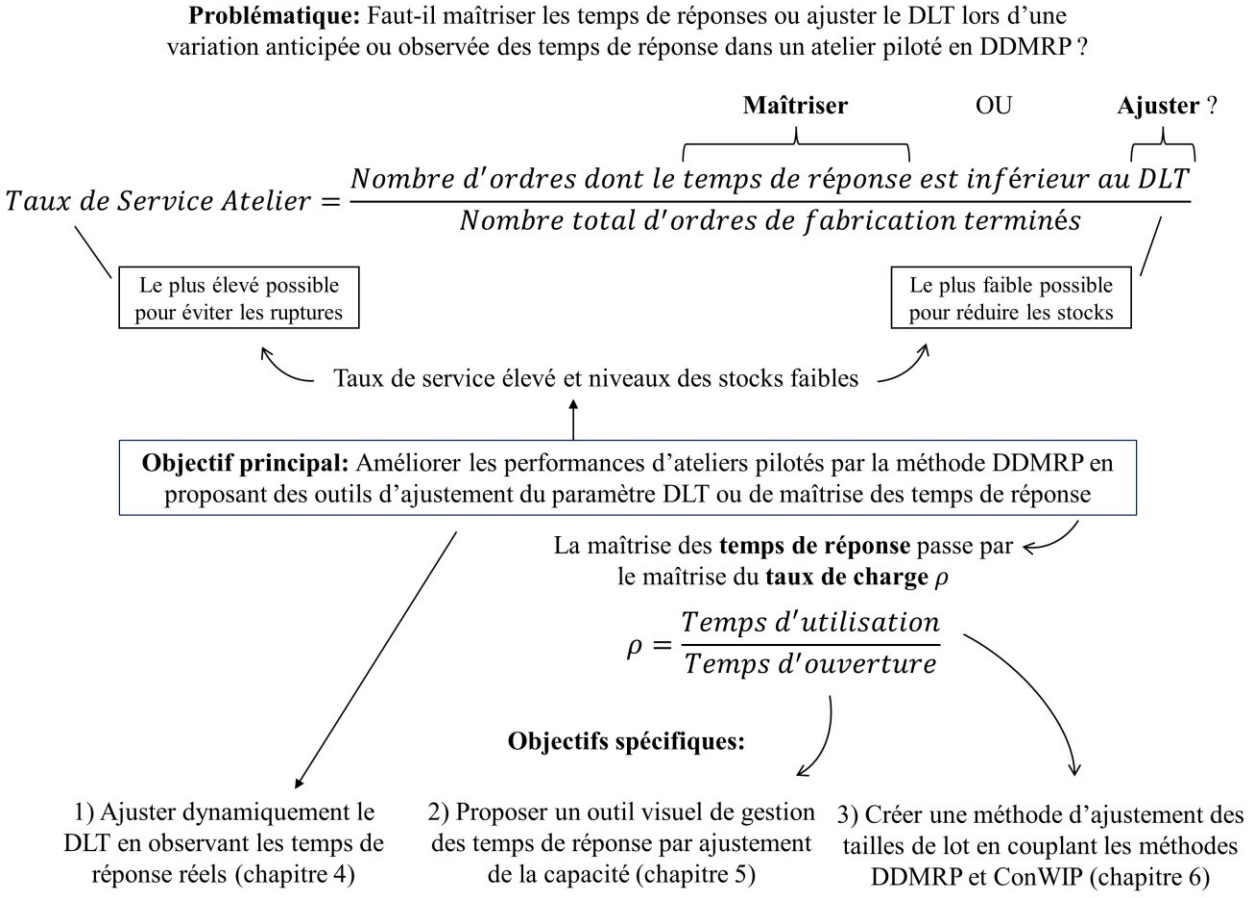


Figure 3.3 : Lien entre la problématique de recherche, l'objectif principal et les objectifs spécifiques

Les chapitres 4, 5 et 6 de ce manuscrit de thèse répondent chacun à l'un de ces objectifs.

## 3.3 Les éléments principaux des modèles de simulation

Nous avons décidé de résoudre nos objectifs de recherche en utilisant la simulation à évènements discrets. La modélisation et la simulation permettent en effet d'étudier des environnements complexes soumis à de fortes variabilités. La simulation à évènements discrets consiste à modéliser un système, qui va évoluer dans le temps en fonction d'évènements venant altérer ce système. On modélise des entités (commande, pièce, client, etc.) qui entrent dans le modèle, effectuent ou subissent des actions et services de la part de ressources (guichet, opérateur, machine, etc.), viennent modifier des variables, puis quittent le système (Babulak and Wang, 2010). Une fois que le modèle est représentatif de la réalité, on peut alors tester toutes nos hypothèses en changeant certains paramètres (comme le DLT), ressources (choix du nombre d'opérateurs) ou encore politiques de gestion.

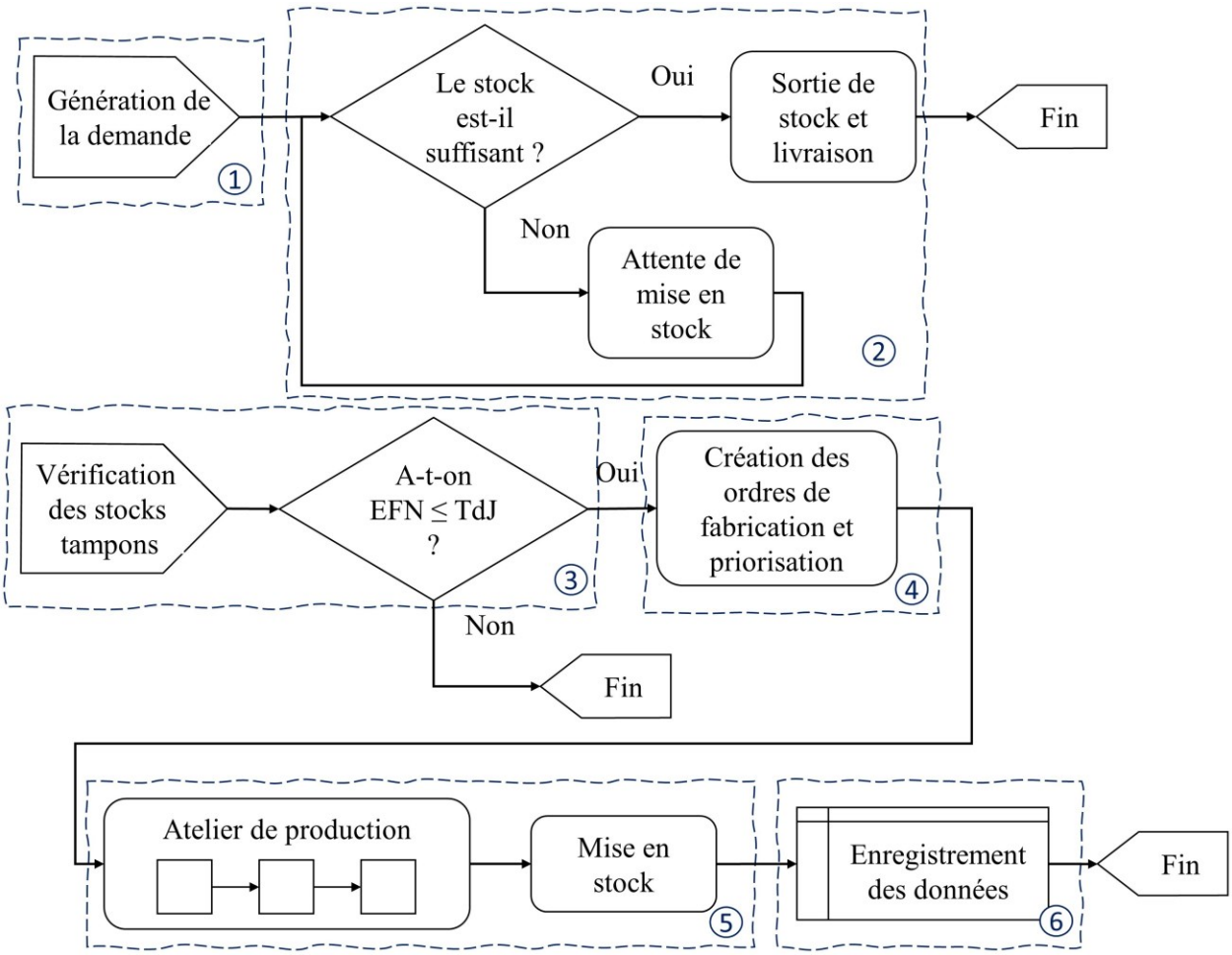


Figure 3.4 : Les principaux éléments de modélisation d'un atelier géré en DDMRP

Dans les différents modèles de simulation réalisés, on retrouve en général les mêmes éléments, présentés en Figure 3.4, à savoir :

1) Un module de **génération de la demande** : les entités « commandes client » sont créées ici. En faisant varier les quantités commandées, le temps moyen entre deux commandes, le produit commandé, ou en utilisant des données externes (sur un fichier Excel par exemple), on peut simuler différents signaux de demande. Par exemple, dans le troisième article (chapitre 6), nous avons modélisé trois signaux différents de demande, correspondant à la même demande moyenne, générant donc la même charge moyenne dans l'atelier (Figure 3.5). En modélisant des signaux représentatifs de la réalité (une demande stable pour les produits communs, des pics de commandes pour des produits saisonniers, ou encore un accroissement de la demande pour un nouveau produit), on peut alors plus facilement généraliser nos résultats.

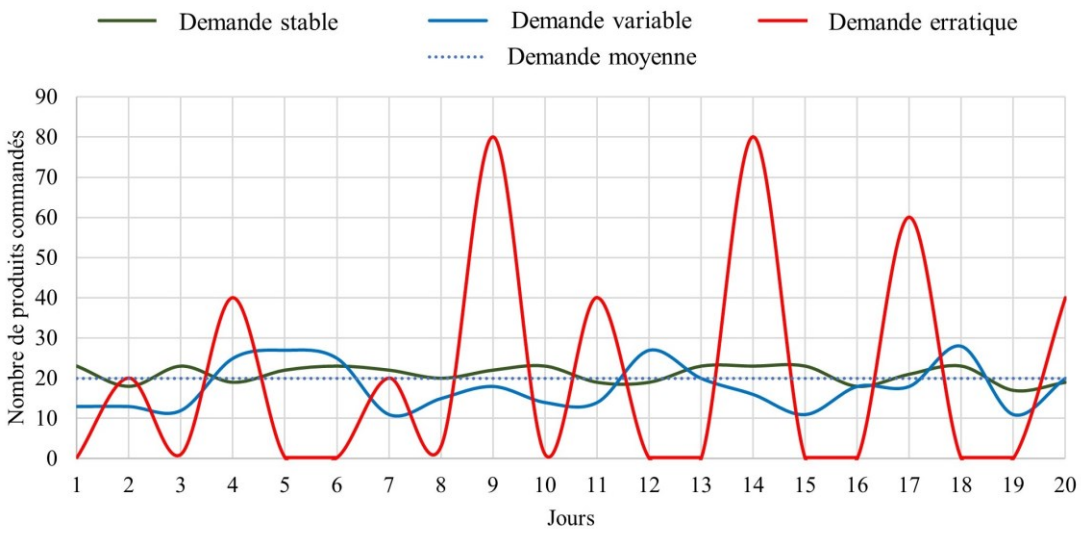


Figure 3.5 : Exemple de signaux de demande modélisés dans le chapitre 6.

2) Un module de **sortie de stock et de gestion des commandes partielles** : lorsqu'une commande client arrive, on vérifie s'il y a suffisamment de stock du produit demandé. Si c'est le cas, on effectue la sortie de stock (et la livraison si nécessaire), et l'entité sort du système. Si la quantité de stock n'est pas suffisante et que l'on gère les commandes partielles, la commande est honorée en partie, et le reste à livrer entre dans une file d'attente. Lorsque le stock sera de nouveau disponible, les commandes en retard seront prioritaires pour être livrées et sortiront du système lorsqu'elles seront complètement honorées ;

3) Un module de **vérification des stocks tampons** : il s'agit ici de modéliser le comportement d'un humain qui viendrait vérifier les stocks tampons DDMRP. On calcule la position de stock par l'équation de flux net et on la compare au Top du Jaune pour chaque produit. Si la position de stock est inférieure au seuil de lancement, un ordre de fabrication va être créé, sinon l'entité représentant la vérification du stock tampon sort du système. Elle sera regénérée le lendemain, lorsque l'humain viendra vérifier les stock tampon à nouveau ;

4) Un module de **création des ordres de fabrication** : pour chaque produit pour lequel un ordre a été créé, on détermine ici la taille de lot à produire (donnée par la différence entre le Top du Vert et la position de stock). On détermine également la priorité des ordres les uns par rapport aux autres selon l'approche DDMRP, en divisant la position de stock par le seuil de réapprovisionnement (le Top du Jaune). Ainsi, plus ce ratio est faible, plus le stock tampon a été consommé, et plus il est prioritaire (Figure 3.6).

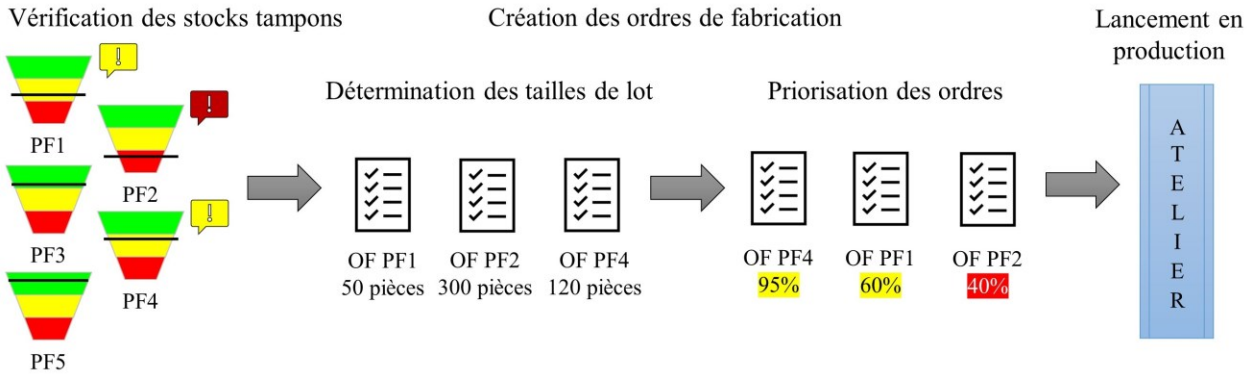


Figure 3.6 : Modélisation du processus de création des ordres (tailles de lot et priorisation).

Après cette priorisation initiale, les ordres de fabrication sont gérés en Premier-Entré-Premier-Sorti (PEPS) dans l'atelier. C'est-à-dire que l'ordonnancement reste inchangé et que les machines traitent les ordres de fabrication dans l'ordre d'arrivée. On met également les variables à jour, notamment l'en-cours de production pour chaque produit ;

5) **L'atelier de production et la mise en stock** : il s'agit ici de modéliser l'agencement des ateliers. On modélise le nombre de machines, le nombre d'opérateurs, les processus de fabrication, les temps de mise en course, les temps unitaires de production, les temps de transfert, etc. On peut introduire différents degrés de variabilité dans la production (des pannes machines suivant des lois exponentielles et des temps de production suivant des lois uniformes ou triangulaires, par exemple). Si nécessaire, on réalise également la sortie de

stock des matières premières et des composants. Lorsqu'un ordre a traversé tout l'atelier et que les produits finis sont fabriqués en quantité voulue, on procède à la mise en stock et à la mise à jour des variables du modèle (en-cours et stock de produits finis, et nombre de tickets ConWIP disponibles si besoin) ;

6) Des modules d'**enregistrement des données** : pour analyser et comparer les modèles, les méthodes ou les politiques de gestion, il est nécessaire d'enregistrer des données tout au long des simulations. Des modules spéciaux existent pour incrémenter ou décrémenter des compteurs, enregistrer des valeurs comme la taille de lot de chaque ordre par exemple, ainsi que différents temps comme les temps de réponse notamment. Cela permet, lors de la génération du rapport final de simulation, de connaître les valeurs moyennes, les écart-types, les intervalles de confiance, les valeurs minimales et maximales, etc.

On retrouve en Figure 3.7[3] ces différents éléments (numérotés de 1 à 6) sur la photo d'une modélisation d'un atelier de production géré en DDMRP, sur le logiciel Arena.

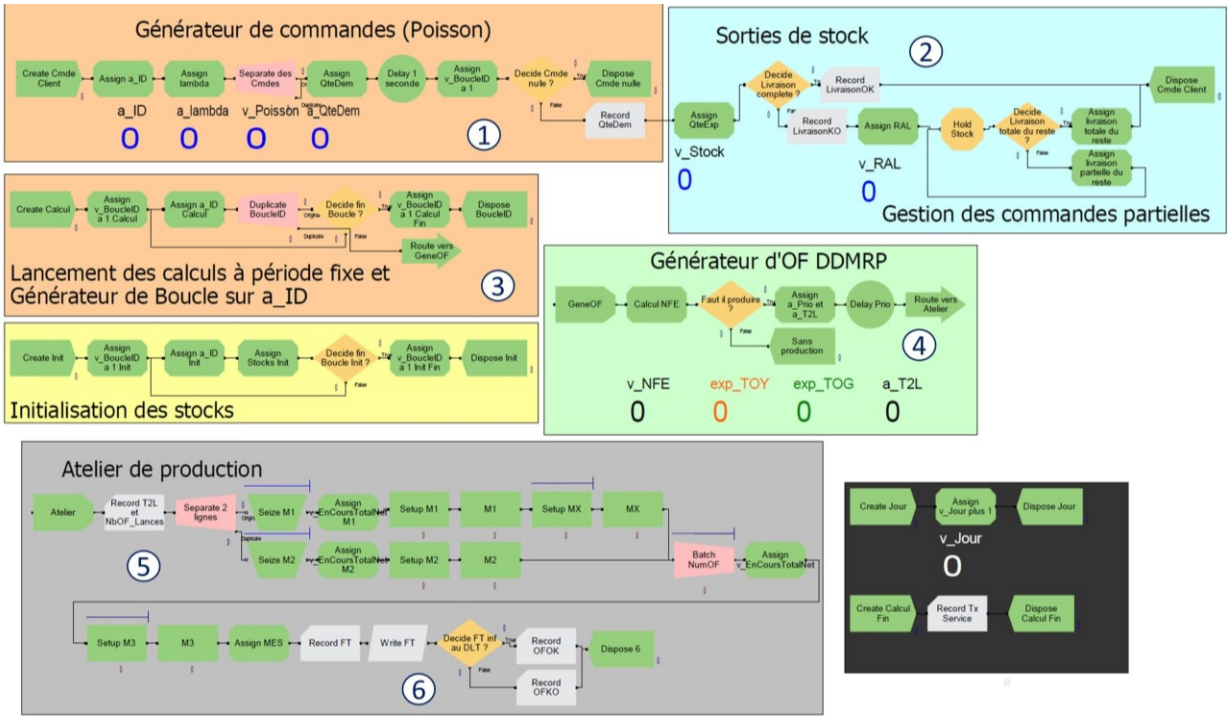


Figure 3.7 : Photo de la modélisation d'un atelier géré en DDMRP sur le logiciel Arena.

[3] Attention la Figure 3.7 est truffée d'anglicismes qui pourraient choquer certaines âmes sensibles.

## 3.4 Les données de sortie

Dans nos modèles de simulation, on utilise principalement les données de sorties suivantes pour comprendre et expliquer nos résultats :

- Le **taux de service client** (Équation 11). Il est généralement défini comme le nombre de commandes client honorées dans les temps divisé par le nombre total de commandes client. Ainsi, si le stock de produits finis est présent en quantité suffisante lorsqu'une commande doit être livrée, le numérateur et le dénominateur sont incrémentés de un, sinon, seul le dénominateur l'est (faisant chuter le taux de service).

$$Taux\ de\ service\ client = \frac{Nombre\ de\ commandes\ client\ honorées\ dans\ les\ temps}{Nombre\ total\ de\ commandes\ client} \quad (11)$$

  On peut également utiliser un **taux de service produit**, en divisant le nombre de produits finis livrés dans les temps (plutôt que le nombre de commandes) par le nombre de produits commandés, ce qui évite d'avoir un taux de service de 50% alors qu'une commande qui demandait 10 produits a été livrée et qu'une autre de 90 n'a pas été honorée (on a alors un taux de service de 10%). Un taux de service client ou produit faible peut être expliqué par une cadence trop faible, des temps de réponse trop longs, un temps alloué trop court, ou encore une prévision de demande mal établie ;

- Le **taux de service atelier**. Comme défini précédemment, il s'agit du ratio entre le nombre d'ordres de fabrication dont le temps de réponse est inférieur ou égal au délai de production et le nombre total d'ordres (Équation 1, page 10). Dans un atelier géré en DDMRP, on remplace le délai de production par le DLT.

  Un taux de service atelier faible est généralement précurseur d'un taux de service client en chute. Il est donc important de le surveiller et de comprendre pourquoi les temps de réponse sont plus longs que le temps alloué pour éviter des ruptures ;

- Les **temps de réponse**. Défini comme le temps s'écoulant entre la reconnaissance d'un besoin et la mise en stock des produits répondant à ce besoin, cet indicateur permet notamment de calculer le taux de service atelier. On cherche à calculer la moyenne, l'écart-type, voire la distribution des temps de réponse pour paramétrer le DLT. De plus, la décomposition d'un temps de réponse en différents temps (attente, production, transfert,

etc.) permet de comprendre pourquoi un temps de réponse est plus long que prévu, et d'où cela vient ;

- Le **taux de charge des ressources**. Noté $\rho$ en général, il s'agit du temps total d'utilisation d'une ressource sur le temps total d'ouverture. Cet indicateur nous permet d'expliquer pourquoi les temps de réponse sont plus longs que prévu si le taux de charge est élevé. Ceci indique une mauvaise gestion capacitaire des ressources de l'atelier ;
- Le **stock d'en-cours**. Il est défini comme le nombre moyen de pièces ou de lots en cours de fabrication (par exemple un lot de 50 pièces en cours de fabrication) ;
- Le **stock de produit finis**, exprimé en nombre moyen de produits finis stockés dans le temps ; et
- La **cadence de l'atelier**. Elle est définie comme le nombre de pièces fabriquées (ou le nombre de produits conditionnés) par heure ou par jour. On utilise surtout cet indicateur pour vérifier que la cadence est maximale dans une boucle ConWIP, puisqu'elle dépend du nombre de tickets. Si la cadence est plus faible que la normale, c'est parce que la ressource goulot s'est désarmée. Il faut alors rajouter des tickets dans la boucle.

## 3.5 Organisation générale du manuscrit et cohérence entre les objectifs spécifiques de recherche et les articles

Notre but, fil rouge de cette thèse, est d'arriver à maîtriser au mieux les temps de réponse ou d'ajuster le DLT en conséquence pour avoir un taux de service élevé et des stocks les plus faibles possible, pour nous aider à répondre à notre problématique. Le lien entre la problématique de recherche, l'objectif principal et les trois objectifs spécifiques a été illustré en Figure 3.3 précédemment.

À la suite de notre revue de la littérature, nous avions décomposé notre problématique en trois objectifs de recherche (page 41), qui sont, pour rappel :

**Premier objectif spécifique de recherche :** Ajuster dynamiquement le DLT en fonction des temps de réponses observés pour l'adapter aux variations de charge. Une baisse de charge réduira le DLT et les stocks, tandis qu'une hausse augmentera le DLT pour aboutir à un taux de service élevé. La méthode proposée adaptera automatiquement le DLT.

**Deuxième objectif spécifique de recherche :** Vérifier le phénomène de la théorie des files d'attente où les temps de réponses augmentent drastiquement lorsque le taux de charge approche les 100%, et proposer un outil visuel de gestion des temps de réponses par des leviers de capacité, destiné au responsable de production d'atelier géré en DDMRP. Cet outil servira à maîtriser les temps de réponses ou ajuster le DLT « manuellement ».

**Troisième objectif spécifique de recherche :** Créer une méthode capable d'ajuster les tailles de lots en fonction de l'état de l'atelier de production, en s'inspirant du fonctionnement de la méthode ConWIP et du système de génération des ordres du DDMRP. La méthode proposée lissera automatiquement la charge.

Ces trois objectifs ont été le sujet d'un article de conférence et de deux articles de journaux.

Le premier article, intitulé *Decoupled Lead Time in finite capacity flowshop: a feedback loop approach* a été publié dans le compte rendu de la 8e édition de la conférence *International Conference on Industrial Engineering and Systems Management* en 2019 à Shanghai. Cet article a été écrit en collaboration avec Guillaume Martin de l'IMT Mines Albi. Il traite de l'ajustement dynamique du paramètre DLT en utilisant une boucle de régulation hebdomadaire prenant en compte les derniers temps de réponse observés ainsi que l'historique des semaines précédentes. Cet ajustement permet d'adapter le DLT à la charge de l'atelier pour optimiser le dimensionnement des stocks. Lors de la conférence, l'article a reçu le prix du meilleur article étudiant.

Le deuxième article, intitulé *Visual Charts Produced by Simulation to Correlate Service Rate, Resource Utilization and DDMRP Parameters*, a été soumis et accepté par le journal *International Journal of Production Research*. Dans cet article, nous nous attaquons au deuxième objectif, en corrélant la distribution des temps de réponse, le taux de charge des ressources et le paramètre DLT. Cette corrélation permet de mieux comprendre comment se comportent les temps de réponse lorsque le taux de charge varie, et de recommander le choix d'un DLT pour un taux de charge donné. On a également vu qu'il existe de nombreux leviers capacitaires (à court, moyen ou long termes), l'idée est donc de relier ces leviers aux temps de réponse et au paramètre DLT. Ainsi, pour une charge donnée (prévue), on peut décider quel levier capacitaire utiliser pour contrôler les temps de réponse. L'utilisation de la simulation permet de créer des abaques représentant différentes situations, pour aider les décideurs à choisir la meilleure option capacitaire en fonction de la situation.

Le troisième article, intitulé *Improvement of the DDMRP production system by coupling reorder point and ConWIP loop*, a été soumis au journal *International Journal of Production Research*. Il s'agit ici de rendre plus résilient un atelier de production piloté en DDMRP. On a vu qu'il existe plusieurs sources de variabilité dans le monde industriel, notamment dans la demande et la production. Les méthodes de gestion et de planification prennent de plus en plus en compte la variation de la demande et les chercheurs proposent de nouvelles solutions (comme des heuristiques) pour déterminer les paramètres optimaux d'une méthode donnée dans une situation donnée. Ces méthodes considèrent rarement la variabilité de la production, et ne font pas le lien entre temps de réponse et taille de lot. Or, on a vu que les temps de réponse sont liés au taux de charge, et que le taux de charge est lié à la taille des lots de fabrication. De plus, une boucle ConWIP permet de limiter l'en-cours de production et ainsi de contrôler les temps de cycle dans cette boucle. Notre objectif est donc de créer un modèle qui ajuste automatiquement la taille de lot des ordres de fabrication en fonction de la demande et de l'état de l'atelier, en utilisant une boucle ConWIP. Ce modèle, prenant en compte la variabilité de la demande et de la production, est plus résistant aux évènements imprévus, donc plus résilient.

La suite du manuscrit est composée des trois articles pour les trois prochains chapitres, suivis par une discussion générale sur les méthodes employées, les résultats et les limites. Enfin, le chapitre 8 conclut et préconise des recommandations.

# CHAPITRE 4 ARTICLE 1 : DECOUPLED LEAD TIME IN FINITE CAPACITY FLOWSHOP: A FEEDBACK LOOP APPROACH

Cet article a été publié et présenté à la conférence *International Conference on Industrial Engineering and Systems Management* en 2019 à Shanghai. La version présentée dans cette thèse est identique à la version publiée.

Les auteurs de cet article sont : Guillaume Dessevre[1], Guillaume Martin[2], Pierre Baptiste[1], Jacques Lamothe[2], Robert Pellerin[1], et Matthieu Lauras[2].

[1]Département de Mathématiques et Génie Industriel, Polytechnique Montréal, Montréal, Canada.
[2]Centre Génie Industriel, IMT Mines Albi, Albi, France.

Dans ce travail de recherche, Guillaume Martin s'est principalement chargé de la génération de la demande client, notamment pour s'assurer que l'atelier n'était ni surchargé ni surcapacitaire.

**Note au lecteur**: lors de la rédaction de cet article, datant du début de l'année 2019, nous commettions encore l'erreur de confondre délai de production et temps de réponse. À l'époque, nous utilisions le terme *lead time* en anglais pour parler ce que nous appelons *flow time* aujourd'hui. Les modifications ont été apportées dans ce manuscrit pour uniformiser les définitions, c'est pourquoi la version présentée est légèrement différente de la version publiée. Ainsi, les temps de réponses correspondent aux *flow times* et les délais de production aux *lead times*.

## Résumé

Récemment étudié, le *Demand Driven Material Requirements Planning* (DDMRP) semble offrir de bonnes performances. Le paramétrage de ses stocks tampons est réalisé par un ajustement dynamique prenant en compte l'évolution de la demande, mais non l'état de l'atelier. Plusieurs articles traitent de la variabilité des temps de réponse et considèrent leurs variations dans le paramétrage des méthodes traditionnelles de gestion de la production tels que le *Material Requirements Planning* et les méthodes à flux tiré. Cet article propose un ajustement dynamique du *Decoupled Lead Time* de la méthode DDMRP, en considérant la variabilité des temps de réponse. En l'appliquant à une ligne de production modifiée où la ressource goulot est représentée par les opérateurs, les résultats montrent que l'ajustement dynamique des zones des stocks tampons réduit les stocks, tout en assurant un taux de service décent. Nous constatons également que cela augmente le taux de charge, jusqu'à une certaine limite à déterminer, qu'il peut être préférable de ne pas franchir, sinon les temps de réponse augmentent drastiquement.



## Abstract

Recently studied, Demand Driven Material Requirement Planning (DDMRP) seems to offer good performances. Its buffer parametrization is done by a dynamic adjustment considering the demand evolution, but not the workshop state. Several papers deal with flow time variability and take into account its variations in the parametrization of traditional material management methods such as Material Requirements Planning and pull flow systems. This paper considers flow time variability and proposes a dynamical adjustment of the decoupled lead time of the DDMRP. When applying it to a modified flow shop with a competence bottleneck, results show that the dynamic adjustment of the buffer sizes reduces stocks, while assuring a decent service level. We also find that it increases workload, to an extent where it can be better not to cross a certain limit that has to be determined, otherwise flow times increase drastically.


Keywords: DDMRP, decoupled lead time, dynamic adjustment, finite capacity, discrete event simulation

## 4.1 Introduction

The Demand-Driven Adaptive Enterprise (DDAE) framework claims to obtain better results than previous methods (such as MRP (Orlicky, 1975)) in dealing with uncertainty in the supply chain (Ptak and Smith, 2018). It does so by promoting flow of goods, information and money through the whole organization and by using agile response to the different sources of variations. Variations in the Demand-Driven environment may come from clients, suppliers, the processes themselves or even from management policies (Ptak and Smith, 2016).

At the core of the framework mentioned above, is the operating model (OM) of the supply chain. The model itself is a collection of objects: machines with capacities, physical buffer stocks, human resources, bill of materials... The model also uses a defined set of parameters for each buffered reference, including (but not limited to): Average Daily Usage (ADU), Decoupled Lead Time (DLT), Lead Time Factor (a measure of uncertainty in the lead times) and Variability Factor (another measure of uncertainty but for variation over the whole process) (Ptak and Smith, 2016).

Previous works have shown how a flow shop can be efficiently controlled by updating only the ADU on a regular basis and with a specific formula based on real demand (Miclo et al., 2018). But this is only one side of the problem: is it possible to control the DLT of a reference based on effective measures of flow times? By applying queuing theory (Shortle et al., 2018) and Little's Law (Little, 1961) to the same system, one could argue that finite capacities could bring about longer idle times. This would impact flow times but to what extent would it change DLT, which is said to be based on lead times (Ptak and Smith, 2016)?

This paper explores the effects of putting lead times under control in the case of a finite capacity flow shop. Both lead time definitions and lead time control methods are considered. The existing literature is reviewed about lead time control at the floor level in general and specifically in a DDMRP environment. Then, the definition of the flow shop simulation model and the associated design of experiments for lead time definitions and control methods are exposed. The paper concludes on the discussion of the results and their application to real word use cases.

## 4.2 Literature Review

This literature review is decomposed as follows: first section focuses on DDMRP in the literature and the use of lead time within the method, the second section is about MRP parametrization and

planned lead time, and finally the third section is about lead time in material management methods using pull flow concepts.

### 4.2.1 DDMRP and decoupled lead time

The management principles of DDMRP were introduced in the third edition of Orlicky's Material Requirements Planning (Ptak and Smith, 2011). Since 2011, the Demand Driven MRP has been the subject of several research studies.

The method presents a new concept, the Decoupled Lead Time (DLT) measured between physical buffers. It is defined as "a qualified cumulative lead time defined as the longest unprotected or unbuffered sequence in a bill of material" (Ptak and Smith, 2016). Figure 4.1 is a representation that differentiates lead time, cumulative lead time and DLT.

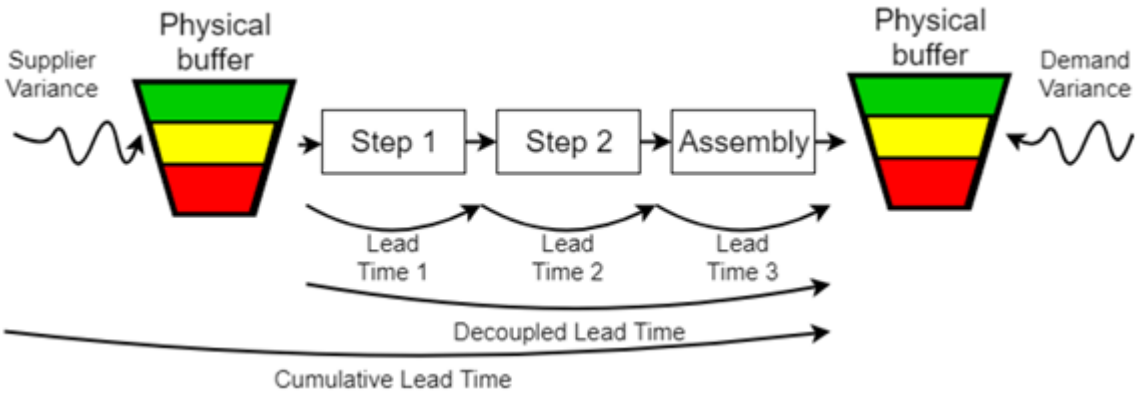


Figure 4.1 : Difference between (manufacturing) lead time, decoupled lead time and cumulative lead time.

The Demand Driven MRP was widely compared with other material management methods such as MRP and Kanban, in particular by Miclo et al. (2016b). Using Discrete Event Simulation (DES), Miclo et al. (2015) started to compare DDMRP with MRP, then MRPII and finally with the method Kanban (Miclo et al., 2016a; Miclo et al., 2018). Shofa and Widyarto (2017) also compared DDMRP and MRP with a DES model in a more complex case, and Ihme and Stratton (2015) used DDMRP to improve a printing ink manufacturing company in terms of stability and product availability.

Recently, Martin et al. (2018) proposed a process map for the demand driven adaptive enterprise model in order to draw an explicit cartography used in deployment, and Baptiste (2018) showed some scheduling opportunities with complex bills of material in a Demand Driven MRP system.

If several research studies dealing with DDMRP were published, none of them considered a variation of the DLT.

### 4.2.2 MRP parametrization with planned lead times

Among the different MRP parameters, there is the planned lead time: difference between the due date and the release date. Across the field, there are nowadays two ways to deal with planned lead times: trying to find the optimal ones or taking into account the variations in the parametrization.

To determine optimal planned lead times, Yano (1987) proposed a solution in a stochastic environment where the goal is to minimize inventory and backlog, Weng (1996) presented a model in which the workshop has stochastic processing times and the demand is lead time-related.

Axsäter (2006) seeks for the optimal release dates in order to minimize inventory costs, and Buzacott and Shanthikumar (1994) worked on lead times and safety stock, concluding on using safety lead times instead of safety stock when forecasting are reliable.

However, the dynamic MRP seems to offer better results than the static one (Ioannou and Dimitriou, 2012). That is why some researchers propose different methods to adapt MRP parameters according to the system state: Dolgui and Louly (2008) developed a Branch & Cut algorithm where the customer demand is known and steady, but the lead times of the components are variable; in another paper, they presented an algorithm to minimize set-up costs, waiting costs and backlog costs (Louly and Dolgui, 2011). They created a decision tool to find optimal MRP parameters while there is supplier variability.

Ioannou and Dimitriou (2012) proposed iterative algorithms to substitute the lead time in the MRP parametrization by taking into account waiting times to get a better estimation of the lead time. Ammar et al. (2016) proposed a method to reduce backlog and inventory when supplier lead times are variables but known in an interval.

### 4.2.3 Pull flow methods parametrization

Parametrization in pull flow systems was also studied considering lead time variations. Rees et al. (1987) used lead time estimations to adjust the number of kanbans; Gupta and Al-Turki (1997) proposed an algorithm to change dynamically the number of kanbans in a stochastic environment (processing times and customer demand) and show its efficiency; Tardif and Maaseidvaag (2001) considered inventory level and customer demand to decide the number of Kanban tickets to add or to remove from the loop. Their system offers better results than a traditional static Kanban system.

Marek et al. (2001) proposed a heuristic to adjust the number of ConWIP tickets: reducing it until the objective can no more be met, then add one ticket. Belisário et al. (2015) used a genetic algorithm to change the number of ConWIP tickets according to circumstances. Takahashi and Nakamura (2002) compared Kanban, reactive Kanban, ConWIP and reactive ConWIP to conclude that reactive Kanban seems to offer better results than the others.

Pull flow methods can also be integrated into a push system: Selçuk (2013) studied the adjustment of the supplier lead times in a Kanban-MRP system.

To conclude, in the traditional material management methods such as MRP, Kanban or ConWIP, adapting parameters to deal with lead time variations is proved to be an efficient regulation. But no one published a research about DLT control in DDMRP.

## 4.3 Methods

The literature review shows that: (i) previous works on the DDAE methodology have not considered a variation of the DLT at the operational level and, (ii) that varying lead time parametrization has given good results, both on pushed and pulled flow studies. Taking these facts into account, we propose to answer the following questions:

(i) Do DLT calculation rules have an effect on system performance?

(ii) Does DLT type of control loop also has an effect on performance?

(iii) Does the system behavior depend on the type of perturbation?

These questions will be answered by running simulations on a modified flow shop environment (Johnson, 1954). The flowshop itself is a version in which products may skip machines according to their production routines. The physical buffers are at the input and output of the flowshop.

Production routines for the flowshop are given in the matrix below (Tableau 4.1). Each machine requires an operator during all the production process (run and setup). These multi-skilled operators are the bottleneck of the line.

Tableau 4.1 : Production routines (machine-product associations).

| | **FP1** | **FP2** | **FP3** | **FP4** | **FP5** | **FP6** | **FP7** | **FP8** | **FP9** |
|---|---|---|---|---|---|---|---|---|---|
| **M1** | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| **M2** | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| **M3** | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| **M4** | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| **M5** | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| **M6** | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| **M7** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **M8** | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |

Both the simulation and the flowshop itself run according to the following hypotheses:

- Daily total demand trend is constant but is applied an uncertainty factor given by a normal distribution of parameters (mu = 1, sigma = 0.05).
- Daily total demand is well under the flow shop's machines total capacity, to create a competence bottleneck and not a machine bottleneck.
- Production times for each part follow a triangular distribution. Changeover times are fixed for each machine, and the moving of production lots is immediate.
- No minimum order quantity is given, orders are placed following the DDMRP method (Ptak and Smith, 2016).
- Production orders are scheduled with a FIFO rule.
- The DLT control loop principle is: Anytime an order goes out of the line, its effective DLT can be measured, filtered, the DLT parameter can be changed and the buffer sizes are adapted each week (see regulation diagram in Figure 4.2).

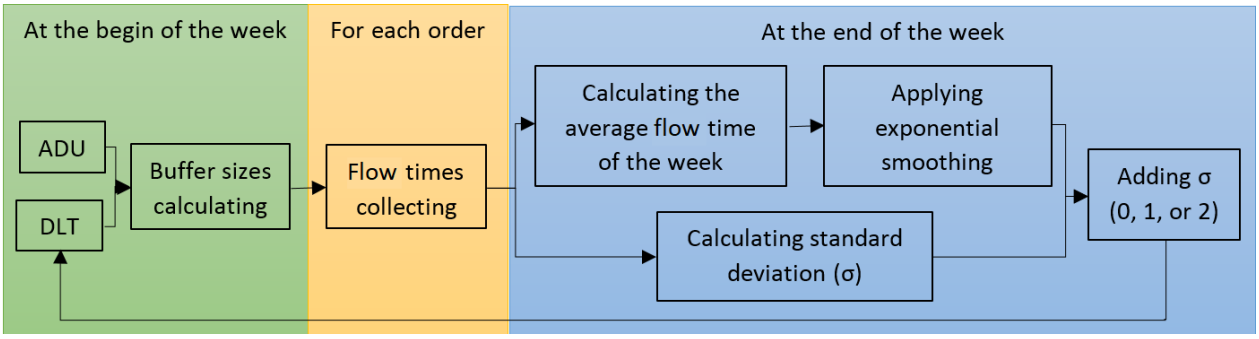


Figure 4.2 : Regulation diagram used in our model.

- In order to ensure statistical reliability, each scenario will be run 50 times on Arena version 15.10.

In order to answer the three questions raised before, we propose the following design of experiments (for a total of 28 scenarios):

(i) DLT parameter will be either fixed, or controlled according to an exponential smoothing method each week (with two smoothing factors: 5% and 10%. We tested with higher smoothing factors, but it was more chaotic).

(ii) DLT setpoint will be fixed to representing the average of flow times (0 sigma), 68% of measured flow times (1 sigma) or 95% of flow times (2 sigma).

(iii) Demand will be alternatively submitted to increasing and decreasing steps every six months. If the first step is an increase, scenarios are denoted +, otherwise it is a decrease (denoted -).

Each scenario performance will be measured according to the same criteria: a weighted function of service level and work-in-progress (WIP). The function is defined as follows:

$$\frac{1}{2} \times (1 - Service\ level) + \frac{1}{2} \times \left(\frac{WIP - Min(WIP)}{Max(WIP)}\right)$$

Maximum and minimum levels of WIP are calculated at the end of each simulation run. Service level is weighted according to the mix of products in the daily demand to take into account each product proportion. Tableau 4.2 below gives the exact proportion for our experiments.

Tableau 4.2 : Product mix in total demand.

| **Product** | FP1 | FP2 | FP3 | FP4 | FP5 | FP6 | FP7 | FP8 | FP9 |
|---|---|---|---|---|---|---|---|---|---|
| **Part of daily total demand** | 6% | 5% | 19% | 16% | 5% | 4% | 5% | 19% | 21% |

Definitions for service level and work-in-progress are taken from the APICS Dictionary (Blackstone, 2013). For more details on how to size the buffers, release production orders and move them, please refer to Ptak and Smith (2011) and Ptak and Smith (2016).

## 4.4 Results

Searching for the best scenarios is done regarding the lowest value of our weighted function, maximizing service rate (minimize 1 – service rate) while minimizing inventory.

Tableaux 4.3 and 4.4 present the results.

Tableau 4.3 : Simulation results with increasing demand.

| **ID** | **Number of operators** | **Filter smoothing factor** | **Setpoint Number of σ** | **Operator utilization** | **Standard deviation** | **Service level** | **Standard deviation** | **Stock + WIP** | **Standard deviation** | **Objective** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1+ | 4 | 0 | Ø | 91% | 22% | 97,6% | 1,0% | 1378,8 | 146,2 | 15,4% |
| 2+ | 5 | 0 | Ø | 73% | 29% | 100,0% | 0,0% | 1375,9 | 144,1 | 14,1% |
| 3+ | 4 | 5% | Ø | 89% | 21% | 90,7% | 5,2% | 1365,2 | 443,8 | 18,5% |
| 4+ | 4 | 5% | + σ | 87% | 22% | 93,6% | 4,1% | 1504,9 | 466,4 | 20,1% |
| 5+ | 4 | 5% | + 2*σ | 84% | 23% | 95,0% | 3,3% | 1779,5 | 512,4 | 25,5% |
| 6+ | 4 | 10% | Ø | 88% | 21% | 91,3% | 4,2% | 1517,7 | 665,6 | 21,6% |
| 7+ | 4 | 10% | + σ | 85% | 22% | 94,7% | 3,5% | 1703,1 | 766,3 | 23,9% |
| 8+ | 4 | 10% | + 2*σ | 82% | 23% | 95,4% | 3,1% | 2275,0 | 945,6 | 36,2% |
| **9+** | **5** | **5%** | **Ø** | **78%** | **31%** | **99,9%** | **0,1%** | **772,8** | **82,4** | **0,9%** |
| 10+ | 5 | 5% | + σ | 78% | 32% | 100,0% | 0,1% | 838,4 | 87,2 | 2,3% |
| 11+ | 5 | 5% | + 2*σ | 78% | 32% | 100,0% | 0,1% | 915,4 | 95,7 | 4,0% |
| **12+** | **5** | **10%** | **Ø** | **81%** | **29%** | **98,6%** | **0,7%** | **734,8** | **97,1** | **0,7%** |
| 13+ | 5 | 10% | + σ | 78% | 31% | 99,9% | 0,1% | 803,4 | 109,0 | 1,6% |
| 14+ | 5 | 10% | + 2*σ | 78% | 32% | 99,9% | 0,3% | 855,6 | 115,2 | 2,7% |

Scenarios with best performances (from 9± to 14±) are:

- Having 5 operators: those with only 4 have not enough capacity when there is a customer demand increase.
- Having a dynamic adjustment, with an exponential smoothing filter of 5% or 10%: it reduces inventory.
- Not considering the standard deviations (1σ and 2σ ): In the best scenario, these standard deviations behave as safety lead times. They slightly increase service rate but mainly inventory. Conversely, with 4 operators these safety lead times enable to smooth capacity limits.

Tableau 4.4 : Simulation results with decreasing demand.

| ID | Number of operators | Filter smoothing factor | Setpoint Number of σ | Operator utilization | Standard deviation | Service level | Standard deviation | Stock + WIP | Standard deviation | Objective |
|---|---|---|---|---|---|---|---|---|---|---|
| 1- | 4 | 0 | Ø | 94% | 13% | 94,6% | 1,3% | 1447,87 | 144,07 | 22,3% |
| 2- | 5 | 0 | Ø | 75% | 27% | 100,0% | 0,0% | 1443,42 | 140,76 | 19,5% |
| 3- | 4 | 5% | Ø | 91% | 18% | 95,8% | 2,1% | 1448,67 | 285,76 | 21,7% |
| 4- | 4 | 5% | + σ | 92% | 18% | 98,9% | 0,6% | 1405,10 | 206,55 | 18,9% |
| 5- | 4 | 5% | + 2*σ | 91% | 18% | 99,8% | 0,1% | 1436,05 | 178,23 | 19,4% |
| 6- | 4 | 10% | Ø | 90% | 18% | 94,6% | 2,3% | 1630,04 | 485,46 | 27,9% |
| 7- | 4 | 10% | + σ | 90% | 19% | 98,9% | 0,6% | 1494,05 | 294,86 | 21,6% |
| 8- | 4 | 10% | + 2*σ | 89% | 21% | 99,9% | 0,1% | 1499,49 | 223,43 | 21,3% |
| 9- | 5 | 5% | Ø | 80% | 28% | 100,0% | 0,1% | 850,71 | 143,99 | 1,3% |
| 10- | 5 | 5% | + σ | 80% | 28% | 100,0% | 0,1% | 950,49 | 134,44 | 4,4% |
| 11- | 5 | 5% | + 2*σ | 80% | 28% | 100,0% | 0,0% | 1053,71 | 124,66 | 7,5% |
| **12-** | **5** | **10%** | **Ø** | **82%** | **26%** | **99,7%** | **0,2%** | **808,09** | **114,15** | **0,2%** |
| 13- | 5 | 10% | + σ | 80% | 27% | 100,0% | 0,1% | 882,57 | 111,57 | 2,3% |
| 14- | 5 | 10% | + 2*σ | 80% | 28% | 100,0% | 0,1% | 968,00 | 116,94 | 4,9% |

The 12th scenario is the best: 5 operators, exponential smoothing with a factor of 10% and no standard deviation. It reduces inventory by 47% compared to the scenario without smoothing (2+) by ensuring a service rate of 98.8%.

Figure 4.3 shows an example of the evolution of Flow Time (FT) and decoupled lead time of products 3 and 6 in one simulation of scenario 12+. In addition, Figure 4.4 shows the evolution of buffer size of product 6 during the same simulation. While flow times follow the demand curve (rectangular), decoupled lead times follow the exponential smoothing (so does the buffer size).

It is an example of the dynamic adjustment of the buffer size considering the evolution of DLT and ADU. Colors represent the different zones of the buffer (green, yellow and red), and the black curve represents the stock of finished product number 6 (FP6). Every six months, the average ADU changes, then does progressively the DLT parameter then does also each buffer zone increases, then finally the measured DLT converges to parameter one.

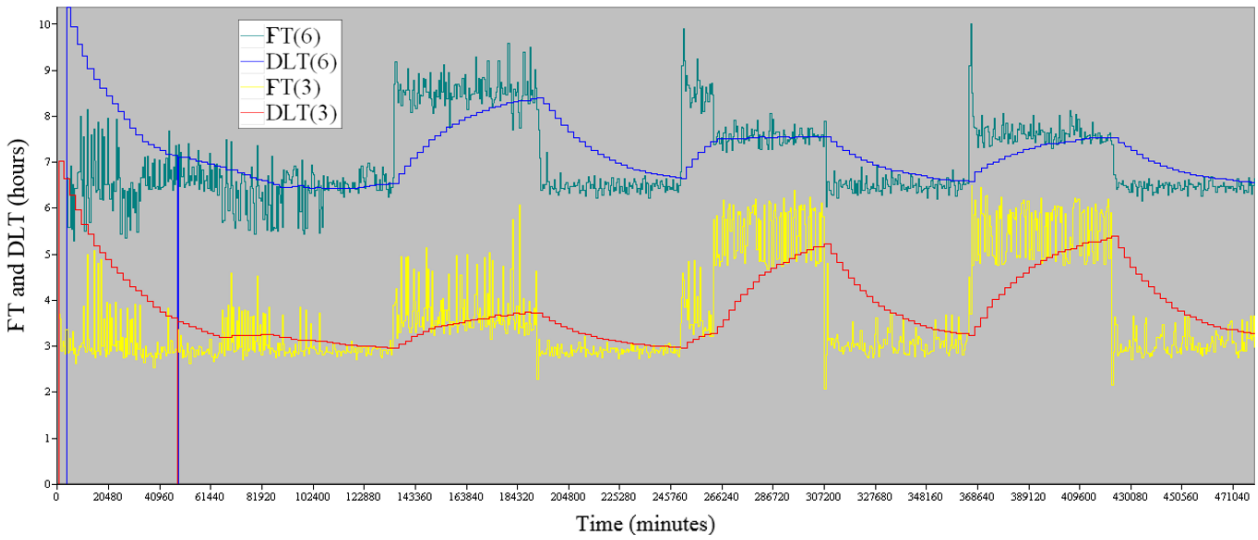


Figure 4.3 : Evolution of flow time and decoupled lead time of products 3 and 6 (respectively in yellow and in red, and in green and in blue).

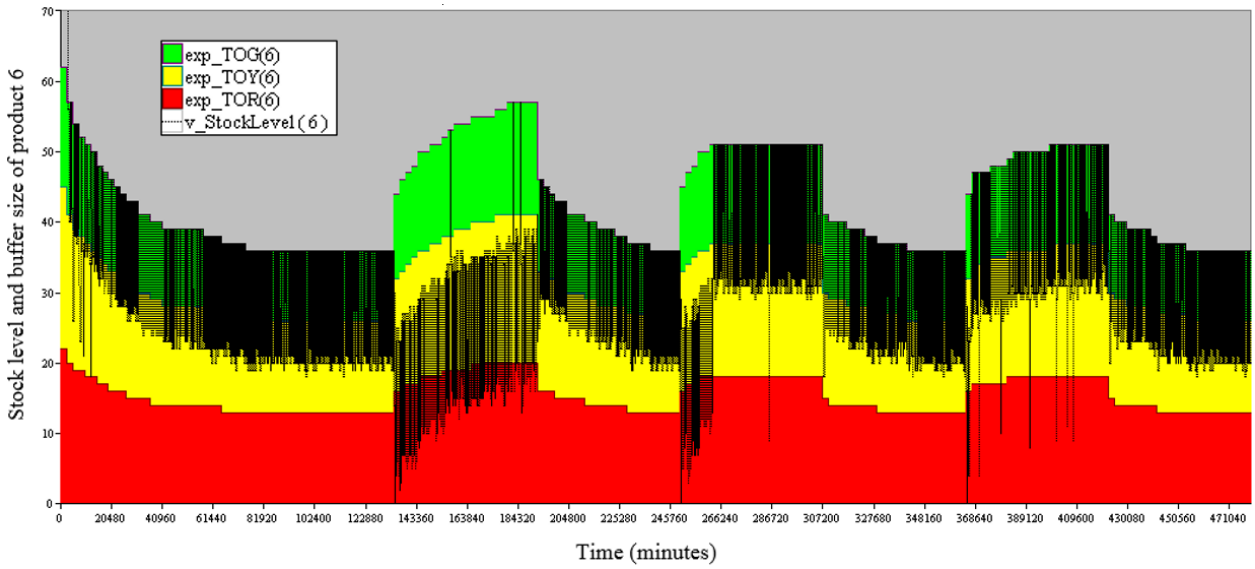


Figure 4.4 : Evolution of buffer size of the product 6 during the first simulation of scenario 12+.

However, it seems to change every time: in Figure 4.3, DLT(6) increases and decreases every six months but less and less. In contrast, DLT(3) increases more the second and third time than the first one. It does not seem to converge to an optimal DLT for each buffer, but for the entire system.

For a better service rate, there is scenario 9+, with a smoothing factor of 5%, a service rate of 99.8% and a reduction of stocks of 44% compared to scenario 2+. An exponential smoothing with a coefficient of 10% considers a little bit more the last DLT than one of 5%. That is why it reduces less the inventory, but it deals with the demand increase better when it comes.

A dynamic adjustment of the DLT enables to decrease inventory, while assuring a decent service rate (99.8% on average). The same trend is noted when customer demand decreases first (scenarios –, as we can see Table IV).

However, we note that operator utilization is:

- Lower with a DLT adjustment when we have 4 operators (85.8% and 90.5% instead of 91% and 94%).
- Higher with a DLT adjustment when we have 5 operators (78.5% and 80.3% instead of 73% and 75%).

This can be explained by looking at the buffer size: in under capacity scenarios (4 operators), the buffer sizes are increased until the demand customer is met (setup times have less impact). The

short-term workload increases, so the DLT increases trying to reach a balance between workload and capacity.

When capacity is enough (5 operators), buffer sizes are reduced, and consequently, lot sizes, and set up times have a bigger impact: for the same volume of products, operators work a little bit more.

When the customer demand step increases, measured DLTs (*i.e.* flow times) rise drastically and quickly (especially in under capacity scenarios). Here the control loop can become inefficient: in Figure 4.5, the scenario 1+ (without DLT control loop) seems steadier than scenario 8+ (with control loop), DLT is better, but real stock goes more frequently in the buffer red zone (too low inventory).

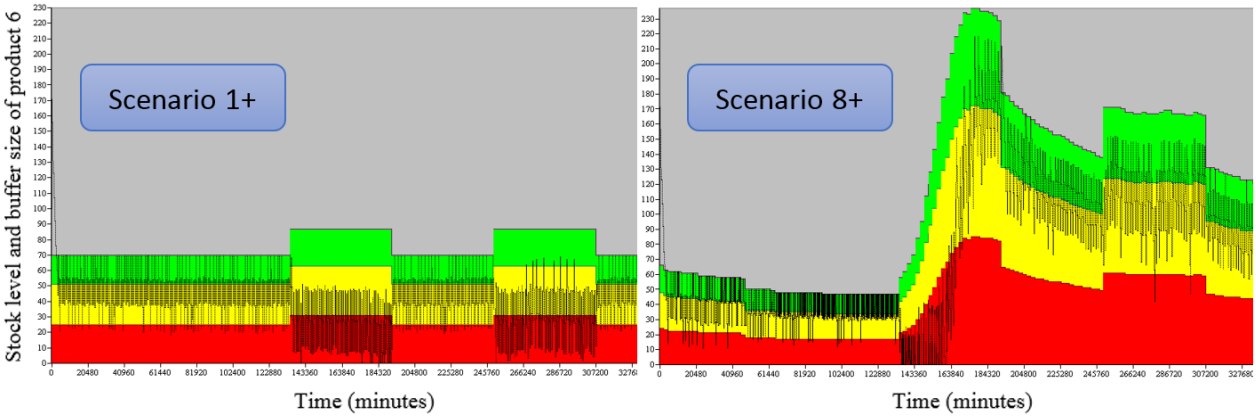


Figure 4.5 : Evolution of the buffer size of product 6 in scenario 1+ and scenario 8+.

The closer to the capacity limit, the higher the DLT rises. Figure 4.6 confirms this, regarding to scenarios 1+, 3+ and 6+ (only 4 operators) :

- In red is the flow time evolution of product 6 without control loop (scenario 1+). In green is the same flow time, with a 5% smoothing factor (scenario 3+) and in blue with a 10% smoothing factor (scenario 6+).
- With an exponential smoothing, just before the demand customer rise, DLT and buffer sizes are so low (and operator utilization is so high) that the system cannot absorb the demand rise, and flow times increase considerably just after the demand step (and DLT too).
- With a smoothing factor of 10%, the inventory is more reduced because the DLT re-decreases faster. But, the measured DLT (*i.e.* flow time) has more effect in the control, so

3 “DLT explosion” are noticed (each time demand rises). With a 5% smoothing, the DLT explosion is avoided once.

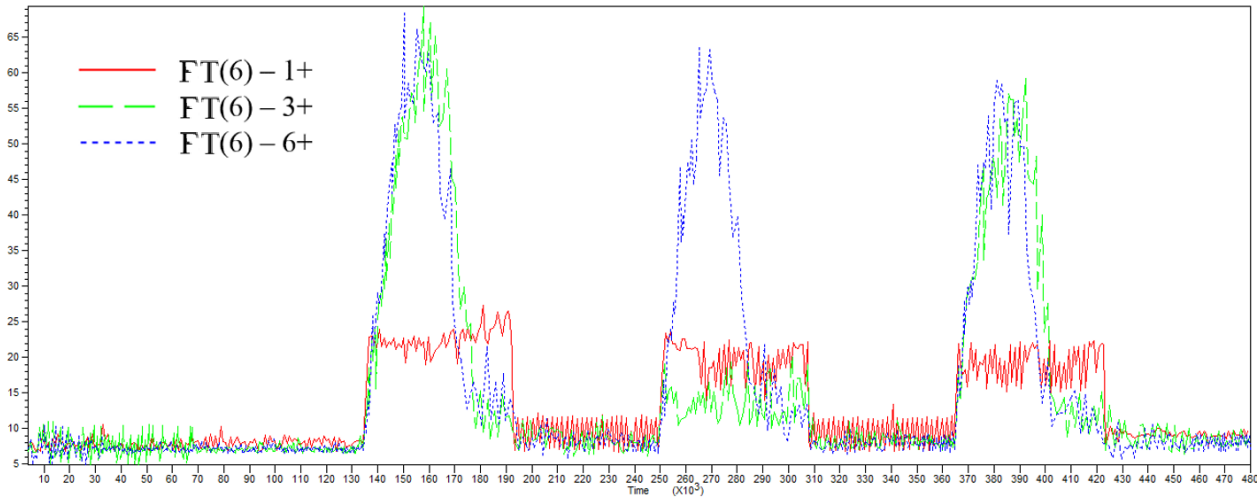


Figure 4.6 : Flow time evolution of product 6 in scenario 1+, 3+ and 6+.

A time shift between green peak and blue peak appears in Fig. 6, suggesting that with a smoothing factor of 10% (in blue) the system reacts faster.

Consequently, exponential smoothing of DLT is a powerful tool, but it seems to be risky in some cases. At last, when the ratio workload/capacity gets away from 100%, the dynamic adjustment of DLT offers better results.

## 4.5 Conclusion

With the precedent results, one can conclude that:

- Parameter Decoupled Lead Time of the method DDMRP has a huge impact on the system performances.
- Considering DLT in the dynamic adjustment of the buffer sizes reduces stocks.
- It also causes an increase of workload.
- In some cases, the system becomes under the limit of capacity and the flow times increase drastically.

An important point to underline, that could be a research subject, is that dynamic adjusting DLT considers the flow times of each product. Without DLT control, if only half of the references see

their ADU double, then the buffer sizes of these references will increase, whereas the size of the buffers of other products to whose ADU will stay the same. However, all these products use the same resources, and if the flow times of half of them increase then the others will increase too, and we have to adapt every buffer. The dynamic adjustment considering DLT allows to take into account this issue.

Nevertheless, it seems to be dangerous to adjust buffers when the ratio workload/capacity is close to 100%. The dynamic adjustment decreases the size of every buffer. As the size of buffers decreases, smaller orders are launched but more often. Then, the set-up consumption increases and so the workload. If the workload increases too much, the line cannot absorb a rise of the customer demand and the flow times in the workshop will “explode”. It seems obvious to determine a limit to the decrease of buffer size when is forecasted a rise of customer demand. Determining this limit can be a future research subject.

Furthermore, we do not consider the variation of each lead time factor (for each reference it is set to 0.75), that could have an impact on buffer sizing.

Finally, in this work we chose arbitrarily to adjust the decoupled lead times with an exponential smoothing and two smoothing factors (0.05 and 0.10). Another study could be to work on different possible methods to adjust the DLT (mean, moving mean, weighted moving mean, exponential moving mean, etc.) and the impact of smoothing factor choice.

# CHAPITRE 5 ARTICLE 2 : VISUAL CHARTS PRODUCED BY SIMULATION TO CORRELATE SERVICE RATE, RESOURCE UTILIZATION AND DDMRP PARAMETERS

Cet article a été soumis au journal *International Journal of Production Research* en avril 2021. À la suite du processus d'évaluation par les pairs, il a été révisé et resoumis en juillet 2021, puis accepté en septembre 2021 avec modifications mineures. La version présentée dans cette thèse est identique à la dernière version soumise.

Les auteurs de cet article sont : Guillaume Dessevre[1], Pierre Baptiste[1], Jacques Lamothe[2], Robert Pellerin[1].

[1]Département de Mathématiques et Génie Industriel, Polytechnique Montréal, Montréal, Canada.
[2]Centre Génie Industriel, IMT Mines Albi, Albi, France.

## Résumé

Le *Demand Driven Material Requirements Planning* (DDMRP) est une méthode récente mêlant flux poussé et flux tiré. Bien qu'elle se revendique comme la solution aux limitations des méthodes traditionnelles, la méthode DDMRP fonctionne à capacité infinie : les ordres de fabrication ou d'approvisionnement sont générés selon une logique de réapprovisionnement de stocks définis comme stocks tampons. Cet article propose une évaluation de la gestion capacitaire à l'aide de graphiques visuels construits par simulation. Ces graphiques corrèlent le taux de charge d'une ressource goulot au taux de service en considérant un des paramètres de la méthode DDMRP, le *Decoupled Lead Time* (DLT). Les graphiques sont des outils d'aide à la décision. Ils permettent d'identifier jusqu'à quel taux de charge les DLT sont représentatifs des temps de réponse des ordres de fabrication, et quel niveau de capacité utiliser. Nous étudions différents ateliers, dont un cas industriel réel. Nos résultats montrent qu'il est préférable de maîtriser les temps de défilement en ajustant la capacité plutôt que d'ajuster le paramètre DLT.

## Abstract

Demand Driven Material Requirements Planning (DDMRP) is a recent method mixing push and pull flow management. Although it claims to be the solution to traditional methods' limitations, the DDMRP method works at infinite capacity: manufacturing or supply orders are launched according to a logic of replenishment of stocks defined as buffers. This article proposes an evaluation of capacity management using visual charts developed by simulation. These charts correlate the bottleneck resource's loading rate to a service rate by considering one of the DDMRP method parameters, the Decoupled Lead Time (DLT). The charts are a decision support tool. They allow identifying to which loading rate the DLTs are representative of the flow times of manufacturing orders and which capacity level to use. We study different workshops, including a real industrial case. Our results show that it is better to control the flow times by adjusting capacity rather than adjust the DLT parameter.


Keywords: decision support; capacity management; decoupled lead time; DDMRP; simulation

## 5.1 Introduction

Demand Driven Material Requirements Planning (DDMRP) is a production planning and management method introduced by Ptak and Smith (2011). In integrating push and pull flows, this method is based on buffers' strategic positioning along the Bill of Material (BOM). DDMRP is one of the production control systems that emerged recently (Bagni et al., 2021).

As shown in Figure 5.1, each stock buffer is sized according to the Average Daily Demand (ADU), which can change over time, and several parameters, including the Decoupled Lead Time (DLT), which is defined as the longest upstream lead time not protected by a buffer, the Lead Time Factor (LTF), the Variability Factor (VF) and the Minimum Order Quantity (MOQ).

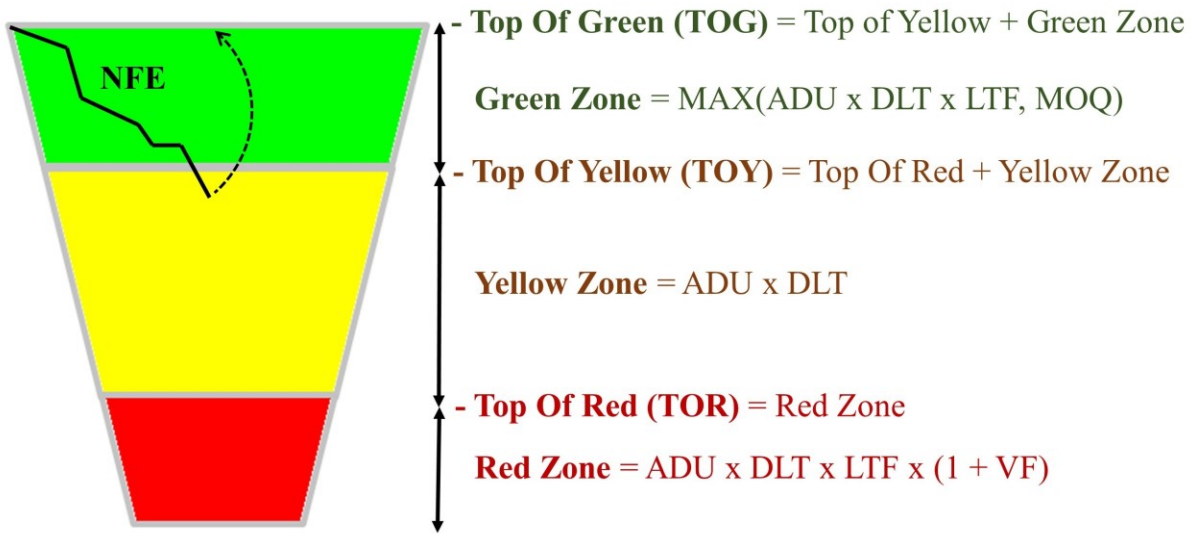


Figure 5.1 : The three zones of a DDMRP stock buffer.

A replenishment order to the TopOfGreen level is placed anytime a TopOfYellow threshold is exceeded:

$$Physical\ inventory + Work\text{-}in\text{-}Process - (Day\ demand + Peak\ demand) \leq TopOfYellow$$

A peak is an original concept of DDMRP. It corresponds to an exceptionally large demand in the backlog over a given time horizon (Ptak and Smith, 2016). It generates a replenishment on demand instead of on consumption.

Dessevre, Martin, Baptiste, Lamothe, Pellerin, et al. (2019) have shown that the choice of DLT can influence shop floor performance, including customer service rate. However, the "DLT - service rate" relationship still deserves to be clarified. Hopp and Spearman (1996) define a shop floor's service rate as the probability that the order flow times are less than or equal to the allocated lead time. In a DDMRP context, the shop service rate can be defined as the probability that these flow times are less than or equal to the DLT.

$$Workshop\ Service\ Rate = P(Flow\ Time \leq DLT)$$

Therefore, the DLT is an important parameter linked to both the service rate and the stock level (because it determines the buffer's zones): a too small DLT will not absorb all the flow times and may cause shortages, degrading both customer service rate and workshop service rate, and a too larger DLT will increase stock levels and the associated costs.

Also, queuing theory shows that flow times in production lines increase drastically when the loading rate increases (Kingman, 1962). However, DDMRP operates at infinite capacity, suggesting that scrolling times in a DDMRP managed shop can increase rapidly as the loading rate increases. To answer the capacity issue, Ptak and Smith (2011) propose a capacity buffer defined as "the protective capacity at both constraint and unconstraint resources that allows these resources to catch up when Murphy strikes", referring to Murphy's law: what can go wrong will go wrong. Unfortunately, a capacity buffer is just an additional amount of capacity to protect the system: it gives no recommendation, and it may be incorrectly sized.

This paper aims to create visual charts that correlate the bottleneck resource loading rate, the workshop service rate, and the DLT parameter by simulating different DDMRP managed workshops, including an industrial case.

The created charts will help to answer the following questions:

- Up to what loading rate of the bottleneck resource can the DLT absorb flow times?
- What percentage of flow time is less than or equal to this DLT?
- How does the service rate behave according to the loading rate?
- What can a production manager do to get under control the flow times and so the service rate?

Answering these questions will allow a better parameterization of the DDMRP method, as it will provide an assessment of the sizing of the DLT, with a visualization of the parameterization choices' consequences. Moreover, our charts can be used to determine the size of the capacity buffer. This paper is an extension of a previous one (Dessevre, Baptiste, and Lamothe, 2020), where we go further in the reasoning: we study here a real industrial case, we compare it with our fictional workshop, we answer another management problem (the choice of the number of shifts that the production manager must decide, explaining how to use the charts) and we show limits of the charts. The two papers are complementary in this research project.

This article is organized as follows. First, Section 2 presents a review of the literature on publications related to the topic. Section 3 then describes the research methodology (the workshops studied, the experimental design, and the simulation parameters). Section 4 presents the results, and finally, Section 5 concludes and proposes avenues for further research.

## 5.2 Literature Review

First, The DDMRP method is a controversial subject. Often restricted to a "consultants' method", it has made its way into the academic world with an increasing number of research articles dedicated to it. The first publications have demonstrated the method's relevance by comparing it to other traditional MRPII and Kanban methods, showing the force of the DDMRP: a better compromise between stock level and service rate, the anticipation of peak demand, dynamic adjustment of buffer sizing, and the ability to work with high product diversity (Ihme and Stratton, 2015; Miclo et al., 2016a; Shofa and Widyarto, 2017; Miclo et al., 2018). Today, the field of research on DDMRP has expanded. While some are interested in its strategic perspective, like Vidal et al. (2020) studying the Adaptive Sales & Operations Planning, others focus on the mechanics and operational parameters. For instance, Martin et al. (2018) propose a decision tree allowing a better parameterization of buffers. Dessevre, Martin, Baptiste, Lamothe, Pellerin, et al. (2019) are interested in DLT and LTF parameters by putting them under control. Recently Lee and Rim (2019) propose an alternative to the safety stock calculation model. The comparison of DDMRP with other methods is still up to date, as Thürer, Fernandes, and Stevenson (2020) compare four production control systems, showing the potential of DDMRP in multi-stage assembly systems. Nowadays, studies on DDMRP are both axiomatic and empirical (Bagni et al., 2021). However, there are still many issues to be addressed scientifically, while more and more

companies are developing DDMRP in many industrial sectors (Bahu, Bironneau, and Hovelaque, 2019). Therefore, researchers aim to study the method in more complex environments (Velasco Acosta, Mascle, and Baptiste, 2019), raising new questioning from particular industrial sectors (Dessevre et al., 2020), and bringing the need of a standardized implementation process for the method DDMRP (Orue, Lizarralde, and Kortabarria, 2020).

Then, we focus on the role of lead time in a workshop in literature. Hopp and Spearman (1996) define "the lead time of a given routine or line is the time allotted for production of a part on that routing or line". They clearly explain that lead time is different from flow time, as the former is a management choice, and the latter is generally random. A good lead time must absorb flow times and their variations, but it must but as small as possible to limit stocks: in a DDMRP context for example, the greater the lead time, the greater the DLT, and therefore the greater the stock buffers (Figure 5.1). Moreover, Christensen, Germain, and Birou (2007) show that lead time is linked to financial performance. That is why a branch in the literature is about "controllable" times, where these studies focus on reducing lead times in procurement (preparation and transport time) and/or in production (changeover time, production time, speed, batch size, etc.). For example, Sarkar, Mandal, and Sarkar (2015) studied two models with different demand distributions, Jha and Shanker (2013) included a constraint on the service rate in their model, Glock (2012) proposed methods for reducing lead time in a single-vendor-single-buyer model, and many others (Pan and Yang, 2002; Ouyang, Wu, and Ho, 2004; Hidayat and Simatupang, 2018; Shin et al., 2016). Thus, many scientists consider lead times as a decision variable and try to size it as best as possible.

Finally, the factor $\rho/(1-\rho)$, where $\rho$ represents the loading rate, is present in the formulas for calculating average flow times from the queue theory of 1-server systems: Kingman (1962), Marchal (1976) or Krämer and Langenbach-Belz (1976). Therefore, the higher the loading rate, the more drastically the times increase. To avoid this issue, it is important to control the loading rate, especially for the bottleneck resource. Ptak and Smith (2011) introduce a capacity buffer, which is an additional amount of capacity in order to absorb variability. They explain that "capacity buffers are not being used […] to maximize a resource's utilization or efficiency. [They] require that a resource maintain a bank a capacity that goes unused" (Ptak and Smith, 2016). But how to determine the size of that bank? And if it is not enough? They answer the former by analyzing the demand variability, and the latter by proposing different long-term methods such as reengineering

the products to manipulate the load or raising price to manipulate the demand. Thereby, there is no answer for short-term capacity issues.

As a conclusion of this review, it is known that (i) flow times are correlated with the loading rate (by the queue theory), that (ii) lead time is a decision variable that can be reduced at a certain cost (controllable times), that (iii) the DLT parameter of the DDMRP method is used in the dimensioning of buffers and that (iv) it is an important parameter related to the service rate of the workshop. One question thus arises: How can we control a service rate in a DDMRP-managed shop subject to load variations? To answer this question, we propose to create charts correlating the service rate, the loading rate, and the DLT parameter.

## 5.3 Methodology

Our research strategy is based on the study of two simulated cases: a fictional flowshop and an industrial case. Both cases are similar: products are buffered and manufactured in a workshop where components are buffered too. In this way, there is always a flowshop with a bottleneck station between two buffers. If both cases give similar results, we might work on the fictional case in future research between two buffers, as it is easier to change fictional parameters (number of stages, products, etc.). Simulation has been chosen because it easily allows to model and analyze complex environments with several sources of variability (Mourtzis, 2020). Furthermore, the objective is not to optimize but to observe a phenomenon, it is therefore not necessary to develop more “fine” tools.

This section describes the workshops studied as well as the simulation parameters used.

### 5.3.1 The fictional flowshop

#### 5.3.1.1 Workshop parameters

The workshop studied is a production line composed of 6 workstations (one machine per workstation) with DDMRP stock buffers at the beginning and end of the line (for components and finished products, as shown in Figure 5.2). This type of workshop was chosen since it corresponds to what can be found between two DDMRP stock buffers in the industry.

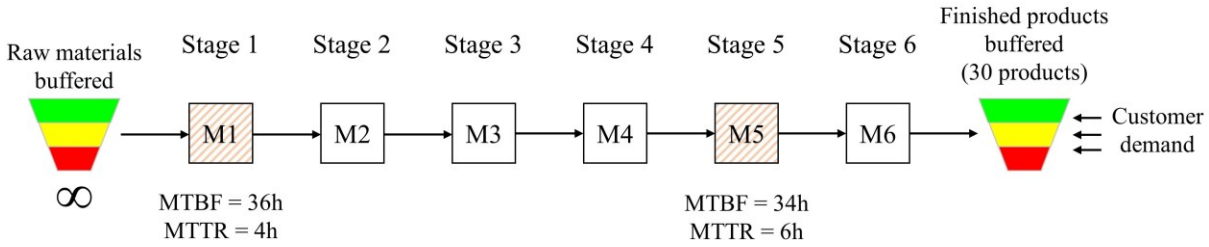


Figure 5.2 : Diagram of the studied production line and positioning of the DDMRP stock buffers.

Since DDMRP performs well in highly variable environments, several sources of variability were introduced into the system (demand, production, etc.). The first machine is subject to 4-hour outages every 36 hours and the fifth machine to 6-hour outages every 34 hours, representing 10% and 15% of the time respectively. These values arbitrarily follow exponential laws to model a non-negligible but not excessive variability in production, close to reality.

Thirty products are manufactured in the workshop. Series changeover times (in hours) and production times (in seconds per part) per machine are presented in Tableaux 5.1 and 5.2. To simulate variability in production times, the changeover times are multiplied by a variable following a triangular law (0.75, 1, 1.25) and the production times by a variable following a triangular law (0.8, 1, 1.2). An operator is required during changeover and production operations for a machine to operate.

Tableau 5.1 : Changeover time (in hours) for each product.

| **Products** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** |
|---|---|---|---|---|---|---|
| 1 to 10 | 0.25 | 0.25 | 0.5 | 0.5 | 0.25 | 1 |
| 11 to 20 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| 21 to 30 | 1 | 1 | 0.5 | 0.5 | 1 | 0.25 |

Tableau 5.2 : Production time (in seconds per part) for each product.

| **Products** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** |
|---|---|---|---|---|---|---|
| 1 to 10 | 5 | 5 | 5 | 10 | 15 | 5 |
| 11 to 20 | 5 | 10 | 10 | 10 | 10 | 10 |
| 21 to 30 | 5 | 15 | 15 | 10 | 5 | 15 |

Each product has a different average order size, and this order size follows a uniform law of ±20% around the average as an arbitrary choice after observing demand signals from the industrial case. The delay between each order of the same product follows an exponential law of expectation of 1 day. Therefore, the Average Daily Usage (ADU) of each product equals the average order size, shown in Tableau 5.3.

Tableau 5.3 : Average order size for each product.

| **Products** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| ADU | 65 | 23 | 55 | 63 | 19 | 38 | 29 | 33 | 61 | 34 |
| **Products** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| ADU | 61 | 62 | 57 | 32 | 52 | 21 | 61 | 42 | 48 | 44 |
| **Products** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **229** | **30** |
| ADU | 35 | 52 | 23 | 56 | 55 | 37 | 32 | 51 | 55 | 52 |

The DLT, LTF, and VF are equal for each product and are valid for 10 days, 50%, and 50%, respectively. This choice of parameterization comes from previous research conclusions (Dessevre, Martin, Baptiste, Lamothe, and Lauras, 2019).

The workshop and operators work 8 hours a day, 5 days a week. Since the components are themselves managed on DDMRP stock buffers, they are considered available at all times in sufficient quantities. Finished product stocks are randomly initialized between 50% and 100% of the Top Green of each buffer.

Because of breakdowns and product mix, the bottleneck is globally located on the M5 machine but can temporarily shift to other machines. Moreover, when the number of operators is reduced, the bottleneck resource becomes the operators.

The workshop manages partial orders: when a customer order arrives, it is delivered in full if possible. Otherwise, it enters a queue and will be given priority when the product's stock in question is available again.

#### 5.3.1.2 Design of experiments and simulation parameters

For the fictional case, the goals of the design of experiments are:

- To verify the average flow time of all production orders represents the average flow time of productions orders for each finished product;

- To compare a case where the bottleneck resource is a machine and the one where it is the operators; and
- To create flow charts that will be compared with the industrial case.

The input and output variables of the design of experiments are presented here. There are two input variables: the average customer demand and the number of operators.

To generate a progressive scale-up, the average size of each order has been uniformly increased step by step. The aim is to have a load/capacity ratio between 70% and 100% (in all demand scenarios, below this ratio, the workshop has overcapacity, above this ratio, it is overused and saturates very quickly).

Also, two cases are studied to test two different types of critical resources:

- Case 1: the bottleneck is a machine (the fifth on the line) and 6 operators are present in the workshop; and
- Case 2: 5 operators are present in the workshop and represent the bottleneck resource.

The output variables of the experimental design are:

- The loading rate of the bottleneck resource;
- The workshop service rate, defined according to Hopp and Spearman (1996) as the ratio between the number of production orders with flow times less than or equal to the allotted time (i.e., the DLT), and the total number of production orders;
- The customer service rate, defined as the ratio of the number of orders filled on time to the total number of orders;
- Flow time distribution, defined as the value where X% of the production orders have flow times less than or equal to this value (where X is 50, 60, 70, 80, 90, and 100); and
- The average flow times.

These variables will allow us to plot the desired charts presented in the results section. 120 scenarios were simulated for each case. Each scenario is simulated 100 times and each replication lasts 110 weeks (550 days), including a 10-week warm-up period (not included in the results). The modelling and simulation were performed on Arena version 15.10 software.

## 5.3.2 The industrial case

### 5.3.2.1 Workshop parameters for the industrial case

The industrial case is made of two consecutive workshops. In the first one, the raw materials are weighed, then mixed, and heated up in a reactor to manufacture semi-finished products in a tank. In the second one, the semi-finished products are packed on a conditioning line to manufactured finished products, which are shampoo bottles filled up by the semi-finished products (Figure 5.3).

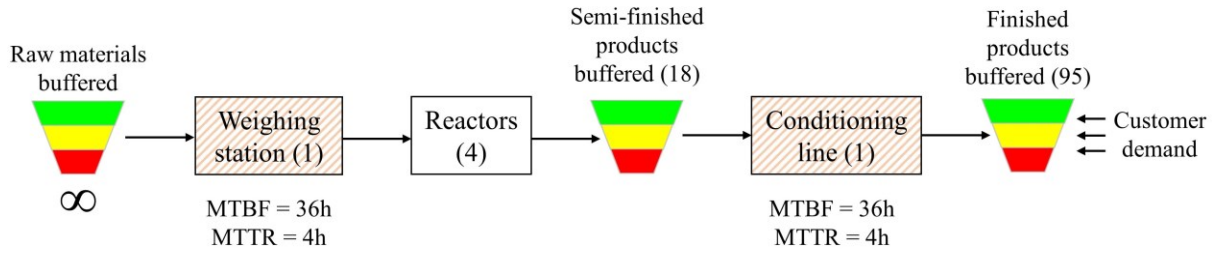


Figure 5.3 : Diagram of the studied industrial case composed of two workshops and positioning of the DDMRP stock buffers.

The raw materials (plant extracts) are buffered and considered available at all times in sufficient quantities. The semi-finished products (shampoos) are buffered, and there are 18 different ones (almond, mint, quinine, etc.). To produce them, the raw materials are weighed at the weighing station, mixed and heated in a reactor, and finally transferred in a tank. There are one weighing station, out of order 10% of the time, and four reactors. The weighing station is the bottleneck resource in this workshop, working 24 hours a day (3 shifts of 8 hours).

The 95 finished products are buffered. They are made of semi-finished products by being operated on a conditioning line. A semi-finished product can make between 2 and 8 different finished products: The differentiation between finished products from the same semi-finished products comes from the bottle size and the linguistic version. The link between semi-finished products and finished products is shown in Tableau 5.4.

Tableau 5.4 : Links between semi-finished products and finished products.

| Semi-finished product | Bottle size | Finished products | Semi-finished product | Bottle size | Finished products |
|---|---|---|---|---|---|
| 1 | 200 | 1 – 3 | 10 | 200 | 50 – 52 |
| | 400 | 4 – 5 | | 400 | 53 – 55 |
| 2 | 200 | 6 – 8 | 11 | 200 | 56 – 58 |
| | 400 | 9 – 11 | | 400 | 59 – 60 |
| 3 | 200 | 12 – 14 | 12 | 200 | 61 – 61 |
| | 400 | 15 – 17 | | 400 | 63 – 64 |
| 4 | 200 | 18 – 20 | 13 | 200 | 65 – 67 |
| | | | | 400 | 68 – 70 |
| 5 | 200 | 21 – 23 | 14 | 200 | 71 – 73 |
| | 400 | 24 – 26 | | 400 | 74 – 75 |
| 6 | 200 | 27 – 29 | 15 | 200 | 76 – 77 |
| | 400 | 30 – 31 | | 400 | 78 – 79 |
| 7 | 200 | 32 – 34 | 16 | 200 | 80 – 82 |
| | 400 | 35 – 37 | | 400 | 83 – 85 |
| 8 | 200 | 38 – 40 | 17 | 200 | 86 – 88 |
| | 400 | 41 – 43 | | 400 | 89 – 93 |
| 9 | 200 | 44 – 46 | 18 | 200 | 94 – 95 |
| | 400 | 47 – 49 | | | |

The conditioning line is out of order 10% of the time. There are six types of finished products, two according to the bottles' size (200mL and 400mL) and up to 3 different linguistic versions. About half of the finished products of the same semi-finished products is made of 200mL bottles. The time to change the size of bottles on the conditioning line is about four times the time to change the semi-finished product. Therefore, the decision to change the bottle size is taken the first day of the week only: on Monday, the conditioning line is set to satisfy the longest queue of conditioning orders between those in 200mL and those in 400mL.

The DLTs are fixed to 10 days for the first workshop (weighing station and reactors) and 15 days for the second (conditioning line). The LTFs are set to 10% for the semi-finished products buffers and 50% for the finished products buffers. The VF are set to 10% for the semi-finished products buffers and 20% for the finished products buffers. The size of a production order for the semi-finished products are predetermined, depending on the formula and the reactor size (6 or 10 tons). Therefore, we cannot launch orders of 4 tons or 13 tons, for example. The finished products buffers have a MOQ of 5 000 bottles (this is a psychological threshold below which operators and

managers do not see the point of launching an order). The workshop manages partial orders the same way as the fictional case.

The second workshop can work 16 hours a day (2 shifts of 8 hours per week), 20 hours a day (1 week with 2 shifts and 1 week with 3 shifts, resulting in an average of 2.5 shifts per week), or 24 hours a day (3 shifts). The production manager's goal is to find the best number of shifts to deal with the customer demand. Thus, we will create visual charts to help decide between the three possibilities of the number of shifts (2, 2.5, and 3).

#### 5.3.2.2 Design of experiments and simulation parameters for the industrial case

For the industrial case, the design of experiments has different goals:

- To compare and validate the fictional case study (the charts must look like the same);
- To help the production manager to find the number of shifts with the visual charts, depending on an expected demand; and
- To compare the average flow times of the same resource (the conditioning line) between the three capacity possibilities (number of shifts), depending on the resource's loading rate.

To do so, we generate a progressive scale-up to have a load/capacity ratio between 70% and 100% for both workshops and for each capacity possibility for the second workshop. For the chart with the three shifts, we simulate the workshop with an average weekly demand from 75 000 bottles a week to 300 000 bottles a week for each shift.

The outputs variables are the same as the ones for the fictional case: the loading rates of the bottleneck resources, the workshop service rates, the customer service rates (named “customer service rate” for the second workshop because it serves customers, and “buffer service rate” for the first workshop because the second one is the customer of the first one), the average flow times and their distributions.

300 scenarios were simulated (60 for the first workshop and 240 for the second one), each of them is simulated 100 times, and each replication lasts 60 weeks, including a 10-week warm-up period.

## 5.4 Results and discussion

This section presents the results of the above scenarios. For space reasons, only the visual charts are presented.

### 5.4.1 The fictional flowshop

First, we verify the average flow time of all Production Orders (PO) represents the average flow time of PO for each finished product. Figure 5.4 illustrates the average flow times of PO for six different finished products (finished products 5, 10, 15, 20, 25, and 30 in different blues) and the average flow time for all PO (in black) depending on three loading rates of the bottleneck resource (70%, 85%, and 95%). It shows an unsubstantial difference between flow times of different finished products. The average flow time all of POs can be used to represent the flow time of POs for all finished products, independently of the loading rate of the resource. We explain it because most of the flow time is queue time in front of each stage (Hopp and Spearman, 1996), then differences of changeover time and production time between products do not affect the average flow times.

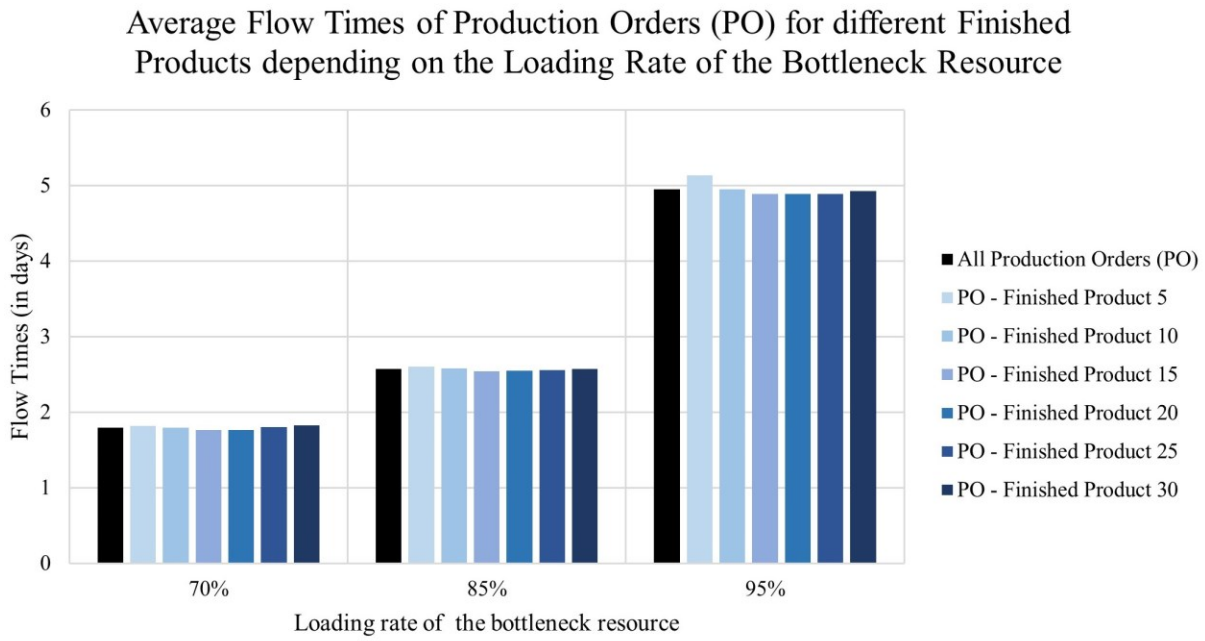


Figure 5.4 : Average Flow Times of Production Orders for different Finished Products depending on the Loading Rate of the Bottleneck Resource.

Now we can create charts using the average flow time to represent all finished products. Both charts in Figures 5.5 et 5.6 have the same axes and legend:

- On the abscissa is the loading rate of the bottleneck resource (machine 5 for figure 5 and the operators for figure 6), ranging from about 70% to 100%;
- On the left y-axis is the flow times in days. The black curve represents the average flow time. The dotted horizontal line represents the DLT of the workshop. The areas correspond to the distribution of flow times (from 50% in light blue to 100% in dark blue); and
- On the right y-axis is the service rate scale, between 0 and 100%. The curve represents the workshop service rate at the square marks and the customer service rate by the curve at the circular marks.

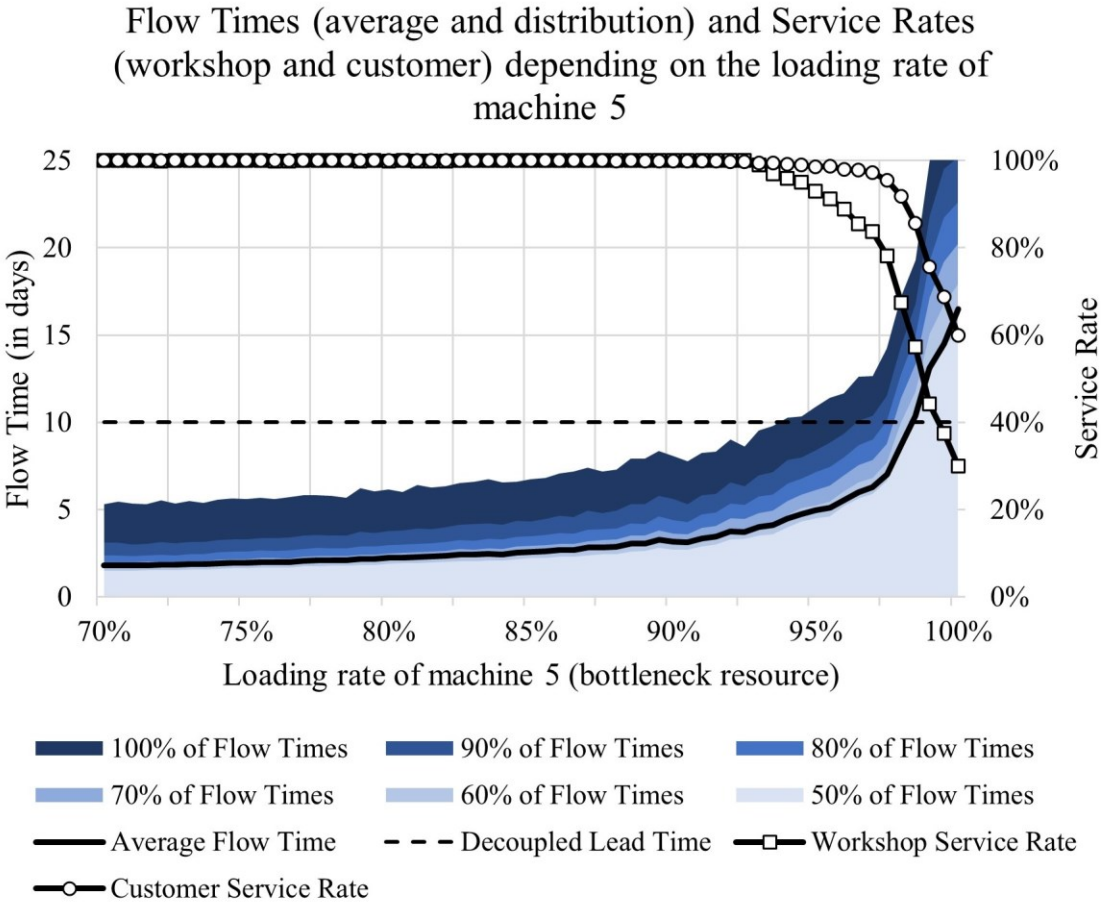


Figure 5.5 : Flow Times (average and distribution) and Service Rates (workshop and customer) depending on the loading rate of machine 5.

From these graphs, it follows that the more the loading rate of the bottleneck resource increases, the more the average flow time increases hyperbolically in the workshop (conclusion known by the queue theory (Hopp and Spearman, 1996)) but also its dispersion. The average flow time coincides with the area representing 50% of all flow time in the workshop (i.e., the distribution median). Therefore, taking a DLT equal to the average flow time represents only half of the manufacturing orders.

The workshop service rate begins to deteriorate sharply when the DLT line crosses the different flow time distribution areas. The customer service rate is slightly out of line with the workshop service rate due to each stock buffer's red zone, acting as a safety stock.

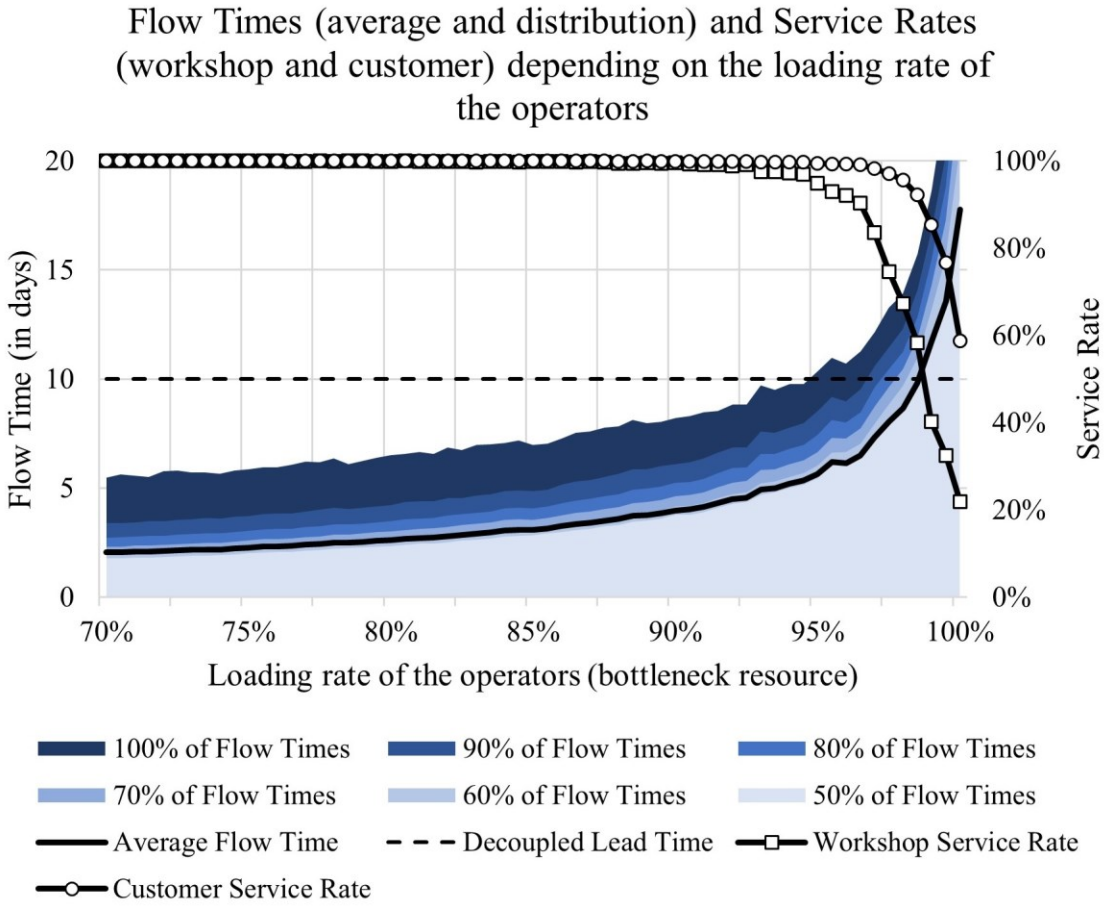


Figure 5.6 : Flow Times (average and distribution) and Service Rates (workshop and customer) depending on the loading rate of the operators.

There is not much difference between the two charts (Figure 5.5 and 5.6), meaning that machines and operators behave the same way as the bottleneck resource.

These charts help answer the first three research questions. In our example, the DLT set at 10 days can absorb 100% of the flow times up to a load rate of about 91%, 90% of the times up to a rate of 94%. This shows that the average remaining flow time of around 3 to 5 days should not be relied upon as long as the load rate is below 94%. A 5-day DLT would not even absorb all the flow time for a 70% charge rate.

Then, for example, for a 97% loading rate in Figure 4, the DLT would only absorb 80% of the flow times. As a result, the workshop service rate drops to 80%, and the customer service rate also deteriorates to around 92%. The production manager, anticipating a ramp-up (or ramp-down) on the shop floor, might then choose to adjust the DLT accordingly, or seek to reduce the loading rate by adapting capacity to control flow times (with a target service rate), or increase/decrease safety stocks to absorb flow time variations.

### 5.4.2 The industrial case

In this case, we verify that the chart looks like the fictional case (to validate our assumptions), and we show how to use the visual charts to help decide on the number of shifts.

Figure 5.7 has the same axes and legend as the previous charts. It represents the average flow time, dispersion, and service rates depending on the weighing station's loading rate (the bottleneck of the first workshop).

The chart is closer to the ones made with the fictional workshop: when the bottleneck resource's loading rate increases, flow times increase too, and service rate decreases. In this case, a DLT set to 10 days covers 100% of the flow times until around 85% of the loading rate. Above this value, the production manager might increase the DLT or seek to reduce the loading rate. To visualize the effect of reducing loading rate, we created a chart for the conditioning line of the second workshop.

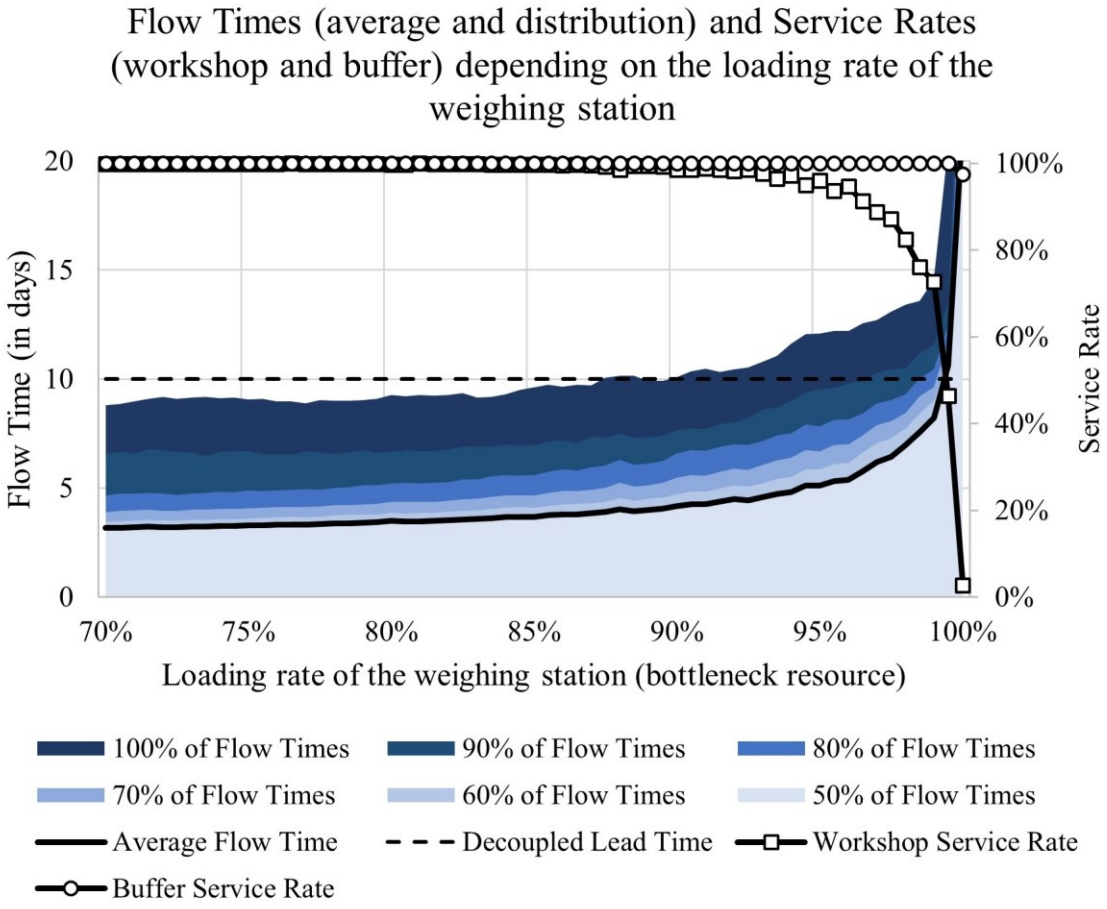


Figure 5.7 : Flow Times (average and distribution) and Service Rates (workshop and buffer) depending on the loading rate of the weighing station.

In Figure 5.8, the average weekly customer demand of the workshop is represented in abscissa, and the curves represent the average flow times (full curves), 80% of flow times (dashed curves), and 100% of flow times (dotted curves), when the conditioning line works with 2 shifts (black curves), 2.5 shifts (blue curves) and 3 shifts (green curves). The red dashed line is the Decoupled Lead Time.

As the flow times increases drastically when the demand increases, it seems better to seek to reduce the loading rate (by changing the number of shifts) rather than changing the DLT (what increases the buffer sizing and reduces the shop's reactivity) : for example, with a demand of 150 000 bottles a week, a DLT of 15 days absorbs all the flow times with 2.5 shifts while it needs a DLT of 30 days to absorb only the average flow times.

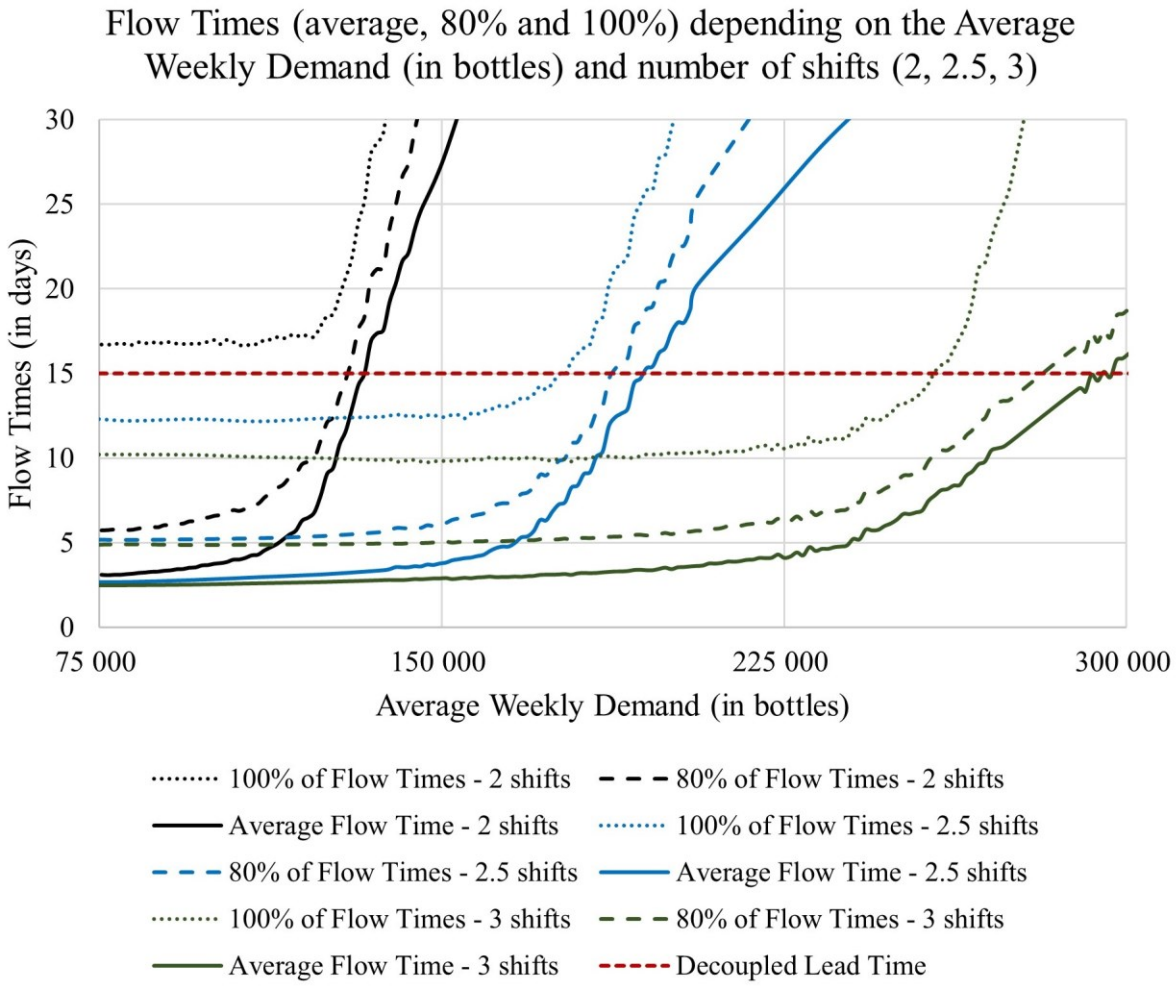


Figure 5.8 : Flow Times (average, 80% and 100%) depending on the Average Weekly Demand (in bottles) and the number of shifts (2, 2.5, 3).

When the average demand increases, the loading rate of the conditioning line increases too, so does the average flow time (and its dispersion). As a result, the production manager could use this chart: if the average demand is not higher than 100 000 bottles a week, 2 shifts and a DLT of 15 days are enough to absorb at least 90% of flow times. If we forecast more than 100 000 bottles a week, the production manager has to increase the DLT, or set 2.5 shifts. Above 200 000 bottles a week, 2.5 shifts are not enough, and we need 3 shifts. Figure 5.9 represents the average flow times depending on the conditioning line's loading rate for the three sets of shifts.

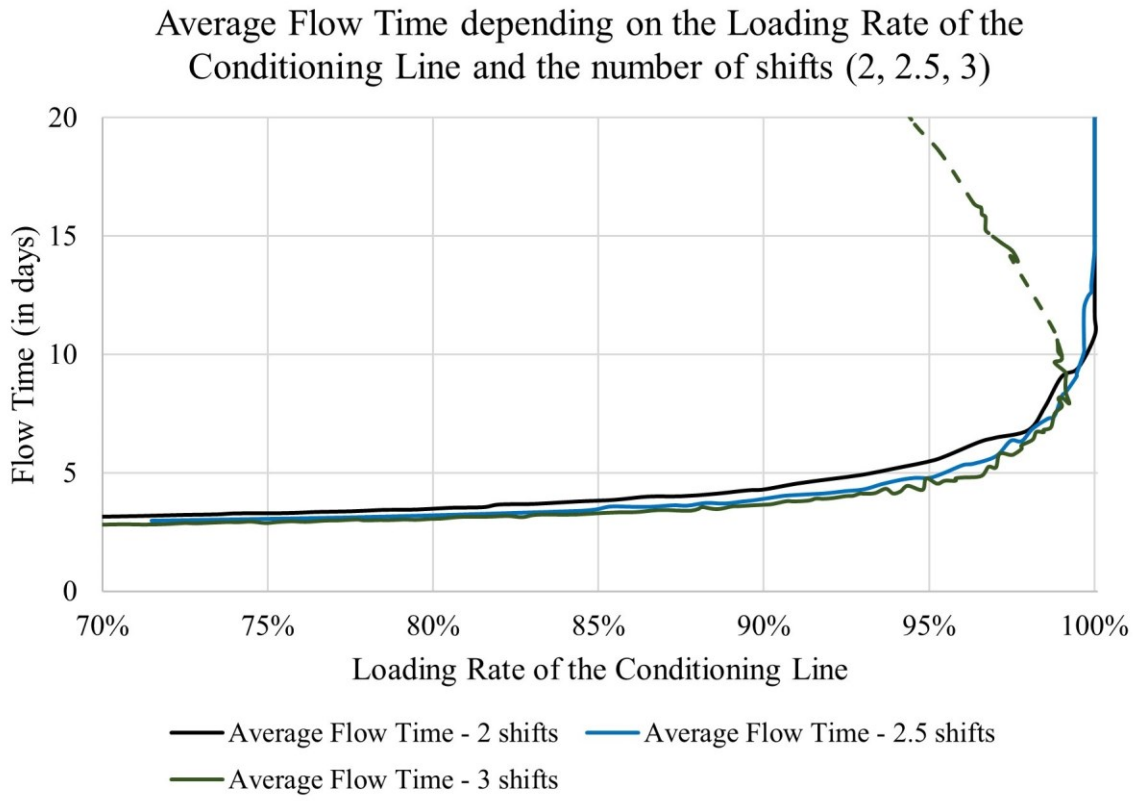


Figure 5.9 : Average Flow Time depending on the Loading Rate of the Conditioning Line and the number of shifts (2, 2.5, 3).

Firstly, under 98% of loading rate, the three curves are very close, meaning that the average flow times is nearly the same with 2, 2.5, or 3 shifts when the resource is used at the same loading rate (for example, the average flow time is around 5 days at 95% of the loading rate, regardless of the number of shifts).

Secondly, the green curve (3 shifts) turns around when it reaches 98%. The conditioning line set with 3 shifts never get a loading rate above 98%, because the first workshop (weighing station and reactors) becomes the bottleneck workshop. Consequently, as demand increases, conditioning orders are waiting for semi-finished products availability, which increases the flowtime while reducing the loading rate of the conditioning stage. To better understand, Figure 5.10 represents the loading rate of the weighing station (black curve), the weighing workshop buffer service rate (blue curve), and the loading rate of the conditioning line (green curve). Above 250 000 bottles a week, the weighing station loading rate reaches 100%. This is why its service rate begins to

decrease (as we saw in Figure 5.7 previously). Therefore, the semi-finished products are slow to arrive, delaying the production orders of finished products. As a result, the conditioning line is waiting, and its loading rate decreases. This last chart represents the relation between the two workshops, depending on the demand. Above 250 000 bottles a week, the average flow times increase because both the loading rates of the conditioning line are near 98%, and the loading rate of the weighing station reached 100%.

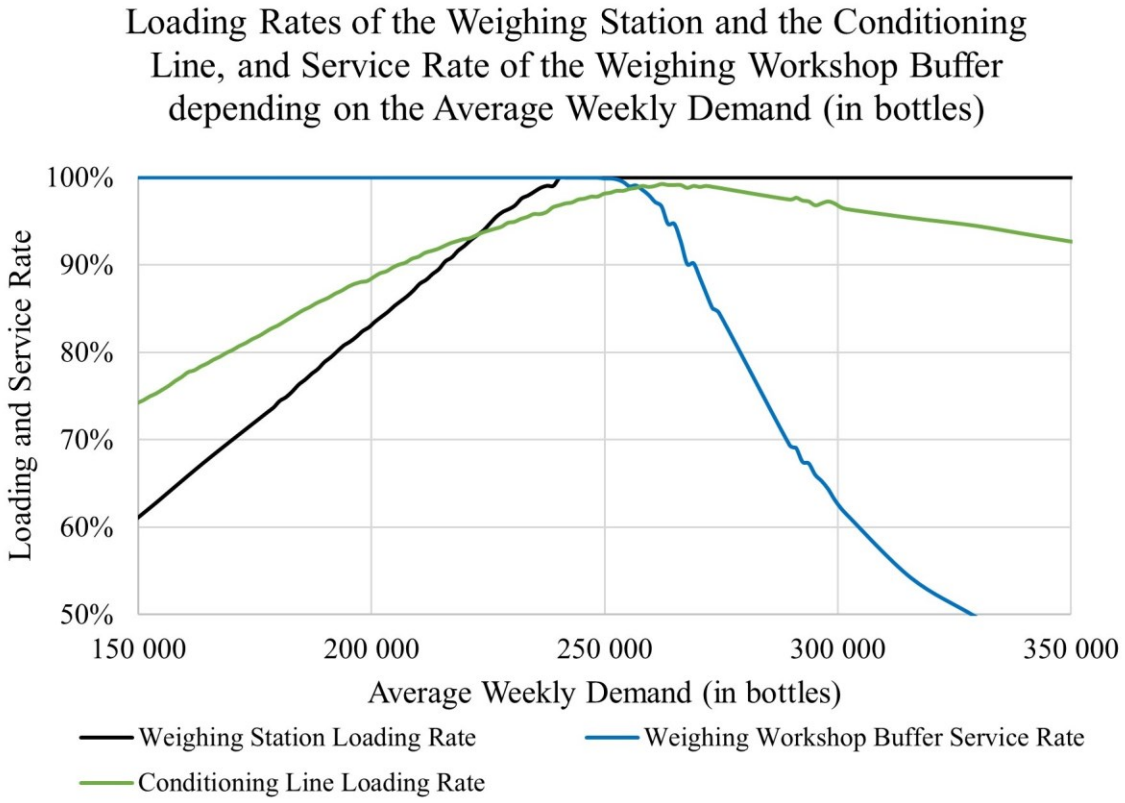


Figure 5.10 : Loading Rates of the Weighing Station and the Conditioning Line, and Service Rate of the Weighing Workshop Buffer depending on the Average Weekly Demand (in bottles).

## 5.5 Conclusion

This paper proposes charts to visually correlate bottleneck resources load rate, flow time distribution, and service rates. They are realized by simulating DDMRP-managed production workshops, including an industrial case. Using the workshop service rate definition as the probability that the flow times are less than or equal to the DLT parameter, we have two choices

to improve the workshop service rate: Control the flow times by adjusting the capacity and/or adjust the DLT parameter.

The graphs we have created help deciding based on the expected loading rate or the expected customer weekly demand. It enables mid-term management of capacity as demand trend changes. These experimental results are essentially very contextual. That is why we simulated an industrial case to validate our fictional one. However, it will be necessary to verify their sensitivity to key workshop parameters: product mix, workshop variability, and technical data. In this study, charts show similar behavior between the fictional case with bottleneck machine, the fictional case with bottleneck operators, and the industrial case with different numbers of shifts.

The customer service rate is then shifted from the workshop service rate due to the safety stock represented by the DDMRP red zone. Therefore, it is possible to keep a correct customer service rate by playing on this safety stock. However, it requires a strong control of the loading rate. The slightest error induces a drastic drop in the service rate. Consequently, it is more interesting to try to control the flow times rather than to adapt the DLT parameter or to play with the safety stock: a minor capacity adjustment (from 95% to 90%, for example) is equivalent to significantly increasing the DLT (from 25% to 50%) without impacting the average stock. This remark supports the implementation of capacity buffers as suggested by Ptak and Smith (2011), and our charts can be used to determine the size of the capacity buffer.

In our industrial case, the production manager's concern is to find which number of shifts is required to face the demand (the decision is made weekly or monthly). The charts correlate the flow time dispersion depending on the forecasted customer demand and the number of shifts working. They help to decide if 2 shifts are good enough, or if more working hours are needed.

Therefore, it remains up to the production manager to choose which capacity adjustment to implement in the workshop, depending on the available levers (increasing the number of operators, working overtime, working several shifts (switching to 3 shifts for example), capacity subcontracting, etc.). Thus, we can imagine different charts similar to those presented here. Each curve would propose a different scenario (6 operators with 2 shifts or 4 operators with 3 shifts, for example, or even a scenario with an increase in DLT and comparing average in-process and finished product stocks). These graphs would be used as a decision aid for capacity adjustments

when forecasting a load increase (or decrease), by setting a customer and/or workshop service rate objective. They could thus be used for tactical or even strategic decisions.

Finally, we showed that the average flow time is not suitable to size the DLT, and it is better to know the dispersion of flow times. By coupling the DDMRP method with a tool reducing the variability of flow times, we could reduce the dispersion, and therefore reduce the DLT, thus reducing stocks while maintaining a high loading rate. Moreover, other types of workshops could also be processed (job shop, open shop, production lines crossovers, etc.) to verify that the same phenomena are identified.

# CHAPITRE 6 ARTICLE 3 : IMPROVEMENT OF THE DDMRP PRODUCTION SYSTEM BY COUPLING REORDER POINT AND CONWIP LOOP

Cet article a été soumis au journal *International Journal of Production Research* en octobre 2021. La version présentée dans cette thèse est identique à la version soumise.

Les auteurs de cet article sont : Guillaume Dessevre[1], Pierre Baptiste[1], Jacques Lamothe[2], Robert Pellerin[1].

[1]Département de Mathématiques et Génie Industriel, Polytechnique Montréal, Montréal, Canada.
[2]Centre Génie Industriel, IMT Mines Albi, Albi, France.

## Résumé

Dans le monde industriel volatil et incertain d'aujourd'hui, les méthodes à flux poussé et à flux tiré ont atteint leurs limites, et les stocks de sécurité sont devenus insuffisants. Les auteurs du *Demand Driven Material Requirements Planning* (DDMRP) ont tenté de résoudre ce problème. Le mécanisme de génération d'ordres du DDMRP est similaire à un point de commande : le stock tampon est réapprovisionné à son maximum lorsqu'un seuil défini est atteint. Cependant, le DDMRP considère les variations de la demande dans la position de stock de ses stocks tampons mais ne considère par l'état de l'atelier pour décider combien et quand générer les ordres de fabrication. Dans cet article, nous proposons de coupler la méthode DDMRP avec une boucle ConWIP générant des ordres de fabrication chaque fois qu'un ticket ConWIP revient et qu'un stock tampon doit être réapprovisionné. Nos résultats montrent que le modèle DDMRP-ConWIP Intégré améliore le flux dans l'atelier en adaptant la taille des lots aux perturbations venant de la demande et de la production. Notre modèle est plus robuste et plus performant que le système DDMRP classique et le modèle à point de commande ($R$, $Q$).



## Abstract

In today's volatile and uncertain industrial world, both push and pull methods have reached their limits, and safety stocks have become inadequate. The authors of the Demand Driven Material Requirements Planning attempted to resolve this issue. The DDMRP order launch mechanism is similar to an order point: the buffer is replenished to its top when a defined threshold is reached. However, the DDMRP considers demand variations in its buffers' net flow position but does not consider the state of the workshop when deciding how much and when to launch production orders. In this paper, we propose to couple the DDMRP method with a ConWIP loop generating production order whenever a ConWIP ticket comes back, and a buffer needs to be replenished. Our results show that the DDMRP-ConWIP Integrated model improves the flow of the workshop by adapting lot sizes to perturbations from both demand and production. Our model is more robust and outperforms the classic DDMRP system and the reorder point model (R, Q).


Keywords: DDMRP, Reorder Point, ConWIP, flow times, robustness.

## 6.1 Introduction

The Demand Driven Material Requirements Planning (DDMRP) is a recent material management method presented by Ptak and Smith (2016), mixing push and pull flow management. It is based on buffers' strategic positioning along the bill of material, that will be replenished when a threshold is reached, which is a comparison of the stock level and the Work-In-Process (WIP), with the qualified demand (defined as the sum of daily demand, the detected demand peaks, and the unsatisfied demand). The mechanic is simple: every time this threshold called "Top of Yellow" (ToY) is reached, a supply or production order is created to replenish the buffer to its top, called "Top of Green" (ToG). This is the definition of a reorder point method, which is applied in various industries.

The problem is it can launch many simultaneous orders if several buffers are in need, overloading the workshop, leading to an increase of flow times and a drop in the service rate. In a reorder point method, the lot size is either fixed, determined by using the economic order quantity formula for example, or variable, considering the real demand (which is what the DDMRP does). This may be interesting for supply orders, where capacity is outsourced, but not for production orders where lot size will affect the load into the workshop. In literature, a Reorder point (R) with a fixed Quantity (Q) is called a (R, Q) model and a Reorder point (R) with a maximum Stock (S) to replenish to the top is called a (R, S) model. Thus, the DDMRP is a (ToY, ToG) model.

The Constant Work-In-Process (ConWIP) is a pull control system using a single card type called ticket to control the amount of WIP permitted in the entire loop (Spearman, Woodruff, and Hopp, 1990). The method is based on Little's Law (Little, 1961), and it aims to control the total amount of work in the workshop by keeping it constant.

$$Work\ In\ Process = Throughput \times Cycle\ Time$$

Therefore, by controlling WIP, we can control cycle time in the ConWIP loop.

This paper proposes a DDMRP-ConWIP Integrated model, where a production order is released only if both a ConWIP ticket is available and a buffer needs to be replenished. This model is different from a classic DDMRP-ConWIP model where DDMRP would create an order (date, quantity, and priority) and ConWIP would simply allow it in the workshop because in our DDMRP-ConWIP Integrated model, the quantity to release and the order priority are decided at the last

moment (when a ticket comes back and so the order enters in the workshop). Thereby, our model considers both demand variation and production variation when creating an order. It allows skipping the "solving capacity problem phase" where most of the solutions are the utilization of alternatives resources, the adjustment of available capacity, or the summarization of lots to reduce the number of setups (Taal and Wortmann, 1997). This phase is time-consuming, and it might be expensive to bring capacity. Our model automatically adapts the lot size according to the state of the workshop, to be more robust (Brandon-Jones et al., 2014).

Why couple a ConWIP loop with DDMRP and not another method? Because DDMRP is a (R, S) method, which is an aperiodic method with variable quantity, and so is a ConWIP loop generating order production orders: the time between two production orders depends on the return of a ticket (and a need to produce), and the quantity launch into production might change between two orders from the same buffer as it depends on its consumption. Therefore, a periodic or a fixed quantity method is not as relevant as the DDMRP method. For the rest of the article, we call our DDMRP-ConWIP Integrated system a (ToY-ConWIP, ToG) model because it requires both reaching the ToY threshold and that a ConWIP be available to trigger a production order, and the buffer is then replenished to its top, the ToG.

In this paper, we seek to verify three hypotheses:

1) Without perturbation coming from production (only demand variation), the three reorder point methods (R, Q), (ToY, ToG) and (ToY-ConWIP, ToG) behave in the same way as the three models consider demand variation;

2) With perturbations coming from both demand and production, our (ToY-ConWIP, ToG) method outperforms the others as it considers both demand variation and production variation; and

3) With our (ToY-ConWIP, ToG) model, the flow time of production orders is reduced.

The paper is organized as follows. Section 2 is dedicated to the review of literature on publications related to the topic. Then, Section 3 describes the research methodology (workshop studied, inventory models' parameters, and experimental design). Section 4 presents the results, and finally, Section 5 concludes and proposes avenues for further research.

## 6.2 Literature Review

The literature review is divided into three subsections: the first is dedicated to the inventory model called reorder point, the next one is about DDMRP, and the last one is for relevant publications on ConWIP.

### 6.2.1 The reorder point method

The main idea of the reorder point method is to trigger a replenishment order (supply or production) whenever the inventory position drops to or below a threshold called "reorder point". One of the first papers on reorder point method was about minimizing inventory, setup and backorder costs by proposing an algorithm to find the optimal lot-size and reorder point (Hadley and Whitin, 1963). One of the main differences among research papers on reorder point is whether the lot size is fixed (often noted Q for Quantity) or variable (noted S, as we replenish to the maximum stock level S). Moreover, the Reorder point is noted R: we have (R, Q) models and (R, S) models.

The publications on (R, Q) models focus on finding the best parameters R and Q that reduce the total cost (Hadley and Whitin, 1963): for example, De Bodt and Graves (1985) study a multi-echelon case with stochastic demand, Nahmias and Wang (1979) specialize in decaying inventories, Chang, Yao, and Lee (1998) and Kao and Hsu (2002) study inventory models with fuzzy demand, etc. The quantity Q was originally fixed, but Axsäter (2005) introduces a (R, *n*Q) model where the replenishment quantity equals an integer multiple (*n*) of the fixed lot-size. His approach enables a setup cost model for integer multiple lot sizes. Jodlbauer and Dehmer (2020) improve this model using demand spike information.

The other widely studied reorder point model is the (R, S) model, where the idea is to replenish the inventory to its maximum stock level S when the threshold R is reached. In that case, Q = S – R is variable. In the literature, the main objective is to find the best parameters R and S depending on the situation: Veinott Jr and Wagner (1965) use a computational approach for finding optimal (R, S) policies, Liu (1990) study the case of Poisson demand and exponential distribution of product life, Liu and Lian (1999) focus on perishable inventory system with a general renewal demand process and instantaneous replenishments, etc. More publications are presented in the literature review of Khanlarzade et al. (2014) about inventory control with deteriorating items and the state

of the art of Chaudhary, Kulshrestha, and Routroy (2018) on inventory models for perishable products.

Recently, Jodlbauer and Dehmer (2020) improved the reorder point method using advance demand spike information which decreases the inventory position. This is exactly what DDMRP does, making it a make-to-stock model with no spike, and a make-to-order model when demand is greater than the spike threshold. By definition, the DDMRP production system is a (R, S) model where R is called ToY and S is ToG: DDMRP is a (Toy, ToG) reorder point model that considers demand spikes in its inventory position.

Most of the papers consider that the replenishment times are variable (often called lead times but which are instead flow times), but the cause of variability is external to the system: there is no relation between the lot size of an order and its flow time. This may be a valid assumption with supply orders where capacity is outsourced, but not with production orders where lot size is linked to the loading rate, and loading rate is linked to flow time (Kingman, 1962).

### 6.2.2 DDMRP

The DDMRP is a more and more well-known material management method, both in the industrial world where many companies have implemented it, and in the academic world with an increasing number of research articles dedicated to it (Bahu, Bironneau, and Hovelaque, 2019). Two main research topics are found in these publications.

The first one is the comparison to other traditional methods, such as MRPII and Kanban, demonstrating the relevance of the DDMRP: a better compromise between stock level and service rate, peak demand anticipation, dynamic adjustment of buffer sizing, and an ability to work in a highly variable environment (Ihme and Stratton, 2015; Miclo et al., 2016a; Shofa and Widyarto, 2017; Miclo et al., 2018; Thürer, Fernandes, and Stevenson, 2020).

The second topic is the study and improvement of the method itself. For example, Vidal et al. (2020) propose an aggregate approach to focus on strategic perspective of the DDMRP, Martin et al. (2019) develop a decision tree to adjust buffer parameters, Dessevre, Martin, Baptiste, Lamothe, Pellerin, et al. (2019) put under control the parameter DLT to improve performances, Lee and Rim (2019) propose an alternative model for the safety stock calculation, and recently Achergui,

Allaoui, and Hsu (2020) develop an algorithm to solve the optimization problem of minimizing storing costs for uncapacitated buffer positioning.

Although studies on DDMRP are both axiomatic and empirical today (Bagni et al., 2021), many issues remain to be tackle scientifically, especially questioning from industrial sectors (Dessevre et al., 2020), complex environments (Velasco Acosta, Mascle, and Baptiste, 2019) and its implementation process (Orue, Lizarralde, and Kortabarria, 2020). More details are available in the systematic review of Azzamouri et al. (2021).

### 6.2.3 ConWIP

According to Jaegler et al. (2018), ConWIP research is divided into four fields.

The first one is sizing characteristics, where there are two main characteristics: number of tickets and lot sizing (Spearman, Woodruff, and Hopp, 1990) and (Hopp and Roof, 1998). These parameters determine the targeted average cycle time or throughput: as the number of tickets determines the amount of WIP in the ConWIP loop, defining the number of tickets is a compromise between throughput, lowering cost, and cycle time. Two approaches are used in the literature to determine these parameters: static card count calculations (Hopp and Spearman, 1996) and (Marek, Elkins, and Smith, 2001), and adaptive WIP level methodology (Hopp and Roof, 1998), (Belisário and Pierreval, 2015), and (Tardif and Maaseidvaag, 2001).

The second field of research is the implementation environment of ConWIP loops. As Stevenson, Hendry, and Kingsman (2005) explain, the choice of a Production Control System (PCS) depends on the demand context and workshop configuration. Make-to-Stock demand and flow shops are the most studied environment on ConWIP.

The third field is the comparison of ConWIP to other PCS. Comparing two methods in several environments is an unavoidable field in research. For example, ConWIP has been compared to push methods (Bahaji and Kuhl, 2008) and (Lavoie, Gharbi, and Kenne, 2010), Kanban (Marek, Elkins, and Smith, 2001), Modified ConWIP (Takahashi and Nakamura, 2002), and (Prakash and Chin, 2015), Base-Stock (Khojasteh, 2015), Polca (Harrod and Kanet, 2013), Cobacabana (Land, 2009), etc. In conclusion, the ConWIP loop is not always the best system, especially in complex environments, but is easier to implement and maintain.

The fourth and last field is the research approach used to study ConWIP, where simulation is the most frequently used. More details on ConWIP publications in the literature reviews of Framinan, González, and Ruiz-Usano (2003) and Jaegler et al. (2018).

Finally, we add to these fields of research studies on coupling ConWIP with other methods and modified ConWIP systems. For example, Bonvik, Couch, and Gershwin (1997) propose a hybrid Kanban/ConWIP system that reduces inventories and Leonardo et al. (2017) analyze a case study implementing that hybrid system; Takahashi and Hirotani (2005) compare ConWIP and synchronized ConWIP in an assembly line, where orders for each stage are released while adjusting the difference in the lead time and they are synchronized at the assembly stage, allowing to reduce inventories; and finally Onyeocha et al. (2015) compared a Hybrid Kanban (HK) ConWIP system and a Base Stock Kanban (BK) ConWIP system while Al-Hawari, Qasem, and Smadi (2018) compared a Base Stock ConWIP model and a Base Stock Kanban ConWIP model to find that BK-ConWIP outperforms HK-ConWIP and Base Stock ConWIP is better for the service rate while BK-ConWIP is better for the WIP level. The major difference between all these methods is the feedback of customer demand: when and where do we get this information? To all the buffered stages (BS-ConWIP)? When a Kanban/ConWIP ticket is seized (BK-ConWIP)? Or when a Kanban/ConWIP ticket is released? Etc. All these methods use customer demand information to release production orders, but they do not use workshop information to generate those orders. More details in the literature review of Prakash and Chin (2015) about fifteen modified ConWIP systems and the introductory overview of González-r, Framinan, and Pierreval (2012) about token-based pull production control systems.

As a conclusion of this literature review, it is known that (i) DDMRP is a recent and promising material management method that can be still improved, especially to address capacity issues; that (ii) the DDMRP production system is a (R, S) reorder point where R = ToY and S = ToG; that (iii) papers about reorder point do not link lot size and flow time (acceptable for supply orders but not for production orders); and that (iv) a DDMRP-ConWIP Integrated model that considers both demand and production variations to adapt the lot sizes to be more robust has never been studied. Thus, our research question arises: Can we make the production order system of the DDMRP method more resilient by integrating a ConWIP loop? To answer this question and validate (or not) our three hypotheses, we compare a (R, Q) model, a (ToY, ToG) model, and our (ToY-ConWIP, ToG) model in several scenarios with different demand and production variabilities.

## 6.3 Methodology

This paper is based on a discrete events simulation, as it easily allows to model and compare different scenarios with great variability (Mourtzis, 2020). Furthermore, the objective is not to optimize but to observe a phenomenon. Then, the mains steps are (i) to model the workshop; (ii) to configure the parameters of each inventory model; (iii) to determine a design of experiments and the simulation parameters; and finally (iv) to analyse the results. These steps are detailed in this section.

### 6.3.1 The workshop parameters

The workshop studied is a production line composed of three stages (one machine per stage), where the first stage is subject to breakdown (explained in the design of experiments) and the second stage is the bottleneck. We decided to study a flow shop because it is the most common workshop found in the industry, especially between two DDMRP stock buffers. Thirty products are manufactured in the workshop. The raw materials are buffered and considered infinite (always available in needed quantity). Setup times (in minutes) and run times (in seconds per piece) presented in Tableau 6.1 were determined to satisfy a utilization rate of 90% of the bottleneck station. These times are triangularly distributed to model production variability.

Tableau 6.1 : Setup times and run times for each stage.

| | **Stage 1** | **Stage 2** | **Stage 3** |
|---|---|---|---|
| **Setup times (minutes)** | TRIA(14.25, 19, 23.75) | TRIA(18, 24, 30) | TRIA(15.75, 21, 26.25) |
| **Run times (seconds per piece)** | TRIA(17.25, 23, 28.75) | TRIA(21.75, 29, 36.25) | TRIA(19.5, 26, 32.5) |

For example, a production order with a lot size of 100 pieces will wait around 24 minutes for the setup of the machine in stage 2, and it will last about 100x29/60 = 48.3 minutes to manufacture it.

The workshop is open 8 hours a day, 5 days a week, and manages partial orders: when a customer order arrives, it is delivered in full if possible. Otherwise, it enters a queue and is prioritized when the stock of the product in question is available again.

### 6.3.2 The inventory models' parameters

In this subsection, we determine the parameters for each inventory method and illustrate how they behave when facing the same customer demand. For the illustration (Figure 6.1), we have only one product with an initial stock of 90 pieces.

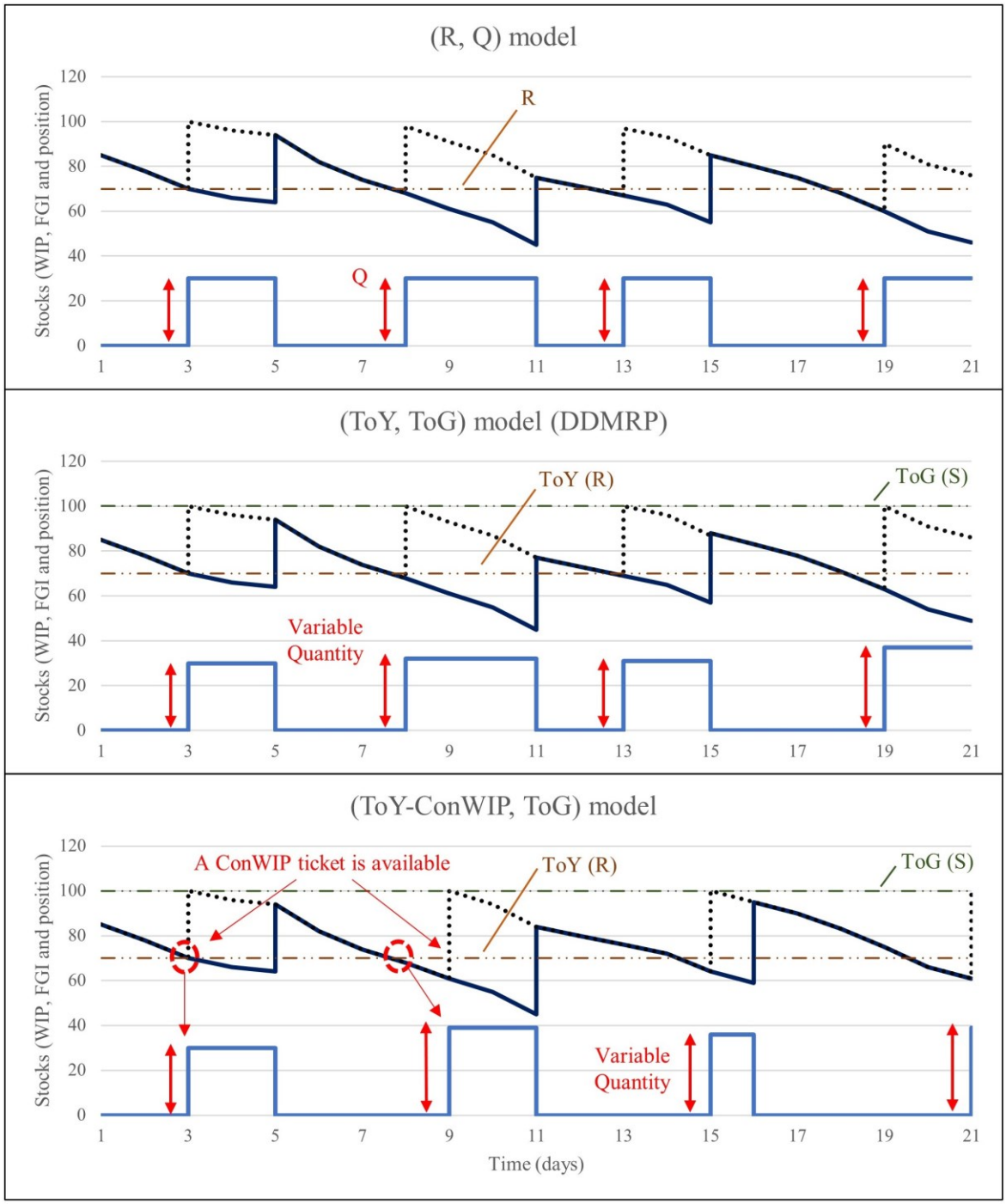


Figure 6.1 : Illustration of the three inventory models and their parameters.

The light blue curve represents the Work-In-Process (WIP), the dark blue curve is the Final Good Inventory (FGI), and the dotted curve is the inventory position. The dashed lines represent the reorder point threshold (R or ToY) and the top of the buffer (ToG).

#### 6.3.2.1 The Reorder Point model (R, Q)

The two parameters to determine for this model are the reorder point threshold R and the lot size Q. In the illustration, Q is fixed at 30 pieces and R is set at 70 pieces (Figure 6.1): every time the stock position falls below 70 pieces, a production order with a lot size of 30 pieces is released.

The main advantage of this method is the lot size is known in advance. Therefore, the supply of raw materials is easier, and the load launched into the workshop is known in advance too. In our model, the lot size Q is fixed at 100 pieces and the reorder point threshold R is set at 140 pieces.

#### 6.3.2.2 The DDMRP model (ToY, ToG)

We must determine the reorder point threshold (ToY) and the maximum inventory (ToG) for this system. In the illustration, ToY is fixed at 70 pieces and ToG is set to 100 pieces (Figure 6.1): every time the stock position falls below 70 pieces, a production order with a lot size to replenish the buffer to its top (ToG) is released.

The main advantage of this method is to consider demand variation. If customer demand is bigger than usual, the stock position will fall lower, and therefore the lot size will be larger than the minimum. The lot size adapts to demand. Unfortunately, it might be more challenging to forecast the raw materials and the load release into the workshop. In our model, the ToY and ToG are respectively set to 140 pieces and 240 pieces. Thus, the threshold R from the (R, Q) model is equal to the ToY in the (ToY, ToG) model, and the minimum lot size for the (ToY, ToG) model correspond to Q, ensuring a fair comparison.

#### 6.3.2.3 The DDMRP-ConWIP Integrated model (ToY-ConWIP, ToG)

The DDMRP-ConWIP Integrated model we propose is a coupling between the DDMRP method and a ConWIP loop. They are both interweave as the ConWIP loop helps to generate the Productions Orders (PO). In a classic ConWIP system, the release sequence of jobs is given by a release list (generated by a MRP system for example), with actual releases occurring only when authorized by a ConWIP ticket. There is no release list in our DDMRP-ConWIP Integrated model:

we generate a PO only when a ConWIP ticket comes back from the workshop, and the lot size is then defined “as late as possible”.

There are three parameters to define for this system: the reorder point threshold (ToY), the maximum inventory (ToG), and the number of ConWIP tickets. In the illustration, ToY and ToG are the same as for the (ToY, ToG) method, i.e. 70 and 100. The only difference is now the PO are launched only if a ConWIP ticket comes back: for example, the buffer needs a replenishment at time t = 3, and as a ticket is available, the PO is launched. The buffer needs another replenishment at time t = 8, but no ticket is available until t = 9. That is why the PO is launched one day after the need (and therefore, the lot size is slightly larger as customers continue to consume the buffer).

To determine the number of ConWIP tickets, we create a chart by simulating the model with 1 ticket, 2 tickets, 3 tickets, etc. We represent both the throughput and the cycle time depending on the number of tickets (Dumoulinneuf et al., 2020). The idea is to find the optimal number of tickets which allows having a high throughput while ensuring low cycle time.

The main advantage of this method is it considers both demand variation and production variation, which is additional information that the other method ignores. The return of a ConWIP ticket gives this information:

- If the workshop is doing well (without perturbation), the tickets come back frequently, and our model launch a lot of PO with small lot size (as a classic DDMRP method would do); and
- If there are disturbances in the workshop (which could compromise flow times of the PO, such as a breakdown, for example), the tickets come back more slowly, and our system launch a few POs with larger lot size. This saves time on setup times, and therefore we decongest the workshop.

To better understand the difference between the classic DDMRP method and our DDMRP-ConWIP Integrated system, Figure 6.2 illustrates the main steps to launch production orders. Finally, when two buffers need to be replenished but only one ticket is available, we prioritize the buffer with the highest priority using the calculation of the DDMRP method: the ratio “inventory position / ToY” (Ptak and Smith, 2016). Thus, the more a buffer has been consumed, the more it has priority.

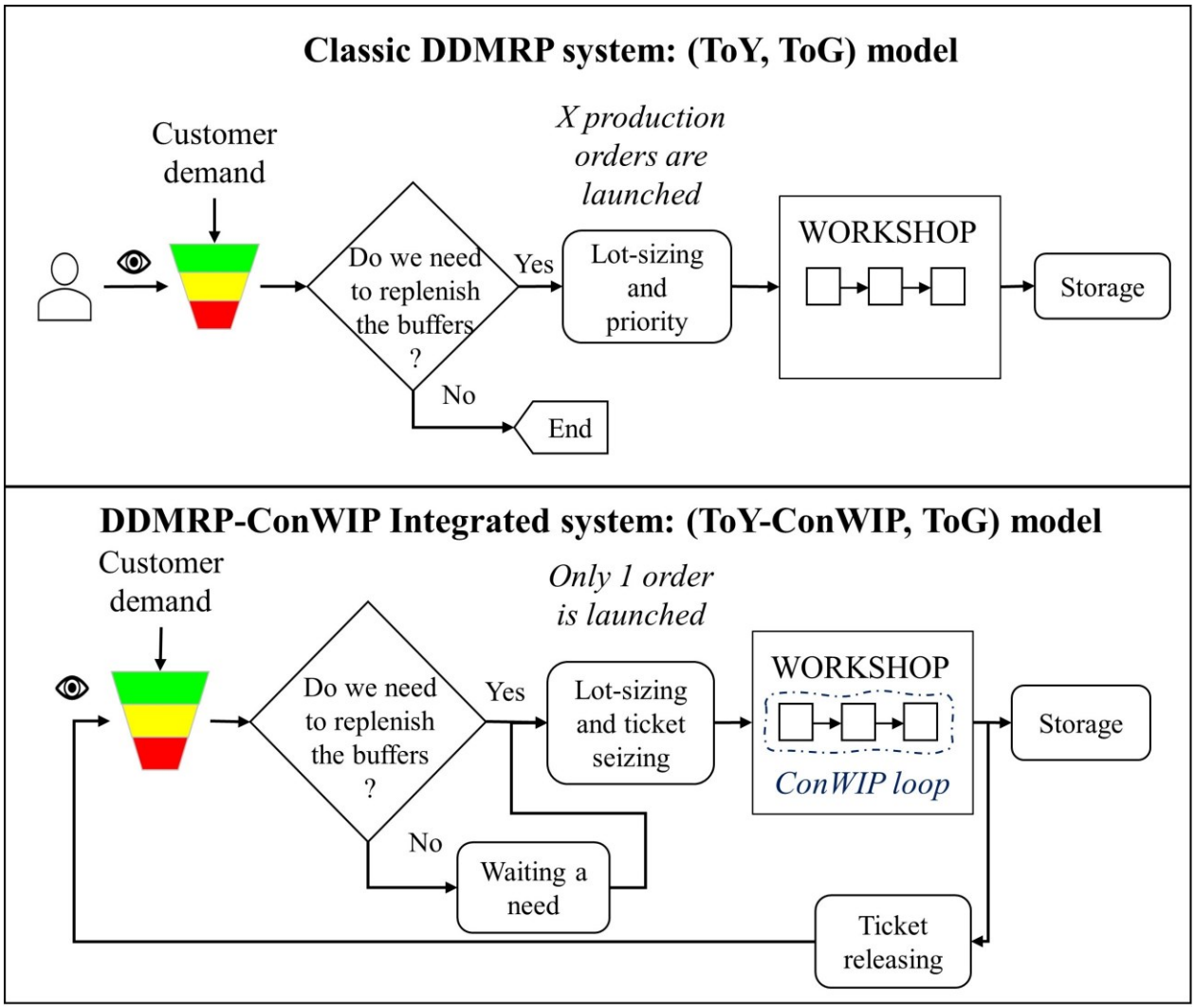


Figure 6.2 : Production order release steps for (ToY, ToG) model and (ToY-ConWIP, ToG) model.

### 6.3.3 The design of experiments

The Design of Experiment (DoE) is used to validate our three hypotheses. Consequently, the DoE has three dimensions:

1) The three reorder point methods: (R, Q), (ToY, ToG), and (ToY-ConWIP, ToG);
2) Three different demand signals: a stable one, a variable one, and an erratic one; and
3) Three different failure times: 0%, 10% and 20%.

The first dimension has been detailed previously; the two last are detailed in the following sub-sections and the outputs. The DoE is made of 3x3x3 = 27 scenarios. Each scenario is made of 100

replications of a simulation that lasts 200 weeks. Modeling and simulations are carried out with the software Arena, version 16.10.

#### 6.3.3.1 The three demand signals

To verify the three methods behave the same way with different variable demand, we model three demand signals with the same average number of products to deliver per day (to assure a loading rate of around 90% for the bottleneck). The demand signal is the same for all products:

- A stable demand, where the daily number of products is uniformly distributed between 16 and 24 ($\sigma \approx 2.6$);
- A variable demand, where the daily number of products is uniformly distributed between 10 and 30 ($\sigma \approx 6.1$); and
- An erratic demand, where the daily number of products has 60% chance to be 0, 10% to be 20, 10% to be 40, 10% to be 60 and 10% to be 80 ($\sigma \approx 28.2$).

The three demand signals are presented in Figure 6.3 with an example of 20 days for one of the thirty products.

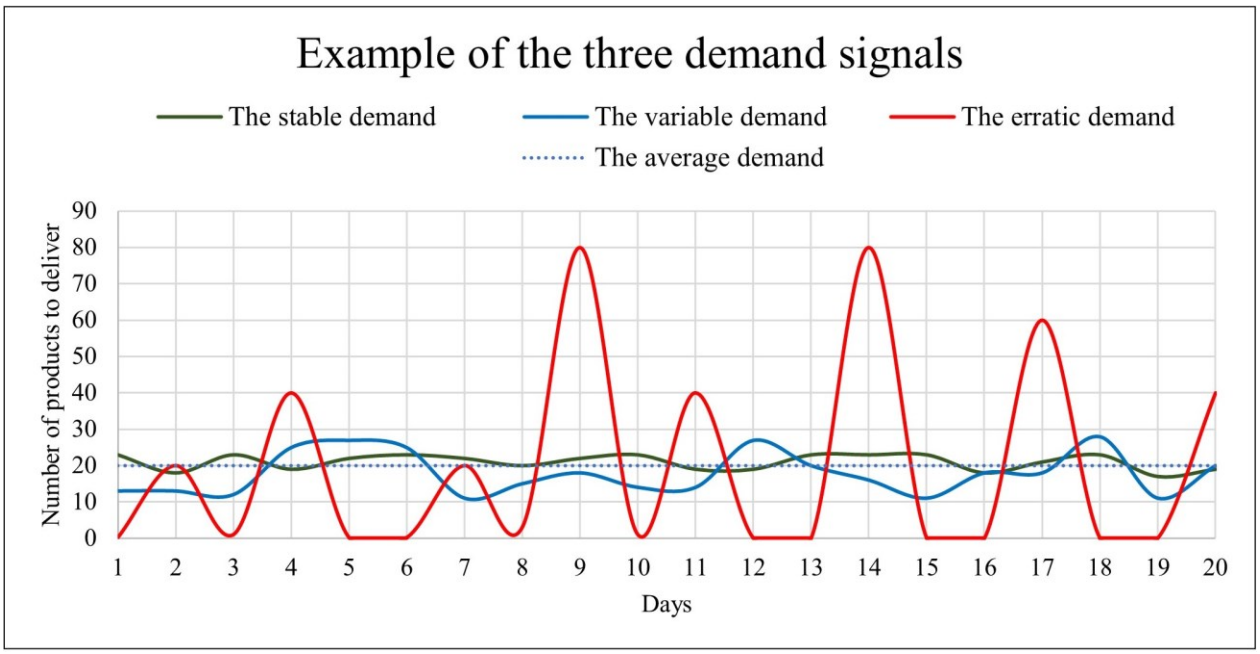


Figure 6.3 : Example of the three demand signals for 20 days.

#### 6.3.3.2 The three failure times

To verify our model outperforms the two others because it considers production variability (breakdown or unavailability), we model three failure times for the first stage in the workshop:

1) The first stage is never down, so the failure state represents 0% of the time;
2) The first stage is subject to breakdowns with a Mean Time Between Failure (MTBF) of 36 hours and a Mean Time To Repair (MTTR) of 4 hours. These two values are both exponentially distributed, representing a failure state of 10% of the time; and
3) The first stage is subject to breakdowns with a MTBF of 32 hours and a MTTR of 8 hours, both exponentially distributed, representing a failure state of20% of the time.

#### 6.3.3.3 The outputs of the DoE

The outputs are the variables we analyze to compare the scenarios and to understand what happened. They are the following:

- The Workshop Service Rate. We define it as the probability an order has its flow time less than or equal to the Lead Time (LT), which is a control variable (Hopp and Spearman, 1996):

$$Workshop\ Service\ Rate = P(Flow\ Time \leq LT)$$

  In our study, the LT of the workshop is equal to a week (5 days). In the DDMRP, the LT between two buffers is called Decoupled LT (DLT). So, for the (ToY, ToG) and (ToY-ConWIP, ToG) models, the Workshop Service Rate is the probability an order has a flow time less than or equal to the DLT;

- The Customer Service Rate;
- The number of PO in production;
- The flow times of PO and their distributions. We define the flow time as the time between recognition of the need for an order (which is when the inventory position reaches the reorder point threshold R or ToY) and the receipt of goods;
- The lot size. For the (R, Q) model, it is fixed at Q pieces. For the two other models, it is at least Q, and it is variable as we replenish the buffers to their ToG; and

- The loading rate of the bottleneck. The target is 90%, but it may differ as the lot size varies for the two last models.

## 6.4 Results

### 6.4.1 Determination of the number of ConWIP

The number of ConWIP tickets in the loop depends on the failure time: the longer the downtime, the more ConWIP tickets are needed to avoid starving the bottleneck (and therefore loosing throughput).

To determine the number of ConWIP tickets, we represent both the throughput and the cycle time (the time spent by the POs inside the ConWIP loop) depending on the number of tickets. Here is an example in Figure 6.4. We determined we need 3 tickets for the scenarios without breakdown, 5 tickets for the scenarios with 10% failure time, and 7 tickets for the scenarios with 20% failure time.

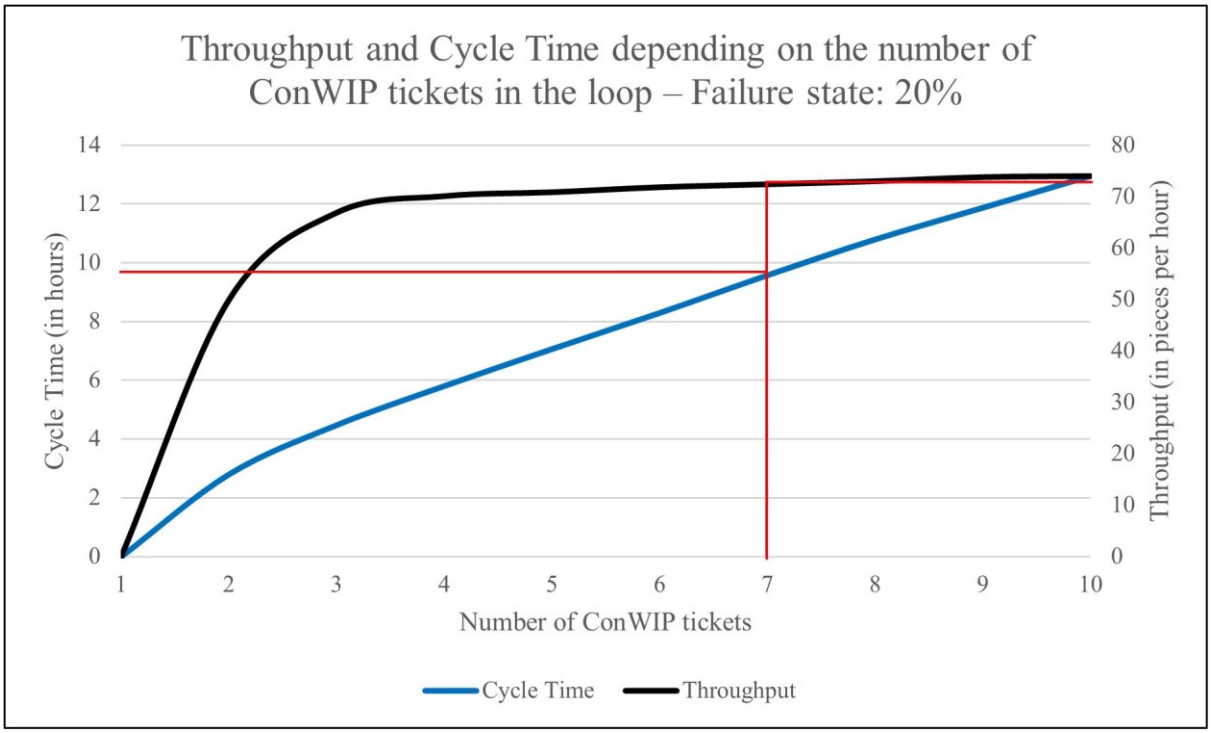


Figure 6.4 : Throughput and Cycle Time depending on the number of ConWIP tickets.

### 6.4.2 Customer Service Rate and Work-In-Process

The results of the DoE are presented in Figure 6.5, where there are the Customer Service Rate and Work-In-Process depending on the inventory model, the demand signal, and the failure time. The results are represented by square for the (R, Q) model, by triangles for the (ToY, ToG) model, and by circles for our (ToY-ConWIP, ToG) model. The colors represent the "failure time" dimension of the DoE: yellow for 0%, orange for 10%, and red for 20%.

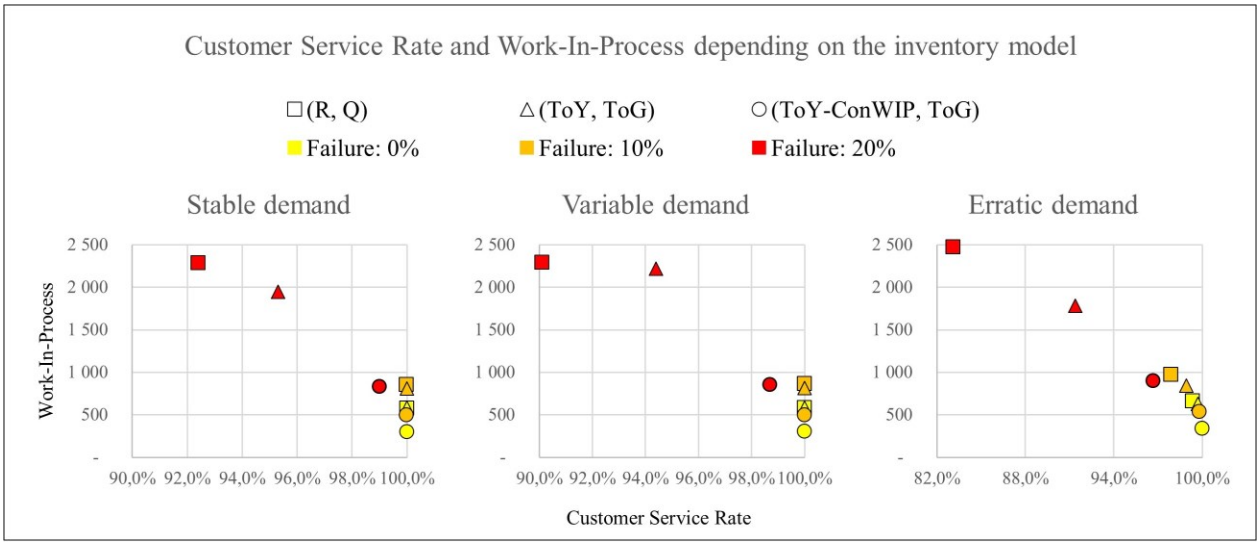


Figure 6.5 : Customer Service Rate and Work-In-Process depending on the inventory model, the demand signal, and the failure time.

We observe the same phenomenon through the three demand signals:

- Without breakdown (in yellow), the (R, Q) and (ToY, ToG) model are similar in terms of WIP (around 580 pieces) and customer service rate. This latter is at 100% for the stable and variable demand signal and drops to 99% for the erratic demand. The (ToY-ConWIP, ToG) model reduces WIP in all three cases (around 300 pieces) while ensuring a customer service rate of 100%;
- When the failure time rises to 10% (in orange), the (R, Q) model stands out from the two others with a customer service rate that drops to 98%. The (ToY, ToG) model follows it with a customer service rate of 99%. The (ToY-ConWIP, ToG) model is still at 100%, with less WIP than the others; and

- When the failure time reaches 20% (in red), the (R, Q) model can no longer follow and its customer service rate drops to 92%, 90% and 83% while the WIP reaches more than 2000 pieces. The (ToY, ToG) model performs a little better with a customer service rate at 95%, 94%, and 91% but the WIP reaches 2000 pieces. Finally, the (ToY-ConWIP, ToG) model outperforms, with a customer service rate at 99%, 99%, and 97% while ensuring a WIP below 1000 pieces.

### 6.4.3 Analysis of the observed phenomenon

To understand what happens in the workshop, we observe the number of POs in the workshop, their flow times and lot sizes, and the consequences on the Workshop Service Rate and Customer Service Rate over time (representing by the number of PO having a flow time above the allotted time and the number of shortages). We only represent the two best models (Figure 6.6). The light blue curves represent the classic DDMRP system (ToY, ToG), and the dark blue curves represent our DDMRP-ConWIP Integrated system (ToY-ConWIP, ToG). The red circles define a numbered causal path: for the (ToY, ToG) model, when the number of PO in the workshop increases (1), the flow times of PO naturally increase too (2) as a consequence of the Little's Law (Little 1961). Therefore, the number of PO above the Decoupled Lead Time increases too (3), causing shortages for the customers (4) because product buffers are replenished late. On the other side, for our (ToY-ConWIP, ToG) model, the number of PO in the workshop never exceeds the fixed number of tickets (7 tickets are allowed in the loop in this example), and the flow times of PO are on average lower. When there are disturbances in the workshop (caused by breakdowns, for example), the flow times increase (5), and the tickets come back more slowly. Therefore the next POs have a bigger lot size than usual (6). The POs' lot size adapts to the state of the workshop, occasionally doubling from the minimum lot size: the lot sizes vary between 100 and 120 with the (ToY, ToG) model (by adapting to variations in demand), and between 100 and 250 with our model (by adapting to variations in demand and production). However, this model is not without risk: When POs wait too long before being launched and lot sizes increase (7), flow times increase too and may exceed the allotted time (8), causing shortages for the customers (9). Nevertheless, the number of shortages is divided by four between the (ToY, ToG) model (4) and our (ToY-ConWIP, ToG) model (9), explaining the differences in the Customer Service Rate.

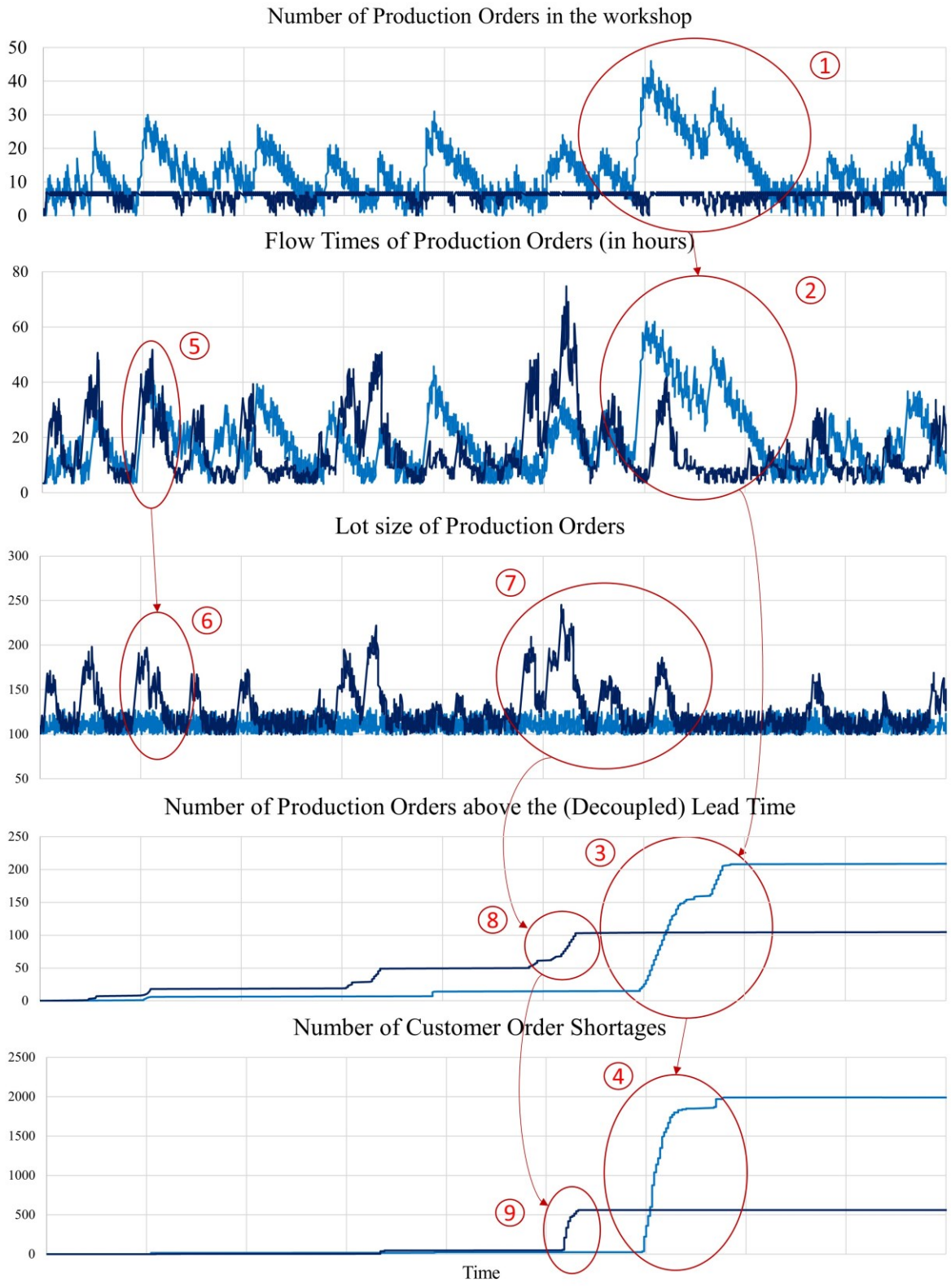


Figure 6.6 : Number of PO in the workshop, Flow Times, Lot sizes, number of PO above the DLT, and number of Customer Order Shortages over time.

### 6.4.4 To go further: the flow times' distributions

To complete our analysis, we compare the flow times' distributions for each inventory model in the scenario with variable demand and 20% of failure time (Figure 6.7).

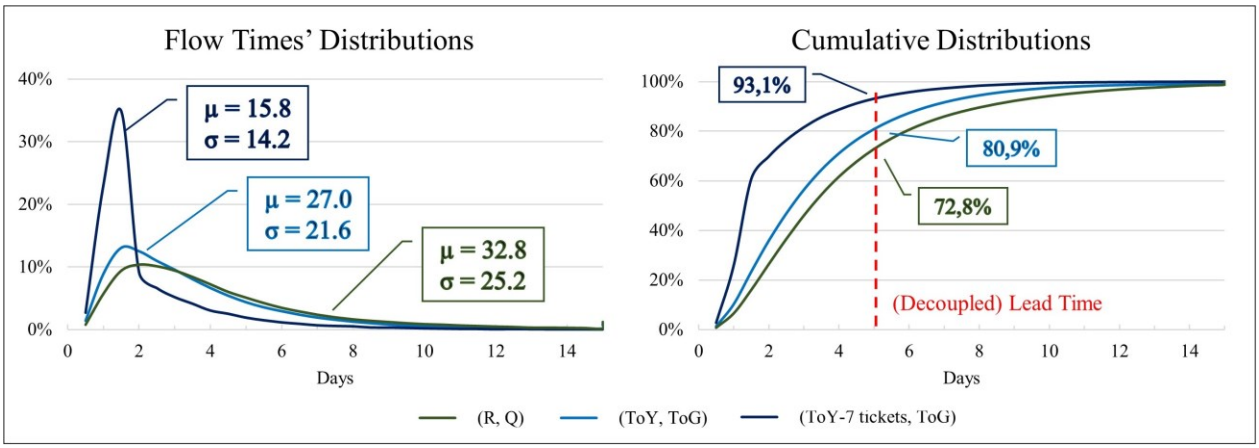


Figure 6.7 : Flow Times' Distributions and Cumulative Distributions.

The dark blue curve represents our (ToY-ConWIP, ToG) model, where the average flow time is 15.8 hours (≈ 2 days) and the standard deviation is 14.2 hours. With the cumulative distribution, we see that 93.1% of the PO has a flow time under 5 days (the allotted time). It corresponds to the Workshop Service Rate. For the (ToY, ToG) model, in light blue, the average flow time is 27.0 hours (≈ 3.4 days), and the standard deviation is 21.6 hours. Because of this dispersion, the Decoupled Lead Time absorbs only 80.9% of the PO (with repercussions on the Customer Service Rate). For the (R, Q) model, in green, the average flow time is 32.8 hours (= 4.1 days), and the standard deviation is 25.2 hours. The Workshop Service Rate drops to 72.8% with this model. The (ToY-ConWIP, ToG) model reduces both the average flow time and its standard deviation, allowing to reduce the allotted time (and so the total stocks).

Finally, as the average lot size varies from a model to another, the loading rate of the bottleneck varies too. Consequently, the loading rate of the second stage is around 90% with the (R, Q) model (which was our target), it is around 87.5% for the (ToY, ToG) model, and about 85% for the (ToY-ConWIP, ToG) model (and the ConWIP tickets are used about 90% of the time). This also explains the difference in flow times between the three models since queueing theory teaches us that the more the bottleneck resource is used, the longer the flow times (Kingman, 1962).

## 6.5 Conclusion and openings

This paper proposes a DDMRP-ConWIP Integrated model that couples the DDMRP production order generation system, close to a reorder point method, with a ConWIP loop generating production orders every time a ticket comes back and at least one buffer need to be replenished. We compared our model with a reorder point model (R, Q) and a classic DDMRP model (ToY, ToG). We found our (ToY-ConWIP, ToG) model outperforms the others in scenarios where different demand and production variabilities were tested. With our model, the WIP is reduced in the workshop, the average flow time and its standard deviation decreased too, and the customer service rate is kept high.

The particularity of the (ToY-ConWIP, ToG) model is the information given by the ConWIP loop: if tickets come back frequently, then there is no perturbation in the system, and we can release a lot of orders with small lot size; but if the tickets are stuck in the workshop, due to production perturbation, then we need to release few production orders with bigger lot size to decongest the workshop.

A potentially negative point is that lot sizes of production orders are less predictable than the classic DDMRP model or a (R, Q) model, so the raw materials must be available. In this study, raw materials were infinite.

In future research, the DDMRP-ConWIP Integrated model we propose could be tested on a real industrial case to approve, or not, our hypotheses and results. More environmental factors must be analysed, like products' families or the share of setup time in production time. In our study, setup time represents about one-third of the production time, and we believe that the greater this share, the more efficient our model (since time savings are made on setup times). Moreover, in our study, we had 30 products and 30 buffers. We believe our model is more relevant when there are more buffers in the system, since the (ToY, ToG) model may launch a lot of production orders simultaneously, leading to an increase of flow times and a decrease of customer service rate, that could be prevent with our (ToY-ConWIP, ToG) model.

Finally, we address the issue where both components and finished products are buffered. In an environment where raw materials are buffered but not the components (i.e., a two-level bill of materials), different rules must be studied and compared with our DDMRP-ConWIP Integrated

model to decide which production ordered must be launched when a ticket comes back, to deal with the intermediate level of the bill of materials.

# CHAPITRE 7 DISCUSSION GÉNÉRALE

> Ah non, mais causer, causer… Non, c'est votre truc à vous ça, hein. Faudrait toujours être en train d'tailler le bout d'gras… (Alexandre Astier, 2006, Livre III, L'attaque nocturne)

Ce chapitre a pour but de discuter de l'ensemble de la thèse en regard des aspects méthodologiques et des résultats en lien avec la revue de la littérature. Il ne reprend pas les discussions et limites de chaque article puisqu'elles ont été présentées dans les chapitres précédents.

## 7.1 Fondement et justification des objectifs de recherche

Dans le premier chapitre du manuscrit, après avoir positionné la méthode DDMRP entre celles à flux poussé et celles à flux tiré, nous nous sommes penchés sur un des paramètres fondamentaux du dimensionnement des stocks tampons : le *Decoupled Lead Time*. Défini comme le temps alloué à un ordre de fabrication pour être complété depuis la reconnaissance de son besoin, le DLT est lié aux taux de service (atelier et client) et aux niveaux de stocks. Le but étant de faire en sorte que les temps de réponse, temps réellement écoulés entre la reconnaissance du besoin et la mise en stock des produits finis, soient inférieurs ou égaux au DLT. Or, les temps de réponse sont incertains, car soumis à des variations internes et externes au système. Notre problématique, mettant en valeur le compromis sur le dimensionnement du DLT (taux de service élevé et faibles niveaux de stocks), était donc la suivante :

**Faut-il maîtriser les temps de réponses ou ajuster le DLT lors d'une variation anticipée ou observée des temps de réponse dans un atelier piloté en DDMRP ?**

Nous avons alors par la suite établi cinq volets de revue de la littérature pour comprendre comment fonctionne le DDMRP, quelles méthodes s'en rapprochent et quelles méthodes pourraient nous être utiles, quels éléments influencent les temps de réponse, et les différentes notions de temps dans le jargon industriel. La revue de la littérature a permis de souligner l'influence du paramètre DLT, l'importance du contrôle des temps de réponse, qui se fait notamment par la maîtrise du taux de charge des ressources goulots, et enfin la pertinence de la méthode ConWIP. C'est ainsi que l'on a défini trois objectifs spécifiques de recherche à accomplir dans cette thèse : (1) ajuster dynamiquement le DLT en fonction des temps de réponse observés dans l'atelier pour l'adapter aux variations de charge ; (2) proposer un outil visuel de gestion des temps de réponse par des

leviers de capacité ; et (3) coupler les méthodes DDMRP et ConWIP pour ajuster les tailles de lot en fonction de l'état de l'atelier de production.

## 7.2 Discussion sur les aspects méthodologiques

Pour accomplir nos objectifs de recherche, nous avons basé notre protocole expérimental sur l'utilisation de la simulation à évènements discrets. Ce choix a été motivé par plusieurs raisons. Premièrement, la simulation est l'un des outils de plus en plus utilisés dans la recherche en génie industriel, comme c'est le cas pour étudier la méthode ConWIP par exemple (Jaegler et al., 2018). L'autre méthode la plus connue en recherche est le développement de modèles mathématiques et leur résolution par des programmes informatiques. Cela dit, la simulation permet de modéliser des environnements complexes à fortes variabilités, comme un cas industriel, qu'il est plus difficile d'intégrer dans un modèle mathématique (Mourtzis, 2020). De plus, la méthode DDMRP est une méthode de gestion récente et peu mature (Azzamouri et al., 2021), c'est pourquoi la simulation est l'outil idéal pour tester et valider différentes hypothèses. Enfin, la simulation est un outil visuel, facile et rapide à maîtriser, permettant de modéliser des systèmes proches de la réalité.

C'est d'ailleurs tout l'intérêt de la recherche, surtout dans le monde industriel, de passer des résultats théoriques aux expérimentations pratiques dans l'industrie. Le choix des ateliers étudiés, des lignes de production dans la plupart des cas, se justifie par le fait qu'il s'agit des ateliers les plus communs dans l'industrie, notamment entre deux stocks tampons DDMRP. C'est aussi pourquoi de nombreuses sources de variabilité sont introduites dans les modèles (les temps de production suivent des lois uniformes ou triangulaires, les temps de panne des lois exponentielles, les signaux de demande sont des combinaisons de plusieurs lois, etc.). C'est ainsi que l'on a vu que certaines formules données par Ptak and Smith (2016), comme le calcul du stock moyen, sont fausses dans la pratique (chapitre 2 sur la définition du DLT).

Le choix d'ajuster le paramètre DLT ou de maîtriser les temps de réponse a été appuyé par la définition du taux de service atelier, inspirée par celle de Hopp and Spearman (1996). Cette définition permet de souligner le compromis entre taux de service élevé, par le contrôle des temps de réponse en-dessous du DLT, et faibles niveaux de stock, par le dimensionnement du DLT. La théorie des files d'attente explique par ailleurs certains éléments influençant la dispersion des temps

de réponse, en particulier le taux de charge des ressources, et c'est pourquoi nous avons proposé d'ajuster la capacité pour maîtriser les temps de réponse.

Enfin, le choix d'utiliser la méthode ConWIP a été motivé par son efficacité et sa simplicité de mise en place, notamment dans des environnements ayant une multitude de produits finis à gérer (Framinan, González, and Ruiz-Usano, 2003; Jaegler et al., 2018). De plus, le couplage des méthodes ConWIP et DDMRP est innovant et pertinent puisque le système de génération des ordres du DDMRP est semblable à une méthode à point de commande, apériodique à quantité variable.

## 7.3 Synthèse des résultats obtenus en lien avec la revue de la littérature

D'un point de vue global, ce projet de recherche a mis en valeur les résultats principaux suivants :

- Dans un atelier, peu importe les temps de production, tant que les ordres utilisent la même ressource goulot et qu'aucune condition de production particulière ne s'applique, leur distribution de temps de réponse sera sensiblement la même. Par conséquent, le temps alloué doit être le même. Dans un contexte DDMRP, cela veut dire que le DLT est unique, et qu'il est dépendant de l'atelier et non des produits ;
- La mise sous contrôle des temps de réponse est plus intéressante que l'ajustement du paramètre DLT pour satisfaire la demande dans les temps. Cette mise sous contrôle passe par la maîtrise du taux de charge de la ressource goulot. Il est plus intéressant de chercher à réduire un peu le taux de charge, plutôt que de doubler le DLT ;
- L'utilisation d'abaques réalisés par simulation est un outil visuel facilement utilisable pour corréler un taux de charge, la distribution des temps de réponse, le temps alloué et les taux de service. Il peut ainsi servir d'outil d'aide à la décision pour choisir le meilleur levier capacitaire dans une situation donnée ; et
- Le couplage des méthodes DDMRP et ConWIP permet de générer les ordres de fabrication au plus tard, prenant en compte l'état de l'atelier. Ce couplage offre de meilleures performances en termes de stock d'en-cours, de temps de réponse, et de taux de service.

Ces résultats viennent de l'accomplissement de nos trois objectifs de recherche, en lien avec la synthèse de la revue de littérature. Nous avons montré l'importance du dimensionnement du

paramètre DLT et les conséquences de son ajustement dynamique, ce qui n'est pas préconisé ni étudié par Ptak and Smith (2016).

De plus, nous avons souligné l'importance de dimensionner ce DLT en fonction de l'atelier, et non des produits. Dans la littérature, et notamment dans Ptak and Smith (2016), aucune recommandation sur le choix du DLT n'est faite (à part le fait de sommer les délais de production sur le chemin critique dans la nomenclature), et les délais de production sont le plus souvent liés aux produits plutôt qu'aux ateliers ou aux ressources utilisées.

Ensuite, nous avons mis l'accent sur la pertinence du taux de charge des ressources, et ses conséquences sur la distribution des temps de réponse et les taux de service client et atelier. Nos travaux soutiennent les résultats de la théorie des files d'attente, et notamment les approximations des temps d'attente lors d'une hausse du taux de charge (Kingman, 1962; Shortle et al., 2018). Nos abaques visuels, qui sont des outils utiles et innovants d'aide à la décision pour choisir les leviers capacitaires, n'ont jamais été publiés dans la littérature, en particulier dans un contexte de pilotage par la méthode DDMRP, bien qu'ils puissent être utilisés dans n'importe quel contexte.

Enfin, le couplage original des méthodes DDMRP et ConWIP pour créer un modèle auto-adaptatif de lissage de charge par ajustement des tailles de lot a permis d'accentuer la pertinence et les performances des deux méthodes étudiées et publiées (Azzamouri et al., 2021; Jaegler et al., 2018), et ainsi d'accomplir notre troisième et dernier objectif de recherche.

En conclusion, l'accomplissement de nos objectifs de recherche et la publication des résultats a permis d'enrichir les études autour du DDMRP et des sujets connexes, tout en proposant des méthodes novatrices pour mieux dimensionner et piloter des ateliers gérés en DDMRP.

## 7.4 Limites des études

La première limite, commune à chaque étude réalisée dans ce projet de recherche, est la typologie des ateliers de production étudiée. Nous avons étudié des ateliers différents (nombre de produits fabriqués, temps de mise en course, temps opératoire, ressources utilisées, etc.), mais tous étaient des lignes de production. Ces ateliers sont caractérisés par une série de machines dont les produits passent les uns après les autres dans le même ordre. Il existe pourtant dans l'industrie d'autres types d'atelier, comme les ateliers multigammes où l'ordre de passage des machines est différent d'un produit à un autre. Nous avons limité nos recherches aux lignes de production car il s'agit du type

d'atelier le plus commun dans l'industrie, et c'est notamment ce que l'on retrouve entre deux stocks tampons DDMRP. De plus, les raisonnements sur les ressources goulots faites sur des ateliers monogammes peuvent s'appliquer à des ateliers multigammes ou à d'autres types d'atelier.

Ensuite, dans toutes nos études, nous avons appliqué la règle de priorisation suivante : lorsque plusieurs ordres de fabrication sont générés en même temps, on les ordonne selon la consommation des stocks tampons des produits finis respectifs. Autrement dit, plus un stock tampon est consommé (en ratio « position de stock / Top du Jaune »), plus son ordre de fabrication est prioritaire. Dans l'atelier, les ordres sont ensuite gérés en PEPS devant chaque machine. Ce choix a été motivé par la règle standard de la méthode DDMRP (Ptak and Smith, 2016). D'autres règles ou l'application de cette règle devant chaque machine pourraient nuancer nos propos. Cela dit, l'ordonnancement des ordres a suivi la même règle lors de comparaisons de méthodes (comme dans le chapitre 6 par exemple), quel que soit le modèle utilisé, pour assurer une comparaison juste.

Enfin, chaque article possède ses propres limites, dont les plus pertinentes sont énumérées ci-dessous.

Nous avons choisi dans le premier article de mettre en place une boucle de régulation pour ajuster dynamiquement le DLT. Le choix de la formule d'ajustement a été arbitrairement un lissage exponentiel dont les paramètres ont été fixés à 5% et 10%. De plus, nous avions à l'époque choisi de paramétrer un DLT par produit, et non un DLT commun à l'atelier, alors que la ressource goulot était partagée par tous les produits.

Dans le second article, nous nous sommes limités au choix du nombre d'équipes à mettre en place pour satisfaire la charge prévisionnelle à produire, alors qu'il existe de nombreux leviers capacitaires (comme faire des heures supplémentaires, par exemple).

Finalement, le couplage des méthodes DDMRP et ConWIP a été étudié sur une nomenclature à un étage. Lors d'une application à des cas plus complexes dont les nomenclatures sont sur plusieurs étages, l'intégration d'une boucle ConWIP peut être plus difficile à mettre en place.

# CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

> Aujourd'hui, y a du dessert. […] Y a des gens qui ont pris la peine de faire un dessert, la moindre des choses c'est de rester pour le manger. Y en a marre de se comporter comme des sagouins avec tout l'monde sous prétexte qu'on a des responsabilités. (Alexandre Astier, Kaamelott Livre I, 2005, La tarte aux myrtilles)

Dans ce projet de recherche, nous nous sommes concentrés sur le paramétrage du *Decoupled Lead Time* et sur la distribution des temps de réponse dans des ateliers pilotés par la méthode DDMRP. Notre objectif principal étant d'améliorer les performances des ateliers de production pilotés par la méthode DDMRP en proposant des outils de maîtrise des temps de réponses ou d'ajustement du paramètre DLT. Le but est donc de s'assurer que les temps de réponse des ordres de fabrication soient le plus souvent inférieurs ou égaux au DLT, pour maintenir les taux de service atelier et client les plus hauts possibles. Conjointement, il est intéressant de dimensionner le DLT aussi faible que possible, pour minimiser les niveaux de stocks. D'où notre problématique sur la maîtrise des temps de réponse ou l'ajustement du DLT. Pour accomplir cet objectif, nous avons étudié différentes approches (la régulation du DLT, l'utilisation de graphiques pour décider des leviers de capacité, la mise en place d'une boucle ConWIP), en utilisant la simulation à évènements discrets.

En réponse à notre problématique, nous pouvons affirmer que dans un atelier de type ligne de production où les produits utilisent la ou les mêmes ressources goulots, un ajustement dynamique du paramètre DLT, tant qu'il est commun à l'atelier et non aux produits, peut servir à améliorer son dimensionnement initial. Il est important de choisir sa méthode d'ajustement ainsi que ses paramètres, la définition du DLT (ne pas prendre la moyenne des temps observés par exemple), de connaître la distribution des temps de réponse et d'identifier les ressources goulots pour surveiller leur taux de charge.

Ensuite, il est plus pertinent de chercher à maîtriser les temps de réponse pour qu'ils respectent le temps alloué, *c.-à-d.* le DLT, plutôt que de continuer à ajuster ce dernier. Pour cela, il faut passer par le contrôle du taux de charge des ressources de l'atelier (machines, opérateurs, etc.). Un taux de charge trop élevé rallonge les temps d'attente, et par conséquent les temps de réponse. L'utilisation d'abaques comme outils d'aide à la décision est une méthode visuelle, facilement réalisable par simulation, et destinée aux responsables de production pour déterminer les meilleures solutions capacitaires à une situation donnée.

Enfin, le couplage des méthodes DDMRP et ConWIP, générant un ordre de fabrication unique lorsqu'un stock tampon est dans le besoin et qu'un ticket ConWIP est disponible en ajustant la taille de lot au dernier moment, rend le système plus robuste. Il permet notamment de réduire les temps de réponse et l'en-cours de production, tout en assurant des taux de service atelier et client plus élevés que le modèle DDMRP classique.

Tous ces résultats ont été présentés, publiés ou soumis à différentes conférences et à des journaux de recherche. En tout, cinq articles de conférences ont été publiés et trois articles de journaux ont été soumis (dont au moins un accepté). L'ensemble des travaux, hormis un article de conférence et un article de journal, est représenté en Figure 8.1.

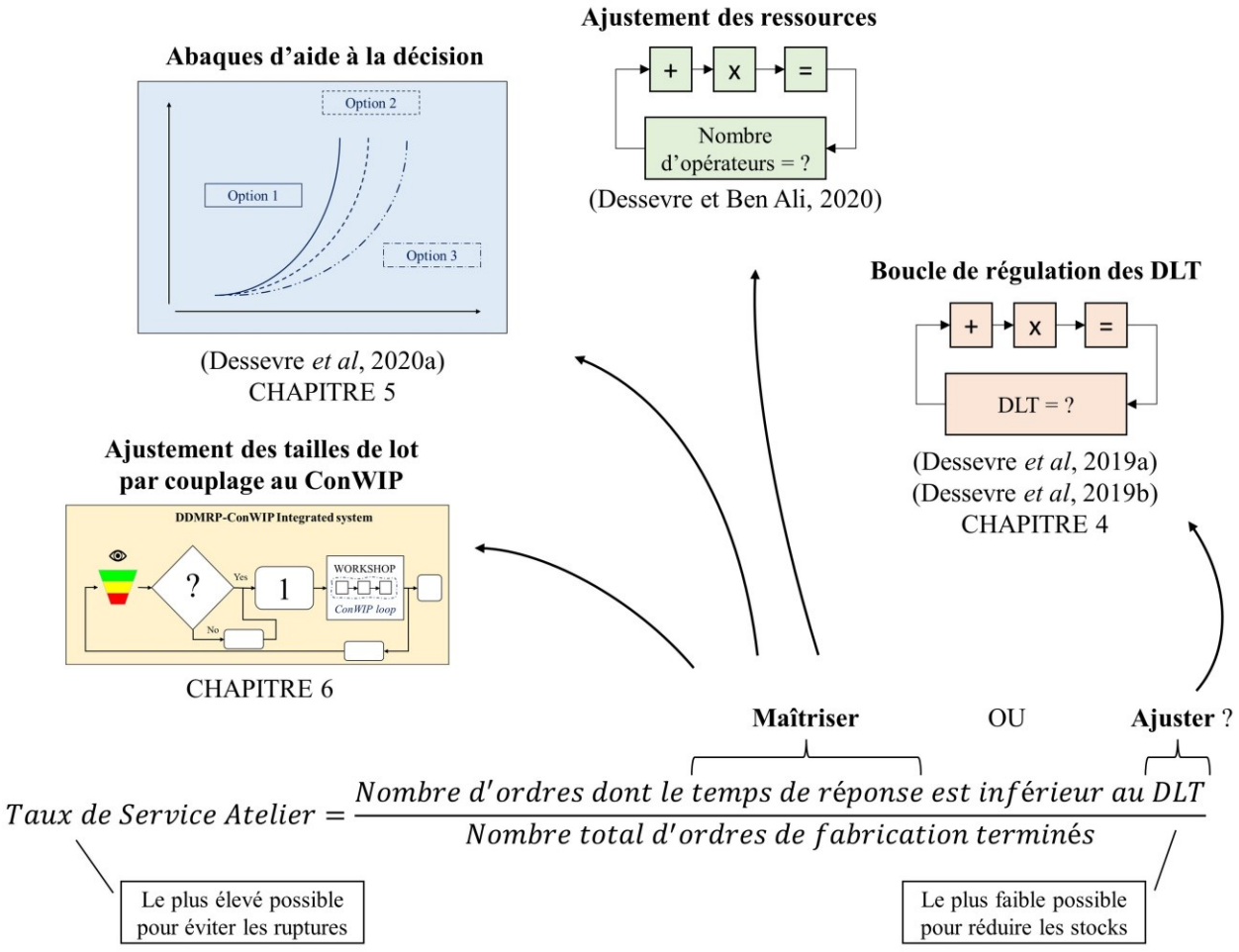


Figure 8.1 : Ensemble des travaux publiés lors de la thèse.

## 8.1 Pistes de recherche et recommandations

Pour les pistes de recherche, chaque article possède ses propres limites et ouvertures à creuser, comme l'ajustement dynamique du DLT en considérant un DLT unique et en prenant en compte la distribution des temps de réponse au lieu de considérer leur moyenne; la création d'abaques par simulation en considérant différents niveaux et leviers de capacité tout en prenant en compte l'aspect financier, non traité dans l'article, pour répondre à des questions comme « est-il préférable de faire deux heures supplémentaires par jour, de passer en 3x8 ou de sous-traiter une partie de la charge ? »; ou encore l'étude du couplage DDMRP-ConWIP sur des nomenclatures à plusieurs étages et l'impact du nombre de tickets ConWIP sur les performances de l'atelier.

Pour ce qui est des recommandations, je pense que l'étude des paramètres de la méthode DDMRP est l'un des sujets les plus importants à attaquer. Lee and Rim (2019) proposent une méthode alternative pour calculer le stock de sécurité des stocks tampons (la zone rouge). En effet, le calcul classique proposé par Ptak and Smith (2016) prend en compte la demande moyenne, le DLT, et les deux paramètres FV et FD (Figure 2.6). Or le FD est utilisé pour définir la taille de lot minimale, mais est-ce judicieux de relier taille de lot et stock de sécurité ? Bien que Ptak and Smith (2016) soumettent l'idée d'utiliser deux FD (un pour la zone verte et un pour la zone rouge), ils préconisent de garder des valeurs proches pour les deux. N'y a-t-il donc pas de meilleures recommandations que celles-ci pour dimensionner le FD ? Les auteurs du DDMRP préconisent de le dimensionner en fonction du DLT (un grand DLT implique un faible FD), mais si tous les produits possèdent le même DLT, quid du FD ? Est-ce que les paramètres et les préconisations actuels de la méthode DDMRP sont pertinents ?

Bien évidemment, l'application sur des cas industriels, théoriques ou pratiques, est l'ultime recommandation possible pour vérifier et entériner nos propos.

Enfin, j'adresse un message à quiconque cherchant un sujet de thèse lié au DDMRP. Lors de la modélisation du cas industriel, utilisé dans le chapitre 5, j'ai été confronté à plusieurs problèmes et questionnements. Parmi toutes les réflexions possibles, celui du calcul de la CMJ (la demande moyenne) et la propagation de la demande client sur tous les stocks tampons de la chaîne logistique, m'a certainement posé le plus de souci. Les questions étant : connaissant le signal de demande client, comment faut-il calculer la CMJ sur les stocks tampons de produits finis dans les filiales proches des clients ? Et pour les stocks tampons du centre de distribution ? Et les composants et

les matières premières ? Est-il plus efficace de propager la demande finale ou de faire des prévisions de demande locales ? Faut-il la décaler du DLT ? Faut-il agréger les signaux pour les composants communs ? Comment propager les pics ? Etc. Bref, de quoi en faire une thèse.

## 8.2 Réflexion personnelle sur les évènements du quotidien

Je termine ce manuscrit sur une réflexion plus personnelle, que j'adresse à tous les lecteurs. Nous avons vu l'importance du taux de charge des ressources dans les files d'attente, que ce soient des ordres de fabrication devant une machine d'un atelier, des clients devant une caisse d'un supermarché, ou encore des véhicules devant le péage d'une autoroute. Et que vous le vouliez ou non, vous êtes également des ressources ayant des tâches à accomplir dans votre quotidien !

Donc la prochaine fois que vous planifiez votre semaine de travail, celle de vos employés, ou encore vos prochaines vacances, gardez toujours en tête la courbe de la théorie des files d'attente en Figure 8.2, et visez un taux de charge autour des 85%. N'hésitez pas à prévoir une demi-journée de libre par semaine pour éviter de faire des heures supplémentaires au bureau, de ramener du travail à la maison, ou de finir en retard.

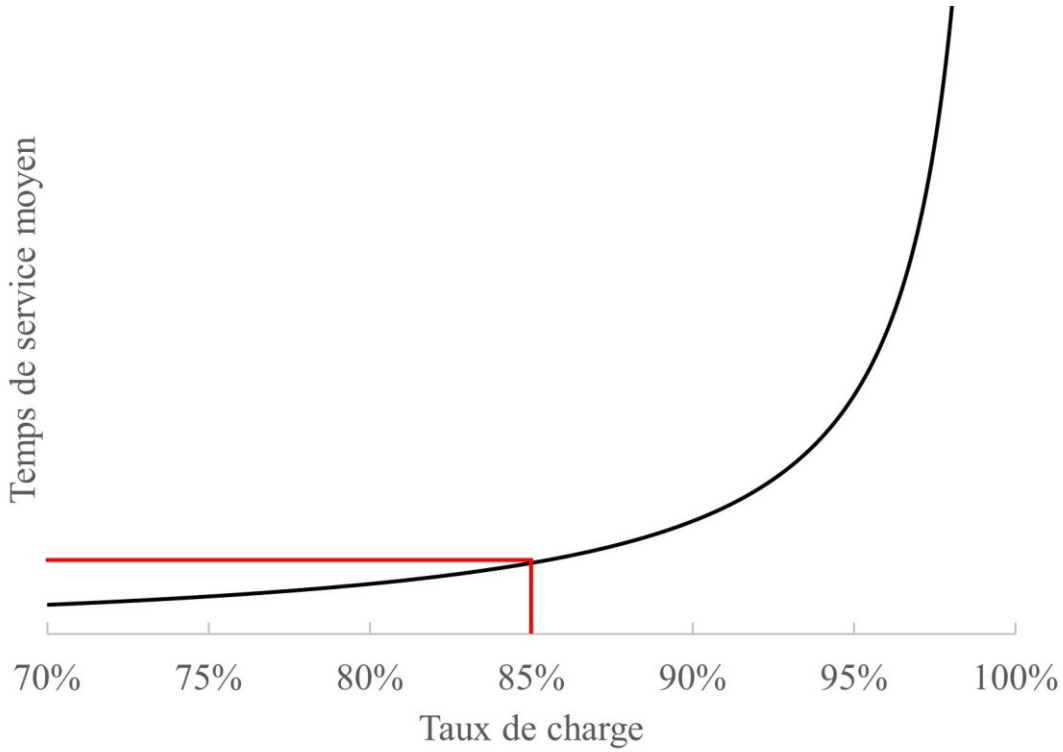


Figure 8.2 : Corrélation entre taux de charge d'une ressource et temps de service moyen.

Car vous n'êtes jamais à l'abri d'évènements imprévus : un bug informatique qui vous prend plusieurs heures à résoudre, ou nouveau courriel urgent à faire passer en priorité, une fausse alarme incendie beaucoup trop longue, un collègue qui propose une pause-café inévitable…

# RÉFÉRENCES


Achergui Abdelhalim, Allaoui Hamid, and Hsu Tiente. 2020. "Strategic DDMRP's Buffer Positioning for hybrid MTO/MTS manufacturing." In *2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 1-8. : IEEE.

Al-Hawari Tarek, Qasem Ahmed Gailan, and Smadi Hazem. 2018. 'Development and evaluation of a Basestock-CONWIP pull production control strategy in balanced assembly systems', *Simulation Modelling Practice and Theory*, 84: 83-105.

Ammar Oussama Ben, Guillaume Romain, and Thierry Caroline. 2016. "MRP parameter evaluation under fuzzy lead times." In *8th IFAC Conference on Manufacturing Modelling, Management, and Control (MIM 2016)*, pp. 1110-15.

Axsäter Sven. 2005. 'A simple decision rule for decentralized two-echelon inventory control', *International Journal of Production Economics*, 93: 53-59.

Azzamouri Ahlam, Baptiste Pierre, Dessevre Guillaume, and Pellerin Robert. 2021. 'Demand Driven Material Requirements Planning (DDMRP): A systematic review and classification', *Journal of Industrial Engineering and Management*, 14: 439-56.

Babulak Eduard, and Wang Ming. 2010. 'Discrete event simulation', *Aitor Goti (Hg.): Discrete Event Simulations. Rijeka, Kroatien: Sciyo*: 1.

Bagni Gustavo, Godinho Filho Moacir, Thürer Matthias, and Stevenson Mark. 2021. 'Systematic review and discussion of production control systems that emerged between 1999 and 2018', *Production Planning & Control*, 32: 511-25.

Bahaji N, and Kuhl ME. 2008. 'A simulation study of new multi-objective composite dispatching rules, CONWIP, and push lot release in semiconductor fabrication', *International Journal of Production Research*, 46: 3801-24.

Bahu Baptiste, Bironneau Laurent, and Hovelaque Vincent. 2019. 'Compréhension du DDMRP et de son adoption: premiers éléments empiriques', *Logistique & Management*, 27: 20-32.

Baptiste Pierre. 2018. "DDMRP: Scheduling opportunities in case of complex BOMs." In *7th International Conference on Information Systems, Logistics and Supply Chain, ILS 2018, July 8, 2018 - July 11, 2018*, 317-25. Lyon, France: INSA Lyon.

Belisário Lorena Silva, Azouz Nesrine, and Pierreval Henri. 2015. "Adaptive ConWIP: analyzing the impact of changing the number of cards." In *Industrial Engineering and Systems Management (IESM), 2015 International Conference on*, 930-37. : IEEE.

Belisário Lorena Silva, and Pierreval Henri. 2015. 'Using genetic programming and simulation to learn how to dynamically adapt the number of cards in reactive pull systems', *Expert Systems with Applications*, 42: 3129-41.

Blackstone John H. 2013. *APICS Dictionary* (APICS: Chicago, IL.).

Bonvik Asbjoern M, Couch CE, and Gershwin Stanley B. 1997. 'A comparison of production-line control mechanisms', *International Journal of Production Research*, 35: 789-804.

Brandon-Jones Emma, Squire Brian, Autry Chad W, and Petersen Kenneth J. 2014. 'A contingent resource-based perspective of supply chain resilience and robustness', *Journal of Supply Chain Management*, 50: 55-73.

Buzacott John A, and Shanthikumar JG. 1994. 'Safety stock versus safety time in MRP controlled production systems', *Management Science*, 40: 1678-89.

Chang San-Chyi, Yao Jing-Shing, and Lee Huey-Ming. 1998. 'Economic reorder point for fuzzy backorder quantity', *European Journal of Operational Research*, 109: 183-202.

Chaudhary Vaibhav, Kulshrestha Rakhee, and Routroy Srikanta. 2018. 'State-of-the-art literature review on inventory models for perishable products', *Journal of Advances in Management Research*.

Christensen William J, Germain Richard N, and Birou Laura. 2007. 'Variance vs average: supply chain lead-time as a predictor of financial performance', *Supply Chain Management: An International Journal*, 12: 349-57.

Dallery Yves, and Liberopoulos George. 2000. 'Extended kanban control system: combining kanban and base stock', *IIE Transactions*, 32: 369-86.

De Bodt Marc A, and Graves Stephen C. 1985. 'Continuous-review policies for a multi-echelon inventory problem with stochastic demand', *Management Science*, 31: 1286-99.

Dessevre Guillaume, Baptiste P, and Lamothe Jacques. 2020. "Corrélation entre taux de service, taux de charge et paramètres du DDMRP: utilisation d'abaques réalisés par simulation." In *MOSIM'20-13ème Conférence internationale de Modélisation, Optimisation et Simulation*, 6 p.

Dessevre Guillaume, and Ben Ali Maha. 2020. "Modélisation et simulation d'un module d'ajustement de la capacité d'un système DDMRP." In *13ème Conférence internationale de Modélisation, Optimisation et Simulation (MOSIM2020), 12-14 Nov 2020, AGADIR, Maroc*.

Dessevre Guillaume, Lamothe Jacques, Pomponne Vincent, Baptiste Pierre, Lauras Matthieu, and Pellerin Robert. 2020. "A DDMRP implementation user feedbacks and stakes analysis." In *ILS 2020-8th International Conference on Information Systems, Logistics and Supply Chain*, 204-11.

Dessevre Guillaume, Martin Guillaume, Baptiste Pierre, Lamothe Jacques, and Lauras Matthieu. 2019. "Étude d'impact du paramétrage des temps de défilement sur la performance d'un déploiement de la méthode DDMRP." In *CIGI QUALITA 2019-13eme Conférence Internationale CIGI QUALITA*. Montréal, Canada.

Dessevre Guillaume, Martin Guillaume, Baptiste Pierre, Lamothe Jacques, Pellerin Robert, and Lauras Matthieu. 2019. "Decoupled Lead Time in finite capacity flowshop: a feedback loop approach." In *IESM 19-8th International Conference on Industrial Engineering and Systems Management*, p. 142-48. Shanghai, China.

Dolgui Alexandre, and Louly MA. 2008. "An Approach for the MRP Parameterization under Lead Time Uncertainty: Branch and Cut Algorithm." In *Proceedings of the 17th IFAC World Congress*, 6-11.

Duenyas Izak, and Patana-Anake Prayoon. 1997. 'Base-stock control for single-product tandem make-to-stock systems', *IIE Transactions*, 30: 31-39.

Dumoulinneuf Sandrine, Faure Lucile, Jaegler Anicia, Antomarchi Anne-Lise, and Burlat Patrick. 2020. 'Pilotage ConWip en contexte mixte MTO/MTS', *Logistique & Management*, 28: 114-24.

Framinan Jose M, González Pedro L, and Ruiz-Usano Rafael. 2003. 'The CONWIP production control system: review and research issues', *Production Planning & Control*, 14: 255-65.

Frein Yannick, Di Mascolo Maria, and Dallery Yves. 1995. 'On the design of generalized kanban control systems', *International Journal of Operations & Production Management*.

Glock Christoph H. 2012. 'Lead time reduction strategies in a single-vendor–single-buyer integrated inventory model with lot size-dependent lead times and stochastic demand', *International Journal of Production Economics*, 136: 37-44.

González-r Pedro L, Framinan José M, and Pierreval Henry. 2012. 'Token-based pull production control systems: an introductory overview', *Journal of Intelligent Manufacturing*, 23: 5-22.

Gupta Surendra M, and Al-Turki Yousef AY. 1997. 'An algorithm to dynamically adjust the number of kanbans in stochastic processing times and variable demand environment', *Production Planning & Control*, 8: 133-41.

Hadley George, and Whitin Thomson M. 1963. "Analysis of inventory systems." No. 658.787 H3.

Harrod Steven, and Kanet John J. 2013. 'Applying work flow control in make-to-order job shops', *International Journal of Production Economics*, 143: 620-26.

Hidayat YA, and Simatupang T. 2018. "Supplier Selection Model Development for Modular Product with Substitutability and Controllable Lead Time." In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 970-75. : IEEE.

Hopp W, and Spearman M. 1996. 'Factory physics: foundations of factory management', *InvinIMcGraw Hill, Chicago, IL*.

Hopp Wallace J, and Roof ML. 1998. 'Setting WIP levels with statistical throughput control (STC) in CONWIP production lines', *International Journal of Production Research*, 36: 867-82.

Hopp Wallace J, and Spearman Mark L. 2011. *Factory physics* (Waveland Press).

Ihme Mathias, and Stratton R. 2015. "Evaluating demand driven MRP: a case based simulated study." In *International Conference of the European Operations Management Association*. Neuchatel, Switzerland.

Ioannou George, and Dimitriou Stavrianna. 2012. 'Lead time estimation in MRP/ERP for make-to-order manufacturing systems', *International Journal of Production Economics*, 139: 551-63.

Jaegler Yann, Jaegler Anicia, Burlat Patrick, Lamouri Samir, and Trentesaux Damien. 2018. 'The ConWip production control system: a systematic review and classification', *International Journal of Production Research*, 56: 5736-57.

Jha JK, and Shanker Kripa. 2013. 'Single-vendor multi-buyer integrated production-inventory model with controllable lead time and service level constraints', *Applied Mathematical Modelling*, 37: 1753-67.

Jodlbauer Herbert, and Dehmer Matthias. 2020. 'An extension of the reorder point method by using advance demand spike information', *Computers & Operations Research*, 124: 105055.

Jodlbauer Herbert, and Reitner Sonja. 2012. 'Material and capacity requirements planning with dynamic lead times', *International Journal of Production Research*, 50: 4477-92.

Johnson Selmer Martin. 1954. 'Optimal two-and three-stage production schedules with setup times included', *Naval Research Logistics Quarterly*, 1: 61-68.

Kao Chiang, and Hsu Wen-Kai. 2002. 'Lot size-reorder point inventory model with fuzzy demands', *Computers & Mathematics with Applications*, 43: 1291-302.

Karmarkar Uday S. 1987. 'Lot sizes, lead times and in-process inventories', *Management Science*, 33: 409-18.

Khanlarzade Narges, Yegane B, Kamalabadi I, and Farughi Hiwa. 2014. 'Inventory control with deteriorating items: A state-of-the-art literature review', *International Journal of Industrial Engineering Computations*, 5: 179-98.

Khojasteh Yacob. 2015. *Production Control Systems: A Guide to Enhance Performance of Pull Systems* (Springer).

Kim JS, and Benton WC. 1995. 'Lot size dependent lead times in a Q, R inventory system', *The International Journal of Production Research*, 33: 41-58.

King Peter L, and King Jennifer S. 2013. *The product wheel handbook: Creating balanced flow in high-mix process operations* (CRC Press).

Kingman JFC. 1962. 'Some inequalities for the queue GI/G/1', *Biometrika*, 49: 315-24.

Koenigsberg Ernest. 1959. 'Production lines and internal storage—A review', *Management Science*, 5: 410-33.

Kortabarria Alaitz, Apaolaza Unai, Lizarralde Aitor, and Amorrortu Itxaso. 2018. 'Material management without forecasting: From MRP to demand driven MRP', *Journal of Industrial Engineering and Management*, 11: 632-50.

Krämer Wolfgang, and Langenbach-Belz M. 1976. 'Approximate Formulae for the Delay in the Queueing System GI/G/l', *Congressbook, 8th ITC, Melbourne*: 235.1-35.8.

Land Martin J. 2009. 'Cobacabana (control of balance by card-based navigation): A card-based system for job shop control', *International Journal of Production Economics*, 117: 97-103.

Lavoie Philippe, Gharbi Ali, and Kenne J-P. 2010. 'A comparative study of pull control mechanisms for unreliable homogenous transfer lines', *International Journal of Production Economics*, 124: 241-51.

Lee Chan-Ju, and Rim Suk-Chul. 2019. 'A Mathematical Safety Stock Model for DDMRP Inventory Replenishment', *Mathematical Problems in Engineering*, 2019: 1-10.

Lee Hau L, and Billington Corey. 1992. 'Managing supply chain inventory: pitfalls and opportunities', *Sloan Management Review*, 33: 65-73.

Leonardo Dênis Gustavo, Sereno Bruno, da Silva Daniel Sant Anna, Sampaio Mauro, Massote Alexandre Augusto, and Simões Jairo Celso. 2017. 'Implementation of hybrid Kanban-CONWIP system: a case study', *Journal of Manufacturing Technology Management*.

Little John DC. 1961. 'A proof for the queuing formula: L= λ W', *Operations Research*, 9: 383-87.

Liu Liming. 1990. '(s, S) continuous review models for inventory with random lifetimes', *Operations Research Letters*, 9: 161-67.

Liu Liming, and Lian Zhaotong. 1999. '(s, S) continuous review models for products with fixed lifetimes', *Operations Research*, 47: 150-58.

Louly Mohamed-Aly, and Dolgui Alexandre. 2011. 'Optimal time phasing and periodicity for MRP with POQ policy', *International Journal of Production Economics*, 131: 76-86.

Marchal William G. 1976. 'An approximate formula for waiting time in single server queues', *AIIE Transactions*, 8: 473-74.

Marek Richard P, Elkins Debra A, and Smith Donald R. 2001. "Manufacturing controls: understanding the fundamentals of Kanban and CONWIP pull systems using simulation." In *Proceedings of the 33nd conference on Winter simulation*, 921-29. : IEEE Computer Society.

Martin Guillaume, Baptiste Pierre, Lamothe Jacques, Miclo Romain, and Lauras Matthieu. 2018. "A process map for the demand driven adaptive enterprise model: Towards an explicit cartography." In *7th International Conference on Information Systems, Logistics and Supply Chain, ILS 2018, July 8, 2018 - July 11, 2018*, 664-72. Lyon, France: INSA Lyon.

Martin Guillaume, Lauras Matthieu, Baptiste Pierre, Lamothe Jacques, Fouqu Anthony, and Miclo Romain. 2019. "Process control and decision-making for Demand Driven Sales and Operations Planning." In *2019 International Conference on Industrial Engineering and Systems Management (IESM)*, 1-6. : IEEE.

Miclo Romain, Fontanili Franck, Lauras Matthieu, Lamothe Jacques, and Milian Bernard. 2015. "MRP vs. demand-driven MRP: Towards an objective comparison." In *Industrial Engineering and Systems Management (IESM), 2015 International Conference on*, 1072-80. : IEEE.

———. 2016a. 'An empirical comparison of MRPII and Demand-Driven MRP', *IFAC-PapersOnLine*, 49: 1725-30.

———. 2016b. "An empirical study of Demand-Driven MRP." In *ILS 2016-6th International Conference on Information Systems, Logistics and Supply Chain*.

Miclo Romain, Lauras Matthieu, Fontanili Franck, Lamothe Jacques, and Melnyk Steven A. 2018. 'Demand Driven MRP: assessment of a new approach to materials management', *International Journal of Production Research*: 1-16.

Mourtzis Dimitris. 2020. 'Simulation in the design and operation of manufacturing systems: state of the art and new trends', *International Journal of Production Research*, 58: 1927-49.

Nahmias Steven, and Wang Shan Shan. 1979. 'A heuristic lot size reorder point model for decaying inventories', *Management Science*, 25: 90-97.

Onyeocha Chukwunonyelum Emmanuel, Wang Jiayi, Khoury Joseph, and Geraghty John. 2015. 'A comparison of HK-CONWIP and BK-CONWIP control strategies in a multi-product manufacturing system', *Operations Research Perspectives*, 2: 137-49.

Orlicky Joseph A. 1975. *Material requirements planning: the new way of life in production and inventory management* (McGraw-Hill).

Orue A, Lizarralde A, and Kortabarria A. 2020. 'Demand Driven MRP–The need to standardise an implementation process', *International Journal of Production Management and Engineering*, 8: 65-73.

Ouyang Liang-Yuh, Wu Kun-Shan, and Ho Chia-Huei. 2004. 'Integrated vendor–buyer cooperative models with stochastic demand in controllable lead time', *International Journal of Production Economics*, 92: 255-66.

Pan Jason Chao-Hsien, and Yang Jin-Shan. 2002. 'A study of an integrated inventory with controllable lead time', *International Journal of Production Research*, 40: 1263-73.

Pergher Isaac, and Vaccaro Guilherme Luís Roehe. 2014. 'Work in process level definition: a method based on computer simulation and electre tri', *Production*, 24: 536-47.

Plossl George W, and Orlicky Joseph. 1994. *Orlicky's material requirements planning* (McGraw-Hill Professional).

Prakash Joshua, and Chin Jeng Feng. 2015. 'Modified CONWIP systems: a review and classification', *Production Planning & Control*, 26: 296-307.

Ptak Carol, and Smith Chad. 2011. *Orlicky's Material Requirements Planning 3/E* (McGraw Hill Professional).

———. 2016. *Demand Driven Material Requirements Planning (DDMRP)* (Industrial Press, Incorporated).

———. 2018. *The Demand Driven Adaptive Enterprise* (Industrial Press, Incorporated).

Rees Loren P, Philipoom Patrick R, Taylor III Bernard W, and Huang Philip Y. 1987. 'Dynamically adjusting the number of kanbans in a just-in-time production system using estimated values of leadtime', *IIE Transactions*, 19: 199-207.

Sarkar Biswajit, Mandal Buddhadev, and Sarkar Sumon. 2015. 'Quality improvement and backorder price discount under controllable lead time in an inventory model', *Journal of Manufacturing Systems*, 35: 26-36.

Selçuk Barış. 2013. 'Adaptive lead time quotation in a pull production system with lead time responsive demand', *Journal of Manufacturing Systems*, 32: 138-46.

Shin Dongmin, Guchhait Rekha, Sarkar Biswajit, and Mittal Mandeep. 2016. 'Controllable lead time, service level constraint, and transportation discounts in a continuous review inventory model', *RAIRO-Operations Research*, 50: 921-34.

Shofa Mohamad Jihan, and Widyarto Wahyu Oktri. 2017. "Effective production control in an automotive industry: MRP vs. demand-driven MRP." In *AIP Conference Proceedings*, 020004. : AIP Publishing.

Shortle John F, Thompson James M, Gross Donald, and Harris Carl M. 2018. *Fundamentals of queueing theory* (John Wiley & Sons).

Spearman Mark L, Woodruff David L, and Hopp Wallace J. 1990. 'CONWIP: a pull alternative to kanban', *The International Journal of Production Research*, 28: 879-94.

Stevenson Mark, Hendry Linda C, and Kingsman Brian G. 2005. 'A review of production planning and control: the applicability of key concepts to the make-to-order industry', *International Journal of Production Research*, 43: 869-98.

Sugimori Y, Kusunoki K, Cho F, and UCHIKAWA SJTIJOPR. 1977. 'Toyota production system and kanban system materialization of just-in-time and respect-for-human system', *The International Journal of Production Research*, 15: 553-64.

Taal Martin, and Wortmann Johan C. 1997. 'Integrating MRP and finite capacity planning', *Production Planning & Control*, 8: 245-54.

Takahashi Katsuhiko, and Hirotani Daisuke. 2005. 'Comparing CONWIP, synchronized CONWIP, and Kanban in complex supply chains', *International Journal of Production Economics*, 93: 25-40.

Takahashi Katsuhiko, and Nakamura Nobuto. 2002. 'Comparing reactive Kanban and reactive CONWIP', *Production Planning & Control*, 13: 702-14.

Tardif Valerie, and Maaseidvaag Lars. 2001. 'An adaptive approach to controlling kanban systems', *European Journal of Operational Research*, 132: 411-24.

Thürer Matthias, Fernandes Nuno O, and Stevenson Mark. 2020. 'Production planning and control in multi-stage assembly systems: an assessment of Kanban, MRP, OPT (DBR) and DDMRP by simulation', *International Journal of Production Research*: 1-15.

Veinott Jr Arthur F, and Wagner Harvey M. 1965. 'Computing optimal (s, S) inventory policies', *Management Science*, 11: 525-52.

Velasco Acosta Angela Patricia, Mascle Christian, and Baptiste Pierre. 2019. 'Applicability of Demand-Driven MRP in a complex manufacturing environment', *International Journal of Production Research*: 1-13.

Vidal Jean, Lauras Matthieu, Lamothe Jacques, and Miclo Romain. 2020. "Toward an Aggregate Approach for Supporting Adaptive Sales And Operations Planning." In *2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*, 1031-38. Bangkok, Thailand: IEEE.

Weng Z Kevin. 1996. 'Manufacturing lead times, system utilization rates and lead-time-related demand', *European Journal of Operational Research*, 89: 259-68.

Wu Kan, Srivathsan Sandeep, and Shen Yichi. 2018. 'Three-moment approximation for the mean queue time of a GI/G/1 queue', *IISE Transactions*, 50: 63-73.

Yano Candace Arai. 1987. 'Setting planned leadtimes in serial production systems with tardiness costs', *Management Science*, 33: 95-106.

# ANNEXE A LEXIQUE ET TRADUCTIONS

Cette annexe reprend les termes les plus importants utilisés dans ce manuscrit avec leur définition et leur traduction anglaise. Plus de détails sont disponibles dans le chapitre 2 (partie 2.1) avec des exemples et des illustrations.

**Délai de production** : temps alloué pour la fabrication d'un ordre, depuis l'identification de son besoin jusqu'à la mise en stock des produits attachés à cet ordre. Le délai de production est un paramètre de gestion, c'est un choix issu d'une prise de décision. Dans les chapitres 4 à 6 correspondant aux articles en anglais, ce terme est traduit par *lead time*.

***Decoupled Lead Time* (DLT)** : il est défini comme le délai le plus long non protégé par un stock tampon dans une nomenclature (Ptak and Smith, 2016). Ce délai représente un chemin critique. Pour déterminer le DLT, on somme les délais de production des différents composants du chemin critique. Il s'agit donc également d'un paramètre de gestion, issu d'une prise de décision portant sur le placement et dimensionnement d'un stock tampon.

**Temps de cycle** : temps s'écoulant entre l'entrée d'un ordre de fabrication dans une boucle de production et sa sortie de la boucle. Autrement dit, temps s'écoulant entre la prise d'un ticket ConWIP et sa libération lorsque la boucle est contrôlée par un ConWIP. Dans les chapitres 4 à 6 correspondant aux articles en anglais, ce terme est traduit par *cycle time*.

**Temps de réponse** : temps s'écoulant entre l'identification d'un besoin de fabrication d'un produit et la mise en stock des produits répondants à ce besoin. Il s'agit donc du temps que met le système à répondre à un besoin. Les temps de réponse sont variables d'un ordre de fabrication à un autre. Dans un contexte DDMRP, le temps de réponse d'un ordre commence lorsque la position de stock atteint le Top du Jaune. Dans les chapitres 4 à 6 correspondant aux articles en anglais, ce terme est traduit par *flow time*.