REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT DE GENIE INDUSTRIEL

# Mémoire

En vue d'obtenir le diplôme
D'ingénieur d'Etat en GENIE INDUSTRIEL



thème

# ANALYSE DES PERFORMANCES D'UN ATELIER FLEXIBLE PAR LES MODELES DE RESEAUX DE FILES D'ATTENTE

Proposé et dirigé par :

Mr. Z HADDAD
Mr. A.AISSANI

Etudié par :

Mr R. KHIAR
Mr M.AIT MENGUELLET

PROMOTION 1998

# *REMERCIEMENTS*



On exprime une reconnaissance toute particulière à Mr Z. Haddad et A. Aissani qui nous ont dirigé pendant l'élaboration de ce travail.

Que les membres de jury trouvent ici l'expression de nos vifs remerciements pour l'honneur qu'il nous font en acceptant de faire partie de notre jury.

On tient également à adresser nos profonds remerciements à toute notre famille pour son soutient moral et physique.

# Dédicace

*A mes très chers parents qu'ils reçoivent toute ma gratitude*

*A mes frères et sœur, particulièrement à mes neveux LILIA et HAKIM.*

*A tous mes amis*

*A la perfection*

*Je dédie ce modeste travail*

*Rabah*

# Dédicace

*A ma très chère mère*

*A mon très cher père*

*A mes frères et sœurs*

*A tous mes amis.*

*Je dédie ce modeste travail*

*MASSINISSA*

# SOMMAIRE

# Sommaire

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

*DEPARTEMENT DE GENIE INDUSTRIEL*

**Intitulé du projet :**
Analyse des performances d'un atelier flexible par les modèles de réseaux de files d'attente

ملخص : في هذا العمل قمنا بتلخيص بعض الإشكاليات لأنظمة الإنتاج المرنة المرتبطة بنمذجتها وتقييم امتيازاتها وإيجاد قيمها الحدية.
لقد قدمنا نظاما تحليليا مركزا على أساس شبكات طوابير الإنتظار للتنبؤ بإمتيازات نظام إنتاج مرن إحداثي جهاز النقل و التحويل.

Abstract :
In this work, we have abstrct some problems of flixible manufacturing systems like their modelisation, performance evaluation and their optimisation.
We present an analytical model for preformence prediction of FMS with a single discrete material-handling device.

Résume :
Dans ce travail, nous avons résume certains problèmes des systèmes manufacturiers flexibles liés à leur modélisation, l'évaluation de leur performances et leur optimisation. Il est question surtout d'un point de vue développé dans la littérature de files d'attente. Nous avons présenté un modèle analytique basé sur des réseaux de files d'attente pour prédire les performances d'un FMS avec un seul appareil de manutention.

# INTRODUCTION GENERALE

# Introduction générale

Le monde de la production industrielle a évolué de manière importante depuis le début de la seconde moitié du siècle.

Deux aspects sont à tenir en compte : accélération de l'automatisation et le remplacement de l'économie d'échelle par l'économie d'envergure.

L'automatisation, facteur de rigidité, doit s'accommoder de la nécessité de produire rapidement des biens très diversifiés en petites ou en moyennes séries. On tente de résoudre le problème en faisant appel aux systèmes flexibles, censés approcher la souplesse des ateliers peu automatisés en utilisant les possibilités les plus récentes d'automatisation, facteur de rendement de qualité.

Toutes erreur de conception se révèle très pénalisante sur le plan financier, d'où la nécessité d'une poussée du système avant son implantation, ou avant toute modification importante de ses caractéristiques(introduction de machines nouvelles, de robots, de stockeurs automatisés, ... ).

Cette analyse fait partie d'une démarche que l'on désigne sous le nom le nom de conception préliminaire. Elle aboutit aux caractéristiques du système de fabrication et du système de gestion à mettre en œuvre pour produire les biens demandés dans les meilleures conditions. pour la conception optimale, on distingue plusieurs approches parmi elle :

Ordonnancement stochastique, approche multicritere, la simulation, les réseaux de files d'attentes.

Dans notre travail, on s'est basé sur les modèles de files d'attentes. Ils ont l'avantage de fournir des résultats rapides car ils ne nécessitent pas le déroulement de l'historique du système.

Ils permettent l'évaluation des performances donnent diverses propriétés qualitatives permettant la simplification pour divers types de contraintes.

# CHAPITRE I

# CONCEPTS DE FLEXIBILITE

# Chapitre I

## Concepts de flexibilité.

L'environnement d'une entreprise, qu'elle soit considérée comme un centre de décision ou un système économique, se veut fortement compétitif. De plus, la firme est, par nature, soumise à des événements perturbateurs internes et externes, de natures et d'impacts aussi complexes que variés. Après un développement rapide de ces deux aspects, la flexibilité s'avère nécessaire pour la viabilité des entreprises.

Les récentes ouvertures massives des économies avancées, combinées à un contexte économique relativement morose, auxquels s'ajoute une diversification des produits demandés par le marché créent un climat de forte concurrence entre industriels supportant des charges de production, entre autre salariales, proches. Les conséquences sur les systèmes productifs de cette économie «de variété », basée sur la segmentation de marchés, sur la recherche de nouveaux créneaux et de nouvelles cibles, et sur la tendance au développement de niches, sont nombreuses et relèvent d'horizons décisionnels divers allant du long terme (décisions stratégiques du type délocalisation ou réorganisation des unités de production, des métiers, des équipes d'opérateurs, des services commerciaux.... etc.) au court terme (décisions opérationnelles permettant la production au plus juste), en passant par des aspects liés à l'innovation, à la réorganisation de la production...

Les systèmes productifs sont également soumis à un grand nombre d'aléas du fait de leurs interfaces actives avec leur environnement (il s'agit alors de perturbations externes telles les variations de la demande, des prix des produits entrants ou sortants ou les délais de livraison incertains) et du fait qu'ils soient composés de sous systèmes eux même source de perturbation (il s'agit de perturbations internes ayant trait par exemple, aux durées incertaines de production et de transport, aux dysfonctionnements des ressources de production ou aux évolutions des coûts de production).

Face à la combinaison de tous ces facteurs, les systèmes productifs doivent évoluer selon un certain nombre de propriétés. F. Roubellat [Roubellat, 1994], de son point de vue d'Automaticien, en décompte trois principales :

- La réactivité, «c'est à dire l'aptitude à s'adapter aux variations ou aux aléas externes ainsi qu'aux aléas internes » ;
- L'efficacité qui reflète la «productivité dans ce contexte de production diversifiée et fluctuante ».

**I-1- La flexibilité : Une notion complexe et ambiguë.**

La flexibilité est une notion multiforme (une cinquantaine d'expressions s'y réfèrent selon [Sethi, Sethi, 1990] et ambiguë. Pour montrer la complexité de cette notion, il suffit de se demander si l'on pense à la même flexibilité lorsque l'on évoque les notions d' «économie de flexibilité », d' «ateliers flexibles », ou même de «flexibilité logicielle ».

Dans certains cas, l'utilisation du terme «flexibilité » traduit en fait la rigidité du phénomène associé. Ainsi, par exemple, la flexibilité de l'emploi qui passe par le recourt à l'intérim et au travail temporaire, se pose en contradiction avec les politiques de développement des capacités de réactivité et d'anticipation des entreprises à plus ou moins long terme. Il en et de même pour la flexibilité technologique qui, apparue sous l'ère de l'automatisation flexible, décrit des sous systèmes de production intégrés en réseaux : Une « rigidification » des liaisons en découle impliquant une fragilisation de la structure globale. La flexibilité caractérise un système par rapport à son environnement. Ainsi, elle traduit, pour un décideur donné, le de pouvoir reconsidérer ses choix à tout moment de manière à pouvoir maintenir l'optimalité de sa décision face à un contexte interne et externe donné [Secke, 1989]. P. Cohendet, Lierena et B. Mutel [Cohendet et al, 1989] l'assimilent au maintien de la cohérence entre l'interne et l'externe du système face aux variations de son environnement.

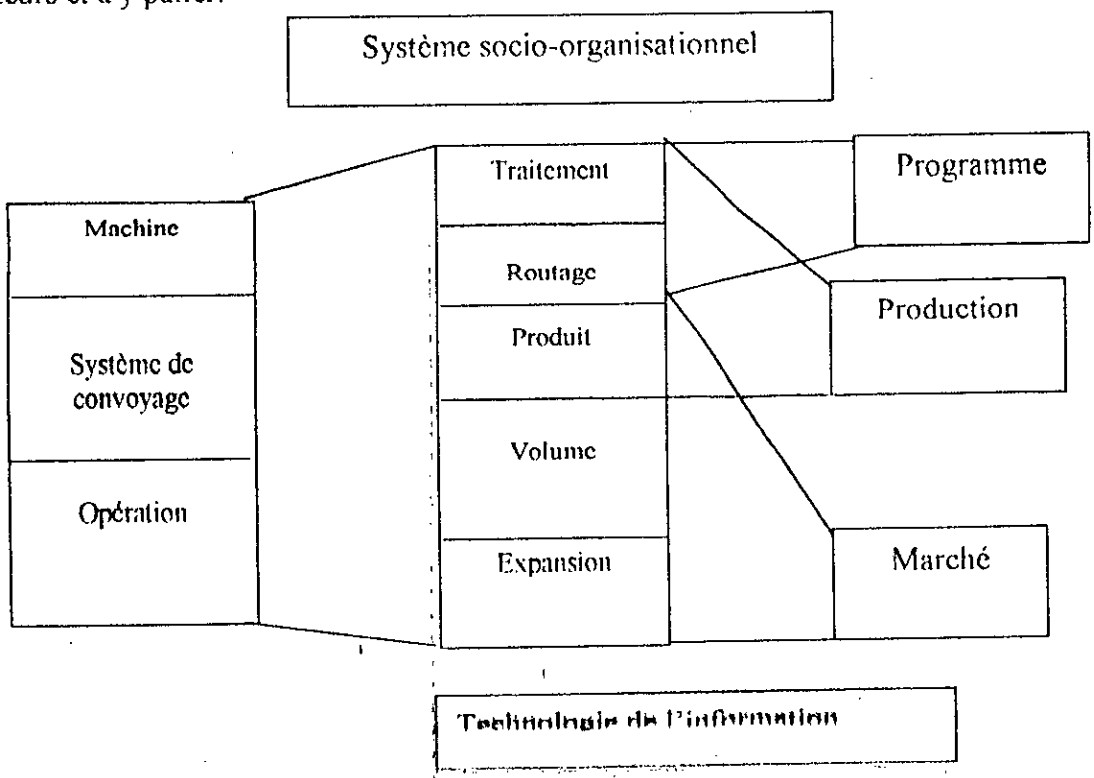
Les concepts de réactivité et d'anticipation, combinés au concept de « maintien dans le temps d'alternatives », mettent en avant la dimension temporelle de la notion de flexibilité. A ce sujet, elle apparaît dans [cohendet et al, 1989]comme statique, relevant l'existence à un moment donné d'un ensemble plus ou moins vaste d'opportunités, ou dynamique et traduisant alors la capacité à réagir continûment dans le temps aux variations de l'environnement. La réactivité y est définie comme une flexibilité dynamique des comportements vis à vis d'un environnement marqué par l'incertitude du processus temporel d'acquisition de l'information. La flexibilité décisionnelle peut être située dans un contexte plus vaste. Elle est considérée dès 1921 comme une perspective économique de recherche intéressante [Lavington, 1921]. La flexibilité organisationnelle constitue elle aussi un vaste domaine de recherche traitant, entre autre, de déférantes formes de flexibilité liées à l'émergence de nouveaux modèles d'organisation tyloriennes et post-tyloriennes : De nouvelles notions sont avancées, basées sur la technologie de groupe [Burbidge,

1992] ou sur le network computer [Massotte, 1996]. Des descriptions plus détaillées de ces nouvelles structures organisationnelles, du type 'product-focussed form' sont développées dans [Kolodny, 1989].

Mieux encore, des analyses plus poussées considèrent que la flexibilité est une notion qui, sur un horizon à long terme, est propre à toutes les composantes de l'unité de production, qu'elles soient physiques, organisationnelles, décisionnelles ou informationnelles.

A ce sujet, A. K. Sethi et S. P. Sethi [Sethi, Sethi, 1990] définissent et mettent en interaction onze types de flexibilité du système de production en mettant en avant l'intérêt d'un système socio-organisationnel adéquatement flexible et d'un système d'information performant(figure1). Ce dernier doit être supporté par un système informatique (hardware et software) flexible, dans la mesure où il ne limite en rien la flexibilité des composantes décisionnelles et informationnelles.

La flexibilité de routage dépend de l'aptitude à produire une pièce en utilisant différents routages à travers le système. La flexibilité de programme traduit l'aptitude du système à fonctionner sans surveillance réelle des opérateurs pendant une période donnée. Elle nécessite l'installation d'un système de capteurs associés à un contrôle informatique apte à détecter les événements perturbateurs et à y palier.




**- Figure-I-1 : Typologie et interaction de flexibilité**

Récemment encore, P. Veltz et P. Zarifian [Veltz, Zarifian, 1993] montre que la flexibilité doit concerner la structure des moyens de production dans son ensemble et doit intégrer l'aptitude de l'organisation ( en tant que système social surplombant la dite structure ) à construire et à développer, dans le temps, une capacité individuelle et collective d'adaptation et d'anticipation. Dans leur typologie de la flexibilité selon des contextes concurrentiels définis, les auteurs présentent la réactivité comme la flexibilité liée à la concurrence par le temps sur la phase de court ou moyen terme comprenant la commande, la fabrication et la livraison.

En conclusion, la flexibilité est devenue, dans cette économie dite «de flexibilité », un avantage managerail qu'il est nécessaire de créer et de gérer de manière à conserver un atout sur la concurrence. Elle relève alors de considérations stratégiques [Lim, 1987] et trouve ses implications dans l'ensemble des autres niveaux tant tactiques qu'opérationnels.

La flexibilité est perçue comme la gestion efficace des modifications des plans de production dont la finalité principale réside en la conservation des atouts managériaux à partir desquels découle une multitude d'opportunités décisionnelles. Elle concerne l'ensemble des ressources, tant humaines que techniques.

## I-2-Gestion et conduite des ateliers.

Avant d'aborder la conduite et la gestion des ateliers, il est nécessaire de rappeler les principales évolutions des systèmes de production, ces derniers étant en étroite relation avec la conception du phénomène de production.

### I-2-1 L'évolution des systèmes de production vers une flexibilité croissante.

L'organisation du travail par la division des tâches est apparue vers le milieu du dix-huitième siècle avec l'apparition de la machine à vapeur. L'extinction de la forme artisanale du travail par la réduction de l'étendue des métiers était une conséquence directe de l'organisation des fabriques autour de machines spécifiques (ateliers), actionnées par l'énergie mécanique. La nécessité de rentabiliser les investissements accentuait les augmentations de rythmes de travail et rendait nécessaire l'organisation des tâches.

L'Organisation Sociale du Travail (OST) introduite par F. Taylor en 1895, et à laquelle on attribue la dégradation de la qualité de travail des ouvriers a le mérite d'engendrer la séparation radicale des parties ' préparation et analyse ' de la partie ' exécution ' des tâches. Basée sur l'optimisation du contexte de production cette méthode met en œuvre des techniques de production de masse nécessaire pour la satisfaction d'une demande sans cesse croissante, le déséquilibre offre / demande reléguant l'aspect qualité en second plan. L'implantation des

ateliers en processus, forme prédominante d'alors, résulte en des regroupements géographiques de postes réalisant les mêmes types d'opérations. Elle va de pair avec hyper-spécialisation du personnel et facilite la supervision.

La production en « job-shop » résulte de l'implantation fonctionnelle des unités de production qui suivit : La réception des produits entrants (matières premières et produits semi-finis) est localisée. Le produit circule entre les regroupements de machines selon son processus de fabrication.

D'autres formes d'organisation des moyens de production, certes moins flexibles s'ajoutent à l'implantation en job-shop. L'implantation d'atelier en position fixe est caractérisée par le déplacement des ressources de production autour de produits de tailles ou de poids importants. Elle s'oppose à l'implantation d'atelier à débit de produits : Dans ce « flow-shop », les produits en cours de fabrication y sont amenés d'un poste à l'autre sur un tapis roulant. Leurs gammes de production sont donc caractérisées par des séquences d'opérations similaires.

Depuis les années 60, l'économie de production a évolué, par étapes successives vers une économie de marché où les entreprises sont contraintes de « fabriquer ce qu'elles peuvent vendre, plutôt que d'essayer de vendre ce qu'elles peuvent fabriquer », puis de variété, dont le maître-mot est la différenciation par l'excellence, pour enfin atteindre celle dite « de flexibilité ». Deux grandes phases sont distinguées : Celle de la recherche de la productivité accrue, passant par la réduction des coûts, précède celle pour laquelle diversification de produits, flexibilité et qualité priment. Ces « grandes transitions contemporaines des systèmes de production » se traduisent par l'évolution de la production des ateliers de type job-shop vers une production de grande variété de produits, en petites ou moyennes séries, ou vers l'implantation des ateliers en îlots ou en cellules de production : Les machines ou les postes de travail effectuant des opérations successives sur une pièce y sont disposés côte à côte, afin d'optimiser les flux de produits. L'implantation en U, permettant de simplifier les trajets des pièces dans les ateliers du fait de la polyvalence de la main d'œuvre et l'implantation en lignes, consistant à mettre en ligne la production de chacune des famille de produits, sont les deux formes les plus répandues des dites structures en îlots.

### I-2-2- Des évolutions conséquentes des modèles de gestion de production.

La fonction production met en œuvre un ensemble de processus physiques à travers des moyens humains, physiques et technologiques propres à l'entreprise considérée, de manière à transformer les 'inputs' (matières premières et produits semi-finis) en 'output' (produits finis). Selon G. Doumeingts, D. breuil et L. Pun [doumeingts et al, 1983], la gestion de la production a pour rôle

d'organiser et de piloter le fonctionnement des processus physiques mis en œuvre afin d'assurer une meilleure utilisation des moyens disponibles et de satisfaire au mieux l'objectif global de production défini en terme de quantité à fabriquer avec une qualité demandéet des délais à respecter ». Complexe du fait qu'elle traite des objectifs aussi difficilement réalisables et intégrables, cette définition met un point en avant la diversité des approches permettant d'avoir une gestion efficace du processus de production.

On distingue deux tendances principales dans l'évolution des concepts et méthodes de gestion. Ces deux tendances ne sont pas tout à fait distinctes : Certains auteurs avancent qu'elles se chevauchent vers la fin des années quatre-vingt. D'autres sont d'avis qu'elles subsistent jusqu'à présent, mues par des travaux de recherches relevant de concepts différents.

La première tendance est relative à la recherche de l'augmentation des performances par le travail direct. Elle résulte en une automatisation du système productif et en une recherche de la polyvalence et de la performance des équipements de production, ces deux objectifs n'étant pas facilement intégrables.

La seconde tendance, quand à elle, considère la composante humaine comme un réseau de preneurs de décisions, impliqués dans le processus de production et non comme un ensemble d'effecteurs perçus, de manière très restrictive, comme des ressources de production et des sources d'informations.

**Première tendance.**

Le début du vingtième siècle a vu se développer des idées de rationalisation du travail. Ainsi, relève-t-on les travaux de F. W. Taylor relatifs à l'Organisation Sociale de Travail (1895), de H. Ford traitant de la standardisation et du travail à la chaîne (1913) ou de Gantt, introduisant, par son diagramme, les premières notions d'ordonnancement (1917).

L'économie de production où « il fallait produire ce qu'il allait être vendu » a trouvé un support conséquent dans le gestion scientifique des activités grâce à l'application de concepts mathématiques dès le début des années trente puis, par la naissance et l'essor de la recherche opérationnelle dans les années cinquante.

Dans les années soixante-dix, l'introduction de l'outil informatique a permis la gestion intégrée de la production grâce aux possibilités d'accès direct aux gros volumes d'information et de leur actualisation en temps réel.

Avec la prédominance des gros systèmes informatiques, les carences de la mini-informatique et les balbutiements encore précaires de la micro-informatique, les bases de données sont physiquement centralisées, partagées et accessibles à tous afin d'atteindre les objectifs de

production consistant en la réduction des coûts et la diminution des temps de cycle. Cette periode de centralisation, retrouvée tant du point de vue fonctionnel qu'organisationnel, est également le théâtre de la multiplication des méthodes basées sur la hiérarchisation des problèmes de gestion de production [Hax, Meal, 1975]. Des algorithmes de gestion de production hiérarchisés sont développés en fonction des besoins industriels exprimés. Parallèlement, le concept de planification des besoins s'est traduit par l'élaboration des méthodes MRP 0, MRP 1, MRP 2 reflétant l'évolution des impératifs économiques : au passage à l'économie où « il fallait vendre ce qui était produit », elles planifiaient les besoins dépendants et indépendants (MRP 0) en fonction de l'évolution de la demande, en tenant compte des contraintes de capacité limitée des ressources de production (MRP 1), pour intégrer ensuite des aspects coûts, devenus d'actualité, du fait que les plannings opérationnels et financiers étaient combinés (MRP 2) [Dallery, 1994].
Nous ne pouvons pas clore ce tour d'horizon sans aborder quelques concepts issus des impératifs de flexibilité et de réactivité. Le premier consiste en l'introduction de marges globales lors de l'élaboration de plans de production. Ces marges correspondent à des degrés de liberté temporels et supplémentaires que s'alloue le gestionnaire afin de pouvoir palier aux perturbations internes et externes.
Le second concept, sans lequel l'exploitation des marges est difficilement mise en œuvre, consiste, dans l'optique systémique des unités de production, en l'établissement de boucles pour le retour d'information entre le système physique et le système de décision.

**Seconde tendance.**

De nombreux travaux témoignent de l'intérêt porté à la composante humaine des systèmes de production. Il apparaît, cependant, que ceux relatifs aux sciences économiques et sociales datent d'une époque où ceux relevant des sciences dites dures s'évertuaient encore à rechercher l'efficacité par la performance technique. Citons, à titre d'exemple les travaux de O.E Williamson [Williamson, 1963] pour lesquels la rationalité des dirigeants d'entreprise ne revient à maximiser un profit mais à développer une politique d'entreprise compatible avec les intérêts personnels des membres de la direction. La fonction de préférence se distingue de la recherche de bénéfice, du fait, entre autre, de l'introduction d'une fonction émoluments.
Il est difficile d'avoir un bref aperçu des analyses réalisées dans le cadre des sciences sociales et sciences économiques. Nous limiterons donc le champ de nos investigations aux quelques approches et concepts, dont l'objectif principal réside en l'introduction explicite du facteur humain dans le processus de prise de décision :

# CHAPITRE II

# GESTION DE PRODUCTION HIERARCHISEE

# *Chapitre II*

## Gestion de production hiérarchisée

### II-1- La gestion de production.

La définition de la gestion de la production découle de cet extrait d'offre de services, trouvée sur le web, de Dubbé, Lunn et Associés Inc, un des plus importants groupe conseil en Gestion des affaires québécois.

« La fonction production est la pierre angulaire de l'entreprise. Qu'il s'agisse d'une société de services ou d'une compagnie manufacturière, le succès d'une entreprise est directement relié à sa capacité de maintenir de façon constante une production de qualité supérieure à moindre coût. Toute déficience sans la dynamique de fabrication ou de livraison du produit peut entraîner des rejets coûteux, des coûts supplémentaires ou des plaintes qui font un tort considérable à l'entreprise ».

#### II-1-1- Algorithmes

La définition de la planification de la production énoncée par [V. Giard , 1988]

> « La planification de la production vise, pour un horizon en général de quelques mois, à optimiser l'utilisation des facteurs productifs disponibles pour la production d'un ou de plusieurs produits répondant à des caractéristiques précises. Il s'agit d'un processus de traitement d'informations aboutissant à une programmation prévisionnelle s'appuyant sur une démarche d'optimisation [Giard, 1988] ».

est limitative dans la mesure où elle ne fait référence qu'aux approches algorithmiques. On distingue en effet deux grandes familles dans la manière d'appréhender la gestion de production [Bonna et al, 1990], [Molet, 1993] :

- Les approches basées sur les techniques algorithmiques. Leur hypothèse de base est la possibilité d'optimiser une fonction technique ou économique sous un ensemble de contraintes formalisables. Elles présentent l'avantage de pouvoir manipuler des données de production agrégées ou désagrégées. Cependant, leur inefficacité croissante est due à la combinaison de deux facteurs. Tout d'abord, la problématique de gestion de production ne se pose plus en terme d'une fonction objectif unique mains en la recherche d'un compromis entre différents critères souvent contradictoires. Les approches multicritères développées alors se heurtent au problème de classement quantitatif de ces critères. Par ailleurs, la

modélisation de la réalité et de leur gestion génère des problèmes combinatoires et aléatoires, ce qui rend toute foemalisation soit réductrice et donc peu portable, soit impossible [Allab, Fink, 1994].

- Les approches utilisant des bases de connaissance. Prenant en compte la complexité de l'environnement manufacturier, elles développent des solutions réalisables et applicables. Elles sont en outre utilisées, en aval des outils basés sur les techniques algorithmiques, afin de remettre les solutions qui en découlent dans le contexte réel de production. Elles présentent par contre l'inconvénient d'être peu portables d'un système de production à l'autre.

Outre les approches utilisées, deux modes de gestion de production s'opposent : le centralisé pour lequel toutes les informations sont prises en compte dans un modèle monolithique unique, et le décentralisé par une forte autonomie des éléments de décision en information et en prise de décision. La gestion de la production hiérarchisée est posée comme une structure intermédiaire, combinant ces deux modes extrêmes.

### II-1-2- Principales architectures de modèles de gestion de production.

Dans un contexte de production hautement concurrentiel, la demande est caractérisée par :

- son évolution éventuellement saisonnière,
- son évolution aléatoire due aux perturbations auxquelles elle est soumise : annulation de commandes, nouvelles commandes à caractère urgent,
- sa traduction en termes quantifiables, tels que les quantités de produits désirés et les délais consentis, ou en termes non quantifiables telle que la qualité moyenne désirée, et
- la fiabilité de son estimation, fortement dépendante de la quantité d'informations disponible.

D'autres paramètres perturbateurs compliquent la tâche de la gestion de la production : il s'agit des événements incertains (occurrence de pannes, rupture d'outils, ...) ou des événements d'incertitude relatifs aux durées d'exécution des opérations sur machines, aux délais de réception des inputs, aux durées de transfert de produits entre machines, etc.

Un système centralisé de gestion, basé sur l'établissement d'un modèle monolithique, détermine un planning prévisionnel en tenant compte simultanément de tous les paramètres, données et informations.

De manière à réagir rapidement face aux perturbations, des méthodes décentralisées ont été développées. Elles permettent le pilotage dynamique de l'atelier par le biais de règles de priorité simples ou complexes. La structure répartie des informations entre centres de décision justifie la

vitesse de réaction face aux perturbations par le fait que par faibles volumes de données, généralement homogènes, sont analysées.

La « gestion de production hiérarchisée » est une expression qui regroupe l'ensemble des structures intermédiaires, combinant des concepts relatifs aux gestion de production centralisée. Elle consiste en l'élaboration de règles simples ou complexes , dépendant de politiques globales de gestion d'atelier.

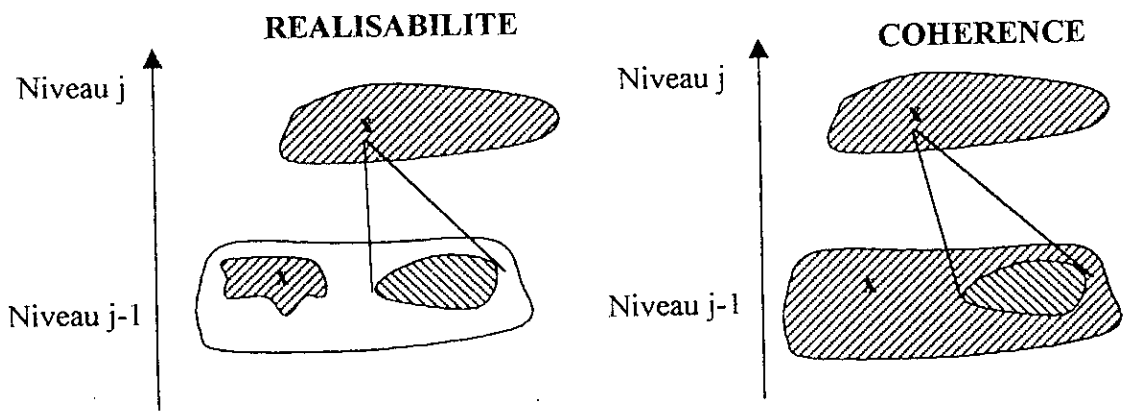
### II-1-3- Quelques notions de gestion de production hiérarchisée.

Les premiers principes de hiérarchisation d'un problème de gestion de production ont été posés par A. C. Hax et H. C. Meal [Hax, Meal, 1975] : un problème global y est décomposé en sous problèmes résolus de manière séquentielle. Les décisions sont prises en cascade, de sorte que les décisions agrégées imposent des contraintes aux décisions détaillées.

La hiérarchisation d'un problème pose deux types de contraintes qu'il faut satisfaire :

**La réalisabilité :** la solution optimale du sous problème $P_j$ génère un ensemble de solutions réalisables non vide pour le sous problème inférieur $P_{j-1}$

**La cohérence :** la solution optimale d'un sous problème $P_j$ génère un ensemble de solutions réalisables comprenant la solution optimale du problème $P_{j-1}$.




Domaine de réalisabilité du niveau considéré

Domaine de réalisabilité généré par la propagation des contraintes associées à la solution optimale du niveau supérieur

x Solution optimale du niveau

**Figure II.1 : réalisabilité et cohérence -**

### II-2-4- Réactivité et flexibilité des structures de modèles de gestion de production hiérarchisée.

La première structure de planification hiérarchisée découle de la classification de la procédure de décision en gestion, proposée dans [Anthony, 1965]. Elle repose sur une caractérisation des décisions selon leur horizon, leur degré d'incertitude, le niveau de détail des informations requises et celui de participation des gestionnaires. Trois catégories principales en résultent :

- Le niveau stratégique : il élabore les politiques et stratégies au plus haut niveau de l'entreprise. Les moyens humains, techniques et financiers sont planifiés sur le long terme en fonction des objectifs techniques et économiques retenus.
- Le niveau tactique : traitant de l'organisation de la production, il défini une utilisation efficace des moyens de production pour la réalisation des impératifs stratégiques.
- Le niveau opérationnel : il intègre deux types de décisions. L'affectation des ressources et l'ordonnancement des opérations sont élaborés à partir des données statiques issues du niveau tactique.

Depuis, les architectures de modèles de production ont évolué en satisfaisant à deux besoins conséquents à l'évolution des systèmes productifs :

**Le besoin de réactivité :** de ce besoin découle l'intégration des boucles pour les retours d'informations entre le système de production et les centres de décision, ou entre ces derniers.

Le premier type de boucles est nécessaire tant en mode de fonctionnement nominal qu'en mode perturbé. En mode nominal, il permet l'actualisation des paramètres des modèles de gestion en fonction de l'évolution de la production dans l'atelier. dans le cas contraire, il répercute les impacts des perturbations au niveau des centres de décision concernés [Al Kazzas, 1989].

La seconde catégorie de boucles assure deux fonctions principales : en plus du maintien de la réalisabilité et de la cohérence entre modèles associés aux niveaux hiérarchiques, elles assurent le bon déroulement du processus down-up d'absorption des perturbations.

**Le besoin de flexibilité :** il conduit vers l'élaboration d'approches génériques pour la modélisation des architectures hiérarchisées. Le nombre de niveaux décisionnels n'est plus fixe, mais dépend des fréquences et impacts des événements, contrôle ou critères affectant le système de production considéré. Ainsi, les modèles de programmation mathématique, qui tiennent compte des aspects relatifs à la demande, aux stocks, ou autres contrôles, intègrent difficilement les événements d'ordre temporel, tels que les délais et retards. Ces derniers, sont généralement pris en compte par des modèles basés sur les graphes PERT ou les réseaux de Petri.

## II-3- Les principaux systèmes de production

### II-3-1- L'atelier «traditionnel »

Dans ce type d'atelier, dont relèvent nombre de P.M.E., l'élément dominant est le compagnon :ouvrier qualifié, sur qui repose la qualité du produit et le respect des délais et qui connaît toutes les astuces permettant de tirer partie des machines implantées le plus souvent au gré des besoins et des possibilités financières de l'entreprise. Ce sont l'implantation des moyens de production et la disponibilité des hommes qui imposent la circulation des produits. Les moyens de manutention y sont souvent rudimentaires : conteneurs ou chariots à roulettes véhiculés au coup par coup par conduite manuelle. Pour peu que les machines soient relativement universelles et les ouvriers polyvalents, un tel atelier est capable de fabriquer une grande variété de produits. Le problème qui s'y pose réside dans la maîtrise des coûts et délais de production, par manque de moyens et de méthodes pour assurer un suivi de production rigoureux. En résumé, il s'agit de systèmes difficilement observables.

### II-3-2- L'atelier spécialisé

Lorsque les séries sont suffisamment importantes, la structure de production (atelier ou cellule) est étudiée et implantée pour le produit. A l'inverse de l'atelier traditionnel, c'est donc la circulation du produit qui impose l'implantation des moyens de production et l'affectation des hommes, et ce de façon durable. Dans ces conditions, rien ne s'oppose à la mise en place de moyens de manutention automatiques pour assurer la circulation du produit. Il est en revanche difficile de convertir des ateliers spécialisés pour d'autres produits que ceux pour lesquels ils ont été conçus.

### II-3-3- La chaîne

Symbole de la grande industrie, la chaîne ou ligne de production représente le système de production dans lequel le produit domine tout : hommes et machines, auxquels il impose sa cadence et sa gamme de fabrication, obligatoirement linéaire. Apparue avec l'avènement du taylorisme, terriblement efficace au temps où toutes les voitures étaient identiques, une chaîne de production devient extrêmement complexe à maintenir lorsque le nombre d'options et de variantes du produit augmente.

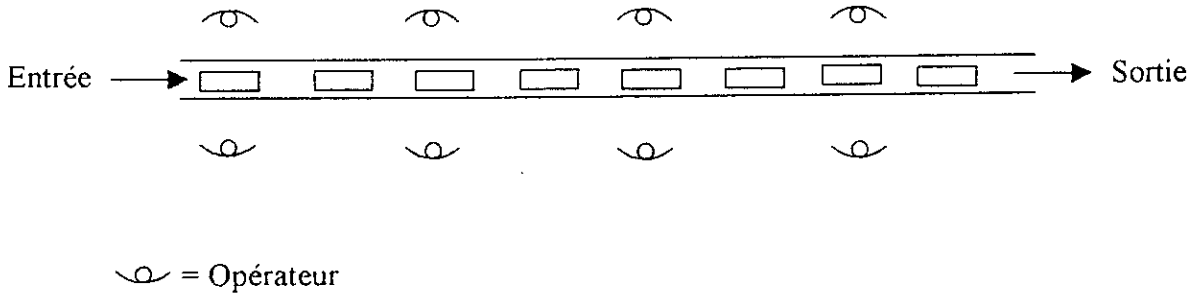



Fig.II 2 Chaîne de production

**II-3-4 La cellule flexible et l'atelier flexible.**

L'atelier spécialisé, la ligne continue ont en commun un avantage : ce sont des à en-cours minimum, et le flux de produit y est parfaitement maîtrisé. Ils sont une bonne réponse pour les productions en séries. En revanche, ils offrent peu de souplesse vis-à-vis des gammes opératoires.

A l'opposé, l'atelier traditionnel, permet d'accepter un nombre élevé de produits différents, de préférence pour de faibles quantités. Par contre, le flux de produit y est mal maîtrisé.

L'atelier (figure 5) ou la cellule flexible (figure 4), tente de concilier les avantages des unes et des autres, ou plutôt d'utiliser les équipements, les techniques et les modes développés dans la grande industrie pour améliorer le point faible de l'atelier non linéaire: la maîtrise des en-cours et du flux de produits.

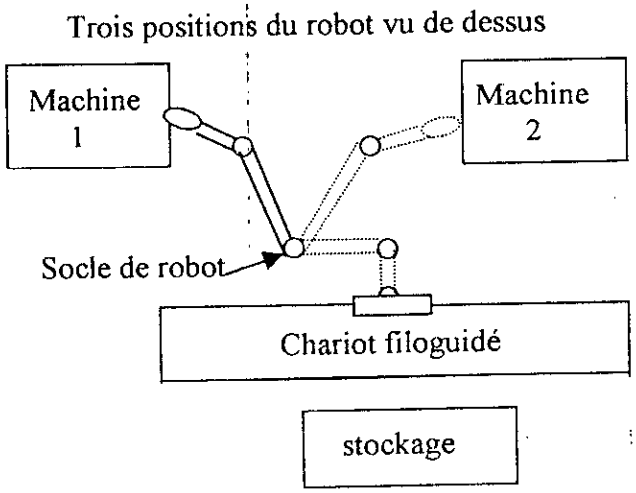



Fig.II 3 Cellule

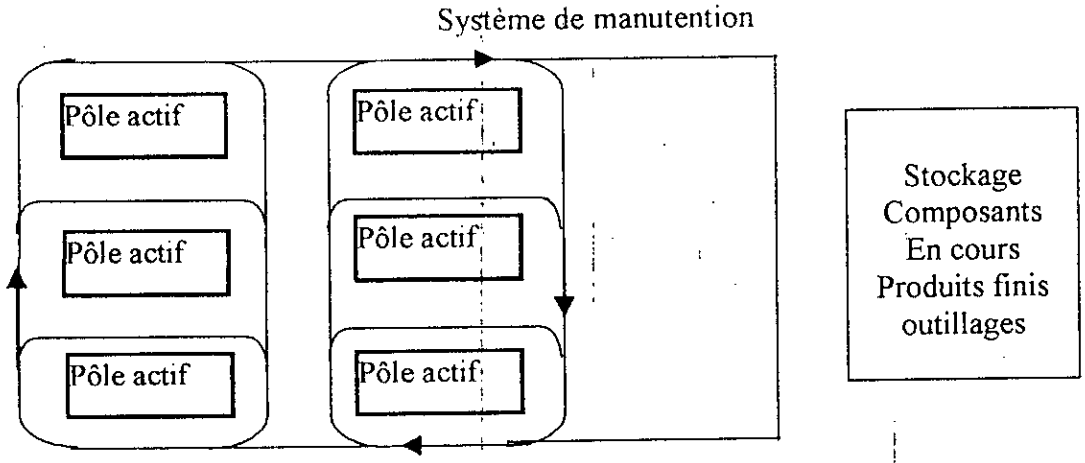



Fig.II 5 Principe d'un atelier flexible

En conséquence, un système de production flexible, atelier ou cellule, peut se décrire comme un certain nombre de pôles actifs (machines-outils, postes d'assemblage, cellules elles-mêmes flexibles, etc.) équipés de moyens si possible polyvalents, travaillant au milieu ou en liaison directe avec un ensemble de moyens de manutention et de stockage dont la mission est d'une part de ne perdre aucune pièce, et d'autre part de distribuer le travail aux postes susceptibles de l'exécuter, selon les directives d'un système de pilotage.

De part son principe, l'atelier flexible est plus proche de l'atelier traditionnel que d'une structure de fabrication en série. Seulement, tout est mis en œuvre pour éviter de retomber dans l'image classique des machines inactives noyées au milieu des accumulations de pièces, qui elles-mêmes séjournent dans l'atelier pendant dix (10) fois le temps nécessaire à leur élaboration (figure 6)

90 % Temps de manutention
et d'attente
10 % Temps de travail
réel sur la pièce

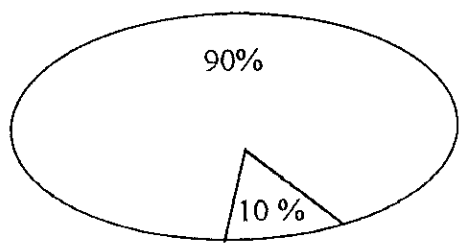



**Fig.II 6 Décomposition du temps de séjour D'une pièce dans un atelier non** Flexible

Les systèmes de production que nous venons de voir peuvent se définir à l'aide de deux paramètres difficilement conciliables à première vue : la productivité et la flexibilité. Le graphe de la figure 7 les site en fonction de ces paramètres. L'approche productique consiste à trouver le meilleur compromis entre les deux.

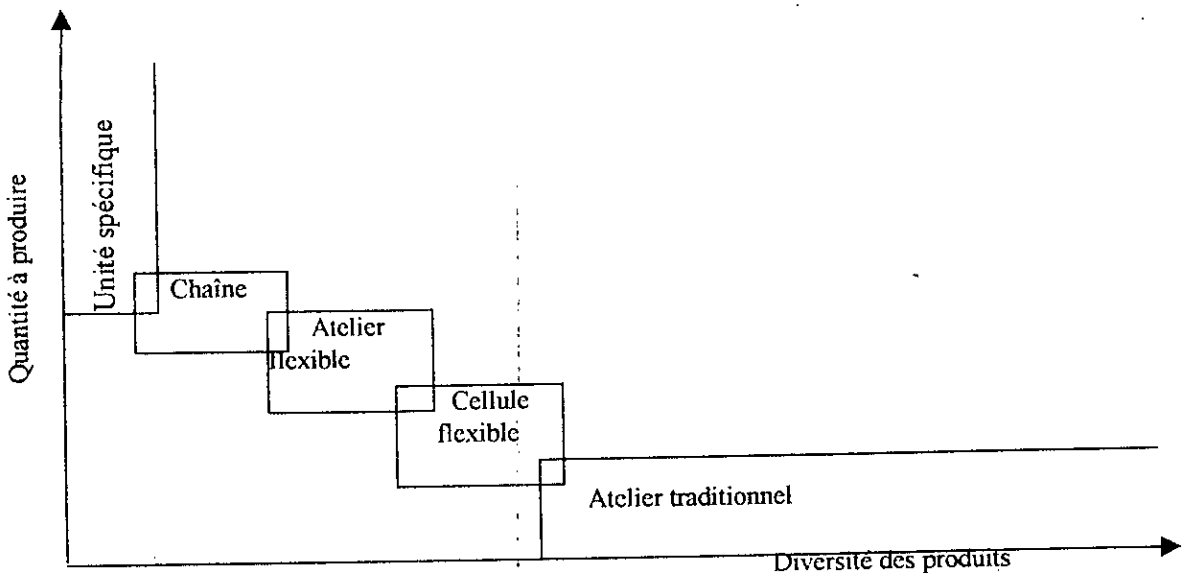



Fig.7 Positionnement relatif des différents systèmes de production

# CHAPITRE III

# LES FILES D'ATTENTE

# Chapitre III

## Les Files d'attente

### III-1- Description d'un système de files d'attente:

Un système d'attente sera défini par le processus d'arrivée des clients, le mécanisme de service (disponibilité, nombre de serveurs), discipline d'attente.

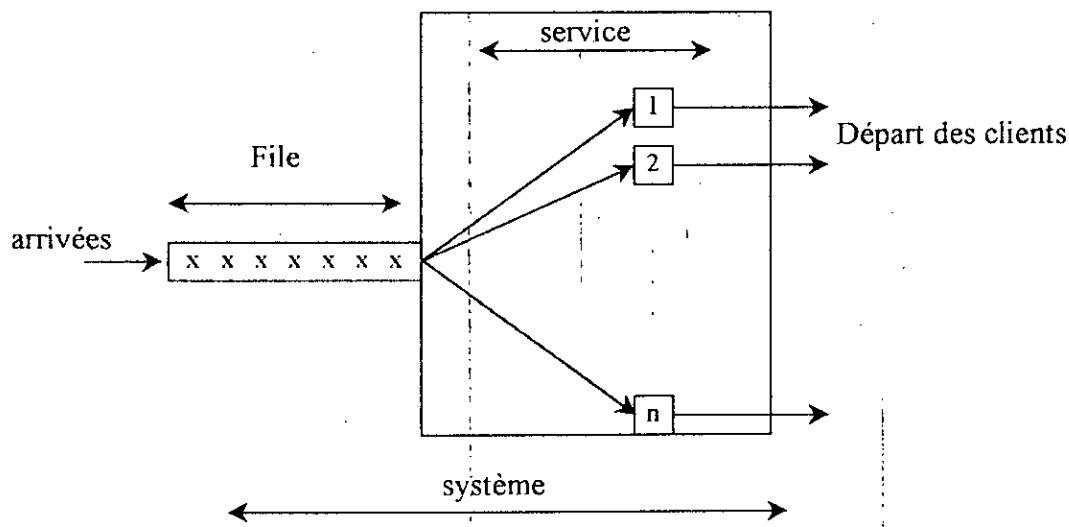



*Figure IV-1. description d'un système de file d'attente*

### III-1-1Processus d'arrivées :

On admet le plus souvent que les clients arrivent indépendamment les uns des autres. Les intervalles de temps séparants deux arrivées de deux clients sont des variables aléatoires indépendantes de même loi.

Si X(t) est le nombre de clients qui arrivent dans [0 , t], alors X(t) est un processus de renouvellement.

- Arrivées régulières: le processus d'arrivée est désigné par D (déterministe)
- Arrivées aléatoires: les instants des arrivées forment un processus de Poisson (des arrivées Marcoviènnes).

- Les arrivées selon la loi d'Erlang d'ordre k: elle peut découler du processus de poisson. Des clients (fictifs) arrivent selon le processus de Poisson et on ne tient que les clients réels d'ordre k, 2k, 3k...désignées par $E_k$.
- Les arrivées selon une loi générale : si la distribution de temps séparants deux événements successifs ne peut être ajustée à une des lois précédentes, alors la loi est dite générale.

**III-1-2- Le mécanisme de service :**

Les durées de service du même serveur sont des variables aléatoires identiquement distribuées

**III-1-2-1 Type de service :**

Service régulier:: Ces toujours le même temps pour effectuer un service donné, la loi est désigné par D (déterministe).

Service exponentiel: La loi est désignée par M (Marcovienne).
Service d'Erlang d'ordre k: Le service consiste en la succession d'opérations élémentaires dans des durées de temps indépendantes.
Service d'une loi générale: La loi est désignée par GI (générale).

**III-1-3- discipline d'attente :**

Les discipline les plus courantes sont:
- Le premier arrivé, le premier servi (FIFO).
- RSS (random selection for service) on sélectionne au hasard les clients pour les servir.
- PR (priority rules).

**III-2- Classification des files d'attentes:**

L'approximation consistant à considérer que le nombre de clients potentiels est infini est acceptable lorsque le nombre de clients dans le système n'a aucune influence sur les clients qui arrivent, ce système est dit ouvert.
Quand le nombre de clients potentiels est fini et ou les clients servis sont réinjectés dans le système dans un délai plus ou moins long, alors le système est dit fermé.

**III-2-1 Les paramètres :**
- $\lambda$:le taux d'arrivée des clients
- $1/\lambda$: l'espérance de l'intervalle de temps séparant deux arrivées successives.

$\mu$: le taux de sortie des clients
- $\rho=\lambda/\mu$: l'intensité de trafic qui s'exprime en Erlang, indique si le système est stable ou non stable.

- $\rho = \lambda/\mu \leq 1$ le système est stable (il tend vers une distribution stationnaire indépendante des conditions initiales).

**III-2-2 Les mesures de performance ou d'efficacité du système.**

$\pi_n$ La probabilité qu'il y ait n clients dans le système.

$L_S$ Espérance du nombre de clients dans le système.

$L_q$ Espérance du nombre de clients dans la file.

$W_s$ Espérance du temps d'attente dans le système.

$W_q$ Espérance du temps d'attente dans la file.

$$L_s = \sum_{n=1}^{\infty} n\pi_n$$

$$L_q = \sum_{n=c}^{\infty} (n-c)\pi_n$$ c: nombre de serveurs en parallèles.

$L_s = \lambda W_S$ formule de Little

$L_q = \lambda W_q$

$\lambda_{eff}$ Taux des arrivées fictives (taux des arrivées des clients qui se rejoignent au système.

Le temps moyen dans le système = temps moyens dans la file + temps moyen de service.

$W_s = W_q + 1/\mu$

$\lambda_{eff} W_s = \lambda_{eff} W_q + \lambda_{eff}/\mu$

$L_s = L_q + \lambda_{eff}/\mu \quad \Rightarrow \quad \lambda_{eff} = \mu(L_s - L_q)$

# CHAPITRE IV

# MODELISATION ET EVALUATION DES SYSTEMES MANUFACTURIERS EN UTILISANT LES MODELES DE FILES D'ATTENTE

# *Chapitre IV*

## Modélisation et évaluation des systèmes manufacturiers en utilisant les modèles de files d'attente

### IV-1- Introduction :

Un FMS est un système manufacturier intégré qui consiste en un certain nombre de stations de travail (assurant des taches telles que l'assemblage, l'inspection, ...) reliées par des systèmes de transport (convoyeurs, véhicules autoguidés, ...) et aptes à dispatcher et orienter les jobs suivant divers itinéraires (routage)dans le système. Les éléments principaux de ce système sont en particulier : les buffers, les palettes, le dispatcher, le système de communication, calcul, contrôle (y compris maintenance et ordonnancement).

Le concept FMS est développé pour tenter de combiner les avantages des systèmes manufacturiers de production de masse (tels que les lignes automatiques de transfert) et les systèmes à grande variété de produits fabriqués en petits lots (tels que les ateliers job shops ). La différence du FMS par rapport aux systèmes traditionnels réside donc dans sa flexibilité (machines versatiles, variétés des types d'opérations et des jobs, routages flexibles en cas d'activité (de surcharges) ou de pannes des machines ). Puisque les machines sont contrôlées automatiquement, le système est flexible pour produire une variété d'articles par une simple modification software. Durant les opérations, le système peut répondre flexiblement à des événements indésirables (pannes, surcharges temporaires, ...) en réorientant les pièces vers d'autres stations. Cette flexibilité du FMS permet d'augmenter la productivité tout en augmentant l'utilisation des machines, et en même temps en réduisant les travaux en cours WIP ( work in process ) et les cycles de productions. Il s'agit donc d'exploiter cette flexibilité de manière adéquate dans le choix des paramètres du système et les opérations de contrôle.

### IV-2- Classification des problèmes FMS :

Plusieurs synthèses tentent de donner une classification des problèmes FMS : [Browne ,84], [Kouvelis ,92] et plus récemment certaines contributions du groupe Bermudes qui tentent d'élaborer des modèles génériques, surtout en relation avec les problèmes d'ordonnancement et de conduite.

Toutes les propositions de classification sont accompagnées de regards critiques à l'égard des modèles mathématiques. Les problèmes formulés sont des problèmes combinatoires NP durs, et leur résolution se fait généralement par des heuristiques ou par simulation. L'approche multicritères intéressante et très en vogue reste encore restrictive.

La plupart des synthèses soulignent l'apport de la théorie des files d'attente et de l'ordonnancement stochastique pour simplifier les fonctions objectif à optimiser, mieux comprendre leurs propriétés et servir de base plus rigoureuse à certaines heuristiques orientées vers certains objectifs.

Citons la synthèse de[Dallery ,92], focalisée sur les lignes de flux et qui comporte en outre une synthèse de synthèses sur la question. Elle est basée essentiellement sur des problèmes d'évaluation des performances et résume diverses propriétés qualitatives permettant des simplifications pour divers types de contraintes : synchrone/asynchrone, panne/maintenance, buffer/blocage.

La synthèse du[Buzacott , 86] montre l'intérêt (et les limites ) des systèmes et réseaux de files d'attentes pour la modélisation FMS.

Enfin, la synthèse de [Buzacott ,92] est consacrée aux propriétés de second ordre (monotonie, convexité ...) est leur ûtilité pour réduire les problèmes d'optimisation combinatoire à des problèmes qui peuvent être résolus par simple allocation marginale.

Ces auteurs tentent de classifier les niveaux de décision pour la conception, la justification et la conduite du FMS. La discussion des modèles est organisée en 4 niveaux hiérarchiques selon le niveau de gestion et la longueur de l'horizon de planification associé aux décisions.

Niveau 1. (Analyse stratégique et justification économique).

La décision d'installer un FMS est prise à haut niveau de gestion où s'élaborent les plans financiers stratégiques. La décision d'implémenter un FMS est souhaitée lorsqu'une variété de produits est fabriquée à volume moyen. Un FMS peut être justifié à ce niveau de planification parce que :

(1) Le temps requis pour installer un FMS peut être de l'ordre de deux années.

(2) Une partie significative du capital peut être mobilisée.

(3) Un risque élevé est à prendre dans l'implémentation en raisons des incertitudes liées à des sources telles que :

- Situation générale de l'économie
- Coût du capital
- Volume des ventes futures

- Ensemble de produits que la compagnie fabriquera dans le futur.

Ces aspects créent un environnement non traditionnel de décision. Les gestionnaires ont historiquement évité des investissements qui nécessitent une mobilisation importante du capital et peuvent occasionner de longues périodes de remboursement. Le niveau d'incertitude (risque) a d'avantage limité l'implémentation de FMS.

Les objectifs pouvant être utilisés pour justifier un FMS sont liés au coût (minimiser l'investissement plus le coût opérationnel ) de la productivité (i.e. maximiser le débit ) et la qualité (améliorer la qualité du produit ).

La première difficulté dans la justification d'un FMS est de quantifier les bénéfices virtuels.

Les facteurs qui doivent être considérés dans l'évaluation d'un FMS incluent :

- Le coût initial de l'installation,
- La flexibilité,
- La fiabilité,
- La maintenabilité,
- Les coûts des tâches,
- L'installation des machines,
- Etc....

Les compagnies candidates à l'implémentation de FMS (celles qui sont caractérisées par des volumes moyens, et des traitements par lots ) introduisent typiquement de nouveaux produits à un rythme plus rapide, et donc, les exigences des futures ressources peuvent ne pas être connues avec un haut degré de confiance. Il apparaît donc important de développer des approches stratégiques adéquates et de justification économique qui définissent explicitement les niveaux de risque de manière à ce que le gestionnaire soit capable de prendre des décisions à ce niveau.

Niveau 2. (conception des installations et équipements )

C'est le niveau où le plan stratégique de financement est considéré en une conception spécifique de l'installation pour réaliser des objectifs multiples. La conception d'une installation FMS nécessite la spécification de :

- Ensemble des actions à produire.
- Nombre et type des machines de production.
- Type de système de transport (convoyeur, véhicule autoguidé, rail, ...).
- Interfaces entre traitement et transport (robot, station chargement/déchargement, ...).
- Nombre et types d'accessoires (palette) utilisés pour le transport et le positionnement des pièces.

- Moyens de stockage des encours (buffers à chaque machine, en position centrale ou combinaison des deux ).
- Composants du système de calcul et de contrôle.
- Configuration du système ( layout ).

Il y a évidemment des interactions entre ces éléments. Pour la conception optimale, on peut opter pour des approches multicritères qui font à l'heure actuelle l'objet de recherches intensives aussi bien théoriques que pratiques.

Suri et Whitney (1984) distinguent 3 niveaux hiérarchiques principaux de décision:

- 1$^{er}$ niveau de décision : décisions stratégiques (i.e. sélection des familles de produit, capacité du système);
- 2$^{eme}$ niveau de décision: décision d'allocation de ressources et des lots.
- 3$^{eme}$ niveau de décision : ordonnancement, dispatching, gestion des outillages, (management tool), décision d'interception.

[Browne ,84] propose 4 catégories de classification: conception (design), planification, ordonnancement, contrôle.

Gershwin (1984), précise cette classification selon l'horizon du temps:

Long terme: investissement et décision de conception initiale;

Moyen terme: décision de conception et de planification;

Court terme: conduite et contrôle en temps réel.

En vertu de ces classifications, on distingue principalement [Kouvelis ,92].

**A. Problème de conception FMS.**

1. Configuration optimale du système (détermination du nombre et type de machines, niveau WIP);
2. Spécification du FMS layout (ordonnancement des stations);
3. Sélection des systèmes de stockage (buffer);
4. Spécification du type et capacité des systèmes de transport;
5. Détermination d'autres ressources (nombre de palettes, ...).

**B. Problème de planification.**

1. Sélection des types de produit ;
2. Groupement des machines ;
3. Charge (allocation des tâches et outillages aux groupes de machines) ;
4. Autres problèmes de planification (allocation palettes/accessoires aux types de

produit, allocation stockage, optimisation du taux de service de la machine...)

**Exemple.** Vinot et Solberg (1985) formulent le problème de configuration optimale du FMS (problème A.1) comme suit :

$$\min z = \sum_{i=1}^{m} k_i . c_i + k_m . n \qquad (1)$$

sous contrainte $TH_m \geq P_0$

où

$z$ = coût global d'investissement et d'exploitation du FMS;

$k_i$ = coût d'investissement et d'exploitation d'une machine de la station i par unité de temps ;

$m$ = nombre de stations;

$c_i$ = nombre de machines (serveurs) à la station i;

$k_N$ = coût de possession du stock par job dans le système;

$n$ = nombre (unitaire) de jobs dans le système;

$TH_m$ = débit du système ;

$P_0$ = débit désiré.

Ce problème, pourtant relativement simple dans sa formulation, est toutefois difficile à résoudre, d'une part à cause de la nature combinatoire du programme en nombre entiers, d'autre part, parce qu'il est difficile à représenter explicitement le débit du système en terme de variables de décisions m, c1, c2, ..., cm. Il est donc difficile d'utiliser les méthodes classiques d'optimisation et de programmation mathématique, même multicritères.

Des problèmes similaires à (1), liés aux problèmes d'ordonnancement, souvent en liaison réelle, sont relatés dans les différents numéros du bulletin [Bermudes, 97].

On y remarque que les praticiens s'orientent le plus souvent dans un premier temps vers la formulation de problèmes complexes et leur résolution par des heuristiques:

- Descente stochastique, recuit simulé, méthode tabou;
- Programmation dynamique, quoique d'un usage peu répandu en pratique;
- Méthodes arborescentes (Séparation et Evaluation, recherche par faisceaux, génération progressive, ...).

Certaines méthodes liées à l'intelligence artificielle (système expert, programmation logique sous contraintes, algorithmes génétiques) sont très en vogue ces dernières années.

Nous verrons que les difficultés mises en évidence peuvent être contournées en exploitant les propriétés fonctionnelles des modèles de réseaux de files d'attente démontrées récemment. Par exemple, la monotonie, la convexité (déterministe ou stochastique) des fonctions-objectif ou des contraintes en termes de variables de décision ou paramètres du système.

Les conclusions de ce type peuvent être utiles pour les problèmes d'ordonnancement, ainsi que pour l'étude des FMS en termes de Réseaux de Petri[Proth, 86].

**C. Ordonnancement.**

Les résultats pour le modèle avec pannes permettent tout comme dans le cas de priorités de simplifier l'étude des critères d'ordonnancement. C'est pour cela que les critères (fonctions-objectif) d'optimalité considérés en théorie d'ordonnancement sont exprimés en terme de temps de complétion.

**D. Version non fiable d'une machine fiable [Dallery et Gershwin].**

Cette approche consiste à transformer une machine fiable complexe en une machine non fiable pour laquelle les résultats connus peuvent être utilisés. Considérons l'exemple simple d'une loi de Cox d'ordre 2 pour le temps de service de paramètres ( $\eta_1$ , $q_1$ , $\eta_2$ )dans le cas d'une machine fiable. On peut la modéliser par une machine non fiable, où $\eta_1$ est le taux de service, $\eta_2$ est le taux de complétion, $q_1$ la probabilité d'avoir une panne (interruption, préemption) à la date d'achèvement de l'opération.

**E. Modèles à 2 machines.**

Comme pour le cas précédant, le temps de blocage peut être assimilé au temps de réparation.

**F. Caractérisation de la loi exponentielle** [Dimitrov ,1994].

Cette question relève priori d'une problématique assez théorique, mais elle n'en possède pas moins des applications importantes en modélisation.

La majorité des problèmes sont basés sur l'absence de mémoire qui caractérise la loi exponentielle: Réseaux de files d'attente, Réseaux de petri stochastiques.

La littérature comporte toute une variété de caractérisation de la loi exponentielle (et donc de l'absence de mémoire).

Les caractérisations les plus connues sont: [Kotz,78]

1. La distribution de probabilité F( ) est exponentielle (F(x) = 1-exp(-$\lambda$ x)) si et seulement si $\lambda(t) \equiv \lambda = \text{cste} \quad \forall t \in R^+$.

2. $F \in exp(\lambda) \Leftrightarrow \overline{F}(x+t) = \overline{F}(x)\overline{F}(t)$ (2)

3. $F \in exp(\lambda) \Leftrightarrow sup\left\{\left|\overline{F}(x+t) - \overline{F}(x)\overline{F}(t)\right| : x,t \geq 0\right\} = 0.$

Les écarts par rapport à l'une de ces propriétés permet de définir les lois nonparamètriques IFR, NBU,...ces caractérisations permettent de mieux définir les critères d'optimalité de politique d'ordonnancement.

**G. Caractérisation par le système non fiable.** [Dimitrov ,94].

Soit T une variable aléatoire $\geq 0$ interprétée comme temps de traitement d'un job qui peut être interrompu par des pannes aléatoires.

- $T \in \exp(\lambda) \Leftrightarrow$ le temps de complétion du job pour les politiques préemptive-résume et préemptive-répeat coïncident en distribution pour une durée de vie X constante du serveur i.e. $P(X=x) = 1, \forall\ 0 \leq x \leq a,\ a > 0$.

*- $T \in \exp(\lambda) \Leftrightarrow$ pour une durée de vie du serveur exponentiellement distribuée et une loi de réparation Y non dégénérée $P(Y=0)<1$, les temps de complétion pour les deux politiques préemptive résume et préemptive-repeat- différent coïncident en distribution.

- $T \in \exp(\lambda) \Leftrightarrow$ les temps de complétion pour les deux politiques préemptive-résume et préemptive-repeat-différent ont des espérances mathématiques égales pour une séquence de durées de vie du serveur ayant des distributions exponentielles de paramètres différents et un point limite fini.

- Soit $P(Y = 0) = 1$ (réparation instantanée) et deux jobs statiquement identiques sont servis sur deux serveurs de durées de vie x constantes i.e. $P(X = x) = 1$, sous les politiques préemptive-résume et préemptive-repeat-différent respectivement alors

$T \in \exp(\lambda) \Leftrightarrow$ soit les distributions des deux politiques coïncident soit leurs espérances mathématiques coïncident.

- (Caractérisation de Huang et Shoung (1993)).

R(x) = fonction de répartition du temps entre panne.

G(x) = fonction de répartition du temps de service du job T.

L(x) = fonction de répartition du temps de complétion $\eta$ du job.

$T \in \exp(\lambda) \Leftrightarrow$ Si pour R(x), G(x), L(x) continues telles que $G'(0^+)$ existe, $T^{\infty} = \eta$ en distribution.

Dans une précédante étude, Dimitrov et Khallil supposait que $T^{\infty} = \eta$ pour des interarrivées exponentielles. Dans la caractérisation précédante, il suffit que R(x) soit continue, strictement croissante et que G(x) soit NBUE pour $E(T)<\infty$.

Il existe une autre caractérisation de [|Galambos ,94] pour d'autres disciplines de préemption.

**IV-3- Systèmes à une station.**

Ils se rencontrent rarement en pratique, mais outre l'intérêt didactique, ils permettent d'approcher des systèmes plus complexes. Leurs propriétés sont bien connues.

**IV-3-1- Système convoyeur overflow.**

L'atelier consiste en c machines parallèles (avec opérateurs) numérotées de 1à c. Les machines sont numérotées dans la direction du mouvement du convoyeur. Un article en mouvement dans le convoyeur sera affecté à la première machine libre et disponible. Si toutes les machines sont occupées ou en panne, l'article est mis au rebut dans une aire affectée à cet usage.

Les articles mis au rebut (overflow):

(i) Sont traités par un opérateur supplémentaire (superviseur);

(ii) Sont traités durant les heures supplémentaires de l'un des opérateurs affectés aux c machines.

(iii) Sont traités à l'issue d'une autre sollicitation du service (rappel), par exemple dans le cas de convoyeurs circulaires.

Le traitement d'un article de l'overflow engendre en général un coût supérieur à ceux des autres. Ainsi à la conception, il est nécessaire de trouver un compromis entre le nombre de machines (ou d'opérateurs) c à affecter, et le taux d'overflow.

Si les opérateurs ont des niveaux de qualification différents, il est possible de réduire le taux d'overflow en affectant correctement les opérateurs aux machines et en ordonnançant convenablement les machines.

Ainsi, les problèmes de conception consistent à déterminer de façon optimale:

- Le nombre de machines en parallèle (ou des opérateurs);
- L'affectation des opérateurs;
- La capacité de production;
- La capacité de l'aire de stockage;
- Le taux d'overflow.

La littérature de files d'attente théorique ou appliquée fournit une quantité de méthodes et résultats pour la résolution de ces problèmes.

Dans les notations de Kandall-Lee, les modèles de files d'attente sont notés A/B/C/K/M , où A code de la loi d'arrivées, B code de la loi de service, C = nombre de serveurs, K = capacité du buffer (s'il y en a), M = capacité de la source.

Les problèmes d'ateliers utilisent des notations similaires étendues à d'autres paramètres et contraintes.

Pour des hypothèses particulières sur l'un (ou plusieurs ) de ces éléments (loi A ou B exponentielle, K et ou M = $\infty$ ), on dispose de résultats analytiques exacts. Pour des modèles plus généraux, il existe des approches numériques ou algorithmiques, utiles pour un objectif d'évaluation des performances (probabilités numériques et moments). Ces dernières ne sont pas de grande utilité pour un objectif d'optimisation car elles ne font pas apparaître les propriétés des fonctions objectif étudiées.

On utilise dans ce but diverses méthodes d'approximation:

- Asymptotiques (type approximation diffusion, petit paramètre,...).
- Comparabilité stochastique.
- Heuristiques (interpolation des systèmes, théorie de l'information, ...).

Elles sont intéressantes car elles conduisent à des formules permettant d'évaluer plus simplement les fonctions objectif (et ou contraintes) et de mettre en évidence des propriétés (monotonie, convexité) par rapport aux variables de décisions et paramètres. les solutions optimales peuvent alors être déduites plus efficacement et plus simplement.

Nous allons illustrer ce type d'approche sur le modèle simple d'une seule station.

### IV-3-2- Système M/GI/C/C avec refus (modèle d'Erlang ).

C'est le modèle le plus simple et le plus intéressant du point de vue didactique.

Hypothèses :

- Les temps de service s des articles sont des v.a.i.i.d., $E(s) = 1/\mu$ .
- Les articles arrivent au convoyeur selon un processus de poisson de taux $\lambda$ .
- Les articles mis au rebut (overflow) sont supposés définitivement perdus.
- Les services sont indépendants du flux d'arrivées.
- Il n'y a pas de buffer (c phases au plus dans le système).

Le système peut alors être modélisé par M/GI/C/C pour lequel les mesures de performance sont bien connues (en régime stationnaire).

Le nombre moyen de clients (articles) dans le système:

$$EN(\lambda,\mu,c) = \frac{\lambda}{\mu}[1 - B(\lambda,\mu,c)] \tag{3}$$

Le débit (taux stationnaire d'articles servis par le système):

$$TH(\lambda,\mu,c) = \lambda B(\lambda,\mu,c) \tag{4}$$

Le taux stationnaire de clients refusés (envoyés au rebut):

$$LR(\lambda,\mu,c) = \lambda - TH(\lambda,\mu,c) = \lambda\left[1 - B(\lambda,\mu,c)\right] \tag{5}$$

Ici, B est donné par la célèbre formule d'Erlang:

$$B(\lambda,\mu,c) = \left\{\frac{(\lambda/\mu)^c}{c!}\right\} / \left\{\sum_{i=0}^{c}\frac{(\lambda/\mu)^i}{i!}\right\} \qquad (6)$$

On sait que pour un tel système, toutes les mesures de performance sont insensibles à la loi de service. On connaît bien également ses propriétés:

| mesures | $\lambda \in R^+$ | $\mu \in R^+$ | $c \in N$ | $(\lambda,\mu) \in R_+^2$ |
|---|---|---|---|---|
| EN | ↗ concave | ? | ↗ concave | concavité ou convexité impossible |
| LR | ↗ | ↘ | ↘ convexe | convexe |
| TH | ? | ? | ↗ concave | ↗ concave |

Ces propriétés permettent de résoudre certains problèmes de conception.

**IV-3-2-1 Nombre optimal de machines parallèles.**

Soit la fonction objectif

$$Z = \min\{\, cv + \psi(LR(\lambda,\mu,c)) : c \in N \,\} \qquad (7)$$

où v = coût unitaire d'une machine (ou d'un opérateur);

$\psi(x)$ = coût de traitement des articles mis au rebut (overflow) lorsque le taux de rebut est x.

Si on suppose que la fonction $\psi$ est croissante et convexe, alors la fonction objectif du problème (7) est convexe en c (voir tableau). On peut donc réduire le problème à un programme d'allocation marginale.

soit $c^* = \arg Z$ la solution optimale du problème (7) , et

$$\Delta\psi(LR(\lambda,\mu,c)) = \psi(LR(\lambda,\mu,c)) - \psi(LR(\lambda,\mu,c+1)) \qquad (8)$$

le bénéfice marginale si on ajoute une machine en plus. En vertu de l'analyse marginale [5]

$$c^* = \min\{c: \Delta\psi(LR(\lambda,\mu,c)) \le v\} \qquad (9)$$

Les propriétés ci-dessus ont été démontrées récemment à l'aide de la théorie de la comparabilité et convexité stochastiques. Elles permettent de donner ainsi un fondement rigoureux à l'algorithme d'allocation marginale qui était au par avant largement utilisé sur la base de considérations heuristiques.

### IV-3-2-2 Affectation des opérateurs.

Ce problème doit être résolu lorsque les niveaux de qualification des opérateurs sont différents. On suppose que les temps de traitement des articles sont des v.a.i.i.d., mais les opérateurs affectés au c machines (ou les machines elles même) fonctionnent à des taux différents. Dans ce cas on modélise le système par GI/M/C/C avec serveurs hétérogènes. Les temps d'interarrivées forment un processus de renouvellement de MTBF = $1/\lambda$. Les serveurs sont numérotés de 1 à c. Le temps de service de la machine i est exponentiel de moyenne $1/\mu_i$, $i = \overline{1,c}$. Les arrivées et services indépendants. Ce modèle est également bien connu, et les principales mesures de performance :

$$TH\left(\bar{\mu},c\right) = \sum_{1+1}^{c} TH_1 \quad ,\bar{\mu} = \left(\mu_1,\mu_2,...,\mu_c\right) \tag{10}$$

$$LR\left(\bar{\mu},c\right) = \lambda - TH\left(\bar{\mu},c\right) \tag{11}$$

$$EN\left(\bar{\mu},c\right) = \sum_{i=1}^{c} EN_i \tag{12}$$

où $TH_i = LR_{i-1} - LR_i$ débit du serveur i

$LR_i = -1/\alpha_i'(0)$ taux d'overflow de la machine i

$EN_i = TH_i/\mu_i$ nombre moyen de clients au serveur i.

Enfin,

$$\alpha_i(s) = \frac{\alpha_{i-1}(s+\mu_i)}{1-\alpha_{i-1}(s)+\alpha_{i-1}(s+\mu_i)} \qquad i = 1,2,....,c. \tag{13}$$

est la transformée de Laplace-Sheltjes de la distribution des temps entre overflow.

On a les propriétés suivantes:

| mesure | $\bar{\mu} \in R_+^c$ | $c \in N, \mu_1 \geq \mu_2 \geq ... \geq \mu_c$ |
|---|---|---|
| TH | croissant en transposition | ↗ concave |
| LR | ↗ en arrangement | ↘ convexe |

Supposons que le taux de traitement de la machine i est $\mu$, i = 1,...,c, et le taux de travail de l'opérateur j est $r_j$, j = 1, 2, ..., c. En d'autres termes, si on affecte l'opérateur j à la machine i, alors le temps moyen pris pour traiter un article est $1/\mu r_j$. Sans perte de généralité, on peut supposer que $r_1 \geq r_2 \geq ... \geq r_c$. Soit $\Pi$ une permutation de {1,2,...,n} et P = {$\Pi$} l'ensemble de

toutes les permutations des c premiers entiers naturels. Un élément $\Pi \in P$ représente une affectation des opérateurs.

Le taux de rebut du système est alors $\eta(\pi) = LR\left(\bar{\mu}(\pi), c\right)$ (14).

Ici, $\mu_i(\pi) = \mu r_{\pi(i)}$ est le taux combiné de service de la machine et de l'opérateur affecté à la $i^{ème}$ position dans le convoyeur.

On cherche l'affectation optimale $\Pi^*$ des opérateurs

$$\Pi^* = \arg\min\{\eta(\Pi): \Pi \in P\} \quad (15)$$

En vertu du tableau, $\Pi^* = <1,2,..,c>$ i.e. les opérateurs les plus qualifiés sont positionnés au début de parcours du convoyeur.

**IV-3-3 Atelier job shop.**

Ce modèle est peu réalisable pour décrire l'atelier job shop, mais nous le présentons à titre didactique .

Les deux principaux problèmes rencontrés dans les job shops sont: l'excèdent du nombre de jobs non achevés et les longtemps de séjour (leadtime). Pour réduire ces coûts, on tente

(a) d'utiliser un contrôle par ordonnancement;

(b) d'augmenter le nombre de machines parallèles, et/ou augmenter la capacité de production (taux de service).

Ici, nous nous focaliserons sur le point (b).

Le modèle le plus simple pour cet atelier est le modèle bien connu M/M/C:

- les jobs arrivent selon un processus de Poisson de taux $\lambda$ ;
- les temps de service sont des v.a.i.i.d. de loi exponentielle de moyenne $1/\mu$ ;
- la capacité de stockage ( buffer) est infinie.

$$. \ EN(\lambda,\mu,c) = \left[\frac{(\lambda/\mu)^c}{c!}\right]\frac{\lambda/c\mu}{(1-\lambda/c\mu)^2} I(\lambda,\mu,c) + \lambda/\mu \quad (16)$$

$$. \ EL(\lambda,\mu,c) = EN(\lambda,\mu,c) - \lambda/\mu \quad ; \quad ET(\lambda,\mu,c) = 1/\lambda \, EN(\lambda,\mu,c) \quad (17)$$

$$. \ EW(\lambda,\mu,c) = 1/\lambda EL(\lambda,\mu,c) \quad (18)$$

où EN(EL) est le nombre moyen stationnaire de clients dans le système ( la file );

ET(EW) est le temps moyen stationnaire de séjour dans le système ( la file ).

$$I(\lambda,\mu,c) = \left[\sum_{i=0}^{c-1}\frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^2/c!}{1-\lambda/c\mu}\right]^{-1} \quad (19)$$

est la probabilité que tous les serveurs soient libres à un instant donné en régime stationnaire. Toutes les mesures de performance sont croissantes, convexes en $\lambda \in (0, c\mu)$; décroissantes convexes en $\mu \in (\lambda / c, \infty)$; décroissantes convexes en $c \in (\lambda / c, \infty)$; EN et EL sont submodulaires et ET, EW concaves en $(\lambda,\mu)$ tels que $(\lambda \in (0,\infty);\ \mu \in (\lambda / c, \infty))$.

Considérons le problème du nombre optimal de machines :

$$c^* = \arg\ \min\{cv + E\Psi(T(\lambda,\mu,c)):\ c \in \aleph\} \qquad (20)$$

où v est le coût unitaire de la machine et $\psi(x)$ le coût d'attente d'un client qui passe x unités de temps dans le système. D'après le tableau, si $\psi$ est une fonction linéaire, alors la fonction objectif du problème ( 20 ) est convexe en c et l'algorithme d'allocation marginale fournit le nombre optimale de machines $c^*$.

**IV-3-3-1 Capacité de production.**

On cherche le taux optimal de service (capacité de production) $\mu^*$ qui minimise le coût :

$$\mu^* = \arg\ \min\ \{c\, g(\mu) + EN_{M/M/C}(\lambda,\mu,c)h : \mu \geq \lambda / c\} \qquad (21)$$

où h est le coût unitaire d'un job en stock et g(x) le coût par serveur de conserver une vitesse de x unités de temps.

Si g est croissante convexe, alors l'algorithme d'allocation marginale ($\mu$) est optimal.

Si $g(x) = vx$, alors $\mu^*$ est la solution positive unique de $\mu = \mathrm{Var}[N_{M/M/C}(\lambda,\mu,c)]h / cv$.

en particulier, si c=1, on obtient le résultat bien connu $\mu^* = \lambda + \sqrt{\lambda h / v!}$

**IV-3-4 Autres approximations.**

Pour M / GI / C . Soit $EA(\lambda,\mu, c)$ l'une des mesures EN, EL, ET ou EW , alors

$$EA(\lambda,\mu,c) = \frac{1 + c_S^2}{2} EA_{M/M/C}(\lambda,\mu,c) + \alpha \qquad (22)$$

Si $\alpha=\lambda/\mu$ si A=N ; $\alpha=0$ si A=L ou W ; $\alpha=1/\mu$ si A=T . $C_S^2$ est le carré du coefficient de variation ( CCV) du temps de service S.

Les mêmes propriétés des mesures de performance énoncées pour M/M/C restent valables.

Pour GI/GI/C .

$$EA(\lambda,\mu,c) = \left(\frac{C_A^2 + \rho^2 C_S^2}{1 + \rho^2 C_S^2}\right)\left(\frac{1 + C_S^2}{2}\right) EA_{M/M/C}(\lambda,\mu,c) + \alpha \qquad (23)$$

où $C_A^2$ est le CCV du temps d'interarrivées. Ces approximations peuvent également être utilisées pour résoudre les problèmes précédents.

### IV-3-5 Utilisation de lois non paramétriques.

Les lois non paramétriques de fiabilité ( IFR, NBU, ...) permettent également d'obtenir des estimations simples et donc de réduire la complexité de certains problèmes.

Rolski prouve que pour GI/GI/C avec une loi d'arrivées NBUE ( New better than used expectation ) , on a $W \leq_d W_{M/GI/1}$ où $W_{M/GI/1}$ est le temps d'attente stationnaire dans un système M/GI/1 de même loi de service et de taux d'arrivées $\lambda/c$; ($\leq d$ est l'ordre en distribution usuel ). En particulier, on en déduit que

$$EW(\lambda,\mu,c) \leq \lambda \frac{E(s)^2 + Var(s)}{2(c-\rho)}, \qquad \rho = \lambda E(s) \tag{24}$$

Voir par exemple, les monographies [ 10, 11 ] où ces aspects sont détaillés pour des réseaux.

### IV-3-6- Dimensionnement du buffer.

En plus des c articles en service, il peut y avoir b articles au plus dans une aire de stockage ( buffer ). Dans le cas le plus simple, on peut utiliser le modèle M/M/C/C+B avec refus. Dans ce cas, le débit moyen (nombre moyen de clients servis ) :

$$TH(\lambda,\mu,c,b) = \lambda\,[\,1-B(\lambda,\mu,c,b)\,] \tag{25}$$

et le taus de refus (overflow )

$$LR(\lambda,\mu,c,b) = \lambda - TH(\lambda,\mu,c,b) = \lambda\, B(\lambda,\mu,c,b) \tag{26}$$

où

$$B(\lambda,\mu,c,b) = \frac{B(\lambda,\mu,c)\left(\frac{\lambda}{c\mu}\right)^b}{\left[1 + B(\lambda,\mu,c)\sum_{i=1}^{b}\left(\frac{\lambda}{c\mu}\right)^i\right]} \tag{27}$$

et $B(\lambda,\mu,c)$a été défini précédemment pour M/M/C/C.

On montre que TH est croissante et concave en $\lambda \in R^+$, $\mu \in R^+$, $c \in N$ et submodulaire et concave en $(\lambda,\mu)$. LR est convexe et croissante en $\lambda \in R^+$, décroissante et convexe en $\mu$, c, b; submodulaire et convexe en $(\lambda,\mu)$.

Certains de ces résultats peuvent être étendus à G/M/C/C+B.

Soit w le profit par client servi (ou par article traité) et $\varphi(b)$ le coût unitaire d'allocation d'une capacité de stockage égale à b. Alors pour le problème de dimensionnement optimal du buffer, la solution

$b^* = \arg\max \{ w.\, TH(\lambda, \mu, c, b) - \psi(b):\ b \in N \}$

est donnée par l'algorithme d'allocation marginale, si $\Psi$ est convexe.

### IV-3-7- Modèle avec rappels.

Comme nous l'avons fait remarqué précédemment les clients refusés (overflow) ne sont pas perdus définitivement puisqu'ils doivent être traité ultérieurement.

Les modèles avec rappels ( initialement développé en téléphonie) permettent de prendre en compte ce phénomène en introduisant une politique de rappels

- constante : les clients sont rappelés à des intervalles périodiques de durée $1/\nu$
- dépendante de la taille de l'overflow: si n articles sont au rebut, le taux de rappels est $\nu n$.
- linéaire: combinaison des deux précédentes etc. ....

La littérature sur les modèles avec rappels s'est enrichie ces dernières années, et on trouvera une compilation des progrès sur la question dans deux synthèses de Yang etTempleton (1987), Falin (1990), ainsi que dans la monographie récente de Templeton et Falin (1997).

La majorité des résultats ne concernent que les modèles de type M/G/1. Les modèles avec arrivées arbitraires, rappels non exponentiels, serveurs multiples, priorités restent encore mal connus; mais si certains approches ont pu être développées (numériques) qui restent toutefois limitées à un objectif d'évaluation de performances.

### IV-3-8- Modèle avec vacances.

Ils sont très utiles pour les systèmes de production et consistent à introduire une politique de vacances du serveur dès que la file se vide. L'oisiveté est ainsi exploitée à d'autres tâches que peut assumer le serveur. Les modèles avec rappels peuvent être assimilés dans un cadre théorique à des modèles avec vacances, même si leur étude nécessite des spécificités que l'on ne rencontre pas dans la littérature systèmes avec vacances.

Les politiques de vacances peuvent être de nature diverse

- le serveur peut retourner à sa tâche initiale à des intervalles périodiques,
- il peut le faire dès que la capacité du buffer atteint un niveau donné.

On peut imaginer des modèles avec vacances et rappels (Artelejo 1997).

## IV-4- Lignes de flux et lignes de transfert.[16,18-23]

les articles à traiter passent par plusieurs phases (m), dans l'ordre 1,2,3,...,m.

La seule différence entre ces deux classes d'ateliers est que dans les lignes de flux on a souvent recours à des opérations manuelles, et il faut en tenir compte dans les formulations des problèmes de conception et d'exploitation. Cependant la plupart des problèmes sont similaires

aux deux. Une bonne synthèse figure dans [Dallery, 1992]. Nous reconsidérons que les lignes de flux avec machines fiables (FLRM).

Les problèmes que l'on peut formuler sont de déterminer de manière optimale:

- le nombre de phases.
- le nombre de machines ou d'opérateurs par phase;
- l'allocation de charge de travail;
- l'affectation des ouvriers;
- l'ordonnancement des serveurs.

Supposons que les articles arrivent selon un processus de Poisson de taux $\lambda$. Il y a $c_i$ serveurs à la station $N^0$ i(en parallèle) et tout exponentiels de moyen de service $1/\mu_i$. On impose les hypothèses d'indépendance habituelles entre services et arrivées. Alors, le nombre moyen de clients dans le système:

$$EN\left(\lambda,\bar{\mu},\bar{c}\right) = \sum_{i=1}^{m} EN_{M/M/C}\left(\lambda,\mu_i,c_i\right) \tag{29}$$

et le nombre moyen de séjour d'un client arbitraire dans le système:

$$ET\left(\lambda,\bar{\mu},\bar{c}\right) = \sum_{i=1}^{m} ET_{M/M/C}\left(\lambda,\mu_i,c_i\right) = \frac{1}{\lambda} EN\left(\lambda,\bar{\mu},\bar{c}\right) \tag{30}$$

où $\bar{\mu} = (\mu_1, \mu_2, ..., \mu_m)$ et $\bar{c} = (c_1, c_2, ..., c_m)$.

Ce résultat signifie que les mesures de performance peuvent être décomposées par phase isolées. Par conséquent, on peut utiliser les résultats précédents concernants une station isolée. Ce type de résultat peut être étendu au cas d'ateliers jobs shops décrits par de »s réseaux de Jackson.

Soit la fonction $\tilde{N}_{M/M/C}: [0,c]xN \rightarrow R^+$ dé finiepar $\tilde{N}_{M/M/C}\left(\frac{\lambda}{\mu},c\right) = EN_{M/M/C}(\lambda,\mu,c)$ car EN ne dépend que de $\lambda/\mu$, la charge moyenne offerte par unité de temps.

Soit $R = c_1 + c_2 + ... + c_m$ le nombre total de serveurs disponibles et $W = \frac{1}{\mu_1} + ... + \frac{1}{\mu_m}$ la charge totale à allouer.

On montre que $\tilde{N}_{M/M/C}\left(\bar{\mu},\bar{c}\right)$ est submodulaire en $(\mu, c) \in [0, c] \times N$, et par conséquent

$$\tilde{N}_{M/M/C}(\mu_1+\mu_2,c_1+c_2) \leq \tilde{N}_{M/M/C}(\mu_1,c_1) + \tilde{N}_{M/M/C}(\mu_2,c_2) \tag{31}$$

Par conséquent, $EN\left(\lambda,\bar{\mu},\bar{c}\right) \geq N_{M/M/C}(\lambda,w,R)$. Cela signifie qu'il est optimal d'avoir une seule phase avec tous les serveurs affectés à cette unique station. Cela nécessite cependant que les serveurs soient entraînés à effectuer toutes les tâches nécessaires au traitement complet d'un job. Il y a don en général plus d'une phase, mais l'objectif reste de choisir le minimum de phases possibles.

Soit $h_i$ le coût d'un job dans le système et $v_i$ le coût d'une machine (ou opérateur) à la phase i, i= 1, ...,m. L'allocation optimale des machines par phase

$$C^* = \arg\min\left\{\sum h_i EN_{M/M/C}(\lambda,\mu_i,c_i) + v_i c_i \quad : c_i \in N, i = 1,...,m\right\} \tag{32}$$

En vertu des propriétés énoncées précédemment de $EN_{M/M/C}$ , la fonction objectif est séparable convexe. Par conséquent, la solution optimale est encore fournie par l'algorithme d'allocation marginale.

Considérons maintenant le problème d'allocation de la charge. Soit la fonction

$\hat{N}_{M/M/C} : [0,1] \times \aleph \to \Re$ définie par $\hat{N}_{M/M/C}\left(\frac{\lambda}{c\mu},c\right) = EN_{M/M/C}(\lambda,\mu,c)$

car EN ne dépend que de $l = \lambda/c\mu$. On vérifie que $\hat{N}_{M/M/C}(\rho,c)$ est croissante convexe en $\rho \in [0,1]$, et submodulaire en $(\rho, c) \in [0,1] \times \aleph$.

Soit à allouer une charge totale de w dans les différentes phases. L'allocation optimale minimisant le nombre de jobs dans le système ( et donc le temps moyen de séjour passé dans le système ).

$$\bar{w}^* = \left(w_1^*, w_2^*, ..., w_m^*\right) = \arg\min\left\{\sum_{i=1}^{m} EN_{M/M/C}\left(\lambda,\frac{1}{w_i},c_i\right) : \sum_{i=1}^{m} w_i = W, \quad \bar{w} \in \Re_+^m\right\} \tag{33}$$

où $w_i$ est la charge allouée à la phase i (i.e. $w_i$ est le temps de traitement d'un job à la phase i). Réécrivons la fonction objectif sous la forme:

$$\sum_{i=1}^{m} \hat{N}_{M/M/C}\left(\frac{\lambda w_i}{c_i},c_i\right) \tag{34}$$

Puisque $\hat{N}_{M/M/C}(l,c)$ est croissante convexe en $l = \lambda w/c$ , alors d'après la théorie classique d'optimisation, la solution optimale vérifie la condition de Karush-Kun-Tuker:

$$\frac{\lambda}{c_i}\hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_i},c_i\right) = \text{constante}, \quad i = 1,...,m \tag{35}$$

où $\hat{N}'$ est la dérivée par rapport à l. On en déduit que pour $c_i < c_j$

$$\hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_i}, c_i\right) = \frac{c_i}{c_j}\hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_j}, c_j\right) \le \hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_j}, c_j\right)$$

Puisque la submodularité de $\hat{N}_{M/M/C}(l,c)$ en (l,c) est équivalente au fait que $\hat{N}'_{M/M/C}(l,c)$ est décroissante en c, alors d'après l'inégalité ci-dessus:

$$\hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_i}, c_i\right) \le \hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_j}, c_j\right) \le \hat{N}'_{M/M/C}\left(\frac{\lambda w_i^*}{c_j}, c_j\right)$$

En vertu de cette inégalité, et puisque $\hat{N}_{M/M/C}(\rho,c)$ est croissante convexe en $\rho$, alors on obtient une caractérisation partielle de l'allocation optimale de la charge

$$c_i \le c_j \Rightarrow \frac{w_i^*}{c_i} \le \frac{w_j^*}{c_j} \quad \text{,pour tout } j = 1, \ldots, m \tag{36}$$

En particulier, $c_i = c_j \Rightarrow w_i^* = w_j^*$ , et donc si toutes les phases ont le même nombre de machines, il est optimal d'allouer la charge totale de manière équitable (i.e. $w_i^* = w/m$ , i=1,...,m).

L'optimalité de l'affectation équitable de la charge est valable pour les fonctions objectif plus générales, par exemple si le coût total des jobs dans le système est

$$\sum_{i=1}^{m} E\psi\left(N\left(\lambda, \frac{1}{w_i}, c\right)\right) \quad \text{ou} \quad \psi\left(\sum_{i=1}^{m} N\left(\lambda, \frac{1}{w_i}, c\right)\right) \text{, ou } \psi \text{ est une fonction convexe.}$$

Par contre, si le coût d'attente est différent du coût de service, alors même dans le cas d'un serveur par phase, l'allocation optimale de la charge n'est équitable.

Si les nombres de serveurs aux différentes stations sont différents, la charge allouée par station ($w_i^*/c_i$) est plus grande à une phase qui possède plus de stations.

L'allocation optimale de la charge peut être maintenant obtenue en utilisant l'algorithme d'allocation marginale.

### IV-4-1- Affectation des ouvriers.

L'allocation de la charge est réalisée en général avant l'affectation des ouvriers. Supposons qu'on affecte un ouvrier par phase de taux $r_i$ à la station i i.e. l'ouvrier peut traiter $r_i$ unités de charge par unité de temps réel. Si $r_1 = \ldots = r_m$ on peut choisir n'importe quelle affectation ( car les ouvriers ont des qualifications identiques ).

Supposons que l'ouvrier j est affecté à la phase i à laquelle on a alloué une charge $w_i$. alors le temps de service suit une loi exponentielle de moyenne $w_i/r_j$ , et le nombre moyen de jobs est $\hat{N}_{M/M/C}\left(\frac{\lambda w_j}{r_j},1\right)$.

Supposons que l'ouvrier $\pi$ (i) est affecté à la phase i. Pour cette affectation

$\pi = < \pi(1), \pi(2), \ldots \pi(m) >$ des ouvriers aux étapes 1,2,3, ...,m, le nombre total de jobs dans le système est

$$EN(\pi) = \sum_{i=1}^{m} \hat{N}_{M/M/C}\left(\frac{\lambda w_i}{r_{\pi(i)}},1\right) \tag{37}$$

On cherche l'affectation optimale $\pi^* = \arg\min \{EN(\pi) : \pi \in P\}$. (38)

Considérons deux affectations $\pi$ et $\pi'$ telles que $\pi(i) = \pi'(i)$, $i = 1, \ldots, m$ ; $i \neq l, l \neq 1$;

$\pi(l) = \pi'(l+1)$ et $\pi(l+1) = \pi'(l)$.

Sans perte de généralité, on peut supposer que $r_{\pi(l)} \geq r_{\pi(l+1)}$. Alors si $w_l \geq w_{l+1}$ on a

$$\frac{w_{l+1}}{r_{\pi(l)}} \leq \min\left\{\frac{w_{l+1}}{r_{\pi(l+1)}}, \frac{w_l}{r_{\pi(l)}}\right\} \leq \max\left\{\frac{w_{l+1}}{r_{\pi(l+1)}}, \frac{w_l}{r_{\pi(l)}}\right\} \leq \frac{w_l}{r_{\pi(l+1)}}$$

d'ou

$$\frac{w_{l+1}}{r_{\pi(l)}} + \frac{w_{l+1}}{r_{\pi(l+1)}} \geq \frac{w_l}{r_{\pi(l)}} + \frac{w_{l+1}}{r_{\pi(l+1)}}$$

Puisque $\hat{N}_{M/M/C}(\lambda w,1)$ est croissante convexe en w, alors

$$\hat{N}_{M/M/C}\left(\frac{\lambda w_{l+1}}{r_{\pi(l)}},1\right) + \hat{N}_{M/M/C}\left(\frac{\lambda w_l}{r_{\pi(l+1)}},1\right) \geq \hat{N}_{M/M/C}\left(\frac{\lambda w_l}{r_{\pi(l)}},1\right) + \hat{N}_{M/M/C}\left(\frac{\lambda w_{l+1}}{r_{\pi(l+1)}},1\right)$$

et par conséquent, $EN(\pi') \geq EN(\pi)$. Il est donc préférable d'affecter l'ouvrier de taux le plus grand ($\pi$ (l)) à la phase la plus chargée ($w_l$). Ainsi, si $\underset{\pi}{w} = (w_{\pi(1)}, \ldots, w_{\pi(m)})$ est un réarrangement des charges dans l'ordre décroissant (i.e. la charge allouée à la phase $\hat{\pi}(i)$ est plus grande que celle allouée à la phase $\hat{\pi}(j)$, pour $i < j$ ) et si les ouvriers sont numérotés dans l'ordre décroissant de leurs taux ( $r_i \geq r_j$ pour $i$ $j$ ) alors $\hat{\pi}$ est l'affectation optimale.

**IV-4-2- Cas de stations (phase) GI/GI/C.**

En utilisant l'approximation (6) à la première phase, le nombre moyen de jobs est

$$\hat{N}_{GI/GI/c}\left(\lambda, c_A^2, \mu, c_S^2, c\right) = \left(\frac{c_A^2 + \rho^2 c_S^2}{1 + \rho^2 c_S^2}\right)\left(\frac{1 + c_S^2}{2}\right) EL_{M/M/c}(\lambda, \mu, c) + \frac{\lambda}{\mu} \tag{39}$$

Le CCV du temps d'interarrivées peut être approché par

$$c_D^2\left(\lambda, c_A^2, \mu, c_S^2, c\right) = 1 + \left(1 - \rho^2\right)\left(\frac{c_A^2 - 1}{1 + \rho c_S^2}\right) + \frac{\rho^2}{c}\left(c_S^2 - 1\right) \tag{40}$$

Par conséquent, le nombre moyen de jobs à la phase i peut être approché par

$$EN_i = \hat{N}_{GI/GI/c}\left(\lambda, c_{A_i}^2, \mu_i, c_{S_i}^2, c_i\right) \tag{41}$$

où $\quad c_{A_i}^2 = c_A^2 \ , \ c_{A_{i+1}}^2 = c_{D_i}^2\left(\lambda, c_{A_i}^2, \mu_i, c_{S_i}^2, c_i\right), \ i = 1, \ldots, m-1 \tag{42}$

Pour $c_i = 1$ , $i = 1, \ldots, m$, l'approximation du nombre total de jobs dans le système est

$$EN\left(\bar{\rho}, \bar{c}_S^2\right) = \sum_{i=1}^{m} \frac{(c_{A_i}^2 + \rho^2 c_{S_i}^2)(1 + c_{S_i}^2)\rho_i^2}{(1 + \rho_i^2 c_{S_i}^2)(2)(1 - \rho_i)} + \sum_{i=1}^{m} \rho_i \tag{43}$$

où $\rho_i = \lambda E(S_i)$.

On montre que si $\rho_1 = \ldots = \rho_m$ , alors $EN(\bar{\rho}, \bar{x})$ est croissante en transposition adjacente en $\bar{x} \in \Re_+^m$.

**IV-4-3- Ordonnancement des stations.**

On suppose que la charge a été allouée et les ouvriers affectés. Soit $\pi = < \pi(1), \ldots, \pi(m) >$ l'ordonnancement des stations i.e. le job qui arrive dans le système va à la station $\pi(1)$, puis $\pi(2), \pi(3), \ldots, \pi(m)$ puis quitte le système.

L'approximation du nombre de jobs dans le système est

$$\tilde{N}(\pi) = \sum_{i=1}^{m} \frac{(c_{A_{\pi(i)}}^2 + \rho_{\pi(i)}^2 c_{S_{\pi(i)}}^2)(1 + c_{S_{\pi(i)}}^2)\rho_{\pi(i)}^2}{(1 + \rho_{\pi(i)}^2 c_{\pi(i)}^2)(2)(1 - \rho_{\pi(i)}^2)} + \sum_{i=1}^{m} \rho_i \tag{44}$$

$$\text{où } c_{A_{\pi(1)}}^2 = c_A^2 \ , c_{A_{\pi(i+1)}}^2 = 1 + (1 - \rho_{\pi(i)}^2)\left(\frac{c_{A_{\pi(1)}}^2 - 1}{1 + \rho_{\pi(i)}^2 c_{S_{\pi(1)}}^2}\right) + \rho_{\pi(i)}^2\left(c_{S\pi(i)}^2 - 1\right) \ ; i = 1, \ldots, m-1 \tag{45}$$

$$\text{L'ordonnancement optimal} \qquad \pi^* = \arg\min\left\{\tilde{N}(\pi) : \pi \in P\right\} \tag{46}$$

**IV-4-4- Contrôle de qualité.**

On suppose toujours qu'on a une ligne de flux à m phase avec buffers illimités. Soit $p_i$ la probabilité qu'un job traité à la station i soit défectueux. Le flux des jobs à l'arrivée est

poissonnien de taux $\lambda$. Une station d'inspection est placée après la phase m. Si un job est défectueux, il est renvoyé à la phase qui a causé le premier défaut, disons i, et sera de nouveau traité aux stations i+1, ..., m. Soit $\lambda_i$ , le taux de jobs qui quittent la phase i, et $\gamma_i$ le taux avec lequel les jobs sont renvoyés par la station d'inspection vers la phase i; $\beta_i$ le taux de jobs défectueux qui arrivent à i+1 en provenance de i. $\lambda_0 = \lambda$; $\beta_0 = 0$. La loi de conservation des flux à une station i permet d'écrire:

$$\lambda_i = \lambda_{i+1} + \gamma_i , \qquad i = 1, ..., m.$$

$$\beta_i = \beta_{i-1} + \gamma_i p_i + ( \lambda_{i+1} - \beta_{i-1} ) p_i \quad , i = 1, ..., m \tag{48}$$

$\gamma_i = \gamma_i p_i + ( \lambda_{i-1} - \beta_{i-1} ) p_i$ , $\quad i = 1, ..., m.$

On vérifie que $\lambda_i - \beta_i = \lambda = \lambda_{i-1} - \beta_{i-1}$ , i = 1, ..., m et $\gamma_i = \lambda p_i /(1-p_i)$. donc

$$\lambda_i = \lambda(1 + \sum_{j=1}^{i} \frac{p_j}{1 - p_j}), \quad i = 1, ..., m \tag{49}$$

Si le nombre de machines affectées à la phase i est $c_i$ et la charge allouée est $w_i$ , alors le système est stable si $\lambda_i w_i/c_i < 1$. Soit

$$\eta_i = \frac{w_i}{c_i}[1 + \sum_{j=1}^{i} \frac{p_j}{1 - p_j}]$$

Alors le débit du système est

$$TH = \min \{ 1/\eta_1 , 1/\eta_2 , ..., 1/\eta_m \} \tag{50}$$

Si les temps de traitement à la phase i ~ exp ($w_i = 1/\mu_i$ ), alors le nombre moyen de jobs dans le système est

$$EN = \sum_{i=1}^{m} \tilde{N}_{M/M/C}(\lambda\eta_i, c_i) \tag{51}$$

Soit $\pi \in P$ et TH ($\pi$) le débit du système si la phase i possède $c_{\pi(i)}$ station, une charge $w_{\pi(i)}$ et une probabilité de traitement défectueux $p_{\pi(i)}$ , i = 1, ..., m.

L'ordonnancement $\pi^*$ maximisant le débit

$$\pi^* = \arg\max \{ TH(\pi): \pi \in P \} \tag{52}$$

est tel que

$$w_{\pi^*(i)} / c_{\pi^*(i)} \geq w_{\pi^*(i+1)} / c_{\pi^*(i+1)} \quad , i = 1, ..., m-1 \tag{53}$$

Nous passons sur la preuve dont l'idée est similaire à la précédente, voir [5]. On utilise le fait que le débit est croissant en transposition adjacente en $w_i/c_i$.

On peut considérer l'ordonnancement qui minimise le WIP ( work-in-process )

$$\pi^* = \arg\min [ EN(\pi) : \pi \in P ] \tag{54}$$

$$\text{où} \qquad EN(\pi) = \sum_{i=1}^{m} \tilde{N}_{M/M/C}(\rho_{\pi(i)}(\pi), c), \ \rho_{\pi(i)}(\pi) = \frac{\lambda w_{\pi(i)}}{c_i}\left[1 + \sum_{j=1}^{i} \frac{p_{\pi(j)}}{1 - p_{\pi(j)}}\right] \tag{55}$$

et on pose $c_1 = c_2 = c_3 = \ldots = c_m \equiv c$.

On montre [5], en utilisant le fait que $\hat{N}_{M/M/C}(\rho, c)$ croissante convexe en $\rho$, que $\pi^*$ est telle que

$$\left(1 - p_{\pi(i)}\right) \geq \left(1 - p_{\pi(i+1)}\right), \ i = 1, \ldots, m-1 \tag{56}$$

**IV-4-5- Capacités des buffers limitées.**

On suppose qu'à la phase i il ne peut y avoir qu'au plus $b_i$ articles, soit un buffer de capacité $b_i - c_i$ , i = 1, ..., m. On cherche un protocole de service pour déterminer quand une station qui vient de compléter son service sur un job et l'a transféré au buffer suivant et va initier le service du job suivant. On considérera deux types de protocoles utilisés dans les lignes de flux:

- **Blocage de production:** chaque station sert toujours un job disponible et la station n'est pas bloquée.
- **Blocage de communication:** le service d'un job à la phase i-1 n'est initié que si un job est disponible et si le nombre de jobs à la phase i plus le nombre de jobs à la phase i-1 est inférieur ou égal à $b_i$. On s'intéresse particulièrement à l'allocation de la capacité des buffers.

On considère le cas $c_i = 1$ i = 1, ..., m et sauf mention contraire $b_1 = \infty$.

On suppose ici que les stations sont distribuées selon une loi exponentielle. L'extension au cas de lois générales est traité dans les références [5 ].

Soit $\hat{B}(\lambda, \mu, b)$ la probabilité de blocage et $\hat{I}(\lambda, \mu, b)$ la probabilité d'oisiveté du serveur dans M/M/1/b de taux d'arrivées $\lambda$ et de taux de service $\mu$ .

Soit $\mu_{iu}$ , $\mu_{id}$ , i = 1, ..., m ($\mu_{1u} = \mu_1$ , $\mu_{md} = \mu_m$) la solution unique [5] de

$$\frac{1}{\mu_{i-1d}} = \frac{1}{\mu_{i-1}} + \frac{1}{\mu_{id}} \hat{B}(\mu_{i-1u}, \mu_{id}, b_i) \qquad i = \overline{2,m} \tag{57}$$

et

$$\frac{1}{\mu_{iu}} = \frac{1}{\mu_i} + \frac{1}{\mu_{i-1u}} \hat{I}(\mu_{i-1u}, \mu_{id}, b_i) \qquad i = \overline{2,m} \tag{58}$$

alors le débit est approximé par

$$TH(\bar{\mu}, \bar{b}) = \mu_1[1 - \hat{B}(\mu_1, \mu_{2d}, b_{2+1}] = \mu_m[1 - \hat{I}(\mu_{m-1}, \mu_m, b_m] \tag{59}$$

où $\bar{\mu}(\mu_1, \mu_2, \ldots, \mu_m)$ et $\bar{b} = (b_1, b_2, \ldots, b_m)$.

**IV-4-6- Allocation du buffer et de la charge.**

On cherche l'allocation de la charge totale w aux m phases. Soit $1/\mu_i = w_i$ la charge de i. l'allocation optimale est

$$\mu^* = arg\,max\left\{TH(\overline{\mu},\overline{b}) : \sum_{1}^{m}\frac{1}{\mu_i} = w\,,\ \mu_i \geq 0, i = \overline{1,m}\right\} \tag{60}$$

vérifie $$\mu_{uu}(\overline{\mu^*}) = \mu_{i+1.d}(\overline{\mu^*}),\ i = \overline{1,m-1} \tag{61}$$

on peut alors obtenir le débit optimal

$$TH(\overline{\mu^*},\overline{b}) = \frac{1}{w}\sum_{j=1}^{m}(1 - k_j - k_{j+1}) = \frac{1}{w}(m-2)\sum_{j=2}^{m}\frac{1}{b_j + 2} \tag{62}$$

on cherche maintenant à dimensionner la capacité des buffers. Si (m-1)b est la capacité de stockage totale à allouer telle que $\sum_{i=1}^{m} b_i = (m-1)b,\ b_i \geq 1$, alors pour maximiser le débit, l'allocation optimale est

$$(\overline{\mu^*},\overline{b^*}) = arg\,max\left\{TH(\overline{\mu},\overline{b}) : \sum_{i=1}^{m}\frac{1}{\mu_i} = w; \mu_i \geq 0, i = \overline{1,m}; \sum_{i=2}^{m} b_i = (m-1)b;\ b_i \in \aleph, i = \overline{1,m}\right\}$$

l'allocation optimale est $b_i^* = b, i = \overline{2,m}$ et le débit (approximé) optimale

$$TH(\overline{\mu^*},\overline{b^*}) = (\frac{mb-2}{b+2})\frac{1}{w} \tag{63}$$

L'allocation de la charge est

$$w_1^* = \frac{(1-k)w}{2mk-2k+m} = w_m^*\quad,\ w_i^* = \frac{(1-2k)w}{2mk-2k+m}\quad,\ i = \overline{2,m-1} \tag{64}$$

**IV-5- Job shop dynamique.**

Le job shop consiste en un certain nombre de différents types de machines où chaque type est capable de réaliser un ensemble spécifique d'opérations de production. En général, le job shop est tel que toutes les machines de même type sont localisées ensemble (layout). Il se peut que les différents groupes de machines soient localisés dans différents bâtiments, pas forcément dans le même site (layout fonctionnel). Ces dernières années, il y a une tendance à passer du fonctionnel layout au cellular layout où les cellules sont formées (en utilisant la technologie de groupe) telles que les articles similaires sont produit dans une cellule.

Un trait caractéristique distinguant les jobs shops est que différent types de jobs avec différents routages (i.e. séquence de machines à visiter par un job) peuvent être produits.

On utilise généralement les modèles ouverts de réseaux de files d'attente. Les problèmes typiques dans un atelier job shop de type fonctionnel layout consistent à déterminer :

- Le nombre de machines pour chaque centre.
- La capacité de production pour chaque centre.
- Le nombre d'ouvriers (opérateurs) dans le système.
- L'allocation des ouvriers aux centres.

Un autre problème additionnel est de choisir entre un atelier de type fonctionnel layout et un atelier de type cellular layout. Pour cela, il faut étudier les effets de :

- la diversité des jobs en routage,
- La diversité des jobs en exigences de traitement,

sur les mesures de performance du système.

Considérons le réseau de Jackson suivant : Les jobs arrivent au job shop selon un processus de poisson de taux $\lambda$.

$\gamma_i$ = proportion de jobs qui rejoignent le centre i à leurs arrivées ($\sum_{i=1}^{m} \gamma_i = 1$ ).

m = nombre de centres.

$p_{ij}$ = proportion de jobs qui après service au centre i rejoignent le centre j.

$1 - \sum_{i=1}^{m} p_{ij}$ = proportion de jobs qui quittent le centre i vers l'extérieur (après service).

Les temps de service au centre i sont des v.a.i.i.d ~ Exp ($\mu_i$), i = 1, ..., m.

Les services et arrivées et routages sont indépendants.

$\mu_i r_i(m)$ = taux de service d'un job au centre i lorsqu'il y a n jobs présents.

Si le centre i possède $c_i$ machines en parallèle, $r_i$ (n) = min (n, $c_i$). Les centres sont FIFO.

En vertu de la propriété de décomposition, chaque centre i représente un modèle M/M/$c_i$ indépendant de paramètre $\lambda_i$ *et* $\mu_i r_i(n)$ *où les* $\lambda_i, i = \overline{1,m}$ sont solution du système d'équations linéaires algébriques suivant :

$$\lambda_i = \lambda\gamma_i + \sum_{j=1}^{m} \lambda_j p_{ji}, \quad i = \overline{1,m} \tag{65}$$

**IV-5-1- Allocation optimale des ouvriers aux centres.**

On a un total de c ouvriers pour les m centres. Chaque ouvrier peut opérer à l'une des $\sum_{i=1}^{m} k_i$ machines, et au plus une machine peut être affectée à un ouvrier.

Il y a $k_i$ machines identiques au centre i. Si $\sum k_i \leq c$ on peut allouer les ouvriers un par un à toutes les $\sum k_i$ machines. Mais en général $\sum k_i \geq c$.

Supposons qu'on ait affecté $c_i$ ouvriers au centre i, alors le nombre total de jobs dans le système (supposé être un réseau de Jackson) est

$$EN = EN_{M/M/C}(\lambda_i, \mu_i, c_i) \tag{66}$$

où $EN_{M/M/C}(\lambda_i, \mu_i, c_i)$ est le nombre moyen de jobs dans un système M/M/C de paramètres $\lambda_i$ $\mu_i$, i = 1, ..., m.

l'allocation optimale des ouvriers est :

$$c^* = arg\,min\left\{\sum_{i=1}^{m} EN_{M/M/C}(\lambda_i, \mu_i, c_i) : \sum_{i=1}^{m} c_i = C, 1 \leq c_i \leq k_i, 1 \leq i \leq m\right\} \tag{67}$$

puisque $EN_{M/M/C}(\lambda_i, \mu_i, c_i)$ est décroissante convexe en $c_i$, alors on peut utiliser l'algorithme d'allocation marginale.

Soit $c_i^*(k)$, $i = \overline{1,n}$ une solution optimal où $C=k$

Si

$$j^* = arg\,max\left\{EN_{M/M/C}\left(\lambda_i, \mu_i, c_i^*(k)\right) - EN_{M/M/C}\left(\lambda_i, \mu_i, c_i^*(k+1)\right) : c_i^*(k+1) \leq k_i, i = \overline{1,m}\right\}$$

est le centre qui garantit la réduction maximale du nombre de jobs grâce à un serveur additionnel, alors on pose :

$$c_{j^*}^*(k+1) = c_{j^*}^*(k) + 1 \tag{68}$$

$$c_{i^*}^*(k+1) = c_{i^*}^*(k), \quad i = \overline{1,m}, i \neq j^* \tag{69}$$

Ces résultats permettent également de trouver la réaffectation des ouvriers disponibles durant l'absentéisme des ouvriers.

Si $c_{i^*}^*(k+1), i = \overline{1,m}$, est une allocation avec $k+1$ ouvriers, et l'un d'eux (disons celui qui est affecté au centre $j$) est absent, soit $j^*$ défini comme précédemment, alors l'allocation optimale des ouvriers est obtenue en conservant la location précédente avec un ouvrier du centre $j^*$ à affecter au centre $j$.

### IV-5-2- Nombre d'ouvriers :

La formulation précédente permet de déterminer le nombre optimal d'ouvriers à engager. Supposant que $h$ est le coût unitaire de recrutement d'un nouvel ouvrier relativement au coût de possession. On cherche :

$$c^* = arg\,min\left\{ hc + \sum_{i=1}^{m} EN_{M/M/C}(\lambda_i, \mu_i, c_i) : \sum_{i=1}^{m} c_i = c, \quad 1 \le c_i \le k_i, \quad i = \overline{1,m} \right\} \qquad (70)$$

ce problème peut être résolu par l'algorithme d'allocation marginale.

Soit $c_i^*(k), i = \overline{1,m}$ l'allocation optimale des ouvriers où $c=k$. alors le nombre optimal d'ouvrier est donné par :

$$c^* = min\left\{ c : h \ge EN_{M/M/C}(\lambda_i, \mu_i, c_i^*(c)) - EN_{M/M/C}(\lambda_i, \mu_i, c_i^*(c)+1), i = \overline{1,m}, \quad c \ge m \right\} \quad (71)$$

### IV-5-3- Affectation des tâches aux cellules de machines :

Supposons qu'il n y a qu'un ouvrier à chaque cellule. La manière dont les tâches sont affectées aux cellules détermine le nombre moyen de visites $v_i$ à la cellule $i$ par un job arbitraire. De quelque manière qu'on alloue les tâches $\sum_{i=1}^{m} v_i = K$ reste constant où $K$ est le nombre moyen de tâches (opérations) requises pour un job similaire.

On cherche à obtenir une allocation optimale des tâches aux cellules de travail de telle manière que le *flow-time* moyen 'temps moyen de séjour) d'un job quelconque où le coût moyen de possession soit minimal. Puisque $\lambda ET=EN$, alors toute allocation minimisant $EN$ minimise $ET$ et vis vers sa. On minimisera $EN$.

Identifions d'abord les valeurs «idéales »de $v_i\ i=1,m$ qui minimisent $EN$. Elles doivent être telle qu'on ne puisse trouver d'allocation des tâches où chaque tâche n'est assignée qu'a une seule cellule. Pour cela, nous trouverons une méthode d'allocation simple pour identifier les tâches qui doivent être allouées à plus d'une cellule.

On suppose qu'une tâche affectée à la cellule $i$ prend un temps d'opération $\sim Exp\ (1/\mu_i)$. avec cette hypothèse, on voit que pour l'allocation d'une tâche qui résulte d'un nombre moyen de visites $v_i$ à la cellule $i$, l'utilisation du serveur à la cellule $i$ est $\rho i=\lambda v_i/\mu_i,\ i=1,m$. par conséquent :

$EN = \sum_{i=1}^{m} \lambda v_i / (\mu_i - \lambda v_i)$ et les valeurs idéales de $v_i$

$$\overline{v^*} = arg\,min\left\{ \sum_{i=1}^{m} \frac{\lambda v_i}{\mu_i - \lambda v_i} : \sum_{i=1}^{m} v_i = K,\ v_i \geq 0, i = \overline{1,m} \right\} \quad \text{.....(72)}$$

La fonction objectif est séparable convexe, alors en utilisant la condition de Kun-Tucker, les valeurs optimales de $v_i$ *sont données par*

$$v_i^* = \frac{1}{\lambda}\left( \mu_i - \left( \sum_{j=1}^{m} \mu_j - \lambda K \right) \frac{\sqrt{\mu_i}}{\sum_{j=1}^{m} \sqrt{\mu_j}} \right),\ i = \overline{1,m} \qquad (73)$$

On suppose que $\lambda K < \sum_{j=1}^{m} \mu_j$ . Dans le cas contraire, aucune allocation des tâches ne conduit à la stabilité du système. $\sum \mu_j - \lambda K$ peut être interprétée comme une mesure de disponibilité pour l'atelier.

On voit de (73) que l'allocation des tâches équitable telle que $v_i / \mu_i = cste$ devient préférable.

Soit $\alpha$={1, ..., L}l'ensemble des L tâches à allouer aux m cellules. Soit $w_i$ le nombre moyen de fois où la tâche i doit traiter un job quelconque. Alors $\sum_{i=1}^{L} w_i = K$ .

Supposons qu'il existe une partition $\{s_1, ..., s_m\}$de $\alpha$ telle que $\sum_{j \in s_i} w_1 = v_i^*, i = \overline{1,m}$ .

Alors l'affectation des tâches $j \in s_i$ à la cellule i donnera une allocation optimale. L'identification d'une telle partition, si elle existe, prendra un temps exponentiel par rapport à L. Par conséquent, on suggère l'allocation simple suivante qui donne le $v_i$* optimale, mais au prix de dépenses d'allocation de quelques unes (au plus m) des L tâches au plus d'une cellule.

Trouver l(k) ; k = 1, ..., m tel que l(k) est la plus grande valeur satisfaisant

$$\sum_{i=1}^{l(k)} w_i \leq \sum_{j=1}^{k} v_j^*,\ k = \overline{1,m} \qquad (74)$$

$$l(m) = L \qquad (75)$$

Si $\sum_{j=1}^{k} v_j^* - \sum_{i=1}^{l(k)} w_j > 0$, alors la tâche l(k)+1 est allouée aux deux cellules k et k+1. La proportion de jobs nécessitant la tâche l(k)+1 à diriger vers la cellule k est

$$\left( \sum_{j=1}^{k} v_j^* - \sum_{i=1}^{l(k)} w_i \right) \Big/ w_{l(k)+1} \tag{76}$$

La fonction restante est orientée vers la cellule k+1 pour traiter la tâche l(k)+1, k = 1, ..., m-1.

**IV-5-4- Réseaux généraux ouverts.**

On utilise comme précédemment les approximations pour le modèle G/G/C

$$\begin{aligned} EN &= \sum_{i=1}^{m} \hat{N}(\lambda_i, C_{A_1}^2, \mu_i, C_{S_i}^2) \\ ET &= \sum_{i=1}^{m} v_i \hat{T}(\lambda_i, C_{A_1}^2 \mu_i, C_{S_i}^2) \end{aligned} \tag{77}$$

où N est le nombre total de jobs dans le système et T le temps moyen de séjour d'un job (mean flow time).

$\lambda$ est le taux d'arrivées externes,

$C_A^2$ , le carré du coefficient de variation (cvv) du temps entre arrivées externes ,

$v_i = \lambda_i / \lambda$ est le nombre moyen de visites d'un job au centre i,

$\lambda_i$ = taux d'arrivées au centre i,

les $\lambda_i$ sont solution du système d'équations linéaires algébriques

$$\lambda_i = \lambda \gamma_i + \sum_{j=1}^{m} \lambda_j p_{ij} \qquad , i = \overline{1,m} \tag{78}$$

On a donc les paramètres suivants $\lambda, C_A^2, p = \|p_{ij}\|, \gamma_i, \mu_i, C_{S_i}^2, c_i, \lambda_i, \quad i = \overline{1,m}$

$$C_{D_1}^2 = 1 + (1 - \rho_i^2)(C_{A_i}^2 - 1)\frac{\rho_i^2}{c_i}(C_{S_i}^2 - 1), \quad i = \overline{1,m} \tag{79}$$

Cette approximation est légèrement différente de celle utilisée précédemment. Ceci donne m équations pour 2m inconnues, $C_{D_1}^2, c_{A_i}^2, \quad i = \overline{1,m}$. Pour obtenir m équations supplimentaires, on procède en deux étapes.

Spitting. Le flux de départ de chaque centre (disons i) est divisé en différents flux selon les routages des jobs.

Chaque branche est un flux tamisé (thinning point process) où un job est retenu avec une probabilité $p_{ij}$ si la destination est j. Il est connu que le flux tamisé obtenu par l'opération p-thinning d'un processus de renouvellement de cvv des temps entre pannes C est encore un processus de renouvellement de cvv $pc^2+1-p$. Par conséquent, le cvv du flux des jobs à l'arête (ij) peut être approximé par $1-p_{ij}+p_{ij}\,C_{D_i}^2$, $i=1,\ldots,m,\ i\neq j$.

Composition. Elle consiste en la superposition des flux dirigés vers un centre.

Pour un processus de renouvellement de MTBF = $1/\hat{\lambda}$ et de cvv $\hat{c}^2$, on a

$$\lim_{t\to\infty} var(N(t))/t = \hat{\lambda}\hat{c}^2 .$$

On approchera le cvv du flux composé des processus d'arrivées (supposés de renouvellement) au centre i par

$$C_{A_i}^2 = \frac{1}{\lambda_i}\sum \lambda_j p_{ij}\left[p_{ji}C_{D_j}^2 + (1-p_{ji}) + \frac{\lambda\gamma_i}{\lambda_i}\left[\gamma_i C_A^2 + (1-\gamma_i)\right]\right],\quad i=\overline{1,m} \qquad (80)$$

ceci donne les m équations manquantes pour déterminer $C_{A_i}^2$ *et* $C_{D_i}^2$, $i=\overline{1,m}$.

Si on considère un job shop dynamique à centres multiples avec une machine par centre, alors le temps moyen passé au centre i par un job arbitraire peut être approché par

$$ET_i = \hat{T}\left(\lambda_i, C_{A_i}^2, \mu_i, C_{S_i}^2, 1\right) \qquad (81)$$

De (79) et (80), on obtient donc $C_{A_i}^2$ *et* $C_{D_i}^2$, $i=\overline{1,m}$ et $\lambda_i$, i =1, ..., m en résultant

$$\begin{aligned}
&\lambda_i = \lambda\gamma_i + \sum_{j=1}^{m}\lambda_j p_{ji},\quad i=\overline{1,m}\\
&C_{D_i}^2 = (1-\rho_i^2)C_{A_i}^2 + \rho_i^2 C_{S_i}^2,\quad i=\overline{1,m} \qquad (82)\\
&C_{A_i}^2 = \sum_{j=1}^{m}\frac{\lambda_j p_{ji}}{\lambda_i}\left(p_{ji}C_{D_j}^2 + (1-P_{ji})\right) + \frac{\gamma_i}{\lambda_i},\quad i=\overline{1,m}
\end{aligned}$$

Ces équations peuvent être utilisées pour approximer le flow time moyen $ET = \sum_{i=1}^{m} v_i ET_i$, et le nombre moyen de jobs dans le système EN = λ ET.

**IV-5-5- Type de routage.**

La littérature comporte divers types de routages, parmi lesquels :

- Routages symétriques : $p_{ij} = p_{ji} \quad \forall i, j = \overline{1,m}$. Dans ce cas toutes les stations sont équitablement visitées i.e. un job qui quitte un centre a la même probabilité de se rendre vers l'une des m-1 centres ou de quitter le système : $p_{ij} = 1/m \quad i \neq j, i, j = \overline{1,m}; \; p_{ii} = 0, i = \overline{1,m}$ on vérifie dans ce cas que $\lambda_i \equiv \lambda \quad \forall i$, et d'après (80) $C_{A_i}^2 \to 1 \; lorsque \; m \to \infty$.

On peut alors simplifier l'approximation précédente en remplaçant $\hat{T}\left(\lambda_i, C_{A_I}^2, \mu_i, C_{S_i}^2, 1\right)$ dans (81) par $\left[\rho_i^2(1+C_{S_i}^2)\right]\left[2\lambda_i(1-\rho_i)\right]+1/\mu_i$, le temps moyen de séjour d'un job dans un système M/G/1. D'où l'approximation :

$$ET = \sum_{i=1}^{m}\left(\frac{\rho_i^2(1+C_{S_i}^2)}{2\lambda_i(1-\rho_i)}+\frac{1}{\mu_i}\right) \tag{83}$$

- Routage uniforme : Tous les jobs suivent la même séquence de centres. On peut donc poser, $\gamma_1 = 1, \; p_{i,i+1} = 1, \quad i = \overline{1,m-1}$ et les jobs quittant le centre m quittent le système immédiatement. Ici aussi, $\lambda_i = \lambda, \quad i = \overline{1,m}$ . Cela correspond à une ligne de flux étudiée précédemment.

- Routage avec serveur central. C'est un modèle intéressant où les jobs allant à un centre pour les futures opérations, se dirigent d'abord vers un serveur spécial pour transport, accessoires...

- routage dynamique : Les jobs sont orientés vers la station qui possède la plus petite file d'attente (en probabilité) (probabilistic shortest-queue routing).

Ces types de routage sont intéressants car ils conduisent à des réseaux réversibles.

**IV-5-6- Diversité des jobs en routage.**

Considérons le cas de job shops chargés, $\rho_i \approx 1, \forall i$. On suppose $c_i = c$ et $S_I = S \; \forall \; i$ i.e. $\rho_i \equiv \rho ES/C \quad \forall i, \; v_i \approx 1 \quad \forall i$.

La matrice de routage P et le vecteur de la première opération $\bar{\gamma}$ reflètent la diversité des routages. On cherche donc la diversité optimale des jobs en routage qui minimise le flow time moyen

$$\left(\bar{\gamma},P\right)=min\left\{\begin{array}{l} ET\left(\bar{\gamma},P\right):\sum_{i=1}^{m}\gamma_i=1;\sum_{j=1}^{m}p_{ji}\le 1,i=\overline{1,m};\\ \left(p_{ij}\ge 0,j=\overline{1,m};\gamma_i+\sum_{j=1}^{m}p_{ji}=1;\gamma_i\ge 0;p_{ii}=0\right),i=\overline{1,m}\end{array}\right\} \tag{84}$$

La dernière contrainte signifie que le nombre moyen de visites à i est égal à 1. On utilise l'approximation (23)

$$\hat{T}(\lambda,C_A^2,\mu,c)=\left(\frac{C_A^2+\rho^2C_S^2}{1+\rho^2C_S^2}\right)\left(\frac{1+C_S^2}{2\lambda}\right)EL_{M/M/C}(\lambda,\mu,c)+E(S) \tag{85}$$

et on trouve

$$ET=\left(\frac{\sum_{i=1}^{m}C_{A_I}^2+m\rho^2C_S^2}{1+\rho^2C_S^2}\right)\left(\frac{1+C_S^2}{2\lambda}\right)EL_{M/M/C}(\lambda,\mu,c)+mE(S) \tag{86}$$

par conséquent, il est équivalent de minimiser $ET$ ou $\sum_{i=1}^{m}C_{A_i}^2$. De (79) et (80), on voit que pour

$\rho_i\approx 1$

$$\sum_{i=1}^{m}C_{A_i}^2=\sum_{i=1}^{m}\left\{\sum_{j=1}^{m}\left[p_{j,i}\left(p_{j,i}\cdot C_S^2+\left(1-p_{j,i}\right)\right)\right]+\gamma_i.\left(\gamma_i.C_A^2+\left(1-\gamma_i\right)\right)\right\}=$$

$$m-\left\{\sum_{i=1}^{m}\left(1-C_S^2\right)\left(\sum_{j=1}^{m}p_{ji}^2+\gamma_i^2\right)+\gamma_i^2\left(C_A^2+C_S^2\right)\right\}$$

supposant que : $C_A^2\ge C_S^2$ *et* $C_S^2\le 1$. *alors,*

$$\left(1-C_S^2\right)+\frac{C_A^2-C_S^2}{m}\le\sum_{i=1}^{m}\left\{\left(1-C_S^2\right)\cdot\left(\sum_{j=1}^{m}p_{ji}^2+\gamma_i^2\right)+\gamma_i^2\left(C_A^2-C_S^2\right)\right\}\le m\left(1-C_S^2\right)+C_A^2-C_S^2$$

d'autre part , si $C_A^2 \le C_S^2$ *et* $C_S^2 \ge 1$, les inégalités ci-dessus sont inversées . la limite inférieure de l'inégalité ci-dessus est atteinte par le routage symétrique des jobs (*i.e.* $p_{ji} = \frac{1}{m}$ , $i \ne j$ ,$i,j = \overline{1,m}$ ;$\gamma_i = \frac{1}{m}$ ,$i = \overline{1,m}$), et la limite supérieure est atteinte pour un routage uniforme des jobs (i.e.$\gamma_1 = 1$, $p_{i,i+1} = 1, i = \overline{1,m-1}$, $p_{ij} = 0$ *ailleurs).* Donc on a la conclusion suivante :

1. Si $C_A^2 \ge C_S^2$, *et* $C_S^2 \le 1$, alors le routage uniforme des jobs minimise le flow time moyen, et le routage symétrique maximise le MFT.
2. Si $C_A^2 \le C_S^2$, *et* $C_S^2 \ge 1$, alors le routage symétrique minimise le MFT et le routage uniforme maximise le MFT.

**IV-5-7- Job shops généraux multiclasses.**

**Hypothèses :**

- r classes de jobs

  - Les jobs de classe l arrivent selon un processus de renouvellement de MTBF = 1/λ(l) et $ccv = C_A^{2(l)}$

  - $p_{ij}^{(l)}$ = probabilité de routage d'un client de classe l du centre i au centre j

$$p^{(l)} = \left\| p_{ij}^{(l)} \right\|_{mxm}, \quad l = \overline{1,r}$$

- Les temps de traitement des jobs de classe l sont des v.a.i.i.d. de paramètres :

moyenne : $1/\mu_i^{(l)}$ $\quad ccv : C_{S_i}^{2(l)}$ , $\quad i = \overline{1,m}, \quad l = \overline{1,r}$

- le service des jobs à chaque centre suit la discipline FIFO.

L'analyse de tels modèles est complexe. Au lieu de développer des méthodes numériques sophistiquées, on préfère de donner des résultats approximatifs simples et facilement interprétables. On distingue deux types d'approximations l'agrégation et la décomposition.

Agrégation.

On suppose que les routages des jobs d'un centre à un autre sont similaires pour toutes les classes de jobs. De plus, le temps de traitement pour toutes les classes de jobs à chaque centre sont similaires i.e.

$\left|\mu_i^{(l)} - \mu_i^{(l')}\right|$ *et* $\left|C_{S_I}^{2(l)} - C_{S_I}^{2(l')}\right|$ ne sont pas très grand $\forall l,l' = \overline{1,r}$ .

On peut donc agréger toutes les classes en une seule et obtenir les paramètres des temps d'interarrivées correspondants au centre i : *moyenne* $1/\lambda$ *et ccv* : $C_{A_i}^2, i = \overline{1,m}$ .

Si $\frac{1}{\mu_i}$ *et* $C_{s_i}^2$ sont la moyenne et le CWV des jobs agrégés au centre i, alors

$$\frac{1}{\mu_i} = \frac{1}{\lambda_i}\sum_{l=1}^{r}\frac{\lambda_i^{(l)}}{\mu_i^{(l)}} \quad i = \overline{1,m} \qquad et \qquad C_{s_i}^2 = \left(\frac{\mu_i^2}{\lambda_i}\right)\sum_{l=1}^{r}\frac{\lambda_i^{(l)}}{\left(\mu_i^{(l)}\right)^2}\left(1 + C_{s_i}^{2(l)}\right) \quad i = \overline{1,m} \qquad (87)$$

ou

$\lambda_i^{(l)}$ = taux d'arrivées des jobs de classe l au centre i.

$\mu_i^{(l)}$ = leurs taux de services .

$\lambda_i = \sum_{l=1}^{r}\lambda_i^{(l)}$ = taux d'arrivées des jobs agrégés au centre i.

la matrice de routage des jobs agrégés $P = \|P_{ij}\|$ avec $P_{ij} = \frac{1}{\lambda_i}\sum_{l=1}^{r}\lambda_i^{(l)} p_{ij}^{(l)}$ $i,j = \overline{1,m}$

ces paramètres des jobs agrégés sont utilises avec (79) , (80)pour obtenir $\lambda_i, C_{s_i}^2, i = \overline{1,m}$

Maintenant , chaque centre est modélisé par un système 6I/6I/C multiclasses avec :

$\frac{\lambda_i^{(l)}}{\lambda_i}$ = la proportion de jobs de classe l , l= $\overline{1,r}$

soit $\hat{N}^{(l)}\left(\lambda_i, \mu_i, C_{s_i}^2, C_{s_i}^{2(l)}, C_{s_i}^{2(l')}, \frac{\lambda_i^{(l')}}{\lambda_i}, l' = \overline{1,m}\right)$, le nombre approximatif de jobs de classe l dans ce système de file d'attente .

Alors, le nombre total de jobs de classe l dans le système est approché par

$$EN^{(l)} = \sum_{i=1}^{m}\hat{N}^{(l)}\left(\lambda_i, \mu_i^{(l')}, C_{A_i}^2, C_{S_i}^{2(l')}, \lambda_i^{(l')}/\lambda_i, l' = \overline{1,r}, l = \overline{1,r}\right) \qquad (88)$$

et le flow time moyen d'un job de classe l est

$$ET^{(l)} = \hat{N}^{(l)}/\lambda^{(l)}, \quad l = \overline{1,r} \qquad (89)$$

Décomposition.

Elle est utilisée lorsque les contraintes de traitement des différentes classes de jobs sont différentes. On s'intéresse à la classe l. la méthode de décomposition assimile le système à une seule classe de réseau ouvert général.

- Les temps de service au centre i sont des v.a.i.i.d. de moyenne $\frac{1}{\mu_i}$ *et de cvv* $C_{S_i}^2$.

- La matrice de routage $P^{(l)}$ de cette classe unique est celle de la classe l.

On approche le centre i par un système M/GI/1 multiclasse-FIFO pour déterminer $\mu_i$ *et* $C_{S_i}^2$.

$$EN^{(l)} = \frac{\lambda^{(l)} \sum_{k=1}^{r} \lambda^{(k)} (1 + C_S^{2(k)}) / (\mu^{(k)})^2}{2(1-\rho} + \rho^{(l)}, \quad l = \overline{1,r} \tag{90}$$

Si on doit représenter la performance de la classe l par celle d'un système M/GI/1 à une classe de temps de service moyen 1/μ et cvv = $C_S^2$ , on doit avoir

$$EN^{(l)} = \frac{\lambda^{(l)}(1 + C_S^2)}{2(1 - \lambda^{(l)}/\mu)} + \frac{\lambda^{(l)}}{\mu} \tag{91}$$

On a deux inconnues $\mu$ *et* $C_S^2$ et une seule équation. Pour obtenir la deuxième équation, on observe que le serveur du M/G/1 multiclasse est oisif une fraction de temps égale à

$$1 - \sum_{\substack{k=1 \\ k \neq l}}^{r} \lambda^{(k)} / \mu^{(k)}$$

Donc, le temps moyen nécessaire au traitement d'un job de classe l avec une contrainte de traitement x est $x \Big/ \left( 1 - \sum_{\substack{k=1 \\ k \neq l}}^{r} \lambda^{(k)} / \mu^{(k)} \right)$. On pose alors

$$\mu = \mu^{(l)} \left( 1 - \sum_{\substack{k=1 \\ k \neq l}}^{r} \lambda^{(k)} / \mu^{(k)} \right) = \mu^{(l)} (1 - \rho + \rho^{(l)} \tag{92}$$

où $\rho^{(l)} = \lambda^{(l)} \Big/ \mu^{(l)}$ *et* $\rho = \sum_{k=1}^{r} \rho^{(k)}$, et

$$C_S^2 = 2\left(\frac{1-\rho}{1-\rho+\rho^{(1)}}\right)\left[\frac{\sum_{k=1}^{r}\lambda_i^{(k)}(1+C_S^{2(k)})\ (\mu^{(k)})^2}{2(1-\rho)} - \left(\frac{\rho-\rho^{(1)}}{1-\rho+\rho^{(1)}}\right)\frac{1}{\mu^{(1)}}\right] - 1 \qquad (93)$$

On se basant sur cette représentation équivalente, la décomposition pour la classe l consiste à remplacer les paramètres de service du centre i pour le modèle à une classe dans les formules (92) et (93)

$$\mu \to \mu_i \quad , \quad \rho \to \rho_i \quad , \quad C_S^{2(l)} \to C_{S_i}^2$$
$$\mu^{(l)} \to \mu_i^{(l)} \quad , \quad \rho^{(l)} \to \rho_i^{(l)} \qquad (94)$$

**IV-5-8- Diversité des jobs en temps de traitement.**

Soit J une collection de classes de jobs. Les jobs de classe l possède au centre i un temps moyen de service $E(S_i^{(l)})$, *et un cvv* $C_{S_i}^{2(l)}$ ,$l = 1,2,..,r$ ,$i = \overline{1,m}$ .

Soit $\hat{J}$ une collection avec $E(\hat{S}_i^{(l)})$, $\hat{C}_{S_i}^{2(l)}$ ,$l = \overline{1,r}$ ,$i = \overline{1,m}$.

Puisque la variabilité en temps de service est causée par la machine ou l'ouvrier du centre, on suppose que

$$C_{S_i}^{2(l)} = \hat{C}_{S_i}^{2(l)} = \hat{C}_{S_i}^2 , \qquad l = \overline{1,r} ; i = \overline{1,m} \qquad (95)$$

On suppose que les matrices de routage et vecteurs de la 1ere opération sont les mêmes pour les deux collections de classe de jobs. Soit $\lambda_i^{(l)}$ = taux d'arrivées des jobs de classe $l$ au centre i . On cherche à étudier l'effet du temps moyen de service des jobs sur le flow time moyen lorsque la charge du centre i est la même pour les deux collections $J$ et $\hat{J}$ .

Si on agrège les classes, alors les jobs agrégés auront un temps de traitement moyen

$$\frac{1}{\mu_i} = \frac{1}{\lambda_i}\sum_{l=1}^{r}\lambda_i^{(l)}E(S_i^{(l)}) \qquad (96)$$

le même pour les deux collections, et un cvv

$$C_{S_I}^2 = \left(\frac{\mu_i^2}{\lambda_i}\right)\left(1 + \tilde{C}_{S_i}^2\right)\sum_{l=1}^{r} \lambda_i^{(l)}\, E(S_i^{(l)})^2 - 1, \quad i = \overline{1,m} \quad \textbf{pour la collection } J \tag{97}$$

$$\hat{C}_{S_I}^2 = \left(\frac{\mu_i^2}{\lambda_i}\right)\left(1 + \tilde{C}_{S_i}^2\right)\sum_{l=1}^{r} \lambda_i^{(l)}\, E(\hat{S}_i^{(l)})^2 - 1, \quad i = \overline{1,m} \quad \textbf{pour la collection } \hat{J} \tag{98}$$

si $$\left\{E(S_i^{(l)}),\ l = \overline{1,r}\right\} \leq_m \left\{E(\hat{S}_i^{(l)}),\ l = \overline{1,r}\right\} \tag{99}$$

où $\leq_m$ est l'ordre en majoration, alors d'après (97), (98) on a $C_{S_i}^2 \leq \hat{C}_{S_i}^2$ (100)

Si (94) est satisfaite $\forall$ i=1, ..., m, on dit que la collection de jobs de classe $\hat{J}$ est plus diversifiée en temps de traitement que la collection J.

Puisque le flow time moyen dans un job shop à une classe est croissant en $C_{S_i}^2$, alors on peut formuler la conclusion suivante :

Le flow time moyen d'un job dans job shop est croissant en diversité des temps de traitements des jobs.

# CHAPITRE V
# ANALYSE DES PERFORMANCES D'UN SYSTEME MANUFACTURIER AVEC UN SEUL APPAREIL DE MANUTENTION

# *Chapitre V*

## ANALYSE DES PERFORMANCES DES FMS AVEC UN SEUL APPAREIL DE MANUTENTION

### V-1- La signification du modèle.

Parmi les différentes configurations FMS; nous attirons l'attention sur la configuration FMS avec un seul appareil de manutention discret(MHD). Quoique; cette configuration semble apparaître comme une configuration limitée ou particulière; elle est toutefois signifiante pour plusieurs raisons :

1- Cette configuration est courante en industrie: plusieurs cas peuvent être trouvés

a- FMS avec un stockeur central et un système de récupération (AS/RS)

b- Une cellule flexible avec robot comme système de manutention

c- Certaines configurations de lignes « electroplating » flexibles

2- Cette configuration simplifie le contrôle du système de manutention:

Les gains dans les systèmes manufacturiers résultent de la simplification des opérations; particulièrement celles liées à la manutention . L'idée dans cette configuration est l'utilisation d'une seule ressource de manutention; ce qui simplifie considérablement la fonction de contrôle qui est appelée à prévenir les problèmes de blocages et de collision

(la première application : les lignes « electroplating » où le remplacement des robots de la configuration initiale par un seul robot plus performant a permis la simplification du contrôle des opérations de manutention).

3- L'utilisation d'outils analytiques est possible pour l'analyse de cette configuration :

Cette configuration de FMS peut être analysée par des modèles analytiques. A priori; les modèles analytiques n'existent pas ; mais comme nous le verrons , ces modèles peuvent être construits par block(ou module). Chaque étape de construction servira de base pour la construction de l'étape suivante .Par ailleurs; l'utilisation de modèles analytiques permettra l'analyse de plusieurs classes différentes en conditions d'exploitations.

4- Cette configuration présente un module(ou block) dans des systèmes plus complexes :Plusieurs systèmes manufacturiers utilisent des cellules pour la réalisation des opérations. Elles

présentent beaucoup d'avantages tels (simplification des flux de matières, réduction des temps d'attente, réduction des encours).La configuration que nous analysons peut représenter une cellule flexible dans un système complexe. Ainsi, pour des systèmes complexes la démarche est l'analyse des modules(cellules) puis l'étude de leurs interactions.

**V-2- Le modèle analytique.**

Comme nous l'avons déjà évoqué; ces configurations de FMS sont devenus très utilisés en industrie. Les investissements engagés lors de leurs conception sont très importants. Ceci implique des efforts d'analyse continus pour assurer l'efficacité des opérations. Cette analyse est indispensable pour l'évaluation des différentes configurations: Différent schémas, différents routages, différents systèmes de manutention.

Une analyse rapide et interactive doit évaluer plusieurs cas de configurations avec une précision admissible. Les modèles analytiques offrent cette rapidité en temps de réponse et un bon degré de précision pour l'estimation des mesures de performances qui aident à l'évaluation des différentes configurations.

Dans notre cas ; nous utiliserons les méthodes d'analyses de réseaux de files d'attentes autant qu'elles offrent un bon compromis entre précision et rapidité tout en tenant compte des aspects stochastiques et de la dynamique du système.

Ces modèles analytiques peuvent être utilisés pour répondre aux questions de type « what if . » telles que: Quel est le nombre de palettes dans le système? , Quel est le meilleur MHD en terme de vitesse et accélération.

Notre objectif est de développer un outil analytique pour analyser les performances du système. Un outil offrant à la fois une bonne précision et une rapidité.

**V-3- Caractéristiques de la configuration.**

La configuration que nous analysons a les caractéristiques suivantes:

- Réseau fermé: Le système est réseau fermé parce que le nombre de palettes est fixe.
- Interférence ou blocage du MHD : Quand l'opération est achevée, la station ne peut décharger la palette avant l'arrivée du MHD.
- Etat de routage dépendant : La destination de la palette dépend de l'état de la prochaine station.

**V-4- Réseaux fermés :**

Plusieurs méthodes d'analyse des réseaux fermés sont basées sur (MVA) analyse des valeurs moyennes.

MVA estime le temps de réponse comme la somme des temps de services et d'attentes.

Elle calcule les mesures de performance en commençant avec une configuration de zéro palettes; et itérativement jusqu'au nombre de palettes dans le système. Pour un réseau fermé avec des distributions exponentielles et un routage selon une loi de Bernouli, MVA calcule exactement les mesures de performance.

**V-5- Manutention.**

Dans plusieurs modèles analytiques des systèmes manufacturiers, le système de manutention est ignoré ou modélisé comme une station centrale avec un ou plusieurs transporteurs avec des temps de retard. Considérer le système de manutention comme un serveur central est approprié dans le cas de MHS continu tel un convoyeur. Cependant, ceci résulte des données de routage. La probabilité pour qu'un produit parte à la prochaine station est indépendante de la station d'où le produit arrive. Dans le cas d'un système de manutention discret, l'analyse par serveur central n'est pas appropriée puisque : le convoyeur n'est pas libre de suite à chaque fois qu'une demande à son service est générée.

Peu de modèles analytiques analysent des ressources de manutention discrète, tel AGVS et AS/RS.

**V-6- LA CONFIGURATION DU SYSTEME**

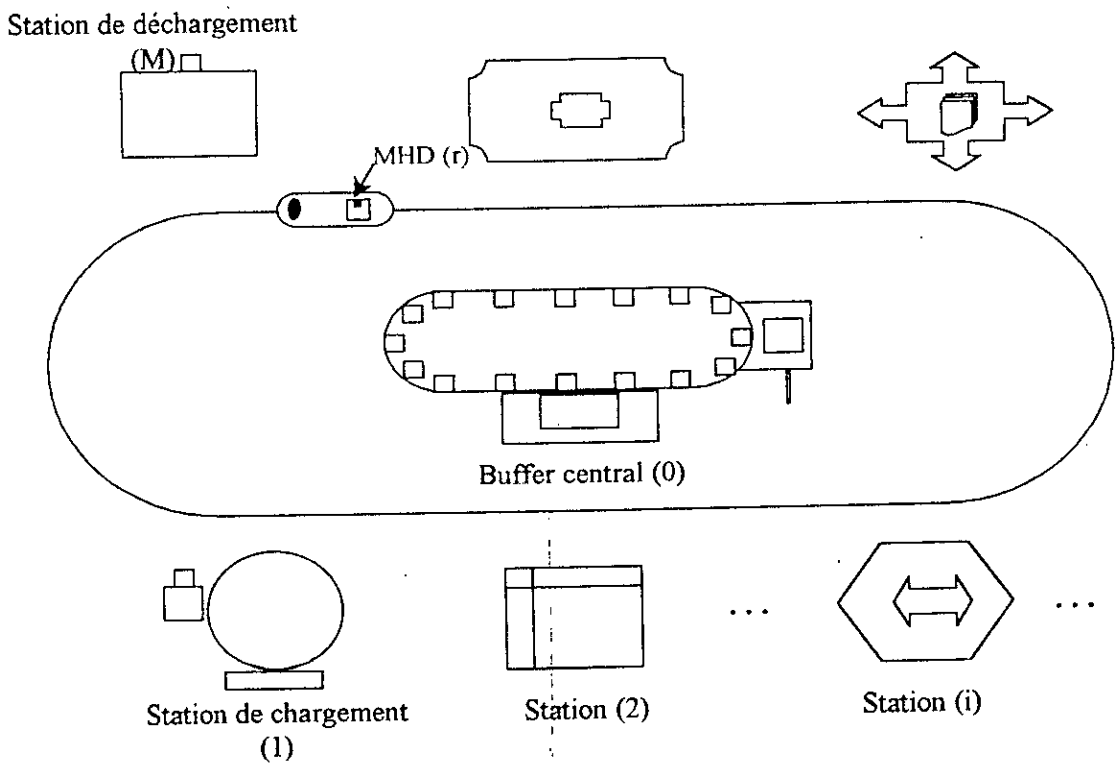



***Figure (IV.1) Configuration du modèle***

palettes circulent dans le système qui est constitué de M stations de traitement (y compris les stations de chargement et de déchargement )est un seul appareil de manutention (MHD).

Les pièces sont chargées sur les palettes et transportées d'une station à une autre station par le MHD dans l'ordre de leur plan de traitement. Le système a un buffer central assez grand pour stocker toutes les N palettes. Quand la palette se dirige vers la prochaine station et celle ci est occupée le MHD la transporte vers le buffer central. Plus tard quand la station sera libre, le MHD la ramène du buffer vers la station.

Après que le traitement est fini à toutes les stations, la pièce est déchargée à la station de déchargement (M) et le MHD déplace les palettes déchargées à la station de chargement(1).

A celle ci une nouvelle pièce sera chargée. Les pièces entrent et quittent le FMS avec le nombre de palettes qui circulent dans le système qui reste constant. Parce que la quantité de pièces produites par le FMS est la même que dans la station de déchargement. On suppose que toutes les pièces ont le même routage. Ce routage consiste en K étapes, Ni dénote la station visitée à la ieme étape. Les opérations de chargement et de déchargement sont inclues explicitement dans ce routage. Le temps moyen de traitement à la ieme étape est désigné par $\theta$i unités de temps et le mouvement de la station i à la station j prend $\delta$ ij unités de temps. Le déplacement de station i au buffer centrale est désigné par $\delta$ i0 et $\delta$ 0i désigne le temps pour le déplacement inverse.

On s'intéresse à l'estimation des mesures de performance tels que le taux moyen stationnaire de pièces finies produites par le système (X) ; l'utilisation de la $m^{ime}$ station ($\rho$ m) ; le temps d'attente d'un job à la m'ieme station (wm) et le temps d'attente d'un job pour compléter tous les traitements (L).

Quand on traite le MHD comme station dans le réseau, on utilise r pour le désigner (r= M+1) ; par exemple $\rho$ r indique l'utilisation du MHD.

On suppose que tous les temps de déplacement et de traitement sont distribués selon une loi exponentielle.

MASI

**V-7- Modélisation analytique pour l'état dépendant de routage :**

Pour le point de départ de notre analyse, nous allons aborder l'état dépendant de routage et la complexité de l'interférence du MHD. Ainsi on supposera que l'utilisation est assez faible que le temps d'attente du MHD peut être négligé. Dans ce cas pour toutes les pratiques les stations ne seront pas bloquées quand elles attendent l'arrivé du MHD. Cependant quand on suppose qu'on connaît préalablement la proportion de temps, que le MHD met pour déplacer la palette au buffer

centrale, la configuration résultante est un réseau de files d'attente fermé où le MHD sera une ressource dans le réseau.

MVA peut analyser exactement la configuration d'un réseau de files d'attente fermé avec routage selon Bernouli et les temps de traitement exponentiels (sans état dépendant de routage et interférence). Elle a besoin des différents temps de traitement et le compte des visites dans les inputs et estime les mesures de performance tels que l'utilisation, les de réponse et la production effective (throughput).

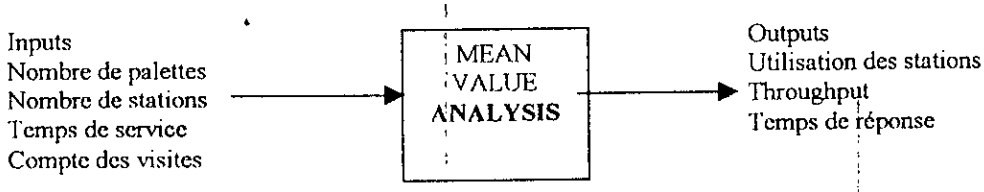



Figure (iv 2). correspondance input-output de la matrice des performances dans MVA.

Le compte de visite $v_i$ est défini par le nombre de visites que la palette fait à station i durant un cycle de traitement ; qui du temps de chargement de la pièce à la station (1) au temps de déchargement de celle ci à la station (M). Dans notre configuration la palette fait K visites aux déférentes stations dans l'ordre du plan de traitement et à la ieme étape passe une durée $\theta_i$ de temps à la station i. On peut introduire cette information détaillée dans le compte de visites $v_m$ et le temps moyen de service $T_m$ pour une visite à une station particulière m comme suit :

$$v_m = \sum_{i=1}^{K} I_{N_i=m} \qquad \forall m = 1,M$$

$$T_m = \frac{1}{v_m} \sum_{1+1}^{K} \theta_i I_m \qquad \forall m = 1,M$$

$$I = \begin{cases} 1 & \text{Si la station m est visitée par la palette dans cette ieme étape} \\ 0 & \text{Si non} \end{cases}$$

Cependant, MVA n'est entièrement adéquate pour modéliser notre configuration du FMS parce que le temps de service moyen du MHD requiert la connaissance des proportions de temps que prenne le MHD pour déplacer les palettes entre les stations et des stations au buffer central et

vise versa pour cela on approxime la probabilité actuelle de l'état dépendant par la probabilité moyenne à long terme. On suppose que :

$$1 - \pi_{i,i+1} = \rho_{N_i,N_{i+1}}$$

Alors :

$$t_{i,i+1} = \delta_{N_i,N_{i+1}}\left(1 - \rho_{N_{i+1}}\right) + \left(\delta_{N_i,0} + \delta_{0,N_{i+1}}\right)\rho_{N_{i+1}}$$

le temps total moyen de service du MHD Tr (pour tous les déplacements d'une palette) est donné par :

$$Tr = \frac{1}{K}\sum_{i=1}^{K}\left\{\delta_{N_1,N_{i+1}}\left(1 - \rho_{N_{i+1}}\right) + \left(\delta_{N_i,0} + \delta_{N_0,i+1}\right)\rho_{N_{i+1}}\right\} \qquad (1)$$

Dans l'équation (1), on définit circulairement $N_{K+1}$ à être la première station dans le plan de traitement, $N_1$. Durant un seul cycle de traitement pour une palette, le MHD est utilisé pour déplacer la palette après qu'elle soit traitée à chaque station ;d'où le taux de visite au MHD est égale au nombre des étapes dans le cycle de traitement, K. Puisqu'on utilise MVA pour évaluer les mesures de performance, on a besoin de faire l'approximation que le temps de service du MHD est distribué selon une loi exponentielle avec la moyenne approximative Tr et, similairement, que les temps de service des stations sont distribués selon une loi exponentielle de moyenne approximative Tm, m= 1..., M.

L'équation (1)calcule le temps moyen de service du MHD. Cependant, de la figure 4 on voit que le temps moyen de service du MHD est nécessaire dans l'input de MVA pour déterminer les déférentes mesures de performance du réseau et l'utilisation des machines. Pour résoudre cette équation on utilise une méthode itérative.

On définira la station goulot comme étant la station de traitement qui a le plus grand nombre de visites et indexée par b : qui est b = max. $_{m \neq M+1}$ (TmVm). On trouve l'utilisation de la station goulot $\rho_m$. Une fois que cette utilisation est trouvée, les autres utilisations seront déterminées successivement vu la nature du réseau fermé.

Algorithme 1. Evaluation des performances du système avec l'état dépendant de routage :

Inputs. K , M , N , $\{(\theta_i, N_i), i = 1...,K\}, [\delta_{ij}, i = 1...,M, j = 1...,M]$.

Etape 0. Calcul du compte de visites et les temps de service :

$$v_m = \sum_{i=1}^{K} I_{N_i=m} \qquad \forall m = 1,M \quad : \text{Vr} = \text{K}$$

$$T_m = \frac{1}{v_m} \sum_{1+1}^{K} \theta_i I_m \qquad \forall m = 1,M$$

Déterminer la station goulot b = max. $_{m \neq M+1}$ (TmVm) :

$$\text{j} = 1,\ X = 0, \quad \tau_i^0 = v_i T_i \qquad \forall i = 1,...,M, \text{w}_\text{r} = 0$$

Etape 1.

$$\tau_i = \tau_i^0 + w_r \qquad \forall i = 1,...,M$$

$$\text{j} = 1\ ; \qquad \rho_{left} = 0, \qquad \rho_{right} = 1$$

Etape 2.

$$\rho_b^{(j)} = \left(\rho_{left} + \rho_{right}\right)/2$$

$$\rho_i = \left(\tau_i / \tau_b\right)\rho_b \qquad \text{i} = 1,\ldots,\text{M}$$

$$T_r = \frac{1}{K} \sum_{i=1}^{K} \left\{\delta_{N_i,N_{1+1}} \left(1 - \rho_{N_{i+1}}\right) + \left(\delta_{N_1,0} + \delta_{0,N_{i+1}}\right)\rho_{N_{i+1}}\right\}$$

Etape 3. Utilisation de MVA pour trouver $\rho_b$ :

$$: L_i(0) = 0 \qquad \forall i = 1,...,M$$

Pour n = 1 à N

faire :

$$W_i(n) = \tau_i + L_i(n-1)\tau_i \qquad \forall i = 1,...,M; \quad X = \frac{N}{\sum_{i=1}^{M} W_i(n) + \tau_r}; L_i(n) = XW_i(n) \quad \forall i = 1,...,M$$

$$\rho_b = X\tau_b.$$

.Etape 4.

Si $\rho_b \le \rho_b^{(j)}$ *alors* $\rho_{left} = \rho_b^{(j)}$; *sinon* $\rho_{right} = \rho_b^{(j)}$.

Etape 5. Si $\left(\left|\rho_b - \rho_b^{(j)}\right| \le \varepsilon \quad ou \quad (j > j_{max})\right.$, aller à l'étape 6. Si non, j=j+1, X=X,

Aller à l'étape 2.

Etape 6. Enregistrer toutes les dernières mesures de performance avec indexe.

Les résultats sont donnés dans le tableau 1.

Validation du premier algorithme :

Le modèle de simulation est construit e utilisant le langage SIMAN [ PEGDEN, 90].

Dans le modèle de simulation, les pièces sont chargées sur un nombre fini de palettes qui sont traitées séquentiellement par 4 (quatre) stations, dans le système les temps de traitement sont supposés distribués selon une loi exponentielle de moyenne 100 Mn . Une palette qui termine le traitement à une station, le MHD la déplace vers la prochaine station si celle-ci est libre au début du déplacement sinon il la déplace vers un buffer central. Ces temps de déplacement sont supposés distribués selon une loi exponentielle.

Dans ce premier modèle, on suppose que les stations qui finissent le traitement des palettes ne seront pas bloquées à cause de l'interdépendance du MHD.

En conclusion le modèle de simulation prend en compte que l'état dépendant de routage et le temps de déplacement dépendant de la destination.

Pour évaluer les performances du modèle analytique on fait varier deux paramètres :

- Le nombre de palettes de 4 à 8
- La vitesse du MHD

Les résultats sont donnés dans le tableau 1.

| exe | Nbre de palettes | Throughput (x) pieces/mn | | |
|---|---|---|---|---|
| | | simulation | Algorithme | Erreur (%) |
| 1 | 4 | 0,00562 | 0.00581 | 3.3 |
| 2 | 5 | 0,00614 | 0.00615 | 0.1 |
| 3 | 6 | 0,00650 | 0.00649 | 0.1 |
| 4 | 7 | 0,00684 | 0.00684 | 0.0 |
| 5 | 8 | 0,00715 | 0.00713 | 0.2 |

***Tableau 1.*** *Validation de l'algorithme pour l'état dépendant de routage*

**V-8- La vue de la palette :**

On examine l'itinéraire de la palette. Une fois que la palette termine le traitement à une station, elle attend l'arrivée du MHD. A l'arrivée du MHD la palette libère la station et occupe le MHD. Après se dirige directement à la prochaine station, si celle ci est libre, si non elle est transportée vers le buffer central où elle attend la libération de la prochaine station ; une fois que la prochaine station est libérée, elle la réserve. Apres elle demande le MHD et elle est transportée vers la prochaine station quand le MHD arrive. Cette étape est répétée jusqu'à ce que la palette atteint la station de déchargement, où la pièce (produit) est déchargée et la palette est transférée à la station de chargement.L'effet de l'interférence du MHD est d'immobiliser les stations jusqu'à l'arrivée du MHD.

On peut considérer un autre scénario équivalent dans lequel, le temps de traitement à chaque station est plus grand que le temps de traitement original d'une quantité égale au temps d'attente du MHD. Cela aura un effet sur l'utilisation des stations, seulement on doit modifier les modèles du MHD. La représentation du réseau de files d'attente pour ce modèle est donnée dans la (figure (iv 3)).

On pose $B_m$ le temps de blocage de la station m quand elle attend le MHD. Une fois que les valeurs de $B_m$ sont connues, on aura le temps total de service des stations. Comme on a encore besoin de justifier pour l'état-dépendant de routage, on peut résoudre ce réseau ; pour cela on utilise une méthode itérative développée dans l'algorithme1. Avec cette méthode et la connaissance des valeurs des $B_m$, on peut estimer les performances du système. Pour faciliter l'analyse, on suppose que $B_m$ est le même pour toutes les stations. Cependant, cela n'est pas une limitation pour le modèle, et la flexibilité des différentes valeurs de $B_m$ peut être exploitée pour d'avantages majorations dans les futurs modèles.

Pour déterminer le temps de blocage des stations, on a besoin en équivalant de savoir le temps que mettent les palettes à attendre le MHD à la station. L'attente au niveau du MHD aura lieu seulement si considère le MHD comme une ressource finie. Dans la représentation de la figure de la palette, le MHD est serveur infini. Donc on a deux cas :

Le MHD est une ressource finie ou un serveur infini.

La solution de ce dilemme est de prendre le second cas, le cas ou le MHD est un serveur fini ce qui admet la file des palettes, de là on peut déterminer le temps d'attente.

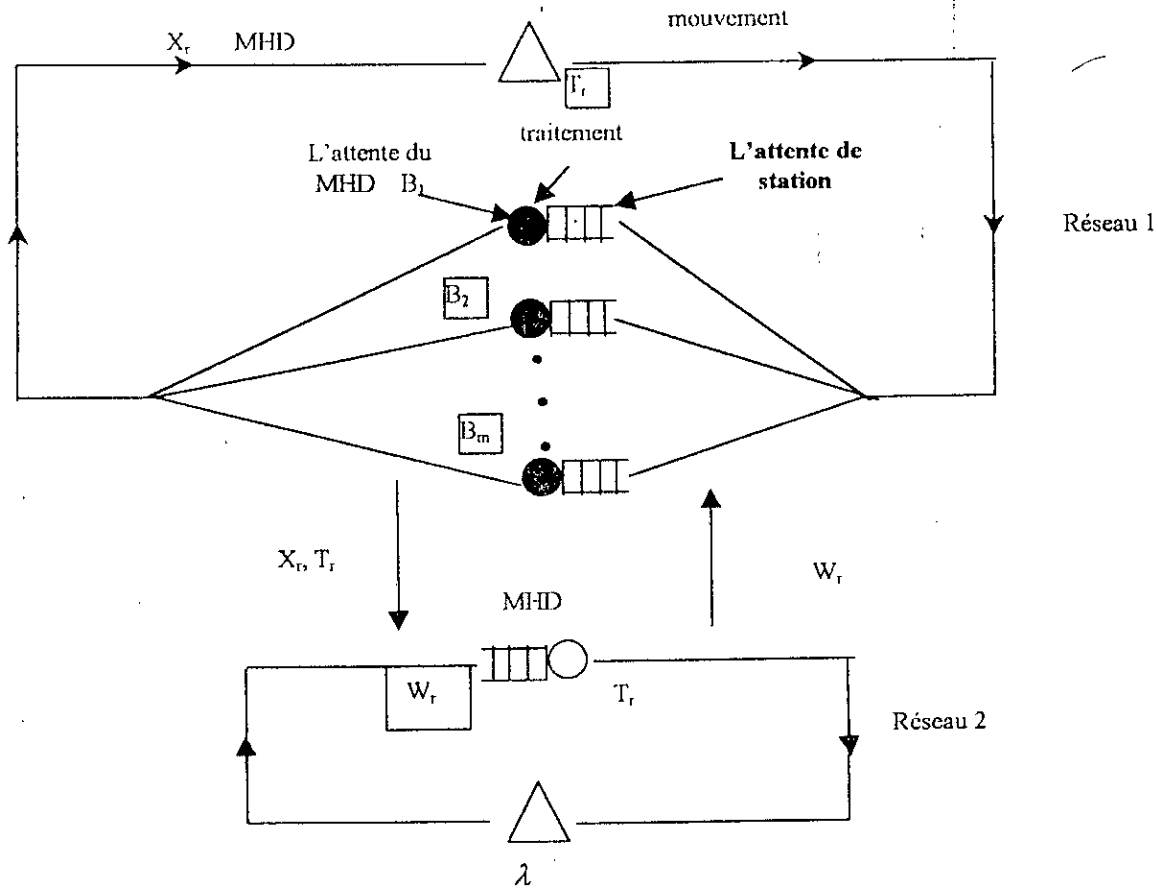



Figure (iv 3). réconciliation des deux cas de figure

## V-9- Modélisation de la vue du MHD.

On observe que les palettes demandent les services du MHD pour être déplacées au buffer central ou aux stations. Comme on peut voir du réseau 1 de la figure( IV 3) que seules les palettes qui ont complété le service aux stations, peuvent demander les services du MHD. Le déplacement du buffer central est possible seulement pour les palettes qui ont réservé les prochaines stations, ce qui complique le processus de déplacement. Cependant, si on connaît le taux de clients effectif à l'appareil MHD ($X_r$), qui est le taux d'arrivée des demandes du service du MHD. De plus les demandes sont faites seulement pour les palettes qui sont dans la boite noir non pour celles qui sont déjà en attente du MHD. Spécifiquement parce qu'il y a N palettes dans le système, si n palettes sont en attente du MHD, alors (N-n)palettes sont en traitement dans la boite noir. Alors le taux de palettes qui arrivent à la boite noir est fonction de (N-n).

Ce ci est modélisé par un réseau de files d'attente qui est donné dans la figure 5 (réseau 2). Dans ce réseau on a une boucle fermée qui est constituée du MHD et du serveur en ligne qui représente la boite noir. Quoi qu'on connaît pas la distribution du taux de service du serveur en ligne $\lambda(.)$. Pour simplifier, on suppose que chaque palette a la même probabilité de demander le

service du MHD une fois qu'elle se dirige à la boite noir. Alors on suppose que $\lambda(i)=i\lambda$. Cela est équivalent à traiter le MHD comme une file M/M/1/N (population finie égale au nombre de palettes dans le système N). Cela explique que seules les palettes dans la boite noir contribuent au processus d'arrivée.

Le problème est que $\lambda$ n'est pas encore connu. Mais si on dispose du taux effectif des palettes ($X_r$) et le temps moyen de service du MHD ($T_r$), on peut déterminer la valeur de $\lambda$ pour laquelle le taux effectif est $X_r$. spécifiquement la valeur de $\lambda$ tel que :

$$X_r = \lambda[N - L(\lambda)]$$

Où $L(\lambda)$ est définie telle que : (Gross et Harris, 1985)

$$\rho = \lambda T_r$$

$$P_0 = \frac{1}{\sum_{k=0}^{N} \rho^k \frac{N!}{(N-k)!}}$$

$$P_k = P_0 \rho^k \frac{N!}{(N-k)!}$$

$$L(\lambda) = \sum_{k=0}^{N} k P_k$$

C'est une équation complexe non linéaire en $\lambda$. Pour la résoudre on utilise la bissection pour trouver $\lambda$ tel que $\lambda_{eff} = X_r$. On utilise la bissection parce qu'elle garantie la précision, puisque $\lambda_{eff}$ est une fonction non décroissante en $\lambda$, limitée entre 0 et $X_r$. La solution est obtenue dans l'algorithme 2.

Une fois que $\lambda$ est déterminé, le temps moyen d'attente dans la file (M/M/1/N) est déterminé à son tour et c'est le temps moyen d'attente pour le MHD $W_r$. Pour Gross et Harris, 1985 on a :

$$W_r = T_r \sum_{k=1}^{N} (k-1) P_k$$

**Algorithme 2** :

Inputs. N, $X_r$, $T_r$.

$$\lambda_{left} = 0, \qquad \lambda_{rigth} = 0, \qquad j = 0$$

étape 1:

$$\lambda = \frac{\lambda_{left} + \lambda_{rigth}}{2}$$

étape 2 :

$$P_0 = \frac{1}{\sum_{k=0}^{N} \rho^k \frac{N!}{(N-k)!}}$$

$$P_k = P_0 \rho^k \frac{N!}{(N-k)!}$$

$$L = \sum_{k=0}^{N} k P_k$$

$$\lambda_{eff} = \lambda(N - L)$$

étape 3 :

$$si \quad \lambda_{eff} < X_r, \quad alors, \quad \lambda_{left} = \lambda$$
$$sinon \quad \lambda_{rigth} = \lambda$$

$$j = j + 1 :$$

$$si \quad \left|\lambda_{eff} - X_r\right| > \varepsilon \qquad alors\ aller\ à\ l'étape\ 1$$
$$sinon: \qquad aller\ à\ l'étape\ 4$$

étape 4:

$$W_r = \tau \sum_{k=1}^{N} (k-1) p_k$$

output. $W_r$..

**V-10- Reconciliation des deux vues.**

La figure 5 montrep l'intéraction entre les deux réseaux qui représentent les deux cas de figure. Une fois qu'on dispose de l'estimation de $B_m$ , on peut déterminer le taux effectif aux différentes stations, en particulier $X_r$ pour le réseau 1, qu'on utiliset pour le réseau 2 pour déterminer le temps d'attente pour le MHD, $W_r$ .

Si on suppose que le temps pour lequel les stations sont bloquées est le même et égal à cette valeur moyenne de $W_r$ , alors l'estimation de $W_r$ doit correspondre au temps de blocage $B_m$ pour le réseau 1.

La solution simultanée des deux réseaux peut être obtenue par un algorithme itératif.

Initialiser les valeurs de $B_m$ et résoudre le réseau 1en utilisant l'algorithme 1 pour obtenir $X_r$ qu'on utilise dans l'algorithme 2 pour déterminer $W_r$ dans le réseau 1 et on répete cette procédure jusqu'à ce que la différence entre le taux effectif des deux réseax soit inferieur à un ε donné.

**Validation du modèle**

On teste ce modèle plus général en comparant avec des résultats de simulation de 24 stations (Electroplating line of major U.S. computer manufacturer).

La manutention consiste en un seul appareil de vitesse 0.82 m/s et d'accélération 0.9 m/sxs.

Pour notre expérience le plan de traitement requiert que les palettes visitent les 24 stations séquentiellement.

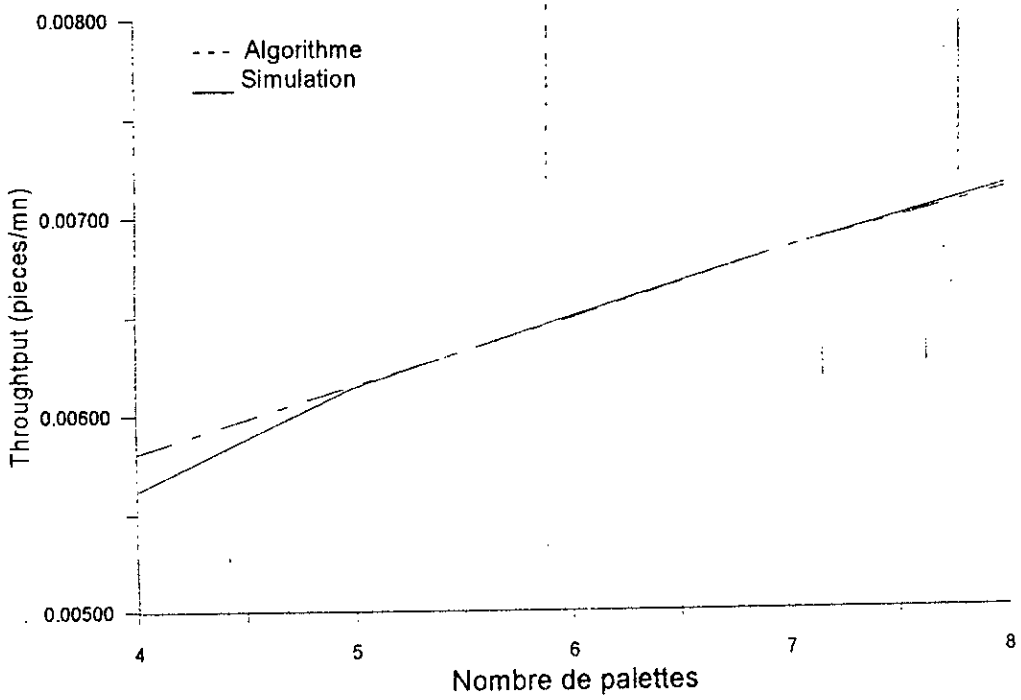
Le modèle de simulation pour cette ligne est aussi construit en utilisant le langage SIMAN. On fait varier le nombre de palettes de 15 à 50. Pour chaque nombre de palettes, on étudie la performance du système, en faisant varier la performance du MHD.

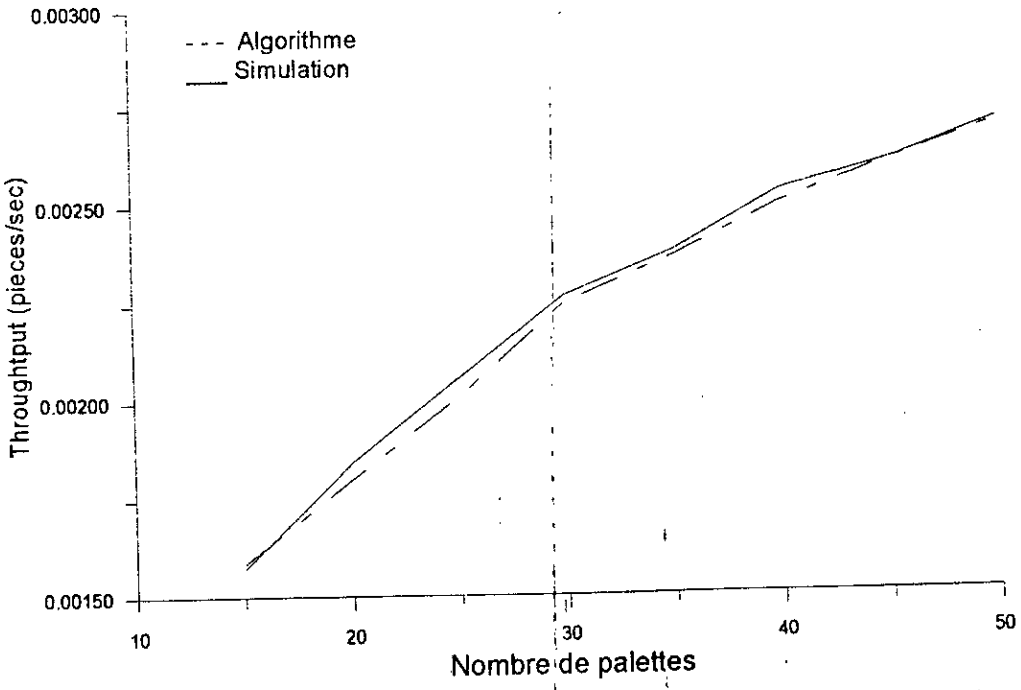
Tous les temps de traitement sont supposés distribués selon une loi exponentielle de moyenne 250 secondes.

Les comparaisons des taux stationnaires de production (Throughtput) sont données dans le tableau 2 et leur graphe est donné dans la figure VI-2.

Les résultats du modèle analytique sont très proches de ceux de la simulation. Pour les 40 exécutions qu'on a effectué, l'erreur absolue ne dépasse pas 3%.

| exé | | $N^{bre}$ de palettes | Throughtput (x) Pieces / Sec | | |
|---|---|---|---|---|---|
| | | | simulation | Algorithme | Erreur ( % ) |
| 1 | 11.5 | 15 | 0.001572 | 0.001585 | 0.8 |
| 2 | 6 | 20 | 0.001843 | 0.001800 | 2.3 |
| 3 | 8.4 | 25 | 0.002058 | 0.002010 | 2.3 |
| 4 | 10.8 | 30 | 0.002271 | 0.002251 | 0.9 |
| 5 | 14 | 35 | 0.002381 | 0.002367 | 0.6 |
| 6 | 17.3 | 40 | 0.002532 | 0.002499 | 1.3 |
| 7 | 21 | 45 | 0.002606 | 0.002605 | 0.0 |
| 8 | 25 | 50 | 0.002709 | 0.002701 | 0.2 |


0.00800
0.00700
0.00600
0.00500
Throughtput (pieces/mn)
Algorithme
Simulation
4
5
6
7
8
Nombre de palettes


0.00300
0.00250
0.00200
0.00150
Throughtput (pieces/sec)
Algorithme
Simulation
10
20
30
40
50
Nombre de palettes

## Conclusion

Pour analyser les performances de cette configuration correctement et efficacement,on a développé un réseau de file d'attente basé essentiellement sur les models analytiques . Cependant les models développés supposent que les temps de traitement sont distribués selon une loi exponentielle. En plus la méthodologie qu'on a développée pour approcher l'état dépendant de routage et l'interférence du MHD peut être utilisée pour le développement de models de distributions générales.

On peut aussi envisager que par la construction de ces models d'analyser des systèmes manufacturiers plus larges.

Dans un tel cas une bonne règle de conception sera de retenir l'utilisation du MHD dans les 60 % pour assurer l'exactitude de l'estimation des déformations en utilisant les models analytiques.

L'utilisation de cette règle avec des models analytiques pour l'analyse des performances peut assister la conception rapide et effective et l'analyse des systèmes manufacturiers.

# CONCLUSION GENERALE

# *Conclusion générale :*

Dans ce travail, on a tenté de résumer certains problèmes d'ateliers flexibles liés à leur modélisation, l'évaluation de leurs performances, leur optimisation. Il est question surtout d'un point de vue développé dans la littérature de files d'attente.

L'avantage de cette approche est d'abord de mieux comprendre les propriétés des modèles FMS. Elle permet de justifier certaines heuristiques et d'en développer de nouvelles.

On a présenté un modèle analytique basé sur les réseaux de files d'attente pour prédire les mesures de performance d'un système manufacturier flexible avec un seul appareil de manutention (MHD ). Cette configuration est significative pour plusieurs raisons : Elle est courante dans l'industrie, elle simplifie le contrôle de la manutention, elle est docile pour la modélisation analytique et elle forme une base de construction pour les modèles plus complexes.

Pour analyser l'état dépendant de routage, on a développé une méthode itérative basée sur l'analyse des valeurs moyennes ( MVA ). Pour analyser l'interférence du MHD on a utilisé les modèles des réseaux de files d'attentes. Dans le premier réseau on ignore la file à l'appareil MHD mais on modélise son interférence en gonflant les temps de service des stations. Le deuxième réseau modélise la file pour le MHD et estime les temps de blocage nécessaires pour le premier modèle. En itérant entre les deux réseaux, on arrive à prédire les performances de cette configuration des FMS. Nos estimations analytiques sont validées en comparaison avec des résultats de simulation.

# BIBLIOGRAPHIE

## *Bibliographie*


**[Aït Hssain 93]** Aït Hssain 93, « conduite hiérarchisée intégrées des ateliers manifacturiers flexibles : une approche mixte objets / réseaux de Petri », thèse de doctorat INPG, laboratoire d'automatique de Grenoble.

**[Allab,94]** S.Allab, G.Finke « deadline scheduling in job categories », 7th International conference of the European capter on combinatorial optimization (ECCO VII), Milan, Italie, Fev 1994.

**[Askin,93]** : Askin.R. Standridge.C.; "Modeling and analysis of manufacturing systems",
John Wiley & Sons edition, 1993.

**[Bermudes, 97]** : " Bullletin de liaison du groupe Bermudes N°1à4", 1997.

**[ Browne, 84]** Browne & Stecke :" Classification of flexible manufacturing systems ". The FMS magazine, avril 1984.

**[Burbidge,92]** J.L. Burbidge, « Change to group technologie, process organization is obselete » International journal for production reserch, vol 30 n°5, 1992

**[Buzacott, 86]** Buzacott & Yao : "Flexible manufacturing systems", a review of analytical
models, 1986

**[Buzacott, 92]** Buzacott & Shanthikumar: "Design of manufacturing systems using queueing
models", 1992.

**[Cohendet,89]** P.Cohendet, P.Llerena, B.Mutel, « Fléxibilité et mises en coherance des données de production » dans ' les nouvelles rationalisations de la production ' CEPADUES EDITIONS.

**[Dallery,92]** Dallery & Gershwin :"Manufacturing flow line systems", a review of models and analytical results,1992

**[Dallery,94]** Y.Dallery, « Gestion de production » cours de D.E.A. d'automatique productique, ENSIE de Grenoble, 1994.

**[Dimitrov,94]** Dimitrov & Khallil :"The service time properties of a server characterize the exponential distribution", 1994.

**[Doumeingts,83]** G.Doumeingts, D.Breuil, L.Pun « La gestion de prduction assistée par ordinateur » Ed Hermès, 1983

**Galambos, 94]** Galambos &Hagwood : "An unveliable server characterization of the exponential distribution, 1994.

**[Hax,75]** A.C.Hax, H.C.Meal « Hiearchical integration of production planning and scheduling » TIMS studies in managment science, vol 1 Logistics, ed, M.A.Geisler, NewYork 1975.

**[Jubin, 94]** Jubin. M : "Ateliers flexibles d'usinage", techniques de l'ingenieur, 1994.

**[Kolodny,89]** H.F. Kolodny, « Product focussed forms to complement flexible technologies » Working paper, Facullty of managment, univ of Toroto, Canada.

**[Kotz,78]** Galambos & Kotz :"characterization of probability distributions",1978.

**Kouvelis, 92]** Kouvelis :"Design and planning problems in flexible manufacturing systems". Journal of intelligent manufacturing ,1992.

**[Lavington,21]** F. Lavington, « The english capital marcket » Methuen, London, Angletaire

**[Lim,87]** S.H. Lim « Flexible manifacturing systems and manifacturing felxibility in the UK » International journal of flexible manifacturing.systems, n°6,1987.

**[Massotte,96]** P.Massotte, « Ordonnancement réactif et informtique » Conférence de l'ENSGI Grenoble 1996.

**[Proth, 86]** Proth. J : "Systèmes flexibles de production", édition Masson, 1986.

**Proth, 86]** Proth. J : "Systèmes flexibles de production", édition Masson, 1986.

**[Sethi,90]** A.K.Sethi, S.P.Sethi, « Flexibility in manifacturing : A Survey » the international journal of flexible manifacturing.systems, n°2,1990.

**[Stecke,89]** K.Stecke, « Algorithm for efficient planning and operation of a particular FMS » International journal of flexible manifacturing.systems, n°1,1989.

**[Veltz,93]** P.Veltz et P.Zarefian, « Modèle systémique et flexibilité », dans ' Les nouvelles rationalisation de la production' CEPADUES EDITIONS'.

**[Williamsom,63]** O.E.Williamson, « A model of rational manaregial behaviour » dans ' A behaaviourical theory of the firm' Ed R. Cyert, J. March, 1963

# ANNEXE

# *Annexe:*

## La flexibilité du système de production selon [ SETHI 1990 ]:

Les différents types de flexibilité sont définis comme suit :

***Flexibilité machine*** : Aptitude à exécuter des opérations de nature différentes sans engendrer des coûts de lancements prohibitifs.

***Flexibilité du système de convoyage*** : Aptitude à transporter les différents produits a travers le système de production de manière à assurer une fabrication efficace et de qualité.

*Flexibilité opératoire* : Relative à un produit, elle traduit l'existence de différentes gammes de production possible (elle apparaît donc lors de la phase de conception du produit ).

*Flexibilité du process* : Traduit la capacité d'un système de production à produire un certain nombre de famille de produits sans majorer les coûts de lancements de manière significative.

*Flexibilité de routage* : Dépend de l'aptitude à produire une pièce en utilisant différents routages a travers le système.

*Flexibilité de produit* : Traduit la facilite avec laquelle de nouveaux produits peuvent êtres introduits dans le plan de production.

*Flexibilité de volume* : Relative au nombre de niveaux de flux de produits sortants pour lesquels la production reste rentable.

*Flexibilité d'extension* : Traduit le facilite avec laquelle la capacité et les <<compétences >>du système peuvent être développées.

*Flexibilité de programme* : Traduit la capacité du système à fonctionner sans surveillance réelle des opérateurs pendant une période donnée. Ajoutés aux flexibilités de routage et de process, elle nécessite l'installation d'un système de capteurs associé à un contrôle informatique apte à détecter les événements perturbateurs et y palier.

*Flexibilité de production* : Relative à l'univers des familles de produits qu'un système de fabrication est apte à produire sans nécessiter pour cela des investissements majeurs.

**Q-NAP** (Queuing Network Analysis System)

Développé à l'INRIA (Institut National de recherche en informatique et en Automatique), initialement pour l'étude des réseaux d'ordinateurs, ce langage s'est ensuite étendu pour servir à l'étude des systèmes de production. Q-NAP2, est un outil logiciel de description et d'analyse quantitative de réseaux de files d'attente.

*Il* existe une grande variété de techniques de résolution de réseaux de files d'attente. Elles diffèrent par leur champ d'application, et lorsque plusieurs méthodes sont concurrentes, par leur coût en fonction de l'application.

Résoudre un modèle à l'aide d'une mathématique, c'est calculer les critères de performance qui caractérisent l'état permanent (i.e. d'équilibre) de ce modèle. Les méthodes mathématiques disponibles sont:

- Méthode analytique exacte,
- Méthode markovienne,
- Méthode analytique approchée.