Le wip concurrent : une proposition de file d'attente du point de vue du produit pour caractériser le temps de cycle

Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, Philippe Vialletelle

11e Conférence Francophone d'Optimisation et Simulation- MOSIM'16
Du 22 au 24 août 2016
Montréal, Québec, Canada
« L'innovation technologique au service des systèmes : défis et opportunités»

LE WIP CONCURRENT : UNE PROPOSITION DE FILE D'ATTENTE DU POINT DE VUE DU PRODUIT POUR CARACTERISER LE TEMPS DE CYCLE

Kean DEQUEANT[1,2], Pierre LEMAIRE[1], Marie-Laure ESPINOUSE[1], Philippe VIALLETELLE[2]

[1]Univ. Grenoble Alpes, G-SCOP, F-38000 Grenoble, France
CNRS, G-SCOP, F-38000 Grenoble, France
{kean.dequeant, pierre.lemaire, marie-laure.espinouse}@g-scop.grenoble-inp.fr
[2]STMicroelectronics, F-38926 Crolles Cedex, France
{kean.dequeant, philippe.vialletelle}@st.com

RESUME : *Nous nous intéressons à des modèles de théorie des files d'attente pour caractériser les temps de cycle (délais de fabrication à différentes étapes de production) de produits dans des productions complexes. Des modèles de théorie de file d'attente sont régulièrement utilisés dans l'industrie pour cela, mais en dehors de leurs hypothèses de modélisation. Nous montrons tout d'abord dans cet article l'incidence d'une utilisation de ces modèles hors hypothèses sur la qualité de l'estimation du temps de cycle. Nous proposons alors un nouveau type de représentation des files d'attente, du point de vue des produits et sans hypothèses sur les équipements. Nous montrons sur un cas d'étude réel d'équipements complexes de microélectronique comment cette nouvelle représentation des files d'attente permet, en plus d'une première caractérisation du temps de cycle, d'extraire des informations fondamentales de n'importe quel groupe d'équipements traitant un même flux de produits. Enfin, nous discutons des étapes à venir pour intégrer cette représentation dans des outils de simulation ainsi que dans des modèles totalement génériques de files d'attente.*

MOTS-CLES : *Temps de cycle – Files d'attente – Microélectronique – Production high-mix, low-volume.*

1 INTRODUCTION

La maitrise des délais de fabrication est un enjeu majeur dans l'industrie microélectronique. Elle permet un équilibre entre les coûts de fabrication, l'obsolescence des produits et le respect des livraisons aux clients. Représenter et comprendre les délais de fabrication (appelés temps de cycle dans le reste de cet article) pour mieux les maitriser à chaque étape de production est donc fondamental pour les acteurs de ce secteur. Dans ce sens, nous cherchons à caractériser le temps de cycle de produits pour leurs diverses étapes de fabrication, et en particulier à développer des méthodes utilisables dans le cas complexe des unités de production à faible volumes et à forte diversité de produits (high-mix, low-volume).

Les principales méthodes de caractérisation du temps de cycle reposent sur la théorie des files d'attente. Celle-ci permet, entre autre, le calcul du temps de cycle pour des systèmes clients/serveurs (ou produits/équipements) sous un certain nombre d'hypothèses. Nous commencerons donc par un état de l'art de la théorie des files d'attente et de son utilisation en production microélectronique. Puis, nous montrerons les limites de l'approche classique dans le cadre de productions à faible volume et forte diversité de produits. Nous proposerons une nouvelle approche, basée sur la modélisation des files d'attente du point de vue des produits au lieu de l'approche classique qui adopte le point de vue des équipements. Nous montrerons que cette approche, en recomposant les files d'attente réellement vues par chaque lot de production à partir de l'historique, permet une meilleure représentation et une meilleure caractérisation du temps de cycle. Nous introduirons alors la notion de WIP concurrent, dont la définition et l'utilisation représentent la principale contribution de cet article, avant de discuter des perspectives qu'offre cette nouvelle notion.

2 ETAT DE L'ART

Les premiers modèles exacts de la théorie des files d'attente ont été développés par Erlang avec les hypothèses de loi d'arrivée exponentielle et de loi de service déterministe (Erlang, 1909). Puis, ces modèles ont été étendus progressivement par Kendall (1953), Pollaczek (1957), et Kingman (1961). Kendall a en particulier introduit la notation standard en théorie des files d'attente (une suite de symboles, souvent limitée aux trois premiers, qui représentent respectivement la loi d'arrivée, la loi de durée de service ou process et le nombre de serveurs qui traitent la file), et Kingman a proposé une formule majeure pour le calcul du temps de cycle moyen pour le cas G/G/1 (lois d'arrivée et de process « générales » et un unique serveur). Le cas général G/G/m, le plus proche de la réalité, n'a pas de solution exacte connue, cependant Hopp et Spearman (2001) ont établi une référence dans l'industrie en popularisant les travaux de Sakasegawa (1977) et de Whitt (1993) avec une approximation pour le cas G/G/m intégrant des

composantes de variabilité (d'arrivées et de temps de process effectif) permettant de calculer le temps de cycle moyen à une étape (la variabilité est la non-homogénéité des paramètres, qui crée des perturbations et augmente le temps de cycle ; elle est souvent mesurée par le ratio entre l'écart-type et la moyenne d'un paramètre [coefficient de variation]). Leurs travaux servent concrètement de base au dimensionnement des usines de production de semi-conducteurs, permettant de s'assurer de garder les temps de cycle sous contrôle. Beaucoup d'autres formules ont été proposées depuis : Huang, Chang, et Chou (2001) traitent du multi-produits et du traitement simultané de plusieurs lots par un même équipement (batching), sous l'hypothèse de loi d'arrivée exponentielle et de loi de process exponentielle. Morrison et Martin (2006) proposent entre autres une extension de la formule de Hopp et Spearman qui inclut le re-traitement des lots lors de pannes d'équipements. Zisgen, Meents, Wheeler, et Hanschke (2008) proposent des formules très complexes pour le cas des arrivées groupées (batch) et du multi-produits, en gardant l'hypothèse d'équipements identiques.

La principale formule établie par Hopp et Spearman, et citée comme référence dans l'industrie par Shanthikumar, Ding, et Zhang (2007) pour le calcul d'un temps de cycle *CT* est:

$$CT = \left(\frac{C_a^2 + C_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e + t_e$$

Avec :

- $t_e = \frac{PT}{A}$
- PT : temps de process moyen.
- A : disponibilité moyenne des équipements.
- C_a: coefficient de variation des temps inter-arrivées.
- C_e: coefficient de variation des temps de process effectifs.
- m : nombre d'équipements en parallèle.
- u : taux d'utilisation des équipements.

Leachman (2012) a proposé une extension de la formule de Hopp et Spearman et a utilisé cette formule pour caractériser le temps de cycle théorique de chaque étape de production et ainsi repérer les pertes opérationnelles de temps de cycle (i.e. non liées à la structure de l'atelier mais imputables à sa mauvaise gestion) dans le but de les corriger. Senderovich, Weidlich, Gal, et Mandelbaum (2015) traitent du cas multi-classes (de priorités) pour des algorithmes de prédiction dynamique du temps de cycle mais avec des hypothèses strictes de politique d'ordonnancement. Schelasin (2013) et Kim, Wang et Havey (2014) proposent un ajustement empirique du facteur de variabilité de la formule de Hopp et Spearman (la première parenthèse de la formule, regroupant les coefficients de variation des temps inter-arrivées et de process effectifs) permettant d'en augmenter la précision, tout en gardant les nombreuses hypothèses de modélisation.

L'utilisation de la théorie des files d'attente pour la modélisation des temps de cycle dans l'industrie microélectronique est cependant sujette à controverse. Etman, Veeger, Lefeber, Adan, et Rooda (2011) dénoncent le manque de précision dû à la complexité des processus de fabrication. Ils citent en particulier la propriété de certains équipements à traiter simultanément plusieurs lots (multi opération série ou parallèle, batching…) ainsi que la difficulté d'accès aux données nécessaires pour nourrir les formules de files d'attente. Shanthikumar, Ding, et Zhang (2007) et Akhavan-Tabatabaei, Ding, et Shanthikumar (2009) critiquent les hypothèses d'indépendance et d'identité des distributions des temps inter-arrivées et des temps de process. Shanthikumar, Ding, et Zhang (2007) critiquent également l'hypothèse d'indépendance entre les arrivées et les temps de process effectifs : les arrêts planifiés sont généralement reportés si beaucoup d'arrivées sont prévues. La spécialisation des équipements est un autre problème qui viole l'hypothèse d'équipements identiques dans les modèles classiques de la théorie des files d'attente. Ce problème est mis en avant par Miltenburg, Cheng, et Yan (2002) et discuté par Shanthikumar, Ding, et Zhang (2007) sans pour autant qu'une solution soit proposée. Kingman (2009), l'un des contributeurs principaux de la théorie des files d'attente, décrit les critiques de Kendall envers la multitude de solutions formelles issues de la variété des modèles à définir pour traiter chaque cas de façon spécifique et donc avec une validité limitée.

La littérature offre des solutions à la plupart des problèmes cités ci-dessus : le batching est étudié de près par Huang, Chang, et Chou (2001), et Leachman (2012) propose également une formule spécifique. Akhavan-Tabatabaei, Ding, et Shanthikumar (2009) proposent un algorithme empirique pour corriger les effets de dépendance entre arrivées et temps de process effectifs. Ignizio (2009) adresse le problème de spécialisation des équipements mais dans une optique d'ordonnancement et non de caractérisation du temps de cycle.

Certaines des critiques envers la théorie des files d'attente sont adressées indépendamment, mais comme le décrit Kingman (2009), ces solutions deviennent de plus en plus nombreuses et traitent des cas de plus en plus spécifiques. Aucune solution n'a été proposée qui prenne en compte l'ensemble de ces problèmes et la caractérisation générique du temps de cycle ne semble donc pour l'instant pas possible dans le cadre d'unités de fabrication à faible volume et forte diversité de produits.

3 POSITION DU PROBLEME

3.1 Enjeux

Opérationnellement, le temps de cycle des produits est maitrisé à partir d'un taux d'utilisation maximum des équipements, qui est déterminé grâce aux courbes opérationnelles, telles que représentées Figure 1 (les différentes courbes correspondent à différentes configurations du système). En pratique ces courbes sont calculées à partir de formules de la théorie des files d'attente comme celle de Hopp et Spearman, dont les hypothèses ne s'appliquent pas à notre cas de figure (voir discussion qui suit). On a donc besoin d'une méthode de caractérisation du temps de cycle applicable à n'importe quel groupe générique d'équipements.

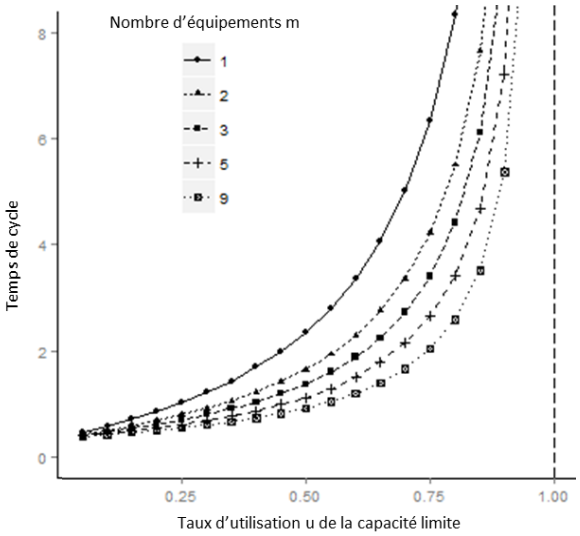

Figure 1 : Courbes opérationnelles montrant l'impact du nombre d'équipements sur le temps de cycle

3.2 Limites des hypothèses standards

Nombre d'auteurs ont soulevé les limites du modèle classique ; il est toutefois courant dans l'industrie de considérer que la réalité est « suffisamment proche » de tel ou tel modèle théorique pour l'appliquer en l'état. Nous allons donc mettre en regard le modèle classique (G/G/m), sur lequel s'applique la formule de Hopp et Spearman, et le modèle générique complexe que nous cherchons à résoudre. Nous montrerons alors, par un exemple simple que, dans le contexte de la production à faible volumes et forte diversité de produits, l'adaptation n'est ni triviale, ni justifiable.

Ainsi, la figure 2 illustre la configuration produits/équipements adressée par le modèle de Hopp et Spearman. C'est le cas classique d'une file d'attente mono-produit traitée par un certain nombre d'équipements en parallèles. Elle est suivie de la figure 3 qui illustre la configuration générique que l'on peut trouver dans une production à faible volume et forte diversité de produits.

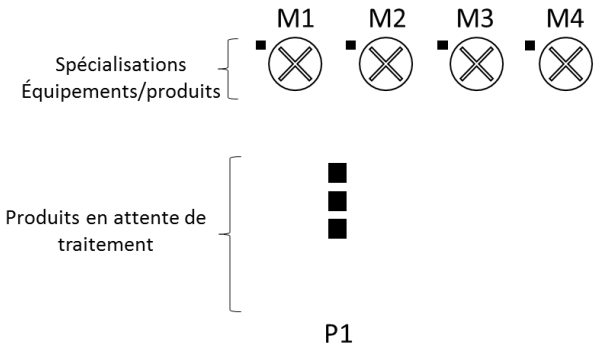

Figure 2 : Modèle de file d'attente classique

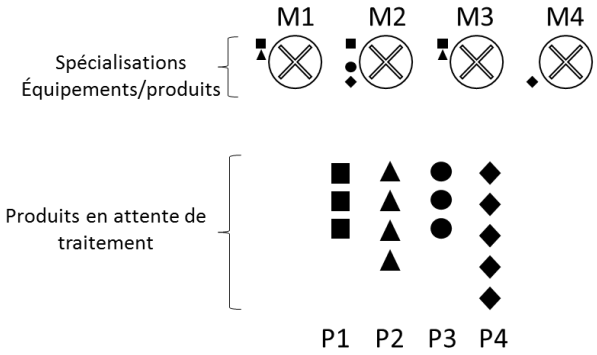

Figure 3 : Modèle de file d'attente générique représentant le cas réel étudié

Dans la figure 3, les équipements M1, M2, M3, M4 sont spécialisés respectivement sur les produits (P1, P2), (P1, P3, P4), (P1, P2) et (P4), et chaque équipement ne traite que les lots des produits pour lesquels il est spécialisé. Le choix des spécialisations des équipements pour cet exemple illustre qu'il n'y a pas de configurations spécifiques : certains équipements peuvent être spécialisés sur un seul produit, d'autres sur plusieurs (voire tous), deux équipements peuvent partager les mêmes spécialisations… c'est pour cette raison que nous parlons de « modèle générique ». Le tableau 1 résume les hypothèses définies dans le modèle classique (cf. figure 2) et si elles sont respectées par le modèle générique (cf. figure 3) :

Hypothèse	Modèle classique	Modèle générique
File d'attente unique mono-produit	OUI	NON
Spécialisation des équipements	NON	OUI
Temps de process des équipements identiques	OUI	NON
Taux de disponibilité des équipements identiques	OUI	NON
Lois d'arrivées des lots : générale	OUI	OUI
Lois de process des lots : générale	OUI	OUI

Tableau 1 : Comparaison des hypothèses entre le modèle classique et le modèle générique

Comme on peut le voir au travers des figures 2 et 3 ainsi que du tableau 1, le modèle générique introduit quatre éléments nouveaux : il n'existe pas de file d'attente unique mais un ensemble de lots de différents produits en attente de traitement ; les équipements sont spécialisés et ne traitent pas l'ensemble des produits du groupe ; les équipements ont des temps de process par produit indépendants ; les équipements ont des taux de disponibilité indépendants. Notons au passage que le périmètre du problème concerne dans les deux cas l'ensemble des équipements liés par une connexité de produits. Dans l'exemple de la figure 3, si un autre équipement venait à traiter l'un des quatre produits de l'exemple, il serait inclus dans le problème. Remarquons également que le modèle classique est un cas particulier du modèle générique où tous les équipements sont spécialisés sur un unique produit.

Notre objectif ici est de montrer la difficulté de transposer le modèle générique illustré par la figure 3 au modèle classique illustré par la figure 2. La formule de Hopp et Spearman (basée sur le modèle classique) inclut le paramètre *m* (nombre d'équipements) qu'il faut définir pour calculer le temps de cycle moyen des produits. A travers les courbes opérationnelles tracées à partir de cette formule (figure 1), on peut voir l'impact majeur de ce paramètre : pour un même taux d'utilisation, plus le nombre d'équipements *m* est élevé, plus le temps de cycle est faible. La question non triviale pour appliquer ce modèle théorique est alors la suivante : si l'on veut calculer un temps de cycle par produit, quelle valeur choisir pour le paramètre *m* de chaque produit ?

Sur l'exemple de la figure 2, le produit P1 étant qualifié sur trois équipements (nb : on parle dans cet article indifféremment de spécialisation d'équipements à des produits et de qualification de produits sur des équipements) et le produit P3 étant qualifié sur un seul équipement, il est implicite qu'à tout autres paramètres égaux P3 aura en moyenne un temps de cycle plus élevé que P1 (par lecture des courbes opérationnelles, figure 1). Le paramètre *m* est clairement différent entre ces deux produits, et on ne peut donc pas simplement choisir *m*=4 (le nombre total d'équipements) pour tous les produits. Tentons de définir le paramètre *m* pour le produit P3 : comme P3 est qualifié sur un seul équipement, une réponse logique serait *m*=1. Seulement, si on choisit *m*=1 pour P3, on considère que P3 a le même temps de cycle qu'il aurait sur une machine spécifiquement dédiée à P3 (tout autres paramètres égaux). Or, dans la réalité, les équipements « s'équilibrent » en temps réel : une suractivité ponctuelle en P3 sera absorbée indirectement par M1, M3 et M4 par une redistribution des produits P1 et P4. Cela signifie que *m*=1 sous-estimerait le temps de cycle dans ce cas. Donc *m*=1 n'est pas non plus la bonne réponse.

Nous venons de justifier que, pour le produit P3 de l'exemple de la figure 3, ni *m*=4 (le nombre total d'équipements) ni *m*=1 (le nombre d'équipements spécialisés sur P3) ne sont des réponses convenables. Or déterminer la valeur du paramètre *m* est crucial puisque, comme le montrent les courbes opérationnelles, le temps de cycle est très sensible à ce paramètre. La spécialisation des équipements a un fort impact sur le temps de cycle et représente donc une limite des modèles « classiques » de files d'attente. A ce titre, elle ne peut être négligée.

3.3 Echelle temporelle et niveau d'agrégation

Le temps de cycle moyen présent dans les différentes formules de modélisation des files d'attente est un temps de cycle « à horizon infini ». Comme on peut le voir dans la construction de ces formules par Whitt (1993), il s'agit du temps de cycle moyen qu'on obtiendrait si les paramètres restaient constants dans le temps et qu'on étendait l'horizon à l'infini. Pour notre recherche de caractérisation du temps de cycle, cela exclut de facto tous les aspects transitoires : « embouteillages » d'encours de production, « résidus » de pannes majeures (qui créent temporairement une file d'attente élevée), phénomènes de changement de mix liés à la montée ou descente en volume d'un produit. Ces périodes qui sont les plus courantes en fabrication à faible volume et forte diversité de produits, ne sont donc pas caractérisées par les modèles classiques de la théorie des files d'attente.

Une autre limite est que, étant basés sur une file d'attente FIFO, c'est-à-dire « Premier entré, premier servi », le temps de cycle calculé correspond à une moyenne sur l'ensemble des produits pour le parc équipement considéré. Les équipements sont regroupés par « type » et l'approche répond essentiellement à la question : quel est le temps de cycle moyen des lots qui passent sur ce groupe d'équipements ? Les modèles classiques de théorie des files d'attente ne permettent donc pas de déterminer le temps de cycle de lots ayant des caractéristiques différentes (comme le type de produit ou la priorité client). En particulier, la notion de priorité « casse » encore plus cette hypothèse de FIFO en permettant à des lots prioritaires de doubler des lots moins prioritaires. Or cette priorité est un aspect fondamental en microélectronique pour gérer la production face à la volatilité des marchés. Des produits à forte priorité auront des temps de cycle moyen plus faibles que des produits à faible priorité, chose que les modèles classiques de théorie des files d'attente ne permettent pas de prendre en compte.

4 NOUVELLE APPROCHE : LE WIP CONCURRENT

Le raisonnement par file d'attente est parfaitement adapté puisque, fondamentalement, c'est à partir de la file d'attente (de la quantité d'encours présente dans le système) que l'on détermine le temps de cycle : loi de Little (1961). Cependant, la spécialisation des équipements et le non-respect du FIFO par l'utilisation de priorités éloigne énormément la réalité industrielle du modèle de file d'attente unique du cas G/G/m traité par Hopp et Spearman.

Pour passer cette difficulté, nous proposons dans cet article d'étudier les files d'attente non pas du point de vue des équipements, ce qui est fait classiquement, mais du point de vue des produits. Plus précisément, chaque lot de production, entre son arrivée sur un groupe d'équipements et sa prise en charge par l'un des équipements du groupe, a « vu » une série de lots être traités consécutivement. Il s'agit, du point de vue de ce lot, de **sa** file d'attente. Faisons l'hypothèse d'un lot *i* ayant comme seule connaissance de son environnement le groupe d'équipements auquel il est associé et les dates de prise en charge d'autres lots par les équipements du groupe. Ce lot *i*, s'il fait l'hypothèse qu'il est dans une file d'attente unique FIFO, peut déduire de ces informations la longueur initiale de la file d'attente dans laquelle il a été placé. Dans la suite de cet article, nous appellerons cette file d'attente individuelle le « WIP concurrent » de chaque lot (WIP faisant référence à Work-In-Process, l'anglais pour encours de fabrication). Ce WIP concurrent sera exprimé non pas en quantité de lots, mais en *temps de process*. Cette unité permet le traitement des cas multi-produits, puisque des lots de produits avec des temps de process différents n'auront pas le même poids sur le WIP concurrent, chose qui n'aurait pas été vraie si l'unité du WIP concurrent avait été un nombre de lots.

Définition :
On définit le ***WIP concurrent*** *d'un lot i sur un groupe d'équipements G comme étant la somme des temps de process des lots traités sur l'un des équipements de G, et dont la prise en charge par un équipement s'est effectuée entre l'arrivée du lot i sur le groupe G et la prise en charge du lot i sur l'un des équipements de G.*

Une première caractérisation que permet le WIP concurrent est de séparer la part de temps de cycle provenant de quantités d'encours à traiter, de la part provenant de variations de capacité. En effet, à l'échelle de chaque lot, le temps passé en file d'attente est défini comme le ratio entre la quantité d'encours de la file d'attente (le WIP concurrent) et la vitesse d'absorption de cette quantité d'encours (la capacité du groupe d'équipements). Pour illustrer ce point, considérons une file d'attente dans un supermarché : si la file fait 10 personnes et que la capacité de la caisse est de 1 personne/minute, vous allez attendre 10 minutes dans la file… On retrouve bien entendu la loi de Little (1961) qui généralise ce ratio à l'échelle de n'importe quel système.

5 ETUDE DE CAS SUR UNE INSTANCE REELLE

5.1 Calcul du WIP concurrent

Le principe de WIP concurrent a été testé sur une instance réelle du site Crolles300 de STMicroelectronics. Le groupe d'équipements considéré est composé de quatre fours à batch. Les équipements peuvent chacun traiter deux lots d'un même produit simultanément, à condition de les commencer au même moment. Les dates d'arrivées des lots ont été extraites directement de la base de données de production pour une période de quatre-vingt jours. Les dates de début de process ont été calculées en ajoutant aux dates d'arrivées l'ensemble des temps non process enregistrés dans le MES (Manufacturing Execution System) pour chaque lot à l'étape considérée. Le WIP concurrent a été calculé pour chaque lot, en respectant la définition donnée précédemment. La figure 4 illustre la relation obtenue entre le WIP concurrent (en abscisse) et le temps de cycle (en ordonnée) pour l'ensemble des lots de cette période ; la figure 4.b est un zoom sur un quart de l'échelle de la figure 4.a.

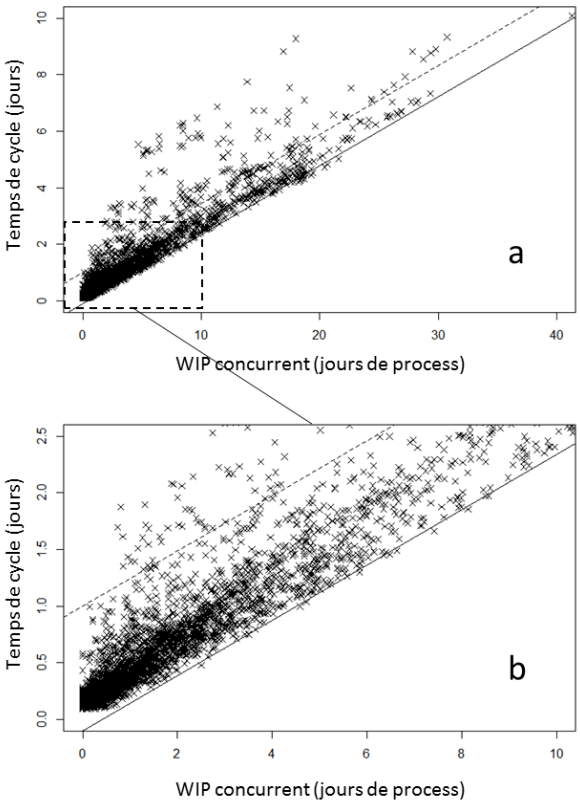

Figure 1 : Relation entre le temps de cycle et le WIP concurrent (avec zoom d'échelle)

On observe sur ces figures une relation linéaire entre le WIP concurrent et le temps de cycle, caractérisée premièrement par un coefficient de détermination R^2 de 0.90, et deuxièmement par une droite frontière basse d'équation $y = -0.1 + x/4.1$ tracée en trait continu sur la figure 4. Si l'on mesure l'écart type des distances des

points à la frontière basse, on remarque également dans notre exemple que 93% des points se trouvent en dessous de deux écarts types de cette frontière basse (traits pointillés de la figure 4). Il ne s'agit pas ici de régression linéaire ni d'intervalle de confiance puisque les hypothèses statistiques ne sont pas respectées, mais simplement d'éléments de caractérisation qui permettent d'extraire de l'information de ces courbes.

L'information essentielle de ce graphique est la relation linéaire entre le WIP concurrent et le temps de cycle. Ce premier élément de caractérisation est extrêmement important puisqu'il montre, sur l'exemple donné, que la majorité de l'information sur le temps de cycle se trouve dans le WIP concurrent. Autrement dit, malgré une très grande complexité de notre groupe d'équipements (4 équipements à batch, 19 produits différents, spécialisation des fours, taux de disponibilité et temps de process différents entre équipements), connaitre ce seul paramètre du WIP concurrent permet une première estimation du temps de cycle de chaque lot. Ce résultat n'est pas direct puisque le temps de cycle dépend théoriquement du WIP concurrent mais aussi de la capacité (la loi de Little donne la relation $WIP = temps\ de\ cycle \times capacité$). Les équipements étant soumis à des maintenances préventives, mais aussi à des pannes aléatoires, on aurait pu penser que les variations de capacité affectaient autant le temps de cycle que les variations de WIP (si on reprend l'exemple du supermarché, on estime intuitivement le temps de cycle par la taille de la file d'attente, mais on se trompe très souvent car on choisit la file la plus lente…). Sur cette instance, il semble que le WIP concurrent soit beaucoup plus impactant que les variations de capacité. On peut à priori expliquer cela par deux phénomènes : le dépassement (de lots moins prioritaires par des lots plus prioritaires) qui peut donner des WIP concurrents très différents au même instant, et la longueur des temps de process (de l'ordre de 4 heures dans notre exemple) par rapport au temps moyen de « réparation » des pannes, ce qui réduit l'impact de leur variabilité sur le temps de cycle observé. Notez que l'information peut être lue dans l'autre sens : pour obtenir un certain temps de cycle (plus ou moins long selon l'avance ou le retard du lot), il suffirait de déterminer le WIP concurrent correspondant, maitrisable par exemple par une gestion astucieuse des priorités.

5.2 Etude de la capacité

La deuxième information essentielle apportée par cette approche est l'estimation de la capacité « limite » du groupe d'équipements. Celle-ci correspond à la capacité réelle maximum du groupe d'équipements, i.e. la limite au-delà de laquelle la charge va s'accumuler et la file d'attente «exploser » sur le long terme. Elle est indispensable pour déterminer les courbes opérationnelles que nous avons vues précédemment (cf. figure 2) puisque l'abscisse correspond à l'utilisation de cette capacité.

Le WIP concurrent va permettre de déterminer plus précisément cette valeur. En effet, nous avons mis en évidence une frontière basse d'équation $y = -0.1 + x/4.1$ dans notre exemple de la figure 4. Si l'ordonnée à l'origine de cette droite est surprenante (et semble montrer un comportement différent proche de l'origine), la pente s'interprète directement par une limite de capacité (pour être exact, l'inverse de la pente s'interprète comme la limite maximale de capacité, c'est encore une fois la loi de Little). Dans notre exemple on détermine une capacité limite de 4.1 *heures de process* / heure, soit un taux de capacité limite de 51% (puisque les 4 équipements à batch de 2 ont une capacité maximale théorique de 8 *heures de process* / heure). Autrement dit, sur la période de 80 jours sélectionnée et avec le mix produit correspondant, ce groupe d'équipements a été capable de produire au maximum et de manière durable à une capacité représentant 51% de sa capacité maximale théorique. Cette information est cruciale pour le calcul des plans de charge et le dimensionnement de la production puisque cette capacité limite ne dépend pas uniquement des pannes équipements : la spécialisation des équipements, l'efficacité du batching, l'efficacité de chargement... sont autant de paramètres qui rentrent en compte dans la définition de cette capacité limite, et l'estimation par le WIP concurrent permet de prendre en compte l'ensemble de ces phénomènes, ce que l'analyse classique (par les caractéristiques des équipements) ne permet pas.

Puisque le WIP concurrent a permis, à travers la frontière basse, un calcul de capacité limite, il serait intéressant de chercher à extraire également la capacité moyenne du groupe d'équipements. Compte tenu de la distribution très inégale des points dans cette courbe (50% des points ont un WIP concurrent inférieur à 1 jour), la régression linéaire n'est pas le meilleur moyen d'obtenir la capacité moyenne. Il est préférable de calculer la capacité observée pour chaque point (ratio WIP concurrent / temps de cycle) et d'en faire la moyenne. La figure 5 montre ainsi l'histogramme des capacités calculées par cette méthode ainsi que la moyenne obtenue:

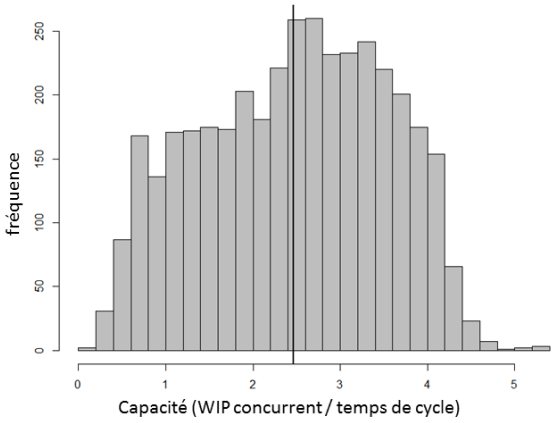

Figure 2 : Histogramme des capacités et capacité moyenne

On observe sur la figure 5 une capacité moyenne de 2.46 *heure de process* / heure, soit un taux de capacité

moyenne de 31%. De nouveau, le WIP concurrent nous a permis d'extraire une information très importante pour le dimensionnement ou les calculs de temps de cycle.

La moyenne montrée en figure 5 (correspondant à un taux de capacité moyenne de 31%) ne semble pas correspondre avec la relation visible en figure 4. En observant la figure 4, il semblerait que la capacité moyenne soit proche de la capacité limite (c'est-à-dire, si « à la main » on effectuait une régression linéaire, on affecterait une pente très proche de la pente limite). L'écart provient en fait de la répartition très hétérogène des points dans la figure 4 : 50% des points ont une abscisse (un WIP concurrent) inférieure à 1 *jour de process* et sont donc quasiment invisibles sur le graphe. Donc, si les points à fort WIP concurrent observent une capacité proche de la capacité limite mais que la moyenne sur l'ensemble de la population est plus faible, cela semble indiquer que les points à faible WIP concurrent « voient » une capacité plus faible, et qu'il y a donc une relation de dépendance entre le WIP concurrent et la capacité (surtout dans le cas d'équipements à batch).

Pour traduire numériquement ce phénomène de dépendance entre le WIP concurrent et la capacité moyenne, nous avons séparé la population de points en 10 groupes de quantiles croissants de WIP concurrents (chaque groupe a un nombre de points identique, et les WIP concurrents de tous les points d'un groupe sont inférieurs à tous les WIP concurrents des lots du groupe supérieur). Pour chaque groupe, nous avons calculé la capacité moyenne (la moyenne des ratios entre WIP concurrent et temps de cycle, de la même manière qu'en figure 5). Les capacités moyennes des 10 groupes ont été tracées en figure 6.

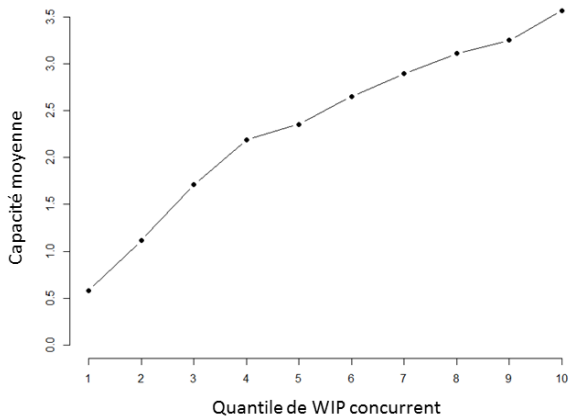

Figure 3 : relation entre le WIP concurrent et la capacité moyenne

On observe sur cette figure 6 que plus le WIP concurrent est élevé, plus la capacité des équipements est forte. Il y a plusieurs raisons possibles à cette dépendance : la première, énoncée par Shanthikumar, Ding, et Zhang, (2007), est le report de maintenances en cas de forts niveaux d'encours (donc de WIP concurrent important). D'autres raisons peuvent être une meilleure efficacité de batching lors de forts encours (la probabilité de trouver un lot avec lequel constituer le batch est plus élevée lorsque le nombre de lots en attente est élevé) et une meilleure réactivité des opérateurs aux pannes lors de forts niveaux d'encours. Par le WIP concurrent, on vient donc de mettre en évidence une forte relation de dépendance entre la capacité des équipements et les niveaux d'encours. Cette dépendance n'est pas du tout prise en compte dans les modèles classiques de la théorie des files d'attente (qui font même l'hypothèse inverse) mais des modèles basés sur le WIP concurrent pourraient intégrer cette dimension.

5.3 Vers une caractérisation du WIP concurrent

Le calcul du WIP concurrent amène à d'autres axes de recherche pour la caractérisation du temps de cycle. Notamment, il serait très intéressant de déterminer la cause de formation du WIP concurrent : dépassements plus fréquent en fort WIP, variation des arrivées, variation de capacité, WIP résiduel…

Le dépassement est la conséquence de files d'attente non FIFO et de la spécialisation des équipements : des lots à plus forte priorité peuvent doubler des lots de priorité inférieure. De plus, un lot en tête de file peut n'avoir aucun équipement libre capable de le traiter et se faire ainsi doubler par un lot qui peut être traité. La variation des arrivées est l'effet de la non-homogénéité dans les arrivées : sur une certaine période de temps, plus de lots peuvent arriver que les équipements ne sont capables d'en traiter pendant ladite période. Une variation de capacité peut également être la cause de formation de WIP concurrent : une plus faible capacité pendant la période précédant l'arrivée du lot (par exemple due à la panne d'un équipement) peut entrainer l'accumulation d'encours. Le WIP résiduel est simplement la présence d'encours antérieur au lot le plus ancien sur le groupe d'équipements du WIP concurrent : l'encours élevé sur une période peut provenir de la période précédente...

Ainsi, la prochaine étape consiste à remonter aux causes de formation du WIP concurrent pour pouvoir caractériser le temps de cycle de chaque lot en fonction de paramètres statiques (type et priorité de chaque lot, capacité limite du groupe d'équipement…) et dynamiques (encours présent à l'arrivé de chaque lot,…).

6 PERSPECTIVES D'UTILISATION

La simulation est une approche complémentaire très intéressante puisqu'elle permet de générer des scénarios et des équipements très complexes. Cependant, le problème de la simulation dans la production microélectronique complexe est la difficulté de cohérence avec la réalité : Le nombre de facteurs à prendre en compte est potentiellement très élevé (batching, temps de setups, mix de produits, mix de priorités, pannes, hétérogénéité des équipements, qualifications des équipements, règles de dispatching, disponibilités des opérateurs…) et la comparaison avec le réel est difficile. En effet, une mesure de temps de cycle moyen ou même une distribution de temps de cycle apporte peu d'information sur le comportement des équipements : pour calibrer la simulation, ce sont les caractéristiques du groupe d'équipement qu'il faut être capable d'aligner. Dans ce cadre, le WIP concurrent peut permettre d'évaluer un certain nombre de paramètres pour calibrer des simulations. Il fournit surtout un nouvel outil de comparaison entre la réalité et la simulation. Concrètement, si la simulation et le passé donnent les mêmes capacités limites (cf. Figure 4), les mêmes histogrammes de capacité (cf. Figure 5) et les mêmes relations de dépendance (cf. Figure 6) sur différentes instances du passé, la simulation sera jugée réaliste et à même de prédire des courbes opérationnelles fiables sur les scénarios du futur. Sinon, les écarts fourniront beaucoup d'indices sur l'origine des désalignements. A noter que le même mécanisme permet de comparer deux situations quelconques, pas seulement une simulation et une réalité, mais aussi, par exemple, l'état de la production à deux dates données.

Si les causes de WIP concurrent parviennent à être déterminées, une autre utilisation pourrait être l'utilisation de ces connaissances pour la projection d'encours de fabrication, à mi-chemin entre la projection à capacité infinie et la simulation à évènements discrets. En effet, il serait possible d'utiliser des paramètres statiques (calculés sur chaque groupe d'équipement comme la capacité moyenne, la capacité limite, la relation capacité/WIP, ou sur chaque lot comme la priorité moyenne des lots concurrents) et des paramètres dynamiques (quantité d'encours présent à chaque instant sur chaque groupe d'équipement dans la simulation) pour construire de manière dynamique le WIP concurrent de chaque lot et ainsi estimer son temps de cycle. Une telle simulation serait évidemment à capacité finie (par la limite de capacité), prendrait en compte l'ensemble des éléments affectant le temps de cycle (par la capacité moyenne mesurée), et aurait un coût de développement et de maintenabilité très faible (puisqu'aucun équipement n'a besoin d'être modélisé individuellement).

Pour des horizons à moyen ou long terme, ainsi que pour du dimensionnement (calcul de taux d'utilisation maximum par exemple) il reste intéressant de pouvoir calculer un temps de cycle moyen sous conditions (études communément nommées « what-if » dans l'industrie) : c'est l'intérêt majeur des formules faisant appel à la théorie des files d'attente. Une deuxième perspective est donc de trouver, à partir des paramètres et des relations établies par le WIP concurrent, une formule adaptée de l'équation de Hopp et Spearman. En effet, le WIP concurrent fait le postulat d'une file d'attente unique « équivalente » : la capacité moyenne, la capacité limite, la relation capacité / WIP concurrent en sont des premiers paramètres. Cette file d'attente unique « équivalente » contourne les deux problèmes majeurs des modèles classiques de théories des files d'attentes (soulevés en parties 2 et 3) : le choix du nombre d'équipements et la dépendance de la capacité aux arrivées. Le problème du choix du nombre d'équipements dans un groupe générique multi-produits avec équipements spécialisés (discuté en partie 3) devrait normalement se retrouver dans les caractéristiques de dépassement (un produit qualifié sur un seul équipement devrait être « dépassé » plus souvent dans la file «équivalente » et donc avoir un WIP concurrent moyen plus élevé, c'est-à-dire un temps de cycle plus élevé). La dépendance de la capacité aux arrivées (discutée dans la revue de littérature en partie 2) est maintenant mesurable (comme montré en partie 5) et est donc intégrable. L'une de nos ambitions et perspective majeure est donc de chercher à développer, à partir du WIP concurrent, une adaptation de la formule de Hopp et Spearman applicable à n'importe quel groupe d'équipement sans hypothèses restrictive.

CONCLUSION

Dans cet article, nous introduisons une nouvelle notion, celle de WIP concurrent, qui est une représentation des files d'attente du point de vue des produits. A partir de l'étude de données réelles provenant d'un parc d'équipements d'une unité de production microélectronique, nous montrons que cette représentation permet d'extraire des caractéristiques fondamentales de groupes d'équipements. Nous montrons en particulier comment obtenir la capacité limite, la capacité moyenne, ainsi qu'une mesure des effets de dépendance entre la capacité et les encours de production, très souvent ignorés dans les modèles de file d'attente de la littérature.

Le principal apport de ces travaux réside en la capacité de notre nouvelle représentation de files d'attente à intégrer l'ensemble des facteurs influents sur le temps de cycle, et ce sans hypothèse préalable. Cela ouvre plusieurs applications: premièrement, ces mesures agrégées permettent de comparer différentes situations, que ce soit la comparaison de modèles de simulation à la réalité (permettant leur calibrage) ou un suivi des performances d'un groupe d'équipement dans le temps. Deuxièmement, cette représentation permettra de reproduire de manière agrégée mais précise les interactions flux/équipements, et donc de développer des simulations rapides et maintenables. Enfin, cela offre une perspective dans le développement de modèles de file d'attente multi-produits totalement génériques, permettant de modéliser le temps de cycle moyen de divers produits à une

même étape de fabrication, et ce quels que soient les types d'équipements, leurs spécialisations, ou les caractéristiques des produits traités.

REMERCIEMENTS

Ce travail a bénéficié d'un financement de l'agence nationale de la recherche technique (ANRT) ainsi que d'un financement de la région Rhône-Alpes.

REFERENCES

Akhavan-Tabatabaei, R., Ding, S., & Shanthikumar, J. G. (2009). A method for cycle time estimation of semiconductor manufacturing toolsets with correlations. *Winter Simulation Conference* (pp. 1719-1729). Winter Simulation Conference.

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B, 20*(33-39), 16.

Etman, L. F., Veeger, C. P., Lefeber, E., Adan, I. J., & Rooda, J. E. (2011). Aggregate modeling of semiconductor equipment using effective process times. *Proceedings of the Winter Simulation Conference* (pp. 1795-1807). Winter Simulation Conference.

Hopp, W. J., & Spearman, M. L. (2001). *Factory Physics: Foundations of Manufacturing Management.* Irwin/McGraw-Hill.

Huang, M.-G., Chang, P.-L., & Chou, Y.-C. (2001, Novembre). Analytic Approximations for Multiserver. *Semiconductor Manufacturing, IEEE Transactions on, 14*(4), 395-405.

Ignizio, J. P. (2009). Cycle time reduction via machine-tooperation. *International Journal of Production, 47*(24), 6899-6906.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 338-354.

Kim, D. J., Wang, L., & Havey, R. (2014, December). Measuring cycle time through the use of the queuing theory formula (G/G/M). *Proceedings of the 2014 Winter Simulation Conference*, 2414-2421.

Kingman, J. F. (1961, Octobre). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society, 57*(04), 902-904.

Kingman, J. F. (2009, Novembre 17). The first Erlang century—and the next. *Queueing Systems, 63*(1-4), 3-12. doi:10.1007/s11134-009-9147-4

Leachman, R. C. (2012). The Engineering Management of Speed. *Proceedings of the 2012 Industry Studies Association Annual Conference*.

Little, J. D. (1961). A proof for the queuing formula: L= λ W. *Operations Research, 9*(3), 383-387.

Miltenburg, J., Cheng, C. H., & Yan, H. (2002). Analysis of wafer fabrication facilities using four variations of the open queueing network decomposition model. *IIE Transactions, 34*(3), 263-272.

Morrison, J. R., & Martin, D. P. (2006). Cycle time approximations for the G/G/m queue subject to server failures and cycle time offsets with applications. *Advanced Semiconductor Manufacturing Conference, 2006. ASMC 2006. The 17th Annual SEMI/IEEE* (pp. 322-326). IEEE.

Pollaczek, F. (1957). Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente à un guichet et par des phénomènes apparentés. *Mémorial des sciences mathématiques, 136*, 1-123.

Sakasegawa, H. (1977). An approximation formula L q≃ α· ρ β/(1-ρ). *Annals of the Institute of Statistical Mathematics, 29*(1), 67-75.

Schelasin, R. E. (2013, December). Estimating wafer processing cycle time using an improved G/G/M queue. *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, 3789-3795.

Senderovich, A., Weidlich, M., Gal, A., & Mandelbaum, A. (2015). Queue mining for delay prediction in multi-class service processes. (Elsevier, Ed.) *Information Systems, 53*, 278-295.

Shanthikumar, J. G., Ding, S., & Zhang, M. T. (2007). Queueing theory for semiconductor manufacturing systems: a survey and open problems. *Automation Science and Engineering, IEEE Transactions on, 4*(4), 513--522.

Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management, 2*(2), 114-161.

Zisgen, H., Meents, I., Wheeler, B. R., & Hanschke, T. (2008, Decembre). A queueing network based system to model capacity and cycle time for semiconductor fabrication. *Proceedings of the 2008 Winter Simulation Conference*, (pp. 2067-2074).