UNIVERSITÉ DE MONTRÉAL

GESTION CONJOINTE DE PRODUCTION ET QUALITÉ APPLIQUÉE AUX LIGNES DE PRODUCTION NON FIABLES

FATIMA ZAHRA MHADA
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE ÉLECTRIQUE)
AOUT 2011

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

GESTION CONJOINTE DE PRODUCTION ET QUALITÉ APPLIQUÉE AUX LIGNES DE PRODUCTION NON FIABLES

présentéee par : MHADA, Fatima Zahra
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. GOURDEAU, Richard, Ph.D., président.
M. MALHAMÉ, Roland, Ph.D., membre et directeur de recherche.
M. PELLERIN, Robert, Ph.D., membre et codirecteur de recherche.
M. GERSHWIN, Stanley B., Ph.D., membre externe.
M. DUFOUR, Steven, Ph.D., membre.

À mon père, à ma mère

À mes soeurs

Et à toute ma famille…

REMERCIEMENTS

Je tiens à remercier toute personne qui m'a aidé de près ou de loin à l'achèvement de ce travail.

Je remercie en premier lieu mon directeur **M .Roland Malhamé** pour sa disponibilité, ses conseils, ses remarques pertinentes et son support moral et financier le long de cette thèse.

Je remercie aussi mon codirecteur **M. Robert Pellerin** pour son apport technique et financier, son expérience industrielle et sa gentillesse.

Je remercie **M. Richard Gourdeau**, **M. Steven Dufour** et **M. Stanley Gershwin** qui m'ont honoré en faisant partie du jury de soutenance de ma thèse.

Je remercie également **M. Ricardo Camarero** pour avoir partagé son expertise sur la méthode des caractéristiques en équations aux dérivées partielles ainsi que **M. Dominique Pelletier** pour la référence fournie.

J'adresse enfin mes remerciements à tous les gens qui travaillent à la section *Automation et systèmes* et au *Gerad* ainsi que tous mes collèges, particulièrement **Javad Sadr**.

J'exprime ma gratitude à **mes parents et mes soeurs** sans qui je n'aurais pas tenu le coup pour leur encouragement et leur amour inconditionnel.

RÉSUMÉ

Cette recherche s'intéresse aux lignes de production non fiables formées de plusieurs machines satisfaisant une demande fixe de produits finis de type unique et comprenant des stocks d'encours à capacité fixe. Deux types de machines sont considérés ici : un type de machine dont une partie de la production est non conforme aux normes de la qualité et un autre type de machine dont la production est 100 % conforme. La thèse est organisée selon trois contributions principales.

L'objectif visé dans la première partie est de développer des modèles d'analyse de performance et des techniques d'optimisation efficaces pour le réglage des paramètres de conception suivant une approche de contrôle de type CONWIP (Constant Work-In-Process). Notre recherche s'inscrit dans le courant des approches de décomposition des ateliers de fabrication. L'analyse de la performance de ces systèmes aléatoires discrets/continus repose essentiellement sur les équations de Kolmogorov et le principe de la demande moyenne. De plus, nous introduisons des blocs de construction formés de paires de stock local-machine globale. La machine globale commune à toutes ces paires permet alors d'introduire une mesure de corrélation importante entre tous les blocs de construction quelle que soit la distance des stocks qui entrent dans leur composition. Ceci permet de créer des liens entre blocs de construction de la décomposition qui se situent au-delà de leurs voisinages respectifs, comme c'est le cas dans d'autres méthodes de décomposition. Cet aspect de corrélation des machines est caractéristique de la stratégie de production CONWIP. De plus, dans notre modélisation globale, la dynamique totale du stock dans la boucle CONWIP est considérée comme étant essentiellement affectée par les statistiques de fiabilité de la machine M_1, et la probabilité de disponibilité des pièces dans le stock $(n-1)$, reflétant ainsi l'opinion que le CONWIP est une forme de Kanban imposée à une collection de machines. Cette approche permet ainsi d'analyser et d'optimiser par décomposition des architectures de production hybrides en vue de développer des principes généraux de choix d'architectures en fonction des caractéristiques de fiabilité et de capacité de production des machines dans la ligne.

La deuxième partie de la recherche se situe dans un courant naissant visant à considérer de façon conjointe le problème d'optimisation de la qualité et de la paramétrisation des stratégies de production. L'originalité de notre démarche consiste à proposer un modèle aléatoire de machine qui conjugue la continuité de la production (réputée mener à des résultats analytiques intéressants et des techniques de simulation par événements discrets efficaces) à une vision continue de la qualité représentée par un état temporaire, non observable et spontanément réversible, soit de production conforme ou alternativement de production non conforme. Nous montrons aussi que la limite de ce modèle, lorsque les oscillations entre états de production conforme et non conforme respectivement deviennent très rapides, est un modèle de Bielecki-Kumar adéquatement modifié et dont la résolution analytique devient possible.

Finalement, nous présentons dans la troisième partie de la thèse des modèles d'approximation et d'optimisation pour les lignes de production avec des machines produisant des pièces qui peuvent être considérées comme conformes ou non conformes, pour minimiser le coût total de stockage et de pénurie, tout en spécifiant la localisation optimale d'une station d'inspection au sein de la ligne. Il est supposé que la fraction entre les pièces non conformes et les pièces conformes est constante pour une machine donnée. En outre, la ligne comprend deux stations d'inspection ; l'emplacement d'une des deux stations est fixe (dédié à l'inspection de pièces finies), tandis que l'emplacement de l'autre station est choisi de manière à optimiser le coût moyen total par unité de temps (coût de stockage, éventuellement coût de pénurie et coût d'inspection). Dans cette partie, nous mettrons l'accent sur l'importance d'examiner conjointement le dimensionnement des stocks et le positionnement d'une station d'inspection, ainsi que sur la relation entre les coûts d'inspection et l'inclusion ou l'exclusion des postes d'inspections.

ABSTRACT

This research is concerned with unreliable production lines. Two types of machines are considered here: a machine for which part of the production is part substandard in quality and a machine whose production is 100% in conformity. The thesis is organized according to three principal contributions.

In the first part of our research and for a given choice of the maximum allowable total storage parameter, the performance of constant work-in-process (CONWIP) disciplines in unreliable transfer lines subjected to a constant rate of demand for parts, is characterized via a tractable approximate mathematical model. For a $(n-1)$ machines CONWIP loop, the model consists of n multi-state machine single buffer building blocks, separately solvable once a total of $(n-1)^2$ unknown constants shared by the building blocks are initialized. The multi-state machine is common to all building blocks, and its n discrete states approximate the joint operating state of the machines within the CONWIP loop; each of the first $(n-1)$ blocks maps into a single internal buffer dynamics, while the n^{th} building block characterizes total work-in-process (wip) dynamics. The blocks correspond to linear n component state equations with boundary conditions. The unknown (shared) constants in the block dynamics are initialized and calculated by means of successive iterations. The performance estimates of interest, mean total wip, and probability of parts availability at the end buffer in the loop are obtained from the model and validated against the results of Monte-Carlo simulations.

In the second part of our research, we address the optimal production control problems for an unreliable manufacturing system that produces items that can be regarded as conforming or non conforming. A new stochastic hybrid state Markovian model with three discrete states, also called modes is introduced. The first two, operational sound and operational defective are not directly observable, while the third mode, failure, is observable. Production of defective parts is respectively initiated and stopped at the random entrance times to and departure times from the defective operational mode. The intricate piecewise-deterministic dynamics of the model are studied, and the associated Kolmogorov equations are developed under the suboptimal class of hedging policies. The behavior of the model is numerically investigated, optimized under hedging policies, and subsequently compared to that of a tractable extension of the two-mode Bielecki-Kumar single machine model, where both conforming and defective parts are simultaneously produced in the operational mode, while the ratio of produced non conforming to conforming parts remains fixed.

Finally, we consider a fluid model of an unreliable production line consisting of n machines and n fixed buffer sizes. These machines produce a single part type with two different quality levels: conforming and non-conforming parts. The ratio of non-conforming parts to conforming ones is assumed to be a constant, which may vary depending on the machine. The production line can contain inspection stations whose function is to reject the non-conforming parts from the system. It is assumed that the production line must meet a constant rate of demand for good parts. The objective is to develop an approximate modeling framework and an optimization algorithm for unreliable transfer line inter machine buffer sizing, so as to minimize, under a constant demand for parts rate, the average long term combined storage and shortage costs, while accounting for parts quality and specifying the optimal location of inspection stations. Decomposition / aggregation methods developed in (Sadr et Malhamé (2004b)) and their dynamic programming based optimization algorithm are adapted to the current model. In addition, numerical results based on the approximate theory, and those obtained from Monte-Carlo simulation, are contrasted.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ v

ABSTRACT vii

TABLE DES MATIÈRES ix

LISTE DES TABLEAUX xiii

LISTE DES FIGURES xiv

LISTE DES ANNEXES xvi

LISTE DES SIGLES ET ABRÉVIATIONS xvii

INTRODUCTION 1

CHAPITRE 1 Revue de la littérature 6

- 1.1 Introduction 6
- 1.2 Ligne de production 6
- 1.3 La gestion de stock et les politiques de production 7
- 1.4 Production/Qualité 8
- 1.5 Production/Qualité/Maintenance 9
- 1.6 Limitations 10
  - 1.6.1 La politique CONWIP 10
  - 1.6.2 Production/Qualité 11

CHAPITRE 2 Méthodologie de recherche et organisation générale de la thèse 12

CHAPITRE 3 Approximate performance analysis of CONWIP disciplines in unreliable non homogeneous transfer lines 15

3.1 Introduction 16

3.2 Mathematical model of the transfer line and statement of objectives. 18

3.3 Background material 21

3.3.1 Forward Kolmogorov equations for modewise constant velocity hybrid-state Markov processes. 21

3.3.2 A transfer line decomposition related approximation: the demand averaging principle. 25

3.4 Approximate CONWIP loop modeling 28

3.4.1 General principles of the decomposition methodology 28

3.4.2 Details and justification of model component dynamics 29

3.5 Iterative calculation of system solution 37

3.6 Numerical results 39

3.7 Conclusion 42

CHAPITRE 4 A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects 44

4.1 Introduction 45

4.2 Model description 46

4.3 Piecewise-deterministic stock dynamics 49

4.3.1 Region $0 < x_1(t) + x_2(t) < z$ 49

4.3.2 Region $x_1(t) + x_2(t) = z$ 52

4.3.3 On line $x_2(t) = 0$ 54

4.4 Steady-state Forward Kolmogorov equations 55

4.4.1 Region $(x_1 > 0) \cap (x_2 > 0) \cap (0 < x_1 + x_2 < z)$ 55

4.4.2 On the line: $x_2 = 0$, $x_1 < z$ 58

4.4.3 On segment $S = \{(x_1; x_2)|(x_1 \in [0; z]) \cap (x_2 = z - x_1)\}$ 59

4.4.4 On the left hand boundary trajectory 61

4.5 Boundary conditions 62

4.5.1 At $(0,0)$ for univariate density $f_2^0(x_1)$ 62
4.5.2 At $(0,0)$ for univariate density $f_3^0(x_1)$ 64
4.5.3 On the half-line $x_1 : 0| \, x_1 < z$ 65
4.5.4 On the left hand boundary trajectory 66
4.5.5 On segment $S = \{(x_1; x_2)|(x_1 \in [0; z]) \cap (x_2 = z - x_1)\}$ 66
4.5.6 At point $x_1 = z, x_2 = 0$ 66
4.5.7 Normalization equation 67
4.6 The numerical algorithm 67
4.6.1 Validation 69
4.7 Comparison with the modified Bielecki-Kumar machine 70
4.7.1 Determining $p_i(x)$ 70
4.7.2 Determining $g_i(x_1 + x_2)$ 71
4.8 Conclusions 74

CHAPITRE 5 Unreliable production lines with defective parts and inspection stations 76
5.1 Introduction 77
5.2 Modeling framework and statement of the optimization problem 78
5.3 Single machine building blocks for transfer line decomposition 81
5.3.1 The BK model 82
5.3.2 Hu model 82
5.3.3 The BK quality modified model 83
5.4 Flow line decomposition 85
5.4.1 The combined wip supply / machine building blocks 85
5.4.2 The demand averaging principle (DAP) 87
5.4.3 Recursive calculation of the q_i's 88
5.4.4 Evaluation of internal buffer i induced costs 89
5.4.5 Evaluation of buffer n induced costs 90
5.5 Approximate model validation 92
5.6 Optimization 92
5.6.1 Numerical procedure 94

5.6.2 A homogeneous machines line 94
5.7 Added value of joint buffer optimization and inspection station positioning . . 96
5.8 Further numerical results of interest 100
5.8.1 Influence of inspection costs per part on the optimal positioning/ necessity of the internal inspection station 100
5.8.2 Assessing the influence of an individual poor quality machine 102
5.9 Conclusions and future work 103

CHAPITRE 6 Discussion générale 105

CONCLUSION 107
6.1 Synthèse des travaux 107
6.2 Améliorations futures 108

RÉFÉRENCES 109

CHAPITRE ANNEXES 113

LISTE DES TABLEAUX

Table 3.1 Machine parameters 39

Table 3.2 Total wip 42

Table 3.3 A 3 machine CONWIP loop ($n - 1 = 3$) 42

Table 3.4 A 4 machine CONWIP loop ($n - 1 = 4$) 42

Table 4.1 The system parameters 69

Table 4.2 The simulation and analytical results. The confidence interval for simulated probability is ±0.005 and for simulated cost is ±0.05 69

Table 4.3 The system parameters 73

LISTE DES FIGURES

Figure 3.1 Transfer line with unreliable machines, storage levels x_i, and constant rate of demand d for parts. All storage levels are bounded above and non negative, except for x_n which can become negative (demand backlog). 19

Figure 3.2 Direction of modewise constant velocity vector associated with scalar continuous state variable $x(t)$ in region R_i in modes $j = 1, ..., M$. v_j^i, $j = 1, ..., M$ are signed velocities. 23

Figure 3.3 Type 1 boundary: abrupt velocity change and zero sojourn time of $x(t)$ at $x = a_i$; probability conservation imposes that probability currents remain continuous across the boundary. 23

Figure 3.4 Type 2 boundary: non zero sojourn time of $x(t)$ at a_i; a differential equation is associated with probability mass $P_{a_i}^j(t)$ sitting at $x(t) = a_i$ and mode j; probability conservation imposes that incoming probability current in mode j contribute to increasing probability mass $P_{a_i}^j(t)$, while probability losses from $P_{a_i}^j(t)$ contribute either to an increase in other probability masses or downwards oriented probability currents at a_i. 24

Figure 3.5 Graphical representation of the steady- state forward Kolmogorov equations associated with CONWIP model building block $\tilde{x}_i(t)$ $i = 0, 1, ..., (n-1)$. 30

Figure 3.6 Convergence of CONWIP loop service level estimate a_{n-1} $(n = 5)$ for the 5 machine transfer line, maximum storage parameter $z = 9$ and an initialization at 1. 40

Figure 3.7 Convergence behavior of conditional coefficients of wip availability estimates $a_{n-1|0j}$ $j = 1, ..., 4$, for the $n = 5$ machine transfer line, maximum storage parameter $z = 9$ and an initialization of all estimates at 1. . . 41

Figure 4.1 The continuous time Markov chain machine model. 47

Figure 4.2 Phase diagrams in mode $\alpha(t) = 1$. 49

Figure 4.3 Phase diagrams in the mode $\alpha(t) = 2$ with $x_2(0) = 0$. 51

Figure 4.4 Phase diagrams in the mode $\alpha(t) = 3$ 52

Figure 4.5 The boundary zone of the area $0 < x_1 + x_2 < z$. 53

Figure 4.6 Boundary $x_1 + x_2 = z$ in mode 1. 60

Figure 4.7 Boundary $x_1 + x_2 = z$ in mode 2. 61

Figure 4.8 Mode 2 related probability transfers near $(0, 0)$. 63

Figure 4.9 Mode 3 related probability fluxes near $(0, 0)$. 64

Figure 4.10 Boundary $x_1 > 0$, $x_2 = 0$ with switching from mode 1 to mode 2. . . . 65

Figure 4.11 Comparison of $p_0(x)$ with $g_0(x)$. 72

Figure 4.12 Comparison of $p_1(x)$ with $g_1(x)$. 73

Figure 5.1 The pseudo-machine $\tilde{M}_{i+1}$, $i = 1, ..., (n - 1)$ 86

Figure 5.2 Flow line with inspection stations. 87

Figure 5.3 The optimal solution. 95

Figure 5.4 $a_n^{des} = 0.95$. 97

Figure 5.5 $a_n^{des} = 0.85$. 98

Figure 5.6 The optimal stock sizes 99

Figure 5.7 Influence of c_I on optimal inspection station position 100

Figure 5.8 Relationship between c_I and the need for an internal inspection station 101

Figure 5.9 Exploring the influence of a single poor quality machine in an otherwise homogeneous transfer line. 103

Figure A.1 CONWIP loop approximate macromachine under the single permissible machine failure assumption 113

LISTE DES ANNEXES

Annexe A Approximate loop mode Markov chain parameters under the single permissible machine failure assumption. 113
Annexe B Derivation of (5.7) and (5.12) 115
Annexe C Derivation of (5.10) 117

LISTE DES SIGLES ET ABRÉVIATIONS

CONWIP	Constant Work-In-Process
WIP	Work-In-Process

Caractères usuels

a_i	Coefficient de disponibilité du stock x_i.
c_I	Coût unitaire d'inspection par unité de temps.
c_n	Coût unitaire de pénurie par unité de temps.
c_p	Coût unitaire de stockage par unité de temps.
d	Taux de la demande des produits finis.
$E[.]$	L'espérance mathématique.
k_i	Taux maximal de production de la machine M_i.
M_i	Machine numéro i.
$\tilde{M}_i$	Machine active numéro i approximativement isolée.
n	Nombre de machines.
$Pr[.]$	La probabilité.
p_i	Taux de panne de la machine M_i.
$\tilde{p}_i$	Taux de panne de la machine active $\tilde{M}_i$.
r_i	Taux de réparation de la machine M_i.
$\tilde{r}_i$	Taux de réparation de la machine active $\tilde{M}_i$.
$u_i(t)$	Taux de production de la machine M_i à l'instant t.
$x_i(t)$	Le stock i a l'instant t.
$x_{i_1}(t)$	Le stock i de bonne pièces à l'instant t.
$x_{i_2}(t)$	Le stock i de pièces défectueuses à l'instant t.
z_i	Seuil critique pour le stock i.

Caractères grecs

$\alpha_i(t)$	Processus binaire de l'état de la machine M_i à l'instant t.

$\tilde{\alpha}(t)$ Processus binaire de l'état de la machine approximative $\tilde{M}_i$ à l'instant t.

β_i La fraction des pièces défectueuses par rapport aux bonne pièces pour de la machine M_i.

$\lambda_i(t)$ Processus binaire de l'existence de la station d'inspection à la sortie du stock i à l'instant t.

INTRODUCTION

Une ligne de fabrication, également connue sous le nom de ligne de production, est composée par des flux de matériaux, des aires de travail et des zones de stockage. Le flux de matières est défini par la séquence aire de travail - zone de stockage - aire de travail, en passant une fois seulement dans une séquence fixe par chaque aire de travail et de stockage. Le flux de matières ou encore le flux des pièces est considéré comme continu. Il entre dans la ligne sous forme de matière première et sort comme produit fini prêt à être livré au client. Les aires de travail sont les lieux où se fait la transformation de la matière : les machines. Ces machines ne sont pas toujours fiables puisqu'elles sont en général sujettes à des pannes. Le temps que les pièces passent dans la ligne de production est donc variable.

Dans ce contexte, les aires de stockage inter-machines jouent un rôle important dans le maintien de la productivité de la ligne, puisque les machines sont en mesure de produire tant qu'elles disposent de pièces sur lesquelles travailler (stock en amont) et d'espace pour entreposer les pièces partiellement usinées (stock en aval).

L'importance de l'étude de ce type de ligne s'explique par le fait qu'elles sont coûteuses à installer (en particulier s'il s'agit de lignes de fabrication flexibles) et qu'elles constituent un mode de production très fréquent dans l'industrie moderne. Dans la littérature, les recherches liées à ce type de lignes ont suscité beaucoup d'intérêt, surtout en ce qui concerne les questions de dimensionnement des aires de stockage et de gestion des stocks, de gestion de la qualité, de gestion de la maintenance, ainsi que du choix de la politique de gestion de production. Nous discutons ci-après brièvement chacune de ces thématiques dans le contexte des chaînes de fabrication.

La gestion de stock : Qu'il s'agisse de matière première, d'encours de fabrication ou de produits finis, le stock sert à minimiser les risques d'interruption de la production liés à des pannes ou à la différence entre les cadences de production des machines et contribuant ainsi à une productivité accrue de la ligne de fabrication. Cependant, ce gain de productivité doit être mis en perspective avec les coûts additionnels qui l'accompagnent (coût de stockage, capital immobilisé). Par ailleurs, une politique zérostock n'est pas toujours la meilleure solution à envisager puisqu'elle fragilise la ligne de fabrication et qu'une incapacité à maintenir la cadence de production visée aura pour conséquence des pénuries de produits finis ou des retards de livraison pouvant mener à court terme à une dégradation de l'image de l'entreprise. Un niveau de stockage d'encours et de produits finis correspond à un « bon » compromis entre coûts effectifs de pénuries ou de retards d'une part, et d'autre part les coûts accompagnant le stockage lui-même.

La gestion de la qualité : Son objectif est le développement de techniques de vérification et de maintien de la qualité des produits, ainsi que la spécification de la norme de qualité à atteindre en fonction d'objectifs économiques ou autres (ex. image d'entreprise à maintenir). L'objectif de maintien de la qualité dépend étroitement d'une habileté à diagnostiquer les sources de problèmes (mesures et techniques de diagnostic) et des stratégies de prévention des problèmes et de correction de ces derniers lors d'un diagnostic. Ce dernier volet est étroitement lié à la prochaine thématique discutée : la gestion de la maintenance.

La gestion de la maintenance : Elle regroupe les maintenances corrective, préventive et la maintenance conditionnelle. La maintenance corrective consiste à n'intervenir que suite à une panne visible de machine. Bien que moins coûteuse a priori en termes de coût de maintenance, elle peut engendrer des coûts indirects importants à cause des longues périodes d'interruption de production qui peuvent être engendrées suite à une panne. C'est le mode de maintenance le plus fréquent. Contrairement à la maintenance corrective qui décrit des règles de comportement face à la panne totale d'un élément du système, la maintenance préventive consiste en des vérifications et remises à niveaux des machines sur une base périodique régulière (une fois par mois par exemple, ou après un certain nombre d'heures d'opération). Elle peut être très coûteuse, mais réduit le nombre d'arrêts non planifiés de la chaîne. Enfin, une orientation en maintenance qui suscite beaucoup d'intérêt est celle de la maintenance conditionnelle. La maintenance conditionnelle est de type réactif suite à certaines observations ou diagnostics sur la performance des machines ; par exemple, une fraction de pièces non conformes à la qualité spécifiée par rapport au total des pièces produites pourrait constituer un critère d'envoi en maintenance. La maintenance conditionnelle constitue un bon compromis entre les deux formes précédentes de maintenance, mais sa conception et son implantation sont plus complexes.

Le choix de la politique de production : Parmi les politiques de production les plus connues, on reconnait les approches dites Kanban et CONWIP, ou encore une combinaison des deux. Tout d'abord, Kanban signifie en japonais étiquette, ou carte. La politique fonctionne de la manière suivante : chaque machine perçoit l'état de stock de l'inventaire qui la suit immédiatement. Ainsi, chaque machine régularise elle-même le stock de son inventaire. De manière générale, une machine en état de marche (hors panne et pénurie) qui suit la méthode Kanban accélérera sa production lorsque la valeur du stock de l'inventaire directement en aval n'est pas jugée suffisamment élevée, et inversement, réduira son taux de production lorsque le stock aura atteint un certain seuil maximal, à ne pas dépasser. Pour sa part, une approche CONWIP agit différemment sur la circulation de l'information. Elle fait en sorte de réguler le stock des machines qui forme la boucle CONWIP. De manière générale, l'information provient de l'inventaire de la dernière machine de la boucle et arrive à la première machine. La première machine, renseignée sur la quantité de pièces produites en aval de la chaîne (ce qui équivaut au nombre de pièces qui sortent du dernier inventaire), va accepter ou non de produire. Le flux en entrée de la boucle CONWIP devient équivalent au flux en sortie. La première machine est, par conséquent, seule capable de réguler le nombre global de pièces dans la boucle CONWIP, les autres machines produisant, lorsqu'elles le peuvent, toujours à leur taux maximal sans se soucier de la régulation des stocks.

Parmi les domaines de recherche cités dans les paragraphes précédents, notre projet de recherche porte sur trois aspects (CONWIP vs Kanban, gestion de la qualité et gestion de la production) regroupés en deux parties distinctes.

La première partie de la recherche proposée est centrée entièrement sur les questions de production et consiste à étendre les méthodes d'analyse des chaînes de fabrication par décomposition tel que développées pour les stratégies de production de type Kanban, au cas des chaînes de fabrication sous la stratégie de production CONWIP.

Il y a une riche littérature qui traite des lignes de production selon différents points de vue. Nous nous sommes intéressés aux méthodes de décomposition comme solution aux problèmes d'optimisation et d'analyse. La plupart de ces méthodes ont été appliquées à la politique de gestion de production Kanban. Nous nous intéressons ici à la politique CONWIP qui est une extension de Kanban puisque la limite de stock n'est plus imposée à la machine (le cas Kanban) mais à un ensemble de machines qui forme ce qu'on appelle la boucle CONWIP.

Les méthodes de décomposition consistent, pour la plupart, à produire des blocs formés par un nombre de machines. Par exemple, un bloc peut être constitué d'une machine et de deux stocks (amont et aval), ou une machine et le stock en amont ou en aval. Dans cette partie de la thèse, nous proposons un nouveau modèle mathématique basé sur une décomposition approximative où le principe de la demande moyenne joue un rôle important. Ce principe d'approximation s'est montré efficace dans des travaux antérieurs sur les lignes de production non-fiables, que ce soit pour les méthodes de décompositions ou les méthodes d'agrégation [Sadr et Malhamé (2004a), Sadr et Malhamé (2004b)]. Il repose sur la simple observation que pour qu'une machine puisse répondre à une demande constante de pièces finies d, il faut qu'elle puisse produire plus que d pendant qu'elle est capable de produire pour compenser le temps où elle ne peut pas le faire à long terme. La moyenne de production doit toutefois être égale à d. Notre décomposition approximative se base elle aussi sur le principe des blocs où dans chacun d'eux on applique les équations de Kolmogorov [Malhamé et Boukas (1991)]. L'objectif de cette décomposition est de calculer différents indicateurs de performance (le stock total moyen, les coefficients de disponibilité) pour la ligne de production non fiable contrôlée par une boucle CONWIP et composée de $(n-1)$ machines et de $(n-1)$ stocks. La méthodologie et les résultats obtenus pour cette partie de la recherche sont détaillés dans le chapitre 3.

Alors que traditionnellement, la question de l'optimisation d'une stratégie de production donnée (ex. dimensionnement optimal des niveaux de kanbans dans une chaîne de fabrication produisant sous la stratégie Kanban) s'est faite en dehors de toute considération de qualité, **nous envisageons dans la deuxième partie de la thèse d'étudier l'optimisation des paramètres de telles stratégies dans un contexte de qualité imparfaite, avec pannes possibles et inspections.** À l'exception d'un nombre limité de travaux, la plupart des modèles étudiés considèrent séparément l'analyse de qualité dans les chaînes de fabrication (position des stations de vérification, les règles pour aller en maintenance préventive, la fréquence d'échantillonnage) et le développement des stratégies de gestion de production (Kanban, CONWIP...).

Or, Inman *et al.* (2003) ont démontré l'existence de plusieurs catégories de décisions qui affectent simultanément la qualité et la productivité de l'industrie manufacturière. **L'objectif de la deuxième partie de la thèse est donc d'évaluer et de développer des stratégies conjointes de gestion de production et de qualité pour des lignes de production non fiables et de valider leur applicabilité dans un contexte stochastique et dynamique**. Plus précisément, notre modèle porte sur une machine non fiable munie d'un stock, qui peut produire des pièces conformes et non-conformes et dont on tire une demande constante de bonnes pièces d à partir du mélange des pièces (voir chapitre 4). L'étude de ce modèle permet d'optimiser la production en tenant compte de la situation de la qualité associée et qui n'est que partiellement observée. Par la suite, nous ferons une extension des principes d'action dégagés vers des cas de chaînes de fabrication (chapitre 5).

Ainsi, les questions qui nous concernent dans cette thèse sont les suivantes :

- **Pour une limite de stock total donnée, quel est le coût de stockage à payer ?**
- **Pour garantir un coefficient de disponibilité requis, quel sera le prix moyen de stockage à payer ?**
- **Quel est l'impact de la prise en considération de la qualité sur la gestion de stock ?**
- **Quel est l'impact de l'introduction d'une station d'inspection dans une ligne de fabrication sur la gestion de stock ?**

Le reste de la thèse est organisé comme suit. Tout d'abord, le chapitre 1 présente une revue de littérature suivi par le chapitre 2 qui présente l'organisation générale du document. Les trois contributions principales décrites précédemment suivent alors au sein des chapitres 3, 4 et 5. Le chapitre 6 vient conclure cette thèse en présentant les principaux résultats obtenus, les limitations de cette recherche et nos recommandations, en plus de discuter de recherches futures potentielles.

CHAPITRE 1

Revue de la littérature

1.1 Introduction

Les lignes de production ont suscité beaucoup d'intérêt dans les trente dernières années, que ce soit dans le domaine des méthodes de décomposition et agrégation, de la gestion de stock, de la gestion de qualité ou de la gestion de maintenance. Chacun de ces éléments est traité de façon successive dans les sections suivantes, avant de conclure le chapitre par une revue critique des principales limitations de ces approches par rapport à notre contexte d'étude.

1.2 Ligne de production

Une ligne de production peut être considérée comme un flux discret ou continu. Le flux discret est le plus représentatif de la ligne, mais il est très difficile à modéliser [Buzacott (1967)]. Le modèle continu est assez utilisé dans les problèmes d'optimisation [Zimmern (1956)] et peut remplacer le modèle discret quand le temps entre les pannes et le temps de réparation sont plus importants que le temps de traitement [Xie (1986)]. Pour le modèle discret, Gershwin et Schick (1983) ont donné une solution exacte dans le cas de trois machines. Toutefois, il est nécessaire de faire des approximations de type décomposition [Zimmern (1956) et Gershwin (1987)], agrégation [Terracol et David (1987) et Ancelin et Semery (1987)] ou les deux [Sadr et Malhamé (2004a)] dans les cas de lignes de production avec un nombre plus important de machines.

Zimmern (1956) a ainsi proposé une méthode de décomposition basée sur des blocs formés de deux machines et un stock intermédiaire. Donc pour une ligne de n machines, on dispose de $(n-1)$ blocs. Cette même méthode a été la base des travaux de Gershwin. La méthode d'agrégation de Terracol et David (1987) et Ancelin et Semery (1987) consiste à agréger une ligne à une seule machine (macro-machine). Ils commencent par remplacer, dès le début de

la chaîne, chaque 2 machines - 1 stock par une machine jusqu'à la fin de la chaîne.
Dans le même esprit, [Sadr et Malhamé (2004a)] ont développé une méthode de décomposition basée sur les blocs « machine-stock » et une méthode d'agrégation qui consiste à remplacer le bloc machine-stock par une macro-machine.

Dallery et Gershwin (1992) ont d'ailleurs présenté une revue de la littérature, pour les modèles développés avant 1992, en y abordant les hypothèses liées à ce type de problème et les méthodes de décomposition et d'agrégation utilisées.

1.3 La gestion de stock et les politiques de production

Dans les chaînes de fabrication, un type de pièce requiert un passage à travers plusieurs stades pour être produit. Si l'atelier est flexible, il pourra être reconfiguré assez rapidement pour produire plusieurs types de pièces. La productivité de la chaîne dépend de celle de chacun de ses éléments. Or, les machines individuelles sont sujettes à des pannes aléatoires. D'autres aléas peuvent également perturber le processus de production : absence d'opérateurs, retards dans les livraisons de matériaux, etc. Ainsi, dans le cas d'absence d'aires de stockage inter-machines et suite à une panne de machine individuelle, les machines en amont seront bloquées. Alors, celles en aval manqueront de pièces et la chaîne s'arrêtera de produire. Dans un tel contexte, on peut voir qu'il est essentiel d'assurer des aires de stockage intermédiaire pour permettre à la chaîne de continuer de produire durant les phases de réparation de machines. Si ces aires de stockage étaient illimitées, la chaîne pourrait atteindre sa limite théorique supérieure de productivité. Toutefois, maintenir des aires de stockage importantes peut représenter des coûts élevés alors que la présence de hauts niveaux de stocks dans la chaîne est souvent associée à l'immobilisation de capitaux importants et sera inévitablement synonyme de temps de transit importants dans l'atelier. Des compromis entre coûts de stockage et coûts de perte de productivité doivent être ainsi envisagés.

Pour toutes ces raisons, il est essentiel de développer des méthodes d'analyse approximative et d'optimisation qui permettent de dimensionner les aires de stockage de façon optimale pour des machines de taux de production et de statistiques de pannes/réparations données.

Il y a un ensemble de publications qui se sont intéressées aux méthodes approximatives de décomposition dans le cas de la politique Kanban. Parmi elles, on cite Chiang *et al.* (2000),

Dallery et Gershwin (1992), Dallery et Bihan (1999), Gershwin (1987), Levantesi *et al.* (2003), et Sadr et Malhamé (2004a). En comparaison avec Kanban, peu de travaux ont été effectués sur les lignes de production en boucle fermée avec des machines non fiables et des stocks à capacité finie. Parmi ceux-ci, Spearman *et al.* (1990) ont introduit une solution de rechange à Kanban dénommée CONWIP (CONstant Work In Process), où la limite du stockage n'est pas sur les stocks encours, mais sur le nombre d'encours total de la ligne de production.
Frein *et al.* (1996) ont pour leur part formulé les hypothèses suivantes :

i) les pannes sont indépendantes de la production comme dans le cas de Dallery et Gershwin (1992),

ii) les pièces circulent dans la boucle de CONWIP à l'aide de pallettes dont le nombre est fixe est égal à N,

iii) le « processing time » est déterministe et identique pour toutes les machines : on parle alors de ligne homogène. Dans un CONWIP normal, la limite est imposée sur le total du stock mais ici la limite est plutôt imposée sur la moyenne totale du stock.

Gershwin et Werner (2007) ont développé une méthode d'approximation pour évaluer la performance de grands systèmes de production non fiables en boucle fermée avec un nombre constant de palettes. Les pièces sont chargées sur les transporteurs et attachée à la palette à la première machine, prêtes à subir toutes les opérations nécessaires. Après l'achèvement des opérations, les pièces finies sont déchargées et les palettes sont libérées et renvoyées à la première machine.
Il est aussi possible de combiner Kanban et CONWIP [Bonvik (1996), Bonvik *et al.* (2000)] pour générer une performance plus élevée que dans le cas de Kanban ou CONWIP séparé. En comparant différentes politiques de gestion pour une ligne de production, Bonvik conclut que la politique formée de CONWIP et du stock fini, donne la meilleure performance en terme de niveau de service.

1.4 Production/Qualité

Dans la littérature, l'intégration de la gestion de qualité avec la gestion de production a suscité beaucoup d'intérêt dernièrement bien qu'il n'existe à ce jour qu'un nombre restreint d'articles à ce sujet. Parmi ceux-ci, Kim et Gershwin (2005) ont mis au point un modèle continu qui étudie l'interaction de la qualité et de la productivité. Le modèle se base sur l'hypothèse qu'une machine continue de produire des pièces défectueuses jusqu'à ce que son fonctionnement soit corrigé. Par la suite, Kim et Gershwin (2008) ont étendu ce modèle à l'analyse de quelques représentations des lignes de production. Dans ces deux articles, le flux des pièces est considéré comme continu alors que lorsqu'on parle de qualité, on parle généralement de flux discret.

A l'opposé, Colledani et Tolio (2005) et Colledani et Tolio (2006a) s'attaquent simultanément à la qualité et aux questions de production dans un cadre entièrement discret. Ils étudient un système de production formé de postes de fabrication et de postes d'inspection non fiables caractérisés par différents modes de pannes. En outre, le comportement des machines peut être contrôlé par des cartes de contrôle (Statistical Process Control - SPC) qui sont utilisées pour générer l'information sur l'état de la machine. Toutefois, seulement des lignes synchrones, caractérisées par des machines ayant le même temps de traitement, ont été examinées et le temps d'inspection n'a pas été directement pris en compte.

Après cela, Colledani et Tolio (2006b) ont introduit l'inspection off-line à ce modèle. Gershwin et Schick (2007) ont aussi présenté une taxonomie dans le domaine d'intégration de la gestion de qualité avec la gestion de production.

1.5 Production/Qualité/Maintenance

Les modèles qui intègrent la production, la qualité et la maintenance portent principalement sur l'étude du cas où la machine passe d'un état où elle produit des pièces conformes à un état où elle produit des pièces non-conformes. Presque tous ces travaux s'appuient sur le fait que la non-qualité est un résultat de la dégradation des machines et c'est ici qu'intervient la notion de maintenance. Dans ce type de problème, la gestion de stock joue un rôle de sécurité pendant les périodes de maintenance, en améliorant les conditions de travail des machines et en diminuant le taux des pièces non-conformes dans le système.

Parmi ces travaux, Ben-Daya (2002) présente une ligne de production qui commence à produire des pièces non-conformes après une certaine période. Le taux de passage à l'état « hors contrôle » se fait par un taux croissant et que la maintenance préventive vise à réduire. Ce modèle est la base des travaux de recherche de Chakraborty *et al.* (2008) où le système de production peut, en plus de passer d'un état « contrôle » à un état « hors contrôle », tomber en panne (aléatoire). Le système est inspecté périodiquement et selon l'état du système, des mesures d'intervention sont prises. L'objectif est donc de trouver la politique d'inspection optimale en tenant compte de l'aspect stochastique du modèle.

Ben-Daya *et al.* (2006) ont aussi développé des modèles intégrés de contrôle des stocks avec ou sans le remplacement des articles non conformes. Les politiques d'inspection sont : pas d'inspection, inspection échantillonnée, inspection à 100 %. L'inspection des pièces se fait quand un lot est reçu chez le client. Après l'inspection, on détermine la fraction des pièces non-conformes, qui est supposée être une variable aléatoire (suit une distribution beta). Donc, la quantité et la politique d'inspection sont des variables de décision dans ce modèle. Contrairement à [Chang (2004)] qui propose un modèle où tous les items sont inspectés, ceux qui sont non conformes sont vendus à un prix réduit.

1.6 Limitations

1.6.1 La politique CONWIP

Les travaux sur CONWIP, basés en majorité sur la théorie développée par Gershwin, se limitent à étudier le cas où le stock total est fixe et le cas où la moyenne de stock total est constante. Donc au lieu d'avoir $\sum x_i(t) = z$, comme dans le cas des travaux basés sur le travail de Gershwin, nous supposerons dans nos travaux que $\sum x_i(t) \leq z$.

Notons que l'hypothèse d'un stock total d'encours toujours égal à z, peut être réconciliée avec l'opération d'un CONWIP, à condition d'inclure dans ce stock les cartes d'autorisation de production qui représenteraient alors le stock en amont de la première machine dans la boucle CONWIP. Un tel formalisme permet alors d'évaluer (Gershwin et Werner (2007)) le taux de production moyen maximal d'une telle boucle an fonction de z, ainsi que les niveaux moyens d'encours de chaque type, réel ou virtuel. Cependant, ce formalisme ne permet pas a priori de faire ces mêmes calculs lorsque la cadence de production est imposée extérieurement, et ne permet pas non plus de calculer le coefficient de disponibilité des encours dans le stock situé au bout de la ligne CONWIP (taux de service de la boucle CONWIP). Ce coefficient est imposé dans notre formalisme, rompant ainsi la symétrie circulaire importante pour l'analyse de Greshwin et Werner.

1.6.2 Production/Qualité

Dans les travaux effectués dans ce domaine, le modèle est basé sur l'hypothèse que la machine produit des pièces défectueuses jusqu'à ce qu'elle subisse des réparations. Danse le modèle de Kim et Gershwin (2005), le flux de production de pièces est considéré continu tandis que la qualité est considérée comme discrète. Cela conduit à certaines difficultés dans l'analyse. Nous développons nos modèles en partie de façon à contourner cette difficulté. Donc, notre modèle est développé dans un cadre de modélisation entièrement fluide pour l'analyse de la gestion intégrée de la qualité/ production et l'optimisation des lignes de fabrication non fiable. De plus, une partie de ces travaux comprend des objectifs de la maintenance corrective, objectifs qui ne font pas partie de nos travaux.

CHAPITRE 2

Méthodologie de recherche et organisation générale de la thèse

Cette thèse est présentée suivant une approche par articles. Ainsi, les trois prochains chapitres présentent les articles publiés ou soumis à des revues avec comité de lecture dans le cadre de cette recherche.

Le premier article intitulé : « Approximate performance analysis of CONWIP disciplines in unreliable non homogeneous transfer lines » a été publié au journal *Annals of Operations Research* en 2011 avec Roland Malhamé comme coauteur. Dans cet article, nous avons développé un modèle mathématique approximatif pour l'évaluation des indicateurs de performance de lignes de transfert contrôlées par CONWIP. Le modèle d'une boucle de CONWIP à $(n-1)$ machines correspond à une chaine de $(n-1)$ blocs élémentaires, dédié à chaque stock dans la ligne, plus un bloc élémentaire dédié à la dynamique du stock total dans la boucle. Dans cet article, nous avons choisi de dimensionner le stock total en supposant que les pièces produites sont toutes conformes aux normes de qualité. Toutefois dans la réalité, la dimension qualité de la production est primordiale puisque l'existence des pièces défectueuses diminue l'utilité des stocks intermédiaires et donc leur capacité à augmenter la productivité de la ligne. Par conséquent, ceci nous pousse à considérer conjointement les problèmes d'organisation du contrôle de la qualité et la gestion de production.

Le problème de modélisation et d'analyse simultanée de la production et de la qualité dans les lignes de production non fiables est d'une grande complexité. La majeure partie de cette complexité est attribuée à la qualité, puisque l'état de qualité d'un système est bien souvent inconnu. Même si les origines de la non-qualité sont connues (pannes de machine, matière première non conforme), les états associés à la non-qualité demeurent partiellement observables. Dans ce contexte, nous considérerons le cas d'une machine qui oscille spontanément (chaîne de Markov) entre un état opérationnel parfait, un état défectueux non observable et un état de panne totale. Nous avons étudié ce modèle dans notre deuxième article intitulé : « A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects » et qui est soumis pour publication au journal *Discrete Event Dynamic Systems* avec Roland Malhamé et Robert Pellerin comme coauteurs.

Sur la base de nos conclusions dans ce deuxième article et dans un contexte de minimisation de coût, nous (Roland Malhamé , Robert Pellerin et moi même) avons analysé des modèles fluides et continus de la production (Kanban) composés par des machines non fiables produisant une fraction de pièces non conformes. En plus de traiter conjointement la qualité et la production, nous avons introduit la notion des stations d'inspection dont le rôle est de rejeter les pièces défectueuses en ne gardant dans le système que les pièces conformes. Cette dernière notion est la base de notre troisième article intitulé : « Unreliable production lines with defective parts and inspection stations » soumis pour publication au journal *IIE Transactions* . L'analyse de performance sera fondée sur la résolution analytique ou numérique des équations de Kolmogorov associées, rendue possible par nos méthodes de décomposition et l'optimisation se fera par des techniques de programmation dynamique.

Ces trois articles font l'objet des trois prochains chapitres (3, 4, et 5).

La démarche scientifique de la thèse se base sur la même méthodologie scientifique suivante :

1. Identification du modèle : Consiste à définir les différentes hypothèses sur lesquelles va reposer notre modèle, ainsi que les paramètres et les indices de performance du modèle.
2. Construction du modèle de simulation Monte-Carlo : Pour mieux assimiler le comportement de notre modèle, et à défaut d'avoir un système réel devant les yeux, on développe des modèles de simulation Monte-Carlo en utilisant le logiciel AWESIM ou le logiciel ARENA. Ce type de simulation est devenu assez utilisé pour mieux analyser et comprendre les différents procédés industriels. De plus, ce modèle va nous fournir les résultats numériques qui auront pour but une comparaison avec les résultats théoriques à venir.
3. Analyser le modèle : À l'aide des résultats de la simulation de Monte-Carlo, nous déterminons les zones intéressantes pour l'analyse, les particularités du modèle, le comportement du modèle dans les frontières de ces zones, etc. En gros, tout ce qui peut être utile pour notre étude théorique.
4. Développer les équations de Kolmogorov : Les équations de Kolmogorov sont des équations aux dérivées partielles qui nous permettent de déterminer les différentes fonctions

de distribution de probabilité qui régissent notre processus stochastique. En plus des équations de Kolmogorov, nous devons déterminer les différentes conditions limites pour pouvoir résoudre notre système d'équations. En effet, à partir de ces distributions de probabilité, il est plus facile de déterminer l'expression théorique de nos indices de performance.

5. Résoudre les équations de Kolmogorov : Quand il s'agit des équations aux dérivées ordinaires (EDO) (article 1), trouver une solution analytique est assez faisable ; contrairement au cas où ces équations sont des équations aux dérivées partielles à deux variables (article 2). La résolution analytique de ces dernières s'avère très complexe, sinon impossible. Pour ces raisons, nous avons développé des algorithmes numériques pour leur résolution.

6. Validation des résultats théoriques : En se basant sur les résultats obtenus par simulation Monte-Carlo, on effectue une comparaison avec les valeurs des résultats théoriques. Dans le cas où l'écart entre les valeurs théoriques et celles de la simulation Monte-Carlo est acceptable, nous validons le modèle et nos hypothèses. Dans le cas contraire, nous rejetons le modèle et nous recommençons les étapes 3, 4, 5 et 6.

7. Exploitation du modèle : Relever tout ce qui parait intéressant comme résultat et comme conclusion.

CHAPITRE 3

Approximate performance analysis of CONWIP disciplines in unreliable non homogeneous transfer lines

Article publié dans *Annals of Operations Research*, pp. 213-233, 2011, et écrit par :

Fatima Zahra Mhada

École Polytechnique de Montréal and GERAD

Roland P. Malhamé

École Polytechnique de Montréal and GERAD

3.1 Introduction

There is a rich literature on approximate decomposition methods for the analysis and optimization of unreliable manufacturing transfer lines (Chiang *et al.* (2000), Dallery et Gershwin (1992), Dallery et Bihan (1999), Gershwin (1987), Levantesi *et al.* (2003), Sadr et Malhamé (2004a)). These papers have analyzed Kanban related production disciplines. The latter dictates that a machine must produce at maximum rate as long as it has not filled the associated downstream buffer up to a *virtual* (Kanban parameter) limit; furthermore, the Kanban level of buffer storage must be maintained whenever possible (i.e. if the machine is operational and not starved). Frein *et al.* (1996) extend the Kanban related decomposition analysis to CONWIP or constant work-in-process systems which in effect implement a Kanban strategy on a group of machines. However, while the CONWIP policy imposes an upper limit on the total amount of wip within the CONWIP controlled loop, in Frein *et al.* (1996), this behavior is somewhat unsatisfactorily approximated by imposing that the total *average* internal wip will always be equal to this upper limit.

Gershwin et Werner (2007)[1] remove the shortcomings of Frein *et al.* (1996) by working with the full capacity of the CONWIP loop. It considers the correlations among population, buffer capacity, blocking probability and starvation probability to limit the radii of influence of individual machines within the loop. However, the actual subject of study is closed loop systems which have a fixed internal population. While this applies also to CONWIP systems, provided one considers the production authorization cards as part of the wip within the system, the circular symmetry of closed loops is disrupted in that the first machine upstream in the loop now plays a special role. This symmetry disruption is not critical when considering problems of calculating maximum throughput as in (Gershwin et Werner (2007)), but can become crucial when the demand rate is externally imposed as in our analysis, as well as a service rate identified with the coefficient of availability of the buffer feeding the part of the line past the CONWIP loop. Thus our problem setting is different, and our analysis accordingly different.

The fact that individual machines are much more dynamically coupled in a CONWIP loop

[1] Cet article a été porté à notre attention après la soumission de notre article

than they are in a pure Kanban type of control architecture, precludes the application of previously well tested Kanban approximation approaches (Chiang *et al.* (2000), Dallery et Bihan (1999), Sadr et Malhamé (2004a), Sadr et Malhamé (2004b)), even modulo some changes, for our approximate CONWIP loop modeling purpose. It is worthwhile noting that (Sadr et Malhamé (2004a), Sadr et Malhamé (2004b)) differ significantly from (Chiang *et al.* (2000), Dallery et Bihan (1999), Levantesi *et al.* (2003)) in that, in the latter, issues of maximum transfer line maximum throughput evaluation (or improvability) are tackled at the outset, while in (Sadr et Malhamé (2004a), Sadr et Malhamé (2004b)), as in the current framework, transfer line performance is considered for the simpler case of *a fixed given required rate of parts production.* Such circumstances allow application in our case of the so-called *demand averaging principle* (discussed in Section 3.3 below), first introduced in (Sadr et Malhamé (2004a)).

In this paper we propose a new mathematical model for the approximate evaluation of the performance of CONWIP controlled loops in unreliable transfer lines under a constant rate of demand for parts d. For $(n-1)$ machines and $(n-1)$ buffers CONWIP controlled loop, the model consists of $(n-1)$ n-state macro machine-single buffer mathematical building blocks, each tagged to a particular buffer: the macro machine is shared by all building blocks, and its state is an approximate representation (single machine failures only allowed at any one time) of the joint operating state of the $(n-1)$ machines in the CONWIP loop (the loop mode); the i^{th} individual building block is in effect a "local" view of the dynamics of the CONWIP loop as construed by *an observer within internal buffer* i, who can measure the i^{th} buffer level and record the loop mode (the "local" hybrid-state), but *does not measure the levels of adjacent buffers.* That observer develops a Markovian representation of its own local state dynamics by making up for insufficient information about the state of its neighbors through the use of an estimate of the average rate of change or *velocity* of its buffer wip *conditional* on the current local state. The computation of these average velocities at building blocks collectively involve a total of $(n-1)^2$ unknown probabilities of wip availabilities in each loop buffer conditional on every one of the $(n-1)$ possible loop failure modes. In addition, an aggregate block of similar mathematical structure is associated to the total wip dynamics within the loop. The blocks correspond to n linear components state differential equations

with boundary conditions. They are individually solvable systems if the associated buffer average velocities for every loop mode are known. The unknown expectations in the buffer blocks dynamics are initialized and calculated by means of successive iterations involving all blocks except the aggregate block. The performance estimates of interest - mean total wip, and probability of parts availability at the last buffer in the loop - are then obtained from the aggregate wip block and validated against the results of Monte-Carlo simulations.
The paper is organized as follows. In Section 3.2, we specify our CONWIP controlled model of an n machine unreliable transfer line, with initial assumptions and mathematical simplifications. Note that throughout the paper, our modeling assumptions and approximations are made very explicit both for reasons of clarity and so as to give the interested reader more tools to build test cases that could better challenge our proposed modeling approximations. Section 3.3 is dedicated to background material on a particular class of hybrid-state (continuous-discrete) Markov processes with scalar continuous component and modewise constant rates of change of the continuous component; also, a very useful approximation and calculation technique in transfer lines subjected to a constant rate of demand for parts, the *demand averaging principle* (Sadr et Malhamé (2004a)) is reviewed. In Section 3.4, we make use of that background material to develop the mathematical form of our n building blocks. Based on observations, themselves inspired in part by Monte-Carlo simulations, our approximate expressions of the individual buffer velocities assumed constant for each particular loop mode are given in terms of the $(n-1)^2$ unknown probabilities of availability of wip in each of the $(n-1)$ buffers when conditioned on each of the $(n-1)$ permissible loop failure modes. In Section 3.5, we describe our recursive algorithm for computing the system solution. In Section 3.6, we compare our theoretical results against Monte-Carlo simulations for a number of transfer lines. Section 3.7 concludes the paper.

3.2 Mathematical model of the transfer line and statement of objectives.

We consider a manufacturing transfer line consisting of n machines, M_i $i = 1, \cdots, n$, each associated with a buffer i and a wip variable x_i. The transfer line (see Figure 3.1) produces a single type of parts and buffer n is subjected to a constant rate of extraction of parts d. Backlog is allowed only at buffer n (in which case $x_n < 0$). Machines M_i, $i = 1, ..., n$ are in a

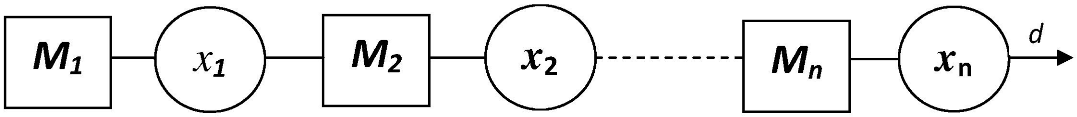


Figure 3.1 Transfer line with unreliable machines, storage levels x_i, and constant rate of demand d for parts. All storage levels are bounded above and non negative, except for x_n which can become negative (demand backlog).

binary state α_i called the i^{th} machine mode. α_i evolves according to a two-state continuous time Markov chain: $\alpha_i = 1$ when the machine is fully operational; $\alpha_i = 0$ when the machine fails. The failure rate of M_i is p_i while its repair rate is r_i. $u_i(t)$ designates the instantaneous production rate of machine M_i and it is bounded above by the maximum production rate k_i. We assume the following wip/finished parts dynamics:

$$\begin{aligned} \frac{dx_i}{dt} &= u_i(t) - u_{i+1}(t), \quad i = 1, ..., (n-1); \\ \frac{dx_n}{dt} &= u_n(t) - d. \end{aligned} \tag{3.1}$$

In (3.1), a fluid model of parts production is considered i.e. $u_i(t)$ can vary continuously from 0 to k_i.

It is assumed that the first $(n-1)$ machines and associated buffers are in a loop subjected to a CONWIP discipline, whereby machines M_i , $i = 2, ..., (n-1)$ produce at maximum rate k_i whenever possible (i.e. if operational and not starved), while machine M_1 also does so, unless the total wip contained in the loop has reached a maximum value designated by z; at that point M_1 is allowed to process parts at a rate that cannot exceed the rate at which parts are extracted from buffer x_{n-1}. In effect, given the CONWIP loop structure, the only design parameter is the maximum permissible total loop wip value z. The role of machine M_n which is outside the CONWIP loop is to secure a parts production rate of d via a pull mechanism. In the rest of the paper, it will be considered as the “customer” of the CONWIP loop.

The objective is to develop an approximate modeling approach such that, given a constant rate of extraction of parts d from the transfer line, one could efficiently estimate for a given

z, the two main performance indices of the loop: mean long term total wip level and the (long term) probability that wip be available at buffer x_{n-1}. In the current context, the latter probability expressed in percentage will be referred to as the *CONWIP loop service level.* In practice, a design objective is to choose z so as to minimize mean loop wip for a given desired level of service at buffer x_{n-1}.

We shall make the following preliminary assumptions:

- *Monotone decreasing maximum production rates:* $k_1 > k_2 > k_3 > ... > k_{n-1}$.
- *Infinite supply of raw material:* M_1 is never starved.

  Furthermore we shall designate by *loop mode* the discrete $(n-1)$ dimensional vector with binary components: $\overrightarrow{\alpha}(t) = [\alpha_1(t), \alpha_2(t), ..., \alpha_{n-1}(t)]^T$.

  The above vector has in theory 2^{n-1} possible values. As n grows, this number can grow very quickly. In order to limit the complexity of computations, we shall approximate the loop mode $\overrightarrow{\alpha}(t)$ by a scalar mode $\tilde{\alpha}(t)$ corresponding to a CONWIP loop macromachine whereby only one failure at a time is allowed. Thus, we make the following important additional assumption:

- *Single machine failure assumption:* Machines can be in a failure mode only one at a time.

  The above assumption is acceptable if for any machine M_i, the mean repair time is much smaller than the joint mean first failure time of the remaining operational machines. More specifically:

  $$\frac{1}{r_i} \ll \frac{1}{\sum\limits_{j \neq i} p_j}, \quad \forall i = \ 1, ..., (n-1). \tag{3.2}$$

  Note that in order to limit the impact of the above approximation (which tends by itself to underestimate the downtime of the transfer line), we show in Appendix A how one can choose to increase the individual machines failure rates from p_i to $\tilde{p}_i$ (r_i is kept unchanged) so that the probability that all machines in the CONWIP loop are operational at the same time and remains unchanged at $\prod\limits_{i=1}^{n-1} \frac{r_i}{r_i + p_i}$.

  The approximate loop mode Markov chain $\tilde{\alpha}$ will have n distinct states corresponding to either a fully operational loop ($\tilde{\alpha} = 1$), or a loop with a single failed machine M_i,

$i = 1, ..., (n-1)$, $(\tilde{\alpha} = 0i)$.

The associated intensity matrix of $\tilde{\alpha}$ is defined as follows (see Appendix A):

$$\tilde{Q} \; = \; \begin{bmatrix} \tilde{q}_{11} & \gamma p_1 & \gamma p_2 & ... & \gamma p_{n-2} & \gamma p_{n-1} \\ r_1 & -r_1 & 0 & ... & 0 & 0 \\ r_2 & 0 & -r_2 & ... & 0 & 0 \\ . & . & . & ... & . & . \\ . & . & . & ... & . & . \\ . & . & . & ... & . & . \\ r_{n-1} & 0 & 0 & ... & 0 & -r_{n-1} \end{bmatrix} \tag{3.3}$$

$$\text{with} \qquad \tilde{q}_{11} \; = \; -\gamma \sum_{i=1}^{n-1} p_i \quad \text{and} \quad \gamma = \frac{1 - \prod_{i=1}^{n-1} \frac{r_i}{(r_i + p_i)}}{(\prod_{i=1}^{n-1} \frac{r_i}{r_i + p_i})(\sum_{i=1}^{n-1} \frac{p_i}{r_i})}.$$

3.3 Background material

In the following, we review some background material essential in developing the building blocks of our approximate model.

3.3.1 Forward Kolmogorov equations for modewise constant velocity hybrid-state Markov processes.

Designate by $(x(t) \; \alpha(t))^T$ a generic hybrid state Markov process with scalar continuous state $x(t)$ and discrete mode $\alpha(t) \in I_M = \{1, 2, ..., M\}$. We assume that for $x(t) \in R_i \equiv]a_i; a_{i+1}[$, the dynamics of $x(t)$ is described by the *modewise constant* velocity v_α^i so that:

$$\dot{x}(t) = \sum_{j=1}^{M} v_\alpha^i \, I[\alpha(t) = j]. \tag{3.4}$$

In (3.4), $I[.]$ is the indicator function. Furthermore, $\alpha(t)$ is assumed to evolve according to a continuous time homogeneous Markov chain defined by :

$$Pr[\alpha(t + \Delta t) = k | \alpha(t) = j] = q_{j_k}^i \Delta t + o(\Delta t), \quad j = 1, ..., M, \quad k = 1, ..., M, \tag{3.5}$$

with

$$q^i_{jj} = -\sum_{k \neq j} q^i_{jk}. \tag{3.6}$$

In addition, we assume that $x = a_i$ and $x = b_i$ are either type 1 switching boundaries past which velocity vectors change abruptly (see Figure 3.3), or type 2 boundaries (see Figure 3.4) associated with a non zero sojourn time of the continuous state $x(t)$ and consequently with the presence of probability masses.

Kolmogorov equations and boundary conditions take the form below .

a. *Inside region R_i (Figure 3.2 ; see (Malhamé et Boukas (1991)))*:

$$\frac{\partial \overrightarrow{f}(x,t)}{\partial t} = -V_i \frac{\partial \overrightarrow{f}(x,t)}{\partial x} + Q_i^T \overrightarrow{f}(x,t) \tag{3.7}$$

where:

$$\begin{aligned} \overrightarrow{f}(x,t) &= [f_1(x,t), f_2(x,t), ..., f_M(x,t)]^T. \\ V_i &= diag[v^i_\alpha], \quad \alpha = 1, ..., M. \\ Q_i &= [q^i_{jk}], \quad j = 1, ..., M, \quad k = 1, ..., M. \end{aligned}$$

Furthermore, $f_j(x,t)$ is the hybrid probability density function associated with $x(t)$, in machine mode j, i.e.:

$$f_j(\lambda,t)d\lambda = Pr[(\lambda < x(t) \leq \lambda + d\lambda) \bigcap (\alpha(t) = j)], \quad j = 1, ..., M.$$

b. *Boundary conditions for type 1 boundaries (Figure 3.3)*

Type 1 boundaries are defined as points in the scalar continuous state space where an abrupt change occurs in the velocity vector associated with a fixed mode j, and where continuous variable $x(t)$ has zero sojourn time. Conservation of probability dictates that probability currents remain continuous across the boundary, thus yielding (Figure 3.3):

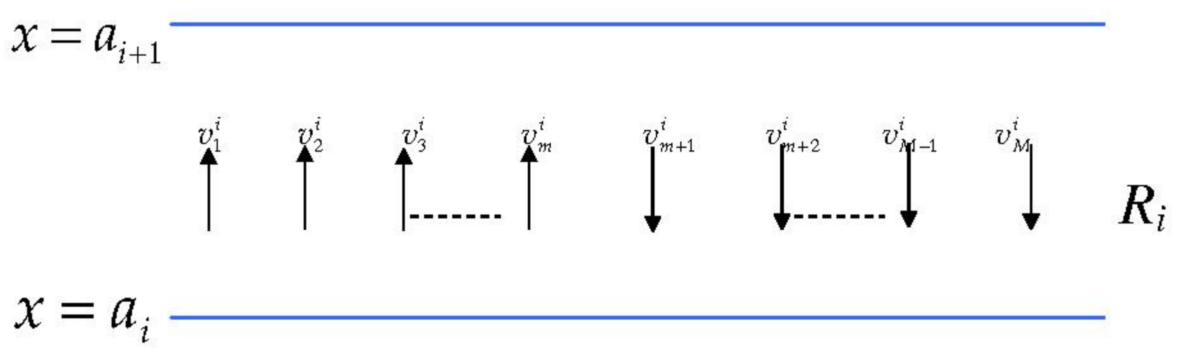


Figure 3.2 Direction of modewise constant velocity vector associated with scalar continuous state variable $x(t)$ in region R_i in modes $j = 1, ..., M$. v_j^i, $j = 1, ..., M$ are signed velocities.

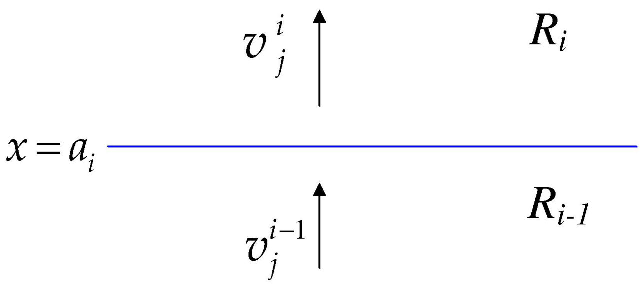


Figure 3.3 Type 1 boundary: abrupt velocity change and zero sojourn time of $x(t)$ at $x = a_i$; probability conservation imposes that probability currents remain continuous across the boundary.

$$v_j^{i-1} f_j(a_i^-, t) = v_j^i f_j(a_i^+, t). \tag{3.8}$$

c. *Boundary conditions for type 2 boundaries (see Figure 3.4)*

Type 2 boundaries are defined as points in the state space where $x(t)$ can spend a non zero amount of time (zero velocity until change to a downwards oriented velocity mode). Thus, there are probability masses $P_{a_i}^j(t)$, $j = 1, ..., m$, associated with events $((x(t) = a_i) \bigcap (\alpha(t) = j))$. For illustrative purposes in Figure 3.4, it is assumed that for $\alpha(t) = 1, ..., m$, the velocity vectors in region R_{i-1} point upwards (and are zero on a_i itself), while they point downwards (away from the boundary) for $\alpha(t) = (m + 1), ..., M$, for some integer m.

In this case, conservation of probability at the boundary dictates that: (i) Probability mass $P_{a_i}^j(t)$ increases upon arrival of probability current $v_j^{i-1} f_j(a_i^-, t)$ or a probability current from

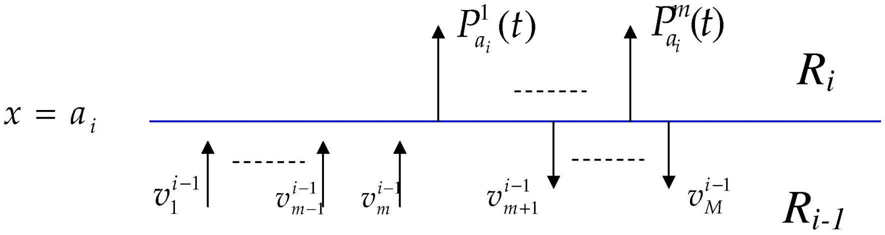


Figure 3.4 Type 2 boundary: non zero sojourn time of $x(t)$ at a_i; a differential equation is associated with probability mass $P^j_{a_i}(t)$ sitting at $x(t) = a_i$ and mode j; probability conservation imposes that incoming probability current in mode j contribute to increasing probability mass $P^j_{a_i}(t)$, while probability losses from $P^j_{a_i}(t)$ contribute either to an increase in other probability masses or downwards oriented probability currents at a_i.

a neighboring probability mass, and decreases as j moves to any other state k, i.e. at rate $q^i_{jj}P^j_{a_i}(t)$, (ii) A leak in probability mass $P^j_{a_i}(t)$ towards a state $k \neq j$ corresponds to a loss rate $P^j_{a_i}(t)q^i_{jk}$. This probability loss in $P^j_{a_i}(t)$ constitutes a gain for $P^k_{a_i}(t)$, if $k = 1, ..., m$, or translates into a downward probability current in mode k if $k = (m+1), ..., M$.

More specifically, one can write:

- *Probability mass dynamics*

$$\frac{dP^j_{a_i}(t)}{dt} = v_j^{i-1}\, f_j(a_i^-, t) + q^i_{jj}\, P^j_{a_i}(t) + \sum_{k=1, k\neq j}^{m} q^i_{kj}\, P^k_{a_i}(t), \quad j = 1, ..., m; \tag{3.9}$$

- *Boundary currents specification*

$$v_j^{i-1}\, f_j(a_i^-, t) + \sum_{k=1}^{m} q^i_{kj}\, P^k_{a_i}(t) = 0, \quad j = m+1, ..., M. \tag{3.10}$$

In matrix form, (3.9) and (3.10) can be written:

$$\frac{d}{dt}\overrightarrow{P}_{a_i}(t) = [V_{i-1}]\,\overrightarrow{f}(a_i^-, t) + [Q_i]^T\,\overrightarrow{P}_{a_i}(t)$$

with

$$\overrightarrow{P}_{a_i}(t) = [P^1_{a_i}(t), P^2_{a_i}(t), ..., P^m_{a_i}(t), 0, 0, ..., 0]^T. \tag{3.11}$$

Note that in a general case, vector $\overrightarrow{P}_{a_i}(t)$ will have non zero entries only whenever the velocity vector associated with the mode points towards the boundary.

3.3.2 A transfer line decomposition related approximation: the demand averaging principle.

The demand averaging principle is an approximation which has proven very useful in our previous work on decomposition/aggregation methods for the approximate analysis of transfer lines (Sadr et Malhamé (2004a), Sadr et Malhamé (2004b)). It stems from the simple observation that if a single part transfer line is constrained (through a pull mechanism) to meet a constant rate of demand for finished parts, say d, provided it succeeds in doing so, and that wip and finished parts available storage is bounded, then *every machine in the transfer line must produce parts at the common long term average rate of* d. Thus d becomes an invariant of the production rate of all machines in the transfer line, as well as of the *extraction* rate of wip from any *intermediate buffer.* In turn, this property is used to achieve approximate decoupling of a machine, from the stream of machines downstream of it in the transfer line. We illustrate the approximation and the calculations for a two-state Markovian machine with unlimited supply of raw parts, feeding into a wip storage space of size z_1. It is assumed that the machine always produces at maximum rate, unless it is down or blocked.
Application of the demand averaging principle (DAP) will be illustrated for machine M_1 and its associated storage variable x_1. Adopting a fluid model of parts production, storage $x_1(t)$ evolves according to:

$$\frac{dx_1}{dt} = u_1(t) - u_2(t), \tag{3.12}$$

where $u_1(t)$, $u_2(t)$ are respectively the production rates of machines M_1 and M_2 at time t, and are respectively bounded by k_1 and k_2 ($k_1 \geq k_2$).
In (3.12), the impact of the machines downstream of M_1 is mediated by $u_2(t)$. Now $u_2(t)$ is constrained by the fact that *its long term average must be equal to d.* If we add to that the observation that when machine M_2 is starved (parts supply $x_1(t)$ unavailable), $u_2(t)$ must necessarily be zero, then one can conclude that the average of $u_2(t)$ over periods of availability of supply $x_1(t)$ must be sufficiently higher than d to make up for starvation periods. More

precisely:

$$\begin{aligned}\lim_{t\to\infty} E[u_2(t)] &= d \\ &= \lim_{t\to\infty} E[u_2(t)|x_1 \text{ available}]\, Pr[\ x_1 \text{ available}] \\ &+ \lim_{t\to\infty} E[u_2(t)|x_1 \text{ not available}]\,(1 - Pr[\ x_1 \text{ available}]). \qquad (3.13)\end{aligned}$$

Equation (3.13) leads to

$$E[u_2(t)|\ x_1 \text{ available}] = \frac{d}{a_1}, \qquad (3.14)$$

where

$$a_1 = \lim_{t\to\infty} Pr[x_1(t) \text{ available}].$$

a_1 is called the coefficient of availability of stock $x_1(t)$.

What the DAP approximation states is the following: *"Of all the $u_2(t)$ processes consistent with constraint (3.14), a constant production rate over the periods of availability of stocks $x_1(t)$ yields sufficiently accurate results for calculations concerning the first order statistics of $x_1(t)$".* Note that it is a powerful decoupling mechanism given that calculations for $x_1(t)$ become *independent* of the part of the transfer line downstream of $x_1(t)$ as long as the transfer line as a whole is able to meet the demand rate d.

Iterative calculation of coefficient of availability a_1

The two-mode machine case

Machine M_1 is a two-mode machine feeding wip buffer 1. Machines downstream of buffer 1 pull wip from that buffer at *long term* average rate $\frac{d}{a_1}$ where a_1 is the (unknown) coefficient of availability of wip x_1.

The following is a description of the algorithm used to calculate a_1. Based on (Hu (1995)) where a closed form expression of the probability distribution of wip for a two-mode Markovian machine under an inventory hedging policy with critical level z_1, constant rate of demand

$\tilde{d}$ for finished parts, and no backlog allowed, one can write:

$$a_1 = 1 - \frac{p_1}{r_1 + p_1} \frac{1 - \frac{r_1(k_1 - \tilde{d})}{p_1 \tilde{d}}}{\frac{1 - r_1(k_1 - \tilde{d})}{p_1 \tilde{d}} e^{-(\frac{(p_1 + r_1)\tilde{d} - k_1 r_1}{(k_1 - \tilde{d})\tilde{d}}) z_1}}, \tag{3.15}$$

where p_1 and r_1 are respectively the rate of failure and rate of repair of machine M_1 and k_1 is the maximum production rate of that machine. Note that in (3.12), by virtue of DAP, one could replace $u_2(t)$ by a constant rate of extraction over periods of availability of wip $x_1(t)$. This would correspond to substituting $\frac{d}{a_1}$ for the constant value $\tilde{d}$ in (3.15). Thus (3.15) becomes an *implicit* equation for the calculation of a_1.

a_1 is obtained through the iterative calculation of a fixed point. More specifically substitute $a_1^{(0)} = 1$ in the right-hand side of (3.15) to obtain $a_1^{(1)}$. This leads to $a_1 < a_1^{(1)}$ (because the true $\frac{d}{a_1} > d$ and a_1 is a monotone decreasing function of demand $\tilde{d}$). If one substitutes $a_1^{(1)}$ in the right-hand side of (3.15), this leads to $a_1 < a_1^{(2)} < a_1^{(1)}$ (because $a_1 < a_1^{(1)} < a_1^{(0)} = 1$ leads to $\frac{d}{a_1} > \frac{d}{a_1^{(1)}} > \frac{d}{a_1^{(0)}}$, and thus yields $a_1 < a_1^{(2)} < a_1^{(1)}$). It is then seen that $a_1^{(i)}$ is monotone decreasing (bounded below if a solution exists, i.e. if the machine is able to sustain demand rate d). Thus, if a solution exists, the iterations will converge to that (unique) solution.

The multi-mode machine case

Note that if machine M_1 has more than two modes, a closed form expression such as (3.15) is no longer available. However, for any demand rate $\tilde{d}$, one can develop the associated (steady-state) forward Kolmogorov equations and their boundary conditions as in the previous section, to which one adds the constraint that total probability must integrate up to 1. The coefficient of availability $a_1(\tilde{d}) = 1 - Pr[x_1(t) = 0]$.

Thus, instead of (3.15), one works for each particular value of $\tilde{d}$ with a numerical system of linear differential and algebraic equations. Iterations proceed as previously, with $a_1^{(0)} = 1$. They converge to the true value (if a solution exists) for the same reasons as before.

3.4 Approximate CONWIP loop modeling

3.4.1 General principles of the decomposition methodology

An exact dynamic model for the computation of $\lim_{t\to\infty} E[\sum_{i=1}^{n-1} x_i(t)]$, the quantity of interest, together with $\lim_{t\to\infty} Pr[x_{n-1} \text{ active}] = a_{n-1}$, would require writing the steady-state forward Kolmogorov equations for the hybrid Markov process $[x_1(t), x_2(t), ..., x_n(t), \alpha_1(t), \alpha_2(t), ..., \alpha_n(t)]^T$, together with all relevant boundary and normalization conditions. They correspond to a system of coupled linear partial and ordinary differential equations of very high order as n grows: n dimensional vector partial differential equations with 2^n vector components with boundary conditions. Computationally, solving such a system would represent a nearly hopeless task. Instead, we choose to work with a dynamic model which is a concatenation of approximate forward Kolmogorov (ordinary) differential equations associated with a sequence of "buffer centric" hybrid-state Markov processes $(x_i(t), \tilde{\alpha})^T$, $i = 1, ..., (n-1)$ and an aggregate buffer $(x_0(t), \tilde{\alpha})^T$ with $x_0(t)$ defined as the total wip stored in the CONWIP loop at time t. Note that $\tilde{\alpha}$ is the approximate loop macromachine discussed in Section 3.2 under the single machine failure assumption.

The word "approximate" is also used in that strictly speaking $(x_i(t), \tilde{\alpha})^T$ cannot be associated with a Kolmogorov equation because, it is *not* Markov all by itself. Furthermore, and as an extension to this remark, we confine our approximate description of $(x_i(t), \tilde{\alpha})^T$ to the class of *modewise constant velocity* hybrid-state Markov processes of Section 3.3. In particular when the real $x_i(t)$ velocity changes over time for a given $\tilde{\alpha}(t)$, because of interactions with $x_j(t)$, $j \neq i$, we replace these instantaneous coupling terms by their *long term averages conditional* on the loop mode $\tilde{\alpha}$ which is *common* to all the modeled hybrid-state *pseudo-Markov* processes.

Thus summarizing, we associate with every hybrid-state pair $(x_i(t), \tilde{\alpha})^T$, herein designated by $\tilde{x}_i(t)$, $i = 0, 1, ..., (n-1)$, a continuous state region dependent of the approximate Markovian dynamic evolution, with constant velocity in any given mode-region pair. These Markovian dynamic building blocks are interrelated through a set of common unknown constants (coefficients of availability of all internal buffer wips conditional on the various loop modes) which can be solved for only when the complete system of equations is considered. Finally, note that machine M_n (see Figure 3.1) is left out of the modeling process. This is because: (i) it is not part of the CONWIP loop; (ii) under the demand averaging principle, for a constant rate of demand for finished parts, the long term first order statistics of x_{n-1} become independent of M_n (recall Section 3.3.2).

3.4.2 Details and justification of model component dynamics

The following hybrid probability density, probability mass, and velocity vectors will be instrumental in characterizing the probabilistic evolution of model building blocks $\tilde{x}_i(t)$, $i = 0, 1, ..., n$:

$$
\begin{aligned}
\overrightarrow{f}^{i}(\lambda,t) &= [f_1^i(\lambda,t), f_{01}^i(\lambda,t), ..., f_{0(n-1)}^i(\lambda,t)]^T, \quad i = 0,1,...,(n-1). \\
\overrightarrow{P}^{i}(t) &= [P_1^i(t), P_{01}^i(t), ..., P_{0i}^i(t), 0, P_{0(i+2)}^i(t), ..., P_{0(n-1)}^i(t)]^T, \quad i = 1,...,(n-2) \\
\overrightarrow{P}^{(n-1)}(t) &= [0, P_{01}^{(n-1)}(t), P_{02}^{(n-1)}(t), ..., P_{0(n-1)}^{(n-1)}(t)]^T, \\
\overrightarrow{P}^{0}(t) &= [0, P_{01}^0(t), 0, ..., 0]^T, \\
\overrightarrow{P}^{iz}(t) &= [0, ..., P_{0(i+1)}^{iz}(t), 0, ..., 0]^T, \quad i = 1,...,(n-2), \qquad (3.16) \\
\overrightarrow{P}^{(n-1)z}(t) &= [P_1^{(n-1)z}(t), 0, ..., 0]^T, \\
\overrightarrow{P}^{0z}(t) &= [P_1^{0z}(t), 0, P_{02}^{0z}(t), ..., P_{0(n-2)}^{0z}(t), P_{0(n-1)}^{0z}(t)]^T, \\
\overrightarrow{v}^{i} &= [v_1^i, v_{01}^i, v_{02}^i, ..., v_{0(n-2)}^i, v_{0(n-1)}^i]^T,
\end{aligned}
$$

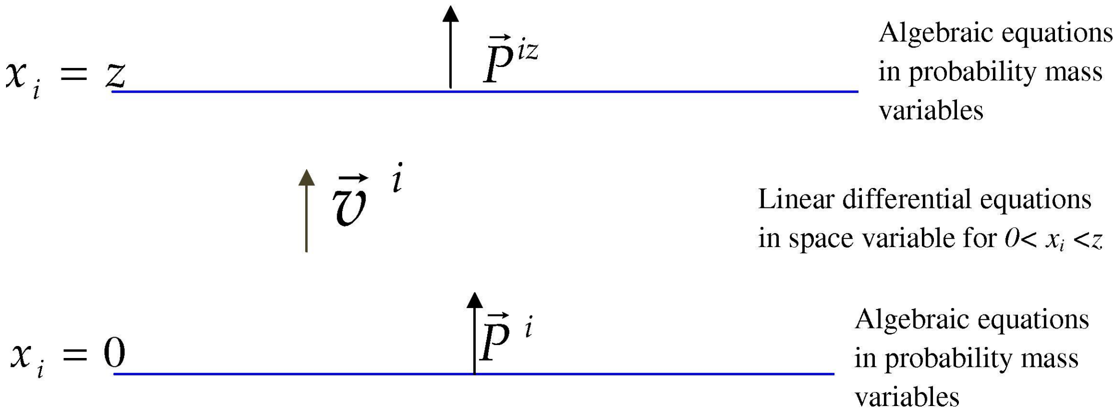
where, in the above, and for $i = 0, 1, ..., (n-1)$, $j = 1, ..., (n-1)$:

$$
\begin{aligned}
f_1^i(\lambda,t)d\lambda &= Pr[(\lambda < x_i(t) \leq \lambda + d\lambda) \bigcap (\tilde{\alpha}(t) = 1)], \\
P_1^i(t) &= Pr[(x_i(t) = 0) \bigcap (\tilde{\alpha}(t) = 1)], \\
P_1^{iz}(t) &= Pr[(x_i(t) = z) \bigcap (\tilde{\alpha}(t) = 1)], \\
f_{0j}^i(\lambda,t)d\lambda &= Pr[(\lambda < x_i(t) \leq \lambda + d\lambda) \bigcap (\tilde{\alpha}(t) = 0j)], \\
P_{0j}^i(t) &= Pr[(x_i(t) = 0) \bigcap (\tilde{\alpha}(t) = 0j)], \\
P_{0j}^{iz}(t) &= Pr[(x_i(t) = z) \bigcap (\tilde{\alpha}(t) = 0j)].
\end{aligned}
$$

v_1^i = constant rate of change of $x_i(t)$ in mode $\tilde{\alpha}(t) = 1$.

v_{0j}^i = constant rate of change of $x_i(t)$ in mode $\tilde{\alpha}(t) = 0j$.

Structure of the building blocks



$x_i = z$
$\vec{P}^{iz}$
Algebraic equations in probability mass variables
$\vec{v}^{\,i}$
Linear differential equations in space variable for $0< x_i <z$
$x_i = 0$
$\vec{P}^{\,i}$
Algebraic equations in probability mass variables

Figure 3.5 Graphical representation of the steady- state forward Kolmogorov equations associated with CONWIP model building block $\tilde{x}_i(t)$ $i = 0, 1, ..., (n-1)$.

The steady-state forward Kolmogorov equations associated with hybrid state pseudo-Markov process $\tilde{x}_i(t)$ have a generic matrix form for $i = 0, 1, ..., (n-1)$. As illustrated in Figure 3.5 above, one must distinguish an open region $(0 < x_i < z_i)$ and its upper and lower boundaries, respectively. Based on the developments in Section 3.3, the equations are as follows:

For $(0 < x_i < z)$:

$$\frac{d\overrightarrow{f}^{\,i}(x_i)}{dx_i} = [V^i]^{-1}\tilde{Q}^T\overrightarrow{f}^{\,i}(x_i). \tag{3.17}$$

At $x_i = z$:

$$\overrightarrow{f}^{\,i}(z^-) = -[V^i]^{-1}\tilde{Q}^T\overrightarrow{P}^{iz}. \tag{3.18}$$

At $x_i = 0$:

$$\overrightarrow{f}^{\,i}(0^+) = -[V^i]^{-1}\tilde{Q}^T\overrightarrow{P}^{i}. \tag{3.19}$$

where in (3.17)-(3.19):

$$\tilde{Q} = \begin{bmatrix} \tilde{q_{11}} & \gamma p_1 & \gamma p_2 & ... & \gamma p_{n-1} \\ r_1 & -r_1 & 0 & ... & 0 \\ r_2 & 0 & -r_2 & ... & 0 \\ . & & & & \\ . & & & & \\ . & & & & \\ r_{n-1} & 0 & 0 & ... & -r_{n-1} \end{bmatrix}, \quad V^i = \begin{bmatrix} v_1^i & 0 & 0 & ... & 0 \\ 0 & v_{01}^i & 0 & ... & 0 \\ 0 & 0 & v_{02}^i & ... & 0 \\ 0 & 0 & ... & ... & v_{0(n-1)}^i \end{bmatrix},$$

$$\tilde{q_{11}} = -\gamma \sum_{i=1}^{n-1} p_i \quad \text{and} \quad \gamma = \frac{1 - \prod_{i=1}^{n-1} \frac{r_i}{(r_i + p_i)}}{(\prod_{i=1}^{n-1} \frac{r_i}{r_i + p_i})(\sum_{i=1}^{n-1} \frac{p_i}{r_i})}.$$

Finally, one must add to (3.17)-(3.19), the normalization equation:

$$\int 1_{n-1}^T[\overrightarrow{f}^i(x)]dx + 1_{n-1}^T[\overrightarrow{P}^i + \overrightarrow{P}^{iz}] = 1 \qquad (3.20)$$

where $1_{n-1}^T = \underbrace{[1...1]}_{n-1}$.

Notice that if one considers the probability masses to be the unknowns, $\overrightarrow{P}^i$ and $\overrightarrow{P}^{iz}$ (recall (3.16)) add up to a total of n unknowns. Starting from vector $\overrightarrow{P}^{iz}$ and computing $\overrightarrow{f^i}(0^+)$ via $\overrightarrow{f^i}(z^-)$ (3.18) and the transition matrix of linear system (3.17), one obtains via (3.19) a linear system of n equations in n unknowns (the total number of unknowns in $\overrightarrow{P}^i$ and $\overrightarrow{P}^{iz}$). However the equations are linearly dependent because $rank[\tilde{Q}]$=$(n-1)$. This is why one still needs normalization equation (3.20) to completely specify the solution.

Now if matrix V^i were completely known, our $\tilde{x}_i$ building blocks would be independently solvable. The difficulty as will become clearer below comes from the fact that the velocity matrix in the real system depends on interactions between buffer x_i and its neighboring buffers x_{i-1} and x_{i+1} for a buffer strictly internal to the transfer line, while buffers x_1 and x_{n-1} are correlated through the CONWIP loop. In order to keep the building block calculations essentially independent, we replace this instantaneous time varying dependence by our best estimate of a long term expectation *conditional* on transfer line mode $\tilde{\alpha}(t)$, which is a random vector observable within all buffer component subsystems in the model. These conditional expectations depend on the $(n-1)^2$ unknown probability of availability of the $(n-1)$ loop wips, conditional on each of the $(n-1)$ allowed loop failure modes associated with $\tilde{\alpha}(t)$. They are the elements that create the true coupling between the blocks of our CONWIP model.

Construction of the velocity matrices.

In what follows, we shall give the details of our construction of the approximate vectors $\overrightarrow{v}^i$ (see (3.16)), defining the diagonal velocity matrices V_i, $i = 1, ..., (n-1)$ in (3.17)-(3.19). We define the following $(n-1)^2$ conditional probability of wip availability coefficients:

$$a_{i|0j} = Pr[x_i > 0|\tilde{\alpha} = 0j], \quad i,\ j = 1, ..., (n-1). \tag{3.21}$$

Furthermore, the following observations are inspired in part by the results of Monte-Carlo simulations:

- *Observation 1*

  After a transient phase, when the machines in the loop are all operational, wip tends to accumulate at downstream buffer x_{n-1} which eventually saturates at maximum level z. All other buffers are active, but *remain at level zero*, because wip is extracted faster than it is produced when loop saturation occurs (since M_1 produces then only at the rate at which wip is extracted from buffer $(n-1)$, while internal machines continue to extract parts at their maximum rate as long as it is possible);
- *Observation 2*

  Once the loop is blocked, it remains so through all machine failures, except that of machine M_1 which plays a critical role in the CONWIP loop. Thus the CONWIP loop is saturated with high probability if machine M_1 is very reliable;
- *Observation 3*

  Under the decreasing production rates assumption, an internal $(i \leq (n-2))$ buffer can saturate at level z only if the machine immediately downstream of it fails;
- *Observation 4*

Given that a string of machines in the loop M_i, M_{i+1}, ..., M_j $(j > i > 1)$ is operational, wip x_k can be positive for at least one k, $i \leq k < (j-1)$ only if $x_{j-1} > 0$. This is because of the assumed fluid parts production model (instaneous propagation of wip through the line), the decreasing maximum production rates assumption, and the fact that under the CONWIP policy, internal machines must produce at maximum rate whenever they can do so;

- *Observation 5*

  Whenever the loop is saturated, M_1 will produce at the rate at which parts are extracted from buffer x_{n-1}. However, according to the demand averaging principle, this rate is considered *constant* at $\tilde{d} = \frac{d}{a_{n-1}}$ where $a_{n-1} = Pr[x_{n-1} > 0]$, on the time intervals for which buffer x_{n-1} is active.

Given the above observations, we now estimate the required velocity vectors. In view of Observation 2, throughout our analysis we shall make the so-called *saturation assumption*: As long as machine M_1 is operational, unless evidence to the contrary is available, the CONWIP loop is considered saturated. The use of $\tilde{d}$ in the equations below is based on Observation 5.

Building block at buffer 1.

$v^1_{\tilde{\alpha}}$ for $0 < x_1 < z$:

- $\tilde{\alpha} = 1$:

$$v^1_1 = \tilde{d} - k_2. \tag{3.22}$$

  The above is because of the saturation assumption, Observation 5, the fact that buffer x_{n-1} is always active whenever all machine loops are operational, and the fact that the CONWIP discipline dictates that M_2 must produce at maximum rate.

- $\tilde{\alpha} = 01$:

$$v^1_{01} = -k_2. \tag{3.23}$$

  The above is because M_1 has failed.

- $\tilde{\alpha} = 02$:

$$v^1_{02} = \tilde{d}\, a_{n-1|02}. \tag{3.24}$$

  The rate of wip extraction is zero because M_2 has failed.

- $\tilde{\alpha} = 0i,\, i = 3, ..., (n-1)$:

$$v^1_{0i} = \tilde{d}\, a_{n-1|0i} - k_2. \tag{3.25}$$

The above wip rate of extraction is because we know that M_2 is operational. The only positive velocity corresponds to $\tilde{\alpha} = 02$ and will be associated with the only non zero entry of probability mass vector $\overrightarrow{P}^{1z}$ (see (3.16)) at $x_1 = z$. All other states are associated with negative velocity vectors and thus correspond to non zero entries of probability mass vector $\overrightarrow{P}^{1}$ at $x_1 = 0$ (see (3.16)).

Building block at intermediate buffer $i = 2, ..., (n-2)$.

$v^i_{\tilde{\alpha}}$ for $0 < x_i < z$:

- $\tilde{\alpha} = 1$:

$$v^i_1 = k_i\, Pr[\bigcup_{j=1}^{i-1}(x_j > 0)|\tilde{\alpha} = 1] + \tilde{d}\, Pr[(\bigcap_{j=1}^{i-1}(x_j = 0))\bigcap(x_{n-1} > 0)|\tilde{\alpha} = 1] - k_{i+1}. \tag{3.26}$$

The first two terms in the right-hand side correspond to the estimated production rate of machine M_i according to whether wip upstream of M_i is strictly positive in some buffer, or alternatively, all such buffers are strictly empty in which case the maximum production rate is adjusted to that of M_1. If we now rely on Observation 4 and the fact that x_{n-1} is always active when $\tilde{\alpha} = 1$, the above equation reduces to:

$$v^i_1 = k_i\, a_{i-1|1} + \tilde{d}\,(1 - a_{i-1|1}) - k_{i+1}. \tag{3.27}$$

- $\tilde{\alpha} = 0j,\, j = 1, ..., (i-1)$:

$$\begin{aligned} v^i_{0j} &= k_i\, Pr[\bigcup_{k=j}^{i-1}(x_k > 0)|\tilde{\alpha} = 0j] - k_{i+1} \\ &= k_i\, a_{i-1|0j} - k_{i+1}. \end{aligned} \tag{3.28}$$

The above follows from Observation 4.

- $\tilde{\alpha} = 0i$:

$$v^i_{0i} = -k_{i+1}. \tag{3.29}$$

The above is because M_i has failed.

– $\tilde{\alpha} = 0(i+1)$:

$$\begin{aligned} v^i_{0(i+1)} &= k_i Pr[\bigcup_{k=1}^{i-1}(x_k > 0)|\tilde{\alpha} = 0(i+1)] + \tilde{d}Pr[(\bigcap_{k=1}^{i-1}(x_k = 0))\bigcap(x_{n-1} > 0)|\tilde{\alpha} = 0(i+1)] \\ &= k_i\, a_{i-1|0(i+1)} + \tilde{d}\,(1 - a_{i-1|0(i+1)}). \end{aligned} \tag{3.30}$$

The first term in the last equation can be written based on Observation 4 while the second term is due to the observation that under the saturation assumption, total storage is equal to z, and since upstream of the failure the total storage is less than z, some positive storage must be present downstream of the failed machine. Thus, from Observation 4, x_{n-1} is necessarily active. The rest follows by recognizing that $Pr[(\bigcap_{k=1}^{i-1}(x_k = 0))] = 1 - Pr[\bigcup_{k=1}^{i-1}(x_k > 0)]$. Finally, note that the rate of extraction is zero because M_{i+1} is down.

– $\tilde{\alpha} = 0j$, $j = i+2, ..., (n-1)$:

$$v^i_{0j} = k_i\, Pr[\bigcup_{k=1}^{i-1}(x_k > 0)|\tilde{\alpha} = 0j] + \tilde{d}\, Pr[(\bigcap_{k=1}^{i-1}(x_k = 0))\bigcap(x_{n-1} > 0)|\tilde{\alpha} = 0j] - k_{i+1}. \tag{3.31}$$

For simplicity, in (3.31), we shall assume *independence* of the events $[(\bigcap_{k=1}^{i-1}(x_k = 0))|\tilde{\alpha} = 0j]$ and $[(x_{n-1} > 0)|\tilde{\alpha} = 0j]$. Thus:

$$Pr[(\bigcap_{k=1}^{i-1}(x_k = 0))\bigcap(x_{n-1} > 0)|\tilde{\alpha} = 0j] = Pr[(\bigcap_{k=1}^{i-1}(x_k = 0))|\tilde{\alpha} = 0j]\, Pr[(x_{n-1} > 0)|\tilde{\alpha} = 0j]. \tag{3.32}$$

(3.31) and (3.32) yield:

$$v^i_{0j} = k_i\, a_{i-1|0j} + \tilde{d}\,(1 - a_{i-1|0j})\, a_{n-1|0j} - k_{i+1}. \tag{3.33}$$

We shall assume the most likely case of negative velocities for all cases except $\tilde{\alpha} = 0(i+1)$. Thus all components in probability mass vector $\overrightarrow{P}^{iz}$ will be zero, except component $P^{iz}_{0(i+1)}$ (see (3.16)), while all components of probability mass vector $\overrightarrow{P}^{i}$ will be non zero except $P^i_{0(i+1)}$ (see (3.16)).

Building block at buffer $(n-1)$.

$v_{\tilde{\alpha}}^{n-1}$ for $0 < x_{n-1} < z$:

- $\tilde{\alpha} = 1$:

$$v_1^{n-1} = k_{n-1}\, Pr[(x_{n-2} > 0)|\tilde{\alpha} = 1] - \tilde{d}. \tag{3.34}$$

Since from the saturation assumption the loop is blocked and $x_{n-1} < z$, then some positive wip must be present upstream of M_{n-1}. From Observation 4, we conclude that $x_{n-2} > 0$. Thus the above equation reduces to:

$$v_1^{n-1} = k_{n-1} - \tilde{d}. \tag{3.35}$$

- $\tilde{\alpha} = 0j$, $j = 1, ..., (n-2)$:

$$\begin{aligned} v_{0j}^{n-1} &= k_{n-1}\, Pr[\bigcup_{k=j}^{n-2}(x_k > 0)|\tilde{\alpha} = 0j] - \tilde{d} \\ &= k_{n-1}\, a_{n-2|0j} - \tilde{d}. \end{aligned} \tag{3.36}$$

(3.36) follows from Observation 4.

- $\tilde{\alpha} = 0(n-1)$:

$$v_{0(n-1)}^{n-1} = -\tilde{d}. \tag{3.37}$$

In this case there is no production because M_{n-1} has failed. We shall consider the more likely case where all velocity vectors except for $\tilde{\alpha} = 1$ are negative. In this case, the only non zero component in probability mass vector $\overrightarrow{P}^{(n-1)z}$ will be $P_1^{(n-1)z}$, while all components but $P_1^{(n-1)}$ will be non zero in $\overrightarrow{P}^{(n-1)}$ (see (3.16)).

Aggregate wip building block.

$v_{\tilde{\alpha}}^{0}$ for $0 < x_0 < z$:

- $\tilde{\alpha} = 1$:

$$v_1^0 = k_1 - \tilde{d}\, Pr[(x_{n-1} > 0)|\tilde{\alpha} = 1]. \tag{3.38}$$

The above equation can be written because M_1 is operational and the loop is blocked.

Furthermore, given that all machines are active, buffer $(n-1)$ will necessarily be active, and thus the equation above reduces to:

$$v_1^0 = k_1 - \tilde{d}. \tag{3.39}$$

- $\tilde{\alpha} = 01$:

$$\begin{aligned} v_{01}^0 &= -\tilde{d}\, Pr[(x_{n-1} > 0) | \tilde{\alpha} = 01] \\ &= -\tilde{d}\, a_{n-1|01}. \end{aligned} \tag{3.40}$$

The production rate above is zero because M_1 is down.

- $\tilde{\alpha} = 0j,\, j = 2, ..., (n-1)$:

$$\begin{aligned} v_{0j}^0 &= k_1 - \tilde{d}\, Pr[(x_{n-1} > 0) | \tilde{\alpha} = 0j] \\ &= k_1 - \tilde{d}\, a_{n-1|0j}. \end{aligned} \tag{3.41}$$

In this case, all velocities are positive except when $\tilde{\alpha} = 01$: Thus all components of the probability mass vector $\overrightarrow{P}^{0z}$ will be non zero except P_{01}^{0z}, while all components of $\overrightarrow{P}^0$ will be zero except P_{01}^0 (see (3.16)).

3.5 Iterative calculation of system solution

Our proposed solution technique relies on the following observations on the structure of the Kolmogorov equations of the various subsystems already alluded too earlier: except for the particular case of the coefficient of availability $a_{(n-1)}$ that we shall discuss separately, provided probabilities $a_{i|0j}$ $i, j = 1, ..., (n-1)$, are assumed to be known, all subsystems, namely those associated with $\tilde{x}_j$, $j = 0, ..., (n-1)$, can be solved *independently*, i.e. as systems of n^{th} order differential equations. Thus, if the vector of these unknown probabilities is initialized sufficiently close to its true value, one could hope that through successive iterations, one would converge to a fixed vector associated with the overall solution.

Further remarks are that the x_0 subsystem (aggregate wip) only depends on $a_{(n-1)}$, and has by itself *no impact* on the rest of the subsystems, while the calculation of $a_{(n-1)}$ itself is carried out via the fixed point (provably convergent) algorithm discussed under the *Demand averaging principle* in subsection 3.3.2. Thus summarizing, the idea of the proposed algorithm is as follows: Start with an initial (reasonably good) guess of the unknown vector of coupling probabilities; solve for a_{n-1} through the fixed point algorithm applied to the $\tilde{x}_{(n-1)}$ dynamics; feed the result to the *separate* $\tilde{x}_j$ subsystems, $j = 1, ..., (n-2)$, to generate a new candidate vector of unknown probabilities; repeat the process until convergence is (hopefully) achieved.

Once convergence in the $\tilde{x}_j$ subsystems, $j = 1, ..., (n-1)$, is achieved, use the result to compute mean total storage in the system by solving for the $\tilde{x}_0$ dynamics.

The details of the algorithm are as follows: We shall refer to $a_{(n-1)|0j}$ $j = 1, ..., (n-1)$ as *last loop buffer coupling probabilities*, and $a_{i|0j}$ $i,\ j = 1, ..., (n-2)$, *as intermediate loop buffers coupling probabilities.*

Initialization of all unknown loop buffer coupling probabilities :

- $a_{(n-1)|0j}(0)$=1 $j = 1, ..., (n-1)$,
- $a_{i|0j}(0)$=0 , $i,\ j = 1, ..., (n-2)$, $i \neq (j-1)$,
- $a_{i-1|0i}(0)(0)$=1.

$\tilde{x}_{(n-1)}$ step: *fixed point calculation of new estimates of last loop buffer coefficient of availability and last loop buffer coupling probabilities.*

- (a) Initialize a_{n-1}=1,
- (b) Calculate $\tilde{d}$ using the most current estimate of a_{n-1},then solve the $\tilde{x}_{(n-1)}$ system of equations to compute $P_{0i}^{(n-1)}(t)$, $i = 1, ..., (n-1)$, $P_1^{(n-1)z}(t)$,
- (c) Calculate the new estimate of a_{n-1}: a_{n-1}=1-$\sum_{i=1}^{n-1} P_{0i}^{(n-1)}(t)$,
- Repeat steps (b) and (c) until convergence.
- Compute new estimates of last loop buffer coupling probabilities.

$\tilde{x}_j$ steps, $j = 1, ..., (n-2)$ *Computation of new estimates of intermediate loop buffer coupling probabilities*: Using the most recent estimates of a_{n-1} and loop buffer coupling probabilities $a_{j|0i}$, $i = 1, ..., (n-1)$, solve the $\tilde{x}_j$, $j = 1, ..., (n-2)$ systems of equations to obtain new estimates of buffer coupling probabilities at each one of the stages.

Fixed loop coupling probability vector computation: repeat the $\tilde{x}_{(n-1)}$ step, and the $\tilde{x}_j$ steps, $j = 1, ..., (n-2)$, until convergence of all loop buffer coupling probabilities.

Computation of total mean wip in the CONWIP loop: given $a_{(n-1)|0i}$, $i = 1, ..., (n-1)$, and a_{n-1}, solve the $\tilde{x}_0$ system of equations and compute $Pr[x_0(t) = z]$ and $E[x_0(t)]$.

3.6 Numerical results

In the following, we compare estimates of a number of transfer line related performance indicators based on our approximate CONWIP model, against those obtained from Monte-Carlo simulations. Percentage errors are reported for the estimated quantities. Two CONWIP loops are tested for various values of the maximal permissible storage parameter z. They include 3 and 4 machines associated respectively with 4 and 5 machine transfer lines, both subjected to a constant rate of demand for parts $d = 1$. The data for all machines is summarized in Table 3.1 below. Notice that in order to challenge the single permissible machine failure assumption and the corresponding machines failure parameters correction (see Appendix A), we have considered machines which can have a relatively high average percentage of down time (between 20 and 30%). Three types of results are reported: results relating to the convergence behavior of model based estimates of buffer subsystems coupling probabilities $a_{i|0j}$ $i,\ j = 1, ..., (n-1)$, and coefficient of wip availability at buffer $(n-1)$ a_{n-1} (service level of the CONWIP loop as defined in Section 3.2); results relating to the estimation of saturation probabilities of the CONWIP loop and the all important total mean wip ; results relating to the coefficients of availabilities of wips at various buffers within the loop, including that of the $(n-1)^{th}$ buffer which characterizes the all important service level of the CONWIP loop. Figure 3.6 and Figure 3.7 display respectively the convergence

Table 3.1 Machine parameters

i	1	2	3	4	5
p	0.15	0.22	0.18	0.2	0.18
r	0.55	0.5	0.45	0.52	0.75
k	3.4	3.2	3	2.9	2.8

behavior of CONWIP loop service level a_{n-1} and conditional coefficients of wip availability $a_{n-1|0j}$ $j = 1, ..., 4$ successive estimates as computed through the $\tilde{x}_{(n-1)}$ building block, itself adjusting to the estimates obtained from other intermediary buffer building blocks for the 5 machine transfer line, and a value of maximum storage parameter $z = 9$. Starting from initial estimates of 1, the values are seen to converge nearly within 4 system wide iterations.

The next table (Table 3.2) summarizes the performance of our approximate CONWIP model based computations insofar as estimating loop saturation probability and total mean loop storage as obtained from the aggregate wip building block. The loop saturation theoretical probability estimate is seen to decrease with CONWIP loop parameter z as one would expect, and the estimation accuracy remains quite good (worst case relative error of about 5 percent). As for the theoretical mean total wip estimate, it tends to increase with z as one

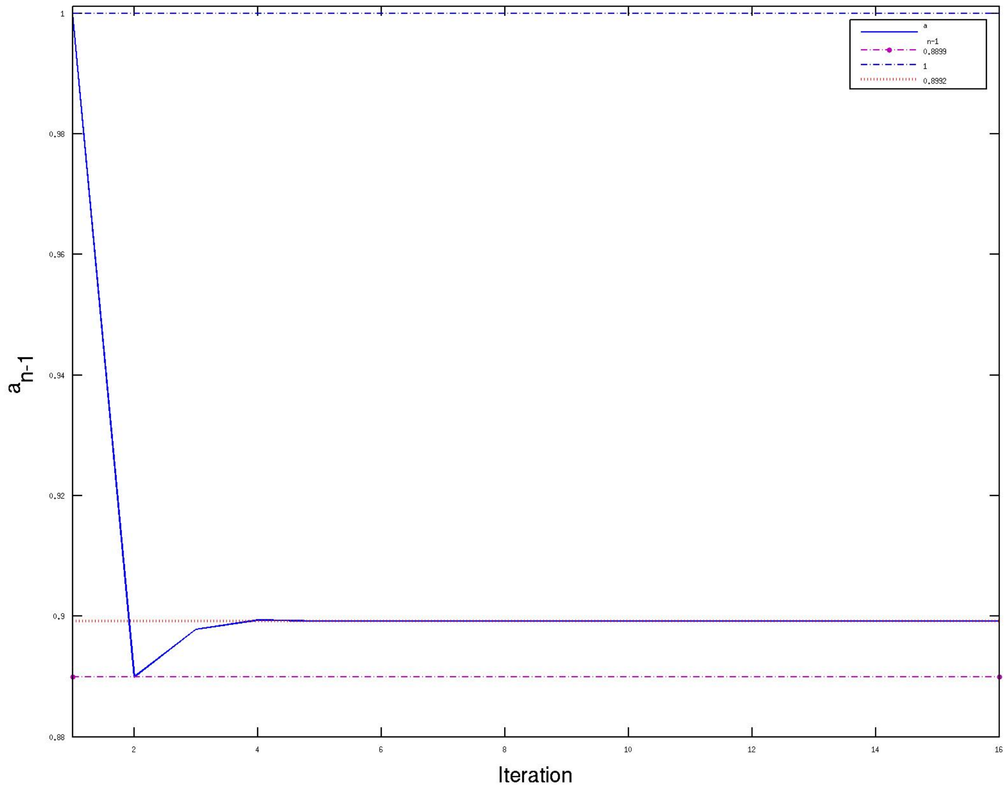


Figure 3.6 Convergence of CONWIP loop service level estimate a_{n-1} ($n = 5$) for the 5 machine transfer line, maximum storage parameter $z = 9$ and an initialization at 1.

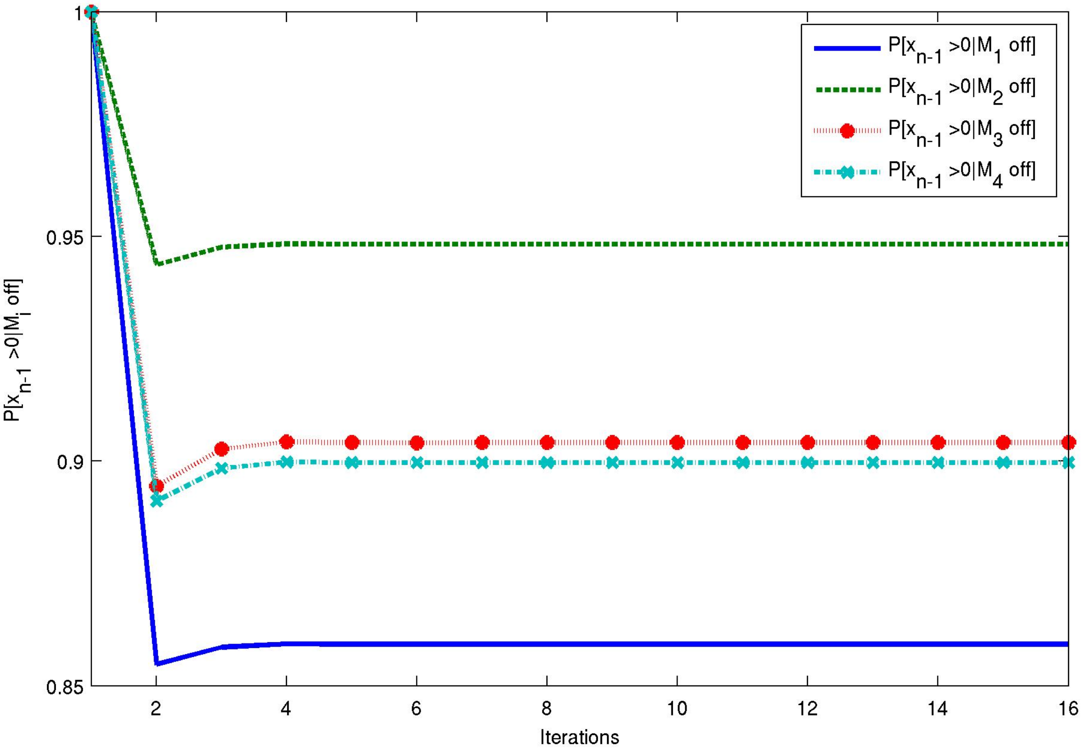


Figure 3.7 Convergence behavior of conditional coefficients of wip availability estimates $a_{n-1|0j}$ $j = 1, ..., 4$, for the $n = 5$ machine transfer line, maximum storage parameter $z = 9$ and an initialization of all estimates at 1.

would expect, and the estimation accuracy is also quite satisfactory except for small values of z for which error sensitivity is clearly expected to be higher. Also, the accuracy of estimation does not seem to be significantly affected by the length of the CONWIP loop.

The next two tables (Tables 3.3 and 3.4) summarize the performance of wip coefficients of availability model based estimation respectively for a 3 machine and a 4 machine CONWIP loop with corresponding transfer lines of 4 and 5 machines again respectively. In both tables the CONWIP loop service level is identified with the wip coefficient of availability of highest index. Theoretical estimates of service levels tend to increase with the level of the CONWIP loop storage parameter z as intuitively expected, and the accuracy of estimation is high for the service level itself and remains good to acceptable for the less important internal wip coefficients of availability. Let us note that for a 5 machines transfer line, Monte-Carlo simulations on a Pentium (R)4 (CPU 2.6 GHz) took about two days of CPU time, while model

Table 3.2 Total wip

		MC Simulation		Theory based estimate		Percentage Error	
$n-1$	z	$P[x_0 = z]$	$E[x_0]$	$P[x_0 = z]$	$E[x_0]$	$P[x_0 = z]$	$E[x_0]$
3	2	0.7755	1.7288	0.7647	1.8234	1.39	-5.47
3	5	0.7351	4.4205	0.717	4.509	2.46	-2
3	8	0.7117	7.3296	0.6906	7.0854	2.96	3.33
4	3	0.7486	2.86	0.778	2.762	-3.90	3.41
4	9	0.7012	8.2863	0.667	8.05	4.92	2.85

based computations require only a few minutes of CPU to produce an answer.

Table 3.3 A 3 machine CONWIP loop ($n-1=3$)

	MC simulation			Theory based estimate			Percentage error		
z	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
2	0.8274	0.6563	0.5965	0.7731	0.6097	0.5804	6.56	7.10	2.70
5	0.8352	0.6648	0.7791	0.8188	0.6576	0.8087	1.96	1.08	-3.80
8	0.8373	0.6899	0.9038	0.8593	0.7091	0.8964	-2.63	-2.78	0.82

Table 3.4 A 4 machine CONWIP loop ($n-1=4$)

	MC simulation				Theory based estimate				Percentage error			
z	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
3	0.830	0.659	0.520	0.702	0.873	0.675	0.501	0.743	5.76	-2.35	3.75	-5.70
9	0.841	0.690	0.533	0.884	0.823	0.689	0.524	0.899	2.17	0.09	1.69	-1.70

3.7 Conclusion

We have developed an approximate mathematical model, amenable to computations, for the evaluation of important performance indicators of CONWIP controlled transfer lines, which can incorporate an arbitrary number of failure prone machines. The model for a $(n-1)$ machine CONWIP loop corresponds to a concatenation of $(n-1)$ building blocks with each of those representing a "buffer centric", or local view of the transfer line, as seen from a particular buffer. The loop mode, i.e. the joint discrete state of the machines in the CONWIP loop is assumed to be shared by all buffer centric subsystems. In addition, a building block is dedicated to the dynamics of the total wip within the loop. In our aggregate modeling, the total wip dynamics is seen to be essentially affected by the reliability statistics of machine M_1, and the probability of wip availability at buffer $(n-1)$, thus reflecting the view of CONWIP as a form of Kanban imposed on a collection of machines.

The individual buffer models correspond to Kolmogorov linear differential equations with boundary conditions, and are coupled by a large vector of unknown probabilities which is initialized and iteratively computed so as to converge to a fixed point. The accuracy of the theory based predictions appears to be quite satisfactory when compared to the results of Monte-Carlo simulations, particularly as far as mean total wip estimation and CONWIP loop service level (with machine M_n viewed as the "customer") are concerned. The availability of the current modeling tool, makes it possible to compute minimal storage requirements in a CONWIP controlled loop given a required service level and a fixed demand rate. It can also become part of a tool for the optimization of hybrid Kanban/CONWIP (Bonvik *et al.* (2000)) architectures in transfer lines.

CHAPITRE 4

A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects

Article soumis pour publication au journal *Discrete Event Dynamic Systems* (Juillet 2011) et écrit par:

Fatima Zahra Mhada

École Polytechnique de Montréal and GERAD

Roland P. Malhamé

École Polytechnique de Montréal and GERAD

Robert Pellerin

École Polytechnique de Montréal and CIRRELT

4.1 Introduction

Inventory control is a recurring issue both in the manufacturing systems environment and in supply chains. In the manufacturing systems environment, it is used to regularize the flow of parts that meet an external demand to prevent shortages or overproduction problems.

The issue is that low inventory reduces the cost of storage but increases the risks of shortages, thus leading to penalties due to unsatisfied demands, while excessive inventories are costly to maintain. Bielecki et Kumar (1988) have established that for a two-state machine, a constant demand, and a cost function including long term storage and backlog costs, the optimal policy is the critical inventory type. This type involves building up and maintaining, whenever possible, a critical level of inventory as a hedge against possible machine failures. Such a policy has also been called a hedging production policy (Gershwin (1994)). However, an attempt to optimize inventory policies without accounting for the quality and acceptability of the produced parts leads to an incomplete analysis. Indeed, in reality, there is always a percentage of manufactured parts that do not meet the quality criteria, either due to poor adjustment of the machine or to the non-compliance of raw materials.

In the literature, the integration of quality management with production management has been the focus of great interest in recent years, but, despite this level of interest, a limited number of articles on the subject exist. Kim and Gershwin (Kim et Gershwin (2005), Kim et Gershwin (2008)) developed a fluid production model aimed at studying the interaction of quality and productivity. The model is based on the assumption that a machine produces defective parts until it undergoes maintenance-related adjustments. This model is subsequently coupled with decomposition techniques developed in (Gershwin (1994)), so as to study the interaction of quality and productivity in transfer lines. However, in both works, the flow of production parts is considered continuous while quality is regarded as a property of discrete parts. This leads to some difficulties in the analysis. We develop our models in this paper partly so as to circumvent the latter difficulty. However, they also carry their own burden of new challenges.

Colledani and Tolio (Colledani et Tolio (2005), Colledani et Tolio (2006a) ,Colledani et Tolio (2006b),Colledani et Tolio (2011)) in contrast to Kim and Gershwin, address quality and production issues simultaneously in a completely discrete framework (discrete time and discrete parts with individual quality attributes); they consider a production system composed of unreliable manufacturing stations and inspection stations with different failure modes. Statistical quality control charts are introduced at inspection stations and act as noisy measurements of the state of quality of the machines. Decomposition methods for the integrated production/quality performance studies of the line are developed.

The Colledani-Tolio work constitutes a powerful practical modeling paradigm however, (i) it is cast in an entirely discrete framework and we wish to work with fluid models of parts; (ii) it includes objectives of corrective maintenance which are beyond the scope of this paper; and (iii) it is too complex at this stage for our objectives of a step by step fundamental understanding of the interactions of parts quality with manufacturing system productivity. Instead, we limit our analysis here to a single unreliable machine producing a single part type, which, in the normal course of things, will produce a fraction of defective parts. For this system, we wish to revisit the optimization of hedging policies.

Our hope is to develop, in this manner, analytical building blocks for more complex architectures.

4.2 Model description

We consider a production system that consists of a machine with two functional states and a breakdown state, fuelled by an infinite supply of raw materials and connected to a finite buffer capacity destined to receive finished parts.

The machine evolves according to a continuous time Markov chain with 3 discrete states $\alpha(t) = 1, 2, 3$ hereby called *modes* and defined as follows:

- $\alpha(t) = 1$: the machine is operational and produces conforming (good) parts $x_1(t)$.
- $\alpha(t) = 2$: the machine is operational but produces non-conforming parts $x_2(t)$.
- $\alpha(t) = 3$: the machine is down.

The failure mode is the only completely observable mode, while the other two modes are observable only insofar as the machine is operational, but the quality state is not directly observable. The machine can evolve between these three states according to the transition diagram in Figure 4.1 with failure rate p, repair rate r, and transfer rates between states 1

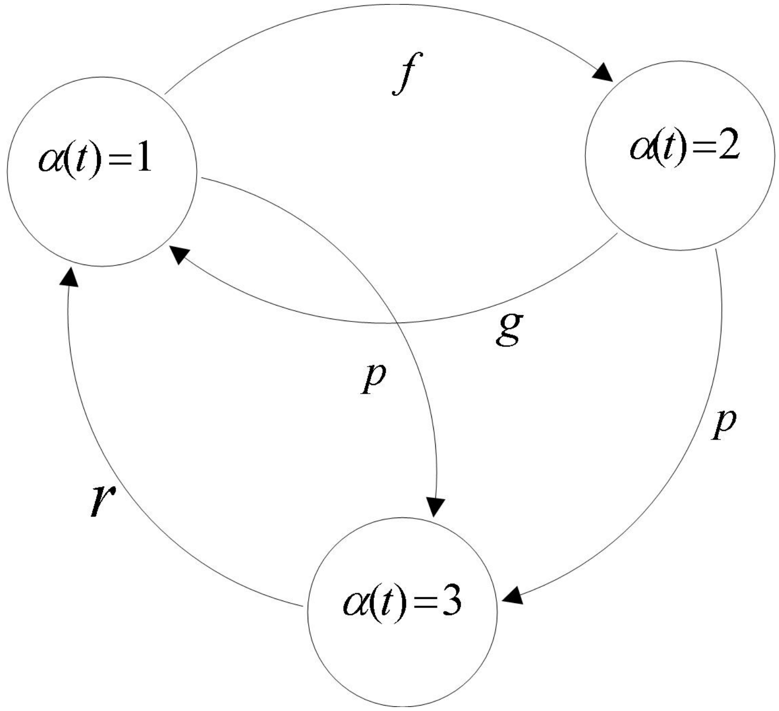


Figure 4.1 The continuous time Markov chain machine model.

and 2 respectively given by f and g. Notice that we consider that production of defective parts is an inherent part of the normal production process and that the machine oscillates spontaneously between states 1 and 2. More precisely, in our modeling framework, preventive or corrective maintenance can only *modify the rates* at which the machine switches between states 1 and 2. Furthermore, we make the following assumptions:

– The rate of demand for good parts is a constant d;

– The maximal rate of production is k;

– The total stock $x(t)$ consists of a *perfectly mixed* stock of good parts $x_1(t)$ and of a stock of non conforming parts $x_2(t)$;

– Because the stocks of good and defective parts are assumed to be perfectly mixed, in order to draw good parts at rate d, it is necessary to draw non conforming parts at the instantaneous rate $\frac{x_2(t)}{x_1(t)}d$;

– The stock $x_1(t)$ can become negative (backlog) unlike the stock of non conforming parts $x_2(t)$ which can never become negative;

– The production of parts is considered to be a continuous process;

- The capacity of storage is limited by z (the only design parameter of the hedging policy);
- The storage cost is c^+ per part per unit time;
- The backlog cost is c^- per part per unit time;
- The parts production rate at t is denoted $u(t)$, with $0 \leq u(t) \leq k$.

The objective is to find the hedging point z for the total stock $x(t) = x_1(t)+x_2(t)$ to minimize the long term average cost:

$$J(x(0), \alpha(0)) = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\Big[\int_0^T \Big(c^+ \max\left(x(\tau), 0\right) + c^- \max\left(-x(\tau), 0\right)|x(0), \alpha(0)\Big) d\tau\Big]. \quad (4.1)$$

Conforming and non conforming part stocks respectively evolve according to:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= u(t)\, I(\alpha(t) = 1) - d; \qquad (4.2)\\ \frac{dx_2(t)}{dt} &= u(t)\, I(\alpha(t) = 2) - \frac{x_2(t)\, d}{x_1(t)}, \end{aligned}$$

with:

$$\begin{aligned} x(t) &= x_1(t) + x_2(t) \quad \text{if} \quad x_1(t) \geq 0; \qquad (4.3)\\ x(t) &= x_1(t) \quad \text{and} \quad x_2(t) = 0 \quad \text{if} \quad x_1(t) < 0. \end{aligned}$$

In the above, I(t) is the indicator function, and the production rate is defined according to the hedging policy objective of producing at the maximum rate when below the hedging point and producing exactly what is required to remain at the hedging point once the hedging point is reached. More precisely:

$$u(t) = \begin{cases} k & \text{if } x_1(t) + x_2(t) < z \\ d + \frac{x_2(t)\, d}{x_1(t)} & \text{if } x_1(t) + x_2(t) = z \\ 0 & \text{if } x_1(t) + x_2(t) > z \end{cases}$$

While, in general, the above control policy may not be the optimal one for the optimal control problem (4.1)-(4.2), in Section 4.7 of the paper, we establish that for a limited case where the switching rates between states 1 and 2 go to infinity while $\frac{f}{p+g}$ goes to some constant ratio β, the limiting model can be subsumed by the Bielecki-Kumar framework (Bielecki et Kumar (1988)), and the optimal policy indeed remains a hedging policy.

4.3 Piecewise-deterministic stock dynamics

By studying (4.2) and (4.3), one can identify distinct regions with smooth dynamics. These regions are:

- The $0 < x_1(t) + x_2(t) < z$ region;
- The $x_1(t) + x_2(t) = z$ region;
- The $x_2(t) = 0$ region.

4.3.1 Region $0 < x_1(t) + x_2(t) < z$

Mode $\alpha(t) = 1$

In this mode, stocks evolve according to:

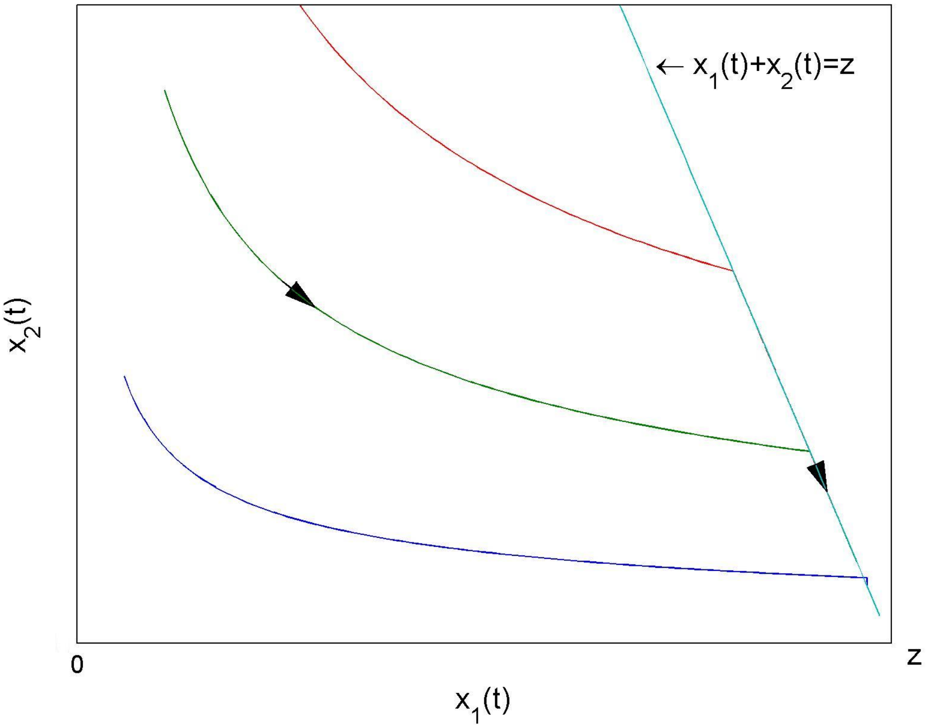


Figure 4.2 Phase diagrams in mode $\alpha(t) = 1$.

$$\begin{aligned} \frac{dx_1(t)}{dt} &= k - d; \\ \frac{dx_2(t)}{dt} &= -\frac{x_2(t)\, d}{x_1(t)}. \end{aligned} \tag{4.4}$$

By solving this system, we obtain $x_1(t)$ and $x_2(t)$.

$$\begin{aligned} x_1(t) &= x_1(0) + (k-d)\, t; \\ x_2(t) &= x_2(0) (\frac{x_1(t)}{x_1(0)})^{\frac{-d}{k-d}}. \end{aligned} \tag{4.5}$$

We note that stock $x_1(t)$ increases while $x_2(t)$ decreases, and the total stock increases until the line $x_1(t) + x_2(t) = z$ is reached.

Mode $\alpha(t) = 2$

In this mode, stocks evolve according to:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= -d; \\ \frac{dx_2(t)}{dt} &= k - \frac{x_2(t)\, d}{x_1(t)}. \end{aligned} \tag{4.6}$$

Thus $x_1(t)$ and $x_2(t)$ are defined by (See Appendix B):

$$\begin{aligned} x_1(t) &= x_1(0) - d\,.t; \\ x_2(t) &= x_1(t)\, (\frac{x_2(0)}{x_1(0)} + \frac{k}{d} \log \frac{x_1(0)}{x_1(t)}). \end{aligned} \tag{4.7}$$

In this case, we note that the stock $x_1(t)$ decreases, whereas the behaviour of $x_2(t)$ is more complex:

- If $k - \frac{x_2(t)\, d}{x_1(t)}$ is negative, stock $x_2(t)$ will decrease until it reaches zero *at the same time* as $x_1(t)$.
- If $k - \frac{x_2(t)\, d}{x_1(t)}$ is positive, stock $x_2(t)$ will increase, and given that $\dot{x_1}(t) + \dot{x_2}(t) = k - d - \frac{x_2(t)\, d}{x_1(t)}$, we can further distinguish two cases:
  1. $\frac{k}{d} - 1 < \frac{x_2(t)}{x_1(t)} < \frac{k}{d}$, so $(\dot{x_1}(t) + \dot{x_2}(t) > 0)$: here stock $x_2(t)$ increases and compensates for the reduction in $x_1(t)$. This implies that the total stock can reach the line $x_1(t) + x_2(t) = z$.

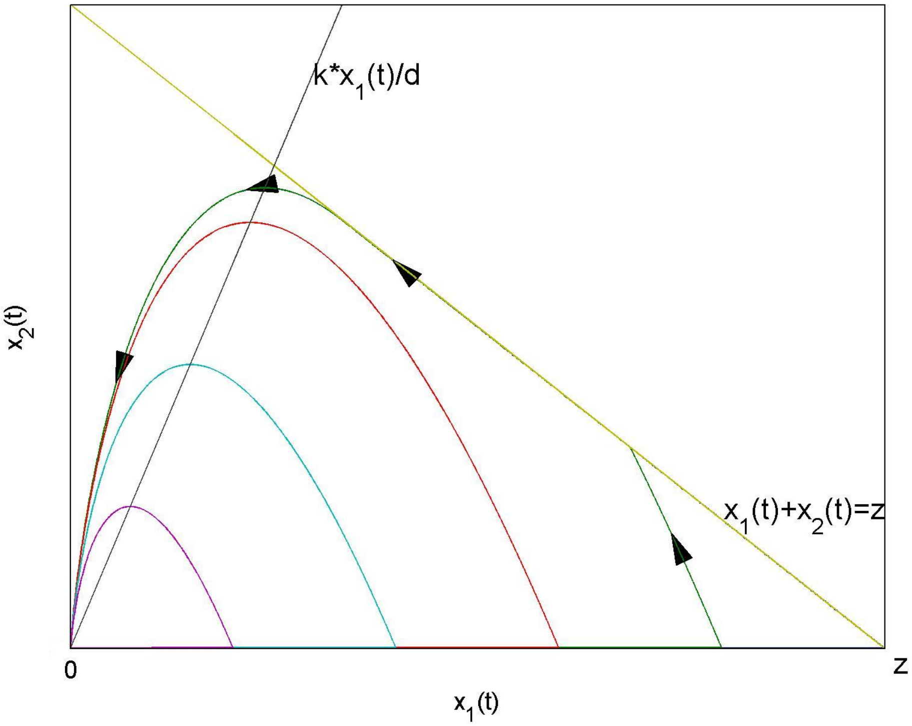


Figure 4.3 Phase diagrams in the mode $\alpha(t) = 2$ with $x_2(0) = 0$.

2. $\frac{x_2(t)}{x_1(t)} < \frac{k}{d} - 1$, so that $(\dot{x_1}(t) + \dot{x_2}(t) < 0)$: here despite the fact that stock $x_2(t)$ increases, this increase does not offset the decrease of $x_1(t)$ and thus the total stock decreases.

Mode $\alpha(t) = 3$

In this mode, stocks evolve according to:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= -d; \\ \frac{dx_2(t)}{dt} &= -\frac{x_2(t)\,d}{x_1(t)}. \end{aligned} \tag{4.8}$$

Stocks $x_1(t)$ and $x_2(t)$ will decrease until they reach zero:

$$\begin{aligned} x_1(t) &= x_1(0) - d\,.t; \\ x_2(t) &= \frac{x_2(0)}{x_1(0)} x_1(t). \end{aligned} \tag{4.9}$$

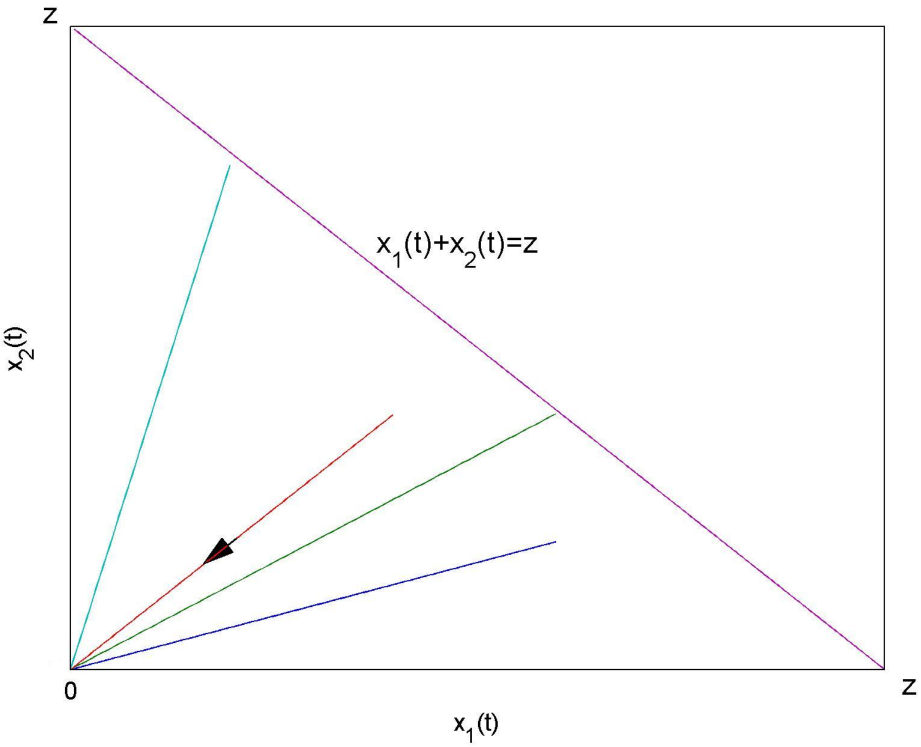


Figure 4.4 Phase diagrams in the mode $\alpha(t) = 3$.

4.3.2 Region $x_1(t) + x_2(t) = z$

As shown in the previous section, there are only two ways to reach this region: mode $\alpha(t) = 1$ or mode $\alpha(t) = 2$. Also, it is only these 2 modes that are defined in this region. Here the production rate is equal to the demand drawn from the entire stock (both conforming and non conforming parts) to maintain the total stock at z, that is, $u(t) = (d + \frac{x_2(t)\, d}{x_1(t)})$. So, the stock dynamics can be expressed as:

For mode $\alpha(t) = 1$**:**

$$\begin{aligned} x_1(t) - x_1(0) + z \log \frac{z - x_1(t)}{z - x_1(0)} + d\,.t = 0; \\ x_2(t) - x_2(0) - z \log \frac{x_2(t)}{x_2(0)} - d\,.t = 0. \end{aligned} \tag{4.10}$$

The above equations are derived in Appendix C. Notice that in this mode, stock $x_1(t)$ will increase *without ever reaching* $x_1(t) = z$, because the rate at which x_1 increases is exactly that at which x_2 decreases and as x_2 approaches zero, the rate $\frac{x_2}{x_1}d$ at which x_2 decreases goes to zero.

For mode $\alpha(t) = 2$:

$$\begin{aligned} x_1(t) &= x_1(0) - d\,.t; \\ x_2(t) &= d\,.t + x_2(0). \end{aligned} \tag{4.11}$$

Thus $x_2(t)$ increases while $x_1(t)$ decreases until the tipping point located at $\frac{x_2(t)}{x_1(t)} = \frac{k}{d} - 1$ is reached. At that point, trajectories leave the $x_1(t)+x_2(t) = z$ line to the area $0 < x_1+x_2 < z$ and start a new trajectory which obeys the following equations (see appendix B):

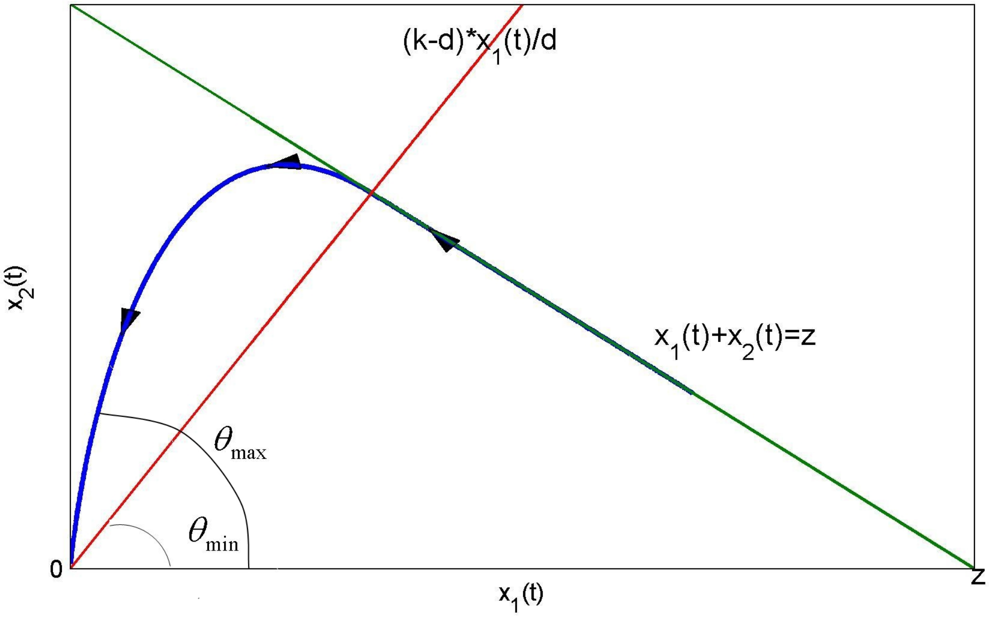


Figure 4.5 The boundary zone of the area $0 < x_1 + x_2 < z$.

$$\begin{aligned} &x_1(0) = \frac{z\,d}{k}; \quad x_2(0) = \frac{z(k-d)}{k}; \\ &x_2(t) = x_1(t)\,(\frac{k-d}{d} + \frac{k}{d}\ln(\frac{z\,d}{k}) - \frac{k}{d}\ln(x_1(t))); \\ &x_1(t) = \frac{z\,d}{k} - d\,.t \end{aligned} \tag{4.12}$$

Note that this trajectory represents an *extreme left hand natural boundary* of the trajectories in the region $(0 < x_1 + x_2 < z) \cap (x_1 \geq 0) \cap (x_2 \geq 0)$, since even if switching towards modes 1 or 3 occurs, motion will take place to the right of this boundary. Also note that the probability that this extreme boundary be a trajectory is a *strictly positive* number. It is thus associated with *a univariate probability density function.*

Despite the fact that $x_2(t)$ increases, the total stock does not increase, and line $x_1 + x_2 = z$ is never reached again directly through that mode.

Note that it is also possible to directly leave the line $x_1(t) + x_2(t) = z$ if the machine enters failure in mode 3, in which case, the trajectories move back to region $0 < x_1(t) + x_2(t) < z$.

4.3.3 On line $x_2(t) = 0$

Mode $\alpha(t) = 1$:

$$\begin{aligned} x_1(t) &= x_1(0) + (k - d)\, t; \\ x_2(t) &= 0. \end{aligned} \tag{4.13}$$

Stock $x_1(t)$ increases to a maximum of z.

Mode $\alpha(t) = 2$:

$$\begin{aligned} x_1(t) &= x_1(0) - d\,.t; \\ x_2(t) &= 0 \quad \text{if } (x_1(0) \leq 0). \end{aligned} \tag{4.14}$$

Note that when in mode 2, the only way that produced defective parts can persist in the system, is if a non zero stock of conforming parts is already present; this is because otherwise, the model dynamics dictate that in an effort to extract good parts at rate d, all non conforming parts would be instantaneously eliminated as soon as they are produced. As a result, in effect, on the half line $x_2(t) = 0$, $x_1(t) \leq 0$, the dynamics associated with mode 2 is essentially indistinguishable from that associated with mode 3. Furthermore, only stock x_1 is defined in this region.

Mode $\alpha(t) = 3$**:**

$$\begin{aligned} x_1(t) &= x_1(0) - d\,.t; \qquad (4.15) \\ x_2(t) &= 0. \end{aligned}$$

Stock $x_1(t)$ decreases continuously.

The above study of various stock dynamics according to mode and region in stock space, will be quite instrumental in developing the Kolmogorov equations describing the probabilistic evolution of the corresponding stochastic hybrid state Markov process; in particular, we shall be able to identify where probability masses, univariate or bivariate probability density functions (pdf) are needed to characterize that evolution.

4.4 Steady-state Forward Kolmogorov equations

Following (Algoet (1989)), we define the steady-state vector current density of probability $\vec{J}_j$ by: $\vec{J}_j = f_j(x_1, x_2)\vec{v}_j(x_1, x_2)$ with $f_j(x_1, x_2)$ the bivariate probability density function associated with x_1 and x_2, in mode $\alpha = j$, i.e., for infinitesimal $d\lambda_1$, $d\lambda_2$ intervals:

$$\begin{aligned} f_j(\lambda_1, \lambda_2)d\lambda_1 d\lambda_2 &= Pr[(\lambda_1 < x_1 \leq \lambda_1 + d\lambda_1, \\ &\qquad \lambda_2 < x_2 \leq \lambda_2 + d\lambda_2)\bigcap(\alpha = j)] \quad j = 1, 2, 3, \qquad (4.16) \end{aligned}$$

and $\vec{v}_j(x_1, x_2)$ is the velocity vector associated with machine mode $\alpha = j$. It is defined as the difference between the production rate vector in mode j and the corresponding region, $\vec{u}_j(x_1, x_2) \equiv [u_{jx_1} \quad u_{jx_2}]^T$, and the demand vector $\vec{d}(x_1, x_2) \equiv [d_{x_1} \quad d_{x_2}]^T$ in the corresponding region. In what follows and for expediency, we shall drop the arguments of functions whenever that does not result in ambiguity.

4.4.1 Region $(x_1 > 0) \cap (x_2 > 0) \cap (0 < x_1 + x_2 < z)$

Velocity vectors and transition matrix

The velocity vectors are given by:

- $\vec{v}_1 = [k - d \quad -\frac{x_2}{x_1} d]^T$;
- $\vec{v}_2 = [-d \quad k - \frac{x_2}{x_1} d]^T$;
- $\vec{v}_3 = [-d \quad -\frac{x_2}{x_1} d]^T$.

The transition matrix is:

$$Q = \begin{bmatrix} -(f+p) & f & p \\ g & -(p+g) & p \\ r & 0 & -r \end{bmatrix} \equiv [q_{\gamma\alpha}].$$

Forward Kolmogorov equations

Stocks x_1 and x_2 are associated with steady-state pdf which satisfy the following Kolmogorov equations (Algoet (1989)):

$$\nabla . \vec{J}_\alpha(x_1, x_2) = \sum_{\gamma=1}^{3} f_\gamma(x_1, x_2)\, q_{\gamma_\alpha}, \quad \alpha = 1, 2, 3. \tag{4.17}$$

The matrix form is given by:

$$V_1 \frac{\partial \vec{f}(x_1, x_2)}{\partial x_1} + V_2(x_1, x_2) \frac{\partial \vec{f}(x_1, x_2)}{\partial x_2} + M(x_1) \vec{f}(x_1, x_2) = Q^T \vec{f}(x_1, x_2), \tag{4.18}$$

with:

$$\begin{aligned} \vec{f}(x_1, x_2) &\equiv [f_1(x_1, x_2), \quad f_2(x_1, x_2), \quad f_3(x_1, x_2)]^T; \\ V_1 &\equiv diag[k - d, -d, -d]; \\ V_2(x_1, x_2) &\equiv diag[-\frac{x_2}{x_1} d, k - \frac{x_2}{x_1} d, -\frac{x_2}{x_1} d]; \\ M(x_1) &\equiv diag[-\frac{1}{x_1} d, -\frac{1}{x_1} d, -\frac{1}{x_1} d]. \end{aligned}$$

Application of a scalar method of characteristics

In an effort to develop a numerical analysis scheme for the resulting system of partial differential equations (P.D.E.s), in lines $i = 1, 2, 3$ of (4.18), we shall treat $f_j(x_1, x_2)$, $j \neq i$ as known exogenous functions, and use the method of characteristics as applied to scalar

P.D.E.s.

– Mode 1: The first line of equation (4.18) is:

$$\begin{aligned}(k-d)\frac{\partial f_1(x_1,x_2)}{\partial x_1}+(-\frac{x_2(t)}{x_1(t)}d)\frac{\partial f_1(x_1,x_2)}{\partial x_2} = & \quad (\frac{d}{x_1(t)}-(f+p))\,f_1(x_1,x_2)\\ & + \quad g\,f_2(x_1,x_2)+r\,f_3(x_1,x_2).\end{aligned}$$

We wish to transform this linear first-order PDE into an ODE along the appropriate curve; we thus define: $\frac{dx_1}{ds}=k-d$ and $\frac{dx_2}{ds}=-\frac{x_2(t)}{x_1(t)}d$ yielding the following parametric equations for the characteristic curve:

$$\begin{aligned}x_1(s) &= x_1(0)+(k-d)\,s;\\ x_2(s) &= x_2(0)\,(\frac{x_1(s)}{x_1(0)})^{\frac{-d}{k-d}},\end{aligned}$$

and the ODE along the characteristic curve is:

$$\frac{df_1(x_1(s),x_2(s))}{ds}=(\frac{d}{x_1(s)}-(f+p))f_1(x_1(s),x_2(s))+g\,f_2(x_1(s),x_2(s))+r\,f_3(x_1(s),x_2(s)).$$

Therefore, on this characteristic, the general solution in terms of the assumed known exogenous functions $f_j(x_1,x_2)$, $j\neq 1$ is:

$$\begin{aligned}f_1(x_1(s),x_2(s)) = & \exp(-(f+p)s)\,x_1(s)^{\frac{d}{k-d}}\Big[x_1(0)^{\frac{-d}{k-d}}f_1(x_1(0),x_2(0)) \\ & + \int_0^s \exp((f+p)\tau)\,x_1(\tau)^{\frac{-d}{k-d}}(g\,f_2(x_1(\tau),x_2(\tau))+r\,f_3(x_1(\tau),x_2(\tau)))d\tau\Big].\end{aligned} \tag{4.19}$$

Similarly, one can show for modes 2 and 3 that:

– Mode 2:

$$\begin{aligned}f_2(x_1(s),x_2(s)) = & \frac{\exp(-(p+g)s)}{x_1(s)}\Big[x_1(0)\,f_2(x_1(0),x_2(0)) \\ & +f\int_0^s \exp((p+g)\tau)\,x_1(\tau)f_1(x_1(\tau),x_2(\tau))\,d\tau\Big]\end{aligned} \tag{4.20}$$

along the characteristic curve described by the parametric equations:

$$\begin{aligned} x_1(s) &= x_1(0) - d\,s; \\ x_2(s) &= x_2(0) - x_1(0)\,\frac{k}{d}\,\ln(\frac{x_1(s)}{x_1(0)}). \end{aligned} \tag{4.21}$$

– Mode 3:

$$\begin{aligned} f_3(x_1(s), x_2(s)) &= \frac{\exp(-rs)}{x_1(s)}\Big[x_1(0)\,f_3(x_1(0), x_2(0)) \\ &+p\int_0^s \exp(r\tau)\,x_1(\tau)(f_2(x_1(\tau), x_2(\tau)) + f_1(x_1(\tau), x_2(\tau)))d\tau\Big] \end{aligned} \tag{4.22}$$

along the characteristic curve described by the parametric equations:

$$\begin{aligned} x_1(s) &= x_1(0) - d\,s; \\ x_2(s) &= \frac{x_2(0)}{x_1(0)}\,x_1(s). \end{aligned} \tag{4.23}$$

4.4.2 On the line: $x_2 = 0$, $x_1 < z$

Denote $f_j^0(x_1)$ the univariate steady-state pdf associated on the half-line defined by $x_1 < z$ and $x_2 = 0$, in mode $\alpha = j$, $j = 1, 2, 3$.

For $x_1 < 0$:

$\vec{f^0}(x_1) \equiv [f_1^0(x_1) \quad f_2^0(x_1) \quad f_3^0(x_1)]$ satisfies the following forward steady state Kolmogorov equations:

$$\frac{d\vec{f^0}(x_1)}{dx_1} = V_1^{-1}Q^T\vec{f^0}(x_1) \tag{4.24}$$

the solution of which can be written

$$\vec{f^0}(x_1) = \exp(V_1^{-1}Q^T x_1)\vec{f^0}(0^-). \tag{4.25}$$

Remark: $f_2^0(x_1)$ exists only for $x_1 < 0$ because, otherwise, switching into mode $\alpha = 2$ would cause the trajectory to leave the line $x_2 = 0$, $x_1 < z$.

For $0 < x_1 < z$**:**

$f_1^0(x_1)$ and $f_3^0(x_1)$ satisfy the following system of steady-state forward Kolmogorov equations:

$$\begin{aligned}(k-d)\frac{df_1^0(x_1)}{dx_1} &= -(f+p)\,f_1^0(x_1) + r\,f_3^0(x_1);\\ -d\frac{df_3^0(x_1)}{dx_1} &= p\,f_1^0(x_1) - r\,f_3^0(x_1).\end{aligned} \tag{4.26}$$

The solution of which is given by:

$$\vec{f^0}(x_1) = \exp\Big(-\begin{bmatrix} -\frac{(f+p)}{k-d} & \frac{r}{k-d} \\ -\frac{p}{d} & \frac{r}{d}\end{bmatrix}(z-x_1)\Big)\,\vec{f^0}(z^-). \tag{4.27}$$

Remark: $f_1^0(x_1)$ is continuous across $x_1 = 0$; this is unlike $f_3^0(x_1)$ given that in mode 3, it is possible to reach the origin both from the half-line $x_1 \geq 0$ and the first quadrant. Thus typically, $f_3^0(0^-)$ will be larger than $f_3^0(0^+)$.

4.4.3 On segment $S = \{(x_1; x_2) | (x_1 \in [0; z]) \cap (x_2 = z - x_1)\}$

Denote $f_j^z(x_1)$ the univariate pdf of x_1 on segment S in mode j, $j = 1, 2$ (these are the only ones that can persist on S).

For a small Δx_1, a part of the bivariate pdf at $(x_1;\, x_2)$ (near segment S) comes to feed the univariate pdf in mode 1 (Figure 4.6) and in mode 2 (Figure 4.7) . Let $B_1(t, x_1, \Delta x_1)$ denote the probability content in mode 1 of the red section at time t. One can write, for a small time increment Δt:

$$\begin{aligned}&B_1(t+\Delta t, x_1, \Delta x1) - B_1(t, x_1, \Delta x_1) \qquad (4.28)\\ &= \left[f_1^z(x_1,t)\frac{(z-x_1)d\sqrt{2}}{x_1} - f_1^z(x_1+\Delta x_1,t)\frac{(z-x_1-\Delta x_1)d\sqrt{2}}{x_1+\Delta x_1}\right]\Delta t\;[1-(f+p)\Delta t]\\ &-\int_{x_1}^{x_1+\Delta x_1}\sqrt{2}(f+p)\,f_1^z(x,t)dx + \int_{x_1}^{x_1+\Delta x_1}\sqrt{2}g\,f_2^z(x,t)dx\\ &+\left[\int_{z-x_1-\Delta x_1}^{z-x_1}(k-d)\,f_1(x_1,y)dy - \int_{x_1}^{x_1+\Delta x_1}\frac{(z-x_1)d}{x_1}\,f_1(x,z-x)dx\right]\\ &\Delta t\;[1-(f+p)\Delta t]\;o(\Delta t)\end{aligned}$$

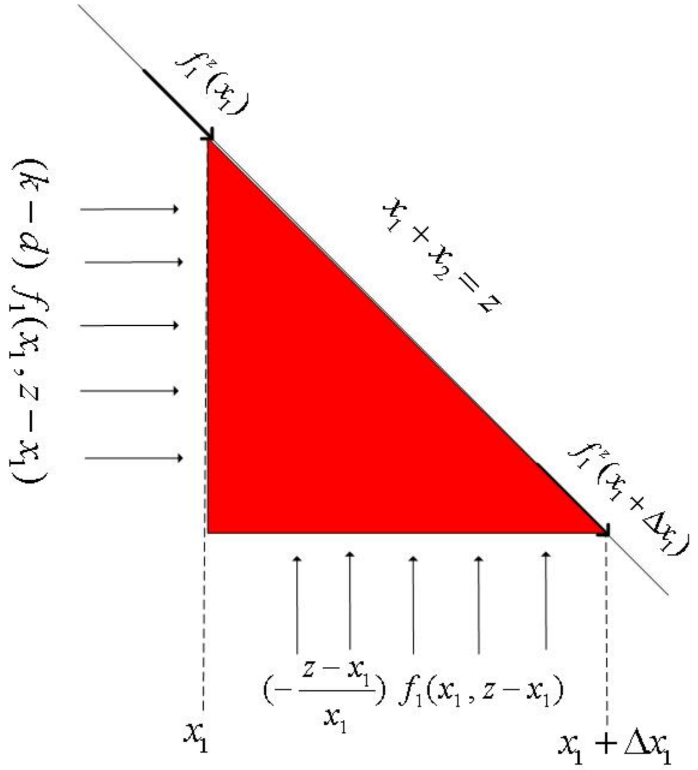


Figure 4.6 Boundary $x_1 + x_2 = z$ in mode 1.

When Δx_1 and Δt tend to zero, (4.28) becomes:

$$\begin{aligned}
&\lim_{\Delta x_1, \Delta t \to 0} (B_1(t+\Delta t, x_1, \Delta x1) - B_1(t, x_1, \Delta x_1)) = \frac{df_1^z(x_1,t)}{dt} \qquad (4.29)\\
&= -\frac{d(f_1^z(x_1)\,\frac{(z-x_1)}{x_1})}{dx_1} d\sqrt{2} - \sqrt{2}(f+p)\, f_1^z(x_1,t) + \sqrt{2}g\, f_2^z(x_1,t)\\
&+(k-d)\, f_1(x_1, z-x_1) - \frac{(z-x_1)d}{x_1}\, f_1(x_1, z-x_1).
\end{aligned}$$

At steady-state:

$$\begin{aligned}
\frac{df_1^z(x_1)}{dx_1}\frac{(z-x_1)d\sqrt{2}}{x_1} &= (-(f+p) + \frac{z}{x_1^2}d)\,\sqrt{2}\, f_1^z(x_1) + \sqrt{2}g\, f_2^z(x_1)\\
&+ (k-d)\, f_1(x_1, z-x_1) - \frac{(z-x_1)d}{x_1}\, f_1(x_1, z-x_1).
\end{aligned}$$

Following a similar analysis for mode 2, $f_1^z(x_1)$ and $f_2^z(x_1)$ can be shown to satisfy the following system of steady-state forward Kolmogorov differential equations:

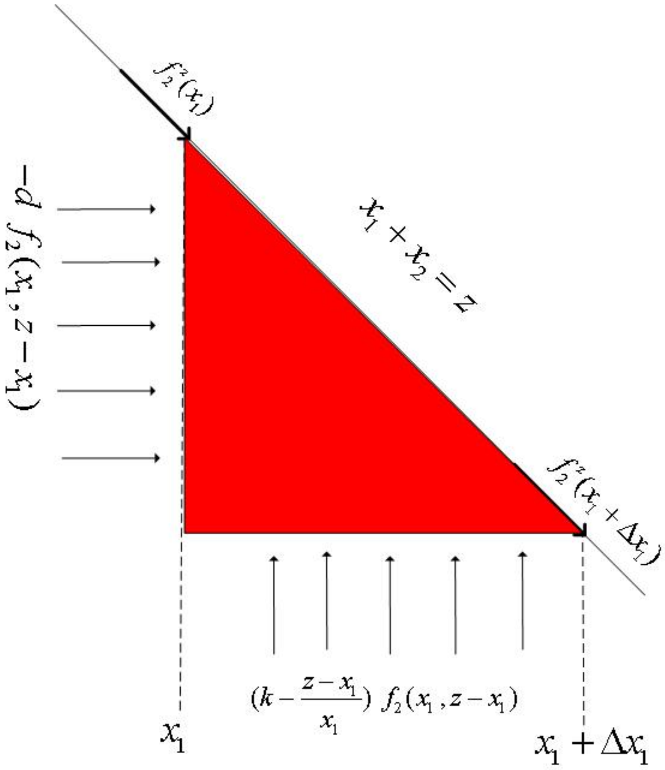


Figure 4.7 Boundary $x_1 + x_2 = z$ in mode 2.

$$\begin{aligned}&\frac{(z-x_1)d}{x_1}\frac{df_1^z(x_1)}{dx_1} = (-f-g+\frac{z}{x_1^2}d)f_1^z(x_1)+gf_2^z(x_1)\\&+\frac{k-d-\frac{(z-x_1)d}{x_1}}{\sqrt{2}}f_1(x_1,z-x_1);\end{aligned}\tag{4.30}$$

$$\frac{df_2^z(x_1)}{dx_1} = \frac{p+g}{d}f_2^z(x_1)-\frac{f}{d}f_1^z(x_1)+f_2(x_1,z-x_1)\frac{k-d-\frac{(z-x_1)d}{x_1}}{\sqrt{2}d}.\tag{4.31}$$

4.4.4 On the left hand boundary trajectory

Let $f_2^{lim}(x_1(s), x_2(s))$ be the univariate pdf associated with the limiting extreme left hand boundary curve discussed in Subsection 4.3.2 above. It is also the continuation of function $f_2^z(x_1)$ on segment S (Figure 4.5). The peculiarity of this function is that it satisfies the Kolmogorov equation (2nd line of system (4.18)), except that it is no longer accessible through modes (1 and 2). Thus, $f_2^{lim}(x_1, x_2)$ satisfies the following steady-state forward Kolmogorov equation:

$$-d\,\frac{\partial f_2^{lim}(x_1,x_2)}{\partial x_1} + (k - \frac{x_2(t)}{x_1(t)}d)\frac{\partial f_2^{lim}(x_1,x_2)}{\partial x_2} = (\frac{d}{x_1(t)} - (g+p))\,f_2^{lim}(x_1,x_2). \qquad (4.32)$$

The univariate pdf on the extreme left hand boundary is accordingly given by:

$$f_2^{lim}(x_1(s),x_2(s)) = \frac{\exp(-(p+g)s)}{x_1(s)}\,x_1(0)\,f_2^{lim}(x_1(0),x_2(0)), \qquad (4.33)$$

with $f_2^{lim}(x_1(0),x_2(0)) = f_2^z(zd/k)$ and from 4.3.2, the extreme left hand boundary satisfies the following parametric equations: $x_1(t) = \frac{z\,d}{k} - d\,t$, $x_2(t) = x_1(t)\,(\frac{k-d}{d} + \frac{k}{d}\ln(\frac{z\,d}{k}) - \frac{k}{d}\ln(x_1(t)))$.

4.5 Boundary conditions

In this section we will present three types of boundary conditions; specifically:

- Boundary conditions involving probability transfers from a bivariate pdf to a univariate pdf (at point $(0,0)$).
- Boundary conditions involving probability transfers from a univariate pdf to a bivariate pdf both at the half-line $(x_2 = 0 \cap x_1 > 0)$, and segment S.
- Boundary conditions involving probability transfers between univariate pdf's and a probability mass at a given point $(x_1 = z, x_2 = 0)$.

4.5.1 At $(0,0)$ for univariate density $f_2^0(x_1)$

In mode 2, whenever $k - (1 + \frac{x_2(t)}{x_1(t)})d < 0$, the total stock level decreases until it reaches the origin $(x_1 = x_2 = 0)$, if the mode persists long enough.

Figure 4.8 shows that for a small Δx_1, support of the associated pdf's is restricted to an angle defined by two rays at θ_{min} and θ_{max} originating at (0,0).

Following the above picture, in the neighborhood of origin $(0,0)$ is an area of transformation of a bivariate pdf $f_2(x_1,x_2)$ into an univariate pdf $f_2^0(x_1)$.

At steady-state, the total probability flux that enters the red triangle in Figure 4.8, call it $Flux_2(\Delta x_1)$, must be equal to that escaping to the left of (0,0) along the half line

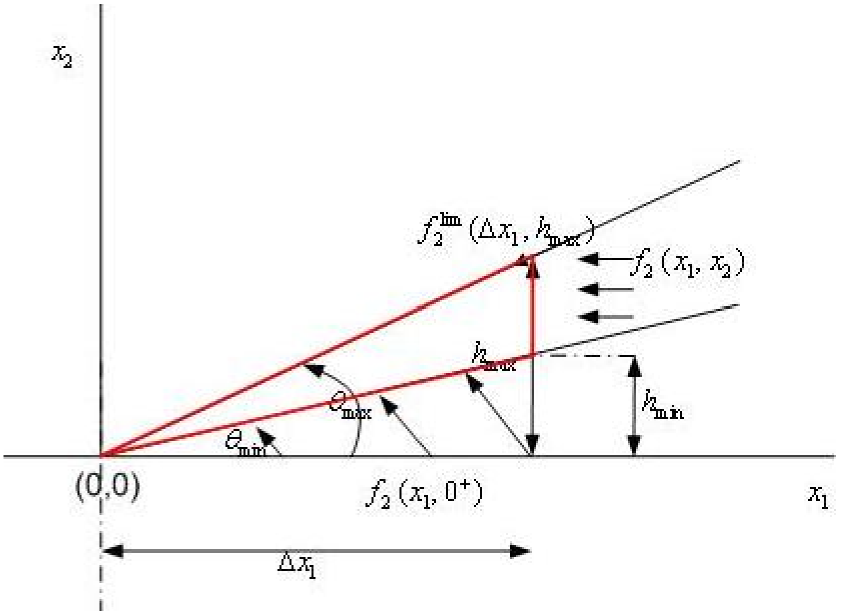


Figure 4.8 Mode 2 related probability transfers near $(0, 0)$.

$(x_2 = 0) \cap (x_1 < 0)$. Thus:

$$Flux_2(\Delta x_1) = d \int_0^{(h_{max}-h_{min})} f_2(\Delta x_1, y)dy + \int_0^{\Delta x_1} d\, f_2(y, y\, tan(\theta_{max}))\, dy + d\, f_2^{lim}(\Delta x_1, h_{max}) \tag{4.34}$$

with: $\tan(\theta_{min}) = \frac{k}{d}$ and $h_{max} = \Delta x_1\, (\frac{k-d}{d} + \frac{k}{d}\ln(\frac{z\,d}{k}) - \frac{k}{d}\ln(\Delta x_1))$. Note that the rightmost term above corresponds to an incoming probability flux along the limiting curve of subsection 4.3.2, which involves a univariate pdf. As discussed above, this incoming flux must be compensated exactly by a net escape probability from the interval of length Δx_1 on the left of $(0, 0)$. As a result, one can write:

$$\begin{aligned} &d\, f_2^0(0^-) - d\, f_2^0(0^+) = \\ &d \int_0^{(h_{max}-h_{min})} f_2(\Delta x_1, y)dy + d \int_0^{\Delta x_1} f_2(y, y\, tan(\theta_{max}))\, dy + d\, f_2^{lim}(\Delta x_1, h_{max}) \end{aligned} \tag{4.35}$$

with $f_2^0(0^+) = 0$. Note that we do not attempt to convert (4.35) into its differential form in view of the fact that singularities occur as Δx_1 goes to zero; this is because as one gets closer to the origin the bivariate pdf approaches a univariate pdf with the same probability mass, and as a result the magnitude of the bivariate pdf increases without bound (bivariate delta function).

4.5.2 At $(0,0)$ for univariate density $f_3^0(x_1)$

In mode 3, the levels of stock x_1 and x_2 will decrease until $(0,0)$ is reached (Figure 4.9). Denoting again this time by $Flux_3(\Delta x_1)$ the total flux crossing into the red triangle of Figure 4.9, and following arguments along the lines of those on Section 4.5.1 arguments, one can write:

$$Flux_3(\Delta x_1) = d \int_0^{(h_{max})} f_3(\Delta x_1, y)dy$$

and

$$df_3^0(0^-) - d\, f_3^0(0^+) = Flux_3(\Delta x_1)$$

Thus:

$$f_3^0(0^-) = f_3^0(0^+) + \frac{Flux_3(\Delta x_1)}{d}. \tag{4.36}$$

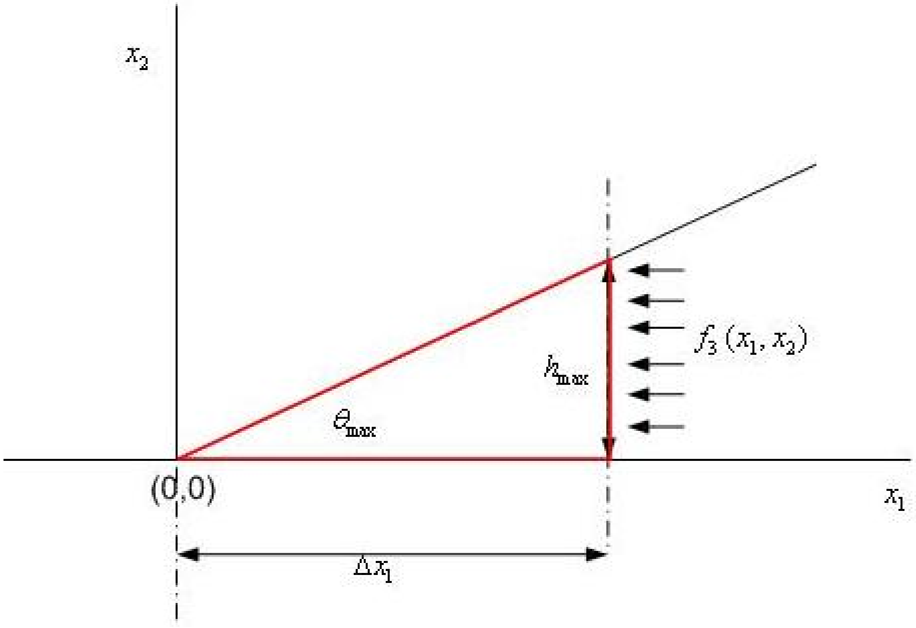


Figure 4.9 Mode 3 related probability fluxes near $(0,0)$.

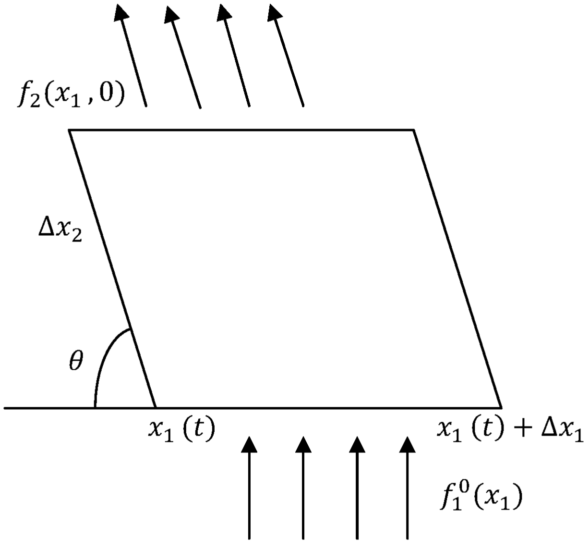


Figure 4.10 Boundary $x_1 > 0$, $x_2 = 0$ with switching from mode 1 to mode 2.

4.5.3 On the half-line $x_1 : 0| \, x_1 < z$

Provided that $x_1 > 0$, this boundary is left towards $0 < x_1 + x_2 < z$ with $x_2 > 0$ whenever switching occurs from mode 1 to mode 2 and this occurs with rate f. Thus, during a small time interval Δt, a fraction $f\Delta t$ of probability is released and starts moving at a vertical velocity $k \sin\theta$ (see Figure 4.10), with the furthermost position reached being $\Delta x_2 = k \sin\theta \quad \Delta t$. At steady state, the boundary condition becomes:

$$f \int_{x_1}^{x_1+\Delta x_1} f_1^0(x) dx \Delta t = (\int_{x_1 - \frac{\Delta x_2}{\tan(\theta)}}^{x_1 - \frac{\Delta x_2}{\tan(\theta)} + \Delta x_1} k f_2(x, \Delta x_2) dx) \Delta t$$

i.e:

$$f f_1^0(x_1) \Delta x_1 = k f_2(x_1, \Delta x_2) \Delta x_1.$$

As Δx_1, Δt and thus Δx_2 go to zero, we obtain:

$$f f_1^0(x_1) = k f_2(x_1, 0). \tag{4.37}$$

4.5.4 On the left hand boundary trajectory

On the left hand extreme boundary trajectory discussed in subsection 4.3.2, there are two possibilities to leave mode 2 : Either towards mode 1 with rate g thus bolstering bivariate pdf $f_1(x_1, x_2)$ at point $(x_1(s)^-, x_2(s)^-)$, or to mode 3 with rate p thus bolstering bivariate pdf $f_3(x_1(s), x_2(s))$ at point $(x_1(s)^-, x_2(s)^-)$.

Therefore, following arguments parallel to those in Section 4.5.3, one can write at steady-state:

$$g f_2^{lim}(x_1(s), x_2(s)) = \frac{k(-k + d + \frac{x_2 d}{x_1})}{\sqrt{d^2 + (k - \frac{x_2 d}{x_1})^2}} f_1(x_1(s)^-, x_2(s)^-); \tag{4.38}$$

$$p f_2^{lim}(x_1(s), x_2(s)) = \frac{kd}{\sqrt{d^2 + (k - \frac{x_2 d}{x_1})^2}} f_3(x_1(s)^-, x_2(s)^-). \tag{4.39}$$

4.5.5 On segment $S = \{(x_1; x_2) | (x_1 \in [0; z]) \cap (x_2 = z - x_1)\}$

The only possibility to leave segment S is when the machine fails i.e. by moving to mode 3 and $f_3(x_1^-, z - x_1^-)$. Hence, we can write :

$$p\left(f_1^z(x_1) + f_2^z(x_1)\right) = \frac{(d + \frac{(z - x_1)d}{x_1})}{\sqrt{2}} f_3(x_1^-, z - x_1^-). \tag{4.40}$$

4.5.6 At point $x_1 = z, x_2 = 0$

Since there is a non zero probability of sitting at point $(0, z)$, in the steady-state, it will be associated with a probability mass which we shall denote P_z. This probability mass fed from mode 1 and depleted towards mode 3 at rate p on the half-line $\{x_1 > 0\ ,\ x_2 = 0\}$, and towards mode 2 at rate f on segment S. Thus at steady-state one can write:

$$d\, f_3^0(z^-) = p\, Pz; \tag{4.41}$$

$$d\, f_2^z(z^-) = f\, Pz; \tag{4.42}$$

$$(k - d)\, f_1^0(z^-) = (p + f)\, Pz. \tag{4.43}$$

4.5.7 Normalization equation

The total probability must be equal to 1, so that:

$$\begin{aligned} P_{Total} &= \Big(\int_{-\infty}^{z}(f_1^0(y)+f_2^0(y)+f_3^0(y))dy\Big)+\Big(\int_{0}^{z}(f_1^z(y)+f_2^z(y))dy\Big) \\ &+ \Big(\int_0^z\int_0^z(f_1(y_1,y_2)+f_2(y_1,y_2)+f_3(y_1,y_2))dy_1\,dy_2\Big)+\;Pz \qquad (4.44) \\ &= 1. \end{aligned}$$

4.6 The numerical algorithm

The above Kolmogorov PDE's with boundary conditions are mathematically intractable, and therefore we resort to numerical analysis for performance evaluation and performance optimization. We now proceed with presentation of the proposed numerical algorithm.
For a specified Δx_1 , it comprises the following steps:

1. Initialization of $Pz = P[x_1 = z, x_2 = 0]$ to value $Pz^0 = 1$,

2. Given Pz^0, we calculate $\vec{f^0}(x_1)$ on the half-line $\{x_2 = 0, x_1 > 0\}$ by relying on analytical solution (4.27) with the following boundary conditions:

$$\begin{aligned} d\,f_3^0(z^-) &= p\,Pz, \\ (k-d)\,f_1^0(z^-) &= (p+f)\,Pz, \end{aligned}$$

   and using (4.37), we calculate $f_2(x_1, 0)\,\forall\,0 \le x_1 \le z$;

3. We calculate $f_2(x_1, x_2)$ defined by (4.20) on the characteristic curves (4.21) whose initial points are defined by $x_2(0) = 0$, $0 < x_1(0) < z$; this is in particular important for obtaining estimates of bivariate density $f_2(x_1^-, z - x_1^-)$ in the neighborhood of segment S for further computations;
   For the first iteration, the unknown bivariate pdf's $f_1(x_1, x_2)$ and $f_3(x_1, x_2)$ are initialized to zero.

4. For the first iteration, we consider univariate pdf $f_1^z(x_1)$ on segment S is zero; otherwise, the most current estimate is used. With the estimate of $f_2(x_1^-, z - x_1^-)$ obtained in step

3, we solve system (4.30 and 4.31) with

$$\begin{aligned} d\,f_2^z(z^-) &= f\,Pz, \\ f_1^z(\frac{z^+d}{k}) &= 0 \end{aligned}$$

to estimate $f_1(x_1, z - x_1)$, $f_2^z(x_1)$ and in particular $f_2^z(zd/k)$.

Using $f_2^z(x_1)$ and $f_1^z(x_1)$ in (4.40), we estimate $f_3(x_1^-, z - x_1^-)$;

5. We use $f_2^z(zd/k)$ to estimate the leftmost boundary univariate pdf $f_2^{lim}(x_1, x_2)$ ((4.33) on the trajectory defined by parametric equations (4.12)).

6. Using (4.38, 4.39), we calculate bivariate pdf's $f_1(x_1^-(s), x_2^-(s))$ and $f_3(x_1^-(s), x_2^-(s))$ along the leftmost boundary curves defined by (4.12);

7. Using the latest estimates of $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$, and using the estimates of $f_3(x_1^-(s), x_2^-(s))$ along the leftmost boundary curve and $f_3(x_1^-, z - x_1^-)$ near segment S as boundary values, we calculate $f_3(x_1, x_2)$ (4.22) along its characteristic curves (4.23).

8. Using the latest estimates of $f_3(x_1, x_2)$ and $f_2(x_1, x_2)$, and the latest estimate of $f_1(x_1^-(s), x_2^-(s))$ along the leftmost boundary curve as boundary values, we calculate $f_1(x_1, x_2)$ from (4.19) along its characteristic curves;

9. With the updated estimates of $f_1(x_1, x_2)$ and $f_3(x_1, x_2)$, we repeat steps 3-8 until a convergence of probability functions within the required precision;

10. We calculate $f_1^0(0^-)$ using $f_1^0(0^-) = f_1^0(0^+)$, $f_2^0(0^-)$ and $f_3^0(0^-)$ using respectively (4.35) and 4.36);

11. We calculate $\vec{f^0}(x_1)$ on $x_2 = 0$, $x_1 < 0$ using (4.25);

12. We calculate P_{Total} using (4.44);

13. We reinitialize Pz^0 to Pz^0/P_{Total} , and we repeat steps 3 to 12;

14. When convergence is achieved within the required precision, we calculate the following values for validation purposes:
    - $Pr[x_1 < 0, x_2 = 0] = (\int_{-\infty}^{0}(f_1^0(y) + f_2^0(y) + f_3^0(y))dy)$;
    - $Pr[0 < x_1 < z, x_2 = 0] = (\int_0^z(f_1^0(y) + f_3^0(y))\,dy)$;
    - $Pr[x_1 + x_2 = z] = (\int_0^z(f_1^z(y) + f_2^z(y))dy) + Pz(t)$;

– The storage cost:

$$\begin{aligned} E^{+} &= \Big(\int_0^z y\,(f_1^0(y)+f_3^0(y))\,dy\Big) + \Big(\int_0^z y\,(f_1^z(y)+f_2^z(y))dy\Big) + z\,Pz(t) \\ &+ \Big(\int_0^z \int_0^z (y_1+y_2)\,(f_1(y_1,y_2)+f_2(y_1,y_2)+f_3(y_1,y_2))dy_1\,dy_2\Big); \end{aligned}$$

– The shortage cost: $E^{-} = \int_{-\infty}^{0} y\,(f_1^0(y)+f_2^0(y)+f_3^0(y))dy$.

4.6.1 Validation

For comparison purposes, we have developed for a particular choice of machine parameters, a Monte Carlo simulation model in Arena V.10.

Table 4.1 details the parameters of our simulation example; the simulation results were found to be close to those obtained from our analytical model for different values of z (Table 4.2).

Let us note that the Monte-Carlo simulations on a Pentium (R) Core Duo CPU (2 GHz) took about 3 hours of CPU time, while model based computations require only a few minutes of CPU to produce an answer.

Table 4.1 The system parameters

r	p	g	f	k	d	c^{+}	c^{-}
0.1	0.05	0.5	0.005	2.8	1	1	10

Table 4.2 The simulation and analytical results. The confidence interval for simulated probability is ±0.005 and for simulated cost is ±0.05

z	5		4		3	
	Simul.	analytic	Simul.	analytic	Simul.	analytic
$Pr[x1 < 0, x_2 = 0]$	0,359	0,355	0,391	0,392	0,415	0,413
$Pr[0 < x1 < z, x_2 = 0]$	0,138	0,140	0,115	0,117	0,093	0,094
$Pr[x_1 + x_2 = z]$	0,415	0,410	0,417	0,419	0,425	0,429
E^{+}	2,885	2,828	2,197	2,215	1,605	1,594
E^{-}	4,867	4,923	5,278	5,222	5,612	5,726

4.7 Comparison with the modified Bielecki-Kumar machine

In this section, we show that in the limiting case where the switching rates between states 1 and 2 of the three mode machine become much faster than other switching rates, while $\frac{f}{p+g}$ goes to a constant ratio β, states 1 and 2 are effectively merged together and become a single operational state where both conforming and non-conforming parts are simultaneously produced so that:

$$\begin{aligned} x_2(t) &= \beta * x_1(t); \qquad (4.45) \\ x_1(t) &= \frac{x(t)}{(1+\beta)}. \end{aligned}$$

In addition, the three modes model converges to a two mode model, essentially equivalent to a modified Bielecki-Kumar machine (Bielecki et Kumar (1988)) which produces both good and bad parts simultaneously, with total conforming and non conforming stock $x(t) = x_1(t)+x_2(t)$, two states ($\alpha = 1$ and $\alpha = 0$), a maximum production rate equal to k , a storage cost c^+, a modified shortage cost $\frac{c^-}{(1+\beta)}$, and a modified demand rate $(1+\beta)d$ (see (Mhada *et al.* (2011)) for further details).

To verify this, designating by $p_i(x)$ the pdf of stock x for the modified Bielecki-Kumar machine model where $i = 0$, 1, indicates the machine mode and compare it with $g_i(x_1 + x_2)$ the pdf associated with total stock in the three-mode machine model for an adequately redefined mode $i = \alpha' = 0, 1$; $\alpha' = 1$ whenever $\alpha = 1, 2$, while $\alpha' = 0$, whenever $\alpha = 3$. Thus $\alpha' = 1$ is the *aggregate* mode of $\alpha = 1$ and $\alpha = 2$ in the three- mode machine.

4.7.1 Determining $p_i(x)$

$p_i(x)$ is defined by:

$$p_i(\lambda)d\lambda = Pr[(\lambda < x \leq \lambda + d\lambda) \bigcap (\alpha = i)] \quad i = 0,\ 1.$$

The matrix form of the solution of the forward Kolmogorov equation is given by:

$$\vec{p}(x) = \exp\Big(\begin{bmatrix} -\frac{p}{k-(1+\beta)d} & \frac{r}{k-(1+\beta)d} \\ -\frac{p}{(1+\beta)d} & \frac{r}{(1+\beta)d} \end{bmatrix} (x-z)\Big)\, \vec{p}(z^-) \tag{4.46}$$

subject to the boundary and normalization conditions:

$$\begin{aligned} &\vec{p}(-\infty) = 0; \\ &(k-(1+\beta)d)\, p_1(z^-) = p\, Pz; \\ &\int_{-\infty}^{z} (p_1(x) + p_0(x))\, dx + \ Pz = 1. \end{aligned}$$

Then the $p_i(x)$ expressions are:

$$\begin{aligned} p_0(x) &= \frac{\delta\, p}{p+r} \exp(\delta\,(x-z)) \quad \text{for} \quad x \le z; \qquad (4.47) \\ p_1(x) &= \frac{\delta\, p\,(1+\beta)d}{(p+r)(k-(1+\beta)d)} \exp(\delta\,(x-z)) \quad \text{for} \quad x \le z; \qquad (4.48) \\ Pz &= \frac{\delta\, d^*}{p+r}, \qquad (4.49) \end{aligned}$$

and where $\delta = \frac{r}{(1+\beta)d} - \frac{p}{k-(1+\beta)d} > 0$ and $d^* = (1+\beta)d$.

4.7.2 Determining $g_i(x_1 + x_2)$

$g_i(x)$ for $i = 1, 0$ is obtained as follows:

– In the area $x < 0$

$$\begin{aligned} g_1(x) &= f_1^0(x_1) + f_2^0(x_1)\, \text{with}\, x = x_1; \\ g_0(x) &= f_3^0(x_1)\, \text{with}\, x = x_1. \end{aligned}$$

- In the area $0 < x < z$

$$
\begin{aligned}
g_1(x) &= f_1^0(x) + \int_0^z (f_1(y, x-y) + f_2(y, x-y))\, dy; \\
g_0(x) &= f_3^0(x) + \int_0^z f_3(y, x-y)\, dy.
\end{aligned}
$$

- In the area $x = z$

$$g_1(z) = Pz + \int_0^z (f_1^z(y) + f_2^z(y))\, dy. \tag{4.50}$$

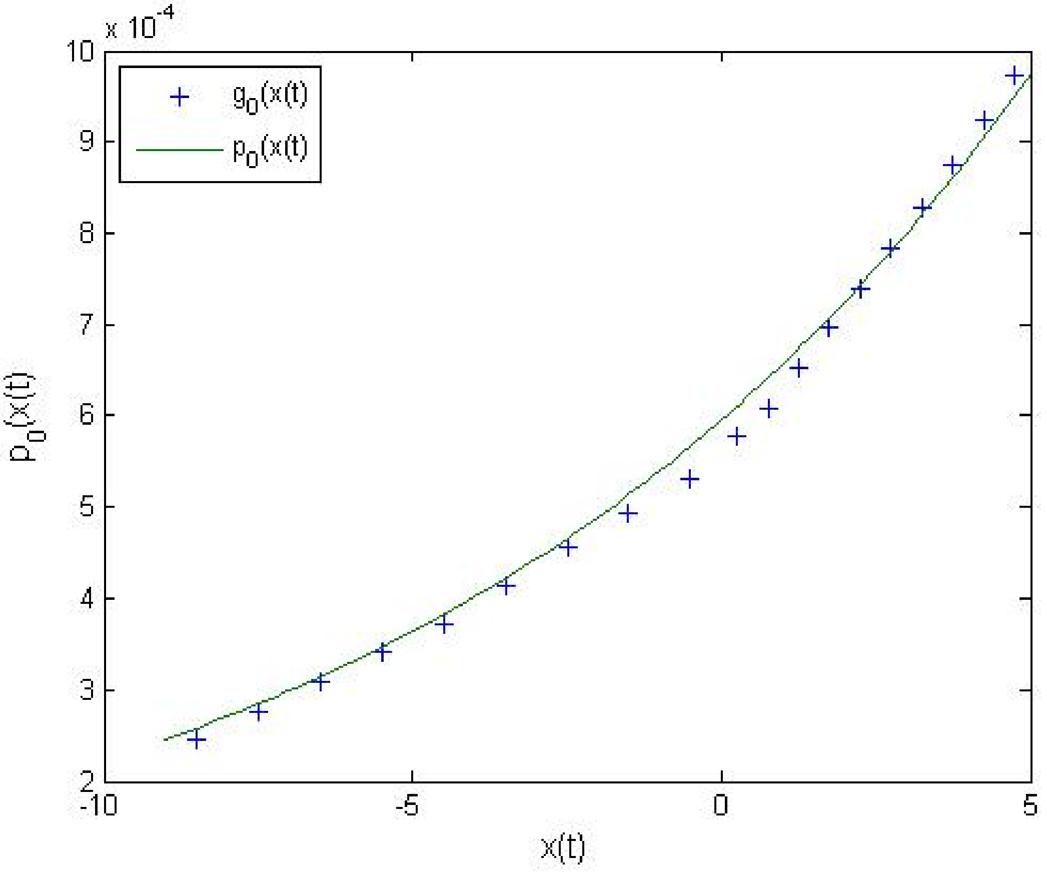


Figure 4.11 Comparison of $p_0(x)$ with $g_0(x)$.

Figure 4.11 and Figure 4.12 illustrate a possible equivalence between our three mode machine model and its aggregation into a two state modified Bielecki-Kumar machine for adequate sets of parameters such as shown in Table 4.3 below. Interest in these results is twofold:

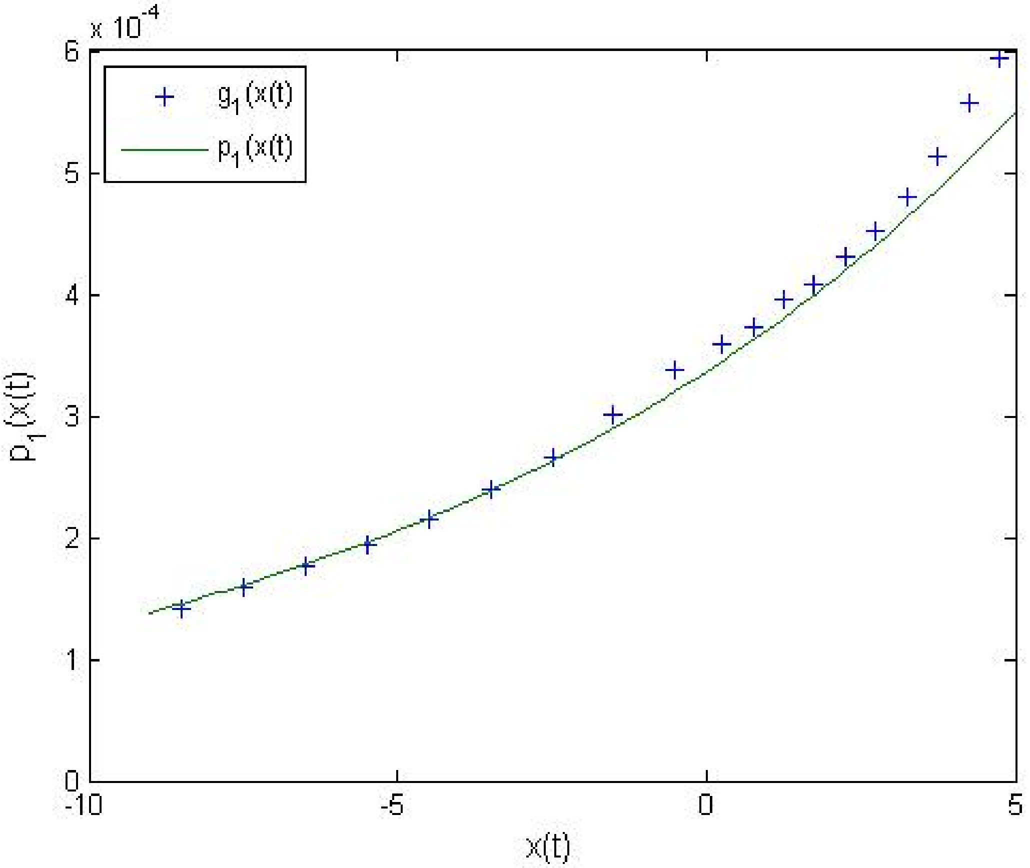


Figure 4.12 Comparison of $p_1(x)$ with $g_1(x)$.

Table 4.3 The system parameters

r	p	g	f	k	d	c^+	c^-
0.1	0.001	1	0.01	2.8	1	1	10

(i) Since hedging policies are optimal for Bielecki-Kumar machines, they also appear to be a good candidate for suboptimal policies in the case of the three-mode machine model; (ii) The Bielecki-Kumar optimal inventory level has a known theoretical expression which can become the basis of an initialization scheme in the numerical search an optimal hedging level (optimal value of z) in the three-mode machine model. Based on the Bielecki-Kumar theory modified to account for the existence of a stream of non conforming parts, one can write the following expressions for initial estimates of the optimal hedging level and the associated optimal cost estimate (see (Mhada *et al.* (2011))):

$$\begin{aligned} z^* = &\quad 0 \quad \text{if} \quad \frac{\frac{k}{(1+\beta)}p(c^- + c^+(1+\beta))}{c^+(1+\beta)(\frac{k}{(1+\beta)} - d) * (p+r)} \leq 1 \\ &\quad \text{and} \quad \frac{\frac{k}{(1+\beta)} - d}{p} > \frac{d}{r}; \\ z^* = &\quad +\infty \quad \text{if} \quad \frac{\frac{k}{(1+\beta)} - d}{p} < \frac{d}{r} \\ z^* = &\quad \frac{(1+\beta)}{(\frac{r}{d} - \frac{p}{\frac{k}{(1+\beta)} - d})}; \\ &\quad \log(\frac{\frac{k}{(1+\beta)}p(c^- + c^+(1+\beta))}{c^+(1+\beta)(\frac{k}{(1+\beta)} - d)(p+r)}) \quad \text{otherwise.} \end{aligned} \tag{4.51}$$

– For $z^* = 0$:

$$J^* = \frac{c^- \, p \frac{k}{(1+\beta)} d}{(p+r)(r\frac{k}{(1+\beta)} - rd - pd)}. \tag{4.52}$$

– For $z^* > 0$:

$$\begin{aligned} J^* = &\quad \frac{c^+(1+\beta)d}{r+p} + c^+(1+\beta)(r/d) - \\ &\quad (\frac{p}{(\frac{k}{(1+\beta)} - d)}) log(\frac{\frac{k}{(1+\beta)}p(c^+(1+\beta) + c^-)}{c^+(1+\beta)(r-d)(r+p)}), \end{aligned} \tag{4.53}$$

where $\beta = \frac{f}{p+g}$.

4.8 Conclusions

We have revisited optimal production rules in an unreliable (subject to failures) manufacturing system under a constant demand for parts and which, as part of its normal operation, produces a mix of conforming and non conforming parts. The non conforming parts are produced whenever the machine enters a non absorbing poor quality state at random times. While the optimal control law may not in general remain a hedging policy, based on an aggregated two-mode associated B-K machine which often acts as as a good approximation, hedging policies are deemed to be a good class of suboptimal policies.

The steady-state Kolmogorov equations together with their boundary conditions have been developed under the class of hedging policies and a numerical scheme for their solution has been proposed, and subsequently validated through Monte-Carlo simulations based on ARENA V.10 software . This numerical scheme can be integrated as part of a hedging policy numerical optimization scheme. Furthermore, the search for the optimal hedging policy can be initialized via the theoretical optimum as given by the B-K theory.

In future work, we shall attempt to use the three-mode model as a buiding block in a transfer line approximate decomposition method, aimed at jointly optimizing buffer sizing and inspection station positioning.

CHAPITRE 5

Unreliable production lines with defective parts and inspection stations

Article soumis pour publication au journal *IIE Transactions* (Août 2011) et écrit par:

Fatima Zahra Mhada

École Polytechnique de Montréal and GERAD

Roland P. Malhamé

École Polytechnique de Montréal and GERAD

Robert Pellerin

École Polytechnique de Montréal and CIRRELT

5.1 Introduction

Quality and quantity modeling in manufacturing systems have long been studied separately, even though they are highly coupled issues. In the literature, it is commonly assumed that product quality is perfect so that the impact of quality failures on the design of production policies is ignored.

Both quantity and quality modeling have the same general objectives: to minimize production cost and to maximize productivity. However, productivity, as analyzed from the quality point of view, requires early detection of quality failures, and therefore dictates lower buffer levels if quality inspections are not carried out at every stage of the production process. Conversely, low buffer sizes may make the production line more vulnerable to machine failures and thus decrease its overall productivity. Therefore, there is a need to integrate both quantity and quality modeling to achieve true optimization of the performance of production lines.

Despite the need for a greater integration of quality and production considerations, there are only a limited number of research results in this field: Kim and Gershwin (Kim et Gershwin (2005), Kim et Gershwin (2008)) study the relationship between quality issues and production policy issues by assuming that machines can enter an (unobservable) quality failure mode which is absorbing, until proper maintenance is carried out. Furthermore, although the assumed production model is fluid, quality remains in their model, an intrinsic property of discrete parts.

Colledani and Tolio (Colledani et Tolio (2005), Colledani et Tolio (2006a), Colledani et Tolio (2006b),Colledani et Tolio (2011)) consider a production system composed of unreliable manufacturing stations and inspection stations with different failure modes. Statistical quality control charts are introduced at inspection stations and act as noisy measurements on the quality state of the machines. Decomposition methods for the integrated production/quality performance studies of the line are developed. Their modeling framework is entirely discrete in the way time flows and in the production process itself.

Anily et Grosfeld-Nir (2006) address the issue of lot sizing and optimal lot inspection policy so as to meet a fixed size order of good parts at minimum cost. This model considers that a unit produced is either conforming or defective. Demand needs to be satisfied in full, by conforming units only. The production process may switch from a “good” state to a “bad” state, at constant rate and the true state is unobservable and can only be inferred from the quality of units inspected. The parts produced can be defective or conforming in both the good production state and the bad one, albeit with different probabilities. The authors prove that the optimal inspection policy has a simple form: continue units inspection only if the good-state-probability is above a problem parameters dependent threshold.

In the current paper, our objective is to develop an entirely fluid modeling framework for the integrated quality/production analysis and optimization of unreliable manufacturing transfer lines. In the model as presented here, machines are either in a good operational mode or a failure mode. Both modes are observable, and machines produce a mix of both conforming and non conforming parts in the operational mode. Statistical control is not considered at this stage.

The rest of the paper is organized as follows: In Section 5.2, our quality imperfect, unreliable transfer line model is presented and our optimization problem stated. In Section 5.3, we recall optimization results for two basic single machine models which act as building blocks in our decomposition methodology. In section 5.4, we extend the Sadr-Malhamé decomposition/aggregation approach for the approximate analysis of unreliable transfer lines, to the current framework. Section 5.5 is dedicated to a Monte Carlo simulation based evaluation of the accuracy of the approximate model of Section 5.4, while Section 5.6 extends the Sadr-Malhamé dynamic programming based algorithm for buffering optimization, to the current framework where positioning optimization for inspection stations is also required. Numerical results are reported. In Sections 5.7 we numerically evaluate the added benefits of completely joint versus partially joint optimization of buffer sizing and inspections station positioning in transfer lines, while in Section 5.8, we study the sensitivity of our optimal solutions to various problem parameters. Section 5.9 is our conclusion.

5.2 Modeling framework and statement of the optimization problem

We consider a single part type transfer line consisting of a series of n unreliable machines M_i separated by n buffers ($i = 1, ..., n$), and subjected to a fixed rate of demand for parts.

The machines can be in either one of two modes: an operational mode and a failure mode. Failures and repairs occur in continuous time according to a two-mode Markov chain. Both modes are observable. However, while in the operational mode, each machine can impart possible defects on the wip that it processes. For simplicity, we consider that the ratio of wip processed by machine M_i with locally imparted defects, to that of processed wip without such locally imparted defects is a constant, machine dependent known fraction β_i, at M_i, $i = 1, ..., n$.

The production line can contain inspection stations located at the exit of wip x_i and the provisioning point for machine M_{i+1}. The presence or absence of these stations is captured by a binary variable λ_i with:

$$\lambda_i = \begin{cases} 1 & \text{if there is an inspection station at the exit of the stock } x_i \\ 0 & \text{otherwise} \end{cases}$$

The following notations and assumptions will be used in the development of our models; for machine M_i, $i = 1, ..., n$:

- α_i : the mode, respectively 1 if machine is operational, and 0 if failed;
- p_i : failure rate;
- r_i : repair rate;
- k_i : Machine production capacity. It is assumed that $k_1 \geq k_2 \geq ... \geq k_n$. This condition guarantees that when both machines M_i and M_{i+1} are neither starved nor blocked, the rate of production of machine M_{i+1} will never be limited above by that of M_i;
- $u_i(t)$: production rate; with $0 \leq u_i(t) \leq k_i$;
- It is assumed that instantaneous wip $x_i(t)$ is a completely homogeneous mixture of good parts denoted $x_{i_1}(t)$ and bad parts denoted $x_{i_2}(t)$. $q_i(t)$ denotes the ratio at time t of bad parts to good parts within wip $x_i(t)$. Note that $q_i(t)$ will in general depend on β_1, β_2,...,β_i and furthermore: $x_i \triangleq x_{i_1} + x_{i_2} \triangleq (1 + q_i)x_{i_1}$. In subsection 5.4.3 below, we show that q_i is a constant which is a function of β_1, β_2, ...,β_i and λ_1, λ_2,...,$\lambda_{(i-1)}$;
- z_i : the inventory level for x_i;
- c_p : the storage cost per time unit and per part;
- c_n : the shortage cost per time unit and per part;

- c_I : the inspection cost per pulled part ;
- It is assumed that at inspection stations, all parts are inspected and furthermore, inspection results are fully reliable. Also, all finished parts are verified at an inspection station;
- Machines M_i, $i = 1, ..., n-1$, cannot have any backlog;
- Machine M_n is operated under either one of the following two constraints:

  1. A finite buffer for finished parts, but unlimited capacity for backlogging demand (this relates to the Bielecki-Kumar optimization model (Bielecki et Kumar (1988)));
  2. Unfulfilled demand cannot be backlogged at M_n; however, the long term probability of availability of conforming finished parts is constrained to be greater than or equal to some fixed value (service level);
- Machine M_1 is never starved;
- The line must satisfy a demand rate d from the stock of good parts x_{n_1}, which means that it must satisfy $(1+q_n)\,d$ from the total stock x_n;
- Wip $x_i(t)$ and finished parts inventory $x_n(t)$ evolve respectively according to:

$$\frac{dx_i(t)}{dt} = u_i(t) - u_{i+1}(t) \quad \text{for} \quad i = 1, ..., (n-1) \quad \text{with} \quad x_i(t) \geq 0; \tag{5.1}$$

$$\frac{dx_n(t)}{dt} = u_n(t) - (1 + q_n(t))\,d; \tag{5.2}$$

- $\tilde{d}_i$, $i = 1, ..., n$, is the long term average number of parts pulled per unit time from the total stock $x_i(t)$. The average unit time cost caused by the existence of defective parts will be proportional to $\tilde{d}_i$, more specifically equal to $c_I\,\tilde{d}_i\,\lambda_i$.

As mentioned above, one inspection station is placed at the end of the transfer line, as a result guaranteeing that parts delivered to the customer are all conforming. However, we are considering the problem of adding another inspection station within the transfer line. Thus the optimization problem studied here is that of the joint placement of an extra inspection station and the sizing of buffer spaces within the transfer line so as to minimize the long term per unit time average global cost of storage, production shortages, and inspection. More precisely, the cost to be minimized over buffer size parameters and inspection machine position is:

$$J_{Tz,\lambda}(x_0^T, \alpha_0^T) \;=\; \lim_{T\to\infty} \frac{1}{T} \Big(\sum_{i=1}^{n-1} E \left[\int\limits_0^T (c_p\, x_i(t) \;+\; c_I\, \tilde{d}_i\, \lambda_i) dt / (x_i(0), \alpha_i(0)) \right] \qquad (5.3)$$

$$+ \;\; E \left[\int\limits_0^T (c_p x_n^+(t) \;+\; c_n\, x_{n_1}^-(t) \;+\; c_I\,(1+q_n)\, d\, \lambda_n)/(x_n(0), \alpha_n(0)) dt \right] \Big)$$

under conditions: $\sum_{i=1}^{n-1} \lambda_i = 1$ and $\lambda_n = 1$ with $x^+(t) = max(x(t), 0)$, $x^-(t) = max(-x(t), 0)$, $x_0 = [x_1(0), ..., x_n(0)]^T$, $\alpha_0 = [\alpha_1(0), ..., \alpha_n(0)]^T$, $z^T = [z_1, z_2, ..., z_n]^T$ and $\lambda^T = [\lambda_1, \lambda_2, ..., \lambda_{n-1}]$. The above formulation will be referred to as the *combined storage-shortage cost minimization problem.* A second optimization problem is one where backlog is *not* allowed. However a service level constraint is imposed dictating that the coefficient of availability of conforming finished products be some fixed desired number a_n^{des}; this, under the additional constraint that the long term average rate of extraction of conforming finished parts be d. Under these constraints, the cost to be minimized is the same as (5.3), with $c_n = 0$. This second formulation will be referred to as *the storage cost minimization problem under service level constraints.*

5.3 Single machine building blocks for transfer line decomposition

In the following, we review analytical and optimization results for two basic single machine models which will serve as building blocks for our transfer line decomposition methodology. In one model (the Bielecki et Kumar (1988) or BK model) backlog is allowed, while in the other (the Hu model (Hu (1995))), no backlog is allowed.

Both models involve a single machine producing a single part type and attempting to respond to a constant demand rate d. The machine state changes in continuous time according to a homogeneous Markov process: the state changes from down ($\alpha(t) = 0$) to up ($\alpha(t) = 1$) at a rate r and from up to down at a rate p. When the machine is up, it can produce at any rate $u(t)$ between zero and a maximum rate k. Storage cost per part per unit time is c_p. The production inventory at time $x(t)$ is characterized by the following differential equation: $\frac{dx(t)}{dt} = u(t) - d$.

In what follows, we will recall both the BK model and the Hu model. Thereafter, we present a BK modified model with the integration of the concept of quality.

5.3.1 The BK model

In the BK model (Bielecki et Kumar (1988)), under the assumption that backlog is allowed, the objective is to minimize the following long run average cost, where c_n is the backlog cost per unit part and unit time:

$$J_T(x(t), \alpha(t)) = E\left[\int_0^T (c_p\, max(x(t), 0) \,+\, c_n\, max(-x(t), 0))/(x(0), \alpha(0))\, dt\right]. \tag{5.4}$$

The optimal control policy is characterized by a single critical inventory level called a hedging point z^*, that the production system must maintain as long as possible a kind of insurance policy against the potential cost of shortages following machine failures. More analytical details will be provided in the quality related extension of the model below.

5.3.2 Hu model

Hu in (Hu (1995)), studies the same production system, under the assumption that backlog is not allowed. The objective is to minimize the sum of the inventory cost and the shortage cost for unsatisfied demand incurred as long as the non negative inventory is zero. This long run average cost is:

$$J_T(x(t), \alpha(t)) = \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T (c_p\, x(t) \,+\, c_n\, I(x(t) = 0,\, \alpha(t) = 0))dt/(x_i(0), \alpha_i(0))\right], \tag{5.5}$$

where c_n is the shortage cost per unit time, and $I(.)$ is the indicator function.
Hu shows that the optimal control policy is still characterized by a single critical inventory level called hedging point z, and the optimal production rate satisfies:

$$u(t) = \begin{cases} k & \text{if} \quad x(t) < z; \\ d & \text{if} \quad x(t) = z; \\ 0 & \text{if} \quad x(t) > z. \end{cases}$$

Finally, the long-run average cost associated with an arbitrary hedging point z has the form:

$$\begin{aligned} J(z) = \quad & \frac{\rho}{(p+r)(1-\rho\,\exp(-\mu(1-\rho)z))}\,[(c_p\,\frac{k\,(1-\exp(-\mu(1-\rho)\,z))}{1-\rho} \\ & -(p+r)\,z\,\exp(-\mu(1-\rho)\,z)) + \,c_n\,(1-\rho)\,p] \end{aligned} \tag{5.6}$$

with: $\rho = \frac{r(k-d)}{p\,d}$, $\mu = \frac{p}{(k-d)}$ and the coefficient of availability of wip, i.e., the steady-state probability that the wip is available is:

$$a = 1 - \frac{p}{(p+r)}\frac{(1-\rho)}{(1-\rho\,\exp(-\mu(1-\rho)\,z))}. \tag{5.7}$$

5.3.3 The BK quality modified model

As is the case for the ordinary BK model, machine M has a discrete state $\alpha(t)$ with two possible values: $\alpha(t) = 1$ (M operational), and $\alpha(t) = 0$ (M has failed). The difference here though is that when M is operational it produces a *mixed parts flow* including both conforming (good) parts and non conforming (defective) parts. It is assumed that the ratio between defective parts and good parts is a constant β. We designate by $x_1(t)$ the inventory of good parts when non negative, and the backlog of good parts otherwise, while $x_2(t)$ and $x(t)$ respectively designate the inventory of defective parts, and the total inventory of parts. Note that $x_2(t)$ cannot become negative; also, according to our model, $x_1(t)$ and $x_2(t)$ must reach zero at the same time.

$$x_2(t) \;=\; \beta\,x_1(t)\,I[x_1(t) \geq 0]; \tag{5.8}$$

$$x(t) \;=\; x_1(t) + x_2(t) = (1+\beta)\,x_1(t) \tag{5.9}$$

where in the above, $I[.]$ is the indicator function.

Finally, when $x(t)$ is negative, it is equal to $x_1(t)$.

The rate of demand for good parts is a constant d, because the mixture of conforming and non conforming parts is assumed to be perfectly homogeneous, the demand rate for total parts is $(1+\beta)\,d$.

Parallel to the ordinary BK model analysis, our objective is to determine the optimal feedback production control policy $\{u_f\}$ minimizing the following long term measure of combined storage and backlog costs:

$$J_{T\{u_f\}} = \lim_{T\to\infty} \frac{1}{T} E \left[\int_0^T \big(c_p\, x^+(t) \;+\; c_n\, x_1^-(t) \big)\, dt \right] \tag{5.10}$$

with $x^+(t) = max(x(t), 0)$, $x_1^-(t) = max(-x_1(t), 0)$.

The cost to be minimized in (5.10) can now be rewritten in terms of the $x(t)$ variable as:

$$J_{T\{u_f\}} = \lim_{T\to\infty} \frac{1}{T} E \left[\int_0^T (c_p\, x^+(t) \;+\; \frac{c_n}{(1+\beta)}\, x^-(t))\, dt \right] \tag{5.11}$$

with the $x(t)$ dynamics:

$$\frac{dx(t)}{dt} = u(t) I\left[\alpha(t) = 1\right] - (1+\beta)\, d. \tag{5.12}$$

As such, it becomes an ordinary BK problem problem with a unit backlog storage cost *decreased to* $c_n^* = \frac{c_n}{(1+\beta)}$ and a constant demand rate *increased to* $d^* = (1+\beta)\, d$. This allows us to conclude that the optimal policy is again a hedging policy with values of z^* and optimal cost J^* specified below, and obtained based on the expressions in (Bielecki et Kumar (1988)).

Optimal hedging point (total finished parts) z^*:

$$\begin{aligned} z^* &= 0 \quad \text{if} \quad \frac{k\, p(\frac{c_n}{(1+\beta)} + c_p)}{c_p(k - (1+\beta)\, d)(p+r)} \leq 1 \quad \text{and} \quad \frac{k - (1+\beta)\, d}{p} > \frac{(1+\beta)\, d}{r}; \\ z^* &= \infty \quad \text{if} \quad \frac{k - (1+\beta)\, d}{p} < \frac{(1+\beta)\, d}{r}; \\ z^* &= \frac{1}{\frac{r}{(1+\beta)\, d} - \frac{p}{(k-(1+\beta)\, d)}} \ln(\frac{k\, p(\frac{c_n}{(1+\beta)} + c_p)}{c_p(k - (1+\beta)\, d)(p+r)}) \quad \text{otherwise.} \end{aligned} \tag{5.13}$$

Optimal cost J^*:

$$J(z^*) = \frac{c_n\, p\, k\, d}{(p+r)(r\, k - r\,(1+\beta)\, d - p\,(1+\beta)\, d)} \quad \text{if} \quad z^* = 0; \tag{5.14}$$

$$J(z^*) = \frac{c_p\,(1+\beta)\, d}{r+p} + \frac{c_p}{\frac{r}{(1+\beta)\, d} - \frac{p}{k-(1+\beta)\, d}} \ln \frac{k\, p(\frac{c_n}{(1+\beta)} + c_p)}{c_p\,(k - (1+\beta)\, d)(p+r)}) \quad \text{if} \quad z^* > 0.$$

For further details and analysis of this model, please refer to (Mhada *et al.* (2011)).

5.4 Flow line decomposition

Decomposition techniques play an important role in the analysis of performance for a given choice of buffer sizes in a transfer line, thus paving the way for parameter optimization. Thanks to such decomposition techniques, a line of n machines is approximately decomposed into n separate machines. To do this, the influence of the universe upstream and downstream of each machine shall be approximately represented. The decomposition technique used in this work is that presented in (Sadr et Malhamé (2004b)). This technique is based on two strategies of approximation:

- The hypothesis of decoupling of any given machine mode, and the binary activity state of the buffer supplying that machine. This assumption is referred to as the *machine decoupling approximation*;
- The demand averaging principle.

5.4.1 The combined wip supply / machine building blocks

The decomposition/aggregation methodology of (Sadr et Malhamé (2004b)) provides a tractable approximation for performance evaluation of failure prone transfer lines under the class of Kanban policies. Furthermore, because of the peculiar resulting unidirectional causality propagation (from upstream to downstream), it allows one to define sequential decision stages (sequential buffer sizing). As a result, dynamic programming becomes an easily implemented optimization tool. The approximation method is based on building blocks involving the aggregation of an on-off process of wip availability, say the indicator function of active x_i, and the on-off machine it supplies, say M_{i+1}, into what is called a pseudo-machine, say $\tilde{M}_{i+1}$ (see Figure 5.1). This virtual machine is never starved, but is more likely to fail than the original one because its failures must account for starvation phenomena. In effect, $\tilde{M}_{i+1}$ is an aggregate representation of the complete transfer line up to machine M_{i+1}, as it appears viewed from the rest of the transfer line downstream. In Figure 5.1, $\tilde{M}_{i+1}$ is a machine with state $\tilde{\alpha}_{i+1}$ which can be zero (failed) or 1 (operational) with respectively repair rate $\tilde{r}_{i+1}$ and failure rate $\tilde{p}_{i+1}$. Using the machine decoupling approximation, it is obtained as the

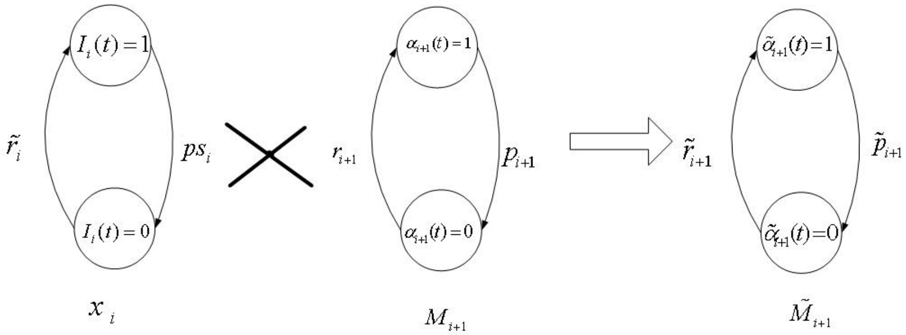


Figure 5.1 The pseudo-machine $\tilde{M}_{i+1}$, $i = 1, ..., (n-1)$

Cartesian product of the Markov chain associated with machine M_{i+1} and that associated with the x_i supply availability indicator $I_i(t)$ ($I_i(t) = 1$ indicates that $x_i(t)$ is active). This four state machine is subsequently collapsed into a further simplified two-state machine as shown to the right of Figure 5.1.

$\tilde{M}_{i+1}$ represents the machines upstream of given buffer $i+1$. Following (Sadr et Malhamé (2004b)) which was initially developed for transfer lines with *only conforming parts*, ps_i, $\tilde{r}_i$ and $\tilde{p}_i$ are given by the following expressions for $i = 1, ..., n$:

$$ps_i = \tilde{r}_i \frac{1-a_i}{a_i}; \tag{5.15}$$

$$\tilde{r}_i = \frac{(ps_{i-1} + p_i)\,\tilde{r}_{i-1}\,r_i}{p_i\,\tilde{r}_{i-1} + ps_{i-1}\,r_i}; \tag{5.16}$$

$$\tilde{p}_i = (\frac{(ps_{i-1} + \tilde{r}_{i-1})\,(p_i + r_i)}{\tilde{r}_{i-1}\,r_i} - 1)\,\tilde{r}_i, \tag{5.17}$$

with: $\tilde{r}_1 = r_1$, $\tilde{p}_1 = p_1$ and where a_i is the coefficient of availability of wip at buffer i, i.e., the steady-state probability that wip is available at buffer i.

The calculation of availability coefficient a_i is based on the so-called demand averaging principle, which can be applied only if the transfer line demand rate is a known constant (under our assumptions, it is the case). However, in the current quality aware context, the above calculation must be adjusted so as to account for the presence of both conforming and non conforming parts, as well as that of inspection stations internal to the line. Details will be given in the next subsection as we discuss the all important demand averaging principle.

The condition of feasibility of demand must be satisfied by each pseudo-machine $\tilde{M}_i$: If $\tilde{M}_i$

is able to meet demand in the long run, then it is necessary that its long term average production rate when always operated at its current full capacity exceed the long term average demand i.e.:

$$\frac{\tilde{r}_i\, k_i}{\tilde{r}_i + \tilde{p}_i} > \tilde{d}_i \tag{5.18}$$

5.4.2 The demand averaging principle (DAP)

The demand averaging principle (DAP) (Sadr et Malhamé (2004b)) is used to approximate the effect of the machines downstream of a given buffer i. It is based on recognizing that for a transfer line with *perfect quality*, if finished parts are being pulled at long term rate d, then *all buffers* in the transfer line will be subjected to that same long term rate of extraction d. Thus rate d is an *invariant* across the transfer line. This observation (Sadr et Malhamé (2004b)) must however be reassessed in the current context of mixed conforming and non conforming parts. Indeed, if there are no inspection stations internal to the line, then while d is the rate at which conforming parts are being extracted, it is actually the *total rate*, both conforming and non conforming parts, at which parts are being extracted from the finished parts buffer that becomes the line invariant. This rate is in fact $(1 + q_n)d$. Given this rate invariance property, the demand averaging principle is the approximation by virtue of which the actual complex stochastic parts extraction process from any buffer in the line is replaced by the simplest process, namely a *constant*, *consistent* with that rate constraint. As a result, and according to DAP, buffer i is considered to be subjected to a constant rate of total parts extraction of value $\frac{\tilde{d}_i}{a_i} = \frac{(1+q_n)d}{a_i}$ *while it is active* and where it is recalled that a_i is the a priori unknown total parts wip availability coefficient at buffer i.

However the presence of an inspection station (see Figure 5.2) will complicate things since

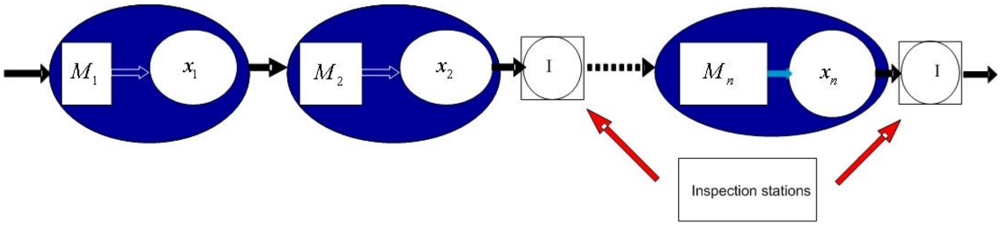


Figure 5.2 Flow line with inspection stations.

an inspection station at the position j, ($\lambda_j = 1$) divides the line in two parts: The part of the line upstream of an inspection station and that downstream of it.

While it is straightforward to include more than one inspection station within the line, the focus here is on the case of a single internal inspection station placement.

Downstream of the inspection station, buffers when active must meet the demand rate : $\frac{\tilde{d_i}}{a_i} = \frac{(1+q_n)d}{a_i}$, $i > j$, while the upstream part of the line should provide a rate $(1 + q_n)d$ of *good* parts to the downstream part i.e. a rate $(1 + q_n)\,(1 + q_j)$ of total parts. Thus buffers upstream of the inspection station, when active, must meet the long term average demand rate : $\frac{\tilde{d_i}}{a_i} = \frac{(1+q_n)\,(1+q_j)\,d}{a_i}$, $i \leq j$. We note that whatever the position of the inspection station in the line ($\forall j = 1, ..., (n-1)$), the demand rate $\tilde{d_i} = (1 + q_n)\,(1 + q_j)\,d$, $i \leq j$ is constant and greater than the demand rate $\tilde{d_i} = (1 + q_n)d$, $i > j$, also constant.

Again summarizing, of all stochastic processes consistent with the above buffer rates of parts extraction constraints, DAP is the approximation whereby the actual rates are taken to be constant while a buffer is active.

5.4.3 Recursive calculation of the q_i's

x_{i_2} is the part of the total stock x_i consisting of non conforming parts and x_{i_1} corresponds to the good / conforming parts. We assume that:

$$\begin{aligned} x_{i_2} &= q_i\, x_{i_1}; && (5.19)\\ x_i &= x_{i_1} + x_{i_2} = (1 + q_i)\, x_{i_1}; && (5.20)\\ x_i &= \frac{(1+q_i)}{q_i}\, x_{i_2}. && (5.21) \end{aligned}$$

Our aim is to calculate $q_{(i+1)}$, the ratio of non conforming to conforming parts in $x_{(i+1)}$. More specifically, $x_{(i+1)_2} = q_{(i+1)}\, x_{(i+1)_1}$. Indeed, only a fraction of the stock of good parts that will be processed by machine $M_{(i+1)}$ remains conforming. This fraction is $\frac{1}{1+\beta_{(i+1)}}$; so for a small time increment δt, the incoming quantity of good parts stored in buffer $i + 1$ is equal to $\frac{1}{1+\beta_{(i+1)}}$ times the quantity of the stock x_{i_1} processed during δt, say δx_{i_1}. The rest of δx_{i_1}, *i.e.* $\frac{\beta_{(i+1)}\,\delta x_{i_1}}{1+\beta_{(i+1)}}$, will be stored in buffer $i + 1$ as non conforming parts. Also, during the same time interval δt, machine M_{i+1} will process a quantity δx_{i_2} of already non conforming parts within wip i, associated with δx_{i_1} and given as $q_i\,\delta x_{i_1}$, unless an inspection station is present at buffer i ($\lambda_i = 1$), in which case δx_{i_2} is rejected, and only good parts are processed by M_{i+1}. If no inspection station is present at buffer i, δx_{i_2} persists as non conforming in wip $i+1$. The resulting net ratio of non conforming to conforming parts within wip $i+1$ is then given by the following recursive calculation:

$$\begin{aligned} q_{i+1} &= \frac{(1-\lambda_i)\,\delta x_{i_2} + \frac{\beta_{i+1}}{(1+\beta_{i+1})}\,\delta x_{i_1}}{(\frac{1}{(1+\beta_{i+1})})\,\delta x_{i_1}} \\ &= \frac{((1-\lambda_i)\,q_i)\,\delta x_{i_1} + \frac{\beta_{i+1}}{(1+\beta_{i+1})}\,\delta x_{i_1}}{(\frac{1}{(1+\beta_{i+1})})\,\delta x_{i_1}} \\ &= (1-\lambda_i)\,q_i\,(1+\beta_{i+1}) + \beta_{i+1}. \end{aligned} \tag{5.22}$$

It is initialized by $q_1 = \beta_1$.

5.4.4 Evaluation of internal buffer i induced costs

The cost of storage in buffer i is given by:

$$J^i_T(z_i, \lambda_i) = \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T (c_p\, x_i(t)\,)dt\right]. \tag{5.23}$$

Under DAP, when active, buffer i is subjected to a constant, coefficient of availability dependent demand, so that wip evolves according to:

$$\frac{dx_i(t)}{dt} = u_i(t) - \frac{\tilde{d}_i}{a_i} \tag{5.24}$$

$$\text{with} \quad u_i(t) = \begin{cases} k_i & \text{if} \quad x_i(t) < z_i; \\ \frac{\tilde{d}_i}{a_i} & \text{if} \quad x_i(t) = z_i; \\ 0 & \text{if} \quad x_i(t) > z_i. \end{cases}$$

The cost and dynamics under (5.24) characterize a Hu machine (Hu (1995)) subject to a constant, a priori unknown, demand $\frac{\tilde{d}_i}{a_i}$. Based on (5.7), and the pseudo-machine $\tilde{M}_{i+1}$

obtained in Subsection 5.4.1, one can write:

$$a_i = 1 - \frac{\tilde{p}_i}{(\tilde{p}_i + \tilde{r}_i)} \frac{(1-\rho_i)}{(1-\rho_i\ \exp(-\mu_i(1-\rho_i)\, z_i))} \tag{5.25}$$

with: $\rho_i = \frac{\tilde{r}_i(k_i - \frac{\tilde{d}_i}{a_i})}{\tilde{p}_i\, \frac{\tilde{d}_i}{a_i}}$, $\mu_i = \frac{\tilde{p}_i}{(k - \frac{\tilde{d}_i}{a_i})}$ and

$$\begin{aligned} J_i(z_i, \lambda_i) \quad = \quad & \frac{c_p \rho_i}{(\tilde{p}_i + \tilde{r}_i)(1-\rho_i\ \exp(-\mu_i(1-\rho_i) z_i))} \\ & \left[\frac{k_i\, (1 - \exp(-\mu_i(1-\rho_i)\, z_i))}{1-\rho_i} - \tilde{p}_i + \tilde{r}_i)\, z_i\ \exp(-\mu_i(1-\rho_i)\, z_i) \right] \end{aligned} \tag{5.26}$$

under constraints: $\sum_{i=1}^{n-1} \lambda_i = 1$ and $\lambda_n = 1$.

Note that for given $\tilde{p}_i$, $\tilde{r}_i$ and z_i, (5.25) constitutes an *implicit* equation for unknown coefficient of availability a_i (a fixed point is obtained by using iterated substitutions starting from initial guess $a_i^{(0)} = 1$ (Sadr et Malhamé (2004b))).

5.4.5 Evaluation of buffer n induced costs

2 distinct models will be used for the analysis of buffer n induced costs: The first model corresponds to a situation where backlog is allowed, but negative excursions are penalized at a cost of c_n per part and unit time. This is the so-called *combined storage-backlog cost minimization problem.* Calculations are carried out using a version of the BK theory (Bielecki et Kumar (1988)), modified to account for the presence of both conforming and non conforming parts; the second model corresponds to a situation where backlog is not allowed. This is the so-called *storage cost minimization problem under service level constraints.* Calculations are carried out using the Hu theory with parameters dependent on the imposed service level constraint.

Combined storage-shortage cost minimization problem

Based on the BK theory (Bielecki et Kumar (1988)) and our previous work (Mhada *et al.* (2011)), the following expressions can be written for peudo-machine $\tilde{M}_n$ associated with storage as well as backlog costs:

$$\begin{aligned} J_T^n(z_n, \lambda_n) &= \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T (c_p x_n^+(t) \,+\, c_n x_{n_1}^-(t))dt\right] \quad (5.27)\\ &= c_p\, z_n + \frac{\tilde{p}_n\, k_n}{\delta\,(\tilde{p}_n + \tilde{r}_n)\,(k_n - \tilde{d}_n)}\,(\frac{c_n}{1+q_n}\ \exp(-\delta\, z_n) - c_p\,(1-\exp(-\delta\, z_n))) \end{aligned}$$

with $\delta = \frac{\tilde{r}_n}{\tilde{d}_n} - \frac{\tilde{p}_n}{k_n - \tilde{d}_n} > 0$ and $\tilde{d}_n = (1+q_n)\,d$.

The optimal hedging point (total finished parts) z^* is :

$$\begin{aligned} z^* &= 0 \quad \text{if} \quad \frac{k_n\,\tilde{p}_n(\frac{c_n}{(1+q_n)} + c_p)}{c_p(k_n - \tilde{d}_n)(\tilde{p}_n + \tilde{r}_n)} \le 1 \quad \text{and} \quad \frac{k_n - \tilde{d}_n}{\tilde{p}_n} > \frac{\tilde{d}_n}{\tilde{r}_n}; \quad (5.28)\\ z^* &= \infty \quad \text{if} \quad \frac{k_n - \tilde{d}_n}{\tilde{p}_n} < \frac{\tilde{d}_n}{\tilde{r}_n};\\ z^* &= \frac{1}{\frac{\tilde{r}_n}{\tilde{d}_n} - \frac{\tilde{p}_n}{(k_n - \tilde{d}_n)}}\ \ln(\frac{k_n\,\tilde{p}_n(\frac{c_n}{(1+q_n)} + c_p)}{c_p(k_n - \tilde{d}_n)(\tilde{p}_n + \tilde{r}_n)}) \quad \text{otherwise} \end{aligned}$$

and the optimal cost J^* is:

$$\begin{aligned} J(z^*) &= \frac{c_n\,\tilde{p}_n\,k_n\,\tilde{d}_n}{(\tilde{p}_n + \tilde{r}_n)(\tilde{r}_n\,k_n - \tilde{r}_n\,\tilde{d}_n - \tilde{p}_n\,\tilde{d}_n)} \quad \text{if} \quad z^* = 0;\\ J(z^*) &= \frac{c_p\,\tilde{d}_n}{\tilde{r}_n + \tilde{p}_n} + \frac{c_p}{\frac{\tilde{r}_n}{\tilde{d}_n} - \frac{\tilde{p}_n}{k_n - \tilde{d}_n}}\ \ln\frac{k_n\,\tilde{p}_n(\frac{c_n}{(1+q_n)} + c_p)}{c_p\,(k_n - \tilde{d}_n)(\tilde{p}_n + \tilde{r}_n)}) \quad \text{if} \quad z^* > 0. \quad (5.29) \end{aligned}$$

Storage cost minimization under service level constraints

Based on the Hu theory (Hu (1995)), and for a given desired coefficient of availability of conforming finished parts a_n^{des}, one can write the constraint:

$$a_n^{des} = 1 - \frac{\tilde{p}_n}{(\tilde{p}_n + \tilde{r}_n)}\frac{(1-\rho_n)}{(1-\rho_n\ \exp(-\mu_n(1-\rho_n)\,z_n))}. \quad (5.30)$$

The required hedging point $z_n(a_n^{des})$ can be readily calculated using (5.30), in terms of (yet to be designed) variables $\tilde{r}_n$ and $\tilde{p}_n$ as:

$$z_n(a_n^{des}) = \frac{1}{-\mu_n(1-\rho_n)}\ \ln\left[\frac{1}{\rho_n}\left(1 - \frac{(1-\rho_n)}{(1-a_n^{des})\,(\frac{(\tilde{p}_n + \tilde{r}_n)}{\tilde{p}_n})}\right)\right] \quad (5.31)$$

with: $\rho_n = \frac{\tilde{r}_n(k_n - \frac{\tilde{d}_n}{a_n^{des}})}{\tilde{p}_n \frac{\tilde{d}_n}{a_n^{des}}}$, $\mu_n = \frac{\tilde{p}_n}{(k_n - \frac{\tilde{d}_n}{a_n^{des}})}$
and the corresponding cost is:

$$\begin{aligned} J_n(a_n^{des}, \tilde{p}_n, \tilde{r}_n) \;\; &= \;\; \frac{\rho_n \, c_p \, \frac{k_n \, (1-\exp(-\mu_n (1-\rho_n) \, z_n(a_n^{des})))}{1-\rho_n}}{(\tilde{p}_n + \tilde{r}_n)(1 - \rho_n \, \exp(-\mu_n(1-\rho_n) z_n(a_n^{des})))} \\ &-\frac{\rho_n \, c_p \, (\tilde{p}_n + \tilde{r}_n) \, z_n(a_n^{des}) \, \exp(-\mu_n(1-\rho_n) \, z_n(a_n^{des}))}{(\tilde{p}_n + \tilde{r}_n)(1 - \rho_n \, \exp(-\mu_n(1-\rho_n) z_n(a_n^{des})))}. \end{aligned} \tag{5.32}$$

5.5 Approximate model validation

In a preliminary verification step, the approximate theoretical expressions derived in Section 4 have been validated against the results of Monte Carlo simulations for a large sample of four machine lines, and the results differed by at most 4%.

With this validation behind us, our objective now is to determine buffer sizes and the one extra inspection station location which solve the combine storage-shortage cost minimization problem. So in the next section we present a dynamic programming based optimization method inspired by the work of (Sadr et Malhamé (2004b)) as adapted it to the current problem at hand.

5.6 Optimization

The objective is to find all z_i and the only $\lambda_i \neq 1$, for $i = 1, ..., (n-1)$ with $\lambda_n = 1$, values that minimize either the combined storage-shortage cost, or the storage cost under service level constraints. It is possible to rewrite the optimization problem as a *dynamic programming* problem for a fixed choice of inspection station positions (λ vector fixed):

$$J^*(\lambda) = \underbrace{inf}_{a_i \in A_i(a_{i+1}, \lambda), \, i=1,...,(n-1)} \left(\sum_{i=1}^{n-1} T^{(i)}(\tilde{r}_i, \tilde{p}_i, a_i, q_i) + T_F(\tilde{r}_n, \tilde{p}_n, q_n) + c_I \sum_{i=1}^{n-1} \lambda_i \, \tilde{d}_i \right) \tag{5.33}$$

$$\begin{aligned} T^{(i)}(\tilde{r}_i, \tilde{p}_i, a_i, q_i) = \quad & c_p \left(\frac{k_i\, \tilde{p}_i}{\sigma_i (k_i - \frac{\tilde{d}_i}{a_i})\,(\tilde{r}_i + \tilde{p}_i)} - \frac{k_i\ (1-a_i)}{\sigma_i (k_i - \frac{\tilde{d}_i}{a_i})} - \frac{1}{\sigma_i}\left[\frac{1}{\frac{\tilde{d}_i}{a_i}} - \frac{(1-a_i)\,(\tilde{r}_i + \tilde{p}_i)}{\sigma_i (k_i - \frac{\tilde{d}_i}{a_i})\,\frac{\tilde{d}_i}{a_i}} \right] \right. \\ & \left. ln\left[\frac{\tilde{p}_i\, \frac{\tilde{d}_i}{a_i}}{\tilde{r}_i\,(k_i - \frac{\tilde{d}_i}{a_i})} - \frac{\sigma_i\, \tilde{p}_i\, \frac{\tilde{d}_i}{a_i}}{(\tilde{r}_i + \tilde{p}_i)\, \tilde{r}_i\,(1-a_i)} \right] \right), \ i = 1, ..., (n-1) \end{aligned}$$

with : $\sigma_i = \frac{(\tilde{p}_i + \tilde{r}_i)\, \frac{\tilde{d}_i}{a_i} - k_i\, \tilde{r}_i}{(k_i - \frac{\tilde{d}_i}{a_i})\, \frac{\tilde{d}_i}{a_i}}$.

For the combined storage-shortage cost minimization:

$$T_F(\tilde{r}_n, \tilde{p}_n, q_n) = \left(c_p\, z_n^* + \frac{\tilde{p}_n\, k_n}{\delta\,(\tilde{p}_n + \tilde{r}_n)\,(k_n - \frac{\tilde{d}_n}{a_n})} \left(\frac{c_n}{1+q_n}\ \exp(-\delta\, z_n^*) - c_p\,(1 - \exp(-\delta\, z_n^*)) \right) \right)$$

where the expression of z_n^* is given in (5.28), thus leading to the alternate cost expressions in (5.29).

For storage cost minimization under service level constraints:

$$T_F(\tilde{r}_n, \tilde{p}_n, q_n) = \frac{\rho_n\ \left[c_p\,(\frac{k_n\,(1 - \exp(-\mu_n (1-\rho_n)\, z_n(a_n^{des})))}{1-\rho_n} - (\tilde{p}_n + \tilde{r}_n)\, z_n(a_n^{des})\ \exp(-\mu_n (1-\rho_n)\, z_n(a_n^{des}))) \right]}{(\tilde{p}_n + \tilde{r}_n)(1 - \rho_n\ \exp(-\mu_n (1-\rho_n)\, z_n(a_n^{des})))}$$

where the expression of $z_n(a_n^{des})$ is given in (5.31); if j is the inspection station location then for $i = 1, ..., n$:

$$\begin{aligned} \tilde{d}_i &= (1+q_n)\, d \quad \text{if} \quad i > j; \\ \tilde{d}_i &= (1+q_n)\,(1+q_j)\, d \quad \text{if} \quad i \leq j. \end{aligned} \tag{5.34}$$

Note that, in (5.33), the decision variables are the coefficients of availability of parts (both conforming and non conforming), instead of the buffer sizes, because their range is bounded; the state space itself is two dimensional at each stage ($\tilde{r}_i$, $\tilde{p}_i$). Also note that buffer sizes z_i can be immediately calculated from the a_i's, once the latter have been obtained. Finally, note that this is a constrained dynamic programming problem in the sense that, at each stage, the range of permissible a_i's is dependent on the state at the next future stage ($\tilde{r}_{i+1}$, $\tilde{p}_{i+1}$). See (Sadr et Malhamé (2004b)) for further details on these constraint sets. In the following, we provide the details of our numerical optimization algorithm, and apply it for the analysis of a homogeneous transfer line.

5.6.1 Numerical procedure

Our numerical algorithm is summarized as follows:
For each choice of the positioning of the internal inspection station ($\lambda_j = 1$; $j = 1, ..., (n-1)$ and $\lambda_i = 0$, $i = 1, ..., n$, $i \neq j$), we first calculate the different values of q_i and $\tilde{d_i}$ based on (5.22) and (5.34).
For each fixed value of the λ vector, we solve a dynamic programming problem (similar to the one in (Sadr et Malhamé (2004b))) to determine the associated optimal buffer sizes, and minimal cost. This algorithm is deployed in two phases: (i) State space generation; (ii) Application of the dynamic programming algorithm.

(i) State space generation: Here, the decision variable a_i, $i = 1, ..., (n-1)$ is discretized between its lower bound defined by ergodicity condition (5.18) and its upper bound of 1. This discretization is then used to generate the sequence of discrete grid points in the two dimensional $[\tilde{r}_i, \tilde{p}_i]$ space, starting from the single point $\tilde{r}_1 = r_1$, $\tilde{p}_1 = p_1$, and using ((5.15)-(5.17)).

(ii) Application of the dynamic programming algorithm: This yields the optimal sequence of a_i decisions, the optimal trajectory on the sequence of $[\tilde{r}_i, \tilde{p}_i]$ planes, and the corresponding optimal buffer sizes z_i, for the given choice of the λ vector.

We add to the total cost of storage and possibly shortage, if the combined storage-shortage cost minimization version of the problem is considered, as calculated by dynamic programming, the inspection cost corresponding to the fixed value of the λ vector.
When all permissible configurations of the λ vector have been explored, we compare the resulting associated optimal costs to determine the position $\lambda_j = 1$, $j = 1, ..., (n-1)$, for which the cost is lowest. This position will represent the solution to our joint optimal buffer sizing and inspection station positioning problem.

5.6.2 A homogeneous machines line

We show an example of a dynamic programming problem solution. The line to be optimized is a 10 homogeneous machines line with $\beta_i = 0.1$, $p_i = 0.2$, $r_i = 0.9$ for $i = 1, ..., n$, with machine production capacity $k_i = 4$, demand rate $d = 1$, and storage, inspection and shortage costs given respectively as $c_p = 1$, $c_I = 2$ and $c_n = 10$. We will study both the combined storage-shortage cost minimization version of the problem, and storage cost minimization under service level constraints.

Combined storage-shortage cost minimization problem

Figure 5.3 presents the optimal solution : The first picture on the left displays the optimal cost as a function of the location of the extra inspection station. The other pictures on the right display the different values of the optimal buffer sizes z_i and coefficient of availability of wip or inventory a_i for the case of an optimally located internal inspection station (following buffer 5 in this case). We notice that for a homogeneous unreliable transfer line, the total minimal cost is a convex function of the position where λ_i is equal to 1, with $\lambda_5 = 1$ identifying

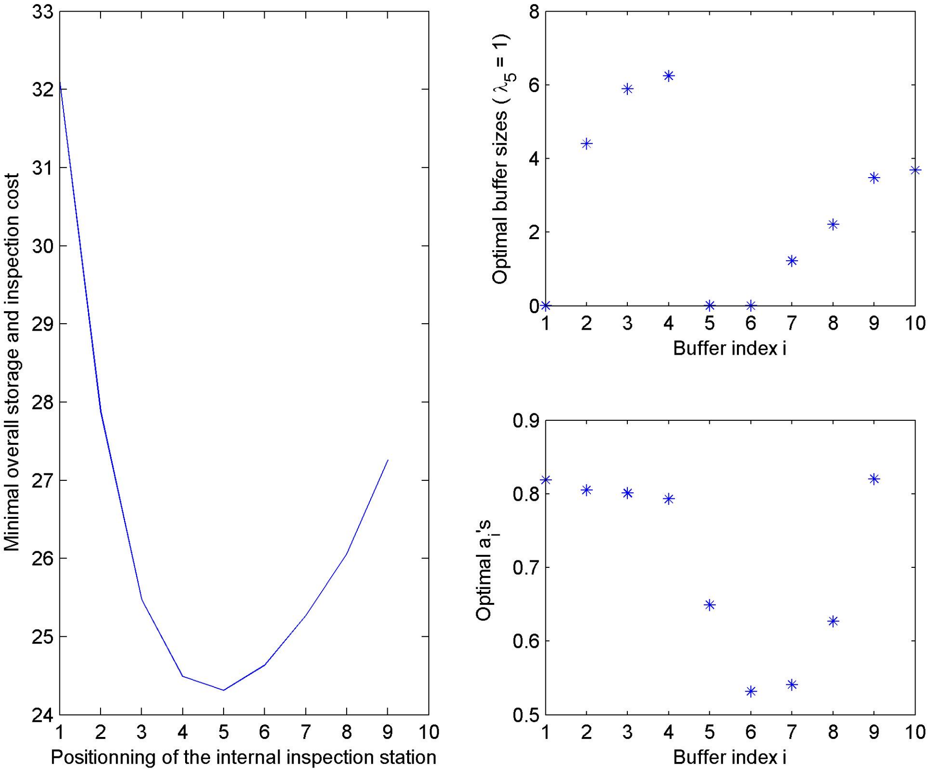


Figure 5.3 The optimal solution.

the center of the line as the optimal inspection station position. This result is rather plausible for a homogeneous line in that the inspection station must not be placed too early as its usefulness would be limited as an instrument of rejection of non conforming parts, and not too late, in that, the more non conforming parts in the system, the higher the storage costs are and the less efficient the transfer line becomes in terms of productivity. Note that in the section of the line upstream of the inspection station, the optimal buffer sizes are greater than those downstream of the inspection station; this is because for the chosen $\beta = 0.1$ (1 bad part to 10 good parts) the upstream section must satisfy a demand rate twice as large as that of the downstream section.

The first machine in the line and the one just after the inspection station (M_6) have almost the same optimal buffer size (since the inspection station resets to zero the q_5 parameter), yet their availability coefficients are very different (since the first machine is never starved while the second one is). The decrease in required buffer size from position $i = 4$, to position $i = 5$, can be explained by the fact that buffer 5 feeds into a machine M_6 whose rate of parts extraction is one half that of machines preceding it in the line.

Storage cost minimization under service level constraints

Figures 5.4 and 5.5, present the optimal solution with a required conforming finished parts availability rate a_n^{des} equal to 0.95 and 0.85.

For both figures, the first picture on the left displays the optimal cost as a function of the location of the extra inspection station. The other pictures on the right present the different values of the optimal buffer sizes z_i and coefficient of availability of wip or inventory a_i for the case of an optimally located internal inspection station ($\lambda_4 = 1$).

We notice a look almost similar to the case presented in section 6.2.1 (for a_i or z_i). The difference is noted at the optimal position of the inspection station (after M_4 in this case) and the $z_n(a_n^{des})$ necessary to meet the desired rate of availability.

5.7 Added value of joint buffer optimization and inspection station positioning

In the following, we shall study the possible gains from a fully joint consideration of buffer sizing and inspection station positioning by comparing the results based on our current mod-

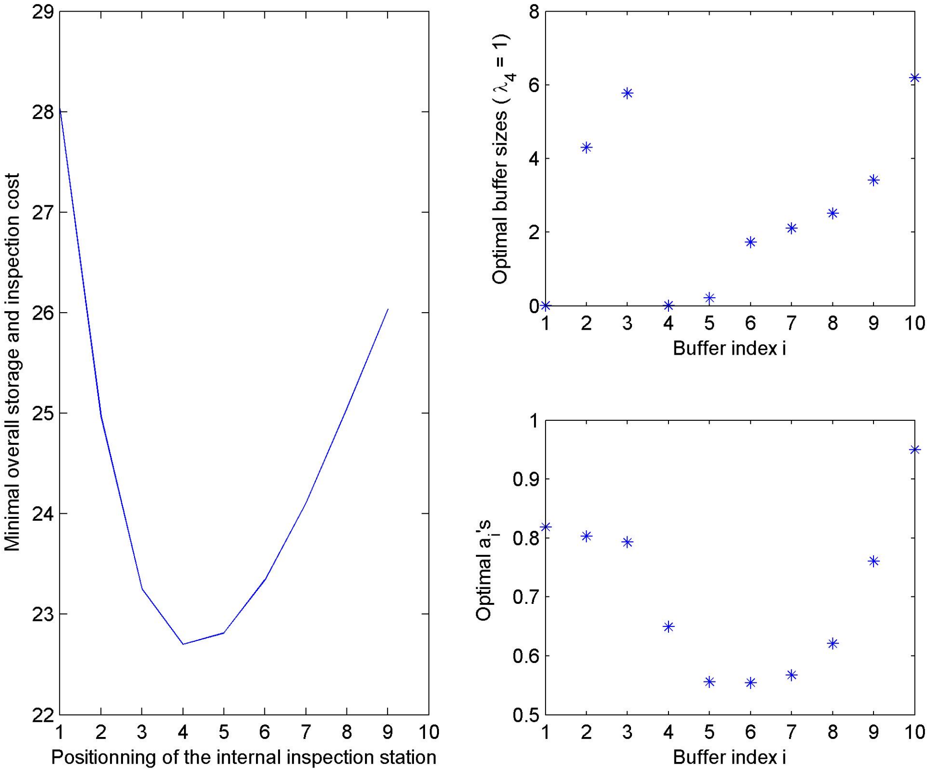


Figure 5.4 $a_n^{des} = 0.95$.

eling and optimization methodology, with those obtained when following only partially joint numerical optimization approaches such as found in Schick et al (Schick *et al.* (2005)) and Colledani and Tolio (Colledani et Tolio (2011)). Indeed, in both of these works, an attempt is made at beating combinatorial complexity of the joint optimization problem at hand, by first picking a *uniform* buffer size allocation in the transfer line, based on which a best positioning of inspection stations is sought in terms of overall productivity.

Thus Schick et al (Schick *et al.* (2005)) simulated an unreliable production line of 15 machines and 14 buffers where the machines are all identical as well as buffer sizes; they subsequently tested the performance of all possible combinations of location of inspection stations, i.e. 2^{14} possible cases (in all the cases considered, there is always an inspection station after the finished goods inventory). Following this study, Gershwin (Gershwin (2006)) showed that for a production line with 15 machines (same model as in (Schick *et al.* (2005))), the number of

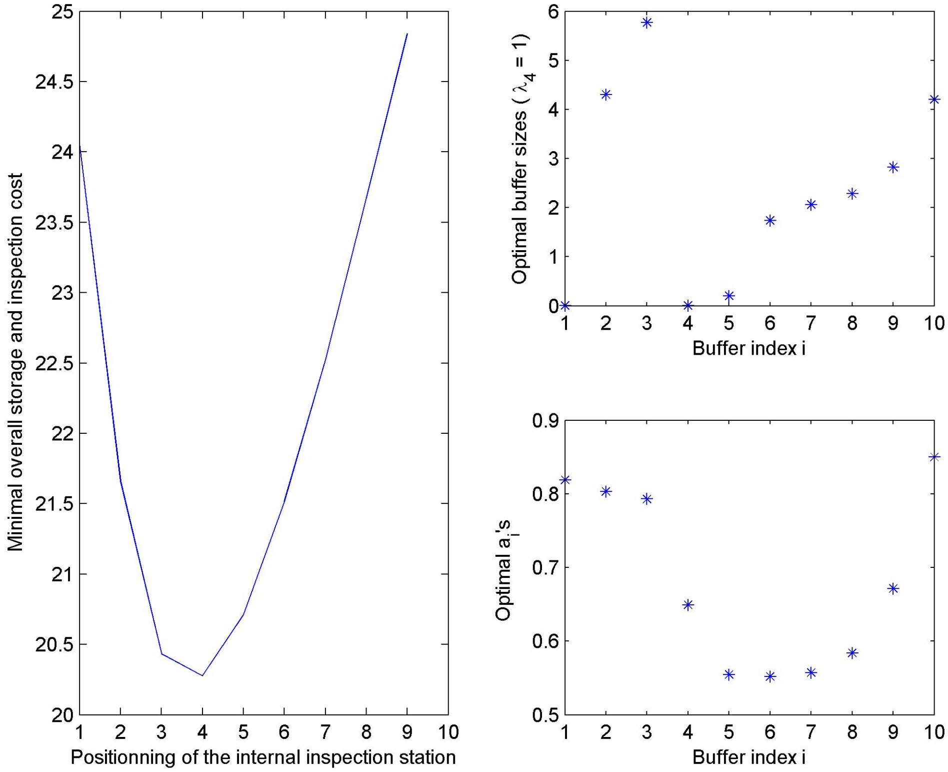


Figure 5.5 $a_n^{des} = 0.85$.

defective parts to be rejected from the system, if the inspection stations are poorly allocated, may be as much as 15% higher than that produced under a more adequate distribution of the same number of inspection stations.

Colledani and Tolio (Colledani et Tolio (2011)) studied production lines in which the positioning of the inspection stations is arbitrary, and then analytically evaluated line performance accounting for statistical control charts parameters. However, when carrying out comparative studies, they fixed the buffering within the transfer lines to a constant.

Following the partial test approach of the above researchers, we will consider a homogeneous production line of 10 machines (same parameters as in Section 5.6.2) where buffers are identically sized and subsequently determine the best and the worst location for a single internal inspection station. For this example, and a uniform buffer size of 5, the results are as follows for the combined storage-shortage cost minimization problem:

- The worst inspection station location is when $\lambda_1 = 1$ i.e. after the first machine;
- The best inspection station location is when $\lambda_9 = 1$ i.e. before the last machine.

In the next step, and by choosing as positioning of inspection stations those identified in the previous step (i.e. the best and worst location), we calculate the optimal buffer sizes and optimal cost.

Figure 5.6 presents the optimal solution i.e. the optimal buffering size distribution i.e. that associated with (i) the case where the inspection station is located after the first machine and (ii) before the last machine. The worst location associated optimal cost is 32.12 and the best location associated optimal cost is 27.26 . When comparing these two optimal costs (32.12, 27.26) with that obtained by our method (24.31), we notice that in the partial approaches of ((Schick *et al.* (2005)) and (Colledani et Tolio (2011))), the cost differential between best and

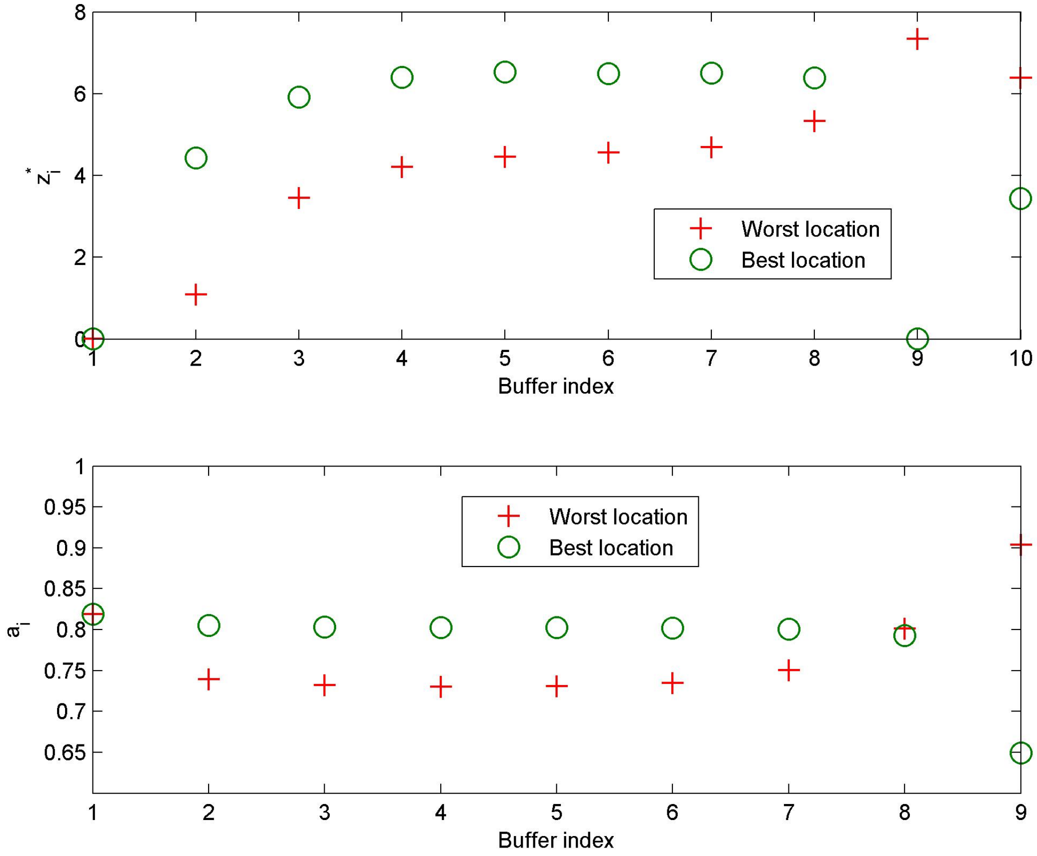


Figure 5.6 The optimal stock sizes

worst is about 15% while based on our fully joint buffering and inspection station positioning methodology, the gain over the worst case configuration is on the order of 25% (based on the results in Subsection 5.6.2). This is an indication of the kind of improvement margins that may be achievable when buffer sizing and inspection station positioning are considered as a single large optimization problem.

5.8 Further numerical results of interest

5.8.1 Influence of inspection costs per part on the optimal positioning/ necessity of the internal inspection station

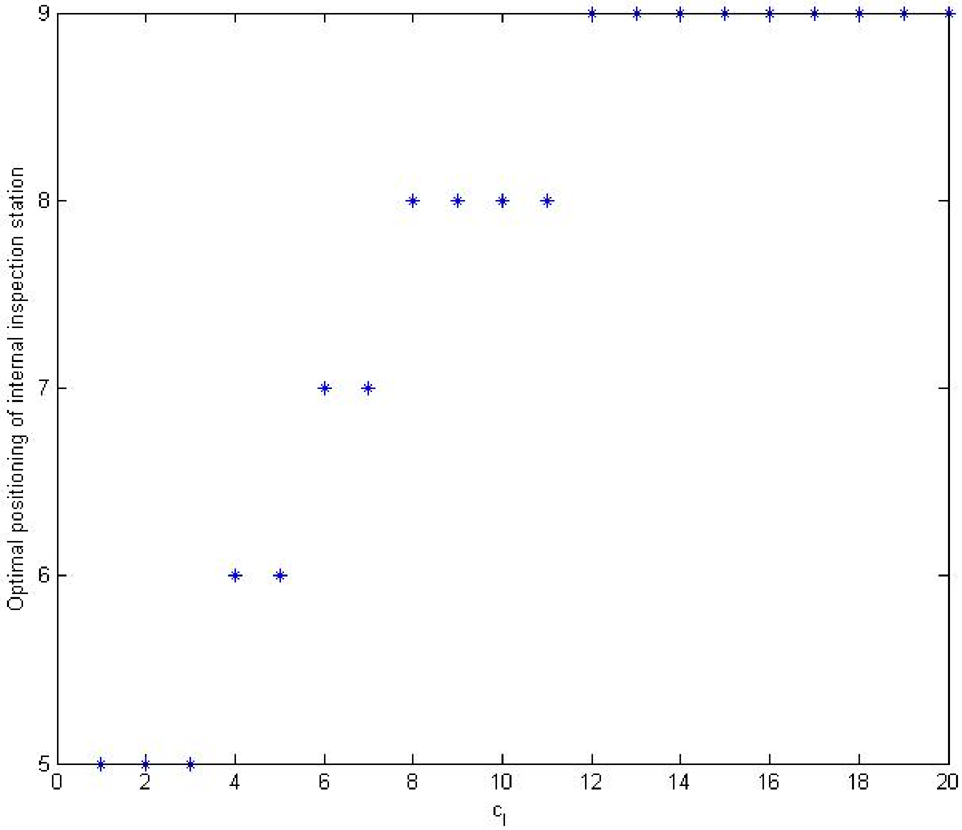


Figure 5.7 Influence of c_I on optimal inspection station position

In this section we will show the effect that an increase in inspection cost c_I has on the optimal positioning of the inspection stations. So keeping c_p and c_n constant, we varied the c_I

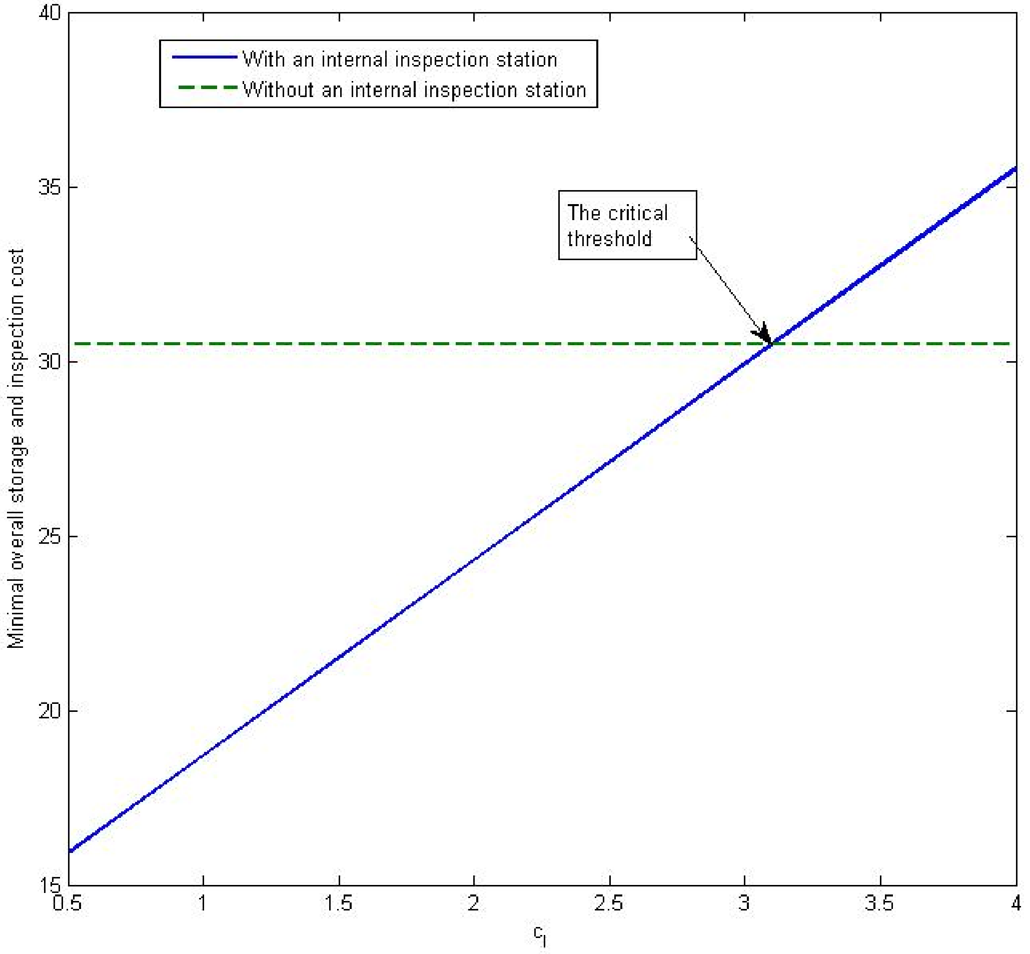


Figure 5.8 Relationship between c_I and the need for an internal inspection station

cost and note every time the optimal internal inspection station position for the homogeneous line described in Subsection 6.2.1.

Figure 5.7 shows that the more we increase the inspection cost, the further the optimal position moves away from the center towards the end of the line An intuitive explanation of this phenomenon may be as follows: Machines up to the first inspection station have to work at a rate sufficient to make up for all the non conforming parts in the system, both as produced by themselves up to the inspection station in question, and as produced in the rest of the line; this is unlike the machines past the inspection station which have to work at a common reduced rate. This means that, although the internal inspection station works as hard no matter where it is located, the more upstream it is located, the less parts it will reject, and thus the harder the second inspection station at the end of line will have to work. Therefore, as inspection costs rise, it pays to push the first inspection station further downstream so as to achieve additional inspection costs reduction for the second inspection station. In effect, this suggests that as inspection costs increase sufficiently, it may be overall more economical to operate with a single inspection station at the end of the line: This intuition is corroborated by Figure 5.8 which indicates that above a certain critical threshold unit inspection cost (intersection of the two cost lines), it is more economical to eliminate the internal inspection station altogether.

Note: We conducted a similar analysis for c_n and c_p, and we noticed no change in the optimal position of the internal inspection station (center of the line).

5.8.2 Assessing the influence of an individual poor quality machine

Figure 5.9 displays the optimal position of the inspection station for the homogeneous line of Subsection 5.6.2.1, when the β_j coefficient of a particular machine M_j is increased while all other β_i's $i \neq j$ remain identical for $j = 1, ..., 10$.

For example for $\beta_j = 0.15$, first notice that, relative to the unperturbed case, the optimal inspection station position moves from $\lambda_4 = 1$, to $\lambda_5 = 1$, as long as the poor quality machine is located at position 5 or higher. The optimal inspection station position reverts back to 4 if the poor quality machine is placed at positions 1 to 4. As one further reduces the quality of the machine to $\beta_j = 0.2$, the optimal inspection station position is further pulled down to 3 as long as the index of the poor machine is 3 or less; it moves and stays at 4 if the poor machine index is 4 or greater. Finally, a further decrease in quality to $\beta_j = 0.25$ induces a motion of the optimal inspection station position at 2, if the poor quality machine index is 1 or 2, while it moves permanently to 3 once the poor quality machine index is 3 or higher.

In conclusion, the introduction of a particularly poor single machine in an otherwise homogeneous line appears to favor configurations with earlier inspections, irrespective of the position of that machine; on the other hand, this influence appears to be most significant when the perturbed machine position is to the left of the optimal inspection station position that would prevail in the unperturbed line. This points to the fundamental asymmetry, or directionality,

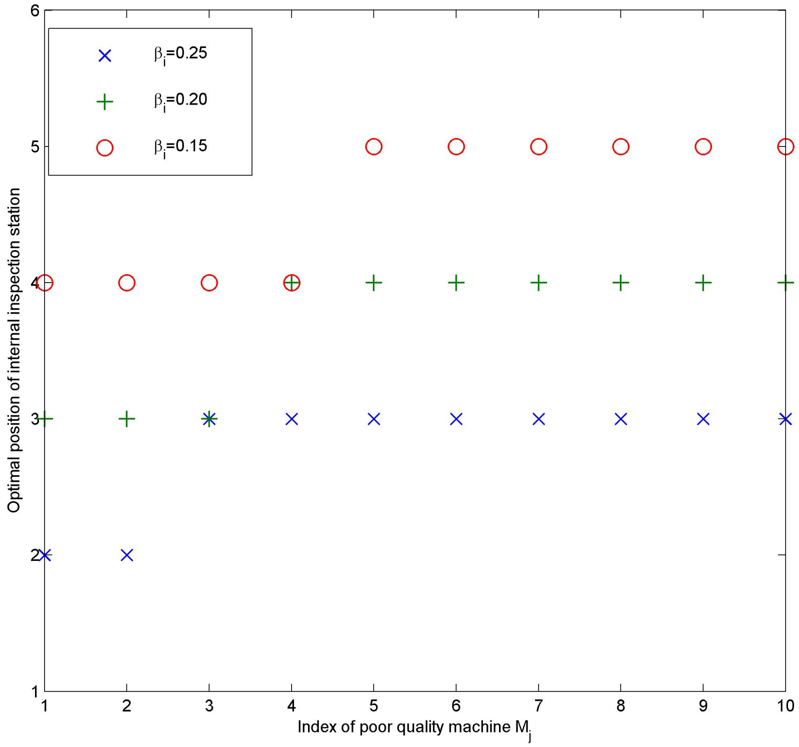


Figure 5.9 Exploring the influence of a single poor quality machine in an otherwise homogeneous transfer line.

of quality effects in a transfer line.

5.9 Conclusions and future work

This paper develops an approximate optimization formalism for unreliable transfer lines where machines produce as part of their normal operation both conforming and non conforming parts.

It is assumed that the proportion between parts correctly machined, and those improperly machined remains constant for a given machine. In addition, the line includes two inspection stations; the location of one station is fixed (dedicated to the inspection of finished parts), while the location of the other station is chosen so as optimize the total per unit time average cost (storage cost cost, possibly shortage, and inspection cost). For a constant rate of demand for finished conforming parts, two versions of the cost minimization problem have been considered: one in which backlog is allowed at a cost; the other where delays in delivery are tolerated as long as their probability of occurrence is less than some number (service level constraint). The importance of the joint consideration of buffer sizing and inspection station positioning is confirmed. Also, inclusion of inspection costs appears to be a useful device for excluding configurations with an excessive number of inspection stations.

In future work, based on our quality aware production model, we shall consider the complex combinatorics of multiple inspection stations in transfer lines, and develop for long homogeneous transfer lines, criteria for specifying the optimal frequency of inspection stations. In addition, we shall explore more complex quality models, including statistical control and the possibility of preventive machine maintenance.

CHAPITRE 6

Discussion générale

Nous observons aujourd'hui un grand intérêt pour les gestions conjointes dans les ateliers de fabrication, que se soit la gestion de vérification de la qualité conjointement avec la gestion de production où la gestion de l'approvisionnement avec la gestion de production. Le problème d'analyse des décisions optimales de production en tenant compte de la situation de qualité associée, dans le cas de l'intégration qualité/production, est toutefois très complexe.
En effet, lorsque nous optimisons les stratégies de production, nous supposons que l'état du système est complètement connu lorsque la décision de production est prise. Dans les modèles de qualité par contre, l'état de qualité de la machine, à l'exception de l'état de panne totale, est inconnu ou dans les meilleurs des cas connus mais à travers des mesures faites sur les pièces produites. Ces mesures sont décalées dans le temps.
Afin de répondre à ce besoin, nous avons débuté dans l'article du Chapitre 4 par le cas le plus simple d'une machine isolée multi-états avec un stock de produits finis partiellement observés, cette machine devant éventuellement servir comme bloc élémentaire dans l'analyse conjointe qualité/ production dans les lignes de transfert. Le mode 1 de la machine correspond à un état opérationnel dans lequel la machine produit des pièces de bonne qualité. Le mode 2 est un mode opérationnel, mais de qualité défectueuse et le mode 3 est celui de la panne totale (seul le mode 3 est observable). Les modes évoluent selon une chaine de Markov à temps continu avec des taux connus. À partir des analyses et des observations obtenues dans cette partie de nos travaux (chapitre 4), nous sommes aptes à conclure que le modèle de base peut être considéré approximativement équivalent à un modèle beaucoup plus simple à analyser, d'une machine à deux états (opérationnel et panne) telle que lorsque la machine est productive, une fraction des pièces produites est non conforme.
Suite à ces observations, nous proposons au chapitre 5 un modèle de machines en tandem dans lequel tous les encours sont observables et tel que la fraction des pièces défectueuses par rapport aux pièces conformes est constante et dépend de la machine. La modélisation est

plus complexe puisque les pièces défectueuses ne sont pas éliminées au fur et à mesure, mais accumulées le long de la ligne, une situation qui nécessite l'inclusion de stations d'inspection pour assurer la qualité requise au client, et améliorer la productivité de la ligne.

Dans le modèle proposé (chapitre 5), nous avons inclus 2 stations d'inspections dans la ligne de production : une à la fin de la ligne pour garantir la conformité de pièces livrées au client et une à l'intérieur de la ligne et dont l'emplacement constitue une variable de décision dans notre modèle. Cette nouvelle variable de décision ajoute de la complexité dans notre modèle puisqu'il s'avère que le positionnement de la station d'inspection et le dimensionnement des stocks sont deux ensembles de décisions fortement interreliées. Ainsi les deux articles rapportés aux chapitres 3 et 4 auront permis d'atteindre, au moins en partie, l'un des objectifs visés dans la thèse, en l'occurence, le développement de modèles (chapitre 4) et d'outils d'analyse (chapitre 5) pour la gestion de la production dans un contexte de qualité imparfaite. Quant à l'article au chapitre 3, il est notre réponse à l'autre objectif annoncé de la thèse : celui d'étendre les méthodes de décomposition à des architectures de production autres que Kanbans, en l'occurence ici, CONWIP.

CONCLUSION

6.1 Synthèse des travaux

Dans cette thèse, nous apportons trois types de contributions :
la première contribution (« Approximate performance analysis of CONWIP disciplines in unreliable non homogeneous transfer lines ») est centrée sur le développement de méthodes de décomposition pour l'analyse accélérée et l'optimisation des paramètres de stratégies de production autres que Kanbans, en particulier CONWIP, dans le but ultime de dégager des principes de choix de structures de stratégies de production adaptées aux divers niveaux de fiabilité et de capacité de production des machines dans une ligne de fabrication. En particulier, nous avons présenté un modèle approximatif d'estimation des paramètres clés dans une ligne de production non fiable sous la politique CONWIP. Ce modèle est basé sur la construction de blocs machine-stock corrélés entre eux par des vecteurs vitesses d'accumulation de pièces et pour lesquels on écrit les équations de Kolmogorov aux dérivées partielles et les conditions limites. Les résultats théoriques obtenus sont très satisfaisants lorsque comparés avec ceux obtenus à partir de la simulation de Monte-Carlo.
La deuxième contribution (« A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects ») est la proposition de modèles a priori à machines uniques où l'on étudie l'impact d'une qualité imparfaite de production sur le dimensionnement de paramètres de stratégies de production. Contrairement aux modèles utilisés récemment pour l'analyse de telles questions et qui présentent un caractère hybride, discret (pour la qualité) et continu (pour la production), le modèle que nous proposons est entièrement continu. Les changements aléatoires de qualité sont représentés par des sauts réversibles d'état de qualité de la machine. L'optimisation de ces modèles est réalisée à partir d'une analyse numérique des équations aux dérivées partielles de Kolmogorov associées. Également, une modification appropriée des paramètres de production et de coût de la machine dite de Bielecki/Kumar s'avère très prometteuse comme outil d'approximation du comportement des équations de Kolmogorov en question.
La troisième contribution (« Unreliable production lines with defective parts and inspection

stations ») se fonde sur les résultats de l'article précédent pour développer un modèle d'approximation pour les lignes de transfert non fiables où les machines produisent dans le cadre de leur fonctionnement normal des pièces à la fois conformes et non conformes. On suppose que la fraction entre les pièces conformes et non conformes est constante, mais dépend de la machine. De plus, la ligne comprend deux stations d'inspection ; l'emplacement d'une station est fixe (dédié à l'inspection des pièces finies) alors que l'emplacement de l'autre station est choisi de manière à optimiser le cout total moyen par unité de temps (ce dernier inclut coût de stockage, coût d'inspection et cout de pénurie lorsqu'applicable).

6.2 Améliorations futures

Comme travaux futurs, et sur la base du modèle de l'article 3 (chapitre 5), il serait intéressant d'étudier des situations plus complexes où il est question de plusieurs stations d'inspection. De plus, il serait indiqué d'enrichir les modèles de qualité associés aux machines pour y inclure des états non réversibles de qualité défectueuse, devant être détectés par contrôle statistique et pouvant requérir une maintenance préventive.

De plus, il pourrait être intéressant d'utiliser le modèle à trois modes du chapitre 4, plus complexe mais plus précis que celui à deux modes utilisé dans l'article du chapitre 5, comme bloc élémentaire dans l'analyse par décomposition des problèmes de dimensionnement de stocks et positionnement de stations d'inspection dans une ligne de transfert. Enfin, il pourrait être intéressant de revisiter le dimensionnement du paramètre de stockage dans une boucle CONWIP et le positionnement de stations d'inspection dans de telles boucles, dans un contexte de qualité imparfaite.

En conclusion, il devient évident que les travaux de thèse présentés ici, pourraient constituer le prélude à de nombreuses recherches futures à l'intersection des questions de gestion de la production, et gestion de la qualité, dans les lignes de fabrication.

RÉFÉRENCES

ALGOET, P. H. (1989). Flow balance equations for the steady-state distribution of a flexible manufacturing system. *IEEE Transactions on Automatic Control*, 34 (8), 917– 921.

ANCELIN, B. et SEMERY, A. (1987). Calcul de la productivité d'une ligne integrée de fabrication : Calif, une méthode analytique industrielle. *RAIRO APII*, 21, 209–238.

ANILY, S. et GROSFELD-NIR, A. (2006). An optimal lot-sizing and offline inspection policy in the case of nonrigid demand. *Operations Research*, 54, 311–323.

BEN-DAYA, M. (2002). The economic production lot-sizing problem with imperfect production processes and imperfect maintenance. *International journal of Production Economics*, 76, 257 – 264.

BEN-DAYA, M., NOMAN, S. et HARIGA, M. (2006). Integrated inventory control and inspection policies with deterministic demand. *Computers and Operations Research*, 33, 1625 – 1638.

BIELECKI, T. et KUMAR, P. (1988). Optimality of zero inventory policies for unreliable manufacturing systems. *Operations Research*, 36 (4), 532–541.

BONVIK, A. (1996). *Performance analysis of manufacturing systems.* Thèse de doctorat, Massachusetts institute of technology, Department of Electrical Engineering and Computer Science.

BONVIK, A., DALLERY, Y. et GERSHWIN, S. (2000). Approximate analysis of production systems operated by a conwip/finite buffer hybrid control policy. *International journal of Production Research*, 30, 2845–2869.

BUZACOTT, J. (1967). Automatic transfer lines with buffer stocks. *International journal of Production Research*, 5, 182–200.

CHAKRABORTY, T., GIRI, B. et CHAUDHURI, K. (2008). Production lot sizing with process deterioration and machine breakdown. *European journal of Operational Research*, 185, 606 – 18.

CHANG, H. C. (2004). An application of fuzzy sets theory to the eoq model imperfect quality items. *Computers and Operations Research*, 31, 2079 – 92.

CHIANG, S., KUO, C. et MEERKOV, S. (2000). Dt- bottlenecks in serial production lines : Theory and application. *IEEE Transactions Robotics and Automation*, 16, 567–580.

COLLEDANI, M. et TOLIO, T. (2005). Impact of statistical process control (spc) on the performance of production systems-part 2 (large systems).

COLLEDANI, M. et TOLIO, T. (2006a). Impact of quality control on production system performance. *Annals of the CIRP*, 55, 453–456.

COLLEDANI, M. et TOLIO, T. (2006b). Performance evaluation of production systems monitored by statistical process control and offline inspections. *Information Control Problems in Manufacturing*, 317–322.

COLLEDANI, M. et TOLIO, T. (2011). Integrated analysis of quality and production logistics performance in manufacturing lines. *International journal of Production Research*, 49, 485 –518.

DALLERY, Y. et BIHAN, H. (1999). An improved decomposition method for the analysis of production lines with unreliable machines and finite buffers. *International journal of Production Research*, 37 (5), 1093–1117.

DALLERY, Y. et GERSHWIN, S. (1992). Manufacturing flow line systems : a review of models and analytical results. *Queueing Systems Theory and Applications*, 12, 3–94.

FREIN, Y., COMMAULT, C. et DALLERY, Y. (1996). Modeling and analysis of closed-loop production lines with unreliable machines and finite buffers. *IIE Transactions*, 28, 545–554.

GERSHWIN, B. et SCHICK, C. (2007). A taxonomy of quality / quantity issues in manufacturing systems. *Analysis of manufacturing systems : AMS*, 1–6.

GERSHWIN, S. (1987). An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research*, 35 (2), 291–305.

GERSHWIN, S. (1994). *Manufacturing systems engineering*. Englewood Cliffs, NJ : PTR Prentice Hal.

GERSHWIN, S. (2006). How do quantity and quality really interact ? precise models instead of strong opinions. *Information Control Problems in Manufacturing*, 1, 33–40.

GERSHWIN, S. et WERNER, L. M. (2007). an approximate analytical method for evaluating the performance of closed-loop flow systems with unreliable machines and finite buffers. *International journal of Production Research*, 45, 3085–3111.

HU, J. (1995). Production rate control for failure-prone production systems with no backlog permitted. *Automatic Control, IEEE Transactions on*, 40, 291 –295.

INMAN, R., BLUMENFELD, D., HUANG, N. et LI, J. (2003). Designing production systems for quality : research opportunities from an automobile industry perspective. *International journal of production Research*, 41, 1953–1971.

KIM, J. et GERSHWIN, S. (2005). Integrated quality and quantity modeling of a production line. *OR Spectrum*, 27, 287–314.

KIM, J. et GERSHWIN, S. (2008). Analysis of long flow lines with quality and operational failures. *IIE Transactions*, 40, 284–296.

LEVANTESI, R., MATTA, A. et TOLIO, T. (2003). Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Performance Evaluation*, 51, 247–268.

MALHAMÉ, R. et BOUKAS, E.-K. (1991). A renewal theoretic analysis of a class of manufacturing systems. *IEEE Transactions on Automatic Control*, 36 (5), 580– 587.

MHADA, F., HAJJI, A., MALHAMÉ, R., GHARBI, A. et PELLERIN, R. (2011). Production control of unreliable manufacturing systems producing defective items. *Journal of Quality in Maintenance Engineering*, 17, 238 – 253.

SADR, J. et MALHAMÉ, R. (2004a). Decomposition/aggregation-based dynamic programming optimization of partially homogeneous unreliable transfer lines. *IEEE Transactions Automatic Control*, 49, 68–81.

SADR, J. et MALHAMÉ, R. (2004b). Unreliable transfer lines : decomposition/ aggregation and optimisation. *Annals of Operations Research*, 125, 167–190.

SCHICK, I. C., GERSHWIN, S. et KIM, J. (2005). New results on the integrated analysis of quality and quantity in production lines. *5th international conference on Analysis of Manufacturing Systems - Production Management*.

SPEARMAN, M., WOODRUFF, D. et HOPP, W. (1990). Conwip : A pull alternative to kanban. *International journal of Production Research*, 28, 879Ű894.

TERRACOL, C. et DAVID, R. (1987). Performances d'une ligne composée de machines et de stock intermédiaries. *RAIRO APZI*, 21, 239–262.

XIE, X. L. (1986). *Comparaison des différentes méthodes pour l'évaluation des performances d'une ligne de production*. Mémoire de maîtrise, Laboratoire d'Automatique de Grenoble.

ZIMMERN, B. (1956). Etude de la propagation des arrêts aléatoires dans les chaines de production. *Revue de Statistique Appliquée*, 4, 85–104.

ANNEXE A

Approximate loop mode Markov chain parameters under the single permissible machine failure assumption.

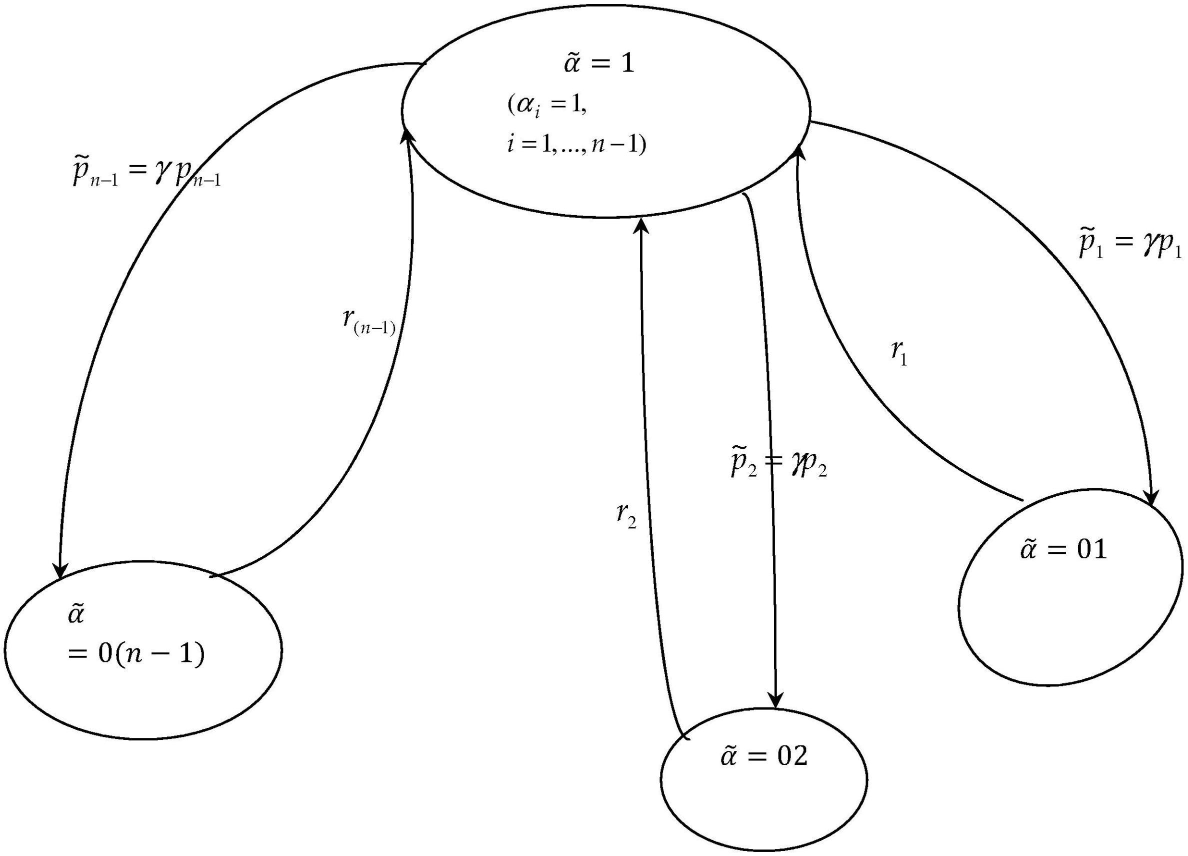


Figure A.1 CONWIP loop approximate macromachine under the single permissible machine failure assumption

In Figure A.1 above, the objective is to compute a common individual machine failure rate amplification factor γ so that the probability that all machines in the loop are operational remains unchanged despite the single permissible machine failure assumption. We thus have the following constraint (from original (n-1) independent machines system) :

$$Pr[\tilde{\alpha} = 1] = \prod_{i=1}^{n-1} (\frac{r_i}{r_i + p_i}). \tag{A.1}$$

Furthermore, the steady-state flow balance equations in Figure A.1 yield :

$$\tilde{p}_i \, Pr[\tilde{\alpha} = 1] = r_i \, Pr[\tilde{\alpha} = 0i], \quad i = 1, ..., n-1. \tag{A.2}$$

(A.2) yields :

$$Pr[\tilde{\alpha} = 0i] = \frac{\tilde{p}_i}{r_i} Pr[\tilde{\alpha} = 1] = \gamma \, \frac{p_i}{r_i} Pr[\tilde{\alpha} = 1], \quad i = 1, ..., n-1. \tag{A.3}$$

Thus :

$$\sum_{i=1}^{n-1} Pr[\tilde{\alpha} = 0i] = 1 - Pr[\tilde{\alpha} = 1] = \gamma Pr[\tilde{\alpha} = 1][\sum_{i=1}^{n-1} \frac{p_i}{r_i}] \tag{A.4}$$

Consequently :

$$\gamma = \frac{1 - \prod_{i=1}^{n-1} \frac{r_i}{(r_i + p_i)}}{(\prod_{i=1}^{n-1} \frac{r_i}{r_i + p_i})(\sum_{i=1}^{n-1} \frac{p_i}{r_i})} \tag{A.5}$$

Figure A.1 and (A.5) yield the Markov chain intensity matrix (3.36).

ANNEXE B

Derivation of (5.7) and (5.12)

The stock dynamics is defined by

$$\begin{aligned}\frac{dx_1(t)}{dt} &= -d \\ \frac{dx_2(t)}{dt} &= k - \frac{x_2(t)\,d}{x_1(t)}\end{aligned}$$

therefore :

$$\begin{aligned}x_1(t) &= x_1(0) - d\,t \\ \frac{dx_2(t)}{dt} &= k - \frac{x_2(t)\,d}{x_1(0) - d\,t}\end{aligned} \tag{B.1}$$

becomes :

$$\frac{dx_2(t)}{dt} + \frac{x_2(t)\,d}{x_1(0) - d\,t} = k$$

by multiplying both sides of the equation by the same term, we get :

$$\begin{aligned}&\exp\left(\int_0^t \frac{d}{x_1(0) - d\,\tau}\,d\tau\right)\left(\frac{dx_2(t)}{dt} + \frac{x_2(t)\,d}{x_1(0) - d\,t}\right) = k\,\exp\left(\int_0^t \frac{d}{x_1(0) - d\,\tau}\,d\tau\right) \\ \text{i.e.}\quad &\frac{d}{dt}\left[\exp\left(\int_0^t \frac{d}{x_1(0) - d\,\tau}\,d\tau\right)\,x_2(t)\right] = k\,\exp\left(\int_0^t \frac{d}{x_1(0) - d\,\tau}\,d\tau\right)\end{aligned} \tag{B.2}$$

and since

$$\begin{aligned}\exp\left(\int_0^t \frac{d\,d\tau}{x_1(0) - d\,\tau}\right) &= \exp\left(-\left[\ln\left(x_1(0) - d\,\tau\right)\right]_0^t\right) \\ &= \exp\left(\ln\left(\frac{x_1(0)}{x_1(0) - d\,t}\right)\right) \\ &= \frac{x_1(0)}{x_1(0) - d\,t}\end{aligned}$$

(B.2) becomes :

$$\begin{aligned} \frac{d}{dt}\left[\frac{x_1(0)}{x_1(0)-d\,t}\,x_2(t)\right] &= k\,\frac{x_1(0)}{x_1(0)-d\,t} \\ \frac{x_1(0)}{x_1(0)-d\,t}\,x_2(t)-x_2(0) &= \frac{k\,x_1(0)}{d}\int_0^t \frac{d\,d\tau}{x_1(0)-d\,\tau} \\ \frac{x_1(0)}{x_1(t)}\,x_2(t) &= x_2(0)\,+\frac{k\,x_1(0)}{d}\,\ln\left(\frac{x_1(0)}{x_1(t)}\right) \end{aligned}$$

We finally retrieve (5.7)

$$\begin{aligned} x_1(t) &= x_1(0)-d\,t \\ x_2(t) &= x_1(t)\,(\frac{x_2(0)}{x_1(0)}+\frac{k}{d}\ln\frac{x_1(0)}{x_1(t)}) \end{aligned}$$

(5.12) is a special case of (5.7) where the initial points are $x_1(0)=\frac{z\,d}{k}$ and $x_2(0)=\frac{z(k-d)}{k}$:

$$\begin{aligned} &x_2(t)=x_1(t)\,(\frac{k-d}{d}+\frac{k}{d}\ln(\frac{z\,d}{k})-\frac{k}{d}\ln(x_1(t))) \\ &x_1(t)=\frac{z\,d}{k}-d\,t \end{aligned}$$

ANNEXE C

Derivation of (5.10)

The stock dynamics in the region $x_1(t) + x_2(t) = z$ is defined by :

$$\begin{aligned} \frac{dx_1(t)}{dt} &= \frac{x_2(t)\,d}{x_1(t)} && \text{(C.1)} \\ x_2(t) &= z - x_1(t) && \text{(C.2)} \end{aligned}$$

Substituting (C.2) in (C.1), we obtain :

$$\frac{dx_1(t)}{dt} = \frac{z - x_1(t)\,d}{x_1(t)}$$

Therefore

$$\begin{aligned} d &= \frac{dx_1(t)}{dt}(\frac{x_1(t)}{z - x_1(t)}) \\ &= \frac{dx_1(t)}{dt}(-1 + \frac{z}{z - x_1(t)}) \\ &= \frac{dx_1(t)}{dt}(-1 + \frac{z}{z - x_1(t)}) \\ &= -\frac{dx_1(t)}{dt} + \frac{z}{z - x_1(t)}\frac{dx_1(t)}{dt} && \text{(C.3)} \end{aligned}$$

Integrating (C.3) with respect to time,we have :

$$\begin{aligned} d\,t &= -(x_1(t) - x_1(0)) - z\ [\ln(z - x_1(t))]_0^t \\ &= -(x_1(t) - x_1(0)) - z\ \ln(\frac{z - x_1(t)}{z - x_1(0)}) \end{aligned}$$

i.e.

$$x_1(t) - x_1(0) + z \ln \frac{z - x_1(t)}{z - x_1(0)} + d\,t = 0$$

Hence, Equation (5.10) is retrieved :

$$\begin{aligned}
x_1(t) - x_1(0) + z \ln \frac{z - x_1(t)}{z - x_1(0)} + d\,t = 0 \\
x_2(t) - x_2(0) - z \ln \frac{x_2(t)}{x_2(0)} - d\,t = 0
\end{aligned}$$