國立臺灣大學電機資訊學院電機工程學系

碩士論文

Graduate Institute of Electrical Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

分析不同變異考量下之簡化模型對排序佳化的影響：以迴流產線產能分配為例

Analysis of How Selecting Simplified Models of Different Variability Affects Ranking for Ordinal Optimization: Re-entrant Line Capacity Allocation Case

張鈞閔

Chun-Ming Chang

指導教授：張時中、陳俊宏 博士

Advisor: Shi-Chung Chang, Chun-Hung Chen Ph.D.

中華民國 104 年 8 月

August 2015

誌謝

首先要感謝台大電機系所這六年來的栽培，特別感謝我的指導教授張時中博士在兩年研究期間耐心指導我做研究的方法，叮囑我們要有左右互搏的思辨精神，您對研究的嚴謹態度對我影響重大，使本論文得以順利完成。特別感謝指導教授陳俊宏博士，每次討論後都覺得豁然開朗、如沐春風，您總能給予精闢的建議，能和您做研究是我莫大的榮幸。謝謝兩位指導教授作為研究路上堅實的巨人的肩膀。本論文由國科會計畫(編號 NSC 102-2221-E-002-206, 102-2219-E-002-012, 103-2221-E-002-220-MY2) 的部分支持下完成，特此致謝。

在此更要感謝學位考試委員台灣大學張時中教授、美國喬治梅森大學陳俊宏教授、美國康乃狄克大學陸寶森教授給予本論文的指正與建議，讓本論文更臻完整。特別感謝美國康乃狄克大學陸寶森教授在擔任訪問學者期間給予我研究上的寶貴意見，與您討論總是獲益良多，謝謝您使得本論文的格局與面向更加寬廣。

謝謝 Lab 207 的學長姐與同學們，謝謝輝哥，身兼實驗室助理幫忙我們處理很多行政事務；謝謝舜丞是我研究生涯最好的榜樣，我會努力向你看齊；謝謝在實驗室共同奮戰的泓捷、名傑、家興、振豪、羅賓、琳茵、冠霖、惠平、登傑。我會記得這些日子一起度過的漫漫長夜和一同散步閒話的小確幸。泓捷總保持正面的態度，謝謝最後這段時間的陪伴與鼓勵；名傑很務實的做好每件事情，相信一定會結出很美好的果實；健談的羅賓希望 Lab 207 是你很好的回憶；坐在我旁邊一年義氣的家興，陪我玩耍、吃飯、睡覺；振豪，要記得我的李星，給酷；琳茵、冠霖、惠平、登傑，祝你們研究順利。最後想要特別謝謝小高這兩年來不僅在研究的教導，在生活中跟你分享事情總能得到很多收穫，祝你之後一切順利。

最要感謝我的家人們，謝謝辛苦工作的父親、母親提供我一個安心無虞的環境，謝謝哥哥時常的關心與照顧，謝謝女友抒珉一直是我重要的依靠。謝謝你們像大海般的溫柔與包容，讓我懷抱遠大志向勇往直前。紙短情長，謝謝你們無私的奉獻與栽培造就今天的我。謹以此論文，獻給所有陪伴我、幫助我的人。

Abstract

Ordinal optimization (OO) focuses on “ranking” in performances among designs instead of their “values” and exploits a goal softening strategy aiming at “good enough” designs with high probability as opposed to an optimal design for sure. Ordinal transformation (OT) is an OO technique that utilizes a simplified model for perform evaluation and ranking to further reduce computational effort. There are often multiple choices of simplified models for a system that capture different levels of details or aspects. The selection of an appropriate simplified model is a key factor for the effectiveness of OT and OO. Thus, how to select simplified models for ranking and how to analyze the goodness of simplified models are significant and challenging problems for OT and OO.

However, there is little literature to theoretically explore the influences of different simplified models on ranking largely because the comparison among various simplified models is often difficult in lack of a common ground. In addition, ranking is a relative index instead of an absolute index. The goodness of ranking is not straight forward to quantify let alone to analyze.

In this thesis, machine capacity allocation for re-entrant lines, an important engineering optimization problem, is adopted as the conveyor problem to investigate the selection of an appropriate simplified model. In particular, Jackson network approximation (JNA) and queueing network analyzer (QNA), two commonly used queueing network approximation models, are studied with the mean cycle time as the performance index. Both models are developed based on parametric decomposition, but JNA has unity SCVs due to its exponential time assumptions while QNA has heterogeneous SCVs. Thus, we compare between QNA and JNA to investigate how

selecting simplified models of different variability affects ranking and analyze the goodness of a simplified model with consideration of heterogeneous SCVs.

A key step in the investigation is the quantification of the goodness of rankings by simplified models. This is difficult since "ranking" is a relative index, not an absolute index. A bound and ranking analysis (BRA) is innovatively developed to quantify and analyze the goodness of rankings by simplified models. BRA consists of two innovations:

i) Analyze the upper and lower bounds of simplified models,

ii) Derive the probability of correct ranking under the assumption of actual cycle time being uniformly distributed between its upper and lower bounds.

The probability of correct ranking between a pair of designs for a single GI/G/m queue is first studied. With the variation of two QNA approximations, the least variation of their upper bound is derived and this helps obtain a higher probability of correct ranking α.

The results and insights from BRA are as follow.

i) Showed that QNA approximation is bounded by known upper and lower bounds proposed by Kingman, Brumelle and Marshall respectively.

ii) Compared with existing literature results, QNA captures the variations of true expected cycle time well because of heterogeneous SCVs but JNA does not.

iii) Obtained a valuable insight from derived α that capturing heterogeneous SCVs benefits the ranking of top designs and improves probability of correct ranking because variability has greater impacts on cycle time while lower utilization.

Based on the above for a single GI/G/m queue, BRA is then extended to general re-entrant lines with multiple workstations. Rank correlation, which measures the concordance of pair-wise comparisons in two quantitative indices, is adopted to quantify the goodness of ranking.

Simulation studies over a five-station re-entrant line demonstrated that rank correlation of QNA always outperforms that of JNA, and the difference is especially significant for top designs. This is consistent with the insight iii) obtained from BRA. Then, in order to investigate the effects of heterogeneous SCVs, the original design space is transformed using true ranking, and in this ordinal space each thirty designs are clustered into a group. After grouping, we found that heterogeneous SCVs contribute to improve differentiation between groups and also make designs in a group better separated, which benefit raise the probability of correct ranking. This is why heterogeneous SCVs benefit rank correlation of a simplified model.

In summary, the contributions of this thesis are as follows.

i) Adopted re-entrant line capacity allocation as the conveyor problem to meaningfully compare two simplified models: JNA has unity SCVs while QNA has heterogeneous SCVs,

ii) Established theoretical foundations, BRA, to analyze the probability of correct ranking and quantify the goodness of different simplified models,

iii) Derived the probability of correct ranking between a pair of designs α, and a valuable insight is that heterogeneous SCVs have greater impacts on top designs,

iv) Simulation studies demonstrated that heterogeneous SCVs contribute to improve differentiation between groups and make designs in a group better separated,

v) Because of iv), QNA always outperforms JNA in terms of rank correlation, and the difference is especially significant for top designs. It is consistent with iii),

vi) Investigate in aspects of both theory and experiment how selecting simplified models of different variability affects ranking for OO and OT.

Keyword: Model selection, ordinal optimization, heterogeneous variability, ranking analysis, re-entrant line capacity allocation

中文摘要

排序佳化(OO)著重在設計間績效值的排名而不是績效值本身，並利用目標軟化的策略以很高機率找到足夠好的設計取代勢必求得最佳設計。排序轉換（OT）是一排序佳化的技術，其利用一個簡化模型的績效評估和排名進一步地減少計算量。在同一個系統中經常有多個簡化模型的選擇，其掌握到系統中不同細節或不同面向。選擇一個適合的簡化模型決定 OT 和 OO 效能的關鍵因素，因此如何選擇用來排名的簡化模型以及如何分析簡化模型的優劣，對於 OT 和 OO 來說都是重要且挑戰的問題。

因為不同簡化模型之間大多缺乏一個共同的依據，使得不同的簡化模型之間的比較是非常困難，以致於鮮少有文獻從理論上探討不同簡化模型對於排名的影響。此外，因為排名是一個相對指標而非絕對指標，使得簡化模型的排名之優劣難以直接量化更遑論分析。

在本論文中，我們選用一個重要的工程優化問題─迴流產線產能分配作為載具以研究如何選擇適合的簡化模型。採用傑克遜網絡近似（JNA）和排隊網絡分析儀（QNA）這兩種常見的排隊網絡近似模型進行研究，並以其平均生產週期時間為績效指標。此兩種模型皆發展自參數分解法，但 JNA 由於指數分配的假設為統一的 SCVs 而 QNA 則有異質的 SCVs。 因此我們對 QNA 與 JNA 進行比較來研究如何考量不同變異的簡化模型選擇對排名的影響，並分析考量異質 SCVs 的簡化模型之優劣。

本研究其中一個關鍵在於量化簡化模型的排名之優劣，因為排名是一個相對的指標而不是絕對指標導致此量化的困難。為此，我們創新地開發了一界限與排名分析(BRA)，用來量化和分析簡化模型的排名之優劣。BRA 有兩項創新之處:

i) 分析簡化模型的上限與下限

ii) 推導正確排序機率，假設真實生產週期在其上、下限間為均勻分佈

首先分析在單一 GI/G/m queue 中正確排序一對設計的機率，從兩個設計間 QNA 近似績效值變化量推導其上限的最少變化量，因而得到更好的正確排序機率 α。從 BRA 的分析可得到下列結果與觀察:

i) 證明 QNA 近似績效值落於分別由 Kingman、Brumelle 和 Marchall 所提出的上限與下限,

ii) 與文獻中的結果比較發現，QNA 因為掌握到異質 SCVs 的特徵，故能掌握到真實的期望生產週期之變化，但 JNA 不行

iii) 從 α 可得知一重要的觀察—因變異對於生產週期的影響在產線於低使用率時特別顯著，所以異質 SCVs 對前若干名設計的影響較大因而有助於提高正確排序機率。

根據上方對單一 GI/G/m queue 的分析結果，將 B&R 分析推展到普遍常見的有多工作站之迴流產線。以排名相關性(rank correlation)作為量化其排名優劣的指標，排名相關性是用來衡量兩個定量指標間成對比較是否一致的統計值。

從實驗模擬發現 QNA 的排名相關性總優於 JNA 的排名相關性，兩者差異在前若干名設計中特別顯著，此結果與 BRA 得到的 iii)觀察一致。為了更加瞭解異質 SCVs 的效果，我們將原本的設計空間依據設計的真實排名轉換到一排序空間，在該排序空間中的每三十個設計都群聚成一組。分組後我們發現異質 SCVs 有助於增加各組間的差異且使得同組內之設計被更好地區隔，這兩者都有助於提升正確排序機率—這就是為什麼考量異質 SCVs 能增進排名相關性。

總結本論文貢獻在於，

i) 以迴流產線產能分配問題為載具比較兩種基於相同理論基礎的簡化模型，有統一 SCVs 的 JNA 和異質 SCVs 的 QNA

ii) 建立 BRA 來分析正確排序機率以此作為量化簡化模型之優劣的依據

iii) 推導正確排序機率 α，和一重要觀察—異質 SCVs 對於前若干名設計有較大的影響

iv) 模擬結果顯示異質 SCVs 有助於增加各組間的差異且使得同組內之設計被

更好地區隔

v) 因為 iv)的結果，QNA 在排名相關性上總優於 JNA，且兩者差距在前若干名設計中尤其顯著，和 iii)的結論一致。

vi) 從學理和實驗的面向分析不同變異考量下之簡化模型對排序佳化和排序轉換的影響

關鍵字: 模型選擇、排序佳化、異質變異、排名分析、迴流產線產能分配

Contents

List of Figures

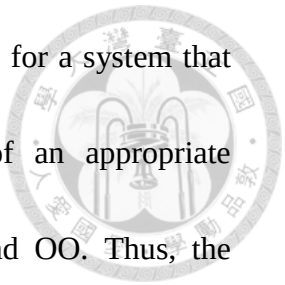
List of Tables

Chapter 1

Introduction

1.1 Motivation

Ordinal optimization (OO) focuses on “ranking” in performances among designs instead of their “values” and exploits a goal softening strategy aiming at “good enough” designs with high probability as opposed to an optimal design for sure. Ordinal transformation (OT) proposed by Xu and Chen et al. [1] is an OO technique that utilizes a simplified model for perform evaluation and ranking to further reduce computational effort. Huang et al. [38] proposes an OT approach to transform the original design space into a new one-dimensional space where all designs are positioned according to their ordinal ranks using the simplified model. The original design space may be high-dimensional, have multiple local optimums spread far apart, and include a mix of integer-valued and categorical variables [38]. After OT, the new design space is one-dimensional, likely to be well-behaved and have some global trends [44]. Therefore, OT has the following important advantages: 1) handles a mix of discrete and categorical decision variables in a high-dimensional design space; 2) OT is a general approach and has the potential to effectively perform optimization when simplified models are appropriate.



However, there are often multiple choices of simplified models for a system that capture different levels of details or aspects. The selection of an appropriate simplified model is a key factor for the effectiveness of OT and OO. Thus, the selection of simplified models is a significant problem for OT. There are few theoretical results on the effects of rankings by different simplified models for OO since it is difficult to meaningfully compare among different simplification methods in a common ground.

Another challenge of selecting simplified models for OO is the quantification of the goodness of rankings by simplified models. This is difficult because “ranking” is a relative index while “value” is an absolute index. Thus, how to theoretically analyze the goodness of rankings by different simplified models is another challenging problem for OO.

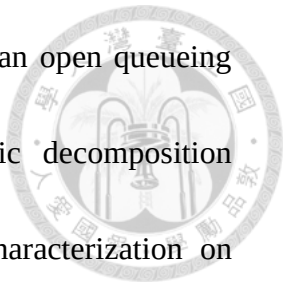
In this thesis, we adopt re-entrant line machine capacity allocation problem as the conveyor problem to investigate selection of simplified models for OO. In a semiconductor fab, there are several products flow through a sequence of processing steps. The production flow of each product consist of a sequence of processing steps , where the flow visits some workstations more than once at different steps of processing. This re-entrant characteristic poses a unique challenge to capacity allocation, since jobs of different products as well as jobs of the same product but at

different processing steps may compete for the finite capacity of a workstation.

In the literature, because capacity allocation is crucially significant to semiconductor fab, there has been much research on this problem. Chang et al. [45] designed a daily target setting system, which helps allocate the capacity to product as well as stages under master production schedule in a horizon of one day. Field applications demonstrated that this allocation method leads to over 20% increase in daily moves and more than 8% decrease of wafers-in-process (WIPs) of a foundry fab case [45]. Therefore, in this thesis, we shall consider the optimal machine capacity allocation for re-entrant lines, an important engineering optimization problem, as the conveyor problem to investigate the selection of an appropriate simplified model.

1.2 Literature Survey

Re-entrant line capacity allocation is regarded as the conveyor problem in this thesis. The determination of number of machines allocated in each workstation of an arbitrary queueing network is a complex problem. There is a mount of literature for the optimal allocation of machines for single node, exponential service and infinite buffer queueing networks, yet not as much literature exists for the case when there are complex topologies like re-entrant lines and general inter-arrival or service time distributions in the network because of the intractability to acquire the exact



performance. Such a complex production line can be modeled as an open queueing network [2][3][4][26]. For open queueing networks, parametric decomposition method is a widely used approximation, but it has different characterization on variability terms owing to various assumptions. Therefore, we analyze how the characterization of variability in simplified models affects ranking using re-entrant line capacity allocation as the conveyor problem to investigate the model selection problem for OO and how to analyze the goodness of ranking by simplified models.

Therefore, we review the related works of optimal machine allocation problem in re-entrant lines, and the commonly used techniques and models which developed for performance evaluation of a queueing network.

1.2.1 Optimal Capacity Allocation of Re-entrant Lines

For closed queueing networks, Shanthikumar and Yao [31] formulate the machine allocation problems as a nonlinear integer program and propose a greedy heuristic to maximize throughput.

Dallery and Stecke [32] define the optimal configuration problem in a closed queueing network. Then, use designed decomposition method to determine the best configuration of each subnetwork that yields the highest throughput for the overall closed queueing network, where the number of stations, the number of machines, and

the workload allocated to each station defines a configuration of each subnetwork.

Bitran and Tirupati [33] propose a greedy heuristic solution approach to minimize the work-in-process (WIP), instead of using conventional convex programming method. An extension of Bitran and Tirupati formulation is presented by Boxma et al. [15], which propose a greedy algorithm for machine allocation problem in multi-server open queueing network with exponential inter-arrival and service processes in order to minimize cost of machine allocation but generate undominated solutions. Later, Frenk et al. [34] propose an improved version of the greedy algorithm proposed by Boxma et al. [15].

Connors et al. [4] develop an open queueing network model for performance evaluation of manufacturing systems characterized by the effect of rework and scrap. A marginal allocation procedure to determine the number of tools needed to achieve a target cycle time is designed based on the performance estimates form the queueing network models.

Bispo and Tayur [35] use simulation-based optimization approach and develop expressions for and validate the appropriate Infinitesimal Perturbation Analysis (IPA) derivatives. These derivatives can be used to determine the optimal parameters of managing re-entrant flow lines.

1.2.2 Performance Evaluation Models

We briefly classify the evaluation models as follows: (1) Exact analysis (2) Approximation models (3) Discrete event system simulation.

Exact Analysis

Exact solutions are known by Jackson[5] for open queueing networks with exponential inter-arrival and service time distributions, and probabilistic job routing which described by a routing matrix. The main result is that the solutions have a product form for the stationary multi-dimensional state probability where each product term is the solution of isolated queueing workstations. Kelly[6] extended Jackson's result to networks with more general routings, i.e. deterministic routing, but inter-arrival and service time distributions still follow exponential distributions. Gordon and Newell [7], and Baskett et al. [8] extend the result to closed and mixed queueing networks, respectively.

Approximation methods

The lack of success in obtaining exact solutions for general queueing networks has motivated researchers to develop approximation models to evaluate network performances. There are several known approaches, i.e. diffusion approximation [2][10], mean value analysis[3][12], and parametric decomposition methods [3] [10] [13][14] [15] [16] [17]. Here, we emphasize on parametric decomposition methods.

The decomposition method can be roughly described in three basic steps:

Step 1: Analysis of interaction between workstations.

Step 2: Evaluation of performance measure at each station.

Step 3: Evaluation of performance measure for the whole network.

It is well-known that AT&T Bell Laboratories uses the parametric decomposition methods as the basis of their Queueing Network Analyzer (QNA) software package [13] [14]. QNA describes open queueing network with non-exponential service time, non-Poisson arrival processes, and non-Markovian routing, for which exact analytical techniques are unavailable. Bitran and Tirupati [3] improve the parametric approximation. Segal and Whitt [14] proposed the refined approximation model for re-entrant lines with deterministic routing of products. Conner et al. [4] develop an open queueing network model for performance analysis of semiconductor manufacturing.

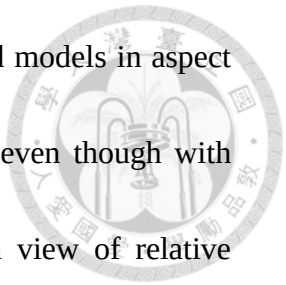
DES Simulation

Currently, the rapid advances in computing power and memory have opened up the opportunities of optimizing by simulation models, which usually called simulation-based optimization. For those complex systems that are intractable to traditional analytical methods, simulation has often been adopted as an evaluation model because it has the advantage of high fidelity and modeling flexibility in coping

with the various system configurations. Examples include the optimization of flow production lines via running a simulation model to find the optimal solution [18] [19]. However, fidelity and flexibility often come with expensive computational cost. Besides, with the stochastic nature of systems, hundreds of Monte Carlo simulation runs are required to reach a statistically significant evaluation.

1.2.3 Selection of Simplified Models for OO

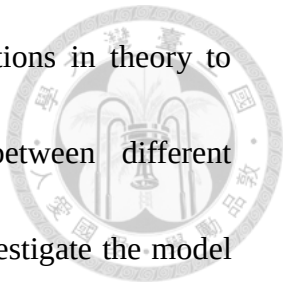
First, simplification is important because it is central to the model selection process. An approach to modelling that is usually suggested to start by building a simple model and then gradually adding details [20], but starting with a simple model implies there are many initial simplifications in the developer's mind. The subsequent process of adding details is the inverse of simplification. Brook et al. [23] studied the simplification in the simulation of manufacturing systems and used an alternative approach to modelling is to start by building a complex model and then try to simplify it. Brook pointed out the advantage of this approach is able to examine model's behavior and indicate which features of the model are important and which are unimportant, and hence the best simplification to perform. Moreover, Johnson et al. [21] develop a procedure for simplifying a detailed model into a fast simulation model that achieves a statistically in distinguishable level of accuracy and precision.



Most of works discuss the accuracy and precision of simplified models in aspect of “value”, and focus on finding a fast enough simplified model even though with endurable biases. But there are few theoretical results discuss in view of relative “orders” provided by simplified models. However, for OO the correlation of ranking among designs by simplified models is much more important than the accuracy of performance of designs.

Xu et al [1] propose an ordinal transformation (OT) framework, which utilizes a simplified model to rank all designs in terms of their approximated performance and then transform the original design space to an ordinal space. There is an illustrative machine allocation example of the OT framework in [1] which uses Jackson network approximation, a model being unity SCVs, as the simplified model of a re-entrant line. However, Kao et al. [24] study the daily target setting and machine allocation problem and [25] investigate the effects of target-induced variability on cycle time performance, these show the significance of variability for cycle time and also expose the deficiency of the model being unity SCVs. Based on the work of Hu [26], it uses queueing network analyzer, a model being heterogeneous SCVs, as the simplified model of a re-entrant line to quickly translate production goals into production flow control parameters.

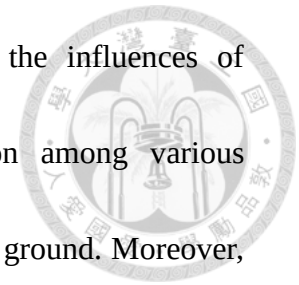
Importantly, both [21] and [23] pointed out the phrase of ‘complexity’ or ‘level of



detail' has not been well-defined, and there are no solid foundations in theory to simplify a system, let alone the theoretical comparisons between different simplifications. Based on the previous works, in this thesis we investigate the model selection problem for OO by the comparison of two simplified models: one is unity SCVs and the other is heterogeneous SCVs. Also, we establish the theoretical analysis of how different variability in simplified models affects ranking and demonstrate the validity of using ranking information in OO by an illustrative experiment.

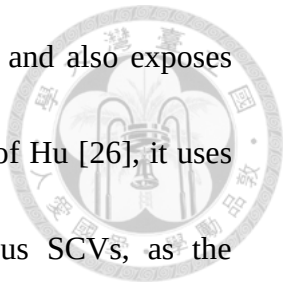
1.3 Scope of Research

Ordinal optimization (OO) focuses on "ranking" in performances among designs instead of their "values" and exploits a goal softening strategy aiming at "good enough" designs with high probability as opposed to an optimal design for sure. Ordinal transformation (OT) proposed by Xu et al. [1] is an OO technique that utilizes a simplified model for perform evaluation and ranking to further reduce computational effort. There are often multiple choices of simplified models for a system that capture different levels of details or aspects. The selection of an appropriate simplified model is a key factor for the effectiveness of OT and OO. Thus, how to select simplified models for ranking and how to analyze the goodness of simplified models are significant and challenging problems for OT and OO.



However, there is little literature to theoretically explore the influences of different simplified models on ranking, since the comparison among various simplified models, however, is often difficult in lack of a common ground. Moreover, because ranking is a relative index instead of an absolute index, the goodness of ranking by a simplified model is also difficult to quantify let alone to analyze.

In this thesis, machine capacity allocation for re-entrant lines, an important engineering optimization problem, is adopted as the conveyor problem to investigate the selection of an appropriate simplified model. For re-entrant lines, queueing networks are commonly used as approximation models, including Jackson network approximation (JNA) and queueing network analyzer (QNA). Both of them are obtained on the same theoretical basis of parametric decomposition method. Parametric decomposition method is a widely used approximation model. It decomposes a queueing network into individual network nodes and use first order and second order parameters, mean and variance, to characterize the stochastic arrival and service processes of each node. Owing to various assumptions of networks, parametric decomposition method has different characterization on variability terms. Xu and Chen et al. [1] use Jackson network as an approximation model of a re-entrant line where assumes the arrival and service processes of each station are all exponentially distributed, a model being unity SCVs. However, [24] shows the



significance of considered variability into cycle time performance and also exposes the deficiency of the model being unity SCVs. Based on the work of Hu [26], it uses queueing network analyzer (QNA), a model being heterogeneous SCVs, as the simplified model of a re-entrant line to quickly translate production goals into production flow control parameters.

Thus, we compare between two simplified models, one being unity SCVs (JNA) and the other being heterogeneous SCVs (QNA), to research how different variability in simplified models affects ranking for OO. It is a meaningful comparison because both these simplified models are developed in the same theoretical basis, and the only difference is the characterization of variability.

A key step in the investigation is the quantification of the goodness of rankings by simplified models. A bound and ranking analysis, BRA, is developed to quantify and analyze the goodness of rankings by simplified models. There are two innovations in BRA: 1) Analyze the upper and lower bounds of simplified models; 2) Derive the probability of correct ranking under the assumption of actual cycle time being uniformly distributed between its upper and lower bounds.

The probability of correct ranking between a pair of designs for a single GI/G/m queue is first studied. With the variation of two QNA approximations, the least variation of their upper bound is derived and this helps obtain a higher probability of

correct ranking α.

In this case, the results and insights from BRA are as follow.

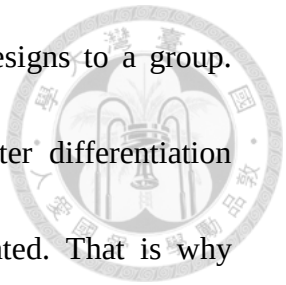
iv) Showed that QNA approximation is bounded by known upper and lower bounds proposed by Kingman, Brumelle and Marshall respectively.

v) Compared with existing literature results, QNA captures the variations of true expected cycle time well because of heterogeneous SCVs but JNA does not.

vi) Obtained a valuable insight from derived α that capturing heterogeneous SCVs benefits the ranking of top designs and improves probability of correct ranking because variability has greater impacts on cycle time while lower utilization.

Based on the above for a single GI/G/m queue, BRA is then extended to general re-entrant lines with multiple workstations. Rank correlation, which measures the concordance of pair-wise comparisons in two quantitative indices, is adopted to quantify the goodness of ranking. In our experiment, α is greater than 0.75 and significantly difference with the probability of 0.5 like tossing a coin.

Finally, a simulation experiment is conducted over a five-station re-entrant line. Simulation results show that rank correlation of QNA always outperforms that of JNA and their difference is especially significant for top designs. This corresponds to the insight obtained from our BRA and discussion about heterogeneous SCVs. Furthermore, we innovatively reverse the logic of ordinal transformation and use true



performance as the ranking index, and then cluster each thirty designs to a group. After grouping, we found that heterogeneous SCVs benefit better differentiation between groups and also make designs in a group better separated. That is why considering heterogeneous SCVs in simplified models improve the rank correlation.

In this thesis, we investigated in aspects of both theory and experiment how selecting simplified models of different variability affects ranking by using re-entrant line capacity allocation as the conveyor problem to compare two simplified models in the same theoretical basis. Finally, we developed the BRA and established theoretical foundations to quantify and analyze the goodness of ranking by simplified models. The contributions of this thesis are as follows.

1. Adopted re-entrant line capacity allocation as the conveyor problem to meaningfully compare simplified models with different level of details, QNA being heterogeneous SCVs and JNA being unity SCVs,
2. Established theoretical foundation to quantify and analyze how selecting simplified models of different variability affects ranking,
3. Derived a lower bound of probability of correct ranking α in the case of single GI/G/m queue with two designs, and the property of α showed that heterogeneous SCVs have greater impacts on top designs,
4. Simulation study demonstrated that heterogeneous SCVs contribute to improve

differentiation between groups and make designs in a group better separated,

5. Simulation result also showed that QNA always outperforms JNA in terms of rank correlation, especially significant for top designs, which corresponds to 3.

1.4 Thesis Organization

The remaining thesis is organized as follows. Chapter 2 defines the machine capacity allocation problem in a re-entrant line and models the re-entrant line into an open queueing network model. Parametric decomposition method is introduced in Chapter 3 and two simplified models with different characterization of variability, QNA and JNA, are also discussed. In Chapter 4, motivated by the deficiencies of traditional simulation-based optimization approaches, ordinal transformation (OT), an OO-based approach, facilitates us significantly reduce the computation effort is described. Then, we analyze the goodness of using QNA or JNA as the simplified model for OT. By bound and ranking analysis (BRA), we investigate the probability of correct ranking in single GI/G/m queue. In Chapter 5, we extend the BRA of single GI/G/m queue to general queueing networks. An experiment of five-workstation re-entrant line with hundreds of designs is conducted in Chapter 6 and the ranking performances of QNA and JNA are compared. Chapter 7 concludes this thesis.

Chapter 2

Conveyor Problem: Re-entrant Line Capacity Allocation

In this chapter, machine capacity allocation problem in re-entrant lines will be defined as our conveyor problem for ordinal optimization. At first, re-entrant lines are characterized by the inclusion of feedback loops that allow products to visit some workstations more than once at different stages of processing. The challenges of machine capacity allocation in re-entrant lines are introduced and the complexity is also investigated. Then, we model the re-entrant lines as an open queuing network (OQN) with multiple product types, shared service workstations, and deterministic routing for each product. At last, the reasons of using re-entrant line capacity allocation as the conveyor problem for ordinal optimization are discussed.

2.1 Problem Description and Complexity Analysis

Re-entrant lines consist of multiple products, shared workstations with several identical machines, and predetermined processing routing through the network for each product. At each workstation, there not only the external arrival flows but also the internal re-entrant flows cause the fierce competition of machine capacity. It is

important to determine how best to allocate service capacity to each workstation so as to optimize various performance measures, such as the total cycle time(or makespan).

In actual manufacturing systems, there are several machine groups with different characteristics. Our research focuses on the machine allocation of one machine group and assume all machines in a machine group are identical and without the quality concern. Even though actual manufacturing systems have various machine groups and the assumption of identical machine here is not realistic, the major effect to this investigation is only the generation of all possible designs.

This scenario is often seen in daily operation planning of flexible manufacturing systems. Managers are responsible for meeting daily production goals and it is important to decide how allocate available machines to each workstation. However, the number of allocation designs depends on how many workstations in the network and how many available machines to be allocated. If there are M workstations and N machines, total number of possible allocation design grows in a combinatorial way. It is known that determining the optimal machine capacity allocations in re-entrant lines is a NP-hard problem. Therefore, how to determine the optimal allocation design of the machine capacity to each workstation in re-entrant lines is a significant and challenging problem and need an efficient methodology to solve [27].

2.2 Mathematical Abstraction of Machine Allocation Problem

Consider a production line illustrated in Figure 2.1. There are M workstations and I products. Machines in one workstation are identical and independent. The process routing of each product among workstations is predetermined. It is assumed that there is an infinite buffer for each workstation. As each workstation can be visited more than once at different processing stages of one product, the line is re-entrant.

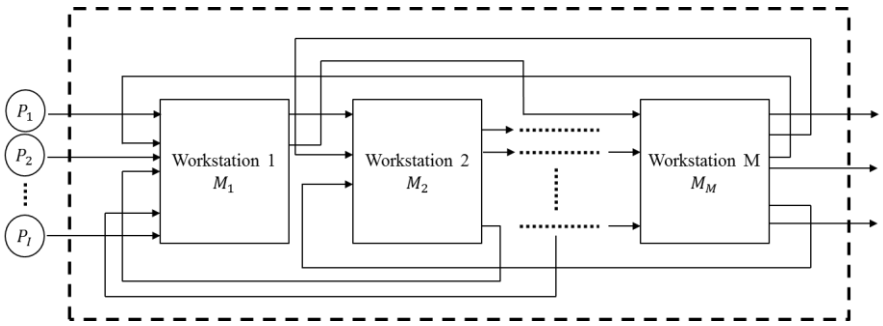
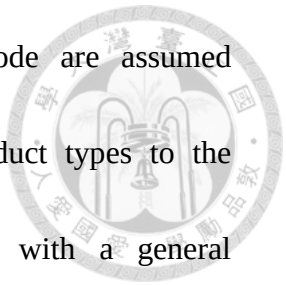


Figure 2.1 Re-entrant Lines

2.2.1 Open Queuing Network Modeling

For analysis of a re-entrant line, we use an open queuing network(OQN) to model the re-entrant line. In OQN modeling, a workstation m of M_m identical machines is represented as a service node of parallel machines with infinite buffer, m=1,2,…,M. The total number of machines is N, $M_1 + M_2 + \cdots + M_M = N$. Each product type i has a total of S_i processing steps, i=1,2,…, I. Let (i,k) be the k-th processing step of type-i product. The process routing of type-i follows a deterministic route $\{(i,k), k = 1,2,\dots,S_i\}$. Step (i,k) is processed by the service node $m_{ik} \in \{1,2,\dots,M\}$. The service time of each node varies with processing steps



and service times of individual product types at a given node are assumed independent of each other. Arrival processes of individual product types to the network are assumed independent and identically distributed with a general distribution. First-Come-First-Serve (FCFS) is the service discipline of each node.

The re-entrant line is modeled as an open queuing network as follows:

(1) *Multiple server nodes*: Each workstation with multiple machines is modeled as one multiple-server node.

(2) *External arrivals*: Products loaded into the re-entrant line for processing constitute the arrival to the network.

(3) *General arrival processes*: Product arrival processes to each node are described by the specific probability distribution of product inter-arrival times. Inter-arrival times of different product types are assumed individual and identically distributed.

(4) *General service processes*: Service processes are specified by general distributions. Service time distribution varies with different processing steps of different products. And the service time distribution at a node is independent of the other.

(5) *Deterministic routing*: Processing routings of individual product types are fixed.

(6) *Service discipline*: The service discipline of each node is FCFS.

2.2.2 Formulation: Nonlinear Integer Programming

Let us formalize the machine capacity allocation problem in re-entrant lines. Suppose we have M workstations, N machines to be allocated, and I product types, our objective is to find an optimal machine allocation design which minimizes the average of mean cycle time of each product.

Definition: Machine allocation design

A machine allocation design is specified by the number of machines in each workstation and represented by a set with M elements. The i^{th} element in this set stands for that there are M_i machines to be allocated at workstation i. And every machine must be allocated to a workstation. Therefore, the machine allocation design indexed by k can be written as $D_k = \{M_1^k, M_2^k, \dots, M_M^k\}$ where $M_1^k + M_2^k + \cdots + M_M^k = N$.

Definition: Machine capacity allocation problem

Our objective is to find an optimal machine capacity allocation design $D_{k^*} = \{M_1^{k^*}, M_2^{k^*}, \dots, M_M^{k^*}\}$, to minimize the average of mean cycle time of each product, where

$$k^* = \underset{k}{\operatorname{argmin}} \frac{1}{I} \sum_{i=1}^{I} \mathrm{MCT}_i(D_k), D_k \in D$$

and $\mathrm{MCT}_i(D_k)$ denotes the mean cycle time of product type i under a specific allocation design D_k.

Because $\mathrm{MCT}_i(D_k)$ is nonlinear with respect to D_k, each element in D_k is the number of machines in each workstation and must be an integer. Therefore, this machine capacity allocation problem is formulated as a nonlinear integer programming problem.

2.3 Conveyor Problem for Ordinal Optimization

For machine capacity allocation problem in re-entrant lines, the original solution space is discrete and high-dimensional, which may have multiple local optimums and be hard to search in such a solution space. The re-entrant behavior poses the challenge to analyze the effects of re-entrant flows since that the interactions or dependencies among workstations are hard to describe and also difficult to exactly analyze the system in re-entrant lines [28]. If we would like to obtain accurate evaluation of a machine allocation design in re-entrant lines, we have to exploit the discrete event simulation (DES) but discrete event simulation suffers from the high computational cost. Due to its nature of complexity and combinatorial solution space, machine capacity allocation problem in re-entrant lines is extremely complicated and the use of simplified models is necessary. Thus, this re-entrant line capacity allocation problem is suitable to be the conveyor problem for investigating the model selection of ordinal optimization.

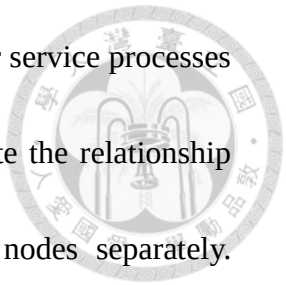
Chapter 3

Parametric Decomposition Method for OQN

In this chapter, we introduce the parametric-decomposition approximation method of OQNs and the related extensions to improve its accuracy of approximations. According to the parametric decomposition method, there are two simplified models developed for OQNs, one is with unity SCVs and the other is with heterogeneous SCVs. The simplified model with unity SCVs is essentially based on the assumption of Jackson networks, and also called “Jackson network approximation (JNA)”. The simplified model with heterogeneous SCVs is developed by Whitt[13], and names as “Queueing Network Analyzer (QNA)”. It is a meaningful comparison because both of them are developed in a same theoretical basis, and the only difference between JNA and QNA is characterization of SCVs. Then, node-level measures and system-level measures of simplified models can be obtained. At last, we summarize how the comparison of these two simplified models relates to the model selection problem in OO.

3.1 Introduction

The parametric-decomposition approximation method first proposed by Reiser and Kobayashi[10] is a useful method to analyze the steady-state performance of



OQNs. The main idea is to approximately characterize the arrival or service processes of each node by two parameters: mean and variability, approximate the relationship among nodes in the network, and then analyze the individual nodes separately. Parametric decomposition method treats each node as an independent GI/G/m queue with m identical machines, infinite buffer for waiting, FCFS discipline and using two parameters to describe its general inter-arrival time distribution and general service time distribution respectively.

A standard decomposition approximation assumes Markovian routing of products after the service at each node in an OQN, which is the basic property of Jackson network. Bitran and Tirupati [14] observed that the SCV of departure of a product calculated under the assumption of Markovian routing is distorted by the presence of other products at a node. Bitran and Tirupati proposed the approximation of the SCV of inter-departure times at each node for each product and showed that the SCV of inter-departure times can be refined as the sum of two terms: the first reflects the queuing effect at the node, and the second captures the effect caused by inter-arrival time distributions of other products. Then, Segal and Whitt[16] proposed the refined approximation of the SCV of inter-departure times for aggregated product flows in re-entrant lines with deterministic routing of products. Numerical results in [16] showed that the refined approximations have relative errors of about 5-20% in

estimating the inter-departure SCVs.

3.2 Class Aggregation

First recall that the notations of a multiple-product OQN model. Each product type i has a total of S_i processing steps, i=1,2,…, *I*. Let (i,k) be the k-th processing step of type-i product. The process routing of type-i follows a deterministic route $\{(i,k), k = 1,2,\dots,S_i\}$. Step (i,k) is processed by the service node $m_{i,k} \in \{1,2,\dots,M\}$. Then, multiple types of products are aggregated into a single product in the OQN model. The aggregation procedure follows the work of Whitt[13] and summarizes in the following.

Define some notations:

λ_i: external arrival rate of product type i;

C_i^2: inter-arrival time SCV of product type i;

$\delta_{ij} = \begin{cases} 1, \text{product type i externally entering the network at node j.} \\ 0, \text{otherwise.} \end{cases}$

λ_E : aggregate external mean arrival rate;

C_E^2 : inter-arrival time SCV of aggregate external arrivals;

λ_{mn} : aggregate mean arrival rate from node m to node n;

τ_m : aggregate mean service time at node m;

C_{sm}^2 : aggregate service time SCV at node m;

τ_{ik} : mean service time at k-th step of product type i;

C_{ik}^2 : service time SCV at k-th step of product type i;

Q=$\{q_{mn}\}$: routing matrix, and q_{mn} is ratio of routings from node m to node n;

$\mathbf{1}$H(x): an indicator function of the set H, $\mathbf{1}\text{H(x)} = \begin{cases} 1, if\ x \in H. \\ 0, otherwise. \end{cases}$

First, we obtain the aggregate external arrival rates by adding up mean arrival rate of each product at node n,

$$\lambda_{En} = \sum_{i=1}^{I} \lambda_i \delta_{in}. \tag{3.1}$$

As the external arrivals of products are independent, the inter-arrival time SCV of aggregate external arrivals is

$$C_{En}^2 = \sum_{i=1}^{I} C_i^2 \frac{\lambda_i \delta_{in}}{\lambda_{En}}. \tag{3.2}$$

The aggregate mean arrival rate from node m to node n is

$$\lambda_{mn} = \sum_{i=1}^{I} \sum_{k=1}^{S_i - 1} \lambda_i \mathbf{1}\{m_{i,k} = m, m_{i,k+1} = n\}, \forall \text{m} \neq \text{n}, \text{m}, \text{n} = 1,2, \dots, \text{M}. \tag{3.3}$$

And the ratio of routings from node m to node n can be calculated as

$$q_{mn} = \frac{\lambda_{mn}}{\sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{m_{i,k} = m\}} \tag{3.4}$$

The aggregate service time of a step at node m is composed of service times of each step of each product that routed to be served by node m. The aggregate mean service time at node m,

$$\tau_m = \frac{\sum_{i=1}^{I} \sum_{k=1}^{S_i} \tau_{ik} \lambda_i \mathbf{1}\{m_{i,k} = m\}}{\sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{m_{i,k} = m\}}\ , \tag{3.5}$$

and the corresponding SCV at node m is

$$C_{sm}^2 = \frac{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_{ik}^2(C_{ik}^2+1)\lambda_i \mathbf{1}\{m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_m^2\lambda_i \mathbf{1}\{m_{i,k}=m\}} - 1. \quad (3.6)$$

3.3 Parametric Decomposition Method

The parametric-decomposition approximation method first proposed by Reiser and Kobayashi[10] is a useful method to analyze the steady-state performance of OQNs. The main idea is to approximately characterize the arrival or service processes of each node by two parameters: mean and variability, approximate the relationship among nodes in the network, and then analyze the individual nodes separately. The decomposition approximation can be comprised of the basic three steps:

(1) analysis of the relationships between arrival, service, and departure processes at a node;

(2) analysis of the dependency among nodes of the network;

(3) approximation of performance measures of the whole network.

Define more notations for each node in OQN:

M_m : number of machines in node m;

λ_{am} : mean total arrival rate to node m;

C_{am}^2 : inter-arrival time SCV at node m;

C_{dm}^2 : inter-departure time SCV at node m;

C_{mn}^2 : inter-departure time SCV for the flow transiting from node m to node n;

Because of the flow relationship among nodes in the network, the total arrival rate of node n is the summation of external arrivals to node n and internal arrivals to node n, represented as

$$\lambda_{an} = \sum_{i=1}^{I} \lambda_i \delta_{in} + \sum_{m=1}^{M} \lambda_{mn} = \sum_{i=1}^{I} \lambda_i \delta_{in} + \sum_{m=1}^{M} \lambda_{am} q_{mn} \quad (3.7)$$

Where δ_{in}=1 if product type i entering the network at node n. Equation (3.7) is known as the ***traffic rate equations*** with λ_{mn} as defined by equation (3.3). In equation (3.7), there are M equalities with M unknown variables $\{\lambda_{an}, n = 1,2,\dots,M\}$, so the λ_{an} can be solved by these M simultaneous equations. After obtaining the total arrival rate to node m, the average utilization of node n can be calculated by,

$$\rho_n = \frac{\lambda_{an}\tau_n}{M_n} \quad (3.8)$$

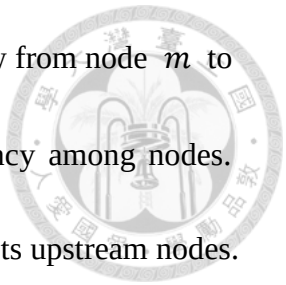
Segal and Whitt [14] pointed out that the resulting utilization of each node is exact. To ensure the stability of networks, the average utilization should be limited below the capacity of the line,

$\rho_n < 1, n = 1,2,\dots,M.$

By utilizing the procedure of Whitt[13], the inter-arrival time SCV of an aggregate arrival process can be obtained as

$$C_{an}{}^2 = 1 - \widetilde{\omega_n} + \widetilde{\omega_n} \frac{\lambda_{En} C_{En}{}^2}{\lambda_{an}} + \widetilde{\omega_n} \sum_{m=1}^{M} \frac{\lambda_{mn}}{\lambda_{an}} C_{mn}{}^2, \quad (3.9)$$

where $\widetilde{\omega_n} = [1 + 4(1-\rho_n)^2(v_n - 1)]^{-1}$ and $v_n = \left[\sum_{m=1}^{M} (\frac{\lambda_{mn}}{\lambda_{an}})^2\right]^{-1}$.



However, the approximate variability parameter of the sub-flow from node m to node n, ${C_{mn}}^2$, is related to the routing criteria and the dependency among nodes. Note that the arrivals of a node are aggregated by the departures of its upstream nodes. And the departures out of a node is split into several sub-flows of different downstream nodes according to the routing matrix $Q=\{q_{mn}\}$. Thus different routing criteria will influence the characteristics of nodes and the properties of networks. In the following we discuss two kinds of routing criteria, Markovian routing and deterministic routing, and obtain the approximate variability parameter of the sub-flow from node m to node n, ${C_{mn}}^2$, under different routing criteria.

3.3.1 Markovian Routing

The Markovian routing means that each product completes service at node m and proceeds to node n with probability q_{mn}, which is independent of the current state and history of the network. The routing matrix $Q=\{q_{mn}\}$ interprets as the independent probabilities of going to node n after completed at node m. The approximate variability parameter of the sub-flow from node m to node n, ${C_{mn}}^2$, under Markovian routing is proposed by Whitt[13],

$${C_{mn}}^2 = q_{mn}{C_{dm}}^2 + (1 - q_{mn}) \tag{3.10}$$

Because of the independency of Markovian routing, if all the external arrival

processes are Poisson and products follow Markovian routings, then all internal arrival processes are also Poisson. If we assume that there is a single product and service time distributions are exponential, then parametric decomposition method is consistent with Jackson network and most significantly provides exact performance measures of Jackson network [13].

3.3.2 Deterministic Routing

As the observation of Bitran and Tirupati in [14] that if in the multiple product types, their arrivals do not follow Poisson distributions and the routings are deterministic, the use of Equation (3.9) to describe the approximate variability parameter of the sub-flow from node m to node n may not perform well due to the independency assumption of Markovian routing. Bitran and Tirupati identified the distortion in the SCV of a given product because of the presence of other products and refer to this distortion as the interference effect. Following the work of Bitran and Tirupati, Segal and Whitt proposed the refined calculation of the approximate variability parameter of the sub-flow from node m to node n,

$$C_{mn}{}^2 = q_{mn}C_{dm}{}^2 + (1 - q_{mn})q_{mn}C_{am}{}^2 + (1 - q_{mn})^2 C_{em}{}^2, \tag{3.11}$$

where C_{em}^2 is an average of the external arrival-process variability parameters,

$$C_{em}{}^2 = \sum_{i=1}^{I} C_i{}^2 \left(\frac{\sum_{k=1}^{S} \lambda_i 1\{(i,k):m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S} \lambda_i 1\{(i,k):m_{i,k}=m\}} \right). \tag{3.12}$$

And Whitt[13] suggested that the SCV of the departure process at node m by

$$C_{dm}{}^2 = 1 + \left(1 - \rho_m{}^2\right)\left(C_{am}{}^2 - 1\right) + \frac{\rho_m{}^2\left(\max\left\{C_{sm}{}^2, 0.2\right\} - 1\right)}{\sqrt{M_m}}. \quad (3.13)$$

The experiments conducted in [14] and [16] if the network is multiple-product and deterministic routing, then apply Equation (3.11) to capture the interaction among stations instead of using Equation (3.10). Numerical results in [16] showed the refined approximations have relative errors of about 5-20% in estimating the inter-departure SCVs.

In this thesis, we focus on re-entrant lines with multiple products, deterministic routing, general (non-Poisson) arrivals, and general service time distributions. Therefore, instead of Equation (3.10), we approximate the variability parameter of the sub-flow from node m to node n, $C_{mn}{}^2$, by Equation (3.11). By substituting Equation (3.11) into Equation (3.9), $C_{an}{}^2$ becomes

$$C_{an}{}^2 = \alpha_n + \sum_{m=1}^{M} C_{am}{}^2 \beta_{mn} \quad (3.14)$$

where

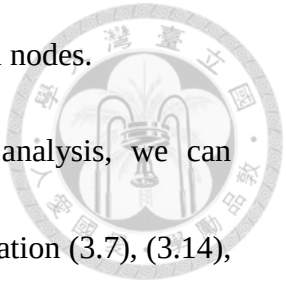
$$\alpha_n = 1 + \widetilde{\omega_n}\left\{\frac{\lambda_{En} C_{En}{}^2}{\lambda_{an}} - 1 + \sum_{m=1}^{M}\left(\frac{\lambda_{mn}}{\lambda_{an}}\right)\left[q_{mn}\rho_m^2 X_m + (1 - q_{mn})^2 C_{em}{}^2\right]\right\},$$

$$\beta_{mn} = \widetilde{\omega_n}\left(\frac{\lambda_{mn}}{\lambda_{an}}\right) q_{mn}[(1 - \rho_m^2) + (1 - q_{mn})],$$

$$\widetilde{\omega_n} = [1 + 4(1 - \rho_n)^2(v_n - 1)]^{-1} \text{ and } v_n = \left[\sum_{m=1}^{M}\left(\frac{\lambda_{mn}}{\lambda_{an}}\right)^2\right]^{-1}$$

$$X_m = 1 + \frac{\left(\max\left\{C_{sm}{}^2, 0.2\right\} - 1\right)}{\sqrt{M_m}}.$$

Equation (3.14) is known as the set of ***traffic variability equations***, which



approximate the relationship among the inter-arrival time SCV of all nodes.

Finally, from the parametric-decomposition approximation analysis, we can obtain the four parameters ($\lambda_{am}, C_{am}^2, \tau_m, C_{sm}^2$) of each node by Equation (3.7), (3.14), (3.5), and (3.6) respectively to describe the characteristics of each node and approximate the performance measures of node-level and system-level as follows.

3.4 Performance Measures

Once we obtain the arrival and service parameters, ($\lambda_{am}, C_{am}^2, \tau_m, C_{sm}^2$), of each node, we can exploit them to calculate many performance measures. In this section, we would describe how to approximate the performance measures by utilizing the results of the parametric-decomposition approximation analysis. Assume that all service nodes are highly utilized, which is usually realistic in industry.

3.4.1 Node Level Measures

According to Whitt [13][16], the expected waiting time approximation of node m with parameter ($\lambda_{am}, C_{am}^2, \tau_m, C_{sm}^2$) as a $GI/G/M_m$ queue based on the heavy-traffic limit theorem is

$$E[W_m(GI/G/M_m)] = \frac{C_{am}^2 + C_{sm}^2}{2} E[W_m(M/M/M_m)] \quad (3.15)$$

where $E[W_m(M/M/M_m)]$ is the expected waiting time for a $M/M/M_m$ queue,

defined as

$$E[W_m(M/M/M_m)] = \frac{\tau_m(M_m\rho_m)^{M_m}\pi_m(0)}{M_m(1-\rho_m)^2 M_m!}$$

Where

$$\pi_m(0) = \left[\sum_{k=0}^{M_m-1}\frac{(M_m\rho_m)^k}{k!} + \frac{(M_m\rho_m)^{M_m}}{(1-\rho_m)M_m!}\right]^{-1}.$$

The expected cycle time of node m is the sum of expected service time and expected waiting time,

$$E[CT_m(GI/G/M_m)] = \tau_m + E[W_m(GI/G/M_m)] \quad (3.16)$$

From the Little formula, the expected number in node m is

$$E[N_m] = \lambda_{am} \times E[CT_m(GI/G/M_m)],$$

and the expected number in the queue of node m is

$$E[Q_m] = \lambda_{am} \times E[W_m(GI/G/M_m)].$$

3.4.2 System Level Measures

The expected cycle time of aggregate flow going through the network is a function of average number of visits per product and expected cycle time of individual nodes in the network. Average number of visits per product to node m is

$$V_m = \frac{\lambda_{am}}{\sum_{n=1}^{M}\lambda_{En}}, for\ m = 1,2,\dots M. \quad (3.17)$$

Therefore, the expected cycle time of going through the network is

$$E[T] = \sum_{m=1}^{M} V_m(\tau_m + E[W_m]). \quad (3.18)$$

The total number of jobs in the whole network can be also obtained,

$$E[N] = E[N_1] + E[N_2] + \cdots + E[N_M],$$

$$= \sum_{m=1}^{M} \lambda_{am}(\tau_m + E[W_m]),$$

$$= \lambda_E \sum_{m=1}^{M} V_m(\tau_m + E[W_m]),$$

$$= \lambda_E \times E[T]. \quad (3.19)$$

Then, expected cycle time of individual products can be obtained because the parametric-decomposition approximation views each workstation as an independent node, the expected total cycle (or sojourn) time for a product is the summation of expected cycle time of each node in the routing of that product. The expected cycle time for product i is

$$E[T_i] = \sum_{k=1}^{S_i}(\tau_{ik} + E[W_{m_{i,k}}]). \quad (3.20)$$

where $E[W_{m_{i,k}}]$ is the expected waiting time of the k^{th} processing step of product i at node $m_{i,k}$. And the cycle time variance for product i can be calculated by

$$Var[T_i] = \sum_{k=1}^{S_i} \tau_{ik}^2 C_{ik}^2 + \sum_{k=1}^{S_i} Var[W_{i,k}]. \quad (3.21)$$

3.5 Two Simplified Models for Re-entrant Line: QNA and JNA

We discuss two simplified models developed according to the parametric decomposition method: one is queueing network analyzer (QNA) and the other is

Jackson network approximation (JNA). JNA assumes the OQN follows Markovian routing using Equation (3.10) and inter-arrival time and service time of each node are exponentially distributed (CV=1). Thus, each node is treated as an independent M/M/m queue. Under these assumptions, the decomposition approximation is equivalent to Jackson network approximation. JNA is a simplified model of OQN which uses one parameter (mean) to characterize each node with unity SCVs.

Unlike JNA, QNA is a free software package which first developed at Bell Laboratories to calculate approximate performance measures for general (non-Markov) open queuing network [13]. QNA describes the variability parameter of network flows by Equation (3.11), (3.12), and (3.13), so each node has its own specific characterization of variability. Thus, QNA is another simplified model of OQN, which uses two parameters (mean and SCV) to characterize each node with heterogeneous SCVs. Most importantly, JNA can be regarded as a special case of QNA while each node assumes unity SCV.

Both QNA and JNA are mathematical models developed based on the same theoretical basis, parametric decomposition method. The major difference between these two simplified models is the characterization of variability parameter. It is a meaningful comparison to investigate how modeling heterogeneous variability in a simplified model affects ranking for ordinal optimization.

Chapter 4

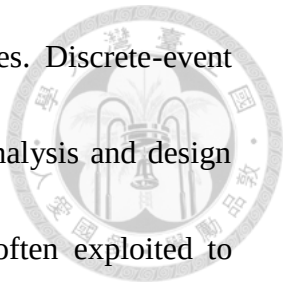
Ordinal Transformation and BRA

In this chapter, we introduce the ordinal transformation (OT) which exploits a simplified model to quickly determine rough performance of designs and their relative orders instead of finding accurate performance. The goodness of ranking by adopting a simplified model for performance approximation is quantified and analyzed in terms of rank correlation.

Queueing network analyzer (QNA) is mainly investigated in the following analysis because QNA utilizes both mean and variance to characterize network flows and JNA is a special case of QNA while assumes unity SCVs. Due to no analytical solution to mean cycle time performance of a general queueing network, we develop a bound and ranking (B&R) analysis to investigate the relation between the bounds of true and approximated performances, and analyze the probability of correctly ranking by a simplified model under some assumptions of true performance between the bounds. In this chapter, we use BRA to take the first step to analyze single GI/G/m queue with 2 designs.

4.1 Ordinal Transformation

In many complex optimization problems, the objective function is a black box



and seems impossible to apply traditional optimization approaches. Discrete-event system (DES) simulation becomes increasingly important in the analysis and design of such complex systems. Simulation optimization methods are often exploited to tackle such type of problems. For those simulation optimization approaches, they have a detailed simulation model to obtain accurate performance measure of the system but such a detailed simulation model usually has high computation cost and may be very time-consuming. With a large-scale design space, it is impossible to evaluate all designs by a detailed simulation model and find the optimal design. Besides detailed DES simulation model, there are actually some simplified (or approximation) models for complex systems, i.e. QNA for OQN, whose computation costs can be ignored in comparison with DES simulation. Such simplified models are much faster but usually have large biases between actual performance measures.

For the sake of finding the optimal design, relative ranking orders among designs are much important than exact differences between their performances. If we find a *good* simplified model whose relative ranking orders among designs are highly correlated with actual ranking orders, then we can make use of such a simplified model to improve the efficiency of searching the optimal design. That is the main idea of OT. In the following, we investigate the basic two steps of OT:

(1) Ranking in terms of approximation by a simplified model

(2) Transform original solution space into ordinal space according to the rankings determined in step (1)

4.1.1 Ranking in terms of Approximations by Simplified Model

We first define some notations below.

D: a design space;

n: total number of designs in D, i.e., $n = |D|$;

x: a design in design space D;

x^*: the optimal design (according to the detailed DES simulation);

$f(x)$: performance of design $x \in D$ evaluated by detailed DES simulation model f(˙);

$g(x)$: performance of design $x \in D$ evaluated by the simplified model g(˙);

$\delta(x)$: $(x) - g(x)$

$F(x)$: ordinal rank of a design $x \in D$ based on $\{f(y), \forall y \in D\}$ in ascending order;

$G(x)$: ordinal rank of a design $x \in D$ based on $\{g(y), \forall y \in D\}$ in ascending order;

$F^{-1}(i)$: the i^{th} best design in D according to $\{f(y), \forall y \in D\}$;

$G^{-1}(i)$: the i^{th} best design in D according to $\{g(y), \forall y \in D\}$;

We represent the relationship between the detailed model and the simplified model by

$$f(x) = g(x) + \delta(x)$$

where $f(x)$ is the accurate performance measures evaluated by detailed model, $g(x)$ is the performance measures approximated by simplified model, and $\delta(x)$ is the bias term. Actually, how accurate the performance measure of the simplified model is not significant because the idea of OT focuses on the ranking orders among all designs. Therefore, instead of obtaining a simplified model with small $\delta(x)$ (accurate performance measures), we would like to find a simplified model whose relative ranking orders among designs in term of an approximate performance measure by the simplified model are highly correlated with actual ranking orders. In short, we desire a simplified model whose rankings among designs are highly correlated with actual rankings,$G(x) \sim F(x), \forall x \in D$, rather than high accuracy in performance, $f(x) \sim g(x), \forall x \in D$. The definition of evaluation between rankings will be defined in subsection 4.1.3.

Next is a small numerical example to illustrate the basic idea and potential benefits of OT. The detailed model is

Example 4.1

$$f(x) = sin^4(2x) - 3sin^3(2x) + sin^2(2x) + 4$$

$$= [sin^2(2x) + sin(2x) + 1][sin^2(2x) - 4sin(2x) + 4],$$

and the simplified model is

$$g(x) = sin^2(2x) - 4sin(2x) + 4$$

The possible designs are $x \in D = \{-0.5, -0.25, 0, \ldots, 2.5\}$, and $n = |D| = 13$. We show the performance measures by f and g in Figure 4.1 and Table 4.1. We can observe that the simplified model, $g(x)$, approximately captures relative performance among designs even through some biases are significant, for example, $f(0.75)=3.007$ and $g(0.75)=1.005$ but both lead to design $x=0.75$ as the best.

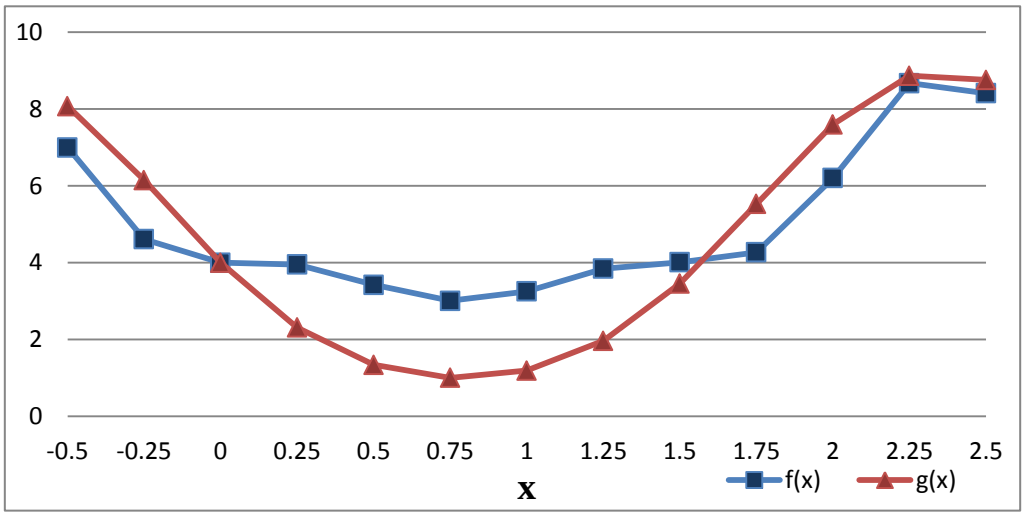


Figure 4.1 An illustrative example of OT

x	-0.5	-0.25	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5
F(x)	11	9	6	5	3	1	2	4	7	8	10	13	12
G(x)	11	9	7	5	3	1	2	4	6	8	10	13	12

Table 4.1 Ranking order among designs of Example 4.1

In Table 4.1, in spite of the large biases, ordinal rank of design x based on $g(x)$ is almost the same as the ordinal rank of design x based on $f(x)$, $F(x) \sim G(x)$.

4.1.2 Transformation to Ordinal Space

We utilize a simplified model to quickly approximate the performances of all designs and rank designs according to their approximated performances. With ranking orders in terms of approximations, we transform the original solution space D to an ordinal space, so this kind of transformation is called ordinal transformation (OT). OT substitutes the original space, x, by the ordinal space of the simplified model, $G(x)$. And we use the numerical results in 4.1.1 to demonstrate OT in Figure 4.2.

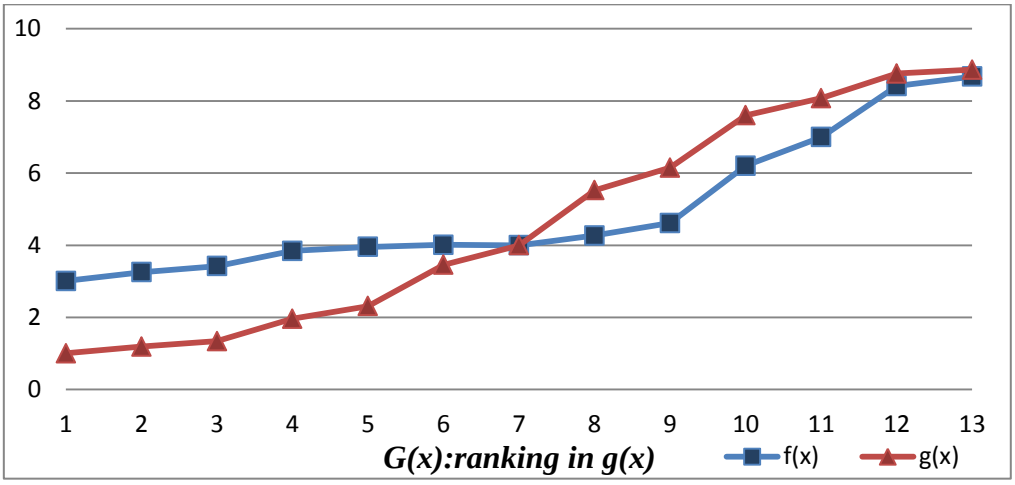


Figure 4.2 Transformation to ordinal space

In the ordinal space, designs with similar performances are ranked and positioned nearby together and either better or worse designs are easily differentiated. Thus, after OT, we can search the optimum in a better space, which increases the efficiency of subsequent optimization processes.

4.1.3 Performance Index: Rank Correlation

The benefit of OT crucially depends on the quality of a simplified model but for

ordinal transformation there is not a standard index to quantify the quality of a simplified model. Here we introduce a meaningful index, rank correlation, to quantify the goodness of ranking performance of a simplified model. Rank correlation is first developed by Kendall [46] to measures the similarity of the orderings of data when ranked by each of the quantities. Rank correlation is a statistic of pair-wise comparisons which corresponds to the idea of OT which comparing the relative order among designs, rank correlation is therefore adopted to measure the concordance of pair-wise comparisons in true and approximated performances.

Definition: Kendall rank correlation coefficient

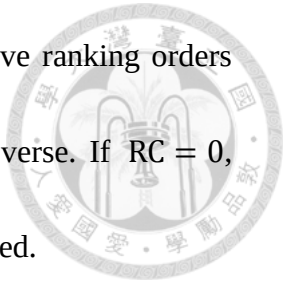
There are N designs, labeled as $x_1, x_2, \dots, x_N$. Let $\big(F(x_i), G(x_i)\big)$ be a rank observations of design x_i in the detailed model and the simplified model. Any pair of observation $\big(F(x_i), G(x_i)\big)$ and $\big(F(x_j), G(x_j)\big)$ are *concordant*, if both $F(x_i) > F(x_j)$ and $G(x_i) > G(x_j)$ or if both $F(x_i) < F(x_j)$ and $G(x_i) < G(x_j)$. They are *discordant*, otherwise.

The Kendall rank correlation coefficient is defined as:

$$\mathrm{RC} = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{\text{Total number of pairs} = \frac{1}{2}N(N-1)}$$

The denominator is the total number of pairs, so the coefficient must be in the range, $-1 \le \mathrm{RC} \le 1$. If $\mathrm{RC} = 1$, relative ranking orders among designs in both the detailed

and simplified model are completely the same. If $RC = -1$, relative ranking orders among designs in the detailed and simplified model are totally reverse. If $RC = 0$, then the rankings in the detailed and simplified model are uncorrelated.



4.2 BRA of QNA and JNA in single GI/G/m queue

Now consider using QNA and INA as simplified models for ranking capacity allocation designs over their mean cycle time performance. It has been pointed out in section 3.5 that JNA is as a special case of QNA, where the inter-arrival and service times of each node assume unity SCVs. In the following discussions of this sub-section, we focus on the analysis of goodness of ranking by using QNA as a simplified model.

Bound analysis investigates the relation between the bounds of true and approximated performance. Under some assumptions of the distribution of true performance, ranking analysis focuses on the conditional probability of correctly ranking given approximated performances. We start our BRA to analyze the probability of correctly ranking by using QNA evaluations from analyzing the case of single GI/G/m queue with 2 designs. BRA is mainly composed of two analyses: (1) bound analysis and (2) ranking analysis.

4.2.1 Bound Analysis of QNA and JNA

We begin with the well-known upper bound of expected waiting time of single GI/G/m queue derived by Kingman [30]:

$$E[\mathrm{WT}_{\mathrm{GI/G/m}}] \leq \frac{{C_a}^2+m^2\rho{C_s}^2+(m-1)\rho^2}{2\lambda(1-\rho)} \tag{4.1}$$

The lower bound of waiting time from Brumelle and Marchal [31][32] is

$$E[\mathrm{WT}_{\mathrm{GI/G/m}}] \geq \frac{\rho^2{C_s}^2-\rho(2-\rho)}{2\lambda(1-\rho)} - \frac{(m-1)({C_s}^2+1)\tau}{2m} \tag{4.2}$$

QNA approximation of expected waiting time $\mathrm{WT}_{\mathrm{QNA}}$ is (Eq. 3.15)

$$E[WT]_{\mathrm{QNA}} \cong \frac{{C_{ai}}^2+{C_{si}}^2}{2} E\left[\mathrm{WT}_{\mathrm{M/M/m}}\right] = \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\mathrm{C}(m,\lambda,\tau)}{m/\tau-\lambda}\right)$$

We know that the cycle time consists of waiting time and service time, so the upper bound of expected cycle time of one GI/G/m queue is

$$UB[\mathrm{ECT}_{\mathrm{GI/G/m}}] = \frac{{C_a}^2+m^2\rho{C_s}^2+(m-1)\rho^2}{2\lambda(1-\rho)} + \tau \tag{4.3}$$

The lower bound of expected cycle time of one GI/G/m queue is

$$LB[\mathrm{ECT}_{\mathrm{GI/G/m}}] = \frac{\rho^2{C_s}^2-\rho(2-\rho)}{2\lambda(1-\rho)} - \frac{(m-1)({C_s}^2+1)\tau}{2m} + \tau \tag{4.4}$$

And, QNA approximation of expected cycle time is

$$E[CT]_{\mathrm{QNA}} = \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\mathrm{C}(m,\lambda,\tau)}{m/\tau-\lambda}\right) + \tau \tag{4.5}$$

In the following, we prove that the upper bound and lower bound of a GI/G/m queue are also the upper bound and lower bound of QNA approximations in terms of expected cycle time.

Theorem 4.1: The upper bound of expected cycle time of a GI/G/m queue is also the upper bound of expected cycle time of QNA approximation.

Proof :

The expected cycle time of QNA approximation, $E[CT]_{\text{QNA}}$ is

$$E[CT]_{\text{QNA}} = \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\text{C}(m,\lambda,\tau)}{m/\tau-\lambda}\right)+\tau$$

where $\text{C}(m,\lambda,\tau) = \left[\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}\right]\left[\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}+\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}\right]^{-1}$.

As $\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} > 0$, the denominator, $\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}+\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}$, must be larger than the nominator, $\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}$. So, $\text{C}(m,\lambda,\tau) < 1$.

Because of $\text{C}(m,\lambda,\tau) < 1$ and $\frac{1}{m/\tau-\lambda} = \frac{1}{\frac{m}{\tau}(1-\rho)} = \frac{\lambda\tau}{\lambda m(1-\rho)} = \frac{\rho}{\lambda(1-\rho)}$,

$$\frac{{C_a}^2+{C_s}^2}{2}\left(\frac{1}{m/\tau-\lambda}\right)+\tau = \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho}{\lambda(1-\rho)}\right)+\tau > E[CT]_{\text{QNA}}$$

The difference is

$$UB[\text{ECT}_{\text{GI/G/m}}] - \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho}{\lambda(1-\rho)}\right)-\tau = \frac{(1-\rho){C_a}^2+(m^2-1)\rho{C_s}^2+(m-1)\rho^2}{2\lambda(1-\rho)} > 0,$$

because there is at least one machine in each workstation, $m \geq 1$, and the utilization of each workstation is limited to smaller than one to maintain the stability of the network. So,

$$UB[\text{ECT}_{\text{GI/G/m}}] > \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho}{\lambda(1-\rho)}\right)+\tau > E[CT]_{\text{QNA}}$$

Q.E.D.

To discuss the lower bound of QNA approximation, let us derive three lemmas regarding function $\text{C}(m,\lambda,\tau)$ in QNA approximation.

Lemma 4.1 : $C(m,\lambda,\tau) > \rho^m$ when m≥ 1 and $1 > \rho = \frac{\lambda\tau}{m} > 0$.

Proof :

$$C(m,\lambda,\tau) = \frac{\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}+\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}} = \frac{\frac{1}{1-\rho}}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m}+\frac{1}{1-\rho}}$$

$$\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m} = \frac{(m\rho)^{m-1}}{(m-1)!}\frac{m!}{(m\rho)^m} + \frac{(m\rho)^{m-2}}{(m-2)!}\frac{m!}{(m\rho)^m} + \cdots + \frac{m!}{(m\rho)^m}$$

$$= \frac{m}{m\rho} + \frac{m}{m\rho}\frac{m-1}{m\rho} + \frac{m}{m\rho}\frac{m-1}{m\rho}\frac{m-2}{m\rho} + \cdots + \frac{m!}{(m\rho)^m}$$

$$< \frac{1}{\rho} + \frac{1}{\rho^2} + \frac{1}{\rho^3} + \cdots + \frac{1}{\rho^m} = \sum_{k=1}^{m}\frac{1}{\rho^k} = \frac{\frac{1}{\rho}\left[1-\left(\frac{1}{\rho}\right)^m\right]}{1-\frac{1}{\rho}} = \frac{\left(\frac{1}{\rho}\right)^m-1}{1-\rho}$$

$$\because \frac{\left(\frac{1}{\rho}\right)^m-1}{1-\rho} > \sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m}$$

$$\therefore C(m,\lambda,\tau) > \frac{\frac{1}{1-\rho}}{\frac{\left(\frac{1}{\rho}\right)^m-1}{1-\rho}+\frac{1}{1-\rho}} = \frac{1}{\left(\frac{1}{\rho}\right)^m-1+1} = \rho^m$$

Q.E.D.

Lemma 4.2 : $\rho^m > 1 - m + m\rho$ when m≥ 1 and $1 > \rho = \frac{\lambda\tau}{m} > 0$.

Proof :

$$\because \mathrm{m} = 1+1+1+\cdots+1 > 1+\rho+\rho^2+\cdots+\rho^{m-1} = \frac{1-\rho^m}{1-\rho},$$

$$\therefore \mathrm{m}(1-\rho) > 1-\rho^m \Rightarrow \rho^m > 1-\mathrm{m}(1-\rho) = 1-m+m\rho.$$

Q.E.D.

Combining *Lemmas 4.1 and 4.2*, we have $C(m,\lambda,\tau) > \rho^m > 1 - m + m\rho$.

Lemma 4.3: $C(m,\lambda,\tau)$ is monotonically increasing with ρ, when m≥ 1 and $1 > \rho = \frac{\lambda\tau}{m} > 0$.

Proof:

$$\mathrm{C}(m,\lambda,\tau)=\frac{\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}+\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}}=\frac{\frac{1}{1-\rho}}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m}+\frac{1}{1-\rho}}=\frac{B(\rho)}{A(\rho)+B(\rho)}$$

where $A(\rho)=\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m}$ and $B(\rho)=\frac{1}{1-\rho}$.

If $1>\rho'>\rho$,

$$A(\rho)-A(\rho')=\sum_{k=0}^{m-1}\left[\frac{(m\rho)^k}{k!}\frac{m!}{(m\rho)^m}-\frac{(m\rho')^k}{k!}\frac{m!}{(m\rho')^m}\right]$$

$$=\sum_{k=0}^{m-1}\frac{m!m^{k-m}}{k!}\left(\frac{1}{\rho^{m-k}}-\frac{1}{\rho'^{m-k}}\right)>0 \;\rightarrow\; A(\rho)>A(\rho')$$

$$B(\rho)-B(\rho')=\frac{1}{1-\rho}-\frac{1}{1-\rho'}=\frac{\rho-\rho'}{(1-\rho)(1-\rho')}<0 \;\rightarrow\; B(\rho)<B(\rho')$$

Because $\mathrm{A}(\rho)>A(\rho')$ and $B(\rho')>B(\rho)$ $\rightarrow$ $\frac{B(\rho')}{A(\rho')+B(\rho')}>\frac{B(\rho)}{A(\rho)+B(\rho)}$

It then follows that $\mathrm{C}(m',\lambda',\tau')>\mathrm{C}(m,\lambda,\tau)$, and

$\mathrm{C}(m,\lambda,\tau)$ is monotonically increasing when $1>\rho>0$.

Q.E.D.

Theorem 4.2: The lower bound of expected cycle time of a GI/G/m queue is also the lower bound of expected cycle time of QNA approximation.

Proof :

The lower bound of expected cycle time of a GI/G/m queue, Eq. (4.4), is

$$LB[E\mathrm{CT}_{\mathrm{GI/G/m}}]=\frac{\rho^2{C_s}^2-\rho(2-\rho)}{2\lambda(1-\rho)}-\frac{(m-1)({C_s}^2+1)\tau}{2m}+\tau$$

$$=\frac{\rho{C_s}^2[1-m(1-\rho)]-\rho[1+m(1-\rho)]}{2\lambda(1-\rho)}+\tau$$

Because $1-\rho>0,\rho>0$ and $-\rho[1+m(1-\rho)]<0$,

$$\frac{\rho{C_s}^2[1-m(1-\rho)]}{2\lambda(1-\rho)}+\tau>LB[\mathrm{ECT}_{\mathrm{GI/G/m}}]$$

From Lemma 4.2, $1>\rho>0$, then $\rho^m\geq 1-m+m\rho$, and

$$\frac{\rho {C_s}^2 \rho^m}{2\lambda(1-\rho)} + \tau > \frac{\rho {C_s}^2 [1-m(1-\rho)]}{2\lambda(1-\rho)} + \tau > LB[E\mathrm{CT}_{\mathrm{GI/G/m}}],$$

From Lemma 4.1, $\mathrm{C}(m, \lambda, \tau) > \rho^m$,

$$\frac{\rho {C_s}^2 \mathrm{C}(m,\lambda,\tau)}{2\lambda(1-\rho)} + \tau > \frac{\rho {C_s}^2 \rho^m}{2\lambda(1-\rho)} + \tau > LB[\mathrm{ECT}_{\mathrm{GI/G/m}}].$$

The expected cycle time by QNA approximation in Eq. (4.5)

$$E[CT]_{\mathrm{QNA}} = \frac{{C_a}^2 + {C_s}^2}{2}\left(\frac{\rho \mathrm{C}(m,\lambda,\tau)}{\lambda(1-\rho)}\right) + \tau \geq \frac{\rho {C_s}^2 \mathrm{C}(m,\lambda,\tau)}{2\lambda(1-\rho)} + \tau > LB[\mathrm{ECT}_{\mathrm{GI/G/m}}]$$

since ${C_a}^2 \geq 0$.

Q.E.D.

From *Theorem 4.1* and *Theorem 4.2*, we conclude that the approximated cycle time by QNA is bounded in the same range of true cycle time performance, which implies that QNA is an appropriate approximation on values.

D-1: Discussion About Bounds

Kingman's upper bound for GI/G/1 is asymptotically tight in heavy traffic, but not in general [29]. Here we compare the bounds of QNA with bounds of another approximation model, JNA which assumes exponential distributions with unity SCV for inter-arrival and service times. Intuitively, QNA bounds derived by Kingman, Brumelle and Marchal are suitable for any situation no matter the level of variability or the network configuration, so JNA bounds are just a special case of QNA bounds while all SCVs are one. When all distributions are assumed unity SCV, QNA bounds are equivalent to JNA bounds.

Thus, if SCVs of actual system are smaller than one, JNA bounds are overestimated and we verify this by comparing with the existing $E_2/M/2$ result in [30].

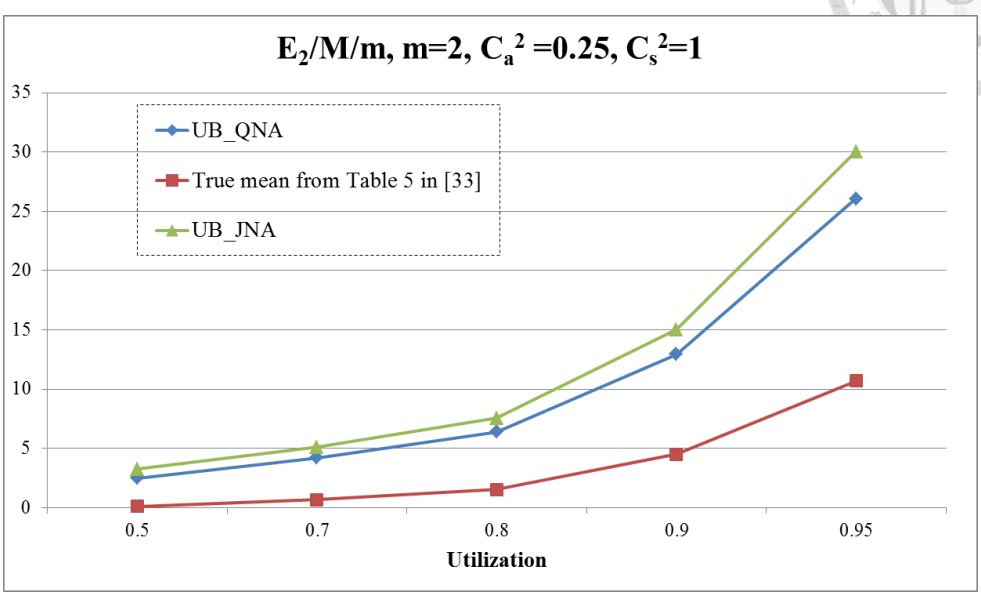


Figure 4.3 Comparison of upper bound between QNA and JNA while SCVs $\leqq 1$ and true value obtained from table 5 in [30]

Figure 4.3 shows the tighter bounds provided by QNA while SCVs smaller than one.

If SCVs greater than one, JNA bounds are underestimated and also verify by the existing $G/H_2/m$ result in [30].

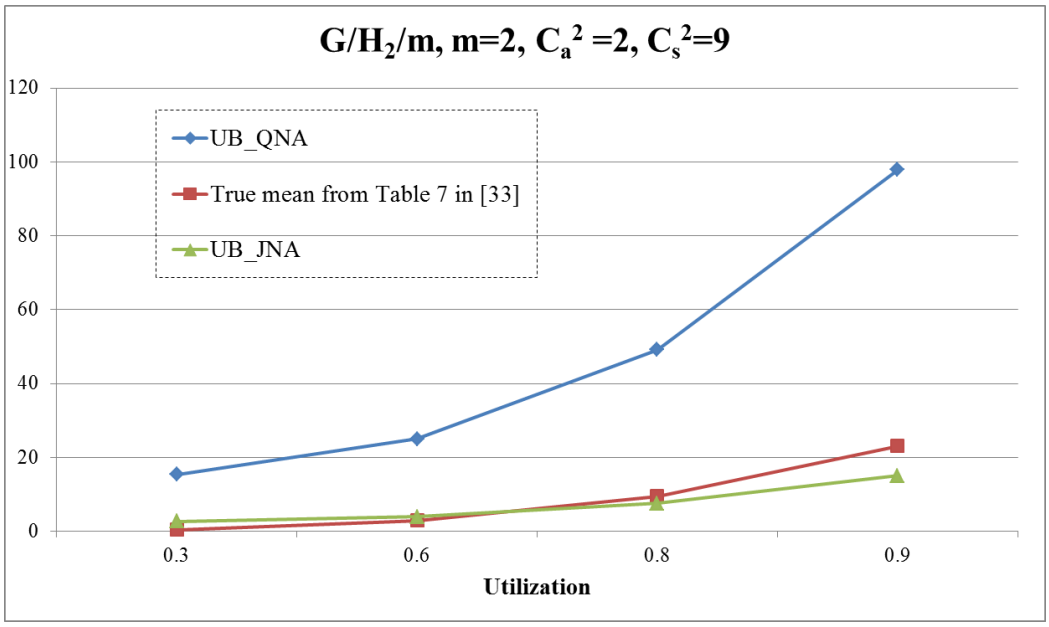


Figure 4.4 Comparison of upper bound between QNA and JNA while SCVs>1

and true value obtained from table 7 in [30]

Importantly, true expected cycle time in [30] is not bounded by JNA bound while C_a^2 =2, C_s^2=9, and utilization greater than 0.8, which shows JNA bound is not consistent for bounding the true performance because of the assumption of unity SCVs. In other words, QNA bounds are suitable for describing the range of true performance even through QNA bounds may be less tight sometimes.

Instead of JNA under the assumption of unity SCVs, QNA captures the heterogeneous SCVs to better characterize network flow, and provides more information for ranking. We show the advantage of QNA bounds with heterogeneous SCVs using the existing result in [30].

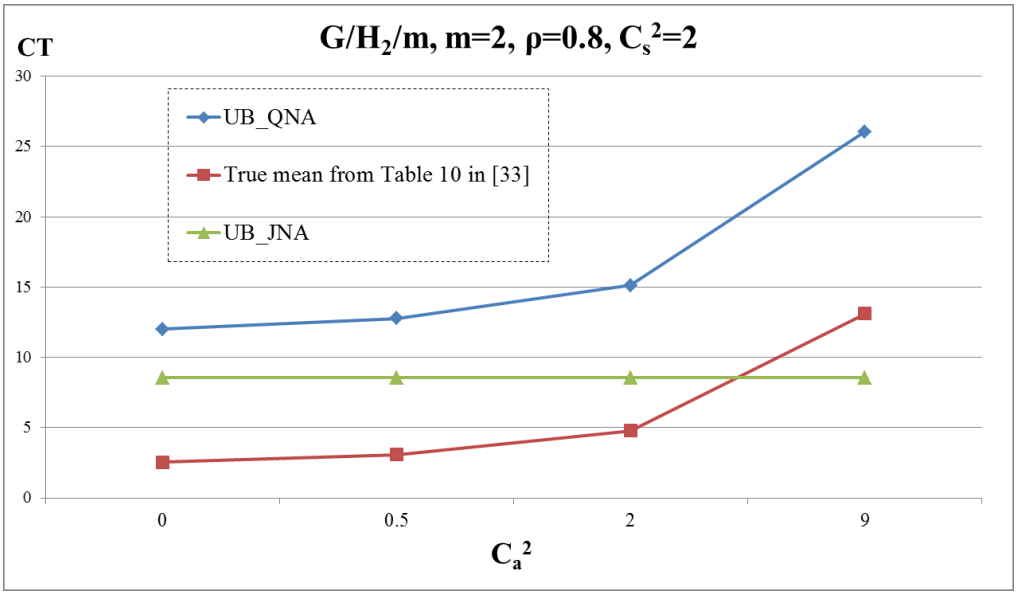


Figure 4.5 Advantage of QNA bounds with heterogeneous SCVs

It is obvious that true expected cycle time is always bounded by QNA upper bound, and most importantly, both of them grows with a similar trend which facilitates us use QNA to infer the true performance. By contrast, JNA bound is flat and this implies

JNA bound does not provide any information about the variation of true performance. The major differences between the models with heterogeneous SCVs and unity SCVs are (1) Better bounds (2) Implicitly useful information about true performance. In the next section, we utilize the bounds to help infer the probability of correct ranking and also show why model selection in OO matters.

4.2.2 Ranking Analysis of QNA and JNA

To exploit the bounds obtained in sub-section 4.2.1 and investigate ranking among designs by QNA approximation of mean cycle times, we first consider the comparison between a pair of designs for a single GI/G/m queue, D_1 and D_2. Let their true mean cycle times be ECT_1 and ECT_2 which are random variables. The approximated mean cycle times of D_1 and D_2 by QNA are, ACT_1 and ACT_2 respectively. Note that QNA approximation describes a node using four parameters $(\lambda_a, C_a^2, \tau, C_s^2)$ to characterize the mean and SCV of inter-arrival and service times [13][16]. We can obtain these four parameters $(\lambda_a, C_a^2, \tau, C_s^2)$ of each node by Eq. (3.7), (3.14), (3.5), and (3.6) respectively in Chapter 3.

Lemma 4.4: There are two designs, D_1 and D_2, and there are n_1 and n_2 machines allocated in node m. the SCV of service time of D_1 is equal to that of D_2 at node m.

Therefore, the SCV of service time of a node is not related to the number of machine allocated in that node.

Proof:

D_1 has n_1 machines allocated at a node m and D_2 has n_2 machines allocated at a node m. Recall that for a node m, its service time SCV, C_{sm}^2, is obtained from Eq. (3.6)

$$C_{sm}^2 = \frac{\sum_{i=1}^{I}\sum_{k=1}^{S_i} \tau_{ik}{}^2 (C_{ik}{}^2 + \mathbf{1})\lambda_i \mathbf{1}\{m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S_i} \tau_m{}^2 \lambda_i \mathbf{1}\{m_{i,k}=m\}} - 1. \quad (3.6)$$

where

$\mathbf{1}$H(x): an indicator function of the set H, $\mathbf{1}\text{H(x)} = \begin{cases} 1, if\ x \in H. \\ 0, otherwise. \end{cases}$

τ_{ik} is the mean processing time at k-th step of product type i and the k-th step of product type i is processed by the service node $m_{i,k}$.

λ_i is the external arrival rate of product type i.

τ_m is the aggregate mean service time of node m obtained from Eq. (3.5)

$$\tau_m = \frac{\sum_{i=1}^{I}\sum_{k=1}^{S_i} \tau_{ik}\lambda_i \mathbf{1}\{m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{m_{i,k}=m\}}\ , \quad (3.5)$$

Eq. (3.5) is a weighted sum of mean processing time of every steps processed by node m and it is not related to the number of machines allocated in node m.

Eq. (3.6) is also a weighted sum of SCV of processing time of every steps processed by node m and also not related to the number of machine allocated in node m.

Therefore, for D_1, the SCV of service time of node m is not a function of n_1.

$$C_{sm}^2 = \frac{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_{ik}^2(C_{ik}^2+1)\lambda_i \mathbf{1}\{m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_m^2\lambda_i \mathbf{1}\{m_{i,k}=m\}} - 1, \text{ for } D_1$$

Also, for D_2, the SCV of service time of node m is not a function of n_2 either.

$$C_{sm}^2 = \frac{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_{ik}^2(C_{ik}^2+1)\lambda_i \mathbf{1}\{m_{i,k}=m\}}{\sum_{i=1}^{I}\sum_{k=1}^{S_i}\tau_m^2\lambda_i \mathbf{1}\{m_{i,k}=m\}} - 1, \text{ for } D_2$$

The above shows that the calculation of C_{sm}^2 is nothing related to the number of machine allocated in node m. Therefore, the SCV of service time of D_1 at node m is equal to that of D_2 at node m.

Q.E.D.

Lemma 4.5: For a single GI/G/m queue, there are two designs, D_1 and D_2 and their approximated cycle times by QNA are ACT_1 and ACT_2 respectively. If ACT_1 is smaller than ACT_2, then $UB(ECT_1) \le UB(ECT_2)$ and $LB(ECT_1) \le LB(ECT_2)$.

Proof:

Assume that D_1 has more machines than D_2. Thus, in a single GI/G/m queue, utilization(ρ) of D_1 is smaller than that of D_2, noted as $\rho_1 < \rho_2$.

From Lemma 4.4, the SCV of service time (C_s^2) of D_1 is equal to that of D_2.

Eq. (3.14) characterizes the SCV of inter-arrival time of each node, which is related to the number of machines allocated in each node. In case of single GI/G/m queue, Eq. (3.14) can be re-written as

$C_a^2 = \alpha + C_a^2\beta \rightarrow C_a^2 = {}^{\alpha}/_{(1-\beta)}$, α and β increases with utilization.

So, the SCV of inter-arrival time($C_a{}^2$) of D_1 is smaller than that of D_2.

The approximated mean cycle time of QNA is, $\text{ACT} = \frac{C_a{}^2+C_s{}^2}{2}\left(\frac{\rho \text{C}(m,\lambda,\tau)}{\lambda(1-\rho)}\right)+\tau$. Because $\frac{\rho}{(1-\rho)}$ is monotonically increasing with ρ and $\text{C}(m,\lambda,\tau)$ is as well according to Lemma 4.3, which induces that $\frac{\rho \text{C}(m,\lambda,\tau)}{(1-\rho)}$ is also monotonically increasing with ρ.

Thus, we know utilization of D_1 is smaller than that of D_2, so $\frac{\rho \text{C}(m,\lambda,\tau)}{(1-\rho)}$ of D_1 is smaller than that of D_2. In addition, the SCV of inter-arrival time ($C_a{}^2$) of D_1 is smaller than that of D_2. So, we obtain that ACT_1 is smaller than ACT_2.

Besides, the upper bound and lower bound of expected cycle time of a GI/G/m queue are positively related to its utilization. The growth of upper bound is faster than ACT but the growth of lower bound is slower than ACT, as shown in Figure 4.3. So if $\rho_1 < \rho_2$ then both upper and lower bounds of expected cycle time of D_1 is smaller than them of D_2, noted as $UB(ECT_1) \leq UB(ECT_2)$ and $LB(ECT_1) \leq LB(ECT_2)$.

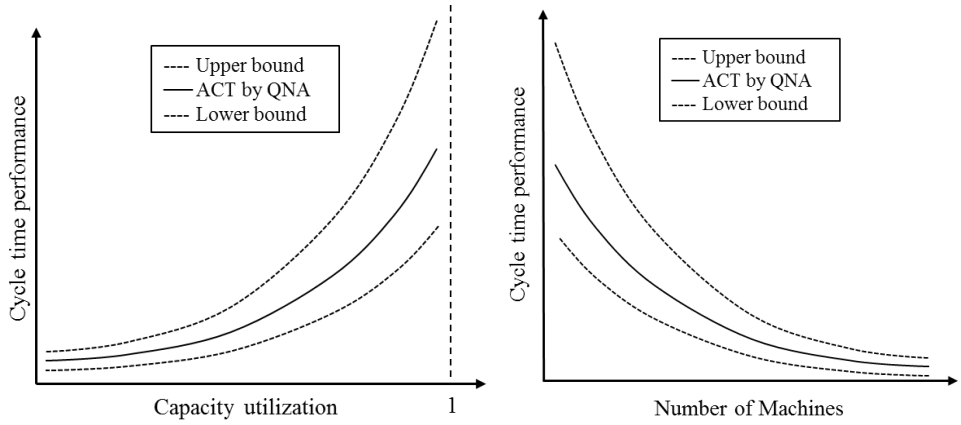


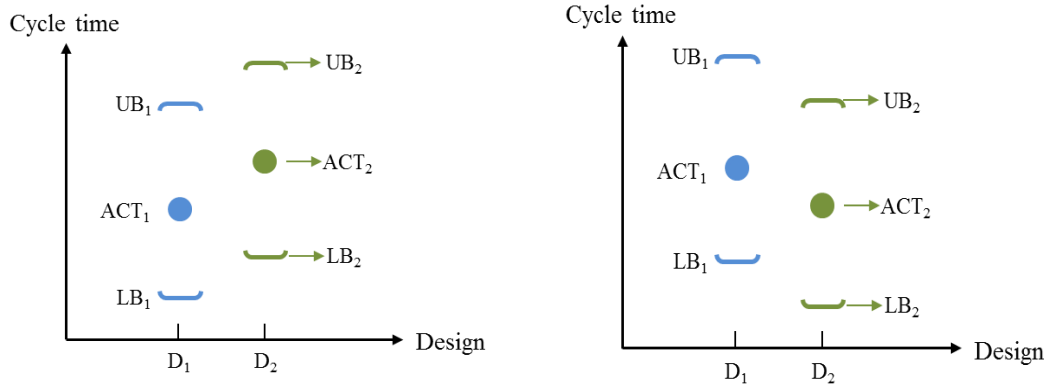
Figure 4.6 UB, ACT, LB w.r.t. utilization and number of machines

We therefore conclude that if ACT_1 is smaller than ACT_2, then $UB(ECT_1) \leq$

UB(ECT_2) and LB(ECT_1) ≤ LB(ECT_2).

Q.E.D.

Based on Lemma 4.5, a simple diagram is shown in Figure 4.7.



(a) Given $ACT_1 < ACT_2$ (b) Given $ACT_1 > ACT_2$

Figure 4.7 Two simple diagrams

Recall the definition of rank correlation coefficient in Section 4.1.3,

$$RC = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{\text{Total number of pairs} = \frac{1}{2}n(n-1)}.$$

There are two kinds of pairs in the definition of rank correlation, concordant pair and discordant pair. Therefore, given two designs, D_1 and D_2, and their approximated cycle time by QNA, ACT_1 and ACT_2, there are two probable events:

(1) (ECT_1, ACT_1) and (ECT_2, ACT_2) is a concordant pair, and

(2) (ECT_1, ACT_1) and (ECT_2, ACT_2) is a discordant pair.

If ACT_1 is smaller/larger than ACT_2 and ECT_1 is smaller/larger than ECT_2, then the pair of (ECT_1, ACT_1) and (ECT_2, ACT_2) is a concordant pair. Otherwise, it is discordant.

Assumption of uniform distributions

Without any prior knowledge we assume that the value lying anywhere between upper bound and lower bound has an equal probability, so distribution of ECT_1 and ECT_2 are assumed to be uniform. Note that F_1 is the CDF of ECT_1 and F_2 is the CDF of ECT_2. We only analyze the case of $ACT_1 < ACT_2$ as shown in Figure 4.7 (a) because the other case of $ACT_1 > ACT_2$ as shown in Figure 4.7 (b) just reverses the notation of ECT_1 and ECT_2.

Given $ACT_1 < ACT_2$, as shown in Figure 4.7 (a), the probability of being a concordant pair is $P\{ECT_1 < ECT_2 \mid ACT_1 < ACT_2\}$.

$$P[ECT_1 < ECT_2 | ACT_1 < ACT_2]$$

$$= \int_{UB_1}^{UB_2} \frac{1}{UB_2 - LB_2} dy + \int_{LB_2}^{UB_1} \frac{y - LB_1}{UB_1 - LB_1} \frac{1}{UB_2 - LB_2} dy$$

$$= \frac{UB_2 - UB_1}{UB_2 - LB_2} + \frac{1}{2} \frac{{UB_1}^2 - {LB_2}^2 - 2LB_1(UB_1 - LB_2)}{(UB_1 - LB_1)(UB_2 - LB_2)}$$

$$= \frac{1}{2} \frac{2UB_2UB_1 - 2UB_2LB_1 - {UB_1}^2 - {LB_2}^2 + 2LB_1LB_2}{(UB_1 - LB_1)(UB_2 - LB_2)}$$

$$= \frac{1}{2} \frac{2UB_2(UB_1 - LB_1) - (UB_1 - LB_1)(UB_1 + LB_1) - (LB_2 - LB_1)^2}{(UB_1 - LB_1)(UB_2 - LB_2)}$$

$$= \frac{1}{2} \frac{(UB_1 - LB_1)[(UB_2 - UB_1) + (UB_2 - LB_1)] - (LB_2 - LB_1)^2}{(UB_1 - LB_1)(UB_2 - LB_2)}$$

$$= \frac{1}{2} \frac{(UB_1 - LB_1)[(UB_2 - UB_1) + (UB_2 - LB_2) + (LB_2 - LB_1)] - (LB_2 - LB_1)^2}{(UB_1 - LB_1)(UB_2 - LB_2)}$$

$$= \frac{1}{2} + \frac{1}{2} \frac{(UB_2 - UB_1)}{(UB_2 - LB_2)} + \frac{1}{2} \frac{(LB_2 - LB_1)(UB_1 - LB_2)}{(UB_1 - LB_1)(UB_2 - LB_2)} \quad (4.3)$$

The probability of being a discordant pair is $P\{ECT_1 > ECT_2 \mid ACT_1 < ACT_2\}$.

$$P[ECT_1 > ECT_2 | ACT_1 < ACT_2]$$

$$= \int_{LB_2}^{UB_1} \frac{y-LB_2}{UB_2-LB_2} \frac{1}{UB_1-LB_1} dy$$

$$= \frac{1}{2} \frac{{UB_1}^2-{LB_2}^2}{(UB_2-LB_2)(UB_1-LB_1)} - \frac{LB_2(UB_1-LB_2)}{(UB_2-LB_2)(UB_1-LB_1)}$$

$$= \frac{1}{2} \frac{{UB_1}^2-2LB_2UB_1+{LB_2}^2}{(UB_2-LB_2)(UB_1-LB_1)}$$

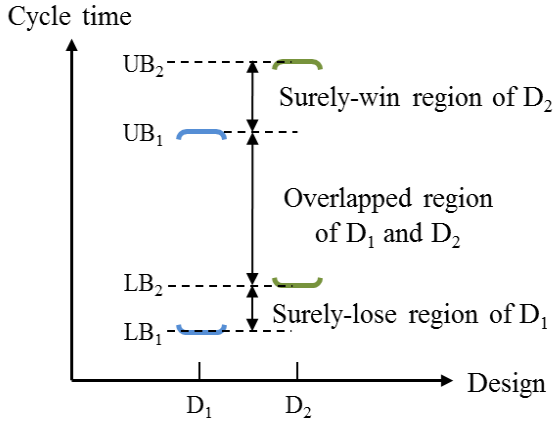
$$= \frac{1}{2} \frac{(UB_1-LB_2)^2}{(UB_2-LB_2)(UB_1-LB_1)} \quad (4.4)$$

$$= 1 - \mathrm{P}[\mathrm{ECT}_1 < \mathrm{ECT}_2 | \mathrm{ACT}_1 < \mathrm{ACT}_2]$$

D-2: Discussion about probability of being a concordant pair, P_c

Equation (4.3) is influenced by two terms: one can be viewed as the ratio of the surely-win region in the range of ECT_2, $\frac{(UB_2-UB_1)}{(UB_2-LB_2)}$ and the other can be viewed as the product of ratio of surely-lose region in the range of ECT_1, $\frac{(LB_2-LB_1)}{(UB_1-LB_1)}$ and ratio of not surely-win region in the range of ECT_2, $\frac{(UB_1-LB_2)}{(UB_2-LB_2)}$. Equation (4.4) can actually split into two terms, one is ratio of overlapped region in the range of ECT_1, $\frac{UB_1-LB_2}{UB_1-LB_1}$, and the other is ratio of overlapped region in the range of ECT_2, $\frac{UB_1-LB_2}{UB_2-LB_2}$.

The maximum of Equation (4.4) is $\frac{1}{2}$ when the range of ECT_1 is totally overlapped with the range of ECT_2, but Equation (4.3) is obviously greater that $\frac{1}{2}$ and the maximum of Equation (4.4) is $\frac{1}{2}$. It shows that if it is known that $ACT_1 < ACT_2$, ranking two designs according to their approximated cycle times by QNA makes sure that the probability of being a concordant pair (P_c) is not lower than 0.5.



D-3: Discussion about the effects of bounds to P_c

We define the ratio of surely-win region in the range of ECT_2, $\frac{(UB_2-UB_1)}{(UB_2-LB_2)}$ as R_W, and the ratio of surely-lose region in the range of ECT_1, $\frac{(LB_2-LB_1)}{(UB_1-LB_1)}$ as R_L, and $R_W, R_L \in [0,1]$ and if $R_W=1$, then $R_L=1$ and if $R_W=0$, then $R_L=0$. The probability of being a concordant pair is determined by R_W and R_L, and Equation (4.3) can be written as $\frac{1}{2}+\frac{1}{2}R_w+\frac{1}{2}R_L(1-R_w)=1-\frac{1}{2}(1-R_w)(1-R_L)$, visualized as Figure 4.8. $(1-R_w)$ is the ratio of overlapped region in the range of ECT_2 and $(1-R_L)$ is the overlapped region in the range of ECT_1. If the overlapped region decreases, $(1-R_w)$ and $(1-R_L)$ decrease and the probability of correct ranking increases. This implies the differentiation between D_1 and D_2 becomes easier.

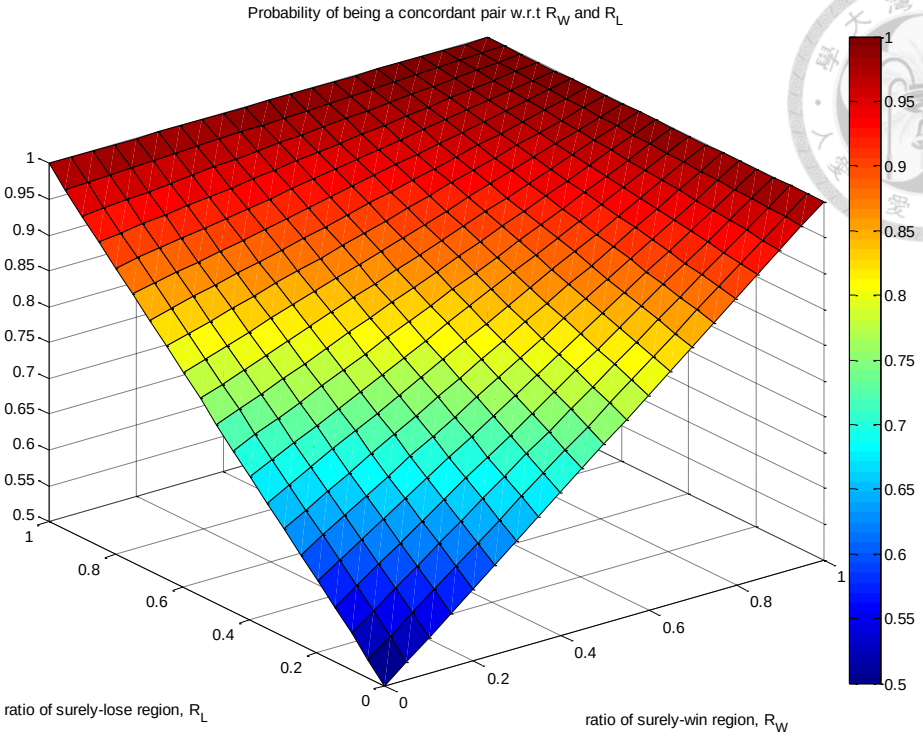
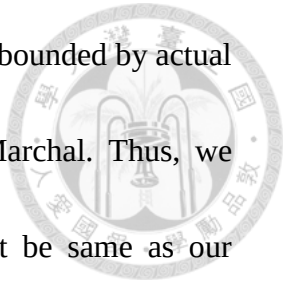


Figure 4.8 Probability of being a concordant pair w.r.t. R_W and R_L

Recall that Figure 4.5 in Discussion D-1 shows that bounds of QNA with heterogeneous SCVs inhibit the similar trend with true performance and better capture the difference due to heterogeneous SCVs. In contrast, flat JNA bound whose R_W equal to zero provides no any information about the probability of being a concordant pair and this causes the probability of 0.5 like tossing a coin. Thus, it is significant that heterogeneous SCVs advantage us to correctly rank among designs with higher probability.

D-4: Discussion about distribution of actual cycle time

In our analysis, because we have no any prior information about the distribution



of actual cycle time, the only information we known is that QNA is bounded by actual upper and lower bounds proposed by Kingman, Brumelle and Marchal. Thus, we assume actual cycle time is uniformly distributed but it may not be same as our assumption. Erlang distribution may be another alternative assumption. Because the support of Erlang distribution is greater than zero, which coincides with the nature of waiting time, and Erlang distribution is concentrated at a specific value, which is more realistic to common situations. Without uniformity, the difference between peaks of any two Erlang PDFs matters. When their peaks are getting closer, probability of correct ranking would be decreasing because of more overlapped region and smaller R_W and R_L.

In the following, we further investigate the useful information hidden behind QNA approximations and therefore derive a lower bound of probability of being a concordant pair, which must be greater than the probability of 0.5, under the assumption of uniform distribution.

Lemma 4.5: For a single GI/G/m queue, there are two designs, D_1 and D_2 and their approximated cycle times by QNA are ACT_1 and ACT_2 respectively. If $ACT_2 > ACT_1$, the difference of approximated cycle times, $\Delta ACT=ACT_2-ACT_1$, leads to the difference of their upper bounds, $\Delta UB \geq \frac{\rho^4-2\rho^3+\rho^2({C_a}^2+\lambda\tau)+(2\rho-1){C_s}^2\lambda^2\tau^2}{\rho^2({C_a}^2+{C_s}^2)}\Delta ACT$.

Proof:

There is a simple relation of ΔUB and ΔACT is that

$$\Delta UB \approx \left(\frac{\partial UB}{\partial ACT}\right)\Delta ACT = \left(\frac{\partial UB}{\partial \rho}\right)\left(\frac{\partial \rho}{\partial ACT}\right)\Delta ACT.$$

1. $ACT = \frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho C(m,\lambda,\tau)}{\lambda(1-\rho)}\right)+\tau$ is monotonically increasing with ρ and always smaller than $\frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho}{\lambda(1-\rho)}\right)+\tau$, so we obtain

   $\frac{\partial ACT}{\partial \rho} \le \frac{\partial}{\partial \rho}\left(\frac{{C_a}^2+{C_s}^2}{2}\left(\frac{\rho}{\lambda(1-\rho)}\right)+\tau\right) = \frac{({C_a}^2+{C_s}^2)}{2\lambda(1-\rho)^2}$ ➔ $\frac{\partial \rho}{\partial ACT} \ge \frac{2\lambda(1-\rho)^2}{({C_a}^2+{C_s}^2)}$

2. Upper bound derived from Kingman is $\frac{{C_a}^2+m^2\rho{C_s}^2+(m-1)\rho^2}{2\lambda(1-\rho)}+\tau$ where $\rho=\frac{\lambda\tau}{m}$, and the partial differential of UB to ρ is $\frac{\partial UB}{\partial \rho} = \frac{\rho^4-2\rho^3+(2\rho-1){C_s}^2\lambda^2\tau^2+\lambda^2({C_a}^2+\lambda\tau)}{2(1-\rho)^2\rho^2\lambda}$.

3. Because both UB and ACT are monotonically increasing, both $\frac{\partial UB}{\partial \rho}$ and $\frac{\partial \rho}{\partial ACT}$ are greater than zero. The difference of their upper bound is obtained by

   $$\Delta UB = \left(\frac{\partial UB}{\partial \rho}\right)\left(\frac{\partial \rho}{\partial ACT}\right)\Delta ACT \ge \frac{\rho^4-2\rho^3+\rho^2({C_a}^2+\lambda\tau)+(2\rho-1){C_s}^2\lambda^2\tau^2}{\rho^2({C_a}^2+{C_s}^2)}\Delta ACT \ge 0. \quad (4.5)$$

Q.E.D.

D-5: Interpretation of Equation (4.5)

Equation (4.5) is related to utilization and variability terms. As Figure 4.9 shown, while utilization becomes lower, the characterization of variability terms has greater effects and this implies that characterization on variability terms really helps us differentiate designs because variability terms press the differences between designs. For machine allocation, a better allocation design leads to lower utilization because its waiting time is smaller, so as its cycle time. Therefore, capturing the characterization

on SCVs is beneficial to recognize designs, especially top designs. That is why QNA has better rank correlation than JNA in top designs.

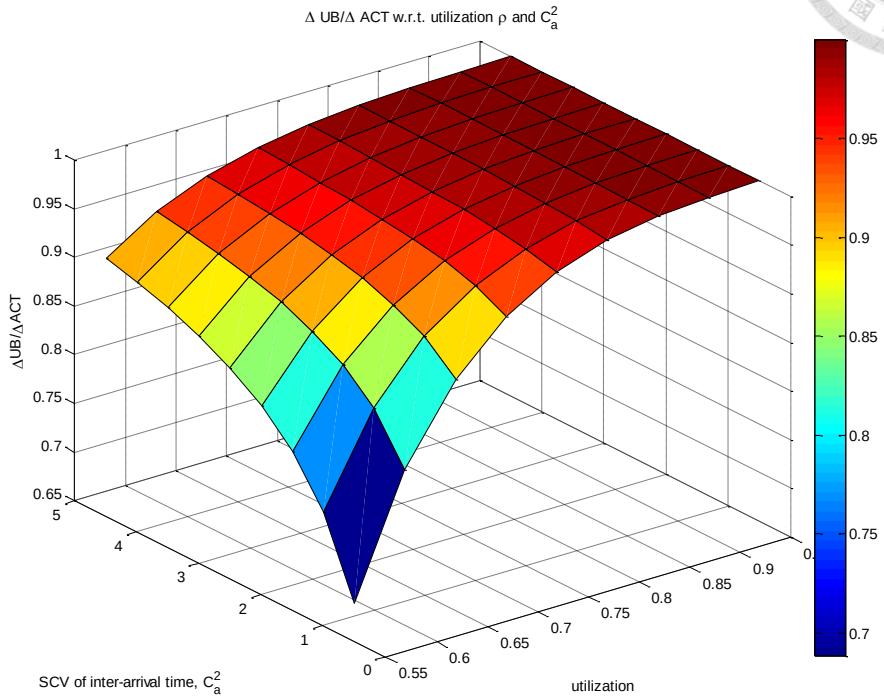


Figure 4.9$\Delta UB/\Delta ACT$ w.r.t. utilization ρ and SCV of inter-arrival time ${C_a}^2$ assumed that λ=1, τ=1, and ${C_s}^2$=1

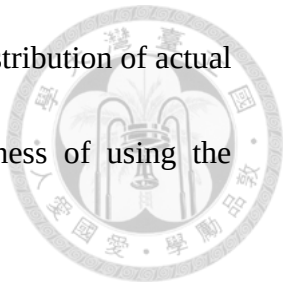
We use Lemma 4.5 to further derive the lower bound of probability of being a concordant pair. Assumed that compared with upper bound, lower bound is quite small, $UB_2 - LB_2 \approx UB_2$, and lower bounds of designs are close, $LB_2 \approx LB_1$. Thus, Equation (4.3) can be re-written as,

$$\mathrm{P}[\mathrm{CT}_1 < \mathrm{CT}_2 | \mathrm{ACT}_1 < \mathrm{ACT}_2]$$

$$= \frac{1}{2} + \frac{1}{2}\frac{(UB_2 - UB_1)}{(UB_2 - LB_2)} + \frac{1}{2}\frac{(LB_2 - LB_1)(UB_1 - LB_2)}{(UB_1 - LB_1)(UB_2 - LB_2)} \approx \frac{1}{2} + \frac{1}{2}\frac{(UB_2 - UB_1)}{(UB_2)}.$$

From Lemma 4.5, $\Delta UB \geq \frac{\rho^4 - 2\rho^3 + \rho^2({C_a}^2 + \lambda\tau) + (2\rho - 1){C_s}^2\lambda^2\tau^2}{\rho^2({C_a}^2 + {C_s}^2)}\Delta ACT$, the probability of being a concordant pair is greater than $\alpha \geq \frac{1}{2} + \frac{\rho^4 - 2\rho^3 + \rho^2({C_a}^2 + \lambda\tau) + (2\rho - 1){C_s}^2\lambda^2\tau^2}{\rho^2({C_a}^2 + {C_s}^2)}\frac{\Delta ACT}{(UB_2)}$

and α must be greater than 0.5 under the assumption of uniform distribution of actual cycle time. The above analysis shows the rationality and goodness of using the ranking information accessed from QNA.



4.3 Summary

In this chapter, we first introduce the central idea of ordinal transformation, transforming the original space to an ordinal space by a simplified model used for ranking. We develop a BRA and take the first step to analyze single GI/G/m queue with 2 designs. BRA applies to the case of QNA as the simplified model of single GI/G/m queue. Our contributions in this chapter are as follow:

(1) Bound analysis shows that QNA is bounded by the upper and lower bounds of cycle time of single GI/G/m queue proposed by Kingman, and Brumelle and Marchal respectively.

(2) Assumed actual cycle time is uniformly distributed between its upper bound and lower bound, QNA approximation is demonstrated to provide the probability of correct ranking P_c, $P_c > 0.5$.

(3) Probability of correct ranking is increasing with increase of the ratio of surely-win and surely-lose region, R_W and R_L.

(4) With the variation of QNA, the least variation of upper bound is derived. This

facilitates us to derive a better probability of correct ranking α,

$$\alpha > \frac{1}{2} + \frac{\rho^4 - 2\rho^3 + (2\rho - 1)C_s^{\,2}\lambda^2\tau^2 + \lambda^2\left(C_a^{\,2} + \lambda\tau\right)}{\rho^2\left(C_a^{\,2} + C_s^{\,2}\right)} \frac{\Delta ACT}{(UB_2)} > \frac{1}{2}.$$

(5) Capturing the heterogeneous SCVs is beneficial to recognize designs, especially top designs.

Finally, the derivations under the assumption of normal distributions of actual cycle times please refer to Appendix A-A.1 for more details. In the next chapter, we extend the ranking analysis to general re-entrant lines with M workstations and N designs.

Chapter 5

Extensions of BRA to General Re-entrant Lines

In this chapter, we extend the discussion of QNA approximation of single GI/G/m queue in Section 4.3 to a general re-entrant line with M workstations. Because congestion measures of QNA for a network are obtained by assuming that the stations are stochastically independent given the approximate flow parameters, the bound and ranking analysis of single GI/G/m queue is generalized to multiple GI/G/m queues using superposition of their upper and lower bounds.

We first consider 2-workstation re-entrant line with a pair of designs and there are four different cases of their bounds due to the superposition of bounds of independent workstations, as shown in Figure 5.2. We analyze the probability of being a concordant pair in these four cases. Based on the analysis of 2-workstation re-entrant line, it is seamless to generalize the result to M-workstation re-entrant lines. In the last, we extend the scope of 2 designs (a pair of designs) to N designs ($\frac{1}{2}\mathrm{N}(\mathrm{N}-1)$ pairs in total), and validate our analysis and conclusions by an experiment of a five-workstation re-entrant line.

5.1 BRA of QNA and JNA for General Re-entrant Lines

Because JNA is a special case of QNA, here we focus on the analysis of QNA.

We first consider a 2-workstation re-entrant line and there are two designs, D_1={M_1, M_2}, which means m_1 machines at workstation 1 and m_2 machines at workstation 2, and another design D_2={M_1-1, M_2+1}. Let their true mean cycle times be ECT_1 and ECT_2 which are random variables.

Let us define a notation, $ACT_{i.m}$, which means the approximated mean cycle time of workstation m given design D_i. Thus, ACT_1= $ACT_{1,1}$+$ACT_{1,2}$ and ACT_2= $ACT_{2,1}$+$ACT_{2,2}$. Actually, D_2 is equivalent to D_1 moving one machine from workstation 1 to workstation 2. Due to decreasing capacity of workstation 1 and increasing capacity of workstation 2, there is growth and decline of approximated mean cycle time, a simple diagram shown in Figure 5.1.

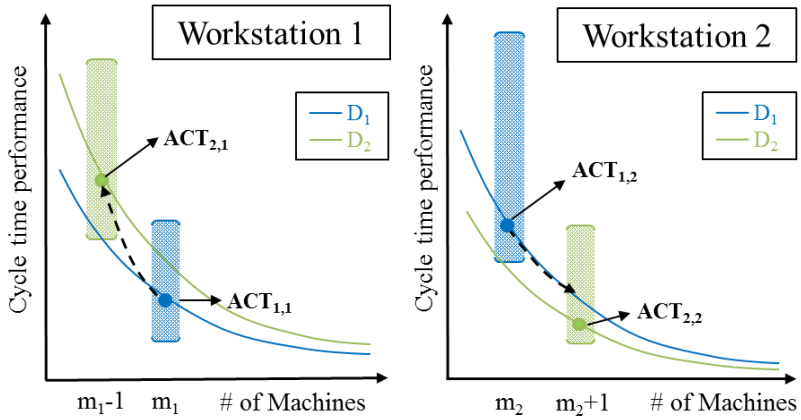


Figure 5.1 ACT of D_1 and D_2 w.r.t. number of machines

Most importantly, because in this case there are interactions among stations, various machine allocation designs bring about different network configuration. The approximated parameters of QNA vary from machine allocation designs, for example, C_a^2 and C_s^2. Therefore, the relation between upper and lower bounds of D_1

and D_2 is not always consistent with Lemma 4.5. Recall that Lemma 4.5 says if $ACT_1 < ACT_2$ then upper and lower bound of D_1 would be smaller than them of D_2 for a single GI/G/m queue.

Even though we know $ACT_1 > ACT_2$, we cannot definitely conclude that the upper bound of ACT_1 must be greater than upper bound of ACT_2. Instead, there are four possible cases of the bounds of D_1 and D_2 as shown in Figure 5.2. We analyze the probability of D_1 and D_2 being a concordant pair in each case.

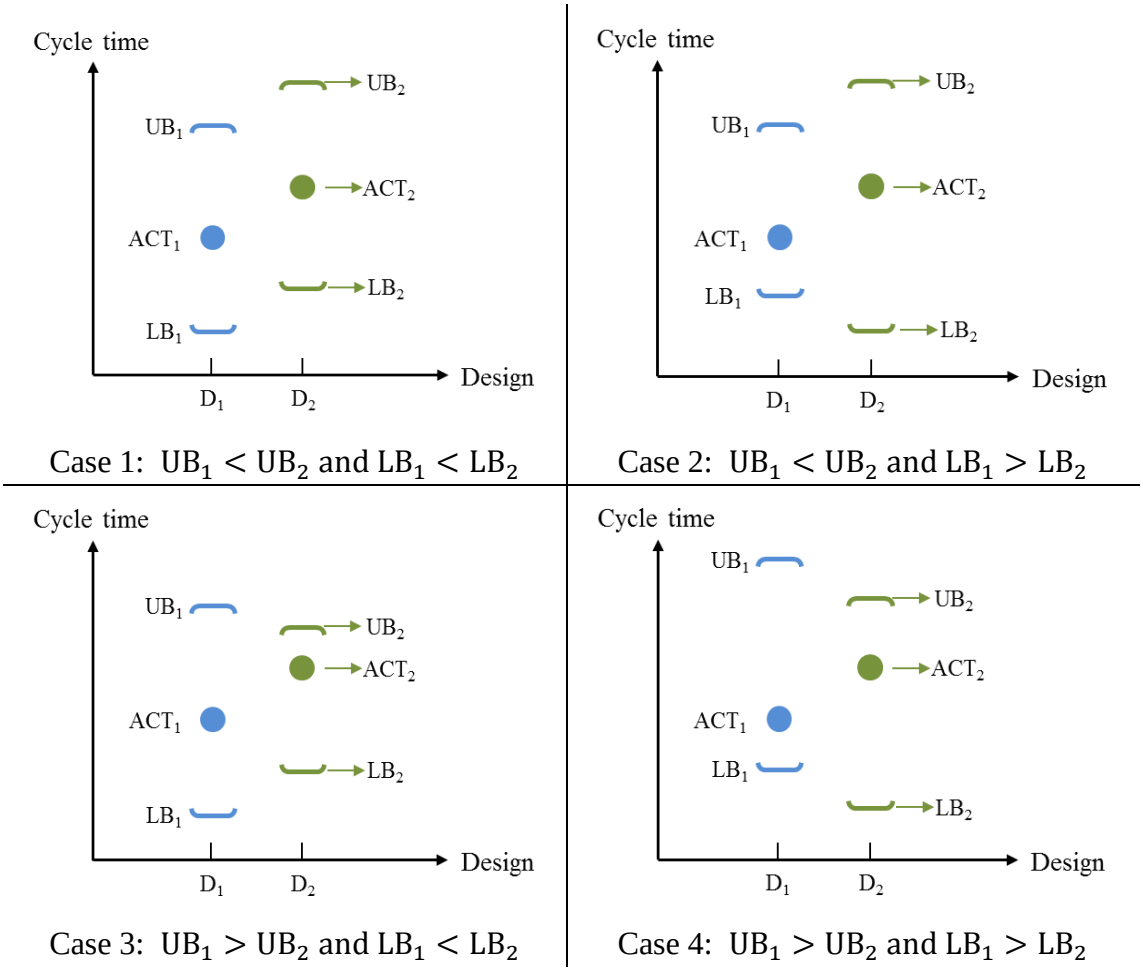


Figure 5.2 Four possible cases of bounds given $ACT_1 < ACT_2$

Note that we derive BRA in single GI/G/m queue under the assumption of uniform distribution, but here the bounds obtained from the superposition of several

stations are not be uniformly distributed anymore. For the ease of BRA, we assume true mean cycle time still follows uniform distribution even though the superposition and discuss the influence on rank correlation later.

Analysis of Case 1

Case 1 is the same as the case of single GI/G/m queue. Thus, analysis of single GI/G/m queue is directly applicable to Case 1. The probability of being a concordant pair is $P\{ECT_1 < ECT_2 \mid ACT_1 < ACT_2\}$ as Equation (4.3) in Chapter 4.

$$P[ECT_1 < ECT_2 | ACT_1 < ACT_2]$$

$$= \frac{1}{2} + \frac{1}{2}\frac{(UB_2 - UB_1)}{(UB_2 - LB_2)} + \frac{1}{2}\frac{(LB_2 - LB_1)(UB_1 - LB_2)}{(UB_1 - LB_1)(UB_2 - LB_2)} \quad (5.1)$$

Analysis of Case 2

Assume that true mean cycle time performance is in uniform distribution. The probability of being a concordant pair is $P\{ECT_1 < ECT_2 \mid ACT_1 < ACT_2\}$. $UB_1 < UB_2$ and $LB_1 > LB_2$ are given.

$$P[ECT_1 < ECT_2 | ACT_1 < ACT_2]$$

$$= \int_{UB_1}^{UB_2} \frac{1}{UB_2 - LB_2} dy + \int_{LB_1}^{UB_1} \frac{y - LB_1}{UB_1 - LB_1} \frac{1}{UB_2 - LB_2} dy$$

$$= \frac{UB_2 - UB_1}{UB_2 - LB_2} + \frac{1}{2}\frac{UB_1 + LB_1}{UB_2 - LB_2} - \frac{LB_1}{UB_2 - LB_2}$$

$$= \frac{UB_2 - UB_1}{UB_2 - LB_2} + \frac{1}{2}\frac{UB_1 - LB_1}{UB_2 - LB_2}$$

$$= \frac{2UB_2 - UB_1 - LB_1}{2(UB_2 - LB_2)} \quad (5.2)$$

Therefore, if $UB_2 - UB_1 > LB_1 - LB_2$ is true, the probability of being a concordant pair is greater than 0.5.

Analysis of Case 3

Assume that true mean cycle time performance is in uniform distribution. The probability of being a concordant pair is $P\{ECT_1 < ECT_2 \mid ACT_1 < ACT_2\}$. $UB_1 > UB_2$ and $LB_1 < LB_2$ are given.

$$P[ECT_1 < ECT_2 | ACT_1 < ACT_2]$$

$$= \int_{LB_2}^{UB_2} \frac{y - LB_1}{UB_1 - LB_1} \frac{1}{UB_2 - LB_2} dy$$

$$= \frac{1}{2} \frac{UB_2 + LB_2}{UB_1 - LB_1} - \frac{LB_1}{UB_1 - LB_1}$$

$$= \frac{UB_2 + LB_2 - 2LB_1}{2(UB_1 - LB_1)} \qquad (5.3)$$

Therefore, if $LB_2 - LB_1 > UB_1 - UB_2$ is true, the probability of being a concordant pair is greater than 0.5.

Analysis of Case 4

Assume that true mean cycle time performance is in uniform distribution. The probability of being a concordant pair is $P\{ECT_1 < ECT_2 \mid ACT_1 < ACT_2\}$. $UB_1 > UB_2$ and $LB_1 > LB_2$ are given.

$$P[ECT_1 < ECT_2 | ACT_1 < ACT_2]$$

$$= \int_{LB_1}^{UB_2} \frac{y - LB_1}{UB_1 - LB_1} \frac{1}{UB_2 - LB_2} dy$$

$$= \frac{1}{2} \frac{{UB_2}^2 - {LB_1}^2}{(UB_2 - LB_2)(UB_1 - LB_1)} - \frac{LB_1(UB_2 - LB_1)}{(UB_2 - LB_2)(UB_1 - LB_1)}$$

$$= \frac{1}{2}\frac{{UB_2}^2 - 2LB_1UB_2 + {LB_1}^2}{(UB_2 - LB_2)(UB_1 - LB_1)}$$

$$= \frac{1}{2}\frac{(UB_2 - LB_1)^2}{(UB_2 - LB_2)(UB_1 - LB_1)} \quad (5.4)$$

Maximum of Equation (5.4) is equal to $\frac{1}{2}$, which means in Case 4 the probability of being a concordant pair is not greater than that of being a discordant pair.

In this section, we respectively discuss the four possible cases of relations of upper and lower bounds with 2 designs. In fact, because there are still only four possible cases of their bounds for a pair of designs, the above analyses are seamlessly applicable to more-than-two workstations re-entrant lines. Thus, the BRA is applicable to a general M-workstations re-entrant line.

Even though there are four possible cases of the bounds of any pairs of designs, we believe that the sum of approximated mean cycle time of workstations is essentially a representative performance index to describe their bounds. At least in single GI/G/m queue, Lemma 4.5 shows higher approximated mean cycle time (ACT) absolutely leads to higher upper bound and lower bound.

We hypothesize that most of pairs belong to Case 1 and we validate the above hypothesis by an experiment of 5-workstation re-entrant line with 415 designs in Chapter 6 and show the statistical result of four possible cases here.

Case	1	2	3	4
Ratio	89.91%	4.13%	3.89%	2.07%

As we can see, most of pairs still hold for Lemma 4.5. For Case 2 and Case 3, there are some specific conditions to make the probability of being a concordant pair greater than 0.5 and the statistical results are as follow.

Case	2	3
Ratio of Pc > 0.5	99.72%	2.06%
Ratio of Pc < 0.5	0.28%	97.94%

The result of statistics shows that most of design pairs belong to Case 1 as our above analysis, and in fact the ratio of pairs whose Pc greater than 0.5 is more than 94%.

Discussion about assumption of uniform distribution after superposition

It is known that sum of uniform distributions is not a uniform distribution anymore. Here we use a sum of two uniform distributions, $\mathrm{X} = U_1 + U_2,\ U_1, U_2$ uniformly distributed between [0,1]. This is a special case (n=2) of Irwin-Hall distribution, and X follows a triangular distribution:

$$f_X(x) = \begin{cases} x, & 0 \le \mathrm{x} \le 1 \\ 2 - x, & 1 \le \mathrm{x} \le 2 \end{cases}$$

If we assume X is still uniformly distributed in its support, [0,2], as same as our BRA, then its probability density function is

$$f_X(x) = \frac{1}{2}, \qquad 0 \le \mathrm{x} \le 2.$$

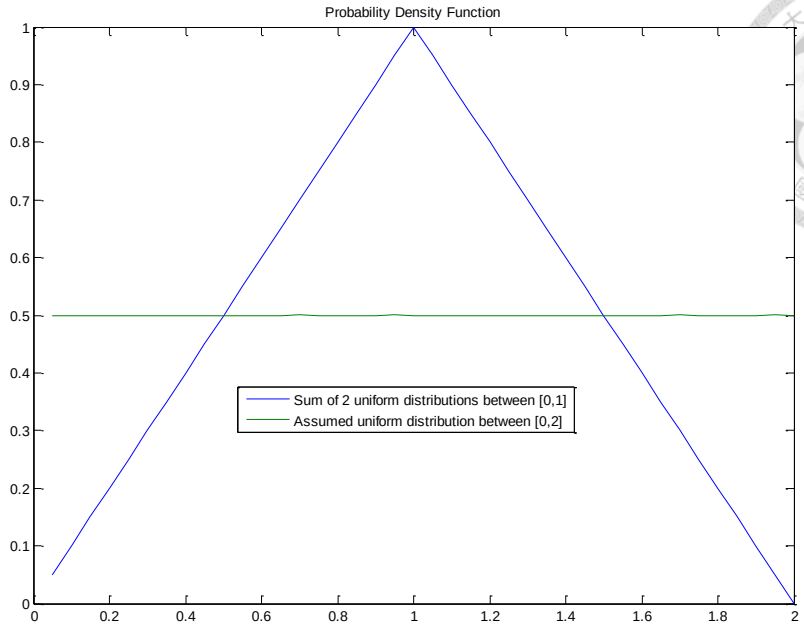


Figure 5.3 Actual p.d.f. and the p.d.f under our assumption

From Figure 5.3, we observe that the p.d.f. under our assumption is higher than actual p.d.f. while $2 \geq x \geq 1.5$ and $0.5 \geq x \geq 0$. Using the notation of D-3 in Section 4.2.2, it implies that we overestimate the probability of surely-win region and probability of surely-lose region, R_W and R_L, so the derived probability of correct ranking is over-optimistic in case of multiple stations. However, the range of true mean cycle time is often large so the value of density is essentially low and the extent of over-estimate has no major influence. The BRA is still applicable for re-entrant lines with multiple stations.

5.2 Extension to N Designs

Let us extend to the general case of N designs. Even though there are N

designs in total, the 2-design analysis is actually also applicable because rank correlation is based on the pair-wise comparisons. For design D_i and design D_j ,their approximated mean cycle times by QNA are ACT_i and ACT_j , where i= 1,2,…, N and j=i+1,⋯, N . (ECT_i, ACT_i) and (ECT_j, ACT_j) still has the probability of being a concordant pair, $P_c(i,j)$, and the probability of being a discordant pair, $1-P_c(i,j)$. Thus, if we rank designs completely in accordance with QNA approximations, then the expected number of concordant pairs is $\sum_{i=1}^{N}\sum_{j=i+1}^{N}P_c(i,j)$, and the expected number of discordant pairs is $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(1-P_c(i,j))$. Since there are $\frac{1}{2}\mathrm{N}(\mathrm{N}-1)$ pairs in total, the expected rank correlation is

$$\mathrm{E}[\mathrm{RC}] = \frac{E[\#\text{ of concordant pairs}]-E[\#\text{ of discordant pairs}]}{\text{Total number of pairs}}$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}P_c(i,j)-\sum_{i=1}^{N}\sum_{j=i+1}^{N}(1-P_c(i,j))}{\frac{1}{2}\mathrm{N}(\mathrm{N}-1)}$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}[2P_c(i,j)-1]}{\frac{1}{2}\mathrm{N}(\mathrm{N}-1)}$$

Because of $\sum_{i=1}^{N}\sum_{j=i+1}^{N}1 = \frac{1}{2}\mathrm{N}(\mathrm{N}-1)$,

$$\mathrm{E}[\mathrm{RC}] = 2\times\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}P_c(i,j)}{\frac{1}{2}\mathrm{N}(\mathrm{N}-1)} - 1$$

We define the average probability of being a concordant pair of all pairs is $\overline{P_c}$, $\overline{P_c} = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}P_c(i,j)}{\frac{1}{2}\mathrm{N}(\mathrm{N}-1)}$. Then, $\mathrm{E}[\mathrm{RC}] = 2\overline{P_c} - 1$.

5.3 Discussion of Variability

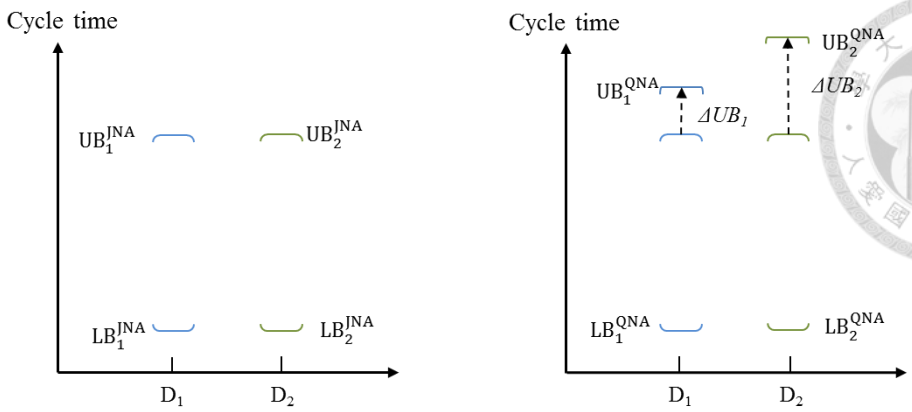
QNA utilizes heterogeneous SCVs to more delicately characterize network flows

but Jackson network approximation (JNA) only utilizes unity SCVs because of its exponential assumption. Here we further discuss about the effects of variability while approximating network performance and also investigate what advantages for ranking can acquire if taking the second order statistics into consideration. JNA ignores the differences in variance terms and uses only their mean values for evaluation and comparison, which intuitively reduces the recognition between designs, especially for those designs whose mean values are similar.

(1) Performances of D_1 and D_2 are similar

We discuss in the case of a single GI/G/m queue and also assume there are two allocation designs, D_1 and D_2. The only difference between D_1 and D_2 is the variance of inter-arrival time. However, because the exponential assumptions of arrival and service processes ignores the unique difference between D_1 and D_2, their JNA performances are the same as shown in Figure 5.4 (a).

For QNA approximation, different network configuration resulted from D_1 and D_2 leads to different characterization of traffic flows which approximated by traffic variability equations as Equation (3.14). QNA utilizes the variance terms to capture more information about arrival and service processes and the minor difference of variabilities is enough to identify which design is better.



(a) JNA performance (b) QNA performance

Figure 5.4 D_1 and D_2 are similar, (a) JNA (b) QNA

Define the following notations for JNA performance and so as QNA:

ΔUB_1 : Derivation of upper bound of D_1, $UB_1^{QNA} - UB_1^{JNA}$

ΔUB_2 : Derivation of upper bound of D_2, $UB_2^{QNA} - UB_2^{JNA}$

Here we ignore the difference of lower bound because that has minor effect in comparison with upper bound. Based on Equation (4.3), if use the JNA performance of D_1 and D_2 for ranking, the probability of being a concordant pair is $\frac{1}{2} + \frac{1}{2}\frac{\left(UB_2^{JNA} - UB_1^{JNA}\right)}{\left(UB_2^{JNA} - LB_2^{JNA}\right)} = \frac{1}{2}$. If use the QNA performance for ranking, upper bounds of D_1 and D_2 change ΔUB_1 and ΔUB_2 respectively, with the variations of bounds, the probability of being a concordant pair is $\frac{1}{2} + \frac{1}{2}\frac{(\Delta UB_2 - \Delta UB_1)}{\left(UB_2^{JNA} - LB_2^{JNA} + \Delta UB_2\right)}$, which is greater than $\frac{1}{2}$. Because of the additional variability information of each design, variabilities press the differences between those designs with similar performances, and this assists to differentiate which one design is better and improve the probability of correct ranking. Clearly, the more we capture of these key factors, the greater is our chance of recognizing some differences between designs. Furthermore, it is significant that JNA

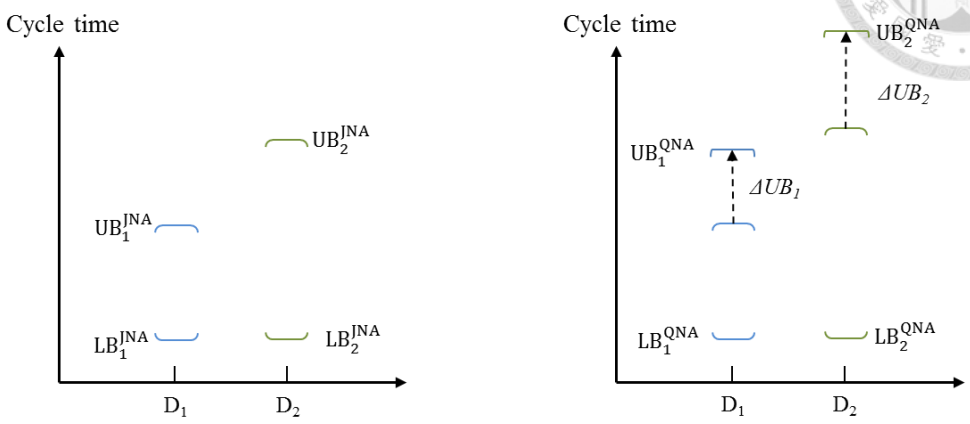
ignores the differences caused by variability and it is a special case of QNA when the true network is exponential distributions (CV=1). Therefore, it is no surprise that QNA has better rank correlation than JNA.

(2) <u>Performances of D_1 and D_2 have distinct differences</u>

Taking variability into consideration is good for pressing the differences between similar designs but it may sometimes deteriorate ranking. For example, there are two designs whose JNA performances have distinct differences as shown in Figure 5.4(a) and the probability of being a concordant pair is $\frac{1}{2}+\frac{1}{2}\frac{\left(UB_2^{JNA}-UB_1^{JNA}\right)}{\left(UB_2^{JNA}-LB_2^{JNA}\right)}$. If use the QNA performance for ranking, upper bounds of both D_1 and D_2 change ΔUB_1 and ΔUB_2 respectively as shown in Figure 5.4(b). With the variations of bounds, the probability of being a concordant pair is $\frac{1}{2}+\frac{1}{2}\frac{\left(UB_2^{JNA}+\Delta UB_2-UB_1^{JNA}-\Delta UB_1\right)}{\left(UB_2^{JNA}-LB_2^{JNA}+\Delta UB_2\right)}$. Thus, if ΔUB_2 is smaller than $\frac{UB_2^{JNA}-LB_2^{JNA}}{UB_1^{JNA}-LB_2^{JNA}}\Delta UB_1$, then the variation of bounds caused by the increased variability would deteriorate the probability of correct ranking and also cloud the judgment on ranking at the same time. Fortunately, in most of cases ΔUB_2 is much greater than ΔUB_1 because the upper bound is exponentially increasing with utilization.

We shortly summarize the above discussion here. First, variability benefit press the differences between similar designs and also improve the probability of correct ranking. Second, at the same time, variability may deteriorate the probability of

correct ranking because of blurring the recognition between designs with distinct differences. But fortunately, the second one is not common.



(a) JNA performance (b) QNA performance

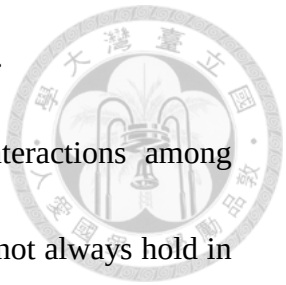
Figure 5.5 D_1 and D_2 have distinct differences, (a) JNA (b) QNA

In this section, we compare JNA and QNA in their ranking performance and also investigate how variability influences the goodness of ranking. In viewpoint of OT, the resulting benefit of variability is significant because it facilitates us to recognize the relative orders of top designs whose performances are similar more precisely. Even though taking variability into consideration makes the rankings of those designs with distinct differences a little blurred at the same time, this is however the secondary concern.

5.4 Summary

In this chapter, we first extend the BRA of a single GI/G/m queue in Section 4.3 to a general re-entrant line with M stations. Because parametric decomposition regards each station as an independent node given the approximate flow parameters,

therefore we generalize BRA by superposition of individual stations.



Unfortunately, unlike a single GI/G/m queue, there are interactions among stations in a general production line which causes Lemma 4.5 does not always hold in case of M stations. There are four possible cases of relations of upper and lower bounds and BRA derives the probability of being a concordant pair of these four cases. Our experimental statistics shows that most of pairs (over 89%) still holds Lemma 4.5 and over 94% pairs have probability of being a concordant pair greater than 0.5. Then, the BRA result is extended from 2 designs to N designs using rank correlation. The effects of variability are also discussed by means of comparing JNA and QNA. From our discussion, variability benefits ranking those top designs because variability presses their differences even though it also blurs some rankings at the same time.

Chapter 6

Machine Capacity Allocation Experiments

Machine capacity allocation is to determine how best to assign machines in each workstation of a queueing network in order to minimize the expected total mean cycle time. We study systems with several types of products and general inter-arrival/service time distributions. For each product type, process flow is re-entrant through systems, which means product routing makes multiple visits to individual workstations and may also compete for the finite capacity of a workstation. For such networks there are no analytical formulas for mean cycle time performance. OO-based method is used to address this kind of problem. Here we conduct an experiment to analyze how the selection of simplified models affects the ranking for OO and compare two simplified models, QNA and JNA, where the main difference is the characterization of variability, one being heterogeneous SCVs and the other being unity SCVs.

Section 6.1 summarizes the flowchart of our experiment. Section 6.2 discusses how to select promising designs after fast evaluations of simplified model for further steps. Section 6.3 describes the simulation model and experiment factors considered in this set of experiments. Section 6.4 shows the experiment results and several

comparisons on ranking between QNA and JNA. Section 6.5 discusses the efficiency.

6.1 Overview

There are two comparable simplified models, QNA and JNA. In the procedures of parametric decomposition method, it needs the following inputs: product routings, product releases, and processing steps of each product. With above information, QNA derives all the means and SCVs of both inter-arrival and service times of a re-entrant line. Parameter $(\lambda_{am}, C_{am}^2, \tau_m, C_{sm}^2)$ of each node can be utilized to estimate the node-level measures and the system-level measures. Due to exponential assumption of JNA, each node is characterized using two parameters $(\lambda_{am}, 1, \tau_m, 1)$ which assumes unity SCVs. Based on these two simplified models, we can quickly take a glance over the whole solution space and rank all designs in terms of approximated performances respectively. According to these rankings, we apply ordinal transformation on the original solution space. Then in the ordinal space, we screen top-ranking designs for further evaluation by DES simulation model. Finally, we simulate all screened designs by detailed DES simulation model and find the optimal one as our best design. This flowchart of experiment is shown in Figure 6.1.

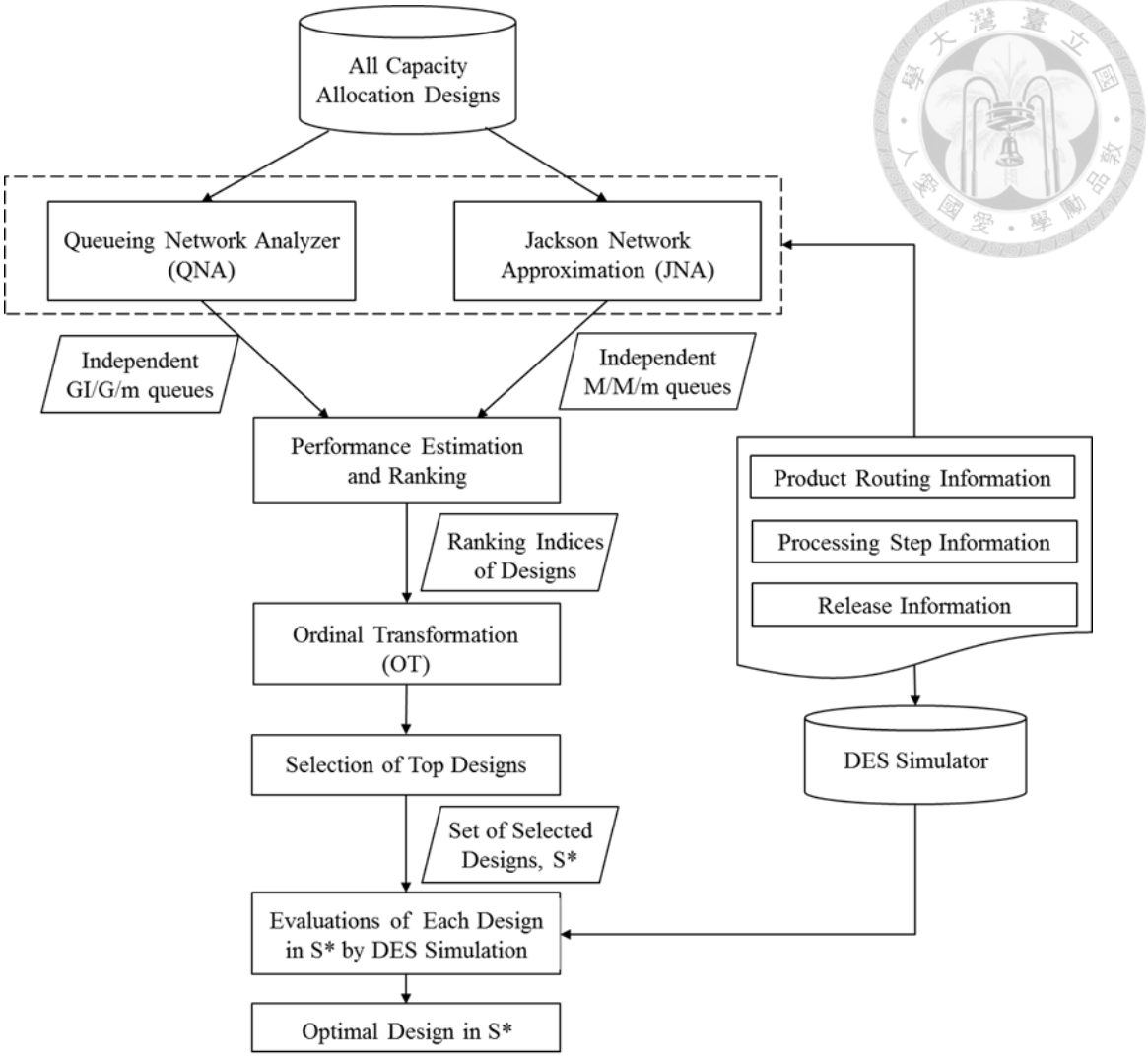


Figure 6.1 Flowchart of Experiment

6.2 Selection of Top Designs in Ordinal Space

After OT, we select top designs in the ordinal space for further evaluations by DES simulation model, and the resulting set of selected top designs is denoted as S^*. This section emphasizes on how to determine the number of selected designs, $|S^*|$.

First, we show a boundary, k, to guarantee the quality of selected designs in simplified model in Lemma 6.1.

Lemma 6.1: Given *n* designs in total and the rank correlation τ between simplified model and detailed model, there exists a boundary, $k=\left\lfloor\sqrt{\frac{(1-\tau)n(n-1)}{4}}\right\rfloor+1$, to make sure that in the top-*k* designs of simplified model there is at least one design also in the top-*k* designs of detailed model.

Proof:

In other words, given rank correlation, find the maximum of $\hat{k}$ that may result in all of the top-$\hat{k}$ designs of simplified model being not in the top-$\hat{k}$ designs of detailed model. Then, the boundary k is $\lfloor\hat{k}\rfloor+1$.

The maximum of $\hat{k}$ occurs when all of the top-$\hat{k}$ designs of simplified model being not in the top-$\hat{k}$ designs of detailed model is the only source of discordant pairs, and causes that the number of discordant pairs= $\hat{k}^2$. From the definition of rank correlation τ in Section 4.1, $\tau=\frac{\text{Number of concordant pairs}-\text{Number of discordant pairs}}{\text{Total number of pairs}}$ and total number of pairs = number of concordant pairs + number of discordant pairs, $\frac{1}{2}n(n-1)$. Thus, the number of discordant pairs is $\frac{1-\tau}{2}\times\frac{1}{2}n(n-1)$.

The maximum of $\hat{k}$ is the positive root of $\hat{k}^2=\frac{1-\tau}{2}\times\frac{1}{2}n(n-1)$, so

$$\hat{k}=\sqrt{\frac{(1-\tau)n(n-1)}{4}} \quad \text{and} \quad k=\lfloor\hat{k}\rfloor+1=\left\lfloor\sqrt{\frac{(1-\tau)n(n-1)}{4}}\right\rfloor+1.$$

Q.E.D.

From Lemma 6.1, we can make sure that in the top-*k* designs of simplified model there is at least one design also in the top-k designs of detailed model, where

$k = \left\lfloor \sqrt{\frac{(1-\tau)n(n-1)}{4}} \right\rfloor + 1$. Therefore, if the cost of computation of k designs is endurable, then we could set the number of selected designs to k, $|S^*| = k$, and guarantee that in those k designs there exists at least one design placed in the top-k of detailed model.

In practice, the number of selected designs depends on how much computation budget we have. If total computation budget is limited and it is not enough to contain the guaranteed k designs, then an intuitive selection strategy is to fully utilize total computation budget. If assume that $C(R)$ is the computation cost of executing R replications of DES simulation and total computation budget is limited to B, then $|S^*|$ is set to $\frac{B}{C(R)}$.

6.3 Re-entrant Network Models and Experiment Factors

Detailed re-entrant queueing network model is introduced and experiment factors are also discussed in this section.

6.3.1 Simulation model: 5-station and 2-product model

This simulation model, as shown in Figure 6.2, has multiple product types, re-entry, failure-prone, and general service and inter-arrival time distributions. There are two types of products, P1 and P2. Each product has a deterministic routing through the system, as shown in Figure 6.3. There are the mean release rates of 1.0

jobs/sec for P1 and 1.25 jobs/sec for P2, and their inter-arrival time distributions are shown in Table 6.1.

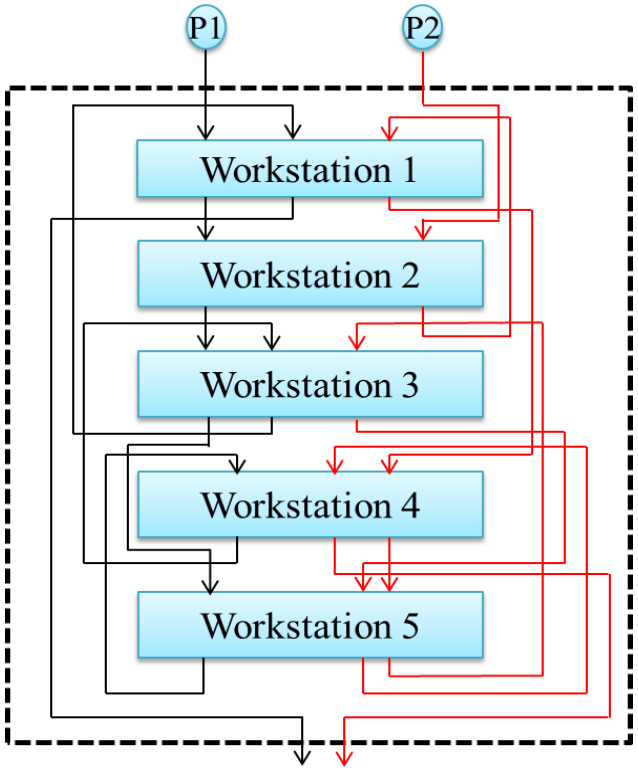


Figure 6.2 A five-workstation and two-product re-entrant Line

Product 1 (P1) Routing
Enter → 1 → 2 → 3 → 5 → 4 → 3 → 1 → Exit
Product 2 (P2) Routing
Enter → 2 → 1 → 4 → 5 → 3 → 4 → 5 → Exit

Figure 6.3 Routing of each product

Table 6.1 Release of each product

Product Release			
Product	Inter-arrival time distribution	Mean (sec)	SCV
1	Log-normal	1	0.8
2	Log-normal	0.8	0.8

Processing times of one product type at the same workstation are different due to various processing steps. The model involves failure-prone processing workstations, each having one or more identical but independent machines. Both times between failures and times to repair follow exponential distributions, more details in Table 6.2. Processing times of products at each processing step are shown in Table 6.3 and all are generally distributed.

Table 6.2 Workstation failure setting

Workstation			
Workstation	MTTF[1] (sec)	MTTR[2] (sec)	Distributions
1	40	10	Both exponential
2	60	10	Both exponential
3	40	10	Both exponential
4	60	10	Both exponential
5	40	10	Both exponential

MTTF[1] : Mean Time To Failure, MTTR[2] : Mean Time To Repair

Table 6.3 Processing steps of each product

Product 1, P1				
Step	At Workstation	Service time distribution	Mean (sec)	SCV
1	1	Log-normal	1	0.5
2	2	Uniform	1.66	0.8
3	3	Erlang Order 2	1	0.5
4	5	Log-normal	1.25	0.5
5	4	Erlang Order 2	1.66	0.8
6	3	Log-normal	1	0.5
7	1	Uniform	1	0.5
Product 2, P2				
Step	At Workstation	Service time distribution	Mean (sec)	SCV
1	2	Uniform	1.66	0.8
2	1	Uniform	1	0.5
3	4	Erlang Order 2	1.66	0.8
4	5	Log-normal	1.25	0.5
5	3	Erlang Order 2	1	0.5
6	5	Log-normal	1.25	0.5
7	4	Erlang Order 2	1.66	0.8

6.3.2 Experiment Factors

Control factors considered in this experiment are number of machines allocated in each workstation given a total number of machines. The corresponding dependent factor is the performance measure of each allocation design, which is expected mean cycle time in our experiment.

Design

Given total number of machines N and number of workstations M, a machine allocation design is specified by the number of machines in each workstation, represented by a set with M elements. In this set, the m^{th} element stands for that there are M_m machines to be allocated in workstation m. Every machine must be allocated to a workstation. Therefore, the machine allocation design indexed by k can be written as $D_k = \{M_1^k, M_2^k, \dots, M_M^k\}$ where $M_1^k + M_2^k + \cdots + M_M^k = N$.

In this experiment, we set N=37 and M=5. Additionally, we limit number of machines in a workstation to be not less than 5 and not more than 10.

Performance Index

In practice, it is common that after promising to undertake the orders, managers break down the market demands into monthly, weekly, or even daily production targets to push the staff forward. Based on the production targets, product releases are determined, so in our experiment the product releases are known. If given product

releases and satisfying machine capacities, then optimal system throughputs are equivalent to the product releases. While meeting the target requirements, operation practitioners pursue mean cycle time reduction because shortening mean cycle time of whole system improves the ability to response to the variation of market demands and also satisfies customer requirements. Thus, we consider expected mean cycle time of whole system as the performance index.

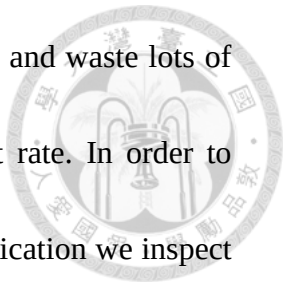
To find an optimal machine capacity allocation design $D_{k^*} = \{M_1^{k^*}, M_2^{k^*}, \dots, M_M^{k^*}\}$, to minimize the average of mean cycle time of each product, where

$$k^* = \underset{k}{\operatorname{argmin}} \frac{1}{I} \sum_{i=1}^{I} \mathrm{MCT}_i(D_k), D_k \in D$$

and $\mathrm{MCT}_i(D_k)$ denotes the mean cycle time of product type i under a specific allocation design D_k.

Simulation Setting

While using DES simulations, simulating each design over detailed DES simulation runs 30 replications with a different random seed and statistics are averaged over all the 30 simulation replications and the simulation horizon is set to 30 minutes. Because of operation variability and uncertain failure, some inadequate allocation designs would make system unstable, which means utilization of a workstation close to 1. Such an unstable system is the last thing managers would like



to see because that unstable system would suffer long queue length and waste lots of capacity, which causes long mean cycle time and low throughput rate. In order to screen those unstable designs out, at the end of each simulation replication we inspect whether this allocation design results in utilization of any workstation approaching to 1. If yes, we label this design as an unstable design and would not consider this design in the following discussion.

CPU Time

In our experiment, we would like to acquire a top-quality solution in limited time so we constrain the maximal CPU time to 10 minutes. Because executing 30 replications of DES simulation for one design takes approximately 1 minute, according to the discussion in Section 6.2, the number of selection of top designs in simplified model is set to 10, $|S^*| = 10$.

6.4 Numerical Results

We first start from the brute force method which simulates all designs using DES simulator. Unfortunately, because of operation variability and uncertain failure, there are 365 designs being labeled as unstable designs. Note that Performances of unstable designs are average cycle time of the collectable data until the simulation ends. In brute force method, these unstable designs are not found before run simulations but

QNA and JNA are easy to identify those designs because in the procedures of parametric decomposition method we estimate the expected utilization of each node (workstation) which assists us with screening out the unstable designs. Most importantly, these unstable designs in QNA and JNA are wholly identical to the unstable designs in DES simulation. The amazing observation implies that parametric decomposition method essentially captures the characteristics of traffics in network and is useful for filtering out the unstable designs without running simulations. The result is shown in Figure 6.4. We would not consider these unstable designs in the following discussion. After removing those unstable designs, total number of allocation designs is 415. Figure 6.5 presents the performances of DES simulation in the original design space.

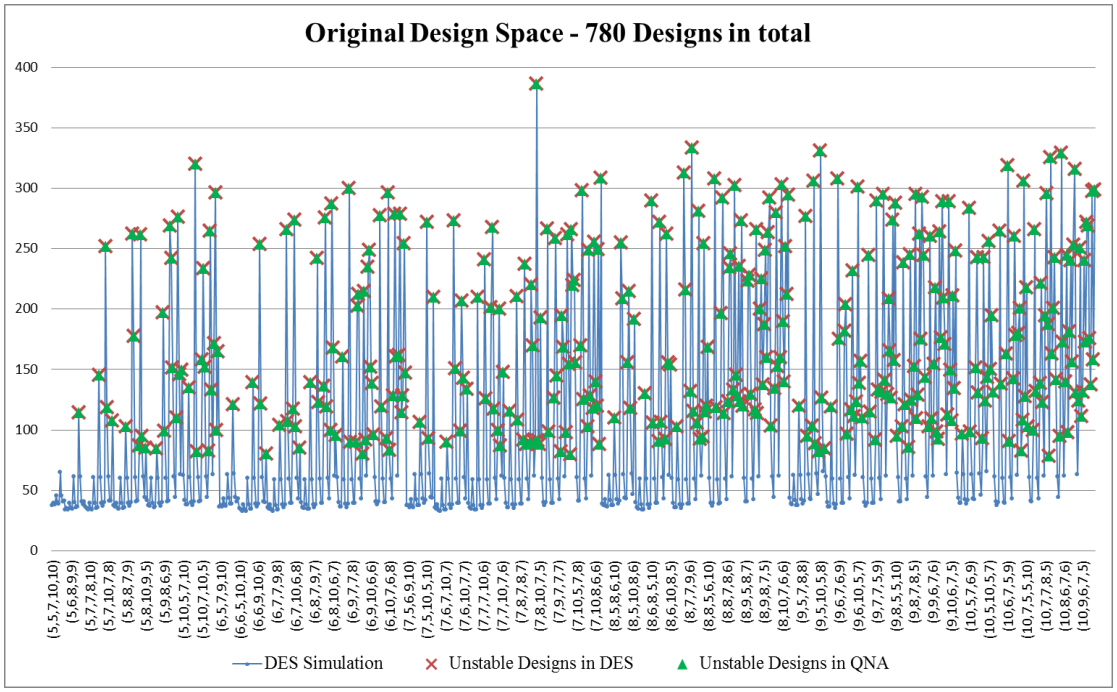


Figure 6.4 Unstable designs labeled in both Simulation and QNA(JNA)

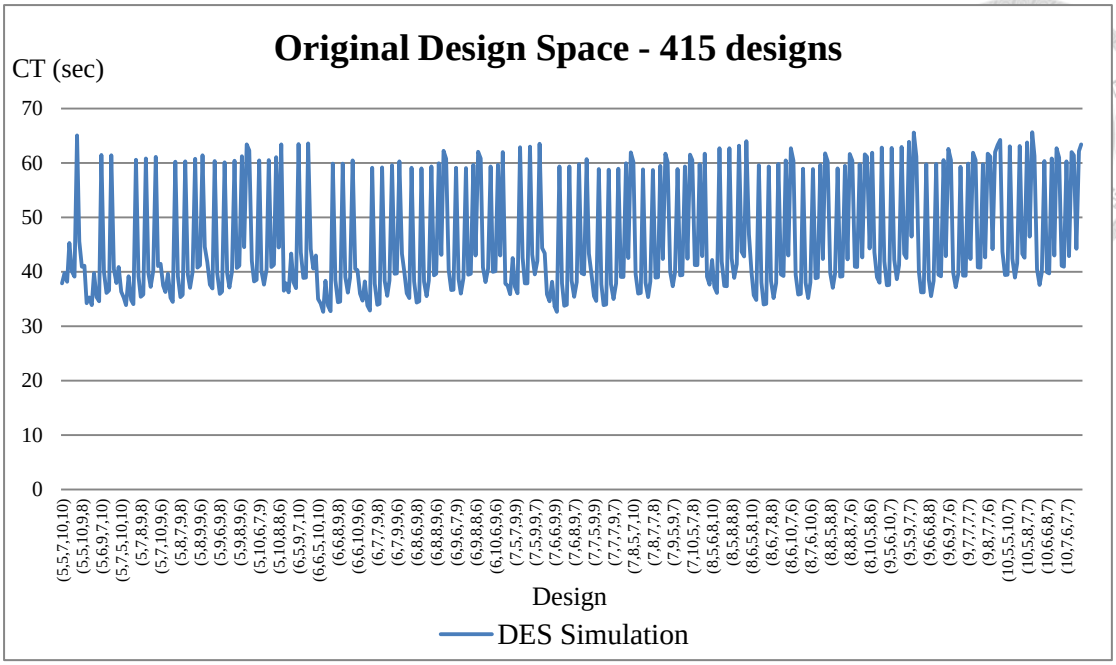


Figure 6.5 Performances of DES simulation in original design space

We approximate the mean cycle time performance of each design by the simplified model and then rank all designs in terms of their approximated mean cycle time. By these rankings of designs, we can apply OT to transform the original design space to the ordinal space, and the result of QNA is in Figure 6.6 and the result of JNA is in Figure 6.7. The ranking among designs by QNA is quite accurate, rank correlation=0.8545, which is better than the rank correlation of JNA(=0.8245).

After OT, designs with similar performances are grouped together, and most of top designs are clustered at the left end of the transformed space which facilitates us more efficiently focus on that region. We pick top-10 designs in the ordinal space for further evaluations by DES simulation. We compare the result of top-10 designs selected by QNA and JNA. The comparison of these two approximation methods is shown in Table 6.4.

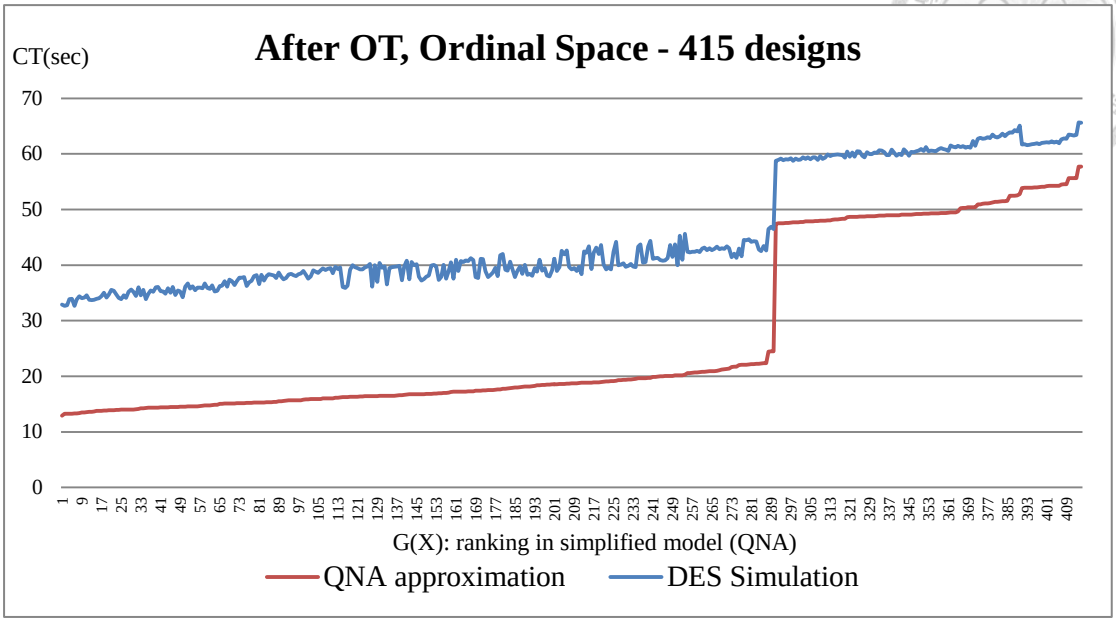


Figure 6.6 Performances of DES Simulation after OT by QNA

and Rank Correlation = 0.8545

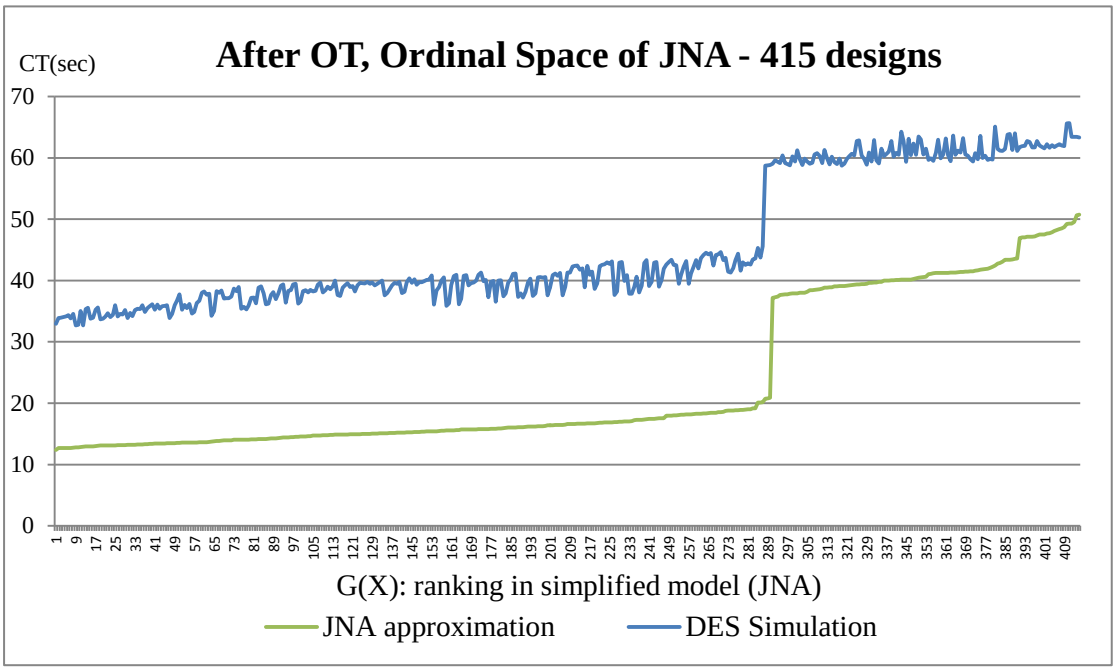


Figure 6.7 Performances of DES Simulation after OT by JNA

and Rank Correlation = 0.8245

Table 6.4 True rankings of the selected top-10 designs in QNA and JNA

True rankings of the selected top-10 designs in QNA		True rankings of the selected top-10 designs in JNA	
(6,7,6,10,8)	**4**	**(6,7,6,10,8)**	**4**
(7,6,6,10,8)	**1**	**(7,7,6,9,8)**	**10**
(6,6,7,10,8)	**3**	(6,7,7,9,8)	13
(7,7,6,9,8)	**10**	(7,7,6,10,7)	14
(6,7,7,9,8)	13	(6,7,7,10,7)	18
(6,6,6,10,9)	**2**	(6,8,6,9,8)	21
(6,7,6,9,9)	**8**	**(6,7,6,9,9)**	**8**
(6,8,6,9,8)	21	(6,8,6,10,7)	25
(7,7,6,10,7)	14	**(7,6,6,10,8)**	**1**
(6,7,7,10,7)	18	**(6,6,7,10,8)**	**3**

Figure 6.8 is the comparison between QNA and JNA in terms of rank correlation of their own top-K designs. The result shows that QNA always outperforms JNA and the difference of their rank correlations is especially large when K is small, which means that QNA is much likely to correctly rank true top designs. This result corresponds to our discussion D-5 in Section 4.2.2 and discussion about variability in Section 5.3. The characterization of heterogeneous SCVs is beneficial to recognize designs, especially top designs.

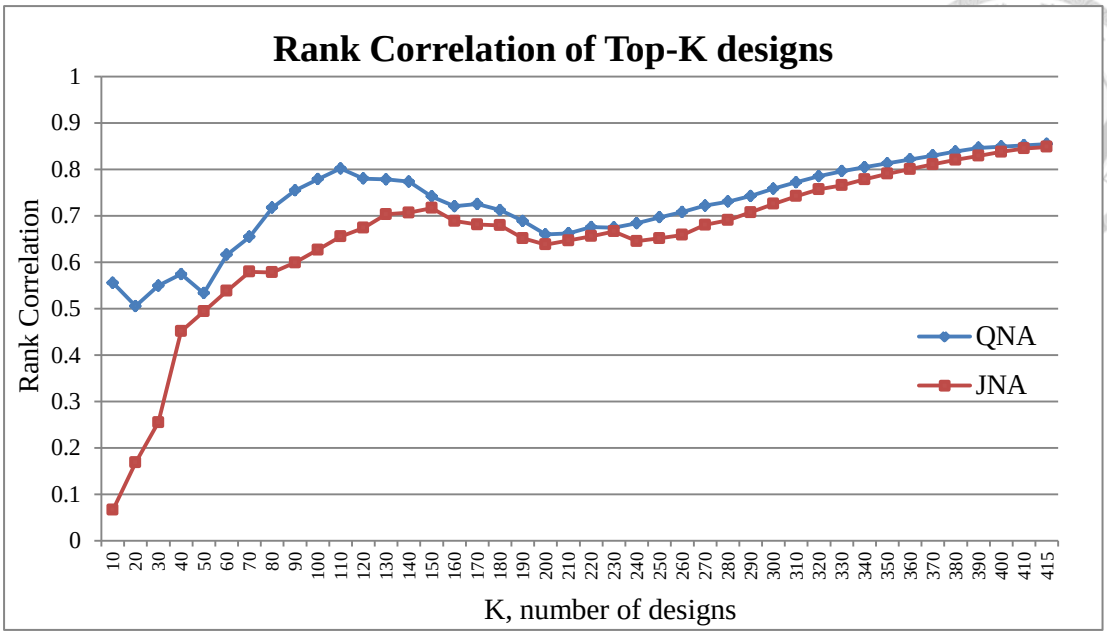


Figure 6.8 Comparison between QNA and JNA in rank correlation of top-K designs

To further investigate why heterogeneous SCVs benefit top designs, we use true performance as the ranking index and apply ordinal transformation. In order to let each group have its representativeness, we regard each thirty designs as a group like black circle as shown in Figure 6.9. After grouping, we found that with heterogeneous SCVs the difference of mean between groups increases, which implies heterogeneous SCVs benefits better differentiation between groups as shown in Figure 6.10. Furthermore, in Figure 6.11 heterogeneous SCVs result in the coefficient of variance of each group increasing, which shows heterogeneous SCVs make designs in a group better separated and improves the recognition between designs in the same group. The above amazing result shows that capturing the characterization of SCVs really advantages their ranking because it is beneficial to better separate designs no matter in a group or between groups.

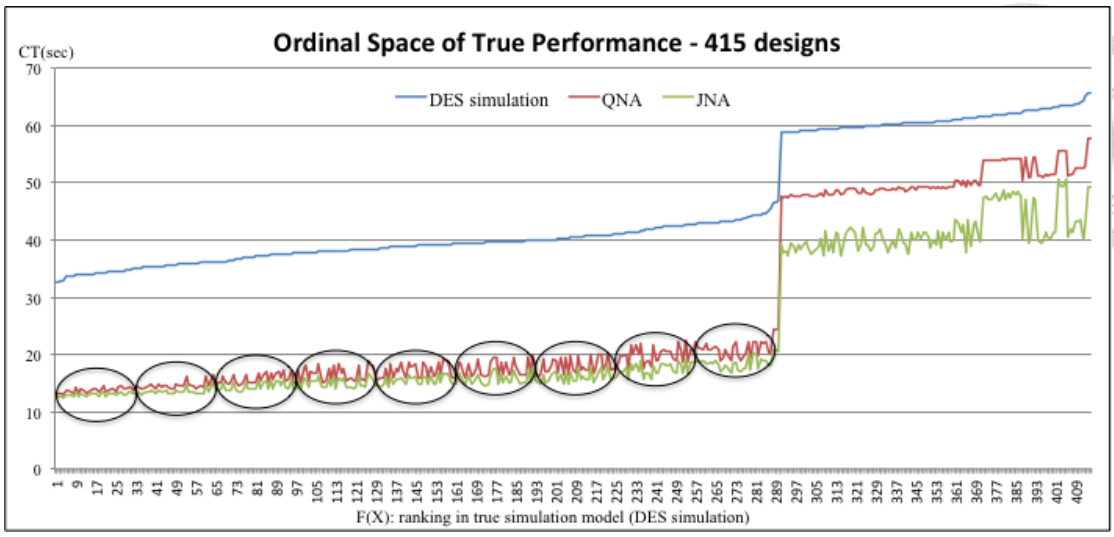


Figure 6.9 Grouping after ordinal transformation using true performance

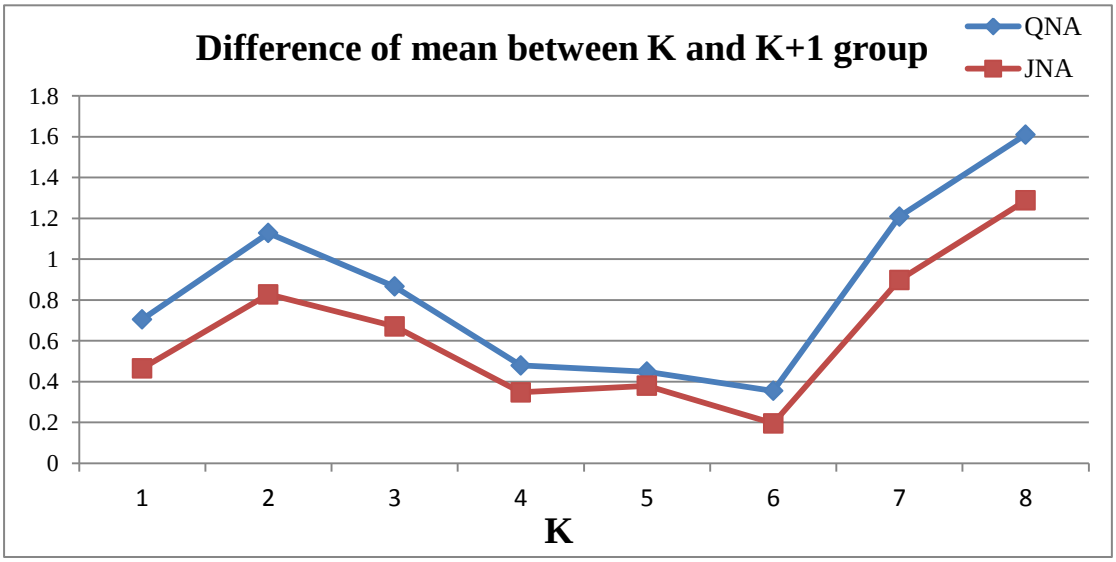


Figure 6.10 Difference of mean between K^{th} and $K+1^{th}$ group

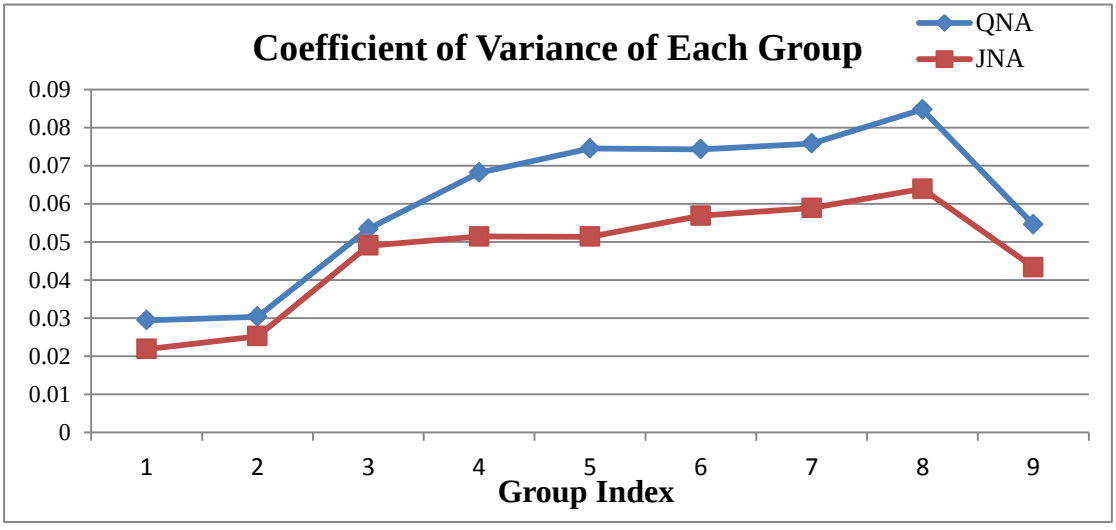


Figure 6.11 Coefficient of variance of each group

Based on our proof in Chapter 4 and insights from above observation, it is no surprise that ranking according to QNA approximations has higher probability of correct ranking and the rank correlation in this experiment is 0.8545 which is better than rank correlation of JNA (0.8245) because of heterogeneous SCVs. In views of computation time, QNA (0.832 sec) has no major increase in the comparison with JNA (0.825) as shown in Table 6.5.

Table 6.5 Comparison of computation time between QNA and JNA

Computation Time (second)	
Evaluation of all designs by QNA	Evaluation of all designs by JNA
0.853	**0.825**

- Compared between QNA and JNA, the benefits of characterization of SCVs:

(1) Increases both difference of mean between groups and coefficient of variance in a group, which causes designs better separated and enhance the probability of correct ranking, so as rank correlation

(2) Improves rank correlation, especially for top designs, which coincides with our analysis in Chapter 4 and supports by our experiment result

(3) Improves ranking performance without major increase of computational time

6.5 Efficiency of Using Simplified Models for OT

The DES simulation is developed in a commercial software, Plant Simulation 8.1.

Each replication takes 2 seconds in average and each design needs 30 replications. For brute force method there are 415 designs in total, which include the unstable designs because brute force method could not identify unstable designs before running simulations. So, computation time of brute force method is 415*30*2 seconds, approximately equal to 7 hours.

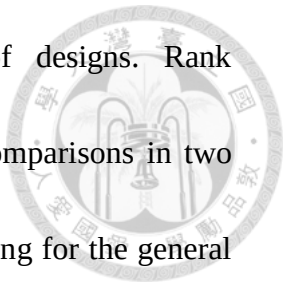
For OT, the mathematical formulas of parametric decomposition method are implemented by Matlab 2010a. Evaluations of all 415 designs by QNA or JNA take less than one second of CPU time, approximately 0.8 second. In the comparison with computation time of DES simulation, computation time of QNA or JNA can be ignored. QNA considers heterogeneous SCVs and takes additional computation time of milliseconds compared with JNA, but acquires a great improvement on ranking of top designs. This deal is actually a real bargain and cost-effective.

Chapter 7

Conclusions

In this thesis, how selecting the simplified models affects ranking for OO is investigated and specify re-entrant line machine capacity allocation problem as the conveyor problem. Parametric decomposition method was exploited to the considered re-entrant line. Based on parametric decomposition method, we compared two simplified models: queueing network analyzer (QNA) and Jackson network approximation (JNA). The major difference is only the characterization of variability terms, QNA being heterogeneous SCVs and JNA being unity SCVs because of exponential assumptions.

To analyze the goodness of ranking by simplified models in theory, we developed a bound and ranking analysis, BRA, and took the first step to investigate the probability of correct ranking in case of single GI/G/m queue with two designs. In single GI/G/m queue, bound analysis showed that QNA is bounded by the upper and lower bound presented by Kingman, and Brumelle and Marchal respectively. In addition, with the variation of QNA, the least variation of upper bound was derived. This facilitates us to derive a better probability of correct ranking α. In our experiment, α is greater than 0.75 which is significantly difference with the probability of 0.5 like tossing a coin. Based on single GI/G/m queue, BRA is extended to general re-entrant

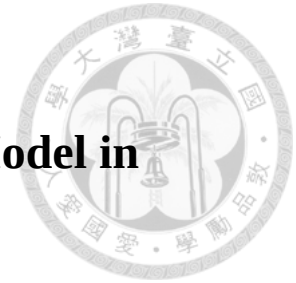


queueing network with multiple workstations and amount of designs. Rank correlation, a statistic to measure the concordance of pair-wise comparisons in two quantitative indices, is introduced to quantify the goodness of ranking for the general cases.

Simulation studies demonstrated that rank correlation of QNA always outperforms JNA, especially significant for top-10 designs, which coincided with our BRA. Then, the original designs space is transformed by true ranking, and cluster each thirty designs into a group in this ordinal space. After grouping, it shows that heterogeneous SCVs benefit differentiation between groups and also make designs in a group better separated. That is why considering heterogeneous SCVs in simplified models improve the rank correlation. In this thesis, we investigated how selecting simplified models of different variability affects ranking for OO and support the validity of using ranking information by simplified models for optimization in aspects of theory and experiment.

Appendix

Ranking Analysis of QNA as Simplified Model in Other Cases



A.1 Ranking analysis of QNA as simplified model for Single GI/G/m queue under the assumption of normal distribution of actual cycle time

We assume that CT_1 and CT_2 are normal distributions, which the means of CT_1 and CT_2 are ACT_1 and ACT_2 fortunately, $\mathrm{CT}_1 \sim N(ACT_1, \sigma_1^2)$ and $\mathrm{CT}_2 \sim N(ACT_2, \sigma_2^2)$. Given $\mathrm{ACT}_1 < \mathrm{ACT}_2$, the probability of being a concordant pair is

$$\mathrm{P}[\mathrm{CT}_1 < \mathrm{CT}_2 | LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]$$

$$= \frac{P[\mathrm{CT}_1 < \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]}{P[LB_1 < \mathrm{CT}_1 < UB_1, LB_2 < \mathrm{CT}_2 < UB_2]}$$

$$= \frac{P[\mathrm{CT}_1 < \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]}{P[LB_1 \le \mathrm{CT}_1 \le UB_1] \times P[LB_2 \le \mathrm{CT}_2 \le UB_2]}$$

$$= \frac{P[\mathrm{CT}_1 < \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]} \quad \text{(A.1)}$$

Define $\mathrm{Z} = \mathrm{CT}_2 - \mathrm{CT}_1$, sum of two independent normal distributions, $\mathrm{CT}_1 \sim N(ACT_1, \sigma_1^2)$ and $\mathrm{CT}_2 \sim N(ACT_2, \sigma_2^2)$, is also a normal distribution, $\mathrm{Z} \sim N(\mu_z = ACT_2 - ACT_1, \sigma_z^2 = \sigma_1^2 + \sigma_2^2)$. The support of Z is $LB_2 - UB_1 \le \mathrm{Z} \le UB_2 - LB_1$.

$$P[\mathrm{CT}_1 < \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]$$

$$= P[Z > 0, LB_2 - UB_1 \le \mathrm{Z} \le UB_2 - LB_1]$$

$$= \mathrm{P}[0 \le \mathrm{Z} \le UB_2 - LB_1]$$

$$= \Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) - \Phi\left(\frac{0-\mu_z}{\sigma_z}\right)$$

Therefore, the probability of being a concordant pair is

$$\mathrm{P}[\mathrm{CT}_1 < \mathrm{CT}_2 | LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]$$

$$= \frac{\Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) - \Phi\left(\frac{0-\mu_z}{\sigma_z}\right)}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]}$$

The probability of being a discordant pair is $\mathrm{P}\{\mathrm{CT}_1 > \mathrm{CT}_2 \mid \mathrm{ACT}_1 < \mathrm{ACT}_2\}$

$$\mathrm{P}[\mathrm{CT}_1 > \mathrm{CT}_2 | LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]$$

$$= \frac{P[\mathrm{CT}_1 > \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]}$$

$$= \frac{P[\mathrm{CT}_1 > \mathrm{CT}_2, LB_1 \le \mathrm{CT}_1 \le UB_1, LB_2 \le \mathrm{CT}_2 \le UB_2]}{P[LB_1 \le \mathrm{CT}_1 \le UB_1] \times P[LB_2 \le \mathrm{CT}_2 \le UB_2]}$$

$$= \frac{P[Z < 0, LB_2 - UB_1 \le \mathrm{Z} \le UB_2 - LB_1]}{P[LB_1 \le \mathrm{CT}_1 \le UB_1] \times P[LB_2 \le \mathrm{CT}_2 \le UB_2]}$$

$$= \frac{P[LB_2 - UB_1 \le \mathrm{Z} \le 0]}{P[LB_1 \le \mathrm{CT}_1 \le UB_1] \times P[LB_2 \le \mathrm{CT}_2 \le UB_2]}$$

$$= \frac{\Phi\left(\frac{0-\mu_z}{\sigma_z}\right) - \Phi\left(\frac{LB_2 - UB_1 - \mu_z}{\sigma_z}\right)}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]} \quad \text{(A.2)}$$

Equation (A.1) – Equation (A.2)

$$= \frac{\left[\Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) - \Phi\left(\frac{0-\mu_z}{\sigma_z}\right)\right] - \left[\Phi\left(\frac{0-\mu_z}{\sigma_z}\right) - \Phi\left(\frac{LB_2 - UB_1 - \mu_z}{\sigma_z}\right)\right]}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]}$$

$$= \frac{\Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) + \Phi\left(\frac{LB_2 - UB_1 - \mu_z}{\sigma_z}\right) - 2\Phi\left(\frac{-\mu_z}{\sigma_z}\right)}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]}$$

$$= \frac{\Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) - \Phi\left(\frac{UB_1 - LB_2 + \mu_z}{\sigma_z}\right) + 2\Phi\left(\frac{\mu_z}{\sigma_z}\right) - 1}{\left[\Phi\left(\frac{UB_1-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{LB_1-\mu_1}{\sigma_1}\right)\right]\left[\Phi\left(\frac{UB_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{LB_2-\mu_2}{\sigma_2}\right)\right]} \quad \text{(A.3)}$$

According to Lemma 4.5, if $\mathrm{ACT}_1 < \mathrm{ACT}_2$, $UB_2 > UB_1$, $LB_2 > UB_1$, and $\mu_z > 0$, then $\Phi\left(\frac{UB_2 - LB_1 - \mu_z}{\sigma_z}\right) > \Phi\left(\frac{UB_1 - LB_2 + \mu_z}{\sigma_z}\right)$ and $2\Phi\left(\frac{\mu_z}{\sigma_z}\right) > 1$. Equation (A.3) must be positive. It implies that in normal distributions ranking two designs according to their approximated mean cycle times by QNA makes sure that the probability of being a concordant pair ($\mathrm{P_c}$) is greater than the probability of being a discordant pair, $\mathrm{P_c} > 0.5$.

References

[1] Xu, Jie, et al. "An ordinal transformation framework for multi-fidelity simulation optimization." *Automation Science and Engineering (CASE), 2014 IEEE International Conference on*. IEEE, (2014): 385-390.

[2] Shanthikumar, J. George, and John A. Buzacott. "Open queueing network models of dynamic job shops." *The International Journal Of Production Research* 19.3 (1981): 255-266.

[3] G. R. Bitran and D. Tirupati , "Multi-product Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference", *Management Science,*Vol.34, No.1(1988): 75-100.

[4] Connors, Daniel P., Gerald E. Feigin, and David D. Yao. "A queueing network model for semiconductor manufacturing." *Semiconductor Manufacturing, IEEE Transactions* on 9.3 (1996): 412-427.

[5] Jackson, James R. "Job shop-like queueing systems." *Management science 10.1* (1963): 131-142.

[6] Kelly, Frank P. "Networks of queues with customers of different types." *Journal of Applied Probability* (1975): 542-554.

[7] Gordon, William J., and Gordon F. Newell. "Closed queuing systems with exponential servers." *Operations research* 15.2 (1967): 254-265.

[8] Baskett, Forest, et al. "Open, closed, and mixed networks of queues with different classes of customers." *Journal of the ACM (JACM)* 22.2 (1975): 248-260.

[9] H. Kobayashi. "Application of Diffusion Approximations to Queueing Networks. Part I. Equilibrium Queue Distributions," *J. Assoc. Comput. Mach*., 21 (1974), 316-328.

[10] Reiser, M., and H. Kobayashi. "Accuracy of the diffusion approximation for some queuing systems." *IBM Journal of Research and development* 18.2 (1974): 110-124.

[11] Reiser, Martin, and Stephen S. Lavenberg. "Mean-value analysis of closed multichain queuing networks." *Journal of the ACM (JACM)* 27.2 (1980): 313-322.

[12] P. Schweitzer and A. Seidmann, “Optimizing processing rates for flexible manufacturingsystems.” *Managment Science*, 37 (4), (1991): 454–466.

[13] W. Whitt, “The Queueing Network Analyzer”, *Bell System Tech. J.* 62 (1983): 2779-2815.

[14] Segal, Moshe, and Ward Whitt. "A queueing network analyzer for manufacturing." *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12* (1989): 1146-1152.

[15] Boxma, Onno Johan, AHG Rinnooy Kan, and Mario van Vliet. "Machine allocation problems in manufacturing networks." *European Journal of Operational Research* 45.1 (1990): 47-54.

[16] W. Whitt, “Towards Better Multi-class Parametric-decomposition Approximation for Open Queueing Networks”, *Annals of Operations Research* 48 (1994) 221-248.

[17] Kouvelis, Panos, Chester Chambers, and Dennis Z. Yu. "Manufacturing operations manuscripts published in the first 52 issues of POM: Review, trends, and opportunities." *Production and Operations Management* 14.4 (2005): 450-467.

[18] Spinellis, Diomidis, Chrissoleon Papadopoulos, and J. MacGregor Smith. "Large production line optimization using simulated annealing." *International Journal of Production Research* 38.3 (2000): 509-541.

[19] Tempelmeier, Horst. "Practical considerations in the optimization of flow production systems." *International Journal of Production Research* 41.1 (2003): 149-170.

[20] Banks, J. and Carson,J. S., 1984, *Discrete-Event System Simulation* (Englewood Cliffs, NJ: Prentice Hall)

[21] Johnson, Rachel T., John W. Fowler, and Gerald T. Mackulak. "A discrete event

simulation model simplification technique." *Simulation Conference, 2005 Proceedings of the Winter*. IEEE (2005): 5-8

[22] Huang, Edward, et al. "Multi-fidelity Model Integration for Engineering Design." *Proceedings of Computer Science* 44 (2015): 336-344.

[23] Brooks, Roger J., and Andrew M. Tobias. "Simplification in the simulation of manufacturing systems." *International Journal of Production Research* 38.5 (2000): 1009-1027.

[24] Kao, Yu-Ting, Chun-Ming Chang, and Shi-Chung Chang. "Do we still need daily production target setting in fully automated fabs?." *e-Manufacturing and Design Collaboration Symposium (eMDC), 2014*. IEEE, (2014): 1-4

[25] Kao, Yu-Ting, Shi-Chung Chang, and Chun-Ming Chang. "Target setting with consideration of target-induced operation variability for performance improvement of semiconductor fabrication." *Automation Science and Engineering (CASE), 2014 IEEE International Conference on*. IEEE (2014): 774-779

[26] Hu, Ming-Der, and Shi-Chung Chang. "Translating overall production goals into distributed flow control parameters for semiconductor manufacturing." *Journal of manufacturing systems* 22.1 (2003): 46-63.

[27] Danping, Lin, and Carman KM Lee. "A review of the research methodology for

the re-entrant scheduling problem." *International Journal of Production Research* 49.8 (2011): 2221-2242.

[28] Park, Youngshin, Sooyoung Kim, and Chi-Hyuck Jun. "Performance analysis of re-entrant flow shop with single-job and batch machines using mean value analysis." *Production Planning & Control* 11.6 (2000): 537-546.

[29] Whitt, Ward. "The Marshall and Stoyan bounds for IMRL/G/1 queues are tight." *Operations Research Letters* 1.6 (1982): 209-213.

[30] Whitt, Ward. "Approximations for the GI/G/m queue." *Production and Operations Management* 2.2 (1993): 114-161.

[31] Shanthikumar, J. George, and David D. Yao. "On server allocation in multiple center manufacturing systems." *Operations Research* 36.2 (1988): 333-342.

[32] Dallery, Yves, and Kathryn E. Stecke. "On the optimal allocation of servers and workloads in closed queueing networks." *Operations Research* 38.4 (1990): 694-703.

[33] Bitran, Gabriel R., and Devanath Tirupati. "Tradeoff curves, targeting and balancing in manufacturing queueing networks." *Operations Research* 37.4 (1989): 547-564.

[34] Frenk, Hans, et al. "Improved algorithms for machine allocation in manufacturing systems." *Operations Research* 42.3 (1994): 523-530.

[35] Bispo, Carlos F., and Sridhar Tayur. "Managing simple re-entrant flow lines: theoretical foundation and experimental results." *IIE transactions* 33.8 (2001): 609-623.

[36] Sadre, Ramin, and Boudewijn R. Haverkort. "Decomposition-based queueing network analysis with FiFiQueues." *Queueing Networks*. Springer US, (2011): 643-699.

[37] Bolch, Gunter, et al. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2006.

[38] Huang, Edward, et al. "Multi-fidelity Model Integration for Engineering Design." *Procedia Computer Science* 44 (2015): 336-344.

[39] J. F. C. Kingman, "Inequalities in the theory of queues." *Journal of the Royal Statistical Society. Series B (Methodological)* (1970): 102-110.

[40] Marshall, Kneale T. "Some inequalities in queuing." *Operations Research* 16.3 (1968): 651-668.

[41] Brumelle, Shelby L. "Some inequalities for parallel-server queues." *Operations Research* 19.2 (1971): 402-413.

[42] Ho, Yu-Chi, R_S Sreenivas, and P. Vakili. "Ordinal optimization of DEDS." *Discrete event dynamic systems* 2.1 (1992): 61-88.

[43] Ho, Yu-Chi. "An explanation of ordinal optimization: soft computing for hard problems." *Information Sciences* 113.3 (1999): 169-192.

[44] Xu, Jie, et al. "Efficient multi-fidelity simulation optimization." *Proceedings of the 2014 Winter Simulation Conference*. IEEE Press (2014): 3940-3951.

[45] Chang, Shi-Chug. "Demand-driven, iterative capacity allocation and cycle time estimation for re-entrant lines." *Decision and Control, 1999. Proceedings of the 38th IEEE Conference on*. Vol. 3. IEEE, (1999): 2270-2275

[46] Kendall, Maurice George. "Rank correlation methods." (1948).