Performance Analysis of Batching in Manufacturing Systems

Sameer Goyal

A thesis
submitted to the Faculty of Graduate Studies
in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy

Department of Mechanical & Industrial Engineering
University of Manitoba
Winnipeg, Manitoba

ISBN 0-612-13148-3

Name _____

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

INDUSTRIAL

SUBJECT TERM

0546

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences

Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

PERFORMANCE ANALYSIS OF BATCHING IN MANUFACTURING SYSTEMS

BY

SAMEER GOYAL

A Thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Contents

Abstract

Manufacturing has become a key issue in the present environment. There has been emerging interest in the use of manufacturing and manufacturing strategy as a competitive advantage. A critical challenge for production managers is to improve productivity in today's global economy. For many firms, the market's need for product differentiation and shorter product life cycles has resulted in diversified production requirements.

One way of achieving the advantages of diversified, low volume production is through advanced manufacturing systems. The installation of these manufacturing systems requires high initial investment. For this reason, many manufacturing facilities are variations of the traditional job shop. The flexibility of a job shop results in a complicated workflow with work queuing up at workcenters and causing congestion. Scheduling has been the conventional method of alleviating operational problems in a job shop. Most manufacturing occurs in batches and scheduling by itself does not present an accurate view of the job shop. An essential determinant of job shop performance is the batching policy employed.

This research is an attempt toward assessing the performance of different batching decisions under various configurations of a manufacturing facility by exploring the interaction of batching policy with variables such as flow time and resource utilization. The research is based on an actual manufacturing facility that is populated with job shops and flow lines. We build analytical models for one- and two-server systems that process work in batches. More complex systems are examined using simulation for various machine configurations under the influence of a number of variables.

The results show that the batch size, the nature of the first node in the system (job shop or flow line), and the scheduling rule are important factors in determining the performance of a facility. There is a trade-off among the various performance

measures under most circumstances. However, some scenarios illustrate that a combination of the independent variables can be effective in containing these tradeoffs. The simulation results are analyzed in detail. We also generalize the results for a generic batch manufacturing system and provide managerial implications. Finally, we provide directions for further research.

Acknowledgment

This thesis represents a long journey in my academic career. In retrospect, it has been an enriching and worthwhile experience. Perhaps the greatest benefit was that of personal development.

I have had the good fortune of working with Dr A S Alfa, my advisor. Dr Alfa has provided insight, expertise, and support throughout this study. More important, he has been a good friend and I hope it has been a learning experience for him as it has been for me. I also thank the other members of the committee, Dr S K Bhatt, Dr D Strong, and Mr O Hawaleshka for their suggestions. My thanks to the external examiner, Dr S D Bhole, for his close scrutiny of the manuscript and his useful recommendations. While the dissertation has benefited from keen examination by the thesis committee and myself, errors have a way of sneaking in, and I take responsibility for these.

I thank Dr Y P Gupta for reintroducing me to academics during the course of this study. I have also had the good wishes and support of some close friends, namely Michael Power, David Chin, Mahesh Gupta, and Ganesh Vaidyanathan. Finally, I thank Computer Services for keeping me gainfully employed during this time.

This thesis is dedicated to my parents and the rest of my family whose love and encouragement made it all possible.

Chapter 1
Introduction

Manufacturing technology has experienced remarkable change in recent years. Conventional manufacturing has relied primarily on two types of equipment. The first—dedicated machinery such as transfer lines—is best suited for mass production of a single part. The process specialization permits low unit costs, but it inhibits flexibility. The second—unautomated general purpose machine tools—is best suited for small batch production of different parts. Costs per unit tend to be high, but the flexibility of the process can accommodate design changes, demand fluctuations, and shifts in product mix. Advanced Manufacturing Systems such as Flexible Manufacturing Systems (FMS) offer a third choice—one with flexibility higher than transfer lines and unit costs lower than general purpose machine tools (Gerwin 1982). An FMS is an integrated system of centrally programmable computers linked with a set of machines that perform such operations as machining, inspection, and assembly (Buzacott 1982).

The installation of advanced manufacturing systems requires high initial investment. Consequently, many manufacturing facilities are adaptations of the traditional job shop. Job shops are complex manufacturing systems with many work centers that process items in batches. The inherent flexibility of a job shop effects a complicated workflow with batches queuing up at work centers and causing congestion. Schaffer (1981) reported that for an average batch-type production shop, a job spends only 5% of its time on the machine, whereas the remaining 95% of the time is spent in transportation and queues (Kekre & Udayabhanu 1988). These queues result from the heterogeneity of the items and cause substantial variability in the arrival and service patterns. Queuing delays also increase manufacturing lead time. This has several consequences such as high levels of WIP and safety stock, increased inventory cost, deterioration and loss of material due to increased delay between production and use,

and poor response to change in demand (Karmarkar *et al* 1985).

Scheduling is the conventional method of alleviating such problems in a job shop. However, scheduling by itself renders an incomplete view of the job shop. An essential determinant of job shop performance is the batching policy employed (Karmarkar *et al* 1985). For small batch sizes, set-up plays a dominant role in the time a batch spends waiting to be processed. As the batch size increases, the same set-up is dispersed over a larger batch, thus reducing the waiting time. A further increase in the batch size causes a larger wait as the batch takes longer to form. The effect of this wait is dominant in that the waiting time starts increasing. This increase is exponential and the system gets saturated quickly, indicating that the waiting time of a batch is convex in batch size.

This research is an attempt toward assessing the performance of batching decisions under given configurations of a manufacturing system. The research is based on a real manufacturing facility. The outcome of the study will provide suitable batching policies with regard to a given objective, such as minimization of flow time.

1.1 Motivation

This research originated as a study of a manufacturing facility in Brampton, Ontario. The facility is owned by a company that is one of the world's leading suppliers of digital telecommunications switching systems. The company designs, builds, markets, and supports a wide range of telecommunications products. The plant undertook a major renovation and reorganization to establish itself as a world class manufacturing facility. Some of the goals of the project were a flexible plant environment, an open concept factory floor, integrated systems, and a simplified material and manufacturing process. An analysis of these issues resulted in widespread improvements in the facility. Concomitantly, several problems also surfaced.

The circuit pack (CP) line at the facility is responsible for populating and testing the (integrated) circuit boards needed for the switching equipment. The line produces over 500 types of CPs/boards and is composed of 8 major stages. Customer

orders drive production at the facility. Work is released in the form of batches. There are no specific rules regarding the work release patterns or batch sizes. Batch sizes vary usually from 5 to 50 units. They are determined by evenly allocating the total requirement for the product over a five-day week. The CP line experiences numerous material shortages and large set-up times in some stages. The problems are caused mainly by unreliable suppliers who sometimes deliver defective components, and by forecast inaccuracies. Lead times of many parts are unduly long because of the remoteness of the suppliers. This causes many problems like high WIP, high safety stocks, schedule changes, and more important, lack of competitiveness.

To gain an insight into some of these problems and to suggest recommendations, a simulation of the CP line was constructed. This involved meeting with the personnel responsible for the CP line and collecting relevant data. The project spanned a period of one year. The facility is primarily a flow line with some stages configured as job shops. The purpose of the simulation was to study the effects of different batching strategies on performance measures such as flow time and resource utilization. Major parameters such as demand, batch size, and set-up time are stochastic. Further, the facility is characterized by material shortages, high WIP, large set-up times in some areas, low resource utilization, and lack of coordination between the manufacturing stages. These elements make the facility typical of many batch manufacturing systems. The results of the simulation suggested scheduling rules and batching policies that were appropriate for one/more of these factors.

1.2 Purpose

The purpose of this research is to evaluate batching decisions under different configurations of a manufacturing facility. These decisions include the batch sizes and the processing order of the batches. Assume n job types are to be processed on a machine. A *batch* is a collection of jobs of the same type. In general, there is more than one batch of the same job type. The *size* of a batch is the number of jobs it contains. A batch is assumed processed upon the completion of the last job in the

batch. A set-up is usually incurred if consecutive batches contain different job types.

While there may be similarities between batching and lot sizing, lot sizing belongs to medium range decisions in the production planning hierarchy whereas the batching problem is a component of short range decisions like scheduling and shop floor control. The interest in this study is in the batching decisions only.

The management of a batch production system can be very involved. The vast variety of the CPs populated at the CP line contributes to high variability in processing times and the arrival patterns, and frequent set-ups. Compounding the circumstances are long lead times of the components, material shortages, defective components, forecast inaccuracies caused by uncertain demand, and changes in the configuration of the facility. These factors induce high safety stocks, high WIP, schedule changes, and an inability of the firm to meet due dates.

Appropriate batching policies can alleviate these problems in various ways. Batching affects queues in the system by influencing the variability in processing times and arrival patterns. More important, it can reduce the time spent by the work in the system by (i) reducing waiting times, (ii) increasing resource utilization, and (iii) reducing set-ups, or a combination of these. Among the problems mentioned above, an important contribution of batching strategies is to enable the firm to meet due dates.

Batching is not a cure-all. Analogous to real life situations, production problems sometimes involve multiple objectives that are often conflicting. The conflict arises because an improvement in one objective can be achieved only by subordinating another. As an example, increasing resource utilization may necessitate higher production rates, which in turn may lead to higher WIP. Meeting due dates may require rush orders, which could lead to frequent set-ups and thereby reduced utilization. As batch sizes are reduced, set-ups cause the workload in the system to increase, resulting in congestion and requiring larger safety stocks to protect against production delays. It seems inconceivable to optimize simultaneously with respect to all the goals. In fact, the multi-item (multi-job type) batching problem is not solvable even for the single machine case (Dobson *et al* 1989).

Smith *et al* (1986) noted several objectives to be of importance to practitioners. These include meeting due dates, maximizing resource utilization, minimizing WIP, and maximizing production rate. Meeting due dates was cited as the most important objective because it can make a firm competitive. The conventional measure of due date realization is tardiness. The goal of meeting due dates has not been examined in the batching studies although small batches can be expected to reduce flow time.

An interesting objective is that of minimizing the flow time of batches. This objective also minimizes WIP. A reduction in flow time shortens the forecast horizon thus reducing the need for safety stock. Lower flow times also accelerate the feedback on quality problems and expedite the introduction of engineering changes and new products (Dobson *et al* 1989). Flow time minimization is achieved easily in a flow line by scheduling batches of size one but small batches require frequent set-ups and they exacerbate the complex work flow in a job shop. The interface between a job shop and a flow line remains to be examined. Experience with the CP line and the simulation supports the intuition that the job shop causes a bottleneck in the operation, which influences WIP and lead times. It would be useful to find a batching strategy that finds a compromise between the job shops and the flow lines. A suitable strategy may also reveal bottlenecks, thus exposing other areas in the line that may be disrupting the work flow.

Minimization of flow time is of little consequence when considered in isolation. Flow time is synergistic with its variance—a reduction in flow time is futile if there is high variability. This observation has important overtones for meeting due dates. A flow time with low variability is more effective in estimating due dates than a flow time with high variability. Flow time is composed of transportation, waiting, set-up, and processing times. It is known (Schaffer 1981) that jobs spend substantial time waiting in queues. Batching can reduce queues in the system by reducing the variability in processing times and arrival patterns, and by increasing resource utilization by reducing set-ups. The issue of flow time variance in the context of batching remains unexplored in the literature.

The issue of demand in conjunction with batching also needs to be examined. The effect of demand is not clear although high demand can affect the performance measures unfavorably. Often high demand entails small orders of varying product types. This implies frequent set-ups and therefore a decrease in resource utilization and an increase in flow time and WIP.

Set-up time is another issue that has received little attention in the batching literature. However, in some cases set-up time can exceed processing time. Set-up time is of seminal importance in gaging the performance of manufacturing facilities that release work in batches. Consider a machine that processes work in batches. Each batch requires a set-up. If the arrival of work is held constant and batch size is increased, the processing time increases linearly because the effect of set-up is diminishing. However, as batch size is reduced, more time is spent on set-ups. At some point the machine reaches capacity resulting in the formation of a queue. Among the lesser known benefits of reduced set-ups are improved quality control and increased capacity. Lower set-ups allow production in smaller batches. This leads to a lower number of defects because defective parts can be identified promptly. Similarly, lower set-ups result in increased capacity. Porteus (1985,1986a,b) performed a detailed analysis of set-up cost reduction.

> The benefits of reducing the set-up cost transcend the benefits identified in the EOQ model alone ... improved quality control, flexibility, and increased effective capacity tend to result (Porteus 1985).

Finally, the effect of processing time variability has been neglected. While the argument is similar to that for flow time variance, processing time variability is a factor that may be regulated. It can affect performance measures such as flow time and flow time variance. A hypothesis in the literature is that large batches reduce "variability" in the system. These beliefs may not be valid in the presence of a highly variable processing time.

While theoretical deliberation abounds in the literature, this study intends to

quantify these effects, examine their logic in the presence of batching, and assess the performance of batching decisions under a variety of configurations.

1.3 Methodology

We will first formulate analytical models for a single-server and a two-server system, that process work in batches. These models can be used to study individual nodes such as bottlenecks, in a manufacturing facility. They can also be used to take a holistic view of a group of machines or facilities. We will then focus on larger systems consisting of facilities that can be characterized as flow lines or job shops. The approach is to acquire an understanding of the interface between a flow line and a job shop. This will be done by initially simulating small systems consisting of job shop and flow line nodes and then continuing with larger systems as the interactions between the dependent and independent variables are understood. The significance of these interactions will be tested by using experimental design. The experience gained will be used to make policy recommendations for generic batch manufacturing systems.

Simulation is an evaluative technique. It provides a performance estimate for a given set of parameters. To find an optimal set, simulation should be interfaced with a generative procedure, ie, one that generates alternative sets of parameter settings. Factorial design is such a procedure. The nature of the problem suggests the use of factorial design with a simulation model to yield a relationship between the independent and dependent variables. The simulation will produce a value for the dependent variables for a given combination of the independent variables. The independent variables will be batch size, scheduling rule, set-up time, demand, and processing time variability. The dependent variables will be resource utilization, flow time, flow time coefficient of variance, and the percentage of tardy batches.

Simulation can capture reality in detail. However, attempts at using a simulation and statistical design to evaluate alternative system configurations can be futile unless the simulation is validated. Queuing network approximations provide an effec-

tive means of modeling such facilities. Whitt (1983a) developed the Queuing Network Analyzer (QNA) to calculate approximate congestion measures for open networks of queues. The model can treat non-Markov networks, ie, the arrival processes are not Poisson and the service-time distributions are not Exponential. QNA can handle large networks quickly because the calculations required are minimal. Congestion measures for the network are obtained by assuming that the nodes in the network are stochastically independent.

The simulation results will be validated by queuing network approximations. Factorial design will be used to test hypotheses and interactions among the independent variables. Regression analysis can then be used to arrive at approximations of the system in the form of relationships between dependent and independent variables.

1.4 Contribution

This research is a step toward analyzing batching decisions for realistic systems, ie, ones with flow lines, job shops, multiple machines, and multiple job types. The results will provide practical managerial implications and insight for similar batch manufacturing systems. The analytical models provide exact results for analyzing single- and two-server systems. The simulation models an actual manufacturing facility. Easily adaptable to other facilities, the model provides an evaluative technique for studying similar manufacturing problems. In addition, the queuing network approximations can be used for modeling the facility and for validating the results of the simulation.

The following are the contributions of this research.

- Formulating analytical models for single- and two-server systems that process work in batches.
- Examining the performance of batching decisions under various facility configurations.
- Examining the physical interface between a flow line and a job shop for improving overall performance.

• Finding empirical evidence for the hypotheses that have been proposed in the literature.

• Making policy recommendations for generic batch manufacturing systems by building on knowledge acquired by analyzing diverse systems.

1.5 Practical Implications

The practical implications of this research are several. The models can produce numerical results to solve specific problems. More important, the research provides a conceptual framework and intuitive insight into manufacturing problems.

Research in manufacturing usually centers on facilities that are comprised primarily of job shops, flow lines, or cells with dedicated machines. Recently, there has been deliberation on flexible manufacturing. However, most manufacturing is a combination of these facilities. The basic premise in this research is that at a macro level a facility can be classified as a job shop or a flow line. Moreover, a series of flow lines can be grouped into an extended flow line and adjacent job shops can be grouped into one job shop. Thus, alternative system configurations can be evaluated by studying the simplified arrangements. This information can be very beneficial in the design and planning stages.

> Conventional lot sizing models use set-up costs as a device to enforce capacity constraints—they are effectively opportunity costs and they seldom represent actual cash consequences associated with set-ups (Karmarkar *et al* 1985).

Set-up costs are fixed one-time costs that are incurred whenever an order is processed. In a production setting these costs result from phenomena such as equipment changeover and transport of materials. These costs cause lot sizes to be higher than they would be otherwise. They necessitate less frequent set-ups and thus increase the lot sizes. The association of these to capacity results from the notion that set-ups may be reduced by investing in technology. This fixed investment is recovered by capital-

izing on the opportunity cost of capital. In other words, set-ups represent the opportunity cost of lost production time. This cost is incurred when production time is a binding constraint.

The intent of this dialog is that in manufacturing facilities with job shops, resource utilization cannot reach one because of queues and therefore the capacity constraint is not literally binding. The cost of set-ups in such a facility is not the opportunity cost of production time but the waiting cost that is incurred by jobs in queues. The models in this research capture this set-up directly. Set-ups consume time and queuing models deal with that time precisely instead of representing them with surrogate costs.

This research also provides a more accurate view of a bottleneck. The conventional view of capacity leads to the concept of a bottleneck machine that determines total output. This machine is the first to saturate as output is increased. However, output is limited by queues and not by a capacity limit. Thus, this concept is of limited use as output cannot be raised to the capacity of the machine. Bottleneck machines are the ones that experience substantial queues, which limit total output. Further, the queues and hence the bottlenecks can shift with a change in the product mix and the batching policy so that the conventional view of a bottleneck is improper. It is often the bottleneck machines that determine the batching policy.

The models in this research can be applied to achieve various ends, eg, scheduling, capacity planning, performance evaluation, and engineering analysis. Scheduling can help by ordering the jobs to save set-ups, especially on bottleneck machines. The models can also help by setting the batch sizes. As discussed earlier, batch sizes that are too small can quickly saturate a machine. A similar argument may be made for overly large batch sizes. The knowledge of lead time for different parts can allow more precise MRP calculations and improve the order release times.

At the planning level, the models provide an insight into the performance of the facility under differing conditions of product mix, demand, and shift policies. Long term decisions such as capacity planning can also be evaluated by moderating parameters such as characteristics of individual machines. More important, the sensi-

tivity of performance measures to variables such as set-up time can be measured easily. If certain set-up time reductions are possible by investing in technology, their impact on performance can be estimated (Karmarkar *et al* 1985).

Thus, the models in this study can help the decision maker in understanding how a performance parameter can be used to select an appropriate decision. The effect of a system configuration on the performance of the facility can significantly influence the design of the system before the production/process planning stages are implemented. This knowledge can be useful in developing a manufacturing strategy for a new product, changing the product mix, and investment in new technology. In short, the management can position the operations strategy of the manufacturing system to complement the manufacturing strategy of the firm.

1.6 Summary

Advanced manufacturing systems have emerged as an offspring of technological innovation in computer and numerical control techniques. These systems manufacture parts with material handling functions, machine operations, and machine tools under computer control (Herald & Nof 1978). The steep initial investment of these technologies has caused most North American manufacturing to adopt traditional methods such as job shops, which process work in batches. A job shop is a flexible operation as it can manufacture many different items. The high part variety results in frequent machine set-ups and a complicated work flow in the shop. The result is that batches spend much time waiting in queues, which causes congestion. Batching policies can significantly affect the performance of a manufacturing facility.

This research started as a simulation that was written for studying an actual manufacturing facility. Familiar problems like long lead times, unstable demand, shortages, and high set-up times were preventing the company from meeting due dates, and causing high WIP and low resource utilization. The main purpose of the simulation was to find a processing order for the batches and batching policies that would alleviate some of the problems.

We propose to assess the performance of different batching policies under given configurations of a manufacturing facility by using experimental design to measure the effect of batch size, demand, processing time variability, scheduling rule, and set-up time (independent variables) on flow time, flow time coefficient of variance, resource utilization, and the percentage of tardy batches (dependent variables). The results will provide suitable batching policies and managerial implications by analyzing the interactions among the variables.

Although simulation can capture reality in detail, inadequate validation of a simulation model can render the results insignificant. Queuing network approximations provide an efficient means for modeling the systems being considered in this study. These approximations can also be used for validating the results provided by the simulation. Upon validation, the simulation results will be used with factorial design to assess the significance of individual factors. These factors will be used to form regression relationships between the dependent and independent variables.

1.7 Organization

The next chapter contains a review of related literature. Analytical models are developed in chapter 3. The research methodology and experimental design are outlined in chapter 4. Chapter 5 contains an analysis of the results. The research concludes in chapter 6.

Chapter 2
Literature Review

Manufacturing technology has been a well discussed issue in recent years. While advanced technologies have attempted to bridge the gap between the conventional job shop and the flow line, the initial investment remains high. For this reason, many manufacturing facilities are adaptations of the traditional job shop. Job shops usually include many work centers that process items in batches. The need to divide items in batches stems from the machine set-up required for each batch. It is estimated that batch production accounts for 60-80% of all manufacturing activities (Chevalier 1986).

Assume n job types are to be processed on a machine. A *batch* is a set of similar jobs and the *batch size* is the number of jobs in the batch. A job is considered processed only when its entire batch is processed completely (Some studies also examine *batch splitting* where part of a batch can be split and delivered. We will not survey that case). There is usually a set-up between consecutive batches, the length of which depends on the two job types. This set-up is needed to configure the machine for the next batch. The general batching problem is to find the batch size and processing order of the batches to optimize a performance measure.

Research in the batching problem is fairly recent. Karmarkar (1987) first examined the implications of batching on manufacturing issues like lead time and WIP. Uzsoy et al (1992, 1994) provided a comprehensive review of the production planning and scheduling models in the semiconductor industry. Bruno and Downey (1978), Monma and Potts (1989), and Potts and Van Wassenhove (1992) reviewed complexity issues. In this paper, we attempt to classify the literature in batching, analyze the proposed models, and provide research directions. We will classify the models based on the number of processing stages—single, dual, and multiple.

2.1 Single-stage Models

In single-stage models, a single job type (product) or multiple job types are processed in batches on a single machine or identical parallel machines. We will classify this section further into single machine models based on (i) mathematical programming, (ii) queuing theory, (iii) case studies, (iv) parallel machine models, and (v) fabrication and assembly models. Depending on model formulation, there may be set-up before every batch is processed. This set-up may also be sequence dependent.

2.1.1 Single machine models based on mathematical programming

The models in this section consider single or multiple part types in a static deterministic environment, ie, the number of jobs and their ready times are known and fixed. Set-up time is also known. Jobs may have due dates and demand may be satisfied by one or more batches with no restriction on the batch sizes. The first three studies optimize some measure of flow time. The authors arrive at the same result for the optimal number of batches for the single product problem.

Santos and Magazine (1985) studied the problem of determining the batch sizes and the number of batches or set-ups required to produce a given set of lots to minimize completion times. Four definitions of flow time are considered:

(i) *Total flow time: item availability*: The completion of processing of the item (in a batch) determines when it is available (batch splitting).

(ii) *Total flow time: batch availability*: The completion of processing of the batch determines when the item is available.

Definitions (iii) and (iv) are similar to (i) and (ii) except that the order for a batch is not released until the time of set-up for that batch. These are termed *total tight flow time*. There are n product types to be scheduled on a machine. Let d_i be the total

batch size required of product i and p_i its processing time per unit. For each batch of product i, a set-up time S_i is incurred. Further, k_i denotes the number of batches and $b_i(j)$ the size of the jth batch of product i, ie, $\sum_{j=1}^{k_i} b_i(j) = d_i$. In (ii), the completion time of jobs is determined by the completion time of the batches. If C_{ij} is the completion time of the jth batch of product i, then $Z = \sum_{i=1}^{n} \sum_{j=1}^{k_i} b_i(j) C_{ij}$. Even the single product problem is complex because the batch sizes can be varied to optimize set-ups. The authors thus present a model for the single product type.

$$\min \sum_{j=1}^{k} b_j \sum_{l=1}^{j} S + b_l p$$

$$\sum_{j=1}^{k} b_j = d$$

$$k,\ b_j > 0,\ j=1,\ldots,k$$

The major result for the single product problem is that the optimum number of batches,

$$k^* = \sqrt{\frac{1}{4} + \frac{2dp}{S}} - \frac{1}{2},$$

and that the optimal batch sizes can be found from

$$\frac{d}{k^*} + \left[\frac{k^* + 1}{2} - j \right] \frac{S}{p},\ j=1,\ldots,k^* .$$

The batch sizes are decreasing, which indicates that large batches are processed first. This increases machine utilization because of less frequent set-ups.

The authors propose the models as a link between lot sizing and machine scheduling by determining the number of batches and the batch sizes to produce the items. Because the models consider the single operation case only, the authors suggest its applicability to the scheduling of a single bottleneck machine. They further mention that scheduling of a bottleneck machine may allow more productive management

of the job shop. Job shops usually employ many machines. Efficient scheduling of a bottleneck machine in a multi-machine job shop may not imply efficient management of the job shop. Bottlenecks are usually not machine specific—they can shift with the items being produced and the batching policy being employed.

Naddef and Santos (1988) additionally examined the multi-product batching problem for one machine. For the single product problem they used the same objective function as Santos and Magazine (1985). Given d jobs with processing time p and set-up time s for the batches, the problem is to find the optimum number of batches k^* and the batch sizes. Let b_i be the size of the ith batch and k be the number of batches. A solution to the d-job batching problem, $B_d = (b_1,\ldots,b_k)$. The objective is to minimize the sum of completion times.

The authors present a theorem that allows one to find an optimum solution to the d-job problem from a solution to the $d-1$ job problem. This theorem leads to a one-pass (greedy) heuristic that always yields an optimum solution to the d-job batching problem. The heuristic considers each job and allocates it among the non-empty batches or the next empty one. This heuristic is not polynomial in the size of the input. A computationally efficient formulation is also presented.

For the multiproduct problem, n job types are to be processed on a machine. For each job type i there are d_i jobs, each with processing time p_i. The set-up time depends on the type of jobs in the batch. This suggests that batches of like items be consecutive to save on set-up. By considering job types individually, the authors propose the one-pass heuristic for the multi-product problem. Each batch is considered as one job with processing time $s_i+d_ip_i$ and weight d_i. These jobs can then be sequenced in the well known SPT ratio $s_i+d_ip_i$. Limited computational testing of the heuristic showed favorable results when the set-up time per product is at least twice its processing time, on average.

The authors examine a single machine and a single job type. For the multi-product problem, the authors cite the well known SPT heuristic and also provide some computational experience with this heuristic. The authors also propose that the heuris-

tic is likely to give better results for a large number of job types than for a small number. However, the simulations consider three to six job types only and any improvement in the heuristic is not evident from the results. The multiproduct batching heuristic ignores the downstream batches and treats each product as a single product problem. This policy is fairly myopic because the downstream batches can have an effect on the optimal batching. Set-up times are not sequence dependent and there is no reference to the impact of the batching policies on WIP and cost.

Dobson *et al* (1987) extended the above results by providing a bound for the multiple product problem. They studied the problem of optimal batching decisions in a closed job shop, which they define as a shop that builds to a schedule derived from a downstream department instead of external demand. The objective is to minimize the flow time of parts through the shop. Two types of flow time are mentioned. In *item flow* (batch splitting), a part can be delivered after being processed. In *batch flow*, a part waits until the remainder of its batch is processed. The authors present IP formulations of both problems for the single product case and heuristics for the multi-product case.

The single product problem is not simple to solve because both the *number* and the *composition* of batches must be resolved. Thus, the authors present a model for a single product type and constant set-up time. This model is extended to include variable set-up times. The multiple product batch flow problem is substantially more complex than the single product problem. The authors resort to heuristics that build upon the results above. Both heuristics produce sub-optimal results due to their myopic nature, ie, they ignore the size of downstream batches when determining the size of the initial batches (larger batches should be sequenced first). To improve these schedules, the authors look at reallocating the quantities among existing batches by reformulating the variable set-up time model. Under restrictive assumptions, the authors derive an arduous expression for the optimal batch sizes. To this effect, two more heuristics are proposed. A lower bound is also presented for assessing the solutions provided by the heuristics.

Extensive computational testing of the heuristics revealed that the correct batch size depends not only on the characteristics of the parts that form the batch, but also on the total work waiting. The average percentage by which the heuristic solutions exceed the lower bound varies from .6-24.1%. When the demand per part type is small, the solution tends to allocate a batch to each part. Better results are obtained for cases with high demand because of the aggregation of batches of similar parts.

The main result of this study is that the optimal policy to minimize flow times is to sequence the batches in order of decreasing size. It is possible that increasing all batch sizes will reduce the variability in the system. However, increasing the batch size of selected items may increase the variability of the arrival process. A limitation of the results is that they apply to the single machine case. Set-up times are also sequence independent.

The above studies optimize some measure of lead time. Zdrzałka (1991) presented heuristics for the single machine scheduling problem with the objective of minimizing makespan. Jobs are partitioned into batches and a unit set-up is incurred when consecutive batches are of different job types. All jobs are available at time zero and the machine processes one job at a time, without preemption. The processing times and delivery dates are known. The three heuristics presented are of polynomial complexity with the worst case performance bounds on makespan of $n-1$, 2, and 5/3, where n is the number of jobs. Zdrzałka (1995) later provided two more heuristics with worst case performance bounds of 3/2. Computational results show that the heuristics are highly sensitive to the parameters like set-up and delivery times.

Unal and Kiran (1992) examined a single machine scheduling problem where a number of part types are processed. The objective is to minimize tardiness. The Batch Sequencing Problem (BSP) is defined as finding a sequence of batches of part types so that the production requirement of all parts is met. The authors show that the BSP can be transformed into an equivalent Feasibility Problem (FP). The FP is finding a feasible sequence given the jobs and their processing times, due dates, and set-up times.

Therefore, BSP has a solution only if FP is feasible and a solution to FP is a solution to BSP. It is shown that tardiness is minimized by ordering the jobs according to their due dates. Thus, given a batching structure, the batch sequence is easy to find. The FP is then reduced to finding a structure that would yield a feasible sequence. The authors also mention that in a feasible sequence, jobs of the same part type should be processed according to due dates.

The authors present a heuristic to construct a sequence that minimizes maximum tardiness. The argument is then used in an exact algorithm that sequences jobs according to due dates. The jobs are grouped into batches and the batch with the last due date is scheduled as late as possible. A feasible sequence minimizes the number of set-ups. Extensive numerical testing of the algorithm showed satisfactory results. The study considers only the single machine case and part dependent set-up times. Set-up times could also be caused by machines, tool changeover, and transportation.

2.1.2 Single machine models based on queuing theory

Karmarkar (1987) first explored the interactions among batching, manufacturing lead times, and WIP by using standard queuing models that investigate congestion phenomena and their effects on waiting times. The models are more suited to manufacturing facilities such as closed job shops where the WIP remains fairly constant. The study considers a single machine in such a job shop, which is modeled as an M/M/1 queue. Items arriving at the machine are alike and have the same batch size. In effect, the machine processes batches of a single item. Let

D = demand (units/year)

P = processing rate of machine (units/year)

Q = batch size

$\lambda = D/Q$ = average arrival rate of batches (batches/year)

τ = set-up time per batch (year)

$\bar{x} = \tau + Q/P$ = processing time per batch (year)

$\mu = 1/\bar{x} = P/(P\tau+Q)$ = processing rate of machine (batches/year).

The M/M/1 assumption precludes the batching policy from causing variability in the arrival or the service process. Standard results from the M/M/1 queue give $\rho = \lambda\bar{x} = D/p+D\tau/Q$. This shows the dependence of ρ on the batch size. The condition $\rho<1$ gives an upper bound on Q as $D\tau/(1-D/P)$. The expression for the average time in system (Karmarkar *et al* 1985b) shows that as batch size (Q) decreases, average time in system (T) grows rapidly as ρ approaches 1. For large values of Q, T becomes approximately linear in Q. The author provides an asymptotic lower bound for T and shows that average waiting time is system is convex in batch size. It is shown that similar results hold for the M/D/1 case.

The formulation assumes only one type of item. The multi-item scenario is considerably more complex. Altering the lot size of an item can cause queuing delays for all other items. This can also affect the lead times of the items. In the multi-item model, the author considers a single stage facility with arrival of batches according to a Poisson process. The time to process a batch varies according to the item. Batches are processed in the order they arrive. The facility is modeled as an M/G/1 queue with the batch processing times specifying the service time distribution. It is shown that items should be batched so that batch processing times are uniform.

The study considers the single product case, which raises the question of the need for batching and set-up. Few job shops are single machine work centers. The models do not take into consideration final product inventories and backorders.

Kekre (1987) examined the impact of product mix on the performance of a manufacturing cell. The model is similar to the above but considers savings in set-up if consecutive batches arriving at the facility are of the same type. The context is a closed job shop with a single work center that processes items in batches. In a closed job shop, the amount of work remains constant and parts are not made to-order. The author studies the effect of increased product mix and a "look-ahead" sequencing rule

on the queuing delay. The cell is modeled as an M/M/1 queue and the results are validated by simulation.

The impact of increased product mix is analyzed by the saving in set-up if consecutive batches are the same type. If λ_i denotes the arrival rate of a batch of item i then $\lambda=\Sigma\lambda_i$ is the total arrival rate. Thus the probability that an arriving batch is of the same type as the previous batch is λ_i/λ. This gives the probability that a set-up is not required for the arriving batch. Similarly, $1-\lambda_i/\lambda$ is the probability that a set-up is needed. The author shows that this policy reduces the wait as compared to a model that does not take set-up saving into account. However, if the product mix is too varied then λ_i/λ is too small to make a significant impact. For the "similar" item case the author shows that queuing delay increases with product mix even if the load on the facility is kept constant. *Similar* items are defined as having identical demand, set-up, and processing time requirements. It is not clear how product mix can be varied using this definition. The increased delay is likely due to more frequent set-ups instead.

The look-ahead sequencing rule arranges the queue so that the batches of different items that are waiting for service are consecutive. This eliminates the need for frequent set-ups. In effect, the item that is being processed currently is given priority. The probability of a set-up is computed by conditionally examining the system with n batches present. A set-up for an arriving batch is needed if all batches in the system are different or if the system is empty and the last batch processed was different. The heuristic is validated by simulation and results show that savings in set-up are small.

The models in this study consider the single machine case and for the most part the single item case. The distinction between *similar* items and identical items is not clear. The facility is assumed to be a closed job shop but in most job shops the arrival of work is dynamic and demand is difficult to estimate.

Karmarkar *et al* (1992) extended the results of Karmarkar (1987) to a multi-item environment and provided qualitative implications. The facility is modeled as an

M/G/1 queue with the objective of choosing batch sizes to minimize queuing delay. The authors develop bounds on the queuing time and optimal batch sizes. These results are used to formulate batch sizing heuristics. The approximations and bounds are used to discuss qualitative implications of batch sizes on queuing delays. The first heuristic is based on the upper bound of the queuing delay in an M/G/1 queue. The assumption is that set-up times are approximately equal. The second heuristic assumes that utilization levels are high. The third heuristic uses the lower bound as an approximation. The expression for the optimal batch size in the second heuristic is similar to the one in Dobson *et al* (1987). Computational testing of the heuristics shows good results.

The results suggest that batch sizes vary directly with set-up times; batch sizes increase at an increasing rate with utilization; batch sizes for all items should have the same run time/set-up time ratio; and this ratio depends on the total utilization of the facility. Further, this ratio should be between 2 and 20. The authors also propose that batching to minimize average wait also tends to minimize the variability in waiting time. The model assumes that there are no set-up costs. Also, the 2-20 heuristic is seldom true in the manufacture of small discrete parts like circuit board assembly where the set-up time frequently exceeds processing time.

2.1.3 Case studies

Seidmann *et al* (1985) examined the relationship between batch sizes and lead times in the context of a unitary manufacturing cell (UMC). A UMC contains several flexible work stations served by a materials handling robot. The cell produces one product at a time. Produced parts are examined in the cell and reworked if necessary. As opposed to single machine sequencing, the UMC operates under continuous load with stochastic processing times and no set-up. The objective of the study is to develop a predictive model for describing the production capacity of the cell. The authors consider batch production and well as "interleaved" production of several products.

Suppose the cell has to manufacture a batch of B identical parts. Since the cell processes one part at a time, the manufacturing times for all parts are independently

and identically distributed. Let T_M and T_R denote the times for processing one part at *manufacturing* (M) and *reworking* (R), and $T_{\bar{M}}$ and $T_{\bar{R}}$ denote the corresponding times for the batch. The total batch time is $\theta = T_{\bar{M}} + T_{\bar{R}}$. Let $h_M(t), h_R(t), h_{\bar{M}}(t), h_{\bar{R}}(t)$, and $h_\theta(t)$ denote the appropriate density functions, where the first two are given and the others are desired. The mean and variance are denoted by μ_M, σ_M, and so on. The authors provide the various means and variances, and the probability density functions for $h_{\bar{M}}(t), h_{\bar{R}}(t)$, and $h_\theta(t)$. They further develop expressions for the coefficients of variation of N and θ and show that large batches display small process variability.

The results allow the prediction of mean and variance of variables such as number of recycles and time spent by a batch in a facility. The effect of a change in batch sizes and rework rate on the operation of the cell can also be explored. The model is valid only under some very restrictive assumptions. Set-up costs are assumed zero and preemptions are not allowed. This implies that there are no machine breakdowns. The authors also do not refer to the issue of sequencing the batches.

Lee *et al* (1993) proposed a batching and sequencing algorithm for minimizing set-up on NC punch presses used in the production of sheet metal components. Similar to the studies in §2.1, the algorithm works in two stages. In the first stage, the algorithm attempts to partition the set of products into the least possible number of batches by considering the tool magazine constraints. The second stage determines a sequence for the batches by using the nearest neighbor heuristic. Multiple solutions to the problem are generated by considering a different batch as the starting batch each time. Given the set of sequences, the one that minimizes the total tool changeover time is selected.

The problem is equivalent to solving the TSP where the objective is to tour the given cities in the shortest distance with no city visited twice. The similarity between the two problems becomes apparent when the distances are replaced by tool changeover times. Thus, given a starting batch, the batch with the minimum changeover time is sequenced next. The difference is that in TSP the tour ends where it begins, while in this problem it ends at the last city. The algorithm assumes that

products require at most one tool magazine. Demand is deterministic and raw material is available at the beginning of the planning horizon.

Chua *et al* (1993) examined the batching problem for a repair shop with limited spares and finite capacity, with the objective of minimizing the backorders. They suggested several batching policies based on the works of Karmarkar *et al* and Dobson *et al* and evaluated them using a simulation of a hypothetical repair shop. On arrival a failed unit is diagnosed and disassembled. Assuming the failure of the unit is caused by a single part, the failed part is removed from the unit and sent for repair. In the meantime, the unit may wait for the failed part to be repaired or a spare part may be supplied if available. If a spare is supplied, then the failed part enters the spare parts inventory upon repair. The amount of capital limits the availability of spares.

The authors formulate a mathematical model of the single product problem whose failure is due to one of the parts. The batching problem is then the determination of batch sizes for the *parts* given the number of spares for each part to minimize the average time in system for the *units*. Because of the intractability of the mathematical model, the authors present six heuristics that are based on the literature in batching, manufacturing, and scheduling.

For the simulation experiments, the authors assume that the unit contains five parts and failure is caused by one of these parts. Besides the batching policy, the factors considered are arrival rate of failed units, ratio of set-up to run time, and initial spares inventory budget. The performance measure is average time in system of units. The results of the experiments indicate the superiority of two heuristics: (i) batch parts of the same type that are waiting for repair and select the one with the smallest weighted batch processing time (WBPT), and (ii) batch parts of the same type and process the batch with the smallest WBPT weighted by the current number of spares (SP-WBPT). Under these two policies, batches are formed as soon as the repair center becomes available. The other four policies can render the repair center idle as batches form.

The study considers a single type of repairable unit and a simple product structure—the repairable unit consists of five parts. WBPT and SP-WBPT minimize time in system by forming smaller batches. This should also increase the frequency of set-ups, especially with a large number of parts, the impact of which is not apparent in the study. The authors do not refer to the issues of normality and independence of the simulation output or report the number of replications of the simulation experiments.

2.1.4 Parallel machine models

Dobson *et al* (1989) extended their single machine models to examine the batching of a single product on parallel heterogeneous machines with the objective of flow time minimization. With multiple machines there is also the issue of work allocation among the machines. In absence of set-ups or if set-ups are equal across the machines, each machine will be allocated work proportional to its processing rate. In the presence of set-ups, it is possible that some machines may not be allocated work at all and the batching policies may be machine specific.

The problem is to allocate work to machines and set batch sizes to minimize the total flow time for processing total work D through a machine center with m heterogeneous machines, ie, choosing the allocations D_i, $i=1,\ldots,n$ and dividing them into batches q_{ij}, $j=1,\ldots,n_i$, $i=1,\ldots,m$, to minimize total flow time. The flow time minimization problem can be written as:

$$\min_{D_i \geq 0} \min_{\substack{q_{ij} \geq 0 \\ \{n_i\}_{i=1}^{m}}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\sum_{k=1}^{j} s_i + q_{ik}/r_i \right) q_{ij}$$

$$\sum_{j=1}^{n_i} q_{ij} = D_i$$

$$\sum_{1}^{m} D_i = D \ .$$

where

m = number of machines

D = total amount of work

s_i = set-up time on machine i

r_i = processing rate of machine i.

D_i = amount of work allocated to machine i

n_i = number of batches to be run on machine i

q_{ij} = size of batch j on machine i.

This problem can be split into m single machine problems, the solution of which is provided in Dobson *et al* (1987). It is further shown that with equal set-up times the number of batches is equal for all machines.

The authors consider the processing of batches of one product on parallel machines. This raises the question of the need for batching. It is assumed that the total amount of work is divisible continuously, which may not be true especially if the machines are heterogeneous. The authors do not provide examples or computational experience with the heuristics developed.

Tang (1990) presented a scheduling model of multiple products on identical parallel machines. This model is more general than the above study, which considers the single item case only. Part types are partitioned into families. On each machine a significant set-up is incurred when the machine changes from processing one part type to a part type that belongs to a different family. In addition, a minor set-up is incurred between part types irrespective of the part families. The processing time of a part is small in comparison with the set-up times. The performance measure is the makespan.

Parts of a type are grouped into a batch, which can be divided into smaller batches (at the expense of minor set-ups) to form a schedule. The order in which the families and part types within a family are processed on a machine is independent of the completion time. Thus, the scheduling problem is to determine the batch size of each part to be processed on each machine. Since this problem is NP-complete, the

author presents a heuristic approach to find a near optimal solution.

Similar to Dobson *et al* (1989), the heuristic consists of two levels—aggregated and disaggregated. Both levels employ modified versions of the Multifit method (Coffman *et al* 1978) and the FFD rule. The output of the aggregated level is the proportion of machine time allocated to each family initially. The disaggregated level uses this information to determine the batch size of each part type to be processed on each machine. Using an IP formulation, the author establishes a tight lower bound on the optimal completion time to evaluate the heuristic. Computational results show that, on average, the modified Multifit method finds schedules that are within 4.5% of optimal.

The author examines a production system that is single stage. He further assumes that set-up times are sequence independent and significantly greater than processing times. Batches are divisible continuously for forming schedules and no machine processes more than one batch of the same part type or family. The completion time is independent of the order in which families and parts within a family are processed on a machine. This assumption is not valid in the presence of set-ups, which are usually sequence dependent.

2.1.5 Fabrication & assembly models

This problem occurs in the production of components on a machine, for assembly into end items. The production process consists of a fabrication stage and an assembly stage. The end items are assembled from several components, which are fabricated on a single machine. Each end item requires components unique to it and components common to all items. The assembly stage is not capacity constrained. The fabrication stage thus represents the capacity bottleneck. The machine incurs a fixed set-up between batches of different component types. There is no limit on the batch sizes. The problem is to schedule the fabrication stage so as to feed the assembly stage efficiently.

Coffman *et al* (1989) first examined this problem in which subassemblies of two types

are made and then assembled into end products. Each product contains one subassembly of each type. The objective is to minimize the total flow time of the products. The authors present rather involved $O(n)$ and $O(\sqrt{n})$ algorithms for the problem, where n is the number of products.

The study is applicable for single stage manufacturing systems only. Set-up time is constant for both subassemblies and all products contain both subassemblies, ie, there are no unique parts in the products. Moreover, the intricacy of the algorithms is such that extension to more than two products requires "a substantially greater effort". Finally, the authors do not present any computational examples.

Baker (1988) analyzed the problem with the objective of minimizing the mean completion time of jobs in fabrication. A *job* is the fabrication of the required quantity of common and unique components. A job is complete when its common and unique components have been fabricated. The author shows that SPT sequencing is optimal under the assumption that the product requiring the shortest run time for common parts will also require the shortest set-up and run time for unique parts. The remaining problem is to determine the number of set-ups and their location in the SPT sequence.

Consider a base schedule that contains no set-ups. A job's completion time will then exceed the completion time in the base schedule by a *delay* caused by set-ups. Thus the completion times can be minimized by minimizing the total delay. A job's delay is determined by the number of set-ups that precede it and by the number of following jobs prior to the next set-up. Suppose that at some stage a set-up is scheduled immediately before a run for the common components of product i and in the following batch parts for jobs i to k are produced. Thus a portion of the schedule is in the form $SC_i \ldots C_kU_i \ldots U_k$, where S = set-up time for a batch of common components, C_j = run time for the common components of product j, and U_j = set-up + run time for the unique components of product j. If v_{ik} denotes the delay incurred in this sequence, then $v_{ik} = (n - i + 1)S - (k - i + 1)(C_{k+1} + \ldots + C_n)$, where n is the total number of products. This definition allows a dynamic programming formulation

for the minimum total delay $F(n)$. Define $F(0) = 0$ and $F(k) = \min_{1 \le i \le k} \{F(i-1) + v_{ik}\}$.

The author gives a numerical example for four jobs and two set-ups.

The model considers only a single stage facility and each product requires only two part types. A job's requirement for common parts is related to its demand for unique components. A common component is common to all products. There is no limit on batch sizes and set-up times are independent of sequence. The assembly process is not capacity constrained. The implicit assumption is that the completion times of the jobs in fabrication determine the completion times of assembled products. Moreover, with many products the dynamic programming formulation becomes complex.

Aneja and Singh (1990) extended Baker's approach to the case where each product requires a unique component and m different common components, each requiring a separate set-up. The authors show that the problem is equivalent to solving m 2-component type problems. The optimal batching decisions for these m problems can be combined to obtain an optimal batching decision for the entire problem. The authors present an $O(mn)$ algorithm for finding a schedule that minimizes the total completion time of all products.

Sung & Park (1993) examined the same problem but with the minimization of mean flow time as the objective. The authors develop a branch and bound (BB) algorithm for the batch splitting instance, ie, an end product is complete when the fabrication of both its common and product dependent components is complete. They also constructed a dynamic programming (DP) algorithm and compared the performance of the two algorithms. The BB algorithm was more efficient in terms of both the memory requirement and execution time, especially for a large number of products.

The batching problem is complex for even the single-product, single-machine case. The only results available are the optimal number of batches and batch sizes. While it

seems that batching is unnecessary with only one product type, the machine may still need set-up for loading parts and maintenance. The need for set-up adds complexity to the problem. While the analytical models include an explicit set-up, the models based on queuing theory denote set-up by congestion and queuing delays. Only heuristics are available for the multiple product case. These heuristics work in two stages—allocating the work among batches and then sequencing the batches. The purpose of studying the this problem is to gain insight into the parallel machine case, which uses the results for the single machine case. The heuristics for the parallel machine case divide the problem into two steps. The first step is to allocate the work among the machines. Once work has been allocated, the machines can be considered independently and the batch sizes can be determined by using the results for the single machine problem.

The fabrication and assembly models are more restrictive. The assumption of infinite batch sizes implies that the machine never runs out of parts, which also precludes machine breakdowns. Set-ups are also assumed constant. The problem occurs frequently in the semiconductor industry where parts are inserted into circuit boards for assembly later in the manufacturing process. However, in real life many such machines may operate in series to minimize set-up, which is sequence dependent. The assumption of an infinite capacity assembly stage becomes less tenable under these circumstances.

2.2 Dual-stage Models

In dual-stage models, processing occurs in two stages. Ahmadi *et al* (1992) examined manufacturing systems that are equipped with batch and discrete processors. A discrete processor processes one job at a time while a batch processor processes a batch of jobs simultaneously. The authors analyzed a class of scheduling problem arising from a two-stage flowshop where the batch processor plays an important role such as a bottleneck.

Let β denote a batch processor, δ denote a discrete processor, and $\rightarrow$ denote

the system configuration. For example, $\beta \rightarrow \delta$ denotes a configuration in which a batch processor is followed by a discrete processor. The total number of jobs is an integer multiple of the batch size. Jobs are arranged in ascending order of their processing times. Given the system configuration and set of jobs to be processed, the scheduling problem is to decide the composition of the batches, the batch sequence on β and the job sequence on δ. The performance measures are the makespan and the sum of completion times. The system configurations are $\beta \rightarrow \delta$, $\delta \rightarrow \beta$, and $\beta_1 \rightarrow \beta_2$.

For the minimization of makespan on $\beta \rightarrow \delta$, the authors show that the optimal policy is the Full Batch–LPT schedule, ie, run full batches on β with the jobs sorted in the LPT order according to δ. This schedule avoids idle time on δ thus minimizing the makespan. The implicit assumption in a Full Batch is that the jobs in the batch are identical and the batch size is determined by the capacity of β. Since $\delta \rightarrow \beta$ is antithetical to $\beta \rightarrow \delta$, an SPT–Full Batch schedule is optimal on $\delta \rightarrow \beta$, ie, process the jobs according to SPT on δ and process full batches on β. This schedule is optimal because the batches are processed as early as possible, thus avoiding idle time on β. A Full Batch–Full Batch schedule is optimal for $\beta_1 \rightarrow \beta_2$. Jobs are processed on β_1 as soon as possible and on β_2 as late as possible without increasing the makespan.

The minimization of completion time is more involved than makespan. In a $\delta \rightarrow \beta$ system, the optimal policy is to process jobs in SPT on δ. The batch dispatching problem is then to determine the batch size and the processing order of batches on β for minimizing the sum of completion times. The authors present a dynamic program to determine an optimal schedule on β. In the $\beta_1 \rightarrow \beta_2$ system it is optimal to schedule full batches on β_1. The problem then reduces to $\delta \rightarrow \beta$. In the $\beta \rightarrow \delta$ system the Full-Batch policy is optimal on β_1, with the jobs in a batch sorted by SPT. The authors describe two heuristics to determine the job content of each batch to minimize the sum of completion times. The performance of the heuristics is compared with a lower bound that is generated by the Lagrangian relaxation of an IP. The heuristics perform "reasonably well". Extensions are presented for the multiple family and the three-machine flowshop problem.

A limitation of the study is the omission of set-ups. This stems from the ass-

umption that the processors process only one family of jobs simultaneously. However, set-ups are not present even in the multiple family case. Other limiting assumptions are that the processing times are known in advance and the batch processing time is independent of the composition of the batch.

The two-stage case is a natural extension of the single-stage case. There is a lack of research on two-stage models. A possible reason is the complexity involved considering that the only results available are for the single-machine, single product case.

2.3 Multi-stage Models

Multi-stage models are the most general case since there can be several machines in a given configuration. These machines can be arranged in series, parallel, or a combination of the two. A common thread that ties most multi-stage models is the use of queuing networks. In this approach, the entire facility is decomposed into its constituent nodes. Each node is then treated as an independent M/M/1, M/G/1, or G/G/1 queue. The route taken by an item or a class of items through the facility determines the amount of work at the node. The performance characteristics of each node, such as utilization and flow time, are computed using these models. Once the results are available for individual nodes, the results for the entire facility are obtained easily using the independence assumption.

Karmarkar *et al* (1985a) extended Karmarkar (1987) to a multi-item multi-machine job shop. The authors contend that in job shops with queues, the lot sizing problem can be formulated as the minimization of inventory cost because capacity constraints manifest themselves in queues and WIP.

$$\min_{Q} \quad \sum_i \sum_j h_{ij} D_i T_{ij}(Q) + \sum_i \frac{h_{if} Q_i}{2}$$

$$\rho_j(Q) < 1 \quad j = 1,\ldots,n$$

$$Q_j \geq 0 \quad i = 1,\ldots,m \ ,$$

where

h_{ij} = holding cost of units of item i waiting or being processed at work center j

h_{if} = holding cost of finished units of item i

D_i = total required output of item i (units/time)

Q_i = batch size for item i

ρ_j = traffic intensity at machine j

T_{ij} = mean time spent by batches of item i at machine j.

The first term in the objective represents WIP costs and the second finished goods cost. The first constraint is the stability condition for the queues at work centers. This constraint is redundant as each T_{ij} approaches infinity as ρ_j approaches one. Similarly the second constraint is also redundant since some ρ_j will exceed one as any Q_i approaches zero, due to increased set-ups. The authors also provide extensions to the model for set-up costs and safety stock costs.

The authors mention several applications of the model in job shop management. They also concede that the model will not handle seasonal demand variations and the cost of holding accumulated inventory. Also, all machines at a work center are assumed identical. The model cannot handle more than one limiting capacity at a work center, ie, a situation in which machine time and tool availability are both constrained.

Karmarkar *et al* (1985b) validated the results of the above study by examining the relationship between lot sizes and lead times. The authors examined these phenomena by developing an analytical model and a simulation model for an actual manufacturing cell. The manufacturing cell was arranged to improve the production of a group of

similar parts that had been experiencing long production lead times, high WIP, and difficulty in coordinating assemblies. The objective of the simulation was to devise appropriate lot sizing policies for the cell. Although the simulation was not validated against the operation of the cell, experience with the cell convinced the users of its validity. Experimentation with simulation led to a reduction of manufacturing lead time and WIP by over 50%.

The authors claim that queuing delays at machines in multi-item manufacturing shops are related directly to lot sizes. If the machine is modeled heuristically as an M/M/1 queue processing identical items, the authors show that the average time T spent in the system by a batch is given by

$$T = \frac{\tau + Q/P}{1 - D/P - D\tau/Q},$$

where

D = total work to be done (units/time)

P = processing rate of the machine (unit/time)

Q = batch size

τ = set-up time per batch.

This queuing model is extended to the multi-item case by modeling the facility as an M/G/1 queue. The case of a manufacturing system with several machines at each work center is modeled as an open network of M/G/c queues. At each stage the queuing model is embedded in an optimization model that determines optimum lot sizes. The most general case is coded as a computer program called Q-LOTS. A comparison of the results obtained by Q-LOTS and simulation with respect to average lead time indicates that Q-LOTS results are 20% better than the simulation approach. This may not be a shortcoming of the simulation as the authors do not test the simulation thoroughly. Another factor that restrained the performance of the simulation is that work is released at a uniform rate in the simulation but not in the analytical model.

Karmarkar *et al* (1987) extended Karmarkar *et al* (1985a,b) by examining the impact of operational issues such as equipment levels, multiple shifts, overtime, and batching policy on the performance of a manufacturing cell.

The cell is modeled as an open network of queues using node decomposition heuristics. The queuing model incorporates as decision variables the number and types of machines, the items to be processed, the production volume of each item, the shift and overtime policy, and the batching policy used in the cell. The model also selects the best batching policy for given capacity and load. The manufacturing cell consists of eight major work centers, two of them with multiple machines. The cell processes 27 parts with different routings and 3-12 operations/part. The available data comprises routings for each part including set-up and processing times. Raw material cost, finished goods cost, and annual demand are also known.

The increase in shift capacity expectedly leads to decreased manufacturing lead time and WIP though the improvements are obtained at a diminishing rate. The effects of batching policy are more interesting. Although the lot sizes being used were fairly small, the model finds improved lot sizes for some combinations of capacity and number of shifts. The model results in up to 40% improvement in performance in terms of lead time and WIP for all cases. However, the results were less striking in the two-shift case where the additional capacity had already alleviated queues and lead times. The effect of capacity changes was also deemed favorable in that the improvement in lead time and WIP outweighed the cost of additional equipment. The batching policies had substantial impact on the queues and bottlenecks although the results are inconclusive. Overall, there was an improvement in performance.

The authors contend that average performance indicators are misleading because they assume full capacity utilization. The model assumes resource availability and takes an average view of the manufacturing facility. Further, to model shifts and overtime, the processing rate of machines is altered. This has the effect of reducing variability in the process, which may affect the optimal batching policy.

Zipkin (1986) extended Karmarkar (1987) to develop cost minimization models for

batch production facilities, by including multiple products, backorders and final-product inventories. The production facility is described by fixed parameters, and decision variables represent batch sizes and safety stocks. Demand and the production process are stationary. The models are supposed to be simple enough to be used for comparing alternative facility configurations, evaluating new equipment, and assessing the effects of changes in demand and product mix. The basic premise behind the models is to represent the inventory of a product by a standard inventory model, and the production facility by a standard queuing model. These two models are then linked by the waiting time in the queuing system.

Numerous simplifying assumptions are made. The production scheduling function is passive and it does not utilize information on inventory for making decisions. Orders are thus processed on an FCFS basis. All costs, demand, and production processes are stationary. Batches of a product are of the same size and the demands of individual products are independent. There is considerable independence among demands during disjoint time intervals. Demand for each product is described by the mean and variance per unit time. All stockouts are backordered. The order processes for the products are independent and nearly renewal processes, the superposition of which constitute the arrival process at the production facility. The average inventory and backorders of each product are represented by its distribution of lead time demand. The time required to process a batch is dominated by the set-up time, not the batch size. The models are also highly mathematical in nature, which presents a limitation in modeling sufficiently realistic systems. The need for a queuing model necessitates the knowledge of the mean and variance of the waiting times, which may not be computable. The effect of batch size on processing time is not considered in detail, which is crucial in modeling the behavior of a batch processing facility.

Bertrand (1985) studied the effect of batch sizes on batch flow times and cost in a multi-product multi-stage manufacturing system. There is more than one routing for a batch and the mean production rate is constant for each product. Batches are served as they arrive, without preemption. The batch flow time model uses a closed shop queu-

ing model (Solberg 1981) to estimate values for the waiting times at the workcenters for given workload, capacity, and processing times. Since batch sizes affects flow times, the author develops a batch size optimization model. Analogous to Zipkin (1986), this model is linked to the queuing model via the in-process inventory. Bertrand presents an iterative procedure that produces batch sizes for minimizing cost.

Bertrand also represents total costs (ordering + carrying + WIP) as a function of the batch sizes and shows that this function is strictly convex, ie, it has a unique minimum. This function can be used to solve for the optimal batch sizes. By examining a simplified situation with homogeneous products, he showed that the consideration of carrying costs creates an upper bound for the optimal batch size for increasing demand. Also, neglecting the carrying costs may result in batch sizes being up to twice as large as optimal. He also assumed that processing time per batch at a work center is proportional to the batch size. However, set-ups can have a significant effect on the batch processing times.

The multi-stage case is the most general. There can be several stages, which can be arranged in an arbitrary configuration. Moreover, the decomposition approach lends considerable flexibility in modeling such facilities. In most cases, the only information needed is the arrival pattern of jobs, the mean and variance of the processing times, the configuration of the facility, and the routing of the jobs. While most of these parameters are available, the mean and variance of the processing times are almost impossible to estimate. The decomposition approach is also heuristic in nature. A major shortcoming is that it ignores the correlation that is added to the output job stream from a machine. This implies that the arrival process to the next stage is not a renewal process, which invalidates one of the assumptions of the decomposition approach.

2.4 Summary

The batching problem seems modest without set-ups. However, even advanced man-

ufacturing technologies incur set-up, which makes the problem intricate. Moreover, since different machines have different characteristics, a batch size optimal for one machine may not be optimal for another machine. The complexity of the problem makes it unsolvable for even the multiple item, single machine case. Only heuristics are available. Most research in batching is on the single stage case, which is not a realistic real-life scenario. The models also assume that the facility is a closed job shop where product variety is small and demand is known and constant. Most job shops are *open* and make parts to-order. This increases the product variety and it is difficult to estimate the load on the machines because demand is not known. The multi-stage models often rely on queuing networks, which is a heuristic approach. Additionally, most studies consider a single performance measure, or measures that are related. This makes it impossible to gage interactions among the measures and study the trade-offs involved. There have also been simulation studies but most models are tested inadequately and the details of the simulation models are minimal. The issues of variable demand, variation of set-up time, flow time variance, and processing time variability are also possible research directions.

Chapter 3
Analytical Models

In this chapter we consider a single server and a two server queuing system with batch service. A batch is defined as a collection of homogeneous items. Customers arrive from an infinite source according to a renewal process and wait to form a batch of size M. The customer who completes the batch (Mth customer) is termed the *super-*customer. The batch is then released for service after set-up. The server services the batch according to the FIFO protocol. A batch may have to wait for conclusion of service of the preceding batch(es) since preemption is not allowed. The interarrival time between individual customers is governed by a Phase-type distribution. The set-up time and service time distributions are also Phase-type. The two models presented are: (i) a PH/PH/1 queue, and (ii) a PH/PH/2 queue, both with individual arrival and batch service. These queues occur frequently in manufacturing systems, traffic control, and communication.

Manufacturing systems: Most manufacturing systems that process discrete parts release work in batches. Large batches minimize machine setup whereas small batches process work quickly. It is usually assumed that the performance of manufacturing systems is determined primarily by the scheduling policy. The batching policy also has a significant impact on the performance measures. This fact is significant because high machine utilization is very desirable.

Traffic control: The timing of traffic lights at an intersection can be optimized for various times of the day by realizing that the movement of traffic is essentially in variable sized batches because of traffic lights.

Communication: The transfer of data in windowed protocols is in batches.

In our case we consider the discrete time versions of the PH/PH/1 and PH/PH/2 queues. We obtain the probability distribution of wait before service of an individual customer in a batch.

In the following, **e** denotes a column vector of 1s, I is an identity matrix of appropriate dimensions, and 0 is a null matrix of appropriate dimensions. The symbol $\otimes$ denotes the Kronecker product of two matrices, where

$$A \otimes B = \begin{bmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{bmatrix}, \text{ if A is } m \times n.$$

3.1 Definitions

A probability distribution $F(\cdot)$ on $[0,\infty)$ is of *phase type* if and only if it is the distribution of time until absorption in a finite Markov process. A discrete phase distribution is defined by considering a $k+1$ state Markov chain of the form

$$P = \begin{bmatrix} T & T^0 \\ 0 & 1 \end{bmatrix},$$

where T is a substochastic matrix, such that $I-T$ is nonsingular, and $T_{ij} \geq 0 \ \forall i,j$. T^0 is an absorption matrix and $T\mathbf{e}+T^0=\mathbf{e}$. The initial probability vector is $(\boldsymbol{\alpha}, \alpha_{k+1})$, with $\boldsymbol{\alpha}\mathbf{e} + \alpha_{k+1} = 1$. We assume that states $1,\ldots,k$ are transient, so that absorption into state $k+1$, from any initial state, is certain (Neuts 1981).

3.2 PH/PH/1 Model

The arrival process of individual customers is phase type with representation (α', T') of order n, where α' is the initial probability vector and T' is an irreducible substochastic matrix of transition probabilities. The super-customer is the Mth customer and it completes the batch. Thus, by definition the interarrival time of the super-customer is

also phase type with representation (α,T) of order M, where $\alpha=(\alpha',\ \mathbf{0})$ and T is a convolution matrix. The service process of individual customers is phase type with representation (β',S'). Thus, the service process of the super-customer is also phase type with representation (β,S) of order m. The mean arrival rate, $\lambda^{-1}=\alpha(I-T)^{-1}\mathbf{e}$ and the mean service time, $\mu^{-1}=\beta(I-S)^{-1}\mathbf{e}$. The state space of the system can be described by $\Delta = \{(0,j),\ 0<j\leq Mn\} \cup \{(i,j,k),\ i>0,\ 0<j\leq Mn,\ 0<k\leq m\}$. The first part denotes an idle server and arrival of the super-customer in phase j. The second part denotes a busy server with i batches waiting for service, arrival in phase j, and service in phase k. The transition matrix P of this system is given by:

$$P = \begin{bmatrix} B_{00} & B_{01} & & & \\ B_{10} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{bmatrix},$$

where

$B_{00} = T$ $\qquad$ $B_{01} = T^0\alpha \otimes \beta$ $\qquad$ $B_{10} = T \otimes S^0$

$A_0 = T^0\alpha \otimes S$ $\qquad$ $A_1 = T\otimes S + T^0\alpha\otimes S^0\beta$ $\qquad$ $A_2 = T \otimes S^0\beta.$

T is a convolution matrix of order M,

$$T = \begin{bmatrix} T' & T'^0\alpha' & & \\ & T' & T'^0\alpha' & \\ & & \ddots & \ddots \\ & & & T' \end{bmatrix},$$

and S is a convolution matrix of order m,

$$S = \begin{bmatrix} S' & S'^0\beta' & & & \\ & S' & S'^0\beta' & & \\ & & \ddots & \ddots & \\ & & & & S' \end{bmatrix}.$$

If we allow a phase set-up time (δ, D) then,

$$S = \begin{bmatrix} D & D^0\beta & & & & \\ & S' & S'^0\beta' & & & \\ & & S' & S'^0\beta' & & \\ & & & \ddots & \ddots & \\ & & & & & S' \end{bmatrix},$$

where $\beta=(\delta, \mathbf{0})$.

If $\lambda/\mu<1$ then P is positive recurrent (Neuts 1981). For a stable system the invariant probability vector $\mathbf{x} = (x_0\ x_1\ x_2\ \dots\ )$ for the transition matrix P can be obtained by $\mathbf{x}P = \mathbf{x}$, $\mathbf{xe} = 1$, and $x_{i+1}=x_iR$, $i>0$. The matrix R is the minimal non-negative solution to the matrix quadratic equation:

$$R = A_0 + RA_1 + R^2A_2.$$

Since we have a complex structure near the lower boundary, x_0 has to be determined as follows. Define matrix $B[R]$,

$$B[R] = \begin{bmatrix} B_{00} & B_{01} \\ B_{10} & B_{11}+RA_2 \end{bmatrix}.$$

The stochastic matrix $B[R]$ has a positive left invariant vector (x_0, x_1), which can be normalized by

$$x_0e + x_1(I-R)^{-1}e = 1.$$

The state of the system upon an arrival of a super-customer is given by the vector $\mathbf{y}$. Let $D = T^0\alpha$, then

$$y_0 = \lambda^{-1}(x_0 \otimes D + x_1(D \otimes S^0))$$
$$y_k = \lambda^{-1}(x_k(D \otimes S) + x_{k+1}(D \otimes S^0\beta)),\ k>0.$$

The wait of a super-customer is given by the vector **w**, where

$$w_0 = y_0 e$$

$$w_r = \sum_1^r y_k\ (e \otimes I)\ G^k(r)e,\ r>0.$$

$G^k(r)$ is the probability that the service of the kth super-customer takes r time units.

$$G^1(r) = S^{r-1}\ S^0\beta$$
$$G^k(k) = (S^0\beta)^k,\ k>0$$
$$G^k(r) = S^0\beta\ G^{k-1}(r-1) + S\ G^k(r-1),\ r>k>1.$$

The probability that the wait of the mth customer in the batch is k time units,

$$w_k^m = \sum_{i=1}^{k} \tilde{w}_{k-i+1}^m\ \beta(m-\varsigma)\ S^{i-1}(m-\varsigma)\ S^0(m-\varsigma),$$

where

$$\tilde{w}_j^m = \sum_{i=1}^{j} w_{j-i+1}\alpha(M-m)\ T^{i-1}(M-m)\ T^0(M-m),$$

$$\alpha(M-m)\ T(M-m)\ T^0(M-m) = \left\{ \begin{matrix} 0 & j>1 \\ 1 & j=1 \end{matrix} \right\} \text{ for } m=M,$$

$$\beta(m-1)\ S(m-1)\ S^0(m-1) = \left\{ \begin{matrix} 0 & k>1 \\ 1 & k=1 \end{matrix} \right\} \text{ for } m=1,$$

and $\varsigma=0$ if there is set-up, 1 otherwise.

3.3 PH/PH/2 Model

The arrival process of the individual customers and the super-customer is identical to

the PH/PH/1 model. The service process of the super-customer is phase-type with representation (β_1,S_1) for server 1 and (β_2,S_2) for server 2. The batch selects server 1 with probability p and server 2 with probability $q=1-p$. The state space of the system can be described by $\Delta = \{(0,j),\ 0<j\leq Mn\} \cup \{(j,k_1),\ 0<j\leq Mn,\ 0<k_1\leq m\} \cup \{(j,k_2),\ 0<j\leq Mn,\ 0<k_2\leq m\} \cup \{(i,j,k_1,k_2),\ i>0,\ 0<j\leq Mn,\ 0<k_1,k_2\leq m\}$. The first part describes idle servers and arrival in phase j. The second part denotes a busy server 1 with arrival in phase j, and service in phase k_1. The third part describes a similar process for server 2. The last part denotes busy servers with i batches waiting for service, arrival in phase j, server 1 in phase k_1, and server 2 in phase k_2. The transition matrix P of this system is given by:

$$P = \begin{pmatrix} B_{00} & B_{01'} & B_{01''} & & & & \\ B_{1'0} & B_{1'1'} & B_{1'1''} & B_{1'2} & & & \\ B_{1''0} & B_{1''1'} & B_{1''1''} & B_{1''2} & & & \\ B_{20} & B_{21'} & B_{21''} & A_1 & A_0 & & \\ & B_{31'} & B_{31''} & A_2 & A_1 & A_0 & \\ & & & A_3 & A_2 & A_1 & A_0 \\ & & & & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$B_{00} = T$

$B_{01'} = p\ T^0\alpha \otimes \beta_1$

$B_{01''} = q\ T^0\alpha \otimes \beta_2$

$B_{1'0} = T \otimes S_1^0$

$B_{1'1'} = T\otimes S_1 + p\ T^0\alpha\otimes S_1^0\beta_1$

$B_{1'1''} = q\ T^0\alpha \otimes S_1^0\beta_2$

$B_{1'2} = T^0\alpha \otimes S_1\beta_2$

$B_{1''0} = T \otimes S_2^0$

$B_{1''1'} = p\ T^0\alpha \otimes S_2^0\beta_1$

$B_{1''1''} = T\otimes S_2 + q\ T^0\alpha\otimes S_2^0\beta_2$

$$B_{1''2} = T^0\alpha \otimes S_2\beta_1$$
$$B_{20} = T \otimes S_1^0 \otimes S_2^0$$
$$B_{21'} = T\otimes S_1\otimes S_2^0 + p\, T^0\alpha\otimes S_1^0\beta_1\otimes S_2^0$$
$$B_{21''} = T\otimes S_1^0\otimes S_2 + q\, T^0\alpha\otimes S_1^0\otimes S_2^0\beta_2$$
$$B_{31'} = p\, T \otimes S_1^0\beta_1 \otimes S_2^0$$
$$B_{31''} = q\, T \otimes S_1^0 \otimes S_2^0\beta_2$$
$$A_0 = T^0\alpha \otimes S_1 \otimes S_2$$
$$A_1 = T\otimes S_1\otimes S_2 + T^0\alpha\otimes S_1^0\beta_1\otimes S_2 + T^0\alpha\otimes S_1\otimes S_2^0\beta_2$$
$$A_2 = T\otimes S_1^0\beta_1\otimes S_2 + T\otimes S_1\otimes S_2^0\beta_2 + T^0\alpha\otimes S_1^0\beta_1\otimes S_2^0\beta_2$$
$$A_3 = T \otimes S_1^0\beta_1 \otimes S_2^0\beta_2.$$

If $\lambda/(\mu_1+\mu_2) < 1$ then P is positive recurrent. For a stable system the probability vector $\mathbf{x} = (x_0\ x_{1'}\ x_{1''}\ x_2\ x_3\ \dots\ )$ for the transition matrix P can be obtained by $\mathbf{x}P = \mathbf{x}$, $\mathbf{x}\mathbf{e} = 1$, and $x_{i+1}=x_iR$, $i>0$. The matrix R is the minimal non-negative solution to the matrix cubic equation:

$$R = A_0 + RA_1 + R^2A_2 + R^3A_3.$$

Since we have a complex structure near the lower boundary, x_0 has to be obtained as follows. Define matrix $B[R]$,

$$B[R] = \begin{pmatrix} B_{00} & B_{01} \\ B_{10}+RB_{20} & A_1+RA_2+R^2A_3 \end{pmatrix}.$$

The stochastic matrix $B[R]$ has a positive left invariant n-vector $(x_0\ x_{1'}\ x_{1''}\ x_2)$, which can be normalized as before.

Let $\mathbf{y} = (y_0\ y_{1'}\ y_{1''}\ y_2\ y_3\ \dots\ )$ be the state of the system at an arrival. Let $D = T^0\alpha$, then

$$y_0 = \lambda^{-1}[x_0(D\otimes I) + x_{1'}(D\otimes S_1^0) + x_{1''}(D\otimes S_2^0) + x_2(D\otimes S_1^0\otimes S_2^0)]$$
$$y_{1'} = \lambda^{-1}[x_{1'}(D\otimes S_1) + x_2(D\otimes S_1\otimes S_2^0)]$$

$$y_{1''} = \lambda^{-1}[x_{1''}(D \otimes S_2) + x_2(D \otimes S_1^0 \otimes S_2)]$$
$$y_k = \lambda^{-1}[x_k(D \otimes S_1 \otimes S_2) + x_{k+1}(D(S_1^0\beta_1 \otimes S_2 + S_1 \otimes S_2^0\beta_2))],\ k>1.$$

Define $\mathbf{z} = (z_0\ z_{1'}\ z_{1''}\ z_2\ z_3\ \dots\ )$, where $z_0=y_0\mathbf{e}$, $z_{1'}=y_{1'}\mathbf{e}$, $z_{1''}=y_{1''}\mathbf{e}$, and $z_i=y_i(\mathbf{e}\otimes I)$, $i>1$. Define $\Delta = \{0,\ 1',\ 1'',\ 2,\ 3,\ \dots\ \}$, where $\{0,\ 1',\ 1''\}$ are absorbing states and the other states are transient. Consider the transition matrix $\hat{P}$ which represents the absorbing Markov chain corresponding to P.

$$\hat{P} = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ S_1^0\otimes S_2^0 & S_1\otimes S_2^0 & S_1^0\otimes S_2 & S_1\otimes S_2 & & & \\ & p\,S_1^0\otimes S_2^0\beta_1 & q\,S_1^0\otimes S_2^0\beta_2 & S_1^0\beta_1\otimes S_2 + S_1\otimes S_2^0\beta_2 & S_1\otimes S_2 & & \\ & & & S_1^0\beta_1\otimes S_2^0\beta_2 & S_1^0\beta_1\otimes S_2 + S_1\otimes S_2^0\beta_2 & S_1\otimes S_2 & \\ & & & \ddots & \ddots & \ddots & \end{bmatrix}$$

Let $z^j = zP^j$, where $z^j = (z_0^j\ z_1^j\ z_{1'}^j\ z_2^j\ z_3^j\ \dots\ )$. If w_j is the probability that the waiting time is j, then

$$w_j = \theta_0 z_0^j + \theta_{1'} z_1^j + \theta_{1''} z_{1''}^j.$$

θ is the probability of getting absorbed in the respective state and is calculated as follows. Arrange $\hat{P}$ with transient states first and then the absorbing states. The transition matrix for the absorbing chain is then truncated to

$$\hat{P}' = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}.$$

Q represents transitions between transient states (2,3) since these states are the only gateways to the absorbing states from the transient states; R represents transitions from transient to absorbing states (0,1′,1″); 0 reflects that we cannot go from absorbing to transient states; and I reflects that we cannot leave an absorbing state. The

probability of absorption into the absorbing states is then:

$$\tilde{z}(I-Q)^{-1}R = (\gamma_0\ \gamma_{1'}\ \gamma_{1''}),$$

where

$$\tilde{z} = \frac{1}{z_2+z_3}(z_2\ z_3)\ .$$

$\tilde{z}$ gives the ratio of customers entering the absorbing states from the transient states. If $u = z_0+z_{1'}+z_{1''}$ $(u \leq 1)$ gives the probability of getting absorbed in the absorbing states, then

$$(1-u)\ (\gamma_0\ \gamma_{1'}\ \gamma_{1''}) = (\tau_0\ \tau_{1'}\ \tau_{1''}).$$

is the probability of absorption starting from the transient states. Thus,

$$\theta_0 = z_0 + \tau_0$$
$$\theta_{1'} = z_{1'} + \tau_{1'}$$
$$\theta_{1''} = z_{1''} + \tau_{1''}.$$

Once the wait before service of the batch has been determined, the wait before service of an individual customer in the batch can be determined as in the PH/PH/1 model. Since the batch can select server 1 with probability p and server 2 with probability $1-p$ the final wait is a linear combination of the two waiting times.

3.4 Examples

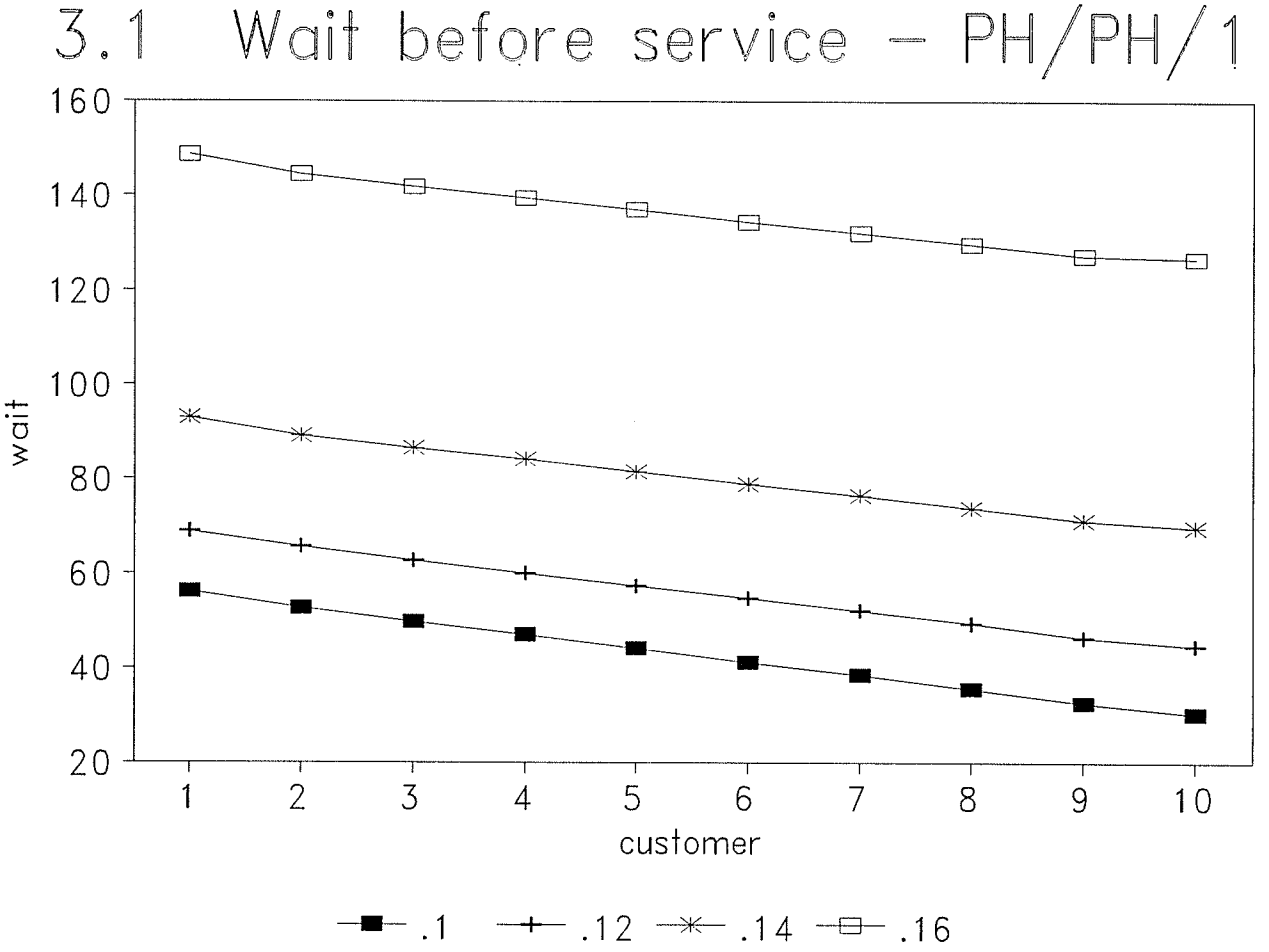
In the following examples the arrival and service time distributions are Geometric and the batch size is 10. We ignore set-up as it simply shifts the plots upwards because of the increased wait for all customers in the batch. We examine the effect of traffic on the wait before service of individual customers in the batch. The corresponding graphs are at the end of the chapter. Graph 3.1 shows the wait before service in the

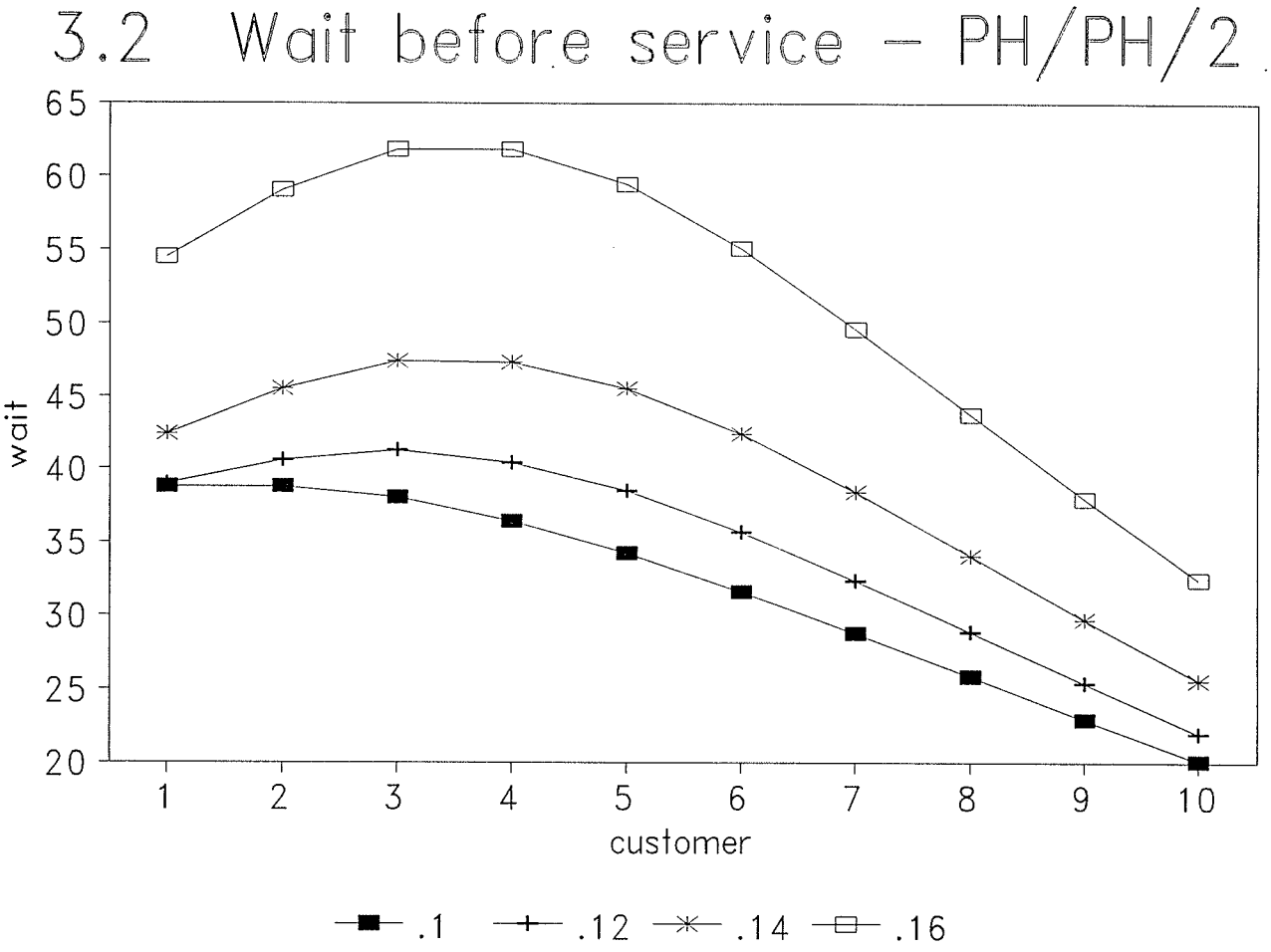
PH/PH/1 queue. The service rate is .2 and the arrival rates are .1, .12, .14, and .16. The wait decreases as the initial customers in the batch have to wait for more arrivals than the trailing customers.

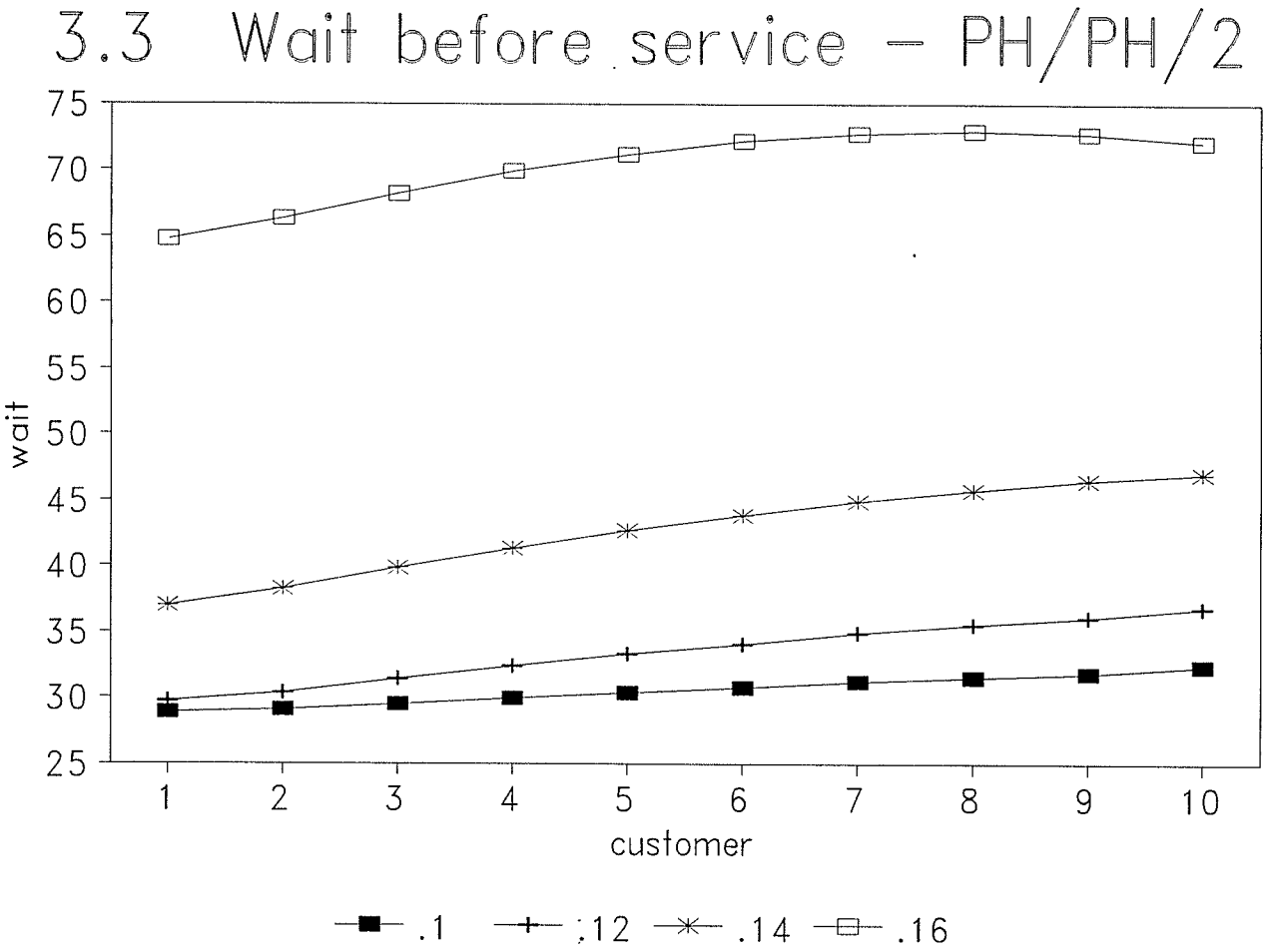
In the PH/PH/2 case, graph 3.2 corresponds to service rates .15 and .05 with probabilities .25 and .75 respectively. The arrival rate varies from .1 to .16. The wait of a customer in a batch is composed of the wait for arrival of the succeeding customers, the wait for service of the preceding customers, and the wait for the preceding batches. The wait for the preceding batches is the same for all customers in the batch and the wait for arrivals is always decreasing. The concave shape can be explained by the wait for service, which supplants the wait for arrivals for the first few customers. In graph 3.3, the service rates are .1 with probability .5 each and the arrival rates are the same as above.

3.5 Summary

We have obtained the probability distribution of the waiting time of a customer in a batch in a PH/PH/1 and a PH/PH/2 queuing system using the matrix-geometric method. The results can be used to obtain exact results for single- and two-server queuing systems. The models can also be applied at a macro level to merge a number of facilities together and study them as one. Ultimately the models can help select appropriate batch sizes. We will now attempt to study larger systems using simulation. In the next chapter we establish the methodology and experimental design for the simulation experiments.


3.1 Wait before service - PH/PH/1
160
140
120
100
80
60
40
20
wait
1
2
3
4
5
6
7
8
9
10
customer
.1
.12
.14
.16


3.2 Wait before service - PH/PH/2
wait
65
60
55
50
45
40
35
30
25
20
1
2
3
4
5
6
7
8
9
10
customer
.1
.12
.14
.16


3.3 Wait before service – PH/PH/2
wait
75
70
65
60
55
50
45
40
35
30
25
1
2
3
4
5
6
7
8
9
10
customer
.1
.12
.14
.16

Chapter 4
Research Methodology

Experimental design is an important step in an empirical study. The design phase is concerned with setting various parameters of the model and deciding on the conditions under which the model will be appropriate. In this chapter we present the approach that will be used to achieve the research objectives. We will review the methodology and describe the statistical models, formulate statistical hypotheses consistent with the research objectives, discuss the issues of verification/validation and data analysis, and outline the research plan.

The objective in this research is to assess the performance of different batching decisions under given configurations of a manufacturing facility by using experimental design to measure the effect of batch size, demand, processing time variability, scheduling rule, and set-up time (independent variables) on flow time, flow time coefficient of variance, resource utilization, and the percentage of tardy batches (dependent variables). Flow time of a job is the time that the job spends in the system. Tardiness is the positive difference between the completion time and the due date of a job.

In a flow line, each job must be processed once on each machine in the same order. In a job shop, not all jobs require processing on all machines and some jobs may require multiple operations on a single machine. Each job may also have a different sequence of operations and it may visit a machine more than once. The basic assumption in this study is that a manufacturing facility can be decomposed into nodes with the characteristics of flow lines or job shops. Further, a series of flow lines can be represented by one flow line node and a group of job shops can be represented by one job shop node. This transformation (which is beyond the scope of this study) can be performed by replacing the similar consecutive nodes by a *super*-node. The set-up and processing time of the super-node can then be moderated to model the bottleneck

node in the individual facility. Alternatively, these two parameters can be moderated so that the super-node models the entire facility. In this case, the parameters would reflect the total set-up and processing time of the jobs processed in the facility. Thus, the value of these two parameters should be greater for a job shop than for a flow line. While this approach does not allow a detailed analysis of the facility, it is useful in the product/process planning stage. These facilities usually experience bottlenecks in the job shops because of the complex workflow. Appropriate batching and scheduling rules can alleviate the bottlenecks thus smoothing the workflow in the facility.

4.1 Methodology

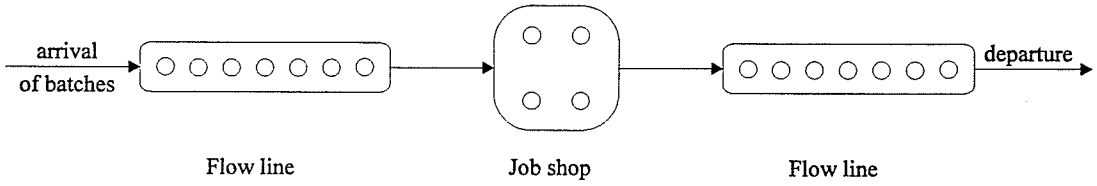
The methodology is composed of three steps. The first step is to (i) model a small system with job shop and flow line attributes, (ii) acquire an understanding of the interactions among the dependent and independent variables, (iii) verify the simulation results with the analytical models, and (iv) iterate for larger systems. The second step is to use experimental design and statistical hypotheses to find significant relationships between the dependent and independent variables. The third step is to synthesize the knowledge thus gained to make policy recommendations for generic systems.

The systems will be modeled using simulation and network models. The most elementary system that can be modeled has two nodes—the first representing a flow line and the second representing a job shop. A variation on this is a system with these nodes in reverse order. More complex systems are three nodes in series with the middle node representing a job shop and the outer nodes representing flow lines and a series of alternating job shop and flow line nodes. This approach will allow an understanding of the interface between a flow line and a job shop and the interactions among the dependent and independent variables. The methodology will also provide general recommendations on policy issues.

Coding a simulation model is a first step in empirical research. Issues such as initialization bias, verification, and validation must be resolved before the simulation model can help quantify relationships between input and output variables, and serve as

a precursor for good experimental design. Following the process description we review these issues, suggest possible resolutions, and discuss the experimental design.

4.2 Process Description



A typical configuration

The facility consists of nodes arranged in series. Each node represents a flow line or a job shop and is modeled by a single server. Because we are assuming that flow lines and job shops can be used to represent a facility, flow line and job shop nodes always alternate. The two types of node are differentiated by set-up and processing time, ie, the values of these parameters are greater for a job shop than a flow line. Batches arrive according to a Poisson process and wait for service. Note that arrival is in batches in contrast with the analytical model in which customers arrive individually and then form batches. Although individual arrivals are more realistic, the situation can also be modeled by batch arrivals, noting the arrival of *super*-customers. These are customers that complete the batch. Individual arrivals were also examined in pilot runs of the simulation model. However, not only is this alternative computing time intensive, it does not affect the performance measures significantly. Each batch incurs a set-up before service. In a real-life situation, set-up may also be incurred within a batch, eg, machine loading and machine breakdown, but these set-ups will be ignored. If the server is busy, the batch waits for service, which proceeds without preemption. Upon completion of service a batch advances to the next node or leaves the system if there are no nodes left. All batches visit all the nodes once, and in the same sequence, starting at the first node.

4.3 Simulation Issues

Simulations can be classified broadly into terminating and non-terminating, depending on whether there is an obvious way of determining run length. A terminating simulation is one for which there is a natural event that stops the simulation. This event is specified before any runs are made and it may also be a random variable. A non-terminating simulation is one that has no such terminating event to specify run length. Performance measures for such systems are often steady state parameters like the mean. Whereas not many real systems reach a steady state, a simulation model may reach steady state. The reason is that the system can change characteristics over time as opposed to a simulation model. Usually one is interested in the long run (steady state) properties of the system, ie, properties that are not influenced by the conditions at time zero. Thus, a simulation can be terminating or non-terminating depending on the objectives of the simulation study (Banks & Carson 1984, Law & Kelton 1991). In this research, our interest is in the steady state behavior of the system.

4.3.1 Verification and validation

A simulation model is only an abstraction of the real system being studied and should be looked at skeptically until its credence can be established. The process of establishing that a simulation model is a credible representation of the real system is called model verification and validation.

Model verification is accomplished by ensuring the correct programming and implementation of the computer model. Techniques include sensitivity analysis—executing the model under different conditions—and assessing the output for accuracy, traces, and reprogramming selected components of the program. Validation consists of ensuring that a model can be considered valid for all its applications (Sargent 1988). Balci and Sargent (1984) proposed various techniques for validating a model. The conclusive validation of a model patterned after a real facility is to compare and contrast its behavior with the facility. In cases where the model is hypotheti-

cal or the real facility is changing constantly, the three-step approach suggested by Carson (1986) can be used. This approach consists of (i) developing a model with high face validity, (ii) validating model specifications and assumptions, and (iii) validating model output.

The simulation model of the real facility was verified by interactive tracing. Additional methods used were sensitivity analysis and reprogramming certain modules of the code. As with verification, there is no predominant technique available for validation (Hoover & Perry 1989). Face validity of the model was ensured by involving practitioners from the manufacturing facility during the coding stage. The legitimacy of model specifications and assumptions was ensured similarly. Most data used were real-time and collected at the facility without human interference. Experimentation suggested that the simulation model was an adequate representation of the manufacturing facility.

4.3.2 Queuing network analyzer

There is always some degree of skepticism involved in validating a simulation model. In this study we will use the Queuing Network Analyzer (QNA) as a validation tool for the simulation and to develop approximations for the performance measures. QNA analyzes networks of queues with the FCFS discipline and no capacity constraints. QNA supports three basic network operations: (i) merging, (ii) splitting, and (iii) departure. External arrival processes can be non-Poisson and the service time distribution can be non-Exponential. It characterizes the arrival process and service time distributions by their first two moments, the mean and variability. The nodes are then analyzed individually as standard G/G/m queues. Congestion measures for the network as a whole are obtained by assuming that the nodes are stochastically independent given the approximate flow parameters (Whitt 1983a).

The first step in the algorithm is to solve for the flow rates and the variability parameters of the internal arrival processes. The second step is to compute approximate congestion measures for each queue separately by regarding it as a G/G/m queue in which the arrival process and the service time distribution are characterized by the

rate and variability parameters. The final step is to calculate congestion measures for the network as a whole (Whitt 1983b).

As an example, consider an open network containing a single node with a single server, and the FCFS discipline (Whitt 1983a). The standard Markov model of this network is the classical M/M/1 queue with Poisson arrival and Exponential service. The expected waiting time (before service), $E[W] = \tau\rho/(1-\rho)$, where τ is the mean service time, ρ is the traffic intensity, and $0 \leq \rho < 1$. QNA uses an approximation for the G/G/1 model to represent this network. The G/G/1 model has a renewal arrival process and both the interarrival and the service time distributions are General. In QNA, the arrival process is represented by a renewal process characterized by two parameters: the arrival rate λ and the variability parameter c_a^2. The service time distribution is also characterized by two parameters: the mean service time τ and the variability parameter c_s^2. The expression for the expected waiting time in QNA is

$$E[W] = \frac{\tau\rho(c_a^2 + c_s^2)g}{2(1-\rho)}, \text{ where } g = g(\rho, c_a^2, c_s^2) = \begin{cases} 1 & c_a^2 \geq 1 \\ <1 & c_a^2 < 1 \end{cases}.$$

When $g=1$, the above formula differs from $E[W]$ in an M/M/1 model by $(c_a^2 + c_s^2)/2$. When the arrival is Poisson and the service time is Exponential, $c_a^2 = c_s^2 = 1$. Thus, in a network with Poisson arrival and Exponential service, QNA gives exact results. QNA is a tool that is adaptable to diverse modeling situations. However, it is based on some restrictive assumptions—it cannot capture the detail of a simulation model.

4.3.3 Output analysis

A discrete event simulation aggregates the confluence of many random variables. Not surprisingly, the output of the model is itself a random variable, which can be misinterpreted easily. This could result in false conclusions about the system that the simulation represents. When analyzing simulation output, it is essential that the data verify the classical assumptions of being independent and distributed identically.

The major issues involved are those of initialization bias, normality, and correlation.

The data collected during the early part of a simulation may be biased by the initial state of the system. Initialization bias is the effect of the "warmup period" on the performance measures. A simple solution is to determine the length of the warm-up period by graphing the performance measure against time and discarding the observations during the transient phase. Schruben (1982) and Schruben *et al* (1983) presented techniques for detecting initialization bias in simulation output.

The variables should also be distributed normally. There are many tests for determining normality. A popular graphical test is the Normal probability plot, where the observations are arranged in increasing order of magnitude and plotted against expected Normal distribution values. The plot should resemble a straight line in the presence of normality. Another option is to examine the histogram for the variable.

Another problem is that of correlation between successive observations arising from a simulation, ie, the observations are not independent of each other. Batching (Conway 1963) is a conceptually straightforward method that transforms correlated observations into fewer (almost) uncorrelated and normally distributed batch means. In this method, one long simulation is performed and the performance measures are recorded periodically and then reset. The time period may be based on either the simulation clock or the occurrence of a certain number of events. If the lapse between successive resets is sufficiently large, the statistics accumulated during each interval may be considered independent (Hoover & Perry, 1989). The method of batch means assumes that initial transient effects have been removed (Schmeiser 1982).

4.3.4 Variance reduction

In many simulation studies, there is often an opportunity to run the simulation in a way that would get more precise estimates, than would be possible otherwise by running the simulation in a normal, straightforward way. A primary means of obtaining more precise estimates of the relevant parameters is by reducing the variance of the point estimates of the parameters. One approach toward reducing variability is to operate the different models under identical random conditions. This eliminates a

source of variability thus causing the performance measures to have lower variance (Thesen & Travis 1988).

Common random numbers and antithetic variates are often used for variance reduction. The method of antithetic variates entails inducing negative correlation between the appropriate input random variables. Whereas the use of antithetic variates draws on many subtleties (Henriksen & Crain 1989, Law & Kelton 1991), the common random number technique is conceptually simple and will be used in this research. The technique requires using the same random number stream across alternate configurations of the model, thus reducing a source of variability. It may be noted that these techniques are limited in use and may even backfire if the random number streams are not synchronized across alternative configurations. It is not known whether the common random number technique will be effective in containing variance.

4.4 Experimental Design

Simulation is an evaluative technique—it can provide only an estimate of the chosen performance parameters. An optimal set of these parameters would require a procedure that can quantify the relationships among these parameters and detect relationships that are significant. Factorial design is one such technique that can be used for examining the impact of these factors on the response variables. In a factorial design, each replication of the experiment examines all possible combinations of the levels of the factors. Factorial design has many benefits. A factorial design (Montgomery 1991): (i) is more efficient than one-factor-at-a-time experiments, (ii) is necessary when interactions may be present to avoid misleading conclusions, and (iii) allows the effects of a factor to be estimated at several levels of the other factors, yielding conclusions that are valid over a range of experimental conditions.

In this research, five levels of batch size, four levels of scheduling rule, two levels of set-up time, two levels of demand, and two levels of processing time variability will be examined. Batch size is 10, 20, 30, 40, or 50. Small batches scheduled

in SPT may lower WIP; large batches scheduled in LPT may increase resource utilization. Note that when the batches are ordered by processing times, *small* and *large* refer to the processing times. Scheduling in SPT and LPT is done by the processing time added to set-up time over all nodes. These batch sizes may be restrictive but the intent is to provide a framework for decision making. Demand is moderated by interarrival time, which is Exponential. Interarrival time varies with batch size. This ensures that the facility processes roughly the same amount of work irrespective of the batch size. The two levels of interarrival time are aimed at achieving 60% and 80% utilization of the bottleneck (job shop) nodes respectively. The higher level ensures that there are queues sufficiently long for the scheduling rules to be effective. The interarrival times are calculated using Little's law since the utilization and service times are known.

Scheduling rule is FIFO, LPT, SPT, or EDD. FIFO is chosen as a base case with which the effect of the other three rules can be compared. LPT has been recommended in many studies (eg, Dobson *et al* 1987, Naddef & Santos 1988) as releasing large batches first reduces variability in the system. SPT is chosen because it may affect flow time and WIP favorably. Meeting due dates has been cited as the most important criterion by practitioners (Smith *et al* 1986). Due dates are decided by the Total Work Content (TWK) method (Conway 1965). This method is preferred by researchers in assigning due dates (Baker & Kanet 1983). According to this method, the due date of a job is its arrival time added to a constant (≥ 1) multiplied by the job's total processing time. The constant used in this research is 20. This value returns a wide range of percent tardy batches and is based on pilot runs of the simulation model. Set-up time is Exponential with a mean of 5 minutes/batch for the flow line nodes and 10 or 20 minutes/batch for the job shop nodes. Processing time is Exponential for the flow line nodes with a mean of one minute/job and Exponential or Normal for the job shop nodes with a mean of two minutes/job. Processing time variability applies to the Normal distribution only and the two levels are N(2,.66) and N(2,.1).

Parameter Settings

	FLOW LINE	JOB SHOP
BATCH SIZE	10, 20, 30, 40, 50	
SCHEDULING RULE	FIFO, SPT, LPT, EDD	
SET-UP	E(5)	E(10),E(20)
DEMAND (resource utilization)	60%, 80%	
PROCESSING TIME/JOB	E(1)	E(2), N(2,.1), N(2,.66)

There are 100 replications of each experiment. Each replicate contains all 160 combinations of the independent variables. There are four dependent variables (resource utilization, flow time, flow time CV, and percent tardy batches). Let $Y_{ijklmrv}$ represent the observation for the ith level of batch size, jth level of demand, kth level of scheduling rule, lth level of set-up, and mth level of processing time variability in the rth replicate for the vth independent variable. The observation can be described by the linear statistical model,

$$Y_{ijklmrv} = \mu_v + \alpha_{iv} + \beta_{jv} + \gamma_{kv} + \delta_{lv} + \zeta_{mv} + (\alpha\beta)_{ijv} + (\alpha\gamma)_{ikv} + (\alpha\delta)_{ilv} + (\alpha\zeta)_{imv} + (\beta\gamma)_{jkv} + (\beta\delta)_{jlv} + (\beta\zeta)_{jmv} + (\gamma\delta)_{klv} + (\gamma\zeta)_{kmv} + (\delta\zeta)_{lmv} + (\alpha\beta\gamma)_{ijkv} + (\alpha\beta\delta)_{ijlv} + (\alpha\beta\zeta)_{ijmv} + (\alpha\gamma\delta)_{iklv} + (\alpha\gamma\zeta)_{ikmv} + (\alpha\delta\zeta)_{ilmv} + (\beta\gamma\delta)_{jklv} + (\beta\gamma\zeta)_{jkmv} + (\beta\delta\zeta)_{jlmv} + (\gamma\delta\zeta)_{klmv} + (\alpha\beta\gamma\delta)_{ijklv} + (\alpha\beta\gamma\zeta)_{ijkmv} + (\alpha\beta\delta\zeta)_{ijlmv} + (\alpha\gamma\delta\zeta)_{iklmv} + (\beta\gamma\delta\zeta)_{jklmv} + (\alpha\beta\gamma\delta\zeta)_{ijklmv} + \epsilon_{ijklmrv};$$

for

(batch size) $i = 1,2,3,4,5;$

(scheduling rule) $j = 1,2,3,4;$

(set-up time) $k = 1,2$;

(demand) $l = 1,2$;

(processing time variability) $m = 1,2$;

(replicate) $r = 1,\ldots,100$; and

(independent variable) $v = 1,2,3,4$;

where

μ_v = overall mean effect;

α_{iv} = effect of the ith level of batch size;

β_{jv} = effect of the jth level of scheduling rule;

γ_{kv} = effect of the kth level of set-up time;

δ_{lv} = effect of the lth level of demand;

ζ_{mv} = effect of the mth level of processing time variability;

$(\alpha\beta)_{ijv}$ = effect of the interaction between α_i and β_j;

$(\alpha\gamma)_{ikv}$ = effect of the interaction between α_i and γ_k;

$(\alpha\delta)_{ilv}$ = effect of the interaction between α_i and δ_l;

$(\alpha\zeta)_{imv}$ = effect of the interaction between α_i and ζ_m;

$(\beta\gamma)_{jkv}$ = effect of the interaction between β_j and γ_k;

$(\beta\delta)_{jlv}$ = effect of the interaction between β_j and δ_l;

$(\beta\zeta)_{jmv}$ = effect of the interaction between β_j and ζ_m;

$(\gamma\delta)_{klv}$ = effect of the interaction between γ_k and δ_l;

$(\gamma\zeta)_{kmv}$ = effect of the interaction between γ_k and ζ_m;

$(\delta\zeta)_{lmv}$ = effect of the interaction between δ_l and ζ_m;

$(\alpha\beta\gamma)_{ijkv}$ = effect of the interaction among α_i, β_j, and γ_k;

$(\alpha\beta\delta)_{ijlv}$ = effect of the interaction among α_i, β_j, and δ_l;

$(\alpha\beta\zeta)_{ijmv}$ = effect of the interaction among α_i, β_j, and ζ_m;

$(\alpha\gamma\delta)_{iklv}$ = effect of the interaction among α_i, γ_k, and δ_l;

$(\alpha\gamma\zeta)_{ikmv}$ = effect of the interaction among α_i, γ_k, and ζ_m;

$(\alpha\delta\zeta)_{ilmv}$ = effect of the interaction among α_i, δ_l, and ζ_m;

$(\beta\gamma\delta)_{jklv}$ = effect of the interaction among β_j, γ_k, and δ_l;
$(\beta\gamma\zeta)_{jkmv}$ = effect of the interaction among β_j, γ_k, and ζ_m;
$(\beta\delta\zeta)_{jlmv}$ = effect of the interaction among β_j, δ_l, and ζ_m;
$(\gamma\delta\zeta)_{klmv}$ = effect of the interaction among γ_k, δ_l, and ζ_m;
$(\alpha\beta\gamma\delta)_{ijklv}$ = effect of the interaction among α_i, β_j, γ_k, and δ_l;
$(\alpha\beta\gamma\zeta)_{ijkmv}$ = effect of the interaction among α_i, β_j, γ_k, and ζ_m;
$(\alpha\beta\delta\zeta)_{ijlmv}$ = effect of the interaction among α_i, β_j, δ_l, and ζ_m;
$(\alpha\gamma\delta\zeta)_{iklmv}$ = effect of the interaction among α_i, γ_k, δ_l, and ζ_m;
$(\beta\gamma\delta\zeta)_{jklmv}$ = effect of the interaction among β_j, γ_k, δ_l, and ζ_m; and
$(\alpha\beta\gamma\delta\zeta)_{ijklmv}$ = effect of the interaction among α_i, β_j, γ_k, δ_l, and ζ_m;

on the vth dependent variable, and ϵ_{ijklmv} is the random error component. We are interested in the main effects and the first order interactions only. While the use of such a model with factorial design can measure the effects, the real power of factorial design lies in hypothesis testing, ie, testing which of these effects affect the dependent variable(s) in a statistically significant sense.

4.5 Research Hypotheses

Hypothesis testing allows the comparison of different configurations of a system to be made on objective terms, with a knowledge of the risks associated with reaching the wrong conclusion (Montgomery 1991). In this research, statistical hypotheses are formulated to analyze differences among treatment levels (levels of independent variables) and their effect on the dependent variables. Two procedures are employed: analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). ANOVA is a univariate procedure as it is used to assess differences among treatment levels on a dependent variable. In other words, ANOVA is used to determine if samples are from populations with equal means. MANOVA is a multivariate procedure as it is used to assess simultaneously differences among treatment levels across several dependent variables. Both test for equality of two or more population means.

MANOVA is an extension of ANOVA, ie, for every F statistic in ANOVA that evaluates treatment effects on a dependent variable, there is a corresponding multivariate statistic (eg, Wilks' Λ) that evaluates the same effect on a set of dependent variables. As statistical inference procedures, both ANOVA and MANOVA are used to assess the statistical significance of differences among treatment levels (Hair *et al* 1987).

There are three main approaches that can be used to analyze data from a multivariate experiment (Hummel & Sligo 1971). One approach is to test each H_0: $\mu_{j1}=\mu_{j2}$ with ANOVA. A second approach is to conduct an overall multivariate test of H_0: $\mu_1=\mu_2$. If H_0 is rejected then one can infer that $\mu_{j1}\neq\mu_{j2}$ for at least one j. Cramer and Bock (1966) recommended that ANOVA can then be run on each variable separately. A third approach by Morrison (1967) follows the rejection of H_0 with simultaneous confidence intervals developed by Roy and Bose (1953). This approach controls overall α for many comparisons. Barcikowski (1983) and Hummel and Sligo (1971) recommended the combination M/ANOVA approach for testing hypotheses because it controls experimentwise error rate (probability that at least one comparison will be declared significant when in fact H_0 is true for all comparisons) better than the other two approaches. According to this approach, an overall multivariate test is conducted. If the test is significant, further post hoc experimentation (eg, univariate tests on each response variable) is conducted.

4.6 Multivariate Hypotheses

Following Barcikowski (1983) and Hummel and Sligo (1971), MANOVA will be used initially to evaluate differences among factor levels on the set of dependent variables. On indication of significance of these differences, ANOVA will be used to test the significance of the factor levels on individual dependent variables.

MANOVA is based on the following assumptions (Stevens 1992):

- The observations on the dependent variables follow a multivariate Normal distribution.
- The variance-covariance matrices for the dependent variables in each group are

equal.

- The non-diagonal terms in the variance-covariance matrix are zero.

Multivariate normality is a more exacting assumption than normality in ANOVA. A necessary condition for multivariate normality is that each variable must be distributed normally. Also, any linear combination of the variables and all subsets of the set of variables must be distributed normally. Various studies (eg, Hopkins & Clay 1963, Mardia 1971, Everitt 1979) suggest that for up to 10 variables and moderate sample sizes, deviation from multivariate normality has only a small effect on type I error. There are several techniques available for checking multivariate normality (see Gnanadesikan 1977) but they are difficult to implement. The assumption of homogeneity of covariance matrices is also very restrictive. This assumption can be checked by applying a Box test (Stevens 1992).

MANOVA provides an overall test of differences among treatment levels. It implicitly tests the composite of response variables that provides the strongest evidence of overall differences between treatment levels. To determine overall significance, statistical software provides test statistics such as Wilks' Λ, Pillai's criterion, Hotelling's trace, and Roy's maximum root criterion. These can be approximated by an F statistic. Six hypotheses will be tested. The first five are related to the five independent variables and the last one to the interaction terms.

$1H_0$: There is no significant overall effect of the five batch sizes on the composite measure of the four dependent variables, ie, $(\alpha_{11},\alpha_{12},\alpha_{13},\alpha_{14}) = (\alpha_{21},\alpha_{22},\alpha_{23},\alpha_{24}) = (\alpha_{31},\alpha_{32},\alpha_{33},\alpha_{34}) = (\alpha_{41},\alpha_{42},\alpha_{43},\alpha_{44}) = (\alpha_{51},\alpha_{52},\alpha_{53},\alpha_{54}) = 0$

$1H_1$: At least one $(\alpha_{i1},\alpha_{i2},\alpha_{i3},\alpha_{i4}) \neq 0 \ \forall i$

$2H_0$: There is no significant overall effect of the four scheduling rules on the composite measure of the four dependent variables, ie, $(\beta_{11},\beta_{12},\beta_{13},\beta_{14}) = (\beta_{21},\beta_{22},\beta_{23},\beta_{24}) = (\beta_{31},\beta_{32},\beta_{33},\beta_{34}) = (\beta_{41},\beta_{42},\beta_{43},\beta_{44}) = 0$

$2H_1$: At least one $(\beta_{j1},\beta_{j2},\beta_{j3},\beta_{j4}) \neq 0 \ \forall j$

$3H_0$: There is no significant overall effect of the two set-up times on the composite measure of the four dependent variables, ie, $(\gamma_{11},\gamma_{12},\gamma_{13},\gamma_{14}) = (\gamma_{21},\gamma_{22},\gamma_{23},\gamma_{24}) = 0$

$3H_1$: At least one $(\gamma_{k1},\gamma_{k2},\gamma_{k3},\gamma_{k4}) \neq 0 \; \forall k$

$4H_0$: There is no significant overall effect of the two levels of demand on the composite measure of the four dependent variables, ie, $(\delta_{11},\delta_{12},\delta_{13},\delta_{14}) = (\delta_{21},\delta_{22},-\delta_{23},\delta_{24}) = 0$

$4H_1$: At least one $(\delta_{l1},\delta_{l2},\delta_{l3},\delta_{l4}) \neq 0 \; \forall l$

$5H_0$: There is no significant overall effect of the two levels of processing time variability on the composite measure of the four dependent variables, ie, $(\zeta_{11},\zeta_{12},-\zeta_{13},\zeta_{14}) = (\zeta_{21},\zeta_{22},\zeta_{23},\zeta_{24}) = 0$

$5H_1$: At least one $(\zeta_{m1},\zeta_{m2},\zeta_{m3},\zeta_{m4}) \neq 0 \; \forall m$

$6H_0$: There is no significant overall effect of the two-factor interactions on the composite measure of the four dependent variables, ie,

$[(\alpha\beta)_{ij1},(\alpha\beta)_{ij2},(\alpha\beta)_{ij3},(\alpha\beta)_{ij4}] = 0$

$[(\alpha\gamma)_{ik1},(\alpha\gamma)_{ik2},(\alpha\gamma)_{ik3},(\alpha\gamma)_{ik4}] = 0$

$[(\alpha\delta)_{il1},(\alpha\delta)_{il2},(\alpha\delta)_{il3},(\alpha\delta)_{il4}] = 0$

$[(\alpha\zeta)_{im1},(\alpha\zeta)_{im2},(\alpha\zeta)_{im3},(\alpha\zeta)_{im4}] = 0$

$[(\beta\gamma)_{jk1},(\beta\gamma)_{jk2},(\beta\gamma)_{jk3},(\beta\gamma)_{jk4}] = 0$

$[(\beta\delta)_{jl1},(\beta\delta)_{jl2},(\beta\delta)_{jl3},(\beta\delta)_{jl4}] = 0$

$[(\beta\zeta)_{jm1},(\beta\zeta)_{jm2},(\beta\zeta)_{jm3},(\beta\zeta)_{jm4}] = 0$

$[(\gamma\delta)_{kl1},(\gamma\delta)_{kl2},(\gamma\delta)_{kl3},(\gamma\delta)_{kl4}] = 0$

$[(\gamma\zeta)_{km1},(\gamma\zeta)_{km2},(\gamma\zeta)_{km3},(\gamma\zeta)_{km4}] = 0$

$[(\delta\zeta)_{lm1},(\delta\zeta)_{lm2},(\delta\zeta)_{lm3},(\gamma\delta)_{lm4}] = 0 \quad \forall\, i,j,k,l,m$

$6H_1$: At least one of $[(\alpha\beta)_{ij1},(\alpha\beta)_{ij2},(\alpha\beta)_{ij3},(\alpha\beta)_{ij4}] \neq 0$

At least one of $[(\alpha\gamma)_{ik1},(\alpha\gamma)_{ik2},(\alpha\gamma)_{ik3},(\alpha\gamma)_{ik4}] \neq 0$

At least one of $[(\alpha\delta)_{il1},(\alpha\delta)_{il2},(\alpha\delta)_{il3},(\alpha\delta)_{il4}] \neq 0$

At least one of $[(\alpha\zeta)_{im1},(\alpha\zeta)_{im2},(\alpha\zeta)_{im3},(\alpha\zeta)_{im4}] \neq 0$

At least one of $[(\beta\gamma)_{jk1},(\beta\gamma)_{jk2},(\beta\gamma)_{jk3},(\beta\gamma)_{jk4}] \neq 0$

At least one of $[(\beta\delta)_{jl1},(\beta\delta)_{jl2},(\beta\delta)_{jl3},(\beta\delta)_{jl4}] \neq 0$

At least one of $[(\beta\zeta)_{jm1},(\beta\zeta)_{jm2},(\beta\zeta)_{jm3},(\beta\zeta)_{jm4}] \neq 0$

At least one of $[(\gamma\delta)_{kl1},(\gamma\delta)_{kl2},(\gamma\delta)_{kl3},(\gamma\delta)_{kl4}] \neq 0$

At least one of $[(\gamma\zeta)_{km1},(\gamma\zeta)_{km2},(\gamma\zeta)_{km3},(\gamma\zeta)_{km4}] \neq 0$

At least one of $[(\delta\zeta)_{lm1},(\delta\zeta)_{lm2},(\delta\zeta)_{lm3},(\gamma\delta)_{lm4}] \neq 0$

4.7 Univariate Hypotheses

Once MANOVA has been used to detect significant overall relationships, ANOVA will be used to assess specific factors and interactions that are significant.

ANOVA is based on the following assumptions (Stevens 1992):

Normality: The observations are distributed normally on the dependent variable in each group.

Homoscedasticity: The population variances for the groups are equal.

Independence: The observations are not correlated.

ANOVA is fairly immune to slight deviations from the above assumptions. Recall that the *nominal* α (level of significance) is set by the experimenter, and is the probability of type I error. It is the proportion of time one is rejecting H_0 falsely when all assumptions are met. The *actual* α is the proportion of time one is rejecting falsely if one or more assumptions are violated. The F statistic is quite robust with respect to the normality assumption, ie, actual $\alpha \approx$ nominal α. The reason is the Central Limit Theorem, which states that the sum of independent observations from any distribution approaches a Normal distribution as the number of observations increases. This also implies that the mean approaches normality, which is what the sampling distribution of F is based on. Wesolowsky (1976) stated that "In large samples lack of normality has no important consequences and in small samples it is difficult to prove."

There are many tests available for assessing normality. Among the graphical

ones are Normal probability plots and histograms. More rigorous tests include the χ^2 goodness of fit, Kolmogorov–Smirnov (KS), and the Shapiro–Wilk (SW) tests. The χ^2 test depends on the intervals used for grouping and it can be used for discrete or continuous data. The KS test has many advantages over the χ^2 test. The KS test works with small samples, no information is lost by grouping data into classes, and it is more powerful than the χ^2 test, ie, at a given α the KS test is less likely to accept a false H_0. However, the KS test applies only to continuous distributions. Shapiro *et al* (1968) showed that the KS test is not as powerful as the SW test. D'Agostino (1986) claimed that "For testing normality, the Kolmogorov–Smirnov test is only a historical curiosity. It should never be used ...".

When group sizes are equal, the F statistic is also robust against heterogeneous variances (Stevens 1992). There are tests available for establishing heteroscedasticity, eg, Spearman rank-correlation test. If heteroscedasticity is established by either of these tests, the appropriate solution is to apply a variance stabilizing transformation to the original model (Berenson *et al* 1983). The independence assumption is the most important one, for even a small violation substantially effects both the significance level and the power of the F test. Dependence among observations can be measured by the intraclass correlation R, where (Stevens 1992):

$$R = \frac{MS_b - MS_w}{MS_b + (n-1)MS_w} .$$

MS_b and MS_w are the numerator and denominator from the F statistic and n is the number of subjects per group.

The univariate hypotheses are based on conjectures proposed in the literature—they are aimed at examining operational issues that were mentioned in §1.2. The population mean of an effect is obtained by combining the means of all the configurations possible for that effect.

$7H_0$: There is no significant difference in population means of mean resource utilization when the system is configured with the five batch sizes, four scheduling

rules, two set-up times, two levels of demand, and two levels of processing time variability, ie, $\alpha_{i1} = \beta_{j1} = \gamma_{k1} = \delta_{l1} = \zeta_{m1} = 0, \forall\ i,j,k,l,m$

$7H_1$: At least one each of $\alpha_{i1}, \beta_{j1}, \gamma_{k1}, \delta_{l1}, \zeta_{m1} \neq 0$

$8H_0$: There is no significant difference in population means of mean resource utilization when the system is configured with any pair of batch size, scheduling rule, set-up time, demand, and processing time variability. In other words first order interactions are not significant, ie, $(\alpha\beta)_{ij1} = (\alpha\gamma)_{ik1} = (\alpha\delta)_{il1} = (\alpha\zeta)_{im1} = (\beta\gamma)_{jk1} = (\beta\delta)_{jl1} = (\beta\zeta)_{jm1} = (\gamma\delta)_{kl1} = (\gamma\zeta)_{km1} = (\delta\zeta)_{lm1} = 0\ \forall\ i,j,k,l,m$

$8H_1$: At least one each of $(\alpha\beta)_{ij1}, (\alpha\gamma)_{ik1}, (\alpha\delta)_{il1}, (\alpha\zeta)_{im1}, (\beta\gamma)_{jk1}, (\beta\delta)_{jl1}, (\beta\zeta)_{jm1}, (\gamma\delta)_{kl1}, (\gamma\zeta)_{km1}, (\delta\zeta)_{lm1} \neq 0$

$9H_0$: There is no significant difference in population means of mean flow time when the system is configured with the five batch sizes, four scheduling rules, two set-up times, two levels of demand, and two levels of processing time variability, ie, $\alpha_{i1} = \beta_{j1} = \gamma_{k1} = \delta_{l1} = \zeta_{m1} = 0, \forall\ i,j,k,l,m$

$9H_1$: At least one each of $\alpha_{i1}, \beta_{j1}, \gamma_{k1}, \delta_{l1}, \zeta_{m1} \neq 0$

$10H_0$: There is no significant difference in population means of mean flow time when the system is configured with any pair of batch size, scheduling rule, set-up time, demand, and processing time variability. In other words first order interactions are not significant, ie, $(\alpha\beta)_{ij2} = (\alpha\gamma)_{ik2} = (\alpha\delta)_{il2} = (\alpha\zeta)_{im2} = (\beta\gamma)_{jk2} = (\beta\delta)_{jl2} = (\beta\zeta)_{jm2} = (\gamma\delta)_{kl2} = (\gamma\zeta)_{km2} = (\delta\zeta)_{lm2} = 0\ \forall\ i,j,k,l,m$

$10H_1$: At least one each of $(\alpha\beta)_{ij2}, (\alpha\gamma)_{ik2}, (\alpha\delta)_{il2}, (\alpha\zeta)_{im2}, (\beta\gamma)_{jk2}, (\beta\delta)_{jl2}, (\beta\zeta)_{jm2}, (\gamma\delta)_{kl2}, (\gamma\zeta)_{km2}, (\delta\zeta)_{lm2} \neq 0$

$11H_0$: There is no significant difference in population means of mean flow time CV when the system is configured with the five batch sizes, four scheduling rules, two set-up times, two levels of demand, and two levels of processing time variability, ie, $\alpha_{i3} = \beta_{j3} = \gamma_{k3} = \delta_{l3} = \zeta_{m3} = 0, \forall\ i,j,k,l,m$

$11H_1$: At least one each of α_{i3}, β_{j3}, γ_{k3}, δ_{l3}, $\zeta_{m3} \neq 0$

$12H_0$: There is no significant difference in population means of mean flow time CV when the system is configured with any pair of batch size, scheduling rule, set-up time, demand, and processing time variability. In other words first order interactions are not significant, ie, $(\alpha\beta)_{ij3} = (\alpha\gamma)_{ik3} = (\alpha\delta)_{il3} = (\alpha\zeta)_{im3} = (\beta\gamma)_{jk3} = (\beta\delta)_{jl3} = (\beta\zeta)_{jm3} = (\gamma\delta)_{kl3} = (\gamma\zeta)_{km3} = (\delta\zeta)_{lm3} = 0 \; \forall \; i,j,k,l,m$

$12H_1$: At least one each of $(\alpha\beta)_{ij3}$, $(\alpha\gamma)_{ik3}$, $(\alpha\delta)_{il3}$, $(\alpha\zeta)_{im3}$, $(\beta\gamma)_{jk3}$, $(\beta\delta)_{jl3}$, $(\beta\zeta)_{jm3}$, $(\gamma\delta)_{kl3}$, $(\gamma\zeta)_{km3}$, $(\delta\zeta)_{lm3} \neq 0$

$13H_0$: There is no significant difference in population means of mean percent tardy batches when the system is configured with the five batch sizes, four scheduling rules, two set-up times, two levels of demand, and two levels of processing time variability, ie, $\alpha_{i4} = \beta_{j4} = \gamma_{k4} = \delta_{l4} = \zeta_{m4} = 0, \; \forall \; i,j,k,l,m$

$13H_1$: At least one each of α_{i4}, β_{j4}, γ_{k4}, δ_{l4}, $\zeta_{m4} \neq 0$

$14H_0$: There is no significant difference in population means of mean percent tardy batches when the system is configured with any pair of batch size, scheduling rule, set-up time, demand, and processing time variability. In other words first order interactions are not significant, ie, $(\alpha\beta)_{ij4} = (\alpha\gamma)_{ik4} = (\alpha\delta)_{il4} = (\alpha\zeta)_{im4} = (\beta\gamma)_{jk4} = (\beta\delta)_{jl4} = (\beta\zeta)_{jm4} = (\gamma\delta)_{kl4} = (\gamma\zeta)_{km4} = (\delta\zeta)_{lm4} = 0 \; \forall \; i,j,k,l,m$

$14H_1$: At least one each of $(\alpha\beta)_{ij4}$, $(\alpha\gamma)_{ik4}$, $(\alpha\delta)_{il4}$, $(\alpha\zeta)_{im4}$, $(\beta\gamma)_{jk4}$, $(\beta\delta)_{jl4}$, $(\beta\zeta)_{jm4}$, $(\gamma\delta)_{kl4}$, $(\gamma\zeta)_{km4}$, $(\delta\zeta)_{lm4} \neq 0$

4.8 Multiple Regression

Multiple regression will be used to further explore the factors and interactions that are declared significant by ANOVA. In multiple regression we are interested in predicting a dependent variable from a set of predictors.

The assumptions necessary for regression analysis are analogous to those of

ANOVA. To assess the validity of the linear model it is assumed that the error term ϵ is distributed normally with mean zero, and the error terms are independent. The homoscedasticity assumption implies that the variation about the regression line is constant for all values of the independent variable. There is the added assumption of linearity in linear regression, ie, there is a linear relationship between the dependent variable and each independent variable. Residual plots can be used to assess violations against the assumptions underlying the regression model.

Autocorrelation occurs when the error terms are correlated. It is detected easily by plotting residuals against time. A regular time pattern shows autocorrelation. Possible causes are omitted explanatory variables and a mis-specified model. Traditional tests include the Durbin–Watson test (Durbin & Watson 1950,1951) for small samples. Another common problem in regression is that of multicollinearity, ie, linear relationships among explanatory variables. An obvious indicator is the correlation matrix calculated from the set of explanatory variables. Other indicators are reversed signs of certain regression coefficients and large standard errors. If the explanatory variables are correlated, their order of entry can make a significant difference toward the variance on y. Multicollinearity can limit R severely and prevent determining the importance of individual predictors since they are confounded due to correlation. Koutsoyiannis (1977) suggested some solutions for autocorrelation and multicollinearity.

Regression analysis is more powerful than ANOVA—it gives all the information that ANOVA does *and* it provides numerical estimates for the influence of each independent variable (Koutsoyiannis 1977). The general linear model with first order interactions is:

$$y = \beta_0 + \sum_i \beta_i x_i + \sum_{\substack{i \\ i<j}} \sum_j \beta_{ij} x_i x_j + \varepsilon_i \ ,$$

where β_0, β_i, and β_{ij} are the parameters to be estimated and ϵ is the prediction error. In general, the linear combination of the x_i that is maximally correlated with y is sought. In this research, there are four variables to be predicted (resource utilization,

flow time, flow time CV, and percent tardy batches) and there are five predictor variables (batch size, scheduling rule, set-up time, demand, and processing time variability).

Resource utilization, *Flow time*, *Flow time CV*, *Percent tardy batches* $= \beta_0 + \beta_1$ *Batch size*$_{1,2,3,4,5} + \beta_2$ *Scheduling rule*$_{1,2,3,4} + \beta_3$ *Set-up time*$_{1,2} + \beta_4$ *Demand*$_{1,2} + \beta_5$ *Processing time variability*$_{1,2}$

Equivalently (with first order interactions):

$$y = \beta_0 + \beta_1 x_1 + \beta_{2i} x_{2i} + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12i} x_1 x_{2i} + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{15} x_1 x_5 + \beta_{2i3} x_{2i} x_3 + \beta_{2i4} x_{2i} x_4 + \beta_{2i5} x_{2i} x_5 + \beta_{34} x_3 x_4 + \beta_{35} x_3 x_5 + \beta_{45} x_4 x_5 + \epsilon;\ i = 1,2,3$$

where

x_1 = batch size,

x_{21} = 1 if work release is SPT, 0 otherwise (FIFO),

x_{22} = 1 if work release is LPT, 0 otherwise,

x_{23} = 1 if work release is EDD, 0 otherwise,

x_3 = set-up time,

x_4 = demand, and

x_5 = processing time variability.

4.9 Summary

The first step in experimental design is to conduct pilot runs of the simulation and check for face validity of the model and its assumptions. The simulation is non-terminating since we are interested in steady state parameters like the mean. Common random numbers are used in an attempt to contain variance. The issues of initialization bias and autocorrelation will be examined by using suitable techniques

like discarding output during the warmup period and batching subsequent runs. Following verification and validation by QNA, the 160 configurations will be replicated 100 times each. The simulation output will be examined for the assumptions of ANOVA, MANOVA, and Regression. The techniques mentioned earlier can be used to transform the data to satisfy these assumptions. Statistical analysis will consist of conducting MANOVA tests. Following a rejection of H_0 (equal population means), ANOVA will be used to test individual relationships. The significant relationships will be explored further by multiple regression. These relationships will be used to form recommendations for batch manufacturing systems. In the next chapter, we present the queuing network model and a qualitative analysis of the results.

Chapter 5
Analysis

The analysis phase is normally used to compare alternative configurations of a system. The analysis that is possible depends on the decisions made in the design stage. We are interested in evaluating the performance of different batching policies under given configurations of a manufacturing facility with flow lines and job shops. The factors that lend to the various configurations range from ones that are regulated easily, such as batch size and scheduling rule, to ones that are not under immediate control, such as set-up time and processing time variability. Further, factors such as demand are completely exogenous. Response variables are resource utilization, mean and CV of the flow time of a batch, and percent tardy batches. We describe the simulation model, formulate the network model, validate the simulation model, and analyze the results.

5.1 Simulation Model

The simulation models are written in GPSS/H (Banks *et al* 1989, Schriber 1974,1991). The discrete-event nature and linear flow of the manufacturing system lend themselves well to a process driven approach. Batches arrive according to a Poisson process. Each batch is assigned parameters corresponding to set-up and processing time at each node. The arrival time of the batch is recorded and the due date is assigned. Depending on the scheduling rule, the batches are rearranged and then released for processing. Prior to service, the server goes through set-up during which time it is unavailable. Following set-up the batch is processed without preemption. After service, the batch advances to the next server and signals the waiting batch(es) that the server is free. Before leaving the system the batch updates statistics such as

the flow time, tardiness, and total number of batches processed.

Each configuration is replicated 100 times. The system reaches steady state during the first 36 hours. The data for the first 48 hours are discarded to control initialization bias. Each replication lasts 5 days of 24 hours each. The long run period and large number of replications ensure high (statistical) power and low variance in the estimates of the performance measures.

Five configurations were simulated. The most elementary system is a flow line node (F) followed by a job shop node (J). Somewhat more involved is the system with these nodes in reverse order, ie, J-F. This system is more complex because it is difficult to predict the behavior of a job shop and the uncertainty accumulates along the system. The third system simulated was constructed by adding a flow line node to the previous systems. This system can be viewed as F-(J-F) or (F-J)-F, ie, a flow line node preceding J-F or succeeding F-J. The fourth system evaluated was constructed by adding a job shop node instead. This system can be considered as (J-F)-J or J-(F-J). The final configuration examined was F-J-F-J-F.

The gradual construction of the systems allowed an understanding of the interface between a flow line and job shop and the interactions among the dependent and independent variables. Further experimentation was deemed unnecessary, especially in view of the computing time overhead. The simulation model was verified by sensitivity analysis and interactive tracing. In the following we present the network model and validate the simulation using the analytical results.

5.2 Network Model

The analytical model is based on queuing networks and uses the Queuing Network Analyzer (QNA) developed by Whitt (1983a). QNA takes a decomposition approach toward analyzing networks. It captures the dependence among nodes and then decomposes the network into individual nodes. Each node is then analyzed as a separate G/G/m queue that is characterized by the first two moments of the arrival and service time distributions. Performance measures for the entire network are obtained by as-

suming as an approximation that the nodes are stochastically independent.

For each network the user specifies the number of nodes and the number of servers at each node. The arrival and service processes at each node are characterized by the mean and variability parameters. Finally, a routing matrix indicates the proportion of customers that go to node j from node i. The input data is as follows:

n = number of nodes in the network

m_j = number of servers at node j

λ_{0j} = external arrival rate to node j

c_{0j}^2 = variability parameter of the external arrival process

τ_j = mean service time at node j

c_{sj}^2 = variability parameter of the service time distribution

q_{ij} = proportion of customers completing service at node i and going to node j

Total arrival rate to node j, $\lambda_j = \lambda_{0j} + \sum_1^n \lambda_i q_{ij}$

Utilization of node i, $\rho_i = \lambda_i \tau_i / m_i$

Arrival rate to node j from node i, $\lambda_{ij} = \lambda_i q_{ij}$

Proportion of arrivals to j from i $(i \geq 0)$, $p_{ij} = \lambda_{ij} / \lambda_j$

The most important step is the system of equations that compute variability parameters of the internal flows (c_{aj}^2) and thus capture the dependency among the nodes before decomposing the network and treating each node independently.

$$c_{aj}^2 = a_j + \sum_1^n c_{ai}^2 b_{ij} \ ,$$

where

$$a_j = 1 + w_j [p_{0j} c_{0j}^2 - 1 + \sum_1^n p_{ij} (1 - q_{ij} + q_{ij} \rho_i^2 x_i)] \ , \text{ and}$$

$$b_{ij} = w_j p_{ij} q_{ij} (1-\rho_i^2) \ .$$

The variables x_j and w_j are included to ease modification of the algorithm.

$$x_j = 1+\frac{\max(c_{si}^2,.2)-1}{\sqrt{m_i}} \ , \text{ and}$$

$$w_j = \frac{1}{1+4(1-\rho_j)^2(v_j-1)} \ , \text{ where } v_j = \frac{1}{\sum_0^n p_{ij}^2} \ .$$

At this point we have decomposed the network into individual nodes. We can now treat each node as a separate G/G/m queue and calculate the congestion measures for the node. Each node is characterized by the number of servers and the rate and variability parameters of the arrival and service processes.

The steady state waiting time (before service) in a G/G/1 queue,

$$E[W] = \frac{\tau\rho(c_a^2+c_s^2)g}{2(1-\rho)} \ ,$$

where

$$g = \begin{cases} e^{\frac{2(\rho-1)(1-c_a^2)^2}{3\rho(c_a^2+c_s^2)}} & c_a^2<1 \\ 1 & c_a^2\geq 1 \end{cases} \ .$$

The probability of delay,

$$\sigma = \rho+(c_a^2-1)\rho(1-\rho)h \ ,$$

where

$$h = \begin{cases} \dfrac{1+c_a^2+\rho c_s^2}{1+\rho(c_s^2-1)+\rho^2(4c_a^2+c_s^2)} & c_a^2 \le 1 \\ \dfrac{4\rho}{c_a^2+\rho^2(4c_a^2+c_s^2)} & c_a^2 \ge 1 \end{cases} .$$

The variability parameter of the waiting time,

$$c_W^2 = \frac{c_D^2+1-\sigma}{\sigma} ,$$

and

$$V(W) = E[W]^2 c_W^2 .$$

We can now calculate the performance measures for the entire network.

Mean time spent by a customer at node i, $E[T_i] = \tau_i + E[W_i]$.

Total time spent by the customer in the network (flow time), $E[T] = \sum_1^n E[T_i]$.

Thus, $V(T_i) = V(W_i) + \tau_i^2 c_{si}^2$ and $V(T) = \sum_1^n V(T_i)$.

The details of the algorithm and some extensions (eg, merging and splitting) that are not relevant to this study have been omitted. However, the algorithm can be modified easily to handle these extensions. In the following, we formulate the model for a case of two nodes. The intent is to find approximations for the performance measures. It is possible to find closed form results for larger systems but the expressions are substantially more onerous.

$$n = 2$$

$$m = (1\ 1)$$

$$\lambda_0 = (\lambda_{01}\ 0)$$

$$c_0^2 = (1\ 0)$$

$$\tau = (\tau_1\ \tau_2)$$

$$c_s^2 = (c_{s1}^2\ c_{s2}^2)$$

$$q = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\lambda = (\lambda_{01}\ \lambda_{01})$$

$$\rho = (\rho_1\ \rho_2)$$

$$p = \begin{bmatrix} 1 & 0 \\ 0 & \lambda_{01}/\lambda_2 \\ 0 & 0 \end{bmatrix}$$

$$v = w = (1\ 1)$$

$$a = (1\ \rho_1^2 c_{s1}^2)$$

$$b = \begin{bmatrix} 0 & 1-\rho_1^2 \\ 0 & 0 \end{bmatrix}$$

$$c_a^2 = (1\ 1-\rho_1^2+\rho_1^2 c_{s1}^2)$$

$$E[W] = \left[\frac{\tau_1\rho_1(1+c_{s1}^2)g}{2-2\rho_1} \quad \frac{\tau_2\rho_2(1-\rho_1^2+\rho_1^2 c_{s1}^2+c_{s2}^2)g}{2-2\rho_2} \right]$$

$$\sigma = (\rho_1 \quad \rho_2+(\rho_1^2 c_{s1}^2-\rho_1^2)\rho_2(1-\rho_2)h)$$

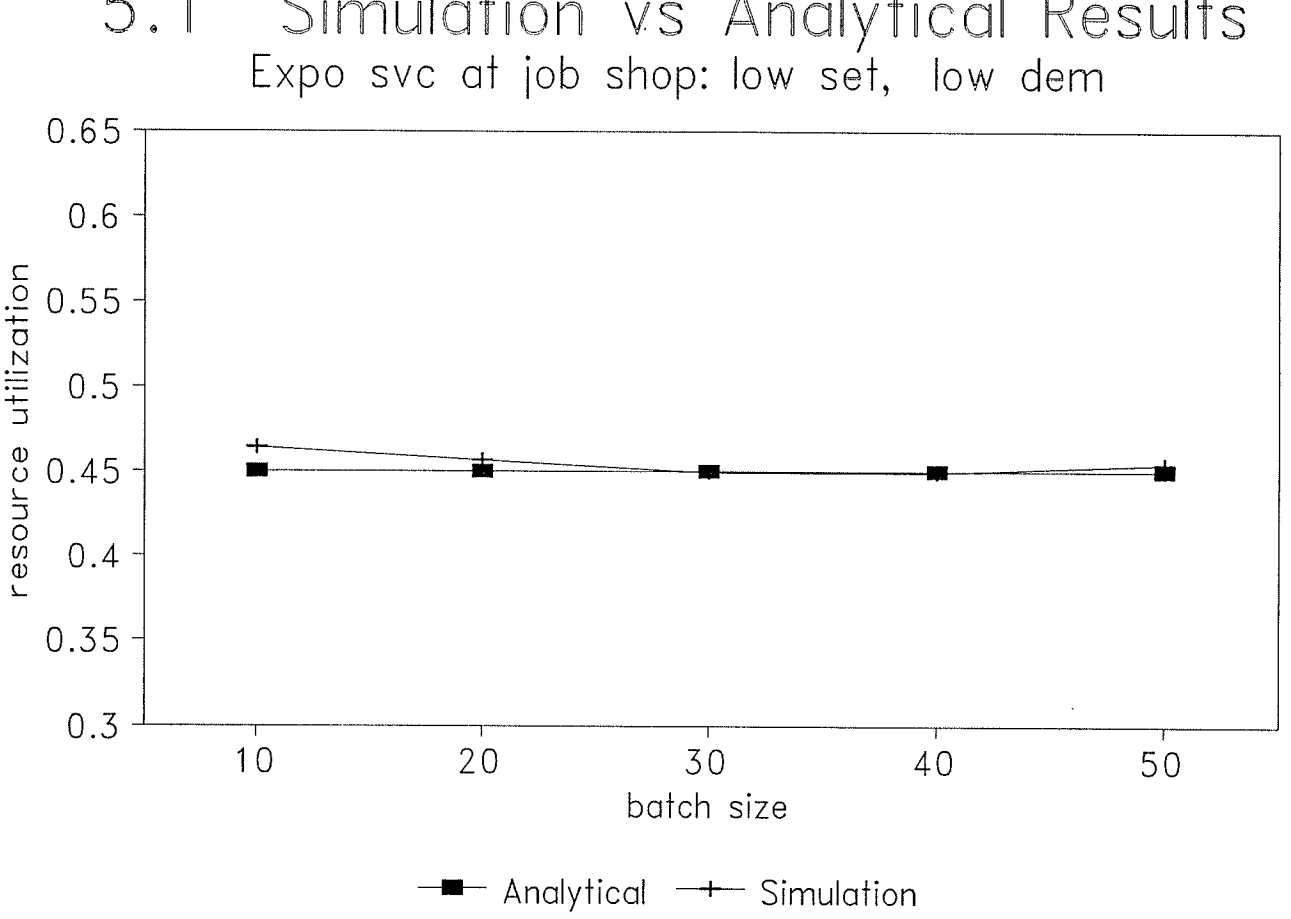
$$c_W^2 = \left[\frac{2-\rho_1}{\rho_1} \quad \frac{2-\rho_1^2+\rho_1^2 c_{s1}^2-\rho_2-(\rho_1^2 c_{s1}^2-\rho_1^2)\rho_2(1-\rho_2)h}{\rho_2+(\rho_1^2 c_{s1}^2-\rho_1^2)\rho_2(1-\rho_2)h} \right]$$

The expressions for $V(W)$, $E[T]$, and $V(T)$ can be computed easily from the above parameters. These expressions are unwieldy and will be omitted.

QNA was coded in Fortran to obtain numerical results. The analytical results compared favorably with the simulation output as shown in graph 5.1. QNA gives approximate results because the internal arrival processes are usually not renewal processes. QNA also assumes that the system is in steady state. Additionally, the reliability of the approximations decreases when c_a^2 increases. Thus, while the utilization estimates from QNA are reliable, the flow time results are not. Jackman and Johnson (1993) found similar results in evaluating queuing network models.

We now analyze the simulation results. The discussion will be segregated by the service time distribution of the job shop node. Within this taxonomy, we will categorize by demand, set-up time, and processing time variability. All five configurations will be discussed individually with respect to the four performance measures. Further, the effect of the four scheduling rules will be examined. Thus, we will observe the following hierarchy for analyzing the results: processing time distribution of the job shop node (Exponential, Normal), demand (low, high), set-up time (low, high), processing time variability (low, high), system configuration (F-J, J-F, F-J-F, J-F-J, F-J-F-J-F), performance measure (resource utilization, flow time, flow time coefficient of variance, percent tardy batches), scheduling rule (FIFO, SPT, LPT, EDD), and batch size (10, 20, 30, 40, 50). The graphs are in the Appendix. The page numbers of the graphs are indicated next to the corresponding sections. The unit of flow time is minutes.

5.1 Simulation vs Analytical Results



5.3 Exponential Service

This section corresponds to Exponential service at the job shop node. The mean service time is two minutes/job.

Low demand, low set-up (p115): The first configuration tested was F-J. Resource utilization (RU) is approximately the same for all scheduling rules. This is because roughly the same amount of work is processed under all scheduling rules. Flow time (FT) increases with batch size because large batches incur more processing time. Flow time CV (FTCV) decreases with batch size because large batches reduce variability by optimizing on set-ups. The results for percentage of jobs that are tardy (PT) are similar to those of flow time.

The second configuration tested was J-F. RU in J-F is somewhat lower than in F-J. A job shop is usually the bottleneck in a system. When a job shop is the first node in a system it delays work for the rest of the system thus lowering utilization. FT is also higher in comparison with F-J for the same reason. FT is highest under LPT and lowest under SPT. Scheduling rules are effective in J-F because the first node is a bottleneck. FTCV is lowest for FIFO and highest for SPT. This is because under FIFO batches are released to the system without reordering whereas the scheduling rules reorder the queue as new batches arrive, thus adding variability to the FT. This effect is compounded in J-F as the bottleneck at the beginning of the facility results in a longer queue as compared to F-J. Thus, FTCV is lower for FIFO in J-F than in F-J and higher for scheduling rules. PT first decreases with batch size and then increases. The decrease is probably due to set-ups whose effect diminishes after a certain batch size (cf §1.2). PT is higher than in F-J because of increased FT.

The third configuration tested was F-J-F. RU is lower than in F-J or J-F because the job shop node restricts work for the following flow line node and RU is averaged over all the nodes. FT is higher than in F-J but lower than in J-F. FTCV is lower than in F-J because of the increased FT. Similarly PT is slightly higher. The

graphs for F-J-F are similar to F-J, and not J-F. This demonstrates that F-J-F is similar to (F-J)-F and not F-(J-F), ie, a flow line node succeeding a manufacturing facility causes less variation in a system than one preceding. This observation can be important in gaging the performance of design changes to the system.

The fourth system tested was J-F-J. As compared with previous configurations, RU, FT, and PT are higher because of the added node. FTCV is slightly lower because of increased FT. Note that, J-F-J behaves like (J-F)-J and not J-(F-J), ie, a job shop node added to the end of a manufacturing line causes less variability.

The last system tested was F-J-F-J-F. This system is similar in behavior to F-J and F-J-F. It is apparent that (i) the first node in the system and the scheduling rule have a major affect on the performance measures and (ii) adding nodes to the end of a system as opposed to the beginning of a system effects less change in the system. Thus, the behavior of large systems can be predicted by gradually adding nodes to the tail of the system even though the physical configuration may require new nodes at the beginning or the middle of the system.

Low demand, high set-up (p119): For systems starting with a flow line, high set-up increases FT due to longer queues. The other performance measures are similar in behavior to the corresponding configurations in the low demand case. This is to be expected as set-up time is a variable for the job shop nodes only—it is fixed for the flow line nodes over all configurations. However, systems with a job shop as the first node perform differently. RU increases significantly for batches of size 20 and then stabilizes. This is because large batches use resources more efficiently by decreasing set-ups. RU also starts decreasing for batches greater than size 30 because of the increased wait for the flow line. FT decreases for batches of size 20 for the same reason. For small batches, FT is lower under LPT than FIFO and EDD because of set-ups. FTCV is still lowest under FIFO. PT follows the same trend as FT but is higher significantly than in the previous case.

High demand, low set-up (p123): For systems with a flow line as the first node, RU,

FT, and FTCV are higher than in the previous configurations. PT is now affected by the scheduling rules with SPT minimizing tardiness. The other performance measures exhibit similar trends as in the previous configurations. Systems with a job shop as the first node behave as in the case above. RU and FT are higher whereas FTCV and PT are roughly the same.

High demand, high set-up (p127): High demand coupled with high set-up further worsens the performance measures, which begin to demonstrate definite trends. RU decreases and FT and PT increase. There is marginal change in FTCV. In systems starting with a flow line, FT increases with batch size while FTCV and PT decrease. In systems statring with a job shop, FT, FTCV, and PT decrease with batch size. Batches of size 40 maximize RU under all scheduling rules.

It seems that demand affects systems that have a flow line as the first node. On the other hand, set-up time affects systems with a job shop as the first node. This effect is measured by the change in the values of the performance measures. A possible reason is that demand may also be affecting the job shops, but this effect is masked by the set-up time, which is not a factor for flow lines.

5.4 Normal Service

This section corresponds to Normal service times at the job shop node. An additional variable in this case is the processing time variability, which is given by the standard deviation of the Normal distribution. Note that the coefficient of variation of a Nor mal distribution is less than that of an Exponential distribution.

Low demand, low set-up, low variability (p131): In systems with a flow line as the first node, RU is almost identical for all batch sizes and scheduling rules. FT and PT increase linearly with batch size while FTCV decreases. With job shop as the first node, RU is again approximately the same under all scheduling rules. FT is mini-

mized by batches of size 20 under SPT probably due to set-up. Large batches favor the other performance measures.

Low demand, low set-up, high variability (p135): The performance measures for the high variability case do not seem significantly different, especially when a flow line is the first node. In the other systems, FT and PT are slightly higher.

Low demand, high set-up, low variability (p139): As in the Exponential case, increasing set-up does not have a significant effect on systems starting with a flow line. When a job shop is the first node, RU and FT are (again) optimal for large batches because of reduced set-ups. FT and PT are significantly higher than in the low set-up case, because of frequent set-ups. Compared with the low set-up case, FT and PT are much higher for small batches.

Low demand, high set-up, high variability (p143): As expected, increasing PTV of the job shop does not have a significant effect on the performance measures. Compared with the low set-up case, FT and PT are much higher for small batches, but still minimized by SPT. FTCV is higher for SPT and LPT and roughly the same for FIFO and EDD. This is probably because of more frequent reordering of the queue before the first node (set-up is added to the processing time for scheduling).

High demand, low set-up, low variability (p147): In systems with a flow line as the first node, RU is similar for all batch sizes and scheduling rules. FT increases with batch size while FTCV decreases. PT is minimized by batches of size 20. In the other systems, RU is maximized by batches of size 40 under LPT. The other performance measures are also optimal for large batches. Compared with the low demand case, the values of all performance measures are higher especially for small batches. As before, demand has a significant effect on systems starting with a flow line.

High demand, low set-up, high variability (p151): Increasing PTV does not affect

the performance measures significantly. In comparison with the low demand case, all performance measures are higher in value.

High demand, high set-up, low variability (p155): Systems beginning with a flow line have similar RU under all configurations. FT and PT are minimum for batch size 20 while FTCV decreases with batch size. When a job shop is the first node, batches of size at least 30 are optimal for all performance measures. Compared with the low demand and low set-up cases, the current system has higher values for all performance measures.

High set-up, high demand, high variability (p159): Once again, PTV has only a marginal effect on the performance measures. An increase in demand or set-up does affect the performance measures—demand influences systems beginning with a flow line and set-up affects systems beginning with a job shop. While both factors influence both kinds of systems, the impact of one factor may be masked by the other.

5.5 Statistical Analysis

We will review some statistical concepts common to ANOVA, MANOVA, and Regression before presenting the statistical analysis. Most of this section draws on Stevens (1992).

5.5.1 ANOVA / MANOVA

The essence of ANOVA is hypothesis testing. Hypothesis testing uses the mean of a sample population to test whether the underlying population mean is different. However, a sampling distribution is an approximation of the population distribution. Thus, there is always a probability of making an error when testing the null hypothesis (H_0: equal population means), which is based on the sampling distribution. This may result in rejection of the null hypothesis when it is actually true. Therefore, one must decide the acceptable risk of making this error. It is desirable to make this

risk small and 5% risk is an acceptable level. Formally, this is equivalent to setting the level of significance (α) at .05, ie, we are willing to take a 5% chance of making a *type I error*, which is the probability of rejecting the null hypothesis when it is true. Type I error is equivalent to saying that the group means differ when they don't.

Type II error (β) is another error associated with conducting statistical tests. This is the probability of accepting H_0 when it is false, ie, saying that population means don't differ when they do. Both type I and type II errors can occur and they are related inversely. Therefore, it is important to maintain a balance between the two. A related issue is that of *power*. The power of a statistical test $(1-\beta)$ is the probability of rejecting H_0 when it is false. Thus, power is the probability of making a correct decision. The power of a statistical test is related directly to: (i) the α level, (ii) the sample size, and (iii) the effect size, which is the magnitude effect of the treatments on the response variables. Power is not an issue for large sample sizes (≥ 100).

ANOVA entails multiple statistical tests and the α level attains much more significance in such cases. *Overall* α for a set of tests is the probability of at least one false rejection when H_0 is true. According to the *Bonferroni Inequality*, if k hypotheses are being tested at α' then overall $\alpha \leq k\alpha'$. If the tests are independent, then overall $\alpha = 1 - (1-\alpha')^k \approx k\alpha'$. Thus, some significant results could actually be type I errors, ie, incorrect results. For example, when conducting a 4-way ANOVA (ABCD), 15 tests are being performed, one for each effect (A, B, C, D, AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD, ABCD). If each effect is tested at the 5% level, then overall $\alpha \approx .54$, ie, there is roughly a 50% probability of making an error.

As mentioned above, *effect size* affects the power of a statistical test. Effect size is the difference a treatment makes on a response variable. Cohen (1977) developed a measure of effect size for the univariate t test. It indicates how many standard deviations the groups differ by. The measure for two groups, $\hat{d} = (\bar{x}_1 - \bar{x}_2)/s$. An effect size of roughly .2 is small, .5 is medium, and over .8 is large. Effect size is

important because it is independent of the sample size unlike the F test. Like most statistical tests, the F test will usually indicate significance given a large sample. However, the difference may not have practical significance (sufficient difference between the means of the response variables). While there is no consensus on a measure of the effect size, a properly designed experiment can contribute toward large effects. Practical significance can also be gaged by examining the means of the response variables.

Finally, in statistical analysis, anomalous data points called *outliers* are common. Outliers are data points very different from the rest. It is important to identify outliers because we want the analysis to reflect majority of the data and not the deviant data points.

5.5.2 Regression

Simple regression attempts to predict a dependent variable from an independent variable. In multiple regression we are interested in predicting a dependent variable from multiple independent variables. We are seeking a combination of predictors that is correlated highly with the dependent variable, not correlated mutually, and capable of explaining a high proportion of variance of the dependent variable. The measure of correlation between the observed and predicted values is called R. Sample size (n) and the number of predictors (k) can determine the generalizability of a regression equation. In general, an n/k ratio (number of subjects/predictor) of 15 or more is desirable.

Evidently, correlation between predictors can limit R severely. This situation is called *multicollinearity*. It can be diagnosed by examining the correlation matrix. More rigorously, the *variance inflation factor* (VIF) for a predictor indicates the strength of the linear association between itself and the other predictors. Myers (1990) remarked that a VIF over 10 is cause for concern. The simplest way of combating multicollinearity is to combine predictors that are correlated highly. Alternatively, all but one of the correlated predictors can be dropped as long as the model is not underspecified.

Another problem in multiple regression is having not enough (underfitting) or too many (overfitting) predictors in a model. Mallow (1973) introduced a measure called Mallow's C_k that can help prevent under/overfitting. It is recommended that for a good model, $C_k \approx k$.

Validation of a regression model indicates how well a regression equation will predict on an independent sample of data. Model validation can be established by data splitting, computing an *adjusted* R^2, and using the *Press* statistic. In data splitting, the sample is split in half. The regression equation obtained from one sample is validated on the other. Adjusted R^2 measures the loss in predictive power. The most commonly used formula is by Wherry and it estimates the amount of variance in y that would be accounted for had the prediction equation been derived from the population from which the sample was drawn. The Press statistic, $R^2_{Press} = 1 - Press/\Sigma(y_i - \hat{y}_i)$. It is a measure of the predictive power of the regression equation on independent data samples. The prediction error for each subject is computed for the regression equation derived from the remaining subjects. Thus, the statistic comprises n validations, each based on $n-1$ samples.

There are no rigorous methods for determining sample size and effect size in k-group MANOVA. Laüter (1978) provided tables for determining the sample size per group required for given values of α and power. The tables assume a knowledge of the effect size. The value of α chosen for this study is .01 for reasons mentioned above. Thus, if the effect size is small and the desired power is .9 then according to the tables the sample size needed is 240 for the 4-group–4-variable case and 260 for the 5-group–4-variable case. However, recall that a reliable regression equation requires at least 15 subjects/predictor. The individual levels of the variables in this study yield up to 15 predictors. This means that we need a sample size of at least 225 for the 5-group case. The sample size in this study is 8000 for the case of Exponential service at the job shop and 16000 for Normal service.

There are four dependent variables (regressands, response variables, variables): resource utilization (RU), flow time (FT), flow time coefficient of variance

(explanatory variables, groups, predictors, regressors, treatments) for Exponential service: batch size (BS), scheduling rule (SR), set-up time (ST), and demand (Dem). Additionally, we are examining processing time variability (PTV) for Normal service at the job shop. There are five levels of BS (10, 20, 30, 40, 50), four levels of SR (FIFO, SPT, LPT, EDD), and two levels each of ST, Dem, and PTV (low, high). Thus there are 80 treatment combinations in the Exponential case and 160 in the Normal case with 100 replications per combination.

5.5.3 Results

Statistical analysis was performed in SAS 6.08 on an Amdahl 5890. Considering that any model drawn on a sample population is only an approximation, our interest is not in the figures provided by SAS—we will be focusing on the results of hypothesis testing.

The simulation output was first verified for the assumptions of ANOVA, ie, (i) the observations are distributed normally on the dependent variable in each group, (ii) the population variances for the groups are equal, and (iii) the observations are independent. In some cases the observations were correlated as indicated by the intraclass correlation (cf §4.7) and the Durbin–Watson statistic. As mentioned earlier, any violation in the independence assumption affects the level of significance and the power of the F statistic. This means that the actual α can be substantially greater that the nominal α. Scariano & Davenport (1987) showed that dependence has a significant effect on type I error. The data were transformed using the *Cochrane–Orcutt* procedure (Cochrane & Orcutt 1949), which removed most of the first-order autocorrelation. The data for some variables were also heteroscedastic, which was corrected by transformations. These transformations are reported in the regression equations.

The output was then considered for the corresponding three MANOVA assumptions. There is no convenient method of assessing multivariate normality. However, the presence of univariate normality of the observations on each variable is a necessary condition for multivariate normality and is a good indicator of multivariate normality (Gnanadesikan 1977). As stated in §4.6, the assumption of homogenous

covariance matrices is very restrictive. It implies that not only are the variances of the dependent variables equal but also their covariances. Contrast this with a univariate t test which requires equal variances for one variable only. It is unlikely that equal variance-covariance matrix assumption would be satisfied literally in practice. Holloway & Dunn (1967) and Olson (1974) conducted Monte Carlo studies to examine the effect of unequal covariance matrices on α and found that equal group sizes keep the nominal α close to actual α except for extreme cases. Because the group sizes are equal in this study and there is no convenient way to verify multivariate normality and heteroscedasticity, the data were not assessed for these assumptions. The third assumption is that of the independence of observations, which was verified in the ANOVA assumptions.

There are three assumptions in multiple regression—independence and normality of error terms with constant variance. The normality assumption was verified by residual plots and the independence assumption was examined in the assumptions for ANOVA. There are also the implicit assumptions of additivity and a linear relationship between the dependent and the independent variables. Finally, outliers were not removed from the datasets because of the enormity of data and our experience that unforseen occurrences are common in a manufacturing environment.

The first set of hypotheses relate to multivariate tests. Hypothesis 1*H* tests whether a simultaneous comparison of the four dependent variables at each of the five levels of BS shows a significant overall effect. Similarly, hypotheses 2*H*-5*H* relate to: SR (2*H*), ST (3*H*), Dem (4*H*), and PTV (5*H*). Hypothesis 6*H* tests whether a simultaneous comparison of the four dependent variables at each level of the first-order interactions shows a significant overall effect. An interaction means that the effect an independent variable has on a dependent variable is not the same for all levels of the other independent variables (Stevens 1992). For brevity, the following results apply to the configuration F-J-F-J-F.

The multivariate hypotheses 1*H*-5*H* were rejected in both the Exponential and Normal cases indicating that the independent variables have a statistically significant effect on the composite of the dependent variables. Hypothesis 6*H* could not be

rejected because the BS×SR, SR×ST, and SR×Dem interactions were not significant in the Exponential case and the SR×ST, SR×PTV, and ST×PTV interactions in the Normal case. Therefore, the corresponding univariate hypotheses were not tested.

Corresponding to the multivariate hypotheses, two types of univariate hypotheses are defined to detect significant main effects and first-order interactions. The population mean of an effect is obtained by combining the means of all configurations possible for that effect. Hypothesis 7*H* tests for a significant difference in population means of mean RU when the system is configured with the five levels of BS, four levels of SR, two levels of ST, two levels of Dem, and two levels of PTV. Hypothesis 8*H* tests for a significant difference in population means of mean RU when the system is configured with any pair of BS, SR, ST, Dem, or PTV. Hypotheses 9*H*-14*H* relate to: FT (9,10*H*), FTCV (11,12*H*), and PT (13,14*H*).

In the Exponential case, 9*H* and 13*H* were rejected indicating that BS, SR, ST, and Dem have a statistically significant effect on FT and PT. However, some of these effects were not practically significant, ie, the effect sizes were small. The effect sizes can be gaged from the graphs and the related discussion in the preceding section. As an example, the effect sizes on FTCV and PT are large among the first three levels of BS and small among the last three levels. Equivalently, batch sizes 10, 20, and 30 have a practical effect, whereas batch sizes 40 and 50 don't. Hypotheses 7*H*, 8*H*, 10-12*H*, and 14*H* could not be rejected. While we cannot conclude that there is no main or interaction effect on the response variables, the experiment was not sufficiently sensitive to detect it.

In the Normal case, the hypotheses related to the main effects, except 7*H*, were rejected. As above, some of the effects were not practically significant. The hypotheses related to interactions could not be rejected. The following table summarizes the main effects and interactions.

Main effects and Interactions

	Resource Utilization	Flow Time	Flow Time CV	Percent Tardy
BS	*e n*	*e n*	*e n*	*e n*
SR		*e n*	*n*	*e n*
ST		*e n*	*e n*	*e n*
Dem	*e n*	*e n*	*e n*	*e n*
PTV		*n*	*n*	*n*
BS×SR				*n*
BS×ST		*e n*	*n*	*e n*
BS×Dem	*n*	*e n*	*e n*	*e n*
BS×PTV				*n*
SR×ST				
SR×Dem			*n*	*n*
SR×PTV				
ST×Dem		*e n*	*e n*	*e n*
ST×PTV				
Dem×PTV				*n*

e - significant when service at the job shop is Exponential

n - significant when service at the job shop is Normal

The final step in the statistical analysis is finding multiple regression relationships for the dependent variables. There are 13 independent variables in the Exponential case and 15 in the Normal case. Each variable corresponds to a level of the independent variables. All variables are dummy (binary) variables since they contain categorical information such as the BS, SR, ST, and Dem level for an observation. Dummy variables are treated exactly like other regressors in multiple regression. The Stepwise procedure was used to select a significant set of predictors. This procedure initially selects the predictor with the largest correlation with y. If this predictor is found significant then another predictor is selected on a similar basis and so on. At each step the importance of a predictor is reassessed. Thus, a predictor that may have been included earlier may be insignificant later.

An examination of the collinearity diagnostics showed no evidence of multicollinearity and all VIFs were under two. Mallow's C_k was used to avoid misspecification and $C_k \approx k$ for all the models. The models were validated using adjusted R^2 and R^2_{Press} and these values were almost identical to R^2 indicating the generalizability of the regression relationships. Since we are dealing with dummy variables, the regression coefficients cannot be used to comment on the magnitude of the effect of the corresponding variables. Note that in such a case the regression equations have one level per category is missing. This is because only $n-1$ dummy variables are needed to specify n categories. If there is only one dummy variable in a model, the coefficient of the dummy variable represents the difference between intercepts of that category and the excluded category. The intercept for the excluded category is the constant term. With more than one dummy variable, an entire group of categories is excluded and the observations are compared with this reference group (Mirer 1988).

Since the regression relationships do not lend themselves well to intuitive deductions and a qualitative analysis was conducted in the previous section, we will condense the present discussion. This does not discount the utility of these expressions in analyzing different scenarios. The transformations used to control heteroscedasticity are indicated in italics.

Exponential:

$RU' = 1.84 - .076\ BS_{10} - .074\ BS_{20} + .57\ Dem_L$ $\quad RU'=1/RU$

$FT' = 6.78 - .91\ BS_{10} - .5\ BS_{20} - .25\ BS_{30} - .19\ BS_{50} + .015\ SR_{FIFO} + .035\ SR_{LPT} - .11\ ST_L - .38\ Dem_L$ $\quad FT'=log(FT)$

$FTCV = .44 - .1\ BS_{10} + .056\ BS_{20} - .026\ BS_{30} - .0065\ SR_{FIFO} - .013\ ST_L + .028Dem_L$

$PT' = 1.89 - .13\ BS_{10} + .16\ BS_{40} + .28\ BS_{50} - .094\ SR_{FIFO} - .22\ SR_{LPT} + .093\ ST_L - .88\ Dem_L$ $\quad PT'=log(PT)$

Normal:

$RU = .56 + .0079\ BS_{40} + .0046\ BS_{50} + .0015\ ST_L - .14\ Dem_L$

$FT' = 6.72 - .95\ BS_{10} - .67\ BS_{20} - .39\ BS_{30} - .17\ BS_{40} + .013\ SR_{FIFO} - .0004\ SR_{SPT} + .03\ SR_{LPT} - .17\ ST_L - .38\ Dem_L - .026\ PTV_L$ $\quad FT'=log(FT)$

$FTCV = .4 + .15\ BS_{10} + .078\ BS_{20} + .042\ BS_{30} + .02\ BS_{40} - .0075\ SR_{FIFO} - .033\ SR_{LPT} - .027\ Dem_L - .006\ PTV_L$

$PT' = 1.03 - .5\ BS_{10} - .73\ BS_{20} - .41\ BS_{30} - .21\ BS_{40} + .028\ SR_{FIFO} + .0014\ SR_{SPT} + .13\ SR_{LPT} - .23\ ST_L - .29\ Dem_L - .15\ PTV_L$ $\quad PT'=log(PT)$

5.6 Summary

We assessed the effectiveness of batching decisions by studying the effect of batch size, scheduling rule, set-up time, demand, and processing time variability on resource utilization, flow time, flow time CV, and percent tardy jobs. There is no variable that optimizes all the performance measure simultaneously. Each variable presents distinct outcomes and implications on the performance measures. While some observations are applicable to all configurations, a segregation by the processing time distribution seems appropriate.

The all-Exponential case causes substantial variability in the process. It is apparent that the first node and the scheduling rule have a significant impact on the performance measures. RU is mostly unaffected by the independent variables except

in high traffic facilities that start with a job shop. It is highest under SPT and lowest under LPT because the first node indicates the performance of the entire facility. Also, in high traffic situations, RU shows an asymptotic relationship with batch size. FT is unaffected by scheduling rules in facilities starting with a flow line. In other configurations, it is highest under LPT and lowest under SPT. FT is also affected by batch size in these cases. Due to set-ups, the FT of batches of size 10 is greater than that of size 20. However, in high traffic situations, SPT and LPT are very effective in containing the FT of small batches. FTCV is affected slightly by scheduling rules when a flow line is the first node. It does decrease with batch size indicating that large batches reduce variability in the process. In systems that begin with a job shop, FTCV is highest under SPT and lowest under FIFO. This is because in FIFO jobs are released to the facility without reordering. EDD is also effective in reducing the FTCV as compared to SPT and LPT. As expected, PT follows the same trend as FT with both being very sensitive to traffic. SPT contains tardiness consistently. EDD is also very effective except for small batches. Small batches are less tardy under LPT than under EDD. High demand leads to more traffic in the system thus magnifying the effect of scheduling rules. Increased traffic also results in a general deterioration in the values of the performance measures. High demand has a detrimental effect on systems that begin with a flow line whereas high set-up acts against systems starting with a job shop.

The Normal case imparts little variability to the job shop. Since processing on the flow lines is Exponential, the variability in the operation of the flow line is greater than that of the job shop. However, the job shop is still a bottleneck due to high set-up. As in the Exponential case, RU is mostly unaffected except in high traffic situations. Due to the reduced variability, FT, FTCV, and PT are slightly lower as compared to the Exponential case. PTV does not have a significant affect on the performance measures.

The preceding analysis forms the essence of this study. We have now completed the steps outlined in the research methodology. We formulated analytical models for one-

and two-server systems with service in batches. We then examined larger systems using simulation models. These were validated by queuing network models. The simulation models allowed an understanding of the interface between a job shop and a flow line and the relationships between the dependent and the independent variables. The simulation output was tested statistically and checked for the assumptions of analysis of variance and hypothesis testing. We then tested the hypotheses proposed in chapter four and built multiple regression models. In the concluding chapter we will summarize the research, make recommendations for batch manufacturing systems, and provide directions for further research.

Chapter 6
Conclusion

Batch production accounts for most manufacturing activity today. The conventional method of appraising the performance of a manufacturing system has been scheduling. It is apparent from this study that batching policies have a significant impact on the performance of a system. The aim of this research is to examine the performance of batching policies under given configurations of a manufacturing facility. The context is manufacturing facilities with flow lines and job shops. As in most studies, we have observed some expected results and a few unexpected ones. In the sequel we provide an abstract of the research and directions for further work.

6.1 Synopsis

We approached this research by exploring analytical models based on phase-type distributions. While the models provide exact results for the wait of an individual customer in a batch, they are suitable for studying individual workcenters such as bottleneck machines. The models can also be used for an aggregate analysis of the entire manufacturing facility. The next step was to build simulation models with alternating flow line and job shop nodes. These nodes are distinguished by the set-up and processing time. The hypothesis is that a manufacturing facility can be decomposed into nodes with these characteristics and consecutive nodes of the same type can be approximated by a single node with appropriate set-up and processing time. Five systems of increasing complexity were simulated. The data gathered were deemed sufficient to extrapolate the results to a general batch manufacturing facility. The simulation model was validated by queuing network analysis. Experimental design and statistical analysis were used as a basis for hypothesis testing and multiple regression.

6.2 Results

The analytical models are based on discrete phase-type distributions, which can model a variety of discrete distributions. The main result from the numerical examples is that the wait before service of an individual customer in a batch is decreasing with the customer and that this decrease may not be linear.

The results from the simulations provide considerable insight into the operation of batch manufacturing systems. The configuration of the facility has a major impact on its performance. A job shop at the beginning of a facility has a detrimental effect on the performance measures as it delays work for the rest of the facility. These types of facilities benefit immensely from scheduling rules because of the queue that forms before the job shop. However, scheduling rules are not beneficial for facilities that begin with a flow line. In this study, batches were scheduled before the first node. It is also possible to apply the scheduling rules before every node but this is not a practical option in most manufacturing systems.

A valuable result is that the gain in performance measures is marginal after batches of size 20, indicating that this batch size "optimizes" the performance measures simultaneously. Thus, it may not be worthwhile to produce in large batches as small batches can enhance quality, flexibility, and capacity (cf §1.2). An interesting result is that SPT is more effective at meeting due dates than EDD. This may be a function of the parameters used in this study. However, pilot runs of the simulation under other parameter values also preferred SPT. As mentioned above, facilities that begin with a flow line do not benefit significantly from scheduling rules. The simulation results show that FIFO performs reliably in such cases. Considering that FIFO does not require knowledge of the processing times and it reduces flow time CV, it provides added benefits as a scheduling rule.

Set-up time has a significant effect on the performance measures, especially for small batches. While the importance of decreasing set-up is stressed often, it is frequently infeasible to reduce set-up without significant investment in technology. In

this study, set-up time is irrespective of batch size. This means that small batches incur the same set-up as large batches. Thus, small batches have a long wait before service. This fact is documented in many analytical studies (cf Chapter 2) and is substantiated by the simulation models in this study. However, in cases where parts are loaded on the machines, set-up time may (also) be a function of batch size. This scenario more closely models sequence-dependent set-up times. The simulation models in this study were used to examine the variable set-up case. The models were also subjected to high traffic and the results from both cases were not significantly different from the ones reported in this study.

The effect of demand is similar to set-up time. This means that increasing demand has a similar effect on the performance measures as increasing set-up. However, unlike set-up, which has a direct effect on the job shops only, demand affects the entire facility. Thus, scheduling rules are slightly effective for systems beginning with a flow line. Finally, while processing time variability has a statistically significant effect on the performance measures, a look at the graphs reveals that this effect is not significant practically.

The main contribution of this research is that batching is a viable option in manufacturing because conventional job shop scheduling theory renders a simplistic and unrealistic view of manufacturing and is inadequate in predicting the behavior of a realistic facility. This statement gains much significance considering that batch production accounts for upto 80% of manufacturing activity (Chevalier 1986).

This research relies on the decomposition of a manufacturing facility into two types of nodes—job shops and flow lines—that are differentiated by the set-up and processing time requirements. The job shop nodes can be considered as bottlenecks, which are present in most manufacturing systems. Thus, the five configurations examined in this study can also be differentiated by the position of the bottleneck(s). An important result of this study is the influence of the system configuration or the position of the bottleneck node(s) on the performance measures. This study is thus applicable to most batch manufacturing systems.

The results can be used for selecting a configuration given the objective of

optimizing a performance measure or for selecting levels of independent variables that optimize a performance measure under a given configuration. It is evident that a bottleneck is more detrimental to a facility when it appears at the front. However, if this is the case, then appropriate batching policies could be used to optimize selected performance measures. For example, if the first node is a job shop and the objective is to meet demand on time, SPT and EDD are both effective, which leaves the manager with the tradeoffs involved in selecting either scheduling rule. If the first node is a flow line then the choice is the same, except that the tradeoffs are different.

6.3 Implications

The models developed in this research are analytical as well as simulation. The analytical models can be used to obtain exact results for single- and two-server systems that process work in batches. These models are useful for aggregate analysis of an entire system. The advantage is that numerical results are available quickly so that alternative configurations are easy to examine. Simulation can then be used to examine a specific configuration in detail. Besides providing numerical results, this research provides fresh intuitive insight into manufacturing problems.

We have assumed that a manufacturing system that is composed of flow lines and job shops can be decomposed so that a series of consecutive flow lines can be aggregated into one flow line and a group of job shops can be aggregated into one job shop. This can be achieved by altering the set-up and processing times of the aggregate nodes. While this approach does not allow a detailed analysis, alternative system configurations can be evaluated quickly by studying the simplified arrangements.

The models in this study capture set-up precisely by including it explicitly. This research also gives a faithful view of a bottleneck machine. A bottleneck is considered to be the machine that determines total output. This machine is the first to saturate as output is increased. However, machines usually saturate because of queues and not a capacity limit. The bottlenecks can shift with queues and are the ones that determine the batching policy.

The models can be used for various purposes. Appropriate scheduling rules and batch sizes can help immensely by saving set-ups, especially on bottleneck machines. A knowledge of flow times can also allow precise MRP calculations and improve the order release times. The models can judge the performance of a facility under differing product mix, demand, and shift policy. Issues such as capacity planning can also be addressed by changing the characteristics of individual machines and if feasible, the configuration of the facility. The interaction of the performance measures with variables such as set-up time can be measured easily. Thus, the impact of investment in technology on set-up time can be evaluated (Karmarkar *et al* 1985).

This research can help the decision maker select a performance parameter that will optimize a given strategy. The effect of system configuration on the performance of a facility can be used to influence the design of the system before the planning stage is implemented. This information can be critical in developing the manufacturing strategy for a new product, changing the product mix, judging capacity, and investment in technology. Thus, the management can align the operations strategy with the long-term manufacturing strategy of the firm.

6.4 Further Research

Research in batching is at an early stage. This is mainly due to the conventional view of scheduling a job shop. This study leads to many avenues in this field. The analytical models could be extended to a network of workcenters. The models could also be modified to consider the correlation that is added to the output stream from a preceding machine. A useful addition to the simulation models will be sequence dependent set-ups and dynamic scheduling rules. A use of other performance measures is also possible. A natural progression would be to combine these measures into a multi-criteria decision making problem. We have considered the effect of the change in a parameter by keeping the other parameters constant. The models can also be used to evaluate configurations in which more than one parameter is altered. It would also be useful to incorporate rework, machine breakdowns, and transport times.

References

Ahmadi J H, R H Ahmadi, S Dasu, and C S Tang, Batching and scheduling jobs on batch and discrete processors, *Operations Research* 40(4) 1992, 750-763

Aneja Y P and N Singh, Scheduling production of common components at a single facility, *IIE Transactions* 22(3) 1990, 234-237

Baker K R, Scheduling the production of components at a common facility, *IIE Transactions* 20(1) 1988, 32-35

Balci O and R G Sargent, A bibliography on the credibility, assessment, and validation of simulation and mathematical models, *Simuletter* 15(3) 1984, 15-27

Banks J and J S Carson, *Discrete-Event System Simulation*, Prentice-Hall, Englewood Cliffs, NJ, 1984

Banks J, J S Carson II, and J N Sy, *Getting Started with GPSS/H*, Wolverine Software Corporation, 1989

Barcikowski R S, *Computer Packages and Research Design, vol 2:SAS*, University Press of America, Lanham, MD, 1983

Berenson M L, D M Levine, and M Goldstein, *Intermediate Statistical Methods and Applications: A Computer Package Approach*, Prentice-Hall Inc, Englewood Cliffs, NJ, 1983

Bertrand J W M, Multiproduct optimal batch sizes with in-process inventories and

multi work centers, *IIE Transactions* 17(2) 1985, 157-163

Bruno J and P Downey, Complexity of task sequencing with deadlines, set-up times changeover costs, *SIAM Journal of Computing* 7(4) 1978, 393-404

Buzacott J A, Optimal operating rules for automated manufacturing systems, *IEEE Transactions on Automatic Control* AC-27(1) 1982, 80-86

Carson J S, Convincing users of model's validity is challenging aspect of modeler's job, *Industrial Engineering* 18(6) 1986, 74-85

Chevalier P W, Group technology as a CAD/CAM integrator in batch manufacturing, *International Journal of Operations and Productions Management* 4(3) 1986, 3-12

Chua R C H, G D Scudder, and A V Hill, Batching policies for a repair shop with limited spares and finite capacity, *European Journal of Operational Research* 66(1) 1993, 135-147

Cochrane D and G H Orcutt, Application of least squares regression to relationships containing autocorrelated error terms, *Journal of the American Statistical Association* 44(245) 1949, 32-61

Coffman E G, M Garey, and D Johnson, An application of bin packing to multi-processor scheduling, *SIAM Journal of Computing* 7(1) 1978, 1-16

Coffman E G, A Nozari, and M Yannakakis, Optimal scheduling of products with two subassemblies on a single machine, *Operations Research* 37(3) 1989, 426-436

Cohen J, *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York, 1977

Conway R W, Some tactical problems in digital simulation, *Management Science* 10(1) 1963, 47-61

Conway R W, Priority dispatching and job lateness in a job shop, *The Journal of Industrial Engineering* 16(4) 1965, 228-237

Cramer E M and R D Bock, Multivariate analysis, *Review of Educational Research* 36(5) 1966, 604-617

D'Agostino R B, Tests for the normal distribution, in *Goodness-of-Fit Techniques* (R B D'Agostino and M A Stephens, eds), Marcel-Dekker Inc, NY, 1986

Dobson G, U S Karmarkar, and J L Rummel, Batching to minimize flow times on one machine, *Management Science* 33(6) 1987, 784-799

Dobson G, U S Karmarkar, and J L Rummel, Batching to minimize flow times on parallel heterogeneous machines, *Management Science* 35(5) 1989, 607-613

Durbin J and G S Watson, Testing for serial correlation in least squares regression, *Biometrika* 37(3-4) 1950, 409-428

Durbin J and G S Watson, Testing for serial correlation in least squares regression, *Biometrika* 38(1-2) 1951, 159-178

Everitt B S, A monte carlo investigation of the robustness of Hotelling's one and two sample T^2 tests, *Journal of the American Statistical Association* 74(365) 1979, 48-51

Gerwin D, Do's and dont's of computerized manufacturing, *Harvard Business Review* 60(2) 1982, 107-116

Gnanadesikan R, *Methods for Statistical Analysis of Multivariate Observations*, John Wiley & Sons, NY, 1977

Hair J F, R E Anderson, and R L Tatham, *Multivariate Data Analysis with Readings, 2nd ed*, Macmillan Publishing Company, NY, 1987

Henriksen J O and R C Crain, *GPSS/H Reference Manual, 3rd ed*, Wolverine Software Corporation, Annandale, VA, 1989

Herald M J and S Y Nof, The optimal planning of computerized manufacturing systems, Report #11, School of Industrial Engineering, Purdue University, 1978

Holloway L N and O J Dunn, The robustness of Hotelling's T^2, *Journal of the American Statistical Association* 62(317) 1967, 124-136

Hoover S V and R F Perry, *Simulation: A Problem Solving Approach*, Addison-Wesley Publishing Company, Reading, MA, 1989

Hopkins J W and P P F Clay, Some empirical distributions of bivariate T^2 and homoscedasticity criterion M under unequal variance and leptokurtosis, *Journal of the American Statistical Association* 58(304) 1963, 1048-1053

Hummel T J and J R Sligo, Empirical comparison of univariate and multivariate analysis of variance procedures, *Psychological Bulletin* 76(1) 1971, 49-57

Jackman J and E Johnson, The role of queuing network models in performance evaluation of manufacturing systems, *Journal of the Operational Research Society*

44(8) 1993, 797-807

Johnson D, Near optimal bin packing algorithms, Report MAC TR-109, MIT, MA, USA, 1973

Karmarkar U S, Lot sizes, lead times, and in-process inventories, *Management Science* 33(3) 1987, 409-418

Karmarkar U S, S Kekre, and S Kekre, Lot sizing in multi-item multi-machine job shops, *IIE Transactions* 17(3) 1985a, 290-297

Karmarkar U S, S Kekre, S Kekre, and S Freeman, Lot sizing and lead time performance in a manufacturing cell, *Interfaces* 15(2) 1985b, 1-9

Karmarkar U S, S Kekre, and S Kekre, Capacity analysis of a manufacturing cell, *Journal of Manufacturing Systems* 6(3) 1987, 165-175

Karmarkar U S, S Kekre, and S Kekre, Multi-item batching heuristics for minimization of queuing delays, *European Journal of Operational Research* 58(1) 1992, 99-111

Kekre S, Performance of a manufacturing cell with increased product mix, *IIE Transactions* 19(3) 1987, 329-339

Kekre S and V Udayabhanu, Customer priorities and lead times in long-term supply contracts, *Journal of Manufacturing and Operations Management* 1(1) 1988, 44-66

Koutsoyiannis A, *Theory of Econometrics, 2nd ed*, The Macmillan Press Ltd, London, UK, 1977

Laüter J, Sample size requirements for the T^2 test of MANOVA (tables for one-way classification), *Biometrical Journal* 20(?) 1978, 389-406

Law A M and W D Kelton, *Simulation Modeling and Analysis, 2nd ed*, McGraw-Hill, NY, 1991

Lee C–Y, S D Liman, and A Wirakusumah, Product batching and batch sequencing for NC punch presses, *International Journal of Production Research* 31(5) (1993) 1143-1156

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

Mallow C L, Some comments on Cp, *Technometrics* 15(4) 1973, 661-676

Mardia K V, The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model, *Biometrika* 58(1) 1971, 105-121

Mirer T W, *Economic Statistics and Econometrics, 2nd ed*, Macmillan Publishing Co, NY, 1988

Monma C L and C N Potts, On the complexity of scheduling with batch setup times, *Operations Research* 37(5) 1989, 798-804

Montgomery D C, *Design and Analysis of Experiments, 3rd ed*, John Wiley & Sons, NY, 1991

Morrison D F, *Multivariate Statistical Methods*, McGraw-Hill, NY, 1967

Myers R, *Classical and modern regression with applications, 2nd ed*, Duxbury Press,

Boston, MA, 1990

Naddef D and C Santos, One-pass batching algorithms for the one-machine problem, *Discrete Applied Mathematics* 21(2) 1988, 133-145

Neuts, M F, Matrix-geometric solutions in stochastic models: An algorithmic approach, The Johns Hopkins University Press, Baltimore MD, 1981

Olson C L, Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association* 69(348) 1974, 894-908

Orlicky J, *Manufacturing Requirements Planning*, McGraw–Hill, NY, 1975

Porteus E L, Investing in reduced setups in the EOQ model, *Management Science* 31(8) 1985, 998-1010

Porteus E L, Investing in new parameter values in the discounted EOQ model, *Naval Research Logistics Quarterly* 33(1) 1986a, 39-48

Porteus E L, Optimal lot sizing, process quality improvement, and setup cost reduction, *Operations Research* 34(1) 1986b, 137-144

Potts C N and L N Van Wassenhove, Integrating scheduling with batching and lot-sizing: A review of algorithms and complexity, *Journal of the Operational Research Society* 43(5) 1992, 395-406

Roy S N and R C Bose, Simultaneous confidence interval estimation, *Annals of Mathematical Statistics* 24(4) 1953, 513-536

Santos C and M Magazine, Batching in single operation manufacturing systems,

Operations Research Letters 4(3) 1985, 99-103

Sargent R G, A tutorial on validation and verification of simulation models, *Proc 1988 Winter Simulation Conference*, San Diego, CA, 1988, 33-39

SAS Institute Inc, *SAS/ETS User's Guide, Version 6, 1st ed*, Cary, NC: SAS Institute Inc, 1988

SAS Institute Inc, *SAS/STAT User's Guide, Version 6, 4th ed, v1*, Cary, NC: SAS Institute Inc, 1989

SAS Institute Inc, *SAS/STAT User's Guide, Version 6, 4th ed, v2*, Cary, NC: SAS Institute Inc, 1989

SAS Institute Inc, *SAS Language: Reference, Version 6, 1st ed*, Cary, NC: SAS Institute Inc, 1990

SAS Institute Inc, *SAS Procedures Guide, Version 6, 3rd ed*, Cary, NC: SAS Institute Inc, 1990

Scariano S and J Davenport, The effects of violations of the independence assumption in the one way ANOVA, *The American Statistician* 41(2) 1987, 123-129

Schaffer G H, Implementing CIM, *American Machinist Special Report 736*, 1981

Schriber, T J, *Simulation using GPSS*, John Wiley & Sons, NY, 1974

Schriber, T J, *An Introduction to Simulation using GPSS/H*, John Wiley & Sons, NY, 1991

Schruben L W, Detecting initialization bias in simulation output, *Operations Research* 30(3) 1982, 569-590

Schruben L W, H Singh, and L Tierney, Optimal tests for initialization bias in simulation output, *Operations Research* 31(6) 1983, 1167-1178

Schmeiser B, Batch size effects in the analysis of simulation output, *Operations Research* 30(3) 1982, 556-568

Seidmann A, P J Schweitzer, and S Y Nof, Performance evaluation of a flexible manufacturing cell with random multiproduct feedback flow, *International Journal of Production Research* 23(6) 1985, 1171-1184

Shanthikumar J G and J A Buzacott, Open queuing network models of dynamic job shops, *International Journal of Production Research* 19(3) 1981, 255-266

Shapiro S S, M B Wilk, and H J Chen, A comparative study of various tests for normality, *Journal of the American Statistical Association* 63(324) 1968, 1343-1372

Smith M L, R Ramesh, R A Dudek, and E L Blair, Characteristics of US flexible manufacturing systems — A survey, *Proc 2nd ORSA/TIMS conference on Flexible Manufacturing Systems* Ann Arbor, MI, 1986, 477-486

Solberg J, Capacity planning with a stochastic workflow model, *AIIE Transactions* 13(2) 1981, 116-122

Stevens J, *Applied Multivariate Statistics for the Social Sciences, 2nd ed*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1992

Sung C S and C K Park, Scheduling of products with common and product dependent components manufactured at a single facility, *Journal of the Operational Research Society* 44(8) 1993, 773-784

Tang C S, Scheduling batches on parallel machines with major and minor set-ups, *European Journal of Operations Research* 46(1) 1990, 28-37

Thesen A and L Travis, Introduction to simulation, *Proc 1988 Winter Simulation Conference*, San Diego, CA, 1988, 7-14

Unal A T and A S Kiran, Batch sequencing, *IIE Transactions* 24(4) 1992, 73-83

Uzsoy R, C-Y Lee, and L A Martin-Vega, A review of production planning and scheduling models in the semiconductor industry. Part I: System characteristics, performance evaluation, and production planning, *IIE Transactions* 24(4) 1992, 47-60

Uzsoy R, C-Y Lee, and L A Martin-Vega, A review of production planning and scheduling models in the semiconductor industry. Part II: Shop-floor control, *IIE Transactions on Scheduling and Logistics* 26(5) 1994

Wesolowsky G O, *Multiple Regression and Analysis of Variance*, John Wiley & Sons, NY, 1976

Whitt W, The queuing network analyzer, *Bell System Technical Journal* 62(9) 1983a, 2779-2815

Whitt W, Performance of the queuing network analyzer, *Bell System Technical Journal* 62(9) 1983b, 2817-2843

Whitt W, Open and closed models for networks of queues, *Bell System Technical Journal* 63(9) 1984, 1911-1979

Zdrzałka S, Approximation algorithms for single-machine sequencing with delivery times and unit batch set-up times, *European Journal of Operational Research* 51(2) 1991, 199-209

Zdrzałka S, Analysis of approximation algorithms for single-machine scheduling with delivery times and sequence independent batch setup times, *European Journal of Operational Research* 80(2) 1995, 371-380

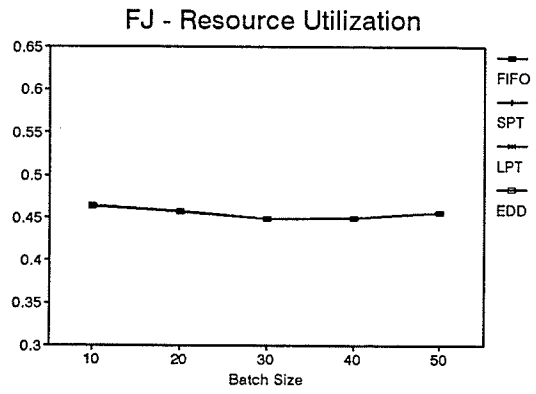
Zipkin P H, Models for design and control of stochastic, multi-item batch production systems, *Operations Research* 34(1) 1986, 91-104

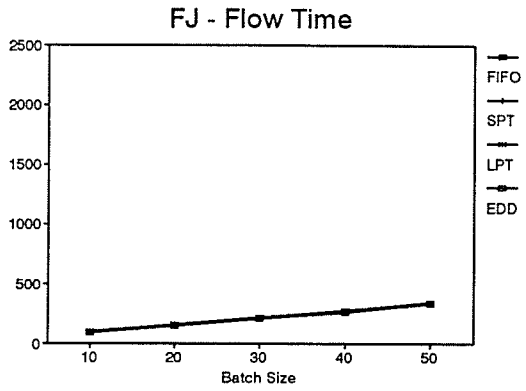
Appendix

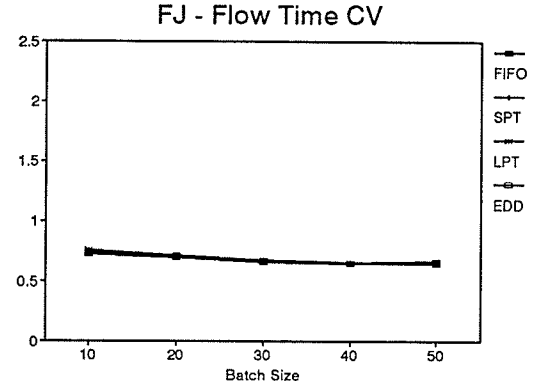
Graphs

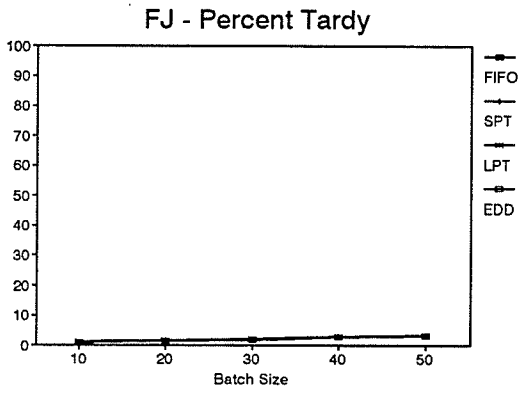
Exponential service at the job shop node

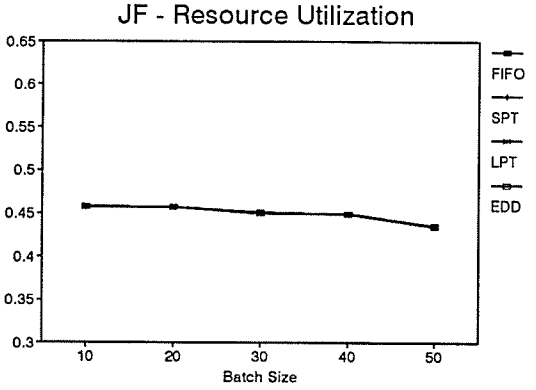
Low demand, low set-up

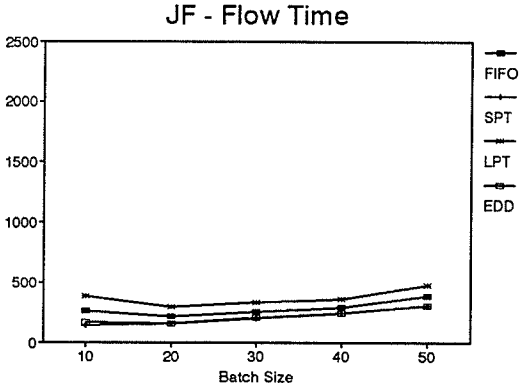

FJ - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

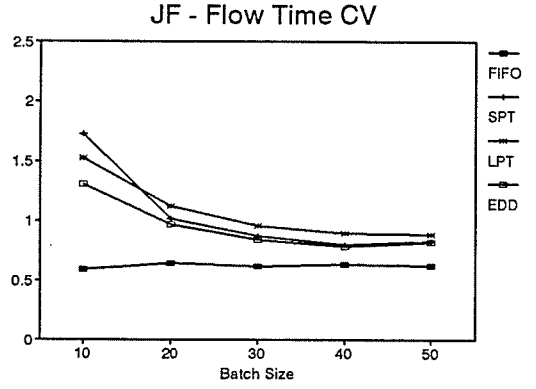

FJ - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

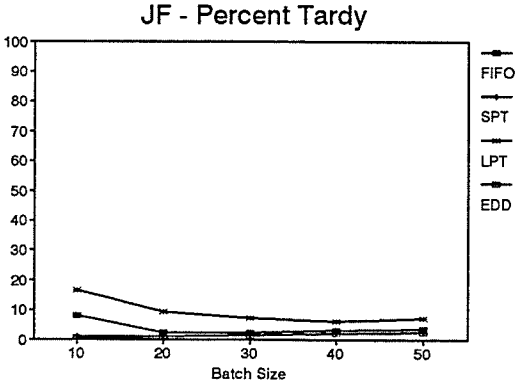

FJ - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

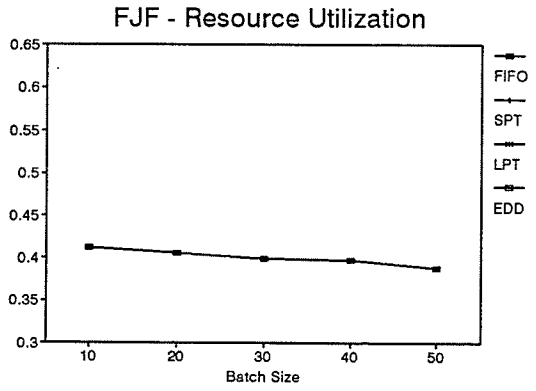

FJ - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

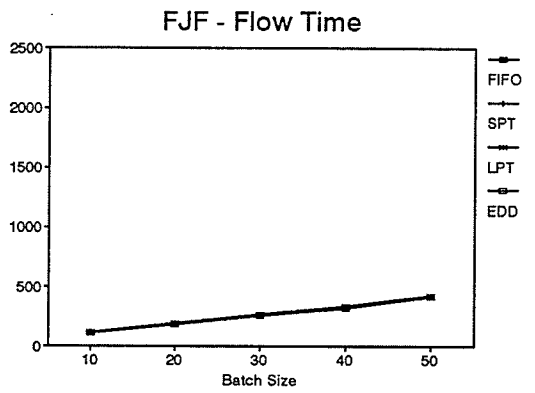

JF - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

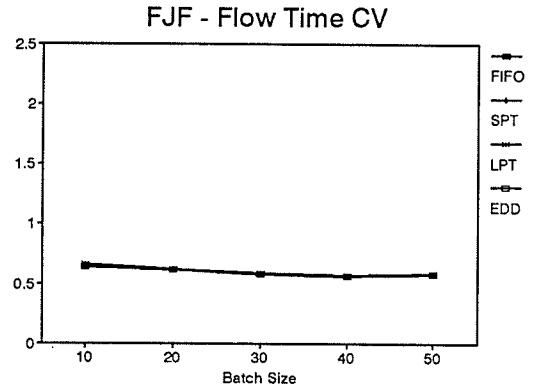

JF - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

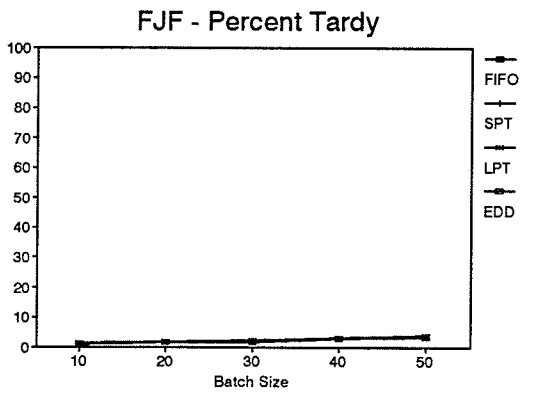

JF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

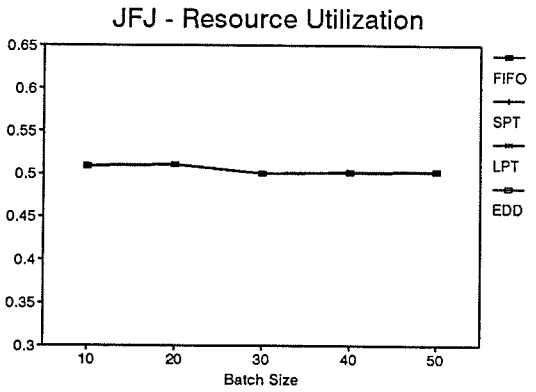

JF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

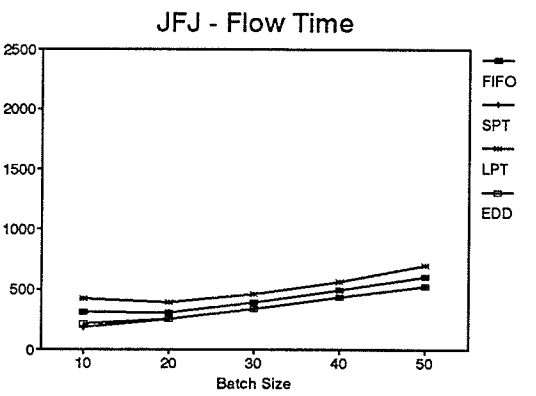

FJF - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

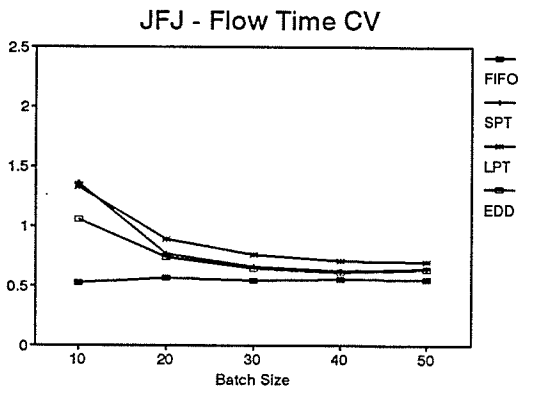

FJF - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

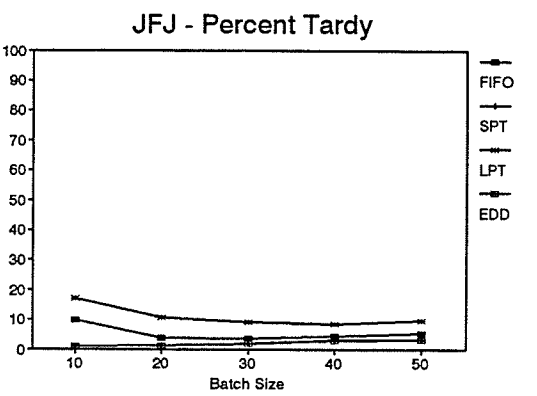

FJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

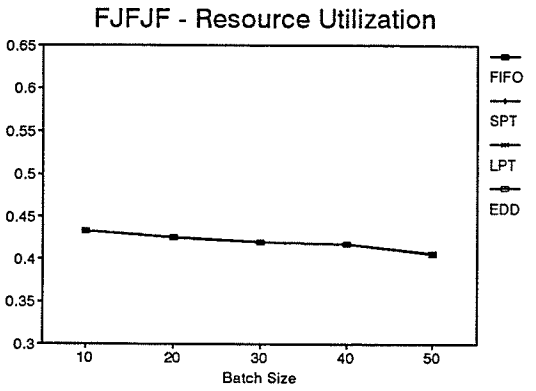

FJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

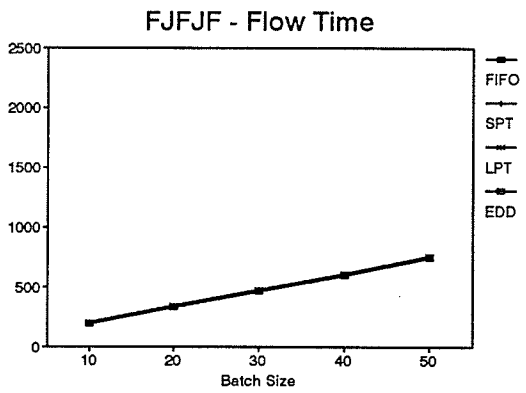

JFJ - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

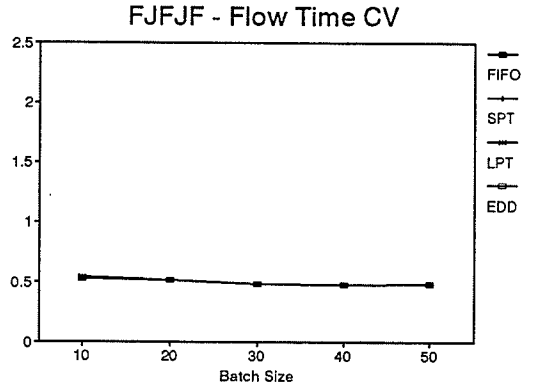

JFJ - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

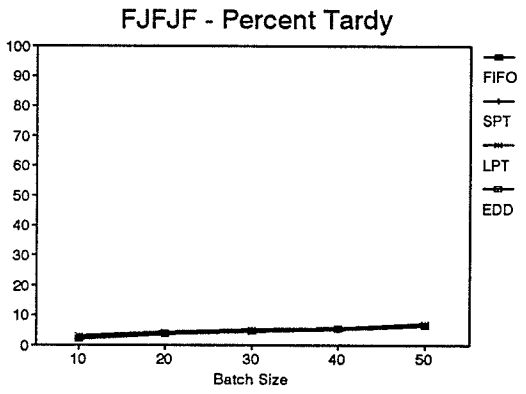

JFJ - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


JFJ - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

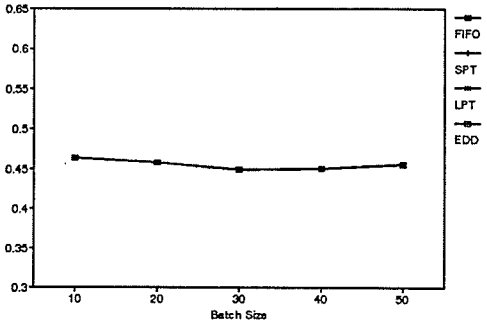

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

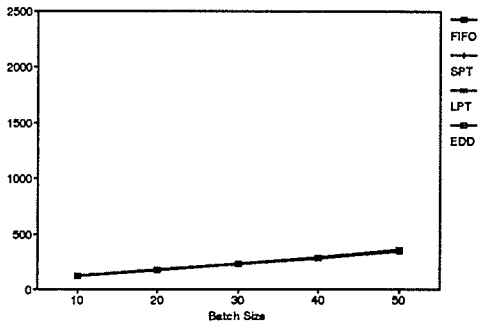
Exponential service at the job shop node

Low demand, high set-up

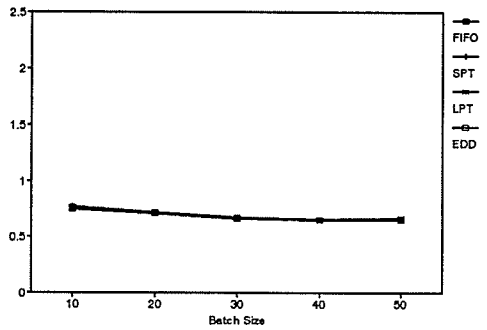
FJ - Resource Utilization



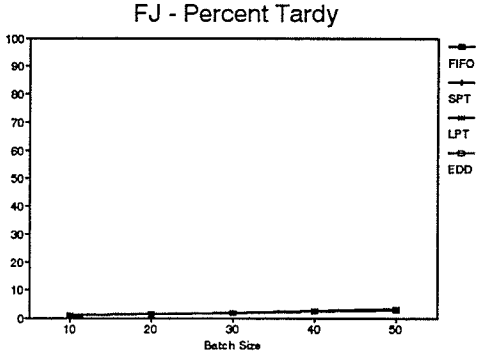
FJ - Flow Time



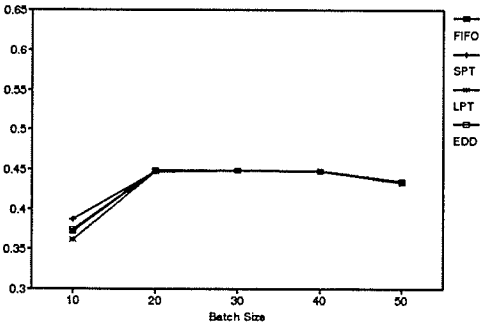
FJ - Flow Time CV



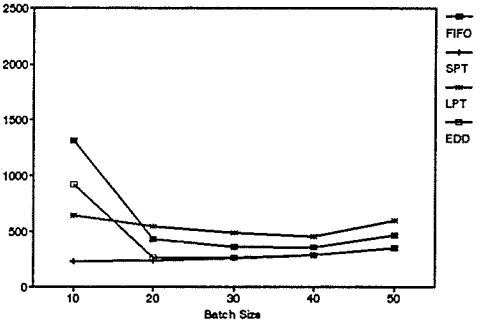
FJ - Percent Tardy



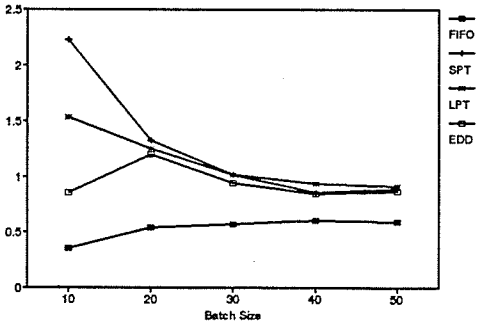
JF - Resource Utilization



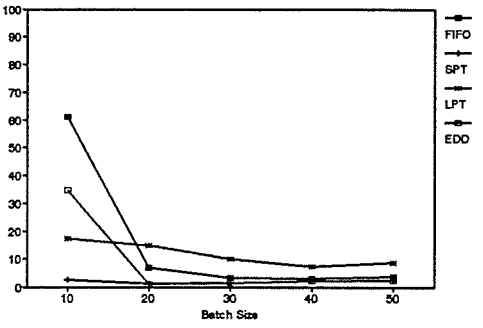
JF - Flow Time



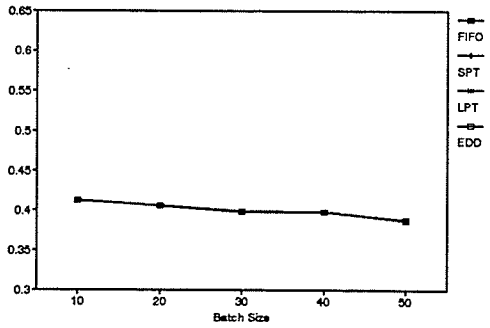
JF - Flow Time CV



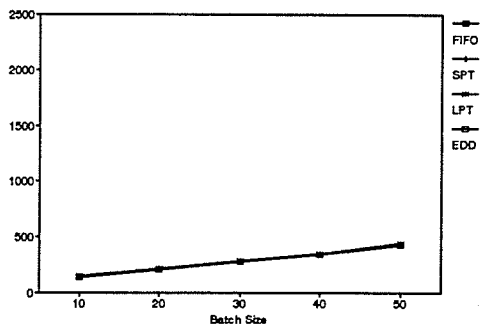
JF - Percent Tardy



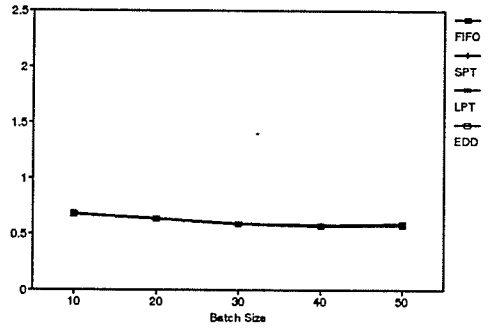
FJF - Resource Utilization



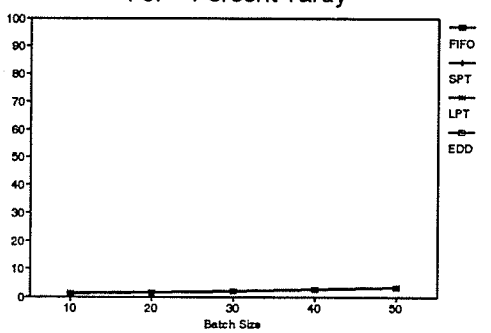
FJF - Flow Time



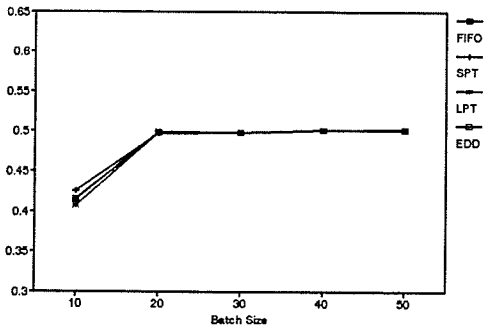
FJF - Flow Time CV



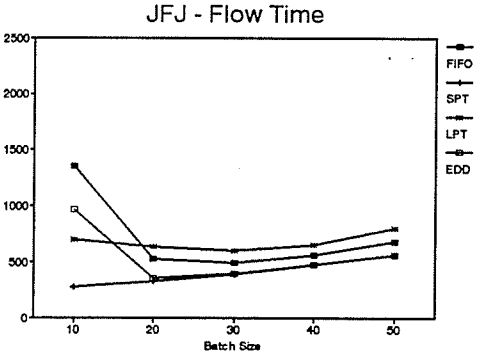
FJF - Percent Tardy



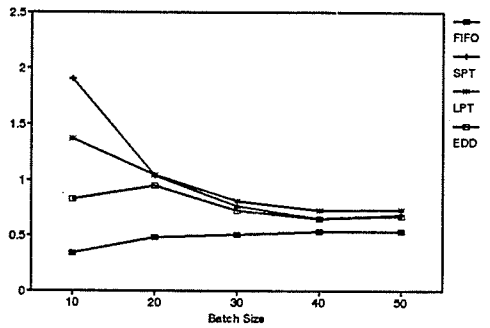
JFJ - Resource Utilization



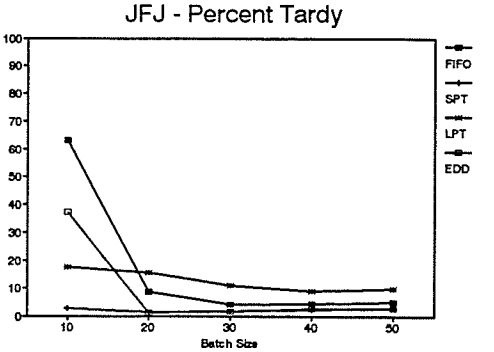
JFJ - Flow Time

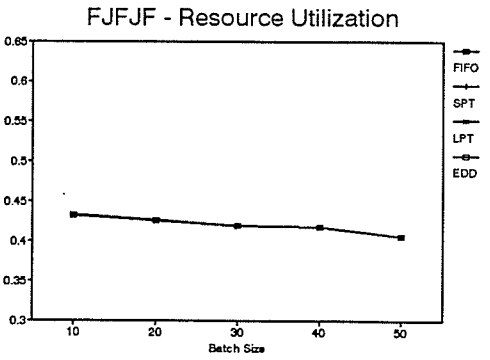


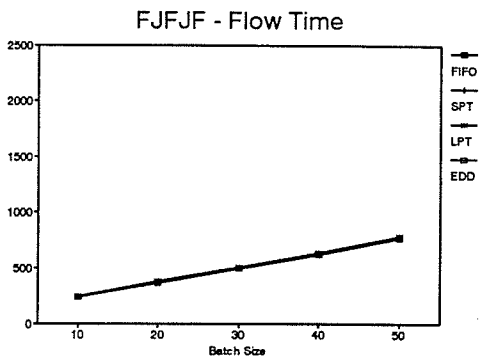
JFJ - Flow Time CV

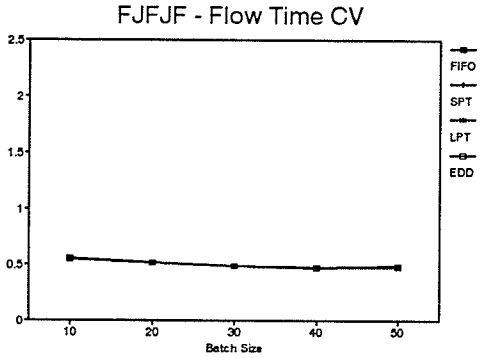


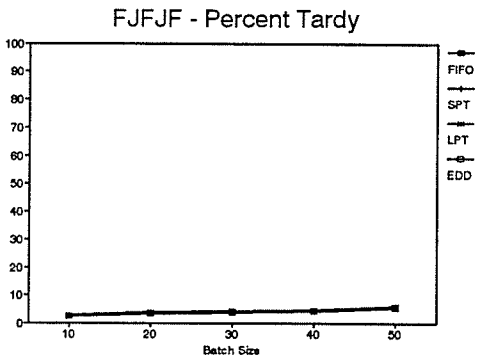
JFJ - Percent Tardy




FJFJF - Resource Utilization
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

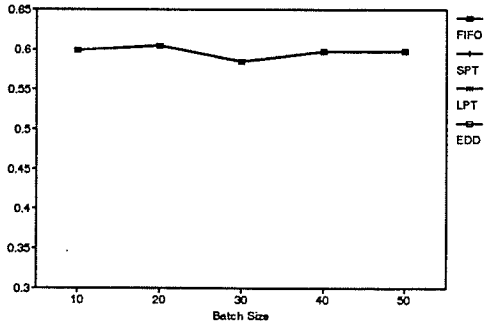

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

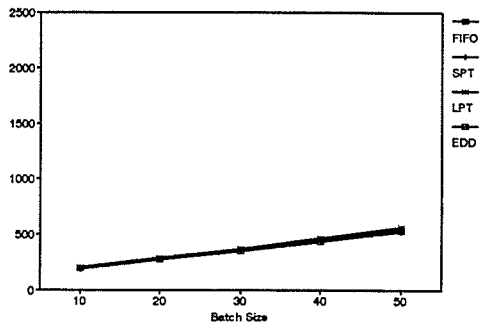
Exponential service at the job shop node

High demand, low set-up

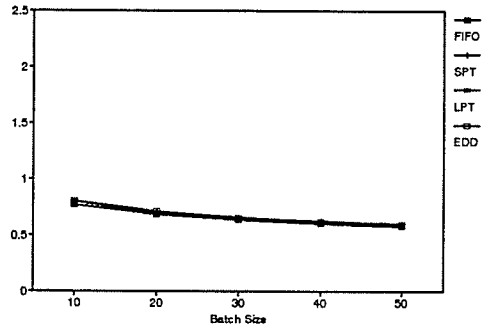
FJ - Resource Utilization



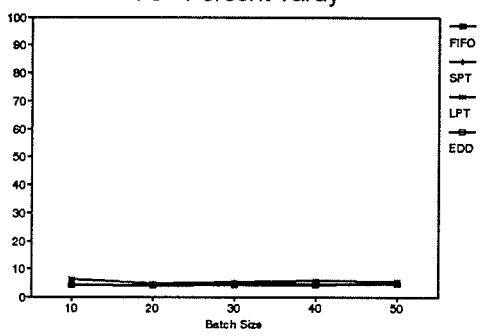
FJ - Flow Time



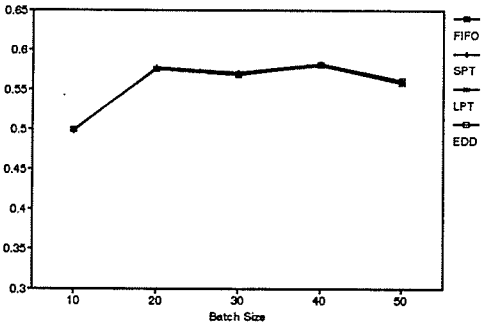
FJ - Flow Time CV



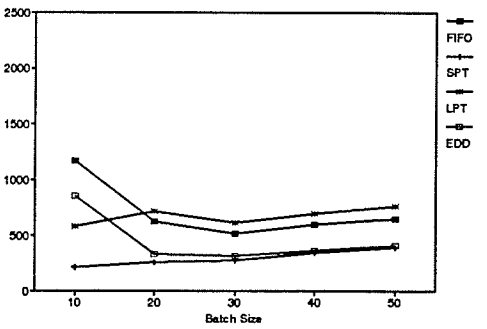
FJ - Percent Tardy



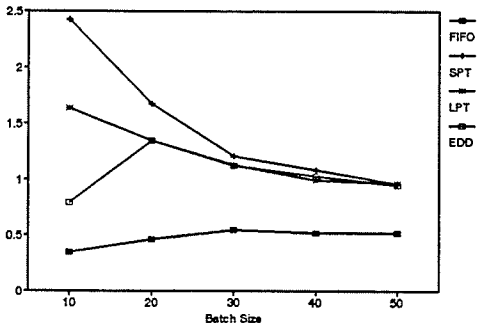
JF - Resource Utilization



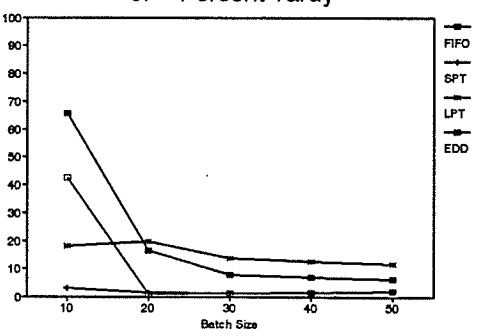
JF - Flow Time



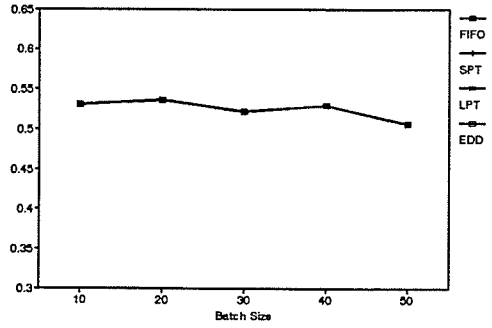
JF - Flow Time CV



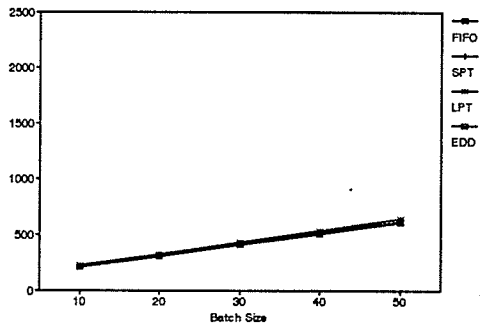
JF - Percent Tardy



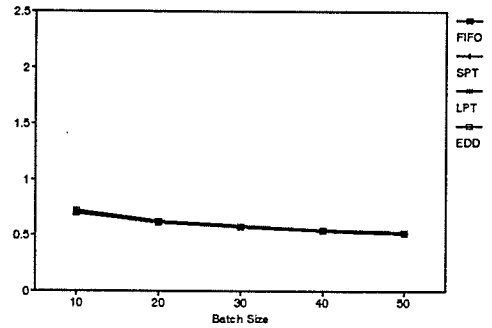
FJF - Resource Utilization



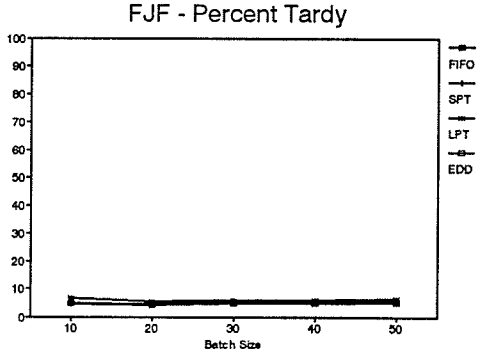
FJF - Flow Time



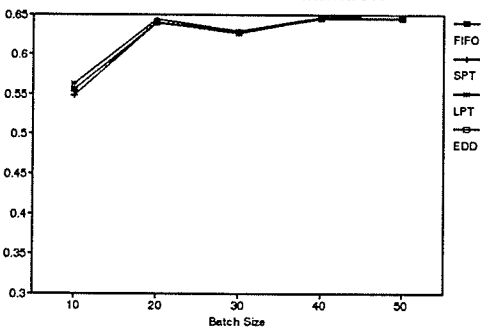
FJF - Flow Time CV



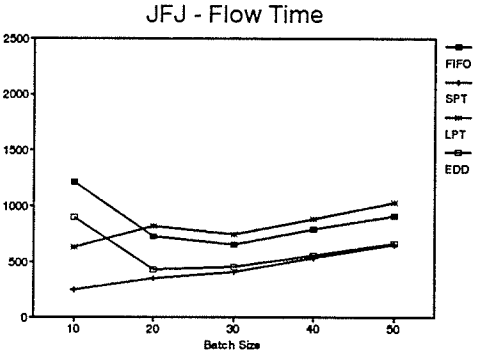
FJF - Percent Tardy



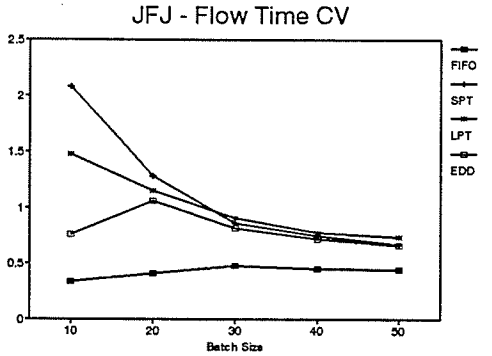
JFJ - Resource Utilization



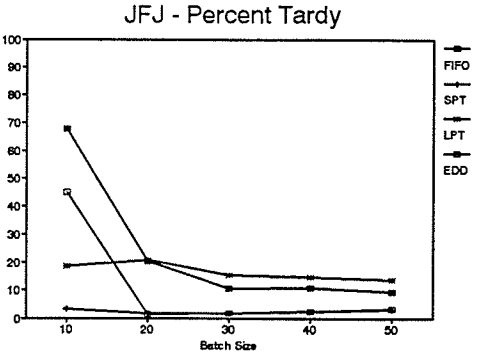
JFJ - Flow Time



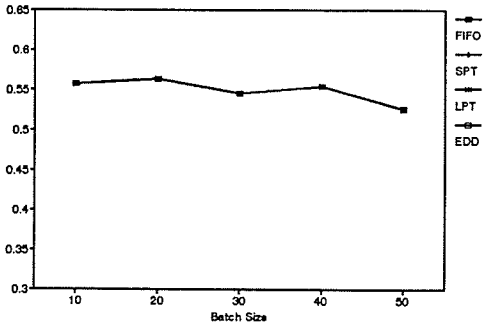
JFJ - Flow Time CV



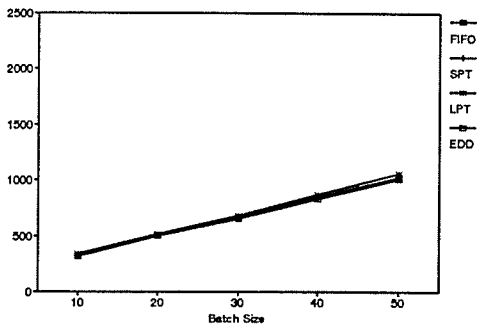
JFJ - Percent Tardy



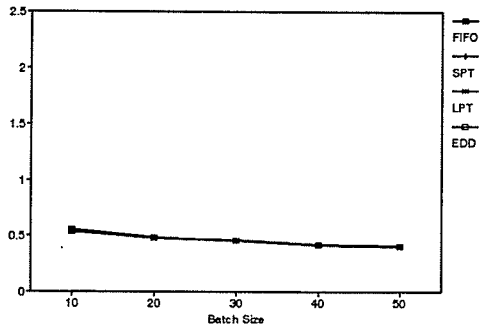
FJFJF - Resource Utilization



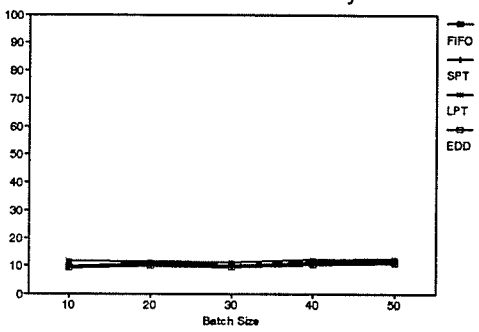
FJFJF - Flow Time



FJFJF - Flow Time CV



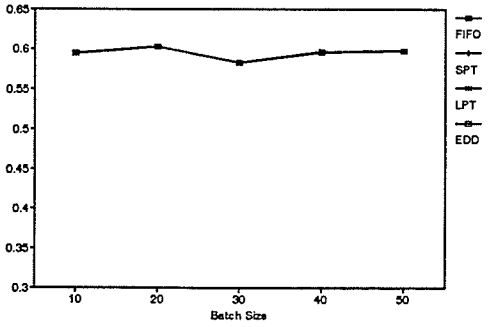
FJFJF - Percent Tardy



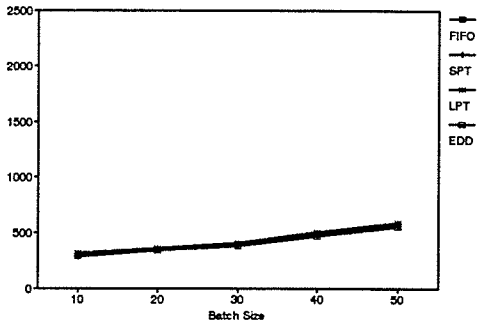
Exponential service at the job shop node

High demand, high set-up

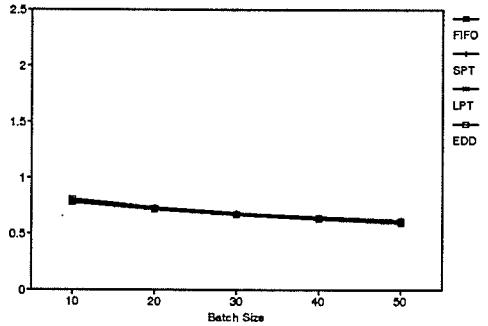
FJ - Resource Utilization



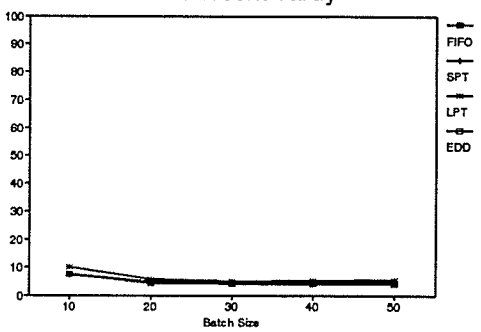
FJ - Flow Time



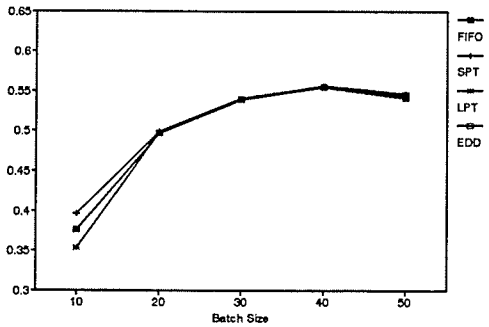
FJ - Flow Time CV



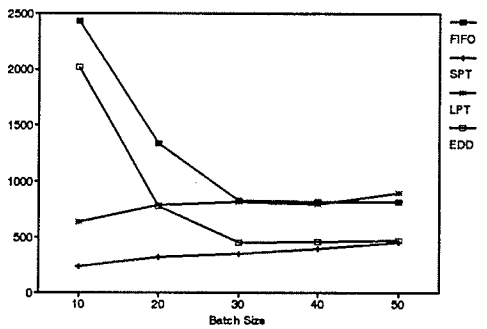
FJ - Percent Tardy



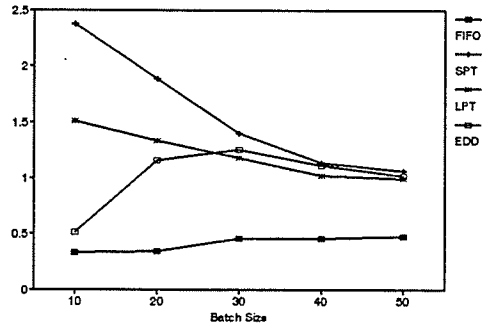
JF - Resource Utilization



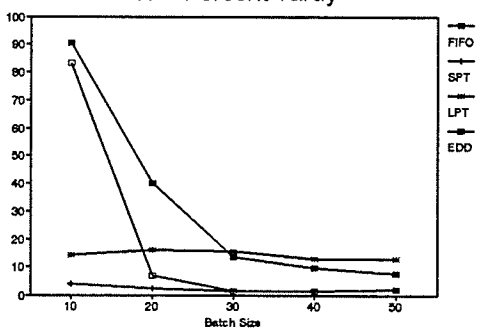
JF - Flow Time



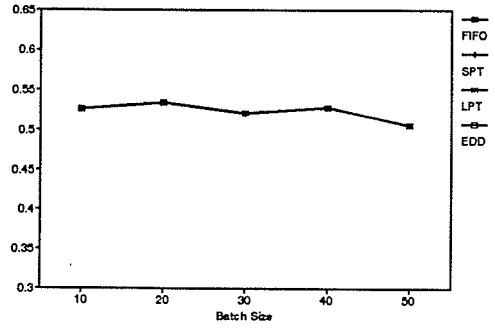
JF - Flow Time CV



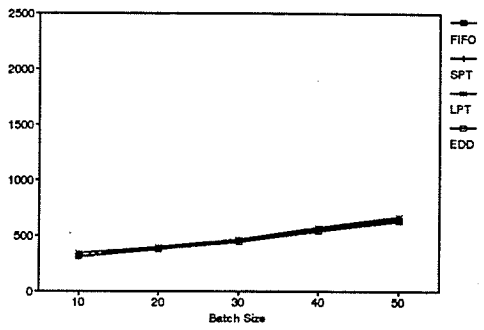
JF - Percent Tardy



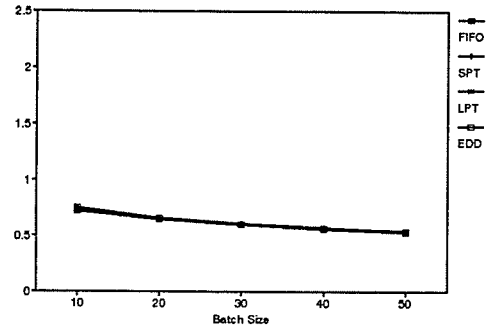
FJF - Resource Utilization



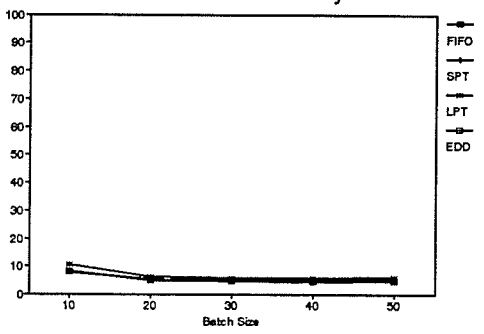
FJF - Flow Time



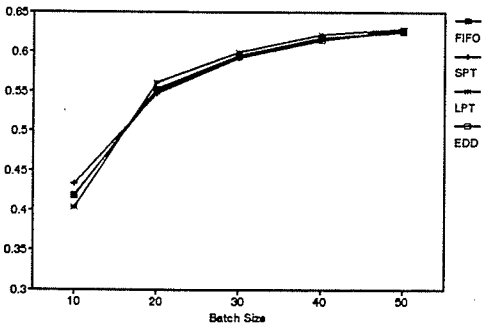
FJF - Flow Time CV



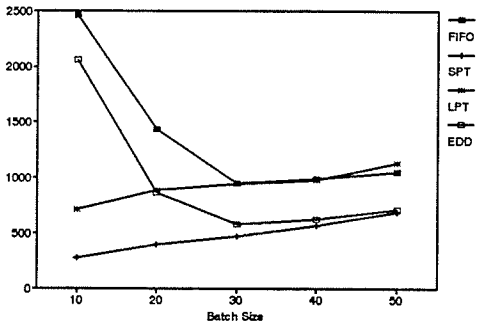
FJF - Percent Tardy



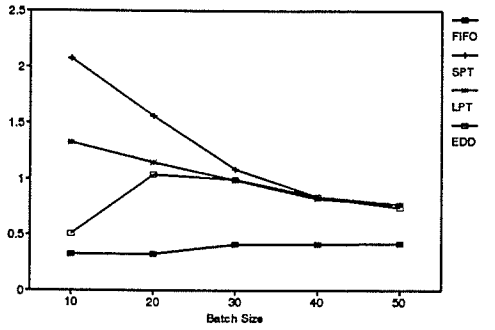
JFJ - Resource Utilization



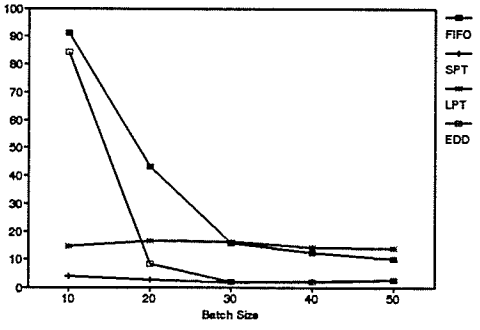
JFJ - Flow Time



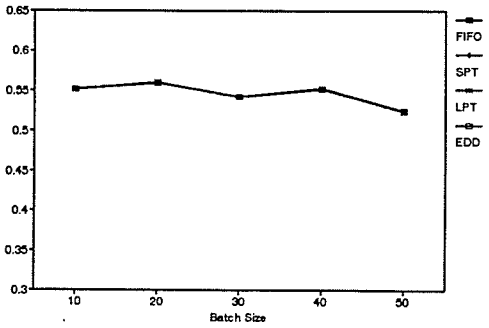
JFJ - Flow Time CV



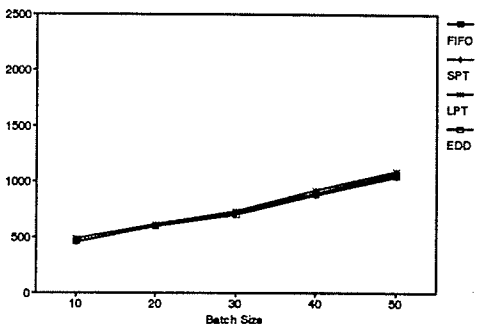
JFJ - Percent Tardy



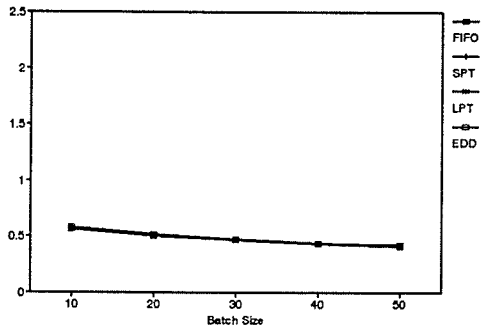
FJFJF - Resource Utilization



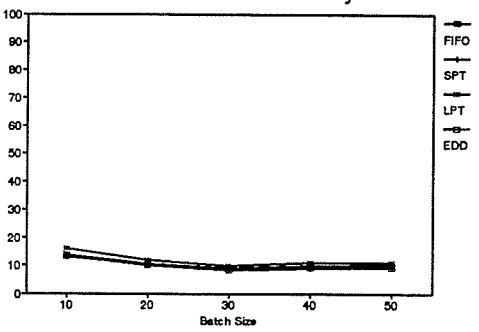
FJFJF - Flow Time



FJFJF - Flow Time CV



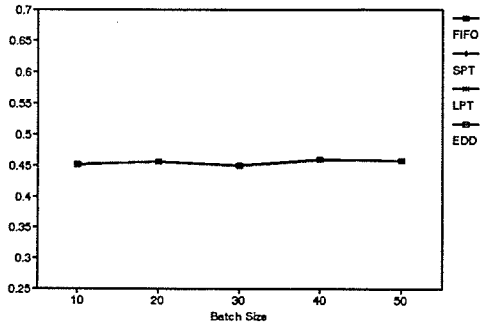
FJFJF - Percent Tardy



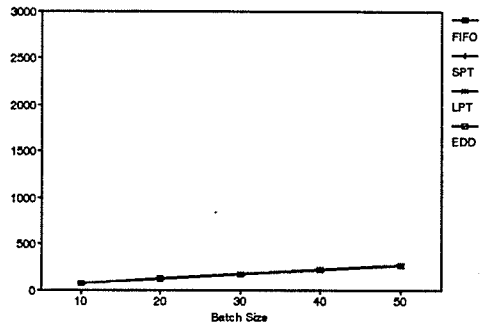
Normal service at the job shop node

Low demand, low set-up, low variability

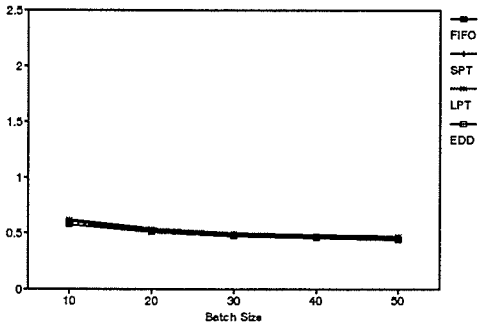
FJ - Resource Utilization



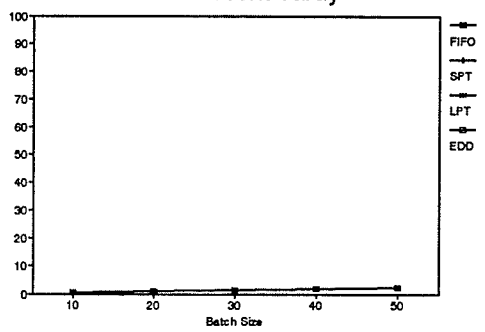
FJ - Flow Time



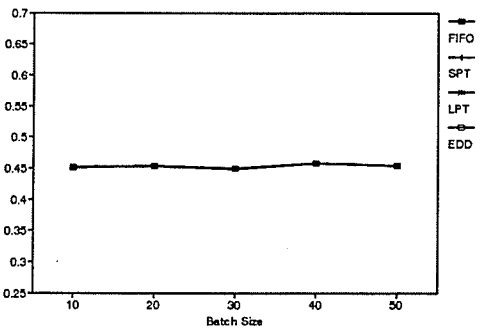
FJ - Flow Time CV



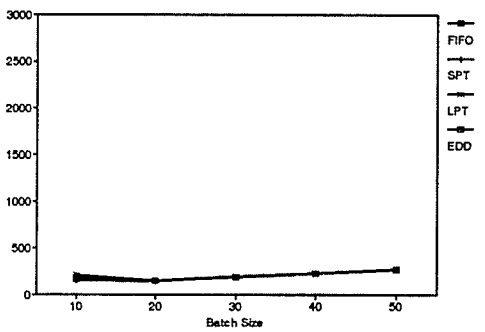
FJ - Percent Tardy



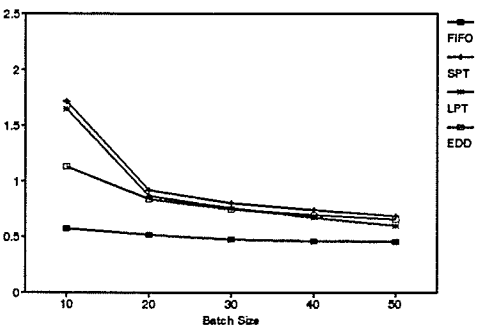
JF - Resource Utilization



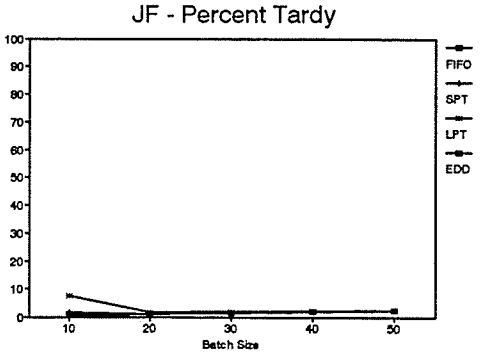
JF - Flow Time



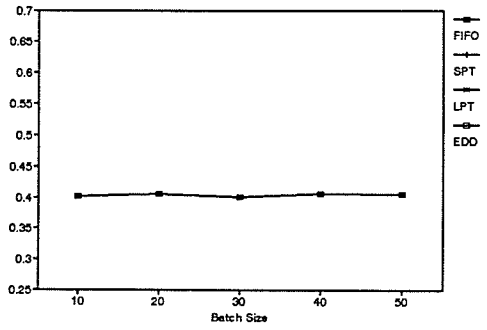
JF - Flow Time CV



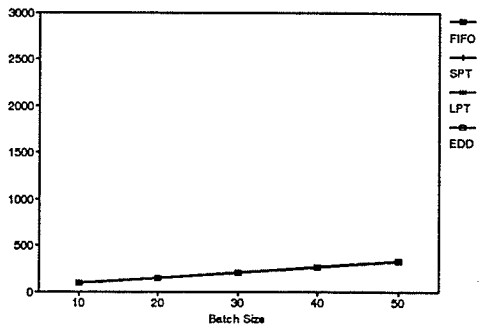
JF - Percent Tardy



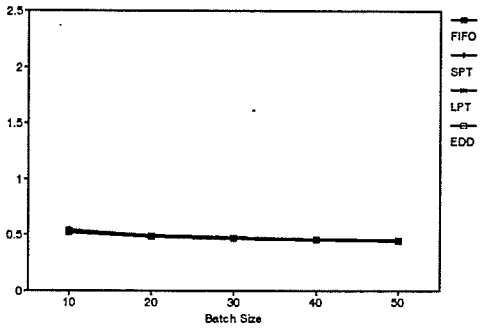
FJF - Resource Utilization



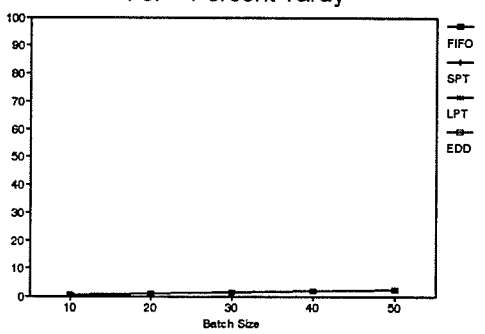
FJF - Flow Time



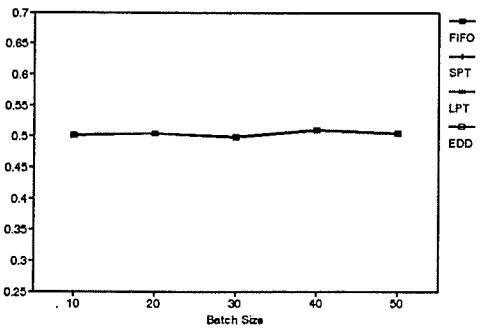
FJF - Flow Time CV



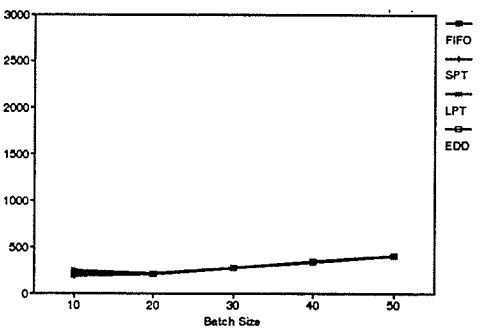
FJF - Percent Tardy



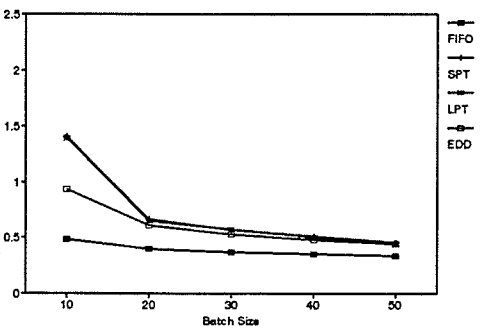
JFJ - Resource Utilization



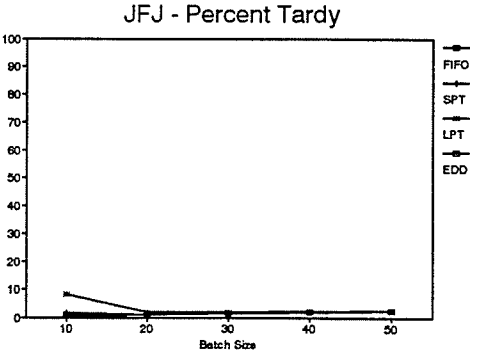
JFJ - Flow Time

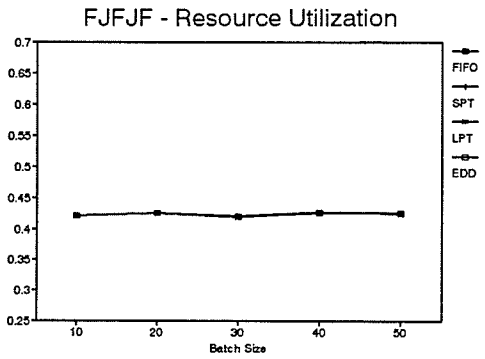


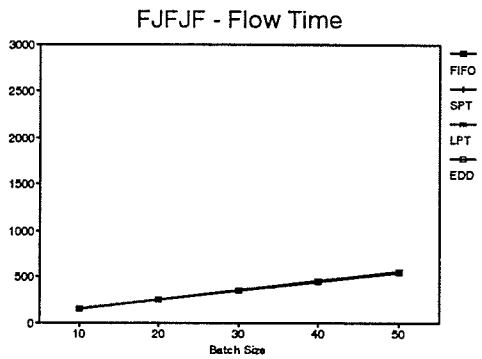
JFJ - Flow Time CV

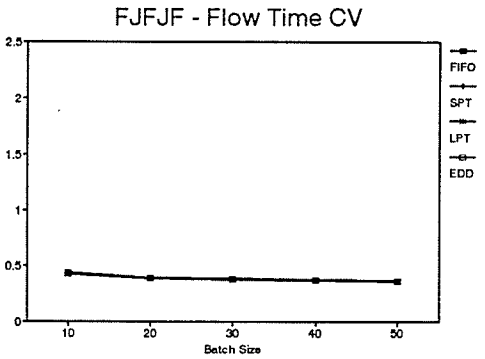


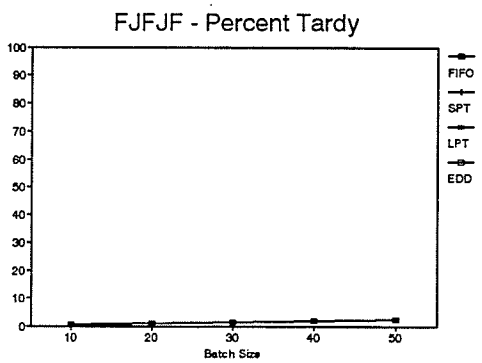
JFJ - Percent Tardy




FJFJF - Resource Utilization
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

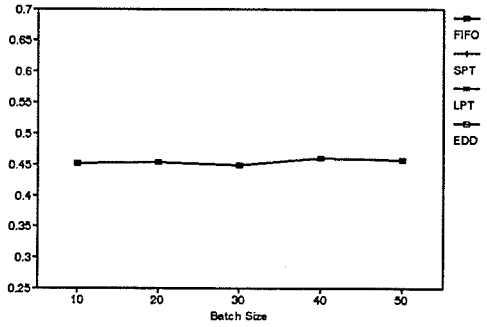

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

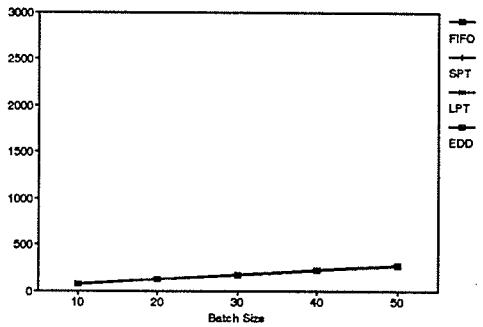
Normal service at the job shop node

Low demand, low set-up, high variability

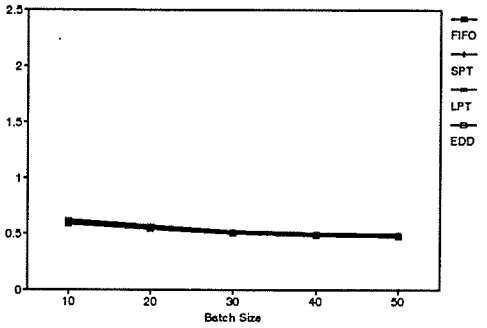
FJ - Resource Utilization



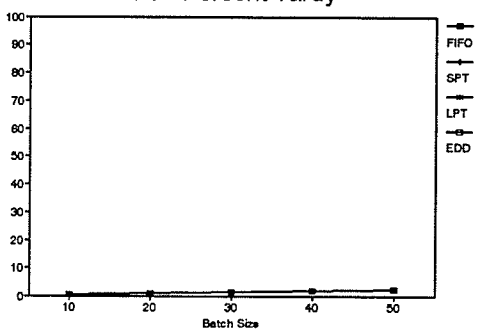
FJ - Flow Time



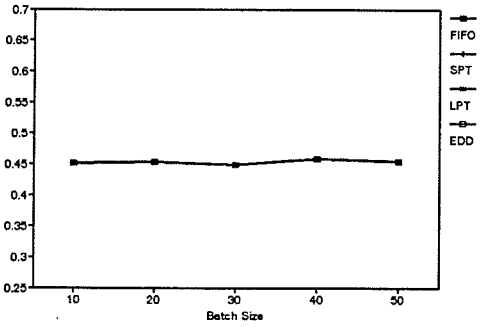
FJ - Flow Time CV



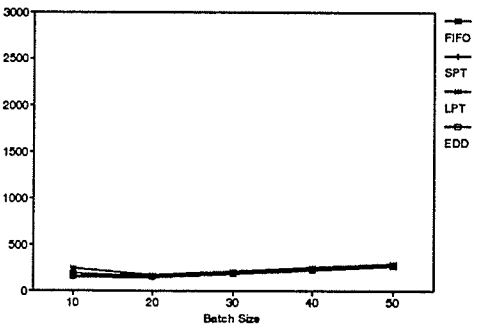
FJ - Percent Tardy



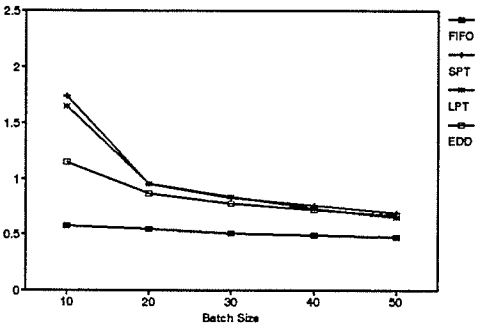
JF - Resource Utilization



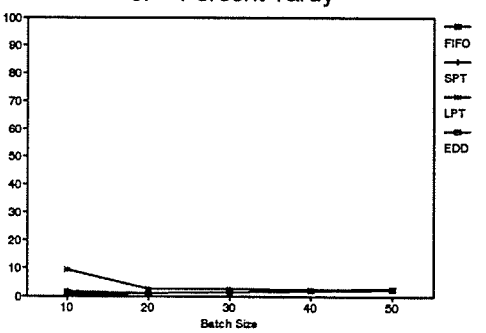
JF - Flow Time



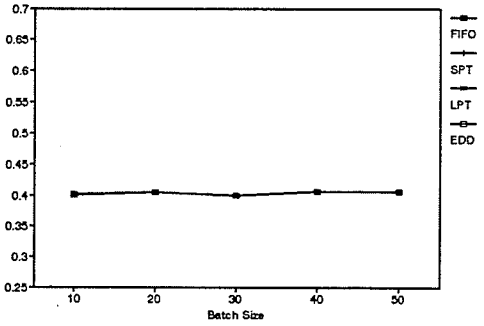
JF - Flow Time CV



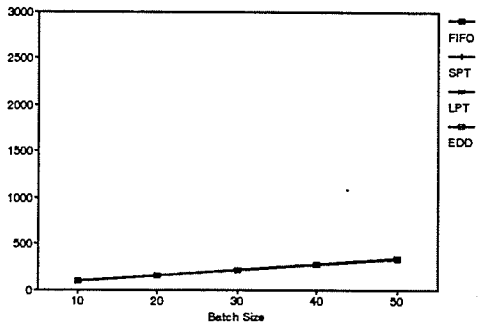
JF - Percent Tardy



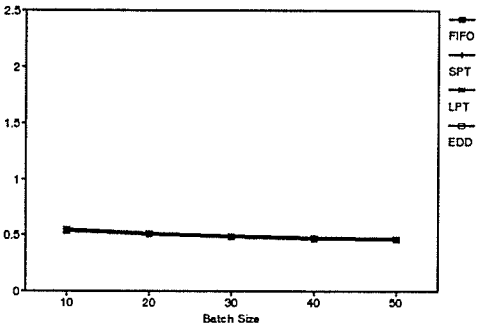
FJF - Resource Utilization



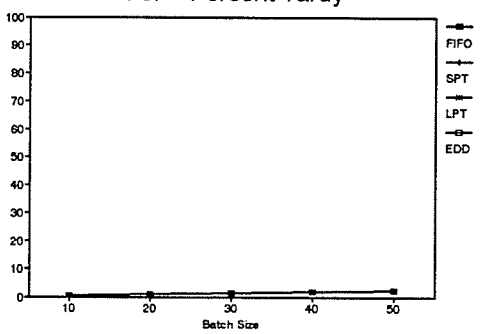
FJF - Flow Time



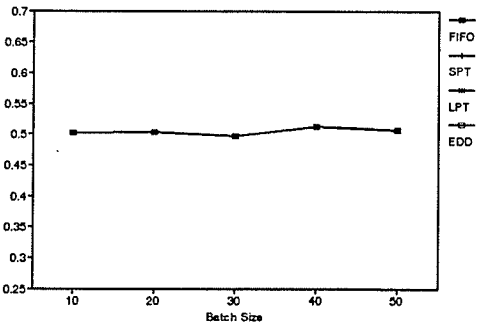
FJF - Flow Time CV



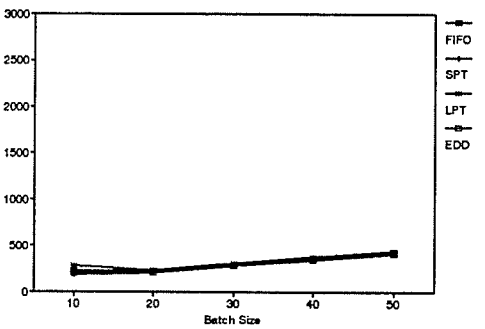
FJF - Percent Tardy



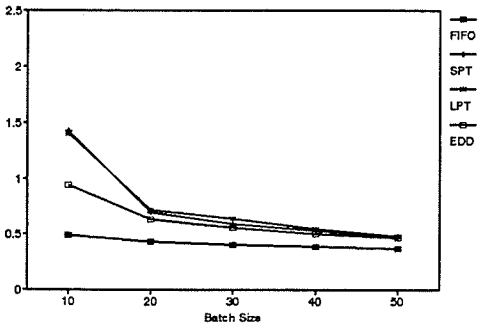
JFJ - Resource Utilization



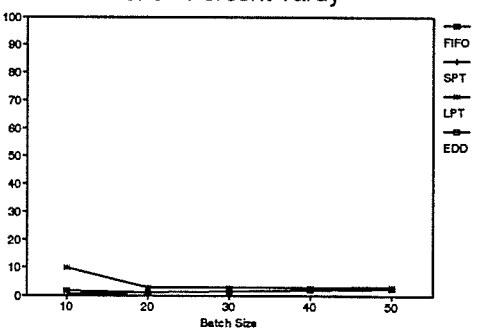
JFJ - Flow Time

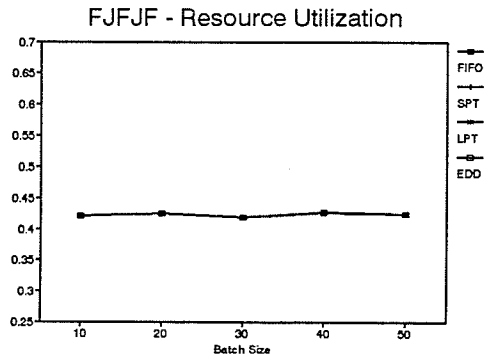


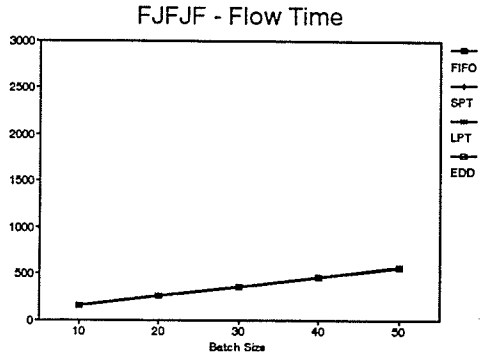
JFJ - Flow Time CV

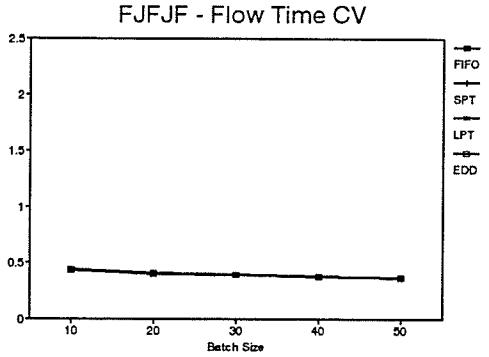


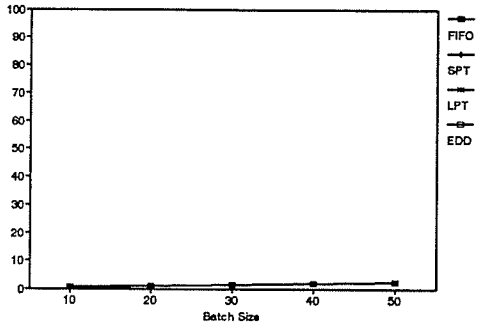
JFJ - Percent Tardy




FJFJF - Resource Utilization
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

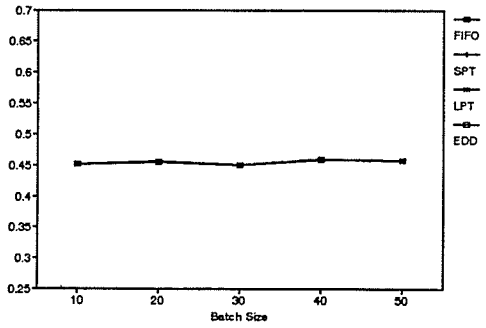

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

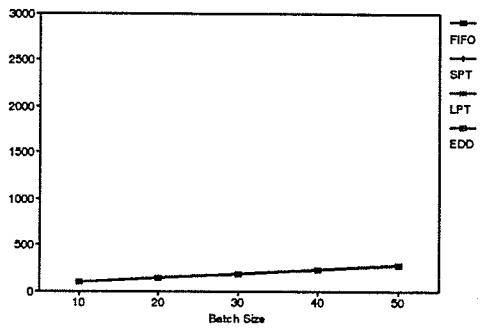
Normal service at the job shop node

Low demand, high set-up, low variability

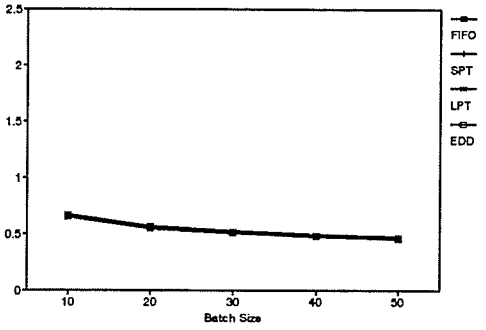
FJ - Resource Utilization



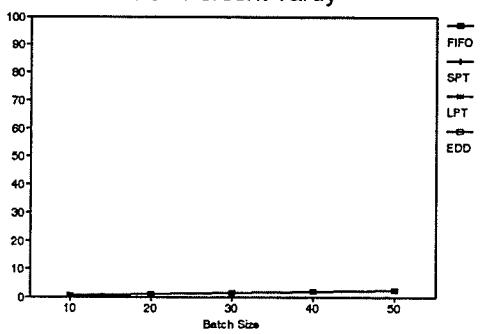
FJ - Flow Time



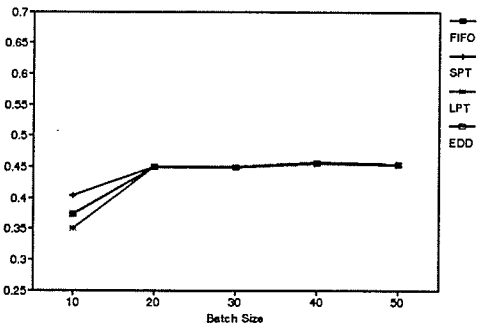
FJ - Flow Time CV



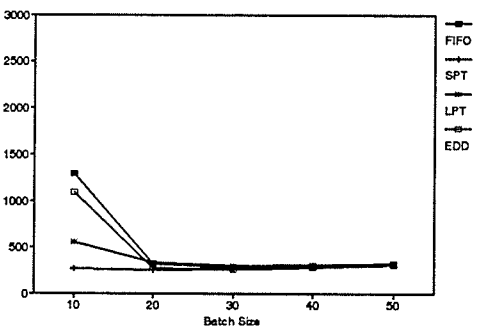
FJ - Percent Tardy



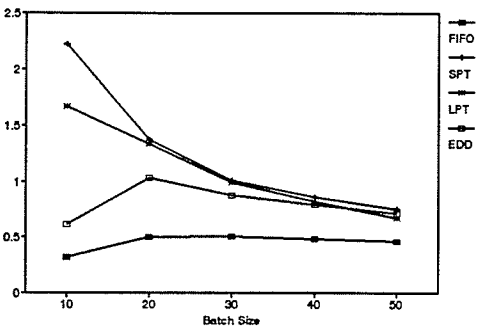
JF - Resource Utilization



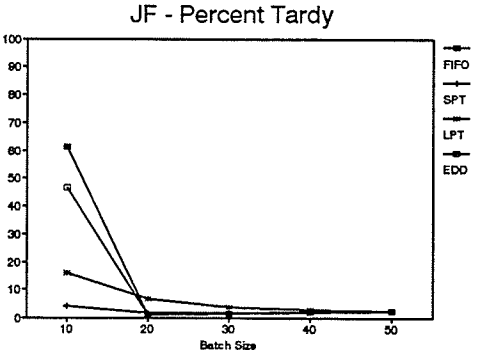
JF - Flow Time



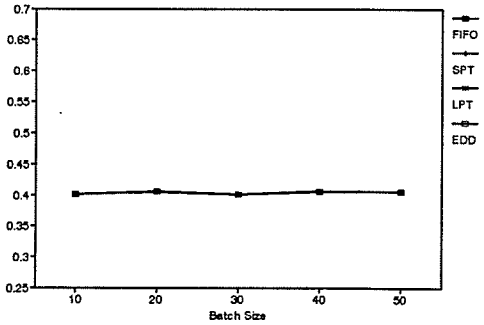
JF - Flow Time CV



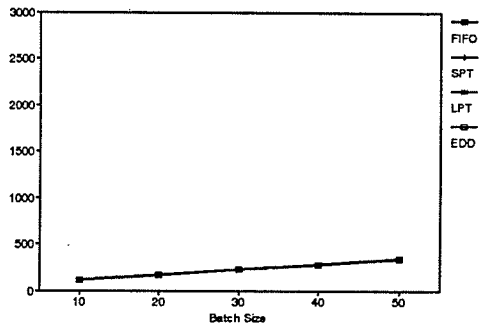
JF - Percent Tardy



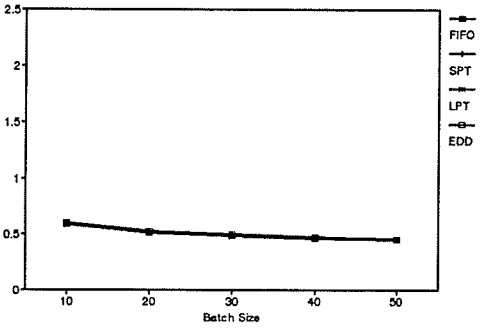
FJF - Resource Utilization



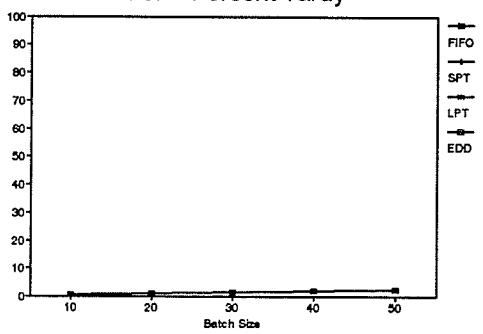
FJF - Flow Time



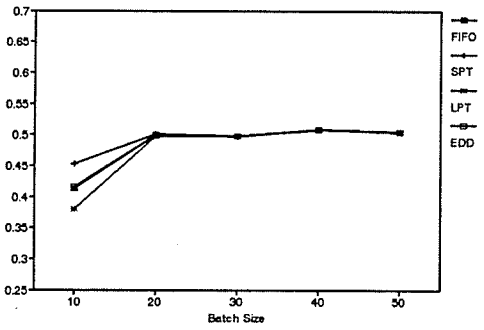
FJF - Flow Time CV



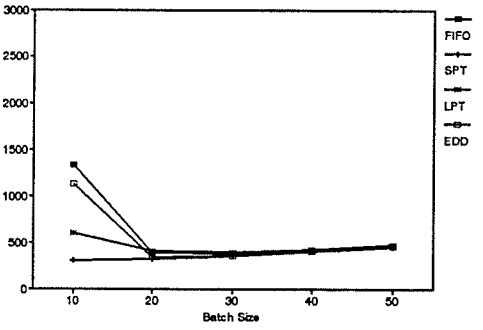
FJF - Percent Tardy



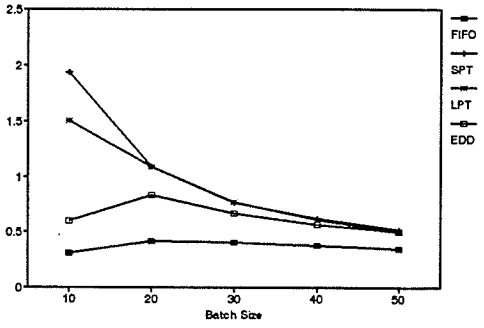
JFJ - Resource Utilization



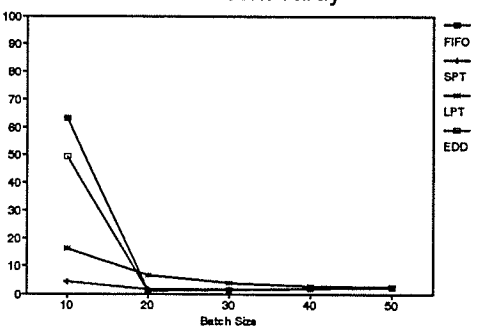
JFJ - Flow Time

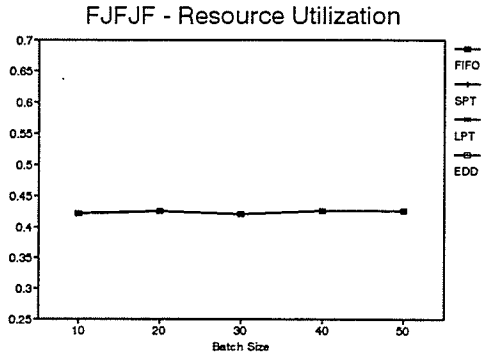


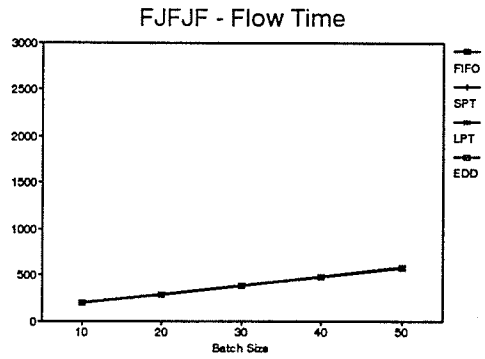
JFJ - Flow Time CV

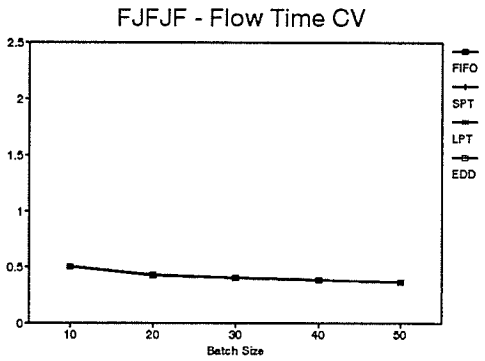


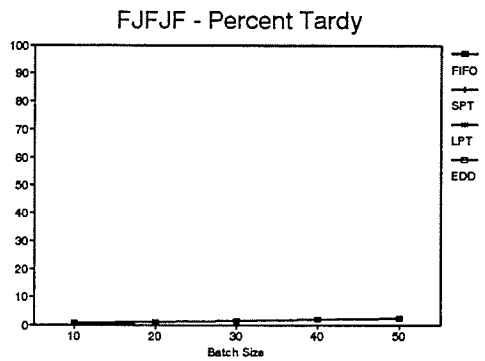
JFJ - Percent Tardy




FJFJF - Resource Utilization
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

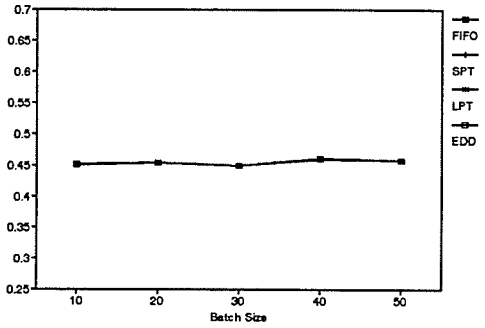

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

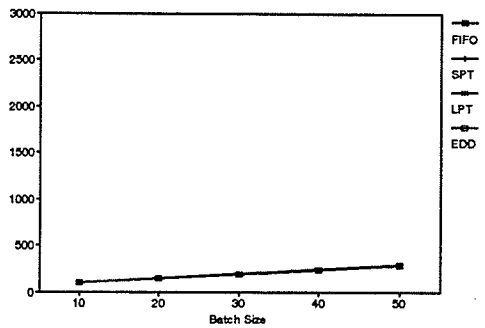
Normal service at the job shop node

Low demand, high set-up, high variability

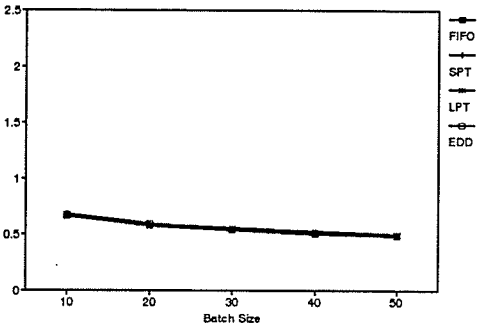
FJ - Resource Utilization



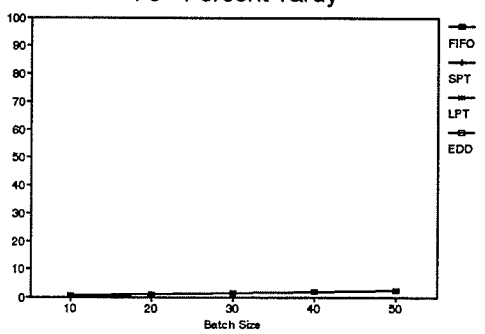
FJ - Flow Time



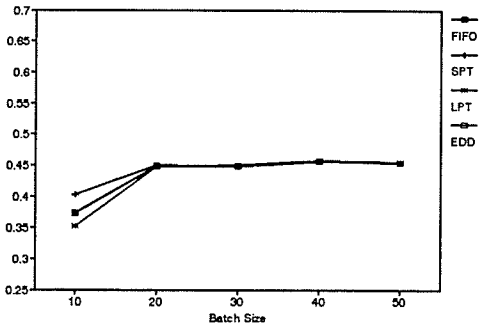
FJ - Flow Time CV



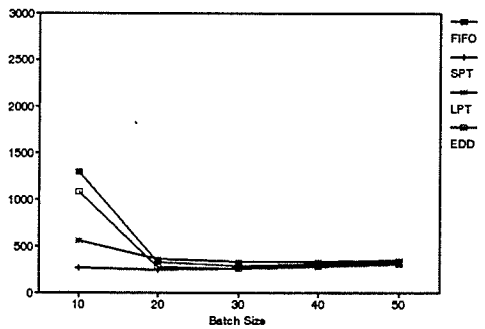
FJ - Percent Tardy



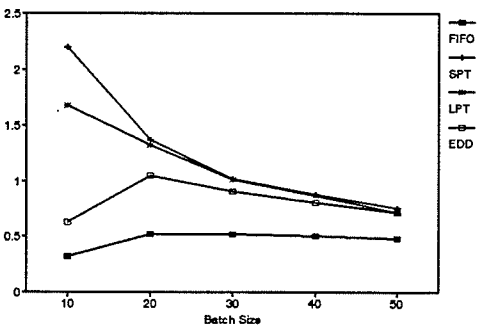
JF - Resource Utilization



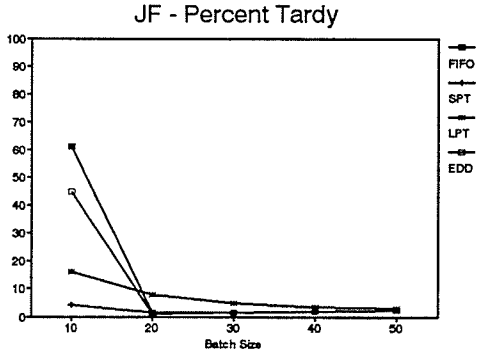
JF - Flow Time



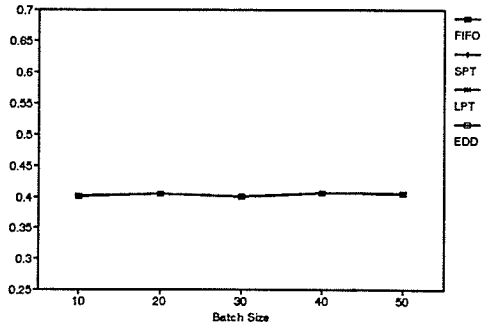
JF - Flow Time CV



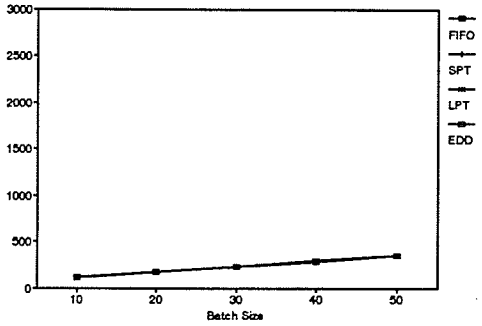
JF - Percent Tardy



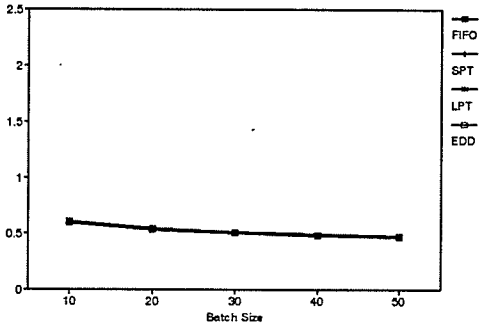
FJF - Resource Utilization



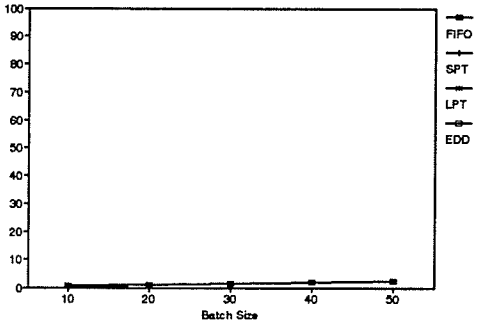
FJF - Flow Time



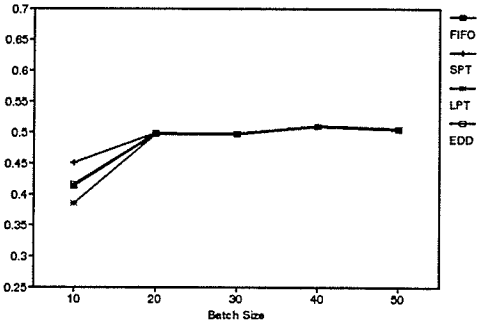
FJF - Flow Time CV



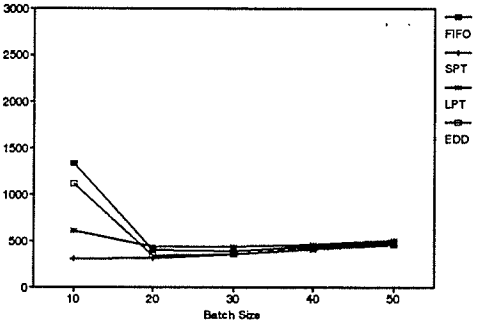
FJF - Percent Tardy



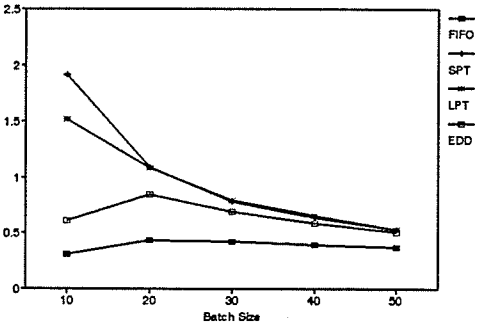
JFJ - Resource Utilization



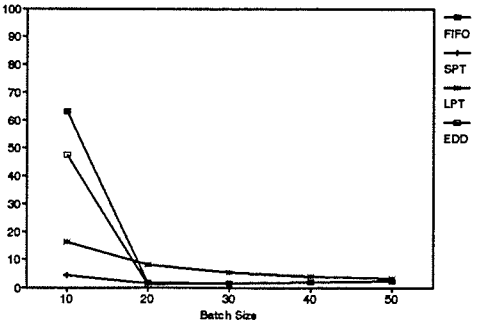
JFJ - Flow Time



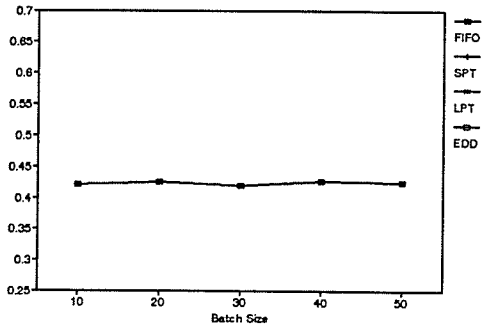
JFJ - Flow Time CV



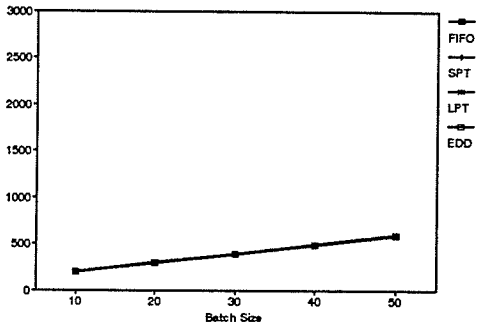
JFJ - Percent Tardy



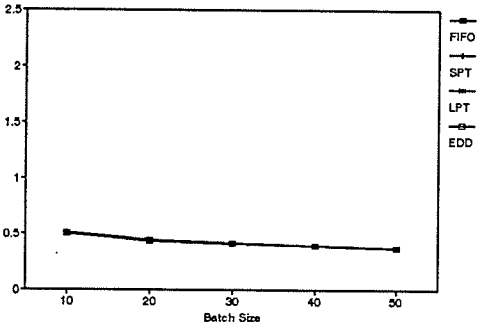
FJFJF - Resource Utilization



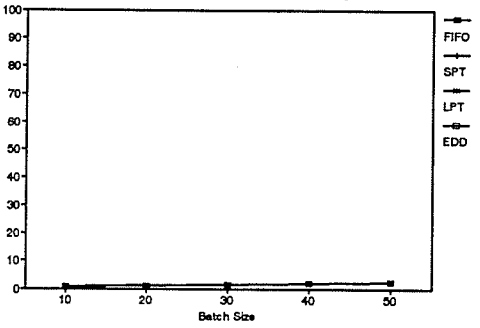
FJFJF - Flow Time



FJFJF - Flow Time CV



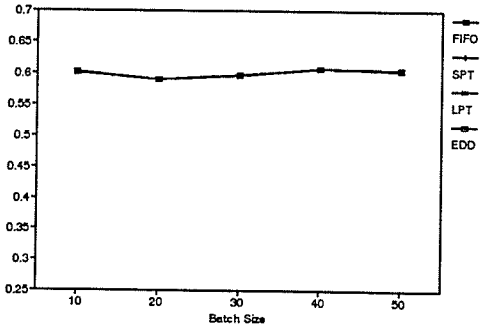
FJFJF - Percent Tardy



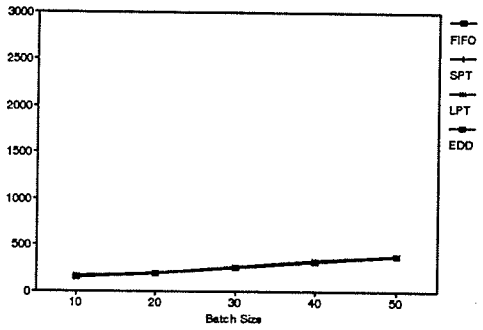
Normal service at the job shop node

High demand, low set-up, low variability

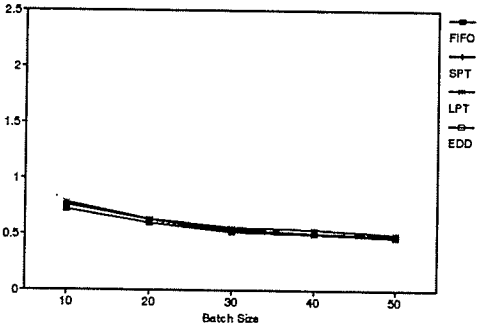
FJ - Resource Utilization



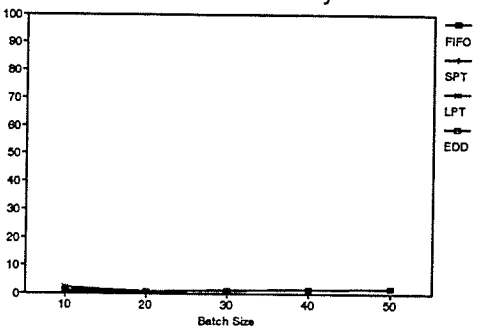
FJ - Flow Time



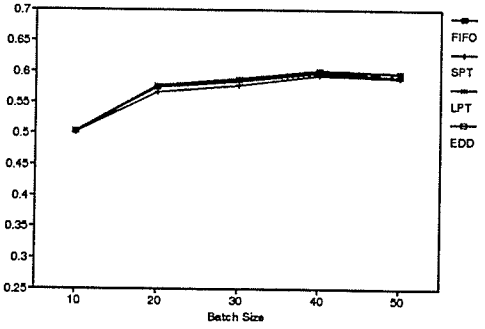
FJ - Flow Time CV



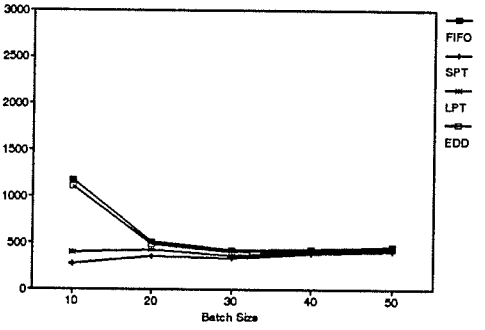
FJ - Percent Tardy



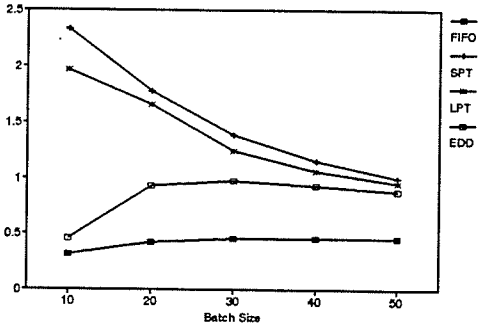
JF - Resource Utilization



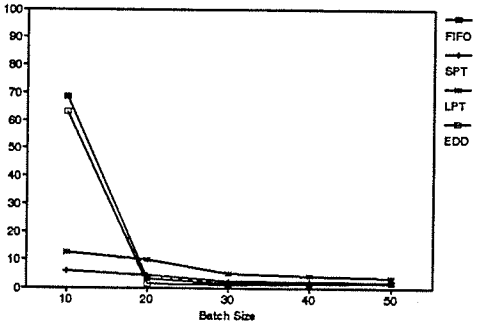
JF - Flow Time



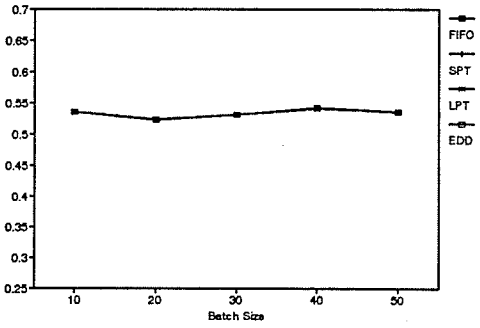
JF - Flow Time CV



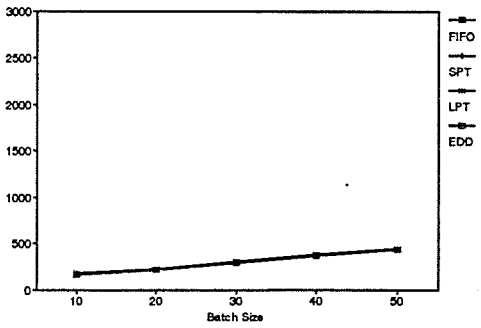
JF - Percent Tardy



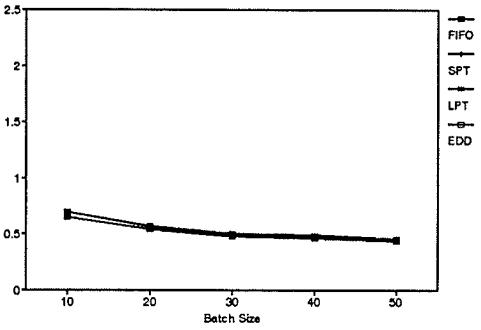
FJF - Resource Utilization



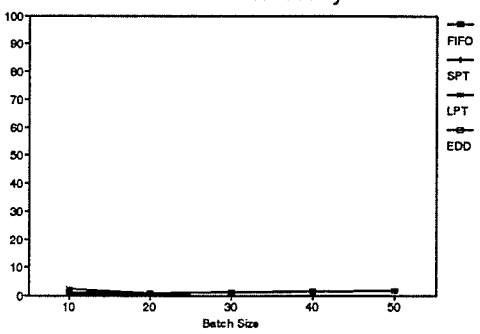
FJF - Flow Time



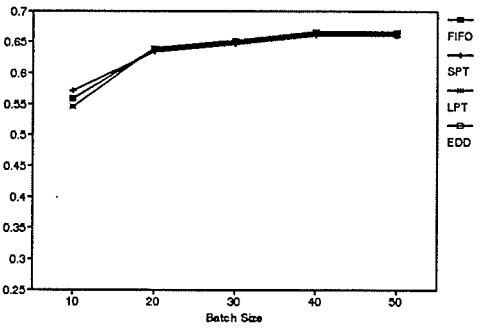
FJF - Flow Time CV



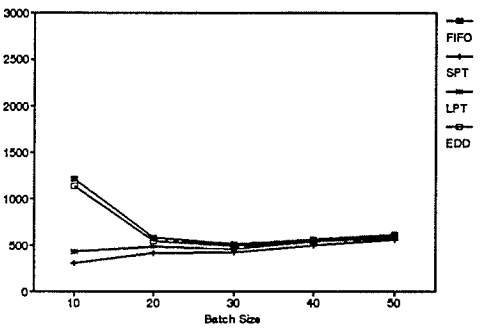
FJF - Percent Tardy



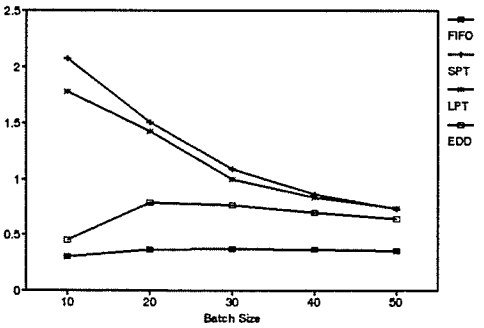
JFJ - Resource Utilization



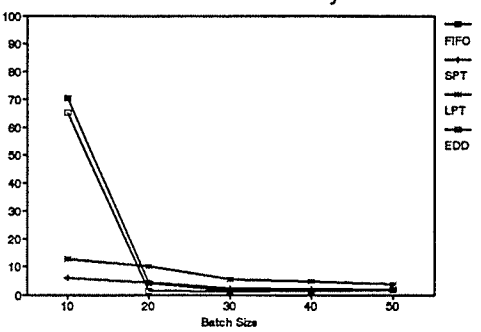
JFJ - Flow Time



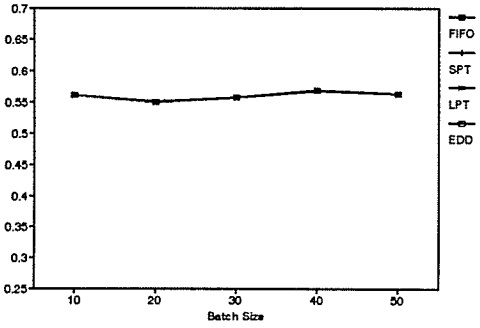
JFJ - Flow Time CV



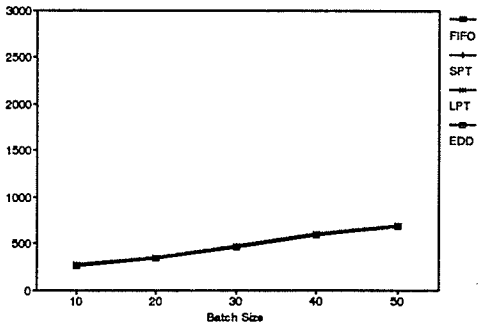
JFJ - Percent Tardy



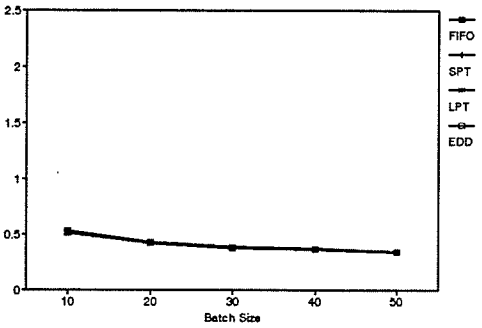
FJFJF - Resource Utilization



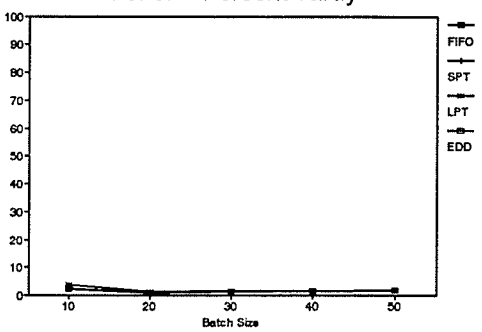
FJFJF - Flow Time



FJFJF - Flow Time CV



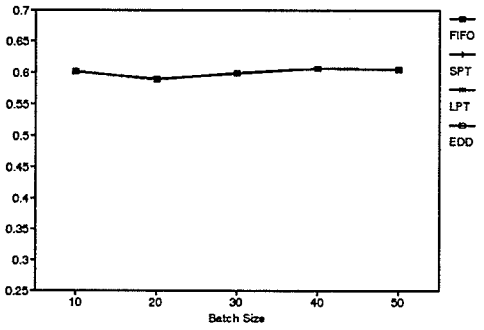
FJFJF - Percent Tardy



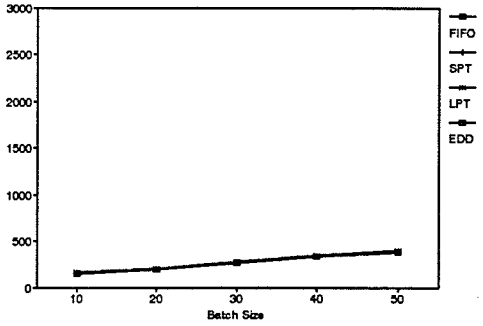
Normal service at the job shop node

High demand, low set-up, high variability

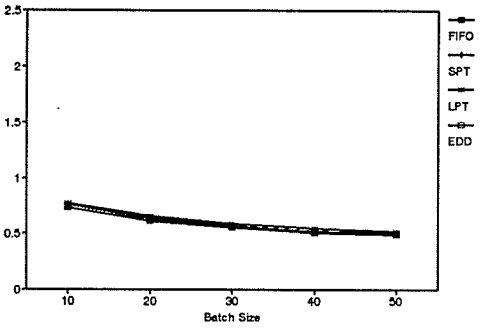
FJ - Resource Utilization



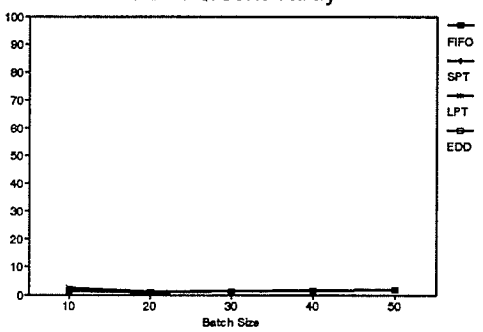
FJ - Flow Time



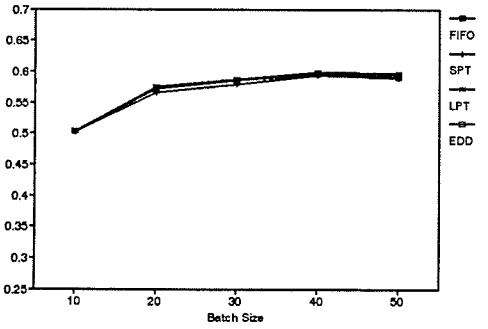
FJ - Flow Time CV



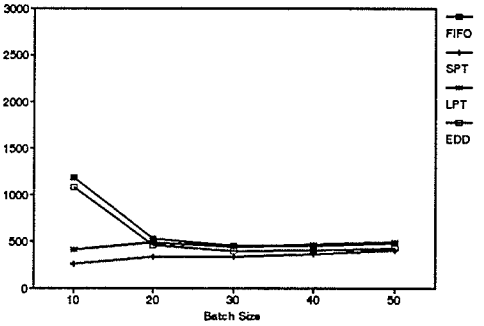
FJ - Percent Tardy



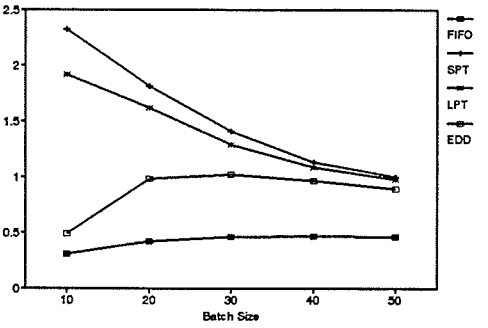
JF - Resource Utilization



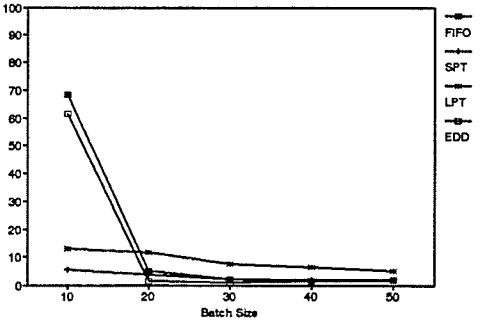
JF - Flow Time



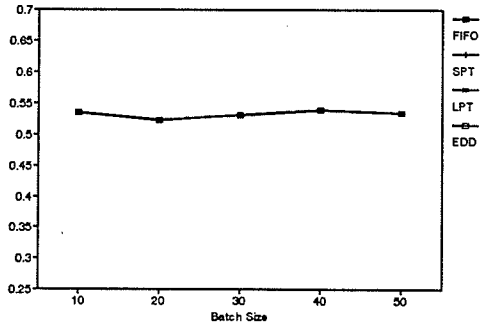
JF - Flow Time CV



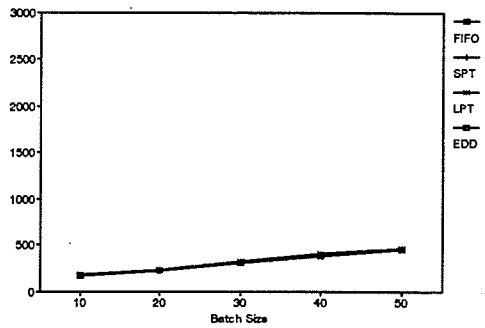
JF - Percent Tardy



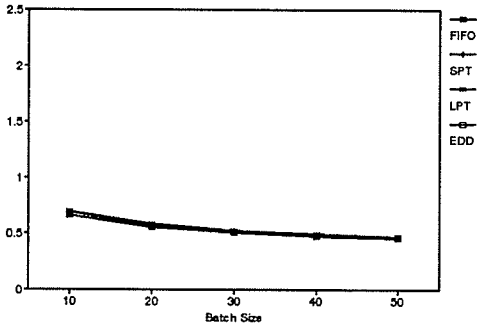
FJF - Resource Utilization



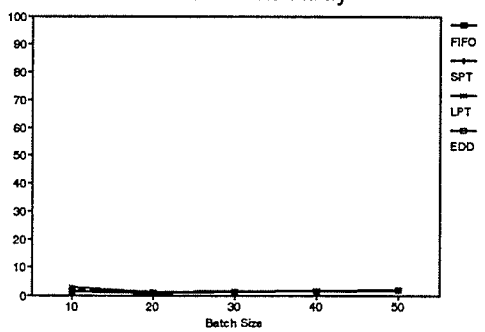
FJF - Flow Time



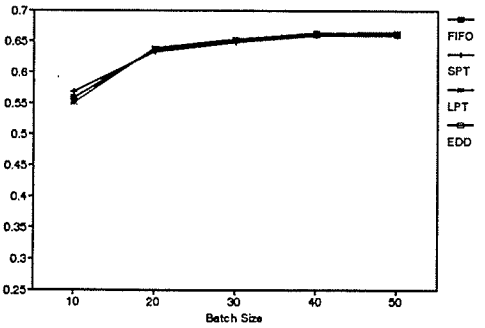
FJF - Flow Time CV



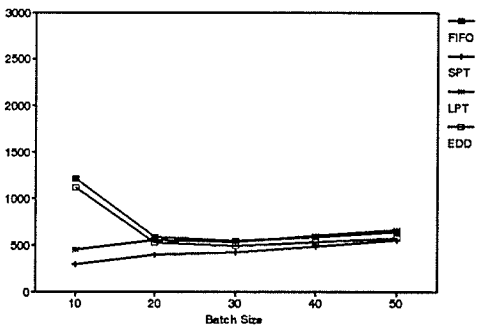
FJF - Percent Tardy



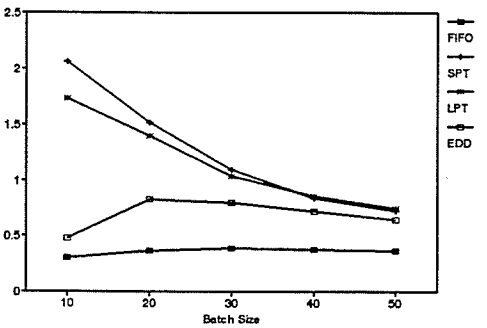
JFJ - Resource Utilization



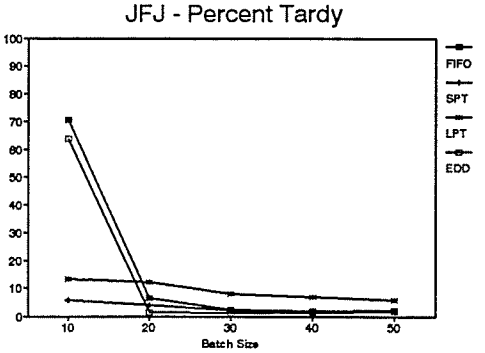
JFJ - Flow Time

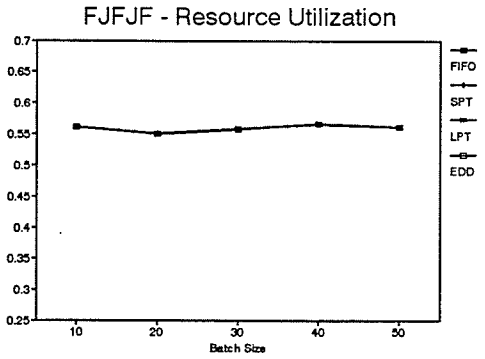


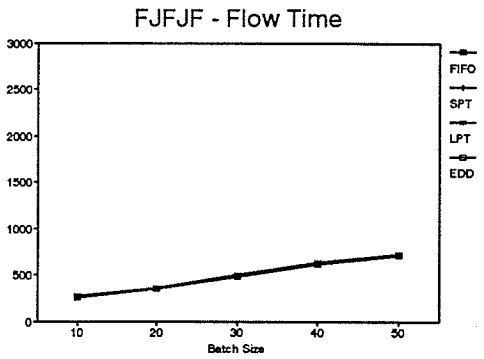
JFJ - Flow Time CV

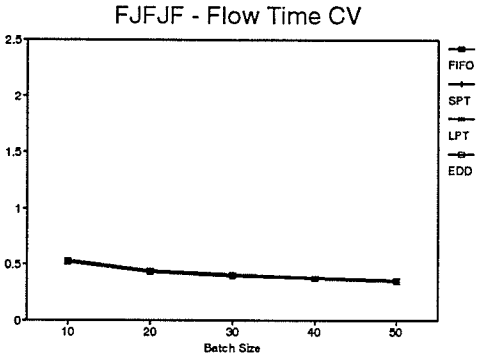


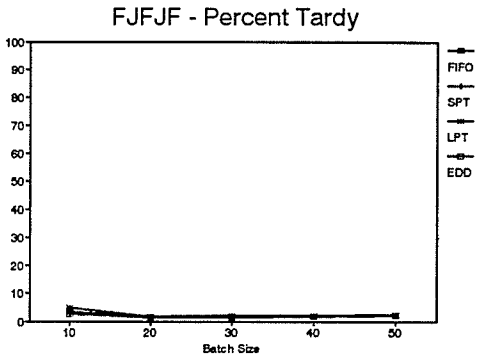
JFJ - Percent Tardy




FJFJF - Resource Utilization
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Flow Time
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

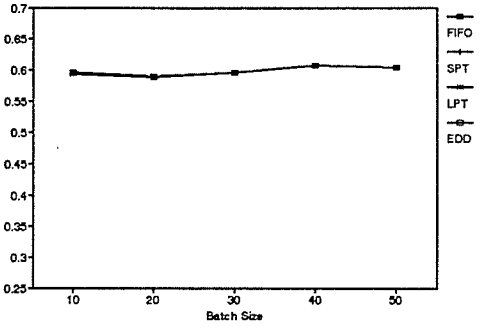

FJFJF - Flow Time CV
2.5
2
1.5
1
0.5
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD


FJFJF - Percent Tardy
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
Batch Size
FIFO
SPT
LPT
EDD

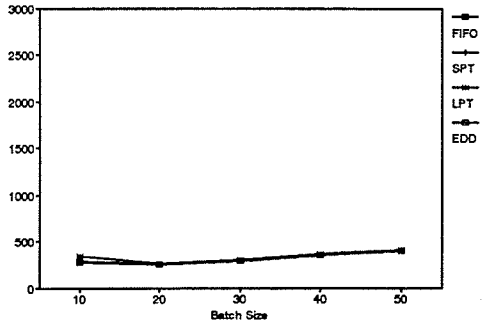
Normal service at the job shop node

High demand, high set-up, low variability

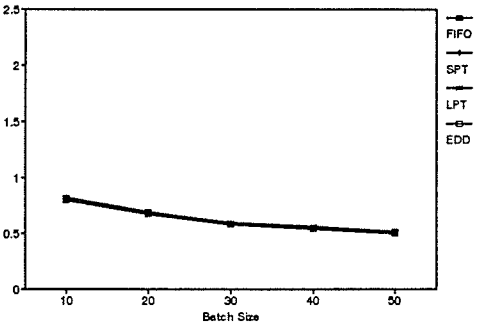
FJ - Resource Utilization



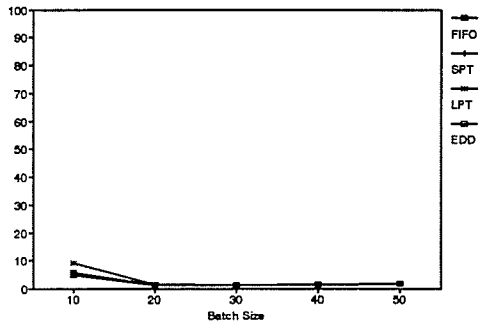
FJ - Flow Time



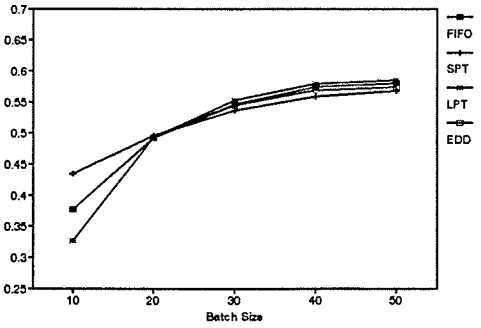
FJ - Flow Time CV



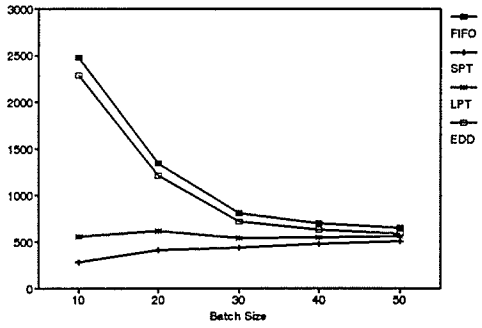
FJ - Percent Tardy



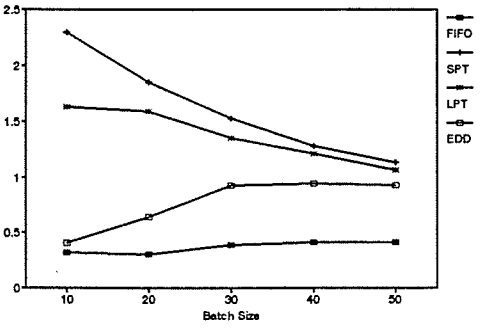
JF - Resource Utilization



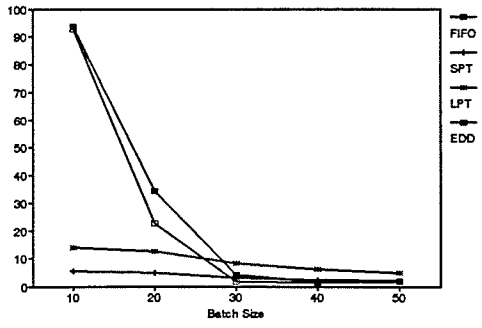
JF - Flow Time



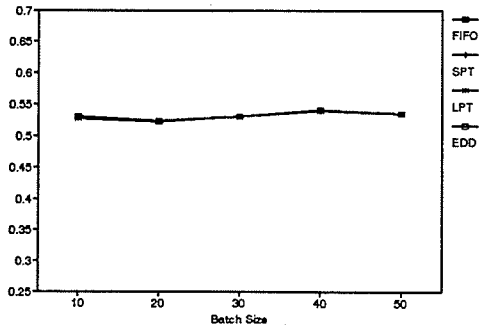
JF - Flow Time CV



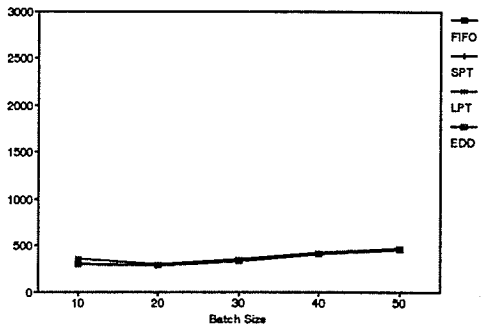
JF - Percent Tardy



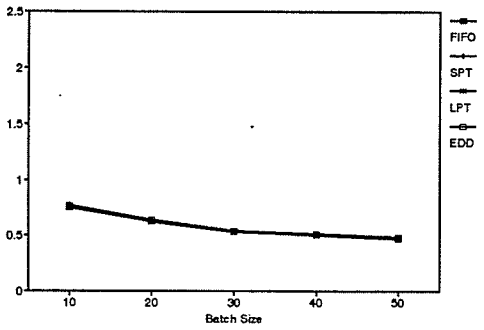
FJF - Resource Utilization



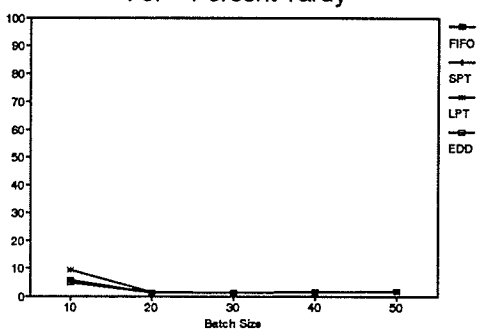
FJF - Flow Time



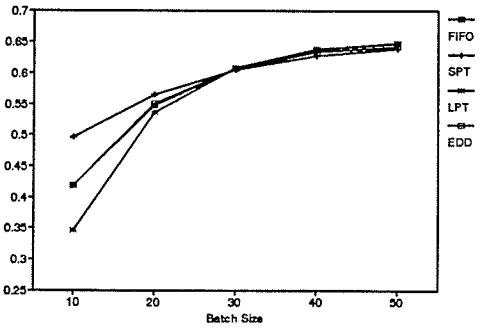
FJF - Flow Time CV



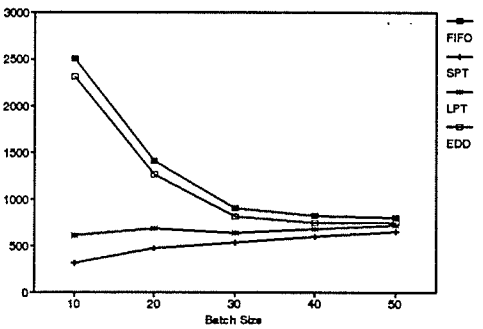
FJF - Percent Tardy



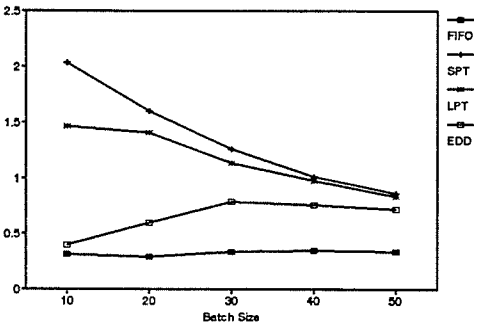
JFJ - Resource Utilization



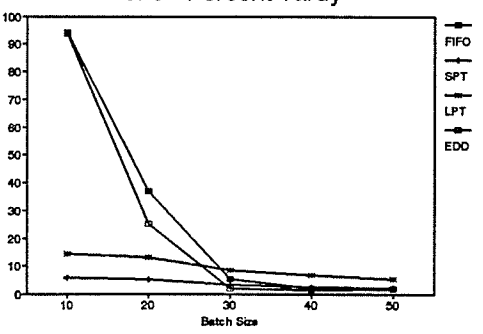
JFJ - Flow Time



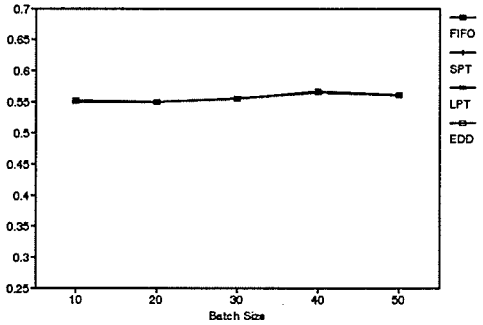
JFJ - Flow Time CV



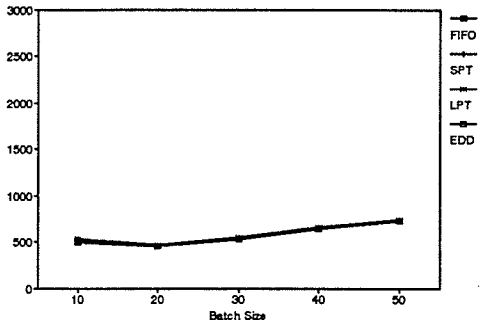
JFJ - Percent Tardy



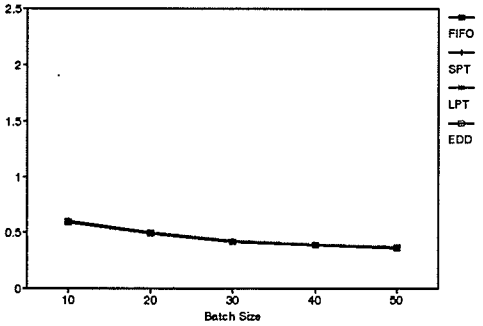
FJFJF - Resource Utilization



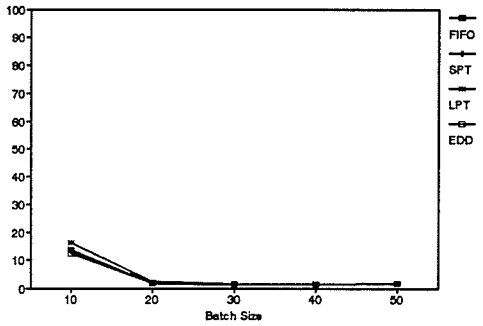
FJFJF - Flow Time



FJFJF - Flow Time CV



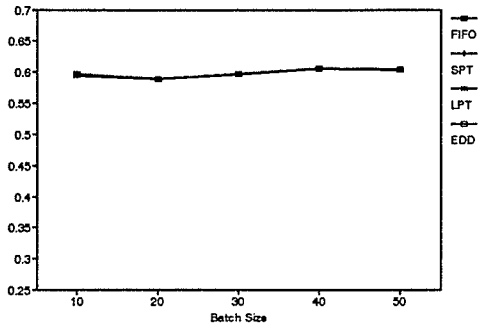
FJFJF - Percent Tardy



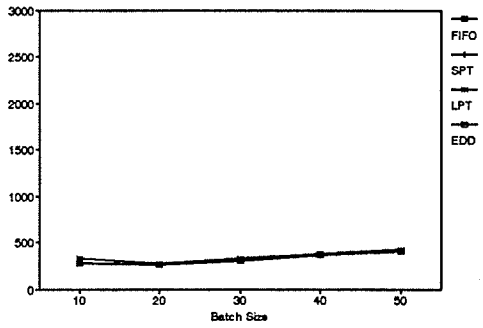
Normal service at the job shop node

High demand, high set-up, high variability

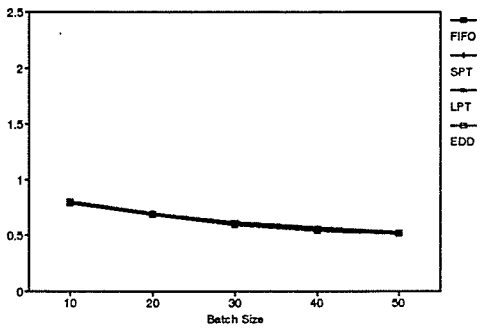
FJ - Resource Utilization



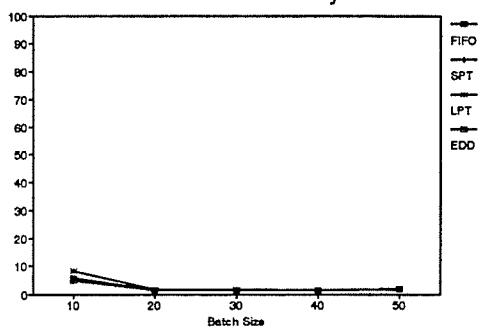
FJ - Flow Time



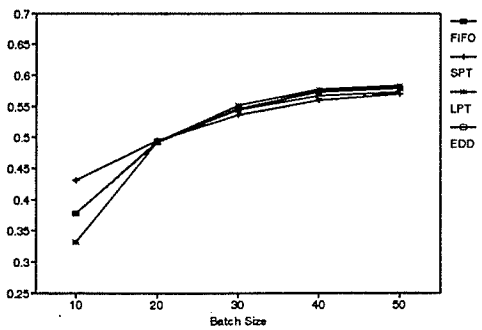
FJ - Flow Time CV



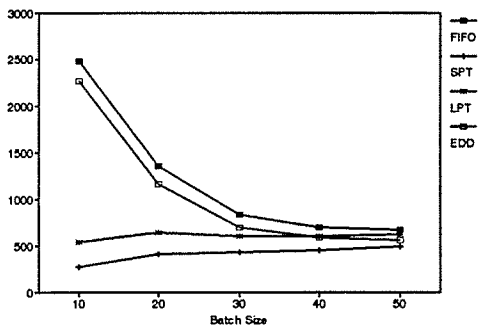
FJ - Percent Tardy



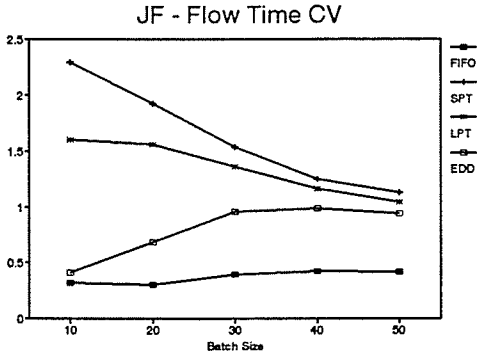
JF - Resource Utilization



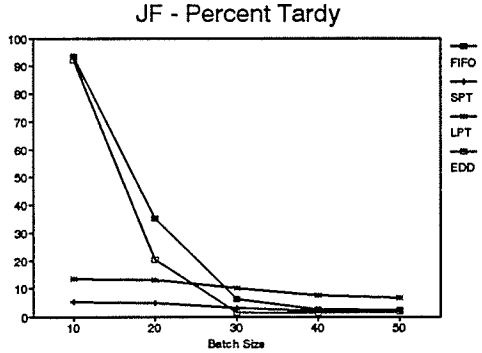
JF - Flow Time



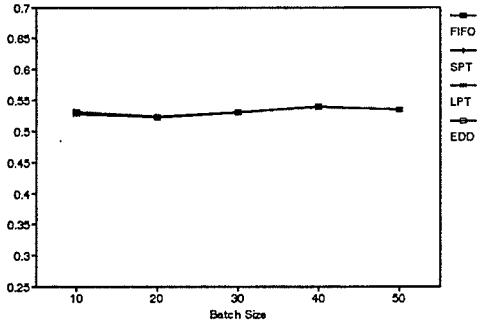
JF - Flow Time CV



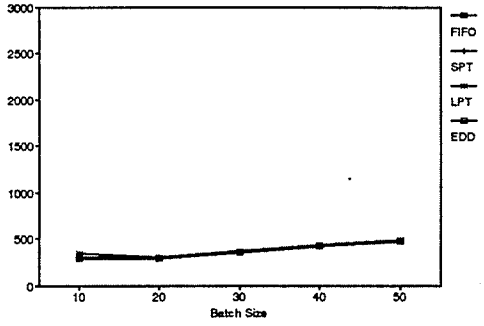
JF - Percent Tardy



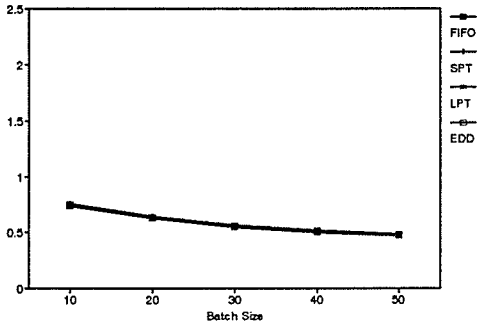
FJF - Resource Utilization



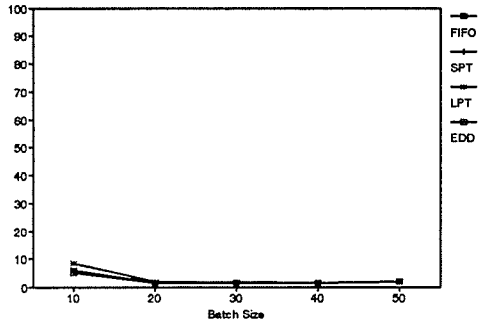
FJF - Flow Time



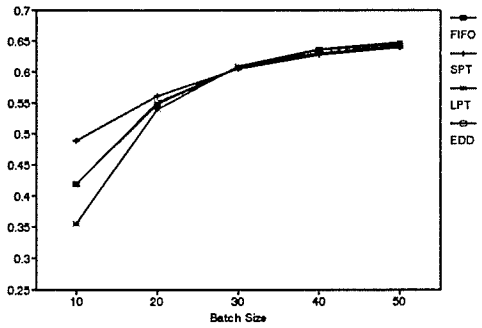
FJF - Flow Time CV



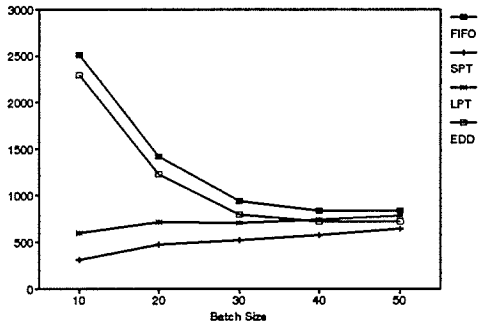
FJF - Percent Tardy



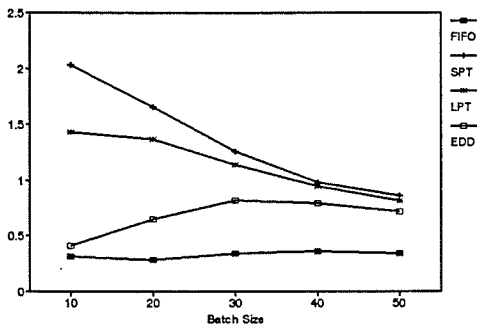
JFJ - Resource Utilization



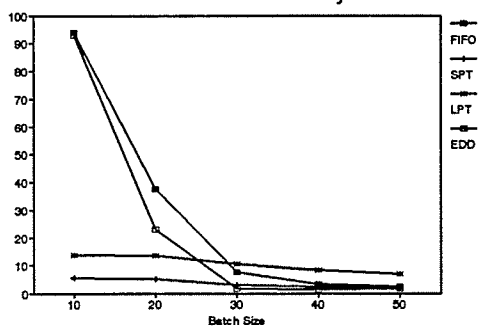
JFJ - Flow Time



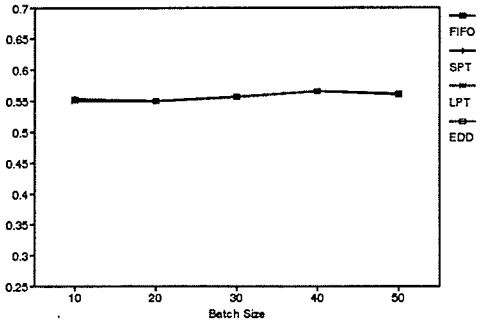
JFJ - Flow Time CV



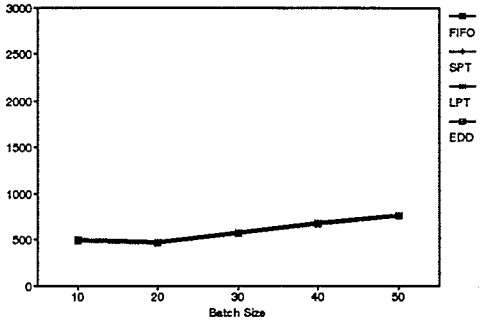
JFJ - Percent Tardy



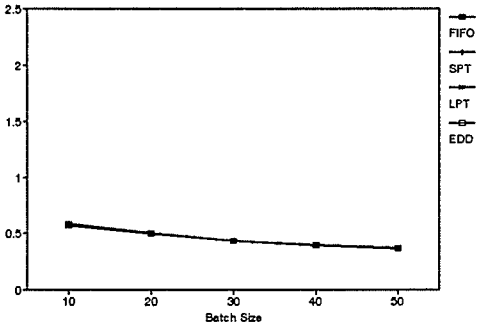
FJFJF - Resource Utilization



FJFJF - Flow Time



FJFJF - Flow Time CV



FJFJF - Percent Tardy