MANAGING CAPACITY IN A SERVICE ENVIRONMENT

by

Kristy Wanliss

24191583

Submitted in partial fulfilment of the requirements for the degree

of

BACHELORS OF INDUSTRIAL ENGINEERING

in the

FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY

UNIVERSITY OF
PRETORIA

October 2009

CONTENTS

1 Introduction 4
2 Background...6
3 The Current Process 9
4 Problem Description..11
5 Literature Review 12
5.1 Capacity Management 12
5.2 Theory of Constraints (TOC) 18
5.3 Queuing Theory 23
5.4 Literature Review Conclusion...28
6 Factors Affecting the Process 29
7 Data Collection and Analysis..31
7.1 Data Source..31
7.2 Data Collection 33
7.3 Data Analysis..36
8 The Recommended Policy 39
9 Policy Analysis..44
10 Conclusion...46
11 References 47
Appendix A 49
Appendix B 54
Appendix C 56

LIST OF TABLES AND FIGURES

Figure 1 Five focusing steps 18
Figure 2 System representation of the health risk management company....................22
Figure 3 Flow of applications near and including doctor assessment stages....................32
Figure 4 The number of received applications versus the number of finalised applications 35
Figure 5 Current Process versus Proposed Process: Time Savings....................45
Figure 6 Histogram of the processing times for Scanning 49
Figure 7 Histogram of the processing times for Capturing....................49
Figure 8 Histogram For PreAssess 49
Figure 9 Histogram Doctor Assessment....................50
Figure 10 Histogram Typing Final Report....................50
Figure 11 Histogram Proofread 50
Figure 12 Histogram Quality and Conformance Review 1 51
Figure 13 Histogram Quality and Conformance Review 2 51
Figure 14 Histogram Corrections Typist 51
Figure 15 Histogram Corrections Checker 52
Figure 16 Histogram Final Scan....................52
Figure 17 Histogram Final Report to Employer....................52
Figure 18 Histogram File Distribution 53
Figure 19 Flow of Short Temporary Incapacity Leave Applications 54
Figure 20 Flow of Long Temporary Incapacity Leave Applications....................55
Figure 21 The Overall view of the simulation model 56
Figure 22 Entity arrivals: Actual arrival times of applications being fed into Arena....................57
Figure 23 The Initial Stages of the System....................58
Figure 24 The middle phase where a time buffer is present 59
Figure 25 The final phases of the processing system....................60

1 INTRODUCTION

The company where the report is based is a health risk management company who provide a recommendation based on an investigation regarding the ill-health retirement and incapacity leave for the contracted clients' employees.

For the health risk managers main contracted client there is a perception that the number of temporary incapacity leave and ill-health retirement applications increase in the third year of the sick leave period. This increase places enormous pressure on the current infrastructure and the agreed to delivery dates.

The work involves medically trained staff to assess applications and report findings for the contracted client within the confinement of a defined company process. The combination of medical practitioner proficiency and the internal process to the company excludes the option of obtaining temporary staff within such a short period of time to address the random submissions of applications and unexpected increase in work load.

Applications arrive at the health risk management company at random intervals of varying quantities. Once the applications have arrived, their arrival needs to be acknowledged within a specified time frame; thereafter the applications enter the system and wait to be processed.

The doctors who make the assessments and recommendations on the applications have been found to be a resource that is not easily scheduled due to personal commitments these doctors may have. The biggest bottleneck in the system is found to be here, where the doctors make an assessment and recommendation regarding the application.

The waiting of applications for processing can be viewed as entities within a system - a queuing system. The entities within the system (the applications) require processing by some or other resource that is part of the system (the medical practitioner). The applications may not (according to a service level agreement with the contracted client) be in the system longer than a specified time period.

Currently the company suffers a backlog and may face penalties for not processing applications and making recommendations regarding them in a timely manner. Capacity constraints within the organisation need to be identified and exploited. Managing capacity, which includes both the planning and control of capacity, can be utilised along with the principles found within the Theory of Constraints to achieve efficiency in the internal processes at this health risk management company.

2 BACKGROUND

The company where the study is based is an independent, health-risk management organisation who (mainly) assesses the applications made for ill-health retirement and temporary incapacity leave for their clients' employees. The necessity for independent health risk management organisations is to ensure the employee is treated fairly if he/she has valid reasons for exhausting the sick leave and also in an attempt to curb the increasing number of staff who attempt to take advantage of the allocated sick leave within their employment contract.

Temporary incapacity leave is paid sick leave that is granted, upon application, to employees who have exhausted their allocated sick leave. Ill-health retirement is granted to those employees who find themselves unable to conduct their work due to severe and/or permanent illness. The assessment of an ill-health retirement application includes the investigation of possible re-employment of the employee within the organisation with a job description that is more suited to the employee/patients condition.

The company's client allocates sick leave according to their own cycle and is not governed by the start of the employees' employment. This cycle is a 3 year cycle and each employee is granted a standard 36 days of leave within this cycle. This is explained with the use of an example:

The company begins a new 3 year cycle at the start of January 2007. A person who commences employment in this month is allocated 36 days sick leave that may be used in the period of January 2007 to December 2009. If the person exhausts this sick leave, he/she may apply for temporary incapacity leave or ill-health retirement depending on the personal situation. If, however, a person commences employment at the start of December 2006, he/she is able to use the full 36 day sick leave period which expires at the end of December 2006 and he/she will be allocated another 36 days sick leave at the start of the next cycle in January 2007. The application is processed and assessed by the health risk manager.

The existence of the client's 3 year sick leave cycle means a similar cycle is also generated within the particular health-risk management company. This cycle is identified by the increase in applications for ill-health retirement and temporary incapacity leave. This increase is due to a number of employees who exhaust their leave toward the end of the cycle. This increase creates an increase in demand on capacity within this health risk management firm.

The company has a service level agreement with its client which states the recommendation made regarding the application needs to be made within specified time frames or else penalties will be incurred. The recommendations are treated as legal documents and recommendations (in the form of a report) and therefore must be of the highest quality and standard. This quality cannot simply be reduced in order to process more applications in a shortened period of time.

Applications are made according to the following needs of the patients:

Long term incapacity leave;
Short term incapacity leave; and
Ill-health retirement.

Applications may require a secondary opinion, where the consultation with and the advice of a specialist doctor is sought. These are referred to as *secondary applications*, while applications that do not require the advice of a specialist doctor are referred to as *primary applications*.

The service level agreement specifies that the applications, along with a recommendation are to be returned within certain time frames. This times frames are listed below:

Long term and short term incapacity leave that is primary: 12 working days;
Long term and short term incapacity leave that is secondary: 40 working days; and
Ill health retirement applications: 90 working days.

The majority of the doctors that are appointed at the company do not work on the premises, or according to any predetermined service level with the health risk manager. The doctors usually are concurrently furthering their studies and do not have a full work day available to process applications. Thus the trend is that doctors arrive at random times to collect a random number of applications for assessment. The applications are returned once the doctor has completed the assessment. This creates an uncertainty for the health risk manager in determining the resources and capacity available at any given time.

It is envisaged that the proposed policy will be one that will enable the company to create sufficient, flexible capacity able to handle increases in demand toward the end of the 3 year cycle. The proposed policy should consider the service level agreements made with the client as well as a basic understanding of the associated costs with increasing capacity.

3 THE CURRENT PROCESS

Applications arrive mostly via courier delivery to the health risk manager. The applications arrive daily and are spread throughout the day, although for the purposes of this study it is assumed that all applications received in a day arrive at once, at the start of the day. Several processes must be completed before a report containing the final recommendation can be returned to the client. Initial processes such as those that involve inspection of whether the application is complete are not included in this study. It is thus assumed that all details and required information is correct and complete at the start of the processes that fall under this analysis.

The application must first be entered into the system after which it must be decided whether the application is a *primary* or a *secondary* application. If it is decided the application is *secondary*, the application and patient are referred to a specialist doctor. If this is the case, the time allowed to make the recommendation is extended.

Primary and *secondary* applications are then assessed by a doctor appointed by the health risk manager and a recommendation is made. The recommendation is given in the form of a report. This report then undergoes an intense process of proofreading (by professionals of both language and medicine) and further capturing before it is returned to the client.

The time needed for the doctor to make an assessment and present a report is largely unknown due to the random collection of applications as well as the random number of applications that the doctor collects. Due to this, the processing of applications made by the doctors seems to be in the vicinity of 30 minutes to 5 hours.

As a result of these long processing times and the uncertainty with regard to the processing time of applications made by the doctors, long waiting times for assessment is common.

The company have made various revisions to the overall processes within the organisation in an effort to improve operations, as well as the installation of an ERP (Enterprise Resource Planning) software system. The ERP system has enabled the company to trace processes and applications and to keep stricter control over the processing of the applications.

The data regarding various stages in the system (other than the doctors' assessment) shows that most staff are not restricted to one stage of processing. It seems that the sequential processing of an application is not uncommon to be made by one processor. This makes the processing times appear untruthful as some items avoid queues which in turn may make other waiting times in queues appear longer than they should. The stages in processing of applications appear to take only a few minutes each, however the waiting time between these stages may take as long as several days.

4 PROBLEM DESCRIPTION

Sick leave is allocated to the contracted client's employees on a three year cycle, which is governed by the employer, and not by the start of the employee's employment. Once the employee has exhausted the allocated sick leave in their cycle, they may apply for temporary incapacity leave or ill-health retirement. These applications are investigated by the health risk manager where a recommendation is made regarding the validity of the claim.

Typically, sick leave starts to run out towards the end of its cycle and many applications come pouring in at the health risk manager in the last third of the three year cycle. This lumpy demand places a large constraint on the capacity of the organisation.

According to the service level agreement between the client and the health risk manager, the application is to be received from the client within a certain time frame and the recommendation is to be given within another specified time frame. As doctors and other medically proficient staff are employed to process the applications and make recommendations, the company's crucial resources are costly and not easily replaceable or duplicated. A balance between the demand for and the availability of resources is required. As doctors, who assess the applications, do not assess a fixed number of applications in any particular cycle, there is a strain on the capacity of the system at this point in the process.

The health risk manager currently cannot cope with the number of applications being made. The capacity required to address the fluctuating demand is unknown. To address the capacity constraints, there are various techniques and methodologies available to optimise and utilise capacity to its full potential.

5 LITERATURE REVIEW

5.1 CAPACITY MANAGEMENT

Capacity can be defined as “the ability to fight off an existing demand” (Armistead and Clark, 1991).This definition makes for a dynamic measure of capacity, which consequently also has a time dimension. This definition indicates that the capacity within this health risk management company is insufficient as the number of applications submitted (input) is larger than the number of applications being assessed (output).

Managing capacity comprises of both capacity planning and capacity control. *Capacity planning* involves the creation of “sufficient, flexible, capable capacity and a valid, best “do-able” resilient plan to accommodate demand”. *Capacity control* is about ensuring that the capacity plan is met (Thacker, 2009).

Capacity is relative to the level of the organisation, of which there may be four levels. Thacker (2009) and Armistead and Clark (1991) define these as:

Level 1: Strategic business planning

Usually the coordination is between business units, it is thus business wide, and the capacity is normally defined in terms of the output per financial year.

In terms of managing capacity, at this level processes need to be simplified and various constraints need to be identified. Constraints at level one can be classified as:

Hard ceilings – Difficult to add capacity
Hard floors – Fixed costs are high and are difficult to remove
Hard walls – The system cannot accommodate simultaneous variety

The health risk management company experiences hard ceilings as it is difficult to add capacity due to the large unknown number of, and the amount of time needed of the doctors who assess the ill health applications.

Level 2: Development, sales and operations management

This is the capacity at the individual process groups, such as a branch or local outlet. Capacity is defined in terms of daily or hourly output rates. At this level capacity must be able to meet peak demand, and not simply the average demand. Capacity planning is required to create a balance between capacity and demand.

Level 3: Workflow or process scheduling

Capacity is coordinated within a work cell, or the work team, and is defined in terms of the teams' daily or hourly output rates (unlike that at level 2 where capacity is defined in terms of the entire branch or outlet's daily or hourly output rates).

As only one office exists for the health risk management's company, we can combine level 2 and level 3. We are also able to investigate the theory of constraints (TOC, discussed later in the paper) that states capacity needs to be defined by the weakest link within the process. This weakest link should define the beat of the system (the drum that sets the beat of the process), and thereafter this process can be studied, with conclusions and decisions being made by using the queuing theory (also discussed later in the paper).

Level 4: Process management

The individual resources must be managed in terms of their capacity; these individual resources usually restrict the overall output of the process.

Capacity control within these four levels requires "meeting the plan, instead of constantly changing it" (Thacker, 2009).

At level one, capacity is controlled typically using budgetary controls. At level two, the overall plan within the local unit or branch needs to be measured and the performance against the plan must be analysed and corrective actions taken. Workflow, at level three, needs to be evaluated much as it is in level 2, but more in shorter terms. Level four requires process management and conformance to requirements (Thacker, 2009).

To appreciate the capacity more, there are a number of other factors to consider. The amount of necessary contact between the client and staff may influence certain decisions that are made concerning capacity. Activities with direct contact between customers and staff are called frontline activities. Activities that are performed without direct contact with the customer are support activities (Armistead and Clark, 1991). Unless contact needs to be made with the client regarding the referral of the employee for a secondary opinion, there is no contact between the client and the health risk manager. As most of these processes are viewed as support activities, the overall process is seen as majority support processes. This poses an important factor to consider: in this service organisation there is little contact with the client which is unlike the majority of studies conducted in the services sector.

The number of stages that need to be completed within the process may also affect the capacity plan. A balanced level between capacities at various stages in the system should exist. The flexibility of the system in respect to managing fluctuations in demand is also critical to understand the capacity of the system. Adding capacity when demand is high may require large capital investment which may not be a viable strategy in profitability terms. Flexibility within the system may not be easily achieved which may require the investigation into other alternatives to balance the supply and demand.

Potential capacity and effective capacity exist. Effective capacity is that which is available at all times within the organisation/system. Potential capacity is that capacity which *could* be made available if time is taken to arrange such capacity for the near future (Armistead and Clark, 1991).

Armistead and Clark (1991) discuss factors that influence (effective) capacity, as summarised below.

> The potential of the delivery system – how “able” is it to deliver the service in terms of the resources required to perform the service;
> The supply of the input to be converted to output; and
> How the condition of the inputs (as a physical measure) affect the system.

“These three aspects of capacity restate one of the basic rules of capacity that variety in the inputs and/or in the demanded outputs from a service delivery system reduces effective capacity.” (Armistead and Clark, 1991)

In an effort to analyse those resources that make up the capacity in the organisation, an approach, presented by Armistead and Clark (1991), could be followed:

Determine an appropriate measure for capacity;
Determine those resources the capacity comprises of;
What factors influence those resources that the capacity is comprised of; and
Identify the bottleneck resource.

A suggested measure for capacity at the health risk management company is the number of applications the system is able to process in a day. This measure inherently considers the number of personnel (or resources) that exist in the process. This capacity measurement will reflect the number of applications that are able to be processed in a period of time by the system's bottleneck, being the doctors who process applications.

The capacity (measured by the number of applications processed in a day) is comprised of various resources. The resources will be the personnel that process applications, and can be separated according to their various responsibilities:

Personnel who capture the applications into the system and determine whether the application is considered as primary or secondary;
Doctors who assess the application, whether they are internal to the company or external;
Typists who capture the assessment and produce a report; and
Proofreaders who ensure the reports conform to a set standard.

The flow of the system and communication between various processing stages will influence the resources. When high fluctuations in incoming applications happen, communication is necessary to ensure that the whole system responds effectively and that there is no slack within the system.

The number of incoming applications will also affect the resources: varying volumes of incoming applications will affect the number of personnel needed in the long term, and it will affect the output of the personnel in shorter time periods.

The quality of the processing will affect the throughput time, as poorer quality in early stages of processing will require correcting at a later stage. This may cause individual applications to iterate through the system until a certain quality is obtained, producing a lower output per resource.

To create balance in capacity management, three strategies presented by Armistead and Clark (1991, (the “Chase”, “Level” and “Coping” strategies), are briefly discussed below.

The Chase strategy is one where the capacity is managed by introducing more/less resources in order to manage the demand. The strategy aims at keeping the capacity as close to the effective capacity as possible.

The Level strategy’s target is to level the demand in order to keep it aligned with the available resources. As all applications must be processed, this is not an applicable strategy to consider for the health risk management company.

The Coping strategy was introduced to overcome problems where managing capacity and balancing demand is at a “capacity breakpoint”. This Coping strategy is effective in situations characterised by being busy or being slack.

In order to reach the point of applying the Coping strategy, the system must first have its own specific mix of Level and Chase strategies. Thereafter, the Coping strategy becomes necessary in four areas as detailed by Armistead and Clark (1991):

> If in the short term the Chase strategy becomes a Level strategy and effective capacity cannot be increased,
> When Chase strategies become Level strategies because resources can no longer further be reduced,
> When the Level strategy cannot reduce demand,
> When Level strategy cannot fill the effective capacity.

In a study investigating the process of admission of patients into an intensive care unit (ICU) of a hospital, authors Buckley, Horrowitz, Kim and Young (1999) uses capacity management to analyse the ICU’s most tightly constrained resource, namely beds. The study aims to develop a model for a hospital to investigate various policies; secondly, to investigate whether the ICU has sufficient capacity (the number of beds forms the measure of capacity in this study); and a third and final aim is to analyse the overall operations of the ICU.

The study uses queuing theory as well as simulation modelling to compliment the capacity management principles. Queuing theory is used to analyse the optimal number of resources (beds) needed in the ICU. The simulation model reinforces those results obtained from the queuing theory and provides a foundation for the continuous improvement needed to keep up with demands.

Many studies based on call centre optimisation involve the application of capacity management principles as well. Call centres operate with a fixed number of employees at a particular time and, like that of the allocation of beds in an ICU, have many of the characteristics of a service organisation: work cannot be stored for busier times, there is a direct link with the customer, and the service product that is offered is intangible. The demand at a call centre has the characteristic that there are large fluctuations in demand (the number of calls). If the caller waits for a response for too long (i.e. is in queue), he or she may wait, phone back at a later stage or abandon the request (call). It is important for the call centre company to develop a certain level of quality of service, and to answer as many calls as possible.

In the study of a banking call centre by Betts, Meadows and Walley (2000), an analysis of the utilisation of staff as well as the range and response flexibility is conducted. Range flexibility is the ability of a company to adapt the changes in demand over a longer period of time. Response flexibility is the ability of the company to quickly adjust to sudden changes in demand. The concept of range and response flexibility in a call centre operation is important. The study suggests that the management of short term flexibility in a call centre is of a somewhat difficult nature.

The spikes in short term demand experienced in the call centre are similar to those experienced at the health risk management organisation which presents difficulties with the correct allocation of staff. The study is similar to the health risk management company in that some of the call centres (such as authorisation centres) have to attend to ALL calls. The health risk management company too must process and make a recommendation regarding all applications that are received.

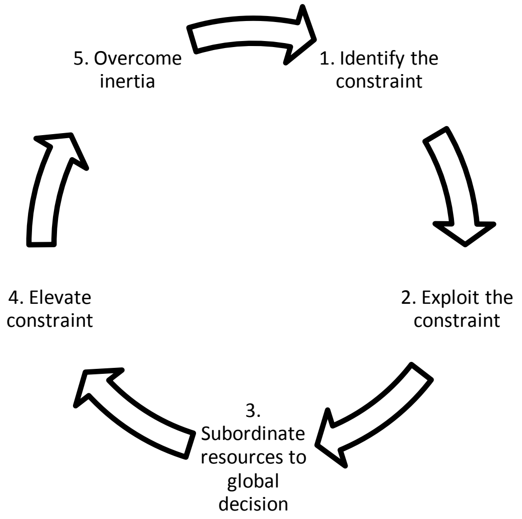
5.2 THEORY OF CONSTRAINTS (TOC)

Developed by Dr. Eliyahu M. Goldratt, this theory was originally developed to be used in a manufacturing organisation, but has application in service organisations too. Theory of Constraints, as well as Materials Requirements Planning (MRP) and Just In Time (JIT) formed three important approaches to achieve competitive advantage in industry (Rahman, 1998).

The theory is one that focuses on methods of ongoing improvement, and provides an approach to investigate, analyse and solve problems. It consists of two parts, the philosophy of TOC and the Thinking Process (TP).

The concept behind TOC is that every system has at least one constraint which represents an opportunity for improvement. This theory provides an approach that considers the whole organisation/system. The Five Focusing Steps of this theory is shown in Figure 1 below. The first step in the ongoing improvement methodology is to identify the system's constraint and then either eliminate or exploit it. The next step is aligning all other aspects of the system with the changes that were necessary to exploit/eliminate the constraint. Constraints must then be elevated to expose them to improvement. If at any stage a constraint happens to be broken (i.e. is no longer a constraint), the process should begin again. It is important to overcome inertia – no policy will be the best for all times (Rahman, 1998).

Figure 1 Five focusing steps


1. Identify the constraint
2. Exploit the constraint
3. Subordinate resources to global decision
4. Elevate constraint
5. Overcome inertia

These steps would correspond to the capacity planning and capacity control that comprises capacity management. Also, the constraint identified through the approach presented by Armistead and Clark (1991) will identify the bottleneck which will be used in the five focusing steps of the theory of constraints.

Along with the philosophy of TOC and the Five Focusing Steps, a Thinking Process (TP) forms the second part of the theory of TOC. It is a generic approach to solving problems while using intuitive knowledge and logic.

All that it involves is that managers make the following three decisions.

> Decide *what* to change, with the purpose of identifying core problems. Tools that can be used to decide what to change are the Current Reality Tree, as discussed later.
>
> Decide what to change *to,* with the purpose of developing a simple and practical solution. The Evaporative Cloud and Future Reality Tree can be used to aid making this decision.
>
> Decide *how* to cause the change, with the purpose of implementing actual solutions. This is aided by the Prerequisite Tree and the Transition Tree.

Determining what to change will be identified once the data is analysed, and the proceeding steps will be followed.

(Rahman, 1998)

The tools mentioned above are explained below.

- Current Reality Tree (CRT)

  The CRT provides a graphical representation of a core problem and the associated symptoms. It is developed working backwards – first identifying the symptoms and attempting to see a common cause amongst them. The CRT should be understood by everyone in all levels of the organisation (McMullen 1998, p56)

- Evaporative Cloud

  Also known as the Conflict Resolution Diagram, the Evaporative Cloud is used to solve problems between two parties, each with different views who have the same ultimate goal. It consists of Objects, Requirements and Prerequisites of the conflict (Yang et al, 2002).

- Future Reality Tree (FRT)

  Once underlying causes have been identified and some actions have been taken, the FRT can be used to identify possible negative outcomes of the changes (Yang et al, 2002).

- Prerequisite Tree (PRT)

  The PRT states which obstacles must be overcome before carrying out a specific action (Yang et al, 2002).

- Transitions Tree (TT)

  The TT describes what actions will be necessary to achieve a certain goal, or change to a certain state (McMullen 1998, p81)

TOC involves the identification of the constraint within the system (as mentioned in the Five Focusing Steps above). This constraint is seen as the bottleneck. The pace that the constraint can produce an output should set the pace for the entire process in order to eliminate inventory and backlog. This bottleneck should set the beat in a Drum-Buffer-Rope process. That is, the drum sets the pace for all other stages to work according to, the buffer is a determined number of safety stock strategically placed throughout to process, and the rope is the communication link between stages of the process to ensure synchronisation between various stages.

As the name suggest, the principles of Drum-Buffer-Rope include the use of a “buffer”. The concept of a buffer is to protect constrained resources from not having sufficient work to do so as to prevent the system from providing optimum output and is usually in the form of work in progress that can be stored before processing at the constrained resource. As the health risk management company is that of a service type, it is impossible to build inventories as buffers at critical stages in the process. Harowitz, Klein and Motwani (1996) suggest the use of time as a buffer in capacity constrained stages of the process in a service organisation.

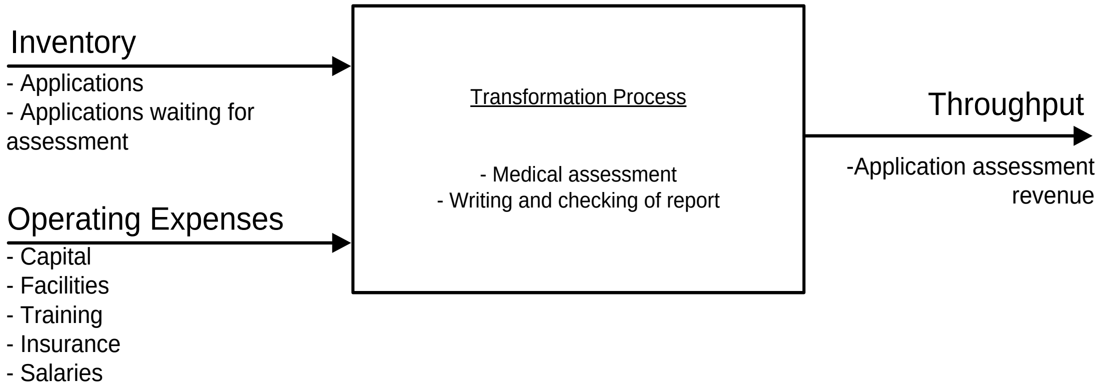
As discussed in the article by Rahman (1998), TOC has been applied and is the topic of discussion in many study fields, such as production, purchasing, accounting, administration, education and quality. There has only recently been an interest into the application of TOC in service industries.

The article by Harowitz, Klein and Motwani (1996) discusses the idea of abstracting the traditional viewpoint of a service organisation to one that may resemble that of a manufacturing environment with the purpose of being able to apply the theory of constraints. The article discusses how those elements, crucial to TOC in manufacturing, also appear in service organisations, for example, the inputs to a manufacturing system would be raw materials and in a service organisation it could be viewed as operating expenses and resources.

With the example of a hospital admissions and discharges waiting times, TOC was used to identify constraints and thus improve the waiting time. This was achieved by implementing drum-buffer-rope methods in the discharge-, room preparation- and admission processes. In another example at a hospital also discussed by Harowitz, Klein and Motwani (1996), inefficiencies existed in the scheduling of operating theatres. The constraint was identified that doctors said they required a certain amount of time, but more often than not used more than that time, creating a delay in operating proceedings and thereby a backlog. After this constraint was identified and elevated, it was realised that a buffer in the form of time was needed between successive surgeries.

The model showing processes in a service organisation presented by Harowitz, Klein and Motwani has been modified according to the health risk management's company and is shown in Figure 2. By showing the processes in terms of inputs, transformations and throughputs, we can more easily apply drum-buffer-rope principles to the system.

Figure 2 System representation of the health risk management company



5.3 Queuing Theory

As applications can be seen as entities in a system where the entity waits in a queue to obtain service from some resource (medical practitioner) within the system, the theory relating to queuing can be very valuable in the development of a model to understand the current situations within the health risk organisation, and to explore various alternatives that may result in a recommended solution to deal with capacity constraints in the organisation.

Queuing theory has the ability to answer the following questions (Winston, 2004):

What fraction of time is each medical practitioner idle?
What is the expected number of applications in the queue?
What is the expected time that an application spends in the queue?
What is the probability distribution of the number of applications present in the queue?
What is the probability distribution of an application's waiting time?
If you want to ensure that only a certain percentage of all applications have to wait longer than a specified time, how many medical practitioners should be employed?

Moreover, by studying the answers obtained in the above questions, ultimate management decisions can be considered (Beasley, 2009), such as:

Would reducing the service time be beneficial to the overall system?
Would introducing more medical practitioners improve the overall system (would the benefits outweigh the additional costs)?
Should pre-emptive priorities be considered (are there applications that are considered as more important, and should they be introduced to the front of the queue)?

The batch size and the arrival times of the applications are variable and the number of medical practitioners may vary from time to time, thus the system can be seen as stochastic. Applications arriving in batches will furthermore be referred to as bulk entries entering the system.

There are various queuing models available that may be applicable to the scenario as described by the health risk management company. These systems are named according to a system –

the Kendell-Lee Notation. This notation describes the queuing system using six characteristics in the following order:

$$1/2/3/4/5/6 \qquad \textbf{(1)}$$

The first characteristic symbolises the probability distribution of the arrival process, while the second represents the probability distribution of the service process. These probability distributions are usually (and for the case of health risk management company) independent, identically distributed random variables with exponential distribution, represented by the capital letter M.

The third characteristic represents the number of parallel servers. For example, a set of parallel servers could be the number of individual medical practitioners that are able to process applications concurrently. A set of servers in series would represent the different stages that an application would need to pass through before leaving the system (where a recommendation is made and referred back to the client/employer). The health risk management company has a combination of both parallel and series servers.

The fourth characteristic describes the queuing discipline. The queuing discipline is the order that applications are processed within the system. The various, possible queuing orders that exist are as follows.

FCFS – first come, first serve

LCFS – last come, first serve

SIRO – service in random order

The fifth characteristic displays the maximum number of customers allowed in the system. The system representing the health risk management company allows an infinite amount of customers into the queue, represented by the infinity symbol (∞).

The final characteristic represents the maximum number of applications in total. For general purposes and for purposes relating to the company under study, this will always be assumed as infinite (∞).

The last three characters of the notation are usually FCFS/∞/∞ and thus will be omitted in this report unless specific mention is given to the queuing discipline to represent various alternatives studied.

In the case of the health risk management company, the queuing system for each individual processing stage can be analysed. The arrivals can be seen as independent, identically distributed random variables with exponential distributions. Each processing stage consists of a number of servers (parallel servers), which will be denoted by *s,* and currently applications are treated on a first come first serve basis. The number of applications in the system is unlimited (as all applications *must* be processed). Thus the appropriate queuing model is represented by

$$\text{M/M/s/FCFS/}\infty/\infty \qquad \textbf{(2)}$$

Terminology and equations related to this model are shown below. The terminology and equations were obtained from Introduction to Probability Models, by Winston (2004).

λ = Average arrival rate of applications;

μ = Average service rate of applications by servers;

$$\rho = \frac{\lambda}{s\mu} \qquad \textbf{(3)}$$

The traffic intensity; $\rho \leq 1$ for the system to exist in a steady state (or in order for it not to explode); and

$P(j \geq s)$ = The steady state probability that all servers are busy.

The probability that no applications are waiting for service (or assessment) (π_0) is:

$$\sum_{i=0}^{i=s-1}\left(\frac{(\mathrm{s}\rho)^{\mathrm{i}}}{\mathrm{i}!}+\frac{(\mathrm{s}\rho)^{\mathrm{s}}}{\mathrm{s}!\,(1-\rho)}\right)^{-1} \qquad \textbf{(4)}$$

This is the probability that all servers in the system are busy:

$$P(j \geq s) = \frac{(s\rho)^{s}\pi_0}{s!\,(1-\rho)} \qquad \textbf{(5)}$$

The expected number of applications in the queue:

$$L_{q=}\frac{P(s \geq j)}{1-\rho} \qquad \textbf{(6)}$$

The expected time the application spends waiting in the queue for service:

$$W_q = \frac{P(j \geq s)}{s\mu - \lambda} \tag{7}$$

The expected number of applications being processed:

$$L_s = \frac{\lambda}{\mu} \tag{8}$$

The expected time that an application spends being processed:

$$W_s = \frac{1}{\mu} \tag{9}$$

The expected number of applications in the system:

$$L = L_q + \frac{\lambda}{\mu} \tag{10}$$

The expected number of applications in the system:

$$W = \frac{L}{\lambda} \tag{11}$$

Queuing systems can be easily and visually represented by simulation models. With the application of simulation to compliment queuing theory, the queuing discipline may also be more thoroughly investigated. For example, it will be possible through the use of simulation software to analyse other forms of queuing disciplines, such as priority rules, where more urgent applications may be pushed to the front of the queue. Another possible application of combining queuing theory with simulation is to investigate the possible redesigning of the processing flow of applications in the system.

Applying the above equations to the problem description as experienced by the health risk management company will aid exploring various alternatives to improve capacity problems within the system.

Queuing theory has been successfully applied in many areas. Much research has been done in the optimal allocation of patient beds in Intensive Care Units (ICU’s). One such study (Buckley et al, 2006) studies the efficient use of beds to ensure the welfare of the patients requiring intensive care treatment and to optimise hospitalisation functions which include minimising costs and ensuring a proper scheduling of staff. Queuing theory was used alongside a simulation model in order to validate the model. The study aimed at answering the following questions:

How often will the capacity be fully utilised (i.e. all beds occupied)?
What is the waiting time for admission to an ICU bed?
How will additional capacity alleviate delays in the allocation of beds to patients?

The study of the optimal number of hospital beds to allocate in a ward conducted by Buckley *et al* with the use of queuing theory and simulation modelling is useful in emphasising the capacity that is required, and making visible where capacity and/or resources require targeted management.

Articles that were researched mostly studied queuing systems with a predetermined, or fixed, capacity and/or number of resources available. The health risk management company can consider the number of staff other than doctors (such as typists and proofreaders) as a resource with known quantity and capacity as these resources are readily available. It is also noted that these resources do not pose as the number one constraint in the organisation. Doctors who assess the application and make the recommendation are not readily available and cannot be scheduled as is possible to do with the other staff.

5.4 LITERATURE REVIEW CONCLUSION

The above studies are restrictive in that they investigate areas where capacity management can be applied where a fixed number of resources exist, and that demand can, in most cases, be controlled. In the case of the hospital intensive care unit, if there is insufficient space for a patient to be admitted, the patient can either join a queue or the patient can be sent to another hospital with available capacity. However, the number of beds still remains fixed.

In the case of the call centre study, if all capacity is full (all operators are busy) the caller has the option to wait, or hang up the phone call and try again later. Although scheduling of operators does exist, the capacity at any given time can be regarded as fixed.

In the case of this particular health risk management company, the doctors are not seen as a fixed capacity but rather as a varying one. It is for these reasons the study presented in this report is justified: to present a solution to manage capacity where resources in one of the process points is unknown or is difficult to determine. In terms of the overall study, the theory of constraints and capacity management will be used in an effort to bring about a policy for the capacity planning and control of the processes at the health risk management company. The constraints that are identified using the approaches of capacity management will be used in conjunction with the theory of constraints in order to elevate the constraints and overcome them. Queuing theory will be used to study these constraints and resources, and in aiding the general capacity plan to manage the process. A simulation of the process will reinforce the recommendation.

6 FACTORS AFFECTING THE PROCESS

In order to determine those factors that affect the efficiency of the process, it is necessary to define efficiency, and then analyse those affecting factors.

“Efficient” is defined by Dictionary.com as “performing or functioning in the best possible manner with the least waste of time and effort; having and using requisite knowledge, skill, and industry; competent; capable”.

If we analyse this definition in terms of the processes at the health risk management company, “performing or functioning in the best possible manner with the least waste of time and effort” would require that the process tasks performed in the system be optimal. System improvements may be studied through the use of simulation modelling where various scenarios may be studied.

“Having and using requisite knowledge, skill, and industry; competent; capable” requires that those people who process and assess ill-health retirement and temporary incapacity leave applications be sufficiently trained and be able to perform their tasks appropriately.

The main company resources that were identified were the staff, such as the doctors, nurses, typists, the ERP system, etc.

Factors that may inhibit the process from functioning efficiently are:

The number of incoming applications;
The resources’ availability; and
The management of the available resources.

The number of incoming applications

All applications that are received must be processed. Due to the nature of the cycle of the allocation of sick leave by the client, there are large fluctuations in the number of applications received in a given period of time. As a result of these fluctuations, queues develop where applications await processing at the various stages. According to the service level agreement, a recommendation must be made to the client within a specified time frame and the build up of queues in the system does not contribute to a speedy recommendation being made.

The resources availability

As discussed in the background of the project, the number of doctors that are available to assess the applications and make recommendations is largely unknown, due to the randomness in the number and frequency of collections of applications. If these doctors are unable to collect enough applications within the time frame allocated, queues arise that eventually result in the entire process being delayed. A solution of scheduling the doctors becomes difficult as the amount of time that individual doctors have to assess applications varies significantly.

Other personnel employed at the health risk management company work on a full time basis and on the company premises. The scheduling of these employees who conduct other stages of the processing of applications is important, but they are not critical resources as are the doctors.

The management of resources

Personnel other than the doctors perform various tasks in the process. Many of these tasks require that the personnel be trained and may even require training in a specific field (such as nursing and typing). Due to various backlogs in the system currently, it is sometimes found that staff proficient in more than one stage of the processing may complete various stages of the process, sometimes even consecutive stages. With various people performing various tasks, applications may get lost and quality may decrease. There are iterative stages within the process (needed for quality assurance and corrective action reasons) that consume much of the resources’ time. This time taken for an application to undergo iterative processes is difficult to deduce, it is however observed that the more senior or experienced the person processing is, the fewer the iterations required. This suggests that the management of resources should include efficient training, or a simpler process that is able to reduce the number of iterations conducted without affecting the quality of the output.

7 DATA COLLECTION AND ANALYSIS

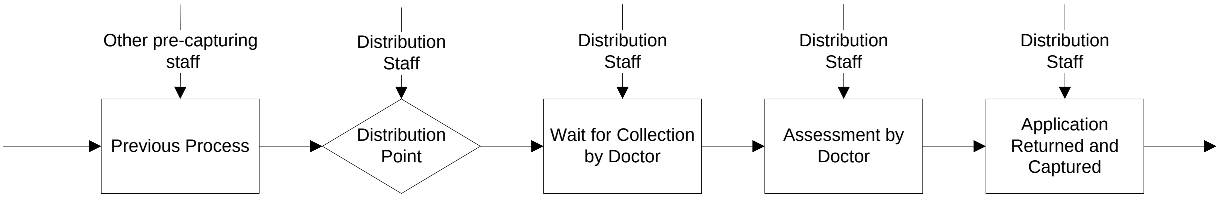
7.1 DATA SOURCE

The factors that affect the system should be analysed in order to determine the limits of those factors. The arrival of the applications should be studied in an effort to understand the quantities of applications that need to be processed. This is important as the number of staff required to process the applications at an acceptable rate is to be determined. The efficiency of the process resources (staff) also needs to be analysed to determine how their availability affects the system. The most efficient way to study the above mentioned factors would be through the use of queuing theory combined with a simulation model. Data required should correspond to the inputs, processing times, number of personnel and other system constraints of the various stages.

The company makes use of an ERP (Enterprise Resource Planning) software package that captures the date and time of every application when a particular stage is completed, as well as the person who completed the process at the particular stage. The system also reports on all applications that are late. This data was extracted from the database, and processing times, etc at the various stages were calculated from this data.

In the collection of data, the first problem encountered was the flow of applications just before and just after the doctor assessment as shown in Figure 3. The doctors do not use the computer system and applications entering and leaving the doctor assessment stages are predominantly captured by one member of staff who is responsible for the distribution of applications. Although the time that the application spent waiting to be collected by the doctor is indicated, this method of processing gives no indication as to how many doctors there were at a specific point in time assessing applications, nor as to how many applications where collected by the doctor to assess.

Figure 3 Flow of applications near and including doctor assessment stages



The system captures the waiting time between the end of the Previous Stage (as shown in Figure 3) and the start time for the process Wait for Collection by Doctor. In Figure 3, the time would be stamped at the end of each process block. The next time stamp that is captured into the system is the time taken from when the doctor returns the bulk load of applications to the offices. The following time stamp is taken after the next stage in the process and so on. It is important to note that the waiting time for a doctor to collect applications is only known due to a process in the computer system that is dedicated to record the time that an application spends waiting for the doctor. This is not the same for other processes in the system.

7.2 DATA COLLECTION

The processing times for all individual stages were computed using a computer program written specifically for the purpose of extracting the information needed. The program calculated the processing times of each application at the various stages as well as the life times at the various stages. It was necessary to compute processing times where the number of resources available was known while the life time of the process was needed for those stages where resources were not known and where processing times were inadequate measures as input to the model.

The process times and the life cycle times were then studied to identify and remove outliers. The removal of outliers is justified by staff not capturing the information correctly on the computer software and where multiple applications are processed and only afterwards captured in succession. The use of these outliers would falsely create very long life cycle times and very short processing times.

The process times remaining after outliers were removed were fed through Arena's Input Analyser, where an appropriate distribution could be fit to the processing times. An expression was generated for use in the relevant process module in the simulation model.

Appendix A contains the histograms associated with the various processes in the system, along with the distribution parameters that can be used to develop an expression for the processing times in the simulation model.

The number of applications received monthly has significantly increased from January 2007, when less than 250 were received monthly, to July 2009, when the number of applications received monthly increased to over 2400. This is shown in Figure 4.

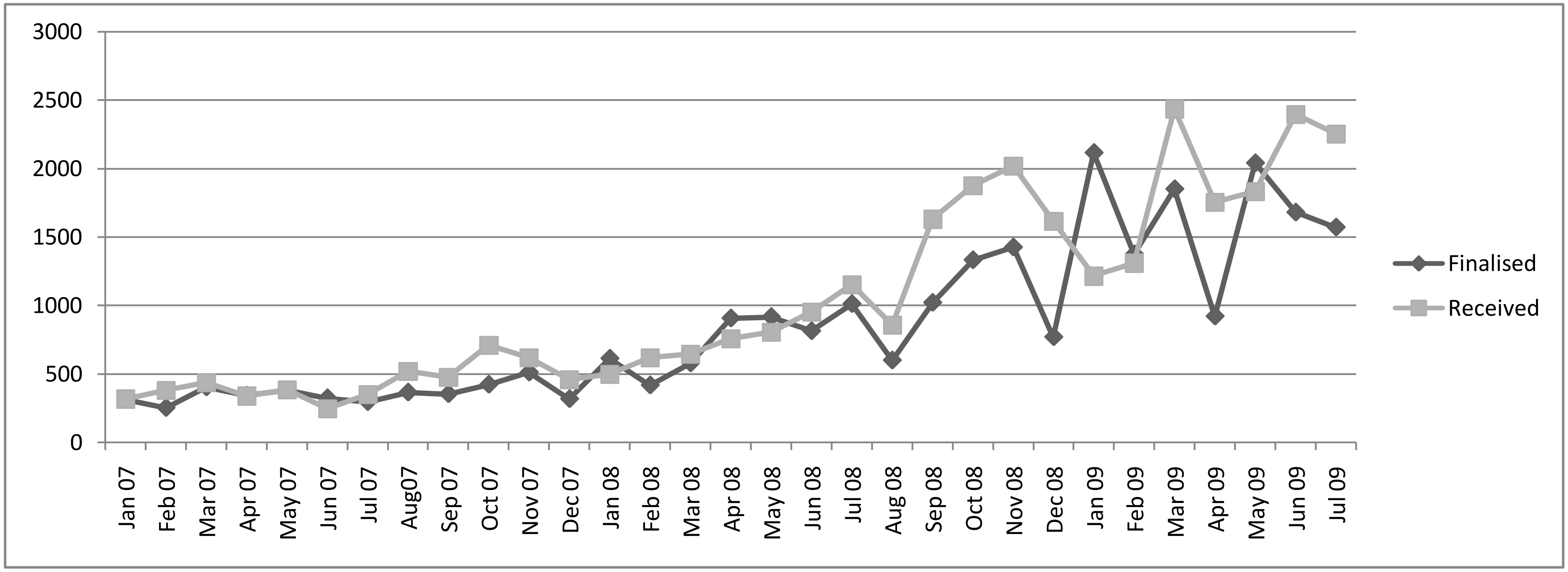
Figure 4 also indicates the number of applications that were received on a monthly basis compared to the number of applications that were completed by the health risk manager. From the graph it is seen that the maximum capacity the health risk management company has been able to deliver is just over 2000 applications in a month. It must be noted that according to the data, this was not the average, but an exception.

The actual arrival of applications (as shown in Figure 4) was used to generate the arrival of applications in the simulation model. This ensures the arrival of entities to the model is accurate and represents the system.

The number of resources at the various stages was determined according to those personnel who completed the most applications per period of time. The approach was used as the processing staff do not only process one stage in the system, instead they process up to 3 different stages. It was found that a person who processes more than one type of processing stage normally dominates in one of those. Where it is unclear which of those processes the person performs more regularly, it was found that it is possible for 2 consecutive stages to be completed successively by one person. In these situations processors did not capture the times between the two different stages, giving one of the processes a null processing time. To overcome this, those processes where combined for use in the model.

The flow of the applications remained the same for ill health retirement applications and temporary long leave applications. The process differs for short leave applications and is somewhat simpler. These processes were further simplified by eliminating any flows between two stages that occurred less than 1% of the time. Visual representation of the process flows are shown in Appendix B

Figure 4 The number of received applications versus the number of finalised applications



7.3 Data Analysis

The simulations is terminated when the last applications arrives (or is created) at the system. The arrival of applications is modelled according to actual arrival times making it necessary to terminate the simulations before resources unrealistically run out of work to process which would decrease the resource utilisation.

A reasonable approximation to the number of applications collected by doctors was also determined using a computer program for that specific use. The program is able to group applications that have the same processor, almost identical task life times and almost identical capture times. These are then grouped (batched) to represent the collection of applications by a doctor. The batches are not the same; however the data was also fed through Arena's Input Analyser to determine a statistical distribution for the batching and collection times. It was found to be exponential, centred at 4 with a mean of 3.75. The batching of applications was rounded to the closest integer value.

The simulation model was able to highlight some important information regarding the processes and the resources of the health risk management company's system.

Table 1 shows large waiting times for the use of many of the resources. The average processing time for the assessment made by the doctor, and the quality and conformance reviews take long. These are all performed by doctors, and thorough detail is required to ensure the process meets the standard.

Although the processing times for those stages with doctors as their resources can be justified, the waiting times are extremely long. For example, the average waiting time for the first quality and control review process is 790799 seconds, or 220 hours or 27 days. The service level agreement for the processing of applications and making a recommendation is 12 days for long and short temporary incapacity leave applications that do not require a specialists' opinion. This means that an average waiting time for the doctor assessment makes the recommendations regarding the application seriously late. The average processing and waiting times were generated in the Arena report after running the simulation.

Table 1 Current average processing and waiting times

Process	**Average Process Time (seconds)**	**Average Waiting Time (seconds)**
Scan	642.11	6990.17
Capture	732.89	3892.92
PreAssess	275	0.701
Specialist Referral	405849.22	3776055.84
Doctor Assessment	98658.91	764129.11
Typing of Report	1205.02	2419.83
File Distribution	103.18	2.10
Quality and Control Review 1	10348.58	871129.68
Quality and Control Review 2	116892.35	790798.68
Corrections Checking 1	122.33	0.03
Corrections Typing 1	123.86	0.03
Proofreading	141.49	0
Final Scanning	158.21	25.26
Sending Report to Client	124.98	25.62
Corrections Checking 2	123.86	0.03
Corrections Typing 2	462.42	2488.49

Table 2 shows the different resources that are used in the health risk management company's system along with the various stages that a resource is capable of processing. The resources' utilisation is also shown in Table 2. The utilisation of resources was generated from the Arena report after running the simulation model.

It is obvious that the doctors, as a resource, are under scheduled, and that resources such as senders and distributors are over scheduled. The utilisation of the specialist is not discussed in this article, as the specialist is any external doctor who may be in the vicinity of the patient to schedule an examination. The scan and capturers as well as the typists have moderate to low utilisation. This could benefit the fluctuations in arrival patterns of the incoming applications.

Table 2 Resources, processes they perform and their utilisation

Resource	Utilisation	Processes included in duties
Scan and Capturer	36%	Scan application
		Capture application
		Final scan application
Capturer	31%	Capture applications
Pre Assessor	11%	Pre assess applications
Specialist	Not included	Not included
Doctor	100%	Assessment of application
		Quality and conformation review 1 & 2
Typist	68%	Typing final report
		Corrections typing
Proof reader	10%	Proofreading
		Corrections checker
Sender	8%	Sending report to client
Distributor	2%	Distributing applications between various process stages

8 THE RECOMMENDED POLICY

The data that was collected and analysed showed that a major bottleneck to the process is the number of doctors that are available to process applications. Following the principles of Theory of Constraints, this bottleneck should be exploited and all other resources appropriately aligned.

As previously discussed the allocation of doctors is of a difficult nature due to other personal commitments made by the doctor. The proposed policy includes a recommendation based on the management of the capacity within the system at the health risk management's organisation, as well as one that proposes a scheduling strategy to be used for the scheduling of resources.

The proposed policy regarding to the scheduling of employees at the various stages is illustrated in Table 3. It is based on queuing theory and addresses capacity at both range and response flexibilities.

The general increasing and decreasing trend within the organisation is well known to exist in cycles of three years with large fluctuations during the winter months and thus a monthly (or daily) arrival rate of applications can be assumed, as well as the percentage of those applications that are either short, or a combination of long and ill. Table 3 shows an interactive table that can be used to fill out these values. The arrival rate that is required as input is the arrival rate of applications, and can be chosen as either monthly or daily quantities. Assume Table 3 contains daily arrival rates.

The input ratio is the ratio of applications that proceed through the specific stage. The input ratio was obtained from the computer program and the associated diagrams that are shown in

The input rate is the product of the arrival rate and the input ratio. The average service rate is that processing rate that was developed from the Input Analysis tool in Arena. The average service time here is the same as that used in Table 1 only it has been converted from seconds to days. The average service rate is the rate of the average service time.

The required number of servers is calculated and displayed in the Required Servers column. This will address the capacity needs in the longer term.

Table 3 Computation of required number of resources at the various stages of the process

Arrival Rate	120
Percentage Shorts	0.88
Percentage Longs/Ill	0.12

Process	**Input Ratio**	**Input Rate**	**Average Service Time**	**Average Service Rate**	**Required Servers**
Scan	1.00	120.00	0.02	44.86	3
Capture	1.00	120.00	0.03	39.34	4
Pre-Assess	1.00	120.00	0.01	104.73	2
Doctor Assessment	0.20	24.00	3.43	0.29	83
Typing Final Report	1.00	120.00	0.04	23.90	6
Proofread	1.06	127.09	0.00	204.26	1
Quality and Conformance Review 1	0.13	15.03	0.36	2.78	6
Quality and Conformance Review 2	0.14	17.26	4.06	0.25	71
Corrections Typist	0.44	53.28	0.02	62.34	1
Corrections Checker	0.44	53.28	0.00	234.15	1
Final Scan	1.00	120.00	0.01	182.28	1
Final Report to Employer	1.00	120.00	0.00	230.40	1
Specialist Referral	0.03	3.60	14.09	0.07	51
File Distribution	0.19	22.40	0.00	279.61	1

Addressing the fluctuations in arriving applications in the shorter term will require management of the system constraints as well as the alignment of the other resources with the bottleneck. This forms the third step of the Five Focusing Steps of Theory of Constraints.

An important principle to note when addressing the recommended policy is that whatever inputs the system receives, it (the system) should be able to produce the same amount as output in the same time frame. This would mean that if the service level agreement states that a recommendation must be made within 12 working days of receiving a applications, 12 days later it should be able to produce at least that same number of received applications.

The bottleneck (identified in the simulation) is the scheduling of the doctors as resources and it can be assumed that this stage in the system creates the “drum beat” to set to pace for all other stages. The doctors are a difficult resource to manage, and thus it is recommended that a time window be allowed to set the pace for the assessment of applications by doctors.

This time window can be determined in the following manner:

All applications undergo similar initial processes and final processes. Initial processes include the scanning, capturing and pre assessment of the application. The final phases include typing of the report, proofreading, checking and sending the final report to the client (the recommendation made by the health risk management company).

These processes are performed by resources who are permanently employed by the company and they have a predictable life cycle time. For that reason it is possible to categorise the stages of the process into 3 phases: the initial, the final and then the intermediary phases (which would involve processes involving the use of doctors as a resource) which are considered as time windows.

A standard processing time for initial and final phases is as follows:

The initial phase consists of scanning, capturing and pre assessing the applications, with average process times of 642, 733 and 275 respectively (obtained from the summary statistics shown in the Arena simulation report). The total value added time for the initial phase is then 0.45 hours, or 28 minutes.

The final phases consists of typing the final report (1205s), file distribution (103s), corrections checking (123s), corrections typing (469s), proofreading (141s), sorting and checking the final report (158) and sending the assessed application to the client (124s). The total value added time for the final process is 2323 seconds, or 0.65 hours or 38 minutes.

The times mentioned above for initial and final phases do not contain any allocations for transfer times. Transfer times do exist and this is taken into account in the following manner:

> All initial phases must be complete within one day of arrival of the application;
> All final phases must be complete the day the recommendation is to be delivered, and must have commenced at least 2 days prior to the delivery date.

Ample time is allowed to process the applications at these phases.

If an application is a long or short and primary one, the service level agreement states that the application must be returned with a recommendation within 12 working days. Of these 12 working days, 1 day will most likely be used for the initial processes although it takes less than an hours' processing time, and 2 days will most likely be used for the processing of the final stages. This leaves a time window of 9 working days for the doctor to collect, assess and return the application to the health risk management company.

If an application is a secondary application that is long or short the service level agreement states that 40 working days are allowed for the recommendation to be made. As the only additional process to a secondary application is the referral of the patient to a specialist doctor for a consultation, the initial, final phase and time window allowed for the doctor assessment may remain the same in duration. An additional time window of 28 working days may be allocated for those processes required when a patient is referred to a specialist doctor.

The application of the policy is the same for ill health applications, except the service level agreement allows 90 working days for the complete assessment of an application. Thus the time window for the specialist referral processes will be extended to 78 working days.

Equation (12) could be used to help manage the return of applications by the doctors:

$$(Number\ of\ Applications) \times (3\ hours\ per\ Application) \div (8\ hours\ per\ day) = roundup(x\ days\ time) \tag{12}$$

For example, if a doctor took 6 applications, he would need to return them in at least $6 \times 3 \div 8 = 2.25$ days, or, rounded up, in 3 days. This allows the distributor an opportunity to manage the doctors and possibly push them to take more applications when there are short increases in the arrival of applications.

This will require a certain amount of commitment from staff who distribute applications, and a high level of management ensuring that individuals meet the deadlines.

The new policy employs methods of capacity management, queuing theory and theory of constraints which complement each other in the recognition of problem areas, constraint identification and the management of resources to continuously improve processes.

The recommendation has the ability to deal with the fluctuations in both the long term and the short term, which should prevent the system from losing control.

9 POLICY ANALYSIS

Table 1 has been expanded as in Table 4 which shows the values for the processes with the initial process and with the recommendation to make it obvious the improvements that are possible with the recommendation. The simulation was run for a year.

Table 4 Current versus proposed average processing and waiting times

Process	**Current Average Process Time (seconds)**	**Current Average Waiting Time (seconds)**	**Current Average Process Time (seconds)**	**Current Average Waiting Time (seconds)**
Scan	642.11	6990.17	638.31	17602.29
Capture	732.89	3892.92	734.08	385.73
PreAssess	275	0.701	276.70	105.54
Specialist Referral	405849.22	3776055.84	459511.23	89822.82
Doctor Assessment	98658.91	764129.11	99957.64	0
Typing of Report	1205.02	2419.83	1198.64	1201.04
File Distribution	103.18	2.10	102.28	4.11
Quality and Control Review 1	10348.58	871129.68	10865.61	0
Quality and Control Review 2	116892.35	790798.68	118533.73	0
Corrections Checking 1	122.33	0.03	122.41	227.12
Corrections Typing 1	123.86	0.03	454.93	979.93
Proofreading	141.49	0	141.53	222.93
Final Scanning	158.21	25.26	158.2	13002
Sending Report to Client	124.98	25.62	124.94	2384.19
Corrections Checking 2	123.86	0.03	123.65	285.9
Corrections Typing 2	462.42	2488.49	462.36	1217.10

Although some of the processing and waiting times are larger than the current process, by reducing the waiting time for doctors one can see the following improvements in the total average time an application spends in the system.

Figure 5 Current Process versus Proposed Process: Time Savings

Total Time in System	Current System (s)	Proposed System (s)	Time Saving
Ill Applications	4971805	725523	85%
Long Applications	1797284	352883	80%
Short Applications	954003	146744	84%

10 CONCLUSION

The report demonstrates that Theory of Constraints principles can be applied to the services industry, and that in many circumstances, certain aspects of a service industry may be modelled or studied using traditional manufacturing type environment tools. It also shows how three industrial techniques can be used to complement one another (Queuing Theory, Capacity Management and Theory of Constraints).

The processing time of applications in the health risk managements' company is able to be reduced by more than 80% when studied with a model that provided enough doctors to process all applications without waiting for service. This may not be possible from a cost and logistics view, it does however provide a recommendation that can be used as a starting point to determine an acceptable number of resources and an acceptable outlay of costs for those resources.

The use of a time window where the resource capacity is largely unknown is not specific to this particular company, but may be applied in any situation where resources are not fully known or understood. The combination of queuing theory with the time window allows both range and response flexibility in the system, and should be studied regularly in the organisation to keep the model updated.

11 REFERENCES

Armistead, CG and Clark, G. 1991, *'Capacity management in services and the influence on quality and productivity performance"*, viewed 19 May 2009, http://dspace.lib.cranfield.ac.uk:8080/bitstream/1826/333/2/SWP5691.pdf.

Beasley, JE, 2009, OR Notes, viewed 22 May 2009, http://people.brunel.ac.uk/~mastjjb/jeb/or/queue.html

Betts, A, Meadows, M, Walley, P, 2000, "Call centre capacity management", *International Journal of Service Industry Management*, vol 11, no2, pp 185-196

Buckley, TA, Horowitz, I, Kim, SC & Young, KK, 2006, "Analysis of capacity management of the intensive care unit in a hospital", *European Journal of Operational Research,* vol. 115, no. 1, pp. 36-46.

Dictionary.com, viewed 30 September, 2009. http://dictionary.reference.com/browse/efficient

Harowitz, R, Klein, D, Motwani, J, 1996. "The theory of constraints in services, Part 2 – examples from healthcare", *Managing Service Quality*, vol. 6, no.2, pp.30-34

McMullen, TB 1998, Introduction to the theory of constraints (TOC) management system, CRC Press.

Rahman, S, 1998, "Theory of constraints: A review of the philosophy and its applications", *International Journal of Operations and Production Management,* vol. 18, no. 4, pp. 335-355.

SOUTH AFRICA, Department of Public Services and Administration, 2005, "*Policy on incapacity leave and ill-health retirement (PILIR)".* Pretoria: Government Printer.

Thacker SM. 2009. *Capacity Management*. SM Thacker & Associates (Consultancy and Training Specialists), viewed 19 May 2009, http://www.smthacker.co.uk/capacity_management.htm.

Winston, WL 2004, *Introduction to probability models,* 4th ed, Brooks/Cole, United States of America.

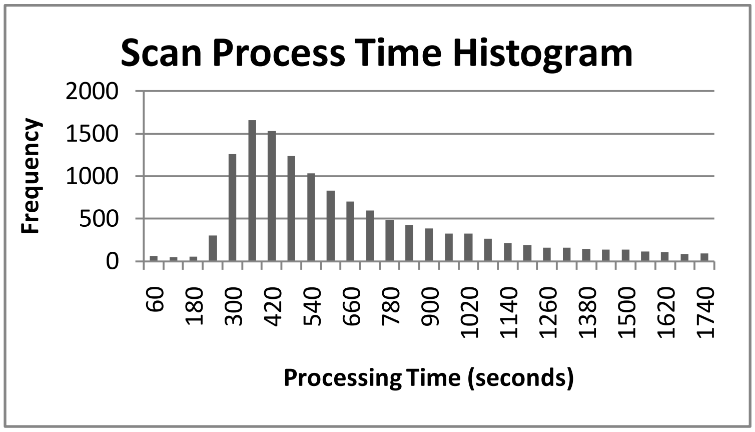
Yang, CL, Hsu, TS & Ching, CY, 2002, 'Integrating thinking into the product design chain', *Journal of Industrial Technology,* vol. 18, no 2, p.3.

Yang, Kai. Design for Six Sigma for Service

Special thanks to my project leader, Mr Olufemi Adetunji, for all his persistence and guidance throughout the duration of the project.

APPENDIX A

Figure 6 Histogram of the processing times for Scanning



The scan processing times have a gamma distribution: 60 + GAMM(237, 2.45)

Figure 7 Histogram of the processing times for Capturing

The capturing processing times have a gamma distribution: 60 + GAMM(832, 0.811).

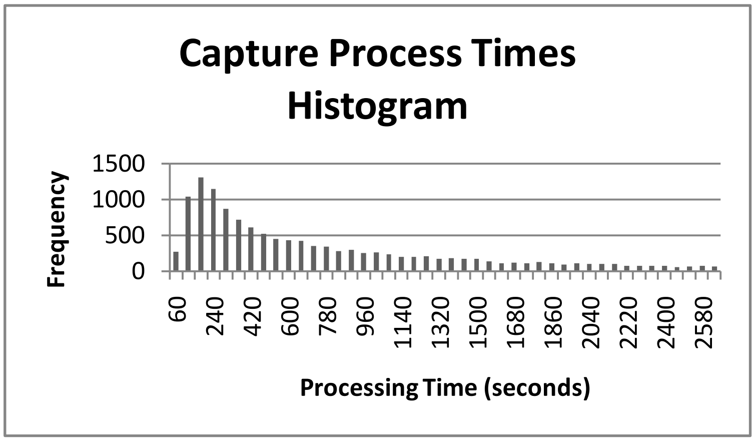
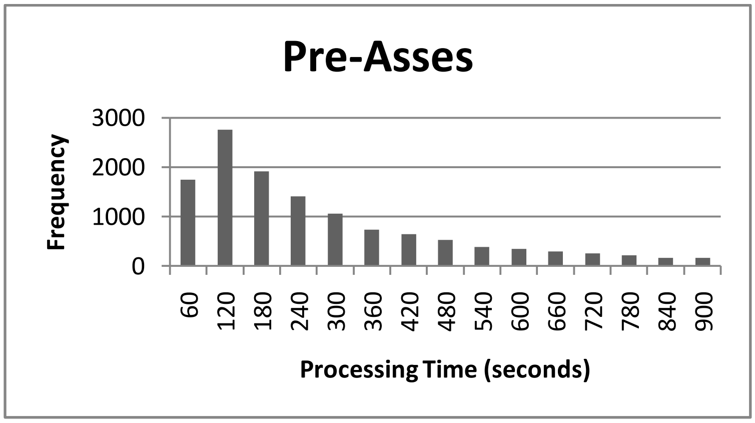
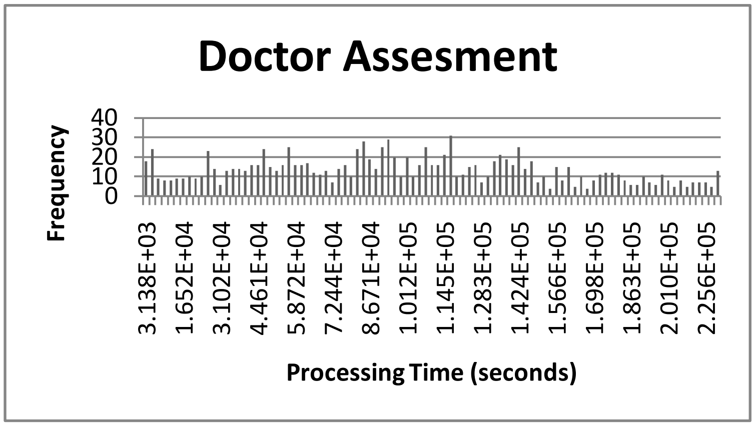


Figure 8 Histogram For PreAssess



The pre assess processing times have an exponential distribution: 60 + EXPO(216)

Figure 9 Histogram Doctor Assessment



The distribution for the doctor processing times is a gamma distribution: (1.95e+003 + 2.32e+005 * BETA(1.31, 1.77))

Figure 10 Histogram Typing Final Report

The time distribution for typing of the final report is a gamma distribution: 60 + GAMM(841, 1.37)

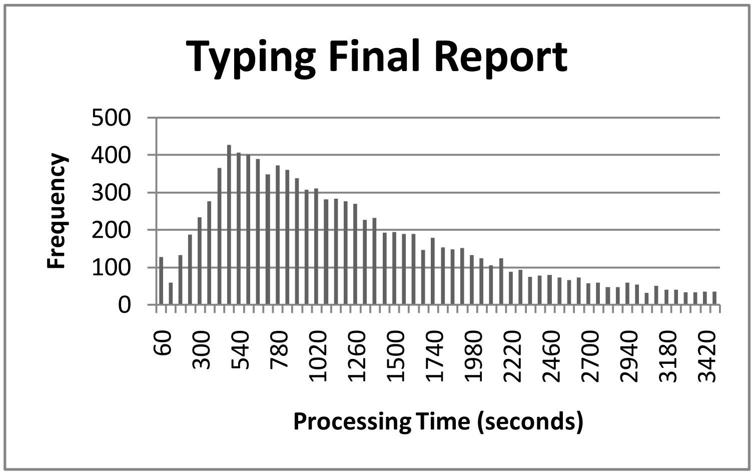
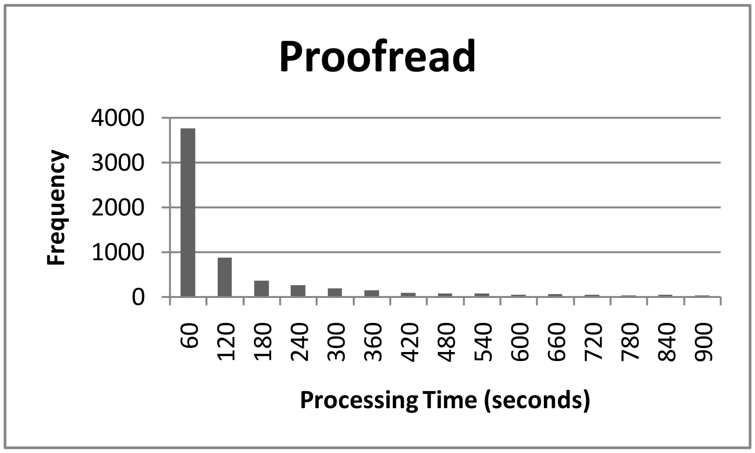
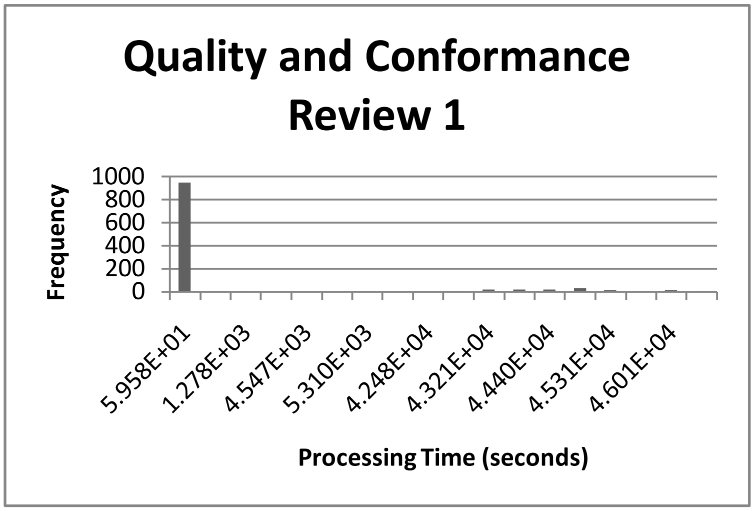


Figure 11 Histogram Proofread



The distribution for the processing times for proofreading is an exponential distribution: 60 + EXPO(81.3)

Figure 12 Histogram Quality and Conformance Review 1



The processing time distribution for the first quality and conformance review is a beta distribution: 3 + 4.74e+004 * BETA(0.21, 0.875)

Figure 13 Histogram Quality and Conformance Review 2

The processing time distribution for the second quality and conformance review is a beta distribution: 6 + 2.67e+005 * BETA(0.914, 1.18)

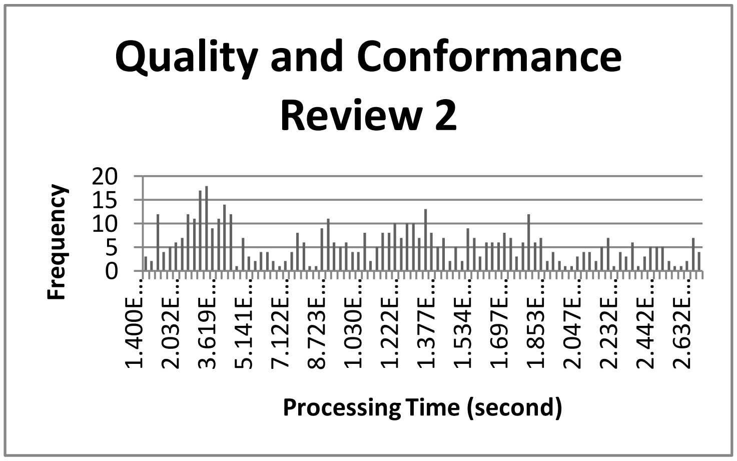
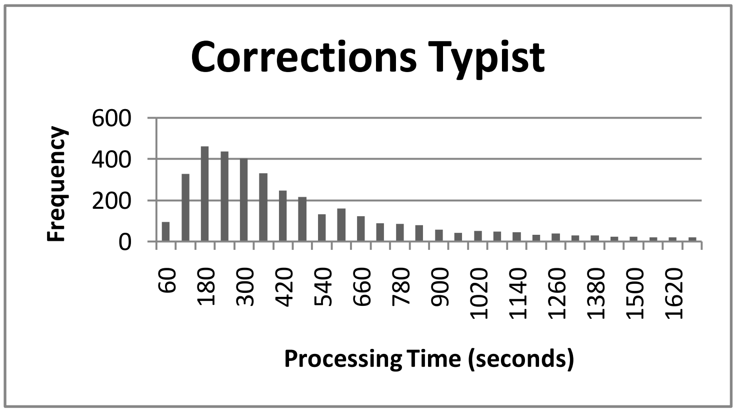
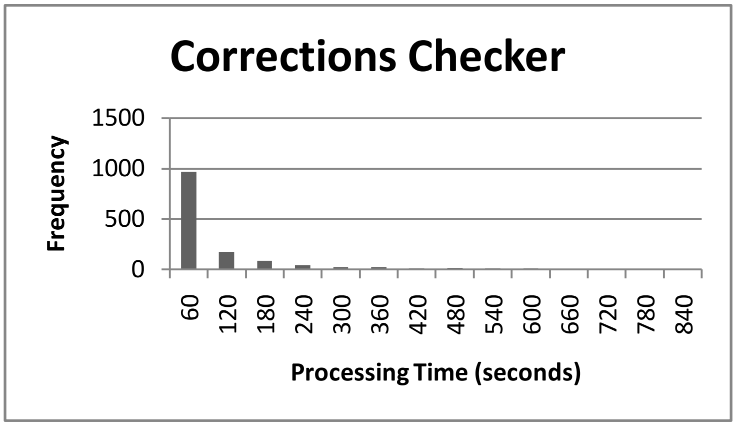


Figure 14 Histogram Corrections Typist



The processing time distribution for the corrections typists is exponential: 60 + EXPO(401)

Figure 15 Histogram Corrections Checker



The processing time distribution for the corrections checker is exponential: 60 + EXPO(62.5)

Figure 16 Histogram Final Scan

The processing time distribution for the scanning of the final report is exponential: 60 + EXPO(98.6)

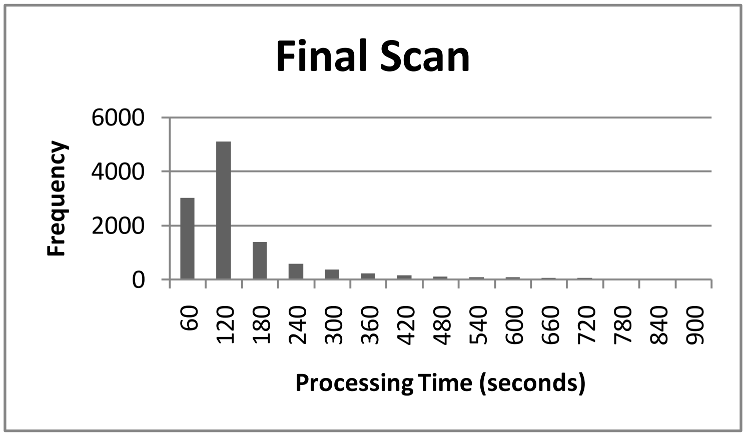
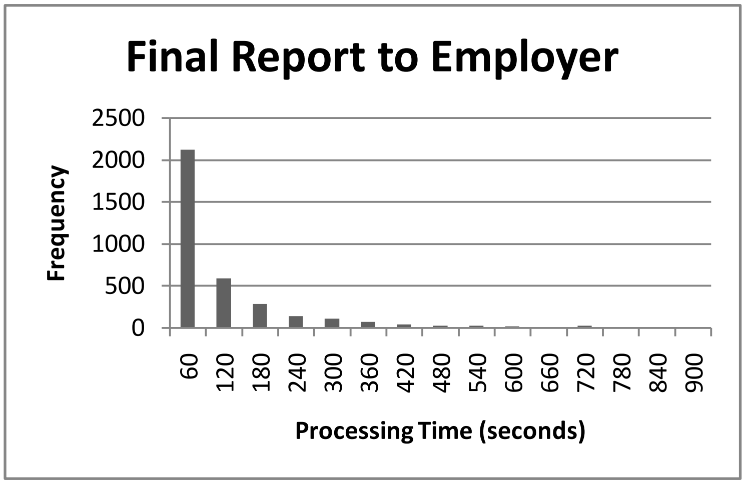
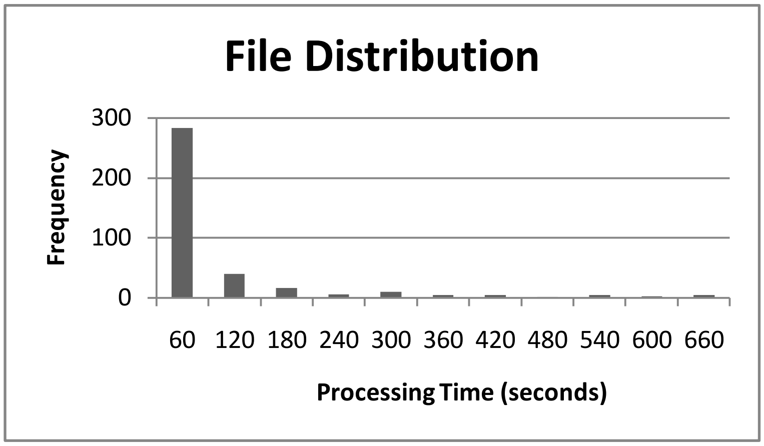


Figure 17 Histogram Final Report to Employer



The processing time distribution for sending the final report to the client is exponential: 60 + EXPO(64.8)

Figure 18 Histogram File Distribution



The processing time distribution for distributing the files is exponential: 60 + EXPO(43.1)

APPENDIX B

Figure 19 Flow of Short Temporary Incapacity Leave Applications

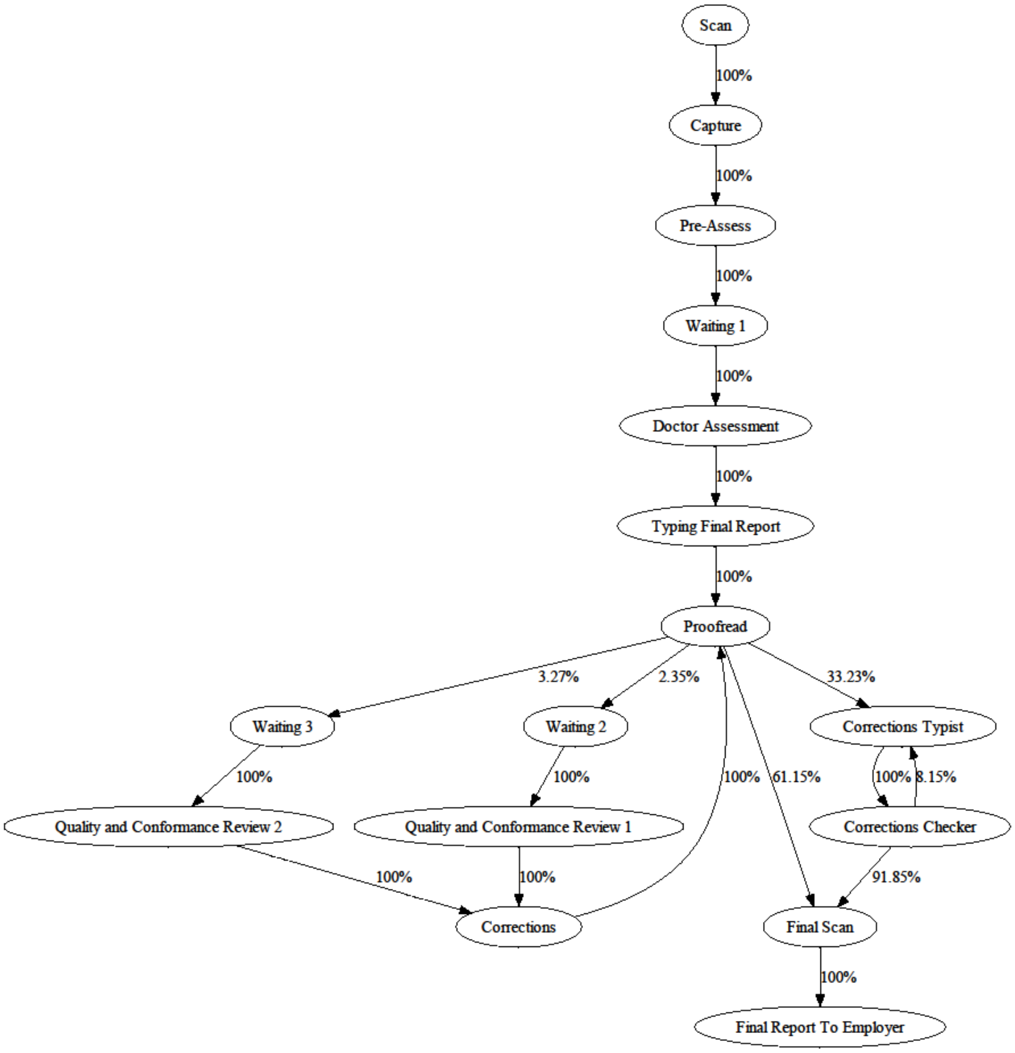
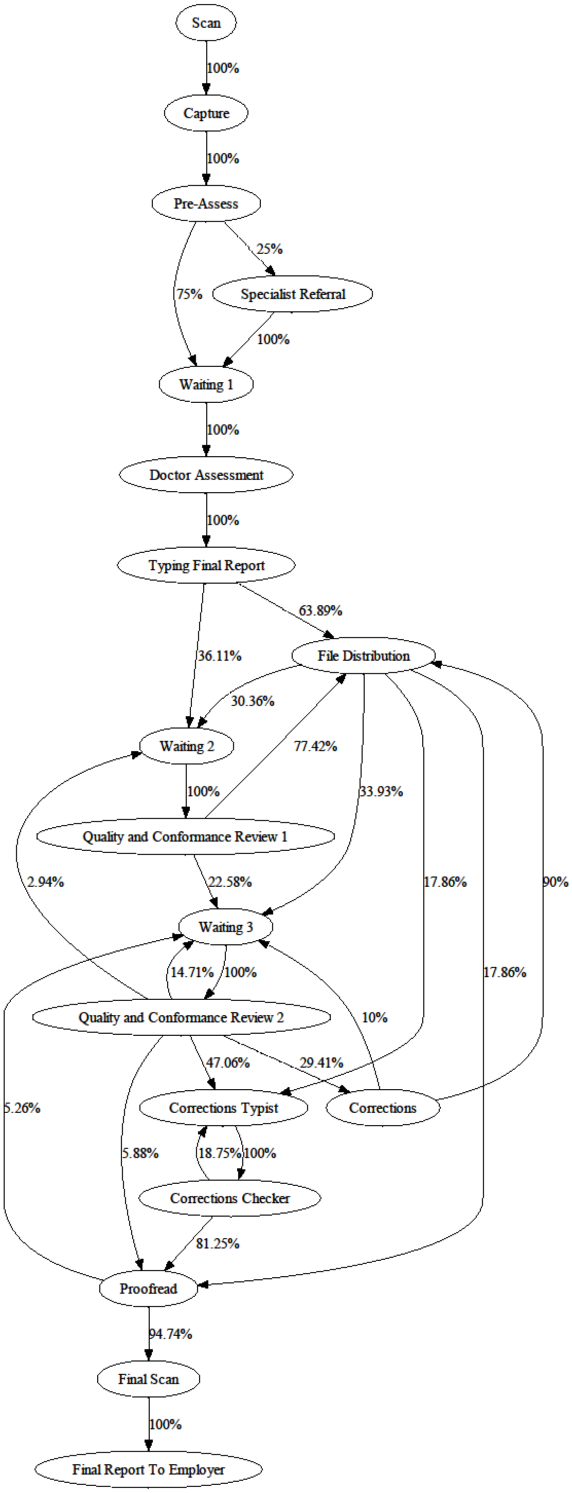
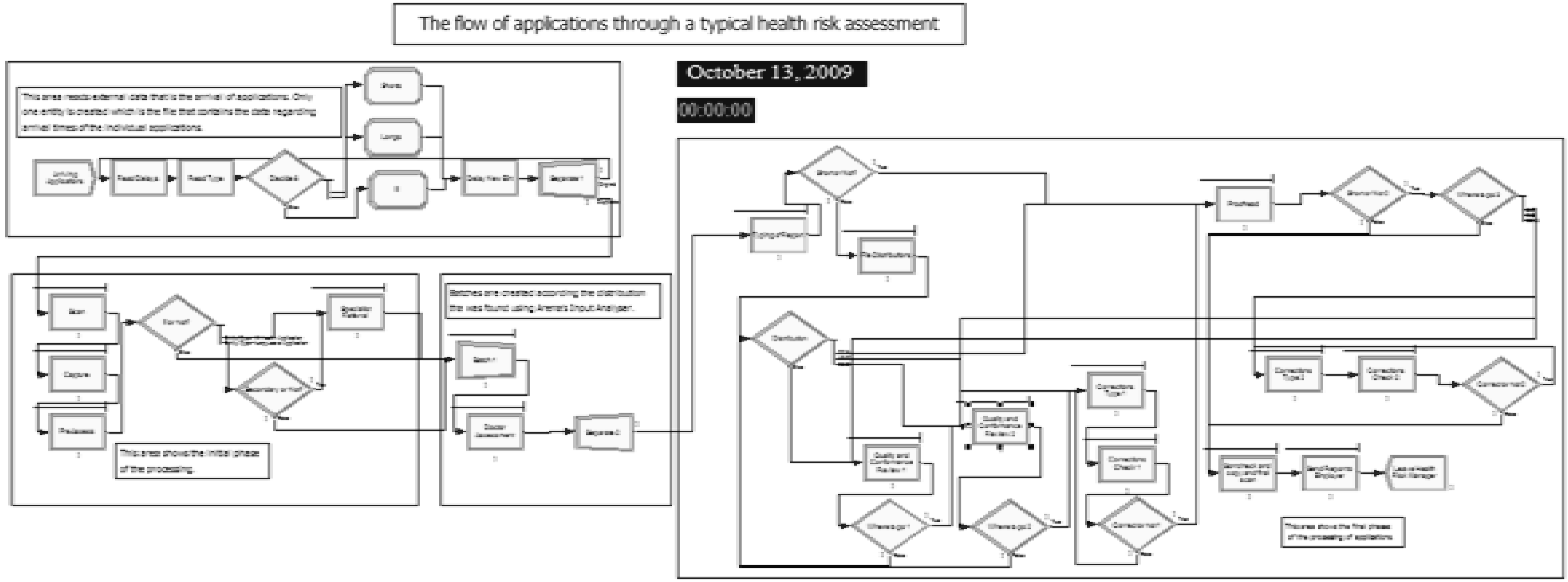


Figure 20 Flow of Long Temporary Incapacity Leave Applications



APPENDIX C

Figure 21 The Overall view of the simulation model



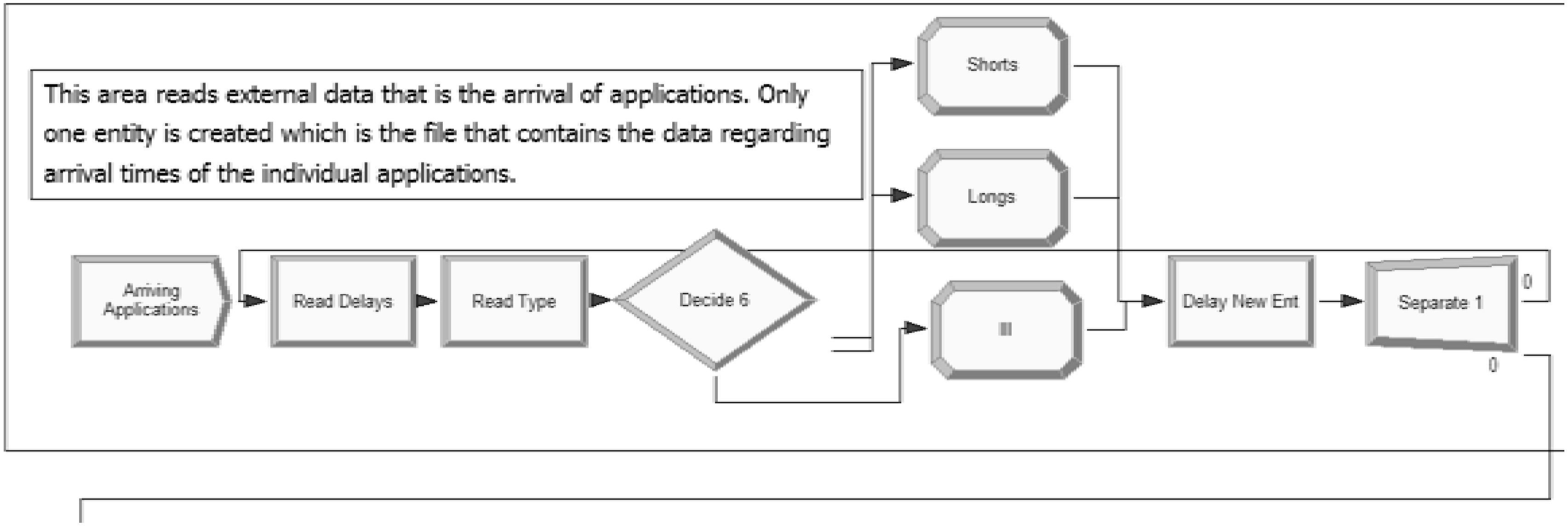
This area reads external data that is the arrival of applications. Only one entity is created which is the file that contains the data regarding arrival times of the individual applications.

This area shows the initial phase of the processing.

Batches are created according the distribution the was found using Arena's Input Analyser.

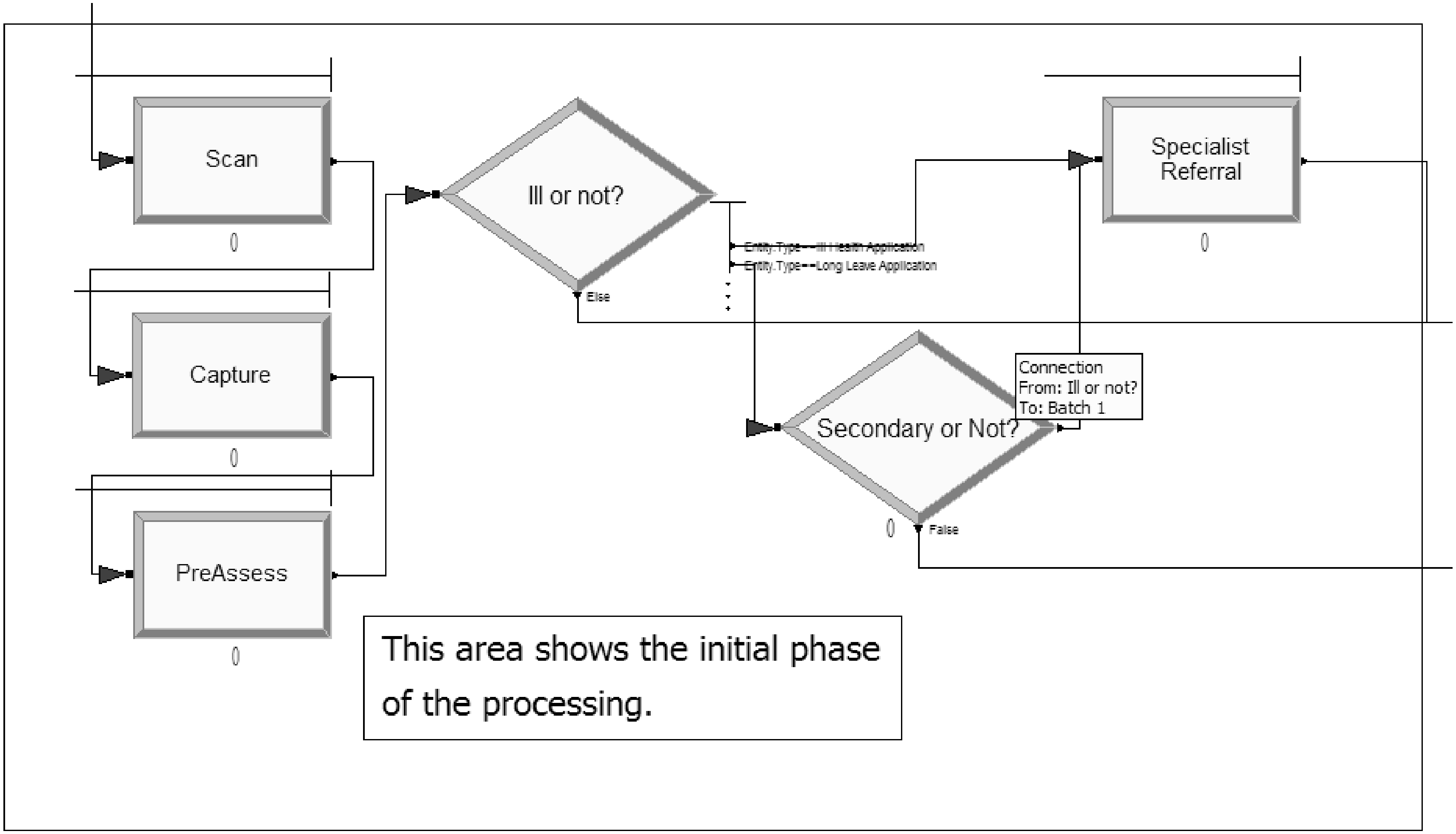
This area shows the final phases of the processing of applications.

Figure 22 Entity arrivals: Actual arrival times of applications being fed into Arena



This area reads external data that is the arrival of applications. Only one entity is created which is the file that contains the data regarding arrival times of the individual applications.

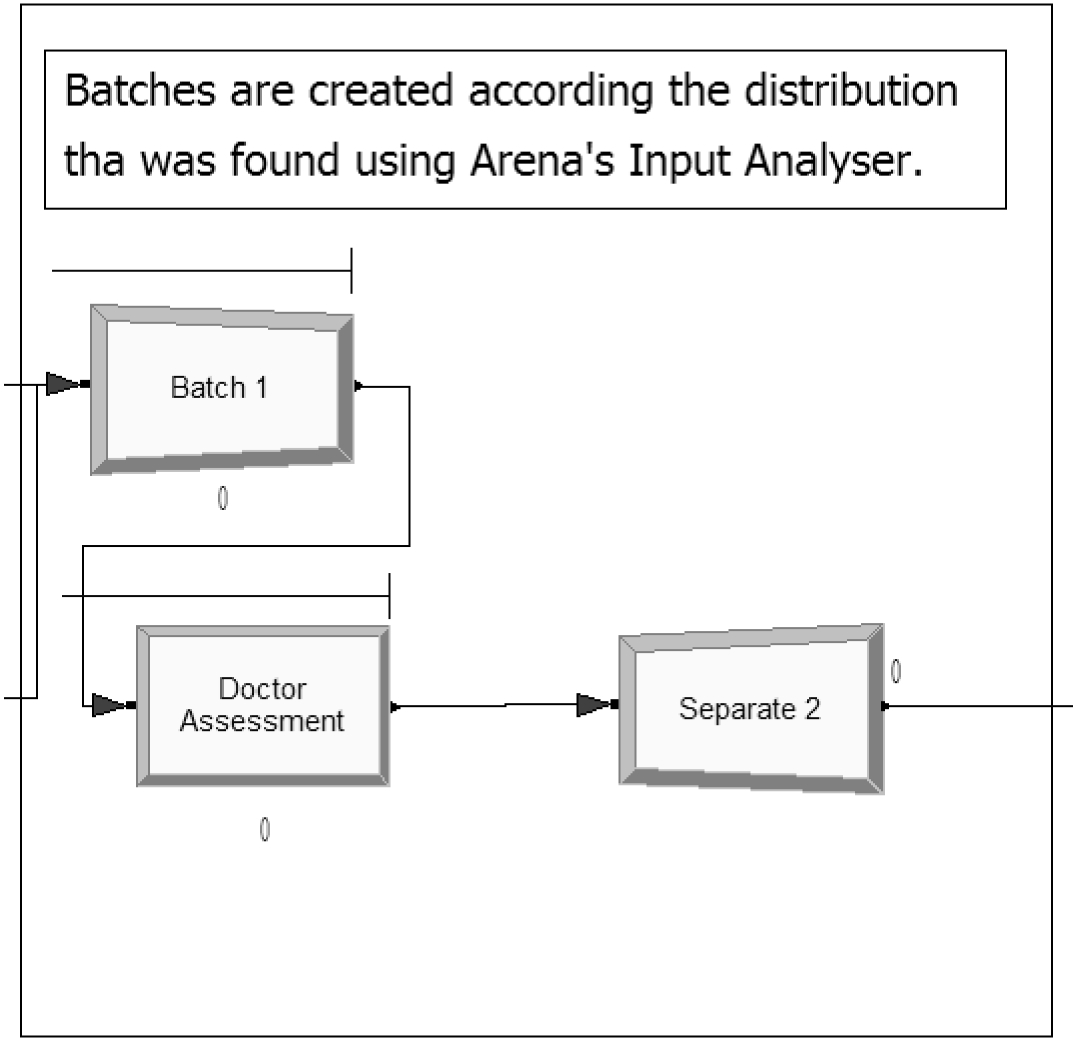
Figure 23 The Initial Stages of the System



Connection
From: Ill or not?
To: Batch 1

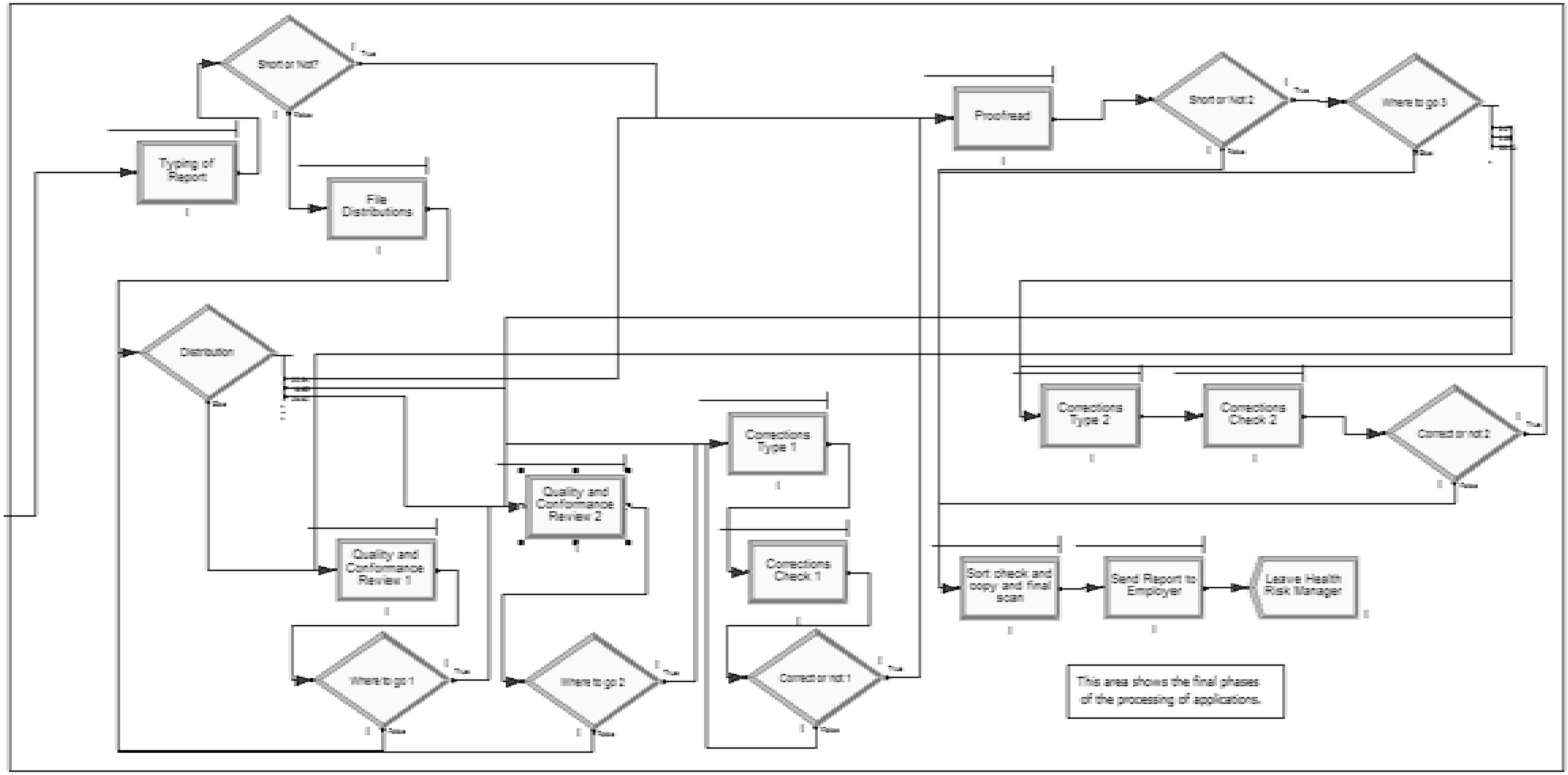
This area shows the initial phase of the processing.

Figure 24 The middle phase where a time buffer is present



Batches are created according the distribution tha was found using Arena's Input Analyser.

Figure 25 The final phases of the processing system


Short or Not?
True
False
Typing of Report
File Distributions
Distribution
Else
Quality and Conformance Review 1
Quality and Conformance Review 2
Where to go 1
Where to go 2
True
False
Corrections Type 1
Corrections Check 1
Correct or not 1
Proofread
Short or Not 2
Where to go 3
Corrections Type 2
Corrections Check 2
Correct or not 2
Sort check and copy and final scan
Send Report to Employer
Leave Health Risk Manager
This area shows the final phases of the processing of applications.