Bachelor's thesis

Business Information Technology

Information systems

2012

Mira Kanervo

# CAPACITY MODELING OF IT SYSTEMS

TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR´S THESIS | ABSTRACT
TURKU UNIVERSITY OF APPLIED SCIENCES

Business Information Technology | Information systems

2012 | 37 + 13

Tuomo Helo

Mira Kanervo

# CAPACITY MODELING OF IT SYSTEMS

This thesis is about capacity modeling of IT systems. Capacity modeling of IT systems is a part of activities that should be made in capacity management. Capacity modeling means performance modeling and throughput forecasting. There are different modeling techniques of which the most popular ones are trending, simulation modeling and analytical modeling. Trending is a technique where historical data is used to forecast future behavior. Simulation modeling is a technique where simulation models are used to emulate the structure of the system. Analytical modeling is a technique where mathematical techniques are used to represent the behavior of computers.

Capacity model is a general term for models where earlier mentioned modeling techniques are used. The main purpose of capacity models are to make performance predictions and to create what-if-analyses.

In the empirical part of this thesis a qualitative research was made using semi-structured interviews which had also aspects of unstructured interviews. The intention was to gather all the expertise inside Nokia IT about capacity models of IT systems and collect requirements for a generic capacity model of IT systems. This thesis was made as an assignment to Nokia.

In conclusion, various capacity modeling techniques should be used in good capacity management and capacity models based on those should be made and used.

KEYWORDS:

Capacity, performance, modeling

OPINNÄYTETYÖ (AMK) | TIIVISTELMÄ
TURUN AMMATTIKORKEAKOULU

Tietojenkäsittely | Tietojärjestelmät

2012 | 37 + 13

Tuomo Helo

Mira Kanervo

# IT-JÄRJESTELMIEN KAPASITEETIN MALLINNUS

Tämä opinnäytetyö käsittelee IT-järjestelmien kapasiteetin mallinnusta. Kapasiteetin mallinnus on yksi niistä toimenpiteistä, joka kapasiteetin hallinnassa pitäisi tehdä. Kapasiteetin mallinnus tarkoittaa IT-järjestelmien suorituskyvyn mallintamista ja suoritustehon ennustamista. On olemassa erilaisia mallinnustekniikoita, joita käytetään. Näistä suosituimpia ovat trendianalyysi, simulaatiomallinnus ja analyyttinen mallinnus. Trendianalyysissa käytetään hyväksi historiallista dataa tulevaisuuden käyttäytymisen ennustamisessa. Simulaatiomallinnuksessa tehdään simulaatiomalleja, jotka emuloivat järjestelmän rakennetta. Analyyttisessa mallinnuksessa käytetään hyväksi matemaattisia tekniikoita tietokoneiden käyttäytymistä esitettäessä.

Kapasiteettimalli on yleisnimitys malleille, joissa näitä aiemmin mainittuja mallinnustekniikoita käytetään. Kapasiteettimallien päätarkoitus on tehdä suorituskyvyn ennustamista ja luoda ”mitä jos” -analyyseja.

Tämän työn empiirisessä osiossa tehtiin kvalitatiivinen tutkimus käyttäen puolijäsenneltyjä haastatteluja, joissa oli piirteitä myös jäsentelemättömistä haastatteluista. Näiden haastatteluiden avulla oli tarkoitus kerätä kaikki asiantuntijuus Nokia IT:n sisällä IT-järjestelmien kapasiteettimalleista ja kerätä vaatimuksia geneeriselle IT-järjestelmien kapasiteettimallille. Tämä työ on tehty toimeksiantona Nokialle.

Yhteenvetona, hyvässä kapasiteetin hallinnassa pitäisi käyttää erilaisia kapasiteetin mallinnustekniikoita ja näihin perustuvia kapasiteettimalleja pitäisi luoda.

ASIASANAT:

kapasiteetti, suorituskyky, mallintaminen

# CONTENT

## PICTURES

## FIGURES

## TABLES

# 1 INTRODUCTION

This thesis is about capacity modeling of IT systems. Capacity modeling is one part of capacity management. Capacity management is important in today's business world because with good capacity management companies can optimize their usage of capacity and achieve great savings.

My thesis aims to answer the following questions:

- What does capacity modeling of IT systems mean?
- What kind of different capacity modeling techniques are there?
- What is capacity model?
- What needs to be taken into account when creating a capacity model?

This thesis is made as an assignment to Nokia and Nokia has a need for a generic capacity model for IT systems. The goal of the empiric part of the thesis is to gather all the expertise inside Nokia IT about capacity models of IT systems and collect requirements for a generic capacity model of IT systems using interviews for data collection.

Nokia is a well-known Finnish company which was founded in 1865. During the years Nokia has evolved from a local paper mill to a global company in telecommunications. Nokia employs approximately 139 000 people globally. Nokia consists of 4 teams: Smart Devices, Mobile Phones, Location & Commerce and Markets. IT is part of the Markets team. (Nokia 2011a; Nokia 2011b; Nokia 2011c.)

In chapter 2 the theory of ITIL is examined. I chose ITIL to be examined for this thesis because it is widely used in Nokia IT. It also works as a good introduction to capacity management. ITIL and the Service Lifecycle are briefly introduced and then the focus is on the parts that are most important for this thesis. These are the Service Design phase and within that the Capacity Management process. The relevant parts of them are introduced.

The theory of capacity modeling of IT systems can be found in chapter 3. First it is defined what capacity is and what kinds of capacities exist. After that the focus is on the modeling of capacity and the different capacity modeling techniques. The term capacity model is also defined. In the end of the chapter it is defined what capacity modeling in practice means.

Chapter 4 is about the empirical part of this thesis. The purpose and the background for the empirical part of thesis are introduced. In this chapter it is described how the qualitative research using interviews for data collection was made and how the data was processed and analyzed.

Chapter 5 includes the results of the interviews. Unfortunately this is confidential data and will not be published in this thesis.

In chapter 6 there are the conclusions of my thesis. This chapter describes what I have learned during this thesis process and how well the research questions defined in the introduction were answered. Also what the best part of this thesis process was and what was the hardest part will be revealed.

# 2 ITIL

## 2.1 Service Lifecycle

ITIL also known as The Information Technology Infrastructure Library is a public framework in which there has been collected best practices in IT service management (Cartlidge et al. 2007, 8; van Bon et al. 2007, 13). ITIL provides a framework for the IT governance and focuses on the continual measurement and quality improvement of IT services from business and from a customer perspective (Cartlidge et al. 2007, 8).

ITIL was developed in the 1980s and 1990s and it has been updated two times, in 2000-2002 for version 2 and in 2007 for version 3, which is still in use (van Bon et al. 2007, 13). The current version of ITIL (V3) consists of five core elements that forms the Service Lifecycle which is represented in Figure 1 (Cartlidge et al. 2007, 8; Lucid IT 2007).

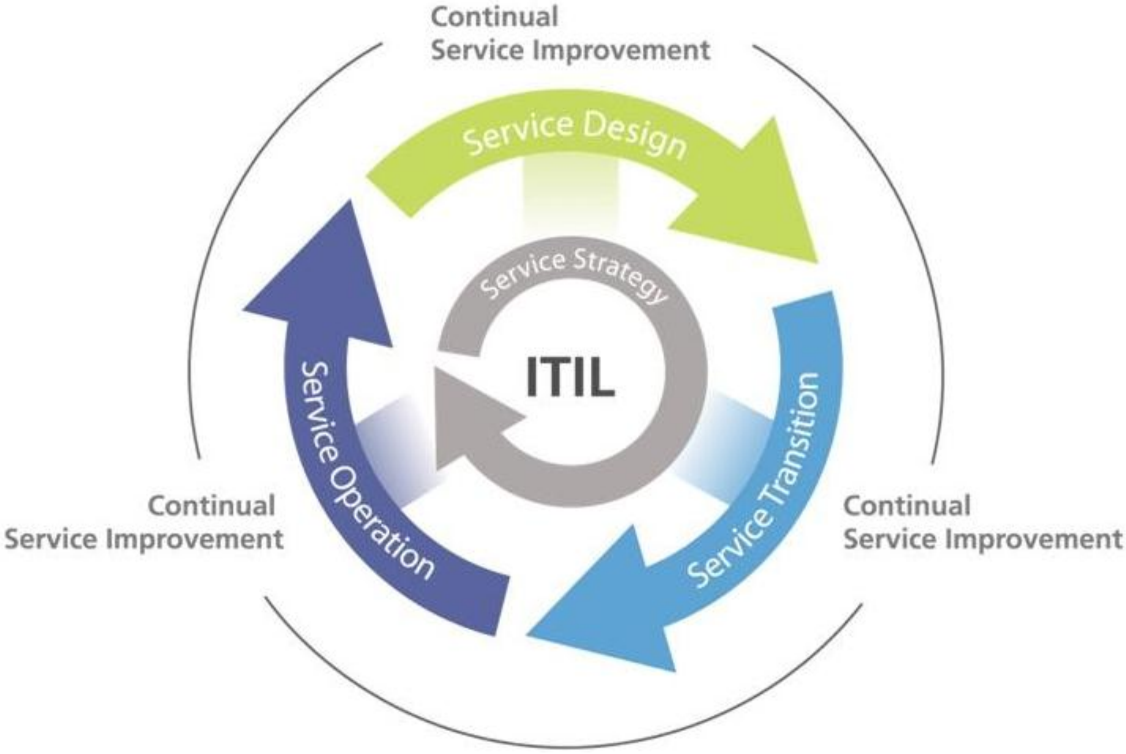


Figure 1. The Service Lifecycle (Lucid IT 2007).

The Service Lifecycle is a model that describes how service management is structured, how various lifecycle components are linked together and how the changes in one component will affect other components and the entire lifecycle. It consists of five phases: Service Strategy, Service Design, Service Transition, Service Operation and Continual Service Improvement. (van Bon et al. 2007, 19-20.)

Service Strategy is a phase where objectives are set and is a centre of the lifecycle that drives other phases (van Bon et al. 2007, 21). It provides guidance on how to view service management as a strategic asset and not only an organizational capability (Office of Government Commerce 2007a, 11).

To provide value to the business, services must be designed in a way that business objectives are taken into consideration. Service Design is a phase where objectives that have been set in Service Strategy phase turn into a blueprint. It provides design and development guidance for the new and changed services and service management practices. (OGC 2007a, 11.)

The role of Service Transition phase is to ensure what was planned in Service Design phase will achieve the expected objectives in implementation phase. In practice it means planning and managing the resources which are required in package, build, test and deploy phases before a release goes into the production. (OGC 2007a, 75.)

Service Operation phase is responsible for delivering and managing services at agreed levels to the business users and customers in everyday use of the services. In practice it means coordinating and carrying out the day-to-day operations of processes that have been planned and implemented in previous lifecycle phases. (OGC 2007a, 93.)

The purpose of Continual Service Improvement (hence CSI) phase is to continually align and re-align and improve the whole service lifecycle so that it would correspond to the changing business needs. CSI is also looking for ways to improve process and cost effectiveness. (OGC 2007a, 125-126.)

In my thesis I will concentrate on the Service Design phase, because it is the most relevant for my thesis.

## 2.2 Service Design

Like earlier was mentioned in chapter 2.1, Service Design is a phase where objectives that have been set in Service Strategy phase turn into blueprint. The main purpose of this phase is to design new or changed services from beginning to the live environment by taking care of all the aspects of design (OGC 2007b, 13).

In Service Design phase there are five aspects of design that need to be considered. These aspects of design are earlier mentioned design of new or changed services, design of service management systems and tools, design of technology architecture and management systems, design of the processes and design of measurement methods and metrics. (OGC 2007b, 14.)

When designing new or changed services, it starts with a set of business requirements that have been extracted from the Service Portfolio which includes all services that are managed by a service provider (OGC 2007a, 216; OGC 2007b, 15). Each of these requirements are analysed, documented and agreed. Then a solution design is produced and compared with strategies that have been defined in Service Strategy phase to ensure that the solution follows corporate and IT policies. (OGC 2007b, 15.)

The purpose of the design of the service management systems and tools is to ensure that new or changed service is consistent with all other services that already exist and that all other services that somehow depend on the new service are consistent. If they are not consistent either the design of the new service or the existing services need to be adapted. To ensure that service management systems and tools are capable of supporting the new service they should be reviewed. (OGC 2007b, 15.)

The purpose of the design of the technology architectures and management systems is to ensure that all existing technology architectures and management systems are consistent with the new service and they are capable of operating and maintaining the new service. If they are not capable of doing that either the architectures or management systems need to be improved or the design of the new service need to be revised. (OGC 2007b, 15.)

The purpose of the design of the processes is to ensure that the processes, roles, responsibilities and skills are capable of operating, supporting and maintaining the new service. If they are incapable of doing those things then the design of the new service need to be revised or enhance the existing process capabilities. (OGC 2007b, 15.)

The purpose of the design of the measurement and metrics is to ensure that required metrics that need to be provide for the new service can be provide with existing measurement methods. If they cannot provide those then the measurement methods need to be enhanced or the service metrics need to be revised. (OGC 2007b, 15.)

If all of these aspects of design are considered and completed during the Service Design phase this will ensure that during the other phases in the Service Lifecycle there will be only a few issues arising (OGC 2007b, 15).

Other important thing in Service Design phase are the processes. Service Design phase consists of seven processes which are Service Catalogue Management, Service Level Management, Capacity Management, Availability Management, IT Service Continuity Management, Information Security Management and Supplier Management (OGC 2007b, 19). True value of these processes will be realized when interfaces between the processes are identified and actioned hence these processes should not be considered in isolation (OGC 2007b, 18).

However, in my thesis I will concentrate on Capacity Management process because it is the most relevant for my thesis.

### 2.2.1 Capacity Management

Capacity Management process is responsible for the fact that all areas of IT have cost-justifiable IT capacity all the time. It needs to match with the current and future business needs. (OGC 2007b, 19.) Capacity Management process extends across the Service Lifecycle but it is included in Service Design phase because designing capacity is such an essential part of good capacity management (OGC 2007b, 79).

Capacity Management process is responsible for providing management for all service and resource capacity and performance related issues. It should be the focus for all capacity and performance issues. It should contain all areas of technology including hardware and software for all components and environments. It should also contain environmental systems capacity, space planning and some aspects of human resources. (OGC 2007b, 79.)

There are several objectives of Capacity Management process:

- Create and maintain Capacity Plan which is up-to-date with current and future business needs.
- Make assessments on how changes in the Capacity Plan will impact the performance and capacity of all services and resources.
- Ensure that agreed performance targets are met or exceeded by service performance achievements.
- Ensure that proactive measurements for improving the performance of services are implemented when it is cost-justifiable.
- Assist with capacity and performance related incidents and problems.
- Provide guidance on all capacity and performance related issues to all other areas of IT and the business. (OGC 2007b, 79.)

Capacity Management needs to understand IT and business environments to ensure that all current and future performance and capacity aspects of services are managed cost-effectively (OGC 2007b, 80).

Capacity Management process enables organizations to decide which components to upgrade, when to do it and how much it will cost by providing information on current and planned resource utilization (OGC 2007b, 80).

When Capacity Management is properly carried out it can forecast business events and impacts before they even occur. It makes sure that there will be no surprises regarding service and component design and performance. (OGC 2007b, 81.)

### 2.2.2 Sub-processes

Capacity Management process is divided to three supporting sub-processes: Business Capacity Management, Service Capacity Management and Component Capacity Management (OGC 2007b, 82-83).

The purpose of the Business Capacity Management is to translate business needs and plans into service and IT infrastructure requirements. It also ensures that the future business requirements for IT services are determined, designed, planned and implemented in timely manner. This can be done by using the data on the current resource utilization by the various services and resources that already exist. With these future requirements can be trended, forecasted, modeled or predicted. (OGC 2007b, 82.)

Service Capacity Management is responsible for identifying and understanding the IT services, their resource use, working patterns and peaks & troughs. It also ensures that the services perform as they should and meet their targets that have been defined in Service Level Agreement (hence SLA) which is an agreement between customer and an IT service provider. (OGC 2007b, 86; OGC 2007a, 215.)

The main purpose of Component Capacity Management is identifying and understanding capacity, performance and utilization of every component that is used in technology which supports the IT services. These IT services are the infrastructure, environment, data and applications. Component Capacity Management ensures the current hardware and software resources are used as optimal as possible to achieve and maintain the service levels that have been agreed. (OGC 2007b, 86.)

### 2.2.3 Proactive / Reactive

In Capacity Management process are of two kinds of activities to handle management: proactive and reactive activities. The more properly managed proactive activities Capacity Management includes the less will be reactive activities needed. (OGC 2007b, 84.)

The proactive activities should include for example:

- Preventing performance issues by taking necessary actions before they even occur.
- Producing trends regarding current component utilization.
- Estimating the future requirements by using thresholds and trends for planning enhancements and upgrades.
- Trending and modeling the predicted changes in services and ensuring that resources needed for those changes are available.
- Ensuring that upgrades are budgeted, planned and implemented as have been agreed.
- Seeking actively service performance improvements wherever it's cost-justifiable. (OGC 2007b, 84.)

The reactive activities should include for example monitoring, measuring, reporting and reviewing the current performance of components and services and reacting to and assisting with certain issues of performance (OGC 2007b, 84).

### 2.2.4 Application Sizing

Application sizing aims to estimate the resource requirements that are needed when implementing a new service or proposing change to an existing service to ensure that it will meet agreed service levels (OGC 2007b, 93).

Application sizing is initiated when designing new service or when making major change in an existing service. It is completed when the application is transferred into the live operational environment. It should include not just applications themselves but all technology areas related to applications. These include infrastructure, environment and data. Application sizing uses also modeling techniques which will be introduced in chapter 3. (OGC 2007b, 93.)

# 3 CAPACITY MODELING OF IT SYSTEMS

## 3.1 Definitions for Capacity

There are different definitions for capacity. According to An encyclopædia Britannica Company Merriam-Webster (2011) it could mean e.g. “the potential or suitability for holding, storing, or accommodating”, “the maximum amount or number that can be contained or accommodated” or “maximum output”.

In capacity management point of view capacity is defined as the maximum throughput that a service can deliver while meeting the agreed performance targets. Capacity and performance are defined like this:

> “Performance and capacity are the two key issues for capacity management. Performance is essentially about the speed at which something can be completed and is typically measured as a response time. Throughput is the count of the number of such things completed in a given time. Capacity is the number of such things that could be completed in a given time and is the maximum throughput.” (Grummitt 2009, 27.)

## 3.2 Existing Capacities

There is no such list that would define what kinds of capacities exist. However there are things that are monitored and can reach their maximum throughput and thus can exceed or run out. In capacity management these things usually are processor utilization, memory utilization, I/O rates, queue lengths, disk utilization, response times and network traffic rates (OGC 2007b, 87-88).

There are also things that require ongoing care and can reach their maximum throughput and thus can exceed or ran out but are not usually perceived as capacity but they are. These things are hardware leasing, software licensing, floor space and facilities infrastructure including power, heating, ventilating, air-conditioning and racks (Betz 2007, 7). Also human resources should be considered as capacity (Betz 2007, 79).

## 3.3 Capacity Modeling

Capacity modeling includes all the techniques of performance modeling and throughput forecasting (Grummitt 2009, 65). Modeling and prediction are important parts of proactive capacity management and they are used to predict the behavior of IT services (OGC 2007b, 92; Grummitt 2009, 6). When wanted to guarantee service levels, it is important to understand the current services and what it takes to create them. It is also important to predict future resources that will be needed. In both of these cases modeling is used. (Potter 2008, 2.)

Capacity modeling is used to tell what workload can be supported with given resources or what service can be given for a planned workload. Workload is the amount of a resource use in a certain period. It usually means the throughput of work for certain groups of users or functions in the organization. (Grummitt 2009, 27 & 65.)

Capacity modeling includes various approaches to the forecasting. For a long time forecasting has been a technique that is used to try to avoid problems. It is used to assess changes in business demand and service workload behavior and to assess the impact of those changes on the resources available to support it. The approach to forecasting depends on how accurate information is required and how much costs are involved when achieving it. (Grummitt 2009, 65.)

There are different types of capacity modeling techniques from making estimates based on information on current resource utilization and experience to making prototypes, full scale benchmarks and pilot studies. All of these have good and bad sides and are suitable for different purposes. It is possible to obtain similar levels of accuracy with all types of modeling but it is dependent on the information that is used when creating it and the skill of the person who is modeling. (OGC 2007b, 92.)

The three most popular capacity modeling techniques are trending, simulation modeling and analytical modeling (Grummitt 2009, 30). Next sub-chapters concentrate on them.

### 3.3.1 Trending

Trending, also known as a trend analysis, is a modeling technique where historical data about resource utilization and service performance is used to forecast future behavior. The historical data is usually analyzed in a spreadsheet where graphical, trending and forecasting facilities are used to show resource utilization over a time and how it is likely to change in the future and define an expected growth. (OGC 2007b, 92; Grummitt 2009, 30.)

For the things that behave in straight lines trending provides a sufficient analysis. In practice it means there are some cases where trending is viable approach and some cases where it is not. There are two key points that decide if it is a good approach. The first one is if that there should not be unknown discontinuities. The other one is that the underlying attributes of a system should be linear. (Grummitt 2009, 67.)

Trending is the most effective when there are a small number of variables and between them is linear relationship (OGC 2007b, 92). It is quite affordable modeling technique but it only provides the future resource utilization estimations so it is not very accurate (OGC 2007b, 92; Potter 2008, 3).

### 3.3.2 Simulation Modeling

Simulation modeling is a technique in which simulation models are based on computer programs that emulate static structure and the different dynamic aspects of a system (Menascé et al. 2004, 36). Simulation modeling is used before making decisions about hardware purchases or load allocation (Potter 2008, 3). Simulation modeling uses a model where the traffic is defined and it is compared to a simulation of the configuration until it finds a solution (Grummitt 2009, 68).

Simulation modeling includes the modeling of discrete events for example running transaction arrival rates against defined hardware configuration. That is the reason it is also known as discrete event simulation (hence DES). (OGC 2007b, 92; Grummitt 2009, 66.)

DES is used to simulate the service, dispatch and arrival of workloads which are for example database queries, transactions and batch jobs. DES also analyses the results of the simulations to produce information about for example utilizations, arrivals or leavers, and queue lengths. (Grummitt 2009, 69.)

Simulation modeling gives a great view into the current and future operations and it can be accurate when predicting the effects of changes on existing applications or when sizing new applications (OGC 2007b, 92; Potter 2008, 3). Downside of this technique is that it often takes a long time to build and execute the model and therefore it is costly (OGC 2007b, 92; TeamQuest 2011, 2).

### 3.3.3 Analytical Modeling

Analytical modeling represents the behavior of the computer system using mathematical techniques, for example, queuing theory (OGC 2007b, 92). In queuing theory congestions and delays of waiting in line are analyzed using a mathematical method (Investopedia 2011).

Analytical models consist of formulas where the traffic and configuration have been defined, and algorithms are used to solve the formulas to provide the needed answers. Analytical models may be used to current performance assessment and future performance prediction. (Grummitt 2009, 30 & 68.)

To generate queuing network models (hence QNM) analytical modeling uses queuing network theory, which is a variation of queuing theory that includes network of queues used to model a system (Grummitt 2009, 66; Encyclopedia.com 2011).

QNM makes assumptions and solves analytical equations based on for example service requirement distributions, I/O device visit ratios, statistical characteristics of workload arrival rates and workload populations. QNM produces queue lengths, utilization levels, throughputs and response times. (Grummitt 2009, 69.)

To be mathematically flexible analytical models include usually a little detail and therefore they tend to be efficient to run but not so accurate as other modeling techniques (Menascé et al. 2004, 37). They also do not take that much time to create but they must be kept updated (OGC 2007b, 92).

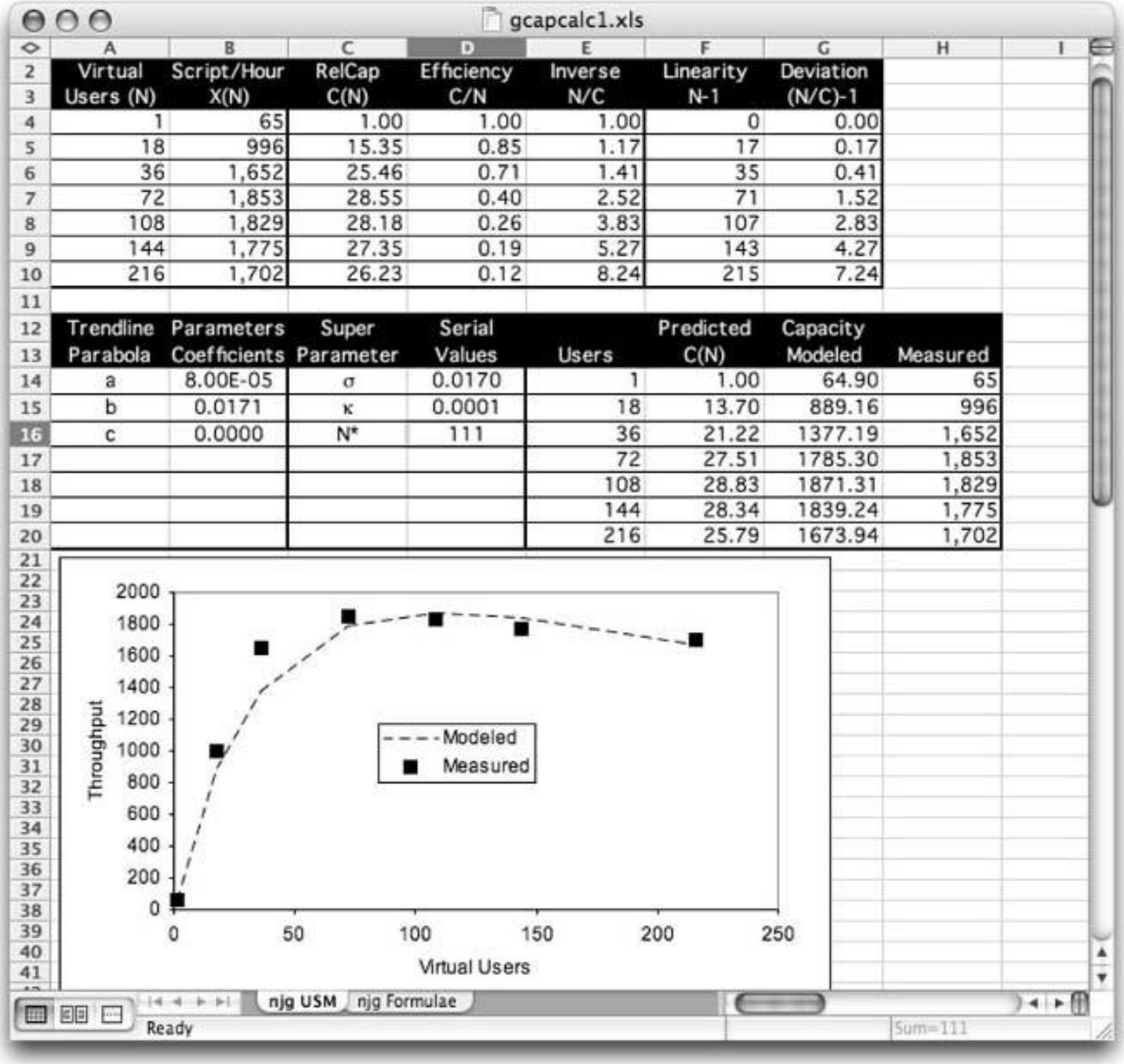
## 3.4 Capacity Models

’Capacity model’ is a general term for the models where modeling techniques that have been introduced above are used. The purpose of capacity models is to make performance predictions and to create what-if analyses to see what would happen in different situations for example how failures and planned changes will reflect to the hardware and variety of workloads (OGC 2007b, 92; TeamQuest 2011, 1).

Capacity models are needed when good capacity management is wanted. When making decisions about capacity, accurate information about the need, changes or problems in capacities are essential and capacity models should offer those. (Savvides 2004, 28.)

Capacity models can be simple spreadsheets or commercial software which is integrated into the service management tools and enterprise system monitoring. Despite the type of capacity model, it is important that it is used the right way. Entire business context should be taken into consideration to make decisions about capacity that are justified. (Savvides 2004, 28.)

Capacity models should include the current set-up that reflects achieved performance and proposed set-up which will reflect the future state (Savvides 2004, 28; OGC 2007b, 92). Capacity models then can be used to specify and justify the changes that should be made in the concrete set-up (Savvides 2004, 28). These aspects can also be found in Picture 1 which illustrates an example of spreadsheet capacity model.



| Virtual Users (N) | Script/Hour X(N) | RelCap C(N) | Efficiency C/N | Inverse N/C | Linearity N-1 | Deviation (N/C)-1 |
|---|---|---|---|---|---|---|
| 1 | 65 | 1.00 | 1.00 | 1.00 | 0 | 0.00 |
| 18 | 996 | 15.35 | 0.85 | 1.17 | 17 | 0.17 |
| 36 | 1,652 | 25.46 | 0.71 | 1.41 | 35 | 0.41 |
| 72 | 1,853 | 28.55 | 0.40 | 2.52 | 71 | 1.52 |
| 108 | 1,829 | 28.18 | 0.26 | 3.83 | 107 | 2.83 |
| 144 | 1,775 | 27.35 | 0.19 | 5.27 | 143 | 4.27 |
| 216 | 1,702 | 26.23 | 0.12 | 8.24 | 215 | 7.24 |

| Trendline Parabola | Parameters Coefficients | Super Parameter | Serial Values | Users | Predicted C(N) | Capacity Modeled | Measured |
|---|---|---|---|---|---|---|---|
| a | 8.00E-05 | σ | 0.0170 | 1 | 1.00 | 64.90 | 65 |
| b | 0.0171 | κ | 0.0001 | 18 | 13.70 | 889.16 | 996 |
| c | 0.0000 | N* | 111 | 36 | 21.22 | 1377.19 | 1,652 |
| | | | | 72 | 27.51 | 1785.30 | 1,853 |
| | | | | 108 | 28.83 | 1871.31 | 1,829 |
| | | | | 144 | 28.34 | 1839.24 | 1,775 |
| | | | | 216 | 25.79 | 1673.94 | 1,702 |

Picture 1. Example of spreadsheet capacity model (Gunther 2007, 10).

Data that will be needed in capacity models can be gathered in a capacity database which will collect and share requirements for performance and expected workloads and demands on IT resources. Current and trend information is stored in this database and it is used to build capacity models. (Savvides 2004, 28.) In chapter 3.4.4 can be found more accurate information about input data that will be used in capacity models.

### 3.4.1 What-if Analyses

Like earlier mentioned, one purpose of capacity models is to create what-if analyses. What-if analyses tell what happens in different situations (Savvides 2004, 28.) What-if analyses for example:

- Help to understand the impact of changes regarding workload growth, server consolidation, virtualized environments and hardware configurations.
- Will tell which systems are likely to run out of capacity and when.
- Predict the impact of changing business events to IT infrastructure.
- Can tell which will be the optimal hardware configuration supporting new application. (TeamQuest 2011, 1.)

### 3.4.2 Performance Prediction

The other purpose of capacity models is performance prediction which means that it helps to predict the performance of system. Performance prediction for example:

- Predicts IT performance on physical and virtualized platforms using techniques that have been introduced earlier in chapter 3.
- Predicts resource utilization, queue lengths, throughputs and response times.
- Determines the impacts of changes that have been made to hardware set-up for example size or number of CPUs, network bandwidth, I/O devices and memory.
- Forecasts how systems will respond to unexpected demand spikes. (TeamQuest 2011, 1.)

### 3.4.3 Inputs for Capacity Models

There are three different types of data that are used as an input for capacity models: machine configuration details, machine performance data and service reporting statistics & service forecasting quantities (Grummitt 2009, 28-29).

Machine configuration details include low-level resource data for example what kinds of host machines there are and what kinds of configuration they have for example what kinds of processors, memory they are including. In cases where virtualization is involved this kind of data tells what the virtualization parameters are. (Grummitt 2009, 28.)

Machine performance data includes application-level statistics for example transaction rates that are provided by middleware. This data is useful because it can create link between the low-level resource data and the high-level business data. (Grummitt 2009, 29.)

Service reporting statistics & service forecasting quantities are the high-level business data that was mentioned above. This kind of data includes for example reports of actual business activity, descriptions of critical scenarios for example seasonal peaks, and forecasts of future levels. (Grummitt 2009, 29.)

### 3.4.4 Things to Be Taken into Consideration when Creating a Capacity Model

Before creating the capacity model, there are two preparatory things to do. First one should get full understanding of the business issues that are involved and the application(s) that will be modeled. The better these things are understood, the better the model will support business functions, which is the ultimate goal. In practice this means getting and using realistic data that will be gathered from business units. Other thing that should be done is getting accurate data on the existing systems because without accurate input data it is impossible to create an accurate model. (Potter 2008, 4.)

It is not reasonable to build a model that takes more effort than gives benefit. The most important thing of the model is its purpose and it is relevant to know what questions should be answered with the model before creating it. (Grummitt 2009, 30.) Also it depends on the purpose of the model how detailed the model is and what aspects of the real system are considered in it (Menascé et al. 2004, 36). Any model is only as accurate as the data that is used in it (Gunther 2007, 7).

A model should be kept as simple as possible to achieve what is wanted (Menascé et al. 2004, 36). The purpose is to discard as much detail as possible but still maintain the essential parts and find the correct balance (Gunther 2007, 8). As Gunther (2007, 8) has paraphrased Einstein: “Keep the model as simple as possible, but no simpler.”

## 3.5 Capacity Modeling in Practice

In good capacity management, various capacity modeling techniques and different capacity modeling activities should be used in different phases. In practice this means different modeling activities should be done in different capacity management sub-processes that have been introduced in chapter 2. 2.3.

| **Business** | **Service** | **Resource/Component** |
|---|---|---|
| Use trend analysis to assess changes to meet new business demands | Use trend analysis to assess changes in service workloads | Use trend analysis to assess likely resource utilizations per device |
| Model systems behavior under varying changes to meet business needs | Model systems behavior under varying business forecasts for workload changes | Model systems behavior under varying workloads and provide tuning recommendations |
| Compare business driver forecasts with actual and reports | Adopt a modeling solution to predict future service levels | Adopt a modeling solution to analyze resource utilization data |
| Use business drivers to assess impact on workload and hence service levels | Use service level requirements (SLRs) as target in modeling exercises | Identify business metrics (drivers) to use as input to forecasts |

Table 1. Capacity Modeling in Practice (Grummitt 2009, 180).

In Table 1 the three different sub-processes are listed on top: business, service and resource/component. Under each sub-process the relevant actions are listed. These sub-processes and actions need to be defined and also assessed in the organization. (Grummitt 2009, 179-180.)

As one can see from Table 1, there are some similarities in the different sub-processes, like using trend analysis and modeling system behavior. But of course these sub-processes have their unique features also that make these suitable for their different purposes. (Grummitt 2009, 180.)

# 4 QUALITATIVE RESEARCH OF CAPACITY MODELS USING INTERVIEWS

## 4.1 The Purpose and the Background of My Empirical Part of the Research

Soon after I had started my internship at Nokia IT it was revealed that there was a need for a generic capacity model for IT systems and it would be possible to do as a thesis. First, the idea was to make it as one thesis, but after further examination it was found out that the subject was too big for just one thesis. It was decided to be divided to two: the preliminary study and then creating the generic capacity model. My responsibility was to do the preliminary study. The intention was to gather all the expertise about the subject inside Nokia IT in one place and also map the requirements for the generic capacity model. This was decided to be done by making qualitative research and using interviews for data collection.

The colleague I am refering to in this chapter, participated in the interviews because he is also making a thesis about this subject. His point of view is to make a generic capacity model based on the results that were gathered with the interviews. He took part in the phases where interview questions and the invite for the interviews were made and in the interviews to get all the information on this subject. It was my responsibility to organize the interviews, ask the questions and then process the data and summarize the results.

## 4.2 Qualitative Research

Qualitative research was chosen for data collection and data analysis technique for this research. The reason why qualitative research was chosen is because it concentrates on generating or using non-numeric data which usually means words. In this case this seemed fit for purpose more than quantitative research which instead concentrates on generating or using numeric data which usually means numbers. (Saunders et al. 2007, 145.)

It was decided that interviews for data collection would be used when making qualitative research. There are various types of interviews. Interviews can be formalized and structured where standardized questions are used for each interviewee or they can be informal, more like unstructured conversations or something between those two. Interviews can be divided in structured interviews, semi-structured interviews or un-structured/in-depth interviews. (Saunders et al. 2007, 311-312.)

A structured interview is a formal type of interview where questionnaires are used. In these questionnaires there are series of identical questions which are predetermined. In these types of interviews it is important that every question is read out as it is written and using the same tone of voice not showing any signs of bias. Then the response is marked on a standardised form which includes usually pre-coded answers. Typical of these types of interviews is that there is not much social interaction between the interviewer and the interviewee. (Saunders et al. 2007, 312.)

A semi-structured interview is a more informal type of interview. Essential in these types of interviews is that there is a list of themes and questions which should be covered. These themes and questions can vary from interview to interview. This means that some questions can be left out in some interviews and in some cases additional questions may be needed. Also the order of the questions can vary if the flow of the conversation requires that. In these types of interviews responses are usually recorded with an audio-recording device or by taking notes. (Saunders et al. 2007, 312.)

An unstructured interview, also known as in-depth interview, is the most informal type of interview. This type of interview is usually used when deeply exploring some area which is in focus. Essential in this type of interview is that there is no list of predetermined questions but there should be a clear idea about the aspects that are wanted to be explored. The interviewee is free to talk about events, behaviour and beliefs which are dealing with the topic. In this type of interview the interviewee is usually the one who directs the interview. (Saunders et al. 2007, 312.)

### 4.2.1 Data Collection

My research was following mainly the structure of a semi-structured interview but there were also aspects of an unstructured interview. There was a set of questions (altogether six questions) that were made in co-operation with my colleague and my line manager. Both have been active in capacity management process.

With these questions, we attempted to explore what was the current state how capacity was taken into consideration when designing a new service or application without a generic capacity model, what elements the capacity model should include and how detailed it should be. We also tried to find out what kind of guidance would be adequate for creating a capacity model and how people who were interviewed could participate in creating the guidance and the model. There was also a question about what capacities are measured at the moment and what capacities should be measured.

When choosing the people for the interviews, we tried to select people who are actively involved with capacity related issues from all the different units. The intention was to get a comprehensive picture of the subject. It was wanted that the interviews were made in Finnish so all interviewees were chosen based also on that criteria. At first there were 10 people who were selected for the interviews. During the interviews three more people were recommended by other interviewees to be interviewed. We decided to ask also these three people to be interviewed, but one of them declined. All in all, there were 12 people who were interviewed.

These 12 people included two capacity managers, one process owner and two process managers in capacity management process, two solution architects and one enterprise architect, a senior manager in performance engineering, two production managers and an IT manager in hosting networks.

An individual invitation to the interview was sent to all interviewees. It included an explanation what we were doing and why their participation was important. They were also told that these interviews and the capacity model based on the results were part of two theses. It was also mentioned that we wanted to record these interviews with an audio-recording device and for people who were located on the same site as we, we hoped that the interviews could be made face to face. The interview questions were also attached to the end of the invitation so people could get to know them beforehand if they wanted.

The invitations were sent via Outlook where it was easy to find a suitable date for all participants (me, my colleague and the interviewee) by using the calendar function. Outlook invitations include three response choices of which the recipient can accept, decline or make one tentative for the event and the answer is sent to the sender of the invitation. With these notifications it was easy to keep track of who had accepted the invitation and based on those make the room reservations for the interviews.

In most cases people accepted the invitation as it was sent. In cases where interviewee responded as tentative or did not answer at all we usually made the room reservation anyway and were prepared for the interview. In all these cases people participated in the interview.

The first interview was made in mid-August and the last one was made in mid-September so it took approximately one month to make the interviews. There were one to four interviews per week depending on the schedules. It was a common setup that my colleague and I were both participating in the interviews except one case which I made by myself. We reserved one hour for every interview which was enough every time. The interviews took from half an hour to almost an hour depending on how much information interviewees had to share.

The interviews were conducted either face to face or in an online meeting using also voice conference. Before the interview started it was confirmed from every interviewee that it was okay that the interview was recorded with an audio-recording device which was in this case my smart phone. For everyone this was okay.

Like I earlier mentioned these interviews were following more the structure of an informal semi-structured or unstructured interview than a formal structured interview. Interviewees were encouraged to share all the information about the subject they had to form a comprehensive picture about the subject. Even though the interviews usually were following the predetermined questions more or less, the ambiance was more like an informal conversation where it was mutually encouraged to ask questions.

In some cases the order of the questions also varied or some questions were left unanswered which are the signs of a semi-structured interview. In some cases it also happened that the interviewee directed the interview because the questions were shared on the screen and this underpins that there were aspects of an unstructured interview.

Because the interviews were recorded with an audio-recording device they needed to be typed to the computer. I usually made that as soon as possible after the interview because making the transcript took time and it was easier to do when there were not many interviews in queue. There was also the aspect that when it was typed as soon as possible there was a lower risk of a data loss. To prevent the data loss I backed up the interview from my phone to my computer after the interview was held.

### 4.2.2 Data Analysis

When every interview was held and transcribed to the computer, it was time to process and analyze the data. I started the data processing by gathering the results of every question into separate documents, one document for one question containing all answers. There was no need to separate what each interviewee has said so it was easier to handle the data first question by question. Certain issues were repeated in the responses so it was easy to pick-up some themes from there and form categories of those.

When mass of qualitative data is handled it is reasonable to organize it into meaningful categories. This is one feature that is used commonly in qualitative data analysis. It helps to analyze and explore the data systematically and accurately. (Saunders et al. 2007, 479.) The categories were based either directly on the issues that were asked in the questions or on the issues that were raised during the interviews. For example there was one question about what capacities are measured or should be measured and based on that there was a category called Measurements. Under that category I then gathered all the issues that were related to measurements.

When there were categories for all the relevant issues and all the relevant information was gathered under those categories, it was time to translate the results from Finnish to English. For some parts the translation was challenging because most of the interviews were like informal conversations so the language was also informal.

When translation process was completed, it was time to gather the information that was used in the final summary. There was no need for further data analysis at this point because the objective of my assignment was to gather the relevant information and present it in a summary. The key target of the summary was also to maintain an objective point of view through the summary.

## 4.3 Conclusions of the Results of the Interviews

After the summary was ready I presented it to my line manager and my colleague. My line manager was also the client of the assignment. The feedback from the client was good. There is now a good basis for creating the generic capacity model based on those requirements and the expertise that was gathered with these interviews. The summary was also mailed to every interviewee.

The results of the interviews can be found in chapter 5, but unfortunately this is confidential data and will not be published in this thesis. However, without revealing anything specific one could say that these interviews were successful. For all the questions various opinions were found and mutually useful information was shared. Also new ideas were raised and things that should be improved were found out.

# 5 THE RESULTS OF THE INTERVIEWS (SECRET)

# 6 CONCLUSIONS

Making this thesis has been a long and educational process. When I first started researching this subject, it was not familiar to me at all and one could say that making this thesis was a journey to the unknown. I have learned a lot from ITIL, capacity modeling of IT systems and making interviews and processing data gathered from them.

At first, the theory part of this thesis should have been only about capacity models but after further examination it was found out that capacity models and capacity modeling of IT systems go hand in hand so it would be necessary to examine both.

In my opinion, this thesis answers quite well to the research questions that were set in the introduction. I had four main research questions and I managed to find answers to all of them. The first one was to find out what capacity modeling of IT systems means. Briefly, capacity modeling of IT systems means performance modeling and throughput forecasting and it is used to predict how IT systems behave.

The second goal was to find out what kind of different capacity modeling techniques exist. There are various techniques that are used but the most popular ones are trending, simulation modeling and analytical modeling.

The third goal was to find out what capacity model is. I managed to find out that ’capacity model’ works as a general term for the models where earlier mentioned modeling techniques are used. The main purpose of capacity models is to make performance prediction and create what-if analyses. Capacity models are used and should be used when making decisions about capacity issues. Capacity models can be simple spreadsheets or commercial software which both should include the current-set up that reflects achieved performance and proposed set-up which will reflect the future state of the system.

The fourth goal was to find out what needs to be taken into account when creating a capacity model. Before creating a capacity model, one should have full understanding of the business issues that are involved and the applications that will be modeled. In practice this means using realistic and accurate data which is derived from business units. The most important thing of the model is its purpose hence it is important to know what questions should be answered with the model before creating it and how detailed the model will be. A model should be kept as simple as possible to achieve what is wanted.

The goal of the empiric part of the thesis was to gather all the expertise inside Nokia IT about capacity models of IT systems and collect requirements for a generic capacity model of IT systems. These were gathered by making interviews. The results of these interviews also supported the third and the fourth goal. In my opinion these interviews were successful because various opinions about the subject were found and mutually useful information was shared. Also the feedback from the client of the assignment was good.

The best part of this thesis was making the interviews. It was interesting to interview different stakeholders and learn from them about this subject and how capacity management is made in a big company. It was also nice to know that the results of the interviews will be actually used. It motivated me through the whole process.

The hardest part of this thesis was finding the sources for the theory part. There is not much written about this subject and if something was found it was usually dealing with this subject from the mathematical point of view which was not relevant for my thesis. Luckily I found some relevant sources to form the theory part of this thesis.

I also had some schedule issues. I started to make this thesis in the beginning of the summer but the interviews could be made just at the end of the summer because of the summer vacations of the interviewees. I also needed to wait for one book that was vital for my theory part almost for two months which slowed down the process.

Like earlier mentioned there will be a follow-up action for this thesis. There will be another thesis where the results of the interviews will be used as a basis for a generic capacity model of IT systems.

# SOURCE MATERIAL

An encyclopædia Britannica Company Merriam-Webster 2011. Capacity. Referenced on 6.9.2011 http://www.merriam-webster.com/dictionary/capacity

Betz, C. 2007. Architecture and Patterns for IT Service Management, Resource Planning, and Governance. Making shoes for the Cobbler's Children. San Francisco: Elsevier.

Van Bon, J.; de Jong, A.; Kolthof, A.; Pieper, M.; Tjassing, R.; Van der Veen, A. & Verheijen, T. 2007. IT Service Management based on ITIL V3. A Pocket Guide. First edition. Zaltbommel: Van Haren Publishing.

Cartlidge, A.; Hanna, A.; Rudd, C.; Macfarlane, I.; Windebank, J. & Rance, S. 2007. An Introductory Overview of ITIL V3. Wokingham: IT Service Management Forum Limited.

Encyclopedia.com 2011. Queuing network. Referenced on 15.9.2011 http://www.encyclopedia.com/doc/1O11-queuingnetwork.html

Grummitt, A. 2009. Capacity Management. A Practitioner guide. First edition. Zaltbommel: Van Haren Publishing.

Gunther, N. 2007. Guerrilla Capacity Planning. A tactical approach to planning for highly scalable applications and services. Berlin: Springer.

Investopedia 2011. Queuing Theory. Referenced on 15.9.2011 http://www.investopedia.com/terms/q/queuing-theory.asp#axzz1Y0JoJBLt

Lucid IT 2007. ITIL Version 3. Referenced on 13.11.2011 http://www.lucidit.com.au/itil_version3news.php

Menascé, D.; Almeida, V. & Dowdy, L. 2004. Performance by Design. Computer Capacity Planning by Example. First edition. Upper Saddle River: Prentice Hall.

Nokia 2011a. The Nokia story. Referenced on 10.12.2011 http://www.nokia.com/global/about-nokia/company/about-us/story/the-nokia-story/

Nokia 2011b. Our people & culture. Referenced on 10.12.2011 http://www.nokia.com/global/about-nokia/company/about-us/culture/our-people-and-culture/

Nokia 2011c. Our structure. Referenced on 10.12.2011 http://www.nokia.com/global/about-nokia/company/about-us/structure/our-structure/

The Office of Government Commerce. 2007a The official Introduction to the ITIL Service Lifecycle. Norwich: TSO.

The Office of Government Commerce. 2007b Service Design. Norwich: TSO.

Potter, R. 2008. Multi-System Modeling. Referenced on 13.11.2011 http://www.teamquest.com/pdfs/whitepaper/msm.pdf

Saunders, M.; Lewis, P. & Thornhill, A. 2007. Research Methods for Business Students. Fourth Edition. Harlow: Pearson Education Limited.

Savvides, A. 2004. Capacity Management. The Computer Bulletin Vol. 46, Issue 4, 28.

TeamQuest 2011. TeamQuest Model. Referenced on 13.11.2011 http://www.teamquest.com/pdfs/products/model.pdf