OPTIMIZATION OF TELECOMMUNICATIONS INCIDENT MANAGEMENT PROCESS

Lappeenranta–Lahti University of Technology LUT

Degree programme in Industrial Engineering and Management, Master's Thesis

2023

Yusuf Abdulahi

Examiner(s): Professor Janne Huiskonen

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Industrial Engineering and Management

Yusuf Abdulahi

Optimization of telecommunications incident management process

Master's thesis

2023

87 pages, 14 figures, 18 tables and 0 appendices

Examiner(s): Professor Janne Huiskonen

Keywords: incident management, queuing theory, simulation

Telecommunications companies have a critical role in providing network infrastructure for their customers. Customers can have high expectations regarding the availability of the service, which means that network outages should be avoided to reduce their negative impact. A major outage is a significant risk for customer operations leading to decreased customer satisfaction and in some cases can require service providers to pay penalties. The incident management process is the manner incidents are solved when detected. Focusing on eliminating inefficiencies from the incident management process can have a substantial effect in reducing the service recovery time and thus limit the impact the incident causes.

This master's thesis investigated ways to optimize the incident management process in a telecommunications company by utilizing various methods. First, a comprehensive literature review was conducted to get an overview of incident management, queuing theory, and simulation. Within incident management, ITIL was presented as a methodology for incident management. Also, Service Level Agreements (SLA) were discussed to obtain information regarding commitments between the service provider and the customer so that the importance of resolving incidents within the expected timeframe could be underlined.

After literature review, queuing theory and simulation were used to first create a conceptual model of the incident management process and then to translate it to a computer-readable form for simulation purposes. The model was created to represent the real-world process so that different ways to prioritize incidents in the queue could be tested without disturbing real-world operations. After multiple rounds and scenarios in the simulation, the results were collected and analyzed. From the results, it became clear that the incident management process can be optimized by sorting the queue not with first-in-first-out discipline but with a priority-based queue that uses incident ticket priority and remaining SLA time to place the ticket into the queue.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT School of Engineering Science

Tuotantotalous

Yusuf Abdulahi

Häiriönhallinnan prosessin optimointi telekommunikaatioalalla

Tuotantotalouden diplomityö

2023

87 sivua, 14 kuvaa, 18 taulukkoa ja 0 liitettä

Tarkastaja(t): Professori Janne Huiskonen

Avainsanat: häiriönhallinta, jonoteoria, simulointi

Teleoperaattoreilla on kriittinen rooli verkkoinfrastruktuurin tarjoamisessa asiakkaille. Asiakkailla on usein suuret vaatimukset palvelun käyttöasteesta tarkoittaen, että verkkokatkokset voidaan nähdä merkittävänä riskinä asiakasyritysten liiketoiminnassa. Laajat häiriöt palveluntarjoajan verkossa voivat johtaa heikentyneeseen asiakastyytyväisyyteen; joissain tapauksissa palvelusopimuksen sisällöstä riippuen palveluntarjoaja voi joutua maksamaan korvauksia. Häiriönhallinnan tarkoituksena teleoperaattoreilla on vastaanottaa ja ratkaista näitä vikatapauksia. Hukan poistamisella kyseisestä prosessista pystytään pienentämään näiden häiriöiden korjaamisaikaa rajoittaen katkoksesta johtuvia seuraamuksia.

Tämä diplomityö tutki tapoja optimoida häiriönhallintaprosessi telekommunikaatioalan kohdeyrityksessä hyödyntäen useita menetelmiä. Kirjallisuuskatsauksen avulla saatiin kokonaisvaltainen käsitys häiriönhallinnasta, jonoteoriasta sekä simuloinnista. Häiriönhallintaa sekä muita siihen liittyviä konsepteja tutkittiin ITIL-viitekehyksen sisällä. Tässä yhteydessä tuotiin myös esille palvelutasosopimukset, jotka korostavat tärkeyttä asiakasodotusten lunastamiselle ja vikojen nopealle korjaamiselle.

Empiirisessä osuudessa hyödynnettiin jonoteoriaa sekä simulaatiota, joiden avulla pystyttiin mallintamaan kohdeyrityksen häiriönhallintaprosessi simulaatioympäristöön. Malli rakennettiin imitoimaan oikeaa prosessia hyödyntäen prosessista saatua dataa, jotta testauksia ei tarvitse heti tehdä käynnissä olevaan prosessiin. Useiden simulaatiokierrosten ja eri skenaarioiden vertailemisen jälkeen tulokset otettiin talteen ja analysoitiin. Tulosten perusteella häiriönhallintaprosessi saadaan tehokkaammaksi hyödyntäen prioriteettijonoa tikettien jakamisessa. Kyseinen prioriteettijono hyödyntää tiketissä olevaa tietoa sen prioriteetista sekä jäljellä olevasta ajasta palvelutason alittamiselle.

ABBREVIATIONS

DES	Discrete Event Simulation
IM	Incident Management
ITIL	Information Technology Infrastructure Library
ITSM	Information Technology Service Management
KEDB	Known Error Database
KPI	Key performance indicator
MI	Major incident management
SLA	Service Level Agreement

Table of contents

1 Introduction

The purpose of this introductory chapter is to provide background and present the structure of this thesis. The chapter consists of multiple subchapters. First, background of the research is introduced where the importance of the topic and the problem is discussed. Then, research objectives are presented with the scope of the research. The research methods are then discussed in order to understand how the set objectives will be met. This introductory chapter also works as a guide by presenting the overall thesis structure.

1.1 Background

In today's digital world, networks play a major role in enabling communication and managing information in addition to providing a base for data-oriented technologies such as Internet of Things, Big Data and Machine Learning. When evaluating the most important aspects in network solutions for companies, the main criteria is network availability meaning that the network should be always online based on a study conducted by Gartner (2014a). In the Gartner study, it was also found that the importance of availability is heavily linked to the impact of lack of availability, which can be noticed as network downtime. Based on surveys done to different types of organizations, an average cost resulted from network downtime is 5 600 dollars per minute extrapolating to over 300 000 dollars per hour (Gartner 2014b). According to another study made from employee productivity perspective, IT downtime cost resulted in seven hundred billion dollars of lost productivity for companies from North America alone (Saarelainen 2016). Thus, minimizing the impact of these outages is essential to both enhance the service quality and save unwanted costs.

As telecommunications network solutions become more complex and larger than before, the criticality of these networks also increases, which is why companies seek high availability from the network provider. To comply with set Service Level Agreement (SLA) made between the service provider and a service owner, effective strategies need to be in place to make sure that the network outages are fixed within the SLA timeframe. (Salah et al. 2019) This type of incident resolution is one part of Information Technology Service Management

(ITSM) activities. ITSM has an overall objective of organizations linking their business objectives and strategy to their IT operations (Pereira et al. 2021).

ITSM uses an industry standard framework named Information Technology Infrastructure Library (ITIL), which defines best practices in information technology management and provides information on helping organizations with identification, planning and deployment of IT services (Dabade 2012). As it can be seen from figure 1, ITIL includes several different aspects that are categorized under service delivery and service support areas. This thesis will especially focus on the incident management (IM) inside service support operational level. Service desk is also tightly linked to the incident management efforts and will be also examined together with IM. With a successful use of ITIL practises, telecommunications companies can minimize negative impact on service quality such as to the availability of networks, which in turn can result in minimizing the damage that network outages have a chance of doing.

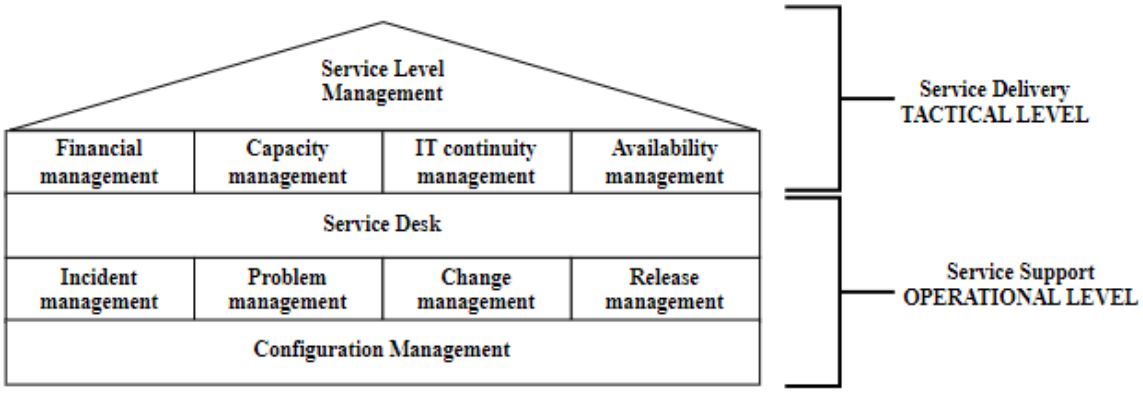

Figure 1: ITIL service management framework (Caster-Steel & Tan 2005)

This thesis is commissioned by Telia Finland, which is a telecommunications company and part of a bigger organization Telia Company. The company has footprint in Baltics and Nordics providing essential digital and network infrastructure to over 25 million customers. In addition to being a leader in telecommunications within this region, Telia also specializes in Media business and provides various ICT services making the company a substantial technology expert with a purpose of reinventing and making society better connected. (Telia 2023)

Telia Finland provides services to different kinds of customer segments. They have solutions for both consumer business and corporate business. Furthermore, they have also a Wholesale business segment, in which this thesis is commissioned. Telia Wholesale is focused on providing product solutions and various services to both national and global telecommunications companies and other service providers. These solutions can vary from a simple standardized product to a more specialized network solution designed for customer needs. Since the Wholesale segment serves other operators, these operators have customers on their own making the impact of these solutions high since there are both "first line customers" that order the solution and end customers that commonly are the users of said solution.

Before the need for this specific thesis in Wholesale business segment was discovered, the department had been foregoing a comprehensive digital transformation together with the rest of the company, which meant that the current processes and way of working had to be reviewed and based on that, developed. This strong state of will developing operating models and information systems has helped the department making daily operations more efficient and effective with the help of principles such as Agile and LEAN. Wholesale Technical Services team is a part of Wholesale department, which is responsible of providing customer support for faults and disturbances in addition to professional services designed to assist customers with Wholesale products. This team seeks to handle tickets according to ITIL incident management process for example to fix network outages within agreed service resolution time.

The current as is process for incident resolution has been examined and a room for improvement has been found. Some of the current information systems used in ticket handling have ineffective processes that rely too much on the experts doing manual work to manage the ticket queue. Moreover, the information systems, mainly ITSM systems, are ongoing development so focus must be to provide value, which requires an effective use of these systems with a logic attached that the incident management team can use to minimize manual work. Telia Wholesale has a strong need for utilizing insights from new academic research and advancing digital transformation by utilizing new technologies to improve current processes. From academic research, especially queueing theory has been deemed promising to facilitate ticket queues and find ways to optimize the incident management process that best serve both the customers and the experts handling the tickets.

1.2 Research objectives

The purpose of this thesis is to provide an optimized ticket handling process for an incident management team. The ticketing systems function by sorting and queuing tickets that are either incoming or previously created. Based on the logic in the ticket handling system, the work queue or multiple queues are created, and the tickets are then assigned to an agent. The goal of this thesis is to thus present a queuing model that can handle tickets effectively and efficiently while reducing the amount of manual work the agents are currently required to do. To do this, the current as-is process must be analysed and as a result, provide a to-be process that has especially focused on the performance of the ticket queue. Additionally, to better reflect the department strategy and its Key Performance Indicators (KPIs) the process should contain a way to keep the customer informed throughout the incident management process to bring more transparency to the ticket handling.

The result will then be an incident management process that has a viable and effective logic attached in sorting incoming or created tickets that can be utilized in ITSM systems. In other words, the result should be ready to use for various ITSM systems and should not only be limited to an IT-system that will be implemented as a ticket handling system to Wholesale Technical Services. Therefore, this thesis will not be IT-system-specific but focuses on giving an overall picture on practices that could help with managing incident ticket work queues while decreasing manual effort and increasing customer-centricity with better customer communication.

The first limitation of this thesis is that it only focuses on telecommunication industry and its incident management process even though ticket handling in incident management has similarities to for example common IT service desk operations in any industry. Since there could be multiple different ways to approach incident management, this thesis only focuses on incident management inside ITIL context. This approach is used in target organisation Telia Finland, and the goals of this thesis should be in accordance with these ITIL principles.

To optimize this incident management process area, multiple methods were selected to help analyse the current process. One of these methods is queueing theory. It can be defined as several mathematical techniques to manage a flow of some object passing through a network. Thus, it can be used in modelling real queue systems and as a result, getting predictions on how the system works under specific conditions. Typical queueing theory approach is

associating cost with delays and with higher service rate. (Newell 2013, 2-5) The performance of the model implemented with queue theory approach can then be experienced with simulation.

To utilize both queueing theory and simulation methods, actual data from the queueing system is needed. One of the methods will then be to use a large amount of ticket data from Telia's ticketing system to provide as accurate environment as possible to simulate approaches to sorting and prioritizing the incident management queues. The end result should be to get insights on what is the optimal solution for ticket queue optimization with the help of real-world data, queueing theory and simulation.

With the help of all these objectives, following research questions were identified:

1. How can the incident management process of a telecommunications company be optimized to handle tickets within agreed SLA of a specific customer?

2. Can queuing theory and simulation be used to identify the most optimal set of prioritization rules for tickets in a telecommunications company's ITSM system?

3. How can real-world data from the ticketing system improve the incident management process in telecommunications?

The first research question is related to the problem of delays in ticket resolution, where a delay not only makes it difficult for achieving set SLAs but also makes network outages and other problems more visible to customer impacting their services. To minimize delays in ticket handling, the current process should be improved to help make the ticket queue effective and avoid unnecessary manual work for the service desk agents, who then can focus on the more important ticket resolution operations.

The second research question focuses on methods for achieving optimized incident management process. The research question answers to the question of how can queue theory as a theoretical approach help optimize the process through achieving effective prioritization rules for the incident management queue. To test the resulted models, simulation approach plays a key role, and the question tries to determine if the simulation helps with finding a clear optimal way to prioritize tickets.

Third research question assists other research questions with focusing on the real-world ticket data and the manner of how it can be exactly utilized. The ticket data includes a lot of

information so finding the crucial information from the data and using it together with queue theory and simulation enables that the simulation is largely based on the real-world service desk operations. This helps with implementing the logic resulted from the thesis when a new ITSM-system is deployed to ticket handling. In other words, using realistic ticket data helps with making the insights on this thesis be applicable and ready-to-use for ticket incident processes based on ITIL.

1.3 Structure of the thesis

The report consists of an introductory chapter, theoretical chapters, and a comprehensive empirical section. The first chapter is introduction where brief topic background and structure are presented. The introduction output includes background, objectives, research questions and thesis outline.

The theoretical chapters are conducted as a literature review. First, overview of telecommunications incident management processes is presented. This chapter gives information of current approaches in incident management and describes its process inside ITIL-framework. To support this, ITIL is also presented to understand the big picture around incident management. Major incident management is then presented to distinguish it from normal incident management. To close the chapter, service level agreements in incident management are discussed.

The next part of the literature review is to present queuing theory. This chapter also contains information on the various applications for queue theory and how it is related to service management. Then, queueing theory approaches in incident management and ticket handling are presented tying the theory to the topic of this thesis. Last, queuing modelling and simulation are discussed to optimize performance of queuing systems.

After the literature review, the overall methodology for solving the research questions is then described. First, data collection and analysis methods are discussed and then the simulation approach. The next chapter presents the simulation results and a model for efficient ticket handling is constructed based on the results. The objective of this chapter is to form an in-depth logic for optimizing the telecommunications incident management processes. The final chapter concludes the thesis by discussing the results. This chapter contains discussion

tying all the topics in the thesis, answers the research questions and provides limitations and future research opportunities.

2 Incident management

This chapter is the first part of the literature review. It gives an overview of incident management process and presents it under ITIL context. First overview of incident management is examined, and the next chapters provide additional important context by first discussing major incident management and service level agreements within incident management.

2.1 Overview of Incident Management

Incident management can be formally understood as a part of Information Technology Infrastructure Library (ITIL) among other similar standards. It is portrayed as a set of activities that are required to help restore operations caused by the incident as swiftly as possible. Based on the ITIL definition, an incident refers to any event that decreases or disrupts the quality of the service and is not part of the regular service. (Cusic & Ma 2010) The overall goal of incident management is to reduce the negative effects of incidents and find ways to resolute them within agreed timescales (Agutter 2019, 65).

The main activities involved in incident management according to Agutter (2019) are planning the practice, set priorities for incidents and use an incident management solution. Types of incidents vary meaning that some of the incidents have higher impact than others. That is why planning the incident management practice with impact at the forefront is required to create varied responses based on the impact that incident in question brings. Setting incident priorities help with this by setting priorities in a way that ensures that the most serious problems are fixed first. Customers shall also be included in setting and agreeing on service levels, which affects the prioritization. Using an incident management solution enables the tracking and management of incidents and could contain helpful wiki type information to help solve the incidents. (Agutter 2019, 65-66)

Kaiser (2020) expands on the understanding of incident management by pointing out the reactive nature of incident management; it reacts to situations as soon as they arise and are not proactive. It is also stated that while the process might be simplistic, the situations vary, and the responses are unique and made on a case-by-case basis. Taking these into account,

incident management plays a critical role in achieving good customer satisfaction. The value that the customer gets from the provided services stops as soon as there is an incident regarding the service. Incident resolution time is now essential in limiting the damage to the customer and fixing the service inside the customer expectations limits damage to customer satisfaction. (Kaiser 2020, 284-286)

To dive deeper into the life cycle of incident management, taking a closer look into the incident management process is crucial. The following activities are included in the process (Cusick & Ma 2010):

- Incident identification and detection
- Incident classification
- Investigation and diagnosis
- Resolution and recovery
- Incident closure

Since the incident management is often considered reactive, a trigger is the starting point of the process. The trigger in the process depicted in figure 2 is call received meaning that the first level support has got a call notifying that an incident is identified. This is not only way the process can be triggered. Kaiser (2020, 291-292) argues that the most used triggers in addition to telephone are monitoring and event management, email and web interface. Monitoring and event management can work as a trigger in an event where an event management system has found an exception in the service or service quality, which brings up a notification that flows through appropriate channels. In telecommunications, a service provider could have a network down in a specific location and the tool identifies this as soon as it happens.

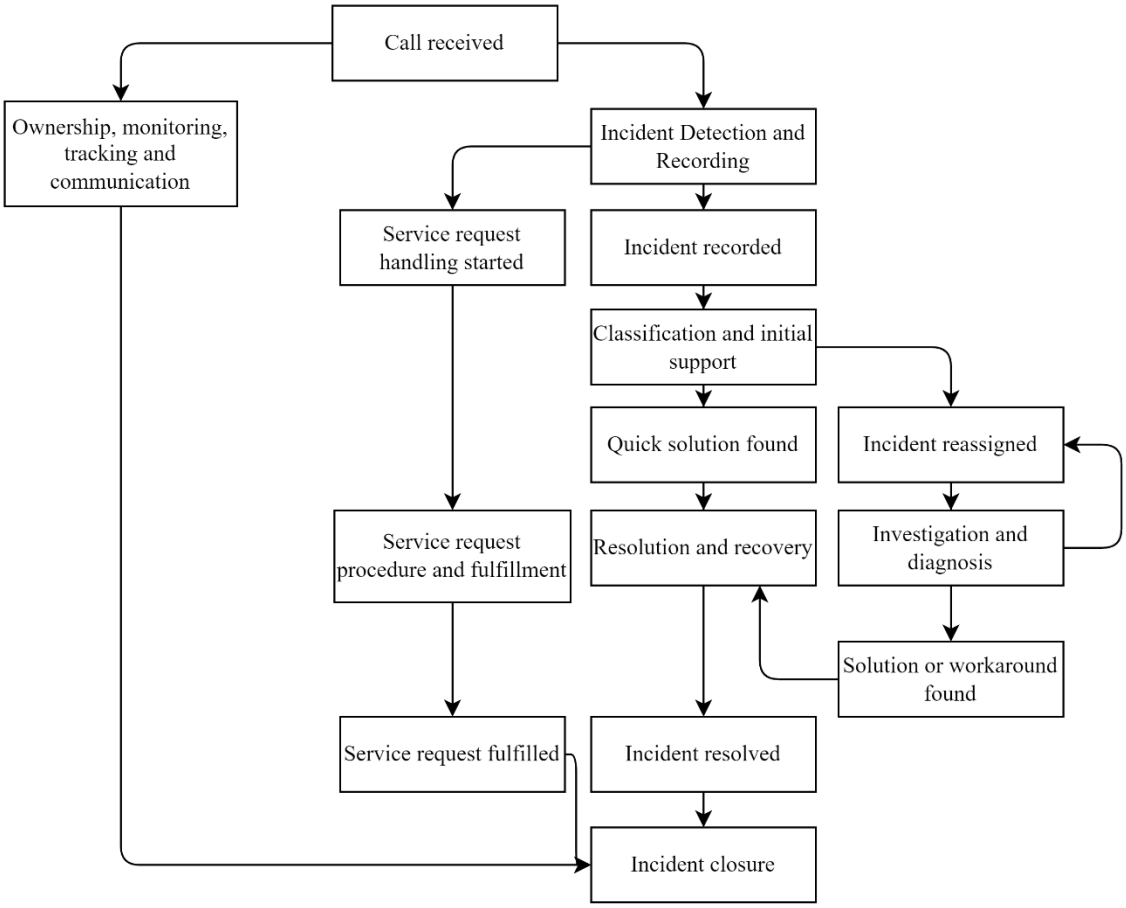

Figure 2: ITIL Incident Management Process (Brenner et al. 2002)

Email or web interface are both forms of communication that do not require calling in the incident but still work similarly if required information regarding the incident is provided. Web interfaces are becoming more popular in ITSM, which means that the customers can raise their own tickets through a web interface without interacting with a service desk. Prerequisites are that the company has a ITSM ticketing tool available to manage tickets and the customer has an end user portal that allows for creating tickets. (Kaiser 2020, 292)

Incident classification starts after the process trigger. First, incident is logged by a service desk agent with information such as the contact details of the user and description of the issue. (Swain & Garza 2022) There exists various ITSM ticketing tools that are designed for logging incident tickets and provide other benefits related to managing the tickets (Kaiser 2020, 293). ServiceNow is an example of this type of tool. Clients can use ServiceNow software to automate popular IT support activities such as incident tracking, password

recovery, troubleshooting and IT system management via simply designed and easy-to-use service portal (Chaykowski & Coatney 2018).

After the incident is logged, starts the incident categorization and prioritisation. The incidents are compartmentalized into appropriate categories and a priority level is assigned to determine when and how the issue is addressed later in the process (Swain & Garza 2022). Categorization is important since the incidents vary by a great degree and the resolvers may also vary based on the ticket category. The importance of incident prioritization can be detected especially in a case where there are many tickets at a given time with limited number of service desk agents.

Since some of the incidents are not as urgent and impactful as others, they can be prioritized lower so the high-priority incidents can be quickly resolved. This priority can be calculated by making it a variable of impact and urgency. Thus, priority number is generated when urgency and impact are known. Impact can be understood as a business factor that refers to losses in earnings, productivity, reputation and violations of regulations or laws. The degree of urgency is referring to how quickly the incident needs to be resolved. Table 1 shows how these two factors relate to calculating priority level of a ticket. (Ghosh 2013; Kaiser 2020, 295-296).

Table 1: Determining the priority of an incident (Ghosh 2013)

Impact	**Urgency**	**Priority**
High	High	**Critical**
High	Medium	**High**
High	Low	**Moderate**
Medium	High	**High**
Medium	Medium	**Moderate**
Medium	Low	**Low**
Low	High	**Moderate**
Low	Medium	**Low**
Low	Low	**Low**

During the investigation and diagnosis phase, the initial investigation is performed by the service desk agent who also has a responsibility in previous steps of the process. For example, if the trigger comes from telephone, the agent can interact with the incident notifier to investigate the nature of the incident and help the user resolve it with some troubleshooting

steps made by the user or the agent. Since the service desk agent cannot resolve all incidents straight away, they are in these cases referred to the next level of support agents. The incident is then deeply analysed to obtain a right course of action to fix it. (Ghosh 2013; Kaiser 2020, 298-299)

Resolution and recovery refer to the stage where technicians continue investigating the incident and trying different solutions based on the diagnosis made previously. After the resolution is found and applied, it is also thoroughly tested to observe that the resolution works. In incident closure, a common practise is to notify the customer that the incident is resolved and keep the ticket open for a few days to allow the customer to easily notify if the problem related to the incident remains. After the incident has been resolved and a few days have passed, the incident ticket is closed. This is the end of the incident management process for one specific ticket. (Swain & Garza 2022; Kaiser 2020, 300-301)

The described process follows ITIL version 3 principles. After the deployment of ITIL v4, the focus has shifted from processes to practises. The newest version ITIL does not invalidate the previous version but gives companies more freedom to design their processes based on their specific requirements. The goal of this approach is to make ITIL more flexible for organizations to use since the operations of organizations greatly vary from each other, which is why a one specific process for incident management is not as effective for everyone. ITIL 4 includes incident management as part of service management and provides guidance on inputs, outputs, key activities and key roles. (Kempter 2022) Kaiser (2020) describes incident management as a part of service operations, which in turn is a part of service management. Other parts of service operations are monitoring and event management and problem management, which are the value adding or value decreasing parts of the service management. (Kaiser 2020, 273-275)

To understand this further, relationship between service management and ITIL should be familiarized. This allows for a broad understanding of the incident management practise and its relation to service management and its other areas since they are heavily linked with each other. IT service management (ITSM) has risen from the need to align business objectives and IT operations, which resulted in “servitization” of IT operations (Conger 2008). ITIL is the most common ITSM framework (Iden & Eikebrokk 2013). According to Pollard et al. (2010), effectively managing IT services requires extra focus on the final stage of the service lifecycle: this stage is also called as continual service improvement. Other parts of ITIL

service lifecycle in addition to service operations and continual service improvement are service transition, service design and service strategy. These stages and their relation to ITIL incident management in service management can be seen from Figure 3.

Figure 3: IT service management contents (BMC Software 2020)

As evident from Figure 3, incident management is linked to multiple different ITIL processes. This can be detected especially by looking the role of other ITIL service operations processes in addition to incident management: event management and problem management. In contrast to incident management and its service restoration approach, problem management is focused with locating and resolving the underlying causes of the incidents. This strategy can raise the level of service quality and mitigate future incidents. Event management on the other hand involves managing and monitoring of events that could change to incidents, which helps with quickly detecting and resolving events via an event monitoring system before the impact to a service owner grows. ITIL defines event management as systematic observation of services and its components to identify and record any changes of event states; event is a change of state for a service or its components. (Kempter 2019; Kaiser 2020, 275-277)

ITIL states that a “problem” is underlying reason for incidents. Thus, primary goal for problem management is to either prevent incidents from happening or minimizing impact of incidents. Incidents and problems are sometimes used under the same context, which is evidently not correct in ITIL. After incidents have been raised and resolutions worked on, a

problem is recorded when the resolution is not possible and the root cause is not known; for the most part, the input for problem management comes from the incident management process. To help reduce probabilities of recurring incidents, incident management teams can utilize problem management practise by using a known error database (KEDB). KEDB contains root causes of known errors and workarounds. Workarounds are temporary fixes, so the KEDB record stops to exist when a permanent solution is found. (Sharifi et al. 2009; Kaiser 2020, 306-310) Therefore, incident management does not only work as an input for problem management but also uses information from problem management making these two processes exceedingly important to each other.

2.2 Major incident management

It is now clear that ITIL incident management is closely related to other ITIL practises, including for example event management and problem management. This interdependence highlights the importance of a collaborative approach to both incident management and IT service management. As depicted in table 1, the highest priority for incident in incident management process is critical, which has both impact and urgency ratings as “high”. While the critical priority incidents can cause large disruptions to the service owner, there is also another category for severe incidents called major incidents. These have their own process outside of standard incident management since major incidents often have specialized requirements for their resolution.

A popular tool for incident management, ServiceNow (2023a) defines in their documentation that a major incident (MI) is an incident that causes huge disruption for the business of service owner. Because of this, a major incident requires a response that overrides incident management process. Major incident management process enables a faster resolution process for incidents with high business impact since they are handled in a separate procedure with shorter timeframes and higher priority. (ServiceNow 2023a)

Often the accountable party for major incidents is a designated incident manager. The responsibility of the manager is to collect all relevant stakeholders together and make sure that the incident is resolved as quickly as possible in a high-pressure environment. A usable practice for major incident situation is to openly communicate the incident and its status to both service provider team and the service owner. This way the service provider incident

management staff and other relevant staff is notified that some key services are down and there may be multiple duplicate tickets coming to the incident management process. On the other hand, transparent communication allows for the service owner to stay updated and to avoid placing more tickets for the same issue to the service provider. (Kaiser 2020, 301-302)

Following a guideline by ServiceNow software documentation, a major incident management process usually contains following phases: identification, communication and collaboration, resolution and post incident review. Starting from the first step, identifying a major incident is possible either with escalating an existing incident or automatically based on set logic and rules. Next, a communication plan has to be constructed to keep relevant stakeholders aware of the incident. This plan includes for example methods for communication and people that must receive the updates. It is consequently crucial to make sure that there are notifications and status updates throughout a major incident lifecycle especially since the impact and urgency of the issue is high for these incidents. (ServiceNow 2023b) If the root cause cannot be solved, a problem ticket is produced and is linked to problem management practise (Kempter 2019).

In ticket resolution phase, the resolution will simultaneously resolve other related incident tickets and final notification is sent with a message that confirms the incident resolved and the service is now ready for normal usage. A service provider should conduct a post-incident review to make sure that the happened incident is properly understood to avoid similar situations in the future. The incident is analysed together with the process that was used to resolve the incident and corrective modifications are made if required. (ServiceNow 2023b) The process for major incident management is thus similar to basic incident management but with more focus on coordination between stakeholders, constant communication and stricter requirements for resolution times because of the high impact of major incidents.

To help detect major incidents, indicators for detecting them should be clear for the incident management team. The incidents are usually identified with its impact on the customers. For example, if business-critical services or infrastructure have been impacted and the estimated recovery time is either long or unknown. (Kempter 2019) According to Kempter (2019) Some key questions to identify characteristics that major incidents contain are:

- Are large number of service providers or key customers ability for service or system usage impacted?

- Does the service outage bring direct or indirect significant costs for the service provider?
- Is the brand and reputation of the service provider negatively affected as a result?
- Is it difficult to stay within agreed service levels? Does it require significant effort?

2.3 Service Level Agreements in Incident Management

While the incident management process by itself can look quite straightforward with its steps, there are various combinations in real world scenario that make the process harder to manage. One of the main challenges in managing incidents arises with large quantities of tickets even when well-staffed. One of the reasons for ticket handling challenges is difficulties in classification the incidents correctly leading to wrong priority for incidents (Jäntti & Cater-Steel 2017). In addition to this, prioritization of incidents is deemed quite ineffective by just utilizing incident priority matrix because of its simplicity and lack of flexibility to reflect real-world scenarios according to Kaiser (2020, 296-297). The concept of prioritizing tickets with impact and urgency with the help of priority matrix is expanded by fusing them with service level agreements.

A Service Level Agreement (SLA) is defined as a written contract that specifies needed services and the expected level of service between a service provider and a customer. A Service Level is interpreted as metric or multiple metrics that elucidate needed or achieved quality of service. (Kaiser 2020, 211-213) IT companies as service providers commonly set service levels of quality based on the cost of the service; a service with stricter SLAs will cost more than a service with SLAs with less or worse guarantees. According to Bianco et al. (2008) following aspects are required to be included for SLA to be rightly specified:

- How will the service be delivered with the promised level of quality?
- Which metrics shall be collected? Who will gather metrics and how?
- The actions that are required if the service is not provided at the expected level of quality, and who are responsible for completing them?

- What are the penalties when the service is not provided with promised level of quality?
- In an instance, when underlying technology changes, how will the SLA evolve to reflect these changes?

There are various ways for setting and agreeing on service levels. A service can include multiple different sides that all may need different service levels. A particular instance in the service can cause huge impact to the customer while others do not, which is why setting service levels to reflect business requirements and business context is vital for achieving great customer satisfaction. To give an example within telecommunications context, the service provider has provided a service that offers network connectivity to the service consumer. Two parties have agreed on that availability of the service must be at least 99.5 percent to stay within SLA. Even though the service provider fulfils expectations, it does not automatically ensure satisfied customer since a rare instance of an outage could happen when the service consumer is doing business critical activities. This is why it is important to break SLA into multiple sections where for example, SLA is different for working hours and outside of working hours. (Kaiser 2020, 213-216)

Other major aspect in defining SLA between service provider and service consumer is to align them to reflect existing business processes. This means that critical stakeholders need to be consulted and the expectations need to be aligned with the service provider organization. SLA must also be worded simply and straightforwardly to ensure that all parties understand the content similarly. (Kaiser 2020, 215) To achieve comprehensive view of managing SLAs effectively, service level management is needed. Service level management (SLM) is the practise of management and maintenance of quality of service. SLM is an important part of ITIL and works as an input to processes within service management context. (Bianco et al. 2008) Adherence to SLA mitigates violation of contractual obligations and legal requirements and business risks (Swain & Garza 2022).

Priority matrix in table 1 has depicted how priority is calculated through impact and urgency. This idea can be expanded by linking SLA to the priority matrix, which ultimately gives a more comprehensive picture of the performance and precision of the organizations SLA compliance efforts. For example, based on the criteria in the priority matrix priority levels can be written down as follows:

P1 = tickets that have priority as “Critical”,

P2 = tickets that have priority as “High”,

P3 = tickets that have priority as “Moderate”,

P4 = tickets that have priority as “Low”.

Now, the performance can be tracked for every priority level separately to get a better picture on which incident tickets comply the SLA the best and which ones especially need improvement. An example SLA compliance report is presented in table 2. The example tracks the performance through comparing the number in each quarter to target numbers where the metric is percent of achieved SLA. A green colour indicates if the target has been met and a red colour indicates that target has not been reached. The report also contains information regarding ticket volume, which helps with getting a broader picture on the incident management workload in a specific quarter. For example, a large quantity of tickets may make achieving SLA target a more difficult. (Ghosh 2013)

Table 2: SLA compliance report template (Ghosh 2013)

SLA Compliance report for <Client Name>

Incidents	Targets	Q1	Q2	Q3	Q4
Incidents					
PI SLA Success (%)	**80**	80	82	100	NA
P2 SLA Success (%)	**78**	84	100	100	100
P3 SLA Success (%)	**82**	78	98	100	98
P4 SLA Success (%)	**95**	96	100	57	100
P1 Ticket Volume		5	8	4	0
P2 Ticket Volume		19	7	17	5
P3 Ticket Volume		264	125	148	101
P4 Ticket Volume		62	9	7	3
Changes (CR)					
% SLA Success	**98.5**	100	100	100	100
Volumes		5	6	3	2

To help achieve SLA and define customer’s expectations, key performance indicators (KPIs) are used. KPIs are measurable values that are utilized for tracking service performance such as availability or response time. In other words, KPIs help measure the service provider’s performance against the SLA and track if the set objectives within the incident management process are met. One set of metrics in incident management context can be described as TTx-

metrics, which are set for different phases of the process. Ultimately, the improvement efforts made to incident management process should be detectable together with these metrics and mitigating incident impacts. (Chen et al. 2020)

There are three different TTx metrics: Time To Detect (TTD), Time To Engage (TTE) and Time To Mitigate (TTM). TTD describes the time it takes from automatic monitoring system to alert when incident first arises, TTE describes the time that a correct team is engaged and TTM describes the overall time it takes to mitigate the impact and resolve the incident. It is important to note that these types of metrics do not work in every incident management context. These examples of metrics make sense especially in cloud incident management where the practice has become faster and automated, where the incident mitigation is proactive, and the goal of the ticket resolution is to resolve the ticket even before the customer is aware of the incident. (Chen et al. 2020; Microsoft 2022) This is thus not as effective in cases where the inputs come from the customer and not from automated alerts.

The most common metric to evaluate the incident resolution performance is to use Mean Time To Resolution (MTTR). This metric simply ties all types of incidents and simply discloses a time it takes from the incident detection to its resolution. Since this is a simplistic metric that struggles when there are too many outliers or many categories of different incidents, some other approaches should be too familiarized. One way to make the metric more flexible is to use it separately for different categories of incidents. For example, different types of incidents or different incident priorities can have their own MTTRs. The KPI metric presented in table 2 is another approach for monitoring SLA performance: percentage resolved. This number describes the percentage of tickets that are resolved within a target SLA time. Some other ideas are to record total number of incidents and cumulative incident time to support other metrics and get a bigger picture into the overall performance of the process. (Churchman 2016; Bartolini et al. 2008)

Ortiz-Rangel et al. (2021) have successfully adopted an approach to fuse a process step to a KPI metric meaning that all identified incident management process steps have their own KPIs. The process is also done in accordance with ITIL and ISO 9001:2015. ISO 9001, an international standard, emphasizes a process-based approach and risk-based thinking. First, the process is established to include five sequential activities: 1) customer ticket assignment, 2) fault identification, 3) technical support, 4) confirmation of service restoration and 5) customer ticket closure. In the first step, Ortiz-Rangel et al. argue that it is critical to fill a

ticket template without errors to prevent that on later stages, the ticket is allocated to a wrong resource or corrective measures taken to resolve the ticket are incorrect. One way to mitigate this risk is to establish work instructions to help the employees to correctly fill out the ticket information. The KPI is thus “ticket without errors”. (Ortiz-Rangel et al. 2021)

Next in fault identification, the ticket is analyzed for 25 minutes by the first support team with the help of comprehensive work instructions. The goal of this stage is to quickly resolve the ticket without escalating it to a more technical staff to reduce the number of tickets they must handle. The third step in technical support means that in cases where technical support is involved, the KPI is the incident solved within a proposed time. In confirmation of service restoration, the KPI reflects the objective of this stage; it is designed to validate that the fix is effective and there are no recurring problems. The final stage of the process focuses on closing the ticket; the KPI is ticket closed within 24 hours. A result sheet with these KPIs is depicted in table 3. (Ortiz-Rangel et al. 2021) The example report is based on a telecommunications company.

Table 3: An example result sheet from January to July (Ortiz-Rangel et al. 2021)

KPI definition			**Implementation period**						
			Month (Data in %)						
Failure Support Process	**Platform**	**KPI**	**J**	**F**	**M**	**A**	**M**	**J**	**J**
Fault identification		0 - 25 minutes in 80% of cases	84.9	75.4	68.3	91.9	89.6	92.95	94.3
Restoration time	Optical fiber	According to the severity and time definition in Quality Plan in 80% of the cases	75.0	75.0	100.0	100.0	100.0	88.4	100.0
	Internet		81.3	83.4	100.0	100.0	90.3	100.0	87.0
	Private Line		72.8	63.0	62.5	88.9	88.4	78.4	100.0
	Long Distance		56.3	64.0	96.7	92.8	85.7	81.8	88.2
Ticket closure		Within 24 hours after service restoration in 80% of the cases	52.6	46.5	63.9	55.4	32.6	26.50	35.2

Kempter (2019) has further identified nine different KPIs designed for incident management in ITIL context. The KPIs are portrayed in table 4. Some of these are like previous KPIs presented on this chapter but some provide extra value for incident managers and other

decision makers interested in getting insights from the performance of incident management process. For example, measuring first time resolution rate is an exceedingly important metric to get a grasp of how first line support solves tickets during the first call between service provider and service owner. When organizations decide to utilize this, two things can be noticed with the metric. First, if the metric indicates a low efficiency of first line support, there may be a need for revision of existing instructions and documentation regarding common ticket solving methods. Second, if the metric points out a high efficiency of solved tickets, the incidents in that context are usually not as complex or broad as incidents that are escalated to more technical teams after initial investigation.

Table 4: ITIL KPIs for Incident Management (Kempter 2019)

Key Performance Indicator (KPI)	**Definition**
Number of repeated Incidents	Number of repeated Incidents, with known resolution methods
Incidents resolved Remotely	Number of Incidents resolved remotely by the Service Desk, which are not required external work at user location
Number of Escalations	Number of escalations for Incidents not resolved in the agreed resolution time
Number of Incidents	Number of incidents registered by the Service Desk
Average Initial Response Time	Average time taken between the time a user reports an Incident and the time that the Service Desk responds to that Incident
Incident Resolution Time	Average time for resolving an incident
First Time Resolution Rate	Percentage of Incidents resolved at the Service Desk during the first call
Resolution within SLA	Rate of incidents resolved during solution times agreed in SLA
Incident Resolution Effort	Average work effort for resolving Incidents

Overall, KPIs provide a way to measure parameters that are included in SLAs, so it is easily identifiable if the service maintains its promises. Swain & Garza (2022) have thoroughly researched factors that affect meeting SLA levels the most. While useful on a general level, the limitation of studying SLAs with an existing dataset from some company is that the results may vary for other organizations that may have different incident management processes or practices. The results from the study indicate that especially incident

prioritization and assignment are areas that impact the SLA the most in the process. First, confirmation of priority level is deemed important. Another example of priorities affecting SLA is that incidents with higher priorities are usually less likely to achieve SLAs compared to lower priority incidents. (Swain & Garza 2022)

In addition to choosing a right priority for the ticket, assigning the ticket correctly mitigates risks of not meeting the SLA. It is useful to monitor if incident tickets assigned to one specific group are consistently performing worse compared to other groups since this can demonstrate that the staff there is either overstaffed or lacks competencies to solve the tickets. Moreover, if the tickets are not meeting SLAs as successfully in some assigned ticket categories, it is then important to find root causes for why some categories of incidents are more difficult to resolve. One major way in achieving SLA regarding the assignment is the number of reassignments. The study points that probability of achieving SLA greatly increases when a ticket is reassigned. This means that reassigning the ticket enables ultimately finding the correct group for resolving the incident. (Swain & Garza 2022)

To get an overall picture of the Service Level Agreements and examples of set performance targets and consequences of failing to achieve them, an example can be presented under the context of services provided to European Union. European Court of Auditors (ECA) occasionally issues tenders for various IT services such as managing data centres or IT solutions (ECA 2019). This is an example of a customer a telecommunications company can have. Table 5 portrays an example of set KPIs between service owner and service provider.

Table 5: KPIs to achieve SLA levels for incident management in an example service (ECA 2019)

#	Scope	Measured Value	KPI	Measurement period	Penalty
SA	Availability	Full-service availability	Availability over 99,5%	Measured monthly	5 * Penalty Unit € (PU)
IM 1	Incident Management	Priority P1 incident	Resolution time must be under 2 hours	Measured monthly	3 * Positive Integer (p) * PU
IM 2	Incident Management	Priority P2 incident	Resolution time must be under 4 hours	Measured monthly	2 * Positive Integer (p) * PU
IM 3	Incident Management	Priority P3 incident	Resolution time must be under 24 hours	Measured monthly	5 * Positive Integer (p) * PU

ECA has provided requirements and specifications for service providers to follow that will serve as an example of real-world Service Level Agreement and its link to incident management and Key Performance Indicators. This is presented in table 5, which also includes the number of penalties each level of incident priority causes. The measured value there is service availability, which is calculated as service availability percentage by using the difference between number of minutes in specified period and number of minutes the service is not working as promised. The service provider should be able to comply with the KPIs to stay within SLA limits and avoid paying penalties to the service owner.

3 Queuing theory

This chapter starts with an overview of queuing theory, in which common practices and concepts are presented. After the overview, utilization of queuing theory in incident management is discussed. In this chapter, also queuing modelling is considered as a way to help optimize real-life queuing systems. Simulation as a related concept is also discussed to provide a specific way to test and predict performance of queuing systems.

3.1 Overview of queuing theory

Queueing theory is the mathematical study of queues or waiting lines. It is heavily researched topic with thousands of papers and books and became popular in the late 1950s. It was first adopted to provide a mathematical approach to telephone traffic lines. The queues represent a flow of objects, in which some of them have restrictions so they cannot pass through the system. An object which cannot pass the queue freely is then stored into an imaginary reservoir waiting for a turn to continue its flow. Length of time needed to continue from the restriction depends on the situation and the inevitable cost of delays depends on the object; some objects can cause higher stress than others when delayed. The goal of queuing theory is to help predict behaviour of real-life systems. Usually in queuing theory, delays are linked with costs and increasing service rates result in increased costs. Instead of reservoirs, restrictions and objects, these concepts also are sometimes depicted as queues, servers, and customers. (Newell 2013, 1-3)

In real life, queuing theory can be detected when a customer enters a waiting area, which is the queuing system in this case. The customers are deemed to wait in the waiting area if there are no available servers to serve them. Service is then started when a server becomes available, and a customer is selected from the queue. Logically, service then ends when customer and server stop their interaction. From this example, arrival times and customer serving times are some observations that can be calculated. Thus, to utilize queueing theory, information about the customer arrivals and the rest of the queuing process are required. Queueing theory has an objective to create formulas, expressions, or algorithms, which are used for performance metrics; examples of performance metrics for queues are average

number of customers and available resources. These can be used to solve various queue related problems such as determining an optimal number of servers needed or creating an effective system architecture. (Gautam 2012, 6-7, 17-18)

Little's Law is one of the fundamental concepts in queuing. It determines the relationship between a flow of queue and stock moving through a stationary system, and shares commonalities between many wating line models (Gonçalves 2022). Little's Law states that under steady state, the average rate customers arrive multiplied by the average amount of time a customer spends in the system is the average number of customers in a queuing system. The formula looks as follows:

$$L = \lambda W \tag{1}$$

where L = average number of customers in a queue, W = average waiting time for the customer and λ = the average rate customers arrive.

It is clearly noticeable how simplistic the law (1) is since it does not require any additional information from the queuing system such as the number of servers or the number of different queues. Since Little's law creates a relationship between three key measures in a queue, it is largely applicable to various use cases. An example of this robustness, Little's law works even while arrival and servicing times are both nonstationary if the observation window just starts and ends when a queue is empty. This is crucial because queues in real-world hardly follow a model where arrivals are constant. For example, lunch restaurants can be crowded at the lunch time but less crowded in other times and there are concrete opening and closing times making them exact circumstances for Little's Law. In addition to the restaurant example, usefulness of the law can be noticed in many other contexts. (Little & Graves 2008)

An excellent example of Little's law in action is to observe its use cases in Operations Management practise. The law has been modified to include relationship between Work in Process (WIP), cycle time and throughput in a following fashion:

$$TH = \frac{WIP}{CT} \tag{2}$$

where throughput (TH) = the average output per unit time, WIP = inventory between start and end of the points and cycle time (CT) = the average time spent in WIP state (Little & Graves 2008; Gallego 2003).

From the formula (2) it is thus easily observable that TH = λ, WIP = L and CT = W. The main difference is that operation management puts more focus on output instead of an arrival time making the average output (TH) equal to average input (λ). (Little & Graves 2008) This naturally requires an assumption that all objects that have entered the system, will also remain there and exit. To make Little's Law practical, it is important to remember that the start time and end time of observation must be zero. Especially in operations management, it is common that WIP is never zero since the staff modifies their service time as slower when WIP is lower; the staff thus keep themselves busy by doing work slower when it seems that there is not a lot of work to be done. Therefore, systems do not always follow the fact that they are empty in some timeframe and these situations require extra conditions for Little's Law to function. (Little & Graves 2008)

Kendall's notation portrays a model that has the following three elements: a, b, c, where "a" is arrival process, "b" is service process and "c" is number of servers. Though this notation is widely popular, an infinite number of operational protocols and other arrangements exist, and these are sometimes extended to the notation. Thus, Kendall's notation has been extended to include three more factors: capacity of the system (d), population size (e) and service discipline (f).

Using the notation, queuing system can be denoted by A/ B/ C/ D/ E/ F, where the numbers represent the factors presented above. It is also common in the literature to define the latter three parameters as K, n and D, where K is the capacity, n is the population size and D is service discipline. A/B/C, the simplest form of system assumes that the size of population is infinite, capacity of the system is infinite and service discipline is first in first out (FIFO). Understanding of Kendall's notation is integral to help describe and analyse queuing systems in a standardized way. (Sztrik 2012; Kardi 2014; Cooper 2010) Each of the factors in Kendall's notation are more closely explained in table 6.

Table 6: Kendall's notation components

A	Arrival time	Represents randomness of customer arrivals where the meaning of customer depends on the system under discussion. Variables can be either the number of arrivals in a time interval or time between two consecutive arrivals. Number of arrivals is a discrete variable and follows a Poisson distribution whereas time between consecutive arrivals follows an exponential distribution. This relationship between Poisson and exponential distribution is heavily utilized in queueing theory. (Lakatos et al. 2013, 191; Tulsian & Pandey 2002)
B	Service time	The service process involves the number of servers, customers that are being served and the duration of this service. Most common type of service time follows Poisson process, which means Poisson as arrival and exponential as service process. The poisson process can be also called as Markovian in this and Arrival process' context. (Kardi 2014; Sztrik 2010, 16)
C	Number of servers	Queue server can be for instace a cashier, a machine, or a staff member. While A and B often are modified to letter M (markovian), number of servers is modified to its number. For example, M/M/1 denotes a Poisson arrival process, exponential distribution as service time and one server. (Lakatos et al. 2013, 191; Kardi 2014)
D	Capacity of system	Represents the maximum number of customers allowed in the queue. The queue can for example be a facility. For example, M/M/1 system assumes that the capacity of the queue is unlimited, and M/M/1/K can be used to portray that the queue has a capacity of K. (Kardi 2014; Sztrik 2010, 17)
E	Population size	Indicates the size of potential customer pool in the system. If the capacity of system exceeds population size, more customers are not accepted for service (Sztrik 2010, 17).
F	Queue discipline	Usually indicates the rule that server follows in receiving customers for the service. Common examples are first-in-first-out (FIFO) where the earliest customer leaves earliest, last-in-first-out (LIFO) where the latest customer leaves earliest, random selection or priority-based service. (Sztrik 2012, 10; Sztrik 2010, 16)

After obtaining the parameters according to queuing theory, optimal design models can be built. To design effective queuing systems, the following three factors are the main components of a design model: the decision variables, benefits & costs, and the objective function. Decision variables include the variables presented in table 6, for example the arrival times, service times and the number of servers. Benefits and costs area brings cost

factors into consideration when designing queuing theory models. For instance, increasing the number of servers to have a more effective queue also results in increased costs since the number of staff increases. Based on the decision variables and benefits & costs, objective function then seeks to optimise some specific system performance measure. (Sztrik 2010, 15-16) Together with managing costs as shown above, other viewpoint of applying queuing theory models in practical situations is to acknowledge utilization and wait time as measurements of interest. To calculate utilization, the following formula applies:

$$\rho = \lambda / (c * \mu) \quad (3)$$

where ρ = utilization, λ = arrival rate, c = number of servers, μ = service rate per server.

This information from the formula (3) can help further calculate for example average wait times and customer numbers in a queue but also probabilities between servers being idle or a number of customers being in the system at a given time. This system information can be utilized for example in analysing the effects of average waiting time while changing the number of servers. (Johnson 2008) It is thus evident that the system designer has a control to optimize the queue with the help of queuing tools by modifying many aspects of a queue. For instance, the arriving customer can be allowed to the queue or turned away, which customer is being served at which point and how the work is allocated to the servers. Additionally, to achieve an effective queue design, changing the number of servers is not mandatory. (Che & Tercieux 2021)

3.2 Queuing theory and modelling in telecommunications incident management

Organizations especially in IT frequently design their organization and structural processes in accordance with their established strategic goals. This need for aligning business level objectives with incident management efforts requires consistently evaluating and optimizing the current processes to come up with the most effective incident management processes. (Bartolini et al. 2012) According to O'Dwyer (2014) queuing theory is a powerful tool that provides relevant mathematical models particularly in technology support functions such as service desk or customer support work. This makes it largely applicable for optimizing incident management since ITIL incident management process includes service desk staff

that for instance receive the incidents in addition to its being comparable to a basic IT service desk process in many ways.

Queuing theory is typically applicable to many fields containing queue sequences and service delivery. In telecommunications context, queuing theory has been heavily used in different use cases such as in network traffic control and load balancing where Little's law has been used. In addition, more comprehensive information can be acquired by using other theoretical tools of queuing theory. To give an example based on previous studies, an effective queuing model for network and traffic jam prediction is (M/M/1): ((C+1)/FIFO) and (M/M/2): ((C+1)/FIFO). For example, the first model states that the process follows a Poisson arrivals and exponential distribution with one server and a capacity of (C+1) and service discipline of FIFO. (Imamverdiyev & Nabiyev 2016)

Organizations with IT service desk have been previously modelled with an open queuing network model and multi-server first-in-first-out approach. An open queuing network allows any number of customers to both arrive or leave the system at any time. The main advantages of these types of models are attractive trade-offs between complexity of the model and accuracy of the model. Additionally, this type of model has been proved effective for mimicking IT support organizations well while capturing metrics such as mean time to incident resolution. However, it is still not as accurate as desired if some in-depth details of the queues have to be taken into account. (Bartolini et al. 2012)

Sojourn times refer to a total time an incident spends in the queue, which include both the time spent in the queue before servicing and time taken by the staff to address and resolve the incident. An open queue network with FIFO-service discipline may not be effective enough for capturing the variability in sojourn times and take account nuances of different support groups that handle the incidents. As another approach, multiple priority queue models offer more complexity but also more accurate modelling of sojourn times in incident management systems. One of the main reasons is that the multiple-priority queue models allow different priority levels for incidents, which is the case in real-world systems if they follow ITIL principles. (Bartolini et al. 2012)

Imamverdiyev & Nabiyev (2016) have proposed a model for Information Security Monitoring systems based on queuing theory, which is largely related to potential approach in incident management in telecommunications context since a potential incident handling

logic is presented in the paper. Information security monitoring relates to acquisition of real-time data to monitor and potentially eliminate cyber-attacks with high efficiency to better manage corporate network via information security service (ISS). Information is received through various sensors, and each received incident is given a priority based on a set of criteria. ISS is then used to organize the tickets based on the priorities with the usage of optimal service discipline. In the study, the selected queuing model for the system is M/G/1, which is a single server system with Poisson-arrivals and general service time distribution. (Imamverdiyev & Nabiyev (2016) This study also utilizes the ITIL incident management priority matrix in defining the priorities linking ITIL practices to ticket handling with queuing theory.

In the study, the M/G/1 model is used to studying mean waiting times in the queue. It is assumed that longer the ticket stays in a system, the more the effectiveness decreases. In the case study, incidents are compartmentalized to three different priority handling methods. Critical incidents are set as absolute priority, medium-priority incidents have relative priority and low-priority incidents do not have a set priority. In other words, when a critical ticket is received, it automatically is prioritized to the top of the queue and handling of low priority tickets stop. When medium priority tickets are received, the current handling does not stop but the high priority incident is selected from the queue after the current ticket is handled. Low-priority tickets are shared in either FIFO, LIFO or Random pick service disciplines. (Imamverdiyev & Nabiyev 2016)

Punyateera et al. (2014) have researched ways to improve internet service provider's (ISP) incident management effectiveness. They have set baseline with the help of studying their current incident management process that follows a FIFO queue and as a result, have simulated the number of staff required to solve a specific number of incidents without affecting output. After studying the current process, an improved process was implemented that utilizes priority queue that compartmentalizes incident tickets to minor cases and major cases. Together with modifying the process to adopt ITIL best practises, this proposed model was deemed more effective since the results indicated that the model could handle more incidents with fewer staff members and the model also enables better resource allocation by adjusting the number of servers for minor and major cases. (Punyateera et al. 2014)

O'Dwyer (2012) points out a possible problem for prioritization of ITIL incident management process for some organizations. The problem arises when the organization has

diverse operations with several distinct technologies that can cause incidents. ITIL incident management has an escalation process where in the incident diagnosis stage, the ticket is passed to the next level of support agents that handle the incidents until resolution. O'Dwyer further states that linear assignment of the ticket from one support group to next is ineffective when the staff lack competencies to route or solve the incidents caused by the number of different technologies they must handle. As a result, incidents are not effectively routed but instead incidents are at risk to be misclassified often leading to extended resolution times. One possible way to escalate tickets is depicted in figure 4.

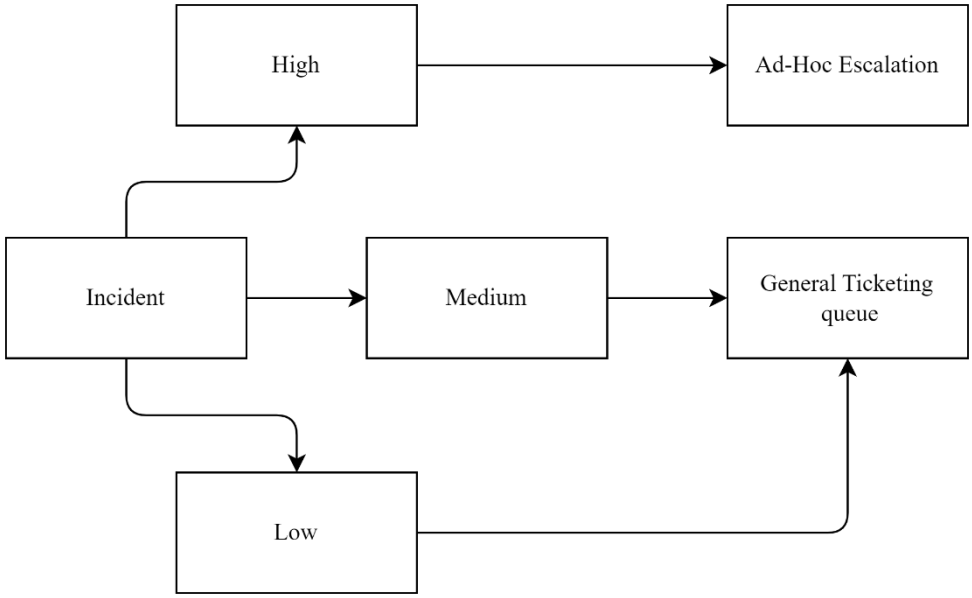

Figure 4: Incident Escalation Process (O'Dwyer 2012)

The incident management ticket queue portrayed in figure 4 provides an example of a multiclass queuing network; it accounts for different routes for completion with distinct service requirements. It also incorporates priority model. The service desk agent first prioritizes the incident, and this classification then determines the escalation process. The pitfall of this type of incident escalation process is shown especially in managing the general ticketing queue. If the process of choosing the tickets is manual in the general ticket queue, the staff could create knowledge-silos by only choosing the easiest tickets and leaving difficult tickets in the queue. Based on this queuing model, new process is formed where a subject matter expert receives the ticket and appropriately escalates it to the correct technology function that are formed in place of support groups. With this approach, tickets

are correctly routed to subject matter experts of their own technology and priorities from low to high separately exist inside each technology function. (O'Dwyer 2012)

Bober (2014) has modelled an IT service desk as illustrated in figure 5 with three levels of support with their own ticket queues. First level is set to receive incoming inputs to the system and tickets are then processed and assigned to the correct level if an immediate resolution is not possible. The first level receives the incidents with FIFO service discipline but after processing the ticket, chosen priority influences the decision of which queue the incident transitions to. This example contains three different priorities: P1, P2 and P3 where P1 has the biggest priority and P2 the lowest. P2 and P3 tickets are escalated to level 3 with priority queue and P1 tickets to level 2 with FIFO queue.

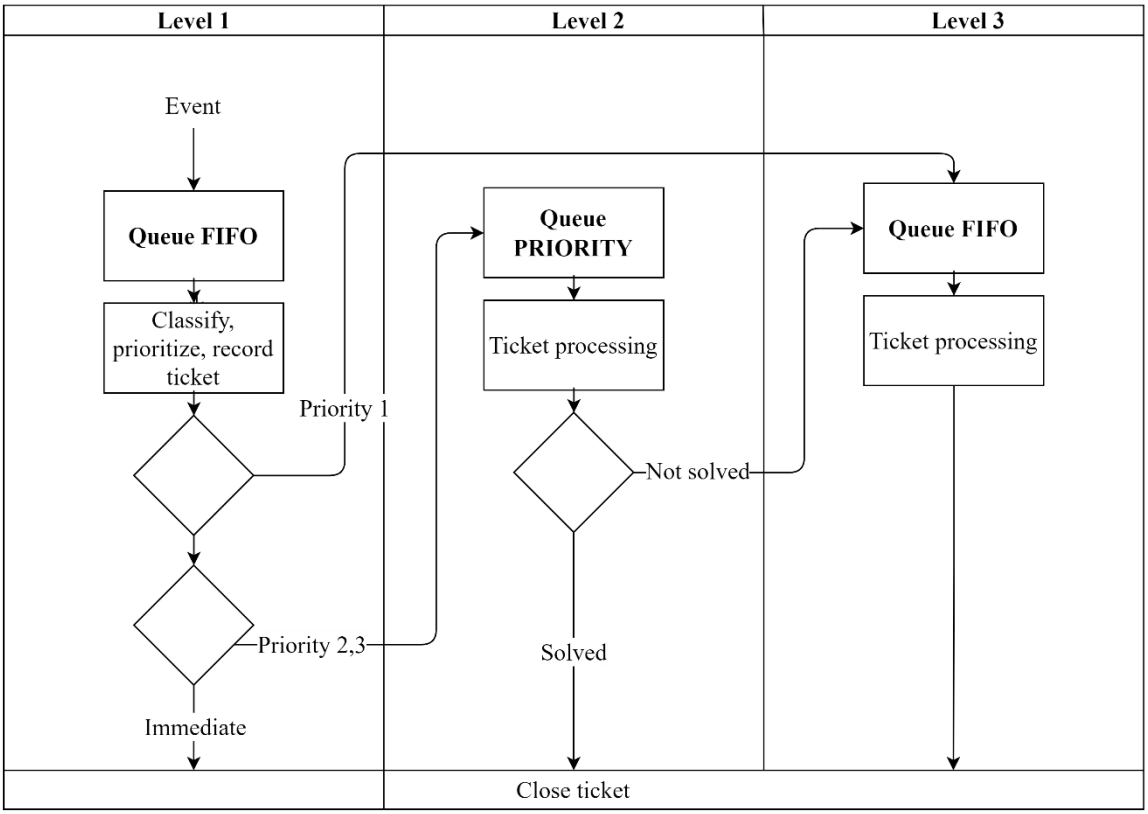

Figure 5: An example model of a service desk (Bober 2014)

Bober (2014) states that complex mathematical methods or queuing methods are not easy to manage to optimize service desk operations at least for organizations or teams lacking the tools or skills to handle these. As an attractive alternative, he presents discrete event simulation to give results notably in two key areas: as a simulation study to find

improvements in the process or as a tool to assist daily operations for managers of service desks. Instantaneous changes in the states of the system are characteristics of a discrete system; For instance, if a ticket enters or exists the system, the quantity of tickets in the environment changes. (Bartsch et al. 2010) Additionally, Sencer & Ozel (2013) developed a simulation-based decision support environment to tackle the problem with simplistic models such as M/M/n called Erlang C queuing model. They recommended that along with mathematical modelling, simulation should be used to generate more reliable results.

It has been discovered that in the queuing systems, customer dissatisfaction usually stems from waiting a long period of time in the queue. Thus, to optimize queue efforts, enough servers are required to provide adequate service while making sure that servers are not idle with low levels of utilization. Waiting time depends on many distinct factors such as the rate the service is given, efficiency of servers, service types and arrival rate of customers (Sarkar et al. 2011). In the quality of service, waiting time and queue length are crucial factors. It is then crucial to determine the optimal queue length, waiting time and a priority by how the customers are served to provide the most effective service when cost of service is also considered. To solve this, simulation is a proven way in modelling these types of situations and evaluate different approaches especially when the objective of the optimization project is to create suggestions for layout changes in the queue. (Madadi 2013)

One way to gather valuable process insights is to approach simulation as presented by Bober (2014). The simulation building based on the queuing model starts from obtaining quantitative data from current ticketing environment. After getting the data from the system, capacity calculation is done to get a utilization rate as a result. The basic metrics are then crucial from the customer's point of view and in this example, chosen metrics required are response time and total service time. After building the model, two use-cases introduced are staff capacity planning and showing the effect of staff skills to response time exceedance. (Bober 2014) Certain procedures should be followed that result in the implementation of a model that is acceptable for use in simulations. Figure 6 depicts phases that can be used to categorize the fundamental modelling and subsequent simulation steps. (Jenčová 2023)

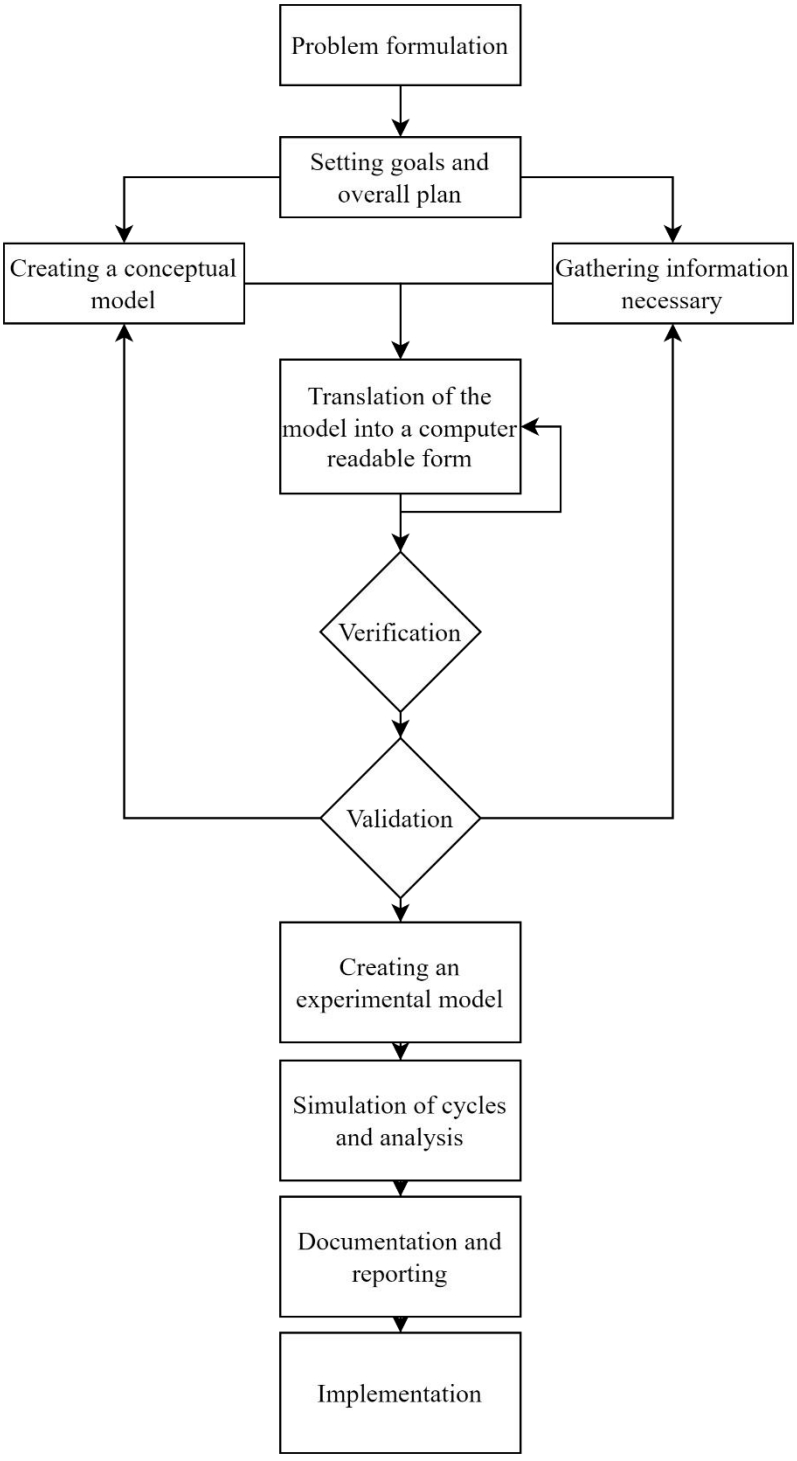

Figure 6: Process for simulation (Jenčová 2023)

To get an idea of simulation approach fundamentals, it is crucial to get familiar with various simulation concepts. While a model is a replica of a real system, an event in the system is a circumstance that modifies the system state. Event can be a customer arrival, customer

service and customer departure. These are examples of endogenous events because they occur in the scope of the simulated system while exogenous events happen outside simulation. A system is described by system state variables that define the system state at any given point in time. For example, these variables can be chosen based on the specific questions asked about the system. The information is moved in a system with entities that are objects that changes over time. For instance, they interact with other objects like a situation where they represent both dynamic entities as customers in a queue and static entities as servers. Each entity has a set of attributes; a customer can have a destination, service level and a time of arrival. (Banks 1999)

A resource is another simulation concept related to entities. It is an entity that provides service to dynamic entities. They can be requested and released by entities while they move through the system. Servers servicing customers are clear examples of resources in a system. A fundamental part of entities' behaviour is managing the activity of events with list processing. Lists in this case are queues in the system and list processing introduces a way for the system to sort the queue according to set rules. FIFO is the most common queue or service discipline used for processing. A discrete-event simulation model (DES) is a simulation model that will utilize these key concepts. State variables are changed in points of time where events happen, and the events are happening because of delays and activity times. DES is conducted by a technique that advances simulated time forward. At every event, the system state is updated together with any resource captures and releases that may take place at that moment. (Banks 1999)

Other approaches for solving queue problems compared to simulation exist but some with substantial disadvantages. There are three examples to make projections of performance based on the existing information: carry out an after the fact analysis based on real values, create a simplistic projection from experience to anticipated future, or create an analytical model based on queuing theory. The first two options have several weaknesses. While the first option does not lead to optimal results since the analysis is always after the fact, the second option is also enduring problems in estimating results under changing system loads. The queuing theory model provides mathematical approach with a set of equations that solve the needed parameters. While it is a relatively good approach for simple queuing problems, several assumptions are required, which lowers the performance in a real-world environment. While the simulation model needs also certain assumptions, the reality can still

be modelled with greater accuracy leading to more valuable answers compared to just using queueing theory formulas. (Stallings 2011)

There are different types of simulation techniques to execute a simulation study. First, what if analysis is a tool for improvement that evaluates how changes in strategic or operational level are influencing the business. Second, system operation analysis helps to gather insight from the system operations and as a result, define changes needed. Third, optimisation searches for a set of system values that leads to the best result under existing conditions and restrictions. (Bober 2014) Like the process in figure 6, Bober (2014) recommends the following procedure for utilizing simulation:

1. The problem to be solved must be clearly defined and simulation technique used to come up with the question that needs answering.
2. The model will be planned, built, and debugged in a way that reflects the initial defined problem.
3. Systems variables need to be defined together with their values and desired responses.
4. Run experiments and save the results.
5. Analyse simulation results and find the answer to the initial problem.

Law (2019) reinforces the simulation procedure by highlighting the importance of validation to create credible simulation models. Validation seeks to discover if the simulation model represents the real system accurately. Techniques for credible models include precise problem formulating since the appropriate level of model detail is impossible to decide without a clear problem description. Another example is to communicate with subject matter experts to gain an understanding of the modelled system and create an assumptions document so that the limitations are clear. A crucial part in the simulation itself is to perform sensitivity analysis to measure which factors in the model are having an impact to the desired results. For instance, value of a parameter, probability distribution choice or an entity transitioning through the system can be examined with sensitivity analysis. The most important test of validation is to investigate if the output data from the simulation model and from the real system match. Based on the feedback, the model is then further refined to represent the real system more accurately. (Law 2019)

The question of simulation credibility is also researched in telecommunications networks context. It has been a popular opinion in scientific community that many published results of telecommunications simulations lack credibility. To address this, several methods have been implemented. First path of building a valid simulation model is to use a realistic conceptual model of the system with appropriate assumptions of the system mechanics and limitations. Furthermore, valid simulation experiments are important where applying appropriate elementary sources of randomness and conducting a proper analysis of simulation output data is critical. It is stated that a stochastic simulation where random processes are being simulated, various statistical methods must be applied to analyse the output data. This can be done for example by examining statistical errors and degrees of confidence for simulation output. (Pawlikowski et al. 2002)

Fixed sample size scenario refers to a situation where the duration of a stochastic simulation is decided for instance by the length of the total simulation run time. Fixed run length is criticized for not producing a confidence interval that produces a desired confidence level. This is why sequential simulation approach is also used where final error can be actively controlled by monitoring confidence interval and adjusting simulation length until a desired level of statistical error is achieved. (Pawlikowski et al. 2002) Another type of simulation called steady state simulation is a scenario where the system is assumed to operate indefinitely. It said to be a good indicator if the simulation stops automatically when results reach steady state. Some additional important considerations in simulations are to choose a credible simulation tool and publish credibility factors for simulation results. (Sarkar & Gutiérrez 2014)

The application of queuing theory in incident management is thus researched topic in the research community with various approaches taken usually based on the case problem at hand. Adoption of different queuing models is common, which is based on the type of situation at hand and a specific queuing model does not fit in all possible scenarios; this also underlines the importance of modelling to effectively create a conceptual model that reflects the real world with valid assumptions and is based on the real-world data. To assist queue analysis, simulation approach can be utilized to provide a way to transition a conceptual model into a computer readable form, which allows for simulating different scenarios and optimizing the system based on the results.

4 Methodology

Theoretical framework of this thesis was conducted as a literature review. The chapter contained theory related to both incident management and queuing theory, which are two vital elements also in this empirical section. The findings of the literature review can thus be applied to this empirical part as well. This section starts with presenting the methodology used in helping the incident management process optimization efforts in the target company Telia. The first part of this chapter discusses the real-world data used and how it will be utilized in developing a conceptual model for simulation purposes. The next part deep dives into approach taken in conducting a simulation study, which will simulate the service desk based on the conceptual model developed. To close the chapter, an approach to evaluating and validating the results is presented.

4.1 Data collection and analysis

In this thesis, real-world data from Telia Wholesale incident management service desk is used to help build a simulation model for optimizing the company's current process. The data was collected from the ticketing system of the service desk that contains tickets from the year 2021 to 2023. This results in almost 7000 rows of data that include a wide range of various incident tickets with information regarding each of them. The data used in this thesis was not limited to a specific year since a usage of wider spectrum of the data was crucial to obtain desired insights and eliminate the cyclical nature of incoming tickets. In other words, utilizing tickets starting from the year 2021 gives a detailed overview of the history of the incident management service desk and its overall performance. The system also contains ticketing data from other IM teams, which can also be used if the quality of Wholesale data is lacking.

To create a simulation model and even conceptual model before it, certain calculations from this data are needed to obtain the right inputs for the simulation model to make sure that the simulation gives results that are based on the historical performance of the service desk. This also helps with validating and later, applying the insights learned in the simulation in Telia Wholesale incident management team's daily operations. The data collected is deemed a

good representation of the ticket flow although it contains some minor inaccuracies and does not provide in-depth information; for instance, data about times taken between changing the statuses of the tickets were not available. All data variables from the Telia Wholesale data are depicted in table 7.

Table 7: Data columns from the ticketing system

Column Name	**Data Type**
inc_ticket_ID	float64
inc_start_date	datetime64[ns]
inc_resolve_date	object
SLA - resolved in time	object
SLA_Service_level	object
Incident_origin	object
Total Incidents #	float64
SLA_repair_time_h	object
SLA_restoration_time_h	object
1st cust.notif. in 15min	object
Efficiency Factor	object
SLA Service Precision [%]	object
inc_status_open	object
inc_ticket_type	object
SLA_interruption_time_h	object
Inter-arrival (days)	float64
Inter-arrival (minutes)	float64

To get a better grasp of the queue's performance, interarrival times and service times are essential. To achieve this, customer arrival pattern was calculated from the data by utilizing the "inc_start_date" column. Interarrival time is the time between consecutive arrivals to the queuing system. In this case, the interarrival time refers to the time between the arrivals of tickets to the ticketing system. This can be calculated by calculating time difference between two consecutive arrivals where the first arrival is labelled as zero since there is no prior ticket to calculate interarrival time. Interarrival times are plotted in figure 7. Since the data contains some extremely high values that distort the view and may be either data errors or special cases, the x axis is limited to the 99^{th} percentile. For customer service time, a key piece of information needed is the time the ticket has been spending in the status "in processing". This status would communicate the time an agent has been manually handling the ticket, but this information is not always available from the dataset, which is why other ways to calculate service times are required.

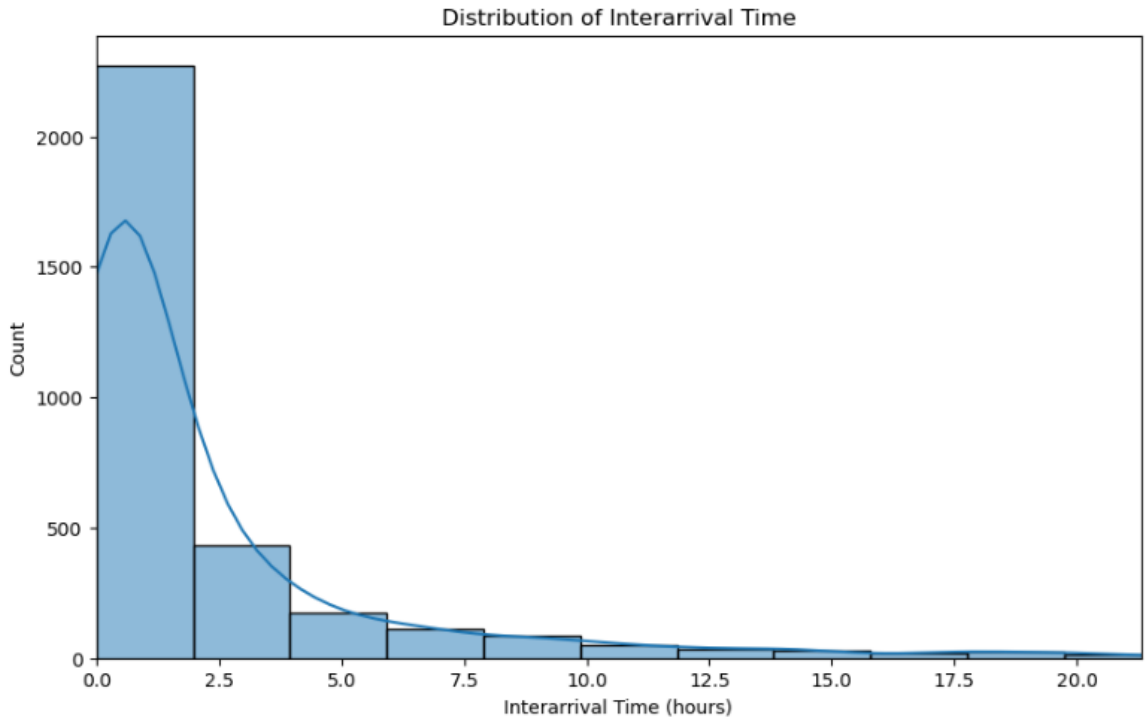

Figure 7: Distribution of interarrival times

Other approaches for calculating customer service time are to utilize the difference between incident end date and start date or either the with columns “SLA_restoration_time_h” and “SLA_repair_time_h” by taking its average. The disadvantage is that these include every work step inside the incident, which means that not all actions made to the ticket were made by the agent in the service desk. For example, some incidents may need external work done. Thus, instead of recording only actions taken by the service desk agent, data used in the model could be the overall time an incident has taken from its creation to resolution, which would still provide a sufficient way to tune the simulation to model this service desk. For this context, SLA restoration column is the most appropriate since the objective is to evaluate the performance of the incident management service desk against its service level agreements.

4.2 Simulation approach

After obtaining key information from the ticketing system to calculate needed parameters for queuing model, the simulation process will continue with modelling a conceptual model

that represents the incident management service desk of Telia Wholesale. The general simulation framework consists of defining problem, creating a conceptual model, realizing it to an executable model, and verifying and validating the model before the simulation results. The objective of this simulation is to introduce and test different queue disciplines and their performance in an environment that closely represents the real ticketing system queue. Queue or service discipline is the service mechanism that will assign tickets to the servers in a controlled fashion. The results should be then used to recommend an optimal approach to sort the tickets in a ticketing system while making sure that the solution can be easily implemented to most ticketing systems. To compare the performance between each queue discipline, the main KPI is to measure their abilities to meet SLAs. For this reason, SLA compliance will be closely followed to compare between different queuing methods. The most optimal queue discipline must thus be effective in resolving the tickets before the limit agreed in service level agreement is reached.

In a thesis by Kilpi (2022), some areas for improvement were found for the current Telia Wholesale incident management service desk, which were procured by conducting interviews among the service desk staff. One of the recommended features were to implement a notification for the service desk agent when SLA time is approaching, which would allow better monitoring of tickets that are in danger of failing to meet set SLA levels. Another potential feature found was to route incidents to different types of priority queues to allow for prioritization among tickets compared to the current FIFO approach. These two areas are thus also essential in this thesis to help take service desk staff expectations into account in building a simulation and later, recommending the optimizations for the current incident management process.

In addition to linking incident management process to its SLAs, another crucial part is to make sure that the process follows practices learned from ITIL. One way to solve the need for routing incidents by their priority is to use a priority matrix, which can be calculated with urgency and impact ratings discussed in ITIL. ITIL provides a clear way for managing a terminology so that all parties have the same understanding of the operations of the incident management service desk. ITIL practices will thus be a major part in designing the to-be incident management process, which will also be reflected in the simulation approach by taking ITIL into account.

The incident management service desk is mapped onto the simulation model by utilizing queuing theory approaches. The system in place is the ticketing system used by the service desk and the queue is the ticketing queue in that system. In this system, instead of customers, tickets are labeled as arrivals. While customers play an important role in incident management, this simulation is focusing on the ticketing system where the customer inquiry has been changed into an incident ticket needing resolving. Servers in the system are incident management service desk agents that process and resolve the tickets. There are multiple servers at any given time meaning that the simulation is not modelled using only a single queue with single server. Other essential simulation component is the service mechanism or service discipline used, which will explain how the tickets are assigned to the servers. A high-level view of the queuing system is depicted in figure 8.

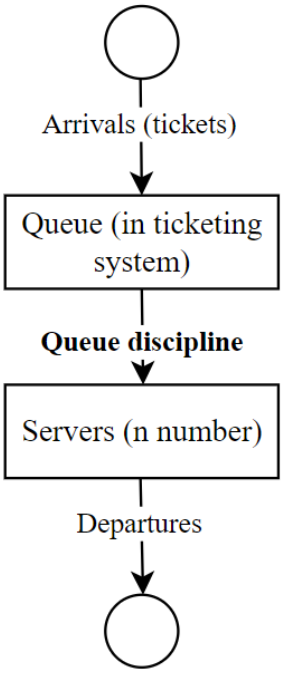

Figure 8: High-level queuing model for the service desk ticketing system

Incoming tickets have a certain distribution as well as service times. Since the problem solving requires many different components, simulation is decided as the best approach compared to just using queuing theory related formulas and calculations such as Little's Law. The ticketing system is a complex system, which will require several assumptions that discussed later. Another justification for simulation is the impossibility of testing the changes in a live environment for a few reasons. First, the potential impact on customers is huge and

brings a lot of risks if the tested changes lower the performance of solving incidents. Second, the team is in the process of changing the current ticketing system, and this current system does not have all capabilities that still are included to the simulation model, which will make testing the system difficult now with current state of ITSM systems used by the team.

To create a conceptual framework of the model, exact information regarding the operations of Telia Wholesale service desk is needed. Therefore, knowledge about customer flow and service desk way of working are crucial. Telia Wholesale provides customers different kinds of telecom services and products and offers technical support when there are incidents in the services provided. These incidents are recorded as incident management tickets that are managed in a ticketing system that is a subset of an ITSM system. Incident management ticket handling is just a one part of the Telia Wholesale Technical Services operations since they are also dealing with various service requests, which are out of scope for this simulation since the SLA limits are usually different or not properly set. The ticketing system handles the incoming ticket mass and are used by the service desk agents to keep track of the current tickets.

Alongside the ticketing system, the team is also using a workload management system with the objective of assigning work to agents in a controlled manner. For instance, the system can handle logic rules that allows for prioritization of incoming tickets and assigning them to specific agents based on their skills. However, at the moment the system is only used for receiving emails with FIFO queue discipline, which limits the possibilities of the system. In addition to emails, customers can reach the team also by calling, which is handled via the same workload management system with FIFO discipline.

To comply more closely with ITIL principles and to take account the recommendations from the service desk staff as Kilpi (2022) has concluded, the incident management process has been slightly modified. The improved process is also used in a simulation since many of the improvements to the process require a more effective use of the current tools such as the workload management system. The current process and its systems do not allow for effective ticket prioritization and allocation, which makes it also difficult to adhere some of the standards set by ITIL. For the simulation however, a baseline has been set to still simulate how the process without improvements will compare against improved scenarios. This will also verify if the environment of the service desk works the best with simple rules and logic

compared taking SLAs and priority levels of the tickets into account to create the most effective business rules.

There is one major change in the process, which is not yet widely used in the technical services' daily operations. To understand this, a typical ticket flow should be described. The incident management process will start when the customer contacts the service desk via appropriate channels. The ticket is then created by the service desk agent by describing the incident and categorizing it. Then, severity is determined by giving the incident ticket an impact rating, which will be used later on the process by the ticketing system. Currently, impact rating field is not playing a major role in the service desk; for instance, it is not used to sort tickets but is only a field that is automatically filled, and this is the biggest change in the process. This impact rating in the new process will be crucial to find the most important incidents among ticket mass, which will help eliminate a manual checking of the ticket queue and opening different tickets to determine the most crucial one.

Another way to prioritize the ticket, urgency rating is automatically calculated by the system. It is calculated with the SLA rating the product has. Priority rating is then calculated with the urgency and impact ratings in the ticket and added to the ticketing system queue. The scope of the simulation starts only after the ticket has been added to the queue and has the needed information specified such as its priority rating and SLA. The ticket is then assigned to an available agent that will open the ticket and change its status to in progress. This status helps in communicating that the ticket has been opened and a service desk agent is processing it and is not available in the queue in that moment to avoid multiple agents working on one ticket simultaneously. The agent then investigates the incident, diagnoses it and attempts to resolve it. Incident tickets often vary by their complexity and can require external work and support from other service desk agents. It is still assumed that every agent has the same skillset, which is also the approach taken in the simulation.

To help translate this into a computer-adjusted model, the different activities of the incident management process should be illustrated with modeling concepts. In the model, the entities are represented as tickets that flow through the system. Each ticket is generated with attributes with a ticket id, priority, SLA-level, SLA-time related to SLA level and arrival time. These attributes are used to manage the lifecycle of the ticket in the simulation. The resources that communicate with these dynamic entities in this model are the service desk

agents. Each agent can handle one ticket at a time and the workload is split between agents so that all have similar utilization.

System state variables in the model are various metrics that will be collected to get a comprehensive view of the system at any given time. Since SLA compliance for how many tickets are solved inside SLA limits are important, information will be tracked regarding SLA compliance levels for each priority level, total number of tickets within each SLA category, waiting time, service time and total number of handled tickets for example. These will be then used to track the performance of the system. List processing will be used to manage a ticket queue that is waiting on an agent to become available to handle the top ticket. Tickets will be sorted according to a set queue discipline. The idea is that the system environment will be the same with the same system state variables to accurately compare between different queue disciplines to gather data how they perform under the same conditions.

Choosing the most optimal queue discipline is one of the main objectives of this study. It is thus exceedingly important to choose the most optimal disciplines that are then represented in the simulation. The choosing rationale will be based on the previous state research, relevance in telecommunications sector and relevance for Telia Wholesale service desk needs. In the literature review, many approaches for sorting the incident management or service desk queues were presented meaning that there is not a one discipline that works best in all situations and choosing the correct discipline is scenario specific. The most known and the most widely used queue discipline is to sort the queue based on the arrival of the ticket. First-In-First-Out (FIFO) or First-Come-First-Served (FCFS) does exactly this by assigning the oldest job in the queue first.

While FIFO is sometimes effective enough and fair with its approach, it still brings a few problems under certain conditions. First, as the literature review has suggested FIFO is not as often used to optimize queue performance and it is even more common to switch from initial FIFO to a more optimal approach. While Telia Wholesale uses FIFO currently, one of the main issues identified is that it does not allow for prioritization between tickets, which makes it harder to stay within SLA limits if the only criteria are to just choose the oldest job created in the queue instead of remaining SLA or priority level. In the simulation, FIFO is thus modelled as the baseline, which means that other queue disciplines are compared against FIFO to get an idea if the current ticket handling process is improved.

A common approach especially in service desks in incident management is to utilize the priority ratings in determining the correct place in the queue for that ticket. Priority queuing has been an effective way to help solve the most critical incidents first and assigning lower priority tickets only after the bigger priority tickets have been solved. It should also work well within ITIL context since it can use priority matrix in determining a correct priority for the ticket after ticket creation. It is exciting to study if the priority-based queuing can perform better also in this simulation study. Another potential approach is to sort the queue based on the SLA-rating since the main KPI for the queue discipline comparison is SLA compliance. This discipline would work by choosing the ticket that has the lowest remaining SLA time. If SLA has been breached, it will still be considered as the lowest remaining SLA time, so no ticket is abandoned after failing.

The logic in FIFO-queue and SLA-queue are straightforward to understand and implement but there are various approaches for priority-based queue. In this instance, there will be four priorities to choose from: critical, high, medium and low, which is a familiar approach from ITIL incident management. In the first scenario, a queue will be sorted by priority first. If there are multiple tickets with the same priority, these will be sorted according to FIFO meaning that the oldest ticket will be assigned first. The second scenario is that after sorting the queue based on priority, the same priority tickets will be sorted according to SLA meaning the lowest remaining SLA time will be chosen first. In other words, the second scenario will use two different ticket attributes to help sort them in the queue and assign them to the agents. The chosen queue disciplines for this study are summarized in table 8.

Table 8: Chosen queue disciplines for the study

Queue discipline	Role	Description
FIFO	Baseline	Tickets are handled in the order they arrive, first-in-first-out.
SLA	Takes ticket SLA into account	Tickets are handled based on their Service Level Agreement. Tickets closest to SLA breach are processed first.
PRIORITY_FIFO	Takes ticket priority into account	Tickets are first sorted based on their priority. Among tickets with the same priority, they are processed in the order they arrive (FIFO).
PRIORITY_SLA	Takes ticket priority and SLA into account	Tickets are first sorted based on their priority. Among tickets with the same priority, they are processed based on their SLA, with tickets closest to SLA breach processed first.

Upon deciding queue disciplines to represent system dynamics, specific parameters within these disciplines are essential to help close the gap between the real world and the simulation model. In the context of this study, parameters are the constants and distributions that influence the model's behavior and make it mimic Telia Wholesale incident management service desk. As discussed earlier in the chapter, real-world data was used to gather insights on the customer arrivals and service times. These two will help determining the correct queuing model to be used to represent the service desk in the simulation. In addition, other parameters are needed to create a scenario for the simulation including what kinds of tickets are created, how long are tickets being handled in the simulation and how many agents are there available to process tickets.

To represent the incident management service desk, M/M/c queuing model was chosen as a reasonable assumption for the arrival times and service times. As figure 9 suggests, when exponential distribution with the same mean has been overlayed with service time, some similarities are shown even though service time does not exactly follow exponential distribution. Both have a single peak with a long tail. Additionally, M/M/c is a common model in queuing theory and often used to model service desks and call centers that have many similarities to the incident management service desk of this study. This approach will also be the most practical to reduce the complexity of the model and allow for a more flexible models with a more complex service time distributions.

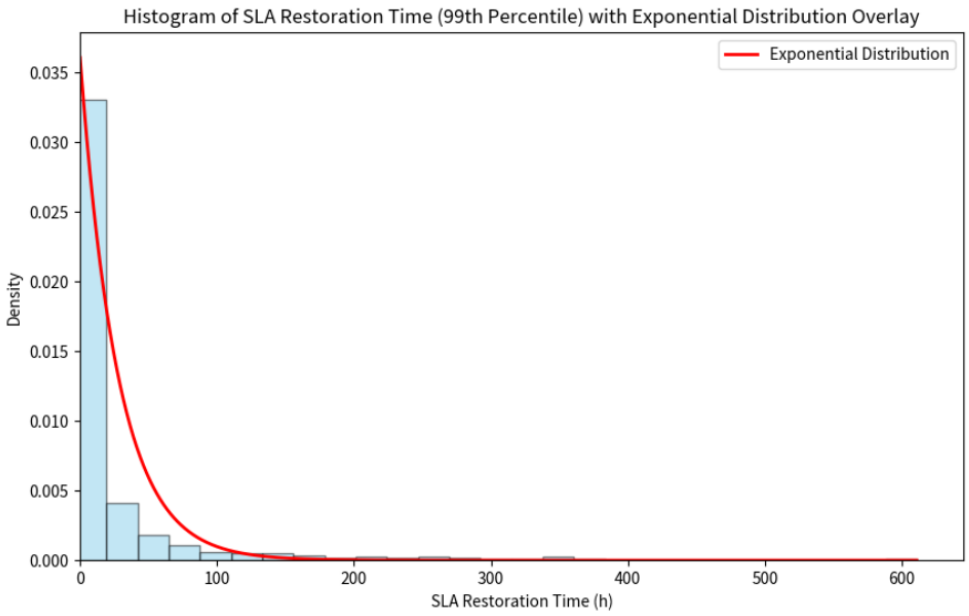

Figure 9: SLA restoration time distribution compared to exponential distribution

For average service time, either SLA restoration time or other handling times from the Telia Wholesale data were not chosen to computer readable model that will be converted to simulation model. The reason lies in Service Level Agreements that were chosen based on a specific product in Telia Wholesale portfolio and its set SLA-levels. In these SLAs, three different service levels are included that will be named as SLA1, SLA2 and SLA3. SLA1 will have restoration time of 4 hours, SLA2 restoration time will be 8 hours and SLA3 will have 12-hour restoration time. However, since the data only contains handling times where actions taken only by the agent are not recorded, average handling times go beyond some SLA-times, which will impact the simulation and comparison of different queue disciplines greatly.

If the average handling time is considerably bigger than set SLA-time, the chosen queue discipline does not matter since all tickets fail to be solved within SLA limits no matter the discipline used. To solve this, for average service time, the real-world data were expanded to take account all incident tickets from the ticketing system since some tickets outside Telia Wholesale incident management service desk contain information regarding agent handling times. Based on that, an average service time of 1.57 hours was chosen as a reasonable assumption that was also validated by the service desk staff as realistic. For distribution of different service levels however, only the Wholesale incident management data was used. For distribution of priorities, the whole ticketing system again was used since the Wholesale service desk does not have sufficient data regarding ticket priorities. These distributions are shown in table 9. SLA-times are assumed to apply to all priority levels equally.

Table 9: Distributions of priorities and SLAs

Distribution	P1	P2	P3	P4
Priority distribution	0,067	0,115	0,403	0,415
SLA distribution (SLA1, 4 hours)	0,06	0,06	0,06	0,06
SLA distribution (SLA2, 8 hours)	0,91	0,91	0,91	0,91
SLA distribution (SLA3, 12 hours)	0,03	0,03	0,03	0,03

The simulation will be built as a discrete event simulation (DES). Besides its strengths in simulating system dynamics event-by-event, discrete event simulation has already been deemed successful for helping companies' decision making and gained use cases from different industries. It is widely used also in the service sector such as modelling and

simulating queuing systems like call centers and healthcare patient waiting lists. (Ing et al. 2010) These contain many similarities to service desk -based queuing modelling. Another advantage of DES is how well it allows for "what-if" analysis, which means the analysis of changing system parameters or variables to create the most effective environment. In other words, it allows for testing different scenarios and the most optimal scenario can be chosen to optimize the real-world system. This is vital for the context of this thesis because the most optimal way to sort the queue and assign tickets to service desk agents requires comparison and different scenarios to gather a comprehensive view of the system and its dynamics.

A common approach for DES is to use a simulation software such as Arena, but this simulation is built with python and its open source SimPy library. It is a framework for process-based DES built on top of regular Python. SimPy can be used to model components active components such as customers or tickets and shared resources to simulate congestion spots with a restricted capacity such as servers (SimPy 2020). For instance, processes can be created to follow specific procedures in a real-world system and the user can run simulations to assess the system's performance under different scenarios and conditions making it a great tool for what-if analysis and process optimization.

4.3 Simulation implementation

The implementation of the model starts with importing required libraries that in this instance are SimPy, random, numpy and pandas. Table 10 describes the main SimPy library elements used in the simulation code. In addition to SimPy, random module is an important part of the simulation that introduces variability to the system to better represent the real-world scenario and its stochastic nature. Numpy module is used for example as calculating summary metrics to allow for comparison between queue disciplines and pandas is used to store the data from tickets and metrics so that they can be comprehensively analyzed later.

Table 10: Main SimPy features used in the code.

Concept	**Description**
Environment	Used to create a new simulation environment. All events and processes take place in this simulation base that is empty by itself
Resource	Used as a container with a specific capacity that processes can request and release
PriorityResource	A subclass of resource that allows for prioritization of requests to the resource
Process	Used to represent an ongoing activity in the simulation
Timeout	Used to represent a waiting time or a delay in the simulation
Request	Used to represent a request for a resource
Yield	Used to suspend a process for a time until an event happens

The constants in the simulation are number of agents, simulation time, ticket arrival rate and service time mean. Additionally, distributions from the real-world data are represented in priority distribution and SLA distribution. The simulation will collect several metrics that describe the ticket handling process from different viewpoints but the most important one is a metric called “SLA compliance”, which directly links the simulation result to the KPI metric used by the team to measure daily service desk performance. Other main metrics collected, and their exact descriptions are discussed in table 11. The objective is to gather information regarding the performance of all the metrics so that a better comparison between the disciplines can be realized. For example, even if the SLA compliance is high, it may be crucial to know why or how efficiently the target has been met.

Table 11: Metrics collected during the simulation

Metric	Description
SLA-compliance	Percentage of tickets that were handled under service level time. They are computed as the total time spent (waiting time+service time) compared to the SLA time. It is also calculated for each ticket service level
Throughput time	The total time a ticket spends in the ticketing system from arrival to departure
Waiting time	The total time a ticket spends in the ticketing system queue before it is assigned to an agent
Overall agent utilization	Average utilization rate of agents, which is calculated by total service time divided by the total available time for the agents
Total handled	The total number of tickets that were handled by the agents after simulation time has ended and the queue is empty

The simulation contains several functions that allow for utilizing the SimPy library and a creation of discrete event simulation model. First, *generate_tickets* function is used as a process to generate the tickets that will arrive to the ticketing system. Each ticket has attributes “id”, “priority”, “sla”, “sla_time” and “arrival time”. Also, interarrival times are generated according to the exponential distribution as discussed. *Ticket_generator* function is used to introduce the tickets to the simulation environment based on their interarrival times. *Ticket_handling* function is used to model the ticket handling of an agent when a ticket is received. Service time is set to follow an exponential distribution. The function also contains the logic for each queue discipline to sort the queue. In this function, waiting time, service time and departure time are recorded.

To gather simulation data, print_results function is defined. The metrics in table 10 are printed in this function and added to DataFrame for storing data. DataFrames are created elsewhere in the simulation, and details of each ticket and metrics data are then stored in print_results function. The main part of the simulation is the run_simulation function that takes queue discipline as an argument and runs the simulation environment for all queue disciplines distinctly. The simulation is executed for all disciplines by using a for loop that iterates over them. Generated tickets list created from *generate_tickets* stay the same for each discipline meaning they are using the same set of tickets within one simulation to allow for direct comparison. At the end of the simulation, metrics data and tickets data stored in pandas DataFrames are added to CSV file for further analysis. The flow and operations of the simulation is depicted in figure 10.

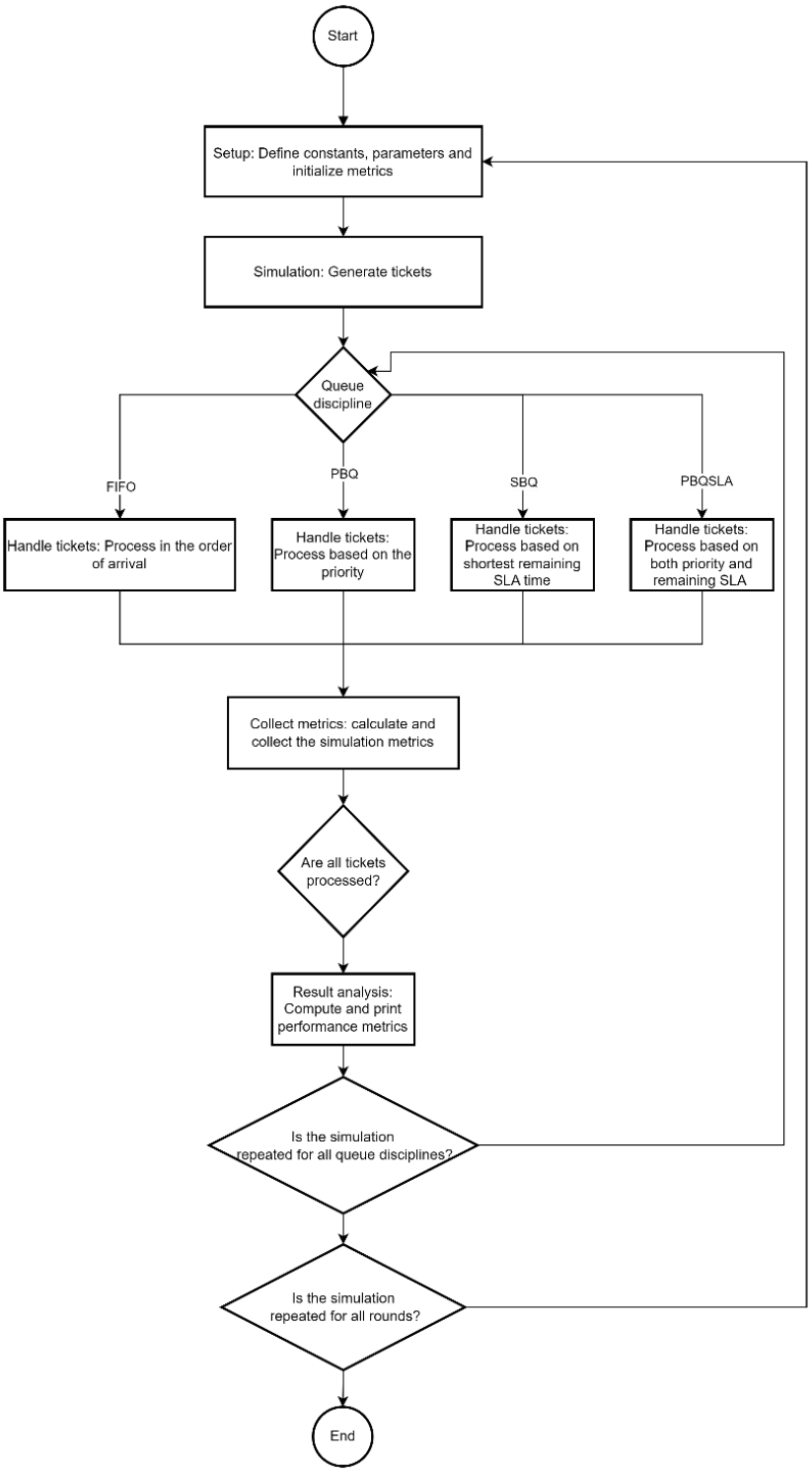

Figure 10: Flowchart of the simulation

In the formulation of this simulation model, several assumptions were made with the objective of gaining a sufficient balance between detail and clarity. The assumptions are as follows:

1. The service desk agents operate identically meaning they have the same skill levels, speed, availability and handling capacity. They also cannot handle multiple tickets at the same time.
2. The simulation assumes that interarrival times and service times are exponentially distributed to follow multi-server queuing model M/M/c. However, it is important to note that the model is only followed in FIFO discipline since the M/M/c model assumes for it to be the service discipline.
3. The simulation assumes that no breaks happen. The simulation does run for a fixed period meaning there are no shift changes, non-business hours or downtime.
4. The rate of arrival of tickets is considered constant with no fluctuations. It does not this take account factors like time of day or seasonal fluctuations.
5. The tickets are handled from start to finish during the service time, which means that a realistic scenario where for example, tickets status is changed and during that time external work is done to fix the cause of the incident, is not modelled. The resolution rate of tickets is also assumed to be 100% so no incidents fail to be resolved.
6. There is no change in priority once the ticket has been created, which could be possible in real-life scenarios to manage the queue more effectively.
7. The simulation ends when the simulation time has ended and remaining tickets in the queue have been handled.

These assumptions also present potential areas of improvement of the simulation model if more complexity needs to be added. However, a model that attempts to capture every detail of a real-world system could become too complex making it less tractable and less understandable. Thus, assumptions are critical to balance the detail and comprehensibility of the model. It also ensures that the simulation model is focusing on the most crucial aspects instead of carefully modeling aspects that are not related to what the simulation is trying to achieve.

5 Results and analysis

In this chapter, the simulation model is run, and its output carefully analysed to gain an understanding of how different queue disciplines perform against each other in two different scenarios. First, simulation results are presented with supportive visualizations. The results are further analysed and a new optimized process for incident management ticketing system is recommended. In the last part of the chapter, verification and validation steps are discussed to achieve better reliability of the simulation model results and its proper utilization to the real-life process.

5.1 Simulation results

After creating a conceptual model and simulation environment, the next step is to run set simulations, gather results and analyse them by using simulation metrics that come as a simulation output. As the simulation now functions and follows the same set of rules as the conceptual or theoretical model, it will be further validated through many repetitions between different seed numbers to decrease the statistical error and random nature of results. What-if analysis is used to represent various simulation scenarios where input values are changed according to our objectives of finding an optimal queue discipline. This way, the comparison between disciplines will be more suited to predict real-world service desk because better captures the inherent randomness and variability of the real service desk ticketing system.

Using a what-if analysis as comparative analysis, the simulation scenarios are created to represent accurate depictions of what could be realistic in the service desk. For example, by changing a ticket load within ticket arrival rate input, the queue disciplines can be tested under peak periods, off-peak periods, and normal periods. The chapter is thus compartmentalized between different service desk scenarios, analysing the set scenarios and after combining the learned information a potential way to optimize the performance of the real-world service desk is recommended. This will largely involve the chosen queue discipline because the selection could also bring requirements to the system and process used. For example, baseline FIFO is a simple way to sort the queue and it does not require

as extensive information compared to SLA or priority disciplines where it is essential to have the needed data in the system to allow for correct queuing.

For each scenario, a total of 100 different simulations were run to gather comprehensive data about the performance of disciplines with different seed numbers. This is done to increase the representability of the data. The seed numbers will affect the arrival time and service time, which means there will be a small difference in number of tickets handled between each round and in service times that also vary according to exponential distribution. The results of the 100 simulation rounds are collected to a separate file, which allows for seamless analysis of the performance between queue disciplines.

Scenario 1:

In the first scenario, the chosen parameters are simulating a balanced situation where the priority distribution, sla distribution and service_time_mean follow the real-world data. The ticket arrival rate is set as four tickets per hour, which is also representative to the real service desk. Seven agents are chosen as servers; this agent quantity is within the limits of available agents for the Telia Wholesale incident management service desk. For simulation time, 23 days are chosen to simulate how tickets are handled in one month timeframe. These input parameters are depicted in Table 12.

Table 12: Simulation scenario 1 input parameters

Simulation parameter/constant	**Value**
NUM_AGENTS	7
SIMULATION_TIME	23 days
TICKET_ARRIVAL_RATE	4 tickets per hour
SERVICE_TIME_MEAN	92.7 minutes

The collective results are combined to table 13. From the first scenario, all disciplines perform quite similarly in the overall sla compliance metric, which means that there seems to be no substantial difference in how well the tickets are resolved before the SLA time is breached between the disciplines. However, SLA-based queue (SBQ) has the best SLA compliance rate, which is near 100% in this specific scenario. It has also great compliance rate for all priority levels as opposed to priority-based queue and performs better than baseline FIFO in all priority levels. From this scenario it is also interesting to notice that FIFO performs quite well and has the second-best overall SLA compliance rate while also

other metrics competing well against other disciplines. Agent utilization metric is the same for all queue disciplines and indicates that there are some idle times for the agents in all disciplines. In other words, there are sufficient number of agents handling the tickets causing high compliance rate.

Table 13: Summary of simulation results for Scenario 1

Queue Discipline	Overall SLA Compliance	Overall Avg Throughput Time	Total Handled	Overall Avg Waiting Time	Agent Utilization	SLA Compliance P1	SLA Compliance P2	SLA Compliance P3	SLA Compliance P4
FIFO	94.12	172.88	2196.75	79.9	0.87	94.03	94.02	94.17	94.11
PBQ	93.76	158.3	2196.75	65.82	0.87	99	98.88	98.71	86.65
PBQSLA	93.35	168.07	2196.75	75.38	0.87	99.02	98.79	98.71	85.73
SBQ	96.29	167.08	2196.75	74.22	0.87	96.33	96.12	96.28	96.34

Since the simulation has been run 100 times, there is also some degree of variability in the results, which cannot be seen from the table above. Also, while the SLA compliance rate provides a great metric for reviewing incident management performance, it is also crucial to delve deeper in the results for each ticket priority level. Since the tickets are recorded to the queue with a priority and SLA, they are not equal, which is why other types of disciplines are compared to FIFO. For this reason, both the variability and the priority level are plotted to figure 11, which portrays how well each discipline handles tickets by priority over the course of the 100 simulation runs.

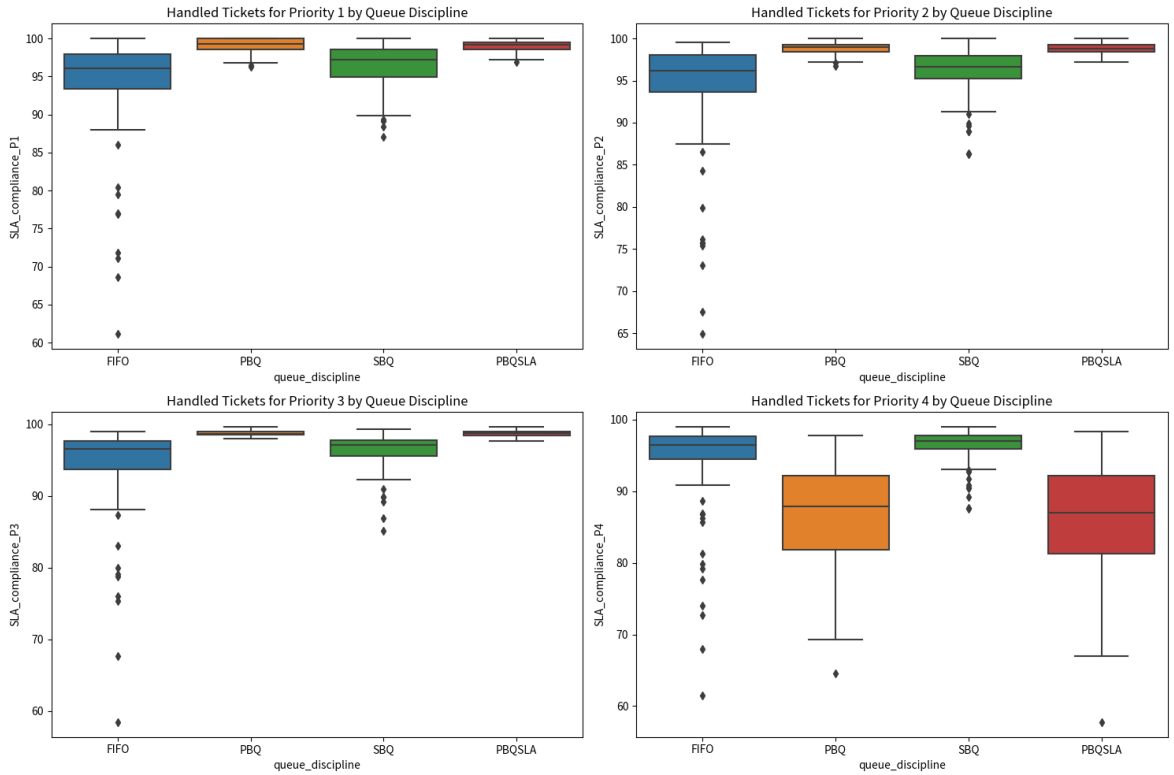

Figure 11: SLA for each priority in the first Scenario

The box plot clearly visualizes the distribution of the data over the course of 100 simulations for each priority level and for each queue discipline. Priority 1 tickets are the most critical and Priority 4 tickets the lowest. The data spread is the highest in FIFO for all priority levels, which means that there is the most variability in FIFO within different simulation runs. The strength of FIFO in considering each ticket as equal is present in the plot since compliance rates for all priority levels are quite high even though there are some outliers. The strengths of priority queues can also be clearly seen from the plot since both the priority-based queue and priority-sla-based queue perform the best P1, P2 and P3 tickets while being considerably worse in handling P4 tickets.

Next, to get more insights on the performance of the service desk and factors that may affect it, correlation analysis is made to help understand which factors are most influential in allowing high SLA compliance while also identifying relationships between the metrics. The colour of each cell represents the strength and direction of the correlation, in which dark blue means strong negative correlation and dark red strong positive correlation. Correlation matrix between the average output metrics of the simulation is presented in figure 12.

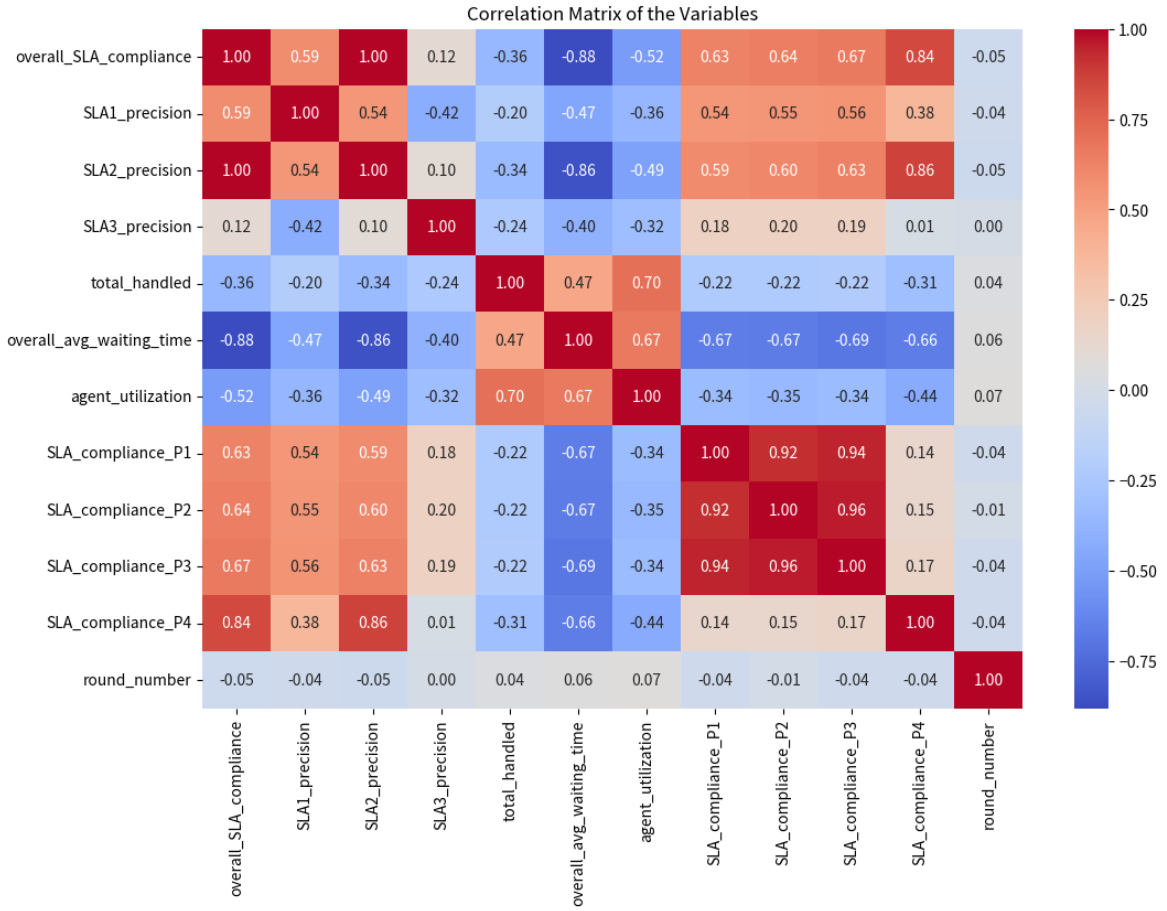

Figure 12: Correlation between the simulation metrics in Scenario 1

Strong positive correlations can be interpreted between overall SLA compliance and compliance at each priority level. This suggests that higher SLA compliance at each priority level constitutes to a higher overall SLA compliance, which is expected since the overall number is the weighted average of the SLA compliance at each priority level. There is also an interesting correlation between waiting time and SLA compliance. The strong negative correlation indicates that as the waiting time increases, the overall SLA compliance decreases. Looking at agent utilization as the other important metric, there seems to be a strong positive correlation between it and number of total handled tickets. This is also expected since the higher number of tickets requires more work from the service desk agents.

Scenario 2:

In the second scenario, the main difference is that the ticket arrival rate has been increased to six tickets incoming in an hour. This is considered to simulate a busy service desk where tickets will accumulate, and the queue discipline should then have a bigger role in deciding

what tickets to handle. The effect is obvious when the average service time of 92.7 minutes is compared to six tickets per hour ticket load with seven agents handling the tickets. Agent utilization then exceeds hundred percent and service desk operates beyond its capacity unlike in the first scenario. This will cause a situation where the agents will not be able to handle all tickets as they come but require a method for choosing a ticket from the list. The simulation parameters are presented in table 14.

Table 14: Simulation parameters for scenario 2

Simulation parameter/constant	**Value**
NUM_AGENTS	7
SIMULATION_TIME	23 days
TICKET_ARRIVAL_RATE	6 tickets per hour (2 ticket increase)
SERVICE_TIME_MEAN	92.7 minutes

As the results from table 15 suggests, overall average waiting time greatly increases with this scenario meaning that the agents are not able to keep up with the ticket load as well as previously. Agent utilization has also increased to almost hundred percent, which implies that the agents have not had any idle time during the simulations. This is naturally not practically possible for the service desk in real life since variables such as shift changes and breaks need to be considered. In addition to overall average waiting time, throughput time also portrays the struggle of handling the tickets as they come and within its SLA limits. They indicate that as the ticket is recorded to the queue, there are too many other tickets in the queue for the new ticket to be serviced quickly, which increases both wait time and throughput time containing the time between ticket departure and ticket arrival.

Table 15: Summary results for Scenario 2

Queue Discipline	**Overall SLA Compliance**	**Overall Avg Throughput Time**	**Overall Avg Waiting Time**	**Agent Utilization**	**SLA Compliance P1**	**SLA Compliance P2**	**SLA Compliance P3**	**SLA Compliance P4**
FIFO	3.92	5365.12	5272.40	0.995	4.04	3.96	3.95	3.86
PBQ	57.65	5364.49	5271.62	0.995	98.90	98.76	96.23	1.85
PBQSLA	58.94	5306.37	5213.72	0.995	98.87	98.88	96.65	4.55
SBQ	9.37	5369.09	5276.29	0.995	9.28	9.51	9.31	9.40

As a result, there can be seen significant changes in how the queue disciplines perform against each other. First looking at the overall SLA compliance, FIFO as a baseline has the worst performance with only under four percent of tickets handled within set SLA time indicating that it has considerable problems with keeping up with the ticket load. Interestingly, SLA-discipline, which in previous scenario was one of if not the most effective discipline, performs almost as bad as FIFO. SLA compliance percent of 9,4% indicates that SLA-queue fails to meet its goals as the number of tickets increases. However, priority-based approaches have the best performance by a significant margin. Both PBQ and PBQSLA have similar average SLA compliance, where an overall of over 50% of tickets are handled within SLA with little variance between simulation rounds. Overall waiting time and average throughput time are quite similar across all disciplines since they process the same sets of tickets, and the selection of a discipline does not affect service time.

The main cause for the better performance of priority-based approaches can be clearly seen from figure 13. There are four graphs each representing a different priority level. In the first scenario, differences were not as evident as in the second scenario. For critical priority tickets, SLA compliance is almost hundred percent across all simulation rounds for priority disciplines, which makes the box plot have only little variance. Almost the same can be figured out from the next two priorities: high and medium. In these two cases, there are a bit more variance across simulation rounds, but they still are near hundred precent SLA compliance levels. As in the first scenario, in this case low priority tickets are handled poorly by priority approaches and SLA based queue handles these tickets multiple times better within SLA limits even though all four have poor performance. Especially priority-based queue has exceptionally bad low priority SLA compliance performance and ranks last across the disciplines.

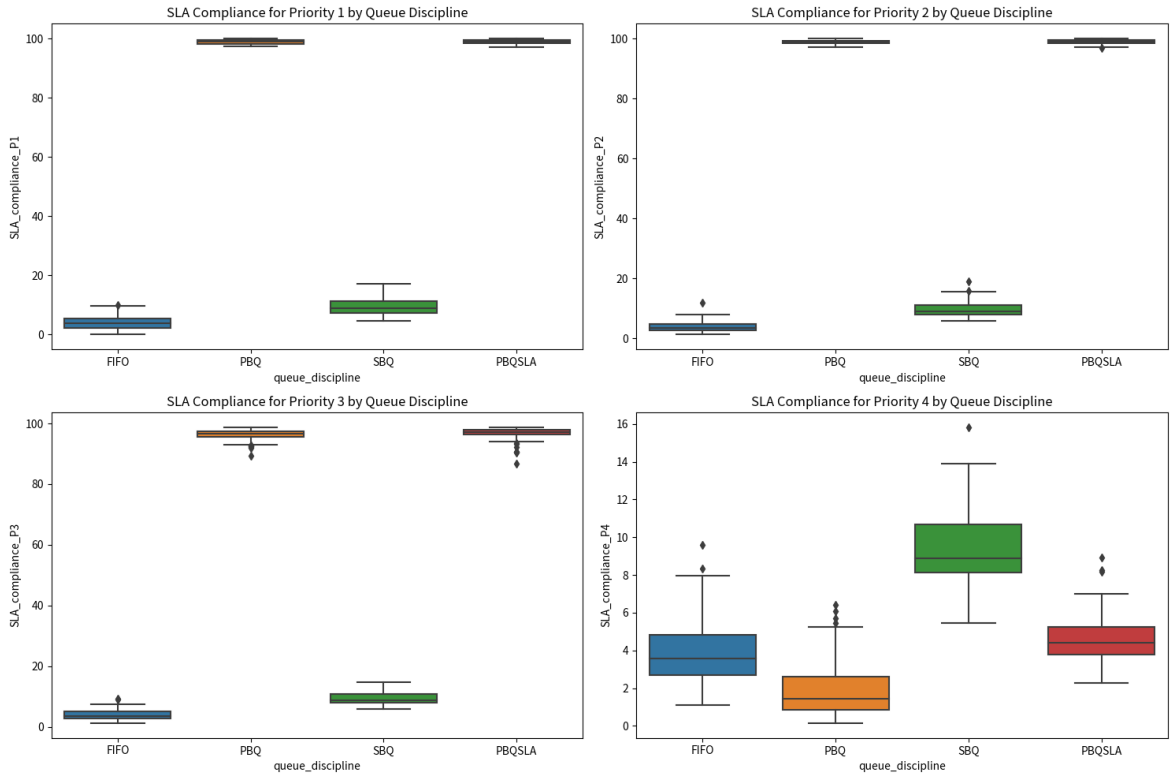

Figure 13: SLA compliance rates for each priority in Scenario 2

Correlation matrix in figure 14 corresponds to the same insights made from other visualizations and contains similarities to the first scenario. For example, SLA compliance metrics across all priority levels have a strong positive correlation to the overall SLA compliance rate and the average waiting time is negatively correlated to overall compliance rate suggesting inefficiencies as the SLA compliance rate decreases. The first difference is in the agent utilization that is not depended on the total number of tickets since all simulation rounds have a high-ticket load with almost no idle time for agents. Behavior difference between scenarios can also be spotted from SLA compliance of P4 tickets. Unlike other priority tickets, P4 tickets have negative correlation to overall SLA compliance suggesting that when the system performs well, it is at the expense of low-priority tasks. This further strengthens the idea of priority-based approaches as the optimal queue discipline for high ticket load scenarios.

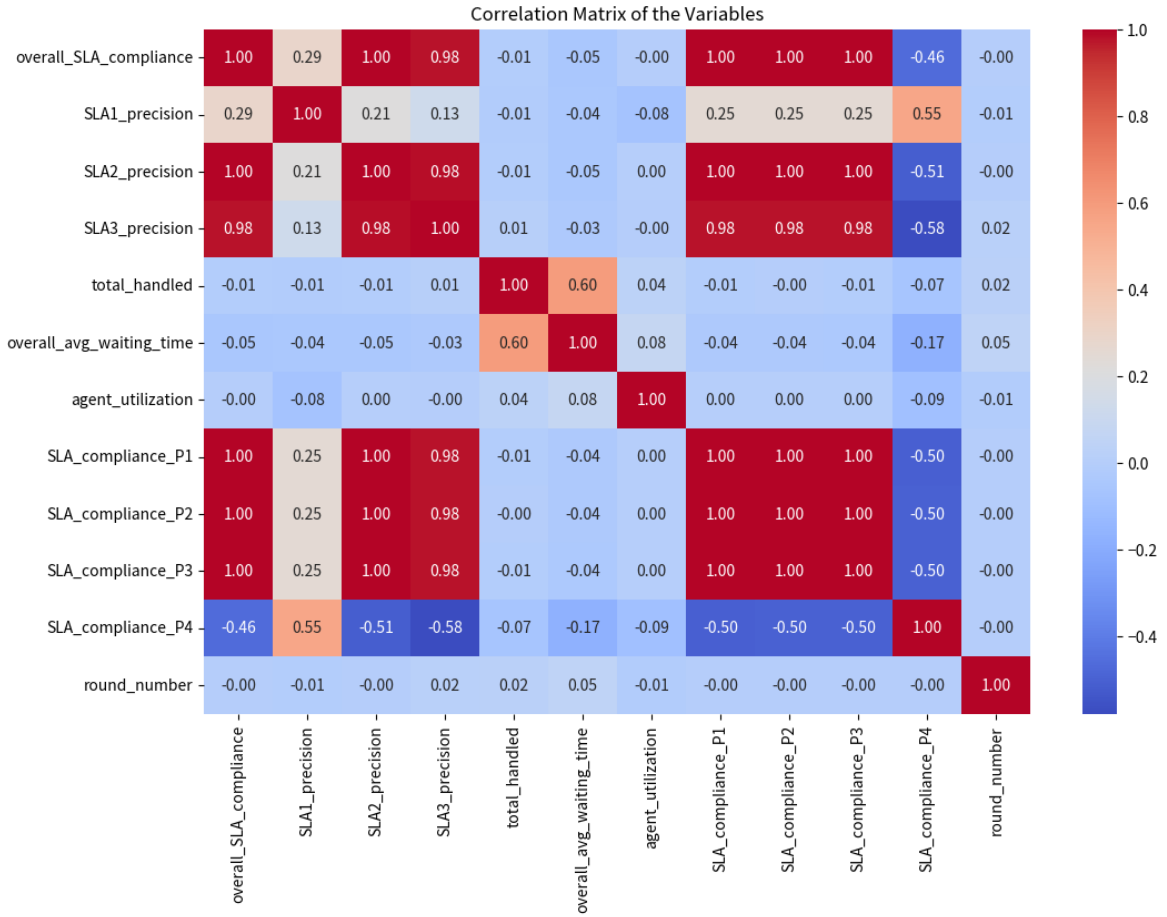

Figure 14: Correlation matrix for Scenario 2

5.2 Results analysis

Incident management service desk was modelled as a queuing model and tested with a discrete event simulation to obtain insights on how the queue should be sorted to achieve the best performing queue that allows for tickets to stay within set SLA limits. For this reason, two different scenarios were introduced in the previous chapter with a minor difference in ticket arrival rate, which affected the results considerably. With both scenarios ran hundred times, the results can be confidently examined to identify the optimal way to optimizing the current incident management service desk, which uses FIFO discipline without proper utilization of SLAs or ticket priorities.

To understand the chosen scenarios more, the nature of the service desk should be further discussed. In Telia Wholesale incident management service desk, the arrival of tickets differs greatly depending on the day, week, or a season. For example, external events such as nature

catastrophes can affect the service desk so that sometimes the number of incident tickets in the queue is far too great for agents to handle them as they come. In some scenarios, the number of incident tickets is far lower, and the service desk works with other service requests that they are also obligated to handle. Changing ticket arrival rate allowed for the testing of how the disciplines perform in different conditions so that an overall best queue discipline could be chosen.

For the first scenario, the results were more closely matched since agents were not over-utilized and could start servicing the tickets quite quickly. This caused a situation where FIFO and SBQ performed well while providing balanced results for the SLA compliance with also lower priority tickets were handled with high compliance rates whereas priority-based approaches had a bit lower compliance since the compliance rate for the lower priority tickets were not on par. For the second scenario however, the selection of a queue discipline affected the outcome greatly even though all disciplines handle the same tickets over the course of the simulation. In other words, the order the tickets were picked from the queue had a strong influence on the results.

More in-depth results were also collected from the simulation runs to get a better picture of how different SLA categories and priority levels were performing within each queue discipline. Table 16 focuses on how average waiting time differs based on both queue discipline and priority level. The table also summarizes results from the two scenarios. FIFO and SBQ work as expected by having a balanced performance over all different priorities while the difference between P1 and P4 tickets in priority-based approaches is substantial in both scenarios. Instead of having a fair and balanced ticket handling mechanism, the average waiting time for these two disciplines starts to slowly increase as the ticket priority decreases until P4 tickets, where the waiting time are a lot longer than FIFO or SBQ has.

Table 16: Average waiting time for each priority level

Queue Discipline	Avg Waiting Time P1 (S2)	Avg Waiting Time P2 (S2)	Avg Waiting Time P3 (S2)	Avg Waiting Time P4 (S2)	Avg Waiting Time P1 (S1)	Avg Waiting Time P2 (S1)	Avg Waiting Time P3 (S1)	Avg Waiting Time P4 (S1)
FIFO	5320.91	5267.65	5270.93	5267.79	78.63	81.27	79.68	79.97
PBQ	14.42	18.95	76.62	12664.79	9.17	11.02	21.14	133.72
PBQSLA	14.31	18.81	76.13	12525.31	9.26	11.21	22.05	155.54
SBQ	5330.01	5263.31	5279.35	5267.69	75.97	75.54	74.22	73.63

One of the reasons to good performance even for high load ticket scenarios for priority-based disciplines is the chosen distribution for priorities. While evident that P4 tickets are left to the queue until there are no other priority tickets available, this priority level takes up over 40% of all tickets because of the distribution used. This leaves more resources to the remaining number of tickets not classified as P4. The agents are then able to better resolve these tickets within SLA limits before continuing to handle the rest of the P4 tickets where SLA is likely already breached.

This is also where minor differences between PBQ and PBQSLA can be found. The wait time for lower priorities is somewhat lower in PBQSLA, which considers remaining SLA time inside all priorities including P4. This causes a situation where if there are no other priorities in the queue anymore, remaining tickets are sorted based on their remaining SLA. As a result, there is a slightly better chance of successfully handling the ticket before SLA is breached as opposed to FIFO, which just chooses the oldest ticket from the queue without taking account the three service levels used in the model (SLA1, SLA2, SLA3). As a result, overall SLA compliance for priority 4 tickets is 4.55% for PBQSLA and 1.85% for PBQ in scenario 2. There are no considerable differences in the first scenario.

Delving more into insights from the output data, also compliance and precision for each service level was calculated. SLA precision for specific service level is defined as how many percentages of incidents are resolved within SLA limits. SLA compliance for each service level indicates how incidents are resolved within SLA limits and distributed among different SLA types. The results indicate the significance of the SLA distribution chosen; SLA2

tickets have been completed the most, which can be seen from SLA2 compliance column from table 17. SLA2 tickets make up most incoming tickets, which makes this behavior expected.

Table 17: SLA specific simulation results

Queue Discipline	Simulation scenario	Overall SLA Compliance (%)	SLA1 Compliance (%)	SLA2 Compliance (%)	SLA3 Compliance (%)	SLA1 Precision (%)	SLA2 Precision (%)	SLA3 Precision (%)
FIFO	1	94.12	4.52	86.62	2.98	75.31	95.2	98.94
PBQ	1	93.76	4.8	85.99	2.97	79.96	94.5	98.57
PBQSLA	1	93.35	5.26	85.4	2.7	87.72	93.86	89.49
SBQ	1	96.29	5.47	88.48	2.34	91.2	97.24	77.37
FIFO	2	3.92	0.10	3.64	0.18	1.60	4.01	5.93
PBQ	2	57.65	2.84	53.00	1.81	47.34	58.27	59.53
PBQSLA	2	58.94	4.43	52.86	1.65	73.89	58.10	54.50
SBQ	2	9.37	5.46	3.88	0.03	91.04	4.27	0.97

An interesting phenomenon with SBQ is its SLA level specific performance. In the first scenario, SLA2 tickets with 8-hour SLA time have the best precision and SLA3 with 12-hour time the worst even though SBQ is not specifically linked to initial SLA level. It equally chooses between the ticket that has the lowest remaining SLA time at that moment. The difference is substantial in scenario 2, in which 4-hour limit SLA1 tickets are resolved with over 90% accuracy compared to drastically lower precision in other SLA levels. One of the reasons is that SLA1 tickets reach their SLA limit twice as fast as the second one. If SLA1 labelled tickets would have been set as the most important incident tickets, utilizing SBQ would be sufficient. In other situations, it will fail to deliver good results in high ticket load and instead starts to behave similarly as ineffective FIFO.

All in all, two scenarios were crucial to get a more comprehensive picture of the queue disciplines since it allowed to identify the strengths and weaknesses of the disciplines in different scenarios. When agent utilization and waiting time is low, there are no drastic differences in which queue discipline to choose from even though SBQ seems to bring the best results by a small margin. On the contrary, when agent utilization and waiting time indicate a busy queue, priority-based approaches perform the best to at least solve the tickets that are the most important. From the two approaches, PBQSLA is the recommended approach and based on the simulation model, the optimal approach for sorting tickets in a queue.

5.3 Verification and validation of the results

Verification and validation procedures has been used throughout making of the simulation model starting from the problem description to checking simulation output results. When the conceptual model that had been validated by subject matter experts was translated into a computer readable form, several verification steps were taken. The code verification was done with structured code reviews where the behavior of the code was checked line by line with another subject matter expert. To verify that M/M/c model was implemented right, in addition to metrics data, also ticket data was added to Pandas DataFrame and exported to a separate file where it was easy to identify that the inter-arrival of tickets and service times correctly followed exponential distribution.

To validate the model outputs, several methods were used to help ensure the reliability and accuracy of the results. The model outputs were reviewed with a project team of knowledgeable subject matter experts throughout making and testing of the model, in which feedback was given whether the output made sense in the real system. A common approach for validation, historical data validation was a clear limitation of the verification and validation step since there are no sufficient data collected from the current process regarding especially real service times.

Sensitivity analysis was conducted to review various simulation setups and their outputs. Input parameters of the model were changed in the queue to view if the outputs were changed reasonably. For example, increasing the number of agents had a clear effect on the collected metrics by decreasing agent utilization and increasing SLA compliance. Another example

was the modification of simulation time where one day simulation resulted in more variability across different simulation runs compared to the 23 days simulation in both scenarios. This parameter sensitivity analysis helped to identify how sensitive results were to change of simulation parameters.

The variability of the results for both scenarios are portrayed in table 18. Since each scenario was run hundred times with different random seed values, some variability in the simulation output is expected because the seed changes the inter-arrival and service time of the tickets around their mean values. The issue with large changes with various seeds is that the results may then be too sensitive to random sequences by different seeds, which could decrease the reliability of the model.

Table 18: Sensitivity analysis for variations between simulation runs in SLA compliance metric.

Queue Discipline	**SLA_compliance_std_Scenario1**	**SLA_compliance_std_Scenario2**
FIFO	6.67	1.56
PBQ	3.03	1.07
PBQSLA	3.40	1.13
SBQ	2.41	1.86

However, as table 18 suggests, the standard deviation of SLA compliance in both scenarios is small given that the SLA compliance is on a scale up to hundred. The most variability can be found in FIFO in the first scenario, which makes sense since its first come-first-serve logic is affected by the change of inter-arrival times between runs. This is not as evident in the second scenario since FIFO performs bad with low SLA compliance in each simulation run.

6 Discussion and conclusions

In the previous chapter, results of the discrete event simulation were collected and then analysed. Based on the simulation results for M/M/c model with real-world related parameters, the optimal approach for sorting queues is to use a priority-based approach that will first sort the queue based on given priorities and then use remaining SLA time to choose between same priority incidents. With this approach, both SLA level and priority level chosen at the beginning of the incident management process are utilized to allocate workload to service desk staff. For the ITSM system, the results indicate that data-driven approach is crucial to optimizing incident management service desk. If sorting is not done by the system itself, that responsibility is in the hands of the service desk staff, which makes the service desk more inefficient since extra time is required for manually sorting through the ticket list.

ITIL incident management has enabled many activities in the simulation model and is thus required for the service desk as a framework to adapt concepts and vocabulary from. An important part of the ITIL incident management is to especially focus on the beginning of the incident management by correctly categorizing and describing the incident. This involves determining both impact and urgency of the incident. Priority rating that is calculated as a result is then used in allocating incident ticket to right person at the right time. Having other incident details also helps with resolving incidents as quickly as possible, especially with ITIL and its known error database approach where previous incidents and their resolutions are referred.

A potential problem with ITIL and its priority matrix is that strict guidelines are needed to ensure that the impact of the incident is recorded correctly. For example, the viewpoint in determining severity can vary based on if it is looked at from service owner's perspective compared to service providers. For example, many users affected, and a strict SLA time may be factors for service providers to prioritize the incident high but from service owners' perspective, a smaller incident in key functionality of their operations could have higher impact than a bigger disruption to service with not as much significance. Challenging aspect is also the difference in scale of operations between customers of the telecom companies, in which smaller customers may have higher relative impact for their incidents compared to large service owners even though the incident affects more users. This requires detailed

guidelines and strategic decisions to ensure that service desk agents can use the same methodology in determining incident impact.

Some ITSM software such as ServiceNow allow for major incidents to route through another process, where incidents are handled with higher priority while enabling better co-operation. These types of incidents should thus be out of scope from the recommended process with PBQSLA and handled separately to ensure their special requirements are met. Other instances of incidents not included in major incident management process can be recorded with impact and urgency, where urgency would largely consist of SLA time of the service. Determining SLA levels for different services is crucial and incident management managers should consider implementing a variety of metrics to help better understand the performance of service desk.

As evident from literature review, there are already several approaches of tracking metrics and managing report sheets for service level agreements. The data would not be just used with reporting KPIs monthly but to understand the root causes of a specific performance of service desk. For example, in the empirical section the simulation model collects SLA compliance as the most vital metric but does not stop by just categorizing all incidents inside one metric; it establishes separate metrics for each priority level and each SLA level so that the differentiation would allow a more precise tracking of incidents and areas for improvement can be better understood. Another example is to report SLA levels based on a specific product to identify if certain products or services are struggling to meet their SLA levels. Another approach for the service desk commonly used in more automated environments is to track different phases of the process to find possible inefficiencies.

There are currently no sufficient data collected from the different phases of the Telia Wholesale incident management process, which is why these recommendations should provide considerable value for the service desk to understand their operations better. This should also allow for correct reporting of service time that is related to just the incident management service desk agents and tracks other work separately. This in turn, would enable even more effective service desk simulations that could be used for staff planning and trying out queue disciplines in different scenarios. In real life, the target company also uses the service desk for service requests, which should be considered in future simulations, especially if staff or capacity planning is the goal.

Regarding customer satisfaction, which is exceedingly important for the incident management service desk, a good approach is to provide transparency with the help of status notifications when the incident is progressing, especially since incidents may need external work and coordination. With the help of implementing ITIL approaches and using an ITSM system that can provide needed platform for managing the incidents, statuses could be added to incidents. For instance, in-progress status may mean that an agent is currently servicing the ticket and waiting on user status may indicate that information is required from the customer to proceed with the incident resolution steps. After certain actions, these statuses would change, and automatic status notifications could be sent to the customer while the ticket is being updated in each step by the staff working on the incident. This relates to the idea that documenting is vital throughout ITIL incident management process.

From the simulation model, it became clear that priority-based approaches perform well in incident management, where incidents vary in their impact and urgency. This way, resources are always focused on the most important work, which is crucial especially in telecom incident management environment, where the incidents have a possibility of largely disturbing the operations of the service owner. A good way to further test queue disciplines and focus only on the high priority tickets is to present a pre-emptive priority approach, where the handling of lower priority tickets is stopped as soon as the bigger priority ticket is recorded to the queue and switched to the higher priority work. For instance, in scenarios where SLA levels are set strictly compared to the mean time of incident resolution, pre-emptive priority would allow for instantaneous servicing of the incident ticket.

6.1 Answering the research questions

The main purpose of this thesis was to provide an optimized way to handle tickets for an incident management team. Based on the theoretical and empirical findings, recommendations for the team were made regarding service discipline used by their ticketing system. Another recommendation was to better utilize ITIL framework to make the service desk more efficient and to enable functionalities of the chosen service discipline. These were made with the help of the research questions, which are now answered.

- **How can the incident management process of a telecommunications company be optimized to handle tickets within the agreed SLA of the specific customer?**

As discussed in the theoretical chapters, there is not only one correct way to optimize incident management operations in telecommunications companies, which has resulted in several methods to help optimize the current performance of incident management (IM) team based on the specifications and requirements of a specific service desk. However, an important finding is to first understand how IM currently functions meaning that it should be clear if there are various incident categories, different skills among staff, kinds of existing service levels and set key performance indicators. A common approach for incident management optimization has been to switch between sorting the queue with FIFO approach to a more priority-based approach since not all incidents are equal with their impact and urgency. There are multiple methods of designing a priority-based queue sorting; it can be sorted as multiple queues with different priorities, priority-FIFO queue, priority-SLA queue, pre-emptive priority queue or some other similar approach.

Findings from empirical section of this thesis also concluded that priority queue performs better overall compared to FIFO or SLA queue sorting. This has been also evident from previous research, which this thesis further strengthens. The optimal strategy for optimizing IM process for Telia Wholesale according to the simulation results is to use PBQSLA queue discipline, in which tickets are first sorted based on given priorities and then using remaining SLA time to choose between incidents of the same priority. Especially remaining SLA time sort within same priority is an effective way to leverage both priority and agreed SLA of specific service. In another words, Service Level Agreement of a specific customer will greatly influence the ticket handling, which will help allocate service desk agents to most important incident tickets before sorting the least important ones. This maximises the SLA compliance metric for high priority tickets and still tries to resolve lower priority tickets by choosing a ticket with the least remaining amount of SLA time.

Implementing PBQSLA queue discipline to an ITSM system is not at itself possible without the use of ITIL principles and appropriate data. First, priority rating of the ticket must be reliable because it is used by the discipline. When ticket is created to the system, it should contain priority rating and SLA level. SLA comes from the initial SLA of the service that is down and priority rating is calculated through priority matrix. In priority matrix, an agent

chooses an impact for the incident and urgency can be either chosen or SLA level can be too used. Implementing ITIL incident management will thus bring clarity to the process, enable usage of PBQSLA and it also provides common terminology for handling incidents.

In addition to queue discipline chosen, attention should be also put to a data-driven approach to better optimize incident management process. Metrics and reports for SLA adherence have a key role of monitoring IM operations and gathering insights on what areas need the most improvements. Metrics should be tracked for separate priorities and SLA levels to get a better picture what kinds of tickets are lacking in performance. Also, metrics should also be collected from specific process parts to get feedback on for example which phase of IM is the bottleneck of the process. These metrics could be presented with an SLA result sheet that should be monitored for continuous improvement.

Overall, this thesis should give clear guidelines for the target company and other telecoms to optimize their current incident management process and help make decisions when switching to a different ITSM system that allow for a more comprehensive managing of incident tickets. As concrete actions for ticket handling within set SLA, a change of queue discipline will have a major impact on SLA compliance as main KPI. ITIL as framework will support this modification and better data collection from the process will enable the company to swiftly respond to the insights learned using metrics. Handling the incidents inside agreed service time is essential to maximise customer satisfaction and to ensure that the downtime from the incidents is as short as possible.

- **Can queuing theory and simulation be used to identify the most optimal set of prioritization rules for tickets in a telecommunications company's ITSM system?**

This thesis succeeded in providing a great example of using queuing theory approaches together with simulation to provide a set of prioritization rules for a telecom company. By familiarizing queuing theory concepts and Kendall’s notation, M/M/c queuing model was chosen with the help of previous academic research from the topic. M/M/c model assumes Poisson arrival processes and exponential distribution service times with a set number of servers. This model was deemed appropriate for the target company and its service desk by analysing the nature of the service desk. This was done by utilizing real-world ticketing data

from the company's ticketing system. The data was historical data from the service desk consisting of previously handled tickets.

Queuing theory by itself provides a good way to solve queuing problems, which also includes service desk that handles incident management tickets. Calculation of utilization or using Little's law for example can be used as good indicators of how the service desk is performing. It allows for swift calculations that can describe the queue, but limitations of this approach are also clear. First, it relies on several assumptions for instance regarding average service rates and inter-arrival times. Incident management operations in real life seldomly have stable rate of incidents arriving and a specific service time for all incidents.

Complimenting queuing theory with simulation, however, provides a way to introduce a more complex conceptual model that takes account more of the inner workings of the service desk. Discrete event simulation was used as the simulation model, which was valuable in simulating a system changing in discrete points of time when certain events happen allowing for capturing complexities of the IM service desk. In the simulation, the queuing model were made more realistic with using distributions of priorities and SLAs from the real service desk together with capable assumptions to service and arrival times.

Simulation and queuing theory have proved themselves capable of giving valuable insights regarding how queue disciplines handle tickets without the need to test these all at production environment, where even small delays in ticket resolution can have major impact on operations of the customer. They have too only scratched surface on what kind of optimizations are possible with the simulation model built for this thesis. Since parameters can be easily modified for simulation runs, it allows for testing multiple components of service desk such as determining optimal SLA levels in addition to testing different queue disciplines made in this thesis. As the environment in real-life operations becomes more complex and more variable, effectivity of simulation models and queuing theory decreases, and it is always important to keep in mind the limitations of the simulation model.

- **How can real-world data from the ticketing system improve the incident management process in telecommunications?**

Whilst discrete event simulation in addition to queuing theory have provided potential improvement suggestions to the existing process, it would not have been possible without the use of real-world data from the ticketing system. Since the target of queuing modelling

and simulation experiments is to model and test an existing system, in-depth information regarding the system in production is essential. System modelling, which queuing theory and simulation are subsets of, requires parameters from the real-world system to accurately represent the system and analyse its behaviour. Arrival rate of tickets, inter-arrival times and service times are examples of parameters that require data from the system to create a conceptual model of the system and to modify it to a computer readable form for simulation purposes of said system.

Use of ticketing data from ITSM or ticketing system also helps with one of the most important steps of the modelling process validation. When high quality data about previous incident tickets exist, simulation output can be easily validated with historical data and if inaccuracies are found, the simulation model can be adjusted to represent the system better. Without data from the ticketing system, simulation model might be too general and does not capture the essence of the specific service desk operations. It could also lead to inaccurate results since educated guesses regarding parameters may be faulty and prone to human error.

Another important use for data in telecom incident management process optimization is to collect data for reporting KPIs and service desk process and performance related metrics for understanding root causes. This understanding is created with more precise tracking of incident tickets, which allows for separate metrics for each IM process phase and SLA or priority level specific tickets. If one area of operations is struggling to meet its SLA levels, it can be better identified with this approach and implement swift actions to solve the inefficiencies. This is not possible with incomplete data. Thus, actions for improving data collection and managing data quality is crucial to better capture the insights from the data to enhance incident management process and its performance tracking.

6.2 Limitations and future research

The goal of the thesis was to optimize telecommunications incident management process for the target company. There are many aspects of the process that can be optimized, which were not included in the scope of the thesis. Thesis was especially focused on Wholesale Technical Services service desk and its specific aspects that were the most important factors for the team based on previous research made to the team’s service desk and to its plans to change the current ITSM system to another in near future. Incident management in

telecommunications is a multi-dimensional process, and because of the set scope of this thesis, it inherently overlooks some potentially crucial factors such as staff training, communication between stakeholders and providing right tools for handling incidents. Most of these suggestions are more human aspects of the incident management process that could be a valid future research opportunity.

Another crucial limitation is the practical aspects of IM process optimization. This thesis recommended an optimal set of prioritization rules and a better use of data and documentation to shift the operations to a more ITIL and continuous improvement approach but did not go in-depth to the selection and implementation of an ITSM system capable of achieving the requirements set in this thesis. These are also exciting opportunities for future research to make sure that the systems are implemented smoothly to current operations and support both the customer satisfaction and employee satisfaction.

Throughout this thesis, limitations regarding the current data from the ticketing system and simplifications made to the simulation model were discussed. While queuing models made with queuing theory and discrete event simulation are useful tools for approximating reality, no model can fully represent and capture every detail of its real-life counterpart. In this thesis, simplifying assumptions were comprehensively presented and justified. The current simulation model can be developed further by making it more complex with taking agent behaviour more into account; this could result in modelling shifts and breaks to the model for example. Because the model does not also capture aspects such as dynamic incident priorities that change over time, they could be further developed.

This thesis highlights the incident management operations and its optimization efforts, but an exciting future research opportunity could be to research other interlinked processes in ITIL context. First, event management and its improvement efforts could help IM process by making the incident management more proactive. The second idea is to further research service requests handling because they are often handled by the team also responsible for incident management. Service requests have their own distinct process and aspects inside ITIL methodology the same way as event management or problem management does, which are presented in the theoretical chapter of this thesis. Exploring these other areas of ITIL to further help optimize incident management process could bring substantial benefits and provide synergies to companies in telecommunications.

References

Agutter, C. (2019) ITIL® Foundation Essentials – ITIL 4 Edition: The ultimate revision guide. 2nd edition. Ely: IT Governance Publishing.

Banks, J. (1999) Introduction to simulation. In *Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future*, Vol. 1, pp. 7-13.

Bartolini, C., Stefanelli, C., & Tortonesi, M. (2008) SYMIAN: A simulation tool for the optimization of the IT incident management process. In *Managing Large-Scale Service Deployment: 19th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, DSOM 2008, Samos Island, Greece, September 22-26, 2008. Proceedings 19*, pp. 83-94.

Bartolini, C., Stefanelli, C., & Tortonesi, M. (2012) Modeling IT support organizations using multiple-priority queues. In *2012 IEEE Network Operations and Management Symposium*, pp. 377-384).

Bartsch, C., Mevius, M., & Oberweis, A. (2010) Simulation environment for IT service support processes: Supporting service providers in estimating service levels for incident management. In *2010 Second International Conference on Information, Process, and Knowledge Management*, pp. 23-31.

Bianco, P., Lewis, G. A., & Merson, P. (2008) Service level agreements in service-oriented architecture environments. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, Technical Report.

BMC Software (2020) ITIL Incident Management: An Introduction. [www-document]. [Cited: 19.2.2023]. Available: https://www.bmc.com/blogs/itil-v3-incident-management/

Bober, P. (2014) Simulation for IT service desk improvement. *Quality Innovation Prosperity*, *18*(1), pp. 47-58.

Brenner, M., Radisic, I., & Schollmeyer, M. (2002) A Criteria Catalog Based Methodology for Analyzing Service Management Processes. In *Management Technologies for E-Commerce and E-Business Applications: 13th IFIP/IEEE International Workshop on*

Distributed Systems: Operations and Management, DSOM 2002 Montreal, Canada, October 21–23, 2002 Proceedings 13, pp. 145-156.

Caster-Steel, A, Tan, W. (2005) Implementation of IT Infrastructure Library (ITIL) in *Australia: Progress and Success factors," 2005 IT Governance International Conference, Auckland*, pp. 14-15.

Chaykowski, K., & Coatney, M. (2018) From Broke To Billionaire: How Fred Luddy Built The World's Most Innovative Company. [www-document]. [Cited: 19.2.2023]. Available: https://www.forbes.com/feature/innovative-companies-service-now/#42396f7cc603

Che, Y. K., & Tercieux, O. (2021). Optimal queue design. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 312-313.

Chen, Z., Kang, Y., Li, L., Zhang, X., Zhang, H., Xu, H., ... & Lyu, M. R. (2020). Towards intelligent incident management: why we need it and how we make it. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1487-1497.

Churchman, M (2016) 6 Essential Steps to Reducing Incident Resolution Time. [www-document]. [Cited: 9.3.2023]. Available: https://www.pagerduty.com/blog/6-steps-reducing-incident-resolution-time/

Conger S, Winniford M, Erickson-Harris L (2008) Service management in operations. *14th Americas. conference on information systems, Toronto AIS*, pp. 3884-3894.

Cooper, R. B. (2010) Queueing notation. *Wiley Encyclopedia of Operations Research and Management Science*, pp 1-3.

Cusick, J. J., & Ma, G. (2010) Creating an ITIL inspired Incident Management approach: Roots, response, and results. In *2010 IEEE/IFIP Network Operations and Management Symposium Workshops*, pp. 142-148.

Dabade, T. D. (2012) Information technology infrastructure library (ITIL). In *Proceedings of the 4th National Conference*, pp. 25-26.

ECA (2019) IT Infrastructures Operations and Evolution services, AO/662 - TENDER SPECIFICATIONS Annex B Service Level Agreement (SLA)

framework. [www-document]. [Cited: 11.3.2023]. Available: https://etendering.ted.europa.eu/cft/cft-document.html?docId=58658

Gallego, G (2003) IEOR 4000: Production Management. [www-document]. [Cited: 23.3.2023]. Available: http://www.columbia.edu/~gmg2/4000/pdfold/throughput.pdf

Gartner (2014a) Network Downtime. [www-document]. [Cited: 31.1.2023]. Available at: https://blogs.gartner.com/andrew-lerner/2014/07/11/network-downtime/

Gartner (2014b) The Cost of Downtime. [www-document]. [Cited: 31.1.2023]. Available at: https://blogs.gartner.com/andrew-lerner/2014/07/16/the-cost-of-downtime/

Gautam, N. (2012) Analysis of queues: methods and applications. CRC Press.

Ghosh, B. (2013) Incident management & service level agreement: an optimistic approach. *In International Journal of Computer Science and Information Technologies*, *4*(3), pp. 461-466.

Gonçalves, P. (2022) Back to basics: fundamental principles of system dynamics and queueing theory. *System dynamics review*. 38 (1), pp. 81-92.

Iden, J., & Eikebrokk, T. R. (2013) Implementing IT Service Management: A systematic literature review. *International Journal of Information Management*, *33*(3), pp. 512-523.

Imamverdiyev, Y. N., & Nabiyev, B. R. (2016) Queuing Model for Information Security Monitoring Systems. *Problems of information technology*, pp. 28-32.

Ing, E., Babulak, E., & Wang, M. (2010). Discrete event simulation: State of the art. *Discrete Event Simulations. London: InTech*, pp. 1-9.

Jäntti, M., & Cater-Steel, A. (2017) Proactive Management of IT Operations to Improve IT Services. *JISTEM Journal of Information Systems and Technology Management*, 14(2), pp. 191–218.

Jenčová, E., Koščák, P., & Koščáková, M. (2023) Dimensioning the Optimal Number of Parallel Service Desks in the Passenger Handling Process at Airports Considered as a Queueing System—Case Study. *Aerospace*, 10(1), pp. 50

Johnson, J. (2008) Simple queuing theory tools you can use in healthcare. *A Presentation at the Hospital Information Management Systems Society*, pp. 1-6.

Kaiser, A. K. (2020) Become ITIL® 4 Foundation Certified in 7 Days: Understand and Prepare for the ITIL Foundation Exam with Real-Life Examples. Berkeley, CA: Apress L. P.

Kardi, T. (2014) Queuing Theory Tutorial. [www-document]. [Cited: 26.3.2023]. Available: https://people.revoledu.com/kardi/tutorial/Queuing/

Kempter (2019) Checklist Incident Priority. [www-document]. [Cited: 14.3.2023]. Available: https://wiki.en.it-processmaps.com/index.php/Checklist_Incident_Priority#Circumstances_that_warrant_the_Incident_to_be_treated_as_a_Major_Incident

Kempter, S (2022) Incident Management. [www-document]. [Cited: 19.2.2023]. Available: https://wiki.en.it-processmaps.com/index.php/Incident_Management#Handling-of-Major-Incidents

Kilpi, F. (2022) Operaattorien verkottamistuotteiden viankorjausprosessin priorisointi palvelutasosopimusten tukena. [www-document]. [Cited: 6.6.2023]. Available: https://www.theseus.fi/bitstream/handle/10024/703514/Kilpi_Feeliks.pdf?sequence=2&isAllowed=y

Lakatos, L., Szeidl, L., & Telek, M. (2013) Introduction to queueing systems with telecommunication applications. Vol. 388, New York: Springer.

Law, A. M. (2019) How to build valid and credible simulation models. In *2019 Winter Simulation Conference (WSC)*, pp. 1402-1414.

Little, J. D., & Graves, S. C. (2008) Little's law. *Building intuition: insights from basic operations management models and principles*, pp. 81-100.

Madadi, N., Roudsari, A. H., Wong, K. Y., & Galankashi, M. R. (2013) Modeling and simulation of a bank queuing system. In *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, pp. 209-215.

Microsoft (2022) What is monitoring?. [www-document]. [Cited: 9.3.2023]. Available: https://learn.microsoft.com/en-us/devops/operate/what-is-monitoring

Newell, C. (2013) Applications of queueing theory. Vol. 4, Springer Science & Business Media.

O'Dwyer, T. K. (2012) A Revised model for the implementation of the ITIL incident management process in broadcast technology operations. *International Journal of Information Technology and Business Management*, 26(1), pp. 1-12

Ortiz-Rangel, D., Rocha-Lona, L., Bada-Carbajal, L. M., Garza-Reyes, J. A., & Nadeem, S. P. (2021) Implementation of Quality Management System ISO 9001 in A Telecom Network Operation Centre–A Case Study. In *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore*.

Pawlikowski, K., Jeong, H. D., & Lee, J. S. (2002) On credibility of simulation studies of telecommunication networks. *IEEE Communications magazine*, *40*(1), pp. 132-139.

Pereira, de Vasconcelos, J. B., Rocha, Á., & Bianchi, I. S. (2021) Business process management heuristics in IT service management: a case study for incident management. *Computational and Mathematical Organization Theory*, 27(3), pp. 264–301.

Pollard, C. E., Gupta, D., & Satzinger, J. W. (2010) Teaching systems development: A compelling case for integrating the SDLC with the ITSM lifecycle. *Information Systems Management*, 27(2), pp. 113–122.

Punyateera, J., Leelasantitham, A., Kiattitsin, S., & Muttitanon, W. (2014) Study of service desk for NEdNet using incident management (Service Operation) of ITIL V. 3. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1-6.

Saarelainen, K. (2016) How and why things happen: anatomy of IT service incidents. Publications of the University of Eastern Finland. Dissertations in Forestry and Natural Sciences, 238.

Salah, Maciá-Fernández, G., & Díaz-Verdejo, J. E. (2019) Fusing information from tickets and alerts to improve the incident resolution process. *Information Fusion*, 45, pp. 38–52.

Sarkar, A., Mukhopadhyay, A. R., & Ghosh, S. K. (2011) Improvement of service quality by reducing waiting time for service. Simulation Modelling Practice and Theory, 19(7), pp. 1689-1698.

Sarkar, N. I., & Gutiérrez, J. A. (2014) Revisiting the issue of the credibility of simulation studies in telecommunication networks: highlighting the results of a comprehensive survey of IEEE publications. *IEEE Communications Magazine*, *52*(5), pp. 218-224.

Sencer, A., & Ozel, B. (2013) A simulation-based decision support system for workforce management in call centers. *Simulation*, *89*(4), pp. 481-497.

ServiceNow (2023a) Product documentation: Major Incident Management. [www-document]. [Cited: 13.03.2023]. Available: https://docs.servicenow.com/bundle/utah-it-service-management/page/product/incident-management/concept/major-incident-management.html

ServiceNow (2023b) Product documentation: Major Incident Management process. [www-document]. [Cited: 14.3.2023]. Available: https://docs.servicenow.com/bundle/utah-it-service-management/page/product/incident-management/concept/major-incident-management-process.html

Sharifi, M., Ayat, M., Ibrahim, S., & Sahibuddin, S. (2009) The most applicable KPIs of problem management process in organizations. *International Journal of Simulation: Systems, Science and Technology*, 10(3), pp. 77-83.

SimPy (2020) Overview — SimPy 4.0.2.dev1+g2973dbe documentation. [www-document]. [Cited: 12.6.2023]. Available: https://simpy.readthedocs.io/en/latest/

Stallings, W. (2011) QUEUEING SYSTEM CONCEPTS. [www-document]. [Cited: 9.5.2023]. Available: https://www.cs.helsinki.fi/group/nodes/kurssit/kj/app_h_queueingsystem.pdf

Swain, A. K., & Garza, V. R. (2022) Key Factors in Achieving Service Level Agreements (SLA) for Information Technology (IT) Incident Resolution. *Information Systems Frontiers*, pp. 1-16.

Sztrik, J. (2010). Queueing theory and its Applications, a personal view. In *Proceedings of the 8th international conference on applied informatics*, Vol. 1, pp. 9-30.

Sztrik, J. (2012) Basic queueing theory. University of Debrecen, Faculty of Informatics, 193, pp. 60-67.

Telia Company (2023) About the company. [www-document]. [Cited: 4.2.2023]. Available: https://www.teliacompany.com/en/about-the-company/

Tulsian, P. C., Pandey, V. (2006) Quantitative Techniques: Theory and Problems. [www-document]. [Cited: 26.3.2023]. Available: https://www.oreilly.com/library/view/quantitative-techniques-theory/9789332512085/xhtml/ch9sec5.xhtml