Journal of
Sensor and Actuator Networks

MDPI

Article

A Virtualization Infrastructure Cost Model for 5G Network Slice Provisioning in a Smart Factory

Jaspreet Singh Walia [1,*], Heikki Hämmäinen [1], Kalevi Kilkki [1], Hannu Flinck [2], Seppo Yrjölä [3] and Marja Matinmikko-Blue [4]

[1] Department of Communications and Networking, Aalto University, 02150 Espoo, Finland; heikki.hammainen@aalto.fi (H.H.); kalevi.kilkki@aalto.fi (K.K.)
[2] Nokia Bell Labs, 02610 Espoo, Finland; hannu.flinck@nokia-bell-labs.com
[3] Nokia, 90650 Oulu, Finland; seppo.yrjola@nokia.com
[4] Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland; marja.matinmikko@oulu.fi
[*] Correspondence: jaspreet.walia@aalto.fi

Abstract: Network slicing is a key enabler for providing new services to industry verticals. In order to enable network slice provisioning, it is important to study the network slice type allocation for different device types in a real industrial case. Furthermore, the costs of the required virtualization infrastructure need to be analyzed for various cloud deployment scenarios. In this paper, a cost model for the virtualization infrastructure needed for network slice provisioning is developed and subsequently applied to a real smart factory. In the model, slice types and devices are mapped such that each factory device is provisioned with one or more slice types, as required. The number of devices to be supported per slice type is forecasted for 2021–2030, and the total costs of ownership, costs per slice type, and costs for every slice type, for each device are calculated. The results are analyzed for three cloud deployment scenarios: local, distributed, and centralized. The centralized scenario was found to have the lowest cost. Moreover, sensitivity analysis is conducted by varying the device growth, the number of factories, the level of isolation between network slices, and resource overbooking. The resulting evaluation and cost breakdown can help stakeholders select a suitable deployment scenario, gauge their investments, and exercise suitable pricing.

Keywords: 5G; industry verticals; smart factory; network slicing; cost model; virtualization infrastructure; network planning; cloud deployment

check for updates

Citation: Walia, J.S.; Hämmäinen, H.; Kilkki, K.; Flinck, H.; Yrjölä, S.; Matinmikko-Blue, M. A Virtualization Infrastructure Cost Model for 5G Network Slice Provisioning in a Smart Factory. *J. Sens. Actuator Netw.* **2021**, *10*, 51. https://doi.org/10.3390/jsan10030051

Academic Editor:
Konstantinos Katsaros

Received: 17 June 2021
Accepted: 26 July 2021
Published: 28 July 2021

1. Introduction

Fifth generation mobile networks or 5G networks are a key enabler for industry verticals that require fast, secure, ultra-reliable, and low-latency communications. Industry verticals consist of diverse device types belonging to different use cases, which need to be provisioned over specialized network slices. Network slicing in 5G is enabled by Network Functions Virtualization (NFV) and Software-Defined Networking (SDN) to logically isolate and provision required virtual network resources as end-to-end virtual network slices. NFV decouples the network functions from their dedicated network equipment and enables their operation as Virtual Network Functions (VNFs) on general-purpose server hardware [1]. Furthermore, SDN enables instantiating the VNFs as service-specific function chains. One of the main advantages of SDN in 5G comes from decoupling the User Plane (UP) and Control Plane (CP) VNFs for independent scalability and management [2]. The Control and User Plane Separation (CUPS) enables the separation and optimal placement of the UP and CP VNFs, which can be centralized in one location or optimally distributed to various locations, as required by the use cases [3]. Additionally, the UP and the CP can be deployed locally, enabling lower latencies and local processing of sensitive data for critical use cases such as manufacturing and public safety. Based on resource requirements derived from the necessary data rate, latency, reliability, availability, and number of connected devices, use-case-specific slices can be provided to various industries.

Mobile Network Operators (MNOs) have traditionally relied on adding more capacity and coverage to serve the increasing number of customers and have primarily served wide-area communications with cellular networks. On the other hand, local area communications have been dependent on wired or wireless connectivity such as Wi-Fi. The fourth industrial revolution is bringing together various industries through the integration of cyber-physical systems, digitalization of assets, Internet of Things, and the virtualization of networking devices over new communication technologies [4]. Local-area communications have become more critical as more industries are automating their processes, thus requiring better connectivity, mobility, and reliability. With a growing number of customers and varying connectivity requirements across different use cases such as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), massive Internet of Things (mIoT), and Vehicle-to-anything (V2X) communications, it has become increasingly important to move towards more reliable wireless connectivity. In the beginning, 5G is rolled out as a non-standalone deployment, utilizing 5G New Radio (NR) in the access network but still relying on the 4G Evolved Packet Core (EPC). Such a non-standalone deployment offers the expected benefits in Radio Access Network (RAN), but a full end-to-end 5G network slicing requires the next generation virtualized 5G core network. A standalone 5G network provides significantly better performance in terms of latency as compared to non-standalone 5G and legacy 4G networks [5]. Thus, for stakeholders transitioning to a full end-to-end sliced 5G network, developing a cost model for the virtualization infrastructure in order to be able to provide network slices is essential.

According to [6], the global network function virtualization market is expected to grow at a Compound Annual Growth Rate (CAGR) of 20.7%, reaching USD 59.1 billion in 2027. The global network slicing market is expected to reach USD 5.8 billion by 2025 and will be led by enterprise and industrial applications, followed by public safety [7]. The total global value of the addressable 5G-enabled market for service providers is expected to reach USD 700 billion in 2030 [8]. Such estimates point to the vast business potential to be unlocked; however, the commercial and technical demand for network slicing can vary for countries, industry verticals, and their use cases. Thus, the virtualization infrastructure cost modeling needs to be adaptable to different scenarios.

The economic benefits of SDN and NFV in providing significant cost savings have been widely studied [9–11]. Different data center topologies have been compared in [12], and savings through virtualization in the Total Costs of Ownership (TCO) for different data center architectures have been studied in [11]. Furthermore, different VNF placement strategies in different data center architectures are studied for VNF service chains, with four VNFs per chain [13]. However, the cost modeling for virtualization infrastructure has not been studied, especially from a market-based approach that considers the perspectives of slice customers and slice providers in provisioning the required one-to-one, one-to-many, many-to-many, or many-to-one slice type allocation. Moreover, the demand for slices is expected to evolve over time, and a demand forecast for the number of devices to be supported per network slice must be conducted. Additionally, different deployment scenarios for the virtualization infrastructure must be compared.

In this paper, a slice demand-based cost model is developed for the virtualization infrastructure for slice provisioning, and the model is applied to a real smart factory as a case study. The smart factory vertical with a multi-tenant, multi-use case ecosystem requires the provisioning of use-case-specific network slices. The allocation of slice types to customers' device types in the model is accomplished for a smart factory, but the cost model is scalable to other verticals as well. The model considers one-to-one, one-to-many, many-to-many, or many-to-one slice type allocation between the slice customer and slice provider, conducts a slice demand forecast over 2021–2030, and compares the costs of three different deployment scenarios. Further sensitivity analysis is performed for the cost model by varying the growth of devices, the number of factory customers, the level of isolation between the slices, and resource overbooking. The rest of the paper provides the relevant background, followed by the developed cost model, case evaluation for a smart factory accompanied with sensitivity analysis of key input parameters, and followed by conclusions.

2. Background

The fifth generation mobile network is considered an innovative and disruptive technology compared to existing communication technologies employed in industrial networks [14]. The 5G Service-Based Architecture (SBA) is derived from the virtualization and separation of network entities into network functions according to the services they perform [3]. The CP and UP network functions can be separated for independent deployment, scalability, and management [15]. The major network functions defined in the 3GPP specifications include the User Plane Function (UPF), Access and Mobility Management Function (AMF), Session Management Function (SMF), Network Slice Selection Function (NSSF), Policy and Charging Function (PCF), User Data Management Function (UDM), and Authentication Server Function (AUSF) [3]. These VNFs can be deployed on off-the-shelf server hardware, in centralized or distributed virtualization infrastructures, and instantiated as service function chains forming the required network slices.

2.1. Network Slicing

The 5G wireless technology enables the provisioning of network slices for use-case-specific connectivity requirements such as the required data rate, latency, reliability, availability, and number of devices to be supported. The NSSF is used to select network slices based on the customer's connectivity requirements. A slice type is identified using the Single-Network Slice Selection Assistance Information (S-NSSAI) identifier consisting of an 8-bit Slice/Service Type (SST) for the type of service, and an optional 24-bit slice differentiator to differentiate among slices of the same SST [16]. The 8-bit SST field gives room for hundreds of different slice types and further differentiation through the slice differentiator. Likewise, one user equipment can request an NSSAI consisting of 8 S-NSSAIs at the same time. Thus, devices from different customers can be provisioned with the requested slice types out of the standardized slice types, or by configuring the slice templates or defining new slice types altogether [17]. As a result, slice allocation between devices and slice types must follow one-to-one, one-to-many, many-to-many, or many-to-one relationships. For the scope of this paper, only the standardized slice types are included in the cost model, and no additional slices are defined. However, any number of slice types can be included in the model.

Network slice provisioning can exercise different levels of isolation between the VNFs of different slices. For example, two use cases—a and b—can be served over the same physical infrastructure without logical isolation, where all VNFs are common; they can also be served with partial/full logical isolation, where some/all VNFs can be dedicated per slice. This is illustrated in Figure 1, where a separate VNF1 is used for each slice for partial isolation, but VNF2 and VNF3 are shared between slices; in the case of full isolation, each VNF is dedicated per slice. The level of isolation between VNFs can depend on use-case-specific technical requirements, data security, and the required flexibility in configuring and managing VNFs for each slice. Furthermore, the slices can consist of VNFs deployed in different physical locations such as the UP VNFs being deployed locally or in the edge cloud, while the CP VNFs can be deployed in an operator's central cloud.

Different isolation options can be applied in the cost modeling. The dedicated/common VNFs translate to dedicated/common VMs, which in turn translate into servers, racks, cabling, switches, and the required software licenses, VNF licenses, electricity, cooling, and maintenance cost allocations per slice.

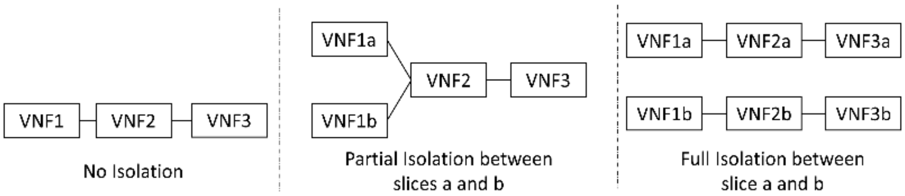

VNF1
VNF2
VNF3
No Isolation
VNF1a
VNF1b
VNF2
VNF3
Partial Isolation between slices a and b
VNF1a
VNF2a
VNF3a
VNF1b
VNF2b
VNF3b
Full Isolation between slice a and b

Figure 1. Isolation in network slice provisioning.

2.2. Virtualization Infrastructure

In this paper, virtualization infrastructure refers to the major cost components of a data center required for the operation of the VNFs for slice provisioning by the virtualized 5G core network. The 3GPP release 16 includes enhancements for the RAN for higher reliability as required by industrial use cases through the duplication of data packets and the addition of corresponding redundant resources [18]. While the RAN and transport dimensioning are not included in this paper, and the focus is limited to virtualization infrastructure for the core network, a similar concept of redundant resources also typical in the factory environment is applied to data center resources. The VNFs can be deployed on different virtualization technologies such as Virtual Machines (VMs), containers, or unikernels with the required network, storage, and computing capacities [1]. A VM is a virtually simulated hardware with required capacities as well as a guest operating system, running in isolation from other VMs in the same physical machine. VMs are a popular approach for provisioning VNFs in the cloud and are suitable for long-running services. On the other hand, containers do not have a guest operating system and allow the applications to directly interact with the underlying system kernel, having faster startup times while still being capable of isolated capacities. Unikernels provide the static linking of applications to small library operating systems and provide similar isolation to that in VMs, but unikernels are less flexible as a recompilation of the entire unikernel image is required for the addition and removal of functionality [1,19]. Thus, different approaches and granularities can be defined for the dimensioning of virtual resources and can be suitably applied in the cost model. Additionally, different levels of isolation between the virtual resources can be defined. For example, in the no-isolation approach, multiple VNFs can run on the same virtual resources; in the partial-isolation approach, some VNFs run on dedicated resources, while some run on shared virtual resources; finally, in the full-isolation approach, each VNF runs on its own dedicated virtual resources.

In the cost model, a VM is selected as the virtual resource to be dimensioned. With the increasing dependence on the cloud for networking services, new virtualization technologies and supporting business models are emerging. For example, microservices can be used to create reusable and independently deployable cloud components and serverless computing with function as a service platforms can then be utilized so that the application developer does not need to maintain virtualization infrastructure, and the resources are invoked only when required by the functions [19–21]. Several large cloud providers such as Microsoft, Google, and Amazon can act as virtualization infrastructure providers and offer such different options for virtualized resources and can also provide high scalability and lower costs due to centralization [19]. The VNFs can be deployed in centralized or distributed virtualization infrastructures that can be dedicated to one actor or may be common for multiple actors. Edge clouds have generated sufficient interest to enable low latency and the faster processing of data closer to customers [22]. Likewise, micro data centers can be considered for local data processing in urban environments [23].

Various business actors need to strategize between centralized and distributed deployments of virtualization infrastructure, considering technical constraints such as data security, latency, and reliability as well as economic costs and benefits. The infrastructure can be owned either by a single actor (network operator, virtualization infrastructure provider, network slice provider, network slice customer) or by a neutral host in order to enable sharing between different actors or even as a joint venture. Furthermore, the virtualization infrastructure can be centralized in an edge/central cloud, distributed to different locations, or deployed locally inside each factory. The decision to centralize or distribute the virtualization infrastructure affects the cost allocation between respective business actors and is dependent on the use-case requirements and the business model for network slice provisioning.

Typically, the data centers involve top-of-rack or access switches, aggregation switches, and core switches. Various deployment architectures such as the two-tier tree, three-tier tree, and fat tree can be suitable for designing the topology of the switches in the data center as well as for different data center sizes [24]. However, in this paper, the scope in terms of switches is limited to the top-of-rack/access switches, and different switch topologies are not evaluated. The top-of-rack switches are assumed to connect the servers in the racks to the rest of the network, and a redundancy of two switches per rack is typically assumed. The servers are the main virtualized resource that host the VMs running the VNFs, thus providing the network, computation, and storage resources required per slice [10. The VNFs can be deployed in the central cloud in the factory, or the UP and CP can be distributed between the factory and the central cloud [22,23]. The major cost components for the virtualization infrastructure are discussed as cost input for the cost model later in the paper.

2.3. Industrial Requirements for Network Slicing

Compared to previous generations, 5G provides higher data rates, better reliability, lower latencies, and support for massive connectivity. Industry verticals such as manufacturing, automotive, healthcare, public safety, and energy consist of many potential slice customers. A smart factory includes multiple device types with use-case-dependent technical and architectural requirements [25,26]. For example, an augmented reality device requires high data rates and low latencies, while a traditional broadband device does not have strict requirements for latency. Similarly, a logistics monitoring device requires low data rates and does not have strict latency requirements; however, the network should be able to support connectivity for a massive number of such devices. Furthermore, a V2X device requires low latencies, and the network should be able to support a varying density of devices depending on the amount of vehicular traffic. Moreover, the number of connected devices has been increasing with the increase in the level of automation. The different device types for different use cases can be provisioned with required slices and managed accordingly over local and wide area networks [27]. The edge cloud must be scalable for reliable service provisioning, and ad hoc networks can be deployed depending on the latency constraints [28,29]. Deploying the virtualized core network functions locally can significantly reduce the latencies; however, depending on the use cases, it is possible to deploy resources up to a maximum geographic distance if it still satisfies the latency requirements, as studied for the augmented reality use case in [30]. Thus, it is possible to deploy the VNFs in a central, distributed, or local cloud, given that the latency requirements can be satisfied.

The reliability and availability of the network are crucial to ensure uninterrupted factory operations. Reliability in communications can be defined in terms of the maximum amount of allowed packet loss in a given time, and availability as the time the connectivity service should be available to the devices [25]. Regarding the smart factory under study, the costs of the communication network make up a very small portion of the overall factory expenses, and the communication network must be reliable as well as available in order to enable uninterrupted communication services for factory processes. Redundant networking and power equipment are employed to ensure high reliability and availability for factory processes; for example, redundant switches, servers, and power backups are utilized so that in the case of a failure, the redundant equipment can continue providing the requested communication services. Additionally, different power areas are defined to isolate power sources and backups. The security of the processes, generated data, and trustworthiness for stakeholders are other crucial aspects in factories and are typically handled by deploying private networks or through logical isolation in order to safeguard critical network infrastructures. Thus, network isolation through network slicing plays a pivotal role, not just for allocating required network resources but also for maintaining the reliability, latency, and security of factory processes.

3. Cost Model

The cost model is defined based on the required slice type allocation between the slice customer and the provider. A network slice customer can be subscribed to single or multiple network slice types. The various stakeholders involved in the network slice provisioning, including customers, network slice providers, virtual network operators, connectivity operators, and virtualization infrastructure providers, can exercise varying levels of control to provision and manage network slices depending on the business models. The VNFs to be implemented in a virtualization infrastructure need to be dimensioned according to the resource requirements per slice type and the number of customer devices per slice type. The cost model is developed based on the slice allocation to be performed per device type for each slice customer; subsequently, based on the forecasted demand, the required virtualization infrastructure costs are calculated.

3.1. Slice Type Allocation

Multiple slice customers are assumed to exist in the market. In the cost model, a slice customer is assumed to have different device types. Slice type allocation must be managed between a slice customer and a slice provider through a slice orchestration platform and should be modified as required. A slice provider needs to allocate the required slice type per device, for each customer, as depicted in Figure 2. The devices and slices can thus have many-to-one, one-to-many, one-to-one, and many-to-many relationships. Therefore, the total number of devices for each device type that needs to be provisioned with a particular slice type(s) dictate the demand for that slice type(s).

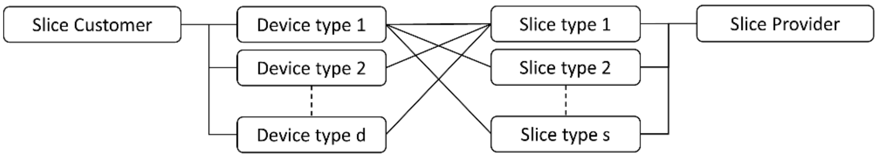


Figure 2. Slice Type Allocation.

3.2. Cost Model

Based on the allocated slice types, the further steps required to calculate the costs of the virtual infrastructure are depicted in Figure 3. The demand for a slice type is determined based on the devices from all customers requiring that specific slice type. While some VNFs can be dedicated to each slice type, some VNFs can be common for different slice types. A typical off-the-shelf server is assumed to have a given amount of cores, RAM, storage, and networking capacity, which can be separated into VMs. The VMs can be of different sizes based on the number of cores and their corresponding capacities.

Based on the slice customer's device growth, a 'Device forecast' is estimated for the time period under study. Device forecasts can be obtained from industry vertical customers' expectations and by following trends in the telecom market. Based on the 'Device forecast' and the allocated slices, the 'Slice type forecast' depicting the demand for network slices is calculated. In the forecast, the slice size is defined as the number of devices to be supported annually. The 'Slice types' are designed by the slice provider in terms of the virtualized resource requirements per slice type, for each device. The defined 'Slice types' along with the 'Slice type forecast' are used to calculate the total virtualized 'Resource requirements' for each slice type. The 'Resource requirements' are then used for 'VM dimensioning', thus converting the slice provider's slice type input into the number of VMs required for each slice type. The 'VM dimensioning' and the server capacity help estimate the 'Number of servers' required. Any required amount of redundancy can be assumed for the servers per slice type in order to calculate the final 'Number of servers', which helps estimate the supporting racks and cabling, and switches.

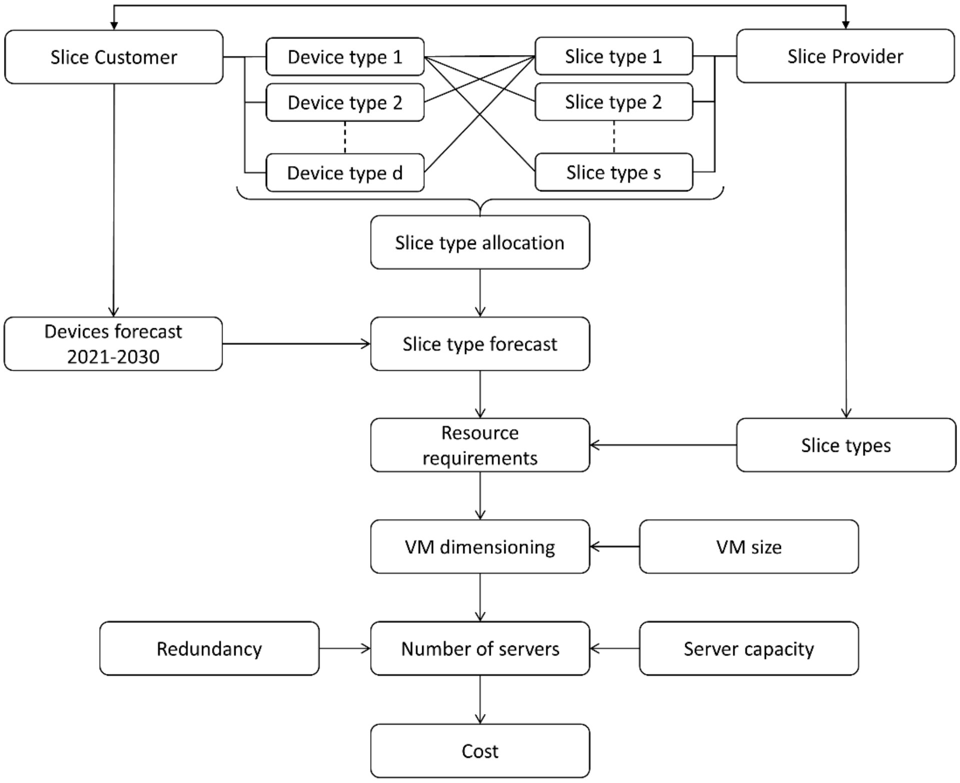


Figure 3. Cost Model for Virtualization Infrastructure.

Furthermore, the following assumptions are made:

- The VMs are assumed to support a maximum throughput for UP traffic and a maximum number of devices for CP traffic;
- Each VM serves only one VNF, while one VNF might require more than one VM based on capacity demand;
- When a slice requires a dedicated VNF, the VM(s) associated to the VNF are dedicated to that slice. On the other hand, for the common VNFs belonging to multiple slices, the associated VM(s) are common for those multiple slices.

In this manner, the VM requirements can be normalized per device for each VNF of each slice type, and then scaled according to the forecasted number of devices each year over 2021–2030. The number of VMs are calculated based on total resource requirements for each CP/UP VNF and the supported CP/UP VM capacity as follows:

$$\text{Number of VMs per VNF} \ = \ \text{Ceil}\left(\frac{\text{Total resource req. per VNF}}{\text{VM capacity}}\right) \tag{1}$$

The number of UP and CP VMs are then converted to the number of required servers, depending on the deployment scenario. Three deployment scenarios are considered, and the number of servers is calculated as follows:

In scenario A, only one virtualization infrastructure is deployed in a central cloud or in an edge cloud, with a strategic location for both UP and CP VNFs from which network slices could be served to multiple factories.

$$\begin{aligned}&\text{Number of servers (A)}\\&= \text{Ceil}\left(\frac{\text{No. of UP VMs}}{\text{Server VM capacity}}\right) + \text{Ceil}\left(\frac{\text{No. of CP VMs}}{\text{Server VM capacity}}\right) + \text{Ceil}\left(\frac{\text{No. of common CP VMs}}{\text{Server VM capacity}}\right)\end{aligned} \tag{2}$$

In scenario B, each factory deploys its own virtualization infrastructure for UP VNFs, while the CP VNFs for all factories are deployed in a central cloud or in an edge cloud.

$$\begin{aligned}&\text{Number of servers (B)}\\&= \text{No. of factories} \times \left(\text{Ceil}\left(\frac{\text{No. of UP VMs per factory}}{\text{Server VM capacity}}\right)\right) + \text{Ceil}\left(\frac{\text{No. of CP VMs}}{\text{Server VM capacity}}\right)\\&+\text{Ceil}\left(\frac{\text{No. of common CP VMs}}{\text{Server VM capacity}}\right)\end{aligned} \tag{3}$$

In scenario C, each factory deploys its own virtualization infrastructure for both UP and CP VNFs, which is open to service providers and MNOs to run their VNFs from in order to serve the factory.

$$\begin{aligned}&\text{Number of servers (C)}\\&= \text{Num. of factories}\\&\times\left(\text{Ceil}\left(\frac{\text{No. of UP VMs per factory}}{\text{Server VM capacity}}\right) + \text{Ceil}\left(\frac{\text{No. of CP VMs per factory}}{\text{Server VM capacity}}\right)\right.\\&\left.+\text{Ceil}\left(\frac{\text{No. of common CP VMs per factory}}{\text{Server VM capacity}}\right)\right)\end{aligned} \tag{4}$$

After calculating the number of servers, the supporting equipment—racks, cabling, and switches along with the software, VNF licenses, energy, cooling, and maintenance can also be calculated as described next.

3.3. Cost Components in a Cloud

The major cost components in a cloud are described below and their average costs available in literature are summarized in Table 1. The Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) contributions of each component are described further.

Table 1. Cost Inputs.

Variable	Cost Item	Cost	Source
Number of Servers, Ns	Server, C_S	EUR 10,000	[31]
Number of Racks, N_R	Racks and cabling, C_R	EUR 20,000	[31]
Number of Switches, N_{SW}	Switch, C_{SW}	EUR 10,000	[31]
-	Power per server, P_S	400 W	[31]
-	Power per switch, P_{SW}	400 W	[31]
-	Electricity price, E	0.1 euro/kWh	[32]
-	Software license per server, C_{SL}	1300 EUR/year	[31]
Number of VNF licenses, N_{VNF}	VNF license, C_{VNF}	85 EUR/year	100 USD/year [33]
-	Maintenance, C_M	10% of CAPEX	Assumption

3.3.1. Servers, Switches, Racks, and Cabling

Servers are assumed to have a lifecycle of five years, thus needing renewal every five years [34]. In the model, the server investments are made based on the forecasted demand. Since server investments are assumed to be made every five years, the total number of servers required to meet the demand of 2025 are bought in 2021, then the total number of servers required to meet the demand of 2030 are bought in 2025. While the computing capacity improves over generations of servers, for simplicity, it is assumed that that servers with the same capacity are bought at inflation-adjusted prices. The average costs of servers, racks and cabling as well as switches (including installation costs) in a telecom operator's cloud are based on [32]. Switches are assumed to have a lifecycle of five years, the same as that of servers, while the CAPEX is calculated in the same manner as the servers. Top-of-rack switches are assumed, thus requiring a minimum of one switch per rack. For redundancy, two switches are assumed per rack, and both contribute to the annual OPEX. A rack housing 12 servers is used for calculation, and the racks and cabling are assumed not to need renewal within the scope of ten years.

$$\text{No. of Racks, } N_R = \text{Ceil}\left(\frac{N_S}{12}\right) \tag{5}$$

$$\text{No. of Switches, } N_{SW} = 2 \times N_R \tag{6}$$

Furthermore, software licenses are assumed per server and VNF licenses are assumed per VNF. The power consumption, maintenance, and software/VNF licenses contribute to the server OPEX. The redundancy for servers is determined according to the slice requirements; different redundancies can be applied during calculation per slice type.

3.3.2. Energy and Cooling

The OPEX also includes the power utilized by servers and switches as well as the power utilized to cool the heat generated by the servers and switches. For this purpose, a power usage efficiency (PUE) of 1.8 is utilized as in [34].

$$\text{Annual Electricity \& Cooling} = \text{PUE} \times \text{E} \times 24 \times 365 \times (N_S \times P_S + N_{SW} \times P_{SW}) \tag{7}$$

This helps estimate the power consumed by the equipment as well as the power required for cooling and the contribution to the annual OPEX. Electricity prices for medium scale industry customers have stayed roughly steady in Finland over the past decade, and a fixed price of 0.1 EUR/kWh is used [32].

3.3.3. Software and VNF licenses

Software licenses are assumed to contribute to annual OPEX and are directly dependent on total servers installed at any time [32]. Fixed operational costs per VNF are assumed and depend on the total VNFs installed; VNF license costs are based on [33]. For example, if a VNF license cost is x EUR/year, and the deployment scenario requires four VNFs, then the total annual VNF license costs are 4x EUR/year.

$$\text{Annual Software costs } = C_{SL} \times N_S \tag{8}$$

$$\text{Annual VNF license costs } = C_{VNF} \times N_{VNF} \tag{9}$$

Apart from the major cost components of a data center described above, slice provisioning will require operational costs arising from personnel management costs and the orchestration of slices. The level of automation in management and orchestration is expected to increase, thus requiring less personnel costs in the future. These personnel costs are not included in the modeling as the focus is limited to the virtualization infrastructure costs.

The costs of the virtualization infrastructure are then calculated as follows:

$$\text{CAPEX} = C_S \times N_S + C_{SW} \times N_{SW} + C_R \times N_R \tag{10}$$

$$\begin{aligned}\text{OPEX } &= C_{SL} \times N_S + C_{VNF} \times N_{VNF} + \text{PUE} \times \text{E} \times 24 \times 365 \\ &\quad \times (N_S \times P_S + N_{SW} \times P_{SW}) + C_M\end{aligned} \tag{11}$$

$$\text{TCO } = \text{CAPEX} + \text{OPEX} \tag{12}$$

The above calculations are performed for each slice type and each deployment scenario, and a TCO breakdown is also provided in terms of costs per slice type and costs per slice type for each device. Furthermore, a sensitivity analysis is performed by varying the device growth over the study period and scaling the number of slice customers, thus making the model applicable to different industry verticals, use cases, and scenarios. The developed cost model is applied to the subsequent smart factory case.

4. Results and Discussion

An existing smart factory is selected as the case study for the evaluation of the virtualization infrastructure costs required for network slices, using the developed cost model.

4.1. Slice Type Allocation and Forecasting

The factory is treated as the slice customer, with a list of device types; slice type allocation is performed for each device type, as presented in Table 2. Four standardized slice types are considered: eMBB, URLLC, mIoT, and V2X [3].

Table 2. Slice Type Allocation for the Smart Factory case.

Device Type	eMBB	mIoT	URLLC	V2X
Production testers	-	x	x	-
Surface mount technology machines	-	x	x	-
Programmable logic controller machines	-	x	x	-
IoT devices, sensors	-	x	-	-
Computers	x	-	-	-
Mobile robots	x	x	x	x
Handheld devices	x	-	x	-
Network cameras	x	-	x	-
Office printers	x	-	-	-
Telepresence (AR/VR)	x	-	x	-
Wide-area logistics vehicles	-	x	-	x

The growth of devices is subsequently estimated in the period between 2021 and 2030, with a CAGR of 6% for a specified list of devices based on inputs from a real electronics assembly factory, while the growth of IoT devices is expected to be higher than the rest of the devices. Thus, for IoT devices, a higher growth rate of 10% is assumed for the selected time period [35]. The calculations are made for ten of such factories. The resulting device forecast is presented in Table 3.

Table 3. Devices forecast (2021–2030) for ten factories with a 6% CAGR (with 10% CAGR for IoT devices).

Device Type	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Production testers	2120	2247	2382	2525	2676	2837	3007	3188	3379	3582
Surface mount technology machines	636	674	715	757	803	851	902	956	1014	1075
Programmable logic controller machines	106	112	119	126	134	142	150	159	169	179
IoT devices, sensors	3300	3630	3993	4392	4832	5315	5846	6431	7074	7781
Computers	3180	3371	3573	3787	4015	4256	4511	4782	5068	5373
Mobile robots	32	34	36	38	40	43	45	48	51	54
Handheld devices	424	449	476	505	535	567	601	638	676	716
Network cameras	106	112	119	126	134	142	150	159	169	179
Office printers	424	449	476	505	535	567	601	638	676	716
Telepresence (AR/VR)	21	22	24	25	27	28	30	32	34	36
Wide-area logistics vehicles	53	56	60	63	67	71	75	80	84	90

The device growth forecast and the slice allocation tables are used to estimate the slice size forecast, which is the number of devices to be supported by a slice type annually, as shown in Figure 4.

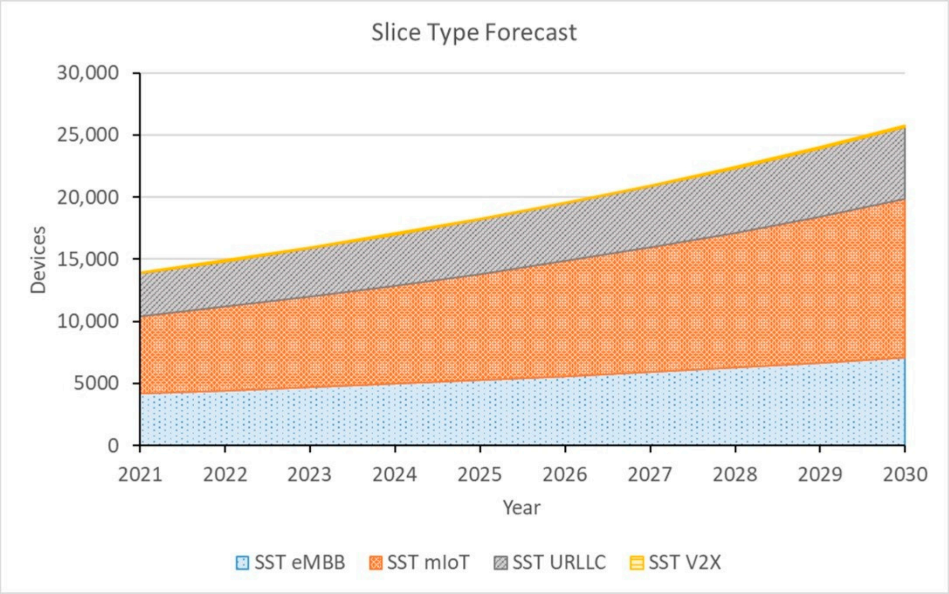


Figure 4. Slice Type Forecast for 2021–2030.

4.2. Dimensioning

The different slice types have different hardware requirements from the virtualization infrastructure. The slice type requirements can be presented in terms of the number of VMs required by a VNF for the UP and the CP. The smart factory vertical requires high availability and a guaranteed level of service; thus, a worst-case scenario with all services running all the time is considered. Alternatively, some other scenarios are also possible when not all services are running all the time, allowing the overbooking of resources, as studied in the sensitivity analysis section. Different configurations for VMs and servers can be used in the model, and for evaluation purposes—up to eight VMs the size of two cores, with a 16 GB RAM, 500 GB storage, and 10 Gbps throughput capacity are assumed to be running on a server with 16 cores, 128 GB RAM, 4000 GB storage, and 80 Gbps throughput capacity. The UP requirements are based on the traffic model for industrial use cases described in [36], which utilizes 3GPP requirements [25]. These requirements are then divided by the throughput capacity of a VM. For the CP requirements, it is assumed that 10,000 devices can be supported per server core (20,000 devices are supported per VM for any CP VNF). It should be noted that the requirements are only representative examples and real input values as well as costs will be case-dependent and will vary per customer segment and from provider to provider. It can also be assumed that resource requirements per device type or slice type are known to operators and slice providers, as they apply such information to build their service portfolios. The assumed input VM requirements for UP and CP are listed in Table 4.

Table 4. Input VM requirements for UP and CP VNFs, per device and for different slice types.

Number of VMs per Device	UP	CP
eMBB	8×10^{-4}	5×10^{-5}
mIoT	10^{-5}	5×10^{-5}
URLLC	9.6×10^{-4}	5×10^{-5}
V2X	9.6×10^{-4}	5×10^{-5}

A partial isolation case is first considered, where some VNFs are dedicated to each slice type and some VNFs are common between slice types. For evaluation, the UPF and the CP functions of SMF and AMF are assumed to be dedicated functions, while other CP functions such as NSSF, PCF, UDM, and AUSF are assumed to be common to multiple slices. The VMs for a dedicated VNF and for a common VNF are then calculated. With regard to sensitivity, full and no isolation cases that encompass all the possible distributions of VNFs per slice are also considered later in the sensitivity analysis section.

The calculated number of VMs for UPFs, dedicated CP VNFs, and common CP VNFs is shown in Figures 5–7 respectively. The number of common CP VNFs is depicted with a black line measured along the secondary y-axis on the right, while the percentage of this common dimensioning for different slice types is shown using the primary y-axis. Due to a small annual growth rate of 6% and only ten-slice customer factories, the VMs required do not increase dramatically. However, the sensitivity analysis for different growth rates and the number of factories will be performed later in the paper. After calculating the total VMs per slice, the total servers per slice are calculated.

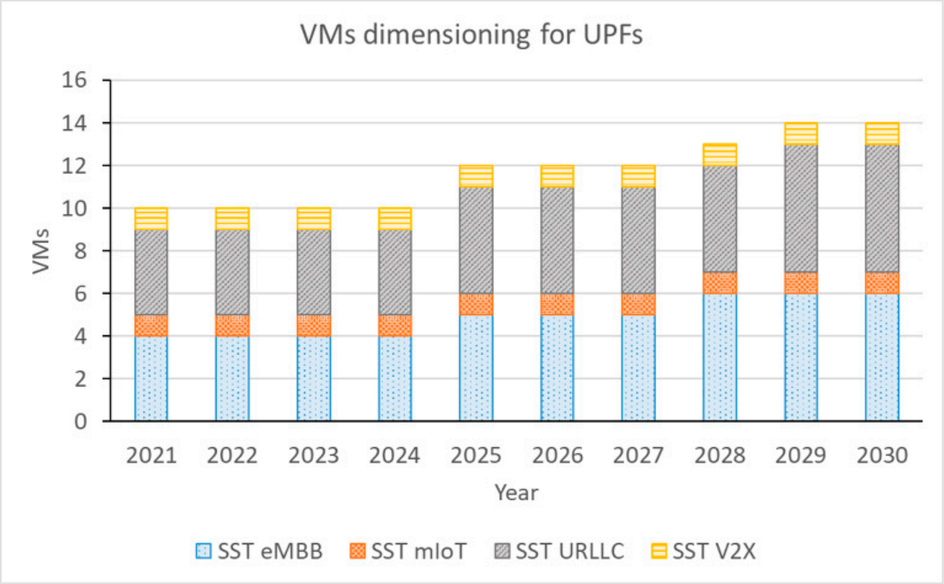


Figure 5. VMs dimensioning for UPFs.

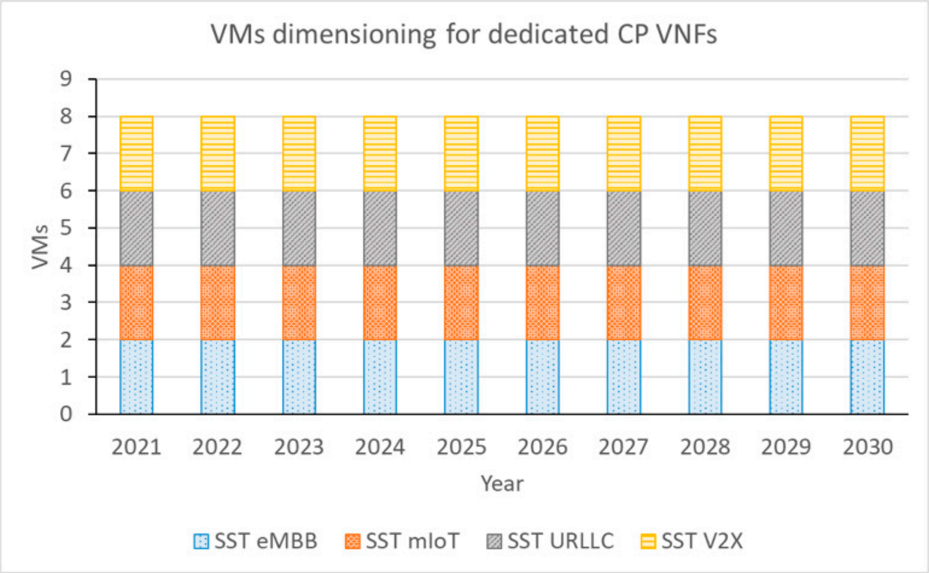


Figure 6. VMs dimensioning for dedicated CP VNFs.

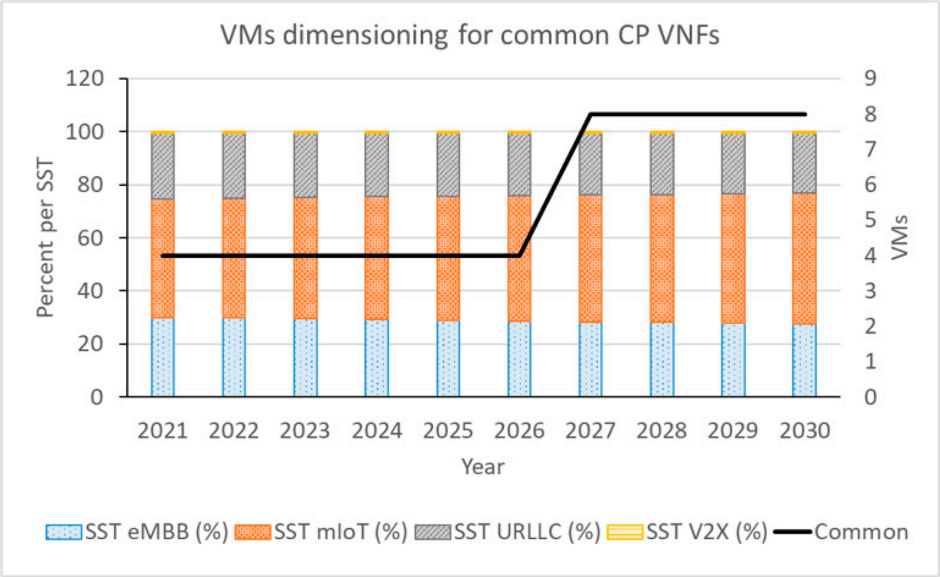


Figure 7. VMs dimensioning for common CP VNFs.

4.3. Cost Calculation and Breakdown

Different redundancies can be assumed per slice type to calculate the number of servers required for the VNFs. A redundancy of one means there are no redundant components and that the minimum number of required components are deployed, while a redundancy of two means two components are deployed with one being redundant and so on. In this evaluation, a redundancy of one is assumed for eMBB and mIoT, while two is assumed for URLLC, V2X, and any other VNFs when deployed on common servers for different slice types.

When calculating the number of servers based on required VMs, the required number of servers differ for the deployment scenario, as described in the cost model. The same server size and costs are utilized for all deployments for direct comparison. The additional infrastructure components such as required racks are based on the number of servers in each deployment type, and the number of switches is calculated as two per rack. Hence, the total CAPEX (racks, servers, switches), OPEX (maintenance, electricity for servers, switches, and electricity for cooling), and the TCO can be calculated.

The costs are compared for the three scenarios, as shown in Figure 8. Scenario A (0.9 M) is significantly cheaper, followed by B (4.8 M), while C (9.6 M) is the most expensive in terms of TCO. While the scenarios differ significantly in total costs, the ratio of cost contributors in each of the scenarios is similar; the CAPEX consists of roughly 70% servers, followed by 20% switches, and 10% racks and cabling; the OPEX consists of roughly 42% maintenance, 37% Software and VNF licenses, and 21% electricity and cooling. This is because the servers and the supporting equipment are proportional to the number of customer devices, which in turn grow at the same rate in each scenario. Thus, the number of servers and supporting equipment required by the different deployment scenarios is different, but the ratio of their contribution to overall costs is similar. The breakdown enables different stakeholders to understand the contribution ratio of different components to overall costs. Furthermore, a comparison between the costs per slice type and the cost per slice type, per device type, for each scenario is also presented, as shown in Figures 9 and 10. Scenario A is the lowest cost deployment scenario, with all resources for multiple factory customers centralized in one cloud. Scenario B is next due to the distribution of UP resources to the factories with less centralization. Scenario C has the highest costs due to the absence of centralization and sharing of resources between the factories.

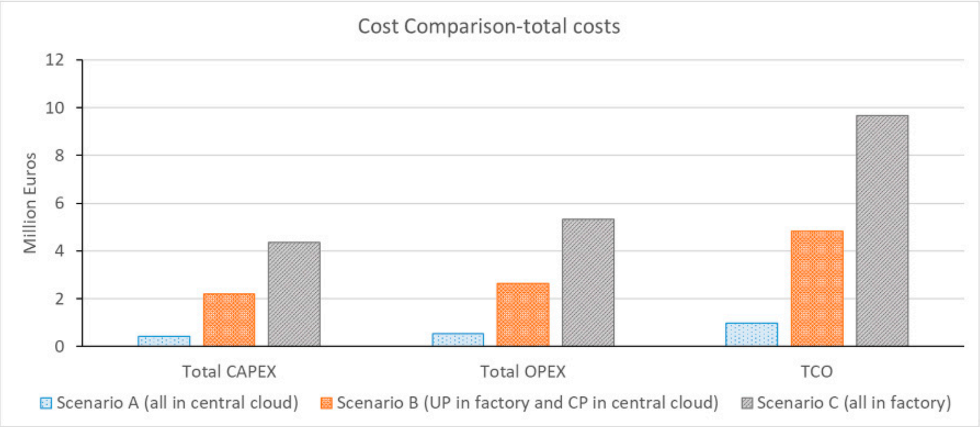


Figure 8. Cost comparison between deployment scenarios.

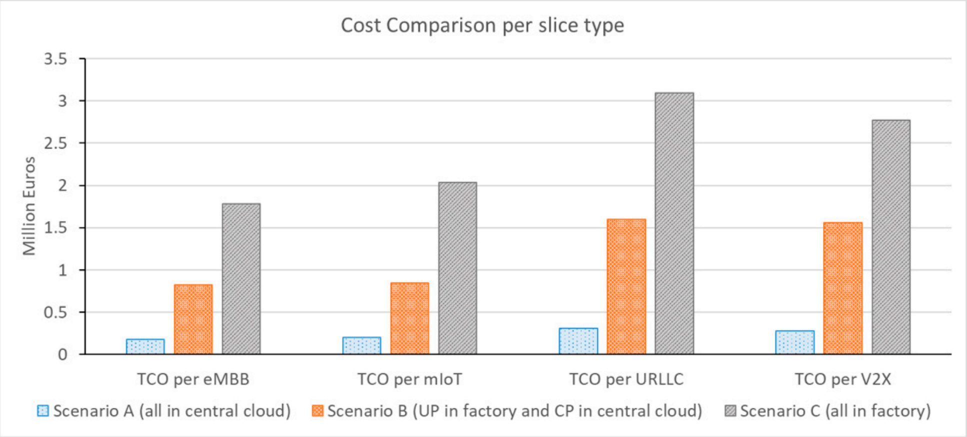


Figure 9. Cost comparison between different slice types and deployment scenarios.

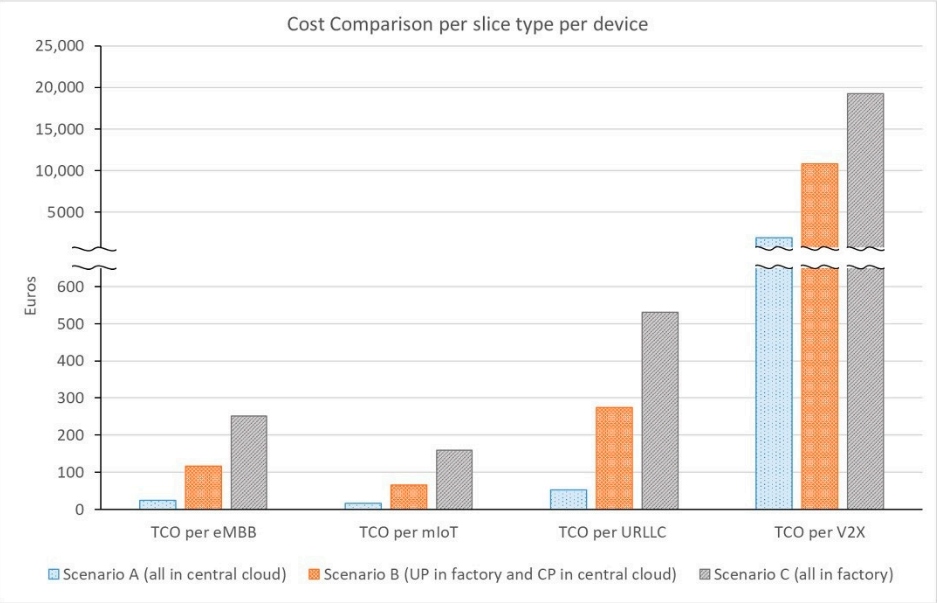


Figure 10. Cost Comparison per slice type, per device type, for different deployment scenarios.

In each of the scenarios, the cost per slice type as well as per slice type, for each device can help stakeholders identify their respective investment requirements in the virtualization infrastructure and also help structure the fees from slice customers, which in turn affect future revenues. Since the VMs are not shared between the VNFs of different slices, and if a slice requires high performance and redundant equipment for high reliability but still consists of a very low number of devices such as the V2X slice, the TCO per slice type, for each device type will be significantly higher than that for other slices, with a higher number of devices, as can be seen in Figure 9. In the case of a very low number of devices, it is preferable to centralize the resources required by multiple slice customers in the same cloud, as in scenario A, in order to minimize costs, given that the performance requirements can still be met. The results show that scale benefits through centralization enable lower costs for scenario A, and the lack of scale benefits in scenario C leads to higher costs. The MNOs, slice providers, virtualization infrastructure providers, or a joint venture between multiple stakeholders and factories can be responsible for the TCO in scenario A and the CP-related infrastructure in scenario B. The factories themselves are most suitable for the TCO in scenario C and the UP part in scenario B. Alternatively, the UP part in scenario B can be deployed by an MNO, a slice provider of virtualization infrastructure in an edge cloud strategically located near the factory premises. The cost allocation in scenario A and B can be based on the business model exercised. For example, if a slice provider provides services to a factory by using a virtualization infrastructure deployed by the MNO, then the slice provider charges the factory for the services provided, and the MNO charges the slice provider for the resources used. Alternatively, a large virtualization infrastructure provider can be responsible for the owning and operating the data centers, and offer different options as a service to MNOs, slice providers, factories, and application developers through infrastructure as well as function as a service model.

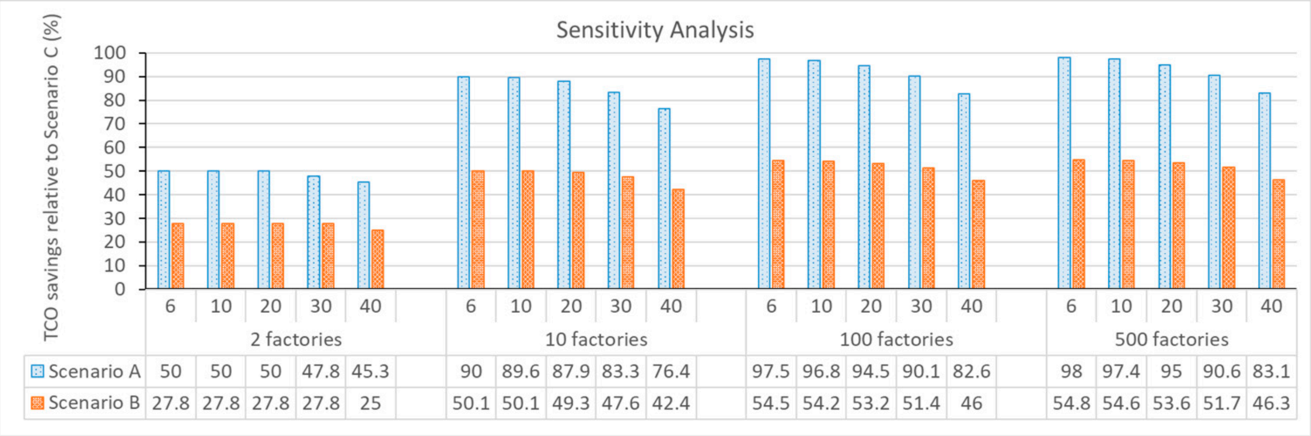
Additionally, the costs can change significantly based on the number of devices, factories, and the device growth. Likewise, there can be significant differences arising from physical location costs; a fully centralized scenario requires one large data center for multiple factories, while a fully local scenario requires smaller data centers in each factory. A fully centralized deployment can offer higher scalability for a larger customer base with significant cost savings. A fully local scenario is expensive but can offer better latency and reliability. Cost savings and technical benefits can be optimized by selecting the optimum deployment scenario for each case. The results above are calculated for one specified scenario of input requirements per slice type and can change for different inputs as dictated by each case. Furthermore, there can be a different level of isolation instead of the partial isolation assumed in the calculation when some VNFs are dedicated to each slice, while some are common between slices. These uncertainties are tackled in the following sensitivity analysis.

4.4. Sensitivity Analysis

The sensitivity analysis in cost modeling needs to account for uncertainties and various extreme scenarios. The inputs for the cost model should be varied from case to case. Firstly, the CAGR and number of factory customers are varied, and the cost savings for each of the three scenarios are compared. The calculations are performed with the growth scenarios ranging from 6% to 40%, and the number of different factories ranging from 2 to 500. Secondly, the cost of isolation is estimated for extreme cases of full isolation and no isolation between the computing resources.

4.4.1. Cost Savings

When the cost savings are calculated for different growth scenarios (6%, 10%, 20%, 30%, and 40%) vs. different number of factories (2, 10, 100, and 500), deployment scenarios A (all in the central cloud) and B (UP in the factory and CP in the central cloud) show significantly lower costs as compared to scenario C (all in factory). The lower costs are attributed to the scale benefits of centralization vs. the full distribution of the major cost components. Furthermore, the scale benefits and relative cost benefits are studied by varying the CAGR for different scenarios of device growth and for the different number of factories, as shown in Figure 11.



	2 factories					10 factories					100 factories					500 factories				
	6	10	20	30	40	6	10	20	30	40	6	10	20	30	40	6	10	20	30	40
▨ Scenario A	50	50	50	47.8	45.3	90	89.6	87.9	83.3	76.4	97.5	96.8	94.5	90.1	82.6	98	97.4	95	90.6	83.1
▨ Scenario B	27.8	27.8	27.8	27.8	25	50.1	50.1	49.3	47.6	42.4	54.5	54.2	53.2	51.4	46	54.8	54.6	53.6	51.7	46.3

Figure 11. Sensitivity analysis—TCO savings in scenario A and B over scenario C, by varying the growth of devices and number of factories.

With a single factory as the slice customer in the market, scenario A and C cost the same and both are lower than scenario B; however, in all other cases, when there is more than one factory, scenario C is the most expensive and scenario A is the least expensive. Figure 10 shows that a percentage of the cost savings relative to scenario C continues to increase with an increase in the number of factory slice customers, indicating that the higher the number of slice customers, the higher the cost savings through scenario A and B. However, for a given number of factories, lower cost savings are possible as the CAGR increases. This is due to the fact that a large factory will have less to gain in terms of cost savings in comparison to a small factory, again highlighting scale benefits. Furthermore, it should be noted that the absolute costs within each scenario are expected to grow with a higher number of factories; the percentage of cost savings relative to scenario C have a weaker trend.

Additionally, scale benefits can be achieved through volume discounts for larger volume purchases. However, some strategic uncertainties exist such as the complexity of human costs and automation levels, which can vary based on the stakeholders involved in the ecosystem. A factory might have more expertise in managing their industrial processes rather than managing the networking infrastructure, while slice providers and network operators are expected to be better suited for network management tasks. However, network slice management for the request and configuration of the slices can be customer-managed, slice provider-managed, or network operator-managed, depending on the business models. The sensitivity analysis above shows that the model is applicable to different market scenarios with different device growth, customer size, and number of customers.

4.4.2. Cost of Isolation

The difference between the cost of full isolation and that of no isolation of computing resources between the different slice types is calculated to find the cost of isolation in each deployment scenario. All VNFs are assumed to be dedicated to all slice types in the full-isolation case, while all VNFs are assumed to be common for all slice types in the no isolation case. Thus, the full and no isolation cases imply no and full sharing of infrastructure, respectively. The results are evaluated again with the same input scenario of 10 factories, the given device types, and 6% CAGR (10% for IoT devices) over the time period, as in chapter 4, but with no additional redundancy per slice types for direct comparison between deployment scenarios A, B, and C. The TCOs obtained for the three scenarios with full and no isolation cases are listed in Table 5.

Table 5. Cost of isolation.

TCO and Cost of Isolation (M Euros)	Scenario A (All in Central Cloud)	Scenario B (UP in Factory and CP in Central Cloud)	Scenario C (All in Factory)
TCO with full isolation	0.75	4.73	7.57
TCO with no isolation	0.36	1.98	2.79
Cost of isolation	0.39	2.75	4.78

The cost of isolation for the three scenarios are approximately EUR 0.39, 2.75, and 4.78 M for scenarios A, B, and C, respectively. The cost of isolation is the lowest when all virtualization infrastructure is in the central cloud, while it is the highest when all virtualization infrastructure is in each factory. The high cost of isolation in deployment scenario C also explains the high costs per slice type for each device type. Furthermore, scenario A can offer higher flexibility and scalability for supporting new and increasing number of services. Thus, involved stakeholders can achieve significant cost savings with higher flexibility and scalability through centralization. However, a local virtualization infrastructure can be preferred for reasons such as trust and security, lower end-to-end latencies, and reliability as well as considering the fact that a communication network might not comprise a huge portion of the overall expenses of a large factory's processes.

4.4.3. Resource Overbooking

The case factory requires high availability for all services; thus, the costs were calculated using the worst-case scenario in which all services are active all the time. However, when services are not used continuously, some VM consolidation can be applied by overbooking the resources. Overbooking implies that less resources are needed compared to the worst-case scenario, while still maintaining the Service Level Agreement (SLA) for every service, as the services are not active all the time [37,38]. For example, if the server supports eight VMs, but the requested number of VMs are not active all the time, then more than eight VMs can be supported per server. Thus, the number of servers and the supporting infrastructure to be dimensioned can be fewer than in a worst-case scenario. However, the statistical analysis of the required number of VMs is complicated and depends on service availability requirements and the statistical properties of service demand. If the services are needed only occasionally, a high overbooking can be applied. On the other hand, if services are active more often, then low overbooking can be applied. For the sensitivity analysis, the original input scenario from chapter four is utilized, and the TCO with different average overbooking ratios for each deployment scenario is calculated, as shown in Figure 12. Here, an overbooking ratio of one means no overbooking is applied, while a smaller value indicates a higher overbooking ratio. Significant cost savings can be achieved in each of the deployment scenarios with resource overbooking. A higher overbooking is suitable and may provide significant cost savings if less services are active at the same time. The risk of overbooking is that if the actual service demand is underestimated, the quality of service (for example, waiting times and available throughputs) would deteriorate, resulting in SLA violation. Thus, overbooking is a reasonable approach when there is sufficient information about service demand based on measurements made in a real environment, making sure that the SLAs are not violated.

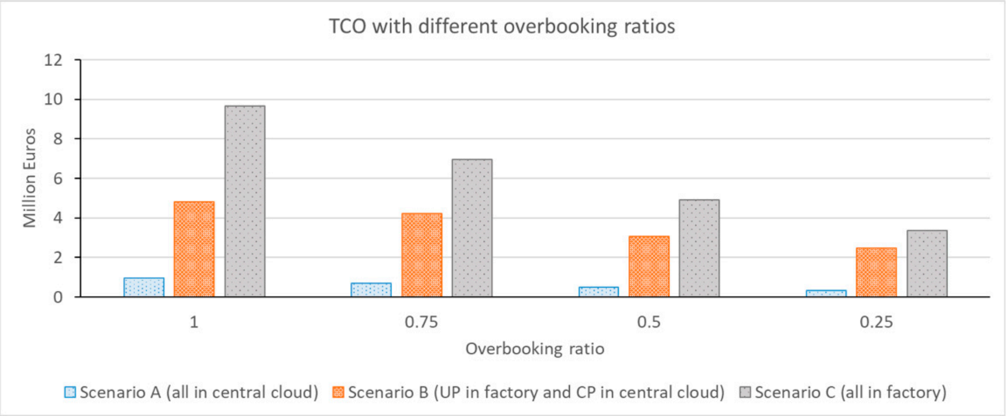


Figure 12. TCO comparison with different overbooking ratios and deployment scenarios.

5. Conclusions

A cost model is developed for virtualization infrastructure for the next-generation virtualized 5G core based on network slice demand and is applied to a smart factory as a case study. The costs for different deployment scenarios are analyzed, with the highest cost savings in a fully centralized scenario, followed by a distributed scenario with a local user plane and a centralized control plane, followed by the most expensive fully local scenario. Cost savings increase with the number of factories and the growth of the number of devices. A large factory with a large number of devices has less to gain in terms of relative cost savings than a small factory by centralizing resources in a service provider's central cloud and thus might prefer a fully local deployment, enjoying scale benefits of its own. While a specified smart factory case is evaluated, a sensitivity analysis is performed by varying the expected growth of devices and the number of slice customers to be provisioned. Additionally, the difference between full-isolation and no-isolation cases is evaluated for the above scenarios, clearly depicting each scenario's different cost of isolation. The results indicate that the model can be applied to different markets, customer sizes, number of customers, and deployment scenarios. Furthermore, the cost breakdown in terms of cost for every slice type and per device offers insights for stakeholders about the required investments for every slice type and device. While a large factory can prefer a completely isolated local network for its reliability instead of cost savings, a large network operator or slice provider serving multiple factories should still consider cost savings as an important driver.

Author Contributions: Conceptualization, J.S.W., H.H., K.K., H.F., S.Y. and M.M.-B.; data curation, J.S.W. and S.Y.; formal analysis, J.S.W., H.H. and K.K.; methodology, J.S.W., H.H., K.K., H.F., S.Y. and M.M.-B.; writing—original draft, J.S.W.; writing—review and editing, J.S.W., H.H., K.K., H.F., S.Y. and M.M.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported and funded by Business Finland in the 5G VIIMA project (grant no. 6430/31/2018).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Yi, B.; Wang, X.; Li, K.; Das, S.K.; Huang, M. A comprehensive survey of Network Function Virtualization. *Comput. Netw.* **2018**, *133*, 212–262. [CrossRef]
2. Barakabitze, A.A.; Ahmad, A.; Mijumbi, R.; Hines, A. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Comput. Netw.* **2020**, *167*, 106984. [CrossRef]
3. 3GPP. *TS 23.501 V17.0.0 System Architecture for the 5G System*; 3GPP: Sophia Antipolis Cedex, France, 2021.
4. Pivoto, D.G.S.; de Almeida, L.F.F.; da Rosa Righi, R.; Rodrigues, J.J.P.C.; Lugli, A.B.; Alberti, A.M. Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review. *J. Manuf. Syst.* **2021**, *58*, 176–192. [CrossRef]
5. Virdis, A.; Nardini, G.; Stea, G.; Sabella, D. End-to-End Performance Evaluation of MEC Deployments in 5G Scenarios. *J. Sens. Actuator Netw.* **2020**, *9*, 57. [CrossRef]
6. Global Industry Analysts, Inc. Network Function Virtualization—Market Study by Global Industry Analysts, Inc. Available online: https://www.strategyr.com/market-report-network-function-virtualization-forecasts-global-industry-analysts-inc.asp (accessed on 9 December 2020).
7. Mind Commerce. *5G Network Slicing by Infrastructure, Spectrum Band, Segment, Industry Vertical, Application and Services 2020–2025*; Mind Commerce: Seattle, WA, USA, 2020.
8. Ericsson. *5G for Business: A 2030 Market Compass*; Ericsson: Stockholm, Sweden, 2019.
9. Dorsch, N.; Kurtz, F.; Wietfeld, C. On the economic benefits of software-defined networking and network slicing for smart grid communications. *Netnomics Econ. Res. Electron. Netw.* **2018**, *19*, 1–30. [CrossRef]
10. Bouras, C.; Ntarzanos, P.; Papazois, A. Cost modeling for SDN/NFV based mobile 5G networks. In Proceedings of the 2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Lisbon, Portugal, 18–20 October 2016; pp. 56–61.
11. Rokkas, T.; Neokosmidis, I.; Xydias, D.; Zetserov, E. TCO savings for data centers using NFV and hardware acceleration. In Proceedings of the Joint 13th CTTE and 10th CMI Conference on Internet of Things—Business Models, Users, and Networks, Copenhagen, Denmark, 23–24 November 2017; Volume 2018, pp. 1–5.
12. Yao, F.; Wu, J.; Venkataramani, G.; Subramaniam, S. A comparative analysis of data center network architectures. In Proceedings of the 2014 IEEE International Conference on Communications, ICC 2014, Sydney, NSW, Australia, 10–14 June 2014; pp. 3106–3111.
13. Herker, S.; An, X.; Kiess, W.; Kirstadter, A. Evaluation of data-center architectures for virtualized Network Functions. In Proceedings of the 2015 IEEE International Conference on Communication Workshop, ICCW 2015, London, UK, 8–12 June 2015; pp. 1852–1858.
14. Morgan, J.; Halton, M.; Qiao, Y.; Breslin, J.G. Industry 4.0 smart reconfigurable manufacturing machines. *J. Manuf. Syst.* **2021**, *59*, 481–506. [CrossRef]
15. 3GPP. *TS 29.244 V17.0.0 Interface between Control Plane and the User Plane Nodes*; 3GPP: Sophia Antipolis Cedex, France, 2021.
16. 3GPP. *TS 38.300 V16.5.0 NR; NR and NG-RAN Overall Description; Stage-2*; 3GPP: Sophia Antipolis Cedex, France, 2021.
17. GSMA. *NG.116-Generic Network Slice Template*; GSMA: London, UK, 2019.
18. Baek, S.; Kim, D.; Tesanovic, M.; Agiwal, A. 3GPP new radio release 16: Evolution of 5G for industrial internet of things. *IEEE Commun. Mag.* **2021**, *59*, 41–47. [CrossRef]
19. Aditya, P.; Akkus, I.E.; Beck, A.; Chen, R.; Hilt, V.; Rimac, I.; Satzke, K.; Stein, M. Will Serverless Computing Revolutionize NFV? *Proc. IEEE* **2019**, *107*, 667–678. [CrossRef]
20. Benedetti, P.; Femminella, M.; Reali, G.; Steenhaut, K. Experimental Analysis of the Application of Serverless Computing to IoT Platforms. *Sensors* **2021**, *21*, 928. [CrossRef] [PubMed]
21. Chowdhury, S.R.; Salahuddin, M.A.; Limam, N.; Boutaba, R. Re-architecting nfv ecosystem with microservices: State of the art and research challenges. *IEEE Netw.* **2019**, *33*, 168–176. [CrossRef]
22. Zhao, Y.; Wang, W.; Li, Y.; Colman Meixner, C.; Tornatore, M.; Zhang, J. Edge Computing and Networking: A Survey on Infrastructures and Applications. *IEEE Access* **2019**, *7*, 101213–101230. [CrossRef]
23. Bruschi, R.; Davoli, F.; Lombardo, C.; Sanchez, O.R. Evaluating the Impact of Micro-Data Center (µDC) Placement in an Urban Environment. In Proceedings of the 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks, NFV-SDN 2018, Verona, Italy, 27–29 November 2018.
24. Herker, S.; An, X.; Kiess, W.; Beker, S.; Kirstaedter, A. Data-center architecture impacts on virtualized network functions service chain embedding with high availability requirements. In Proceedings of the 2015 IEEE Globecom Workshops (GC Wkshps), San Diego, CA, USA, 6–10 December 2015.
25. 3GPP. *TS 22.104 V18.0.0 Service Requirements for Cyber-Physical Control Applications in Vertical Domains*; 3GPP: Sophia Antipolis Cedex, France, 2021.
26. 5GACIA. *Key 5G Use Cases and Requirements—White Paper*; 5GACIA: Frankfurt am Main, Germany, 2020.
27. Walia, J.S.; Hämmäinen, H.; Kilkki, K.; Yrjölä, S. 5G network slicing strategies for a smart factory. *Comput. Ind.* **2019**, *111*, 108–120. [CrossRef]
28. Grasso, C.; Schembra, G. A Fleet of MEC UAVs to Extend a 5G Network Slice for Video Monitoring with Low-Latency Constraints. *J. Sens. Actuator Netw.* **2019**, *8*, 3. [CrossRef]

29. Tonini, F.; Khorsandi, B.; Amato, E.; Raffaelli, C. Scalable Edge Computing Deployment for Reliable Service Provisioning in Vehicular Networks. *J. Sens. Actuator Netw.* **2019**, *8*, 51. [CrossRef]
30. Siriwardhana, Y.; Porambage, P.; Liyanage, M.; Walia, J.S.; Matinmikko-Blue, M.; Ylianttila, M. Micro-Operator driven Local 5G Network Architecture for Industrial Internet. In Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakech, Morocco, 15–18 April 2019.
31. Lahteenmaki, J.; Hammainen, H.; Zhang, N.; Swan, M. Cost modeling of a network service provider cloud platform. In Proceedings of the 2016 IEEE International Conference on Cloud Engineering Workshops, IC2EW 2016, Berlin, Germany, 4–8 April 2016; pp. 148–153.
32. Statistics Finland Energy Prices. 2021. Available online: https://www.stat.fi/til/ehi/2020/04/ehi_2020_04_2021-03-11_tie_001_en.html (accessed on 30 March 2021).
33. Dieye, M.; Ahvar, S.; Sahoo, J.; Ahvar, E.; Glitho, R.; Elbiaze, H.; Crespi, N. CPVNF: Cost-Efficient Proactive VNF Placement and Chaining for Value-Added Services in Content Delivery Networks. *IEEE Trans. Netw. Serv. Manag.* **2018**, *15*, 774–786. [CrossRef]
34. Barroso, L.A.; Clidaras, J.; Hölzle, U. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 2nd ed.; Morgan & Claypool Publishers: San Rafael, CA, USA, 2013; Volume 24, ISBN 9781627050098.
35. Statista Lot Connected Devices Worldwide 2019–2030. 2020. Available online: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/ (accessed on 31 March 2021).
36. 5GACIA. *A 5G Traffic Model for Industrial Use Cases*; 5GACIA: Frankfurt am Main, Germany, 2019.
37. Tomás, L.; Tordsson, J. Improving cloud infrastructure utilization through overbooking. In Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference, Miami, FL, USA, 5–9 August 2013; ACM Press: New York, NY, USA, 2013; p. 1.
38. Son, J.; Dastjerdi, A.V.; Calheiros, R.N.; Buyya, R. SLA-Aware and energy-efficient dynamic overbooking in SDN-based cloud data centers. *IEEE Trans. Sustain. Comput.* **2017**, *2*, 76–89. [CrossRef]