ACCEPTED FROM OPEN CALL

Definition and Application of the Scalability Index: Example of 5G Network Slicing

Tobias Hossfeld, Poul E. Heegaard, and Wolfgang Kellerer

The authors illustrate and assess the scalability index within the context of a 5G core network slicing use case.

Abstract

Scalability is of key interest for communication networks and systems, and frequently mentioned in research. However, a framework for benchmarking the scalability of two or more systems is missing. Therefore, we establish a Scalability Index (SI) to measure one system's scalability in comparison to another system, or a hypothetical one, such as a linear or ideal system. In this article, we illustrate and assess the SI within the context of a 5G core network slicing use case. We provide practical insights on utilizing the SI, and offer recommendations and guidelines for its application.

Introduction

The assessment of communication networks and systems encompasses various facets, including performance, efficiency, elasticity, flexibility, and scalability. Notably, scalability is a frequently cited concept in the literature, often expressed with vague statements like "*The system scales well" or "Our approach exhibits superior scalability compared to previous methods.*" These statements lack precision and fail to offer substantive insights. Literature often provides performance curves as a mean of a scalability analysis. For the comparison of two systems or system configurations, the corresponding performance curves — also known as system functions — are then provided, as illustrated exemplary in Fig. 1. In that example, the two system functions intersect or overlap. Therefore, it remains unclear which system really scales better by just comparing the performance curves visually.

In our publication [1], we establish a rigorous definition of scalability. Our objective is to provide a *single-value* scalability index (SI), enabling quantifying, comparing, and ranking the scalability of two or more systems. Hence, it is necessary to condense the system function into a single value. This is accomplished by employing integral measurements, that is, by calculating the area under the curves. The SI is then defined as the ratio of the integral measurements. Such a desired single-value scalability index is only provided in a few works of literature. In fact, the term scalability is misused and partly considers different aspects like efficiency or performance and the corresponding efficiency- or performance-curves, which may only give qualitative insights. Our framework [1] allows a generalization to a single-value scalability index.

The key contributions of this article are as follows. For completeness, we revisit the previously introduced scalability index [1] and differentiate scalability from related terms (efficiency, elasticity, flexibility). We apply the SI to evaluate the scalability of a 5G core network slicing approach based on a performance study from literature [2]. In this use case, we demonstrate how the scalability index (SI) supports decisions on the configuration of systems, particularly when decision-making is challenging due to system dynamics and the parameters being investigated. For the 5G network slicing use case, different configurations of 5G network slicing are considered and how they affect the average response time and throughput. A combined "productivity" measure [3] aggregates the two measures as well as the cost for the configuration into a key quality indicator (KQI). The KQI is the target measure of the scalability analysis for varying network load. As we will see, the KQI curves are intersecting and the decision for a concrete configuration is difficult in practice. In contrast, the SI guides the decision process and clearly identifies the best configuration in terms of KQI for a given network load scenario. Finally, we offer insights into practical guidance for utilizing the scalability index, particularly with regard to defining reference systems, target functions, and weighted parameter ranges.

Definition of the Scalability Index

In our previous work [1], we defined the scalability index as a single-value quantitative measure of a system. To this end, an arbitrary target measure of interest is considered, like the average response time of the system, over a certain parameter, for example, the demand level or the traffic load in a system. As a result, we obtain a system function, also referred to as a performance curve or target measure function, see Fig. 1. Thereby, the parameter range may be weighted to define the importance or the frequency of some parameter settings in practice. Finally, a reference system is required to relate the scalability of the system under test (SUT) to the reference system. For example, as a reference system, a linear function or the optimal performance curve may be used to test for linear scalability or to quantify how far away the SUT is from the optimal system, respectively. In practice, we are especially interested to benchmark two systems or two system configurations against each other by using one of them as

Tobias Hossfeld is with University of Würzburg, Germany; Poul E. Heegaard is with NTNU — Norwegian University of Science and Technology, Norway; Wolfgang Kellerer is with TU Munich, Germany.

ISSN: 0163-6804
Digital Object Identifier: 10.1109/MCOM.001.2300530

a reference system. The components involved in measuring and quantifying scalability include:

- A system function $f(x)$ quantifies an *arbitrary target measure of interest* for the system $\mathcal{F}$. The system function is also referred to as target measure function or performance curve.
- An arbitrary *reference system* $\mathcal{H}$ with a target reference function $h(x)$ is used for comparison.
- A *parameter range* $x_0 \le x \le x_1$ is weighted using a weight function $w(x)$ to define the importance or frequency of parameter settings.
- *Weighted integral measurements* F and H of the target measure for the system and the reference consider the weighted parameter range under investigation. The integral measurement is simply the weighted area under the corresponding system curve.

Definition 1 (Scalability): *The ability of a system to perform as well or better as a reference system regarding a target measure within a defined weighted parameter range.*

To quantify the scalability as a single value SI, we relate the integral measurements of the system under test and of the reference system. The auxiliary variable γ is employed, where $\gamma = 1$ denotes the "goodness" of the target measure, while $\gamma = -1$ signifies its "badness." For example, average response times indicate "badness," since a higher average response time means worse system performance. Throughput or availability are examples for "goodness," since greater throughput or availability means better systems. A formal definition of the SI is as follows.

Definition 2 (Scalability Index): *Quantification of the scalability of a system $\mathcal{F}$ with respect to a reference system $\mathcal{H}$ as the ratio of the integral measurements F and H over the weighted parameter range and the goodness indicator γ.*

$$SI = \left(\frac{F}{H}\right)^{\gamma} = \left(\frac{\int_{x_0}^{x_1} f(x) \cdot w(x) dx}{\int_{x_0}^{x_1} h(x) \cdot w(x) dx}\right)^{\gamma} \quad (1)$$

When using the ideal system as reference, a SI of 1 means the same performance of the SUT as the ideal system over the weighted parameter range. The worst SI is 0. When benchmarking two systems, a SI larger than 1 shows that the system $\mathcal{F}$ scales better than the system $\mathcal{H}$.

As we have shown in [1], our SI represents a broad and encompassing perspective on scalability metrics, encompassing and extending existing approaches in the literature. Our definition aligns closely with the one proposed in [4], which also accounts for an optimal reference system. However, in practical communication networks and systems, determining the optimal system behavior can be challenging. As shown for our 5G network slicing use case, we emphasize the need to adapt the target measure to suit the context, incorporating factors like KQIs, as well as the weighted parameter range, for example, for accommodating different load scenarios.

Performance, Efficiency, Elasticity, Flexibility, Scalability

Performance in communication networks and systems refers to the quality, speed, and efficiency of data transmission and reception, typically measured by factors like throughput, latency, and packet loss. To derive the system curves, we assess the performance in a quasi-stationary system, for example, long term (steady state) response times in a system with nearly constant arrival and service rates. One important aspect is *efficiency*, for example, the ratio of data plane traffic to overhead, the ratio of demand to supply, or energy efficiency as the ratio of completed requests to the maximum possible with the same energy consumption [5]. For distributed systems, additional efficiency metrics involve evaluating the workload handled by each machine. In contrast, scalability encompasses more than just the ability to function; it also involves efficient operation while maintaining an acceptable quality of service across various configurations [3].

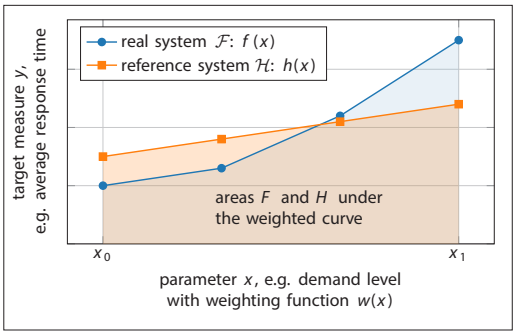

FIGURE 1. Scalability index as ratio $(F/H)^{\gamma}$ of the areas under the weighted target measure curves. The goodness indicator shows whether the target measure reflects "goodness" ($\gamma = 1$, e.g., throughput) or "badness" ($\gamma = -1$, e.g., response times). The system $\mathcal{F}$ under test is related to the reference system $\mathcal{H}$.

Elasticity considers the system's dynamic changes and measures its ability to adjust rapidly. Reference [6] describes elasticity as the system's ability to automatically allocate and release resources to closely match the current demand. In communication networks, elasticity refers to the network's ability to adjust its operation and redistribute resources in response to changes in traffic patterns and service requirements. This adaptation can involve reallocating computational and communication resources, especially in software-defined and virtualized networks like 5G systems [7].

Figure 2 visualizes elasticity by quantifying the system's resource demand and supply over time. In this example, a 5G network function (NF) is considered. An NF instance runs in a dedicated virtual machine. An elastic system adjusts the number of NF instances according to the resource demand. The number of NF requests is the resource demand, while the number of activated NF instances reflects the system's supply. During the considered time frame in Fig. 2, the response time to execute a NF is measured for all NF requests and aggregated into a single performance measure, for example, average response time. In addition, the load is quantified as the number of requests per time frame. Observing the system over a significantly long period of time (e.g., one hour) leads to quasi-stationary results. Then, the tuple (*load*, *avg. response time*) quantifies the system's performance for that load and depicts a single point in the related performance curve in Fig. 3. Measurements of several hours yield different measurement points, resulting in the performance curve (Fig. 3) depending on the

Performance in communication networks and systems refers to the quality, speed, and efficiency of data transmission and reception, typically measured by factors like throughput, latency, and packet loss. To derive the system curves, we assess the performance in a quasi-stationary system, for example, long term (steady state) response times in a system with nearly constant arrival and service rates.

Note that the time dynamics and the elasticity of the system are captured by applying some statistics like the average response time or the 99%-quantile of the response time.

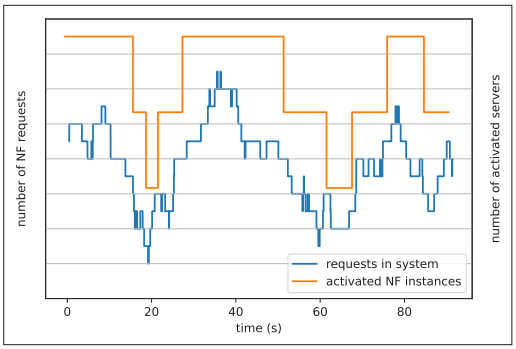

FIGURE 2. Elasticity considers the system's resource demand (number of NF requests) and supply (number of NF instances) over time.

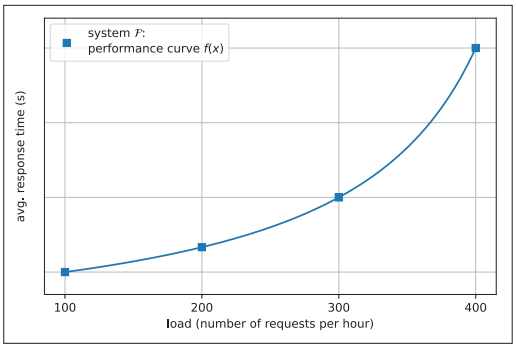

FIGURE 3. The performance curve quantifies the system's quasi-stationary performance by considering a sufficiently long time interval (here: one hour). The entire system behavior in Fig. 2 results in a single point (load vs. average response time) in the performance curve. Performance curves are evaluated on longer time scales than elasticity and capture several points.

load. Such performance curves are important to understand how the system behaves and what is the performance for specific areas of the operating range. The performance curve is also the required input for our scalability analysis.

Note that the time dynamics and the elasticity of the system are captured by applying some statistics like the average response time or the 99%-quantile of the response time. In practice, the probability that the response time is above a certain threshold may be relevant to quantify service-level agreement (SLA) violations. In any case, an appropriate target measure of interest needs to be defined.

Flexibility is a vital property examined in communication network systems and refers to the network's ability to adapt, modify, or reconfigure its infrastructure, protocols, or services in response to changing requirements, conditions, or demands. [8] introduces a model to measure network flexibility by quantifying its ability to handle a specific subset of potential changes. The flexibility framework considers the resulting adaptation of the system to any of the changes, the adaptation time, as well as the required efforts and costs. Furthermore, the adaptation of the system may lead or not lead to a proper system, for example, not being able to fulfill certain SLAs. While elasticity considers how the system's resources are adapted according to changing service demands over time, flexibility considers all possible changes of the system. A change of the system can happen on different time-scales, for example, due to changing service demands. However, also other changes like failures in the network may be considered to test the flexibility of the system to such changes.

Scalability, in essence, refers to a system's ability to handle increasing workloads over an extended period by employing additional resources or adapting configurations, even if it means adjusting quality of service (QoS). In cloud computing, it specifically pertains to expanding software service capacity as needed due to prolonged increased demand. While elasticity addresses immediate, on-demand scaling, scalability assesses performance over a longer period, considering various demand scenarios [4, 9], as depicted in Fig. 3. Therefore, scalability's primary goal is not to measure the speed, frequency, or precision of scaling actions [6], but considers a quasi-stationary system with steady-state performance. Hence, scalability must take into account the system's performance or any other desired target function across various and pertinent scenarios. Figure 3 serves as the essential input for a scalability assessment.

Use Cases of a Scalability Analysis

Literature often provides such performance curves for a scalability analysis. To address the challenge of comparing system scalability when their performance curves intersect, we propose a single-value SI. In communication networks, dealing with changing needs, diverse communication methods, and large-scale operations present considerable challenges. Meeting strict quality demands that can change dynamically is essential. These advancements were explored in [10], focusing on adaptive and scalable communication networks. Assessing a solution's adaptivity, the elasticity [6] quantifies the system's ability to adjust in a timely way to dynamic changes, while the flexibility framework [8] quantifies to which degree all possible changes are handled. For benchmarking the scalability of alternative solutions, our SI framework introduces a novel single-value index applicable to communication networks. The use cases for a scalability analysis range from the performance of the control plane and data plane, the KPIs and KQIs of communication services, as well as their costs like energy consumption. For instance, [11] analyzes the workload-to-overhead ratio for control plane efficiency in software-defined networks. To fit this into a scalability framework, the workload-to-overhead ratio is the target measure, depending on the system load. [12] examines big data processing system scalability in cloud environments, including various scalability types like linear, sublinear, and super-linear. *Linear scalability* is a vital analysis method, focusing on systems with a linear relationship between the target metric and the parameter. Other exemplary use cases and their scalability analysis are investigated in [1]:

- Scalability of an IoT load balancer depending on system load
- Availability of a communication system depending on the number of nodes and system structure
- Scalability comparison of different location selection mechanisms in fog computing regarding delays and energy consumption
- Comparison of time-sensitive networking (TSN) mechanisms in terms of the number of deployed streams while guaranteeing upper delay bounds.

Scalability of 5G Core Network Slicing

For demonstrating the application of the scalability index, we consider 5G core network slicing based on an approach and its performance evaluation results by Arteaga, Ordoñez, and Rendon [2]. The 5G architecture defines several network functions (NFs) which may run on virtual machines (VMs). The intention behind virtualized NFs is to have greater flexibility and scalability. To examine the latter, we apply the scalability framework on the session establishment of 5G network slicing. Creating network slices in the 5G core network entails assembling network functions to execute the signaling processes. For the session establishment, the NF Repository Function (NRF), the Session Management Function (SMF), and the Access and Mobility Management Function (AMF), compose the corresponding service chain, see [2] for more details.

The performance of the 5G session establishment is investigated depending on the network load in terms of number of user equipments (UEs). In particular, the throughput of session establishment as well as the average response times are analyzed for different system configurations. These configurations consider the number of processors which can be used to handle the NRF, SMF, and AMF. The more resources, that is, processors per network function instance, are provided, the higher the maximum throughput that can be achieved. This *vertical scaling* ("*scaling up*") improves the performance, as expected. Depending on the number of processors, up to 23000, 42000, and 85000 concurrent users can be dealt with one, two, and four processors when a single NF instance is used before the session establishment starts to drop incoming requests. Similarly, the average response times improve when more processors are used per NF instance. Scaling up leads to the expected performance improvement.

The same is true for *horizontal scaling* ("*scaling out*") by adding more instances for each of the network functions NRF, SMF, AMF. To be more precise, in every configuration (number of NF instances and number of processors), the throughput increases with the number of users until it reaches its maximum. This results from parallel processing, and many processors support a high number of concurrent users in a short time.

System Curves

The average response times also improve when scaling out the number of instances for NRF and SMF, see Fig. 4. The average response times are smaller for the same number of UEs if more NF instances are available. Up to a certain load, which is roughly 80,000 UEs for two AMF instances, the scale-out improves the average response time. However, more load can now be handled by the system with two AMF instances. As a result, the NRF and SMF need to handle more traffic load. Exceeding that point (e.g., 90,000 UEs), the SMF and NRF get the bottleneck again. As a result, the average response time of the system with two instances per NF under high load (e.g., 90,000 UEs) is the same as the average response time of the system with only one instance per NF (e.g., for 45,000 UEs). In other words, the maximum average response time when having two

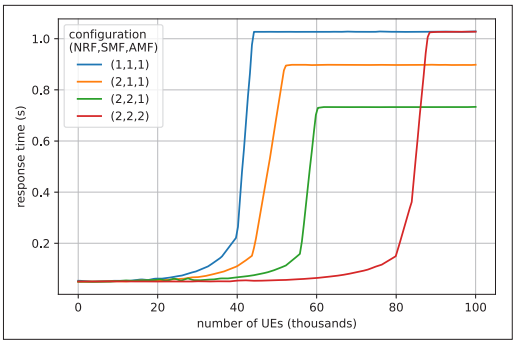

FIGURE 4. For 5G network slicing, average response times for the 5G session establishment are depicted depending on the number of NF instances and the number of UEs (load). Two processors are used per NF instance. A configuration is specified by the number of instances for NRF, SMF, AMF.

NF instances per NRF, SMF, *and* AMF is as high as having a single instance only. Scaling out the AMF means that the AMF sends a high number of requests to the other NFs. To avoid these bottlenecks, SMF and NRF must also be scaled-out accordingly. This discussion shows that the analysis of the performance curves is crucial to understand the performance for specific areas of the operating range. Based on the performance curve, it can be concluded for which concrete load situation (e.g., 90,000 UEs) which configuration works best. Thus, for a *single parameter value*, the performance curve provides the answer to which configuration works best.

If for any arbitrary parameter value, one configuration leads to a better performance than another configuration, then there is no need to conduct the scalability analysis for deciding the better configuration. In other words, if two performance curves are not intersecting, the better configuration is immediately clear, see for example the configuration (1, 1, 1) and the configuration (2, 2, 1) in Fig. 4.

Weighted Parameter Range

However, in real systems, there is a *parameter range of interest* and the best configuration needs to be selected for the entire range, not only for a single value. The SI index guides the decision which configuration works best over the entire parameter range. From the performance curve in Fig. 4, it is difficult to decide whether the configuration (2, 2, 2) outperforms the configuration (2, 2, 1) for high load scenarios, since the curves are intersecting. Hence, for the decision of which configuration is appropriate, it is of utmost importance to consider the load situation and the entire parameter range. In our scalability analysis framework, the parameter under investigation is the network load expressed by the number of UEs. Thereby, the probability of occurrence of a particular load situation is considered. This can be achieved by using the probability of a network load as a weight function in the analysis.

Appropriate Target Measure

The average response time is, however, only one aspect of interest. If the processing capacity of the NF instances is reached, incoming requests will be dropped, which affects the throughput of NF requests. To deal with multi-objectives, we

The performance of the 5G session establishment is investigated depending on the network load in terms of number of user equipments (UEs).

The productivity measure is a key quality indicator which assesses the quality of the system in terms of throughput and response time, while taking into the account the costs of the system, which is the number of processors per instance.

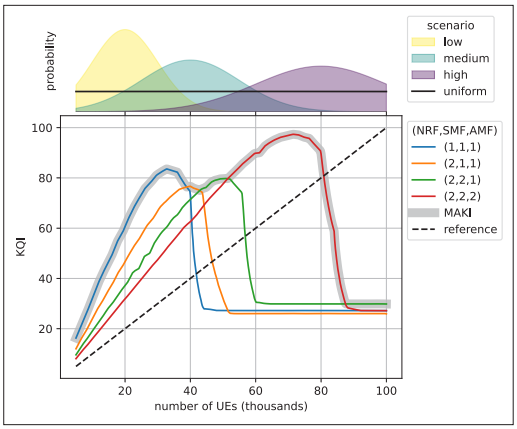

FIGURE 5. For 5G network slicing, the KQI is a function of response time, throughput, and costs. The number of NF instances are varied, and two processors are used per NF instance. As a reference system, a linear function is considered to investigate linear scalability. The MAKI approach [13] represents a perfect adaptive system, which is selecting the best configuration based on the current load.

may define an appropriate target measure which combines several measures of interest. As the target function of their analysis, Arteaga, Ordoñez, and Rendon [2] use the so-called *productivity*, which was proposed by Jogalekar and Woodside [3]. The productivity of a system configuration is defined as the value-throughput of the system divided by the costs for that particular configuration. The value-throughput is the product of the system's throughput, which is the number of session establishments per second for this use case, and the value of that function. In our case, the value is quantified as the relation between a predefined target response time and the average response time. This value therefore reflects to which degree an SLA is fulfilled. As an alternative, we simply consider an SLA that contains multiple performance and dependability attributes and their threshold values. The probability of SLA violations is a relevant KQI in practice.

The productivity measure is a key quality indicator which assesses the quality of the system in terms of throughput and response time, while taking into the account the costs of the system, which is the number of processors per instance. Figure 5 shows the KQI depending on the number of UEs for different configurations. Note that we normalized the KQI to have values between 0 and 100. In particular, the number of instances per NF is varied, while two processors are used per instance. The critical point is now to decide which configuration is best. Just looking at those KQI functions for the different configurations does not allow deriving the best configuration easily. For a concrete number of UEs (*single parameter value*), there is always one configuration which is best. However, the number of UEs may change over time. In our scalability analysis, we therefore consider the probability that a load situation will manifest and therefore consider the probabilities over a parameter range. The scalability index then clearly advises which configuration scales best. While [2] draw their conclusions based on such KQI curves, we advance the analysis by providing this single-value scalability index SI.

Reference System

As a reference system, we consider linear scalability. However, for the decision of which system is best for a certain load scenario, the reference system is not relevant, but we can simply rank the configurations according to the SI. The ingredients of the SI are as follows and are visualized in Fig. 5.

Scalability Analysis

- Target measure: KQI curve in Fig. 5
- Parameter range: number of UEs [0; 100,000]
- Weight: probability of the number of UEs for a particular load scenario (low, medium, high, uniform load)
- Reference system: linear system (dashed line)

Figure 6 shows the scalability index for different load scenarios, which are illustrated in Fig. 5. As we can see, the SI allows drawing conclusions for the different load scenarios. The higher the SI, the better the system scales in comparison to the linear reference system. A SI of 1 means that the system scales like the linear system, and we observe linear scalability over the weighted parameter range. We emphasize that the SI considers the entire (weighted) parameter range. For example, the KQI curves in Fig. 5 show that the configuration (2, 2, 2) has a lower KQI than the configuration (2, 2, 1) for the largest parameter value x = 100,000. Still, the SI over the entire parameter range in the high load scenarios indicates to use the configuration (2, 2, 2). Therefore, the SI is beneficial in practice to decide in such situations where curves are intersecting and weighted according to the probabilities or importance of parameter settings.

In addition, we compare the SI to a perfectly adaptive system. In general, an adaptive system aims at changing the configurations dynamically based on the current load (Fig. 2), such that the best KQI is achieved [10]. The goal of the so-called MAKI approach ("Multi-Mechanisms Adaptation for the Future Internet") is to deploy adaptive and scalable systems, see for example [13]. The key concept of MAKI are transitions, which is a paradigm in communication systems where existing communication mechanisms like services and protocols are replaced with functionally equivalent ones to improve quality. The goal is to achieve a smooth and consistent quality in the communication system. Transitions and system adaptations aim to optimize the overall system performance. Assuming immediate and perfect transitions between the system configurations, the maximum KQI over the different configurations for a given network load is obtained, see Fig. 5. Obviously, a perfectly adaptive system yields the best scalability and has the highest SI, see Fig. 6. The perfect system is a desired reference system to understand how far a realization of a system is away from the perfect system. The SI quantifies then the gap to the perfect system.

Conclusions and Practical Guidelines

Our definition of the *scalability index* enables the quantification of how a system or communication network scales concerning a reference system, facilitating comparisons with optimal systems, system benchmarking, and ranking of systems. In assessment of scalability, several factors must be

considered, and we offer practical guidelines by reexamining the components of this scalability index.

The Target Measure: The system function, which quantifies the target measure based on a specific parameter, plays an important role in evaluating system scalability, as diverse target measures can lead to different scalability assessments. Thus, defining the target measure, including considerations of SLAs, is a fundamental and primary step in assessing system scalability. In practice, it may be important to assess multiple target functions of interest like delays, energy consumption or availability using the SI to understand how a system scales across different dimensions. Combining these measures into a single utility function, such as Productivity [3], can lead to different results and conclusions than just considering a single objective. In literature, there are several approaches how to tackle this multi-objective problem. These include Pareto optimization techniques and strategies for exploring the Pareto front. For instance, multi-objective ranking methods are common, such as the weighted sum method. Another approach is the constraint method, which incorporates constraints to signify the significance of each objective. Still, KQIs may be available in practice and then the recommended choice of the target measure. In real-world scenarios, there are additional considerations to address. It's essential to integrate information into the target function that accounts for instances where the system might not operate properly with certain parameter settings or configurations.

The Parameter Range: The assessment of scalability requires the consideration of all relevant parameter settings to ensure not only the ability to operate but to do so efficiently and with the desired quality of service across the specified range of configurations, making it essential for the scalability index to encompass the *parameter range of interest* for meaningful conclusions. In real scenarios, where complete information about system behavior over the entire parameter range may be lacking, methods like interpolation of measurement points, such as piecewise linear functions, or predictions of future system behavior become crucial for obtaining the required system function input to compute the scalability index. *If the system behavior is not known for the parameter range of interest, then the SI cannot be computed.*

The Weighting Function: By assigning weights to parameter settings and system configurations, scalability can be quantified as an overall measurement of the *weighted target measure across the entire parameter range*, allowing for the inclusion of factors such as parameter importance, likelihood, or associated costs. The weighting function also allows capturing the impact of problematic parameter settings or configurations appropriately. However, creating an effective weighting function often demands extensive system expertise, as elements like cost estimation can be challenging in practical scenarios. When such information is unavailable, treating all parameter settings equally with the same weight is a reasonable approach.

As a best practice, we suggest using the probabilities of parameter settings to occur in a system — if available. However, if probabilities of parameter settings are unknown, the importance of the settings may be reflected by the weight

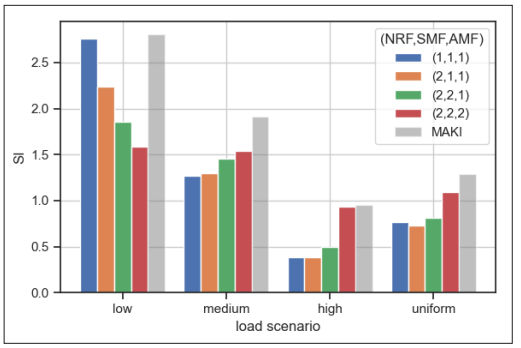

FIGURE 6. Scalability index (SI) for the 5G network slicing use case and different number of instances per NF. Different load scenarios are considered and the probability per load scenario follows a normal distribution with mean and standard deviation given in brackets (mean, std): low (20, 10), medium (40, 15), high (80, 20) load. In addition, a uniform load distribution over the entire parameter range is considered.

function. But this requires expert knowledge of the system and the importance of parameter settings. If probabilities or importance of parameters are unknown, it is recommended to weight all settings equally. Note that the SI is a general framework which allows including arbitrary scenarios (probabilities known, importance of settings due to expert knowledge, no further information).

The Reference System: A *reference system* acts as a benchmark for evaluating the system under test, providing a standard for comparing the (weighted) target measure. Ideally, the optimal system is chosen as the reference, but often, it's either unknown or too complex to determine. In practice, a key question is whether the system scales linearly with a linear function as the reference. Determining the slope and offset of this function can be aided by domain expertise, with the offset often derived from the system's idle state and the slope representing the desired or acceptable system behavior.

However, the primary target of our SI framework is the comparison of a system $\mathcal{F}$ with another similar system $\mathcal{G}$. Then, the scalability index can be computed by using $\mathcal{G}$ as a reference and show the scalability improvement or deterioration of $\mathcal{F}$ in relation to $\mathcal{G}$.

In Summary: The SI assesses a system's scalability potential, considering its ability to scale up or down over time, including elements like elasticity within the scalability analysis. Calculating the SI requires a comprehensive understanding of how the system operates and its functions. When dealing with a black box system, gaining insights into the system's behavior to derive the system function becomes necessary, which can be achieved through experiments, simulations, or stress tests. Typically, the goal is to compare a system with another similar one, but in the absence of a comparative system, a linear or an ideal system in theory can serve as a reference for comparison.

In practice, it may be important to assess multiple target functions of interest like delays, energy consumption or availability using the SI to understand how a system scales across different dimensions.

Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under grant number 316878574 (DFG SDN-App 2).

References

[1] T. Hossfeld, P. E. Heegaard, and W. Kellerer, "Comparing the Scalability of Communication Networks and Systems," *IEEE Access*, 2023; DOI: 10.1109/ACCESS. 2023.3314201.

[2] C. H. T. Arteaga, A. Ordoñez, and O. M. C. Rendon, "Scalability and Performance Analysis in 5G Core Network Slicing," *IEEE Access*, vol. 8, 2020, pp. 142,086–100.
[3] P. Jogalekar and M. Woodside, "Evaluating the Scalability of Distributed Systems," *IEEE Trans. Parallel and Distributed Systems*, vol. 11, no. 6, 2000, pp. 589–603.
[4] A. Al-Said Ahmad and P. Andras, "Scalability Analysis Comparisons of Cloud-Based Software Services," *J. Cloud Computing*, vol. 8, no. 1, 2019, pp. 1–17.
[5] P. Tran-Gia and T. Hossfeld, Performance Modeling and Analysis of Communication Networks, A Lecture Note, Würzburg University Press, 2021; ISBN: 978-3-95826-153-2; DOI: 10.25972/WUP-978-3-95826-153-2.
[6] N. R. Herbst, S. Kounev, and R. H. Reussner, "Elasticity in Cloud Computing: What It Is, and What It Is Not," *ICAC*, vol. 13, 2013, pp. 23–27.
[7] D. M. Gutierrez-Estevez *et al.*, "The Path Towards Resource Elasticity for 5G Network Architecture," Proc. 2018 *IEEE Wireless Commun. and Networking Conf. Workshops*, IEEE, 2018, pp. 214–19.
[8] P. Babarczi *et al.*, "A Mathematical Framework for Measuring Network Flexibility," *Computer Commun.*, vol. 164, 2020, pp. 13–24.
[9] S. Lehrig, H. Eikerling, and S. Becker, "Scalability, Elasticity, and Efficiency in Cloud Computing: A Systematic Literature Review of Definitions and Metrics," *Proc. 11th Int'l. ACM SIGSOFT Conf. Quality of Software Architectures*, 2015, pp. 83–92.
[10] R. Steinmetz *et al.*, "Adaptive and Scalable Communication Networks," *Proc. IEEE*, vol. 107, no. 4, 2019, pp. 635–38.
[11] M. Karakus and A. Durresi, "A Scalability Metric for Control Planes in Software Defined Networks (SDNS)," *Proc. 2016 IEEE 30th Int'l. Conf. Advanced Information Networking and Applications*, IEEE, 2016, pp. 282–89.
[12] Y. Li *et al.*, "Scalability and Performance Analysis of BDPS in Clouds," Computing, 2022, pp. 1–36.
[13] B. Alt *et al.*, "Transitions: A Protocol-Independent View of the Future Internet," *Proc. IEEE*, vol. 107, no. 4, 2019, pp. 835–46.

Biographies

Tobias Hossfeld is Full Professor and head of the Chair of Communication Networks at the University of Würzburg, Germany, since 2018. From 2014 to 2018, he was head of the Chair "Modeling of Adaptive Systems" at the University of Duisburg-Essen, Germany. He is a member of the editorial board of IEEE Commun. Surveys & Tutorials, ACM SIGMM Records, Springer Quality and User Experience, and a senior member of the IEEE.

Poul E. Heegaard is Full Professor at the Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), where he also has acted both as head of department and head of the research group in Networking. He was previously Senior Scientist with SINTEF Digital (1989-1999) and then Telenor R&I (1999-2009). He is a senior member of the IEEE.

Wolfgang Kellerer is a Full Professor with the Technical University of Munich (TUM), heading the Chair of Communication Networks at the Department of Electrical and Computer Engineering. Before, he was for over ten years with NTT DOCOMO's European Research Laboratories. He currently serves as an associate editor for IEEE Trans. Network and Service Management and as area editor for Network Virtualization for IEEE Communications Surveys and Tutorials.