An Architecture Concept for Short- and Long-term Resource Planning in the Industry 4.0 Environment

Arne Neumann*, Marvin Illian†, Tobias Hardes†, Lukas Martenvormfelde*,
Lukasz Wisniewski*and Jürgen Jasperneite*‡
*Institute Industrial IT - inIT, TH OWL, Lemgo, Germany
†Software Innovation Lab, Paderborn University, Germany
‡Fraunhofer IOSB-INA, Lemgo, Germany
{arne.neumann,lukas.martenvormfelde,lukasz.wisniewski}@th-owl.de
{marvin.illian,tobias.hardes}@upb.de
juergen.jasperneite@iosb-ina.fraunhofer.de

***Abstract*—The manufacturing domain is exposed to a continuous change of the requirements towards the IT infrastructure and the flexibility in Industry 4.0. In order to achieve a highly reliable production system, predictive maintenance and additive sensing have been implemented and will be complemented by further applications such as Augmented Reality. As the applications may be required ad-hoc at any time, the dynamic resource utilization of networking and computational resources needs to be managed. In the long-term, the planning of the infrastructure affects the available resources and thus the efficiency and reliability of the short-term resource management. This paper suggests an architecture that combines short- and long-term aspects of the resource utilization and previews how the infrastructure and opportunity costs can be optimized by the joint approach.**

***Index Terms*—industry4.0, architecture, 5G, planning**

I. Introduction

Virtualization is a key component of Industry 4.0. It enables offloading of computation and control tasks from dedicated to generic hardware, thereby transforming the physical factory to a virtual one. This is beneficial because of different reasons, such as lower hardware costs or increasing the reliability of applications [1]. Together with the growing complexity of control and supervision of manufacturing processes, this poses new challenges to the information and communication technology in factories.

Over the lifetime of a factory, various applications require sufficient and temporally flexible resources for network, computation, and data storage. With the decreasing lot sizes and much shorter life cycle of products, these requirements become even more volatile. Moreover, with the introduction of diverse Automated Guided Vehicles (AGVs) or different Augmented Reality (AR) or Virtual Reality (VR) applications, there is an increasing need for ad-hoc, mobile, and Quality of Service (QoS) communication that have to be managed in real-time. An often applied concept (to meet all application requirements) is the overprovisioning of various resources.

Wireless technologies which are improving in reliability and allow mobility, start to complement legacy networks. In particular, 5G comes into play to combine radio networks and edge computing resources. Hence the management of these resources is characterized by heterogeneity and variability.

In this paper, we propose an approach to address two management problems on different time scales. First, we address the problem of short-term planning. Short-term planning is essential to decide how resources shall be used for current or upcoming applications. Moreover, it allows fast detection and handling of situations with excessive load. Second, we address the problem of long-term planning. Optimization of the infrastructure can be achieved with long-term planning. This optimization can take place in terms of costs incurred, capacity utilization, or even manual interventions.

This research has been funded by the Ministry for Economic Affairs, Innovation, Digitization and Energy of the State of North Rhine-Westphalia (grant number 005-2008-0061).

II. Related Work

In the Industry 4.0 environment, a complex application mix and a resulting multi-layered resource management problem often arise [2]. An application mix can vary in terms of the necessary resources, such as required data rate or non-volatile memory capacity. It can also vary in terms of priorities or costs. Accordingly, emerging applications can only be planned at very short notice. This could be, for example, an AR based assistance system to diagnose a machine malfunction. Certain infrastructure adjustments, such as local non-volatile memory expansions, are only scalable in the long term. These different planning periods have a corresponding effect on the architecture of the overall system.

A. Architectures

In network and resource management architectures the trend is moving away from statically coupled management entities towards a set of flexible management functions or services. Related architecture models, such as the Industrial Internet Reference Architecture (IIRA) or the Reference Architectural Model Industry 4.0 (RAMI4.0), define different viewpoints or dimensions of networked systems at a generic level. Thereby they suggest basic requirements and principles for management systems, for example, trustworthiness and concepts for life cycle management. These requirements have been already picked up by researchers, for example in the project 5G

18th IEEE International Conference on Factory Communication Systems (WFCS 2022)

NORMA. 5G NORMA is targeting a mobile network infrastructure that dynamically adapts to variations of the traffic demands over time and location and to a changing network topology [3]. The project proposes a four-layer architecture where the control layer and the management and orchestration layer as layers 2 and 3 are related to our approach. These layers are characterized by:

(i) the management of physical as well as virtualized network functions and the virtualization infrastructure
(ii) network slicing and consequently a split into inter-slice and intra-slice functions.

While this concept is related to fifth generation of mobile telecommunications technology (5G) networks only, our proposal integrates additional resource types such as CPU or (non)-volatile memory resources.

In the project TACNET4.0, this approach was further refined in order to manage heterogeneous network technologies forming different network domains [4]. Here, a multi-domain manager and orchestrator was introduced, providing a network slice-oriented perspective to the industrial applications. Following this architecture, TACNET4.0 realized different QoS requirements and a resource-oriented perspective to manage the heterogeneous network technologies. Our proposal will complement this work by additionally focusing the long-term resource planning.

B. Management Tool

Proactive resource management and application scheduling has been discussed in various publications in the context of cloud computing. Singh and Chana [5] provide an extensive overview of resource scheduling mechanisms that can be optimized for execution times, operational costs, energy consumption, or any mixture of those. The combination of 5G New Radio (NR) and the associated Mobile Edge Cloud (MEC) resources brings another dimension in the management of the available resources. 5G and networking aspects, in general, are not covered in the work.

In the networking domain and particularly in 5G networks, Self-Organizing Network (SON) mechanisms aim for self-configuration, self-optimization, and self-healing of the network. SON was introduced by the 3GPP with the release of LTE and is continuously developed with the upcoming releases for 5G and beyond. In [6], the evolution from heuristical approaches in LTE to Machine Learning driven SON has been studied. However, the majority of the discussed features are not implemented in the current 5G releases and are expected to be rolled out with later 5G releases or the 6G specifications and thus they are far from implementation in Industry 4.0 environments.

The 3GPP specified a couple of mechanisms to manage the QoS of the 5G communication in the TS 23.501 [7]. Among others, the standard includes a QoS flow management as well as network slicing in the Radio Access Network (RAN) and the 5G core which can be used as easy access to the 5G management.

C. Planning Tool

Long-term planning can include the expansion of the local edge cloud, but also the acquisition of cloud resources such as third-party Virtual Machines (VMs).

Hadary *et al.* [8] propose a service named *Protean* to allocate Microsoft Azure VMs to a large number of servers. The idea of *Protean* is to use a flexible rule-based resource allocation theme that ensures turnaround times (time from job submission to completion) of only a few milliseconds. Various constraints and criteria (balanced ratio of used to unused cores) are taken into account during implementation. The authors evaluate the approach utilizing simulations and production tests. They show that user-specific requirements are met in $85 - 90\,\%$ of all cases. However, this approach is specific to Microsoft's Azure. In contrast, our proposal is a generic approach that can accommodate different cloud providers.

Calheiros *et al.* [9] propose a prediction model based on Auto-Regressive Integrated Moving Average (ARIMA). The authors aim to predict future workload using real traces of requests to web servers. The prediction is used to dynamically allocate VMs in an elastic cloud environment. Using simulations they show a prediction accuracy of up to $91\,\%$. Generic QoS parameters, response time, and rejection rate were used to validate the approach. However, this approach only relies on a single data source, whereas our approach can handle an arbitrary number of data sources.

Wang and Zuo [10] examine the problem of workflow scheduling including server configuration selection. The aim is to maximize the utilization of individual VMs and minimize application execution costs. They show that the presented approach is more efficient than the currently available standard solution like the PSO-based scheduling algorithm or the GA-based scheduling algorithm. However, the work is limited to fixed workflows. The approach we present here is completely flexible and can also be extended as desired.

In summary, there are already different solutions for medium and long-term planning. However, the approaches are often limited to only a small portion of the possibilities that can arise in an Industry 4.0 environment. Our architecture proposal, on the other hand, is completely flexible in this respect and can cover any scenario.

III. Architecture

A. Overview

The overall goal of the management approach is to minimize costs. Costs occur when resources are purchased, rented and operated and when applications are not executed, stopped, or throttled. We identified three types of costs to be considered:

- **Renting Costs** from renting cloud resources at an external provider. These can be obtained from public cloud pricing Application Programming Interfaces (APIs).
- **Licensing Costs** for frequency spectrum usage. Costs for local 5G spectrum e.g. in Germany, can be calculated based on a defined formula. However, this is a country-specific regulation and can vary in different countries.

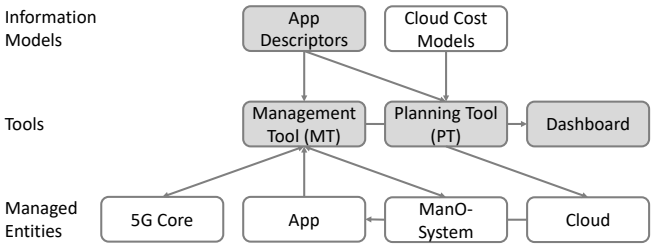

Fig. 1. Architecture for planning and management

- **Opportunity Costs** emerging from interrupting, stopping or throttling an application that would otherwise bring a monetary benefit.

Optimization requires management and control activities on different time scales, leading to the following two different types of actions. First, processes such as gradual convergence to capacity limits of a resource require pro-active, long-term planning. This planning is intended to guide the allocation of resources. Second, sudden events such as peak loads due to burst traffic or launch requests of new applications require reactive, short-term handling. This managing shall realize the most cost effective utilization of the allocated resources, depending on the dynamics in the number and characteristics of applications to be served simultaneously.

Our concept implements planning and management in separate tools, namely the Planning Tool (PT) and Management Tool (MT), as shown in Figure 1 in which the elements of our approach are highlighted in grey. The details of the main architecture elements are given in the following subsections. The tools exchange the following information. Both PT and MT receive information on resource consumption, duration and costs on a per-instance basis of the applications via so called App Descriptors (ADs). The MT also receives the total amount of each type of resources and monitors resource-related statistics about their utilization in a first step. In the second step, the utilization can be monitored individually for a single application. These statistics are forwarded to the PT along with notifications about running applications and their operation modes. Also, the PT requests cost information of cloud resources via the cloud providers' APIs. All output information of the PT is displayed on a dashboard as investment decisions need a personal approval. The approach targets on managing various types of resources that are necessary for realizing services in the Industry 4.0. These types are wired and wireless network for data transport, computation resources for data processing, and volatile and nonvolatile memory for data storage. Available management tools for specific resource types such as existing Management and Orchestration (ManO)-systems need to be integrated for reaching the overall optimization goal in diverse infrastructure deployments.

B. App Descriptors

The proposed architecture assumes that the application requirements are known a-priori and stored in ADs which need to be provided by the developer or owner of the applications. Furthermore, any application may have several operation modes that increase the flexibility, and each application has an execution deadline.

Our approach adapts and extends standardised descriptor approaches [11][12] by incorporating multiple application states with different resource requirements and adding cost models. All resource types and requirements are specified for each operation mode within the AD, such as the number of CPUs and the (non-)volatile memory sizes. Networking properties define the network type, uplink and downlink rates, and the allowed latencies. ADs contain minimum required resources which are assumed to be static over the runtime and satisfy an application's QoS. The QoS specification is highly dependent on the application type. Hence, there is no generic metric for it. To determine the priority of the applications, the opportunity costs of each operation mode of tasks are specified in the AD. A more cost intense application outage has a higher priority which reflects economical decision. These costs may be specified as constant or as accumulating over time depending on the task itself. Also, the ADs contain information whether the application can be preempted or not.

C. Management Tool

The MT makes decisions on the short-term utilization of the available resources and the admission control of the requested applications. It manages all available resource types, i.e., the processing, (non)-volatile memory and the wireless and wired network resources. The MT aims to minimize the costs arising from uncompleted or delayed tasks. In the sense of the proposed architecture, a decision is called a short-term decision, whenever the infrastructure could not be extended within the allowed time for the execution of the task.

Two scenarios for the MT, the overload and underload scenario, can be distinguished. In the overload scenario, the available resources are insufficient for the scheduled applications which therefore can not be executed in time. Thus, the resources assigned to the application are withdrawn and the task is stopped in order to operate the more important tasks. In the underload scenario, the MT primarily acts as a load-balancer which aims to distribute the resource utilization as even as possible to reduce idle periods as well as peaks in the resource utilization. Furthermore, the MT schedules the task resources ahead of the deadline to avoid deadline violations caused by the execution of high-priority ad-hoc tasks.

The MT exploits the ManO systems of the available resources. In more detail, the 5G core of a 5G non-public network offers an API to get status information such as the number of connected devices and their respective QoS. The API further allows to set configuration parameters for the 5G system. To avoid high complexity, the configuration of the 5G system is based on the QoS flows which are independent of the manufacturer of the system so that the MT can be extended for other versions. Furthermore, the admission control of the tasks is executed through the ManO-system managing the edge and cloud resources.

D. Planning Tool

The PT acts proactively. It aims to provide planning recommendations for infrastructure adaptations. There is no active intervention in the process – planning recommendations get reported on a dashboard as guidance. The final decision is always made by a human being.

Infrastructure adaptations consist of adaptations of different resource types which have different utilization characteristics and provisioning times. They can be distinguished into two main categories, self-deployed resources and rented resources.

- **Self-Deployed Resources** that require a hardware installation, like 5G base stations or server hardware for self-deployed compute infrastructure. Provisioning times are typically on the order of weeks or months. Self-Deployed Resources typically induce Capital Expenditure (CAPEX) and Operating Expenditure (OPEX).
- **Rented Resources** that can typically be rented from a cloud provider, like AWS, Azure or Google Cloud. They are already deployed in a remote location and can be rented for use. They can be dynamically adjusted comparably quickly (typically on the order of minutes).

Planning recommendations have to minimize costs. The PT extrapolates the infrastructure load into the future to ensure that the failure of the MT is prevented or moved into the future as far as possible. To do so, we have identified the following required input for the PT: information of required resources to run an application and cost models. A task schedule provides information about the application mix which is run at any point in time. Together they allow to compute an estimation of the required resources. However, this is insufficient as virtualization overhead and overheads from the orchestration framework induce additional load. Therefore, the current load data has to be another input for the PT. This allows to compute the total overhead at any point in time which can be taken into consideration in the extrapolation.

IV. Conclusion & Future Work

This paper proposes an architecture for the joint short-term management of the resource utilization and the long-term planning of infrastructural changes. Knowledge of the resource utilization and applications can be used to detect trends and estimate the future requirements on the infrastructural growth. This bears the potential to overcome the static overprovisioning of resources commonly used nowadays. Moreover, the joint management of different resource types enables the lowering of investment costs to implement Industry 4.0.

An implementation of the proposed architecture is planned as a part of the future work. It will consist of a private 5G network as well as computational resources in edge and core cloud deployments to test and validate the efficiency of the proposed planning and management tools. Further networking technologies need to be integrated in the concept as well as the prediction of required non-volatile memory which is largely unexplored. However, networking technologies differ greatly from each other, for example, not all of them allow configuration of the communication quality based on packet flows. This variety is a major challenge for the general applicability of the approach. Finally, a precise estimation of all costs on the different time scales and the resource requirements of future applications remain key elements for a successful integration of the proposed tools in industrial processes.

References

[1] M. Karrenbauer *et al.*, "Future industrial networking: from use cases to wireless technologies to a flexible system architecture," *at - Automatisierungstechnik, Oldenbourg Wissenschaftsverlag*, vol. 67, no. 7, pp. 526–544, 2019. DOI: 10.1515/auto-2018-0141.

[2] A. Issa, B. Hatiboglu, A. Bildstein, and T. Bauernhansl, "Industrie 4.0 roadmap: Framework for digital transformation based on the concepts of capability maturity and alignment," *Procedia CIRP, Elsevier*, vol. 72, pp. 973–978, 2018. DOI: 10.1016/j.procir.2018.03.151.

[3] C. Mannweiler *et al.*, "5G NORMA network architecture," Tech. Rep. Deliverable 3.3, Oct. 2017.

[4] M. Gundall *et al.*, "Introduction of a 5G-Enabled Architecture for the Realization of Industry 4.0 Use Cases," *IEEE Access*, vol. 9, pp. 25 508–25 521, 2021. DOI: 10.1109/ACCESS.2021.3057675.

[5] S. Singh and I. Chana, "A survey on resource scheduling in cloud computing: Issues and challenges," *Journal of grid computing, Springer*, vol. 14, no. 2, pp. 217–264, 2016. DOI: 10.1007/s10723-015-9359-2.

[6] J. Moysen and L. Giupponi, "From 4G to 5G: Self-organized network management meets machine learning," *Computer Communications*, vol. 129, pp. 248–268, 2018. DOI: 10.1016/j.comcom.2018.07.015.

[7] 3GPP, "System architecture for the 5G System (5GS)," Tech. Rep. TS 23.501, version 16.3.0, 2020.

[8] O. Hadary *et al.*, "Protean: VM Allocation Service at Scale," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, USENIX Association, Nov. 2020, pp. 845–861.

[9] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 449–458, Oct. 2015. DOI: 10.1109/TCC.2014.2350475.

[10] Y. Wang and X. Zuo, "An Effective Cloud Workflow Scheduling Approach Combining PSO and Idle Time Slot-Aware Rules," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 1079–1094, May 2021. DOI: 10.1109/JAS.2021.1003982.

[11] ETSI, "Network Functions Virtualisation (NFV) Rel.2; Management and Orchestration; VNF Descriptor and Packaging Specification," European Telecommunications Standards Institute, Sophia-Antipolis, France, Tech. Rep. RGS/NFV-IFA011ed241, Feb. 2018.

[12] OASIS, "TOSCA Simple Profile in YAML Version 1.3," Organization for the Advancement of Structured Information Standards, Burlington, Massachusetts US, Tech. Rep. TOSCA-Simple-Profile-YAML-v1.3, Feb. 2020.