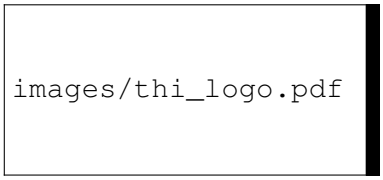


Technische Hochschule Ingolstadt

Specialist area Computer Science
Bachelor's course Computer Science

Bachelor's thesis

Subject:	Conception, Implementation, and Evaluation of a Highly Scalable and Highly Available Kubernetes-Based SaaS Platform on Kubernetes Control Plane (KCP)
Name and Surname:	David Linhardt
Matriculation number:	00122706
Issued on:	2025-04-09
Submitted on:	2025-09-09
First examiner:	Prof. Dr. Bernd Hafenrichter
Second examiner:	Prof. Dr. Ludwig Lausser

Abstract

Contents

Glossary

1 Introduction

1.1 Problem Statement and Motivation

1.2 Objectives and Scope

1.3 Structure of the Thesis

2 Fundamentals

2.1 Kubernetes and Multi-Tenancy

Kubernetes as the Foundation for Cloud-Native Applications As the de facto standard for deploying and managing *cloud-native applications*, Kubernetes plays a pivotal role in modern cloud architecture (Poulton and Joglekar 2021, p. 7–8). Kubernetes works as an application orchestrator for *containerized, cloud-native microservice* apps, meaning it can deploy applications and dynamically respond to changes (Poulton and Joglekar 2021, p. 3). It offers a platform for declarative configuration and automation for containerized workloads, enabling organizations to run distributed applications and services at scale (Kubernetes 2024; Red Hat 2024).

The Importance of Multi-Tenancy in Modern SaaS Platforms Multi-tenancy plays a fundamental role in modern cloud computing. By allowing multiple tenants to share the same infrastructure through virtualization, it significantly increases resource utilization, reduces operational costs, and enables essential features such as VM mobility and dynamic resource allocation (AlJahdali et al. 2014, pp. 345–346). These benefits are critical for cloud providers, as they make the cloud business model economically viable and scalable. In the context of modern SaaS platforms, multi-tenancy goes even further by enabling unified management, frictionless onboarding, and simplified operational processes that allow providers to add new tenants without introducing incremental complexity or cost (AWS 2022, pp. 9–11).

However, while multi-tenancy is indispensable for achieving efficiency, scalability, and cost-effectiveness, it simultaneously introduces complex security challenges, especially in shared environments where resource isolation is limited. In particular, the potential for cross-tenant access and side-channel attacks makes security in multi-tenant environments a primary concern (AlJahdali et al. 2014, pp. 345–346). As such, understanding and addressing multi-tenancy from both operational and security perspectives is essential when designing and securing modern cloud-native platforms (AWS 2022, pp. 9–11; *Information technology - Cloud computing - Part 2: Concepts* 2023, p. 4).

The Challenges of Multi-Tenancy and the Need for Solutions Multi-tenancy introduces a spectrum of technical and security challenges that need to be addressed.

[1]: *Residual-data exposure*. Shared infrastructures may expose tenants to data leakage and hardware-layer attacks. Because hardware resources are only virtually partitioned, residual data left in reusable memory or storage blocks, known as *data remanence*, can be inadvertently leaked or deliberately harvested by co-resident tenants (Zissis and Lekkas 2012, p. 586; AlJahdali et al. 2014, pp. 344–345).

[2]: *Control and transparency*. By design, SaaS moves both data storage and security controls out of the enterprise's boundary and into the provider's multi-tenant cloud, depriving organizations of direct oversight and assurance and thereby heightening concern over how their critical information is protected, replicated and kept available (Subashini and Kavitha 2011, pp. 3–4). To complicate matters further, the customer might have no way to evaluate the SaaS vendors security measures, meaning the pricing and feature set will most likely determine which service is used in practice, often disregarding security concerns (Everett 2009, p. 6; Khorshed, Ali, and Wasimi 2012, p. 836).

[3]: *Scheduling*. In multi-tenant architectures multiple tenants utilize the same hardware, thus creating the need for fair scheduling to ensure cost-effectiveness and performance (Simić et al. 2024, p. 32597).

[4]: *Performance Isolation* A single tenant is able to significantly degrade the performance of other tenants working on the same hardware, if *performance isolation* is not given (Krebs and Mehta 2013, p. 195). As noted by Krebs and Mehta 2013, p. 195 "A system is said to be performance isolated, if for tenants working within their quotas the performance is within the (response time) SLA while other tenants exceed their quotas (e.g., request rate)"

[5]: *Automation*.

A secure solution, keeping multi-tenancies while also addressing security concerns is desperately needed (AlJahdali et al. 2014, p. 346).

Kubernetes Control Plane (KCP) as a Promising Approach

Background: The Evolution of Kubernetes Kubernetes, an open-source container orchestration platform developed by Google, emerged from the need to manage the complexities of containerized applications effectively and to support large-scale deployments in a cloud-native environment (Google Cloud 2025; Kubernetes 2024). It was originally developed at Google and released as open source in 2014 (Google Cloud 2025). Kubernetes was conceived as a successor to Google's internal container management system called Borg, and designed to streamline the process of deploying, scaling, and managing applications composed of microservices running in containers (Verma et al. 2015, pp. 13–14; Bernstein 2014, p. 84). Since its inception, Kubernetes has gained traction among organizations because it provides robust features such as automated scaling, self-healing, and service discovery, which have made it the de facto standard for container orchestration in the tech industry (Damarapati 2025, pp. 855–858). As noted by Moravcik et al. 2022, p. 457 by 2021 almost 90% of organizations used Kubernetes as an orchestrator for managing containers and over 70% of organizations used it in production (Shamim Choudhury 2025). The widespread adoption of Kubernetes is further underscored by Red Hat's latest (2024) report, which no longer asks survey respondents if they use Kubernetes for container orchestration, but rather **which** Kubernetes platform they use (Red Hat, Inc. 2024, p. 27). According to Damarapati 2025, pp. 855–856, Kubernetes has seen unprecedented industry adoption due to its vendor neutrality, strong community support, and flexible, extensible architecture in combination with enterprise readiness caused by high availability, disaster recovery and security.

Moreover Kubernetes enables faster time-to-market by providing a unified, declarative control plane that abstracts away infrastructure, guarantees consistent environments from development to production, and automates operational tasks such as scaling, rolling updates, and self-healing—advantages that translate directly into competitive delivery speed increasing its appeal to organizations of every size (Damarapati 2025, pp. 858–859).

Over the years, Kubernetes—and the many orchestration solutions inspired by or built on it—has evolved to handle an increasingly diverse range of workloads, supporting everything from conventional applications to emerging *edge-native* deployments (Biot et al. 2025, p. 21; Biot et al. 2025, pp. 1–4). Edge-native deployments are applications intended to run on computing

resources located at or near the data source — the network *edge* — rather than in a central cloud (Satyanarayanan et al. 2019, p. 34). This adaptability reflects its fundamental design, which focuses on modularity and extensibility, allowing developers to customize their orchestration needs.

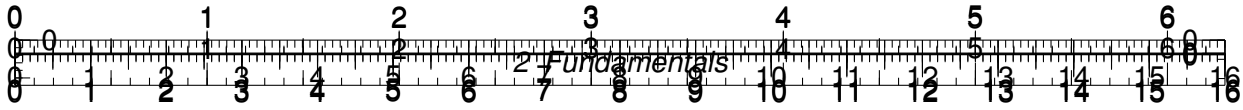
Overall, the history of Kubernetes showcases a transformative journey driven by the evolving demands of software architecture and the necessity for efficient application management in an increasingly complex technological landscape.

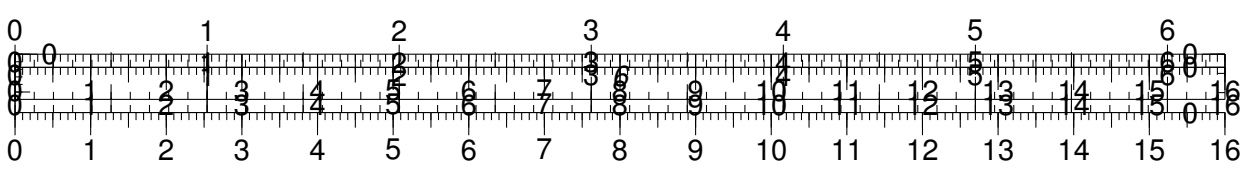
Background: Containerization as an Enabler of Kubernetes *Containerization* is a way to bundle an application's code with all its dependencies to run on any infrastructure thus enhancing portability (AWS 2025; Docker 2025). The lightweight nature and isolation can be leveraged by cloud-native software by enabling vertical and horizontal autoscaling facilitated by quick container boot times, along with self-healing mechanisms and support for distributed, resilient infrastructures (Kubernetes 2025b; Kubernetes 2025c; AWS 2025; Davis 2019, pp. 58–59) Furthermore it complements the microservice architectural pattern by enabling isolated, low overhead deployments, ensuring consistent environments (Balalaie, Heydarnoori, and Jamshidi 2016, p. 209).

Background: The Role of Microservices in Cloud-Native Architectures *Microservices* play a pivotal role in cloud-native architectures by promoting agility, scalability, and maintainability of applications. By decomposing applications into independent, granular services, microservices facilitate development, testing, and deployment using diverse technology stacks, enhancing interoperability across platforms (Waseem, Liang, and Shahin 2020, p. 1; Larrucea et al. 2018, p. 1) and help prevent failures in one component from propagating across the system, by isolating functionality into distinct, self-contained services (Davis 2019, p. 62). This architectural style aligns well with cloud environments, as it allows services to evolve independently, effectively addressing challenges associated with scaling and maintenance without being tied to a singular technological framework (Balalaie, Heydarnoori, and Jamshidi 2016, pp. 202–203). Furthermore, the integration of microservices with platforms like Kubernetes enhances deployment automation and orchestration, thus providing substantial elasticity to accommodate fluctuating workloads (Haugeland et al. 2021, p. 170). Additionally, migrating legacy applications to microservices can foster modernization and efficiency, thus positioning organizations favorably in competitive landscapes (Balalaie, Heydarnoori, and Jamshidi 2016, p. 214). Overall, the synergy between microservices and cloud-native architectures stems from their inherent capability to optimize resource utilization and streamline continuous integration and deployment processes.

Background: Kubernetes Resource Isolation Mechanisms Kubernetes employs several resource isolation mechanisms, primarily through the use of *cgroups* (control groups) and *namespaces* to limit resource allocation for containers. Cgroups are a Linux kernel feature that organizes processes into hierarchical groups for fine-grained resource limitation and monitoring via a pseudo-filesystem called *cgroupfs* (Kubernetes 2025a; Project 2024). *Namespaces* are a mechanism for isolating groups of resources withing a single cluster and scoping resource names to prevent naming conflicts across different teams or projects (Kubernetes 2025d). However, these mechanisms may not always provide sufficient isolation necessary for multi-tenant architectures, because the logical segregation offered by namespaces does not address the fundamental security concerns associated with multi-tenancy (Nguyen and Y. Kim 2022, p. 651) and research indicates that the native isolation strategies can lead to performance interference, where containers that share nodes can experience significant degradation in performance due to CPU contention (E. Kim, Lee, and Yoo 2021, p. 158). Specifically, critical services may be adversely affected when non-critical services monopolize available resources, which undermines the quality of service in multi-tenant environments (Li et al. 2019, p. 30410). Moreover, while Kubernetes allows for container orchestration and resource scheduling, it can lead to resource fragmentation, further exacerbating the issue of performance isolation (Jian et al. 2023, p. 1). A common approach in multi-tenant scenarios is to deploy separate clusters for each tenant, which incurs substantial overhead—particularly in environments utilizing virtual machines for isolation (Şenel et al. 2023, pp. 144574–144575). In summary, although Kubernetes offers essential isolation mechanisms, the complexities of resource sharing and performance consistency in multi-tenant applications highlight the need for enhanced strategies to ensure robust resource management and performance isolation (Nguyen and Y. Kim 2022, p. 651; Jian et al. 2023, p. 2; E. Kim, Lee, and Yoo 2021, p. 158).

Relevance to SaaS and this Thesis





References

AlJahdali, H., A. Albatli, P. Garraghan, P. Townend, L. Lau, and J. Xu (2014). “Multi-tenancy in Cloud Computing”. In: *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, pp. 344–351. DOI: 10.1109/SOSE.2014.50.

AWS (2022). *AWS Whitepaper - SaaS Architecture Fundamentals*. AWS. URL: https://docs.aws.amazon.com/whitepapers/latest/saas-architecture-fundamentals/re-defining-multi-tenancy.html (visited on 05/01/2025).

AWS (2025). *What is Containerization?* URL: https://aws.amazon.com/what-is/containerization/ (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519121803/https://aws.amazon.com/what-is/containerization/.

Balalaie, A., A. Heydarnoori, and P. Jamshidi (2016). "Migrating to cloud-native architectures using microservices: an experience report". In: pp. 201–215. DOI: 10.1007/978-3-319-33313-7_15.

Bernstein, D. (2014). "Containers and Cloud: From LXC to Docker to Kubernetes". In: *IEEE Cloud Computing* 1.3, pp. 81–84. DOI: 10.1109/MCC.2014.51.

Biot, F., A. Fornés-Leal, R. Vaño, R. Simon, I. Lacalle, C. Guardiola, and C. Palau (2025). "A novel orchestrator architecture for deploying virtualized services in next-generation iot computing ecosystems". In: *Sensors* 25 (3), p. 718. DOI: 10.3390/s25030718.

Damarapati, A. (Jan. 2025). "Containers vs. Virtual machines: Understanding the shift to Kubernetes". In: *World Journal of Advanced Engineering Technology and Sciences* 15.1, pp. 852–861. ISSN: 2582-8266. DOI: 10.30574/wjaets.2025.15.1.0305. URL: http://dx.doi.org/10.30574/wjaets.2025.15.1.0305.

Davis, C. (2019). *Cloud Native Patterns - Designing change-tolerant software*. Shelter Island, NY: Manning. ISBN: 9781617294297.

Docker (2025). *Use containers to Build, Share and Run your applications*. URL: https://www.docker.com/resources/what-container/ (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519121103/https://www.docker.com/resources/what-container/.

Everett, C. (June 2009). "Cloud computing - A question of trust". In: *Computer Fraud & Security* 2009.6, pp. 5–7. ISSN: 1361-3723. DOI: 10.1016/s1361-3723(09)70071-5. URL: http://dx.doi.org/10.1016/s1361-3723(09)70071-5.

Google Cloud (2025). *What is Kubernetes?* URL: https://cloud.google.com/learn/what-is-kubernetes (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519121940/https://cloud.google.com/learn/what-is-kubernetes.

Haugeland, S., P. Nguyen, H. Song, and F. Chauvel (2021). "Migrating monoliths to microservices-based customizable multi-tenant cloud-native apps". In: pp. 170–177. DOI: 10.1109/seaa53835.2021.00030.

Information technology - Cloud computing - Part 2: Concepts (2023). Standard.

Jian, Z., X. Xie, Y. Fang, Y. Jiang, T. Li, and Y. Lu (2023). “Drs: a deep reinforcement learning enhanced kubernetes scheduler for microservice-based system”. In: DOI: 10.22541/au.167285897.72278925/v1.

Khorshed, M. T., A. S. Ali, and S. A. Wasimi (June 2012). “A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing”. In: *Future Generation Computer Systems* 28.6, pp. 833–851. ISSN: 0167-739X. DOI: 10.1016/j.future.2012.01.006. URL: http://dx.doi.org/10.1016/j.future.2012.01.006.

Kim, E., K. Lee, and C. Yoo (Jan. 2021). “On the Resource Management of Kubernetes”. In: *2021 International Conference on Information Networking (ICOIN)*. IEEE, pp. 154–158. DOI: 10.1109/icoin50884.2021.9333977. URL: http://dx.doi.org/10.1109/icoin50884.2021.9333977.

Krebs, R. and A. Mehta (Sept. 2013). “A Feedback Controlled Scheduler for Performance Isolation in Multi-Tenant Applications”. In: *2013 International Conference on Cloud and Green Computing*. IEEE, pp. 195–196. DOI: 10.1109/cgc.2013.36. URL: http://dx.doi.org/10.1109/cgc.2013.36.

Kubernetes (2024). *Concepts / Overview*. URL: https://kubernetes.io/docs/concepts/overview/ (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519122217/https://kubernetes.io/docs/concepts/overview/.

Kubernetes (2025a). *About cgroup v2*. URL: https://kubernetes.io/docs/concepts/architecture/cgroups/ (visited on 05/02/2025). Archived at https://web.archive.org/web/20250519120201/https://kubernetes.io/docs/concepts/architecture/cgroups/.

Kubernetes (2025b). *Concepts / Workloads / Autoscaling Workloads*. URL: https://kubernetes.io/docs/concepts/workloads/autoscaling/ (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519121534/https://kubernetes.io/docs/concepts/workloads/autoscaling/.

Kubernetes (2025c). *Kubernetes Self-Healing*. URL: https://kubernetes.io/docs/concepts/architecture/self-healing/ (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519121258/https://kubernetes.io/docs/concepts/architecture/self-healing/.

Kubernetes (2025d). *Namespaces*. URL: https://kubernetes.io/docs/concepts/overview/working-with-objects/namespaces/ (visited on 05/02/2025). Archived at https://web.archive.org/web/20250519115910/https://kubernetes.io/docs/concepts/overview/working-with-objects/namespaces/.

Larrucea, X., I. Santamaria, R. Colomo-Palacios, and C. Ebert (2018). “Microservices”. In: *IEEE Software* 35.3, pp. 96–100. DOI: 10.1109/MS.2018.2141030.

Li, Y., J. Zhang, C. Jiang, J. Wan, and Z. Ren (2019). “Pine: optimizing performance isolation in container environments”. In: *Ieee Access* 7, pp. 30410–30422. DOI: 10.1109/access.2019.2900451.

Moravcik, M., M. Kontsek, P. Segec, and D. Cymbalak (2022). “Kubernetes - evolution of virtualization”. In: *2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pp. 454–459. DOI: 10.1109/ICETA57911.2022.9974681.

Nguyen, N. T. and Y. Kim (Oct. 2022). “A Design of Resource Allocation Structure for Multi-Tenant Services in Kubernetes Cluster”. In: *2022 27th Asia Pacific Conference on Communications (APCC)*. IEEE, pp. 651–654. DOI: 10.1109/apcc55198.2022.9943782.

Poulton, N. and P. Joglekar (2021). *The Kubernetes Book*. 2021 Edition. No ISBN provided. Independently published, p. 243.

Project, T. L. D. (2024). *cgroups(7): Linux control groups*. 6.10. Online; accessed 2025-05-02. Linux man-pages project. URL: https://man7.org/linux/man-pages/man7/cgroups.7.html.

Red Hat (2024). *What is Kubernetes?* URL: https://www.redhat.com/en/topics/containers/what-is-kubernetes (visited on 05/01/2025). Archived at https://web.archive.org/web/20250519122615/https://www.redhat.com/en/topics/containers/what-is-kubernetes.

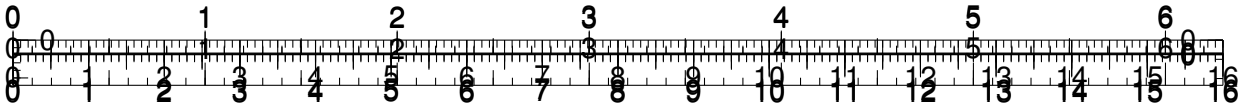
Red Hat, Inc. (2024). *The State of Kubernetes Security Report: 2024 Edition*. Red Hat. URL: https://www.redhat.com/en/resources/kubernetes-adoption-security-market-trends-overview (visited on 05/19/2025).

Satyanarayanan, M., G. Klas, M. Silva, and S. Mangiante (July 2019). “The Seminal Role of Edge-Native Applications”. In: *2019 IEEE International Conference on Edge Computing (EDGE)*. IEEE, pp. 33–40. DOI: 10.1109/edge.2019.00022. URL: http://dx.doi.org/10.1109/edge.2019.00022.

Şenel, B., M. Mouchet, J. Cappos, T. Friedman, O. Fourmaux, and R. McGeer (2023). “Multitenant containers as a service (caas) for clouds and edge clouds”. In: *Ieee Access* 11, pp. 144574–144601. DOI: 10.1109/access.2023.3344486.

Shamim Choudhury (2025). *Kubernetes adoption, security, and market trends report 2021 - by RedHat*. URL: https://www.javelynn.com/cloud/kubernetes-adoption-security-and-market-trends-report-2021 (visited on 05/19/2025). Archived at https://web.archive.org/web/20250519115027/https://www.javelynn.com/cloud/kubernetes-adoption-security-and-market-trends-report-2021.

Simić, M., J. Dedeić, M. Stojkov, and I. Prokić (2024). “A Hierarchical Namespace Approach for Multi-Tenancy in Distributed Clouds”. In: *IEEE Access* 12, pp. 32597–32617. ISSN: 2169-3536.



DOI: 10.1109/access.2024.3369031. URL: http://dx.doi.org/10.1109/access.2024.3369031.

Subashini, S. and V. Kavitha (Jan. 2011). “A survey on security issues in service delivery models of cloud computing”. In: *Journal of Network and Computer Applications* 34.1, pp. 1–11. ISSN: 1084-8045. DOI: 10.1016/j.jnca.2010.07.006. URL: http://dx.doi.org/10.1016/j.jnca.2010.07.006.

Verma, A., L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes (2015). “Large-scale cluster management at Google with Borg”. In: *Proceedings of the Tenth European Conference on Computer Systems*. EuroSys ’15. Bordeaux, France: Association for Computing Machinery. ISBN: 9781450332385. DOI: 10.1145/2741948.2741964. URL: https://doi.org/10.1145/2741948.2741964.

Waseem, M., P. Liang, and M. Shahin (2020). “A systematic mapping study on microservices architecture in devops”. In: *Journal of Systems and Software* 170, p. 110798. DOI: 10.1016/j.jss.2020.110798.

Zissis, D. and D. Lekkas (2012). “Addressing cloud computing security issues”. In: *Future Generation Computer Systems* 28.3, pp. 583–592. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2010.12.006. URL: https://www.sciencedirect.com/science/article/pii/S0167739X10002554.

List of Figures

Appendix