GEKO: GPUs energieeffizient für KI-Inferenz orchestrieren

Project description (VHB) for software campus 2025

Valentin Carl

September 15, 2025

Applicant (Projektrahmenvorhaben)	Technische Universität Berlin
Project lead	Valentin Carl Technische Universität Berlin FG Skalierbare Softwaresysteme (E-N 17) Einsteinufer 17 10587 Berlin E-Mail: nc@3s.tu-berlin.de
Academic supervision	Prof. Dr.-Ing. David Bermbach Technische Universität Berlin FG Skalierbare Softwaresysteme (E-N 17) Einsteinufer 17 10587 Berlin E-Mail: db@3s.tu-berlin.de
Industry partner	Dr. Sripriya Adhatarao Huawei Technologies Duesseldorf GmbH Advanced Wireless Technologies Lab Riesstraße 25 80992 München. E-Mail: sripriya.srikant.adhatarao@huawei.com
Project start **Project duration**	March 1, 2026 12 Months

software campus

HUAWEI

3S

TECHNISCHE UNIVERSITÄT BERLIN

Contents

1 Task Definition and Motivation

Function-as-a-Service (FaaS) is a serverless computing model in which developers write small, stateless functions that are invoked by a cloud platform in response to external requests. In this model, the platform manages nearly all aspects of execution, including resource allocation, auto-scaling, and the runtime environment. Especially in edge environments, where resources are scarce, FaaS has proven to be a suitable paradigm for sharing hardware between applications and allocating resources only when they are actually needed. At the same time, the rapid growth of cloud platforms and data centers has made their energy consumption a critical concern. Over the past decades, global data center electricity use has steadily increased without signs of slowing down. In 2023, data centers in the United States alone consumed $176\,\mathrm{TWh}$, accounting for $4.4\,\%$ of the country's total electricity use, with projections estimating a rise to $6.7\,\%$ to $12\,\%$ by 2028 [She+24]. Despite significant improvements in Power Usage Effectiveness (PUE), rising demand consistently outpaces efficiency gains, and data centers already account for roughly $1\,\%$ of global electricity consumption [Gan+23; Mas+20; Sha24]. This trajectory directly conflicts with the Paris Agreement's target of limiting global warming to well below $2\,^{\circ}\mathrm{C}$, which requires rapid and substantial reductions in greenhouse gas emissions. These developments create a responsibility for both developers and cloud platforms alike to consciously consider and continuously improve the environmental footprint of their infrastructures [Chi21].

The energy demand of artificial intelligence workloads is a particularly pressing issue. Modern inference tasks are heavily GPU-bound, and while GPUs provide the necessary computational performance, they are also associated with high energy costs. This creates a fundamental tension between society's growing reliance on AI-powered services and the urgent need to reduce the carbon footprint of digital infrastructures. To address this challenge, we focus on FaaS as the underlying paradigm. The serverless model provides fine-grained elasticity, centralized infrastructure management, resource sharing across applications, and a large degress of control to the cloud platform, which are particularly valuable both in large-scale cloud data centers and in resource-constrained edge environments. These properties make FaaS a natural foundation for implementing energy-aware orchestration strategies that can adaptively control GPU usage. At the same time, a lot of work remains to be done in order to improve the currently poor energy efficiency of contemporary FaaS platforms [Sha23]. The scientific question that motivates this project is therefore: How can GPU resources for serverless inference be orchestrated in an adaptive and energy-efficient manner, without compromising performance?

The social relevance of this question lies in the sustainability of digital infrastructures. As AI models become important components of everyday applications, from medical diagnostics to large language models, operators must reconcile latency and throughput demands with climate responsibility. Practically, today's serverless platforms provide limited support for GPU execution, in general, and fine-grained GPU management, in particular: In most deployments and when available, GPUs remain powered even during idle times, wasting significant amounts of energy. Addressing this inefficiency not only has the potential to greatly affect environmental impact of serverless infrastructure but also lowers operational costs in cloud-scale environments. This issue is even more pressing for Germany, as the country already experiences substatially increased levels of warming compared to global trends and currently even the most pessimistic RPC8.5 scenario [DWD25].

Practical application examples can be found in inference-as-a-service offerings. Applications such as real-time medical image analysis or conversational AI systems must respond elastically to fluctuating demand. A dynamic orchestration approach will ensure that GPUs are powered on only when required, while predictive scheduling mechanisms mitigate cold-start overheads. This will enable cloud providers to deliver sustainable, latency-sensitive services at scale without sacrificing user experience. The goal of the **GEKO** ("**G**PUs **e**nergieeffizient für **K**I-Inferenz **o**rchestrieren") project is therefore to develop serverless platform architectures and programming abstractions that will enable deploying, managing, and using serverless functions with GPU support for AI applications in an energy-efficient manner.

1.1 Focus and objectives

FaaS has emerged as a promising abstraction for building scalable and elastic applications. However, despite its advantages, current FaaS platforms remain highly energy-inefficient [Sha23]. This inefficiency stems from two key factors: the strong variance in request loads, which often leads to overprovisioning or idle resources, and the expensive software-level isolation required to execute short-lived functions securely [GF20; Sch+23]. As a result, serverless applications today are far from exploiting their potential for sustainable operation.

At the same time, FaaS has the ideal prerequisites to serve as the core programming model for energy-efficient AI inference [Pat+21]. Its fine-grained elasticity, centralized control of resources, and abstraction from application logic make it a natural fit for orchestrating energy-aware scheduling and adaptive GPU usage. Yet, realizing this potential requires foundational research, since existing platforms offer only limited support in this direction. Notably, most public FaaS services do not provide GPU support at all, leaving no basis for exploring efficient orchestration of inference workloads.

A key reason why FaaS is particularly well suited for sustainable computing lies in its platform-centric model. Instead of requiring every developer to solve sustainability challenges individually, the serverless abstraction concentrates responsibility for efficiency at the platform level. This enables resource sharing, workload consolidation, and energy-aware scheduling to be implemented once and leveraged by all applications running on the platform. In principle, this makes FaaS one of the strongest candidates for aligning large-scale digital infrastructures with sustainability goals, provided that the necessary system mechanisms exist.

The focus of this project is therefore to establish the groundwork for sustainable, scalable GPU-based inference in serverless environments. We aim to provide the missing system-level mechanisms that allow GPUs to be integrated into FaaS platforms and managed adaptively with respect to workload demands. In doing so, the GEKO project seeks to bridge the gap between today's energy-inefficient serverless platforms and a future in which FaaS is the foundation of sustainable AI infrastructure.

1.2 Scientific and/or technical objectives of the project

Building on the motivation outlined above, this project aims to develop the foundations for sustainable and scalable GPU-based inference in serverless platforms. The overarching goal is to transform Function-as-a-Service from an energy-inefficient abstraction into a viable basis for sustainable AI infrastructure. To achieve this, we focus on system-level mechanisms that enable adaptive GPU orchestration, efficient resource sharing, and transparent integration of energy-aware scheduling policies into the serverless execution model.

The concrete objectives of this project are threefold. First, we seek to design and implement the missing platform mechanisms that allow GPUs to be exposed as first-class resources in FaaS environments. Second, we aim to develop orchestration strategies that adapt GPU allocation dynamically to workload fluctuations, minimizing idle energy costs while preserving performance. Third, we plan to evaluate the effectiveness of these strategies across a diverse set of inference workloads and deployment settings, thereby quantifying their impact on both energy efficiency and quality of service.

From these objectives, the following technical and research questions emerge:

(1) How can GPUs be exposed and managed as first-class resources in FaaS environments, given the short-lived and highly dynamic nature of serverless functions?

(2) What orchestration strategies can dynamically adapt GPU allocation to workload fluctuations, ensuring high utilization while minimizing idle energy costs?

(3) How do the proposed mechanisms perform across diverse AI inference workloads, and what trade-offs emerge between energy efficiency, performance, and scalability?

To address them, the project will create an open-source prototype of a serverless platform that integrates GPU support and implements energy-aware orchestration mechanisms. This prototype will be designed to be usable and extensible by both the research community and industry practitioners, providing a practical foundation for future work on sustainable AI infrastructures.

1.3 Relation of the project to funding policy objectives/funding program

The project GEKO is closely aligned with the strategic goals of the BMFTR and the objectives of its "Hightech Agenda Deutschland". In particular, it addresses two central themes emphasized in the funding policy: advancing artificial intelligence as a key enabling technology and promoting sustainable digital infrastructures in line with Germany's climate commitments.

First, GEKO contributes to strengthening Germany's technological leadership in AI by addressing the high energy demand of inference workloads. The project develops innovative system-level mechanisms for energy-efficient orchestration of GPU resources in serverless environments. Consequently, it complements existing AI research that predominantly focuses on model efficiency and instead tackles the infrastructure and platform perspective, which are equally important for the practical use of AI technology. This is directly in line with the Hightech Agenda's objective to expand AI research across the full technology stack and secure digital sovereignty in Europe. Second, GEKO has a strong relation to the BMFTR's climate and sustainability goals. The project's central objective, i.e., reducing the carbon footprint of AI workloads in cloud platforms, directly supports Germany's contribution to achieving the climate targets of the Paris Agreement. By focusing on platform-level orchestration and resource sharing, GEKO demonstrates how sustainability can be built into digital infrastructures rather than being left to individual developers or applications. This systemic approach has the potential to deliver significant energy savings at scale, thereby making a measurable contribution to sustainable digitalization.

Finally, the project strengthens education and innovation transfer. Embedded in the Software Campus program, GEKO provides graduate students with the opportunity to develop practical expertise in systems research, cloud platforms, and sustainable computing. In parallel, the project fosters leadership and soft skills that are crucial for future roles in academia and industry. The planned open-source prototype of a serverless GPU platform ensures that the results are not only of academic value but also accessible to the wider research community and industrial stakeholders, thereby accelerating innovation transfer and supporting Germany's role as a leading hub for sustainable AI. In this way, GEKO not only advances fundamental research but also facilitates direct knowledge transfer from academia to industry, ensuring that insights into sustainable AI infrastructures quickly translate into practical innovations in the German and European technology sector.

2 State of the Art in Science and Technology

Function-as-a-Service is a serverless programming model in which developers express applications as small, stateless functions that are invoked on demand by the cloud platform. This abstraction frees developers from concerns about scalability and infrastructure management, while enabling large-scale resource sharing across many tenants. Combined with a pay-per-use billing model, FaaS has quickly become a central paradigm in both industry and research for building elastic applications [Jon+19]. Importantly, this model also has untapped potential as a cornerstone for sustainable computing: If orchestrated carefully, shared resources can be provisioned more efficiently at platform level than if each developer had to optimize for sustainability individually. The GEKO project builds on a solid foundation of scientific groundwork laid by a small but active research community in the field of sustainable serverless computing. This includes early explorations into how serverless abstractions could become enablers for energy-aware resource management and carbon accounting. A cornerstone in this emerging area is the articulation of a broader vision for sustainable serverless computing, which frames FaaS as a potential driver of greener cloud services and highlights key research gaps that remain open [Sha23].

In today's practice, however, serverless computing is far from energy efficient. Current FaaS deployments follow two dominant paths. On the one hand, organizations operate open-source serverless platforms on top of Kubernetes or similar orchestration systems. While this approach offers flexibility, it suffers from heavy reliance on virtualization and container isolation, which introduces substantial overheads; these, in turn, increase per-request energy consumption between 15× and 30×, depending on the virtualization technique used [Sha23]. Recent studies show that software-based isolation layers such as gVisor can significantly degrade efficiency, limiting the potential for sustainable operation [You+19]. On the other hand, large public cloud providers offer commercial FaaS services, which remain largely opaque to researchers and offer only limited resource control (EXAMPLES). In particular, GPU support is mostly absent from these platforms, making it difficult to realize AI workloads in a serverless fashion. Despite these limitations, the research community has begun to address sustainability in serverless computing. Early work has explored quantifying the energy overheads of function isolation and assigning carbon intensity metrics to individual function invocations [SF24]. Other studies have examined spatio-temporal scheduling, where functions are steered to data centers with lower grid carbon intensity (EXAMPLES, viele). However, these approaches often conflict with the latency requirements of serverless workloads and cannot fully exploit hardware-level optimizations [Suk+24]. The majority of existing work still focuses on performance-oriented goals, such as reducing cold start latencies, rather than systematically reducing the energy footprint of the platform (EXAMPLES, viele). In the context of artificial intelligence, the gap is even more pronounced. Modern inference workloads are increasingly GPU-bound, but the lack of GPU integration in public serverless platforms is a contributor to preventing FaaS from being used for scalable AI inference. While research in AI sustainability has made progress on model-level and hardware-level optimizations (EXAMPLES), the platform dimension, i.e., deciding how, when, and which hardware is activated for inference, remains largely unexplored. As a result, serverless computing today does not yet realize its potential as a key abstraction for sustainable AI.

Against this background, the GEKO project addresses a clear research gap. Unlike public FaaS platforms, GEKO builds on an open-source foundation that allows transparent investigation of GPU integration and orchestration. Unlike existing open-source solutions, it focuses not only on enabling GPU support but also on making the serverless platform itself responsible for hardware usage decisions. This system-level focus avoids the mismatch between application-level assumptions and platform-level realities, ensuring that resources are shared and utilized as efficiently as possible. As a result, GEKO directly advances the state of the art in sustainable FaaS and establishes the groundwork for scalable, energy-efficient AI inference.

3 Partners and Previous Work

Das beschriebene Vorhaben ist in die Förderung im Software Campus mit Huawei Technologies Deutschland GmbH als Industriepartner sowie der Technischen Universität Berlin als akademischer Partner integriert. Der Industriepartner wird durch das Vorhaben nicht gefördert, bietet jedoch ein paralleles Mentoring und wertvolle Perspektiven aus der Industrie.

3.1 Research partner – Technische Universität Berlin

Die Technische Universität Berlin ist eine der renommiertesten technischen Universitäten in Deutschland. Mit einer breiten Palette von Studiengängen und einer starken Fokussierung auf Forschung und Innovation ist die Technische Universität Berlin eine der führenden Einrichtungen in Europa für technische Bildung und Wissenschaft.

Im Projekt wird die Technische Universität Berlin durch das Fachgebiet Mobile Cloud Computing vertreten, das 2017 gegründet wurde und Teil des Einstein Center Digital Future Berlin ist. Unter der Leitung von Prof. Dr.-Ing. David Bermbach erforscht das Fachgebiet das Software-technische Design und die Experiment-getriebene Bewertung verteilter IT-Systeme im Kontext moderner Anwendungsdomänen. Der Fokus liegt hier derzeit im Bereich der Datenmanagementsysteme und Anwendungsarchitekturen im Cloud-, Edge- und Fog-Computing, insbesondere bezogen auf Fragen des Placements von Daten und Anwendungskomponenten.

Wie in section 2 beschrieben konnte die Forschungsgruppe um Prof. Dr.-Ing. David Bermbach bereits zahlreiche Forschungsergebnisse zu Softwarearchitekturen und Anwendungsparadigmen für Edge und In-Network Computing in massiven LEO-Satellitennetzwerken auf international renommierten Konferenzen und Workshops veröffentlichen. Als Teil des durch das BMBF geförderten Verbundprojektes *6G NeXt* wird zudem bereits die Expertise des Fachgebiets zur Erforschung von Mobilfunk-, Telekommunikations- und Netzwerktechnologien der sechsten Generation (6G) beigetragen und ausgebaut.

3.2 Industry partner – Huawei Technologies Deutschland

Huawei Technologies ist ein weltweit führender Anbieter von Informations- und Kommunikationstechnologie (IKT) mit Präsenz in über 170 Ländern und einem umfassenden Portfolio an Telekommunikationsprodukten und -lösungen. In Deutschland ist Huawei seit 2001 aktiv und hat einen bedeutenden Einfluss auf die Wirtschaft und das Wachstum des Landes, indem es 2018 eine Bruttowertschöpfung von fast 2,3 Milliarden Euro und Beschäftigungseffekte für mehr als 28.000 Menschen generierte. Huawei spielt eine entscheidende Rolle bei der Digitalisierung und der Einführung von Technologien wie 5G/6G, die die Grundlage für Smart Citys, Industrie 4.0 und nachhaltige Mobilität schaffen.

Für das Projekt bringt Huawei Technologies Deutschland entscheidende Erfahrungen in der Forschung und Entwicklung von 6G-Technologien und nicht-terrestrischen Netzwerken mit und kann zudem wertvolle Perspektiven aus Seiten der Industrie geben. Besonders das Advanced Wireless Technologies Lab hat bereits signifikante Expertise in den Projekt-Kernforschungsbereichen Serverless"=Computing [Elz+22], Edge- und In-Network-Computing [Bra+22] und nicht-terrestrischen und LEO-Satellitennetzwerken [Cau+22].

3.3 Relationship between research partner and industry partner

Im Rahmen des Projekts SPENCER gibt es eine enge Zusammenarbeit zwischen der Technischen Universität Berlin und Huawei Technologies Deutschland, insbesondere zwischen dem Fachgebiet Mobile Cloud Computing (Technische Universität Berlin) und dem Advanced Wireless Technologies Lab (Huawei Technologies Deutschland GmbH). Diese Zusammenarbeit beinhaltet vordergründig einen bilateralen wissenschaftlichen und technischen Austausch über die gesamte Laufzeit des Projekts hinweg. Die hohen Fachkompetenzen beider Partner steigern hier die Erfolgsaussichten für das Projekt: Als Mentor und Ansprechpartner bei Huawei wird Dr.-Ing. Osama Abboud, Senior Researcher des

Advanced Wireless Technologies Lab am Huawei-Technologies-Standort München, die Anforderungsanalyse, Forschung und Entwicklung im Projekt aktiv mitgestalten und wertvolles Feedback geben, das zur kontinuierlichen Qualitätssicherung dient. Als Leiter des Fachgebiets Mobile Cloud Computing und fachlicher Mentor des Mikroprojektleiters wird Prof. Dr.-Ing. David Bermbach die wissenschaftliche Qualitätssicherung übernehmen und die Umsetzung von Publikationen unterstützen.

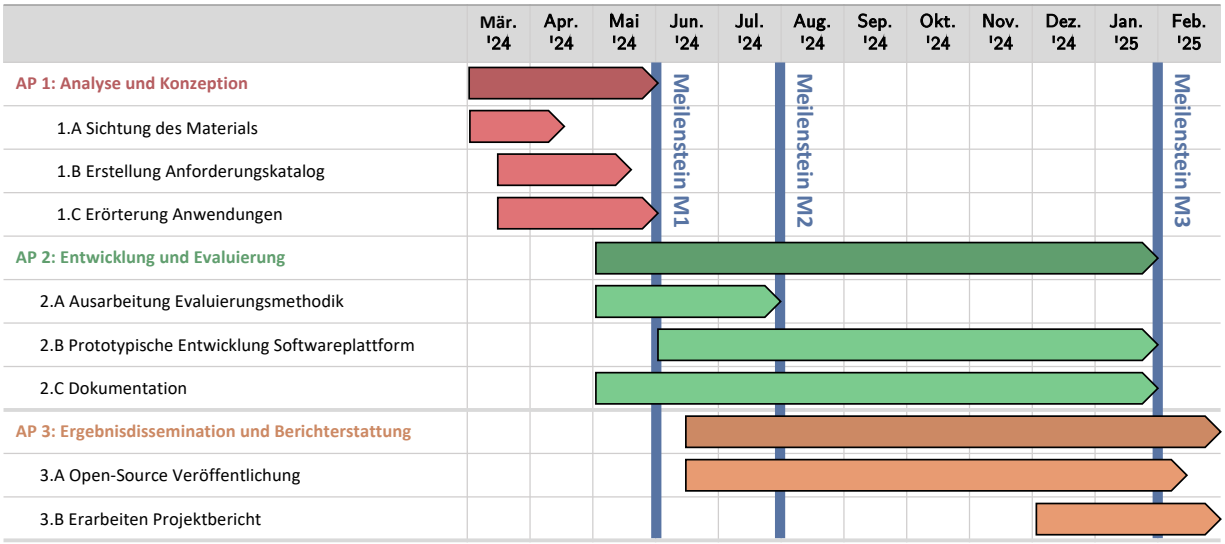

Figure 4.1: Zeitplan des Projekts SPENCER.

4 Detailed Description of the Work Plan

Im Folgenden werden die geplanten Arbeitspakete und Meilensteine (Section 4.1), Zeit- und Ressourcenplanung (Section 4.2) und Finanzplanung (Section 4.3) im Detail beschrieben.

4.1 Work packages and milestones

Das Projekt SPENCER ist in die drei Phasen 'Analyse und Konzeption', 'Entwicklung und Evaluierung' und 'Ergebnisdissemination und Berichterstattung' unterteilt. Der Zeitplan der Arbeitspaketbearbeitung findet sich in figure 4.1.

4.1.1 AP1: ...

In der Analyse- und Konzeptionsphase liegt der Fokus auf einer umfassenden Untersuchung aktueller Publikationen wie Studien, Patente und Forschungsberichte sowie der Analyse technologischer Entwicklungen und wirtschaftlicher Rahmenbedingungen (Punkt 1.A). Durch diese systematische Sichtung sollen fundierte Erkenntnisse gewonnen werden, die als Grundlage für die weitere Ausarbeitung des Projekts dienen.

Die enge Zusammenarbeit mit dem Industriepartner ermöglicht es, den Rahmen, die Grundlagen und die Qualitätsdimensionen des Projekts zu definieren. Hierbei werden sowohl funktionale als auch nicht-funktionale Ziele identifiziert und in einem präzisen Anforderungskatalog festgehalten (Punkt 1.B). Dieser Katalog dient als Leitfaden für die nachfolgenden Entwicklungs- und Evaluierungsphasen des Projekts.

Ein besonderes Augenmerk liegt dabei auf möglichen Anwendungen aus Domänen, die von Edge-Computing profitieren können (Punkt 1.C). Die Identifikation und Betrachtung dieser Anwendungsfelder für die Bereitstellung global verteilter Software trägt dazu bei, zukünftige Entwicklungen und Chancen im Bereich des Edge-Computing zu antizipieren.

Als entscheidenden Meilenstein des ersten Arbeitspakets ist geplant, am Ende eine umfassende Grundlagenstudie zum Ist-Zustand vorzulegen (Meilenstein M1). Diese Studie wird nicht nur einen detaillierten Anforderungskatalog umfassen, sondern auch eine tiefgehende Analyse aktueller Grundlagen und Entwicklungen im Bereich des global verteilten Edge-Computing bieten. Die Ausarbeitung wird zudem eine präzise Auswahl von exemplarischen Anwendungsfeldern einschließen.

Die Grundlagenstudie wird als ein strategisches Dokument fungieren, das nicht nur den aktuellen Stand des Wissens und der Anforderungen widerspiegelt, sondern auch als Leitfaden für die nächsten Phasen des Projekts dient. Durch die klare Definition von Rahmen, Qualitätsdimensionen und konkreten Zielen legt die Studie den Grundstein für eine effektive und zielgerichtete Umsetzung des Projekts. Ihr Abschluss markiert somit nicht nur das Ende der Analyse- und Konzeptionsphase, sondern gleichzeitig den Startpunkt für die aufbauenden Schritte in der Entwicklungs- und Evaluierungsphase.

4.1.2 AP2: ...

In der Phase der "Entwicklung und Evaluierung" erfolgt zunächst die gezielte Auswahl von drei Softwareanwendungen, innerhalb derer die Erfüllung der zuvor definierten Anforderungen aus Anwendungsperspektive intensiv untersucht werden kann. Dies unterliegt dem Ansatz eines Test-Driven-Development-Prozess, der sicherstellt, dass die im Anforderungskatalog festgelegten Anforderungen systematisch in Evaluierungsmethoden umgesetzt werden, die während des Entwicklungsprozesses kontinuierlich Anwendung finden. Hierzu gehören beispielsweise Tests für funktionale Aspekte sowie Performance-Benchmarks für nicht-funktionale Anforderungen (Punkt 2.A). Die entwickelteten wiederverwendbaren Evaluationsmethodiken und -Umgebungen stellen somit den zweiten Meilenstein des Projekts dar (Meilenstein M2). Sie bilden nicht nur die Grundlage für die darauf folgende Implementierung und Integration, sondern ermöglichen auch eine fundierte Evaluation und potenzielle Anpassung der entwickelten Lösung im Hinblick auf die ursprünglichen Anforderungen.

Durch eine kontinuierliche und strenge Anwendung dieser Evaluierungsmethoden an Prototypen wird so anschließend agil eine Serverless Edge-Computing Plattform entwickelt. Diese Prototypenentwicklung orientiert sich an bestehenden Open-Source Software-Plattformen, die iterativ weiterentwickelt werden, um schließlich in einem finalen Software-Prototypen zu konvergieren (Punkt 2.B). Die Entwicklung dieses Software-Prototypen umfasst die Entwicklung zweier Hauptkomponenten: Erstens wird eine Serverless-Laufzeitumgebung entwickelt, die verschiedene Anwendungsdienste parallel aber isoliert auf begrenzten Resourcen ausführen kann und in der Lage ist, Resourcen schnell zwischen Diensten auszutauschen um Elastizität sicherzustellen. Zweitens wird parallel aber in enger Absprache ein Subsystem zur Migration von Anwendungsdienstinstanzen und Dienstzustandsdaten entwickelt, um einen Transfer von Diensten entgegen der Mobilität der unterliegenden Satelliten-Infrastruktur zu gewährleisten. Der Software-Prototyp, der die definierten Anforderungen erfüllt, stellt damit den dritten Meilenstein des Projekts dar (Meilenstein M3).

Die erfolgreiche Umsetzung dieses Entwicklungsprozesses erfordert zudem die Erstellung einer detaillierten technischen Dokumentation (Punkt 2.C). Diese Dokumentation bildet nicht nur die Grundlage für die interne Weiterentwicklung der Plattform, sondern stellt auch sicher, dass das erarbeitete Wissen und die Funktionsweise der entwickelten Software auch über das Ende des Projekts hinaus transparent und nachvollziehbar sind.

4.1.3 AP3: ...

In der Phase der "Ergebnisdissemination und Berichterstattung" stehen zwei entscheidende Aktivitäten im Mittelpunkt: Zum einen erfolgt die Open-Source Veröffentlichung der entwickelten Softwarelösung (Punkt 3.A). Dies ermöglicht anderen Fachleuten und Organisationen, von den erzielten Erkenntnissen und der entwickelten Serverless Edge-Computing Plattform zu profitieren. Die Open-Source Veröffentlichung gewährleistet einen offenen Zugang zu Quellcode, Dokumentation und relevanten Ressourcen, was nicht nur die Nachvollziehbarkeit und Überprüfbarkeit der Ergebnisse fördert, sondern auch eine Grundlage für zukünftige Weiterentwicklungen und Innovationen in der breiteren Gemeinschaft schafft. Für diese Veröffentlichung ist es notwendig, die Code-Qualität der bestehenden Prototypen sicherzustellen und die Dokumentation der Software entsprechend zu formatieren. Dieser Prozess läuft kontinuierlich zur weiteren Entwicklung und Erweiterung der Software-Prototypen.

	AP1	AP2	AP3	$\sum$
Projektleitung	1 PM	1 PM	2 PM	**4 PM**
Stud. Hilfskraft 1	1 PM	4,5 PM	0,5 PM	**6 PM**
Stud. Hilfskraft 2	1 PM	4,5 PM	0,5 PM	**6 PM**
Stud. Hilfskraft 3	1 PM	4,5 PM	0,5 PM	**6 PM**
Stud. Hilfskraft 4	1 PM	4,5 PM	0,5 PM	**6 PM**
Stud. Hilfskraft 5	1 PM	4,5 PM	0,5 PM	**6 PM**
Stud. Hilfskraft 6	1 PM	3 PM	2 PM	**6 PM**
Stud. Hilfskraft 7	1 PM	3 PM	2 PM	**6 PM**
$\sum$	**8 PM**	**29,5 PM**	**8,5 PM**	**46 PM**

Table 4.1: Übersicht Planung Personenmonate (PM) pro Arbeitspaket (AP)

Im letzten Quartal des Projekts wird ein Projektbericht erarbeitet, der den gesamten Projektablauf detailliert dokumentiert und Einblick in die Analyse-, Entwicklungs- und Evaluierungsphasen gibt (Punkt 3.B). Der Bericht soll eine umfassende Bewertung der erreichten Meilensteine bieten, Herausforderungen bei der Umsetzung des Projekts und deren Lösungsansätze analysieren und Empfehlungen für mögliche Weiterentwicklungen geben. Die Verfassung dieses Berichts erfordert eine präzise Zusammenstellung von technischen Details, methodischen Ansätzen, Ergebnissen und Schlussfolgerungen, um einen ganzheitlichen Einblick in die Projektergebnisse zu gewähren.

4.2 Time and ressource plan

Mit der inhaltlichen Bearbeitung des Projekts SPENCER sollen sieben studentische Hilfskräfte von je 80 Monatsstunden (entspricht 6 Personenmonate pro Person pro Jahr oder 2 Zeitmonate pro Personenmonat) betraut werden (genauere Finanzplanung in section 4.3.1). Die Zuordnung von Personenmonaten zu Arbeitspaketen findet sich in table 4.1. Neben der Projektleitung, die jeweils die Anleitung der Arbeitspakete und Mitarbeitenden übernimmt, werden fünf studentische Hilfskräfte vordergründig mit der inhaltlichen Bearbeitung des Arbeitspakets AP2 betraut, während zwei weitere studentische Hilfskräfte teilweise auch die Umsetzung von Arbeitspaket AP3 übernehmen. Alle studentischen Hilfskräfte sind in allen Arbeitspaketen involviert, um den wissenschaftlichen Erfolg des Projekts zu sichern: Jede studentische Hilfskraft soll also sowohl an Konzeption (AP1), Entwicklung (AP2) und Ergebnisdissemination (AP3) beteiligt sein.

Alle studentischen Hilfskräfte erbringen je einen Personenmonat zu Arbeitspaket AP1, das so in Summe acht Personenmonate verlangt. Fünf studentische Hilfskräfte erbringen je 4,5 Personenmonate zu Arbeitspaket AP2 und sind damit vordergründig mit der Entwicklung, Evaluierung und Dokumentation des Software-Prototypen beschäftigt. Zwei weitere studentische Hilfskräfte erbringen einen reduzierten Umfang von drei Personenmonaten zur Umsetzung des Arbeitspakets AP2. Indes erbringen diese zwei studentischen Hilfskräfte eine erhöhte Anzahl von je zwei Personenmonaten zu Arbeitspaket AP3 und sind so auch für die Open-Source-Veröffentlichung des Software-Prototypen verantwortlich. Die übrigen fünf studentischen Hilfskräft erbringen für dieses Arbeitspaket AP3 jeweils 0,5 Personenmonate, die vordergründig in das Erarbeiten des Projektberichts fließen, an denen alle Mitarbeitenden beteiligt sein sollen.

Im Übrigen entfallen mit Planungs-, Koordinierungs- und Leitungsaufgaben je ein Personenmonat für Arbeitspakete AP1 und AP2 auf die Projektleitung. Da Arbeistpaket AP3 zusätzliche Planung

Finanzposten	**Positions-Nr.**	**Einzelposten**	**Ausgaben**
Personal	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
	0822	Studentische Hilfskraft á 80 Monatsstunden (TV Stud III)	14.400,00 €
		$\sum$ **Personalausgaben**	**100.800,00 €**
Dienstreisen	0846	Software Campus	1.720,00 €
	0846	Industriepartner	3.776,00 €
	0846	Konferenzen	4.630,00 €
		$\sum$ **Dienstreisen**	**10.126,00 €**
		$\sum$ **Ausgaben**	**110.926,00 €**

Table 4.2: Übersicht der Finanzplanung

und Koordinierung erfordert, werden hierfür zwei Personenmonate der Projektleitung kalkuliert. Es ergeben sich so in Summe acht Personenmonate für die Bearbeitung des Arbeitspakets AP1, 29,5 Personenmonate für die Bearbeitung des Arbeitspakets AP2 und 8,5 Personenmonate für die Bearbeitung des Arbeitspakets AP3.

4.3 Financial planning

Basierend auf der Arbeits- und Ressourcenplanung in sections 4.1 to 4.2 stellen wir nun die finanzielle Planung des Projekts SPENCER vor. Dies stellt eine Vorkalkulation nach bestem Wissen und Gewissen dar, der der Grundsatz eines sparsamen und angemessenen Mitteleinsatzes zugrunde liegt.

Eine Übersicht der Finanzplanung kann table 4.2 entnommen werden. Es liegen nur die Finanzposten Personal (section 4.3.1) und Dienstreisen (section 4.3.2) vor. Es sind keine Fremdaufträge für Forschung und Entwicklung geplant. Es wird eine Gesamtsumme von 110.926,00 € beantragt.

4.3.1 Personalausgaben

Wie in section 4.2 beschrieben werden sieben studentische Hilfskräfte mit der inhaltlichen Bearbeitung des Projekts betraut und erhalten dafür je eine Anstellung über 80 Monatsstunden nach dem an der Technischen Universität Berlin geltenden Tarifvertrages für studentische Beschäftigte III. Die Personalausgaben belaufen sich somit auf 14.400,00 € pro studentischer Hilfskraft pro

Bezeichnung	**Ziel**	**Kategorie**	**An-/Abreise**	**Gebühr**	**Hotel**	**Tagegeld**	**Dauer**	**Gesamt**
SWC Summit	Deutschland	Software Campus	100,00 €	–	80,00 €	28,00 €	3 Tage	344,00 €
SWC Training 1	Deutschland	Software Campus	100,00 €	–	80,00 €	28,00 €	3 Tage	344,00 €
SWC Training 2	Deutschland	Software Campus	100,00 €	–	80,00 €	28,00 €	3 Tage	344,00 €
SWC Training 3	Deutschland	Software Campus	100,00 €	–	80,00 €	28,00 €	3 Tage	344,00 €
SWC Training 4	Deutschland	Software Campus	100,00 €	–	80,00 €	28,00 €	3 Tage	344,00 €
Besuch Industriepartner Q2 (PL)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q2 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (PL)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
Besuch Industriepartner Q4 (Stud. HK)	München, Deutschland	Industriepartner	100,00 €	–	80,00 €	28,00 €	2 Tage	236,00 €
IEEE IC2E 2024	Paphos, Zypern	Konferenz	415,00 €	850,00 €	125,00 €	35,00 €	6 Tage	2.100,00 €
ACM/IFIP Middleware 2024	Hongkong, SAR, China	Konferenz	483,00 €	750,00 €	145,00 €	61,00 €	7 Tage	2.530,00 €
$\sum$		Software Campus						1.720,00 €
		Industriepartner						3.776,00 €
		Konferenz						4.630,00 €
$\sum$								**10.126,00 €**

Table 4.3: Übersicht der geplanten Reisen und deren zugehöriger Ausgaben

Jahr. Dies deckt jegliche Lohnausgaben und Steuern des Beschäftigungsverhältnisses ab und enthält sowohl den Rentenversicherungsbeitrag als auch den gesetzlichen Beitrag zur Unfallkasse Berlin. In Summe belaufen sich die Personalausgaben damit auf 100.800,00 €. Es sei an dieser Stelle darauf hingewiesen, dass die Einstellung von studentischen Hilfskräften der Stärkung der wissenschaftlichen Erfolgsaussichten dienen soll. Je nach Qualifikationsstand sollen alle Mitarbeiter:innen im Rahmen von Folgeprojekten an das Fachgebiet Mobile Cloud Computing gebunden werden, beispielsweise durch die Möglichkeit zur Promotion (vgl. section 5.3). Die technische Ausstattung der studentischen Beschäftigten wird durch das Fachgebiet Mobile Cloud Computing der Technischen Universität Berlin gestellt.

4.3.2 Dienstreisen

Obschon digitale Kommunikationsformate, etwa E-Mail, Chat oder Videotelefonie-Schalten, flexible Möglichkeiten zum Projekt-bezogenen Austausch der Stakeholder bieten, sind Treffen in Person dem Erfolg des Projekts unerlässlich. Dies erfordert regelmäßige Dienstreisen der mit dem Projekt betrauten Mitarbeiter:innen. Konkret beinhaltet dies drei Arten von Reisen, deren Finanzplanung in diesem Abschnitt genauer beleuchtet wird: Summits und Trainings im Rahmen des Software Campus (Absatz 4.3.2.1), Treffen mit dem Industriepartner (Absatz 4.3.2.2) und der Besuch internationaler wissenschaftlicher Konferenzen (Absatz 4.3.2.3). Eine Übersicht über die Finanzplanung für Dienstreisen ist table 4.3 zu entnehmen.

4.3.2.1 Software-Campus-Summit und -Trainings

Für Projektleiter:innen von Mikroprojekten im Rahmen des Software Campus ist die Teilnahme an einem 'Summit' im Jahr und insgesamt vier Trainings verpflichtend. Diese Veranstaltungen finden an verschiedenen, bisher jedoch noch nicht bekannten Orten in Deutschland statt und werden jeweils

zwei Übernachtungen in Anspruch nehmen. Für die Teilnahme an diesen Veranstaltungen durch die Projektleitung wird jeweils eine An- und Abreise von insgesamt 100,00 € veranschlagt, etwa mit der Deutschen Bahn. Zudem werden auf Grundlage der Empfehlungen und Regeln der Technischen Universität Berlin pauschal Übernachtungsausgaben (etwa in Hotels) von 80,00 € pro Nacht kalkuliert. Auf Grundlage des Bundesreisekostengesetzes in aktueller Fassung wird ein Tagegeld von 28,00 € geplant. Es fallen keine Teilnahmegebühren an.

4.3.2.2 Industriepartner-Besuche

Über die Laufzeit des Projektes von einem Jahr sind zwei Besuche beim Industriepartner Huawei Technologies Deutschland in München geplant, an dem Projektleitung und studentische Beschäftigte teilnehmen, um die aktuelle Projektlage in einem Intensiv-Workshop zu besprechen und weitere Schritte zu planen. Diese Besuche dienen auch der Dissemination von Projektergebnissen innerhalb der Organisation des Industriepartners, etwa mit Vorträgen. Ähnlich zu den Veranstaltungen, deren Finanzplanung in Absatz 4.3.2.1 umrissen ist, werden pauschale An- und Abreiseausgaben von 100,00 € pro Person pro Besuch veranschlagt. Zudem werden erneut pauschale Übernachtungsausgaben von 80,00 € pro Nacht und Tagegeld von 28,00 € berechnet, indes ist nur eine Dauer von zwei Tagen (eine Übernachtung) pro Besuch geplant. Es fallen keine Teilnahmegebühren an.

4.3.2.3 Besuch von wissenschaftlichen Konferenzen und Workshops

In der Informatik im Allgemeinen und im Forschungsgebiet Rechnersysteme und verteilte Systeme im Besonderen stellen Veröffentlichungen auf internationalen Konferenzen und Workshops den primären Publikationsweg dar, da sie eine zeitgemäße Dissemination von Forschungsergebnissen ermöglichen. Für die Veröffentlichung von Publikationen in Tagungsbänden von Konferenzen und Workshops ist der Besuch dieser Veranstaltungen in Verbindung mit einem Vortrag über die Forschungsergebnisse verpflichtend. Traditionell wird diese Aufgabe durch die/den Erstautor/in einer Publikation übernommen. Basierend auf der bisherigen Publikationsrate des Projektleiters (vgl. sections 2 and 3.1) werden für das Projekt SPENCER insgesamt zwei Besuche von Konferenzen veranschlagt. Es sei darauf hingewiesen, dass international anerkannte Tagungen nur international abgehalten werden.

Die geplanten Konferenzen stellen international anerkannte Tagungen dar, auf denen bereits Vorarbeiten der Projektleitung zum Projektthema vorgestellt wurden (vgl. section 2). Die in table 4.3 angeführten An- und Abreiseausgaben basieren auf aktuellen Angeboten, die in appendix A.1 aufgeführt sind. Es wird mit zeitgemäßer An- und Abreise einen Tag vor respektive einen Tag nach der Veranstaltung kalkuliert. Die angeführten Tagungsgebühren beziehen sich auf Gebühren der Tagungsinstanzen 2023. Die kalkulierten Übernachtungsausgaben beziehen sich basierend auf den Empfehlungen und Vorschriften der Technischen Universität Berlin auf die *Allgemeine Verwaltungsvorschrift über die Neufestsetzung der Auslandstage- und Auslandsübernachtungsgelder vom 13.10.2022* (ARVVwV) und stellen somit erwartbare Höchstausgaben dar. Es kann nicht ausgeschlossen werden, dass spezielle Hotelangebote für Tagungsteilnehmende angeboten werden, die zum jetzigen Zeitpunkt jedoch noch nicht vorliegen. Auch das veranschlagte Tagegeld bezieht sich auf die ARVVwV in aktueller Fassung.

Die *IEEE International Conference on Cloud Engineering* (IEEE IC2E) ist eine bedeutende Veranstaltung für Forschende, die sich mit Cloud-Computing und verwandten Themen befassen. Die Konferenz konzentriert sich auf die neuesten Trends in Cloud-Engineering, einschließlich Cloud-Architekturen, Dienstgüte, Sicherheit und Skalierbarkeit. Die IEEE IC2E bietet eine Plattform für den Wissensaustausch und die Diskussion über bewährte Praktiken, um die Leistung und Zuverlässigkeit von Cloud-Infrastrukturen zu verbessern. Die IEEE IC2E 2024 findet vom 24. bis 27. September 2024 in Paphos, Zypern statt (vgl. Konferenzwebsite).

Die *ACM/IFIP International Middleware Conference* (ACM/IFIP Middleware) ist eine führende Konferenz im Bereich Middleware-Technologien und -systeme. Die Veranstaltung bringt Wissenschaftler:innen, Ingenieur:innen und Industrieexpert:innen zusammen, um aktuelle Entwicklungen in den Bereichen

Software-Middleware, verteilte Systeme und Dienstorientierte Architekturen zu präsentieren und zu diskutieren. Middleware spielt eine entscheidende Rolle bei der Integration von Anwendungen und Diensten, weshalb die Konferenz einen bedeutenden Beitrag zur Entwicklung von Middleware-Lösungen leistet. Die ACM/IFIP Middleware 2024 findet vom 11. bis 15. Dezember 2024 in Hongkong, SAR, China statt (vgl. Konferenzwebsite).

5 Utilization Plan

In diesem Abschnitt werden die Chancen der Verwertung der entstandenen Erkenntnisse, Lösungen und Software umrissen.

5.1 Economic prospects of success

Obschon die wirtschaftliche Verwertung der im Projekt SPENCER gewonnen Erkenntnisse und Lösungen nicht der Fokus des Projekts ist, bieten sich verschiedene Gelegenheiten zur industriellen Nutzbarkeit der Projektergebnisse. Zunächst werden jegliche im Rahmen des Projekts entstandenen Software-Artefakte als Open-Source-Software veröffentlicht (vgl. AP3 Punkt 3.A in section 4.1.3). Dank einer Verwendung von Lizenzmodellen, die sowohl kommerzielle als auch nicht-kommerzielle Nutzung ermöglichen, steht diese Open-Source-Software der Allgemeinheit sowohl zur direkten Nutzung als auch zur Weiterentwicklung zur Verfügung.

Auch der ungeförderte Industriepartner Huawei Technologies Deutschland GmbH kann jegliche Projektergebnisse in eigenen Produkten und Dienstleistungen nutzen. Zusätzlich profitiert Huawei Technologies Deutschland GmbH zudem durch den mit der Kooperation mit dem Fachgebiet Mobile Cloud Computing der Technischen Universität Berlin entstehenden Wissenstransfer.

5.2 Scientific and technical prospects of success

Im Vordergrund des Projekts SPENCER steht indes die wissenschaftliche Verwertung. Diese Verwertung ist in zwei Dimensionen zu verstehen: Zunächst bietet das Projekt exzellente Möglichkeiten zur Veröffentlichung von Projektergebnissen in international renommierten wissenschaftlichen Tagungen (vgl. Absatz 4.3.2.3). Dies wird unter anderem durch die hohe Anzahl bereits durch die Projektleitung publizierter wissenschaftlicher Arbeiten von hoher Qualitätsgüte im Forschungsgebiet belegt (vgl. section 2). Geplant sind insgesamt zwei Publikationen im Rahmen des Projekts, für die jeweils die Projektleitung die Erstautorschaft übernimmt. Diese zwei Publikationen entfallen auf die internationalen wissenschaftlichen Konferenzen *IEEE IC2E 2024* und *ACM/IFIP Middleware 2024* (vgl. Absatz 4.3.2.3). Zusätzlich wird ein Projektbericht als technischer Bericht ohne Qualitätssicherung veröffentlicht (vgl. AP3 Punkt 3.B in section 4.1.3)

Zusätzlich bietet das Projekt SPENCER hervorragende Möglichkeiten zur Qualifikation des wissenschaftlichen Nachwuchses. Wie in section 4.3.1 umrissen, sollen die studentischen Beschäftigten auch nach Ablauf des Projekts in Anschlussprojekten an das Fachgebiet Mobile Cloud Computing gebunden werden und je nach Qualifikation beispielsweise Möglichkeit zur Promotion an der Technischen Universität Berlin erhalten (vgl. auch section 5.3). Die Grundlagen des wissenschaftlichen Arbeitens auf Promotionsniveau werden dazu bereits im Rahmen des Projekts SPENCER vermittelt. Darüber hinaus kann das Projekt auch den Rahmen für Qualifikationsarbeiten von Studierenden der Technischen Universität Berlin auf Bachelor- und Master-Niveau bieten, etwa indem Abschluss- und Projektarbeiten zu Projektthemen vergeben und betreut werden, deren Ergebnisse wissenschaftlich verwertet werden können.

5.3 Scientific and economic connectivity

Das Forschungsgebiet massive Satellitennetzwerke ist, wie in section 2 beschrieben, noch verhältnismäßig jung. Das Projekt SPENCER ist eines der ersten Forschungsprojekte, das das spezifische Thema des Edge- und In-Network-Computing in massiven LEO-Satellitennetzwerken erst erschließen. Damit bietet das Projekt eine vortreffliche Grundlage für Folgeprojekte, deren genaue Themen erst im Laufe der intensiven wissenschaftlichen Auseinandersetzung mit dem Themenkomplex im Rahmen des Projekts eruiert werden können. Entsprechende Projekte ließen sich beispielsweise erneut im Rahmen des Programms Software Campus oder entsprechende Förderaufrufe zu Themengebieten wie Mobilfunktechnologien der sechsten Generation (6G) durch Bundesministerien fördern.

6 Bibliography

[Bra+22] Florian Brandherm, Julien Gedeon, Osama Abboud, and Max Mühlhäuser. "BigMEC: Scalable Service Migration for Mobile Edge Computing". In: *Proceedings of the 2022 IEEE/ACM 7th Symposium on Edge Computing* (Seattle, WA, USA). SEC. New York, NY, USA: IEEE, Dec. 2022, pp. 136–148. DOI: 10.1109/SEC54971.2022.00018.

[Cau+22] Màrius Caus, Musbah Shaat, Ana I. Pérez-Neira, Malte Schellmann, and Hanwen Cao. "Reliability Oriented OTFS-based LEO Satellites Joint Transmission Scheme". In: *Proceedings of the 2022 IEEE Globecom Workshops* (Rio de Janeiro, Brazil). GC Wkshps. New York, NY, USA: IEEE, Dec. 2022, pp. 1406–1412. DOI: 10.1109/GCWkshps56602.2022.10008593.

[Chi21] Andrew A. Chien. "Driving the cloud to true zero carbon". en. In: *Communications of the ACM* 64.2 (Jan. 2021), pp. 5–5. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3445037.

[DWD25] DWD, Geschäftsbereich Klima und Umwelt. *Klimastatusbericht Deutschland Jahr 2024*. Offenbach, 2025.

[Elz+22] Mohamed Elzohairy, Mohak Chadha, Anshul Jindal, Andreas Grafberger, Jianfeng Gu, Michael Gerndt, and Osama Abboud. "FedLesScan: Mitigating Stragglers in Serverless Federated Learning". In: *Proceedings of the 2022 IEEE International Conference on Big Data* (Osaka, Japan). Big Data. New York, NY, USA: IEEE, Dec. 2022, pp. 1230–1237. DOI: 10.1109/BigData55660.2022.10021037.

[GF20] Samuel Ginzburg and Michael J. Freedman. "Serverless Isn't Server-Less: Measuring and Exploiting Resource Variability on Cloud FaaS Platforms". en. In: *Proceedings of the 2020 Sixth International Workshop on Serverless Computing*. Delft Netherlands: ACM, Dec. 2020, pp. 43–48. ISBN: 978-1-4503-8204-5. DOI: 10.1145/3429880.3430099.

[Gan+23] Anshul Gandhi, Dongyoon Lee, Zhenhua Liu, Shuai Mu, Erez Zadok, Kanad Ghose, Kartik Gopalan, Yu David Liu, Syed Rafiul Hussain, and Patrick Mcdaniel. "Metrics for Sustainability in Data Centers". en. In: *ACM SIGEnergy Energy Informatics Review* 3.3 (Oct. 2023), pp. 40–46. ISSN: 2770-5331. DOI: 10.1145/3630614.3630622.

[Jon+19] Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti, Chia-Che Tsai, Anurag Khandelwal, Qifan Pu, Vaishaal Shankar, Joao Carreira, Karl Krauth, Neeraja Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson. *Cloud Programming Simplified: A Berkeley View on Serverless Computing*. en. arXiv:1902.03383 [cs]. Feb. 2019. DOI: 10.48550/arXiv.1902.03383.

[Mas+20] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. "Recalibrating global data center energy-use estimates". en. In: *Science* 367.6481 (Feb. 2020), pp. 984–986. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aba3758.

[Pat+21] Panos Patros, Josef Spillner, Alessandro V. Papadopoulos, Blesson Varghese, Omer Rana, and Schahram Dustdar. "Toward Sustainable Serverless Computing". In: *IEEE Internet Computing* 25.6 (Nov. 2021), 42–50. ISSN: 1941-0131. DOI: 10.1109/mic.2021.3093105.

[SF24] Prateek Sharma and Alexander Fuerst. "Accountable Carbon Footprints and Energy Profiling For Serverless Functions". en. In: *Proceedings of the ACM Symposium on Cloud Computing*. Redmond WA USA: ACM, Nov. 2024, pp. 522–541. ISBN: 979-8-4007-1286-9. DOI: 10.1145/3698038.3698531.

[Sch+23] Trever Schirmer, Nils Japke, Sofia Greten, Tobias Pfandzelter, and David Bermbach. "The Night Shift: Understanding Performance Variability of Cloud Serverless Platforms". In: *Proceedings of the 1st Workshop on SErverless Systems, Applications and MEthodologies*. SESAME '23. New York, NY, USA: Association for Computing Machinery, May 2023. DOI: 10.1145/3592533.3592808.

[Sha23] Prateek Sharma. "Challenges and Opportunities in Sustainable Serverless Computing". en. In: *ACM SIGEnergy Energy Informatics Review* 3.3 (Oct. 2023), pp. 53–58. ISSN: 2770-5331. DOI: 10.1145/3630614.3630624.

[Sha24] Prateek Sharma. "The Jevons Paradox In Cloud Computing: A Thermodynamics Perspective". In: (2024). arXiv: 2411.11540 [cs.DC].

[She+24] Arman Shehabi, Sarah Smith, Dale Sartor, Richard Brown, Magnus Herrlin, Jonathan Koomey, Eric Masanet, Nathaniel Horner, Inês Azevedo, and William Lintner. "United States Data Center Energy Usage Report". en. Tech. rep. LBNL–1005775, 1372902. June 2024, LBNL–1005775, 1372902. DOI: 10.2172/1372902.

[Suk+24] Thanathorn Sukprasert, Abel Souza, Noman Bashir, David Irwin, and Prashant Shenoy. "On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud". en. In: *Proceedings of the Nineteenth European Conference on Computer Systems*. Athens Greece: ACM, Apr. 2024, pp. 924–941. ISBN: 979-8-4007-0437-6. DOI: 10.1145/3627703.3650079.

[You+19] Ethan G Young, Pengfei Zhu, Tyler Caraza-Harter, Andrea C Arpaci-Dusseau, and Remzi H Arpaci-Dusseau. "The True Cost of Containing: A gVisor Case Study". In: *11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19)*. 2019.

A Appendix

A.1 Angebote Dienstreisen

Die hier aufgeführten Angebote für Dienstreisen umfassen Vergleichsangebote für Flugreisen mit zeitgemäßer An- und Abreise für die IEEE IC2E 2024 (appendix A.1.1) und die ACM/IFIP Middleware 2024 (appendix A.1.2).

A.1.1 An-/Abreise IEEE IC2E 2024

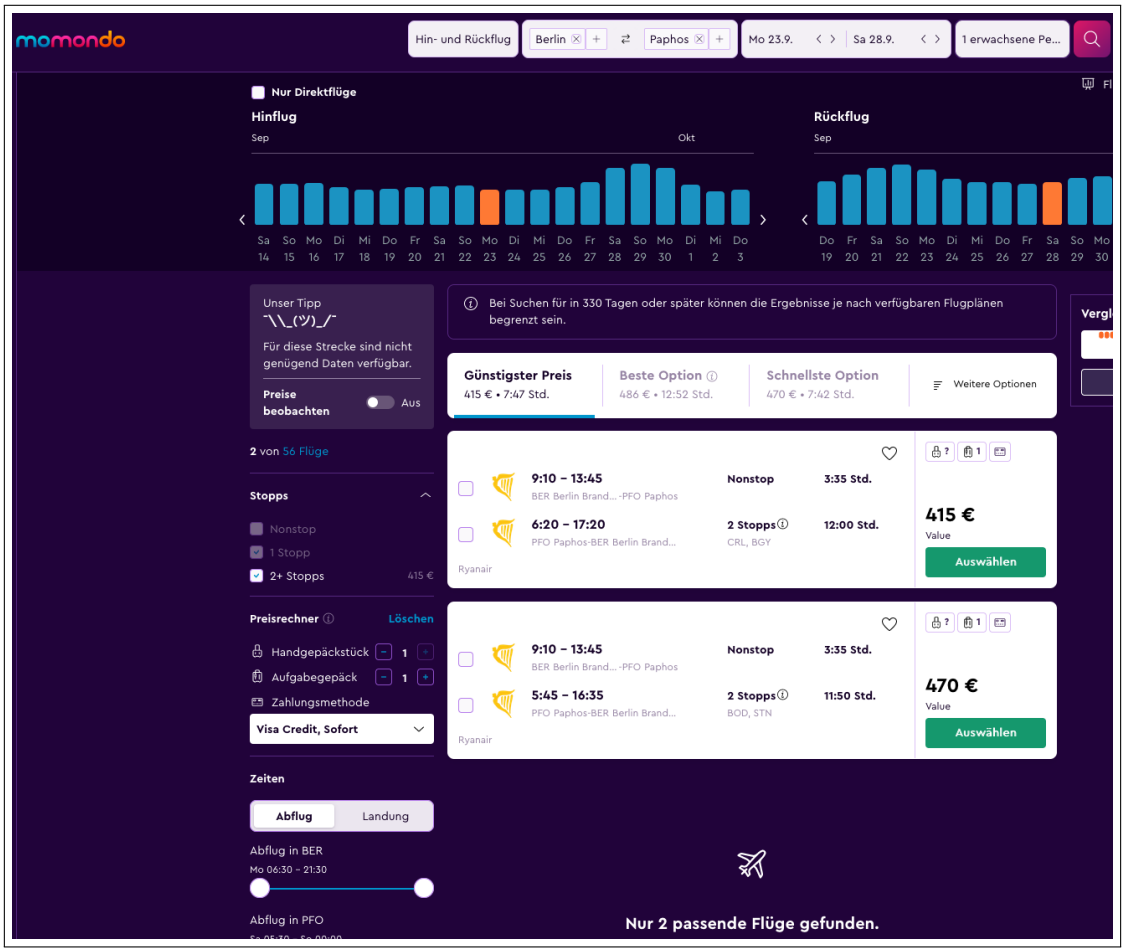

Figure A.1: Vergleichsangebote für die zeitgemäße An- und Abreise zur IEEE IC2E 2024 vom 24. bis 27. September 2024 in Paphos, Zypern. Zuletzt abgerufen am 16. Oktober 2023.

A.1.2 An-/Abreise ACM/IFIP Middleware 2024

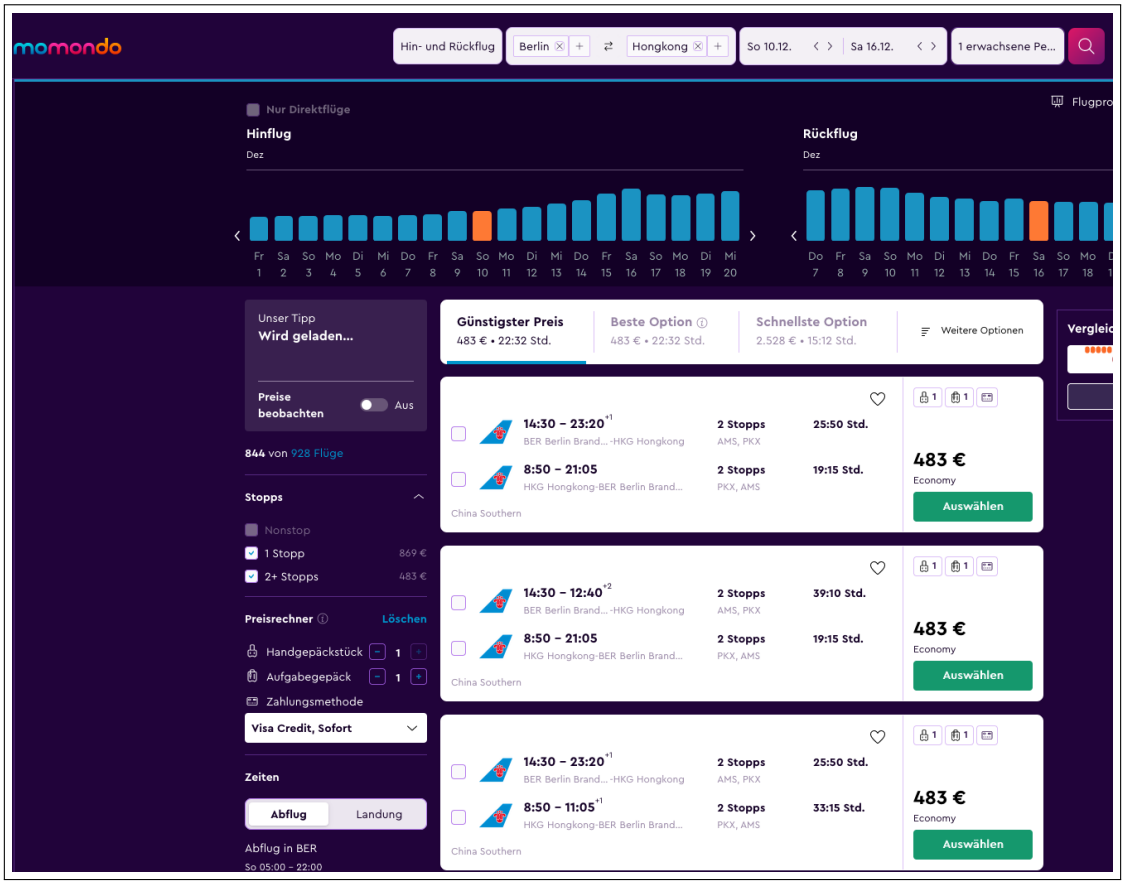

Figure A.2: Vergleichsangebote für die zeitgemäße An- und Abreise zur ACM/IFIP Middleware 2024 vom 11. bis 15. Dezember 2024 in Hongkong, SAR, China. Da die Reisedaten mehr als ein Jahr in der Zukunft liegen, sind noch keine Vergleichsangebote möglich. Es wird daher ein Vergleichsangebot für eine zeitgemäße Anreise für einen fiktiven Aufenthalt in Hongkong, SAR, China vom 11. bis 15. Dezember 2023 präsentiert. Zuletzt abgerufen am 09. November 2023.