CAPE – European Open Compute Architecture for Powerful Edge

Martin Kaiser*, Lennart Tigges*, Jens Hagemeyer*, Christian Klarhorst*, Björn Voß*, Fred Buining†, Bola Fakhoury†, János Lazányi‡, René Griessl§, Muhammad Shahzad§, Yiannis Georgiou¶, Salim Mimouni¶, Pedro Velho¶, Michael Mercier¶, Eva Trungel‖, Julian Gajewski**, Stefan Krupop**, Micha vor dem Berge**, Deepak M. Mathew††, Skipis Dimitrios‡‡, Arnidis Iordanis‡‡, Orestis Vantzos‡‡, David Georgantas‡‡, Gautier Rouaze[x], Christoph Bühler[xi], Guido Salvaneschi[xi], Brandon Lewis[xii], Angela Hauber[xii]
*Bielefeld University, Germany †Hiro-MicroDataCenters, Netherlands ‡PCB Design, Hungary §paraXent GmbH, Germany ¶Ryax Technologies, France ‖CAD-Terv, Hungary **christmann informationstechnik + medien GmbH & Co. KG, Germany ††Fraunhofer Institute for Industrial Mathematics ITWM, Germany ‡‡Independent Power Transmission Operator S.A., Greece [x]ETA-Lab, EPFL, Switzerland [xi]University of St. Gallen, Switzerland [xii]PICMG, USA

***Abstract*—CAPE is a European-funded project targeting to reshape edge-cloud computing by defining edge micro data centers as a new unit of computing. Fully committing to open source, CAPE develops a fully Composable Infrastructure (CI) for high-performance edge server hardware platforms grounded in open, forward-looking standards. Together with an open-source software stack covering the Edge-Cloud Continuum, this holistic approach boosts power and energy efficiency while reducing resource overprovisioning. Completely based on open standards, CAPE strengthens the digital sovereignty Europe needs in a challenging future. This work gives an overview of the current architectural blueprint of the project, focusing on integrating game-changing technologies like Compute Express Link (CXL) for compute and memory disaggregation, pushing open source cluster management, and AI-assisted deployment software stacks using Infrastructure from Code (IfC). The proposed approaches and benefits for future Edge-Cloud data centers are demonstrated within three use cases, ranging from Smart Grid and Edge-AI to Satellite Data Processing.**

I. Introduction

This work gives an overview of the current state of the European multi-partner project CAPE – European Open **C**ompute **A**rchitecture for **P**owerful **E**dge, started in December 2024. The consortium consists of 12 parties from 6 different European countries. With 7 SME partners, and 4 academic partners, the project has a business-driven focus, while being committed to open-source hardware, software, and standards. The project aims to reshape Edge-Cloud computing, is well connected to related projects and is additionally supported by the international standardization body PICMG as an associated member [1].

A. State of the Art Edge-Cloud computing

The ever-increasing demand for computing power in all areas of life poses new challenges for operators of SMEs and large data centers, which are the drivers for innovations at the hardware and software levels. Over 10 years ago, server architectures with x86 CPUs, limited RAM, and possibly PCIe-based accelerators like GPGPUs and FPGAs were monolithic. After virtualization became the norm, data centers evolved with techniques for software-defined infrastructure towards hyperconverged infrastructures (HCI), decentralizing compute, network, and storage resources, reducing the overall complexity, and increasing flexibility, performance, resilience, however, also affecting costs [2–4].

The next step is disaggregated HCI (dHCI), an extension of HCI from local cluster systems towards processing on distributed resources, further reducing overprovisioning [5]. Although there are now established techniques like Infrastructure as Code (IaC), the programming of workloads targeting both edge and cloud is still very complex, often resulting in the underestimated challenges of resource utilization and overprovisioning, leading to "cloud waste" [6].

With its strict demands for data security and privacy in the EU, on-premise edge computing is increasingly essential. Ideally, applications should run cost-effectively on highly trusted, local edge servers, while scalable cloud solutions can handle peak loads. This required dynamic presents data center operators and application developers with major challenges if they have to support the entire spectrum of the Edge-Cloud Continuum [7, 8]. In particular, small and medium-sized enterprises, the backbone of the European economy, lack the know-how to maintain the necessary hardware platforms and provide the software infrastructure.

B. CAPE's vision and goals

Figure 1 shows how this project aligns with the European strategy and which challenges are addressed. Several research projects are related to this project, some of which have been completed, and some are still ongoing. Instead of focusing on one particular item, from hardware platforms to software stack, CAPE covers most of these aspects.

This publication incorporates results from the CAPE project, which received funding from the European Union's Horizon research and innovation programme under grant agreement No. 101189899 and was co-funded by the Swiss National Science Foundation (SNSF, No. 200429), and by Armasuisse Science and Technology.

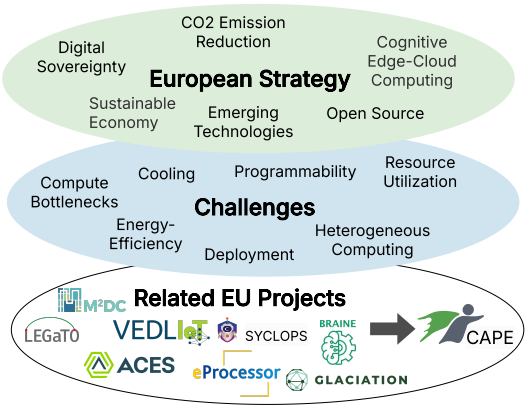


Fig. 1. CAPE's alignment with European strategy, addressed challenges and related European projects

To overcome the increasing challenges such as to increase compute power and resource utilization, the idea of dHCI has been further developed into a Composable Infrastructure (CI) [9]. In CI, all hardware resources, including all kinds of computing resources, storage, memory, and the communication infrastructure can be composed, creating a dynamic pool, to meet the needs of each individual workload. The server architecture we are developing is in line with the idea of CI, and is based on a modular approach, allowing data center operators to tailor or "compose" a set of heterogeneous compute resources, like FPGAs, GPGPU or RISC-V-based CPUs with integrated accelerators. All compute nodes are interconnected by a switched PCIe fabric with CXL, which can be reconfigured at run time, allowing dynamic allocation of resources within a single server chassis or complete cluster.

Figure 2 shows the "Big Picture" of CAPE driving multiple rising technologies to ease the flexible deployment of applications in cloud-to-edge environments. The project strongly focuses on developing novel hardware platforms for modular, heterogeneous computing and future-proof open standards for high-performance edge computing. The hardware components are highly integrated into the building infrastructure, using waste heat recovery and passive cooling to minimize the carbon footprint.

To ensure the highest level of trust, the architectural blueprint of CAPE, as many components as possible, like boot firmware and management, are developed as open source. A comprehensive software stack based on open source frameworks is developed to counter the challenges regarding programmability and deployment of applications on different hardware platforms in the Edge-Cloud Continuum. The necessary elasticity in development is achieved through AI-driven automated deployment strategies using both established technologies such as IaC and innovative approaches such as IfC. The approaches and results are evaluated using three challenging use cases demanding efficiency, acceleration, resilience, and data security.

II. Edge Hardware Platforms

This section deals with the hardware platform developed within CAPE. After giving an overview of existing Computer-On-Module form factors and state-of-the-art modular server architectures, the key enabler technology for Composable Infrastructures, Compute Express Link (CXL), and its advantages are explained in detail. Then, the server architectures developed in CAPE are described, focusing on their communication infrastructure, followed by a section dealing with the hierarchical board management system and firmware. The section is closed with aspects of environment integration and waste heat recovery.

A. Modular, heterogeneous hardware platforms

Currently, most commercially available edge servers consist either of standard 19" rack inserts based on standardized E-ATX mainboards or customized housings that have been hardened for special use, e.g., as telco base stations towards robustness and passive cooling. These hardware platforms are specifically assembled for a single purpose and often lack extensibility. For HPC and cloud computing, the Open Compute Project (OCP) has established many standards for hyperscalers in order to reduce material costs with standardized components and increase interoperability across manufacturers [10].

Multiple research projects demonstrated the high benefits of a modular microserver approach, especially when used with heterogeneous computing devices and extended by hardware accelerators [11–15]. In these projects, different baseboard carriers were developed with support for three up to 27 Compute-On-Module (COM) high-performance microservers in a 3 HU 19" rack insert. The microservers are interconnected by an Ethernet as well as a switched PCIe Infrastructure based on PCIe Gen 3, allowing for high-speed low-latency communication for Host-to-Host communication between arbitrary nodes [12].

In CAPE, we will use the experience gained from previous projects to upgrade these platforms to the next level. We are focusing on open interfaces right from the start in order to establish an open ecosystem and open the architectural blueprints for other manufacturers.

B. Composable Infrastructure with CXL

CXL technology is seen as a game changer when it comes to interconnecting computing nodes with each other. CXL is a standardized protocol based on the PCIe physical layer with cache-coherent protocols, enabling multiple processors, accelerators (GPUs, FPGAs), and memory extension modules to share a unified memory space [16]. In its current version, the efforts of the competing standards CCIX, OpenCAPI, and GenZ, have been combined, resulting in a unified protocol supported by all major manufacturers and software providers for cloud, HPC, and HPDA.

In terms of Composable Infrastructure, CXL serves as an enabler technology, offering several advantages compared to the switched PCIe-based hardware platforms described in the previous section. By interconnecting multiple CXL fabrics, it

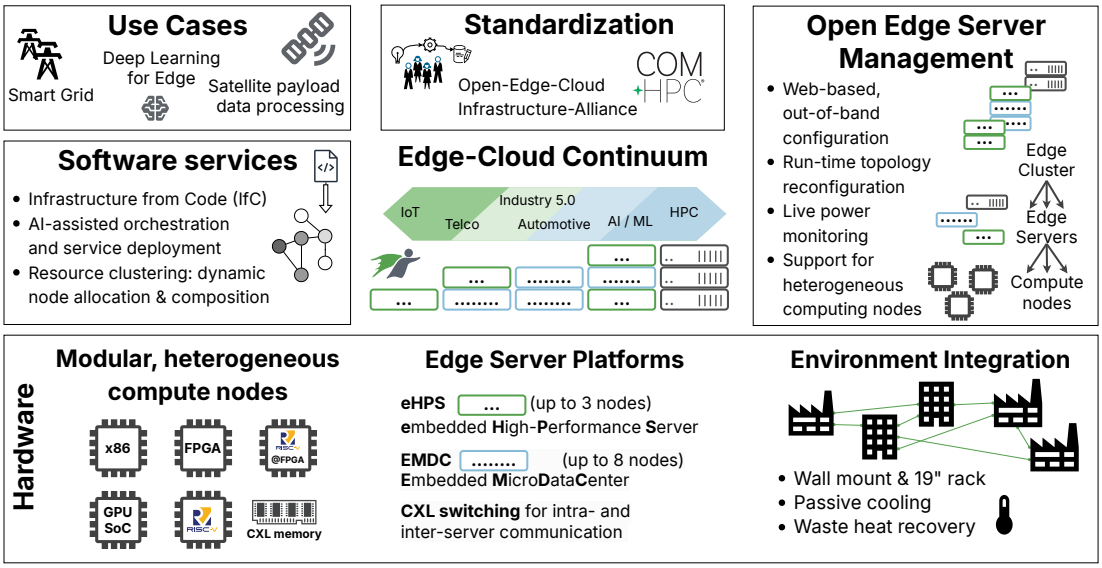


Fig. 2. "Big picture" of CAPE using Composable Infrastructure within the Edge-Cloud Continuum

is possible to scale out across one server's boundaries and create resource pools spanning multiple racks of hardware components.

The first benchmarks related to the connection of CXL memory indicated a memory access latency of 200 ns - 400 ns, which is higher than NUMA accesses between two CPU sockets [17], but lower than typical Infiniband accesses of around 2 µs - 6 µs. For scale-out, multiple hops between CXL switches need to be considered, positioning CXL between the cost-intensive Infiniband and commodity high-latency Ethernet. In contrast to PCIe, where memory sharing typically necessitates complex separately managed coherency and RDMA transfers, with CXL it can be achieved directly and efficiently by configuring the driver and accessing the memory directly from the hosts memory space.

At the time of writing, HPC data centers in particular are the early adopters of this technology, which is due to the fact that the manufacturers of CXL-capable switches have focused on rather expensive data center switches with a large number of ports. Nevertheless, CAPE will adapt the CXL to the edge computing sector at an early stage, as particularly large price reductions can be expected after a successful market launch in the HPC sector.

C. CAPEs Edge Server System Architectures

Two edge server baseboards will be developed, forming the base of CAPEs modular hardware platforms: An embedded High-Performance Server (eHPS) and Embedded Micro Data Center (EMDC) (see Figure 3). Both platforms share a modular approach based on the COM-HPC form factor, supporting a mix of COM-HPC Server and client modules. The communication infrastructure is based on PCIe switches with CXL support. It can be extended by PCIe expansion cards with additional hardware accelerators, with Network Interface Cards (NIC), or with additional CXL memory. In parallel, these systems can be tightly coupled via PCIe-based fabric interconnects, allowing for horizontal CXL scale-out.

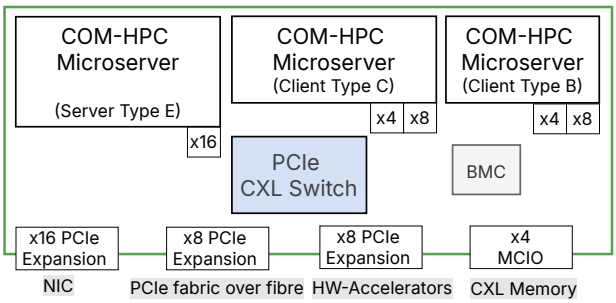
The eHPS is based on an E-ATX form factor and can be integrated into a commodity chassis with 2 or 3 height units of a 19" rack. With three mixed COM HPC modules, high-performance and power-efficient embedded modules can be integrated into the system, whereby the targeted power envelope remains below 800 W. The basic architecture concept has already been validated in previous projects. Various changes will be made in CAPE, in particular, the PCIe infrastructure will be equipped with the latest available technology. This new hardware revision raises the TRL from the current level 5 to a market-relevant level.

Although the architecture of both systems is quite similar, the EMDC is positioned at a different group of edge computing markets, like telecom operators, as it focuses on high-density computing, passive cooling, and symbiotic integration in buildings (see subsection II-F). One EMDC chassis comes with up to eight COM-HPC modules, allowing for combining multiple server-grade Processors with 8 DIMMs of RAM each, reaching more than 1 TB of CXL-pooled memory. The overall power dissipation of one chassis will be slightly less than 3 kW, but three chassis can be combined within a 4-height unit insert for a 19" rack.

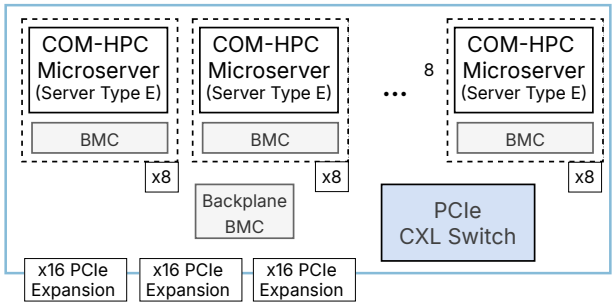
D. Heterogeneous Computer-On-Module microservers

The open COM-HPC standard, primarily targeting the embedded market has been proven to be best suited for modular edge servers [14]. COM-HPC products have already been established on the market for over 3 years. Therefore, various of x86-based or ARM-based modules are available from different manufacturers [1].

COM-HPC comes in three variants, but the *Client* and the *Server* are the most suitable for demanding edge computing requirements [18] as they support up to 49 or 65 PCIe lanes.



(a) embedded High-Performance Server (eHPS) with three mixed COM-HPC (Client and Server) microservers in two or three 19" HU



(b) Embedded Micro Data Center (EMDC) with eight COM-HPC (Server Type E) microservers. Three EMDC baseboards fit in four 19" HU

Fig. 3. Basic architecture of both CAPEs hardware platforms and their communication infrastructure based on switched PCIe with CXL. Scale-out and further interconnect is realized by PCIe expansion slots via PCIe fabric over fibre or additional NICs

With a power envelope of 150 W to 300 W, they are suitable for air cooling as well as complete passive cooling. The Client and Server variants range from $120\,\mathrm{mm} \times 120\,\mathrm{mm}$ (Type B) to $200\,\mathrm{mm} \times 160\,\mathrm{mm}$ (Type E), which is significantly smaller than standard E-ATX server mainboards.

During the development of the standard, emphasis was also placed on supporting non-x86 architectures. In CAPE further non-x86 modules will be developed, for example an FPGA module based on the AMD Versal technology. To extend the range of heterogeneous accelerators, a GPU-SoC module is also being developed providing architectures with a high energy-efficiency, especially for inference workloads. A COM-HPC module carrying multiple CXL memories is planned to further evaluate the memory pooling technology.

E. Board Management and Firmware

In standard servers, the typical out-of-band management is realized as a 1:1 relationship between a dedicated Board Management Controller (BMC) and the CPU. However, CAPE edge servers containing multiple microservers in one chassis need simultaneous management and therefore customized off-the-shelf solutions. One aspect of our work will be to create an open, decentralized, and resilient management system spanning multiple microservers per chassis as described in [19].

The eHPS utilizes a fully custom management system capable of managing multiple nodes with one BMC in a 1:N relation (see Figures 3 and 5). The BMC is based on a mid-range ARM processor and is running a custom Linux, which provides the necessary interfaces to peripherals, their firmware configuration and a web interface for the user. Besides controlling and monitoring all the nodes in the system, it also provides iKVM access as well as serial consoles to each node. The current implementation based on our prototype is already capable of configuring the PCIe switch for Host-to-Host communication and will be extended for CXL.

Unlike the eHPS, the EMDC relies on the openBMC project as management software [20]. However, openBMC requires a 1:1 relationship between BMC and nodes, which is why adapter solutions with dedicated BMCs for each COM-HPC modules are used. The central carrier board management is implemented via a dedicated BMC baseboard, which configures both the chassis and the CXL switch (see Figure 3(b)).

F. Mechanical Integration and building-level cooling

Unlike the conventional air-cooled eHPC, the EMDC focuses on high integration density and symbiotic building integrations. At the microserver level, the planned EMDC cooling solution uses a fully passive, two-stage approach designed for modularity and scalability. Each module houses 8 compute nodes, and configurations with 1, 2, or 3 rows – corresponding to 8, 16, or 24 nodes – can be deployed. Heat is passively transferred from key components within each node by phase-change technologies, such as vapor chambers, loop heat pipes, or thermosyphons [21, 22]. This initial loop operates without active elements like fans or pumps, carrying heat to the module edge. From there, a secondary interface links to a dedicated cooling module, which in turn connects to the building's chilled water or HVAC infrastructure. By separating the compute and cooling functions, the design simplifies maintenance, reduces energy consumption, and eliminates many airflow or noise issues typically found in dense systems.

Development efforts focus on choosing effective passive cooling methods, evaluating thermal performance across varied loads, and ensuring mechanical compatibility. Particular attention is given to the efficiency of phase-change heat transfer, the reliability of modular thermal interfaces, and ease of assembly. These elements form a robust foundation for node-level passive cooling and align with broader data center efficiency targets.

At the building scale, current research addresses cooling strategies that dissipate extracted heat with minimal reliance on active HVAC systems. Although two-phase cooling devices like loop-heat pipes or vapor chambers handle localized chip-level fluxes efficiently, any removed heat must ultimately be released to the outside environment. Traditional approaches for edge devices frequently exhaust heat indoors, increasing demands on HVAC or chiller units, an energy penalty not always captured by standard efficiency metrics (e.g., PUE) [22]. To mitigate this issue, this work follows the ideas of Lu et al., presented in [23] and explores passive

cooling panels that combine radiative and evaporative cooling principles, targeting a heat rejection capacity of approximately $600\,\mathrm{W\,m^{-2}}$ near ambient temperature. These panels may be directly connected to the EMDC water cooling loop or, where higher loads occur, to standard HVAC condensers, increasing their efficiency. While rooftop installations are a primary focus, investigations also consider placing panels on window-free facades, an option that can further reduce solar gains across the building envelope. Ongoing efforts include reducing evaporative water consumption through moisture-sorption in hygroscopic materials and assessing long-term coating performance under real outdoor conditions, with particular attention to UV-induced degradation. By limiting the need for chiller-based solutions, these passive approaches aim to enhance overall energy efficiency in edge data centers.

III. Software Stack

CAPE introduces a comprehensive software stack designed to simplify the deployment and management of applications across the edge-to-cloud continuum. The challenges from the developers' point of view, CAPE's solution approach, and the key technologies IfC and openMPMC are addressed below.

A. Edge-Cloud Continuum: Opportunities and challenges

As modern computing infrastructures grow increasingly heterogeneous and distributed, the Edge-Cloud Continuum paradigm has emerged as a foundational principle in the CAPE project. In general, edge servers are inherently limited in their resources and their capacity to handle compute-intensive tasks or accommodate workload spikes. To mitigate this, the cloud serves as an elastic extension of the edge, capable of dynamically offloading and thereby absorbing peak demand and facilitating unsuitable work for the edge, like large-scale data analytics or backup operations.

This hybrid Edge-Cloud approach is particularly relevant for real-time use cases such as in Smart Grids, where single energy grid substations require continuous surveillance (see subsection IV-B). To monitor and control the entire power grid as the superset of hundreds of substations, all data from every substation must be aggregated and centrally processed.

However, developers face major challenges when adapting applications to the edge or cloud. Depending on the targeted edge or cloud server, different programming models and interfaces must be used, although the common goal is, from the applications' point of view, the same, like, e. g., attaching databases. Hybrid use of resources across distributed environments intensifies these challenges. In an edge server with a Composable Infrastructure, in particular, the number of map-able architectures is many times greater than in the cloud, despite its resource constraints. It is, therefore, inevitable to significantly simplify the programmability to fully leverage the resources provided by both edge and cloud.

B. CAPEs Software Architecture

The architecture of CAPE's software stack is fully open source, and structured around a multi-cluster orchestration layer that dynamically allocates and provisions resources based on the application requirements (Figure 4). The stack integrates workflow-based application design, and cloud-agnostic orchestration tools such as Kubernetes, the Ryax orchestrator, and the Hiro scheduler to enable seamless resource provisioning and optimizations. By leveraging Infrastructure as Code (IaC) platforms such as Nitric, Terraform, and Pulumi, CAPE provides a unified framework that abstracts the complexities of heterogeneous cloud and edge environments. A key innovation is the incorporation of AI-driven automation, where open-source Large Language Models (LLMs) and lightweight flavors of them (SLMs), assist in IaC code generation, and orchestration decisions, making the system accessible to both expert (deployment path (1)) and non-expert users through intuitive low-code interfaces (paths (2) and (3)).

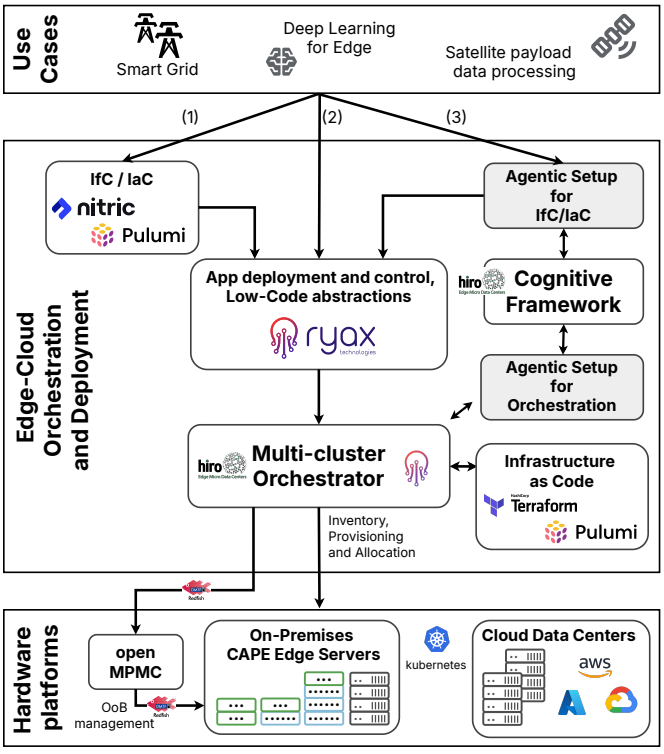


Fig. 4. Overview of CAPEs Software stack

Expert users define their applications using IaC tools, which are translated into Ryax abstractions, e. g. workflows, actions, and services, through specialized plugins. The Ryax workflow orchestrator in collaboration with the Hiro distributed scheduler, then deploys these services across available Edge-Cloud or on-premises infrastructure, automatically provisioning the needed resources via integrations with systems such as Harvester. For non-expert users, the web-based low-code interface of Ryax platform simplifies the application design while maintaining the same underlying orchestration capabilities.

The architecture employs AI agents powered by optimized open-source LLMs, assisting in IaC code generation and orchestration tuning. Additionally, the architecture incorporates automated testing and monitoring tools to maintain repro-

ducibility and performance across diverse hardware components, including RISC-V and CXL-based systems.

C. Infrastructure from Code as an enabler technology

With the need for more complex infrastructure setups, engineers have to address the complexity of managing infrastructure configurations. Tools like Terraform [24] and Pulumi [25] allow engineers to define software infrastructures either with a domain-specific or a fully fledged programming language. Terraform and Pulumi focus on infrastructure management, whereas other frameworks from industry (e. g., Ansible [26]) focus on configuration management of a single servicing instance (like a server).

While Pulumi and Terraform allow engineers to define infrastructure in a reproducible and automated way, they still have to create the digital twin of the infrastructure in the IaC program. Instead, with Infrastructure *from* Code (IfC), engineers are given an SDK which is used in their normal coding workflow. In IfC, the infrastructural code is derived from the application's source code. As an example, an engineer can import an API from the provided SDK and create a new instance of this object. During compilation (and ultimately deployment), the code is analyzed to create the corresponding object that hosts an API application on the targeted cloud provider. An example of this approach from industry is nitric [27].

Within CAPE, we allow engineers to select their preferred mechanism of defining infrastructure. A provider for nitric will be developed, enabling direct communication with the CAPE hardware platforms. This includes a provider for Pulumi to support the declarative approach of IaC that IfC still uses under the hood.

D. Unified for Server Management by openMPMC

As part of the CAPE project, the openMPMC (Open Multi Platform Management Controller) provides an open-source web-based interface for managing multiple local servers in a secure and user-friendly way. The system provides a web interface for intuitive manual operation and exposes a Redfish-compatible API for integration with external tools and automation systems.

A key feature of openMPMC is its bidirectional use of the Redfish standard: It acts as a Redfish client to interact with hardware Baseboard Management Controllers (BMCs), retrieving detailed hardware inventory, sensor data, and executing remote management actions such as power on, off, reset, setting boot order, or mounting ISO images [28]. Simultaneously, it serves as a Redfish endpoint itself, allowing third-party systems to query and control resources managed by openMPMC through a standardized API.

The focus of openMPMC is the integration of the CAPEs platforms eHPS and EMDC featuring multiple compute modules. Additionally, the system is also intended to support existing third-party hardware, enabling broader applicability beyond the project-specific designs. This ensures a uniformly managed heterogeneous infrastructure within the same interface.

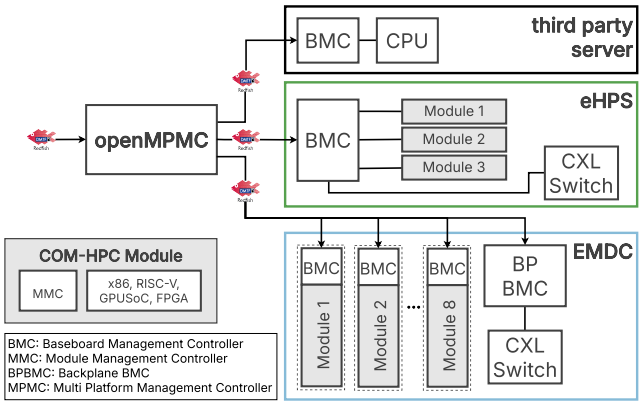


Fig. 5. openMPMC System architecture supporting eHPS and EMDC servers

openMPMC also includes a comprehensive resource composition management subsystem allowing users to configure the assignment of PCIe- or CXL-attached devices to microservers or to configure high-speed inter-node communication, either via the graphical interface or the provided Redfish API. For remote systems with special access conditions, an external agent runs locally in those networks and connects to a central openMPMC instance, providing secure remote communication and control.

IV. Use Cases

CAPE integrates three use case partners who have collaborated to define the requirements for both hardware and software components. Their applications provide a foundational benchmark for evaluating the progress and effectiveness of CAPE's efforts.

A. Deep Learning for Edge

Unlike cloud data centers, edge data centers are typically constrained by tight power budgets and limited memory, making it challenging to deploy state-of-the-art Deep Learning models without significant optimization or architectural support. Transformer-based architectures, currently dominating modern NLP and vision tasks, often use the pre-trained language model BERT-base, which contains 110 million parameters, consumes $420\,\mathrm{MB}$ of memory just for weights, and requires $12.5\,\mathrm{GFLOPS}$ per inference [29].

Without specialized acceleration and memory management, such models require heavy quantization and pruning, often leading to degraded accuracy or can not be executed at all within the edge constraints. Moreover, deep learning inference involves high-volume data movement between memory and compute units, which further stressing the limited bandwidth and power envelope available on edge platforms.

To address these challenges, RISC-V and CXL offer promising architectural solutions. RISC-V utilizes an open, modular architecture based on minimal instruction sets (e. g.,

RV32I/64I), with optional standard and custom extensions (e.g., vector, MX) tailored for energy-efficient AI computations. This flexibility enables designers to include only the necessary features, thereby reducing area and power consumption. As a key technology of CAPE, CXL supports coherent memory sharing between CPUs, accelerators, and memory expanders, significantly reducing data duplication and memory overhead, which is crucial when deploying large AI models in constrained environments. This use case will demonstrate the benefits of RISC-V and CXL on COM-HPC FPGA modules in accelerating deep learning kernels on edge.

B. Smart Grid

There are three major trends currently affecting energy grids and energy ecosystems in general: a) the decarbonization of the energy production, introducing renewable, but often fluctuating energy sources in the production, b) the subsequent transition towards decentralized energy production, increasing the grid's complexity, and c) the digitalization of the grid assets. These trends are causing an exponential increase in the data produced in energy grids.

For CAPE's use case partner IPTO (Independent Power Transmission Operator) the implementation of edge computing within their infrastructure is not only inevitable but also offers significant advantages. IPTO's energy grid covers the whole of Greece, and is comprised of 423 substations, 908 transformers of various types, and $13.690\,\mathrm{km}$ of high voltages cables, which collectively produce hundreds of thousands of Supervisory Control and Data Acquisition (SCADA) signals.

Processing data directly at the source allows for continuous anomaly detection e. g., voltage drops, frequency deviations, equipment malfunctions, or unexpected power surges, leading to predictive maintenance and shorter response times for repairs. This close monitoring is particularly beneficial for integrating renewable energy sources such as wind and solar, which are highly stochastic with no inherent inertia, and therefore require real-time adjustments to maintain grid stability.

By enabling a decentralized or localized control, the independence of individual grid segments (e. g. islands) can be enhanced, ensuring seamless functionality without excessive reliance on a centralized infrastructure or external networks. Such areas may comprise multiple substations producing tens of thousands of signals, corresponding to a sustained rate of $1\,\mathrm{MB/s}$ to $10\,\mathrm{MB/s}$. In order to process these increasing volumes of data, IPTO relies on several algorithms, including classic machine learning algorithms, such as One-Class SVM [30] or Isolation Forest [31], as well as modern pre-trained deep models for time-series analysis like MOMENT [32].

Not only is the choice of a suitable ruggedized hardware platform a major challenge, but also the programming of the infrastructure. Here, CAPE's strengths are utilized regarding heterogeneous hardware platforms and the novel IfC approaches are extensively evaluated.

C. Satellite Data Processing

In this use case, we focus on multiple applications for ground-based satellite data processing to achieve pseudo real-time performance. Unlike satellite onboard processing, which is heavily constrained by power, weight, and radiation tolerance, ground-based infrastructure allows the integration of the latest technologies without such limitations. Traditional workflows, often involving delays of hours or days, are no longer sufficient for modern applications that demand actionable insights in near real-time.

One major application is super-resolution imaging, where deep learning models enhance the clarity and detail of satellite imagery. These models, often based on convolutional neural networks or generative adversarial networks, require significant computational power, making them well-suited to GPU-accelerated servers with large shared memory. Another critical area is Synthetic Aperture Radar (SAR) processing, which involves complex signal processing tasks such as matched filtering and back projection. Processing SAR data in near real-time benefits from parallel workloads and memory-efficient architectures, enabled by technologies like CXL.

Ground-based servers are also instrumental in change detection and anomaly recognition, particularly for applications such as monitoring infrastructure, forests, or disaster zones. These use cases require comparing large volumes of temporal data, which is accelerated by modern AI inference engines and high-speed interconnects. Furthermore, the fusion of multimodal data – optical, thermal, SAR, and hyperspectral – into a unified analysis stream demands bandwidth and computational versatility. Environmental modeling and real-time geospatial analytics, including land use classification and object detection, similarly benefit from system architectures supporting large in-memory databases.

In particular, CAPE's adoption of CXL opens several new possibilities for this use case, by low-latency access to pooled memory and inherent support of hardware accelerators. CAPE's software stack will be used to dynamically distribute applications between the edge and the cloud, easing programmability and simplifying the evaluation of different scenarios.

V. Summary and Outlook

This work gives an early overview of the CAPE project, focussing on the architectural aspects of the planned hardware and software developments. CXL and IfC have been identified as the key technologies for CAPE and have the potential to advance the concept of Composable Infrastructure, which will influence and reshape the complete market from HPC to SME.

The next steps in CAPE are to finalize a detailed roadmap for planning and developing the hardware platforms. In parallel, the software stack is developed, and the use cases are prepared for evaluation. We are convinced that our approach can only be implemented with open standardization. To consolidate these efforts, CAPE considers the foundation of an Open Edge Cloud Infrastructure Alliance (OECIA), strengthening and directing these efforts.

References

[1] PICMG, "Open Standards for Embedded Computing Applications," 2025. [Online]. Available: https://www.picmg.org

[2] A. C. Azagury, R. Haas, D. Hildebrand, S. W. Hunter, T. Neville, S. Oehme, and A. Shaikh, "GPFS-based implementation of a hyperconverged system for software defined infrastructure," *IBM journal of research and development*, vol. 58, no. 2/3, pp. 6–1, 2014.

[3] A. J. Shetty and K. Ganashree, "Comprehensive review of datacenter architecture evolution," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, pp. 5–6, 2020.

[4] G. Silva-Atencio and M. Umaña-Ramírez, "The Evolution and Trends of Hyperconvergence in the Telecommunications Sector: a Competitive Intelligence Review," *Dyna*, vol. 90, no. 227, pp. 126–132, 2023.

[5] C. J. Wang and B. Kim, "Automotive Big Data Pipeline: Disaggregated Hyper-Converged Infrastructure vs Hyper-Converged Infrastructure," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1784–1787.

[6] B. Everman, M. Gao, and Z. Zong, "Evaluating and Reducing Cloud Waste and Cost —- A Data-driven Case Study from Azure Workloads," *Sustainable Computing: Informatics and Systems*, vol. 35, p. 100708, 2022.

[7] R. Rossini and L. Lopez, "Towards an European Open Continuum Reference Stack and Architecture," in *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2024, pp. 1–5.

[8] "Open Continuum / EUCloudEdgeIoT Task Force 3 - Functional View of the Continuum Reference Architecture: Minimum set of functionalities," Jun. 2024.

[9] L. M. Herger, K. El Maghraoui, I.-H. Chung, C. Choudary, K. Tran, and T. Deshane, "Toward an Enterprise-ready Composable Infrastructure as a Service," in *2021 IEEE International Conference on Services Computing (SCC)*. IEEE, 2021, pp. 116–125.

[10] OCP, "Open Compute Project," 2025. [Online]. Available: https://www.opencompute.org

[11] R. Griessl, M. Peykanu, J. Hagemeyer *et al.*, "A Scalable Server Architecture for Next-generation Heterogeneous Compute Clusters," in *2014 12th IEEE International Conference on Embedded and Ubiquitous Computing*. IEEE, 2014, pp. 146–153.

[12] A. Oleksiak, M. Kierzynka, W. Piatek *et al.*, "M2DC – Modular Microserver DataCentre with heterogeneous hardware," *Microprocessors and Microsystems*, vol. 52, pp. 117–130, 2017.

[13] B. Salami, K. Parasyris, C. Adrián *et al.*, "LEGaTO: Low-Energy, Secure, and Resilient Toolset for Heterogeneous Computing," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020.

[14] M. Kaiser, R. Griessl, L. Tigges, J. Hagemeyer *et al.*, "VEDLIoT: Very Efficient Deep Learning in IoT," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022, pp. 963–968.

[15] P. Szántó, T. Kiss, and K. J. Sipos, "Energy-efficient AI at the Edge," in *2022 11th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 2022, pp. 1–6.

[16] "Compute Express Link (CXL) Specification, Revision 3.2," October 2024.

[17] A. Lerner and G. Alonso, "CXL and the Return of Scale-Up Database Engines," *Proc. VLDB Endow.*, vol. 17, no. 10, p. 2568–2575, Jun. 2024. [Online]. Available: https://doi.org/10.14778/3675034.3675047

[18] PICMG, "COM-HPC Module Base Specification rev. 1.2," 2023. [Online]. Available: https://www.picmg.org

[19] B. Scherer, J. Lazányi, B. Kardos, and Á. Radics, "Distributed Supervision of an Edge Micro Datacenter," in *24th International Carpathian Control Conference*, 2023.

[20] Linux Foundation Project, "Openbmc," 2024. [Online]. Available: www.openbmc.org

[21] *Thermal-Hydraulic Characterization of Thermosyphon Cooling System for Highly Compact Edge MicroData-Center. Part I: Design and Experiments*, ser. International Electronic Packaging Technical Conference and Exhibition, vol. ASME, 2023.

[22] E. Minazzo, J. R. Thome, J. B. Marcinichen *et al.*, "A New Edge Micro Data Center and its Passive Thermosyphon Cooling System at tue: Cooling System Thermal Performance Tests," in *40th Semiconductor Thermal Measurement, Modeling & Management*, 2024.

[23] Z. Lu, A. Leroy, L. Zhang, J. J. Patil, E. N. Wang, and J. C. Grossman, "Significantly enhanced sub-ambient passive cooling enabled by evaporation, radiation, and insulation," *Cell Reports Physical Science*, vol. 3, 2022.

[24] HashiCorp, "Terraform," Aug. 2024. [Online]. Available: www.terraform.io/

[25] "Pulumi - Infrastructure as Code, Secrets Management, and AI," Pulumi. [Online]. Available: www.pulumi.com

[26] Red Hat, "Ansible," Aug. 2024. [Online]. Available: www.ansible.com

[27] Nitric Inc., "Nitric," 2025. [Online]. Available: www.nitric.io

[28] Distributed Management Task Force, "Redfish Scalable Platforms Management API (Redfish)," 2023. [Online]. Available: www.dmtf.org.com

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[30] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support Vector Method for Novelty Detection," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[31] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 8th IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[32] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A Family of Open Time-series Foundation Models," 2024.