DOI: 10.26314/GARR-Conf25-proceedings-15

Real-World Federation of Autonomous Kubernetes in an Interconnected Continuum

Giuseppe Zangari[1], Fulvio Risso[2]

[1]ArubaKube, [2]Politecnico di Torino

Abstract. High-performance computing (HPC) and GPU clusters often suffer from inefficiencies of underutilized resources. Studies have shown that many HPC nodes and accelerators run well below full capacity, with CPUs and memory frequently only half-used and GPU memory largely untapped. Such underutilization translates into sunk costs and idle investments, even as other organizations struggle with insufficient compute capacity. Peaks in demand can overwhelm local clusters—researchers and engineers face queue backlogs and delays when their on-premises resources are saturated. This combination of underused hardware in one place and unmet needs in another highlights a critical inefficiency in the status quo. This paper explores how a federated Kubernetes-based approach can turn these inefficiencies into opportunities. By leveraging Kubernetes and Liqo, independent clusters can securely and transparently share compute resources while maintaining full autonomy over their infrastructure. The solution enables organizations to "burst" workloads to remote clusters on demand, resolving capacity shortfalls without costly hardware over-provisioning. At the same time, it allows those remote clusters to share or utilize their idle cycles, improving overall utilization. This federated model preserves cluster sovereignty: each participant retains control through policies and isolation, ensuring that sharing does not compromise security or autonomy. In essence, the AGER initiative demonstrates a real-world "computing continuum" that mitigates waste and scarcity by interconnecting cloud and HPC resources across institutional boundaries. This federated continuum unlocks innovation and operational value. Liqo and Kubernetes provide the cloud-native, secure foundation for this continuum, enabling seamless resource sharing "without borders" and establishing a new paradigm of collaborative computing at scale

Keywords. Computing-Continuum, Liqo, HPC, Efficiency, Offloading

Introduction

Despite widespread cloud and edge computing adoption, the global computing landscape remains fragmented. Organizations operate isolated clusters—on-premises HPC systems, private clouds, or edge nodes—that run independently. This isolation causes resource fragmentation: surplus capacity in one cluster cannot meet demand in another, leading to underutilization and unmet needs. Studies on NERSC's Perlmutter supercomputer show that most jobs used only a fraction of allocated resources, with ~50% of GPU jobs consuming just a quarter of GPU memory (Li et al. 2023). Meanwhile, organizations lacking HPC/GPU capacity face slowdowns and job queues, delaying critical R&D work. This imbalance highlights structural inefficiencies in modern research and enterprise computing.

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

A federated cloud-native infrastructure addresses this by connecting isolated clusters into a computing continuum (Iorio et al. 2023), conceptualize this as "liquid computing," where applications dynamically find execution venues across federated resources. This approach improves performance and flexibility while preserving decentralization and ownership: no single party controls all resources. Each participant—university, corporate cloud, or edge site—remains autonomous, sharing resources under its own policies.

Complementing this is Europe's focus on data sovereignty and federated data sharing. Marino et al. (2023) propose infrastructure-level data spaces, where clusters securely exchange and process data under agreed rules. Using Kubernetes-based federation (Liqo), flexible, on-demand data spaces span multiple domains, ensuring that providers retain sovereignty over infrastructure and data. Initiatives like Gaia-X further stress the importance of federation with autonomy and security.

Within this context, the AGER initiative demonstrates a real-world federated cloud-native infrastructure. AGER links independent Kubernetes clusters across multiple organizations into a resource continuum, operationalizing Iorio's vision with open-source tools. Using Liqo, each cluster can peer with others, securely advertising and consuming resources without altering internal configurations. Workloads flow to available capacity, embodying the "liquid computing" model.

AGER spans diverse environments—university HPC clusters, industrial research sites, and cloud providers—forming a nationwide Kubernetes continuum in Italy across Turin, Bologna, and Bergamo. Its mantra, "research without walls," reflects its ability to run workloads across sites seamlessly, bypassing traditional scheduling and cluster boundaries. AGER remains policy-first and cloud-agnostic, with each site defining sharing rules. This paper details AGER's method and value across academic, industrial, and enterprise contexts, showing federated Kubernetes infrastructure as a practical model for innovation and resource efficiency.

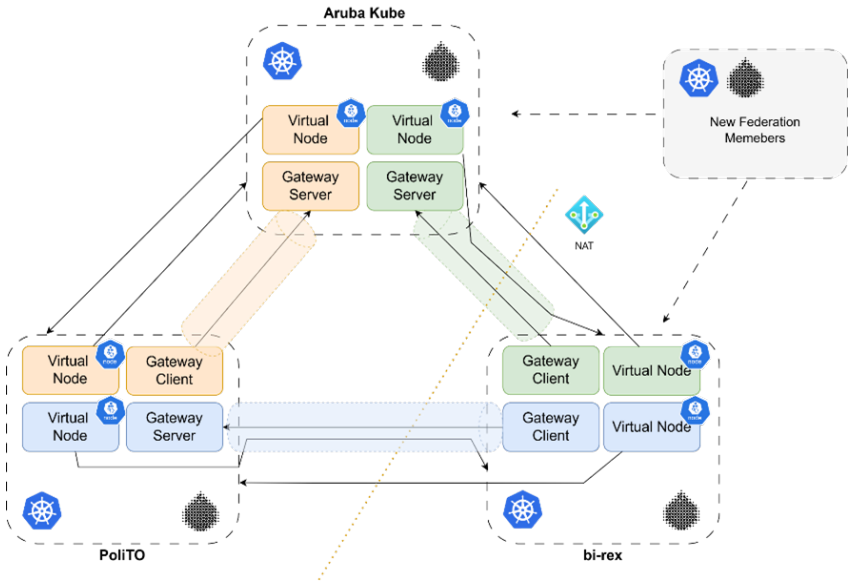
1. Federation Framework

AGER links independently managed Kubernetes clusters into a federated resource continuum using Liqo, an open-source extension designed for seamless multi-cluster Kubernetes federation. On each participating cluster, Liqo deploys a lightweight operator that handles federation tasks with minimal overhead. This operator:

- Advertises idle capacity (CPU, GPU, and memory) by creating a virtual node within the local Kubernetes API, representing the resources available from peered clusters. This abstraction allows local schedulers to see remote resources as if they were native nodes.
- Establishes encrypted network tunnels (typically over WireGuard or equivalent backends), ensuring that inter-cluster traffic remains private and secure. Importantly, Liqo retains the original service accounts, network policies, and namespace isolation, ensuring that identity and access control behave consistently even when pods are offloaded across organizational boundaries.

- Respects local quotas, priority classes, and preemption policies, so that no exported capacity jeopardizes critical home workloads. Clusters can dynamically adjust or revoke resource offers at runtime if local demand surges, offering real-time governance over shared capacity.

Fig. 1
AGER high level architecture



Scheduling remains native and fully transparent: Cluster A's scheduler operates as usual, and when it cannot place a pod locally (due to resource exhaustion or scheduling constraints), it targets the virtual node representing Cluster B. Liqo intercepts this scheduling decision and handles offloading the pod to Cluster B, ensuring that it runs in a sandboxed namespace mapped to the originating tenant. From an operator and developer perspective, the pod appears local—logs, metrics, monitoring hooks, and debugging tools (like kubectl logs and kubectl exec) function exactly as if the pod were on-premises.

Critically, federation is opt-in and namespace-scoped. This means that each organization retains strict control over what resources are shared, with whom, and under what conditions—key for addressing compliance, sovereignty, and governance mandates often imposed in both academia and enterprise. Policies can restrict federation by namespace, resource type, or workload class.

Joining the federation requires no disruptive changes (Marino et al. 2023): a single Helm chart installation of Liqo and a secure token exchange between clusters is sufficient. No "lift-and-shift," migration, or workload reconfiguration is necessary. Existing CI/CD pipelines, deployment scripts, and monitoring frameworks remain fully compatible, making AGER's approach a low-friction, production-ready solution for CTOs seeking scalable, policy-governed, and secure multi-cluster resource sharing across heterogeneous infrastructure environments.

2. AGER across sectors

AGER's federated Kubernetes continuum is not just a technical advancement—it represents a strategic enabler for innovation-driven organizations facing compute, budget, and time-to-market pressures. By breaking down infrastructure silos, AGER empowers institutions and enterprises to dynamically access, trade, and optimize distributed resources without compromising data governance or operational autonomy. This model fosters cross-institutional collaboration, accelerates research and product cycles, and transforms underutilized capacity into a business asset. The following use cases illustrate how federation drives measurable impact across academic research, industrial operations, and enterprise digital transformation.

2.1 Academic and Medical Research

Genome analytics, climate simulation, and large-language-model training surge unpredictably. With AGER, a university hospital can burst oncology pipelines to a spare resources of a national supercomputing centre during peaks, then re-claim resources when demand subsides. Turnaround time may drops from days to hours, grant-funded GPUs avoid idleness, and multi-institution collaborations proceed without data exfiltration using policies.

2.2 Industrial Optimization

Manufacturers, energy firms, and media studios face cyclical compute spikes. Instead of over-provisioning, they federate with AGER. A car maker, for exam-ple, runs crash-simulation sweeps on partner clusters overnight, returning re-sults before the morning stand-up. Capital expenditure falls, idle hardware may gains revenue as a traded asset, and production schedules are insulated from HPC bottlenecks.

2.3 Industrial Optimization

Global enterprises juggle dozens of Kubernetes deployments across clouds and edges. AGER federation converts these silos into a single elastic plane; latency-sensitive microservices drift to edge nodes while batch analytics migrate to available AGER resource. Governance domains remain intact because federa-tion respects jurisdictional boundaries encoded in policies. The net effect is lower total buffer capacity, predictable spending, and faster feature roll-outs

3. Conclusion and future work

AGER proves that federated Kubernetes can reconcile autonomy with collabora-tion. By exposing surplus capacity as a service, it elevates idle hardware from sunk cost to strategic asset, compresses time-to-insight in research, smooths in-dustrial production cycles, and sharpens enterprise competitiveness. Future re-search should explore fine-grained brokering—e.g., sub-GPU sharing—and in-tegrate market pricing to incentivise broader participation. Standardised trust frameworks (Gaia-X, IDSA) can further institutionalise policy exchange, ena-bling federations that span hundreds of clusters on a continental scale. The journey toward a durable computing continuum has begun; the next step is to

mainstream it, making compute-as-commons as ubiquitous as the internet itself.

Bibliographic References

Li J., Michelogiannakis G., Cook B., Cooray D., & Chen Y. (2023). Analyzing Re-source Utilization in an HPC System: A Case Study of NERSC's Perlmutter. Lecture Notes in Computer Science, 13948, 297-316.

Iorio M., Risso F., Palesandro A, Camiciotti L., Manzalini A. (2023) Computing without borders: The Way Thowards Liquid Computing, IEEE Transaction on Cloud Computing (vol. 11, no. 3), pp 2820-2838

Marino J., Camiciotti L., Cheinasso F., Olivero A., Risso F. (2023), Enabling Compute and Data Sovereignty with Infrastructure-Level Data Spaces, ESAAM '23: Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum (October 2023), pp 77-85

Authors

Giuseppe Zangari giuseppe.zangari@arubakube.cloud

Giuseppe Zangari (born in 1982) graduated from the Politecnico di Torino and holds an EMBA from the Graduate School of Management at Politecnico di Milano. He has held various leadership positions in global software organizations like Nokia and Pirelli, in Italian SMEs and in Politecnico di Torino, leading the development of business effective solutions with technologies ranging from IoT to cloud computing and AI. He is an expert of digital transformation, a startup mentor, and he also served as Innovation Lead. At ArubaKube, he is responsible for maximizing the software project's value, serving concurrently as Product and Business Development Lead.

Fulvio Risso fulvio.risso@polito.it

Fulvio Risso is full professor at Politecnico di Torino. Born in Saluzzo, Italy on November 15, 1971, he shares his birthday with the announcement of the Intel 4004 chip. Fulvio completed his BSc in Computer Engineering from Politecnico di Torino in July 1995 and got his PhD in Computer Engineering from the same institution in January 2000. His academic journey has been marked by significant contributions in the field of cloud computing, edge computing, network functions virtualization, and software-defined networking. He greatly contributed to open-source software, starting many successful project such as WinPcap, the de-facto packet capture library for Windows, and many others. He recently started the ArubaKube spin-off of Politecnico di Torino, where he serves as Chief Innovation Officer.