CONFERENZA GARR 2024
Navigare la complessità

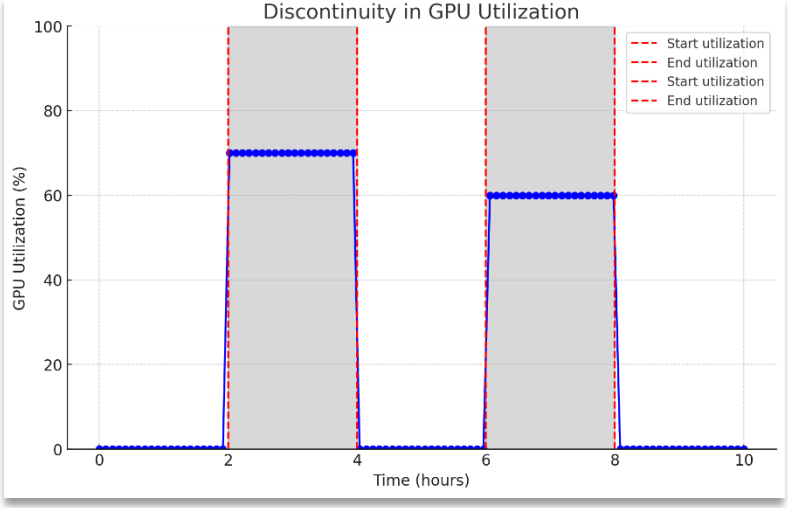
GPU Offloading in MLOps: Multicloud Ecosystem for Flexible AI/ML Deployments

Fulvio Risso Politecnico di Torino

Giuseppe Zangari ArubaKube

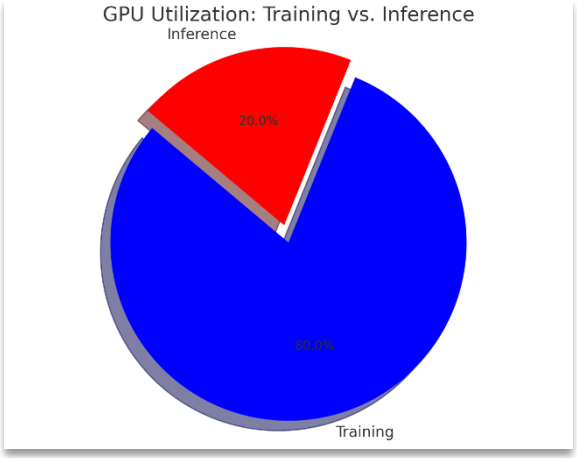
Challenges in traditional GPU adoption

Discontinuity in utilization



Hypotetic scenario

Training VS inference



Hypotetic scenario

Overall GPUs shortage

GPUs needs is rising every day, with expected **CAGR 33% 2023 – 2028** of the market, but finding them is becoming harder and harder**.**

Tough Supply condition

GPUs request a significant **upfron investment**, with a **lead time of 3-4 months**. Or, substantial pay-per-use fees.

Unclear ROI

The ROI for GPU investments remains uncertain due to market volatility, underutilization risks, rapid technological changes, and many AI use cases still not hitting the market.

https://www.knowledge-sourcing.com/report/graphics-processing-unit-gpu-market

A possible solution: effective dynamic GPU allocation

It has to be easy to implement

The dynamic topology should be easy to implement, ideally without changing application layers

It has to scale

Dynamic is the opposite of static, so scalability needs to remain

Security must come first

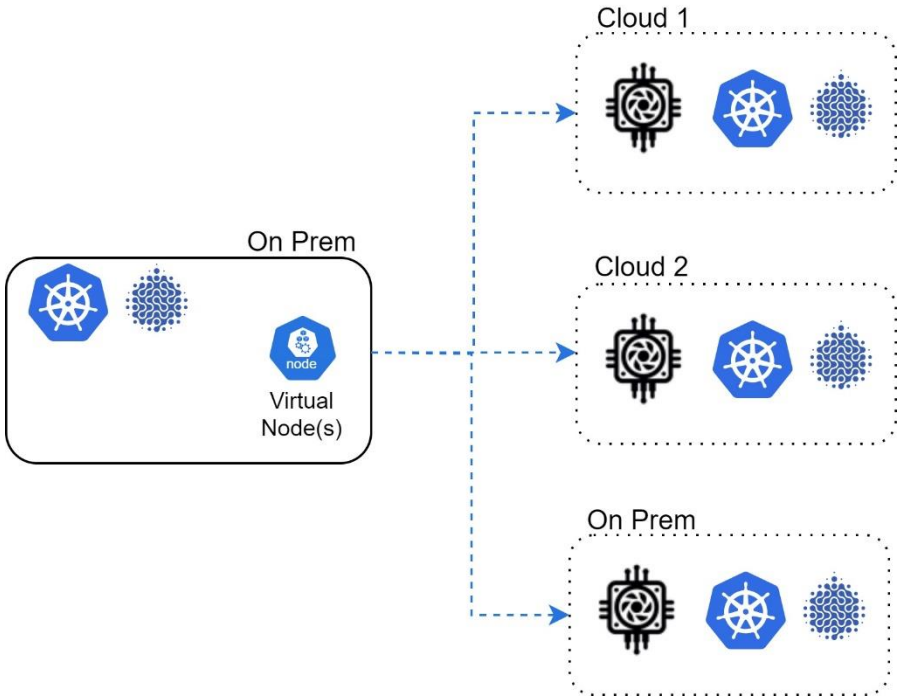
The dynamicity shouldn’t impact security

…and maintanable

Dynamism should be easy to operate and maintain, ensuring minimal downtime and reduced complexity for IT teams.

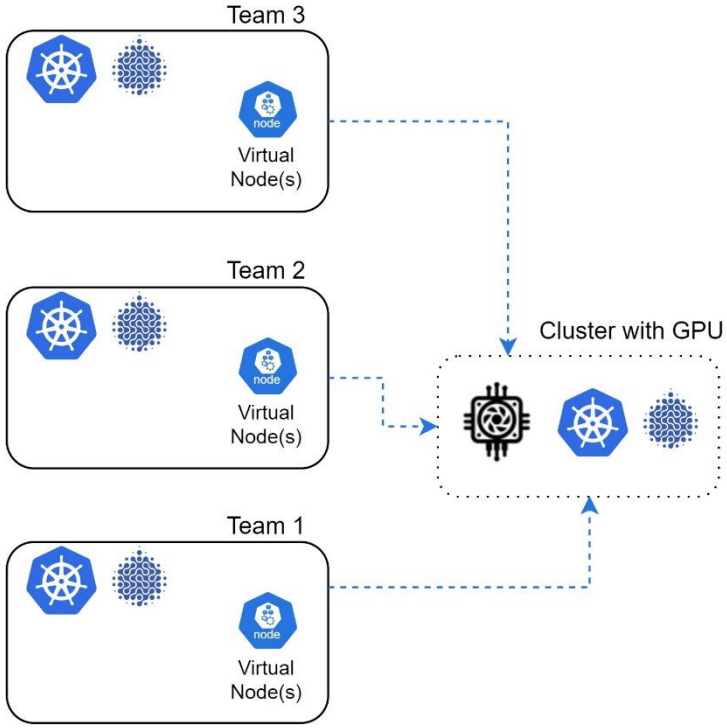
GPU offloading topologies for dynamic allocation

Multicloud GPU offloading



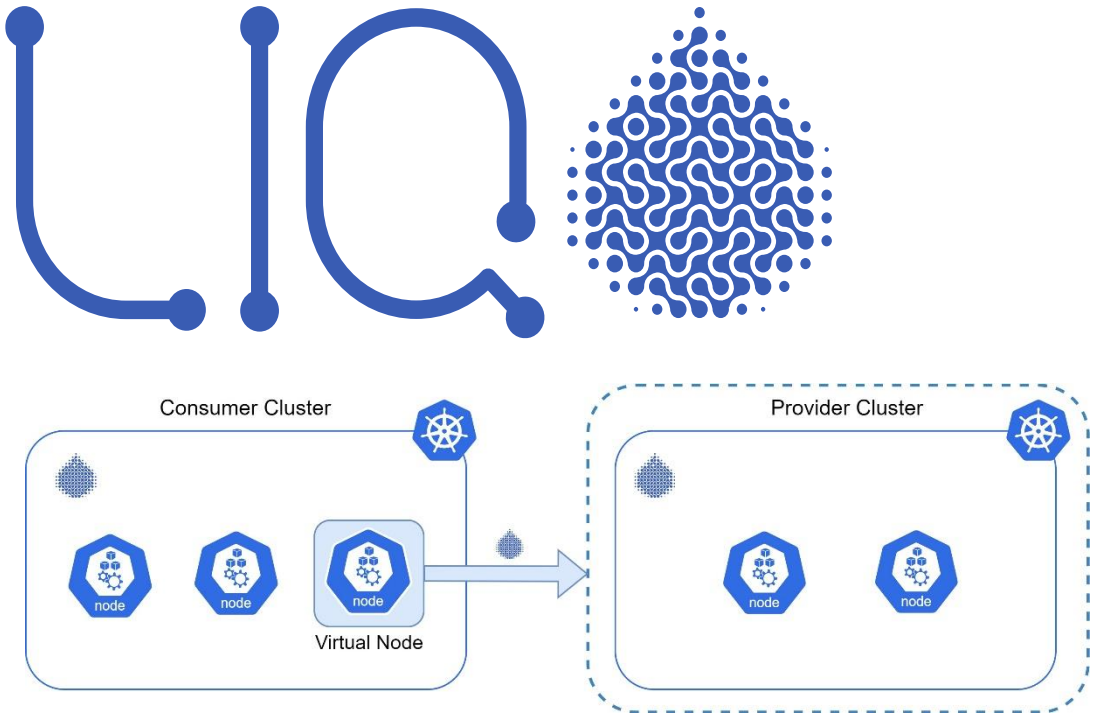
On-premises consumers can alternatively **use GPUs located on remote Kubernetes clusters**. Thanks to **Liqo**, this can be achieved without any changes to the application.

"Star" GPU offloading



With this topology, it's possible to enforce **centralized GPU policies** while maintaining team-level **security and data segregation**. Thanks to **Liqo** each team can share the same GPU resources, ensuring privacy and security of their data.

Liqo: a remote cluster seen as a local kubernetes node



Peering

Automatically establish peer-to-peer resource and service consumption relationships between independent and heterogeneous clusters.

Offloading

Offload workloads seamlessly to remote clusters, without requiring any modification to Kubernetes or the applications themselves.

Network fabric

A transparent network fabric, enabling multi-cluster pod-to-pod and pod-to-service connectivity, regardless of underlying configurations and CNI plugins.

Storage fabric

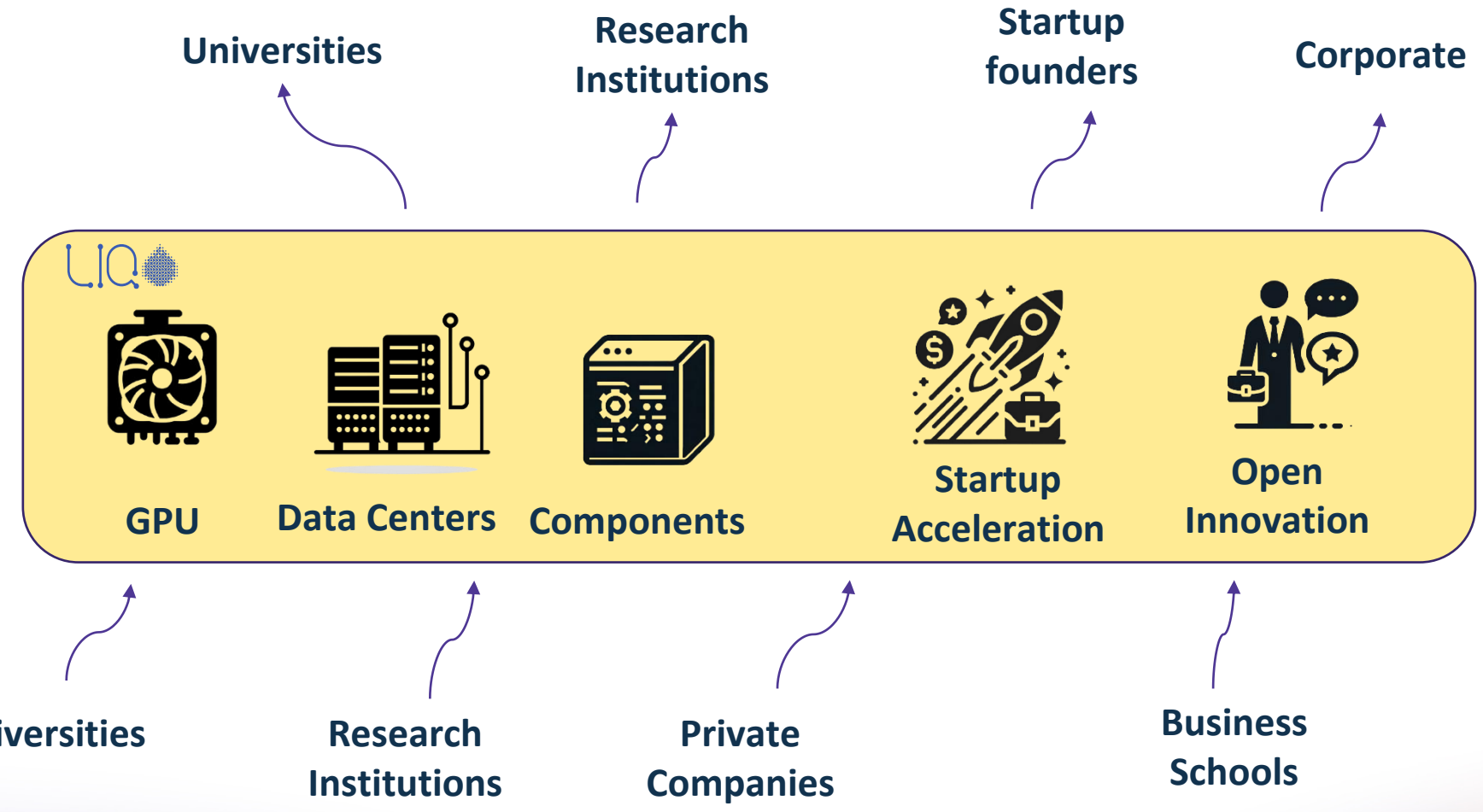
A native storage fabric, supporting the remote execution of stateful workloads according to the data gravity approach.

Active Development
With ArubaKube and PoliTo, Liqo is costantly updated

>1000 GitHub Stars
Liqo's community is always growing, 42 global contributors and 329 slack users

Open Source
Liqo is open source and always will be.

Accelerate AI innovation with an Italian AI platform



Enabled by Liqo, built on Italians components and on datacenters

Let’s innovate together

Fulvio Risso

Giuseppe Zangari

aruba.it KUBE

LIQ