POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Informatica

Tesi di Laurea

Edge-to-Cloud Multi-Cluster Orchestration for Smart Grid Monitoring Services

hello world!

Relatori
prof. Fulvio Risso
ing. Stefano Galantino
firma dei relatori

. . . . . .
. . . . . .

Candidato
Riccardo Medina

firma del candidato

. . . . . .

Anno Accademico 2023-2024

A mio padre

† A mio nonno Pino

Summary

In recent times, the introduction of edge/fog computing platforms, which operate on the principle of processing data in locations other than the central node (directly at the production node or at intermediate nodes), has enabled the development of monitoring systems and data consumption with greater speed and precision.

Within the context of the energy network, these platforms can be particularly suitable for the rapid (or automated) deployment of applications and services.

In fact, the geographically distributed compute infrastructure in combination with the transparency of virtualized (or cloud native) platforms opens up unprecedented flexibility for application deployment.

Furthermore, they facilitate the implementation of new functionalities, for example to ensure the infrastructure's operation in case of disconnection from the backbone network.

This is achieved by temporarily relocating essential software services locally and subsequently realigning the data with the remote "main" instance once the connection is restored (island mode operation).

Based on the feasibility of local solutions using the Kubernetes platform in a distributed environment, this thesis presents a potential model to extend their edge/fog computing features to the entire electrical network, employing the innovative open-source project Liqo as the main technology for the management of the distributed cluster architecture.

In developing this model, significant emphasis was placed on ensuring a high degree of resilience.

This was accomplished through the careful selection of a topology that not only supports the seamless integration of existing distributed database systems within a multi-cluster environment, without requiring any additional modifications, but it also provides the flexibility to distribute workload across any node, unrestricted by fixed architectural constraints but allowing the addition of logical constraints depending on the desired hierarchical architecture.

After examining the topology, the study progressed to creating a possible implementation in the domain of the electrical networks, analysing its behaviour in the event of faults and assessing its scalability.

Comparing the obtained results with the initial solutions, the architecture adopted in this thesis demonstrates its capability to integrate and extend the local solutions features to the entire network, without significant increases in latency despite the greater complexity. It also introduces new functionalities such as island mode operation, enabling

separate management of the two network parts during disconnection and automatic resources reconciliation once the link is restored.

Acknowledgements

I candidati ringraziano vivamente il Granduca di Toscana per i mezzi messi loro a disposizione, ed il signor Von Braun, assistente del prof. Albert Einstein, per le informazioni riservate che egli ha gentilmente fornito loro, e per le utili discussioni che hanno permesso ai candidati di evitare di riscoprire l'acqua calda.

Contents

List of Tables 9

List of Figures 10

1 Introduction 11
- 1.1 Energy section . . . 11
- 1.2 Thesis objectives . . . 12
- 1.3 Overview . . . 13

2 Kubernetes 15
- 2.1 Basic concepts . . . 15
- 2.2 K3s . . . 16

3 State of art 19
- 3.1 Smart Grid components . . . 19
  - 3.1.1 Area Control Center . . . 19
  - 3.1.2 Station . . . 19
  - 3.1.3 Phasor Measurement Units . . . 20
  - 3.1.4 Phasor Data Concentrator . . . 20
  - 3.1.5 Grid State Estimation . . . 21
- 3.2 Multi-master station architecture . . . 21
- 3.3 Actual challenge . . . 21

4 Liqo 23
- 4.1 Basic concepts . . . 23
  - 4.1.1 Network fabric . . . 23
  - 4.1.2 Peering . . . 24
  - 4.1.3 Offloading . . . 25
  - 4.1.4 Storage fabric . . . 26
- 4.2 Distributed DB interaction . . . 26

5 General topology Partial Mesh Star 29
- 5.1 Physical/net architecture . . . 29
- 5.2 Logic complex hierarchy using labels and affinity . . . 30
  - 5.2.1 Independent Groups . . . 31

5.2.2 Dependent Groups 31
5.3 Partial Mesh Star Analysis 32

6 Possible implementations **35**
6.1 Logical domains 35
6.2 Multi-level logical domains 36
6.3 Final consideration 38

7 Domain peering evaluation **39**
7.1 Test Environment 39
7.1.1 Crownlabs 39
7.1.2 Nodes configuration 39
7.1.3 Software configuration 40
7.1.4 Cluster configuration 41
7.2 Latency 41
7.3 k3s reaction time 42
7.4 Stream reaction time 43
7.4.1 Pod failure 44
7.4.2 Cluster failure 44

8 Conclusion and future work **47**

List of Tables

7.1 Kubelet and Controller Manager list of parameter changes 40
7.2 Network average latency 42
7.3 Latency between pod on different nodes, but on the same cluster 42
7.4 Latency between pod on different clusters, peered with Liqo 42
7.5 Average Liqo Latency 42

List of Figures

1.1 Electric energy net general overview 12
3.1 Smart Grid abstract informatic model 20
4.1 Out-of-band control plane peering 25
5.1 Target Telecommunications Architecture Foreseen in the 2023 Development Plan for e-distribution 30
5.2 Independent Groups Scheme 31
5.3 Dependent Groups Scheme 32
6.1 Logical domains scheme 36
6.2 Data Stream comparison in case of failure 37
6.3 2 Level logical domains 38
7.1 Configuration test environment 41
7.2 Reaction to set virtual node as Not Ready in case of remote cluster disconnession 43
7.3 Box plot regarding stream downtime from last old data to the first new data in case of pod failure 44
7.4 Box plot regarding stream downtime from last old data to the first new data in case of cluster failure 45

Chapter 1

Introduction

In recent years, advancements in technology have significantly increased the capacity for data collection across all sectors. However, these improvements have also resurrected longstanding issues associated with managing large volumes of data, such as inefficiencies in transportation networks and data processing centers.

The adoption of edge/fog computing paradigms has addressed these challenges by decentralizing data processing. Edge computing involves processing data directly at the source, while fog computing processes data at intermediate nodes within the network.

This approach enhances the speed and accuracy of data monitoring and consumption systems by reducing the burden on transportation and central processing nodes.

Moreover, these paradigms facilitate the implementation of new functionalities, such as enabling infrastructure to operate independently (island-mode) when disconnected from the main network.

They also increase flexibility by allowing data flows to be rerouted to alternative destinations in response to failures.

Since data is partially processed at distributed nodes, there is less reliance on highly specialized or memory-intensive destination centers, thereby supporting the creation of multiple destination points.

1.1 Energy section

In Italy, the national electrical grid can be divided into four main areas, as highlighted in Figure 1.1.

- **Production:** This area encompasses all energy production facilities, historically dominated by large power plants such as fossil fuel and hydroelectric plants, as well as imported energy. Nowadays, smaller-scale and more variable production outputs from renewable energy sources have been introduced.
- **Transmission:** This area includes the infrastructure responsible for long-distance transmission of produced energy, using high-voltage alternating current. Its primary function, known as "dispatching"[1], is to balance consumption levels with supply

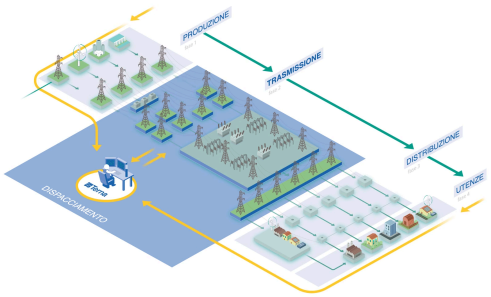

Figure 1.1. Electric energy net general overview

levels since energy cannot be efficiently stored. Managed in Italy by Terna[2], this infrastructure is highly automated to handle power plant failures or service interruptions.

- **Distribution:** This area comprises the infrastructure that transports electrical energy to end-users, passing through primary substations (transforming high-voltage electricity to medium voltage) and secondary substations (from medium voltage to low voltage). It is divided into zones managed by independent distribution companies responsible for maintenance. The Smart Grid model, developed over the past decade, aims to automate this infrastructure similarly to the transmission area to enhance energy control and usage.

- **Consumers:** This area involves delivering electricity to the final customer, determining economic costs, and the characteristics of the supplied energy. These aspects are managed by various sales companies that negotiate agreements with distributors and end-users.

Each area plays a crucial role in the overall operation and management of Italy's electricity network, ensuring efficient and reliable energy supply across the country.

1.2 Thesis objectives

The Smart Grid model converts the entire traditional distribution network into an intelligent information network. It integrates edge/fog computing principles and introduces new functionalities such as supporting island-mode operation and securely managing various energy sources through centralized automation.

Current approaches propose using the Kubernetes platform with multi-master clusters to manage individual stations. While these solutions introduce edge/fog computing and local centralized control, they face challenges in scalability, security, and cannot support critical features like island-mode operation across the entire network.

Building on the feasibility of local Kubernetes solutions in distributed environments, this thesis aims to develop a model that extends edge/fog computing capabilities across the entire power grid.

The innovative open-source project Liqo will be utilized to manage the distributed cluster architecture, enhancing resilience by enabling the system to withstand failures of entire clusters.

1.3 Overview

This thesis endeavors to create a scalable model for the entire network, building upon successful local solutions.

Chapter 2 introduces Kubernetes, an open-source platform, with a focus on its lightweight version, k3s, as the foundational technology.

Chapter 3 delves into the local solution based on Kubernetes, detailing its components and addressing scalability and functionality challenges.

Chapter 4 introduces the Liqo project, pivotal for extending the local model across the entire network, elucidating its core concepts. It examines how this technology interacts with existing distributed database systems designed for single-cluster environments, emphasizing the adjustments needed for seamless integration.

From Chapter 5 onward, the thesis delves deeper into its core discussions. Chapter 5 initiates by outlining the rationale behind selecting the partial mesh star physical topology, exploring various hierarchical configurations within this framework.

Chapter 6 then presents two viable implementations of this model, tailored to the structure of energy distribution grids.

Chapter 7 critically evaluates the first implementation chosen for its robust resilience, contrasting its performance against the initial local solution.

Finally, Chapter 8 provides a reflective analysis of the findings and proposes future research directions.

Chapter 2

Kubernetes

In this chapter, we will briefly describe Kubernetes technology, which has been used as the foundation for local solutions studied in recent years. This thesis integrates Kubernetes with Liqo technology.

As stated in its official documentation[3], Kubernetes is a portable, extensible, open source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation.

Kubernetes emerged as a platform designed to automate the management of containerized applications, ensuring periodic checks to maintain alignment between the actual operational state and the defined ideal state through a declarative language.

A decade since its release as an open-source project, Kubernetes stands as one of the most extensively utilized platforms worldwide. This project focuses on k3s, a lightweight variant of Kubernetes tailored for operation in resource-constrained environments.

2.1 Basic concepts

The foundational principles underlying the architecture of Kubernetes are articulated as follows:

1. Implementation-agnostic APIs: Each Kubernetes object can be implemented differently depending on the version being used, yet the interface used to manage these objects remains consistent across all versions.

2. Completely declarative specification: Kubernetes facilitates the use of a declarative language instead of the traditional imperative approach, simplifying application management by specifying the desired state directly rather than detailing how to achieve that state from various starting points.

3. Control loop-oriented approach: Kubernetes employs components known as controllers that cyclically monitor whether the current state aligns with the desired state. If discrepancies are identified, these controllers initiate actions to minimize the gap between the states.

These principles are the cornerstone of Kubernetes, facilitating efficient management and orchestration of containerized applications in a variety of computing environments.

Every object within Kubernetes is meticulously crafted to adhere to foundational principles, starting with its smallest operational unit: the pod. A pod may consist of one or more containers and its configuration is defined in its respective YAML file. This file may reference other configuration files using key-value pairs, such as Secrets or ConfigMaps, useful in case these configurations are repeated multiple times. Pods are typically instantiated through the implementation of various Kubernetes controllers.

The Deployment controller, commonly used for stateless applications, describes the desired application state while managing scalability through the ReplicaSet. For stateful applications, the StatefulSet controller is normally utilized, managing the pod-to-volume binding and ensuring properties such as unique network IDs that the stateful application required to function properly.

While controllers oversee the lifecycle of pods, pod discovery is entrusted to Services. These Kubernetes objects target all pods matching their selector criteria, facilitating exposure both within and outside the cluster. ClusterIP services expose pods solely within the cluster, whereas NodePort or LoadBalancer services extend pod accessibility externally.

Pods are instantiated on physical or virtual machines known as nodes, which serve either as master or worker nodes based on their role. A master node not only executes various Kubernetes components, as previously discussed, but also hosts the cluster's control plane such as the scheduler, controller manager, and API server. Conversely, a worker node is dedicated solely to executing the workloads of Kubernetes objects.

The cluster, comprising these nodes, can be structured as a single-master or multi-master configuration. In a multi-master setup, control plane components are replicated across all master nodes, and decisions are made via a consensus mechanism facilitated by the etcd quorum process, which utilizes the Raft algorithm [4]. This setup necessitates an odd number of master nodes to prevent split-brain [5] scenarios and reduce decision-making delays. These structural and operational principles form the backbone of Kubernetes architecture, facilitating scalable and efficient container orchestration in diverse computing environments.

2.2 K3s

K3s is a lightweight variant of Kubernetes tailored for operation in resource-constrained environments, as described in the official documentation[6]:

- Edge
- Homelab
- Internet of Things (IoT)
- Continuous Integration (CI)
- Development
- Single board computers (ARM)

- Air-gapped environments
- Embedded K8s
- And as the official page says, situations where a PhD in K8s clusterology is infeasible

K3s efficiently utilizes approximately half the memory resources compared to Kubernetes (K8s) by implementing several optimizations. These include eliminating legacy libraries, opting for lightweight alternatives such as SQLite instead of the standard etcd for database management, and containerd instead of Docker as the container runtime. Additionally, internal mechanisms have been adjusted to reduce memory consumption.

K3s gets its name from the fact that it uses about half the memory, hence it was jokingly named K3s as it consists of 5 letters(K+3+s), which is half of the 10 letters in Kubernetes(K+8+s).

The installation process is streamlined with a simple script that weighs less than 100 MB. Despite its compact size, this script effectively manages many of the complexities typically associated with Kubernetes environments, such as automatically configuring TLS certificates.

This revision clarifies the optimizations made in K3s to reduce memory usage and highlights the streamlined installation process while maintaining readability and technical accuracy.

Chapter 3

State of art

In this chapter is presented the current implementation to support edge and fog computing paradigms within the energy production and monitoring network, based on the Smart Grid model (translating hardware components into an IT network) and the use of the Kubernetes platform. The limitations that the implementation studied in this thesis aims to eliminate will also be highlighted.

3.1 Smart Grid components

Currently, the Smart Grid model of the energy monitoring network consists of the Area Control Center, which is the central hub for control and decision-making mechanisms, and the production and distribution stations, which are divided into primary and secondary. To manage the network, three main applications are used: Phasor Measurement Units (PMU) for measurements, Phasor Data Concentrator (PDC) from the openPDC project for aggregating data from various PMUs, and Grid State Estimation (GSE) for monitoring the network based on the data provided by the previous applications. A graphical example of the Smart Grid model is shown in the figure 3.1.

3.1.1 Area Control Center

A computing node, in the optic of Smart Grid, that represents an Operational Distribution Center, thus housing the control and management logic of the entire network. This node typically manages the high-level PDC, where data streams from other high-level PDCs located at primary stations are aggregated, and the GSE application, which uses the data from the previous PDC to control the network.

3.1.2 Station

A computing node, in the optic of Smart Grid, that represents an energy production or distribution station. Primary stations primarily use a high-level PDC to aggregate data streams from their subnet of secondary stations but can also manage some PMUs.

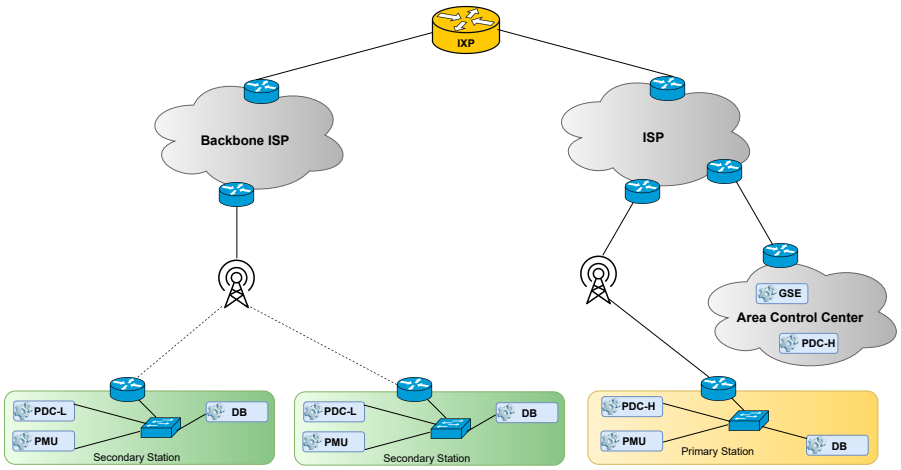

Figure 3.1. Smart Grid abstract informatic model

Secondary stations mainly handle the PMUs or aggregate the data streams produced by those through low-level PDCs.

3.1.3 Phasor Measurement Units

PMUs provide measurements of fundamental electrical quantities, such as voltage and current, in the form of phasors, including information on the amplitude and phase of the measured quantities. These measurements, synchronized via GPS and sampled at a frequency of 50 samples per second, enable precise monitoring of rapid changes in the electrical system caused by the dynamism of distributed energy resources.

PMUs offer a detailed perspective of system dynamics, overcoming the limitations of more traditional Remote Terminal Units (RTUs), which have an update period of several seconds and are not synchronized. The use of PMUs is expected to enhance the observability and reliability of the distribution system.

3.1.4 Phasor Data Concentrator

A phasor data concentrator (PDC) is designed to receive streaming synchrophasor data from phasor measurement units (PMUs) installed on power transmission lines and align these data using GPS timestamps (i.e., it "concentrates" the data based on time). The output of a PDC is a time-synchronized data set that is forwarded to one or more software applications.

openPDC is a flexible platform for high-speed time-series data processing, both in real-time and historically. It does not have significant computational power requirements, so it can be installed anywhere within the synchrophasor infrastructure, including on fanless computers located in substations.

3.1.5 Grid State Estimation

State estimation is a technique that allows for the reconstruction of network states, such as nodal voltages, based on available measurements and the electrical network model. Unlike traditional meters, PMU measurements, which include the phase relative to an absolute reference, simplify the state estimation problem by making it a linear system and significantly reducing the computational load. The objectives of state estimation include the recognition and reduction of measurement errors, the identification of topology errors, the estimation of unmeasured network quantities, and the determination of network parameters through redundant measurements.

3.2 Multi-master station architecture

The current state of research has advanced to managing a single node/place of the Smart Grid through a multi-master architecture [7][8], leveraging the resilience gained from the ability to withstand the failure of a master node, albeit at the cost of increased resources required for replicated control-plane components.

In the cluster, the configuration of applications (for example, the configuration of a PDC if the cluster represents a station) is stored in a high-availability distributed database system. This enables the quick and automatic redeployment to another node in the cluster in case of a failure, without the need to reconfigure the parameters for that application. In this way, the clusters support the automatic local resolution of failures for both applications and the nodes on which they are deployed.

Depending on the location, each cluster represents a point of edge (if managing a secondary station) or fog (if managing a primary station) computing within the Smart Grid, but the overall control architecture is manually established between each pair of clusters, which exist as completely independent separate entities.

3.3 Actual challenge

While this approach of a multi-cluster kubernetes is effective for managing a single station, it proves suboptimal when applied to an entire electrical control system. This is due to both the complexity involved in managing a large number of nodes (with stations alone numbering in the tens of thousands, whereas Kubernetes officially supports up to 5000 nodes [9]) and the fundamental inability to function in isolation. In fact, if a segment of the network underlying a master node becomes isolated from the rest of the architecture, it becomes unmanageable as the master node loses the necessary consensus to initiate new workloads (new pods to manage the isolated entities) and can only partially manage existing workloads (because it can't reschedule workloads if it fails). The only additional failure scenario that the translated architecture can address is when an entire secondary station becomes isolated from the network, allowing the possible PDC pod affected to be relocated to another station. However, this advantage does not outweigh the drawbacks in terms of complexity,lack of scalability, and resource demands inherent in the overall architecture. These challenges can be effectively addressed by adopting Liqo technology.

Chapter 4

Liqo

Due to the rapid adoption of containers as the development environment for applications, there is now a well-established trend towards using orchestration platforms to automate the lifecycle management of containerized applications.

Among the various implementations of these platforms, Kubernetes has gained predominant traction, to the extent that multinational corporations with dedicated cloud departments (such as Google, Amazon, Microsoft, Alibaba...) have developed proprietary solutions based on it.

Recently, a trend similar to the one observed with container adoption has emerged, in which there is a growing need for a system that can automate relationships between various clusters managed by these platforms, whether in the cloud or on-premise.

In this chapter will be summarized the Liqo project, designed to address this necessity, described by its creators [10] as "an open-source project that enables dynamic and seamless Kubernetes multi-cluster topologies, supporting heterogeneous on-premise, cloud, and edge infrastructures."

4.1 Basic concepts

The Liqo technology facilitates the creation of a unified virtual network across diverse clusters, enabling the execution of application pods on remote clusters as if they were local. This system is founded on four primary characteristics: network fabric, peering, offloading, storage fabric

4.1.1 Network fabric

The network fabric is the subsystem of Liqo that seamlessly extends the Kubernetes networking model across multiple independent clusters, enabling pods on different clusters to communicate smoothly even when address NAT is applied.

The control plane of this subsystem resides in the network manager, instantiated as a pod responsible for managing network parameters. It handles tasks both during cluster

peering and inter-cluster communications, as example featuring an interface used by the reflection logic for IP address translation.

Interconnecting two clusters involves deploying a secure VPN tunnel using WireGuard, typically established at the end of the peering process based on negotiated parameters. This functionality is implemented by the Liqo gateway component, operating within the cluster as a privileged pod. It also manages routing tables and configures necessary NAT rules to resolve address conflicts.

Although initialized within the cluster's network, this pod utilizes a separate network namespace and policy routing to avoid conflicts with Kubernetes' existing Container Network Interface (CNI) plugins.

Traffic from local nodes/pods directed to a remote cluster is routed through an overlay network, based on VXLAN, managed by a DaemonSet component. This component is responsible for routing entries and ensures proper handling of traffic across the VPN tunnels.

4.1.2 Peering

Standard peering is the process that establishes a unidirectional link between two different Kubernetes clusters, enabling the sharing of resources and services. Through this connection, the consumer cluster can initiate processes using resources provided by the provider cluster, but not vice versa.

In this context, the consumer cluster initiates an outgoing peering towards the provider, which reciprocates with an incoming peering from the consumer. This linkage is not exclusive, supporting possible bidirectionality and the scenario where a cluster can act as a consumer for some peerings and as a provider for others.

The peering process unfolds through the following steps:

1. Authentication: Each cluster uses a pre-shared token to verify its identity, which has some permissions for Liqo-related resources and negotiations.

2. Parameter Negotiation: The two clusters exchange sets of parameters necessary for finalizing the peering, including network information such as their respective CIDRs or as the amount of resources shared by the provider.

   Some of these parameters can be modified directly or using dedicated plugins, for example is possible adjusting the available resources that the provider cluster shows to the consumer cluster.

3. Creation of the Virtual Node: Within the consumer cluster, a virtual node is created to represent the resources shared by the provider cluster. Processes instantiated using the provider cluster's resources appear to be located within this virtual node, maintaining transparency in the offloading process and adhering to standard Kubernetes practices without requiring API modifications.

4. Configuration of the Network Fabric: The two clusters configure their respective network fabrics and establish a secure VPN tunnel using the previously negotiated parameters (address remapping, endpoints, etc.).

Each connection can be differentiated based on how Liqo's control plane traffic is managed: whether it passes through the VPN tunnel alongside pod traffic (in-band control plane peering) or uses traditional communication channels (out-of-band control plane peering).

In the former case, it is required to expose only the Liqo VPN endpoint to the pod of the remote cluster. However, this setup requires control over both clusters to negotiate network parameters through Liqo CTL tool [11], resulting in a static peering configuration that requires manual intervention for updates.

In the latter case, while to the remote pods must be expose not only the Liqo VPN endpoint but also the Kubernetes API and Liqo authentication service endpoints (as shown in the Figure 4.1), it offers the flexibility to connect clusters across different domains using a pre-shared token and enables dynamic peering, so that an automatic renegotiation of parameters occurs in response to configuration changes.

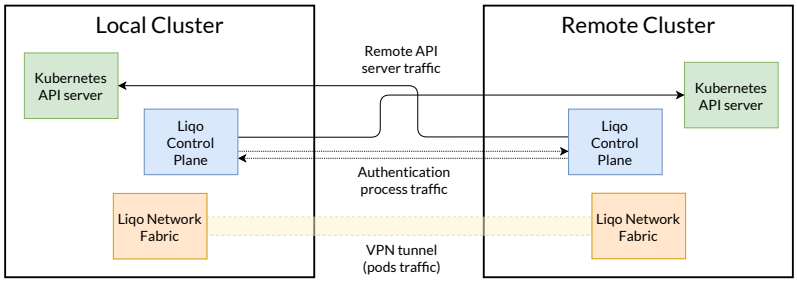

Figure 4.1. Out-of-band control plane peering

4.1.3 Offloading

Offloading is the method enabling transparent extension of the local cluster into a remote cluster, allowing Kubernetes scheduler to seamlessly schedule workloads in the remote cluster when it's deemed optimal. The virtual node is managed by an extended version of the Virtual Kubelet project, which replaces the traditional kubelet if the node isn't physical.

In the context of Liqo, it interacts with the Kubernetes API servers of both clusters to manage artifact propagation (pods, services, config maps) and reconcile state in case of changes on the negotiated configurations. It also performs configurable periodic health checks to assess reachability of the remote API server, marking the node as unready in case of repeated failures and triggering standard Kubernetes evacuation strategies.

An instance of the virtual kubelet is created for each remote cluster to ensure isolation and segregation of authentication tokens.

The offloading process comprises three stages:

1. Namespace Extension: The local cluster's namespace is extended into the remote cluster by creating a gemini counterpart namespace, which will host both offloaded pods and resources required for pod reflection.

2. Resource Reflection: Selected arctifacts from the control plane are reflected in remote clusters to ensure the operational functionality of the pods. Supported resources for reflection include service exposure (Ingress, Services, EndpointSlices), persistent storage (PersistentVolumeClaims, PersistentVolumes), and configuration data (ConfigMaps, Secrets).

3. Pod Offloading: After the scheduler schedules a pod on the virtual node, the corresponding virtual kubelet creates a mirrored pod object in the remote cluster. This object is managed by the custom resource ShadowPod [12], serving as a representation of the pod to maintain service functionality even if connectivity with the remote cluster is lost.

Both stateless and stateful pods are supported, with the latter utilizing either the storage fabric or relying on an external volume provider.

4.1.4 Storage fabric

The storage fabric is the Liqo subsystem responsible for managing the creation of remote volumes for stateful applications. Its operation revolves around delaying the binding of a volume to a pod until it has been scheduled to a node, ensuring volumes are always created where their associated pod is scheduled.

Subsequent scheduling adheres to a data gravity principle, transparently rescheduling the pod to the node where the physical volume resides. These behaviors are implemented through Liqo's virtual storage class, utilizing reflection mechanisms when pods are scheduled on virtual nodes to create the mirrored PVCs in remote clusters. Alternatively, it relies on the real storage class when pods are scheduled on local physical nodes.

4.2 Distributed DB interaction

At present, most of the distributed database systems doesn't support general multi-cluster architecture, primarily due to their reliance on internal headless services for direct pod-to-pod communication. These services return the IP address of the corresponding pod directly when queried, using their DNS entry, unlike regular services that route requests via kube-proxy.

Liqo employs an address remapping mechanism to facilitate seamless communication between clusters; however, this approach results in incorrect IP resolutions for pods scheduled on remote clusters when queried by headless services.

To enable the use of these architectures, Liqo developers currently recommend[13] leveraging the peering process, which exposes the address ranges of the two clusters in two different ways:

1. Connecting a cluster to all others via peering while forcing a pod of the distributed system onto it: This approach ensures that the service in that cluster is aware of all real address ranges, allowing replication through the forced pod, which becomes a critical point.

2. Creating a full mesh of peering between various clusters: This ensures that each headless service knows the addresses of all others, and this is the solution adopted in this research

Some distributed database architectures, such as those implemented by the Percona XtraDB Cluster Operator, may introduce additional complexity. After receiving the translated IP of a remote pod, they may encounter difficulties establishing a connection, primarily because their cluster logic operates with standard Kubernetes component independently of Liqo. This requires the implementation of distinct CIDRs across clusters, ensuring that traffic is routed through Liqo components to establish connections correctly.

Chapter 5

General topology Partial Mesh Star

In this chapter will be described the process that led to the selection of the partial mesh star architecture, a model capable of applying the logical paradigms of edge/fog computing to a multi-cluster architecture while ensuring the possibility of deploying high-availability systems. Initially, the structural choices will be discussed, based on both the multi-cluster environment and the requirements of the adopted technologies (Percona, Liqo, etc.).

Potential use cases will then be described, demonstrating the flexibility of the logical hierarchical architectures that can be implemented. Finally, we will evaluate the characteristics and various limitations that this model entails.

5.1 Physical/net architecture

The first step was to consider how to abstract a logical model from the initial real-world situation. The electric power control and monitoring network, as shown in the figure 5.1, can be schematized using both hierarchical tree topology graphs and peer-to-peer topology graphs.

Peer-to-peer should be discarded because, although the nodes representing the stations can be both data providers and receivers, the node representing the Area Control Center needs to exercise centralized control over all other nodes. Additionally, not all nodes have sufficient processing capabilities, especially if they represent secondary stations.

Among the various tree topology models, the star model is the only one that can be physically implemented using the standard version of Liqo. Indeed, it does not allow the offloading of an already offloaded namespace to prevent critical situations such as circular offloading. This means that all multi-level hierarchical topologies cannot be physically implemented without making customized changes to the technology's code. Moreover, distributed HA database systems tend to need to be in a single namespace, and multi-namespace solutions via operators do not support multi-cluster technologies as they cannot know the namespaces of other clusters.

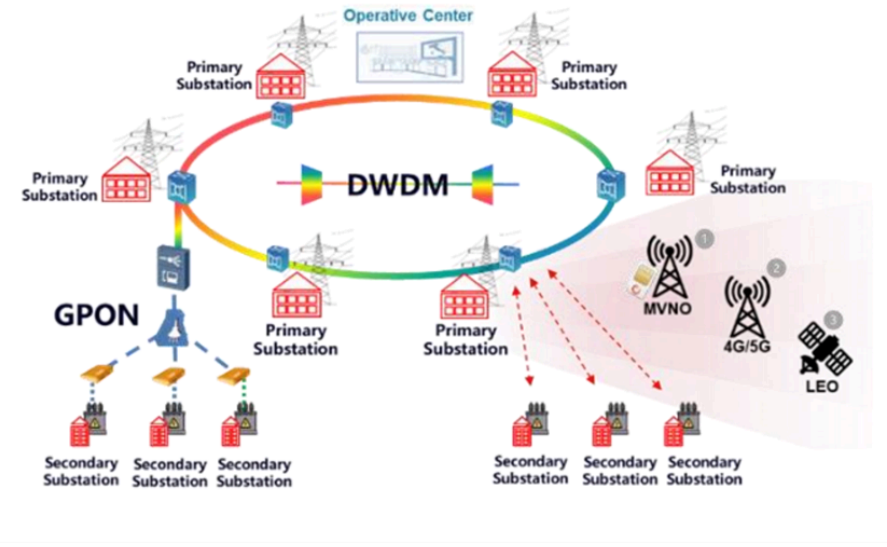

Figure 5.1. Target Telecommunications Architecture Foreseen in the 2023 Development Plan for e-distribution

The partial mesh version of the star model, which allows direct connections between leaves, is necessary for the correct transparent multi-cluster functioning of distributed database systems that rely on headless services. Each cluster that uses the same database system will need, in addition to having the offloaded namespace of the database, a direct connection with all other clusters, thus creating a partial mesh topology (partial because the connections do not necessarily have to be bidirectional).

5.2 Logic complex hierarchy using labels and affinity

A simple partial-mesh star topology offers only a two-level physical separation: a central node and the leaves. It lacks the necessary flexibility to manage the various real-world scenarios encountered in a monitoring and distribution network. Therefore, it is necessary to introduce a strategy to construct a complex logical topology on the existing physical model. This strategy is based on the use of Kubernetes' native label and affinity mechanisms.

Each cluster will be identified by a group of labels that specify its position in the desired logical topology and can be used by the scheduler to distribute the workload according to the intended logic. The node affinity mechanism can be used both to distinguish between different clusters and to differentiate the various nodes within a cluster, as it allows specifying different labels as targets. This way, one can define the label that identifies the cluster as well as the labels of the individual nodes within the cluster.

Pod affinity, on the other hand, is used to enforce coexistence conditions between pods

on the same node. These mechanisms also offer a degree of flexibility, as they provide both "required" and "preferred" options. The "preferred" option allows the scheduler to prioritize the specified target for pod placement while still considering alternative targets if the preferred one is unavailable.

The following subsections will discuss some basic logical topologies, from which one can start to build their desired configuration.

5.2.1 Independent Groups

The leaf clusters of the partial-mesh star model are segmented into independent groups by assigning each node within the cluster a label that uniquely identifies its respective group. This method establishes distinct logical areas, as peering between clusters is only necessary within the same group (and only in case of using a distributed database system), to which separated workloads can be allocated. To enhance the delineation of these divisions, the root node could assign a separate namespace to each group, thereby also increasing security between them.

These groups are not mutually exclusive, provided there is no logical contradiction among the identifying labels. Consequently, a cluster may simultaneously belong to multiple groups, as demonstrated by leaf C in Figure 5.2.

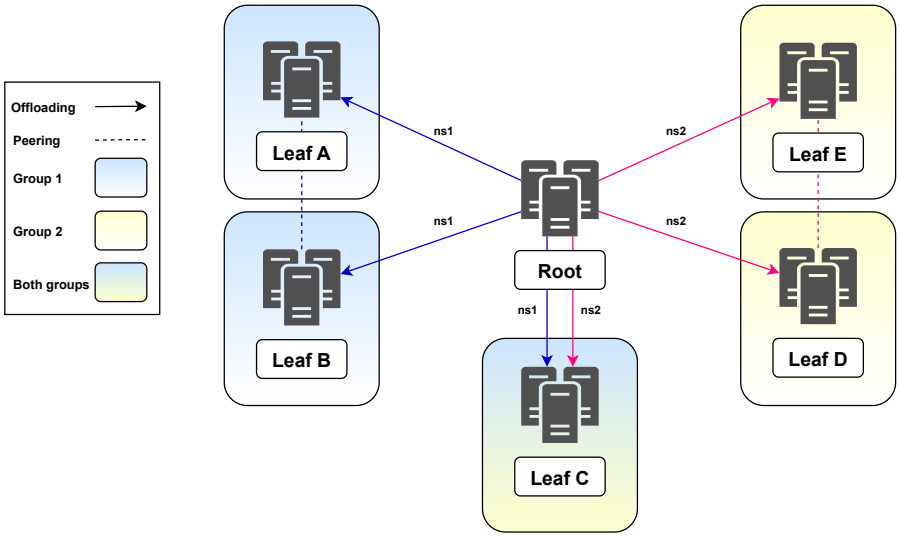

Figure 5.2. Independent Groups Scheme

5.2.2 Dependent Groups

The leaf clusters of the partial-mesh star model are divided following a logical hierarchy by assigning a label indicating their position within the hierarchy. This approach creates

dependent logical areas, as peering is necessary even between clusters belonging to different groups if a distributed database system is used. Figure 5.3 illustrates the worst-case scenario, where the database domain encompasses all leaf clusters. This topology supports new behaviors, such as allowing not only the selection of which groups to schedule workloads within the same domain.

This mechanism allows for the creation of multiple logical hierarchies within the same physical network, each with its own set of labels. Within a given logical hierarchical structure, a cluster can belong to only one group. However, when considering multiple structures, a cluster can belong to different groups.

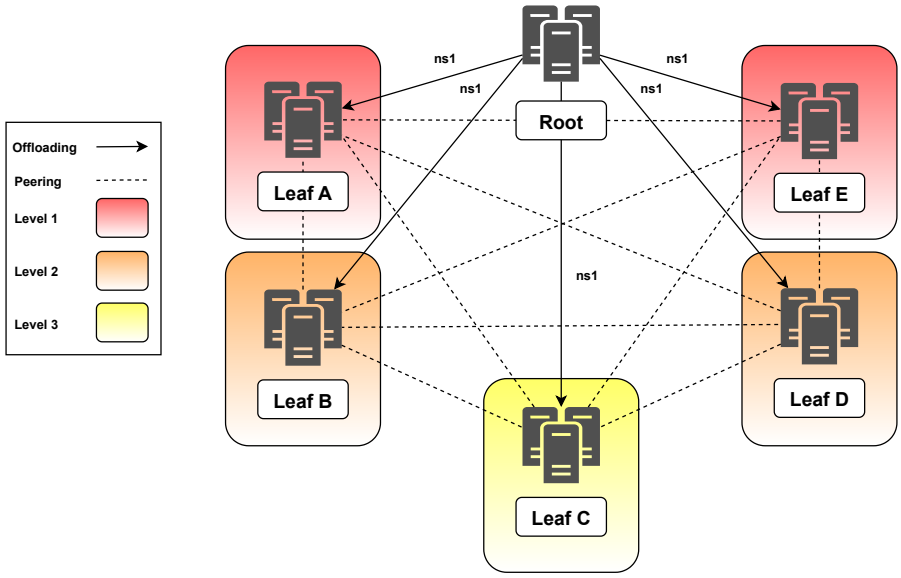

Figure 5.3. Dependent Groups Scheme

5.3 Partial Mesh Star Analysis

As previously illustrated, the partial mesh star topology allows for the use of systems not originally designed for multi-cluster environments, such as distributed HA databases. However, this feature results in a quadratic increase in the number of peerings required between clusters within the database domain. This is because it requires at least a unidirectional peering full mesh. The number of links can be determined using the formula (5.1):

$$\text{Link} = \frac{N(N-1)}{2} \tag{5.1}$$

where N is the number of clusters.

The increase in the number of peerings only affects the time needed to set up the entire architecture during its creation, as the consumption of additional resources is negligible (CITA PAPER Marco?).

The label mechanism offers substantial flexibility in selecting the logical architecture to overlay on the physical infrastructure. However, a drawback is the linear increase in setup time as the number of clusters expands. This characteristic renders the partial mesh star topology ideal for systems with relatively stable physical topologies, facilitating quick adjustments in logical configurations. While significant logical topological changes are supported, they require a corresponding setup time.

Chapter 6

Possible implementations

This chapter will discuss the potential implementations of the partial mesh star topology within the context of a computer network dedicated to energy monitoring. The network primarily consists of the Area Control Center, primary stations, and secondary stations, each managed by its own Kubernetes cluster.

6.1 Logical domains

The Area Control Center occupies the central position in the star topology, establishing unidirectional peering with offloading to every other entity in the topology, whether it is a primary station or a secondary station. This allows the Area Control Center to manage all application deployments without the need to delegate them to other nodes.

The remaining clusters are divided into groups, typically consisting of a primary station and its associated secondary stations. These groups represent a logical domain of applications with their own high-availability distributed database system and, therefore, do not have interconnections among them. Within a group, the clusters form a full mesh of unidirectional peerings for the database system's operation, and they share the same offloaded namespace from the Area Control Center. This implementation is depicted in Figure 6.1.

This architecture allows for the highest degree of resilience, as the only critical point is the Area Control Center, and the effects of a failure or disconnection of this node are negligible when compared to environmental constraints:

1. In the event of a physical failure of the central node, deployments would be lost, making the reconciliation process with the entire network impossible. However, this is negligible because without the central node's logic, the network would not be observable by default.

2. In the event of a complete disconnection from the network, active workloads would continue to function, but the reconciliation process for stateful applications as the database system will not occur since, by the point of view of the deployments, there

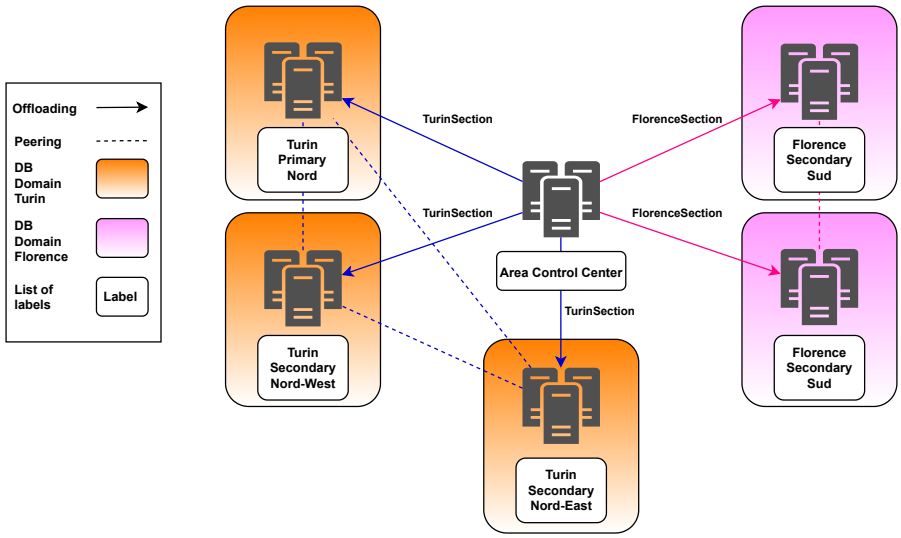

Figure 6.1. Logical domains scheme

would not be enough surviving replicas to maintain the system.Yet, this is also negligible as it falls under the same scenario as before.

In contrast, failures or disconnections in the leaf clusters are fully supported. From the central node's perspective, their load is simply redistributed to other affiliated clusters. Meanwhile, in the leaf clusters, if they are merely disconnected, the workloads continue to operate with the capability to instantiate new applications, allowing isolated operation until reconciliation is achieved.

This can be observed in Figure 6.2, which illustrates the data stream seen by an instance of PDC lower and its directly superior PDC higher, shortly before and shortly after the disconnection of the cluster hosting the PDC lower and its data sources. PDC lower continues to receive data from the sources, operating in an isolated environment, while PDC higher stops receiving the data stream from the isolated source.

The limitations of this architecture pertain to scalability, as each cluster requires its own distinct CIDR for the transparent operation of high-availability distributed database systems. Additionally, each peering creates a virtual representative node in the central cluster, limiting the number of possible clusters to 5000, according to the official Kubernetes documentation.

6.2 Multi-level logical domains

The implementation described in this section leverages a star topology twice, once with a partial mesh version and once with a full version, as shown in Figure 6.3. This follows the division of stations into primary and secondary, although it could be adapted to n

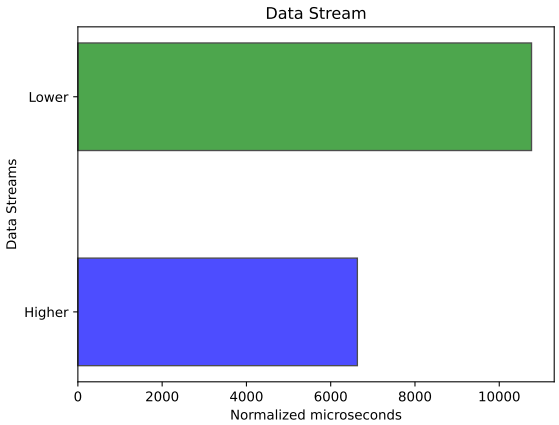

Figure 6.2. Data Stream comparison in case of failure

subdivisions.

The first topology used is a partial mesh star topology used to connect the Area Control Center (central cluster) with all primary stations (leaf clusters). The central cluster handles the deployment of high-level applications along with their respective distributed database systems, offloading the corresponding namespace through peering to the respective group of primary stations.

The groups of primary stations are composed of a main primary station, where workload is preferably directed (using labels), while the others in the group primarily serve as backups in case of failure of the main station. This means that primary stations can be in multiple groups, one where they are primary and others where they function as backups, leveraging the topology seen in Chapter 5 section 5.2.1.

The main primary station of each group also serves as the central cluster in the second star topology, connecting not only to the backup primary stations but also to all secondary stations under its jurisdiction. In our implementation, this will be a full mesh star topology, but a partial mesh could also be used if the secondary stations do not share the same distributed database system.

In this second topology, the main primary station handles the deployment of low-level applications along with their respective high-availability distributed databases, consequently offloading the namespace to its secondary stations. The data stream for monitoring, which passes through two different namespaces (from the low-level to the high-level), relies on external exposure services such as load balancers and ingress, enabling access to the high-level application whether it resides in the primary station or, due to a failure, in one of the backup primary stations.

This architecture enhances scalability limits compared to the previous implementation by reducing the number of peerings managed by the Area Control Center and requiring distinct CIDRs only within the secondary topologies associated with a primary station, allowing for reuse across different types. However, this benefit is balanced by a decrease

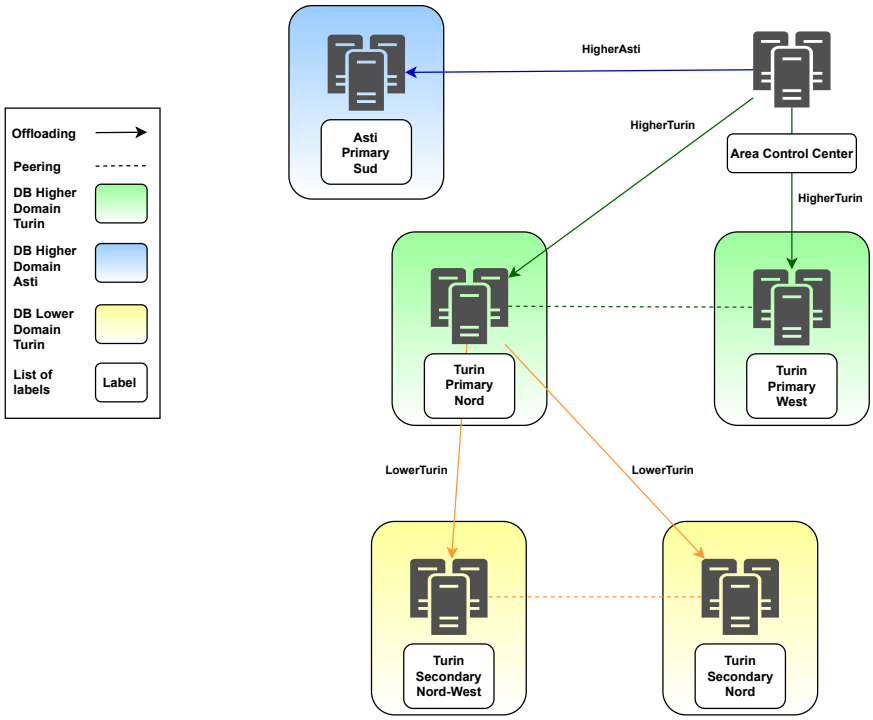

Figure 6.3. 2 Level logical domains

in overall resilience, as a failure or disconnection of the primary station results in the loss of deployment for low-level applications, which is not a feature supported by the Liqo technology, necessitating system reset upon reconnection.

6.3 Final consideration

This thesis focuses on achieving the highest degree of resilience; therefore, the following chapters will focus on testing the first implementation. It is important to note that these two implementations are not mutually exclusive; they can be implemented simultaneously within the same physical network, in cases where different parts of the network require varying degrees of resilience.

Chapter 7

Domain peering evaluation

In this chapter, we present the analyses conducted on the logical domain grouping implementation, chosen for its higher resilience compared to the multi-level implementation.

7.1 Test Environment

The test environment was created by leveraging the functionalities of Crownlabs, an open-source platform associated with the Politecnico di Torino, which was developed during the years of the Coronavirus spread.

7.1.1 Crownlabs

Crownlabs is an open-source project created to provide students with access to laboratory systems and services during the challenging times of the coronavirus pandemic, which imposed severe travel restrictions. In fact, the name derives from the virus itself and its initial purpose, as "Crown" translates to "Corona" in Italian and labs measn laboratories.

The authors of this project were a group of volunteers primarily composed of MSc students who, within just a few weeks of very hard work, as described on the project website [14], managed to deliver a functioning version, to address the university places closures mandated by the Italian government.

Nowadays, Crownlabs continues to be supported by students, and its functionalities have expanded: it not only allows the remote use of laboratory machines through a web browser, enabling both personal exercises and group work, but also leverages the Politecnico di Torino's data center to instantiate and use virtual machines transparently within an internal network.

It is precisely this latter functionality that has been utilized, as the Kubernetes clusters used were composed of these virtual machines.

7.1.2 Nodes configuration

Each virtual machine representing a node in a cluster possesses the following characteristics, chosen to simulate low computational capacity typical of devices found in energy

monitoring and distribution stations.

- **Operating system:** Ubuntu server 20.04 LTS.
- **CPU:** 4 core.
- **RAM:** 8 GB.
- **Disk memory:** 25 GB.

7.1.3 Software configuration

Di seguito vengono specificate le versioni delle piattaforme utilizzate:

- **K3s:** v1.24.17+k3s1.
- **Liqoctl:** v0.10.2.
- **Liqo:** v0.10.2.
- **PDC:** v2.4.
- **Database:** Percona XtraDB operator v1.11.0.
- **Database connector:** MYSQL connector v8.2.

It should be noted that certain parameters of the k3s kubelet and manager controller have been adjusted to decrease the cluster response time in case of failure, as detailed in the table 7.1.

Option	Value	Description
node-status-update-frequency	10s -> 5s	Specifies how often kubelet posts node status
node-monitor-grace-period	40s -> 20s	Specifies the amount of time in seconds that the Kubernetes Controller Manager waits for an update from a kubelet before marking the node unhealthy. Must be N times more than kubelet's nodeStatusUpdateFrequency, where N means number of retries allowed for kubelet to post node status
pod-eviction-timeout	300s -> 5s	This parameter specifies how long Kubernetes waits before evicting pod from a node marked as "NotReady"

Table 7.1. Kubelet and Controller Manager list of parameter changes

7.1.4 Cluster configuration

Due to the limit of 5 virtual machines, the system was organized into 5 Kubernetes clusters, each comprising a single node. As shown in Figure 7.1, the topology is a fully-meshed star topology where the root cluster occupies the central position, hosting all deployments of the PMU, PDC, and database applications. Through Liqo peering, it offloads the test namespace to the leaf clusters.

The leaf clusters are connected by unidirectional peering for the transparent operation of the distributed Percona database system, and they will be the only locations where the pods of the aforementioned applications can be scheduled.

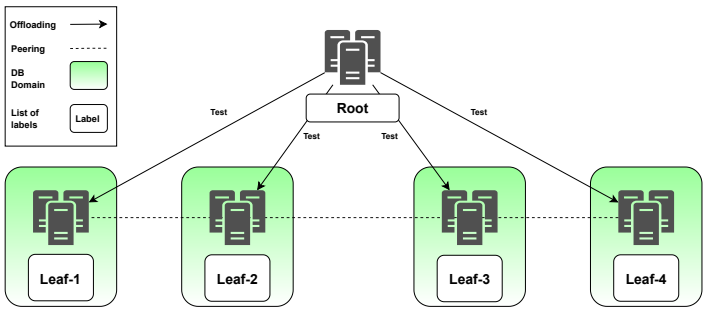

Figure 7.1. Configuration test environment

7.2 Latency

In this section, we demonstrate the latency increase due to the overhead generated by Liqo technology, using 4 virtual machines, where one is consistently used as the secondary member (Root), while the others represent the primary member for each test (Leaves).

Considering that data in our architecture is transmitted using TCP protocols, which utilize acknowledgments (ACKs), latency is measured as the round-trip time of a packet. The following tests were conducted using the ping command, with approximately 1000 iterations per test, individually executed from the Leaf machines to the Root machine.

Firstly, to establish a baseline for the tests, the network latency was calculated by averaging the mean of three ping values obtained from the virtual machines towards the root machine, as shown in the table 7.2.

After establishing the network latency, we proceed to calculate the latency between a pod located on different Leaf nodes and a pod in the Root node, where the Leaf nodes and the Root node belong to the same cluster, shown in Table 7.3, or belong to different clusters peered with Liqo, shown in Table 7.4.

Virtual machine	Latency
Leaf-1	0.689 ± 0.447
Leaf-2	0.760 ± 0.771
Leaf-3	0.715 ± 0.468
Avg	0.721 ± 0.581

Table 7.2. Network average latency

Cluster Node	Latency
Leaf-1	0.735 ± 0.294
Leaf-2	0.992 ± 0.568
Leaf-3	0.936 ± 0.483
Avg	0.888 ± 0.267

Table 7.3. Latency between pod on different nodes, but on the same cluster

Remote Node	Latency
Leaf-1	1.269 ± 0.724
Leaf-2	1.400 ± 0.829
Leaf-3	1.368 ± 0.903
Avg	1.346 ± 0.821

Table 7.4. Latency between pod on different clusters, peered with Liqo

The increase due to Liqo is measured by subtracting the average latency from Table 7.3, which is the sum of network latency + Kubernetes overhead, from the average latency shown in Table 7.4, which is the sum of network latency + Kubernetes overhead + Liqo overhead. This calculation yields the result shown in Table 7.5.

Average Liqo Latency
0.458 ś 0.864 ms

Table 7.5. Average Liqo Latency

The latency between pods on different Kubernetes clusters without multi-cluster technologies was not tested, as the goal is to demonstrate the latency increase using Liqo across different clusters compared to using a single Kubernetes cluster connecting all nodes.

7.3 k3s reaction time

In the upcoming test, two clusters of virtual machines connected via unidirectional Liqo peering were utilized: the consumer cluster and the provider cluster. The objective was to show the consumer cluster's response time in the event of disconnection of the virtual node representing the provider cluster, for any reason.

The test involved two scripts. The first script disabled the network interface on the virtual machine running Liqo in the provider cluster and recorded the timestamp. The second script executed a loop on the consumer cluster, running 'kubectl get node' every 0.4 seconds, and appending the output with a timestamp. (A shorter interval wasn't feasible due to the command execution time.)

The results are depicted in graph A. 7.2.

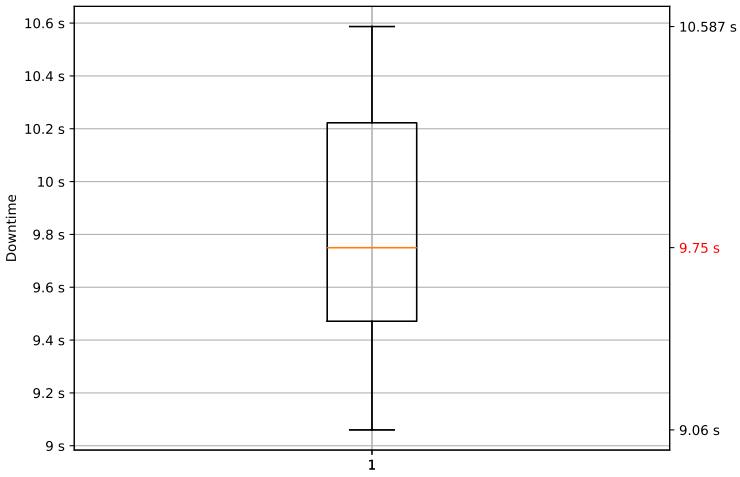

Figure 7.2. Reaction to set virtual node as Not Ready in case of remote cluster disconnession

7.4 Stream reaction time

The following tests demonstrate the downtime of a data stream from a PMU in the following scenarios:

1. Internal failure of a PDC pod, resulting in the rescheduling of the application.

2. Fault/disconnection of the cluster hosting a PDC pod, resulting in the application being rescheduled to another cluster.

Downtime is calculated from the timestamp of the last data frame of the old stream to the timestamp of the first data frame of the new stream, encompassing the time required for rescheduling the PDC pod, retrieving configurations from the database system, and reconnecting to the data stream.

These tests examine a data stream originating from a PMU that traverses through two PDC pod, one considered low-level and the other considered high-level, before reaching its intended application.

The use of the data frame timestamp is crucial due to the PMU's real-time production of data frames at 33-millisecond intervals, ensuring precise downtime calculations and analysis.

7.4.1 Pod failure

The failure scenario of the PDC pod was simulated by customizing the liveness probe mechanism, intentionally triggering a failure check after the pod had been running for 60 seconds.

The test results, displayed in graph 7.3, illustrate the median duration required for the lower-level PDC application to resume normal operation. This duration encompasses the time from detecting the PDC pod failure to its subsequent recovery, including the processes of restarting the pod, retrieving configurations from the system database, and re-establishing connection to the data stream.

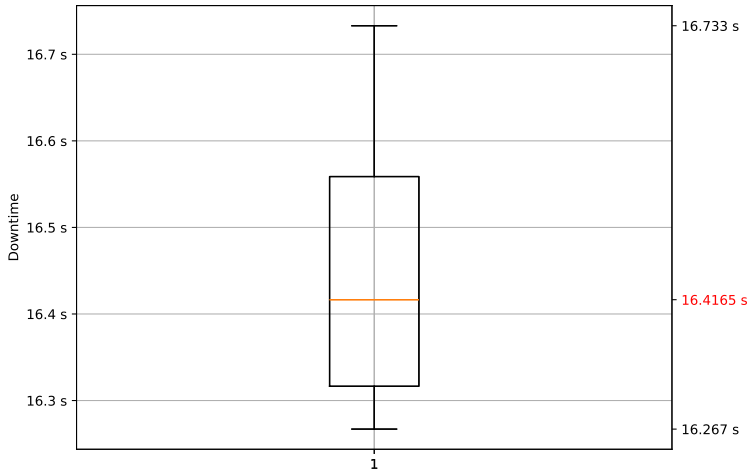

Figure 7.3. Box plot regarding stream downtime from last old data to the first new data in case of pod failure

7.4.2 Cluster failure

The failure scenario of the cluster containing the lower-level PDC pod was simulated by directly disabling its network interface.

The test results, displayed in graph 7.4, illustrate the median duration required for the lower-level PDC application to resume normal operation. This duration includes the time required for the cluster to detect that the virtual node hosting the PDC is unreachable (9.75 seconds as shown in graph 7.2), the waiting time before it can be rescheduled to another node (5 seconds as indicated in table 7.1), and the time necessary for the pod to restart (16.41 seconds as depicted in graph 7.3).

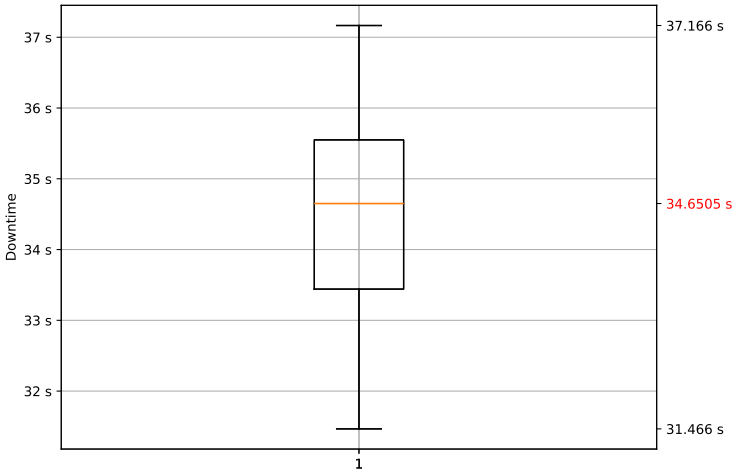

Figure 7.4. Box plot regarding stream downtime from last old data to the first new data in case of cluster failure

Chapter 8

Conclusion and future work

The introduction of Liqo technology in research to implement the paradigms of Edge and Fog computing within the Smart Grid model has not only increased scalability by condensing multiple nodes into a single virtual node but has also enabled the introduction of new functionalities that were previously difficult to implement.

For example, in case of disconnection from the central network, the two parts of the network can operate autonomously, with the capability to deploy new applications until reconnection with the central network (island-mode operation).

The topology humorously called the "winning" topology for real implementation is the partial mesh star topology. Despite not being a complex architecture, it meets all constraints, stemming from the transparent operation of non-multi-cluster native applications, such as distributed database systems, and from design constraints like seeking the lowest possible power consumption and high resilience.

These results were described in the previous chapter, comparing them with the baseline solution values and demonstrating their similarity, without significant latency increases despite increased complexity.

Obviously, this solution does not represent a panacea for all issues; for instance, it is still quite limited in terms of scalability. A potential research direction could be to further explore the possibility of introducing a higher hierarchical level, as demonstrated in the second implementation in Chapter 6, while striving to maintain the high level of reliability demonstrated in this thesis.

This work was conducted using versions utilized in previous research to compare results and demonstrate the efficiency that Liqo technology enables. From this point, implementations on new versions could be developed, which provide access to new functionalities that could optimize the entire structure.

For example, the new versions of k3s and Percona allow the use of the spread operator, which would reduce the complexity of creating the logical hierarchical infrastructure.

Other future researches could focus on investigating the security implications of deploying such decentralized systems. Moreover, collaboration with industry partners could facilitate the transition from theoretical research to practical, real-world applications.

In conclusion, this thesis has demonstrated that integrating Liqo technology into Smart Grid models significantly enhances scalability and functionality. While challenges remain,

the groundwork laid here provides a solid foundation for future advancements. The continued evolution and optimization of these technologies promise to drive significant improvements in the efficiency and resilience of critical infrastructures.

Bibliography

[1] Wikipedia. Dispacciamento, 2024. URL https://it.wikipedia.org/wiki/Dispacciamento. (cit on page kubernetes-section0).

[2] Terna. Terna, 2024. URL https://www.terna.it/it. (cit on page kubernetes-section0).

[3] Kubernetes documentation. Kubernetes overview, 2024. URL https://kubernetes.io/docs/concepts/overview/. (cit on page kubernetes-section0).

[4] RAFT documentation. Raft algorithm, 2024. URL https://raft.github.io/. (cit on page kubernetes-section1).

[5] Sidero Labs. Split brain scenario, 2024. URL https://www.siderolabs.com/blog/why-should-a-kubernetes-control-plane-be-three-nodes/. (cit on page kubernetes-section1).

[6] K3s documentation. Best environment for k3s, 2024. URL https://docs.k3s.io/. (cit on page kubernetes-section2).

[7] Andrea Cazzaniga Fabrizio Garrone Roberta Terruggia Riccardo Lazzari Stefano Galantino, Fulvio Risso. An edge-based architecture for phasor measurements in smart grids. *2022 AEIT International Annual Conference (AEIT)*, pages 1–6, 2022.

[8] Sebastiano La Terra. Analysis of the resilience of monitoring services in smart grid. 2022. URL https://webthesis.biblio.polito.it/24583/.

[9] K3s documentation. Large cluster consideration, 2024. URL https://kubernetes.io/docs/setup/best-practices/cluster-large/. (cit on page kubernetes-section2).

[10] Liqo documentation. Liqo definition, 2024. URL https://docs.liqo.io/en/v0.11.0-rc.3/. (cit on page kubernetes-section2).

[11] Liqo documentation. Liqo ctl tool, 2024. URL https://docs.liqo.io/en/v0.11.0-rc.3/installation/liqoctl.html. (cit on page kubernetes-section2).

[12] Giuseppe Alicino. Prototyping a cloud resource broker. 2021. URL https://webthesis.biblio.polito.it/21145/.

[13] Riccardo Medina. Issue distributed database, 2024. URL https://github.com/liqotech/liqo/issues/2386. (cit on page kubernetes-section2).

[14] Crownlabs authors. Crownlabs history, 2024. URL https://crownlabs.polito.it/about/. (cit on page kubernetes-section2).