Collection and Modelling of Real Network Data for Cloud-Edge Scenarios

Proposal

The performance of cloud-edge distributed systems strongly depends on the characteristics of the network connecting nodes, sites, and regions: latency between different geographic locations, available bandwidth, temporal variability of delays, congestion, packet loss, and the topological structure of the network. These factors directly influence operational decisions such as service placement, replication across multiple sites, or migration of components between cloud and edge.

In the state of the art, datasets and real measurements related to WAN networks, edge/MEC infrastructures, and network topologies exist, but these data are often separated by application context, collected using different methodologies, and described through non-uniform formats and metrics. Consequently, it is difficult to use them in an integrated way to build realistic network scenarios for studying distributed systems and resource management strategies.

The goal of this thesis is to collect, analyse, and integrate real network data by building a coherent, clean, and structured dataset that homogeneously describes topologies, nodes, links, and network metrics in cloud-edge scenarios. The dataset must make it possible to represent, for example, relationships between sites, delays between nodes, link capacities, and performance variability over time, so that it can be used as a basis for simulations and studies on orchestration and distributed systems.

The work includes the analysis of public datasets related to network topologies and metrics, the definition of a unified data model to describe nodes, links, and network properties, data cleaning (normalization of measurement units, handling missing values, format conversion), and, where possible, temporal alignment of measurements from different sources. A significant part of the thesis will be dedicated to the statistical characterization of networks, for example by studying latency distributions, differences between edge–cloud and intra-cloud links, and performance variability over time. The result will be a documented and reusable dataset, accompanied by an analysis highlighting its main properties and limitations.

Tools and Technologies

- Python
- Pandas / NumPy
- Public network and topology datasets
- Matplotlib / Seaborn

- JSON / CSV / Parquet

Objectives

1. Analyse and select relevant data sources on network topologies and metrics.
2. Design a unified data schema.
3. Implement data integration and cleaning pipelines.
4. Perform statistical analyses on latency and bandwidth.
5. Produce a documented and reusable dataset.
6. Demonstrate a concrete use case that uses the produced dataset.

Prerequisites

- Good knowledge of Python.
- Basic knowledge of computer networks.
- Interest in distributed systems and data analysis.