Heterogeneity-aware Deep Learning Workload Deployments on the Computing Continuum

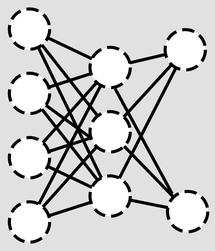
Thomas Bouvier, Alexandru Costan, Gabriel Antoniu

Heterogeneity-aware Deep Learning Workload Deployments on the Computing Continuum

PhD student: Thomas Bouvier **Advisors**: Alexandru Costan, Gabriel Antoniu
Univ Rennes, INSA, Inria, CNRS, IRISA — Rennes, France

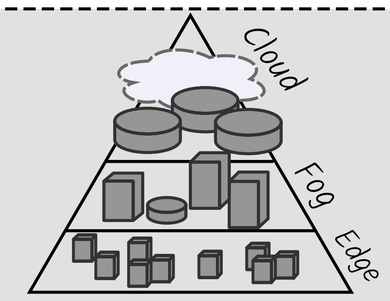
Deep Learning (DL)

- Technique to **build knowledge by training** models over large datasets
- Outperforms human experts in many domains



Computing Continuum

- Shift from centralized clouds towards **multi-tier processing units**
- Spans over **clouds** and **fog** mini-clusters to **edge** devices

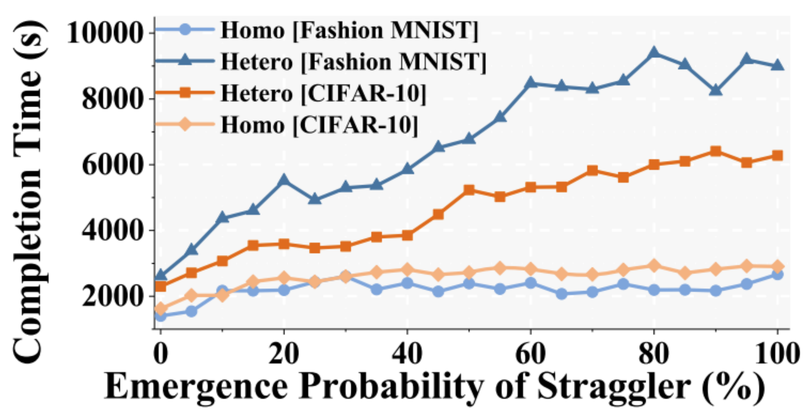


Problem statement

How to perform efficient DL training on the Computing Continuum?

Challenges

- **Network heterogeneity** caused by low-speed, high-latency wide-area networks
- **Compute heterogeneity** caused by GPUs emerging alongside traditional CPUs
- Geo-distributed computing typically leads to stragglers in **large scale scenarios**



Zhou et al., "Falcon: Towards Computation-Parallel Deep Learning in Heterogeneous Parameter Server"

PhD objective

Devise a **middleware layer** to distribute DL training workloads in heterogeneous environments

Deterministic heterogeneity

- Different compute resources
- Slow networks links

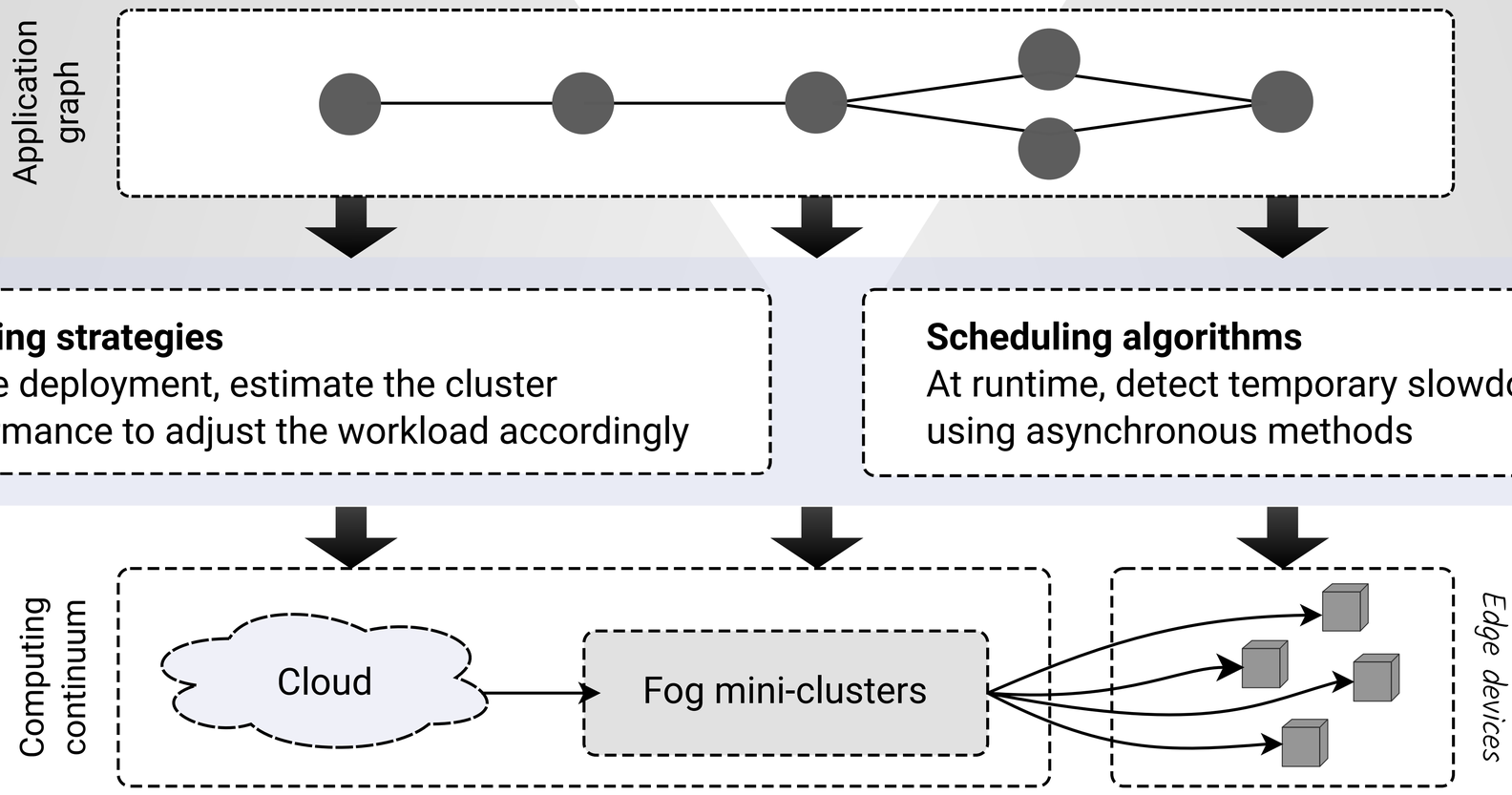
Approach

Distinguish heterogeneity caused by hardware differences and real-time events

Dynamic heterogeneity

- Temporary I/O bottlenecks
- Temporary network slowdowns

Heterogeneity-aware DL workload deployment



Evaluation

- Conduct **large scale** experiments on the Grid'5000 testbed
- Optimize **makespan, cost and fairness** objectives
- Validate with synthetic workloads and **real life applications**

Perspectives

- Prototype a framework distributing DL training efficiently
- Apply the devised techniques to **incremental learning workloads**