Heterogeneity-aware Deep Learning Workload Deployments on the Computing Continuum

Thomas Bouvier, Alexandru Costan, Gabriel Antoniu

To cite this version:
Thomas Bouvier, Alexandru Costan, Gabriel Antoniu. Heterogeneity-aware Deep Learning Workload Deployments on the Computing Continuum. IPDPS 2021 - 35th IEEE International Parallel and Distributed Processing Symposium, May 2021, Virtual / Portland, United States. hal-03270129

HAL Id: hal-03270129
https://hal.science/hal-03270129
Submitted on 25 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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

---

**PhD objective**  
Devises a middleware layer to distribute DL training workloads in heterogeneous environments

---

**Deterministic heterogeneity**  
- Different compute resources  
- Slow networks links

**Dynamic heterogeneity**  
- Temporary I/O bottlenecks  
- Temporary network slowdowns

---

**Approach**  
Distinguish heterogeneity caused by hardware differences and real-time events

---

**Heterogeneity-aware DL workload deployment**

**Application graph**

- Mapping strategies  
  Before deployment, estimate the cluster performance to adjust the workload accordingly

- Scheduling algorithms  
  At runtime, detect temporary slowdowns to mitigate using asynchronous methods

---

**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

---

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