Amory Hoste

Please contact me on Linkedin for a version with full contact details.

SUMMARY

Senior Systems Research Engineer specializing in high-performance AI and cloud infrastructure, with a focus on low-level LLM inference optimization (inference engine, kernels, networking).

WORK EXPERIENCE

Huawei R&D UK, Senior Systems Research Engineer **May 2023–Present**

Edinburgh, UK

Large scale LLM Inference optimization for Huawei Ascend NPUs.

Led multiple key projects to production integration and supervised two research interns. Currently working on long-context LLM inference and sparse attention.

- Developed lightweight NPU Peer-to-Peer (P2P) Transfer Library, increasing KV cache transfer bandwidth by 2.3x, significantly outperforming existing NPU libraries for both RoCE and HCCS.
- Wrote high-performance NPU kernels for several critical scenarios including Mixture of Experts Dispatch/Combine, Large Recommendation Model Embedding Retrieval and KV Cache Transfer.
- Contributed support for LLM Prefill-Decode (PD) Disaggregation and P2P KV Cache Sharing on vLLM-Ascend to the open-source LMCache-Ascend project.
- Improved Ascend 910B point-to-point bandwidth by 5.57x over single-path baseline by developing a software-based multipath transfer library tailored for its mesh-based topology.
- Developed a QoS aware NPU-sharing mechanism, improving resource utilization by enabling colocation of smaller models while maintaining SLOs.
- **Awards:** 2x President's Award - Significant Business Contribution, European Research Institute Excellent Contributor Award, 2012 Labs Outstanding Contributor Award, Quality Star Award

Huawei R&D UK, Systems Research Engineer **Nov 2021–May 2023**

Edinburgh, UK

Performance & resource efficiency optimization of Huawei cloud workloads.

- Developed a distributed Kubernetes scheduler optimized for real-time, high-throughput scheduling decisions, utilizing eBPF for fine-grained, low-overhead monitoring.
- Designed and implemented custom scheduling algorithms to maximize resource utilization and ensure performance isolation for colocated cloud workloads.
- Created a comprehensive benchmark suite and load generator to evaluate new algorithms and architectures against representative production scenarios.
- **Awards:** Future Star Award

Imec IDLab, Research Intern **Summer 2018 & 2019**

Ghent, Belgium

- Built web archival and automated quality analysis tools for the Royal Library of Belgium.
- Developed a fragmented R-tree index to enable efficient geospatial querying of linked data.

EDUCATION

ETH Zurich **Sep 2019–Sep 2021**

MSc Computer Science. Grade: 5.71/6 (Top 10% of class).

- Focus on (Distributed) Systems and High Performance Computing.
- Thesis: Optimization of Serverless Cold Start Latencies through Function Snapshots.

Ghent University **Sep 2016–Jun 2019**

BSc Computer Science. Grade: 808/1000 (1st of class).

TECHNICAL SKILLS

Programming Languages:	Python, C/C++, Go
ML & Inference:	vLLM Internals, PyTorch, Kernel Development, RDMA/RoCE, CUDA
Cloud:	Kubernetes, Container Runtimes, eBPF, Serverless, DevOps & Observability