OOKAMI PROJECT APPLICATION

Date: 02/17/2022

Project Title: Novel Computer Technologies in Education and Research

Usage:

- Testbed: ✓
- Production

Principal Investigator:

- University/Company/Institute: Friedrich Schiller University Jena
- Mailing address including country: FSU Jena – Institute of Computer Science, Fuerstengraben 1, Jena, 07743 Germany
- Phone number: +49 3641 946 371
- Email: alex.breuer@uni-jena.de

Names & Email of initial project users:

- Alexander Breuer (alex.breuer@uni-jena.de)
- Antonio Noack (antonio.noack@uni-jena.de)

Usage Description:

This project will harness Ookami for educating the next generation of High Performance Computing (HPC) scientists, basic architecture research, and the optimization of demanding workloads. Our team is developing efficient and large scale numerical solvers, and is very active in the optimization of Machine Learning (ML) and HPC backends. For example, recent work[1] illustrates the JIT code generation of tensor kernels for a broad range of microarchitectures, including the Arm server processors Graviton2 (N1), Altra (N1) and A64FX, in the context of ML and HPC. Regarding our teaching efforts, we recently

[1] The pre-print “Tensor Processing Primitives: A Programming Abstraction for Efficiency and Portability in Deep Learning & HPC Workloads” is available from: `https://arxiv.org/abs/2104.05755v4`.

introduced the Arm-architecture in a Master-level class targeting HPC. The 2021 class[2] tackled, e.g., the optimization of small matrix multiplications for AWS Graviton2 and Ampere Altra. The SVE-part of the class used ArmIE due to lacking hardware-access for teaching at the time. Within this project, we will make Ookami's A64FX-nodes available to our students (about 5-10 in 2022) and shift the focus of the class to SVE. Further, we are in the process of designing a new class on "Efficient Machine Learning" which will be given in the summer term 2022 for the first time. A part of this class will cover ML-backends and infrastructure. Possible topics include Tensorflow extensions, JAX, or lowering to ML-graphs to MLIR and JITted kernels. Here, we envision also using Ookami's GPU and Skylake nodes. Further, we plan to cover ML for scientific applications where our students apply there knowledge to large(r) scientific data sets, e.g., the identification of seismic facies or benchmarks of MLPerf's Science working group.

Computational Resources:

- Total node hours per year: 5,000
- Size (nodes) and duration (hours) for a typical batch job: 1-16 nodes for 6 hours (development), larger for scaling tests
- Disk space (home, project, scratch): 100GB, 1TB, 2TB

Personnel Resources (assistance in porting/tuning, or training for your users):

No assistance for porting/tuning is required. We are interested in working together with researchers/engineers related to Ookami and potentially sharing experiences or teaching materials.

Required software:

Typical cluster-infrastructure, i.e., compilers, MPI libs, etc.

If your research is supported by US federal agencies:

- Agency: N/A
- Grant number(s): N/A

---

[2] The entry-page of the class is located at `https://scalable.uni-jena.de/opt/hpc`