OOKAMI PROJECT APPLICATION

Date: 05/10/2021

Project Title: Mathematics of Arrays (MoA) Dense GEMM

Usage:

- **Testbed.** MoA design of the algorithm guarantees that it can be mapped to any architecture optimally. We want to test it on A64FX as we have been doing on NVIDIA V100s (XSEDE's Expanse resources at SDSC). Our goals are to show we can predict, validate, port, and scale, our new design for Matrix Multiplication to Ookami by viewing the data and machine as arrays.
- **Method of experimentation.** In this project, we would like to understand the A64FX architecture in terms of scalability and data movement through the memory hierarchy. Thus, the experiments are exploratory: we run the same code slowly incrementing matrix sizes (from small to very large) and measuring performance. During the initial phase of the project, we are not considering matrix specific performance optimizations (speeding up) of the code.

Principal Investigator:

Lenore M. Mullin PhD, Professor Emeritus
University at Albany, SUNY
2641 S Vance Ct. Apt 203, Arlington, VA 22206 USA
703-521-0728 and 518-424-7802
lmullin@albany.edu and lenore.mullin@gmail.com

Names & Email of initial project users:

Katarzyna (Kasia) Świrydowicz PhD, kasia.swirydowicz@pnnl.gov

Manu Shantharam PhD, mshantharam@sdsc.edu

Personnel Resources:

Ideally, it would be best to have access to an individual with expertise in using the machine, e.g. compilers, and other resources for GPUs and CPUs. Especially needed is expertise on specific use of OpenACC, OpenMP, and OpenMPI, as well as other resources like these.

Required software:

OpenACC, OpenMP, OpenMPI, FORTRAN and C compilers, ...

Testbed Project

1. Initial research is on an XSEDE machine, Expanse at SDSC, to show optimality for matrix multiplication based on MoA. MoA, A Mathematics of Arrays, is a formal theory to design and verify designs and mapping to architectures. Initially, designs are formulated and reduced by hand to a DNF, a semantic normal form, and the ONF, an operational normal form specific to an architecture. Eventually, all can be mechanized, e.g. PythonMoA, is a start. However, all hand designs are essential, automated efforts must be able to *target* them.

2. Research will make effective use of the key A64FX architectural features (notably SVE, the high-bandwidth memory, and NUMA characteristics). This is ideal. From an MoA point of view, shapes and indexing are key and variable length registers, are very similar to thoughts of an MoA machine In MoA, shape is essential for analysis and optimizations based on compositional indexing.

   The following **computational resources** will be used:

   - Total node hours per year: 2,000
   - Size (nodes) and duration (hours) for a typical batch job: The jobs will be single node runs and we expect each of these runs to be less than an hour.
   - Disk space (home, project, scratch): default storage size. We would like to request the shared project space of 500 GB for our collaborative work.

   We would like to request access to the following resources as well (default computational time) to compare and contrast our methodology on different architectures: dual socket AMD Rome and dual socket Thunder X2.

   The long term goal of MoA research is not only formally designed software but scalable and verified software, using an MoA instruction set and MoA operating system based only on tensors and arrays. Ookami has the beginnings of such a machine with variable length vector registers. We want to explore how this helps in our analysis. In addition, our methods of experimentation are unique, in that that we view the computer as a computational laboratory, and run experiments similar to the way experiments are run in a physics laboratory, e.g. controlled and reproducible experiments.