OOKAMI PROJECT APPLICATION

Date: 09/08/2021

Project Title: Program Transformation for Automatic GPU-Offloading with OpenMP

Usage:

- Testbed

Principal Investigator: Tony Curtis

- University/Company/Institute: Stony Brook University
- Mailing address including country:

  Institute for Advanced Computational Science,
  Stony Brook University,
  Stony Brook,
  NY 11794-5250,
  USA

- Phone number: (631) 632-4629
- Email: anthony.curtis@stonybrook.edu

Names & Email of initial project users:

- Alok Mishra <alok.mishra@stonybrook.edu>
- Smeet Chheda <schheda@cs.stonybrook.edu>
- Abid M. Malik <amalik@bnl.gov>

Usage Description:

We are working on designing and developing a compiler framework that can automatically discover OpenMP kernels, recommend several potential OpenMP variants for offloading that kernel to a GPU, and using a novel static Neural Network-based compile-time cost model, to predict and return the optimal variant. We divide our framework into 3 modules, each of which functions independently.

- **Module 1** detects and analyzes an OpenMP kernel and suggests several variants, by applying various potential code level transformations, for offloading that kernel to a GPU.
- **Module 2** defines COMPOFF, a tool that statically estimates the Cost of OpenMP OFFloading using Neural Networks (for the first time in OpenMP). Our results on Seawulf show that using COMPOFF, one can predict offloading cost with an accuracy ranging from 95–99%.
- **Module 3** uses the analysis and prediction from the other modules to modify the source code and returns newly generated code that supports GPU offloading.

Our preliminary findings indicate that this framework can assist scientists and compiler developers in porting legacy HPC applications to new heterogeneous computing environments. We need to extend our framework to support various GPU architectures, *including the NVIDIA V100 GPUs used on Ookami.*

Computational Resources:

- Total node hours per year: estimate 1000
- Size (nodes) and duration (hours) for a typical batch job: runs can range from using a single GPU to all available GPUs on a single node. Runs often range from a few minutes to a a few hours (e.g. running large benchmark applications for data collection).
- Disk space (home, project, scratch): 40GB, 4TB, 4TB

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

None anticipated.

Required software:

Probably none extra.

If your research is supported by US federal agencies:

- Agency: Department of Energy
- Grant number(s): ECP (17-SC-20-SC)