

OOKAMI PROJECT APPLICATION

Date: October 17, 2022

Project Title: Parsing and Learning Algorithms for Computational Linguistics and Biology

Usage:

Testbed

Production

Principal Investigator: Jeffrey Heinz

University/Company/Institute: Stony Brook University

Mailing address including country: IACS, Stony Brook University, Stony Brook, NY 11794

Phone number: 310-254-5926

Email: jeffrey.heinz@stonybrook.edu

Names & Email of initial project users:

Jeffrey Heinz, jeffrey.heinz@stonybrook.edu

Logan Swanson, logan.swanson@stonybrook.edu

Usage Description:

Existing parsing and learning algorithms for computational linguistics and biology will be run on Ookami to examine any performance benefits. Specifically, we will examine the performance of a parsing algorithm for the HP model of protein folding (Hockenmaier 2007, Bahar et al. 2017) and the performance of a learning algorithm for phonotactic constraints (Chandlee et al. 2019, Rawski 2021). While the underlying mechanisms of these algorithms differ, they both involve searching a very large space and both can be vectorized to take advantage of parallel processing. Our goal is to optimize the existing algorithms and implementations for Ookami to how far these algorithms can scale.

Computational Resources:

Total node hours per year: 15,000

Size (nodes) and duration (hours) for a typical batch job:

Disk space (home, project, scratch): 1TB

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

Required software: Haskell Platform

If your research is supported by US federal agencies:

Agency:

Grant number(s):

Production projects:

Production projects should provide an additional 1-2 pages of documentation about how
(a) the code has been tuned to perform well on A64FX (ideally including benchmark data comparing performance with other architectures such as x86 or GPUs)

(b) it can make effective use of the key A64FX architectural features (notably SVE, the high-bandwidth memory, and NUMA characteristics)

(c) it can accomplish the scientific objectives within the available 32 Gbyte memory per node