Initial experiences with the Ookami A64FX testbed

Andrew Burford\textsuperscript{1}, Alan C. Calder\textsuperscript{1}, David Carlson\textsuperscript{1}, Barbara Chapman\textsuperscript{1}, Firat Coşkun\textsuperscript{1}, Tony Curtis\textsuperscript{1}, Catherine Feldman\textsuperscript{1}, Robert J. Harrison\textsuperscript{1}, Yan Kang\textsuperscript{1}, Benjamin Michalowicz\textsuperscript{1}, Eric Raut\textsuperscript{1}, Eva Siegmann\textsuperscript{1}, Daniel G. Wood\textsuperscript{1}, Robert L. Deleon\textsuperscript{2}, Mathew Jones\textsuperscript{2}, Nikolay A. Simakov\textsuperscript{2}, Joseph P. White\textsuperscript{2}, Dossay Oryspayev\textsuperscript{3}

\textsuperscript{1}Institute for Advanced Computational Science, USA
\textsuperscript{2}Center for Computational Research, USA
\textsuperscript{3}Brookhaven National Laboratory, USA
Ookami - 狼

• Ookami is Japanese for wolf

• A computer technology testbed supported by NSF

• Available for researchers worldwide

  (excluding ITAR prohibited countries & restricted parties on the EAR entity list)

• Usage is free for non-commercial and limited commercial purposes
What is Ookami

- 174 1.8Ghz A64FX compute nodes each with 32GB of high-bandwidth memory and a 512 GB SSD
  - Same as in currently fastest machine worldwide, Fugaku
  - First deployment outside Japan
  - HPE/Cray Apollo 80

- Ookami also includes:
  - 1 node with dual socket AMD Rome (128 cores) with 512 Gbyte memory
  - 2 nodes with dual socket Thunder X2 (64 cores) each with 256 Gbyte memory and 2 NVIDIA V100 GPU
  - Intel Sky Lake Processors (32 cores) with 192 Gbyte memory

- Delivers ~ 1.5M node hours per year
Fugaku #1
Fastest computer in the world

First machine to be fastest in all 5 major benchmarks:

- Green-500
- Top-500 – 415 PFLOP/s in double precision – nearly 3x Summit!
- HPCG
- HPL-AI
- Graph-500

https://www.r-ccs.riken.jp/en/fugaku
A64fx at a Glance

- ARM V8 64-bit
- 512-bit SVE
- 48 compute cores
- 4 NUMA regions
- 32 (4x8) GB HBM @ 1 TB/s
- PCIe 3 (+ Tofu-3) network
A64fx NUMA Node Architecture

- Supports high calculation performance and low power consumption
- Supports Scalable Vector Extensions (SVE)
- **4 Core Memory Groups (CMGs)**
  - 12 cores (13 in the FX1000)
  - 64KB L1$ per core
    - 256b cache line
  - 8MB L2$ shared between all cores
    - 256b cache line
  - Zero L3$
  - 8 GB HBM at 256GB/s


Diagram is the „1000“ chip. We have „700“ chip, i.e. no assistant cores and no Tofu interface
SVE
(Scalable Vector Extensions)

• Enables Vector Length Agnostic (VLA) programming
  • VLA enables portability, scalability, and optimization
  • The actual vector length is set by the CPU architect
    • Any multiple of 128 bits up to 2048 bits
    • May be dynamically reduced by the OS or hypervisor
  • Predicate-centric architecture
  • SVE was designed for HPC and can vectorize complex structures
    • Gather-load and scatter-store; horizontal reductions
    • SVE begins to tackle traditional barriers to auto-vectorization
  • Support from open source and commercial tools
2017 analysis of XSEDE workload revealed 86% of all jobs need less than 32 GB / node

These 86% of jobs correspond to 85% of the total XSEDE cpu-hour usage

“Programmability of a CPU, performance of a GPU”
Satoshi Matsuoka (Head of RIKEN, home of Fugaku)

- Blazing fast memory
- Easily accessed performance
- New technology path to exascale
What else

- CentOS 8 operating system
- DUO Authentication
- High-performance Lustre file system (~800TB of storage)
- Slurm workload manager
- Compilers: GNU, Arm, Cray, Nvidia, Fujitsu (soon)
- Continuous growing stack of preinstalled software
  - MPI implementations
  - Toolchains
  - Math libraries
  - Performance analysis & debugging: (arm Forge, Cray, GNU, TAU, ..)
Initial Experiences

- Most applications run out of the box
- Obtaining high performance is more complex
Minimod

- seismic modeling mini-app developed by Total
- extracts the stencil computation from a production seismic imaging application
- stencil is used to numerically solve the acoustic wave equation
- benchmark to test new and emerging hardware and programming models for geophysics applications
SWIM

- Part of the SPEC CPU2000 Benchmark suite
- weather forecasting benchmark (FORTRAN OpenMP)
- solves the shallow-water equations using finite differences
• Ookami is monitored with XDMoD
• XDMoD software modify to monitor A64FX-specific metrics
• application kernels are used to proactively monitor HPC resource performance by daily benchmarks
• goal is to see how the performance of benchmarks and real applications change as the compiler toolchains improve
Getting Accounts

• Submit a project request (templates on our website)
  • **Testbed:**
    • Porting and tuning software
    • Benchmarking
    • Limited production calculations to demonstrate capability
    • Significantly less than 15,000 node hours per year
    • First two project years
  
  • **Production:**
    • Less than 150K node hours per year
    • Lower priority during the first two project years

• **Requests must include:**
  Title, date, PI, usage description, computational resources, grant number (if funded)
Getting Accounts

• Getting access:
  • Create a project request and submit it through ticketing system:
    https://iacs.supportsystem.com/
  • Requests will be reviewed & published
  • If you are not affiliated to SBU: Fill a volunteer demographic form

https://www.stonybrook.edu/ookami/
Current Status

• ~ 30 testbed projects (USA & Europe)

• ~ 100 users

• Several trainings & webinars

• Talks about Ookami in this session:
  • Lessons Learned: An In-depth Look at Running FLASH on Ookami
    Alan C. Calder, Catherine Feldman, and Benjamin Michalowicz
  • Performance Engineering using SVE
    Robert J. Harrison
Get in Contact

• https://www.stonybrook.edu/ookami/

• Bi-weekly Hackathon
  • Tue 10am – noon EST
  • Thu 2pm – 4pm EST

• Slack Channel for users #OOKAMI

Acknowledgement:

• The whole Ookami team

• NSF (grant grant OAC 1927880)