minimod on Ookami

Eric Raut[1], Tony Curtis[1], Barbara Chapman[1], Robert Harrison[1], Eva Siegmann[1]
Jonathan Anderson[2], Mauricio Araya-Polo[2], Jie Meng[2]

[1] Institute for Advanced Computational Science, Stony Brook University
[2] TotalEnergies

Ookami - 狼

- A computer technology testbed supported by NSF (grant OAC 1927880)
- 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

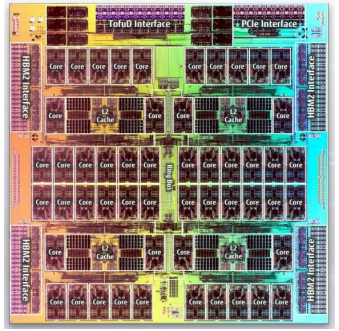
Stony Brook University

NSF

University at Buffalo
The State University of New York

What is Ookami

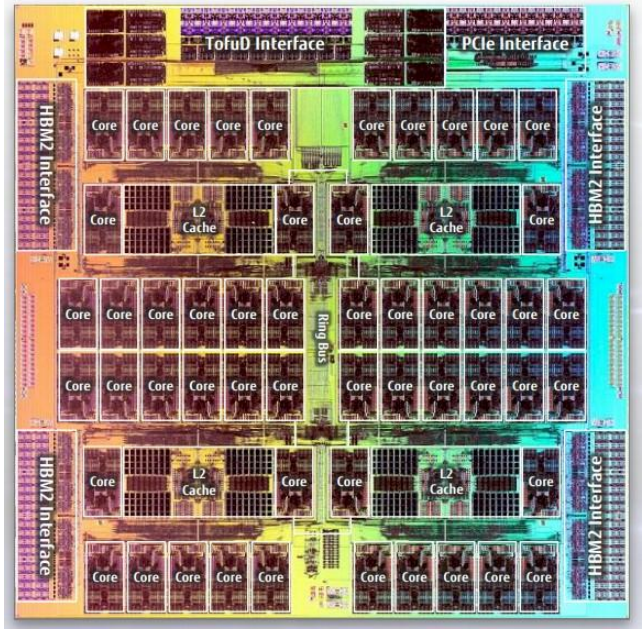
- 176 1.8Ghz **A64FX** compute nodes each with 32GB of high-bandwidth memory and a 512 GB SSD
  - Same as in currently fastest machine worldwide, Fugaku
  - First deployment outside Japan
  - HPE/Cray Apollo 80
- Ookami also includes:
  - 1 node with dual socket AMD Milan (64 cores) with 512 GB memory and 2 NVIDIA V100 GPUs
  - 2 nodes with dual socket Thunder X2 (64 cores) each with 256 GB memory
  - 1 node with dual socket Intel Skylake (36 cores) with 192 GB memory
- Delivers ~1.5M node hours per year



A64FX NUMA Node Architecture

OOKAMI

- 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



Fugaku #1
Fastest computer in the world

OOKAMI

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

- 432 racks
- 158,976 nodes
- 7,630,848 cores
- 440 PF/s dp (880 sp; 1,760 hp)
- 32 Gbyte memory per node
- 1 Tbyte/s memory bandwidth/node
- Tofu-2 interconnect

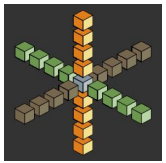
https://www.r-ccs.riken.jp/en/fugaku

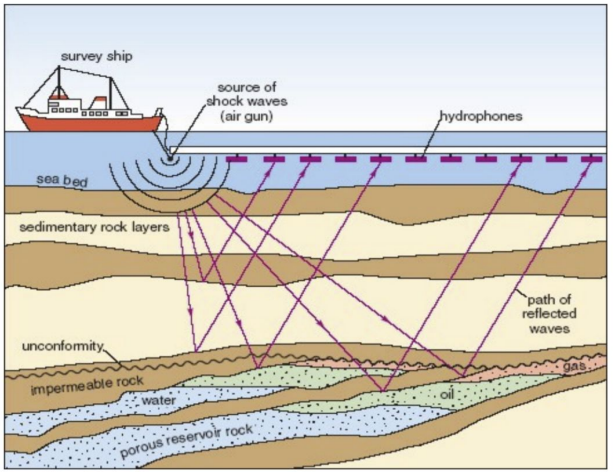
Minimod:
Geophysics Exploration

OOKAMI

- Wave equation important to many geophysics applications
- One simple solution method is finite difference (FD); involves stencil computation

$$\frac{1}{v_p^2}\frac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - \nabla^2 p(\mathbf{x},t) = f(\mathbf{x},t),$$





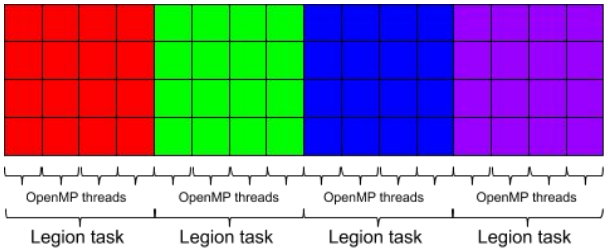
- *Minimod*: wave propagation mini-app developed by **TotalEnergies**
  - Extracts stencil computation from larger application
  - Designed to test new and emerging programming models and hardware

arXiv:2007.06048 [cs.DC]

Legion

- Data-centric task-based programming model
- Based on “logical regions” containing application data
- Automatically extracts parallelism
- Supports GPUs

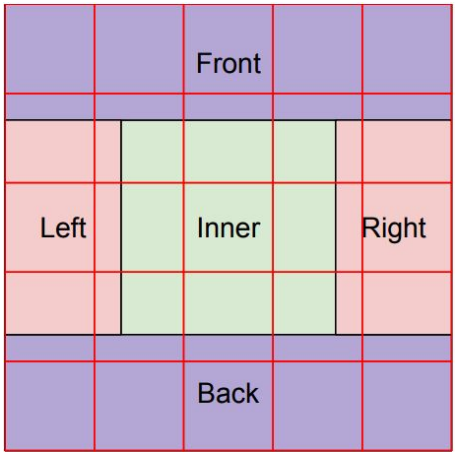
- Ported ASF to use Legion instead of MPI for inter-node parallelism



Minimod
Wavefield Solution


```
   Data: u^{n-1}, u^{n-2}: wavefields at previous two timsteps
   Result: u^n: wavefield at current timestep
 1 for i ← xmin to xmax do
 2     if i ≥ x3 and i ≤ x4 then
 3         for j ← ymin to ymax do
 4             if j ≥ y3 and j ≤ y4 then
 5                 // Bottom Damping (i, j, z1...z2)
 6                 // Inner Computation (i, j, z3...z4)
 7                 // Top Damping (i, j, z5...z6)
 8             else
 9                 // Back and Front Damping (i, j, zmin...zmax)
10             end
11         end
12     else
13         // Left and Right Damping (i, ymin...ymax, zmin...zmax)
14     end
15 end
```

Inherent load imbalance due to boundary conditions



Blocks contain both inner and boundary calculations

Raut et al., 2020

Experimental Setup

- We evaluated the weak scaling and strong scaling of ASF-Legion compared to ASF-MPI.
    - "Weak scaling": fixed grid size **per node**, increasing number of nodes
    - "Strong scaling": fixed **total** grid size, increasing number of nodes
- One MPI rank / Legion processor per NUMA region
- Default Legion mapper

System	Hardware Specs		Software
Ookami	CPU	Fujitsu A64FX	GCC 10.2.1
	CPU cores	12x4	OpenMPI 4.0.5
	Memory	32 GB	GASNet 2020.3.0
	L2	8 MB (x12 cores)	
	L1	64+64 KB	
	Lithography	7nm	
	Interconnect	InfiniBand HDR	
	TDP	160 W (full node)	

Table 1. Hardware and software configuration of the experimental platforms.
* IBM Spectrum MPI (SMPI)

Experiment 1: Weak Scaling

OOKAMI

Nodes	ASF-MPI [s]	ASF-Legion [s]	Parallel Efficiency [%] ASF-MPI, ASF-Legion
1	20.4	22.3	100.0% , 100.0%
2	20.7	24.0	99.0% , 93.0%
4	20.6	24.6	99.2% , 90.8%
8	20.7	25.7	98.7% , 87.0%
16	20.6	47.9	99.3% , 46.6%
32	21.1	-	97.1% , -

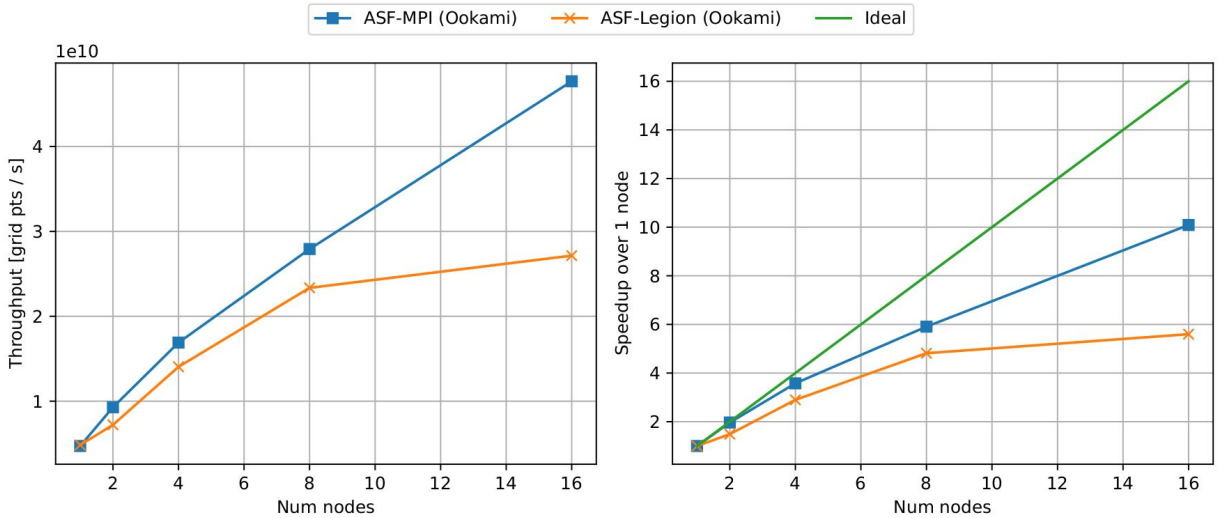
Table 3. Weak scaling: throughput (stable around the reference is ideal) with 1-32 nodes; grid size $1024^3/16$ per node; on Ookami.

- Grid size 1024^3/16 per node, with grid size increasing along the x-axis with more nodes
- Problems occur in Legion scaling beyond 8 nodes. 32-node run did not complete in time.

Raut et al., 2021

Experiment 2: Strong Scaling

Higher is better



- Compares throughput and speedup over one node for each method (grid size 1024^3)
- ASF-MPI throughput is competitive.
  - ASF-Legion suffers a performance penalty even with just a few nodes
  - Larger issue running at 16 nodes

Raut et al., 2021

Get in Contact

OOKAMI

- Available for researchers worldwide (usage is free)
- Active Slack channel
- Ticketing system
- Office hours twice a week
- Regular webinars about various tools

www.stonybrook.edu/ookami/

eva.siegmann@stonybrook.edu