NSF OAC Awards: 1927880 and 2137603

First Impressions of the NVIDIA Grace CPU Superchip and NVIDIA Grace Hopper Superchip for Scientific Workloads.

Nikolay Simakov*, Matthew Jones*, Thomas Furlani*,
Eva Siegmann + and Robert Harrison +

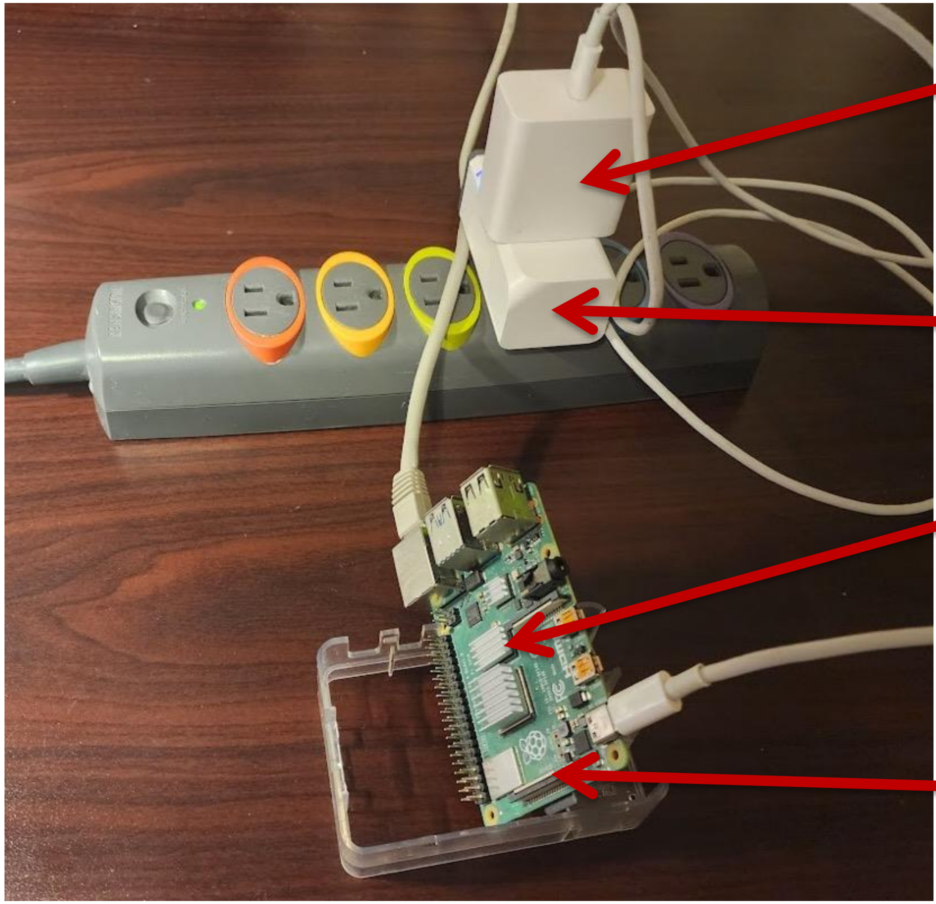
*Center for Computational Research, SUNY University at Buffalo
+Institute for Advanced Computational Science, Stony Brook University

University at Buffalo Center for Computational Research
Stony Brook University
iACS Institute for Advanced Computational Science

HPC-Asia 2024, Jan 25-27, 2024, Nagoya, Japan

International Workshop on Arm-based HPC: Practice and Experience (IWAHPCE-2024)

First Personal Experience with HPC Application on ARM



USB-C interface
provides enough power

Smart outlet provides
Power measurements

Raspberry Pi 4

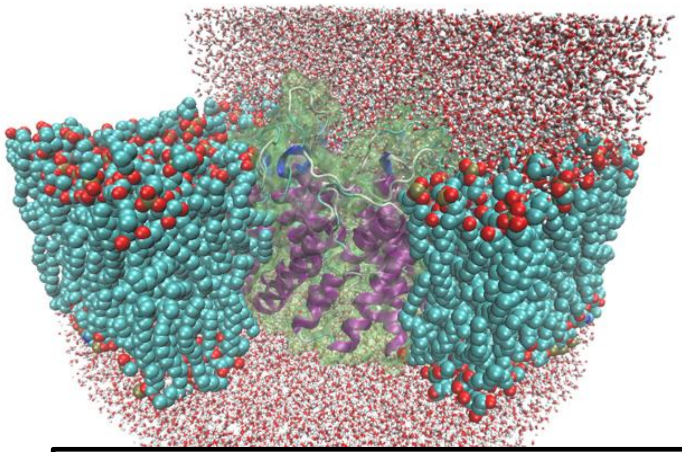
Vertical placement for
Efficient cooling

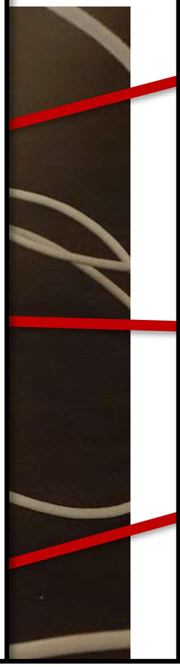
First Personal Experience with HPC Application on ARM

Gromacs

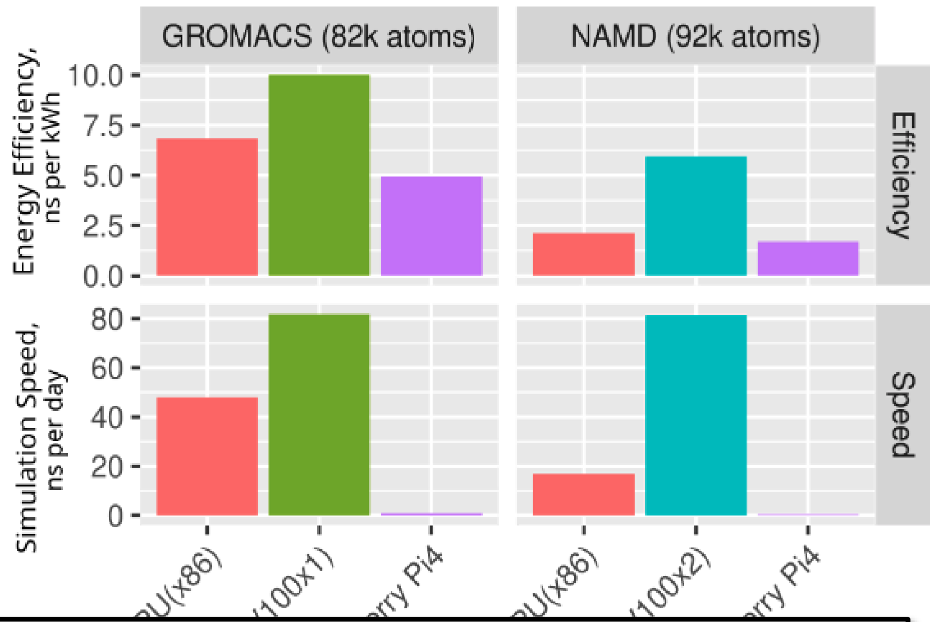
Membrane protein system

~82k atoms





Performance



nd
el
i4

Results:

Good: everything compiled and ran

Bad: Raspberry Pie 4 is neither fast nor energy efficient in compute intensive application like molecular dynamics

What is the performance state of modern high end ARM CPUs?

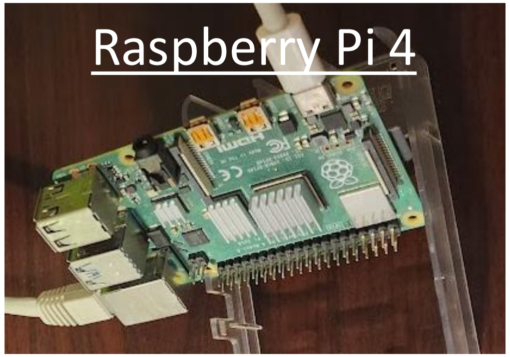
Outline

- Motivation and Introduction
  - Growing popularity and use of ARM-based systems in HPC
  - Energy efficiency is crucial for increasing IT and HPC demands

- Leveraging the XDMoD QoS/Application Kernel Technology for Benchmarking
- Results
  - Comparing ARM performance to x86 performance for range of applications
- Conclusions

XDMoD
METRICS ON DEMAND

Arm Processors are going to HPC Market

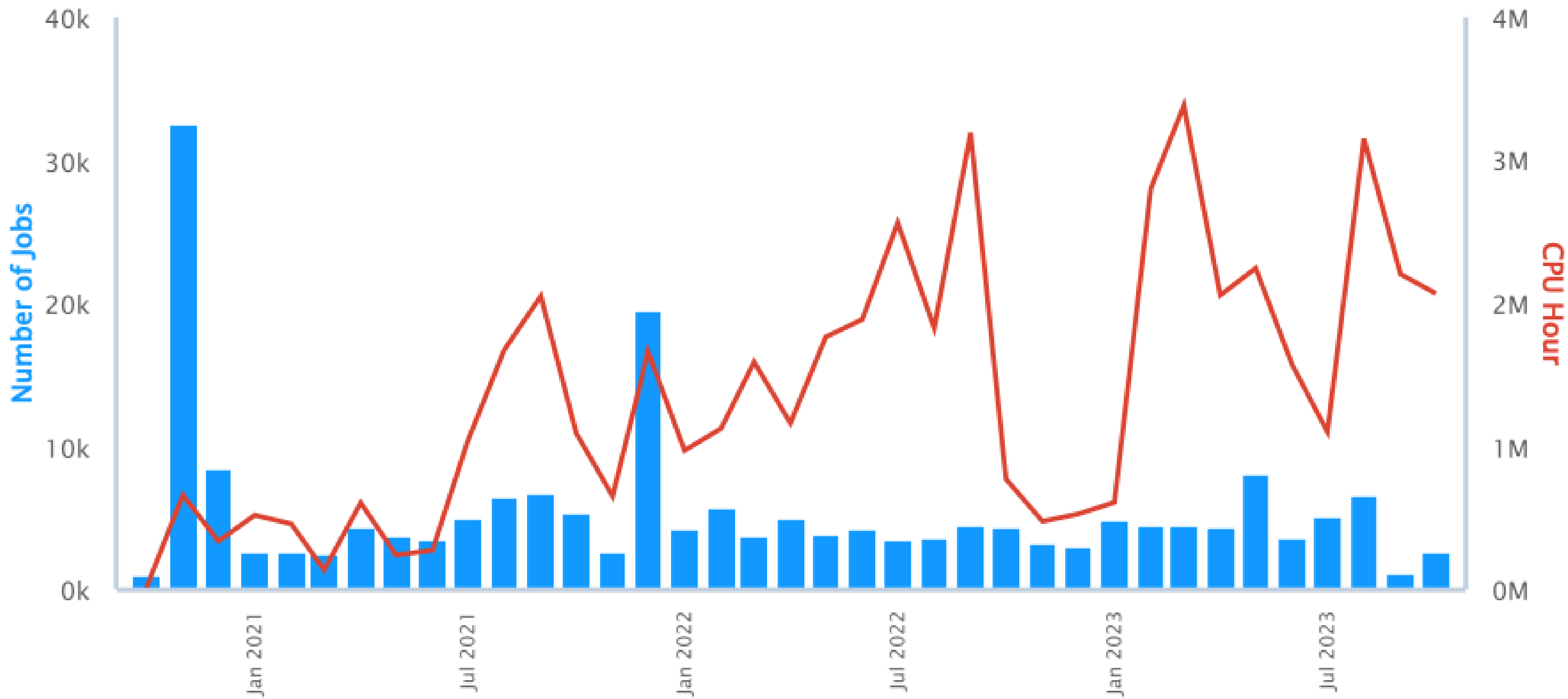


- **Energy efficiency is crucial for our ever-increasing demand for compute power**
  - Information-Communications-Technologies (ICT) ecosystem uses about 10% of world electricity generation. Projected to reach 20% by 2030*
  - Exascale computing is not sustainable without adequate energy efficiency. Frontier, 1.1 Eflop/s machine, consumes 21 MW (17,000 households).
- **Arm CPUs are successfully used in many products, including energy consumption-sensitive products.**
  - Embedded systems and mobile computing devices, like smartphones and tablets
  - Linux server products such as file and web servers
- **More recently, Arm CPUs have been adapted to HPC workloads, and some are specifically designed for scientific calculations**
- **What is their performance for HPC workloads?**
  - What is the performance state of modern high-end ARM CPUs?
  - How do they compare in performance to x86 systems
  - Are we ready for broader adoption of ARM in the HPC community?
- **NVIDIA recently released Grace CPU Superchip which has Arm**

*[https://www.nature.com/articles/d41586-018-06610-y]

Ookami Utilization (Arm Fujitsu A64FX)

Ookami – an Arm Fujitsu A64FX machine with SVE support (512 bit wide)

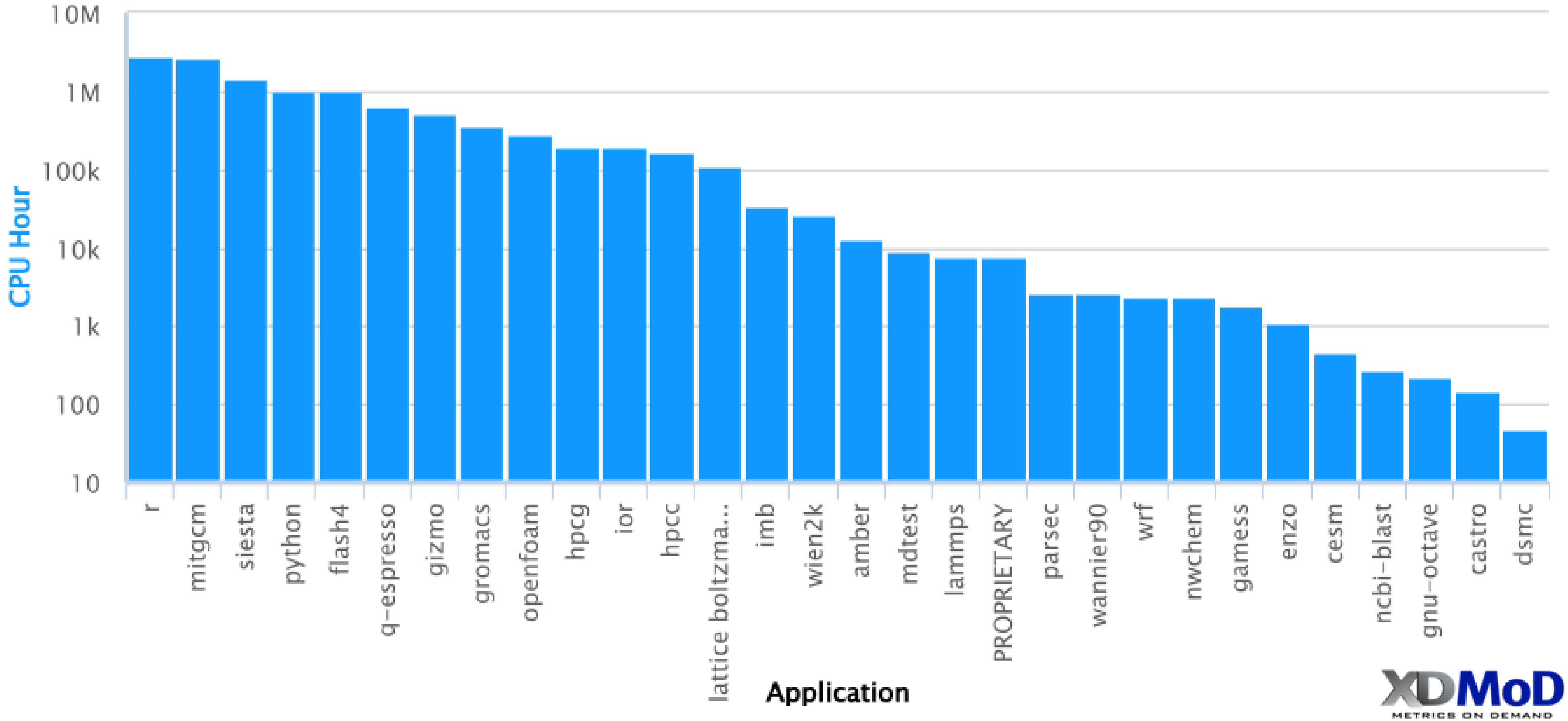


XDMoD
METRICS ON DEMAND

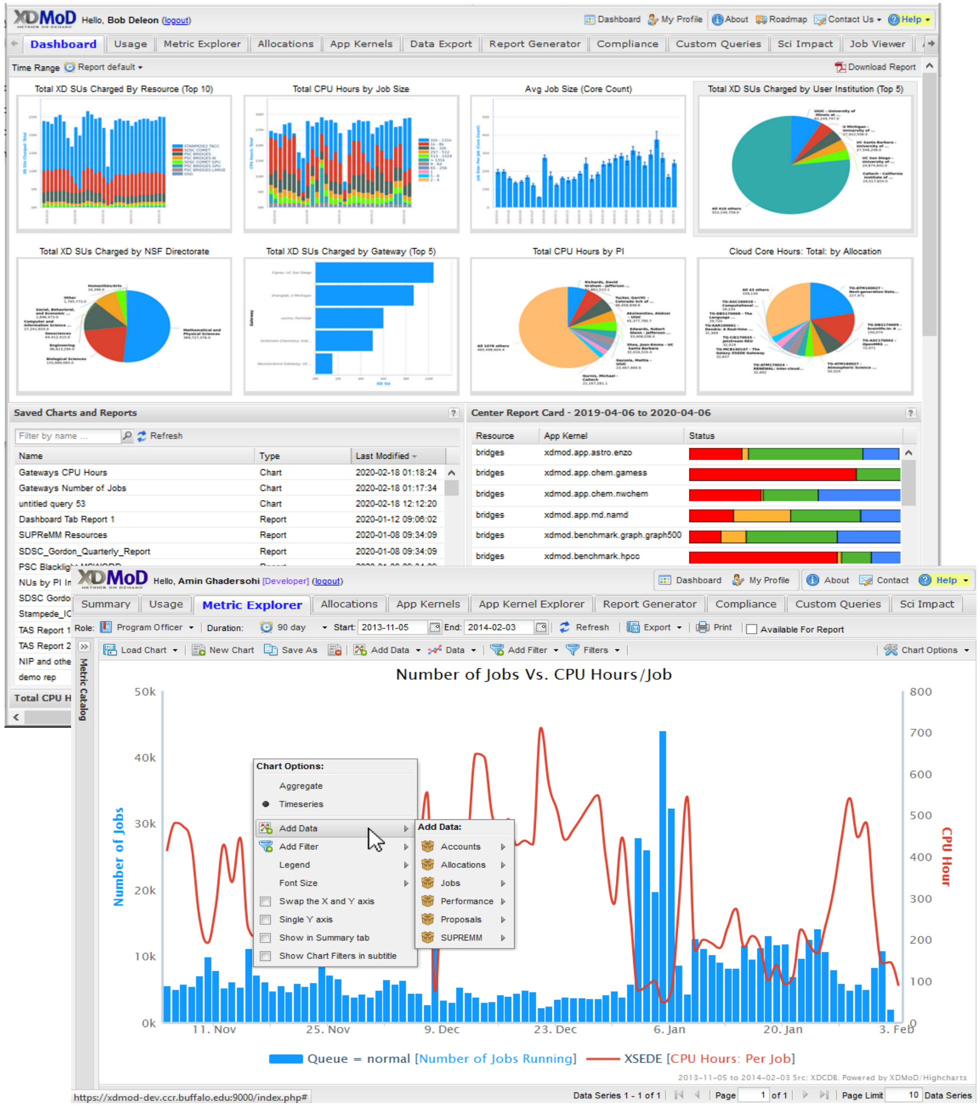
Application usage on Ookami (Arm Fujitsu A64FX)

Ookami – an Arm Fujitsu A64FX machine with SVE support (512 bit wide)

Determine what are the mostly widely used applications (2020-10-01 to 2023-10-31)



XDMoD: A Comprehensive Tool for HPC System Management

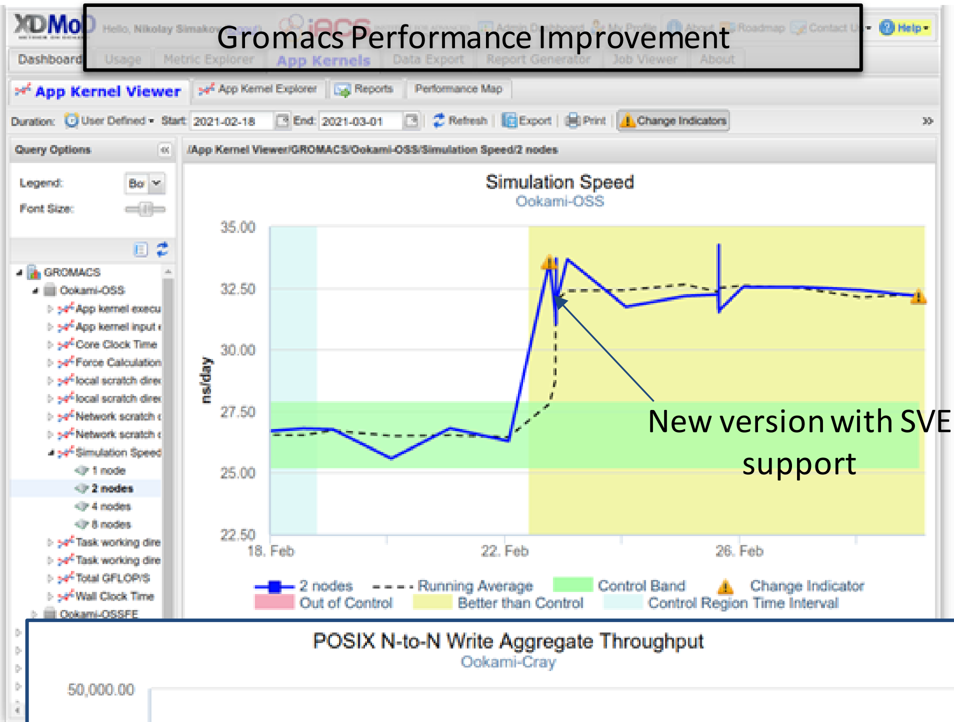


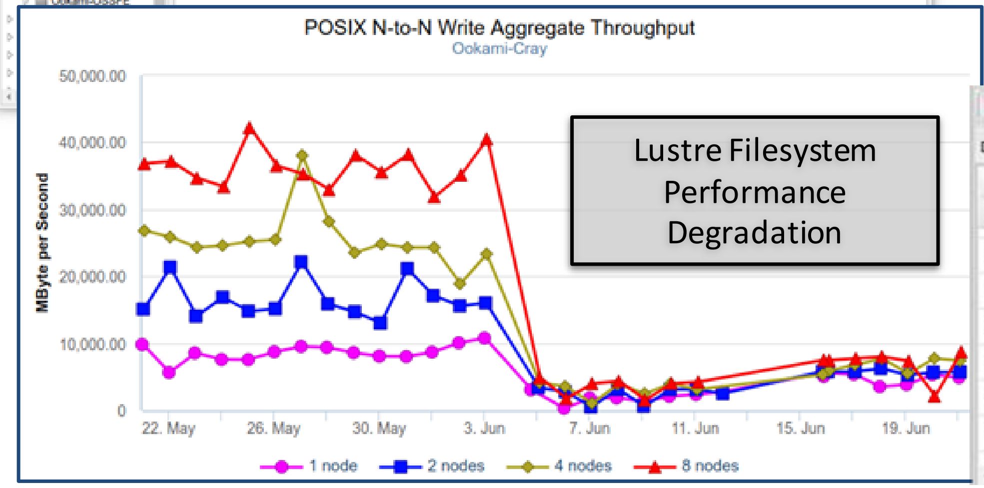
- **Goal: Optimize Resource Utilization and Performance**
  - Provide detailed information on utilization
  - Measure Quality of Service
  - Enable data driven upgrades and procurements
  - Measure and improve job and system level performance
- **NSF ACCESS Measurement and Metrics Service (MMS),**
  - Following XD Net Metrics Service (XMS) and prior 5 year TAS award
  - Develop & deploy **XDMoD** (**XD** **M**etrics **o**n **D**emand) Tool
- **Open XDMoD: Open Source version for Data Centers**
  - Used to measure and optimize performance of HPC centers
  - 300+ academic & industrial installations worldwide

XDMoD METRICS ON DEMAND

University at Buffalo Center for Computational Research

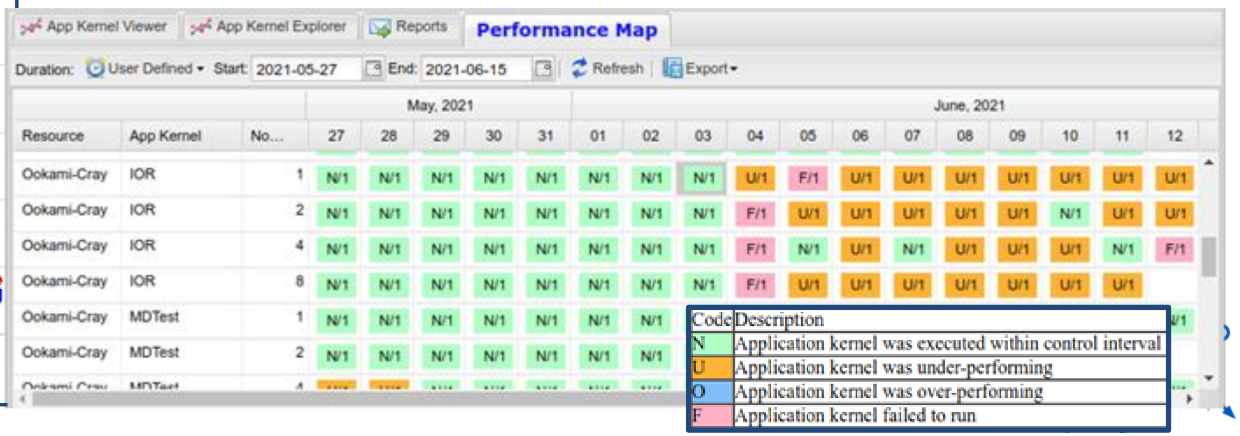
QoS and Performance Monitoring with Application Kernels



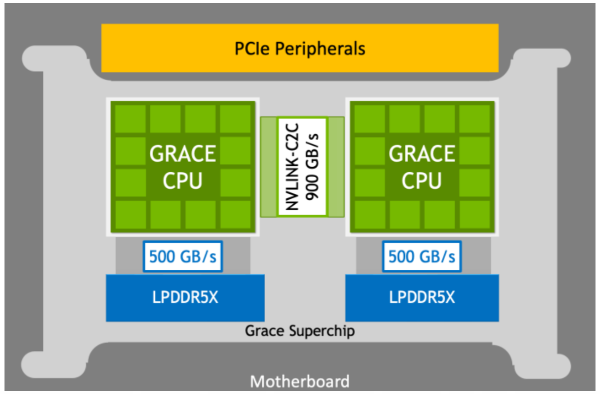


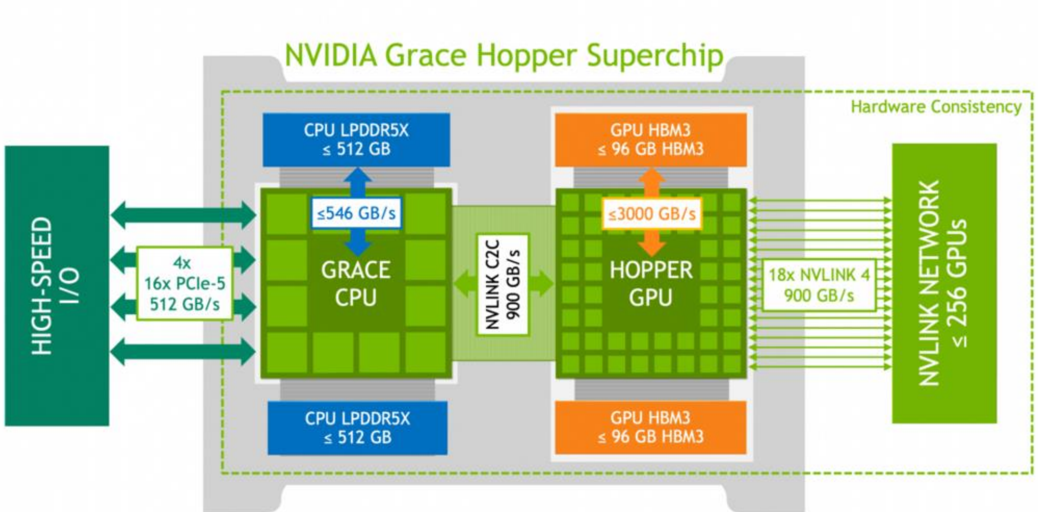
Application kernels module allows **continuous performance monitoring** by periodic execution of applications and benchmarks.

- Computationally lightweight benchmarks or applications
- Run periodically or on demand to actively measure performance
- Measure system performance from User's perspective
- Proactively identify underperforming hardware and software



NVIDIA Grace CPU Superchip and Grace-Hopper Superchip





- NVIDIA Grace CPU Superchip
  - Neoverse V2 Cores: Armv9 with 4x128b SVE2
  - 144 Cores (2 x 72 cores)
  - 32-channel LPDDR5X with ECC, Up to 1 TB/s
  - FP64 peak: 7.1 TFLOPS
- NVIDIA Grace-Hopper Superchip
  - Grace
    - 72x Arm Neoverse V2 cores
  - Hopper
    - 144 SMs and 3x higher FP32 and FP64 throughout compared to the NVIDIA A100 GPU.
    - 96 GB of HBM3 memory, up to 3000 GB/s

https://developer.nvidia.com/blog/nvidia-grace-cpu-superchip-architecture-in-depth/
https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/

Tested Compute Resources – CPU Only

Resource	CPU	CPU Arch/Core Name	SIMD ISA	SIMD Width, bits	# SIMD Units	Cores per Node	Freq, GHZ base/turbo	Memory
Arm								
SBU Ookami	Fujitsu A64FX	v8.2-A	SVE	512	2	48	1.8	HBM
Amazon-Graviton3-64	Amazon Graviton 3	v8.5, Neoverse V1	SVE	256	2	64	2.5	DDR5
NVIDIA Grace CPU Superchip	NVIDIA Grace	v9.0-A, Neoverse V2	SVE2	128	4	144	≥3.2	LPDDR5X
x86 AMD								
Purdue-Anvil	EPYC 7763	Zen3(Milan)	AVX2	256	2	128	2.45/3.5	DDR4
SBU-Milan	EPYC 7643	Zen3(Milan)	AVX2	256	2	96	2.3/3.6	DDR4
x86 Intel								
SBU-Skylake	Xeon Gold 6148	Skylake-X	AVX512	512	2	40	2.4/3.7	DDR4
TACC-Stampede 2 SKX	Xeon Platinum 8160	Skylake-X	AVX512	512	2	48	2.1/3.7	DDR4
TACC-Stampede 2 ICX	Xeon Platinum 8380	Ice Lake	AVX512	512	2	80	2.3/3.4	DDR4
SBU-SPR	Xeon Max 9468	Sapphire Rapids	AVX512	512	2	96	2.1/3.5	DDR5/HBM2e

Tested Compute Resources – CPU and GPU

Resource	CPU	GPU
vast.ai	AMD Ryzen 9 7950X (16 Cores Used)	NVIDIA RTX 4090
runpod	x86 (14 HT? Cores Used)}	NVIDIA RTX 6000 Ada
runpod	AMD EPYC 7773X (24 HT Cores Used)	NVIDIA L40
runpod	x86 (16 HT? Cores Used)	NVIDIA H100-PCE
google-g2-standard-16	Intel AVX512 Capable (16 HT Cores Used)	NVIDIA L4
amazon-g5.4xlarge	AMD EPYC 7R32 (16 HT Cores Used)	NVIDIA A10g
SBU-A100	Intel IceLake	NVIDIA A100
NVIDIA Grace Hopper Superchip	NVIDIA Grace	NVIDIA Hopper

Tested Applications and Benchmarks

- The application kernels used for this study span a variety of computational domains and paradigms:
  - HPCC – multiple benchmarks, including LINPACK and FFT
  - HPCG – High-Performance Conjugate Gradients
  - GROMACS - biomolecular simulation
  - Open Foam - partial differential equation solver
  - AI Benchmark Alpha - AI benchmark

- Other Applications (unused in this study but important for performance monitoring)
  - Intel MPI Benchmark - network
  - IOR and MDTest – parallel filesystem performanceGraph500
  - NWChem - ab initio chemistry
  - Enzo - adaptive mesh refinement

HPCC: HPC challenge benchmark

HPC Challenge Benchmark combine multiple benchmarks together

- High Performance LINPACK, which solves a linear system of equations and measures the floating-point performance
- Matrix-matrix multiplication
- Fast Fourier Transform
- Stream: memory bandwidth
- Parallel Matrix Transpose
- MPI Random Access

HPCC: HPC challenge benchmark

CPU/System	Cores	Matrix Multiplication GFLOPS	Matrix Multiplication GFLOPS/Core		LINPACK GFLOPS		LINPACK GFLOPS /Core	FFT GFLOPS		FFT GFLOPS/ Core
ARM Fujitsu A64FX, SVE 512b (SBU-Ookami, FJ)	48	1978	41.2	± 0.2	1177	± 19	24.5	24.4	± 0.9	0.51
ARM Amazon Graviton 3, Neoverse V1, SVE 256b (AWS)	64	1158	18.1	± 0.0	965	± 1	15.1	71.0	± 0.7	1.11
x86 AMD EPYC 7643 Zen3(Milan), AVX2 (SBU)	96	2775	28.9	± 0.9	1493	± 16	15.6	42.6	± 1.0	0.44
x86 AMD EPYC 7763 Zen3(Milan), AVX2 (Purdue Anvil)	128	3046	23.8	± 1.6	2176	± 100	17.0	54.7	± 4.8	0.43
x86 Intel Xeon Gold 6148, Skylake-X, AVX512 (SBU)	40	1559	39.0	± 8.1	981.22	± 109	24.5	33.4	± 2.4	0.84
x86 Intel Xeon Plat. 8160, Skylake-X, AVX512 (TACC-Stampede 2)	48	2122	44.2	± 1.7	1158	± 34	24.1	35.8	± 1.9	0.75
x86 Intel Xeon Plat. 8380, Ice Lake, AVX512 (TACC-Stampede 2)	80	3824	47.8	± 0.6	1713	± 5	21.4	76.4	± 2.0	0.96
x86 Intel Xeon Max 9468, Sapphire Rapids, DDR mode (SBU)	96	4787	49.9	± 2.7	2211	± 182	23.0	129.0	± 15.1	1.34
x86 Intel Xeon Max 9468, Sapphire Rapids, HBM mode (SBU)	96	5392	56.2	± 4.2	2862	± 36	29.8	143.1	± 24.4	1.49
NVIDIA Grace CPU Superchip ES, ARMPL	144	4089	28.4	± 0.1	3124	± 12	21.7	5.5	± 0.1	0.04
NVIDIA Grace CPU Superchip ES, OpenBLAS, FFTW	144	4461	31.0	± 0.1	3120	± 15	21.7	134.2	± 1.7	0.93

- **In Matrix multiplication and LINPACK wider SIMD has higher performance**
- **In Matrix multiplication and LINPACK wider SIMD NVIDIA Grace performed similar or better to AMD Millan in per core performance. Adding high core counts lead to higher per node performance in LINPACK**
- **For FFT per core performance of Grace is similar to Skylake-X and per node is between different memory modes for Sapphire Rapids**

HPCG: The High-Performance Conjugate Gradients

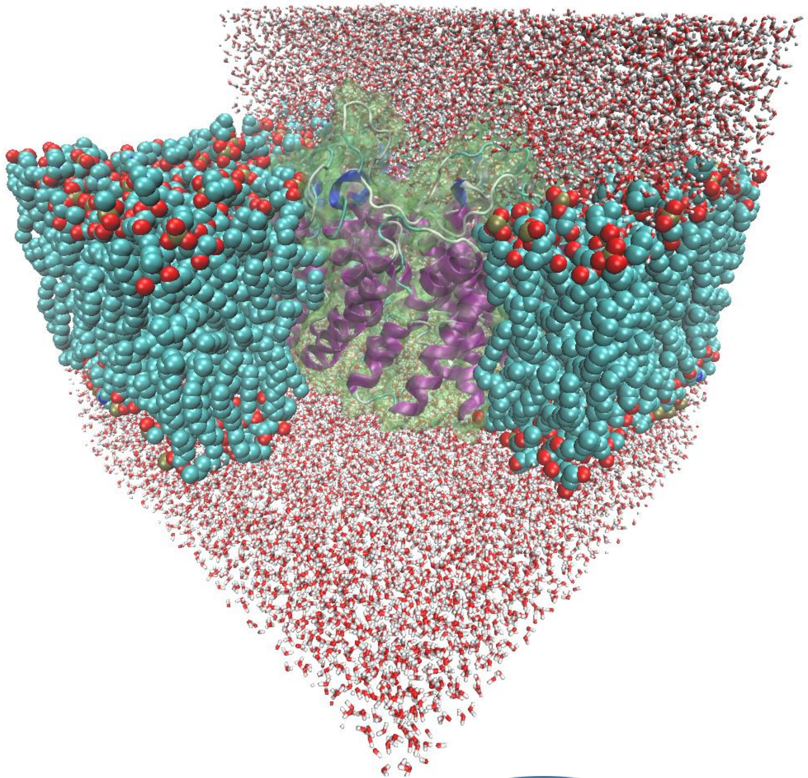
- The High-Performance Conjugate Gradients (HPCG) benchmark is an alternative to the HPL benchmark (used in HPCC) and utilizes methods and patterns commonly used in many PDE solvers
- Unlike HPCC, HPCG does not rely on external libraries but requires vendors to optimize their own version of HPCG.
- Thus for x86 machines, we used the Intel version of HPCG, for the A64FX Cray version and for AMD the reference version.

HPCG: The High-Performance Conjugate Gradients

CPU/System	Cores	HPCG Version	HPCG GFLOPS		HPCG GFLOPS/ Core
ARM Fujitsu A64FX, SVE 512b	48	Cray	64.4	± 2.8	1.34
x86 Intel Xeon Gold 6148, Skylake-X, AVX512 (SBU)	40	Intel	36.4	± 0.3	0.91
x86 AMD EPYC 7643 Zen3(Milan), AVX2 (SBU)	96	Intel	53.0	± 2.0	0.55
x86 Intel Xeon Max 9468, Sapphire Rapids, DDR mode (SBU)	96	Intel	83.6	± 1.1	0.87
x86 Intel Xeon Max 9468, Sapphire Rapids, HBM mode (SBU)	96	Intel	197.5	± 2.1	2.06
NVIDIA Grace CPU Superchip ES	144	Unoptimized	106.5	± 0.1	0.74

- **Even unoptimized, Grace show higher performance per core than AMD Millan and per node performance is higher than Sapphire Rapids in DDR mode**

GROMACS: Molecular Dynamics of Biomolecular Systems



FAST. FLEXIBLE. FREE.
GROMACS

GROMACS is molecular dynamics simulation of biomolecular systems

Application computational characteristics:

- Solve ODE (second Newton law)
- Particle interactions
  - Short range/long range
- FFT
- Has GPU acceleration, but at this time only one work efficiently

Test case:

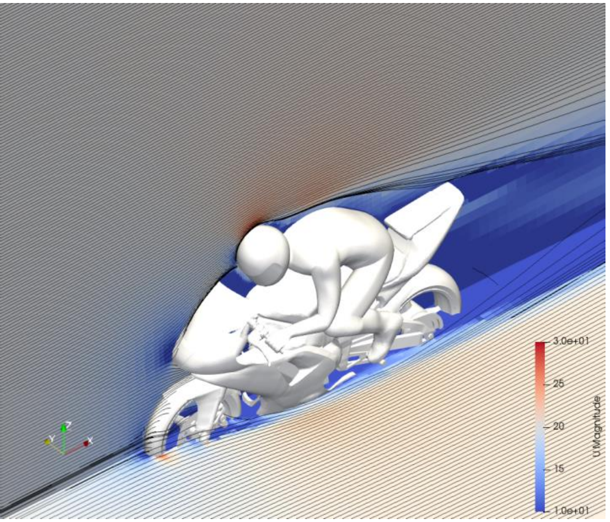
- Membrane protein
- 82k atoms system

GROMACS: Molecular Dynamics of Biomolecular Systems

CPU/GPU	Cores	Speed, ns/day (larger better) MEM 82K Atoms	RIB 2M Atoms	PEP 12M Atoms
CPU-GPU Calculations				
AMD Ryzen 9 7950X (16 Cores Used)/NVIDIA RTX 4090	16	284.82	13.85	3.82
x86 (14 HT Cores Used)/NVIDIA RTX 6000 Ada	14	245.40	19.30	2.37
AMD EPYC 7773X (24 HT Cores Used)/NVIDIA L40	24	160.23	15.09	2.53
x86 (16 HT Cores Used)/NVIDIA H100-PCE	16	183.92	15.81	2.88
Intel AVX512 Capable (16 HT Cores Used)/NVIDIA L4	16	142.04	8.90	0.98
AMD EPYC 7R32 (16 HT Cores Used)/NVIDIA A10g	16	160.53	8.55	1.76
Intel IceLake/NVIDIA A100	64	242.62	21.41	2.42
NVIDIA Grace Hopper Superchip ES	72	429	46.4	4.59
CPU Only Calculation				
ARM Fujitsu A64FX, SVE 512bit (SBU-Ookami, Fujitsu)	48	22.8		
ARM Amazon Graviton 3, Neoverse V1, SVE 256bit (AWS)	64	71.4		
x86 Intel Xeon Gold 6148, Skylake-X, AVX512 (SBU)	40	51.40	4.77	0.42
x86 AMD EPYC 7643 Zen3(Milan), AVX2 (SBU)	96	95.31	10.33	0.92
x86 Intel Xeon Max 9468, Sapphire Rapids, DDR mode (SBU)	96	203.64	13.88	1.18
x86 Intel Xeon Max 9468, Sapphire Rapids, HBM mode (SBU)	96	206.10	13.52	1.20
NVIDIA Grace CPU Superchip ES	144	171	12.7	0.977

- **Grace-Hopper shows outstanding GPU performance**
- **The Grace CPU is faster than Millan and 13-20% slower than Sapphire Rapids**

OpenFOAM: Toolbox for numerical solvers (CFD)



- OpenFOAM is a library and collection of applications for the numerical solution of PDE. Used often in computation fluid dynamics.
- Test case incompressible airflow around motorcycle
- Application computational characteristics:
    - Unstructured grid

OpenFOAM®

OpenFOAM

Resource	Cores	Run time, minutes, (smaller better)					
		Meshing		Solving		Total	
x86 Intel Xeon Gold 6148, Skylake-X, AVX512 (SBU)	40	11.47	±1.46	26.94	±0.34	39.45	±1.57
x86 AMD EPYC 7643 Zen3(Milan), AVX2 (SBU)	96	7.15	±0.89	14.98	±0.84	23.43	±0.59
x86 Intel Xeon Max 9468, Sapphire Rapids, DDR mode (SBU)	96	6.89	±0.43	9.90	±0.30	18.39	±0.67
x86 Intel Xeon Max 9468, Sapphire Rapids, HBM mode (SBU)	96	6.87	±0.67	6.42	±0.18	14.87	±0.72
NVIDIA Grace CPU Superchip ES	144	5.46	±0.01	7.11	±0.01	13.87	±0.00

- Grace CPU shows the highest performance

Regional Ocean Modeling System (ROMS)

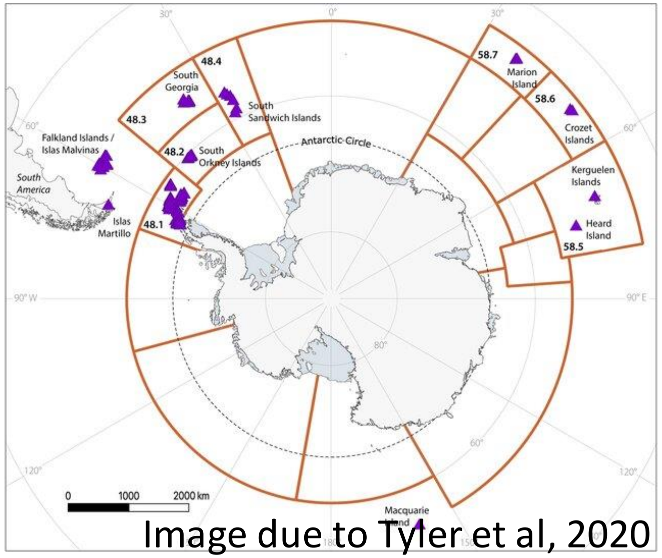


Image due to Tyler et al, 2020

Image due Liam Quinn

- Regional Ocean Modeling System (ROMS) is an ocean model widely used in the scientific community. It is a free-surface, terrain-following, primitive equations ocean model.
- The test case we use for this work simulates the flow around the west Antarctic Peninsula, important for well-being of gentoo penguins

Regional Ocean Modeling System (ROMS)

Resource	Cores	Nodes	Run time, minutes, (smaller better)
ARM Fujitsu A64FX, SVE 512b (SBU)	64	2	141.6
x86 AMD EPYC 7643 Zen3(Milan), AVX2 (SBU)	96	1	108.9
x86 Intel Xeon Max 9468, Sapphire Rapids, DDR mode (SBU)	96	1	57.6
x86 Intel Xeon Max 9468, Sapphire Rapids, HBM mode (SBU)	96	1	30.6
NVIDIA Grace CPU Superchip ES	144	1	46.0

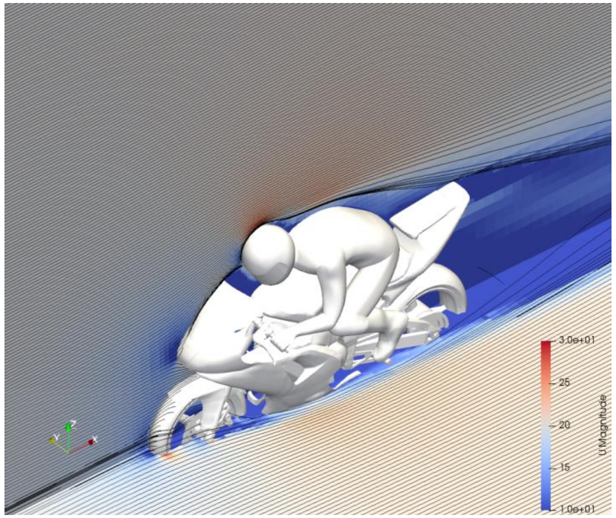
- Grace CPU performance is right between different memory modes of Sapphire Rapids

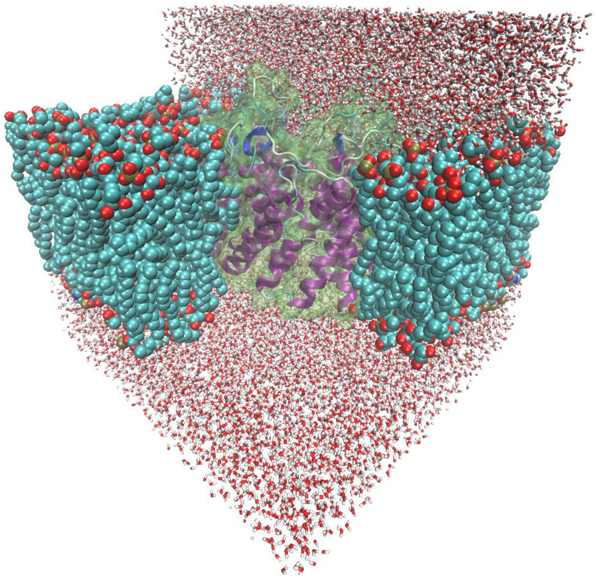
Thought on Energy efficiency

Performance in Gromacs	Cores	Simulation Speed, ns/day		Energy Efficiency, ns/kWh		Power, W	
MEM, 82K Atoms							
ARM Fujitsu A64FX	48	22.8	± 0.3 (10)	9.1	± 0.4 (10)	105	± 5 (10)
Intel Skylake	40	51.4	± 1.2 (10)	8.8	± 0.4 (9)	245	± 9 (9)
Intel Sapphire Rapids DDR	96	203.6	± 4.8 (22)	9.6	± 0.4 (11)	853	± 35 (11)
Intel Sapphire Rapids HBM	96	206.1	± 5.2 (10)	9.5	± 0.4 (10)	859	± 32 (10)
Intel IceLake/NVIDIA A100	54	236.5	± 10.8 (11)	13.9	± 0.8 (11)	707	± 9 (11)
RIB, 2M Atoms							
Intel Skylake	40	4.8	± 0.01 (8)	0.86	± 0.02 (7)	230	± 5 (7)
Intel Sapphire Rapids DDR	96	13.88	± 0.05 (10)	0.58	± 0.01 (10)	997	± 17 (10)
Intel Sapphire Rapids HBM	96	14.49	± 0.05 (10)	0.62	± 0.01 (10)	972	± 8 (10)
Intel IceLake/NVIDIA A100	64	21.41					

- In our previous study of Intel Sapphire Rapids (submitted for publication), the node-level power consumption during compute-intensive applications was almost **1kW**
- The reported TDP for the NVIDIA Grace CPU Superchip (two Grace SoC including memory) is capped at **500W.**
- Thus, given that in many tested situations, the NVIDIA Grace CPU Superchip performs comparably or better than Intel Sapphire Rapids, adopting the NVIDIA Grace CPU Superchip can significantly improve the energy efficiency of CPU only nodes, possibly over two times for certain applications.

Conclusions





- We have tested several engineering and quality samples of the NVIDIA Grace Hopper Superchip family.
- The application building process was not significantly different from the traditional x86 setup
- In numeric benchmarks, the per-core performance is similar or faster than AMD Milan CPUs and a higher core count often results in highest per node performance
- In scientific applications performance, the NVIDIA Grace CPU Superchip shows comparable (Gromacs), faster (OpenFOAM), or right between HBM and DDR mode of Intel Sapphire Rapids.
- The combined CPU-GPU performance of the NVIDIA Grace Hopper Superchip for Gromacs is significantly faster than any tested x86-NVIDIA GPU system
- Based on the specified TPD we expect to see significant performance-per-watt improvement for NVIDIA Grace CPU Superchip over x86 solutions, exceeding two times for certain applications

- **Overall we believe the new NVIDIA Grace Hopper Superchip and NVIDIA Grace CPU Superchip is a solid, high-performance solution for the HPC centers.**

Acknowledgements

NSF OAC Awards: 1927880 and 2137603

- This work is supported by the National Science Foundation under awards OAC 1927880 and 2137603.
- This work used compute resources at Stony Brook University, SUNY UB CCR, the XSEDE/ACCESS (CCR120014) and CloudBank .
- We want to thank NVIDIA for providing early access to NVIDIA Grace Hopper Superchip systems and Filippo Spiga, John Coyne and Ian Finder for their help in getting access to the system.

UB Slurm Simulator

Visit our poster at HPC-Asia-24
Slurm Simulator Development: Balancing Speed, Accuracy, and Maintainability.
Nikolay A. Simakov, Robert L. DeLeon

XDMoD METRICS ON DEMAND

XDMoD –HPC resources analytics and performance monitoring

University at Buffalo Center for Computational Research

Stony Brook University

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

AI-Benchmark-Alpha (Tensorflow)

Images are taken from https://ai-benchmark.com/alpha

- AI-Benchmark-Alpha includes multiple machine learning tasks utilizing deep neuron networks. Tests includes classification, image to image mapping, image segmentation, image inpainting, sentence sentiment analysis and text translation.
- It is relatively light-weight
- Utilize Tensorflow for computation

AI-Benchmark-Alpha (Tensorflow)

CPU/System	Cores	Larger Better AI Score	Larger Better Inference Score	Larger Better Training Score
CPU Only Calculation				
ARM Fujitsu A64FX, SVE 512b (SBU-Ookami)	48	1034 ± 3	535 ± 2	499 ± 2
ARM Amazon Graviton 3, Neoverse V1, SVE 256b (AWS)	64	4850 ± 31	2708 ± 21	2143 ± 13
x86 AMD EPYC 7763 Zen3(Milan), AVX2 (Purdue Anvil)	128	3079 ± 26	1992 ± 16	1087 ± 13
x86 Intel Xeon Plat. 8160, Skylake-X, AVX512 (TACC-Stampede 2)	48	3606 ± 20	2292 ± 18	1314 ± 4
x86 Intel Xeon Plat. 8380, Ice Lake, AVX512 (TACC-Stampede 2)	80	8805 ± 27	3725 ± 20	5081 ± 14
x86 Intel Xeon Gold 6130, Skylake-X, AVX512 (UB-HPC)	32	3233 ± 253	1941 ± 165	1292 ± 88
x86 Intel Xeon Gold 6330, Ice Lake, AVX512 (UB-HPC)	56	10197 ± 53	4398 ± 31	5799 ± 29
NVIDIA Grace ES, Single SoC	72	9004 ± 13	5613 ± 10	3391 ± 6
CPU-GPU Calculations				
x86 Intel Xeon Gold 6130, NVIDIA V100x2 (UB-HPC)	32	32628 ± 433	15656 ± 278	16972 ± 163
x86 Intel Xeon Gold 6330, NVIDIA A100x2 (UB-HPC)	56	59323 ± 378	29691 ± 290	29631 ± 152

- Grace CPU has performance similar to Intel Ice Lake