UCSD Triton Cache

Zhiheng Feng, Francisco Gutierrez, Aarush Mehrotra, Gauri Renjith, Shijie Wang, Zixian Wang
Flavia Tedjarutjianta, Ferrari Guan (support)
Mary Thomas[1], Bryan Chin[2] (Mentors)
Robert Sinkovits[1], Marty Kandes[1], Mahidar Tatineni[1], Christopher Irving[1], Andreas Goetz[1], Miro Hodak[3], Julio Maia[3], Ritoban Roy-Chowdhury[2],Danny Vo[2](Advisors)

San Diego Supercomputing Center[1], University of California, San Diego[2], Advanced Micro Devices[3]

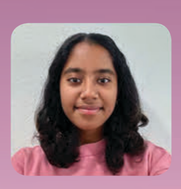
Francisco Gutierrez
Major: Computer Engineering
Role: Infrastructure, HPL, Repro, Captain
Year: Senior
Skills: Parallel Computing, Systems, Linux, Embedded
Has been leading the student efforts for HPC at UCSD for almost 2 years. Through the supercomputing club, he maintains infrastructure, organizes student competitions (SBCC), and is a project leader. SD Local too. 🚣

Zixian Wang
Major: Computer Science
Role: MLPerf, Infrastructure, Mystery App
Year: Senior
Skills: MLSys, LLM, NLP
Zixian led the team winning MLPerf last year. He is experienced in porting and deploying large-scale training infrastracture for ML models such as Mamba 8B, Llama 70B on AMD GPUs. Zixian is experienced in modifying Megatron-LM for customized training and inferencing Mamba 8B model as part of his research.

Shijie (Jacob) Wang
Major: Computer Science
Role: NAMD, MLPerf, Repro, Infrastructure
Year: Senior
Skills: ML, Parallel Computing, Linux
Shijie has experience modifying source code of ML models like stable diffusion XL for customized inferencing and porting MLPerf benchmark to AMD GPU. He also has a solid background in operating system and computer security from CTF.

Zhiheng (Henry) Feng
Major: Computer Engineering
Role: HPL, MLPerf, ICON
Year: Senior
Skills: HPC, Parallel Computing, ML, Linux
Zhiheng is familiar using Linux system and software for HPC like MPI, and he has experience running HPL and HPCG on cluster. Also, he is familiar with ML models and algorithms.

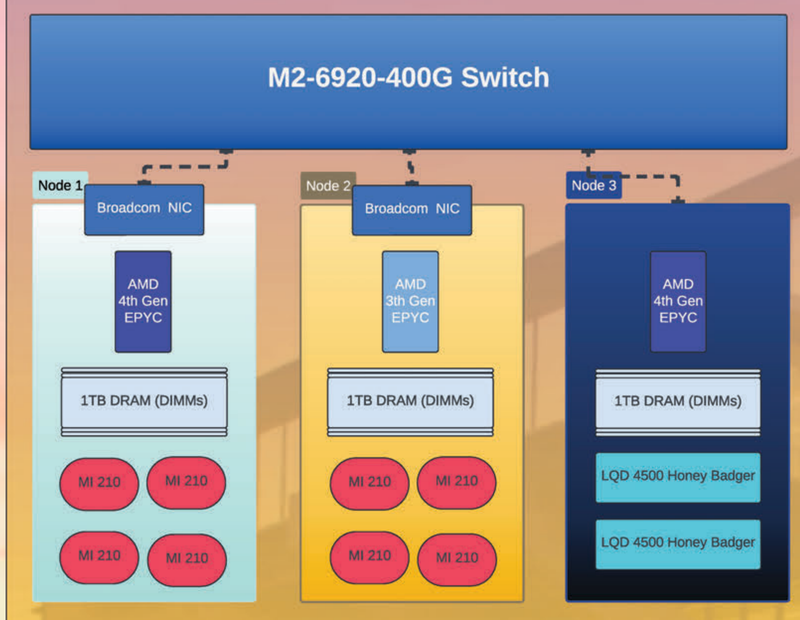
Aarush Mehrotra
Major: Math-CS & Economics
Role: HPL, MLPerf, ICON
Year: Junior
Skills: ML, HPC, Parallel Computing
Aarush has a deep background in data science and supported the SC23's team in porting the MLPerf Benchmark to AMD GPUs. With support from SDSC and UCSD resources, he has discovered a love for HPC and is looking forward to learning new things at this year's conference.



Gauri Renjith
Major: Computer Science with a Specialization in Bioinformatics
Role: NAMD, Repro
Year: Sophomore
Gauri has experience with using HPC for genomics research and using machine learning algorithms to analyze biological image data. She is excited for this opportunity to understand HPC at a deeper level and explore its scientific applications and use-cases for biological analyses.

DEI
✓**Diverse Applicant Pool**: Applications more than doubled since 2021; **36% of applicants** this year were from traditionally underrepresented groups
✓ **Major and minor in diverse disciplines:** Computer Science, Computer Engineering, Mathematics, Bioinformatics, Economics
✓ **Broad range of interest:** Computer Architecture Research, Data Science, Numerical methods and Networking, Deep Learning, High-Performance Networking, System Administration, Containerization, Web Development, User Interface Design, Biological Data Analysis
✓ **Speak Multiple Languages**: English, Mandarin, Cantonese, Hindi, Spanish, French

Hardware



CPU: AMD EPYCS Genoa/Milan,
GPU: CDNA 2 (MI210s), Non-heterogenous nodes
Storage: Liqid IO Honey Badger,
Networking: Broadcomm 400Gbps, MICAS Switch

Software Stack

Monitoring: Prometheus, Grafana
Deployment & Management: Ansible
OS: Rocky Linux
Storage: RAID0, NFS
Communication: OpenMPI, ROCE
Tools: GNU Tools, HIPcc, ROCm
Libraries: ROCm, OpenMP,

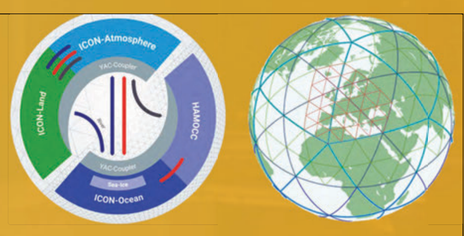
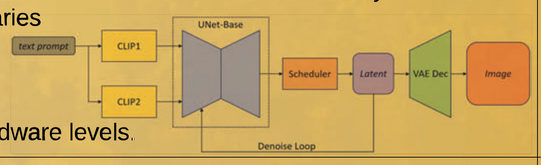
Python
Docker
Singlarity
Spack

How we prepared:
- The team is sponsored by AMD, including: Hardware & HPC Fund credits that provides the team access to AMD EPYC CPUs & INSTINCT GPUs. This enables the team to gain familarity with ROCm software support on the benchmarks & applications.
- The team attended SDSC summer institutes (CIML, and HPC/Data Science), in which students learned HPC concepts: scalable AI; running jobs on Expanse (consisting of multinode jobs using AMD EPYC and Nvidia V100 GPUs); utilizing specific programming interfaces; and frameworks such as CUDA or MPI.
- Cluster Management using the UCSD Supercomputing Club's test cluster in order to understand how a more barebones cluster and network setup would function.
- Strong mentoring: work with mentors on domain-specific knowledge; "home team" students work with team each year, and are now on the team this year.
- Regular meetings: advanced training sessions; progress syncs; work together on running, compiling and understanding each benchmark and application.

Aspiration of Our Team:
- Our team is comprised of enthusiastic individuals possessing an extensive range of expertise and experience across computer systems and in-depth knowledge of computer science.
- More than half of the team has past experience in HPC gained from both Supercomputing Training and the Supercomputing Club.

With a team full of strong and supportive mentors and advisors, we are confident that we can present creative works and bring out the best of our cluster.

Benchmarks, Applications, and Optimization Strategies

TOP 500 The List.	Required Libraries: MPI, UCX, ROCm, OpenBlas Building HPL: We adapted the build script from AMD Infinity Hub to suit our system's architecture and configuration. The script was modified to ensure compatibility with our environment and hardware setup. Running HPL: To achieve optimal performance, we followed the recommended parameter configurations from AMD Infinity Hub. For multi-node running, we need to test all the settings and find the best asymmetric parameters for our hardware setup.	ICON	We built ICON both using Spack for dependencies and building them ourselves to familiarize ourselves with the program and what it needs. This was done on AMD's HPC Fund and on personal systems. We will use sample input data from the ICON team for SCC24 to perform test runs and optimize the program for our hardware. 
MLPerf	The team is experienced with the MLPerf infrastructure's setup. This enables the team spending more time to optimize SDXL Model for the community. The team will be looking into different libraries and distributed, parallel algorithms for the most performance extracted from the system in both software and hardware levels. 	NAMD	With access to HPC Fund, a GPU cluster operated by AMD, we were able to run and test NAMD application in an environment similar to our competition hardware. We are benchmarking our build to understand the performance we can expect during the competition. We are excited to explore NAMD's GPU -resident mode that can scale well on our three nodes and fully utilize the computational resources of our cluster.
Time and Power Strategies	Keeping a Raspberry Pi to monitor and log power and data can alert the team to keep under the rules of the competition. While not using a scheduler to run, the number of applications and benchmarks at a time is low enough to coordinate between teammates. We hope to leverage the number of systems to better lower the latency beteween nodes, as well as the high network speed and the beefy GPU fitted systems.	**Reproducbility Challenge**	We plan to use DataLife tool to analyze the I/O traffic for our benchmarks and applications, which will help us optimize workflow by reducing I/O access and latency.