



HARDWARE AND SOFTWARE

- We are familiar with Clemson's Palmetto cluster, so we are configuring our cluster software stack to be nearly identical to it
- Cluster contains 6 compute nodes, 3 CPU-only and 3 GPU nodes
- Infiniband HDR is used to ensure low latency node to node communication
- DDR5/4800 RAM increases memory bandwidth for improved system responsiveness
- GPU acceleration powered by NVIDIA A100/H200 and OpenACC for CPU offload

| Hardware  | Specs                                               | Software         | Details                                          |
|-----------|-----------------------------------------------------|------------------|--------------------------------------------------|
| CPU nodes | Model: Three nodes<br>96 core AMD EPYC x 2          | System           | OS: Rocky Linux (8.6)<br>Kernel: 4.18.0-372.32.1 |
| GPU nodes | Model: Two nodes A100<br>40G x 4, One node H200 x 8 | OpenMP           | OpenMPI (4.1.4)                                  |
| Memory    | 256 GB DDR5<br>4800 MT/s                            | Compilers        | gcc (12.1.0)                                     |
| Storage   | 3.84 TB SSD                                         | GPU Acceleration | CUDA (12.0.1)                                    |
| NIC       | Infiniband HDR (200 Gb/s)<br>Mellanox ConnectX-6    | Network Stack    | Mellanox OFED<br>23.04-1.1.3.0                   |

TEAM INTRODUCTION

TechTigers brings Clemson students from different backgrounds, academic majors, and perspectives. Their interests and experiences bring diverse sets of skills to the team.



**Keerthi Surisetty**  
**Major:** Computer Science  
**Minor:** Artificial Intelligence  
**Application Lead:** SST  
**Bio:** Keerthi is a Data/AI Intern with Clemson's Energy Visualization and Analytics Center who has also interned at Citi and N.I.W.C. Atlantic. She leads Girls Who Code and serves as Industry Liaison for Clemson Forge.



**Jacob Davis**  
**Major:** Computer Science and Mathematical Sciences  
**Application Lead:** ECE  
**Bio:** Jacob develops HPC applications for NASA GSFC and takes graduate coursework in algorithms and GPU programming. He also worked on ML and EDA with VIPR-GS and Cadence.



**Ethan Prevuznak**  
**Major:** Computer Engineering  
**Application Lead:** MLPerf  
**Bio:** Ethan is the president of the Zeta-Iota chapter of Eta Kappa Nu and a member of IEEE. He interned at Argonne National Lab and takes graduate coursework in Machine Learning.



**Sam Quan**  
**Major:** Computer Engineering  
**Minor:** Physics  
**Application Lead:** Reproducibility  
**Bio:** Sam is a previous intern at Los Alamos National Lab in the Supercomputer Institute. He researches compression at Clemson and has won several quantum and startup competitions.



**Marcus McAbee**  
**Major:** Computer Engineering  
**Minor:** Mathematics  
**Application Lead:** Sys Admin  
**Bio:** Marcus has researched ML signal denoising, and data compression. He worked with Clemson's RCD team to benchmark local LLMs and is interested in OS and HPC system administration.



**Tolga Bilgis**  
**Major:** Computer Engineering  
**Minor:** Mathematical Sciences  
**Application Lead:** HPL  
**Bio:** Tolga is a sophomore who competed in the 2025 Single Board Cluster Competition as the HPL lead. He is a member of IEEE and is interested in large-scale AI/ML infrastructure.

BENCHMARKS

HPL – Tolga Bilgis

- HPL evaluates the performance of computers, particularly HPC clusters, by solving dense linear systems using double-precision arithmetic
- By building on the expertise gained from constructing miniature Raspberry Pi clusters, we can optimize our parallelization strategies since HPL was our main focus on these machines
- The A100 GPUs we opted for are based on the Ampere architecture, packed with high memory bandwidth, significant CUDA cores, and are designed for HPC workloads
- Nvidia provides many optimized HPC benchmark libraries such as BLAS which we will be able to leverage with our hardware



HPL-MxP – Marcus McAbee

- HPL-MxP aims to measure performance with the same approach as HPL, but evaluates the performance of a system with lower priority of correctness
- This is accomplished by reducing the computations to FP16, then refining the solution to double precision to retain accuracy. These operations mimic the behavior of ML software
- Lower precision computations make the benchmark faster, but also more energy efficient as FP16 computations consume less energy
- HPL-MxP can also utilize NVIDIA Tensor cores which are optimized for lower precision matrix operations, enabling massive throughput increases from standard HPL with our A100 nodes



MLPERF – Ethan Prevuznak

- MLPerf provides a standard of evaluation and performance for machine learning and artificial intelligence systems
- Measures how fast a system can process inputs and produce results using a trained model
- Using experience from prior submissions that utilized an AMD CPU, A100 GPUs, CUDA, and other similar hardware and software systems, we will work to optimize our systems to match those top results.



Llama2-70b

- Q&A model with input text. Measures both throughput and accuracy of response.

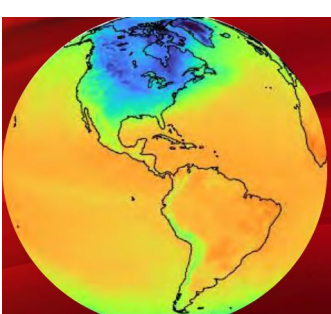
Power and Time Management

- We aim to implement various power management techniques to maximize our performance under the power constraints
- Tools and strategies such as DVFS and under-volting will allow for our system to run in a much more efficient manner
- We will also be utilizing NVIDIA SMI for precise GPU management and monitoring
- Our team will create a rough plan for running applications, as well as utilize Slurm job scheduler to balance workloads on our cluster and maintain perform within the power budget
- To fully utilize our system, we will build optimized CPU only and CPU+GPU builds for applications that support it

APPLICATIONS

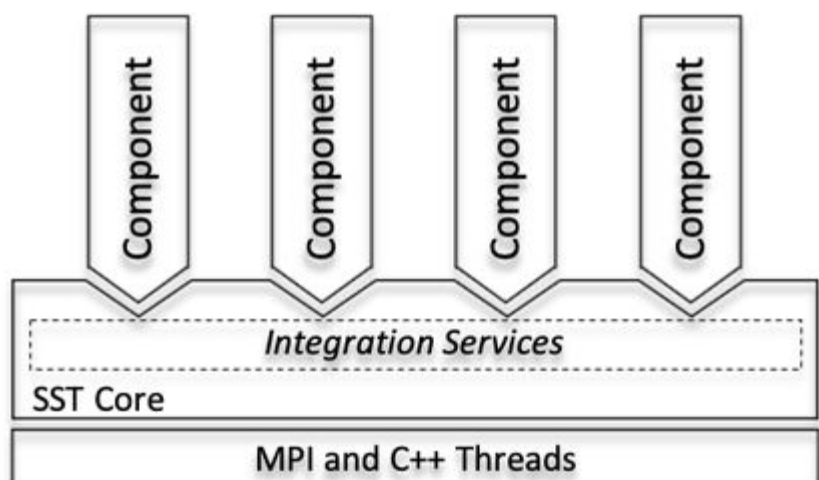
ECE – Jacob Davis

- ECE is a climate emulator that surpaasses existing climate emulators by a factor of 245,280
- Supports GPU acceleration via CUDA and parallel computation with MPI
- Cluster nodes have multiple Nvidia A100/H200 GPUs, so we focus on leveraging CUDA
- Splitting the workload across CPU and GPU allows the simulation to take advantage of all our cluster's resources
- We are focusing on leveraging the PaRSEC runtime system to optimally distribute mixed-precision tasks



SST – Keerthi Surisetty

- SST is a parallel discrete-event simulator used to model large-scale HPC platforms
- Our goal will be to determine the parallel efficiency when using multiple cores
- By testing for the ideal number of cores, we can achieve optimal performance of the application, decreasing computation and communication time across cores



Reproducibility - Sam Quan

- To successfully reproduce results from “Asynchronous Distributed-Memory Parallel Algorithms for Influence Maximization”, we will:
  - Deploy HCLib to run the novel parallel algorithms for the GenerateRR and SelectSeeds kernels
  - Recreate Strong Scaling results on Synthetic and Real graphs and scaling results of influencer counts and approximation ratio
  - Reproduce the memory access advantage of the GenerateRR actor method vs Ripples
  - Adapt scaling differences, memory distribution, and software environment of the thousand core systems of the paper vs our cluster

Why Will We Win?

- Team members meet weekly to discuss our competition strategies and progress on application optimization
- We attend webinars to learn more about the applications
- We utilized Clemson's Palmetto cluster to quickly iterate and test applications prior to our cluster's arrival
- Team members are enrolled in a course in which they run various HPC benchmarks

