ShanghaiTech University

School of Information Science and Technology

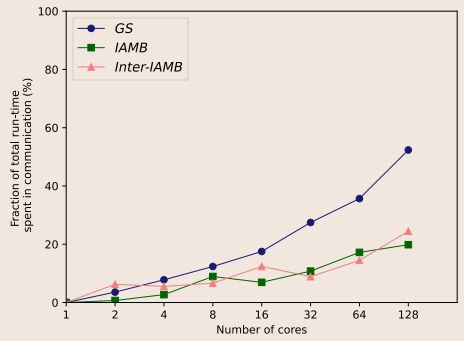
Application

Cardioid

- A **balance** between OpenMP threads and MPI processes for best scaling on cloud architecture
- Optimize for **communication** and **memory**.
- **Profiling** to optimize hot code.

RamBLE

- Deployed several **shell scripts** for compiling and running to save time and budget for teammates.
- Start new jobs **immediately** after old ones have finished reading files to save time for other tasks.
- The reproducibility challenge is not memory-bounded, therefore we can **run small instances simultaneously** to maximize budget efficiency.



??? Mystery

- **Analysis** whether a CPU-intensive application or a GPU-intensive application.
- **Flexible deployment** with different compilers and library using Spack.
- **Reserved resources**. utilize the MIG technology to isolate memory to better run apps simultaneously.

Quantum Espresso

(Plain Wave Self-consistency Field)

- **Tuned parameter** of nk and nd for k point replication and FFT divition for better performance.
- Reduce the temporary file output and utilize the **molecule symmetric** to modify the input file.
- Tried to migrate the Cufft to **VkFFT** and tried to make the code heterogeneous by **data deviding**.

HPCG & HPL

- **Auto tuning** the parameter utilizing Bayesian Optimization Methods.
- Take advantage of NVLink on A100 card and **affine core tailor** to this topology.
- Use **As More As Possible A100 nodes** bounded by teams' races and budgets.

IO500

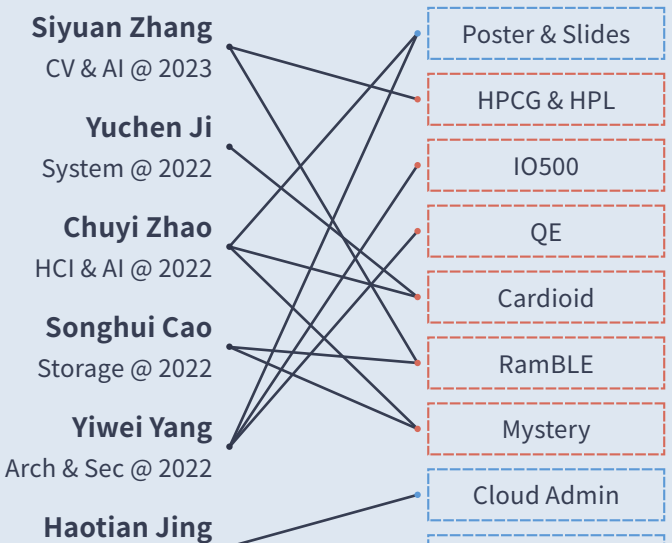
- According to MadFS's idea, we implemented **BadFS** - yet another highly scalable distributed file system for full-speed metadata read.
- Compared to MadFS
  - Designed a **lockfree hashmap** to store data compared to RocksDB with huge write amplication.
  - Write combing the **small data** and store them in the **Memory**.
  - Run on **high IOPS** storage with RAID0 if possible and **high bandwidth** Ethernet cards.

Team

"Diversity + Collaboration."

A diverse team in a diverse school.

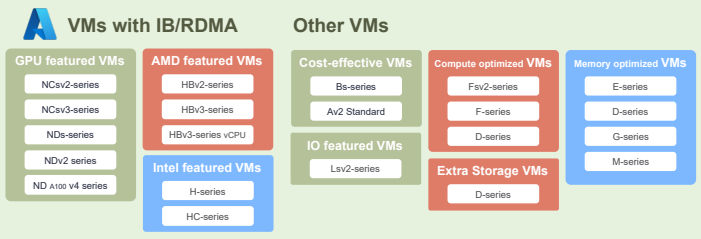
- **Gender diversity**, our most favorable female team member, frequent female STEM activity.
- **Vulnerable Groups**, open all our machines to general students, support students with financial aid.
- **Originality diversity**, come from different provinces, we visited one team member's birthplace in rural place as social practice.



We will fight together against COVID-19!

Cloud

Scalable Azure SKU choice



Cloud Arch Spec

Compute-bound Chassis

- **HB120rs_v2** for data computation and transmission node every dollar.
  - 120 7742 vCores on 2 vSockets.
  - 200Gbps Mellanox HDR InfiniBand Card.
- **E64_v4** for Intel CPU-optimized Applications.
  - 64 8272CL Intel vCores on 2 vSockets.
  - 30 Gbps Ethernet Cards.

Storage optimized Chassis

- **L16s_v2** for IO500 based on BadFS.
  - 160 GiB SSD and 2x 1920 GiB NVMe that can fully utilize the PCIe3.0x128 Lane and SATA Lane.
  - 6.4 Gbps Ethernet Card.
- **HB60rs**
  - We couldn't neglect the Test result and the bandwidth provided by both memory and IB Card.
- **Edsv5**
  - 3rd Gen Intel Xeon Platinum 8370C (Ice Lake).

RDMA-enabled Chassis

- **E64as_v4**
  - 1024 GiB SSD that can be placed as backup
  - for BadFS client Node.
- **H16r** for short time testing with not bad computing power every dollar.
  - 16 Haswell Intel vCores on 1 vSocket.
  - 65Gbps Mellanox FDR InfiniBand.
- **HC44rs**
  - 44 8168 Intel vCores on 2 vSockets.
  - 100Gbps Mellanox FDR InfiniBand Card.

GPU-enhanced Chassis

- **ND96asr_v4** - 8xA100 with NVLink
  - Performance Monster and also Budget Drainer.
- **NC24s_v3** - 4xV100
  - Second-class Performance Monster and Budget Drainer.

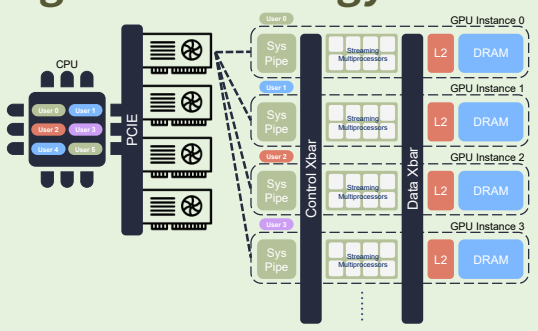
Cheap Chassis

- **DS13 v2** Best Configured for BeeGFS Cluster
  - 8 core 64GiB RAM 112GiB Disk.
- **E8s v3** Best Configured for NFS
  - 8 core 64GiB Disk.

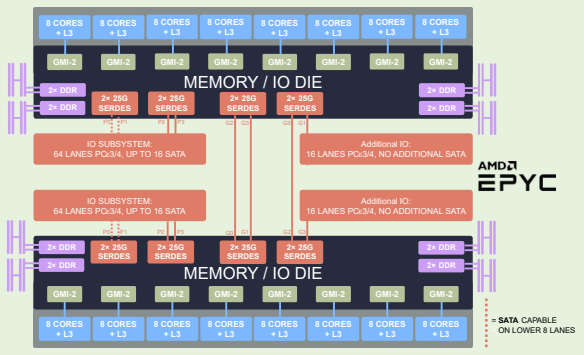
Oracle SKU choice

- **BM.GPU4.8** - A100 with NVLink for GPU-bound chassis.
- **BM.HPC2.36** for compute bound chassis.

Management strategy

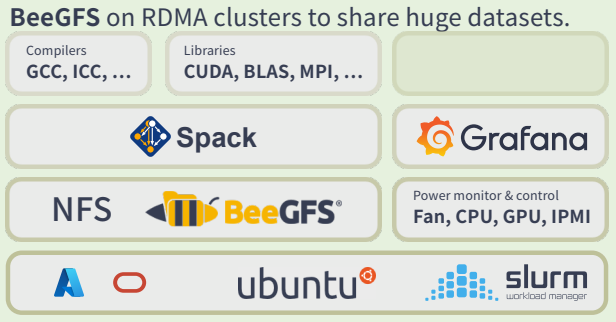


Cloud Optimized GPU



Cloud Optimized CPU

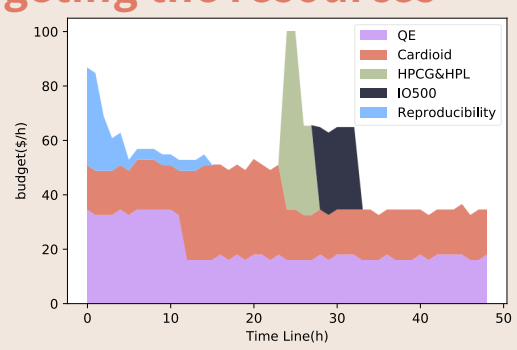
- **Cyclecloud 8.0.2**, with pre-configured slurm/BeeGFS template and supports Hyper-V virtualization.
- **Ubuntu 20.04** with tuned kernel across all chassis installed by cloud-init, different hardware is auto-configured on instantiation.
- **Spack** for flexible deployment on different OS and Arch.
  - GCC 11.2.0 / Intel LLVM Compiler 2021.3.0 / AMD AOCC Compiler 3.1.0 / NVHPC 21.9
    MKL 2021.3.0 / CudaToolkit 11.4/ AOCL 3.1.0 (libflames, amdnetlib)
  - NSight System / Compute, Arm Forge, Intel Vtune Profiler
- **NFS** on Cheap Chassis to share installed package.
- **BeeGFS** on RDMA clusters to share huge datasets.



Cloud Optimized Software

Preparation

Budgeting the resources



- Give **as many budgets as possible** for applications to run. QE and Cardioid, according to last year's experience, may last for the whole competition.
- Give **the rest for the benchmark**. Identify whether the tasks are feasible within the cost, once the budgets cut off or deviate from the graph on the left, we would give up the task.
- Give **high priority to IO500**, so that we may top the board.

Strict Panel Control

- **Alert the slack** on Slurm state change.
- Set a **long-term and instant budget ceiling** for Grafana panel.
- **Terminate the VM** once it turns idle.

Supports & Sponsors

上海科技大学 ShanghaiTech University
信息科学与技术学院 School of Information Science and Technology

ORACLE

Mellanox TECHNOLOGIES

NVIDIA

AMD