Team Peking University

Chun Fan (advisor), Zhenxin Fu (advisor), Yueyang Pan, Zhewen Hao, Ruihan Li, Junyi Guo, Jiaqi Si, Wenyang He
{chunfan, fuzhenxin, haozhewen, pyyjason, lrh2000, jeremyguo, sigongzi, hwy}@pku.edu.cn

About Peking University

Peking University (PKU) is a major research university in China. Founded in 1898, Peking University has always been the pioneer of novation and improvement, playing a significant role “at the center of intellectual movement” in China. Peking University ranks as one of top academic institutions in China, Asia and worldwide. Here, top Chinese students seek opportunities for academic excellence and dedication to the society.

HPC at Peking University

- Peking University launched its first high performance computing university-level public platform in 2018.
- Supercomputing cluster "Weiming no. 1", “Weiming Teaching no.1”, “Weiming Teaching no.2” and “Weiming Life Science no. 1" are available for students and faculties now.
- The total number of computing cores on the public platform reached 11620, and the peak capacity exceeds **999TFLOPS** with huge storage > 9649.6TB.

Team Members

We have seniors and juniors in our team, all obsessed with supercomputing.

PAN Yueyang *Team leader*
Senior, School of EECS
Major: Turing Class
Skills: System Building
Focus: Mystery & cluster management

HAO Zhewen
Senior, School of EECS
Major: Computer Science
Skills: CUDA, System building
Focus: HPL, HPCG, IO500 & cluster management

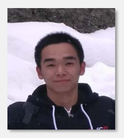
SI Jiaqi
Senior, School of EECS
Major: Computer Science
Skills: Algorithms
Focus: Reproducibility task

GUO Junyi
Senior, School of EECS
Major: Computer Science
Skills: PyTorch, C++, python
Focus: Cardioid

LI Ruihan
Junior, School of EECS
Major: Computer Science
Skills: Algorithms, Parallel Programming
Focus: Quantum

HE Wenyang
Junior, School of EECS
Major: Turing Class
Skills: ACM/ICPC, Algorithms
Focus: Quantum

Prof. FAN Chun *Advisor*
Research Interest: High Performance Computing, Massive cluster management
Publication in Top-tier journal
Position: The chief director of the system management sector in the PKU Computing Center.



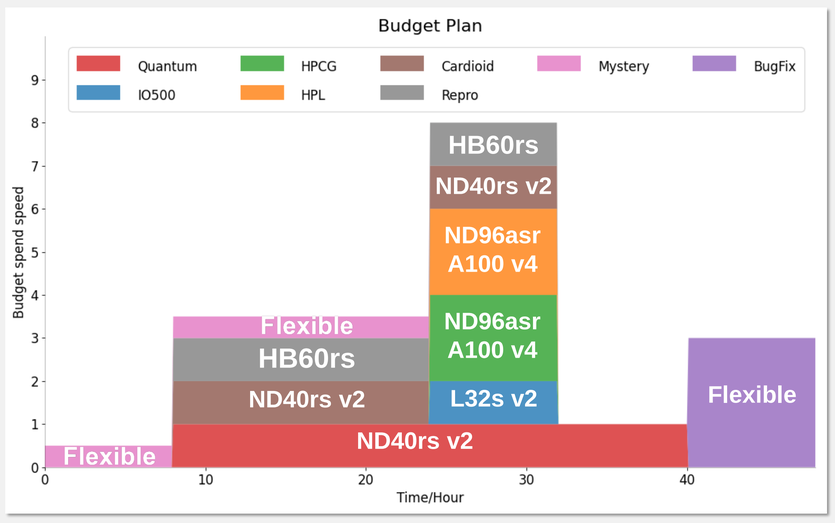
FU Zhenxin *Advisor*
Research Interest: Natural Language Processing.
Publication in Top-tier conferences: AAAI, IJCAI, CIKM, EMNLP, etc.
Position: Junior Engineer of the system management sector

Team Diversity

- **Diversity of the gender**. We provide a platform for both male and female to show their scientific talents and actively encourage them to cultivate interest in computer systems, computer architecture and HPC areas. Our mentors also provide them with constructive suggestions about the career development.
- **Diversity of the growing environment.** Our members come from places with different development levels, from the bustling metropolitan Shanghai and the idyllic county in Sichuan.
- **Diversity of the birthplaces**. The birthplaces of our team members cover both eastern and western China.

Resource and Administration

Budget Plan



Management tools and strategy

- Use **Slurm** for job scheduling and hide configuration details for end users.
- Use **Spack** to manage packages of multiple versions, configurations, and compilers coexisting on the cluster.
- Leverage the **Azure Cyclecloud** for resource management and budget alarming
- Use fast **NVMe** disks for consistent data storage during the contest.
- Schedule benchmark **in the middle** of the contest to adjust the resource usage based on the previous budget consumption.
- Deal with the mystery application **at the beginning** to provision a more detailed resource plan as well as to leave time for analyzing and optimization.
- Leave some **flexible budget** wisely for the occasions when the application takes an unexpected long time to debug and run.

Hardware and Software Configuration

Centralized Management

Spack

slurm workload manager

Slurm for job scheduling. **Spack** for package management. **Cyclecloud** for VM allocation and budget alarming

Possible SKU choices and reasons

- Use **ND A100 v4 series** with advanced NVLink interconnects for HPL&HPCG. Also both the Cardioid and QE could benefit from the computation capability of the most advanced A100 GPUs for acceleration.
- Use **HB-series** because AMD-based CPU VMs have high cost efficiency and provide enough computational cores for the Repro task
- Use **Lsv2-series** with fast NVMe storage for IO500 benchmark. We build our storage server with a replication for safety in the competition.
- Use **HC-series** with powerful Intel Xeon CPUs which are capable of dense computation. This prepares for the mystery application if it couldn’t be GPU boosted.
- **Infiniband connection**. Since the latency of communication is becoming a dominating factor of performance, we adopt IB to maximize the interconnect speed and to optimize memory-bounded applications.

Software Configuration

- CycleCloud CentOS 7
  - Opensource and free Linux distribution that features enterprise-level stability. Wide suits for a variety of platforms and HPC-related software.
- NVIDIA CUDA 11.4 toolkit
  - The most advanced platform necessity to compile and run Nvidia GPU programs.
- Intel One API
  - Optimized code generation via ICC for the Intel x86 architecture CPUs.
  - Support for profiling of both serial and multithreaded applications with Vtune
- AMD AOCC Compiler Suite
  - High level of advanced optimizations and multi-threading support on AMD CPUs.
- Nvidia HPC SDK
  - The advanced compilers (OpenACC, NVCC), libraries (cuFFT, cuBLAS) and software tools (nvprof) essential to maximizing the performance of Nvidia GPUs
- Mellanox OFED with UCX and OpenMPI 4.0
  - Official driver + user libraries for Mellanox 200G HDR IB cards.
- Spack
  - A user-friendly and flexible package dedicated to the HPC clusters
- Slurm
  - A highly-scalable and fault-tolerant cluster management and job dispatching toolkit

intel Software · NVIDIA · AMD · CentOS · Mellanox TECHNOLOGIES

Optimizations and Strategies

General Strategy

- Run profiling to find the application hotspots and data transfer bottlenecks with **ARM Forge**, Intel **VTune** and **NVIDIA Visual Profiler**.
- Achieve the best performance by choosing the most **performance-cost** efficient SKUs and the **fastest** libraries on the SKUs.
- **Tune the performance** around the hotspots by migrating them to computational extensive devices and rewriting the code on the critical paths.
- **Validate** our work.

HPL / HPCG

- Tune the **key parameters** of HPL to change the problem size and get maximum performance according to the cluster configuration
- Increase the **interconnection bandwidth** by using the RDMA technology and NV-Link.

IO500

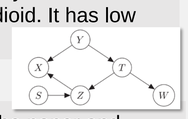
- Try different possible combinations of parameters based on the configuration details of the cluster
- Use the fast **BeeOND filesystem**.

Mystery Application

- Leverage **spack** to choose the suitable compiler and math libraries
- **Wise management** of resource: reserve enough budget for both GPU and CPU VM types.
- Pick out **the best SKU** based on profiling results of the experiment at the beginning.
- Optimization and **analysis**: I/O performance, CPU affinity

Cardioid

- Create a **Makefile** script for fast deployment on the cloud cluster.
- Try different **CMake** configurations using appropriate version of ICC, GCC, MPI and CUDA to build MFEM. Use **OpenMPI** for multi-nodes
- Promote the computation intensity of **MFEM** which is a modular parallel library of the finite element method crucial in Cardioid. It has low GPU utilization now.

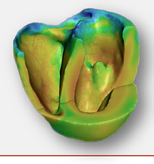


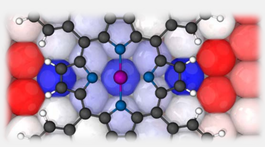
Reproducibility Task

- Comprehensively understand the paper and make comparison of the three algorithms (**GS, IAMB, Inter-IAMB**) to explain their adaption to different datasets
- Write scripts to promise robust **automatic experiment** process
- Analyze **dataset features** to better describe the results

Quantum Espresso

- Speed up by leveraging **OpenMPI** and **offloading** GPU-friendly computation based on profiling results.
- Tune runtime options to adjust **parallelization levels** and achieve efficient hierarchy of processor groups.
- Distribute **3D FFT calculation** computations and minimize the overhead.





Our Preparation

Learning Internally and Externally

- Invite experts in academia and industry to deliver talks on HPC-related topics, free and open for everyone willing to promote HPC education inside the campus.
- Design homework labs for the new members in the HPC team, including CUDA, MPI, OpenMP, profiling and paper reading.

Interdisciplinary learning

- Self-teaching the background knowledge in the application e.g. biochemistry and physics.
- Collaborate with the engineers to solve technical problems.
- Open tutorials by the professors to introduce latest Interdisciplinary methods and progresses.

Responsibility Distribution

- An **application leader** is responsible for dispatching **the application members** to learn background information, profile bottlenecks and brainstorm optimization methods for hotspots
- Our mentor gives advice about the system configuration and maintenance.
- The team leader is responsible for organizing meetings and solving configuration problems.
- Group meetings on fixed time

Support from our Vendors

- Provide training hardware and technical assistance:
  - SITONHOLY: access to HPC hardware for training as well as their expertise in the domain.
  - Peking University: university-sponsored access to online meeting platforms and regular training sessions in particular domains.
- Discussion on HPC software solutions with experts in industry.

Learn from the competition!

■ The **passionate members** who cooperated together in the competition build up solid friendship
■ The members gain deep insights in the areas including **Parallel Computing**, **System Administration, Cluster Management** and **Computer Networking** from the experience.
■ We establish **extensive connections** with our vendors in the industry.

Special Thanks to

ORACLE CLOUD · Azure · acm · 思腾合力 SITONHOLY 加速AI梦想