Tsinghua University Team
SC24 Student Cluster Competition

Kai Yang, Zhiyu Xue, Yifan Li, Yuqing Yang, Yucheng Wang, Chengyu Shi, *Jidong Zhai, Wentao Han and Yuyang Jin*

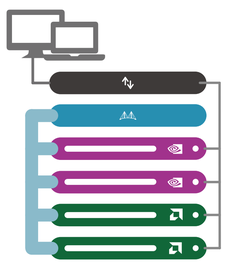
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

we will win with the power of

well-designed architecture

network and cluster overview

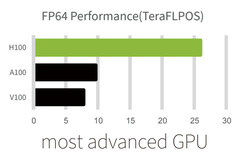
- ***GPU nodes* x2 + *CPU nodes* x2:**
  - 2 x CPU per node
  - 6 x GPU per GPU node
  - reduce cross-node GPU communication
- ***NVIDIA Mellanox ConnectX-6:***
  - 200 Gbps bandwidth with < 0.6 µs latency
- ***Gigabit Ethernet:***
  - on-board 1000MBase-T ethernet adapter
  - stable for management and monitoring
- ***Delicate PCI-E channel allocation:***
  - adapted to requirements of different applications
  - each **GPU** uses one x16 slot, while SSD uses one U.2
  - each **IB** adapter uses one x16 slot



cutting-edge computing hardware

Powered by ***NVIDIA® H100 Tensor Core GPU!***

2x GPU nodes
***H100* x6**
✓ benchmarked
✓ apps optimized
✓ implemented to be GPU memory friendly



most advanced GPU

2TB/s
memory bandwidth
350 Watts
power consumption
74 GFLOPS/Watts
exceptional efficiency

AMD Zen4 EPYC™ 9004 Processors with AVX-512

4x *nodes*
***Epyc 9754* x2**

:-) 12 Memory Channels
:-) PCIe® 5.0 x128
:-) 128 Cores with AVX-512

For each *GPU node* and *CPU node*:

- ***DDR5 4800 MHz 64 GB* x24:** reasonable capacity
  - 16 per socket to work in octo channel for full bandwidth
- ***Intel® SSD DC S3610 100 GB* x1:** high-performace OS and libraries
- ***Hot plug, Redundant power supply* x2:** load balance & high efficiency

Also, on the first node we have:

- ***Intel® SSD DC P4618 (6.4TB)* x2:** blazing fast shared storage
  - up to 6 GB/s sequential read and 5 GB/s sequential write

software choices

- ***Debian GNU/Linux 12 'bookworm':*** stable and reproducible
  - always up to date without expensive subscription
  - friendly to HPC applications with native supports
  - keep track with modern Linux kernel 6.1.27
- ***ZFS on Linux:*** robust, efficient file system
  - periodical snapshots, fast recovery from misoperation
  - pool-based storage, easy for managing and sharing
- ***Spack:*** flexible package manager for HPC
  - manage multiple versions of compilers, utilities and libraries
  - Intel® oneAPI Toolkit 24; GCC 9.2/10.2/12.2
  - Mellanox® HPC-X; OpenMPI 4.1.6
  - NVIDIA® CUDA 11.8/12.1; CuDNN 8.7
- ***Modified Telegraf:*** controlling and monitoring on cluster resources

we will win for

specialized optimizations

general methodology

Profiling: find out the hotspots of an application
- ARM® Allinea MAP, Intel® VTune™ Amplifier, NVIDIA® Profiling Tools

Code optimizing:
- overlapping communication and computation
- cache optimization, data alignment, branch prediction

Compiling: with recent optimizing compilers and tuning options

Tuning: mpitune, CPU affinity, NUMA binding, etc.

HPL

Tune program input parameters:
- fully understand and tune the problem size for each possible scale
- script-driven automatic optimizing

Runtime Power Control:
- use scripts to manage runtime power limit of each hardware

MLPerf

High performance backend: pytorch or tensorflow

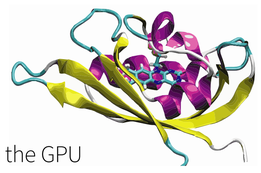
Parameters: change parameters like batch_size, gpu_copy_streams, etc.

NAMD

Bottleneck optimization:
- targeted optimization based on profiling results

Options:
- modify compile options and runtime options
- adjust the process ratio for different functions on the GPU

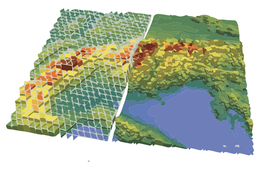
Environment:
- test on different hardware to find the best platform


Simulation Result from NAMD

ICON

Compile options / libraries:
- try different libraries and compile options

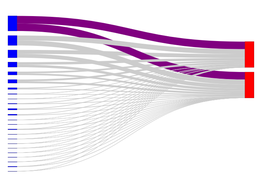
Targeted optimization:
- focus on I/O
- MPI : change worker/IO processes number
- try OpenACC on atomosphere programs


ICON Calculation Scenario Overview

reproducibility challenge(datalife)

Reproduce the DFL DAG graph:
- understand the task-file pair interaction
- use visualization tools to reproduce graph

Tune performance:
- use DFL-Gs patterns to identify problems and conquer


DAG of Belle II Monte Carlo

mystery benchmark

Control power consumption:
- start from a small scale to control power

Tune options:
- adjust compile flags and runtime settings

mystery application

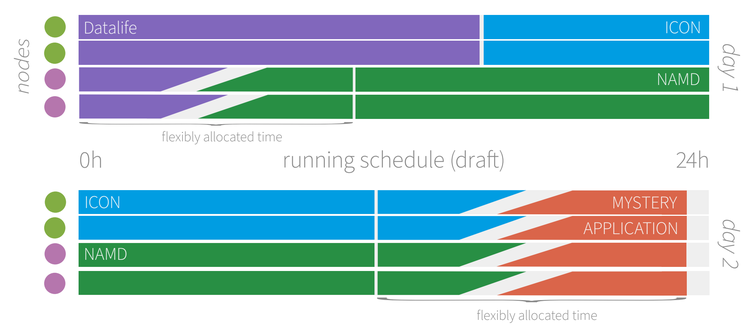
Different compiling methods: thanks to ***Spack***

Analysis: scalability, communication-intensity and GPU-friendliness

we will win since we have

sophisticated tactics

scheduling in 47h allotted time & budget



- ***Reproducibility Challenge (datalife):***
  - high priority for reproducing graph or profiling to start report writing
  - needs CPU nodes, may also need GPU nodes
- ***ICON:***
  - use the appropriate number of nodes for extensibility
  - use the remaining half of the CPU node time
- ***NAMD:***
  - the GPU node time is assigned with priority
  - start all GPU nodes to run in parallel
- ***Mystery Application:***
  - run after thorough analysis and careful optimization
  - decide the number of nodes based on analysis
  - choose run time based on optimization and scoring rubric

controlling over power consumption

Monitoring:
- ✓ realtime with *Grafana*
- ✓ cluster metrics from SNMP
- ✓ server data from IPMI
- ✓ fine-grained CPU & GPU metrics collecting with agents
- ✓ quicker decision making
- ✓ know which is consuming
- ✓ balance efficiency
- ✓ more metrics like bandwidth, throughput, IOwait, etc.

Controlling:
- ***P-state:*** adjust with *tuned* using *tuned-adm*
- ***C-state:*** controlled by P-state, put CPU more often into C6
- ***Turbo Boost:*** enable/disable boost with */proc/sys* settings
- ***Enhanced SpeedStep:*** reduce CPU frequency with *cpupower*
- ***GPU Power:*** balance energy with *nvidia-smi*
- ***Daemon:*** disable useless daemon to reduce overhead


members participating in ISC24

we will win with a team of

diversity and collaboration

creating a diverse team

- ✓ ***Gender diversity (with female team member)***
- ✓ ***Birthplace diversity (particularly in China)***
  - members from 6 different provinces
  - with different cultural backgrounds
- ✓ ***Diversity of experience***
  - different years of study
  - experienced HPC contestants and newbies
- ✓ ***Diversity of personal interest***
  - take different courses: economy, algorithm, OS, AI, database, etc.
  - interested in different fields: math modeling, deep learning, network, etc.
  - have different social works: student union, open source mirror, etc.
  - different extracurricular activities: photography, swimming, bangdi, etc.
  - plan on different futures: doctoral/master study, work, startup, etc.
- ✓ ***Different majors (along with secondary majors)***
- ✓ ***Encourage females and students with interdisiplinary knowledge to join us***


Team Photo (with awards!)

collaboration based on different majors and skill sets

***Kai Yang:* Mathematics & CS**
- team leader, Mathematics

***Yucheng Wang:* Computer Science**
- performance analysis

***Chengyu Shi:* Computer Science**
- algorithm, systems

***Yuqing Yang:* Computer Science**
- architecture, web development

***Zhiyu Xue:* CS & Economy**
- system tuning, supercomputer application

***Yifan Li:* Computer Science**
- bioinformatics, asynchronous parallel frameworks

...More backup members in Chemistry, Physics and other majors

Leader & Presentation
HPL
MLPerf
Mystery Benchmark
NAMD
ICON
Reproducibility
Mystery Application
Cluster Management
Poster and Slides

they support us

Our School:
- **Tsinghua University**
- Motto: Self-Discipline and Social Commitment
- Spirit: Actions Speak Louder than Words

Our Sponsors:

正定私募基金
DEFINITE CAPITAL MANAGEMENT

Corresponding author: zhaijidong@tsinghua.edu.cn