Tsinghua University Team
SC20 Virtual Student Cluster Competition

Chen Zhang, Jiajie Chen, Yutian Wang, Zeyu Song, Mingshu Zhai, Runxin Zhong, *Wentao Han, Lin Gan and Jidong Zhai*

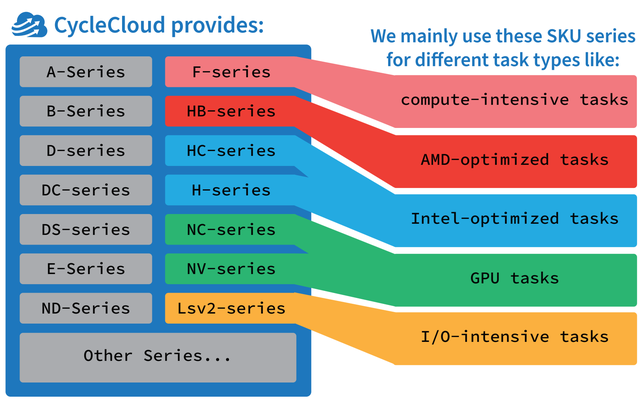
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

we will win powered by
fully-utilized cloud resources

cloud overview

- ✓ various computing hardwares for different HPC workloads
- ✓ monitoring and visualization toolchain
- ✓ built-in templates and custom images
- ✓ autoscaling on top of commonly used schedulers

SKUs choices



- ■ *F-series*:
  - · Intel® Xeon Platinum 8272CL processors for Fsv2
  - · high CPU-to-memory ratio
  - · the best value in price-performance in the Azure portfolio
- ■ *HB-series*:
  - · 120 AMD EPYC 7742 processor cores for HBv2
  - · up to 200 Gb/sec Mellanox® HDR InfiniBand
  - · the best performance, scalability, and consistency for CPU computing
- ■ *HC-series & H-series*:
  - · 44 Intel® Xeon Platinum 8168 processor cores for HC44rs
  - · high CPU frequencies with large memory per core
  - · the best for applications that rely on Intel® toolchain
- ■ *NC-series & NV-series*:
  - · various single, multiple, or fractional NVIDIA® Tesla GPUs
  - · ability to fully utilize GPU performance
  - · the best for compute-intensive GPU-accelerated applications
- ■ *Lsv2-series*:
  - · 19.2TB (10x1.92TB) NVMe SSD M.2 device for L80sv2
  - · up to 3.8M Read IOPS and 20000MBps throughput
  - · the best for IO500 and other I/O-intensive tasks

software choices

- · ***CycleCloud Ubuntu 18 (with Azure tuned kernel)***:
  - · friendly to HPC applications with native supports
  - · support for Hyper-V™ NVMe Direct technology
- · ***Resource manager***:
  - · instantiate the VMs directly instead of through schedulers
  - · use the official CycleCloud cli and web interfaces to orchestrate
- · *Spack*: flexible package manager for HPC
  - · manage multiple versions of compilers, utilities and libraries
  - · Intel® Compilers 18/19/20; GCC 6.3/8.2/9.2
  - · Intel® MPI 2018/19/20; Mellanox® HPC-X; OpenMPI 1.10.7/3.1.2/4.1.0
  - · NVIDIA® CUDA 8/9/10/11; cuDNN 5/6/7

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 improving:
- · overlapping communication and computation
- · cache optimization, data alignment, branch prediction

Compile: with recent optimizing compilers and tune options
Tuning: mpitune, CPU affinity, NUMA binding, etc

LINPACK & HPCG

Tune program input parameters:
- · fully understand and tune the problem size for each possible scale

Test different GPU cards to choose the most efficient SKU
Use all possible GPU nodes for close to linear scalibility

IO-500

Use a custom file system called MadFS:
- · a highly-scalable distributed file system for metadata operations
- · inspired by GekkoFS but rewritten in Rust

Use storage-optimized node with NVMe storage
Tuned paramaters: process numbers, file size, files per process,...

reproducibility challenge (MemXCT)

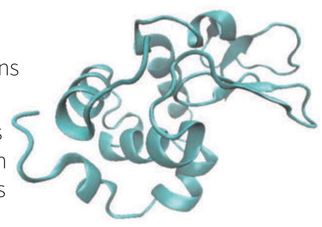
Comprehensive understanding of the paper and code
Experiments:
- · Tested under different architectures
- · Use existing and self-constructed data

Necessary code modification and porting

Gromacs

Understand the basic theory:
- · the underlaying biological interaction
- · the procedure it takes to run simulatons

Tuned performance:
- · domain decomposition, PME settings
- · GPU communication, thread partition
- · experiments on different SKU settings



CESM (Community Earth System Model)

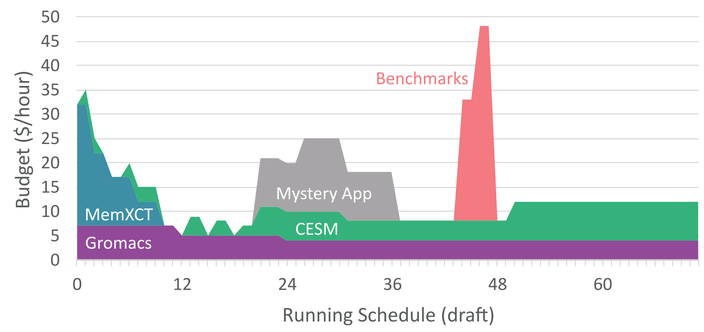
Porting: with different choice of compilers and MPI libraries
Parallel IO: use parallel-netcdf
Load ballance:
- · carefully decide the resources for each component
- · combination of parallel and serial execution
- · choose between MPI processes and OpenMP threads

mystery application

Different compiling methods: thanks to *Spack*
Resource preservation: budget co-scheduling with other applications
Analysis: scalability, communication-intensity and GPU-friendliness

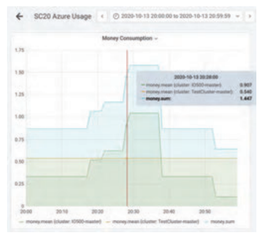
we will win since we have
sophisticated tactics

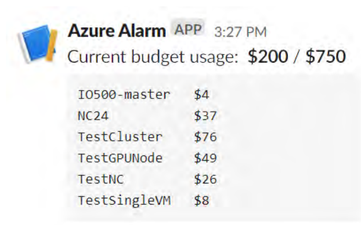
scheduling in 72h allotted time & budget



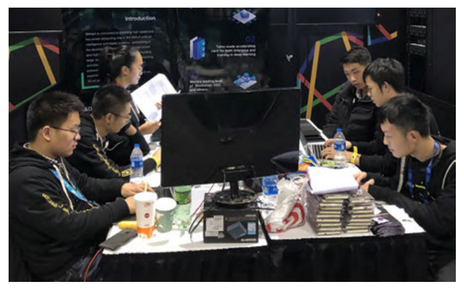
- ■ *Reproducibility Challenge (MemXCT)*:
  - · high priority for performance evaluation to start report writing
  - · finish scalability tests first to gradually release nodes
  - · needs both CPU nodes and GPU nodes
- ■ *Gromacs*:
  - · a single GPU node for each task
  - · start all tasks once the competition starts
- ■ *CESM*:
  - · do some experiments first to find the best mapping
  - · allocate all remaining budget once other tasks finish
- ■ ***Mystery Application***:
  - · run after thorough analysis and careful optimization
  - · decide the SKU based on analysis
  - · choose run time based on optimization and scoring rubric
- ■ *Benchmarks*:
  - · run at last thanks to the predictable score, budget and run time
  - · start this stage once all small tasks finish
  - · find the best budget allocation between tasks and benchmarks

controlling over budget consumption





Realtime money consumption with Grafana
An alarm bot deployed to team slack
A knapsack based best strategy finder



(members working together in SC19)

we will win with a team of
diversity and collaboration

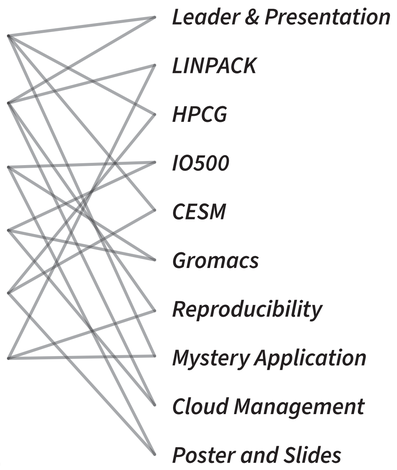
create a diverse team

- ✓ ***Gender diversity (with a female team leader)***
- ✓ ***Birthplace diversity (particularly in China)***:
  - · members from 6 different provinces
  - · each with different cultural backgrounds
- ✓ ***Diversity of experience***:
  - · different years of study
  - · experienced HPC contestants and newbies
- ✓ ***Diversity of personal interest***:
  - · took different courses: architecture, algorithm, OS, AI, database, etc.
  - · interest in different fields: math modelling, deep learning, network, etc.
  - · have different social works: student union, open source mirror, etc.
  - · different extracurricular activities: taekwondo, cyalume dance, etc.
  - · plan on different future: doctoral/master study, work, startup, etc.
- ✓ ***Different majors (along with secondary majors)***
- ✓ ***Encourage females, sophomores and students with interdisiplinary knowledge to join us***

Team Photo (with awards!)

collaboration based on different majors and skill sets

- · ***Chen Zhang***: **Computer Science**
  - · team leader, DL framework
- · ***Jiajie Chen***: **Computer Science**
  - · FPGA, hardware programming
- · ***Yutian Wang***: **Art & CS**
  - · network, cloud native computing
- · ***Zeyu Song***: **Computer Science**
  - · performance tuning, ML
- · ***Mingshu Zhai***: **Computer Science**
  - · algorithm, graph mining
- · ***Runxin Zhong***: **Mathematics & CS**
  - · GPU programming, NAS
- ***...More backup members in Chemstry, Physics and other majors...***



they support us

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

Our Sponsors:

BITMAIN intel Mellanox TECHNOLOGIES

Corresponding author: zhaijidong@tsinghua.edu.cn