Tsinghua University Team in SC18 Student Cluster Competition

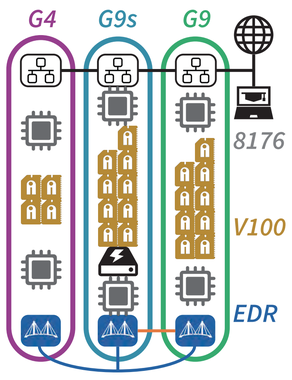
Jiping Yu, Liyan Zheng, Jia'ao He, Xinjian Yu, Chenggang Zhao, Chenyao Lou and Jidong Zhai

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

we will win powered by a

well-designed architecture

networking and cluster overview



3 nodes: G4, G9s, G9
- 168 CPU cores, 2 sockets per node
- 22 GPU accelerators, *4*+*9*+*9* distributed

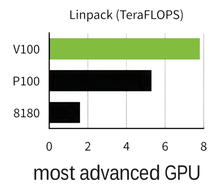
Infiniband EDR
- 100 Gbps bandwidth with 0.7 µs latency
- 2 subnets, better concurrent task control

10 Gigabit Ethernet chain:
- stable for management and monitoring
- lower power consumption than a switch

cutting-edge computing hardware

Thanks NVIDIA® for kindly o□ering *Tesla® V100*!

equipped on all 3 nodes
V100 x (*4*+*9*+*9*)
- □ benchmarked
- □ apps optimized
- □ implemented to be GPU memory friendly



most advanced GPU

900 GB/s memory bandwidth

250 Watts power consumption

26 GFLOPS/Watts exceptional e□iciency

Intel® Xeon Platinum Processors are deployed with AVX-512

equipped
8176 x (*2*+*2*+*2*)

Best performance with 165W TDP

:-) 2.2x acceleration from E5-2699v4
:-) hexa channel to achieve 126 GB/s
:-) 28 cores with AVX-512

Other hardware configuration:

- *DDR4 2666 MT/s 16 GB x* (*12*+*12*+*12*): reasonable capacity
  - 6 per socket to work in hexa channel for full bandwidth
- *Intel® SSD DC S3610 100 GB on G4 & G9*
  Intel® SSD DC S4600 960 GB on G9s
  - high-performace OS and libraries
- *Intel® SSD DC P3608 4 TB on G9s*: blazing fast shared storage
  - up to 5 GB/s sequential read and 3 GB/s sequential write
- *Hot plug, Redundant Power Supply 1000 W x* (*2*+*4*+*4*): load balance

so ware choices

- *Debian GNU/Linux 9 'stretch'*: stable and reproducible
  - always up to date without expensive subscription
  - friendly to HPC applications with native support
  - keep up with lastest Linux kernel 4.9.110
- *Modified Telegraf*: control and monitoring on cluster resources
- *Spack*: flexible package manager for HPC
  - manage multiple versions of compilers, utilities and libraries
  - Intel® Compilers 17/18/19; GCC 6.3/8.2
  - Intel® MPI 2017/18/19; Mellanox® HPC-X; OpenMIP 1.10.7/3.1.2
  - NVIDIA® CUDA 8/9/10; CuDNN 5/6/7
- *ZFS on Linux*: robust, e□icient file system
  - periodical snapshots, fast recovery from misoperation
  - pool-based storage, easy for managing and sharing

we will win for

specialized optimization

general methodology

Profiling: find out the hotspots of each application
- Allinea MAP, Intel® VTune™ Amplifier, NVIDIA® Profiling Tools
- allow us to pay attention to time consuming parts

Code improving:
- vectorization with SIMD intrinsics, o□loading to accelerators
- overlapping communication and computation
- cache optimization, data alignment, branch prediction

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

LINPACK & HPCG

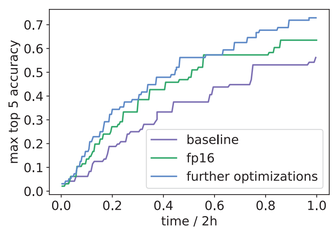
Tune program input parameters:
- fully understand and tune settings including sizes and algorithms
- script-driven automatic optimizing

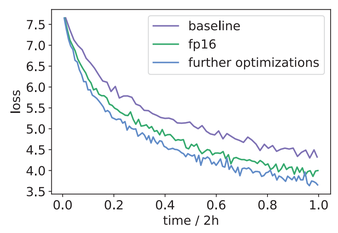
Optimize communication and balance threads to GPUs ratio

reproducibility (SeisSol)

Port to CORE-AVX-512: reproduce on a brand new architecture
Allinea MAP: trace hardware counters and analyze performance
Hyperthreading: study on its impact on performance

Horovod





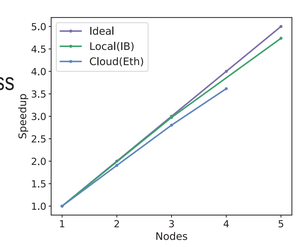
Data compressor: compress data before synchronizing the gradients
- compress data from `float64/32` to `float16`
- strike a balance between speed and precision

Relax Synchronization: add pacman_relax_allreduce method
- directly return the results of last round
- tune `batch_size` and `learning_rate` for this method

OpenMC

Hotspot analysis: find out what makes it slow and try to improve
- locate the bottleneck: memory access
- eliminate atomic operations and random memory access

Compiler mystery: ICC built version runs more than 10 times slower than GCC



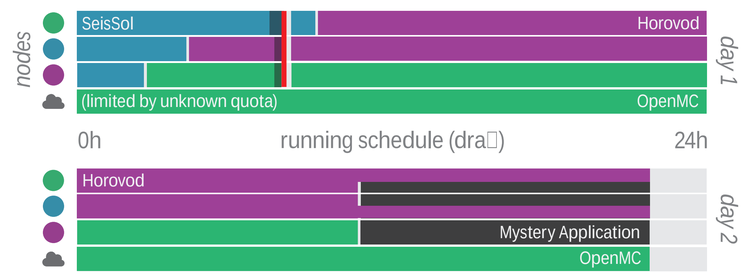
mystery application

Di erent compiling methods: thanks to *Spack*
Resource preservation: preserve *G4* and an IB subnet for it
Analysis: communication-intensity and GPU-friendliness

we will win since we have

sophisticated tactics

scheduling in the allotted time



■ *Reproducibility (SeisSol)*:
- high priority for performance evaluation to start report writing
- finish scalability tests first to gradually release nodes

■ *Horovod*:
- allocate most local cluster time of *G9s*, *G9*
- occupy most local resources, leave the rest for mystery application

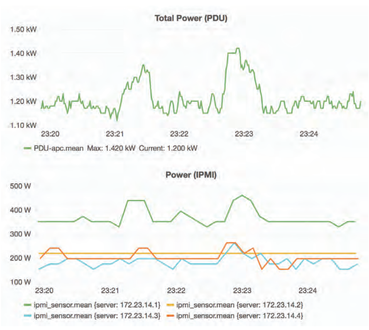
■ *OpenMC*:
- allocate most of the cloud cluster quota and most local time of *G4*

■ *Mystery Application*:
- run last a□er deep analysis and optimization
- decide to use the cloud or local cluster based on analysis
- choose run time based on optimization and scoring rubric

■ *Power Shut-o* :
- implement checkpointing in all applications
- lose tolerable amount of work

controlling over power consumption

Monitoring
- □ realtime with *Grafana*
- □ cluster metrics from SNMP
- □ server data from IPMI
- □ fine-grained CPU & GPU metrics collecting with agent
- quicker decision making
- knowing what is consuming
- balanced e□iciency
- 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 o□en into C6
- *Turbo Boost*: disable boost with */sys* settings
- *Enhanced SpeedStep*: reduce CPU frequency with *cpupower*
- *GPU Power*: balance energy with *nvidia-smi*
- *Daemon*: disable useless daemon to reduce overhead

more secrets…… ;-)

(members monitoring & controlling the power in ISC 18)

we will win with a team of

diversity and collaboration

create a diverse team

□ *Diversity of genders*:
- members of di□erent genders
- male and female candidates in preparing stage

□ *Diversity of birthplaces (particularly in China)*:
- members from 13 di□erent provinces
- each with di□erent cultural backgrounds

□ *Diversity of experience*:
- di□erent years of study
- experienced HPC contestants and newbies
- one backup member still in *high school*

□ *Diversity of personal interests*:
- took di□erent courses: architecture, algorithm, OS, AI, database, etc.
- interest in di□erent fields: math modeling, neural networks, blockchain, etc.
- have di□erent social works: student union, Linux user group, etc.
- di□erent extracurricular activities: ballroom dancing, triathlon, cycling, etc.
- plan on di□erent future: doctoral/master study, work, startup, etc.

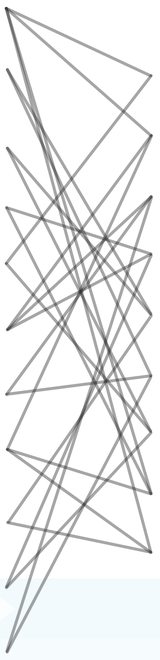
□ *Diversity of majors*:
- computer science, economics, math, interdisplinary science, physics

□ *Actively encourage females, sophomores and students with interdisiplinary knowledge to join us*

collaboration based on di erent majors and skill sets

- *Jiping Yu*: Computer Science
  - operating system, database
- *Yu Chen*: Computer Science
  - compiling, algorithm, CUDA
- *Liyan Zheng*: Economics & CS
  - algorithm, so□ware engineering
- *Shizhi Tang*: Computer Science
  - artificial neural networks
- *Jia'ao He*: Computer Science
  - web development, physics
- *Chenyao Lou*: Interdisplinary Science
  - network, algorithm, systems
- *Chenggang Zhao*: Computer Science
  - MPI, architecture, research
- *Shengqi Chen*: Math & CS
  - devops, network, LaTeX
- *Chen Zhang*: Computer Science
  - discrete optimization, design
- *Yuxian Gu*: Computer Science
  - mathematical modeling
- *Xinjian Yu*: Computer Science
  - distributed systems, blockchain

...More backup members...



they support us

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

Our Sponsor:

BITMAIN

Corresponding author: zhaijidong@tsinghua.edu.cn