Tsinghua University Team in SC17 Student Cluster Competition

Beichen Li, Guanyu Feng, Jiping Yu, Ka Cheong Jason Lau, Lei Xie, Yu Chen and Jidong Zhai

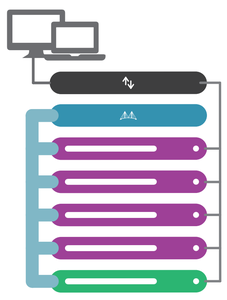
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

we will win powered by a

well-designed architecture

networking and cluster overview

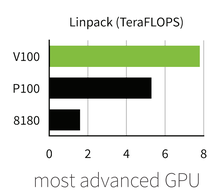
- ***Scaling nodes* x4:**
  - balanced power comsumption
  - equipped with parallel accelerators
- ***Special node* x1:**
  - highest base frequency
  - fully utilize power for sequential parts
- ***InfiniBand EDR:***
  - 100 Gbps bandwidth with 0.7 μs latency
- ***Gigabit Ethernet:***
  - Broadcom BCM5720 1000MBase-T
  - stable for management and monitoring



cutting-edge computing hardware

Thanks NVIDIA for kindly offering ***Tesla® V100***!

4x scaling nodes
***V100* x2**
✓ 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 efficiency

Intel® Xeon Platinum Processors are deployed with AVX-512

4x scaling nodes
***8176* x2**

1x special node
***8180* x2**

$\frac{base\ freq.}{TDP} = 0.01273$ Balanced Energy Efficient

$\frac{base\ freq.}{TDP} = 0.01220$ Highest Performance

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

For each *scaling node* and *special node*:

- ***Samsung® DDR4 2666 MT/s 16 GB* x12**: reasonable capacity
  - 6 per socket to work in hexa channel for full bandwidth
- ***Intel® SSD DC S3610 100 GB* x1**: high-performace OS and libraries
- ***Hot plug, Redundant Power Supply 1100 W* x2**: load balance

Also, on the first node, we have:

- ***Intel® SSD DC P3700 1.8 TB* x2**: blazing fast shared storage
  - up to 2.8 GB/s sequential read and 1.9 GB/s sequential write

software choices

- ***CentOS Linux release 7.4.1708***: stable and reproducible
  - always up to date without expensive subscription
  - friendly to HPC applications with native supports
- ***Cluster-optimized Linux kernel 3.10.0***:
  - P-state supported to reduce power comsumption
  - most suitable power management and memory allocation
- ***Modified Telegraf***: control and monitoring on cluster resources
- ***Intel® C++ Compilers 2018***: speedup applications for AVX-512
- ***CUDA 9***: built for Volta GPUs, faster GPU-accelerated libraries
- ***Intel® MPI Library 2018***: high performance MPI-3.0 on multiple fabrics
- ***PGI 17.10, GCC 4.8.5, MVAPICH 2-2.3a GDR, OpenMPI 3.0, HPC-X...***

we will win for

specialized optimization

general methodology

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

Code improving:
- vectorization with SIMD intrinsics, offloading to accelerators
- 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

CUDA 9.0: support Tesla® V100 to accelerate computation

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

Optimize communication and balance threads to GPUs ratio

reproducibility (LAMMPS)

1-click run: benchmarking script on given input with checkpointing
Port to CORE-AVX-512: reproduce on a brand new architecture
Try AVX-512CD: even faster, investigate more than required
Reproducible everywhere: besides, tested on KNL, GPU and cloud

Born

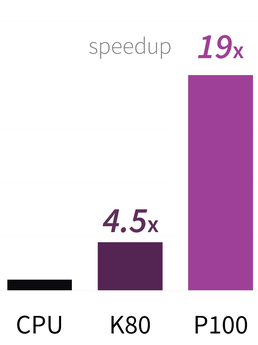
GPU ready: implement prop class to accelerate with GPUs

Save GPU memory: distribute data to multiple GPUs with MPI

Better access: GPU cache-friendly memory accessing pattern

Overlapping: hide I/O and communication time on computation

With a same power consumption ——



MrBayes

MrsBayes: to manage MrBayes, automatically read scores: stddev & psrf
- interactive program to calculate and backup

AVX-512: improve efficiency without violating coding convention
- add module in line with the framework on a consistent precision

Full advantage of resources: no matter it is local or on the cloud
- try parameter configurations to choose the one with best scores

Study Bayesian phylogenetic methods: to understand parameters

cloud component

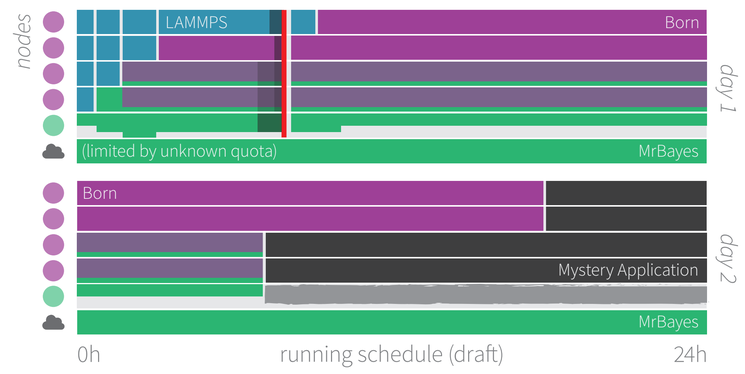
Tradeoff between H16, A11 and F16s: budget matters in decision

MrBayes: much less performance loss compared to other applications

we will win since we have

sophisticated tactics

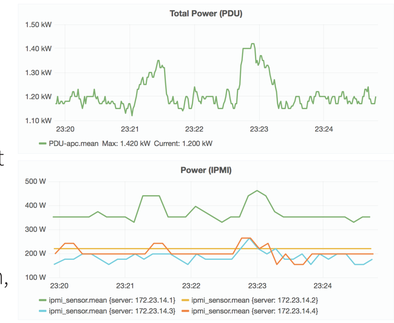
scheduling in 48h allotted time



- ***Reproducibility (LAMMPS)***:
  - high priority for performance evaluation to start report writing
  - finish scalability tests first to gradually release nodes
- ***Born***:
  - allocate most of the local cluster time
  - use only GPU resources: can overlap with MrBayes
- ***MrBayes***:
  - allocate most of the cloud cluster quota
  - comsume idle power resources and local CPU time
- ***Mystery Application***:
  - run last after 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-off***:
  - implement checkpointing in all applications
  - lose no more than 30 mins of work

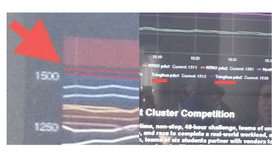
controlling over power consumption

Monitoring
- ✓ realtime with *Grafana*
- ✓ cluster metrics from SNMP
- ✓ server data from IPMI
- ✓ fine-grained CPU & GPU metrics collecting with agent
  - quicker decision making
  - 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
- ***Cgroups***: fine-grain resource limiting for processes
- ***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
- ***MrBayes***: consume extra powers but not going over limit

more secrets…… ;-)



(a great success in SC15 controlling the power)

we will win with a team of

diversity and collaboration

create a diverse team

✓ ***Gender diversity***:
- members of different genders
- a male designer* and a female programmer

✓ ***Birthplace diversity***:


Different Place of Birth

✓ ***Diversity of experience***:
- different years of study
- experienced HPC contestants and newbies

✓ ***Diversity of career interest***:
- took different courses: coding, architecture, algorithm, OS, etc
- interest in different fields: computer graphics, robotics, network, etc
- have different social works: student union, Linux user group, etc
- plan on different future: doctoral study, work or startup

✓ ***Different majors***

✓ ***Encourage females, sophomores and students with interdisiplinary knowledge to join us***

* particularly underrepresented in Chinese universities

collaboration based on different majors and skill sets

- ***Beichen Li***: **Economics & CS**
  - leadership, communication, delivery
- ***Guanyu Feng***: **Computer Science**
  - algorithm, CUDA, MPI
- ***Jason Lau***: **Information Art and Design**
  - design, computer architecture
- ***Lei Xie***: **Computer Science**
  - academic research, numerical analysis
- ***Jiping Yu***: **Computer Science**
  - operating system, database
- ***Yu Chen***: **Computer Science**
  - compiling, algorithm
- ***Qian Xie***: **Interdisciplinary Sciences**
  - LaTeX, mathematical modeling
- ***Xinjian Yu***: **Computer Science**
  - network, system architecture

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