TEAM SUPERNOVA

Nanyang Technological University, Singapore

Students Meiru Hao, Xinye Zhang, Youlin Shen, Liu Cao, Weiliang Heng, Bian Wu

Advisors Prof. Bu-sung Lee (NTU), Mr. Siyuan Liu (NSCC) , Mr. Jianxiong Yin (NVIDIA), Mr. Jeff Adie(NVIDIA), Dr. James Chen (NSCC), Mr. Paul Hiew (NSCC)

Meet the Team and School

Nanyang Technological University, Singapore

Inaugurated in 1991, NTU has grown to become a research-intensive and comprehensive university. In NTU, the High Performance Computing Centre provides HPC resources to the entire NTU research community. What's more, NTU offers a HPC specialization to computer science and computer engineering students.

NTU HPC Team History

A Truly Diverse Team

- **Balanced gender ratio**
  - on par of our university gender ratio
- **Mixed nationalities**
  - Singaporeans and international students from different countries
- **Various skill sets and interest**
  - each application assigned to multiple members with different skill sets
  - collaborate and complement each other
  - brainstorm together for optimization ideas
- **Diversity of experience**
  - mix of seniors and juniors in HPC
  - members with OI and ACM ICPC expereince
- **Making efforts towards a more diverse HPC coomunity in NTU**
  - hold weekly HPC workshop in NTU which opens to all majors
  - even attract exchange students from other universities

Team Members in SC18 Student Cluster Competition

Meiru Hao
- leadership
- SC17, ISC18
- OpenMC, HPL

Xinye Zhang
- system admin, CUDA
- ISC18
- Horovod, HPL, HPCG

Bian Wu
- compiling, design
- ISC17, ISC18
- reproducibility, mystery

Weiliang Heng
- system admin, network
- ISC18
- cloud, reproducibility

Youlin Shen
- algorithm, mathmatics
- new member
- OpenMC, mystery

Liu Cao
- CUDA, algorithm
- new member
- Horovod, HPL, HPCG

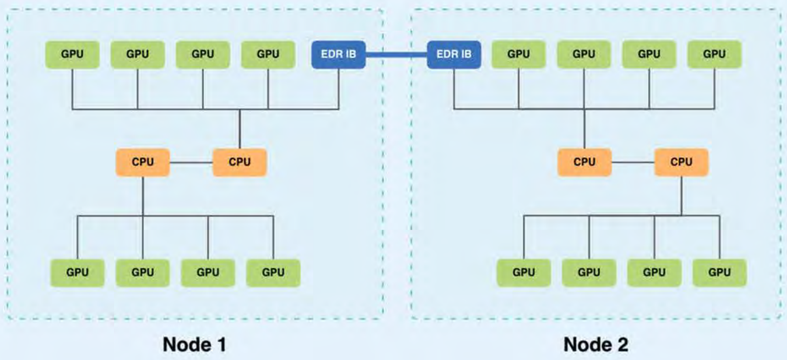
Our Preparation

- We held weekly meeting to update application optimization progress and exchange HPC knowledges.
- We consulted domain experts on seismic imaging, deep learning to better understand the applications.
- We attended some conferences and events, such as the Singapore Supercomputing Asia conference, to accumulate experience in HPC.
- We attended HPC workshops organised by National Supercomputing Center which includes MPI and OpenMP programming and performance analysis.
- We maintains a small training cluster with multiple nodes and GPU support for practising before competition hardware arrives.

Hardware and Software

Cluster Overview

- **2 * Supermicro 4029GP-TRT nodes**
  - highly flexible, enables fine power control
  - 8 GPU slots, high compute power density
- **2 * Intel Xeon Gold 6148 (20 cores) per node**
  - 20 cores with AVX-512, Great power efficiency
- **8 * NVIDIA Tesla V100 per node**
  - 32GB VRAM, beneficial for Horovod
  - set larger batch size to make full use of computation power
- **256GB 2133MHz DDR4 per node**
  - fully utilize CPU bandwidth for best memory access performance
- **480GB SSD per node**
  - reasonable capacity and high speed for OS and and libraries
- **100Gb/s Mellanox InfiniBand EDR**
  - low latency with high bandwidth



Hareware Choice

- **High-density GPU cluster**
  - Our hardware aims at improving the throughput of GPU applications. (HPL, HPCG, Horovod)
- **CPU application**
  - tackle OpenMC with cloud component
  - reproduce SeisSol with our 80 cores
- **Networking**
  - we save the power of the InfiniBand switch by having just two nodes connecting back to back
- **Intel 6148 + 32G V100**
  - Our team own *Intel 6148 (Skylake, 20 cores * 2)* and *Intel E5-2699 V4 (Broadwell, 22 cores * 2)*
  - Thanks to Nvidia's sponsorship, we have the choice of *16G V100* and *32G V100*
  - We carefully compared the 4 combinations and found that *Intel Xeon 6148 with 32G V100* has the best performance/power ratio.

Software Choice

- **CentOS 7.3**
  - high-quality Linux distribution that gives HPC complete control
- **NVIDIA CUDA Toolkit**
  - Highly optimized GPU-accelerated libraries
- **Intel Parallel Studio 2018**
  - Intel compiler performs better and is relatively easier to use than GCC
  - Intel MPI is a high performance MPI implementation
- **Mellanox HPC-X 2.1.0**
  - optimised message communications based on OpenMPI for Mellanox InfiniBand
- **Arm Map and Arm Report**
  - Arm Report provides straightforward idea and Arm Map provides detailed profiling result
- **Lmod**
  - ease of management for multiple versions of libraries
- **Ansible**
  - to automate server configuration & software installation

Application Optimization

General Methodology

- **Compilation**
  - compare different compilers with various optimization flags
- **Libraries**
  - compare different version of dependencies to avoid incompatibility
- **Profiling**
  - profile with Arm Forge and Intel Vtune in order to find hotspot and find a direction for optimisation
- **Domain Knowledge**
  - study relative researches and bring doubts to professors and industry engineers in respective field

Bechmarking

HPL solves a dense system of linear equations with LU factorization.

HPCG solves a sparse one with preconditioned conjugate gradient.

√ **prepared tuning scripts** including automatic GPU & CPU frequency
√ **fully understood input parameters** e.g. N, TRSM
√ **use latest binaries** provided by NVIDIA
√ **carefully tuned** parameters with power monitors

Horovod

Horovod is a distributed training framework running on top of Tensorflow, Pytorch and Keras. It implements "ring-all-reduce" algorithm to effectively average gradients and communicate between multiple GPUs.

√ **consulted domain experts** from Nvidia, Sensetime and NTU
√ **use FP16** for communication
√ **fine-tuned** on several deep models
√ **budget-aware** power monitor
√ **added GPUDirect RDMA support**

OpenMC

OpenMC is a Monte Carlo particle transport simulation code focused on neutron criticality calculations. It supports hybrid usage of MPI and OpenMP. By the nature of Monte Carlo simulation, and also showed by our profiling result, OpenMC is bounded by intensive memory access.

√ **investigated the memory access pattern** and consulted the author
√ **prepared local & cloud version** to make better use of resources
√ **compared different MPI implementation** including IntelMPI and OpenMPI. We also tried HPCX on our cluster.

Reproducibility (SeisSol)

To tackle the extreme scale of 2004 Sumatra megathrust earthquake, an end-to-end optimization of SeisSol has been made in terms of code generation, wave propagation and dynamic rupture optimization, local time stepping, and asynchronous I/O (Uphoff et al., 2017). In the reproducibility challenge, we aim to reproduce the speedup and scalability of the optimized version of SeisSol on our cluster.

√ **prepared 1-click scripts** for both running and graph-plotting
√ **reproduced on different arch** and compare the performance
√ **consulted professor** from the Department of Geophysics to better understand seismic activities and analysis

Running Strategies

Cloud Component

- **Prepared everything in cluster-init**
  - to save time in competition
- **Compared compute optimised, memory optimised and HPC series**
  - on HPC series we can use RDMA
- **Conducted sufficient beforehand testing**
  - so that we are able to choose the most suitable virtual machine type based on the performance/cost ratio after the budget is announced onsite

Power Shutoff

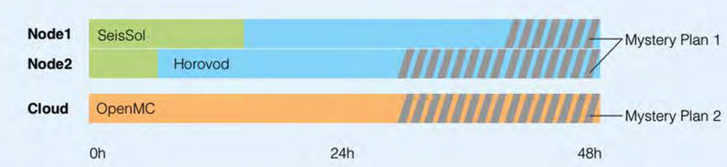
- **Horovod and SeisSol**
  - built-in checkpoint
  - we'll set the checkpoint to every 30 minutes
- **OpenMC**
  - not be affected by power shut-off because it will be run on cloud

Power Consumption

- **Power Monitoring**
  - use ipmi tool to monitor rough power control
  - use power meter when we were tuning HPL and HPCG precisely
  - watch nvidia-smi and keep GPUs at a balanced temperature
- **Power Control**
  - control GPU frequency with nvidia-smi
  - control CPU frequency with cpupower
  - fine-tune power through disabling some of other CPU/GPU features

Runtime Schedule

- At the beginning of the 48-hour competition, we'll run SeisSol locally in order to get the data output for writing report.
- Then we'll run Horovod on our 2 local nodes and compile mystery locally.
- We will decide whether to run mystery locally or on cloud onsite. If we run mystery locally, it will be run on second day and we'll balance mystery and horovod based on the scoring scheme.
- We'll run OpenMC on cloud from the beginning of the competition and give OpenMC most of cloud budget.



Why We Will Win?