Analysis of Characteristic Features of HPC Applications

Lian Jin
HPC Engineer, Inspur
HPC applications

How to make these applications much more efficient?

- Application optimization
- System optimization

Understanding Application features

Software tools and applications: abinit, ADF, Q-CHEM, ANSYS, LAMMPS, VASP, GROMACS, CASTEP, CPMD, Materials Studio, Mathematica, Quantum ESPRESSO, OpenFOAM.
Inspur’s Strategy

- Professional research team of HPC application

High End Talents

- Tao Yu Ph.D: Petroleum
- Yu Liu Ph.D: Material/Physics
- Wenjing Lv Ph.D: Computational Mathematics
- Lian Jin Ph.D: Life science
- Miao Wang Ph.D: Weather\Climate
- Bowen Chen Ph.D: CAE/CFD

More Support teams

- > 50 Institutes from CAS
- Tsinghua University
- etc.
Inspur’s Strategy

- Application features analyzing
- Hardware optimization
- Software optimization
- GPU
- MIC
- Special machines
- Input parameters optimization
- Codes optimization
- Physical Modelling innovation
- Algorithm innovation

The way to achieve excellent performance
System Platform Optimization

**preparation**
Platform, Application, workload

---

**Application tuning**
- Application features analyzing
- Tuning parameters
- Tuning parallelization mode
- Tuning workloads

**Features collection**
- Performance data of CPU/Memory/IO/Net work and so on.

**Platform tuning**
- Platform features analyzing
- Tuning hardware configuration and parameters
- Tuning middleware

---

**Purpose**
Provide real time features and performance analysis.
Design the most optimized system solution.

Optimize both your hardware and software based on these features
Balance a variety of computing requirement and build a high-scalable, high-efficient HPC system to maximize the existing application’s performance.
Methods of optimizing computing-intensive applications

Parameters: precision, convergence condition, boundary condition
Code optimization: vectorization, loop optimization, function optimization
Model: pseudopotential, DFT, QMC, CI

OS: Linux, reduced kernel, vectorization support
Compiler: ifort/icpc O2/O3 SSE AVX ...
Library: MKL, IMSL, Goto...

CPU: Boost frequency
HT: off
System Platform Optimization

Methods of optimizing network-intensive applications

Devices: FDR IB, TH-Express, RDMA
HT: on

Compilers: Intel MPI mpich2 openmpi platform mpi
Parameter: I_MPI_ADJUST_BCAST, I_MPI_ADJUST_GATHER, I_MPI_ADJUST_GATHERV ...
MPI environment: mpitune

Input parameters: mode (eg: spin polarization, particles, bands, base number), granularity
Teye Application Features Analyzer

- Analyse HPC application features
- Diagnose application bottleneck
- Help to improve application performance
- Reduce the optimization cost

Teye server

- Large scale cluster and database support
- Asynchronous monitoring support
- Low hardware resources requirement
- Refresh every second

High performance

- Monitoring system level resources
- Monitoring Xeon CPU microarchitecture indexes
- Monitoring PCI-E devices bandwidth
- Monitored data analysing support

CT scanning CPU usage < 1‰

4096 cores support

Teye system diagram

PCI-Express

Monitoring PCI-E Bandwidth

- PCIe_total_bw_GB
- PCIe_read_bw_GB
- PCIe_write_bw_GB

Monitroing PCI-E Bandwidth
Diagnose and Optimize Application

- Running application on supercomputer
- Clients
  - Experiment scientists
  - Theoretical scientists
- Quick Results
- Profiling by Teye
- Application analysis
- Inspur application features database
- Highly optimized jobs
- Applications
- Diagnose and Optimize Application
CASE A. Optimize Gromacs and understand its features on HPC systems based on Intel xeon processors.

Software environment:
- Gromacs 4.6.1, ACP receptor in water 84035 atoms
- Intel composer xe 2013

Hardware environment:
- 1 management node, 10 I/O nodes, 9 blade calculation boxes
- Intel Xeon E5-2670 processors
- Infiniband
Application Diagnosing

Micro structure features

Before Optimization
Vectorization rate and CPI

After Optimization
Vectorization rate and CPI

Analyzation:
These figures show us that Gromacs using single precision floating point processing and it is floating point intensive application. Besides, Gromacs can be optimized for the AVX instructions.
Application Diagnosing

Memory and network features

Before Optimization

Memory bandwidth (GB/s)

After Optimization

Memory bandwidth (GB/s)

Analyzation:
These figures show us that Gromacs costs very low memory bandwidth but require large network bandwidth. So, Gromacs is network bandwidth intensive application.
I/O features

Before Optimization

After Optimization

Improvement

Efficiency

Analyzation:
These figures show us that Gromacs is not I/O intensive application and its output model is not parallelized.
Computationally intensive: Higher CPU frequency leads to higher performance.

Memory: Performance is nearly independence with the Mem BW.

Storage IO: Low requirement.

Network: The high speed, low latency network will give a good parallel performance and high efficiency.
Software environment:
- WRF3.4.1
- Intel composer xe 2013
- 1km

Hardware environment:
- 1 management node, 10 I/O nodes, 20 blade calculation boxes
- Intel Xeon E5-2670 processors
- Infiniband

CASE B. Optimize WRF and understand its features on HPC systems based on Intel xeon processors.
Application Diagnosing

Micro structure features of WRF

Total SP GFlops

Clock per instruction

SSE Vectorization rate

AVX Vectorization rate

Analyzation:
These figures shows us that WRF using single precision floating point processing. Besides, WRF is not highly optimized for the AVX instructions.
Memory and network features

- **Total Bandwidth (GB/s)**
- **Memory read (GB/s)**
- **Infiniband send (MB/s)**
- **Infiniband receive (MB/s)**

Analyzation:
These figures show us that WRF is memory bandwidth intensive and network bandwidth intensive application.
Application Diagnosing

I/O features

- **Server node**
  - Disk read (MB/s)
  - Disk write (MB/s)

- **Computing node**
  - Nfs read (MB/s)
  - Nfs write (MB/s)

**Analyzation:**

These figures show us that WRF is some kind of I/O intensive application but its input and output mode is not parallelized.
Application Diagnosing

- High memory Bandwidth
- Compiler options, 3%~5%
- Intel MKL DFT, 3%
- Quilt IO, ~10%
- Hybrid DM+SM, ~5%
- Parallel file system, ~10%
- nproc_x/nproc_y parameters optimization, 3%~10%

Total 10%~20%
- **Computationally intensive**: Higher CPU frequency higher performance
- **Memory band-width sensitive**: The requirement of memory capacity is defined by the size of workload and more compute resources less memory bandwidth requirement.
- **Storage IO**: High requirement for large workload.
- **Network**: The high speed, low latency infiniband network will give a good parallel performance and high efficiency.
**Application Diagnosing**

**CASE C.**
Profile VASP features and understand its behavior while running on a supercomputer.

- **VASP cases**

<table>
<thead>
<tr>
<th></th>
<th>CASE 1</th>
<th>CASE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ions</td>
<td>190</td>
<td>96</td>
</tr>
<tr>
<td>Base sets</td>
<td>1.41 million</td>
<td>0.37 million</td>
</tr>
<tr>
<td>Bands</td>
<td>564</td>
<td>2000</td>
</tr>
<tr>
<td>K-points</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Algorithm</td>
<td>CG</td>
<td>DIIS</td>
</tr>
</tbody>
</table>

- **Platform**

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>Memory</th>
<th>Network</th>
<th>IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td>E5-2692v2</td>
<td>DDR3 1600</td>
<td>FDR</td>
<td>Raid0</td>
</tr>
<tr>
<td>Notes</td>
<td>422.4GFlop/s</td>
<td>102.4GF/s</td>
<td>56Gbit/s</td>
<td>~800MB/s</td>
</tr>
</tbody>
</table>
## Floating point speed (GFlops)

### Case 1

<table>
<thead>
<tr>
<th></th>
<th>X87</th>
<th>SSE vectorized</th>
<th>SSE scalar</th>
<th>AVX vectorized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Notes:**
  - The intensity of FP is determined by the **number of bands**. The continuity is determined by **algorithm**.
Vectorization rate

Case 1

SSE vectorization rate
AVX vectorization rate

Notes:
Vectorization rate is determined by the number of basesets and its maximum value is determined by the number of bands.

Case 2
### Application Diagnosing

#### Memory bandwidth (GB/s)

**Case 1**
- Total MemBW
- Read MemBW
- Write MemBW

**Case 2**

**Notes:**
The intensity of memory bandwidth and its extremum value is determined by the number of basesets and algorithm.

<table>
<thead>
<tr>
<th>Case</th>
<th>Total MemBW</th>
<th>Read MemBW</th>
<th>Write MemBW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Application Diagnosing

Infiniband bandwidth (MB/s)

Case 1

Notes:
Same as that of memory bandwidth. One should note that the continuity of data exchange would influence the application’s scalability crucially.
Application Diagnosing

- IO and NFS bandwidth (MB/s)

**Case 1**

- IO read
- IO write
- NFS write
- NFS read

**Notes:**
VASP is not a IO intensive application and its IO is not parallelized.

**Case 2**
- **Computationally intensive:** Higher CPU frequency higher performance
- **Memory bandwidth sensitive:** Higher memory bandwidth higher performance
- **Storage IO:** Low requirement
- **Network:** The high speed, low latency network will give a good parallel performance and high efficiency.
# Application Diagnosing

<table>
<thead>
<tr>
<th>App field</th>
<th>App</th>
<th>CPU</th>
<th>Mem_used</th>
<th>Mem bandwidth</th>
<th>IO</th>
<th>Network</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>GROMACS</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Weather</td>
<td>WRF</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>MS</td>
<td>VASP</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

*note:* Full signal means high requirement.
Thank You!