CMA HPC Update

Supporting meteorological service

Min Wei, Chunyan Zhao and many colleagues
National Meteorological Information Centre
China Meteorological Administration
Contents

• HPC Systems
• Model-Supportive Software Systems
• Conclusions
Contents

• HPC Systems
• Model-Supportive Software Systems
• Conclusions
History

CPU (3008 nodes)
8PFLOPS; 18PB
CPU + GPU (24 nodes)
KNL (24 nodes)
IBM HPCS

<table>
<thead>
<tr>
<th>System</th>
<th>Installation Time</th>
<th>Peak Performance (TFLOPS)</th>
<th>Storage Capacity (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Flex System P460</td>
<td>2013</td>
<td>Production Subsystem: 527</td>
<td>2109.38</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>Research Subsystem: 527</td>
<td>2109.38</td>
</tr>
</tbody>
</table>

Beijing: 1 PFLOPS
Shenyang: 75 TFLOPS
Shanghai: 50 TFLOPS
Wuhan: 75 TFLOPS
Guangzhou: 400 TFLOPS
Chengdu: 25 TFLOPS
Lanzhou: 25 TFLOPS
Urumqi: 25 TFLOPS
Resource utilization

IBM HPCS accounts
• 578

IBM HPCS utilization
• Maintain high both in system availability and CPU utilization, 70% to 80% on average peaking at 95%.
GRAPES & BCC_CSM

- **GRAPES** = Global/Regional Assimilation PrEdiction System
- **BCC_CSM** = Beijing Climate Center Climate System Model

<table>
<thead>
<tr>
<th>Forecast Range</th>
<th>GRAPES-GFS</th>
<th>GRAPES-MESO</th>
<th>GRAPES-TYM</th>
<th>GRAPES-MEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Global</td>
<td>East Asia</td>
<td>West Pacific</td>
<td>East Asia</td>
</tr>
<tr>
<td>H-resolution</td>
<td>0.25°</td>
<td>0.1°</td>
<td>0.12°</td>
<td>0.15°</td>
</tr>
<tr>
<td>V-resolution</td>
<td>60L 3hPa</td>
<td>50L 10hPa</td>
<td>50L 10hPa</td>
<td>50L 10hPa</td>
</tr>
<tr>
<td>Forecast Time</td>
<td>00, 12 UTC 240 h</td>
<td>00, 12 UTC</td>
<td>00, 12 UTC 15 members</td>
<td>00, 12 UTC</td>
</tr>
</tbody>
</table>
Benchmark

- GRAPES-GLOBAL model (Parallel)
- GRAPES-MESO model (Parallel)
- GRAPES-4DVAR four-dimensional variational model (Parallel)
- BCC_CSM climate system model (Parallel)
- BCC_AGCM atmosphere model (Parallel)
- GRAPES-SVD singular vector analysis of regional ensemble forecast system (Serial)
- WRF model (Parallel)

- IOzone Benchmark
- IMB Benchmark
- Job scheduling
- Public domain meteorological software packages
PI-Sugon

2017.9

2018.1 2018.6

1  Contract  2  Arrival  3  Installation  4  Power up  5  Service  6  Pre-Operation
Architecture

Two subsystems: hot standby
- Independent computing, shared storage
- General processor, for each system
  - Computing nodes: ~1500
  - Total CPU cores: ~50000
  - Intel Xeon Gold 6142 (16 Core, 2.6GHz)
- 8 PFLOPS peak performance
- 18 PB storage capacity
- 100Gb/s InfiniBand EDR network
- Parastor 300 parallel file system

New technology test and development subsystem
- CPU + GPU (24 nodes)
- Intel KNL (24 nodes)
Software Stack

Parallel application

Parallel program development environment
  IDE / Parallel debug / Performance monitor

Parallel compiling environment
  C/C++ / Fortran
  Loop vectorization / Code optimization
  Math lib

Parallel OS environment
  Job / Resource / Power / Network / Storage ...

PI-Sugon system
<table>
<thead>
<tr>
<th></th>
<th>IBM</th>
<th>PI-Sugon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Improvement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak Performance</td>
<td>~1PFLOPS</td>
<td>~8PFLOPS</td>
</tr>
<tr>
<td>Storage Capacity</td>
<td>~4PB</td>
<td>~18PB</td>
</tr>
<tr>
<td>Inter-Connection</td>
<td>QDR 40Gb/s</td>
<td>EDR 100Gb/s</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>AIX 7.1.0.0</td>
<td>RedHat Enterprise 7.4</td>
</tr>
<tr>
<td>adios</td>
<td></td>
<td>plapack</td>
</tr>
<tr>
<td>blas</td>
<td></td>
<td>plasma</td>
</tr>
<tr>
<td>boost</td>
<td></td>
<td>wgrib</td>
</tr>
<tr>
<td>esfm</td>
<td></td>
<td>wgrib2</td>
</tr>
<tr>
<td>ferret</td>
<td></td>
<td>gcc</td>
</tr>
<tr>
<td>fftw</td>
<td></td>
<td>RogueWave</td>
</tr>
<tr>
<td>geos</td>
<td></td>
<td>compiler wave</td>
</tr>
<tr>
<td>GotoBlas2</td>
<td></td>
<td>parallel tools</td>
</tr>
<tr>
<td>grads</td>
<td></td>
<td>pgi</td>
</tr>
<tr>
<td>grib_api</td>
<td></td>
<td>compilers</td>
</tr>
<tr>
<td>gsl</td>
<td></td>
<td>intel parallel</td>
</tr>
<tr>
<td>hdf</td>
<td></td>
<td>studio xe</td>
</tr>
<tr>
<td>hdf5</td>
<td></td>
<td>cluster edition</td>
</tr>
<tr>
<td>hdfdços</td>
<td></td>
<td>gnu</td>
</tr>
<tr>
<td>hfeos</td>
<td></td>
<td>intel</td>
</tr>
<tr>
<td>hfeos5</td>
<td></td>
<td>parallel</td>
</tr>
<tr>
<td>hypre</td>
<td></td>
<td>mvapich</td>
</tr>
<tr>
<td>ioapi</td>
<td></td>
<td>python</td>
</tr>
<tr>
<td>jasper</td>
<td></td>
<td>mpi</td>
</tr>
<tr>
<td>lapack</td>
<td></td>
<td>mpich</td>
</tr>
<tr>
<td>libpng16</td>
<td></td>
<td>r</td>
</tr>
<tr>
<td>ncl_ncarg</td>
<td></td>
<td>redhat</td>
</tr>
<tr>
<td>nco</td>
<td></td>
<td>software</td>
</tr>
<tr>
<td>ncview</td>
<td></td>
<td>tools</td>
</tr>
<tr>
<td>netcdf</td>
<td></td>
<td>slurm</td>
</tr>
<tr>
<td>nlopt</td>
<td></td>
<td>slurm</td>
</tr>
<tr>
<td>openblas</td>
<td></td>
<td>gridview</td>
</tr>
<tr>
<td>parallel-netcdf</td>
<td></td>
<td>sugon</td>
</tr>
<tr>
<td>petsc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scalapack</td>
<td></td>
<td></td>
</tr>
<tr>
<td>udunits</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GRAPES model suite
Climate models

![Bar chart comparing BCC_CSM and BCC_AGCM with IBM and PI-Sugon]
Contents

• HPC Systems
• Model-Supportive Software Systems
• Conclusions
• High performance computer management software
  • Refined resource management system
  • Operational monitoring system

• Numerical model supporting software
  • Code management system
  • GRAPES Integrated Setting Experiment Tool (GISET)
  • GRAPES Interactive Data Analytics Tool (GIDAT)
Refined resource management system

• Resource management of IBM & PI-Sugon systems
• Unified management of national and regional resources
• Real-time and historical statistical analysis of system resource usage and utilization
• Computing resource and storage resource usage accounting
• Model & job statistical analysis
• Planning: intelligent resource management
  • Resource data mining
  • Decision support analysis
Operation monitoring system

- Monitoring of IBM & PI-Sugon systems and software, audio alarm
- Unified management of national and regional resources
- Real-time monitoring and historical statistical analysis of failure
- Automatic reporting of system availability & statistical analysis of failure
- Fault handling workflow & fault knowledge database
- Model job monitoring
- Real-time monitoring of memory, CPU utilization and jobs
- Planning: intelligent job management
  - Model application feature analysis and data mining
  - Decision support analysis
Code management system

- Code management system (since 2010)
  - Perforce application on IBM and PI-Sugon HPC
  - GRAPES-GFS, GRAPES-MESO, BCC_CSM code repository
  - National & regional distributed design for GRAPES-MESO collaboration
  - Code version control and integration control
  - Code updates 17,000+, code integration 1,000+, version release and bug fix 500+

- Planning: git-based code management
GRAPES Integrated Setting Experiment Tool (GISET)

- Experiment construction
- Experiment scheduling (ecFlow)
- Experiment sharing, statistics, compare
- Integrated code and experiment data management

- Design and implementation based on C/S mode
- Coded by python
- Back-end services run on servers
GRAPES Interactive Data Analytics Tool (GIDAT)

- On-line plotting of diagnostic data
- Interactive analytics function
- Access the datasets by data service API
Contents

• HPC Systems
• Model-Supportive Software Systems
• Conclusions
What have we achieved?

A new HPC solution has been deployed

- Architecture: CPU cluster; GPU cluster; Intel KNL cluster; network, environment, redundancy……
- Majority of migration work completed
- Testing novel architectures
- Collaboration: CMA members; vendors; universities
Next steps

• Efficient and portable code
• Test new architectures and programming models
• Software support services
Thank You for listening!