DMI-HIRLAM on the NEC SX-6

Maryanne Kmit
Meteorological Research Division
Danish Meteorological Institute
Lyngbyvej 100
DK-2100 Copenhagen Ø
Denmark

11th Workshop on the Use of High Performance Computing in Meteorology
25-29 October 2004
• Danish Meteorological Institute (DMI)
• Applications run on NEC SX-6 cluster
• The NEC SX-6 cluster and access to it
• DMI-HIRLAM - geographic areas, versions, and improvements
• Strategy for utilization and operation of the system
DMI’s mission:

- Making observations
- Communicating them to the general public
- Developing scientific meteorology

DMI’s responsibilities:

- Serving the meteorological needs of the kingdom of Denmark
- Denmark, the Faroes and Greenland, including territorial waters and airspace
- Predicting and monitoring weather, climate and environmental conditions, on land and at sea
Applications running on the NEC SX-6 cluster

Operational usage:

- Long DMI-HIRLAM forecasts 4 times a day
- Wave model forecasts for the North Atlantic, the Danish waters, and for the Mediterranean Sea 4 times a day
- Trajectory particle model and ozone forecasts for air quality

Research usage:

- Global climate simulations
- Regional climate simulations
- Research and development of operational and climate codes
Cluster specifications

- **NEC SX-6** (nec[12345678]) : 64M8 (8 vector nodes with 8 CPU each)
  - **Desc.** : Multi cpu vector nodes. Multi node access via IXS. GFS clients. No interactive access.
  - **Processor specs** : 64 * 8 Gflops
  - **Memory specs** : 32 * 6 + 64 * 2 Gbyte RAM

- **NEC SX-6i** (neci[12]) : 2M2 (2 vector nodes with 1 CPU each)
  - **Desc.** : Single cpu vector nodes. No multi node access via IXS. GFS clients. No interactive access.
  - **Processor specs** : 2 * 8 Gflops
  - **Memory specs** : 2 * 8 Gbyte RAM

- **NEC TX7** (asama[12]) : 16M2 (2 scalar nodes with 8 CPU each)
  - **Desc.** : Nodes used for interactive access, file manipulation and scalar workloads. GFS clients.
  - **Processor specs** : 16 * 1300 MHz Intel ItaniumII
  - **Memory specs** : 16 * 2 Gbyte RAM

- **NEC EXPRESS5800** (azusa[12]) : 8M2 (2 scalar nodes with 4 CPU each)
  - **Desc.** : Nodes used for GFS servicing. No interactive access.
  - **Processor specs** : 8 * 800 MHz Intel Itanium
  - **Memory specs** : 8 * 2 Gbyte RAM
User access to the cluster

Research usage:

- No interactive access to SX vector nodes
- Job submission is done from IA64
- All interactive work is done on the IA64 scalar front ends
- All SX and IA64 nodes see the same file systems as if they were local file systems
- Fair share scheduling via *ERS-II* is used for non-operational queues

Operational usage:

- Jobs are run in batch
- Submitted via cron
- Resubmit themselves upon completion, waiting until their next scheduled run
Geographic areas - Through mid June 2004

<table>
<thead>
<tr>
<th>lat_{min} (south)</th>
<th>lon_{min} (west)</th>
<th>lat_{max} (north)</th>
<th>lon_{max} (east)</th>
<th>lat_{south pole of rotation}</th>
<th>lon_{south pole of rotation}</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>−37.525°</td>
<td>−63.725°</td>
<td>47.523°</td>
<td>26.725°</td>
<td>00.000°</td>
</tr>
<tr>
<td>E</td>
<td>−28.677°</td>
<td>−54.275°</td>
<td>13.473°</td>
<td>−13.625°</td>
<td>00.000°</td>
</tr>
<tr>
<td>D</td>
<td>−15.177°</td>
<td>−36.675°</td>
<td>−06.727°</td>
<td>−27.625°</td>
<td>00.000°</td>
</tr>
<tr>
<td>N</td>
<td>−05.277°</td>
<td>−29.075°</td>
<td>26.073°</td>
<td>−12.500°</td>
<td>00.000°</td>
</tr>
</tbody>
</table>
Geographic areas - Since mid June 2004

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{\text{lon}}$</td>
<td>610</td>
<td>496</td>
</tr>
<tr>
<td>$n_{\text{lat}}$</td>
<td>568</td>
<td>372</td>
</tr>
<tr>
<td>$n_{\text{vert}}$</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Resolution$_{\text{horis}}$</td>
<td>0.15°</td>
<td>0.05°</td>
</tr>
<tr>
<td>$\Delta t_{\text{dynamics}}$</td>
<td>360s</td>
<td>360s</td>
</tr>
<tr>
<td>$\Delta t_{\text{physics}}$</td>
<td>360s</td>
<td>360s</td>
</tr>
<tr>
<td>Forecast length</td>
<td>60h</td>
<td>54h</td>
</tr>
<tr>
<td>Host model</td>
<td>ECMWF</td>
<td>T</td>
</tr>
<tr>
<td>Boundary upd</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Output freq</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lat$_\text{min}$ (south)</th>
<th>lon$_\text{min}$ (west)</th>
<th>lat$_\text{max}$ (north)</th>
<th>lon$_\text{max}$ (east)</th>
<th>lat$_\text{south pole of rotation}$</th>
<th>lon$_\text{south pole of rotation}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>-37.527°</td>
<td>-64.325°</td>
<td>47.523°</td>
<td>27.025°</td>
<td>00.000°</td>
</tr>
<tr>
<td>S</td>
<td>-01.027°</td>
<td>-13.674°</td>
<td>17.523°</td>
<td>11.075°</td>
<td>-40.000°</td>
</tr>
</tbody>
</table>


Still running 3D-VAR; Model now based on HIRLAM Reference system 6.3:

- Analysis of near surface temperature and relative humidity
- Digital filter initialization instead of nonlinear normal mode initialization
- Semi-Lagrangian advection instead of Eulerian advection
- 6th order horizontal diffusion instead of 4th order
- Integrated Soil Biosphere Atmosphere (ISBA) scheme instead of a three layer surface model

Porting of and modifications to HIRLAM Reference system:

- Began porting and adapting HIRLAM Reference system 1 1/2 years ago
- Reference HIRLAM parallelised using MPI
- Some work on OpenMP parallelisation
- Adaptation of our script system
- Grib-Asimof file conversion
- HIRLAM GRIBfile Server (HGS, soon to be operational)
HIRLAM GRIBfile Server (HGS) in pre-operational DMI-HIRLAM

Why: More data with DMI-HIRLAM-T

- 60 hour forecasts run 4 times daily
- 57Mb Interpolated boundary files every 3 hours
- 330Mb Output files every hour
- Time steps involving input and output processing are several times longer than those without

How: HGS and DMI-HIRLAM-T

- Originally written by Jan Boerhout, NEC
- Jussi Heikonen, CSC and Kalle Eerola, FMI, MPI version for output only
- Generalized by Ole Vignes, Norwegian Meteorological Institute
- Optimised by Jan Boerhout, NEC and DMI staff
- Written using Fortran 95
- Asynchronous I/O
- Asynchronous GRIB encoding and decoding
- Presently using 2 MPI tasks for input and output processing

Dramatic improvement: 60 hour forecast wall clock time decreases by roughly 20%
Next step: Implement Jan Boerhout’s optimised version

Performance depends on

- Amount of input and output processing required
- Number of processors used
- Performance of the file system used
- Amount of memory required for buffering files
Operational and scheduling issues

- HIRLAM-T should not start before 1:40 after analysis time and a 36 hour forecast must be available in the grib database 2:15 after analysis time
- Presently we use 3 nodes for HIRLAM production runs
- We want to utilise these 3 nodes for running other applications when not running operational HIRLAM
- We want to use the cluster as efficiently as possible
The queueing system NQS-II and scheduler ERS-II

The queueing system starts operational jobs immediately, but:

- Time critical, operational queues not controlled by the scheduler
- For multinode operational runs, we must specify node numbers as arguments to the qsub command
- Production job suspends other running jobs
- Recently solved problem: Jobs submitted to queue(s) in which job(s) has been suspended will remain queued until the suspended job is resumed

Future set-up?:

- All queues controlled by the ERS-II scheduler
- Production job needn’t suspend other jobs
- Production queues’ priorities much higher than other queues; priorities can be from 1.0 to 100.0
- Looks promising, but start up not instantaneous
- Tuning required to ensure priorities will not be affected by past usage
- Have yet to test this across the entire cluster
• NEC SX-6 cluster
• Use of NEC SX-6 cluster
• DMI-HIRLAM on our NEC SX-6
• Queueing system and scheduler