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

Outline

- 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 - the Danish Meteorological Institute

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

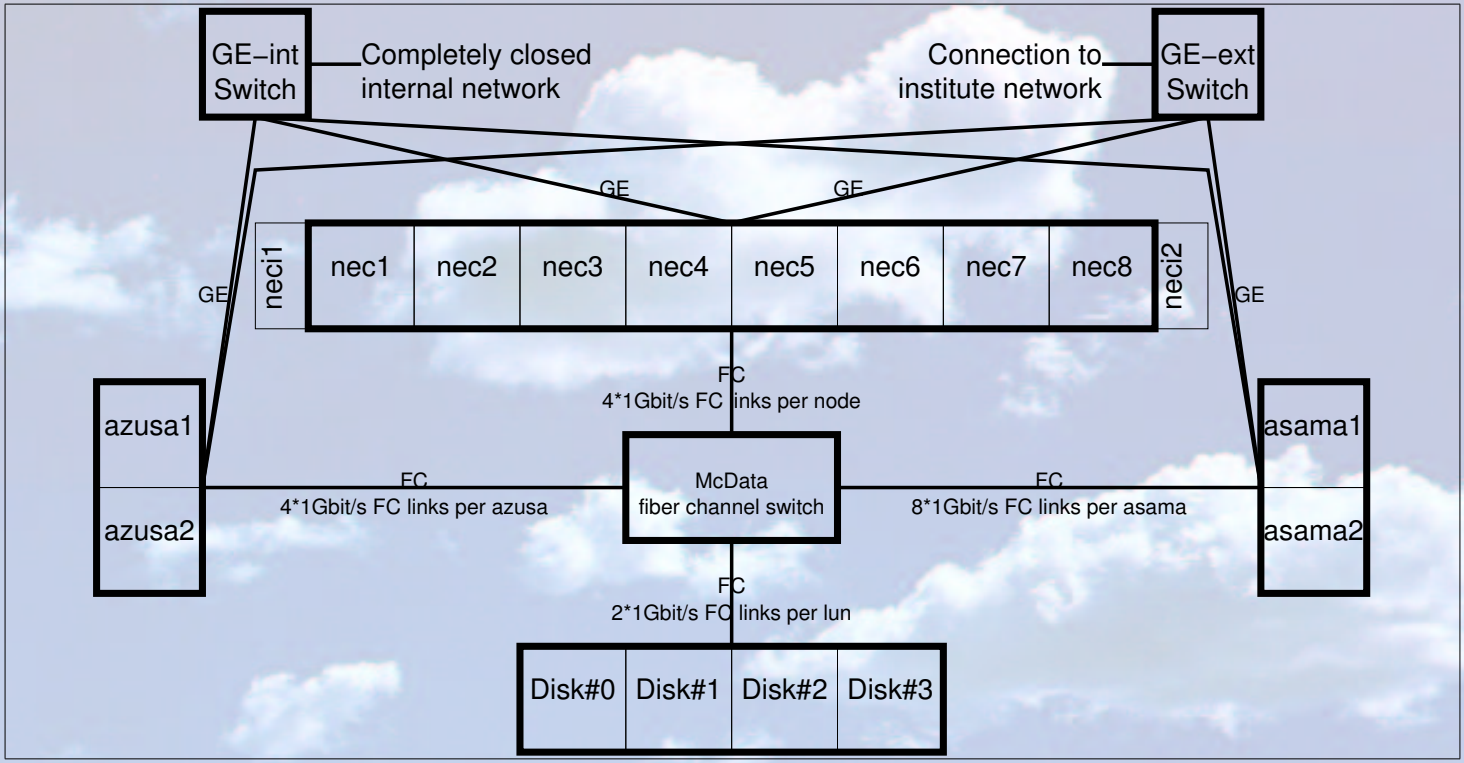
Operational usage:

- 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 interconnect



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

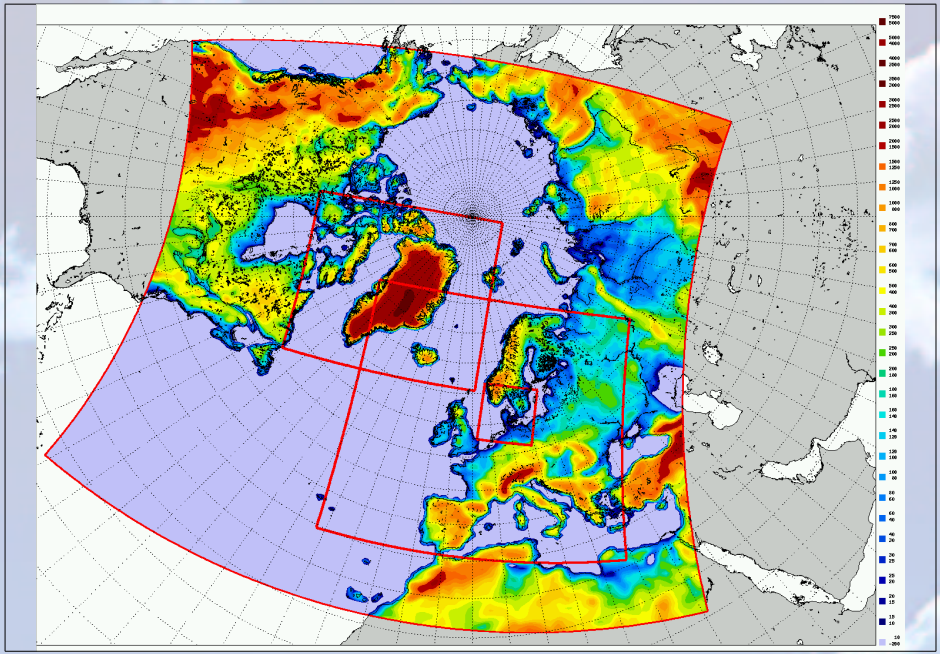
Research usage:

- 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

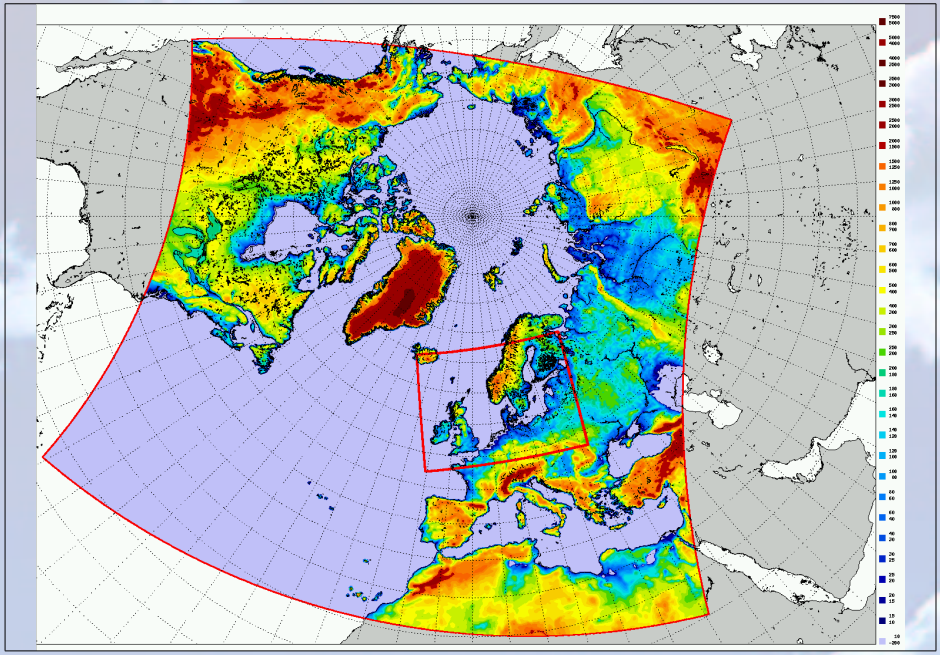
Geographic areas - Through mid June 2004



	G	E	D	N
n_{lon}	202	272	182	194
n_{lat}	190	282	170	210
n_{vert}	40	40	40	40
$\text{Resolution}_{\text{horis}}$	0.45°	0.15°	0.05°	0.15°
$\Delta t_{\text{dynamics}}$	150s	60s	25s	60s
$\Delta t_{\text{physics}}$	450s	360s	150s	360s
Forecast length	60h	54h	36h	36h
Host model	ECMWF	G	E	G
Boundary upd	3	1	1	1
Output freq	1	1	1	1

	lat_{min} (south)	lon_{min} (west)	lat_{max} (north)	lon_{max} (east)	$\text{lat}_{\text{south pole of rotation}}$	$\text{lon}_{\text{south pole of rotation}}$
G	−37.525°	−63.725°	47.523°	26.725°	00.000°	80.000°
E	−28.677°	−54.275°	13.473°	−13.625°	00.000°	80.000°
D	−15.177°	−36.675°	−06.727°	−27.625°	00.000°	80.000°
N	−05.277°	−29.075°	26.073°	−12.500°	00.000°	80.000°

Geographic areas - Since mid June 2004



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

	lat_{min} (south)	lon_{min} (west)	lat_{max} (north)	lon_{max} (east)	$\text{lat}_{\text{south pole of rotation}}$	$\text{lon}_{\text{south pole of rotation}}$
T	−37.527°	−64.325°	47.523°	27.025°	00.000°	80.000°
S	−01.027°	−13.674°	17.523°	11.075°	−40.000°	10.000°

DMI-HIRLAM

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 horisontal 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%

HIRLAM GRIBfile Server (HGS) in pre-operational DMI-HIRLAM, cont.

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

Summary

- NEC SX-6 cluster
- Use of NEC SX-6 cluster
- DMI-HIRLAM on our NEC SX-6
- Queueing system and scheduler