The Irish Meteorological Service

mogall

Running the Hirlam NWP Model

on a 6 x dual-xeon Linux cluster

at Met Éireann (Irish Meteorological Service)

(James Hamilton -- Met Éireann)

MET éireann

The Irish Meteorological Service

MET éireann

NWP COMPUTER SYSTEMS SINCE 1978

DEC TOPS-2050

- Scalar system – 1 cpu

SGI Challenge-L with two processors

- Parallel system – shared memory – 2 cpu

SGI PowerChallenge with six processors

- Parallel system – shared memory – 6 cpu

IBM RS/6000 SP : 9-nodes each with 4-CPU's

- Parallel system – distributed memory – MPI – 36 cpu

LINUX Cluster : Six Twin Xeon Processors [3.2GHz]

- Parallel system – distributed memory – MPI – 12 cpu

The Irish Meteorological Service

OPERATIONAL HIRLAM … IBM RS/6000 SP

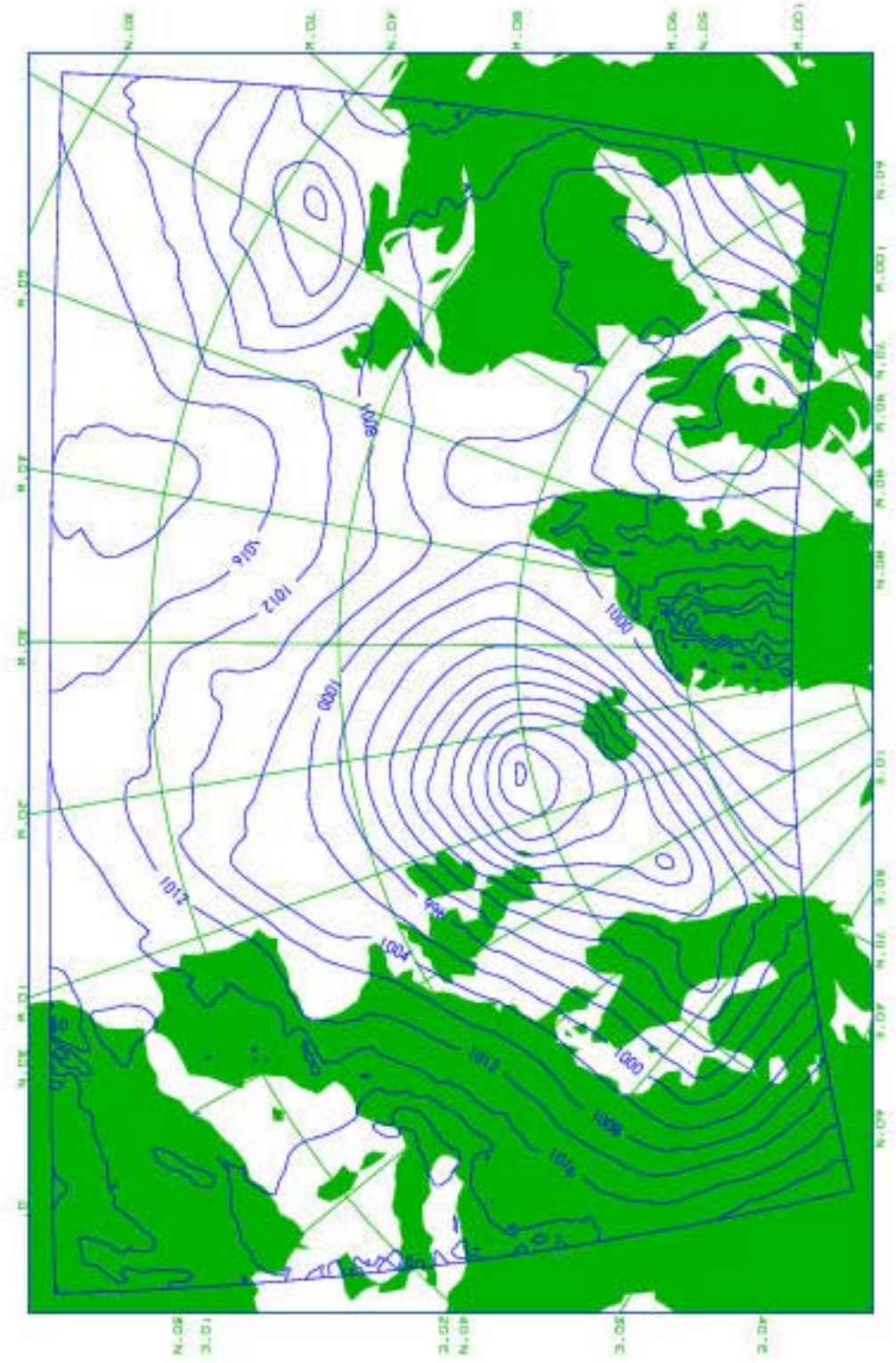
HIRLAM 5.0.1 with 3DVAR

- 3-Hour assimilation cycle with 48-hour forecasts every 6-hours
- Rotated lat/long 0.15x0.15 grid with 438x284 grid points
- Hybrid [eta] coordinates with 31-levels
- CBR vertical diffusion scheme; Sundqvist condensation scheme
- STRACO cloud scheme; Savijarvi radiation scheme
- Digital filter initialisation
- Two time-level Semi-Lagrangian semi-implicit scheme
- Use of ‘frame’ boundaries from ECMWF

MET éireann

MET éireann

The Irish Meteorological Service



The Irish Meteorological Service

OPERATIONAL COMPUTER SYSTEM

IBM RS/6000 SP – WinterHawk

Compute Nodes : 9

- Each has 4 cpu's – total of 36 cpu's
- Each has 2 Gigabytes of memory – total of 18 Gigabytes

Master Node : 1

- Node has 2 cpu's
- Node has 4 Gigabytes of memory

SPS Network Switch / 375MHz Power3-II processors

MET éireann

The Irish Meteorological Service

EXPERIMENTAL COMPUTER SYSTEM

Linux Cluster : Dell Poweredge 1750 : Twin Xeon

Compute Nodes : 6

- Each has 2 cpu's – total of 12 cpu's
- Each has 2 Gigabytes of memory – total of 12 Gigabytes

Master Node : 1

- Node has 2 cpu's
- Node has 4 Gigabytes of memory

Dolphin SCI HBA 4-port network card [2d torus]

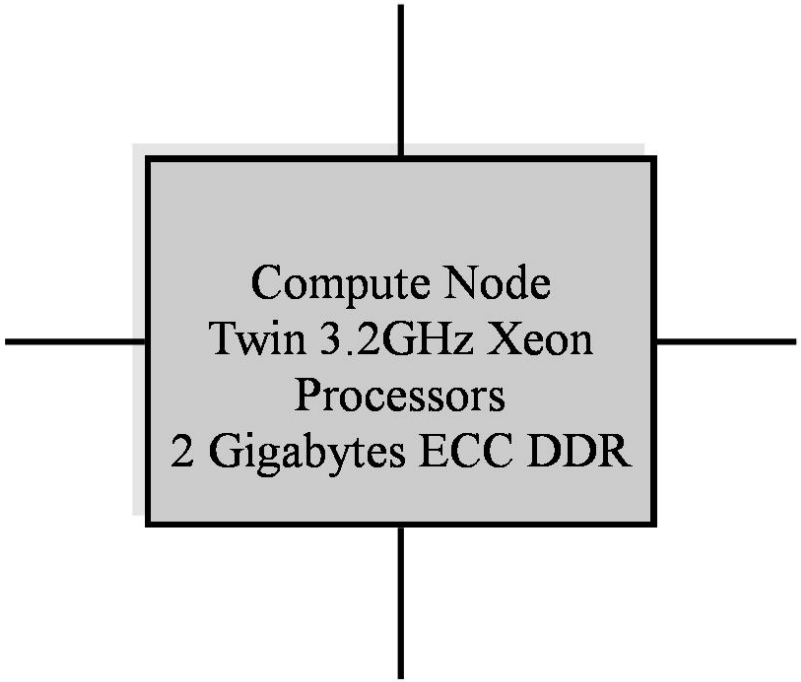
MET éireann

The Irish Meteorological Service

EXPERIMENTAL COMPUTER SYSTEM

Linux Cluster : Dell Poweredge 1750

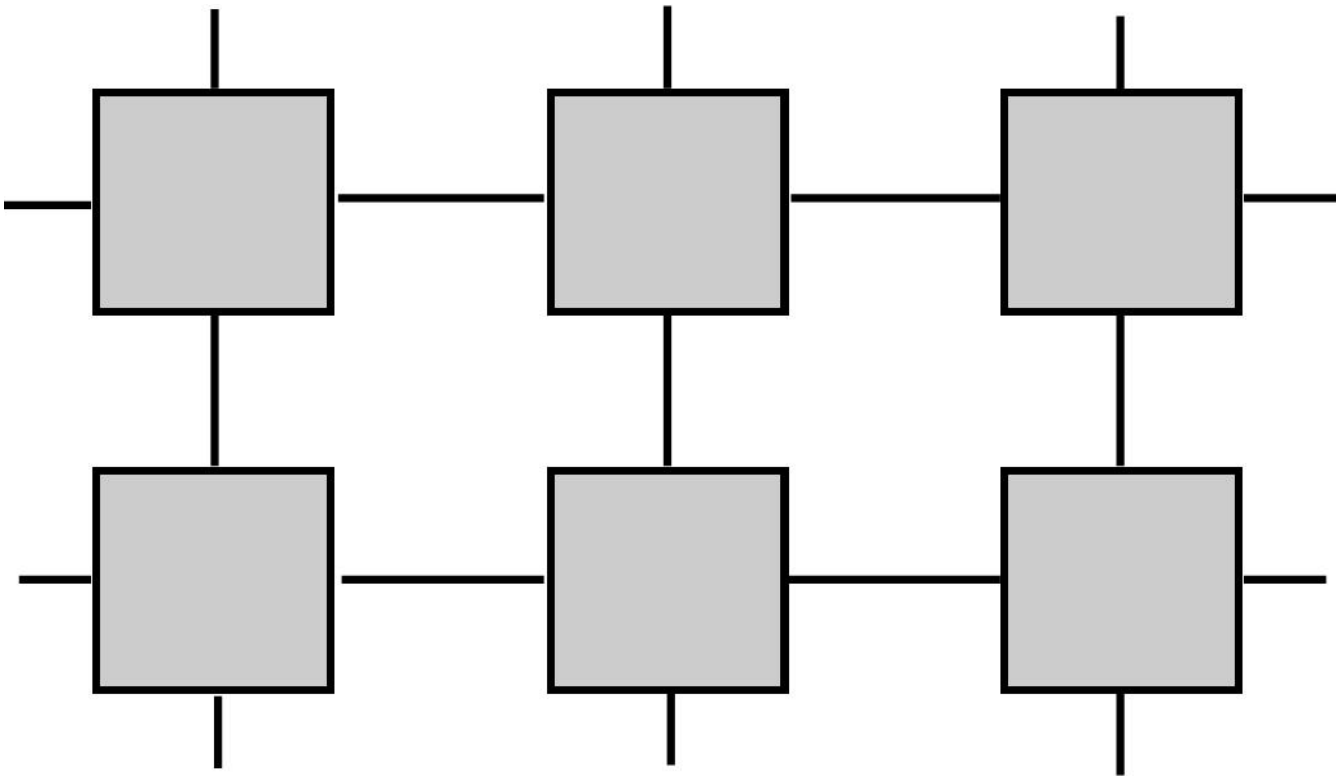
Dolphin SCI HBA 4-port network card



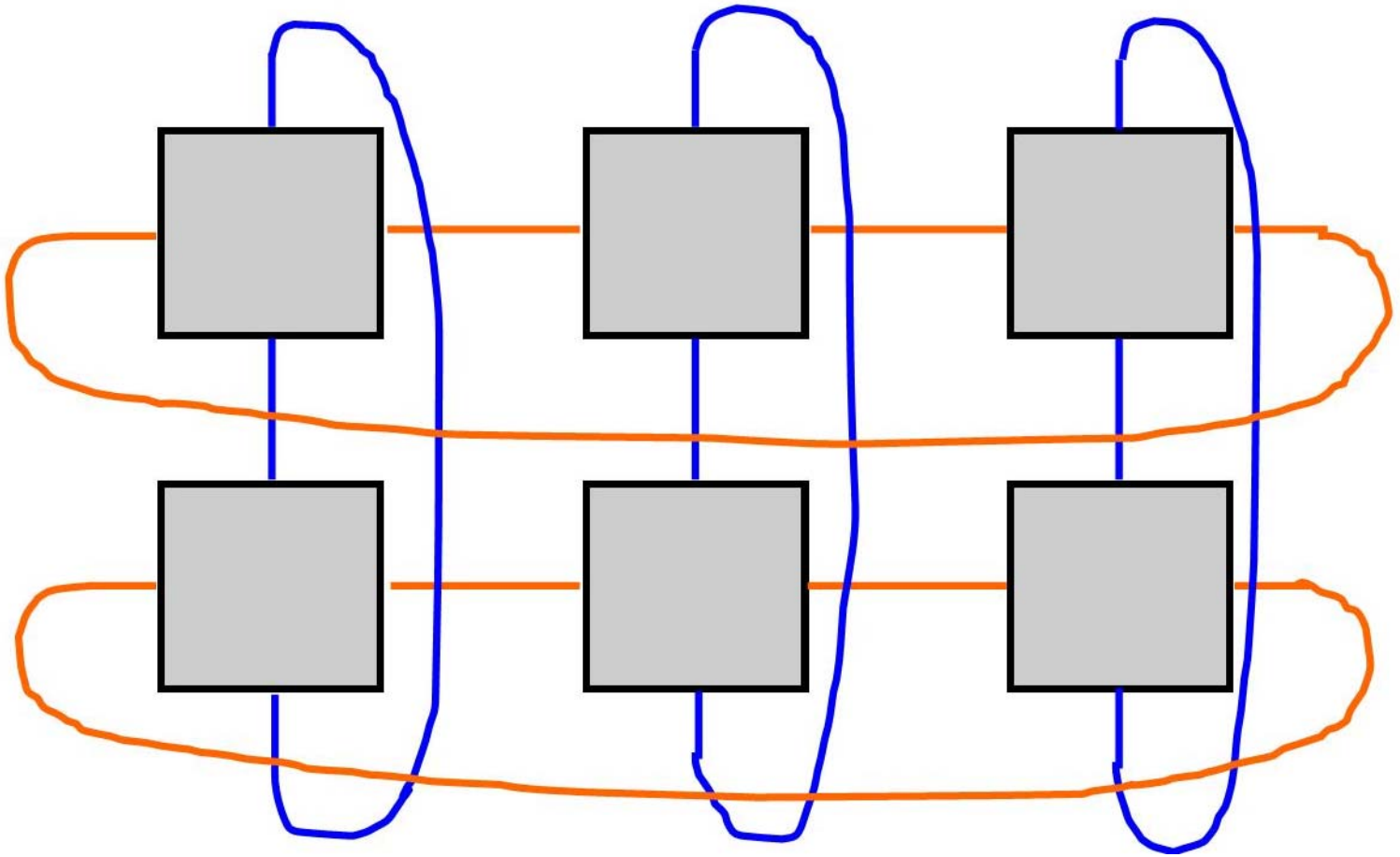
MET éireann

LINUX CLUSTER – 2-dimensional torus

Each node connected to 4 other nodes via Dolphin cards



LINUX CLUSTER – 2-dimensional torus



The Irish Meteorological Service

CLUSTER SYSTEM : HARDWARE

Dell Poweredge 1750

- 6 x Twin Intel Xeon 3.2 GHz compute nodes
- 1 x Twin Intel Xeon 2.8 GHz master node
- Rack mounted
- Gigabit ethernet for all 7 nodes [6-compute, 1-master]
- Local disks, CD-writer, tape-drive

Dolphin Network cards

- Two-dimensional torus linking compute nodes

MET éireann

The Irish Meteorological Service

CLUSTER SYSTEM : SOFTWARE REQUIREMENTS

Operating System

- Reliable Linux operating system with commercial support

Compilers

- Hirlam needs Fortran-90 [will not work with g77]

MPI Software

- Reliable MPI system for Dolphin network cards [and ethernet]

The Irish Meteorological Service

CLUSTER SYSTEM : SOFTWARE

Operating System

- Redhat ES 3.0 on compute nodes
- Redhat WS 3.0 on master node

Compilers

- PGI cluster development kit [includes Fortran]
- Intel Fortran Compiler

MPI Software

- Scali MPI connect [Dolphin network cards]
- Scali TCP connect [Ethernet]
- Scali Manage

MET éireann

The Irish Meteorological Service

INSTALLING HIRLAM ON CLUSTER … I

Compiling Hirlam : Forecast Model

- Hirlam already supports MPI via pre-processor flag [MPI_SRC]
- Needs to link with scali MPI libraries

Compiling Hirlam : 3DVAR analysis

- 3DVAR supports MPI via pre-processor flag
- Various IBM optimisations must be switched off [via flags]
- Needs to link with scali MPI libraries

Running the Hirlam programs : 12-cpu's

- /opt/scali/bin/mpirun –np 12 […hirlam program…]

MET éireann

The Irish Meteorological Service

INSTALLING HIRLAM ON CLUSTER … II

HDF [Hierarchical Data Format] used by Hirlam

- Precompiled version would not work with RedHat ES/WS 3.0
- Recompiling solved this problem

Compiling Hirlam : 3DVAR analysis

- 3DVAR supports MPI via pre-processor flag
- Various IBM optimisations must be switched off [via flags]
- Needs to link with scali MPI libraries

Compiler Comparison

- So far have only used PGI compilers
- Plan to use Intel compilers

MET éireann

INSTALLING HIRLAM ON CLUSTER … III

Initialising Jobs, Sharing Data

- Compilations etc. are done on the master node
- The master node disks are nfs-mounted on the compute nodes
- Each job is started on the master node
- Compute nodes each have their own disk with a copy of Linux

Job Control – selecting nodes

- Scali software allows specification of nodes and processes / node
- Can select one cpu or two cpu per node
- Can run (say) 6 cpu job on 3-nodes or 6-nodes [or mixture]

The Irish Meteorological Service

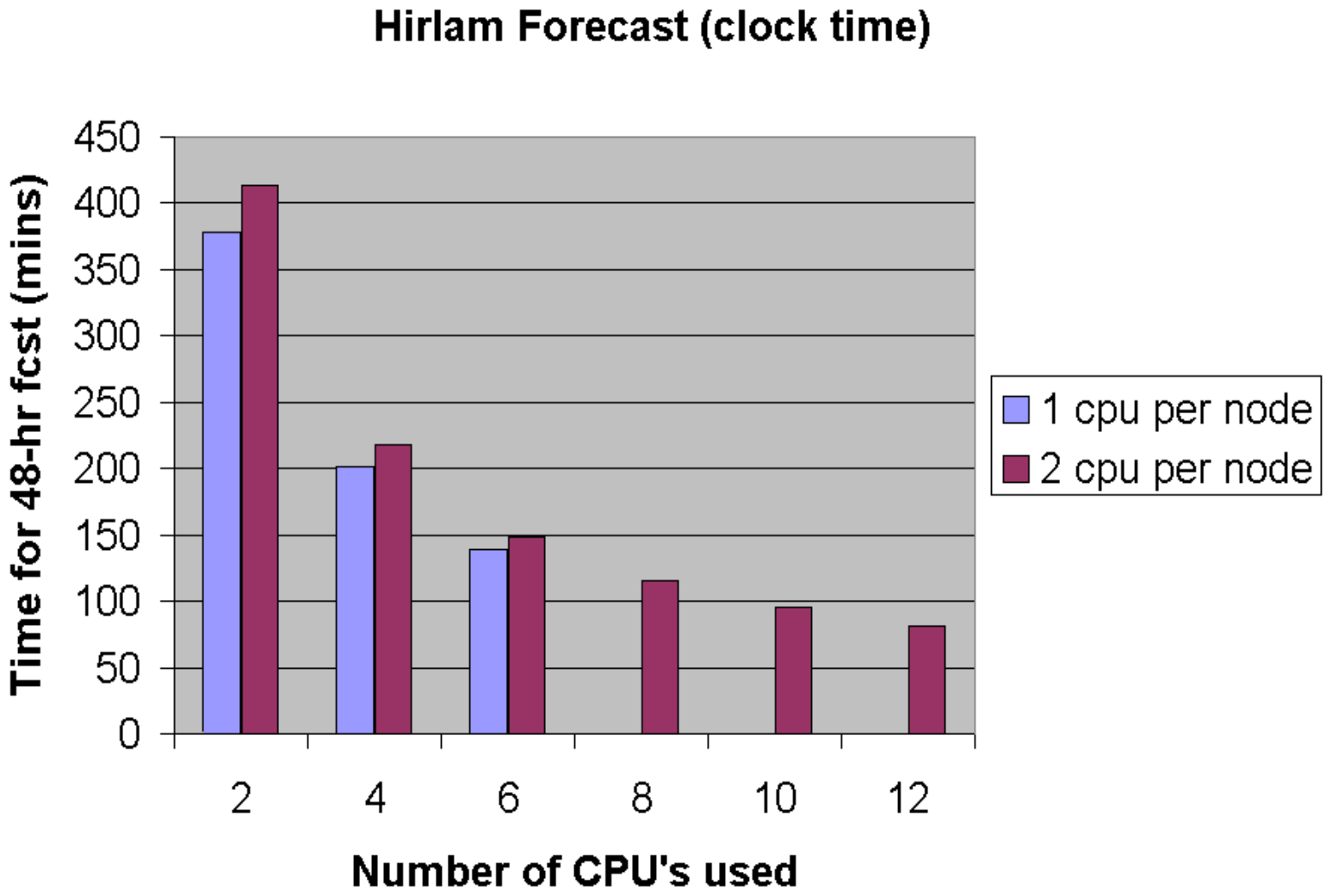
TIMING RESULTS for HIRLAM Model

HIRLAM forecast model

- Results for different numbers of processors
- Model runs as sole job on entire cluster
- Can use one cpu or two cpu per node
- Results obtained using PGI compilers
- Dolphin network cards / Scali MPI

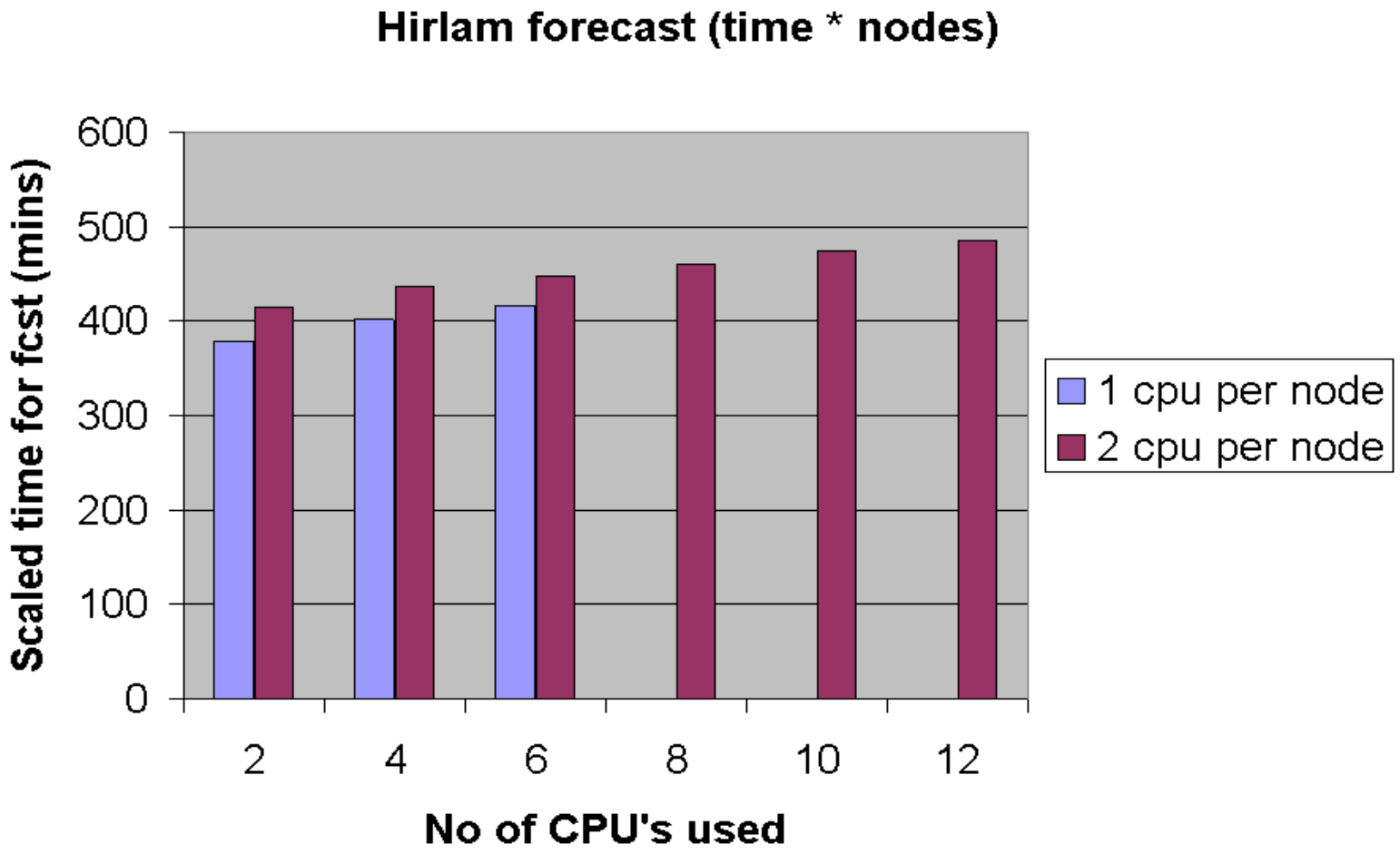
MET éireann

The Irish Meteorological Service



MET éireann

The Irish Meteorological Service



MET éireann

TIMING RESULTS for HIRLAM Model

Conclusions

- Using 1-cpu per node is faster than 2-cpu per node
- Only small loss of efficiency with more nodes
- Full cluster takes 81 mins vs 62 mins for IBM

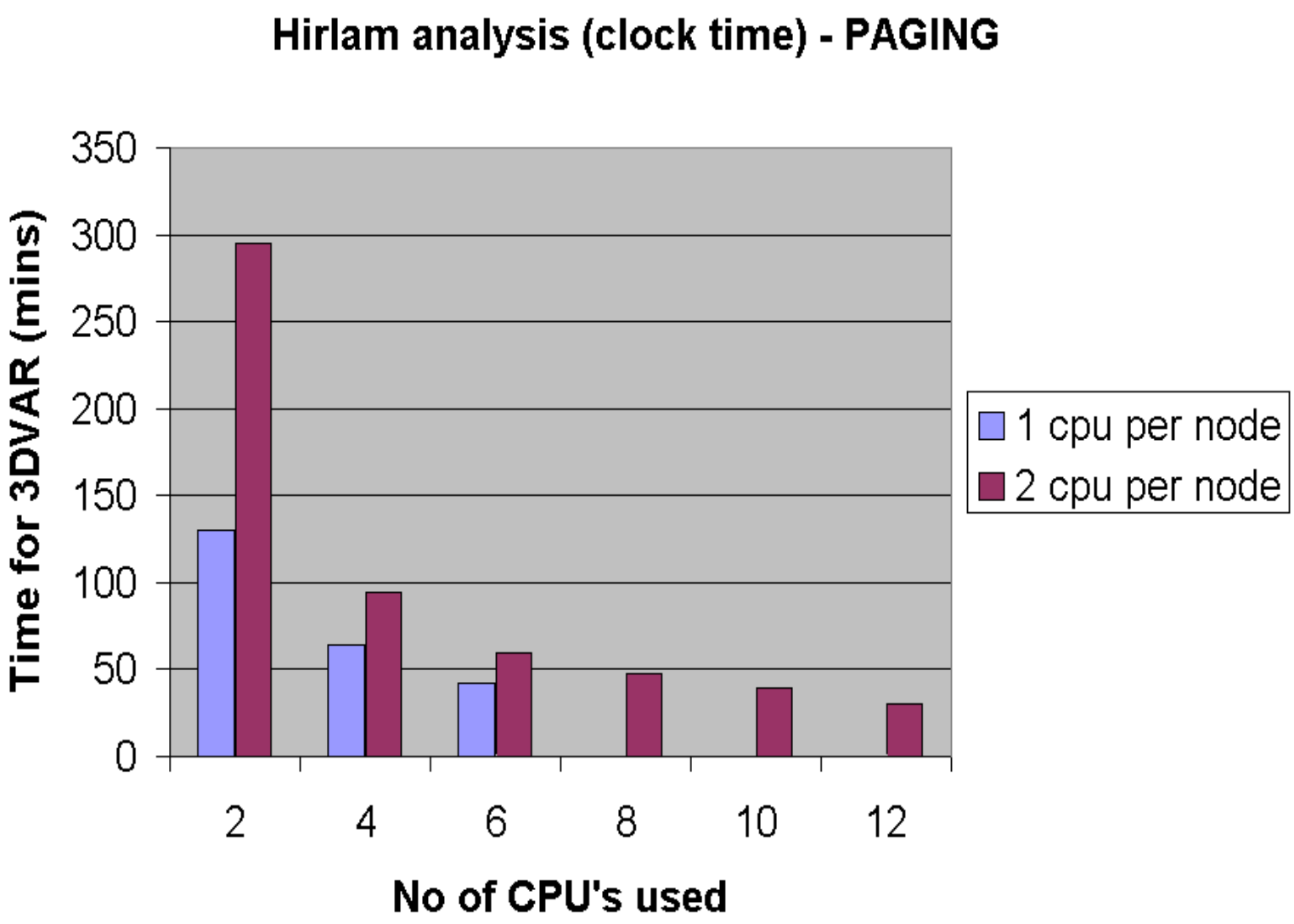
TIMING RESULTS for HIRLAM 3DVAR Analysis

HIRLAM 3DVAR analysis

- Results for different numbers of processors
- Analysis runs as sole job on entire cluster
- Can use one cpu or two cpu per node
- Results obtained using PGI compilers
- Many IBM optimisations [special library calls] removed
- Dolphin network cards / Scali MPI

HIRLAM 3DVAR PAGING

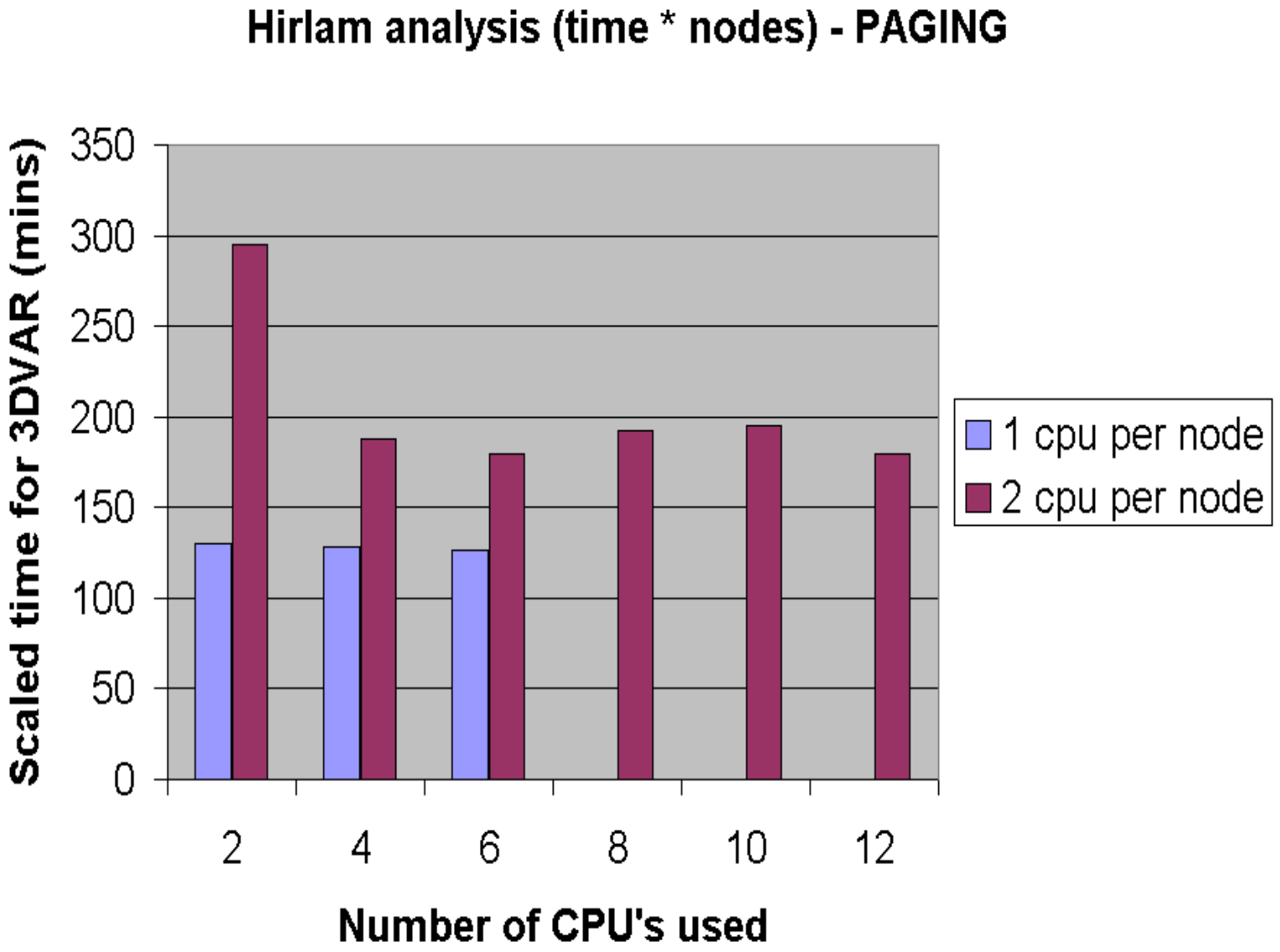
- Running with two processors on one node gives PAGING
- Increasing memory per node should fix this


Hirlam analysis (clock time) - PAGING
Time for 3DVAR (mins)
350
300
250
200
150
100
50
0
2
4
6
8
10
12
No of CPU's used
1 cpu per node
2 cpu per node

MET éireann

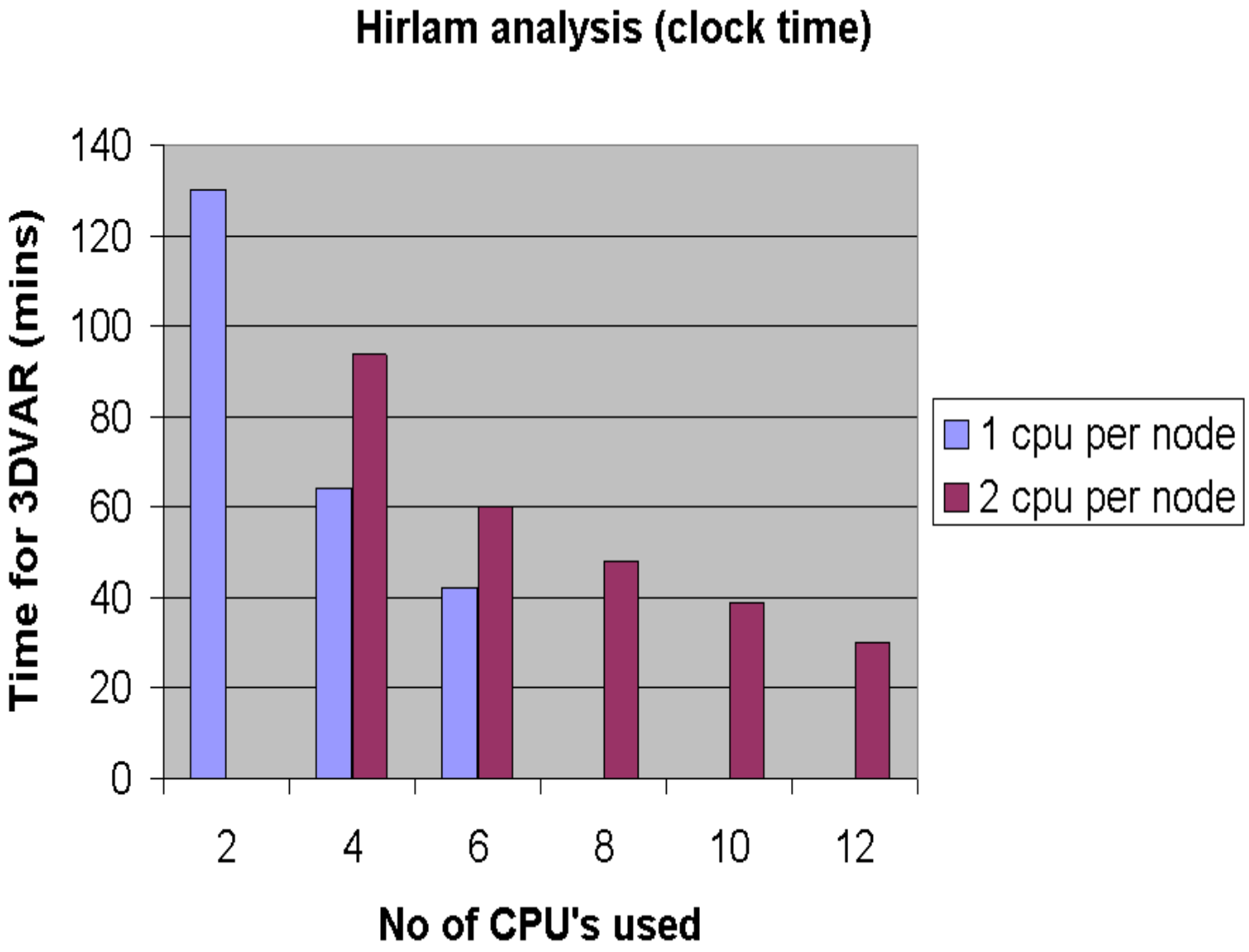
The Irish Meteorological Service

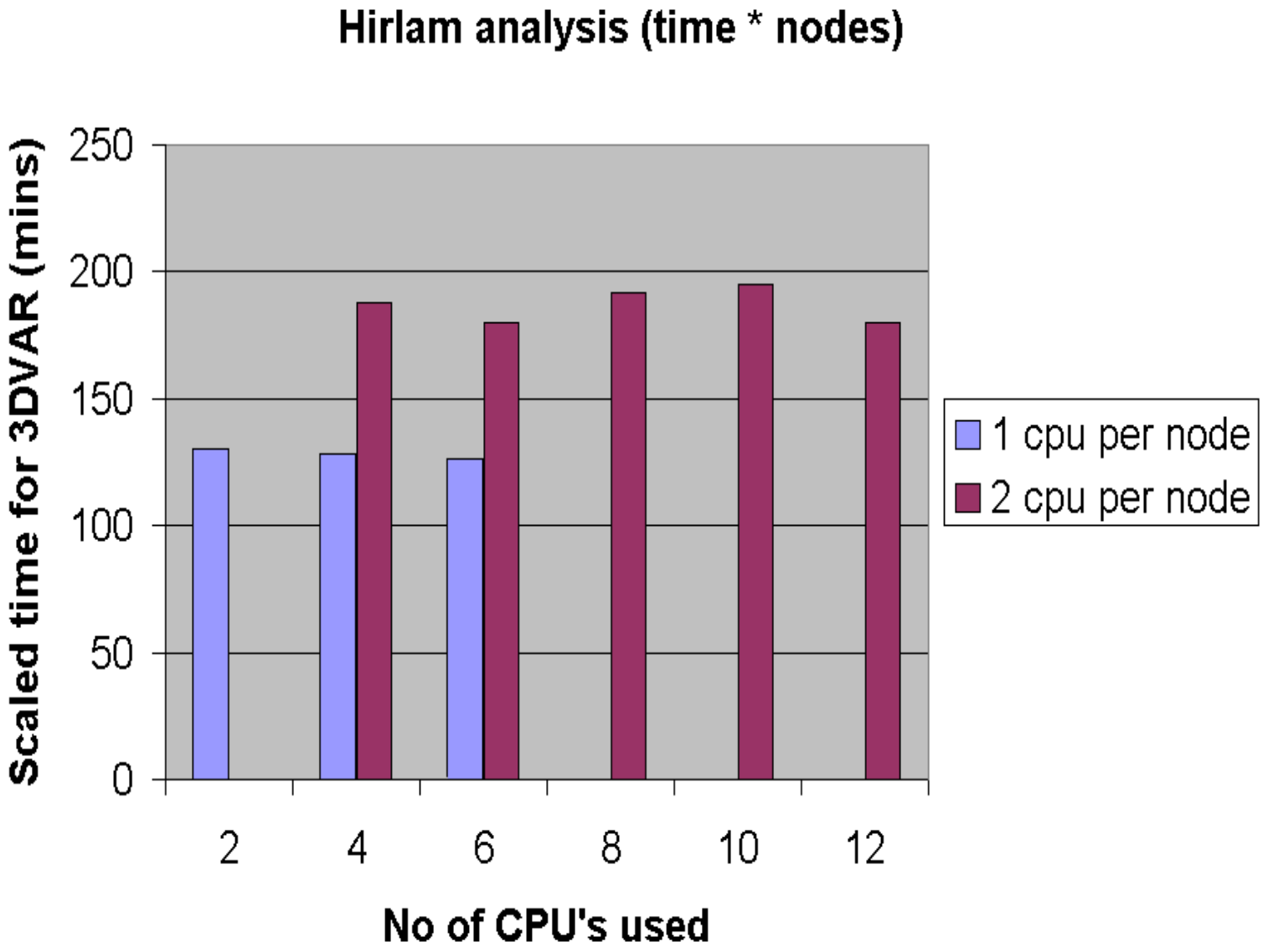
MET éireann



The Irish Meteorological Service

MET éireann




Hirlam analysis (time * nodes)
Scaled time for 3DVAR (mins)
250
200
150
100
50
0
2
4
6
8
10
12
No of CPU's used
1 cpu per node
2 cpu per node

MET éireann

The Irish Meteorological Service

TIMING RESULTS for HIRLAM 3DVAR Analysis

Conclusions

- Using 1-cpu per node is faster than 2-cpu per node
- No loss of efficiency with more nodes
- Full cluster takes 30 mins vs 15 mins for IBM

MET éireann

The Irish Meteorological Service

MET éireann

GENERAL CONCLUSIONS

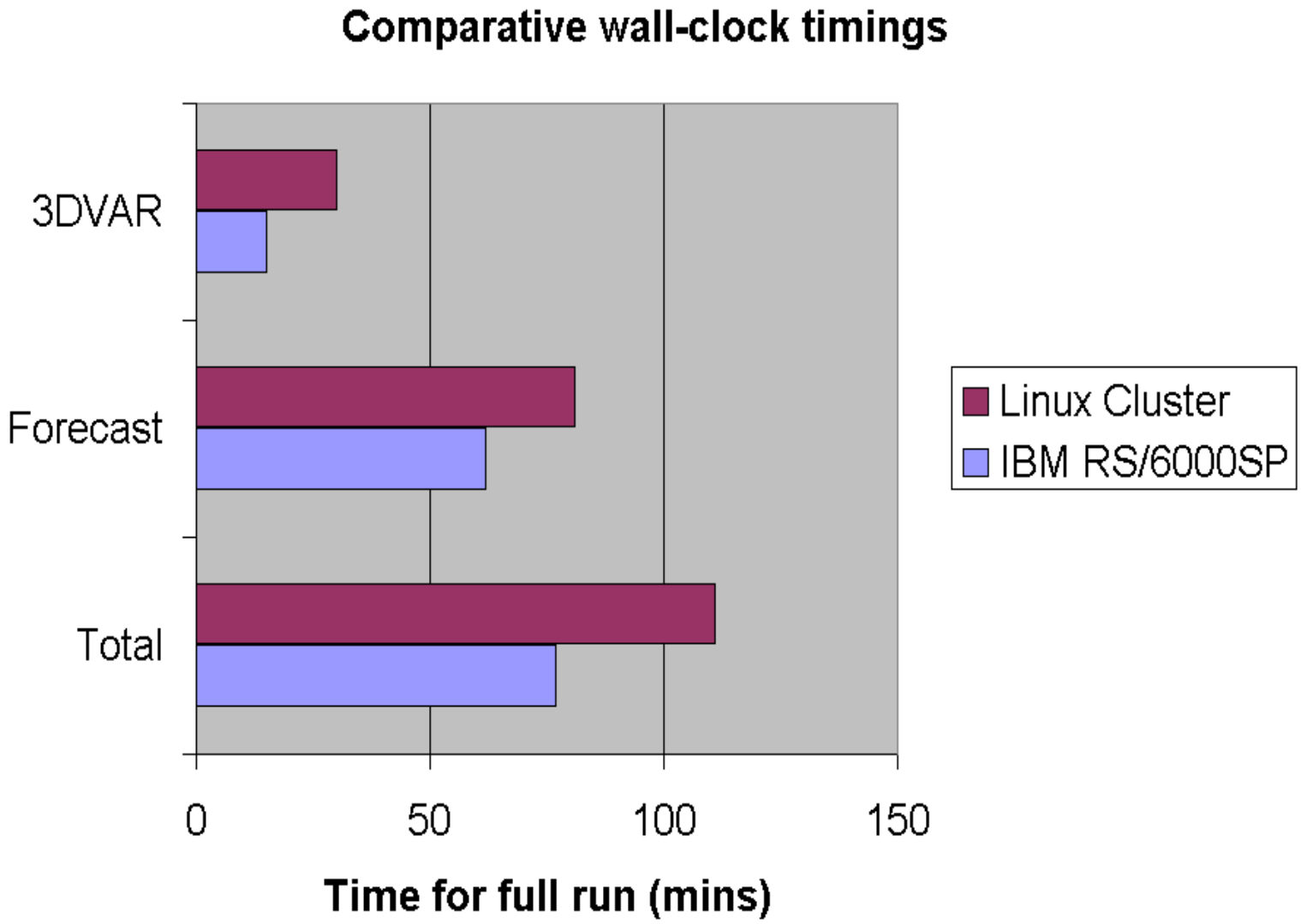
Analysis [3DVAR] is more cpu bound than forecast

- Using 1-cpu per node is faster than 2-cpu per node in both cases
- Difference is larger for 3DVAR
- Some loss of efficiency for forecast as number of cpu's increase
- With 3DVAR little or no loss in efficiency

Analysis [3DVAR] is less efficient than forecast

- Forecast takes 81 mins versus 62 on IBM
- Analysis takes 60 mins versus 15 on IBM
- Analysis on IBM uses special FFT libraries and maths libraries

The Irish Meteorological Service



MET éireann

The Irish Meteorological Service

FINAL CONCLUSIONS

Cluster has been very reliable …

- Cluster has run for over a year with no breakdowns
- Hardware / software have been of high-quality and easy to use

Plans to upgrade cluster …

- We have decided to purchase 3 more nodes [should equal IBM]

Plans to upgrade mainframe …

- We hope to replace our mainframe in 2006

Development plans for cluster …

- Upgrade Hirlam to latest version
- Investigate Intel compiler and faster maths libraries [e.g. fftw]

MET éireann

The Irish Meteorological Service

ACKNOWLEDGEMENTS

Colleagues at Met Eireann …

- James Brennan, Shay McLoughlin

Hirlam community ….

- Hirlam is MPI enabled

Help with specifying the cluster ….

- Lars Muller – SHMI [Norkopping]
- Niclas Andersson, Torgny Faxen – NSC [Linkopping]
- Aarne Mannick, Gerald Cats, Per Under [Hirlam]

Help with the installing cluster ….

- David Hutton – Scali / Greg Moore – Dell

MET éireann

The Irish Meteorological Service

EXPERIMENTS with LINUX CLUSTER …

Running Operational Hirlam on Cluster [6-node]

- Hirlam 5.0.1 with 3DVAR – identical to operational system
- Forecast takes 62 mins on IBM/SP; 81 mins on cluster
- 3DVAR takes 15 mins on IBM/SP; 30 mins on cluster
- Full run takes 77 mins on IBM/SP; 111mins on cluster
- [3DVAR highly optimised for IBM/SP e.g. FFT routines]
- Expect 9-node cluster to be as fast as IBM/SP

Experience with Cluster Very Positive

MET éireann

The Irish Meteorological Service

Thank You

MET éireann