Scalability of the Met Office Unified Model

Andy Malcolm, Maff Glover & Paul Selwood
Contents

Table of Contents

• HPC at the Met Office
• What is the UM and what is it used for
• UM atmosphere forecast scalability results
• Coupled model results
• Recent improvements
• Conclusions
1989-2003 : Cray YMP,C90,T3E

2003-2008 : NEC SX6/8  ~5TFlop peak

2009-12 : IBM p575 Power6

- Operational from August 2009
- 145 TFlop peak capacity (7744 cores)
- 2 identical systems (2*106 node) for resilience plus small system (30 node) for Collaboration with UK Universities

2012-> : IBM Power 7

- ~3 faster than Phase 1 measured by benchmark application speedup
- At least 25000 cores with total Capacity approaching 1PFlop
The Unified Model
The Unified Model

• Supports all atmospheric modelling. Spatial and temporal scales cover climate and seasonal requirements through to global and local weather prediction requirements

Climate modelling: input into IPCC reports
(Coupled Atmosphere-Ocean models)
1 year – 100 year, low resolution

Seasonal forecasting:
(Coupled Atmosphere-Ocean models)
For commercial and business customers
1 month - 1 year low resolution

NWP
Atmosphere model
Public Weather, Aviation, Commercial
6 hours to 2 weeks high resolution
DePreSys

TIGGE

ensemble

GloSea4

MOGREPS-R

ensemble

Coupled atmos/ocean

Earth System

Timescale

36hrs 48hrs 5 days 15 days 6 months 10 years 30 years >100 years

UKV

1.5km

4km

12km

24km

40km

80km

150km

300km

Atmospheric grid length

Met Office

1.5km

4km

12km

24km

40km

80km

150km

300km

UK4

NAE

MOGREPS-R ensemble

Global

GloSea4

TIGGE ensemble

DePreSys

HadCM3

HadGEM1

HadGEM2

HadGEM3

HadGEM3-RA

Regional atmosphere-only

Coupled atmos/ocean

Global atmosphere-only

Regional atmosphere-only

36hrs 48hrs 5 days 15 days 6 months 10 years 30 years >100 years

In transition to Production

Production system

Atmospheric grid length

Timescale

Complexity

Global

Regional

Atmospheric grid length

In transition to Production

Production system
Met Office NWP production system
Operational NWP Models: Nov 2010

Global
- 25km 70L
- 4DVAR
- 60h forecast twice/day
- 144h forecast twice/day
- +24member EPS at 60km

NAE
- 12km 70L
- 4DVAR
- 60h forecast
- 4 times per day
- +24member EPS at 18km

UKV
- 1.5km 70L (variable resolution)
- 3DVAR (hourly)
- 36h forecast
- 4 times per day
Scalability (N512 global) – Mar 2010

N512 Scaling

- N512 Full I/O
- N512 No I/O
- Ideal
Global Model Dynamics Problems

- Lat-Long grid causes problems
- ADI preconditioner scales poorly
- Communication on demand in the advection is fairly costly and introduces imbalance
- Polar filtering is communication dominated and imbalanced
- Polar re-mapping in wind advection introduces load imbalance
- Constant pole requirement introduces communication
Machine comparison

Through the PRACE initiative we were able to compare the UM on IBM Power 6 and an Intel Nehalem cluster (Juropa).

<table>
<thead>
<tr>
<th></th>
<th>IBM Power 6</th>
<th>Juropa (Intel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores per Node</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Clock Frequency</td>
<td>4.7 GHz</td>
<td>2.93 GHz</td>
</tr>
<tr>
<td>Interconnect</td>
<td>DDR Infiniband</td>
<td>QDR Infiniband</td>
</tr>
<tr>
<td>Filesystem</td>
<td>GPFS</td>
<td>Lustre</td>
</tr>
</tbody>
</table>

© Crown copyright    Met Office 2010
PRACE results (MPI & OpenMP)
PRACE results - percentage difference to IBM MPI only
Coupled models
HadGEM3-AO components

- Atmosphere (Unified Model)
- Ocean (NEMO)
- Sea Ice (CICE)
- Coupler (Oasis3)

One executable

Used for climate integrations, seasonal forecasting (GloSea4)
NEMO Scalability
NEMO – Scalability curves

Scalability of NEMO/CICE

0.25 degree resolution

Speedup

IBM Power6 nodes

© Crown copyright   Met Office 2010
NEMO – comparison with MPI

NEMO-CICE 0.25 degree resolution
Percentage change relative to MPI-only case

© Crown copyright   Met Office 2010
Load balancing and all that

- Component speed depends on
  - Cores given
  - Number of threads
  ... and more ...

- Coupled model speed
  - Only runs as fast as the slowest component
  - Don’t want one component waiting for another
  - During optimisation work, constant need to rebalance.
An extra dimension ...
Coupled model scaling

Scaling of HadGEM3-AO

Top-performing Atm/Ocean balance, 8 coupling tasks
Individual components

Scaling of Individual Model Components
Inset: nodes for individual cmpts vs. coupled model total

- Atmosphere
- Ideal
- Ocean

Coupled turnover

© Crown
Recent improvements
QPOS (moisture/tracer reset)

- Current algorithm gathers information onto 1 processor, does work, then scatters data.
- Anti-scales.
- Alternative algorithms coded up
- Used in PS24 – saved 7 minutes on an operational forecast

© Crown copyright  Met Office 2010
• Old algorithm anti-scales
• Correction over orography includes 6 gathers to pe0, 20 iterations of an SOR solver and a scatter
• Revised algorithm coded – uses Jacobi algorithm.
• Can use many more iterations (100’s) and still be cheaper.
Original scalability (N512) – Mar 2010
Improved scalability (N512) – Nov 2010

The diagram shows the relationship between speedup and the number of cores for different configurations:

- **Original - Full I/O**
- **Improved - Full I/O**
- **Original - no I/O**
- **Improved - no I/O**
- **Ideal**

The graph illustrates how speedup improves with the number of cores, with the ideal configuration showing the best performance.
I/O Server

- In the currently released UM a synchronous Output Server is available (24 hour forecast, 768 processors, run time improves from 933 to 856 seconds)
- We will have an asynchronous Output Server (giving further savings) in the version to be released before Christmas.
UM

- now recommend using 2 threads and SMT on most runs
- ~6% speedup in forecast models. (sometimes more)

NEMO (in GloSea4 coupled model)

- N96L38-O1L42 and
- N96L85-O1L75 with OMP+SMT, run in same time.
  (Free level increase.)
Conclusions
• We have improved scalability and run times for the UM.
• There is still scope for improvement in both global and coupled models.
• The Lat-Long grid causes problems.
• ENDGAME (next dynamical core) hopes to address some of the issues but different grid structures may be needed.
Questions and answers