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

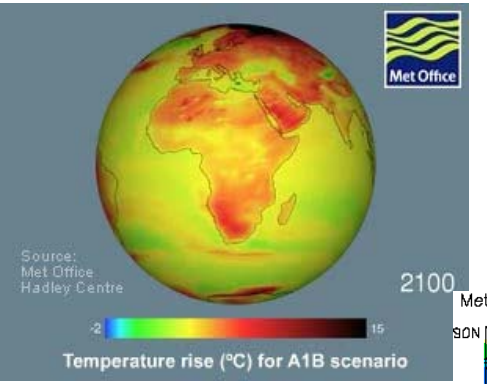
Met Office HPC

- 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

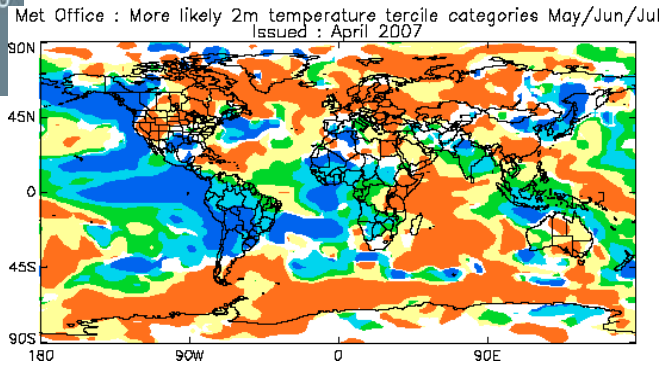
The Unified Model

The Unified Model

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

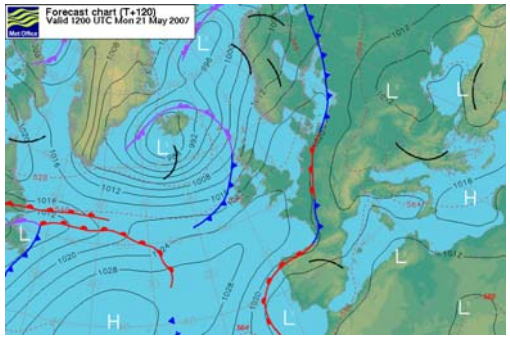


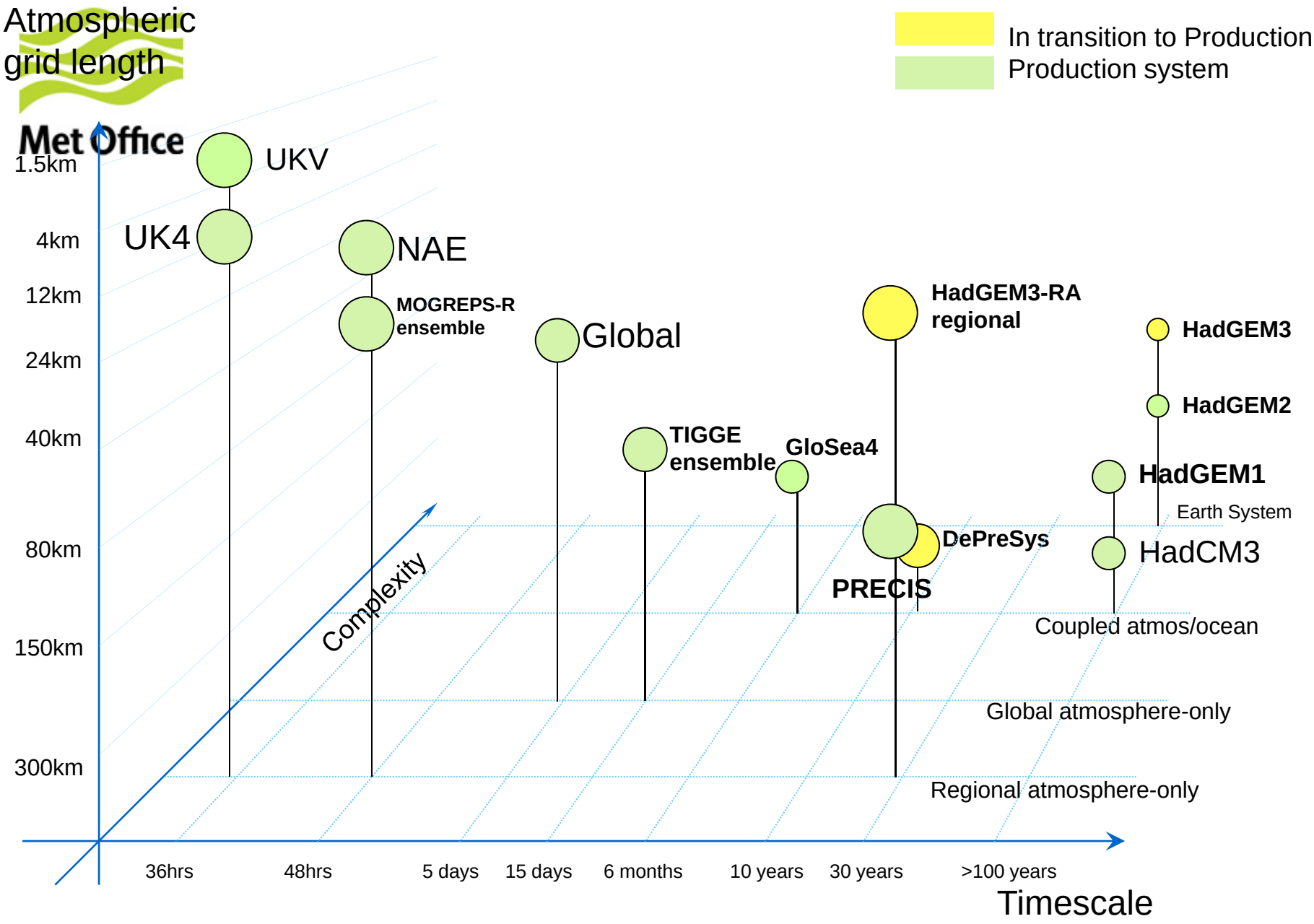
Climate modelling: input into IPCC reports
(Coupled Atmosphere-Ocean models)
1 year – 100 year, low resolution



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




Atmospheric grid length
In transition to Production
Production system
Met Office
1.5km
4km
12km
24km
40km
80km
150km
300km
UKV
UK4
NAE
MOGREPS-R ensemble
Global
TIGGE ensemble
GloSea4
HadGEM3-RA regional
DePreSys
PRECIS
HadGEM3
HadGEM2
HadGEM1
Earth System
HadCM3
Coupled atmos/ocean
Global atmosphere-only
Regional atmosphere-only
Complexity
36hrs
48hrs
5 days
15 days
6 months
10 years
30 years
>100 years
Timescale

Met Office NWP production system

Operational NWP Models: Nov 2010

Met Office

Global

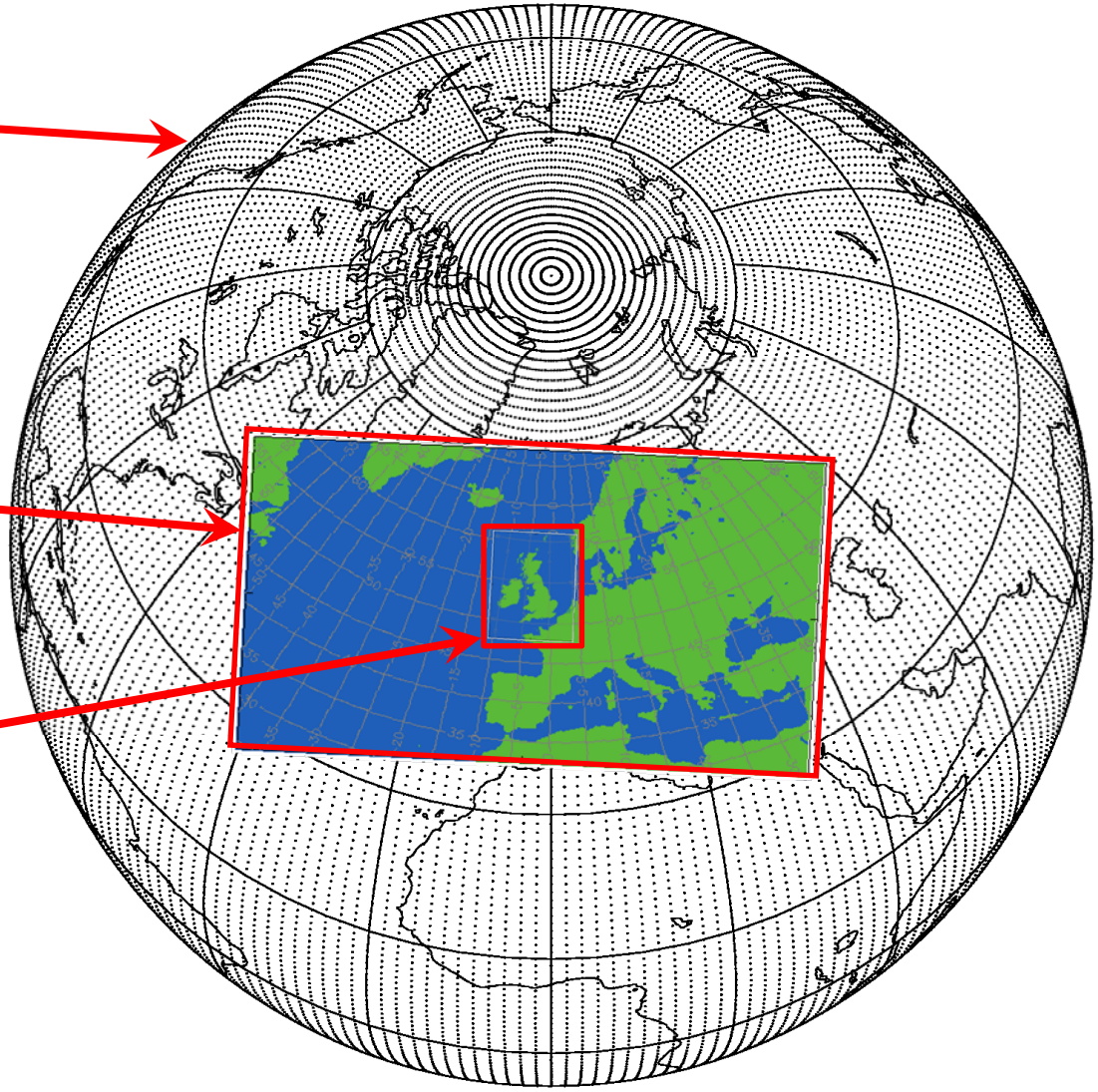
- **25km 70L**
- **4DVAR**
- **60h forecast twice/day**
- **144h forecast twice/day**
- *+24member EPS at 60km*

NAE

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

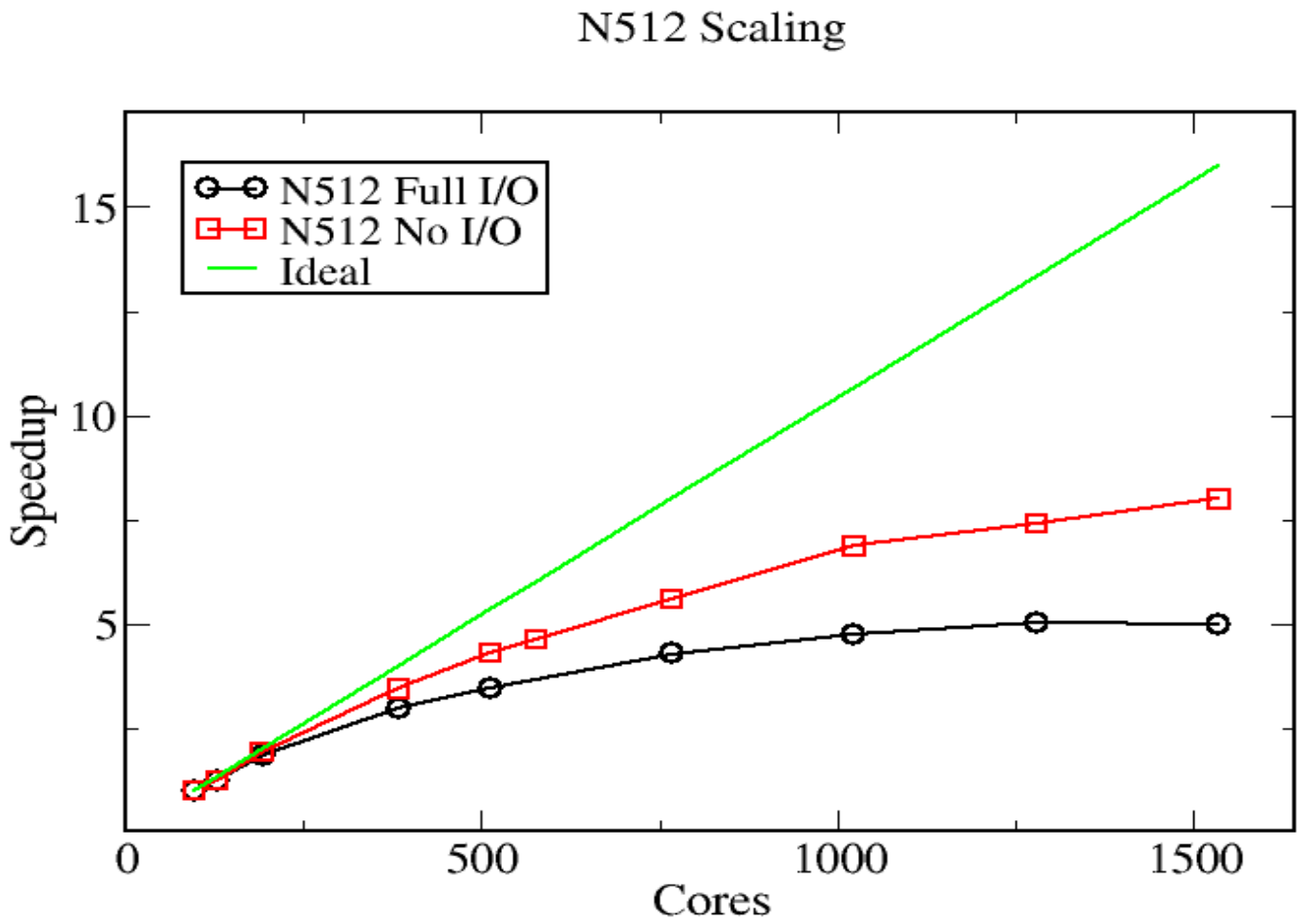
UKV

- **1.5km 70L** (variable resolution)
- **3DVAR** (hourly)
- **36h forecast**
- **4 times per day**

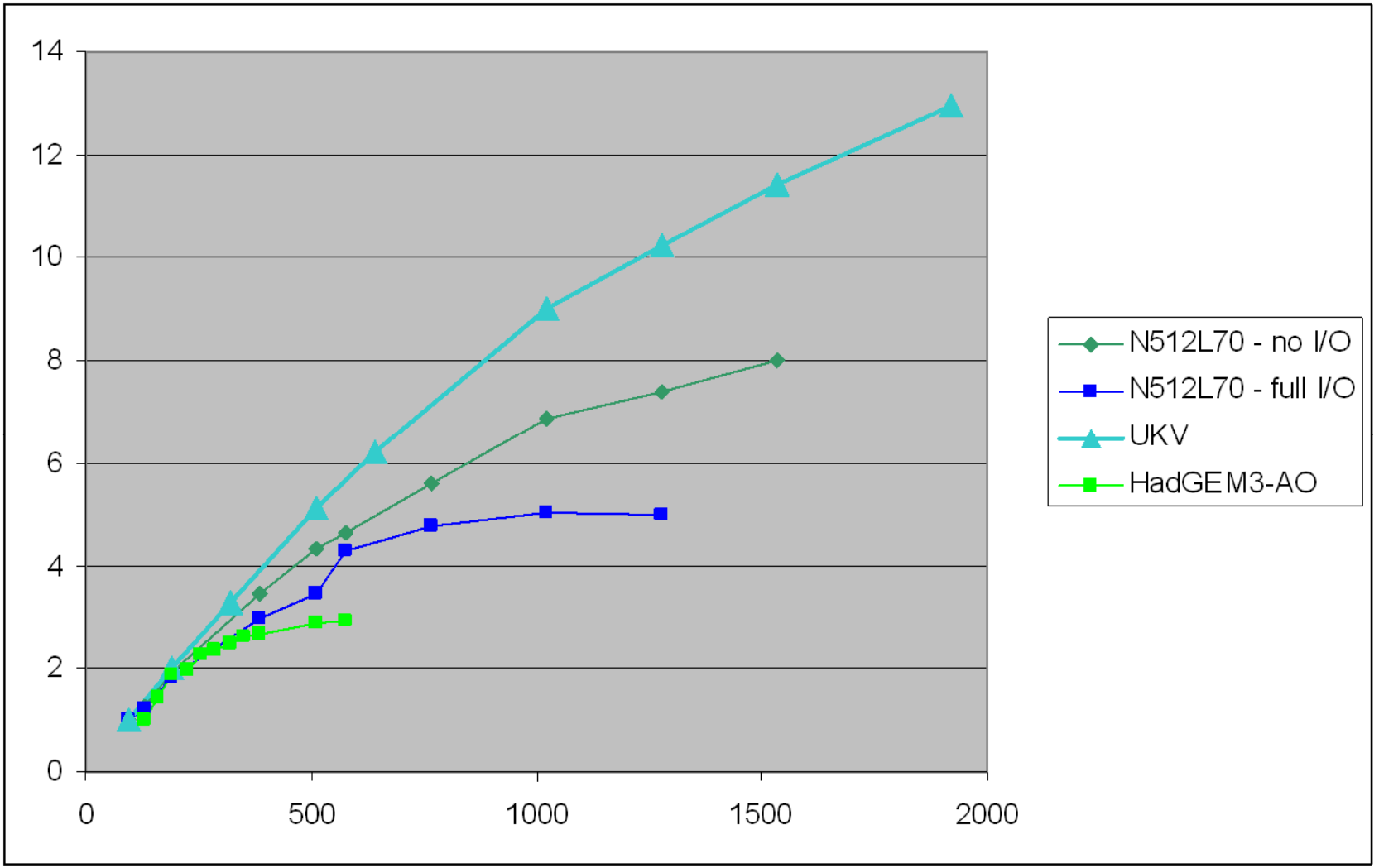


Scalability results

Scalability (N512 global) – Mar 2010



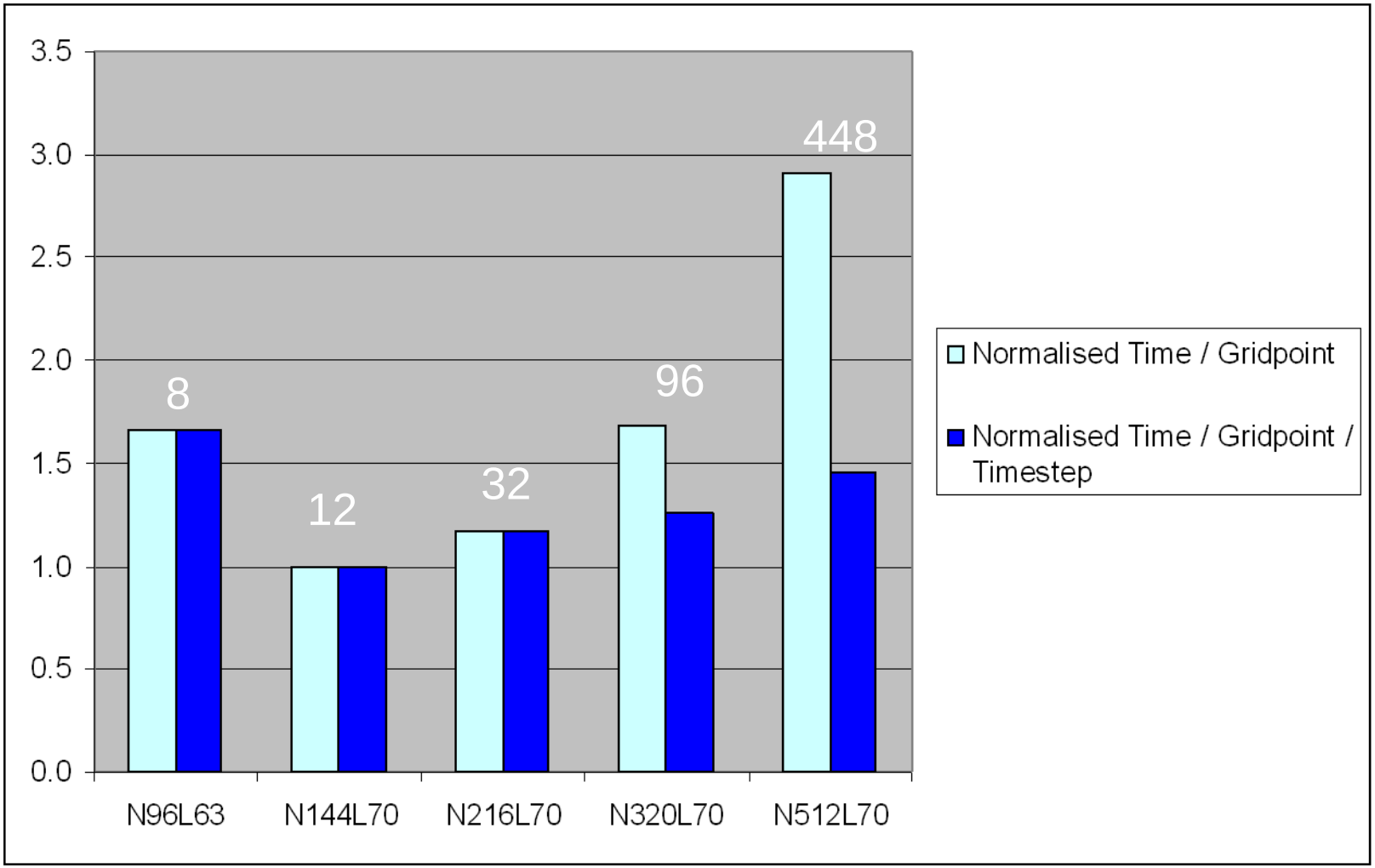
Strong Scaling – Mar 2010



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

Global Model Weak Scalability



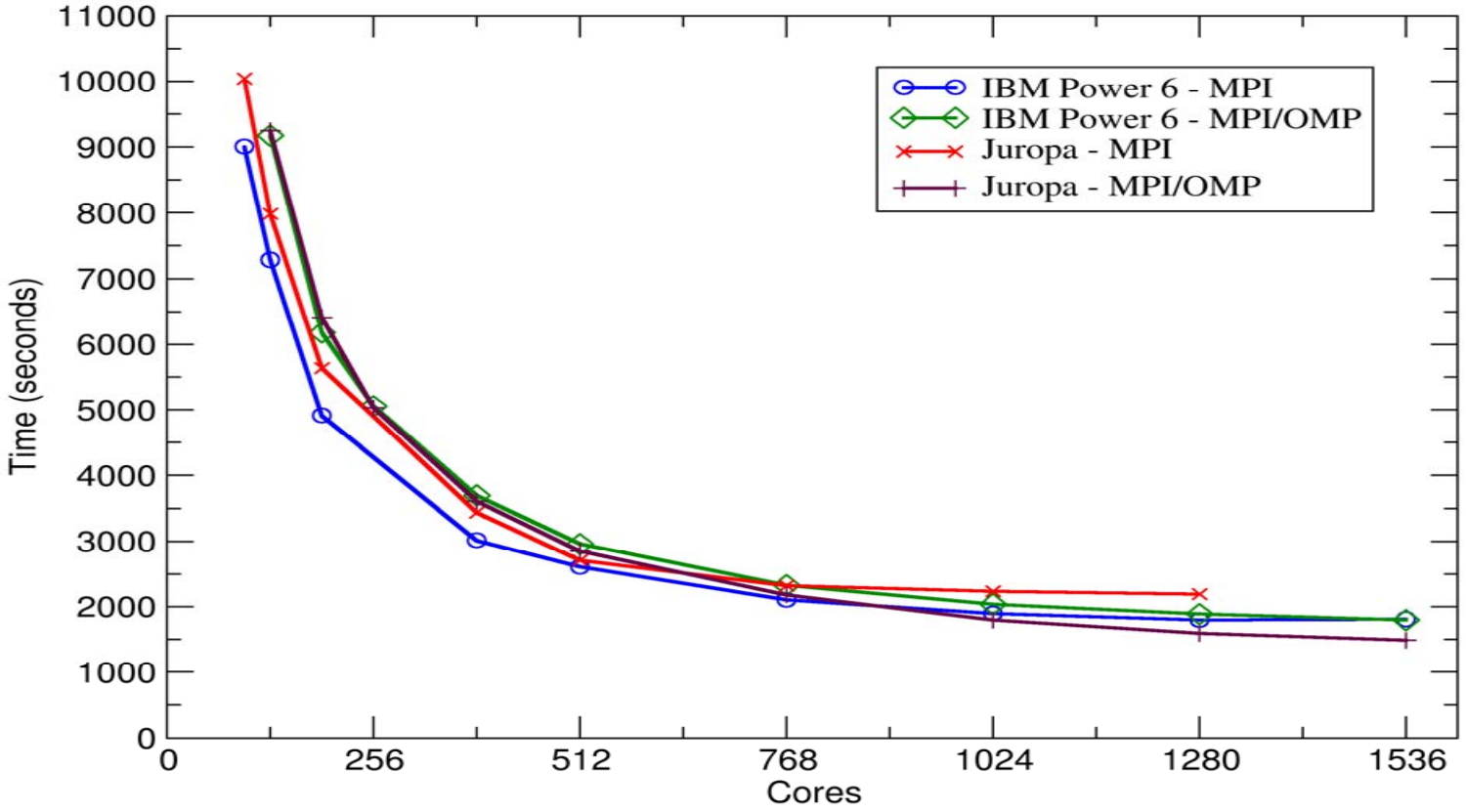
Models

Machine comparison

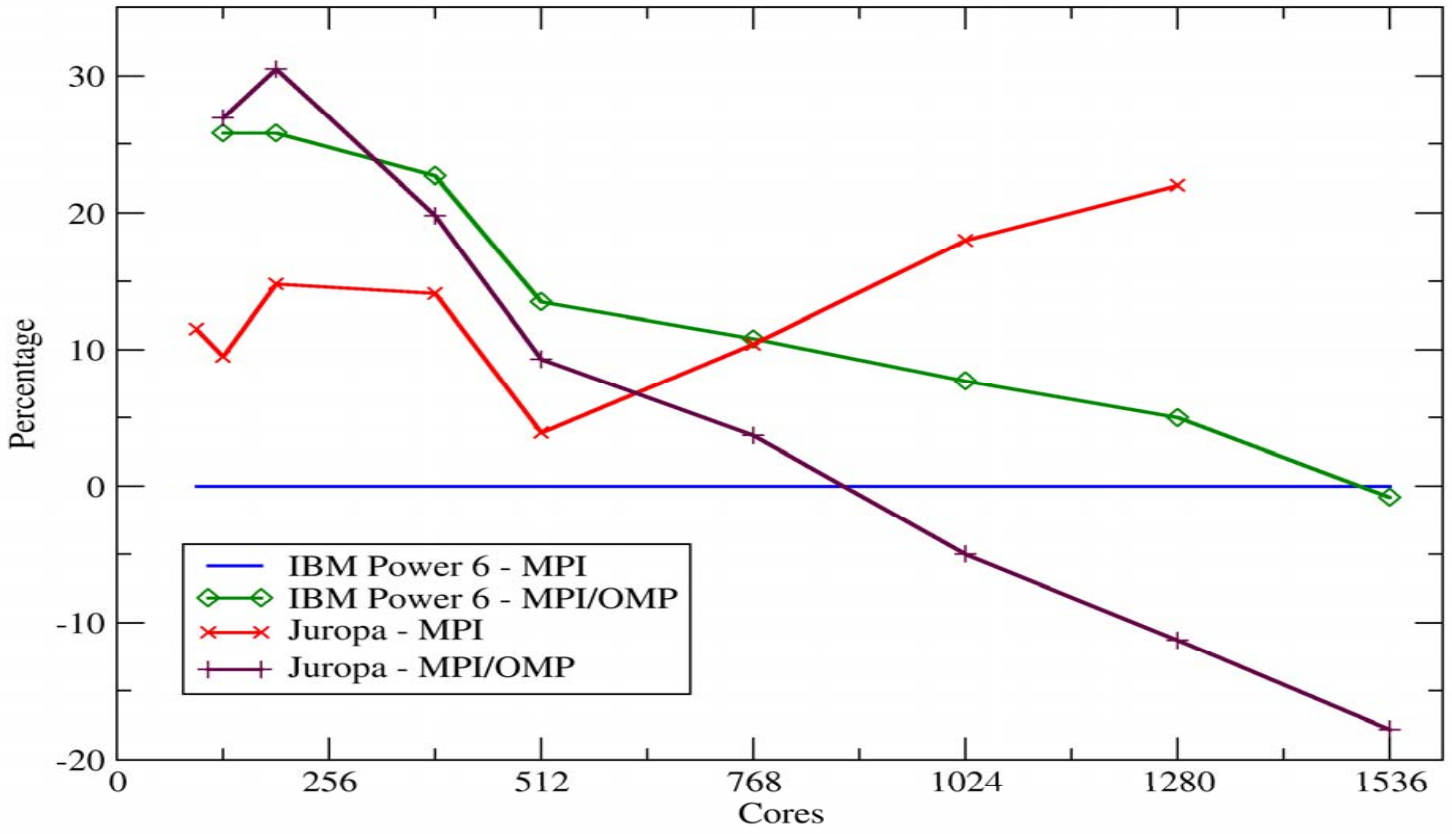
Through the PRACE initiative we were able to compare the UM on IBM Power 6 and an Intel Nehalem cluster (Juropa).

	IBM Power 6	Juropa (Intel)
Cores per Node	32	8
Clock Frequency	4.7 GHz	2.93 GHz
Interconnect	DDR Infiniband	QDR Infiniband
Filesystem	GPFS	Lustre

PRACE results (MPI & OpenMP)



PRACE results -
percentage difference to IBM MPI only



Coupled models

HadGEM3-AO components

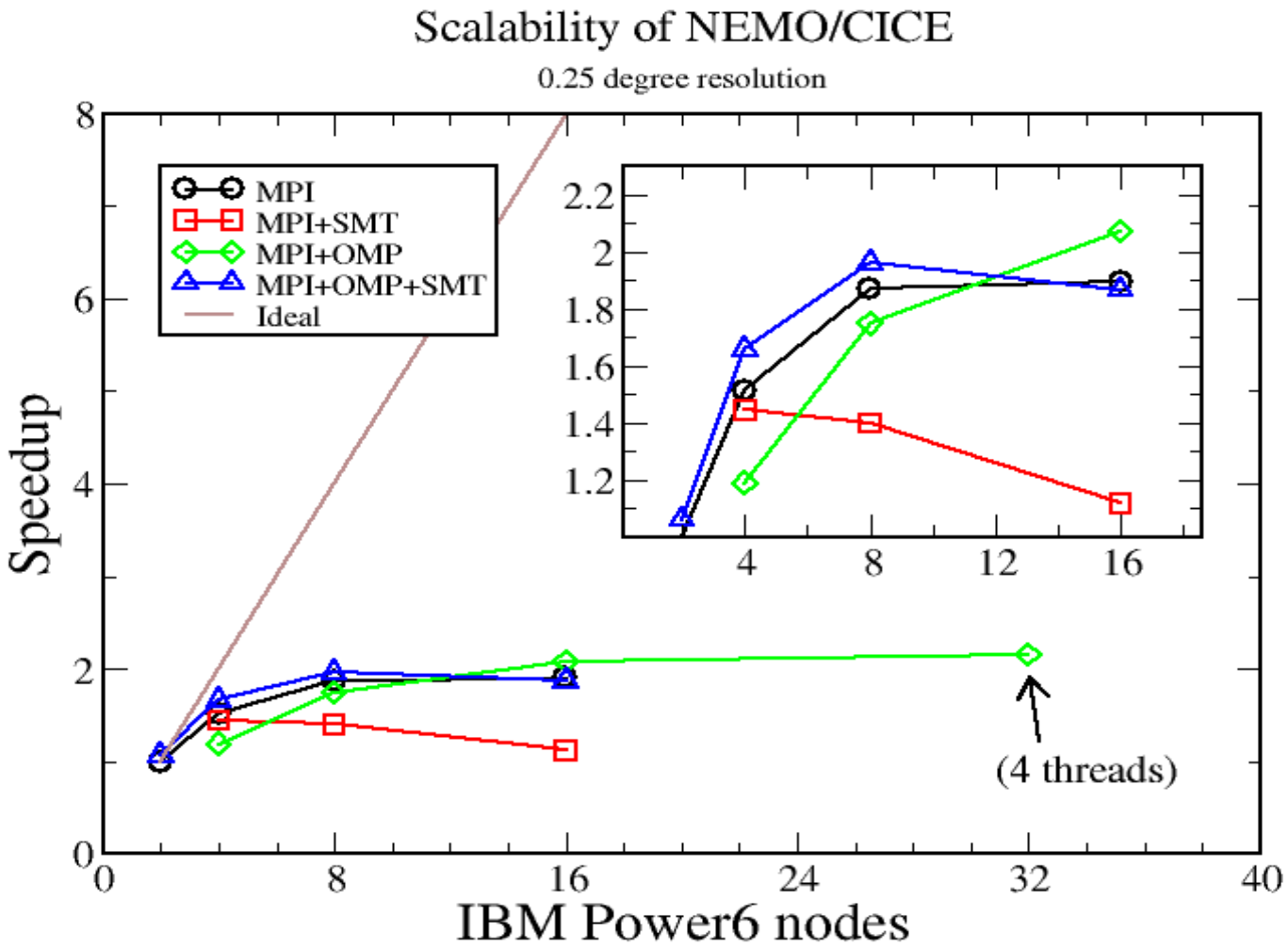
- Atmosphere (Unified Model)
- Ocean (NEMO)
- Sea Ice (CICE)
- Coupler (Oasis3)

One executable

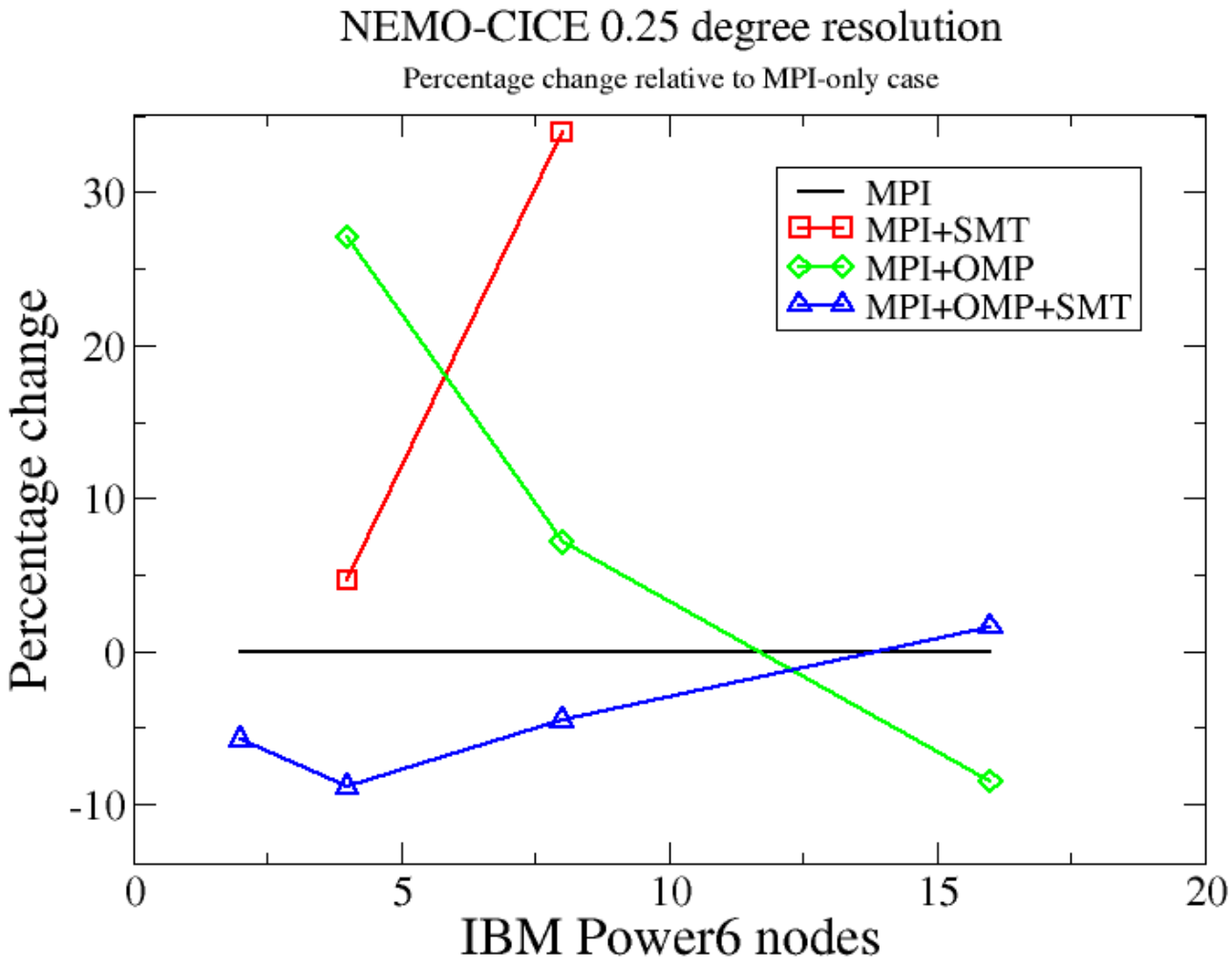
Used for climate integrations, seasonal forecasting (GloSea4)

NEMO Scalability

NEMO – Scalability curves



NEMO – comparison with MPI

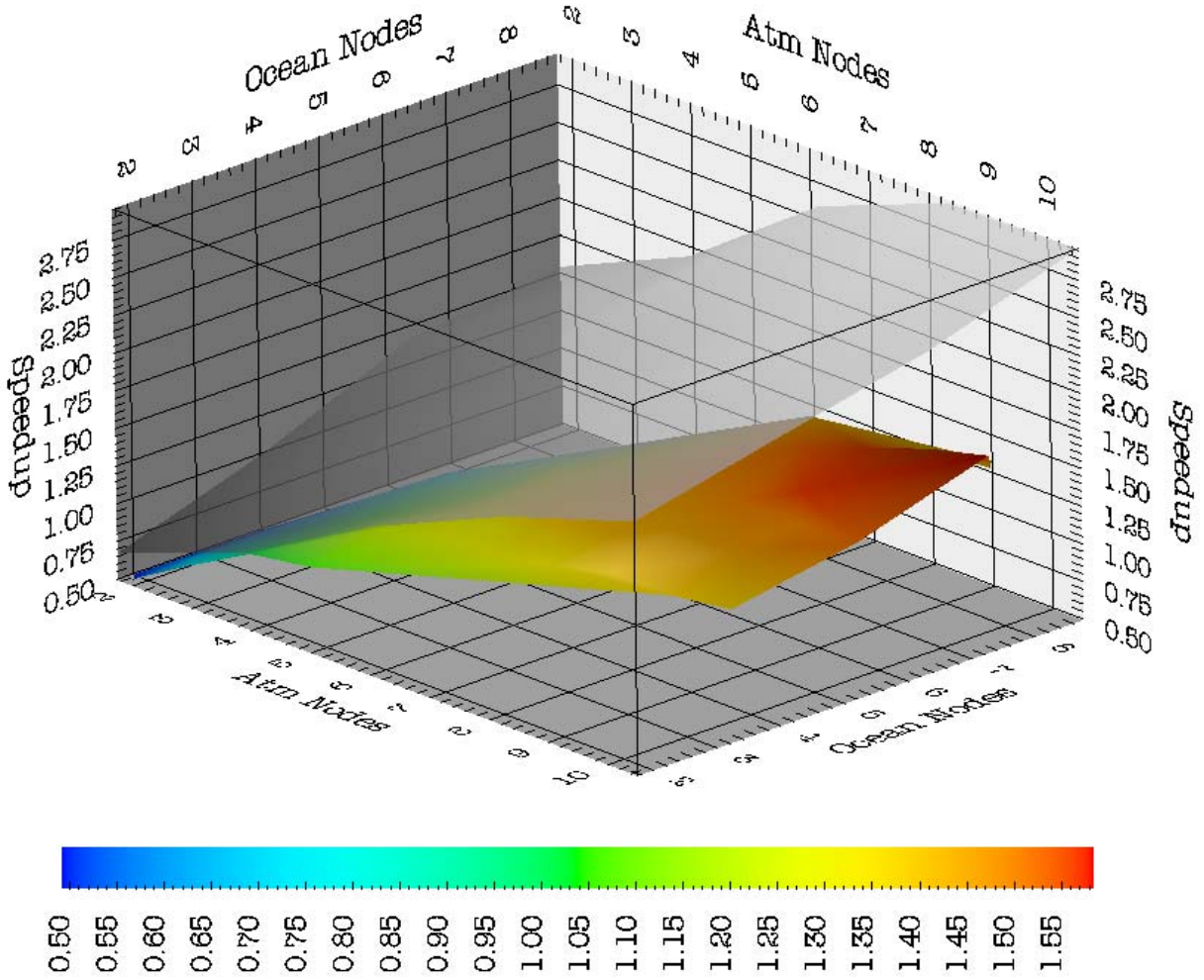


Load balancing and all that

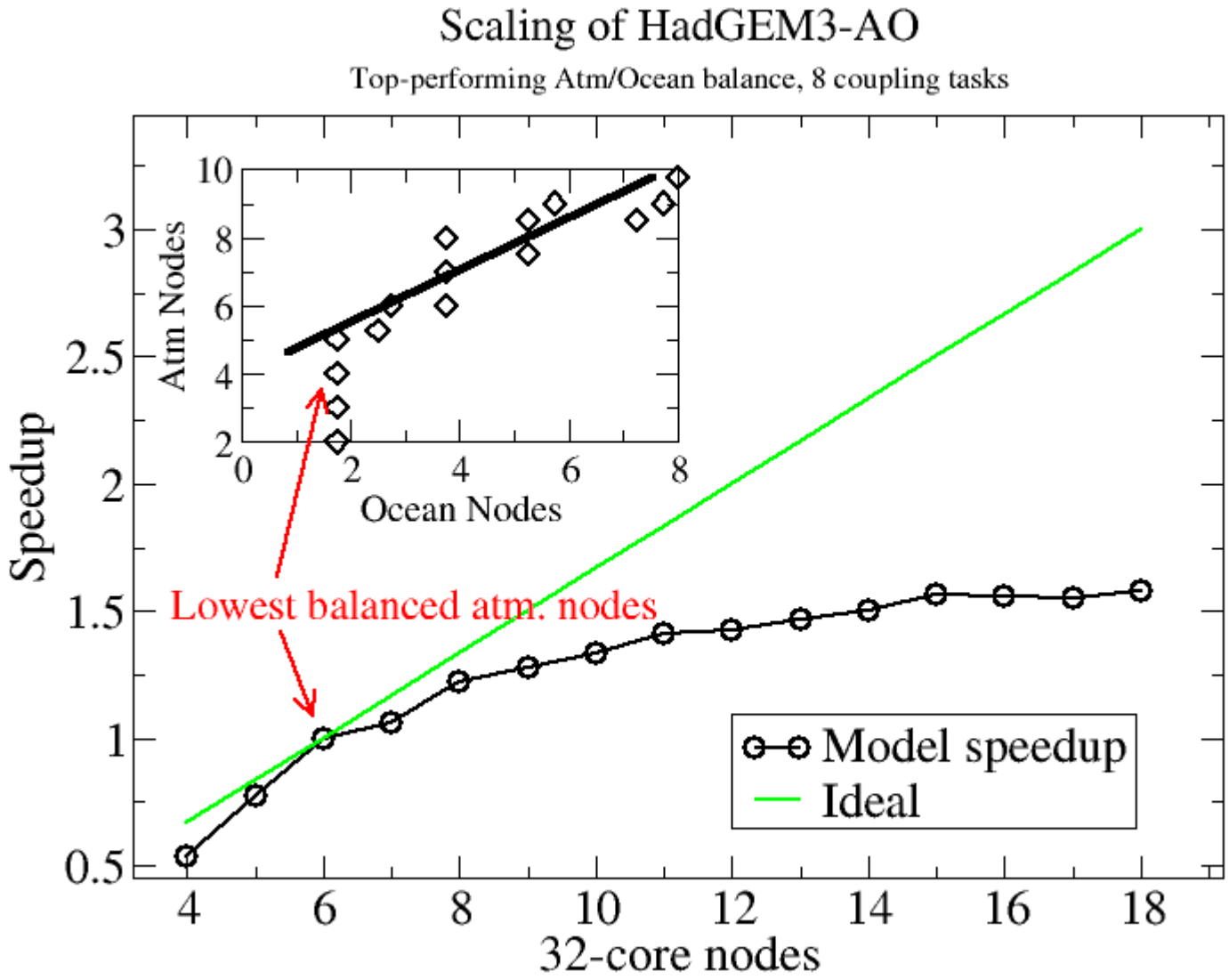
- Component speed depends on
  - Cores given
  - Number of threads
  ... and more ...

- 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.

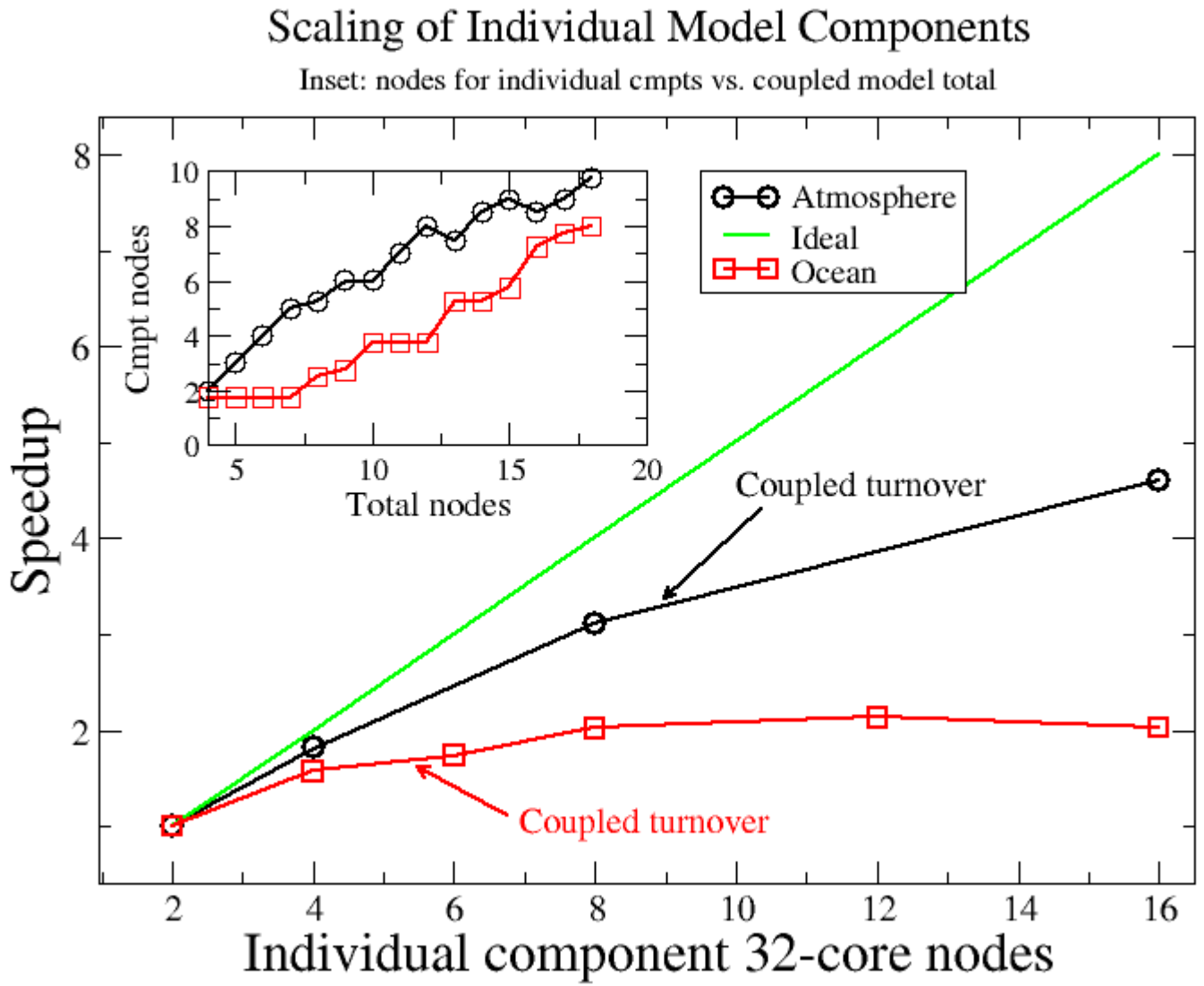
An extra dimension ...



Coupled model scaling



Individual components



Recent improvements

QPOS (moisture/tracer reset)

- Current algorithm gathers information onto 1 processor, does work, then scatters data.
- Anti-scales.

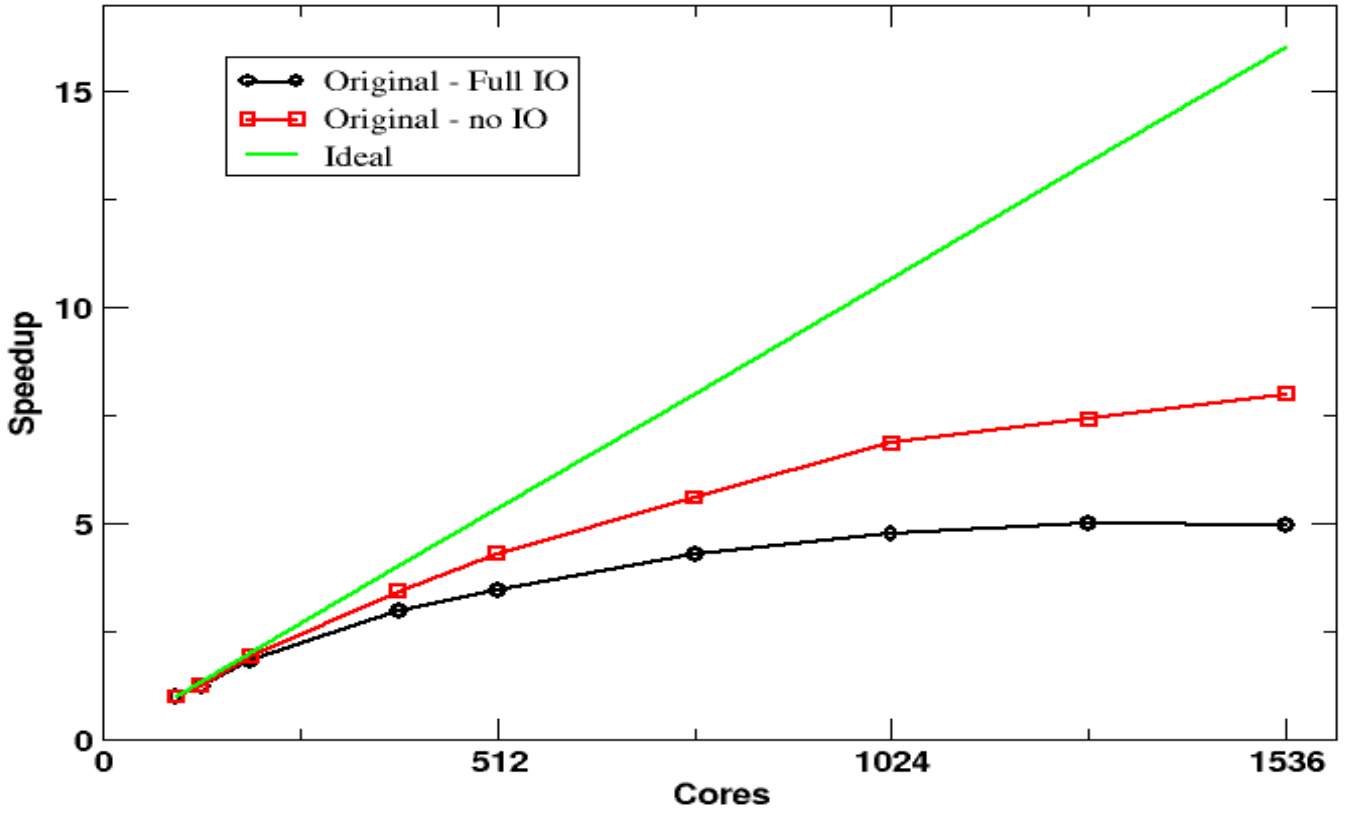
- Alternative algorithms coded up

- Used in PS24 – saved 7 minutes on an operational forecast

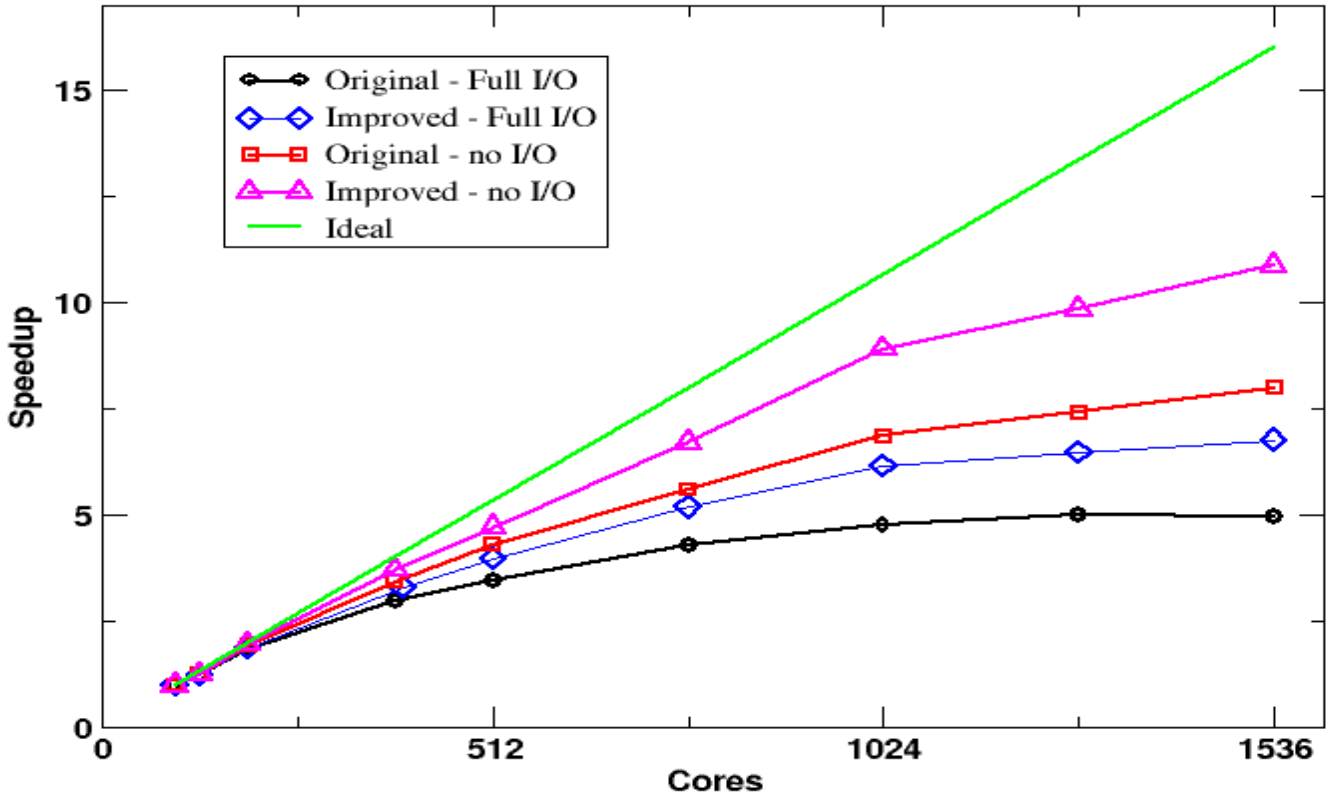
PMSL

- 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



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.

OpenMP

UM

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

NEMO (in GloSea4 coupled model)

- N96**L38**-O1**L42** and
- N96**L85**-O1**L75** 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