Met Office

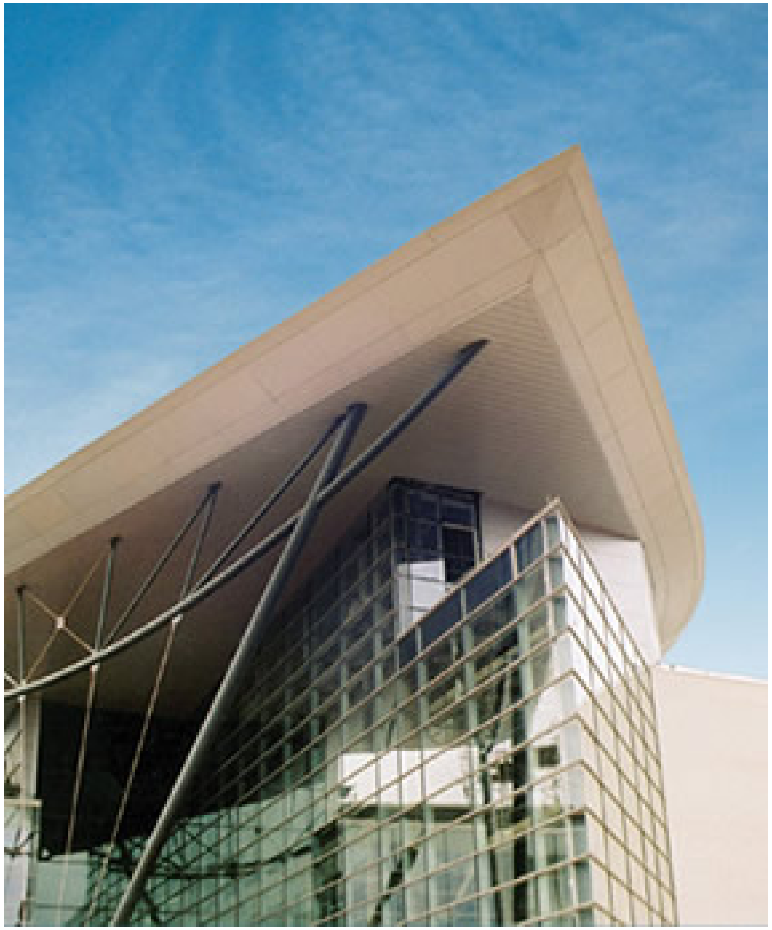
An Overview of HPC at the Met Office

Paul Selwood

Introduction

The Met Office

- National Weather Service for the UK
- Climate Prediction (Hadley Centre)
- Operational and Research activities
- Headquarters relocated to Exeter from Bracknell 2003/4
- 150th Anniversary in 2004



Met Office Supercomputers

- NWP – 19 node SX-6
- Climate – 15 node SX-6

- January 2005 – 16 node SX-8 installed.
  - 8 CPU/node
  - 64GB FCRAM/node

- September 2006 – 4 additional SX-8 nodes for Climate

- November 2006 – 1 additional SX-8 node due for NWP

SX-8; differences from SX-6

- Passed acceptance and reliability tests at first attempt.
  - Easiest supercomputer install we've had!

- Code compiled for SX-6 runs well on SX-8.
  - To allow operational backup on SX-6.
  - SX-8 specific code (sqrt vector pipe) gives little improvement

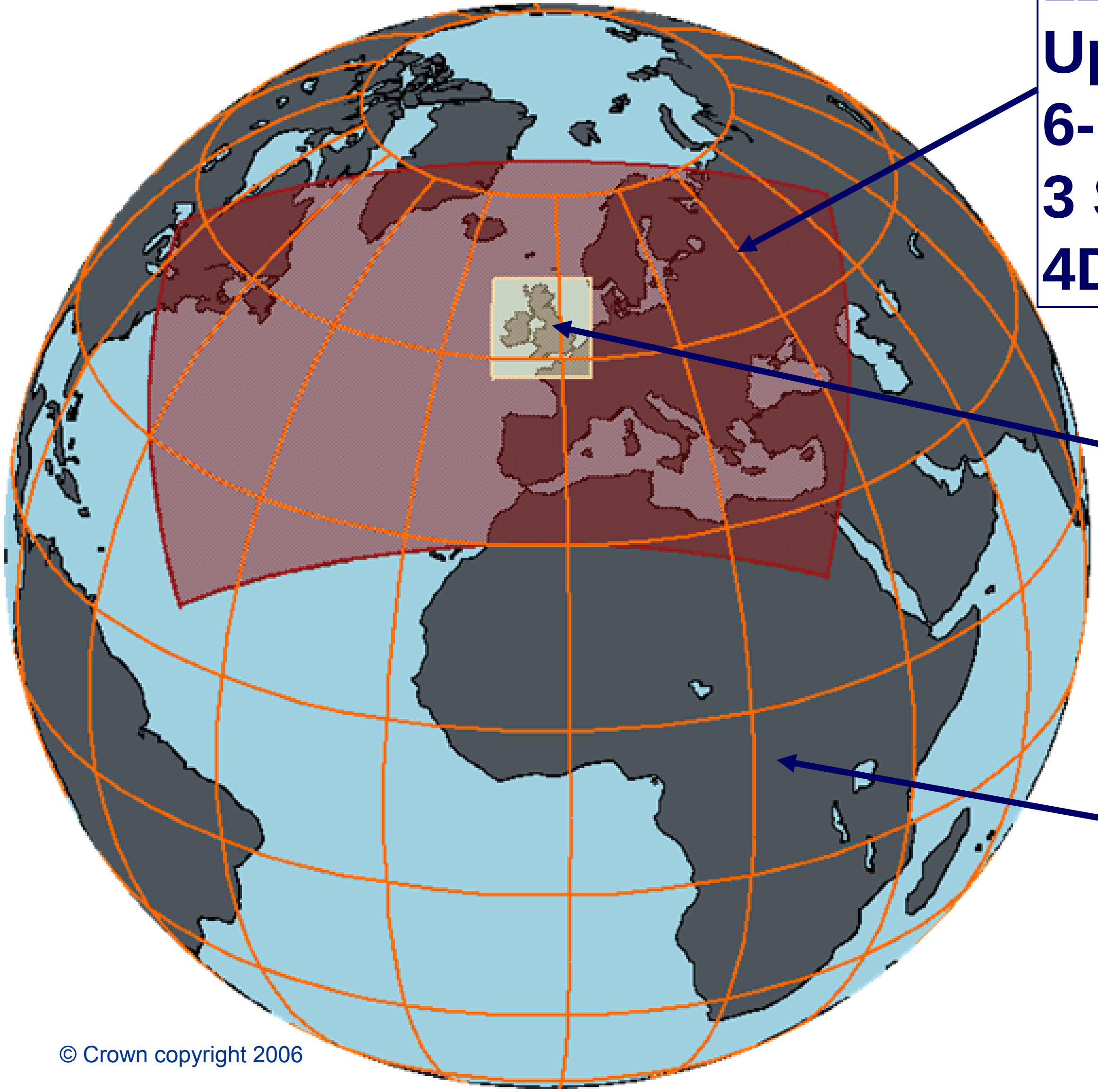
- SX-8 : SX-6 ratio ~2.1 for computation.
- Bank caching on SX-8 gives much better look-up table performance.
- Need to double size of cache for local I/O, else saturation gives variable performance.

The Unified Model

The Unified Model

- Climate and Forecast model
- Atmosphere, Ocean and Coupled (also sea-ice, atmospheric chemistry, aerosols, river transport, …)
- Atmosphere
  - Non-hydrostatic, semi-Lagrangian, semi-implicit, Arakawa C grid, Charney-Phillips vertical coordinate
- ~ 700K LOC
- MPI parallelisation

Deterministic Forecasts



12km – 38 levels
Up to 48hr f/c
6-hourly update
3 SX-8 nodes
4D-Var

4km
Up to 36hr f/c
1 SX-8 node
3D-Var

40km - 50 levels
Up to 7 day f/c
6-hourly update
3 SX-8 nodes
4D-Var

MOGREPS

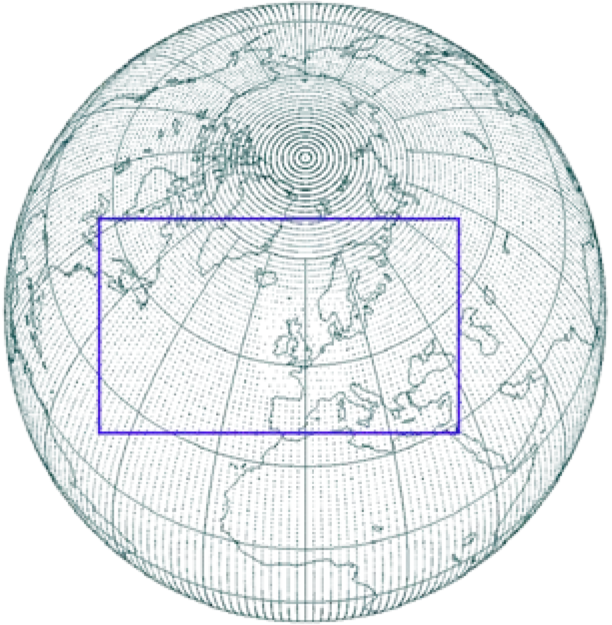
- 24 members
- Run on 3 nodes of SX-6

Global

- Run to T+72
- N144 (~ 90 km)
- Uses Ensemble Transform Kalman Filter (ETKF) for generating initial perturbations
- Stochastic physics – random perturbation of parameterisation schemes

LAM

- Run to T+36
- 24 km
- North-Atlantic Europe
- Takes initial and boundary conditions from global model
- Stochastic physics



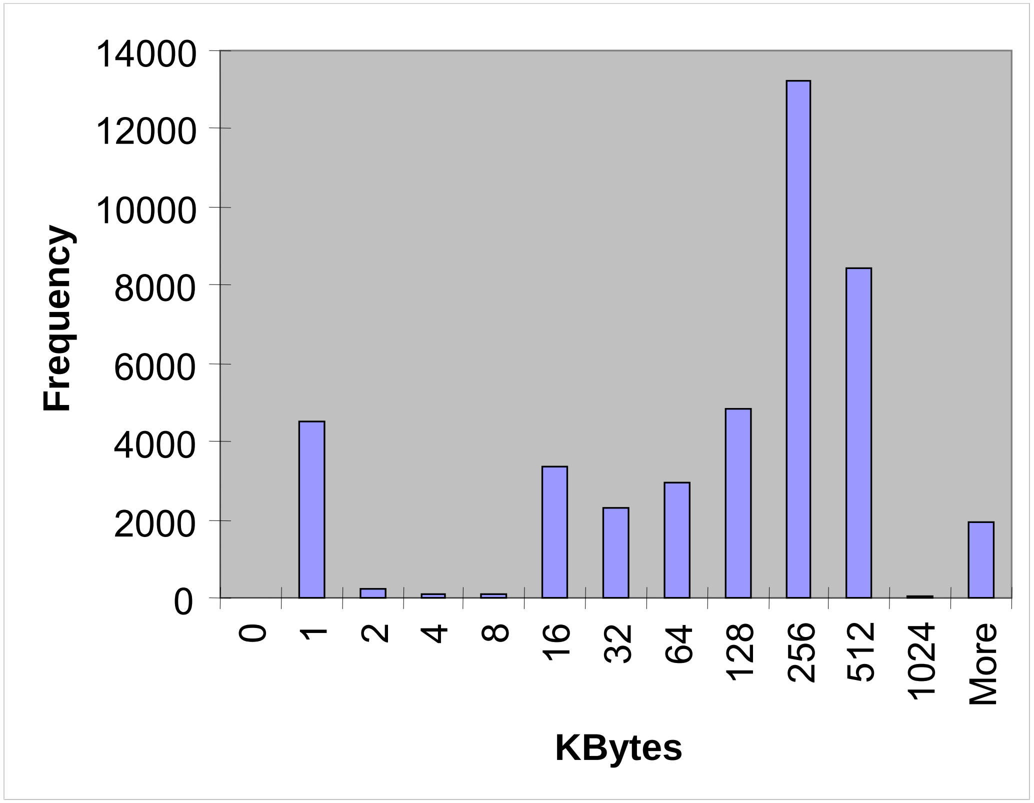
THORPEX - TIGGE

- The Met Office medium-range ensemble forecast system is running on ECMWF *hpcd* (soon *hpce*).
- Based on short-range MOGREPS-Global system, extended to 15 days.
- Resolution N144 (0.833° x 1.25°), 38 levels
- 24 members (control + 23 perturbed), run twice a day (0 and 12 UTC).
- Initial data, with ETKF perturbations, created at Met Office and copied to ECMWF.

Unified Model Performance

I/O – the problem

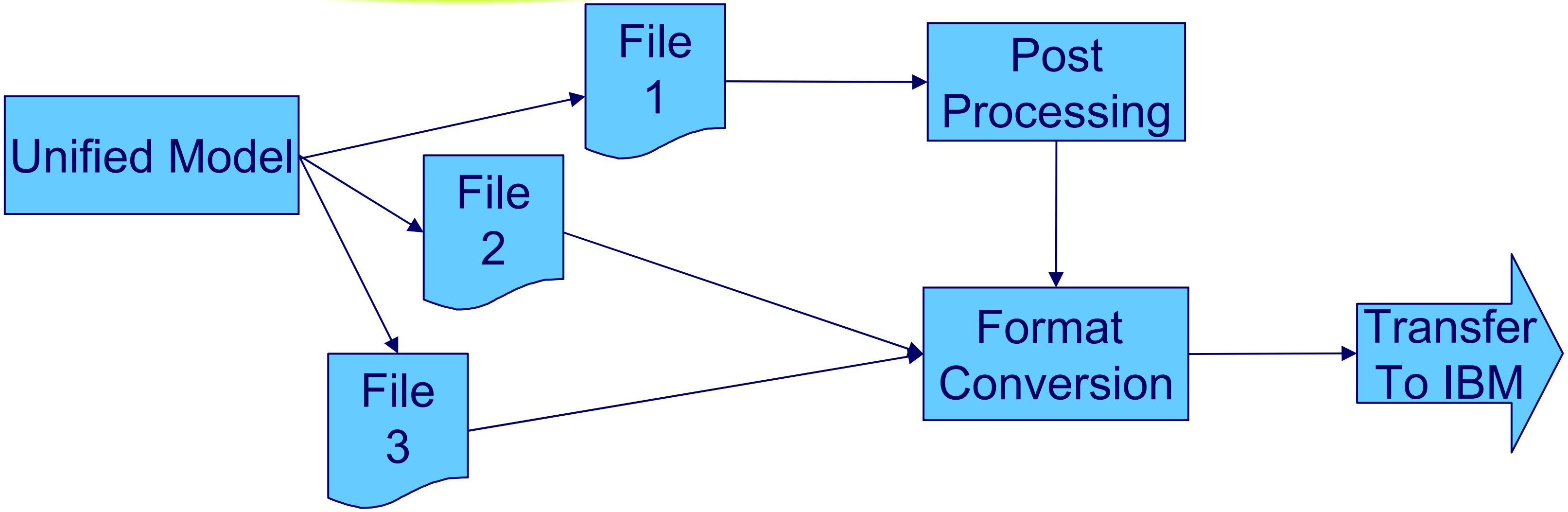
- Unified Model I/O initially very slow
- Route to GFS disk depends on packet size
    - < 64KB nfs (slow)
    - >= 64KB GFS (fast)
- Application buffering improved I/O rates from 40 MB/s to 140MB/s

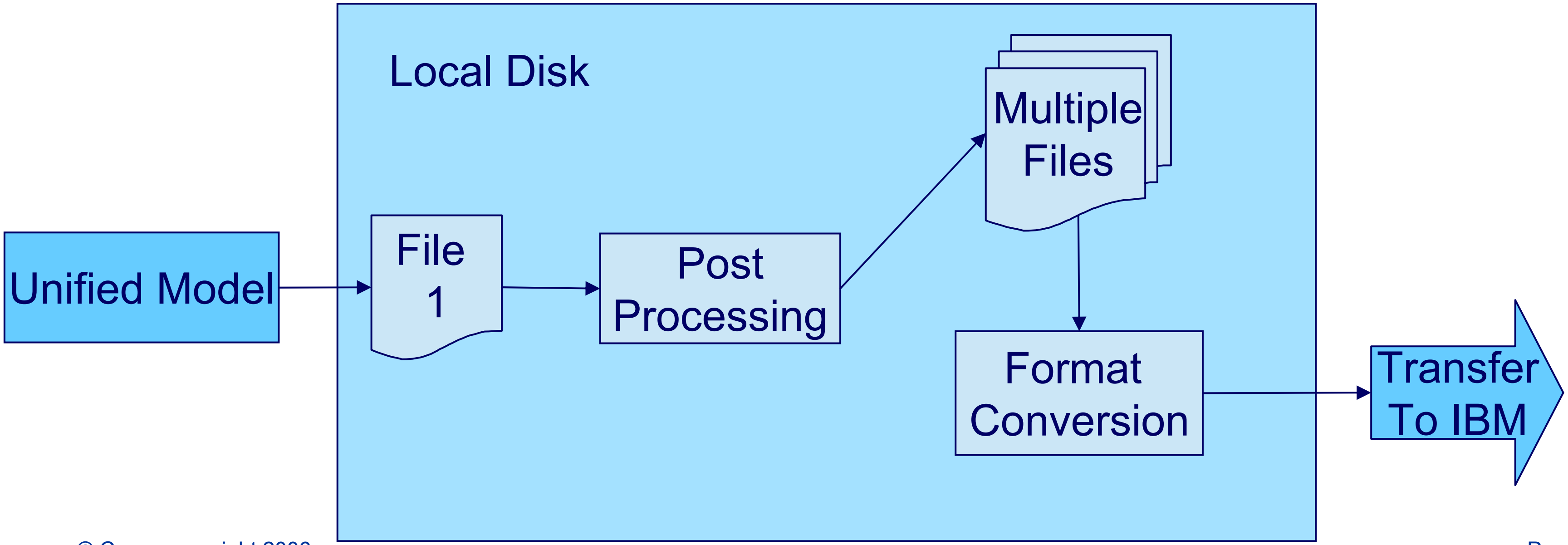


I/O – Local disk

- Local disks have cache, GFS doesn’t
- Application can see > 1 GB/s transfer rate
- Only used for operational work
- Only enough disk space for certain output streams
- Needs careful data management

Local I/O


Unified Model
File 1
File 2
File 3
Post Processing
Format Conversion
Transfer To IBM


Local Disk
Unified Model
File 1
Post Processing
Multiple Files
Format Conversion
Transfer To IBM

I/O Server processes

- Unified Model typically only does output on certain timesteps
- I/O Server process can process the output asynchronously
- Initial work on NEC had little benefit
  - 1% improvement for 15% cost
  - Small numbers of CPUs

- Bob Carruthers (IBM) extended code for multiple server processes and improved scheduling
  - 8% improvement for 3% cost
  - Need to re-evaluate on NEC

I/O in Scripting

- Analysis of full operational suite showed inefficiencies in scripting I/O

- cp on SX node has poor blocksize
- dd allows user control of blocksize
- Typically save 60-70% of cost (30s per file)

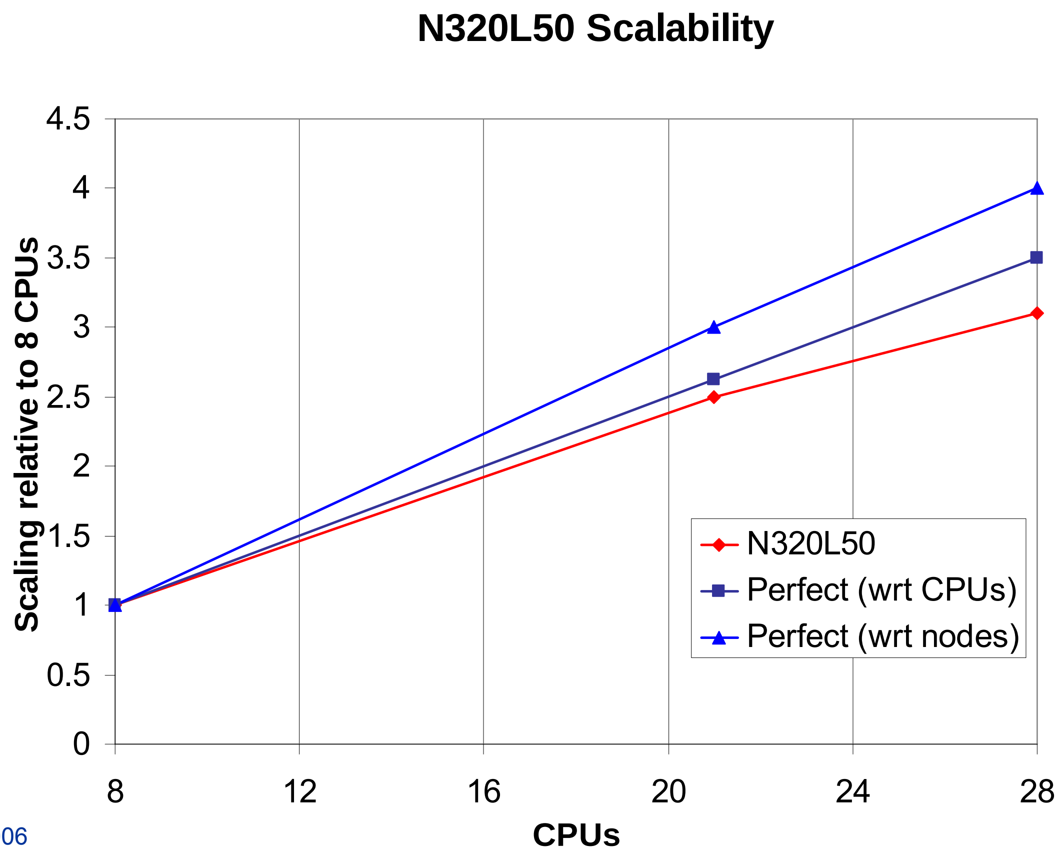
- Multiple appends to a file taking 2-4 minutes
- Buffer via perl array with a single write takes cost to 2-4 seconds.

Communications – improvements made

- T3E coding practices
  - Unnecessary barriers removed
  - Naïve SHMEM → GCOM conversion improved (1326s improved to 25s – overloaded MPI buffers?)
- Gathering/Scattering 3D fields level-by-level
  - Optimised by copying into temporary buffers and doing one communication per CPU pair
  - Halves cost of these communications
- >6000 halo exchanges in 6 hour forecast
  - 1500 & many other communications removed from a single diagnostic calculation!
  - Amalgamating communications only minor benefit

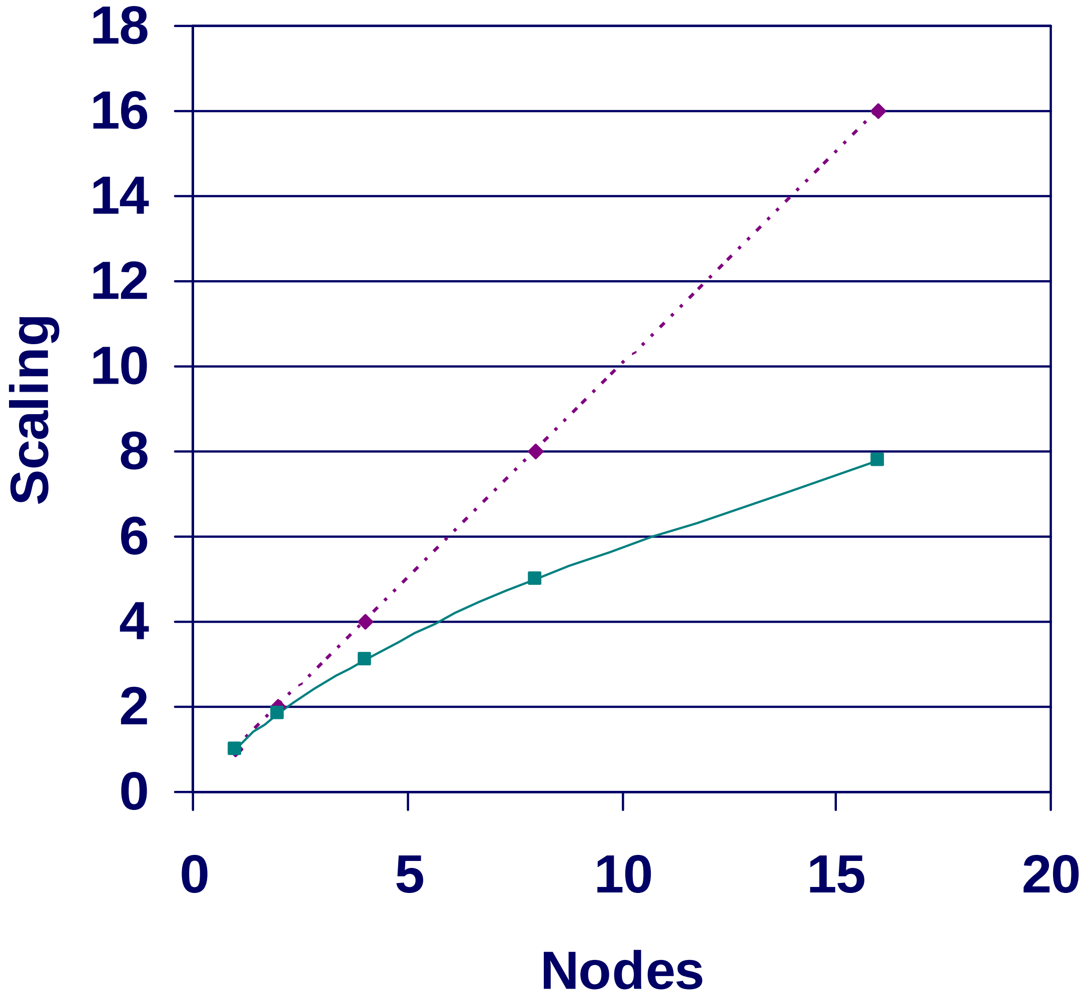
40km Global Model

- 40km Operational Global Model
  - 640 x 481 x 50
  - 7 day forecast in ~45 minutes on 3 SX-8 nodes



North Atlantic/Europe Model Scalability

- 1 day forecast
- 7 cpu/node
- Limited output



UM TIGGE on IBM - optimisation

- Port to IBM from NEC straightforward
- Tuning of physics segmentation (like NPROMA)
- Paul Burton & Deborah Salmond made improvements to communications.
- John Hague (IBM) implemented OpenMP for 90% of runtime.

UM TIGGE on IBM - timings

	32	64	128
hpcd	265	125	67
hpce	129	70	49
hpce (2 thread)	133	70	39

Procurement and RAPS

HPC Procurement timeline

- Business case for new supercomputer and mass-storage accepted
- Possible partnership with UK academia
- Expected tender – July 2007
- Award contract – late Summer 2008
- Delivery of first hardware – Winter 2008/9
- Acceptance – Spring 2009

RAPS

- First Met Office RAPS release
- Initially a Unified Model benchmark
  - N512L76 main resolution
  - 1 day forecast, with or without I/O
  - N320L50 and N48L38 supporting resolutions
- Available now – standard Met Office benchmarking licence
- Plans for a Data Assimilation benchmark in Spring 2007
  - Observation processing
  - 4D-Var

RAPS performance on NEC SX-8

- 2, 4, 5 nodes run
- N512L76 noio case
- 7 cpu/node
- ~70 GB memory required

CPUs	Time
14	7140
28	4200
35	3420

Questions?