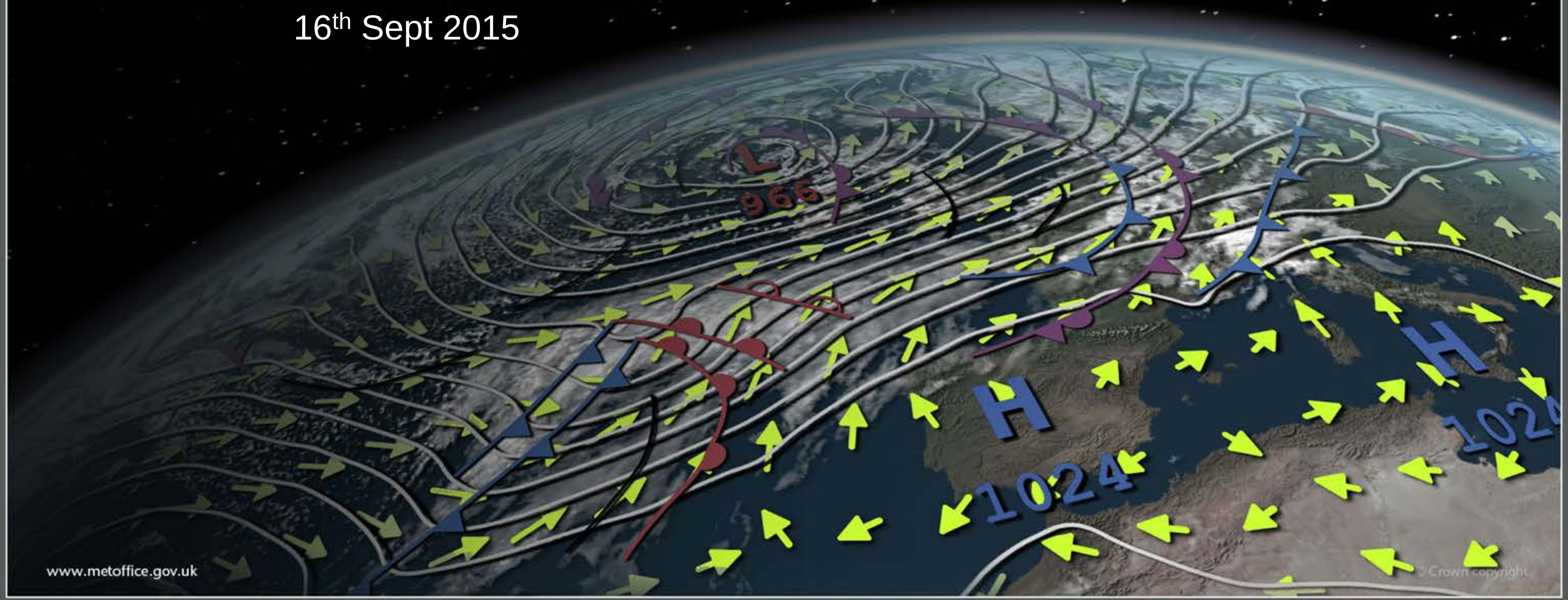
Met Office

An Update on the Met Office HPC

Adam Voysey

16th Sept 2015



Contents

- History & Timescale of Upgrade
- Where are we now?
- Porting: IBM to Cray
- Future
- Conclusions

History & Timescale of Upgrade

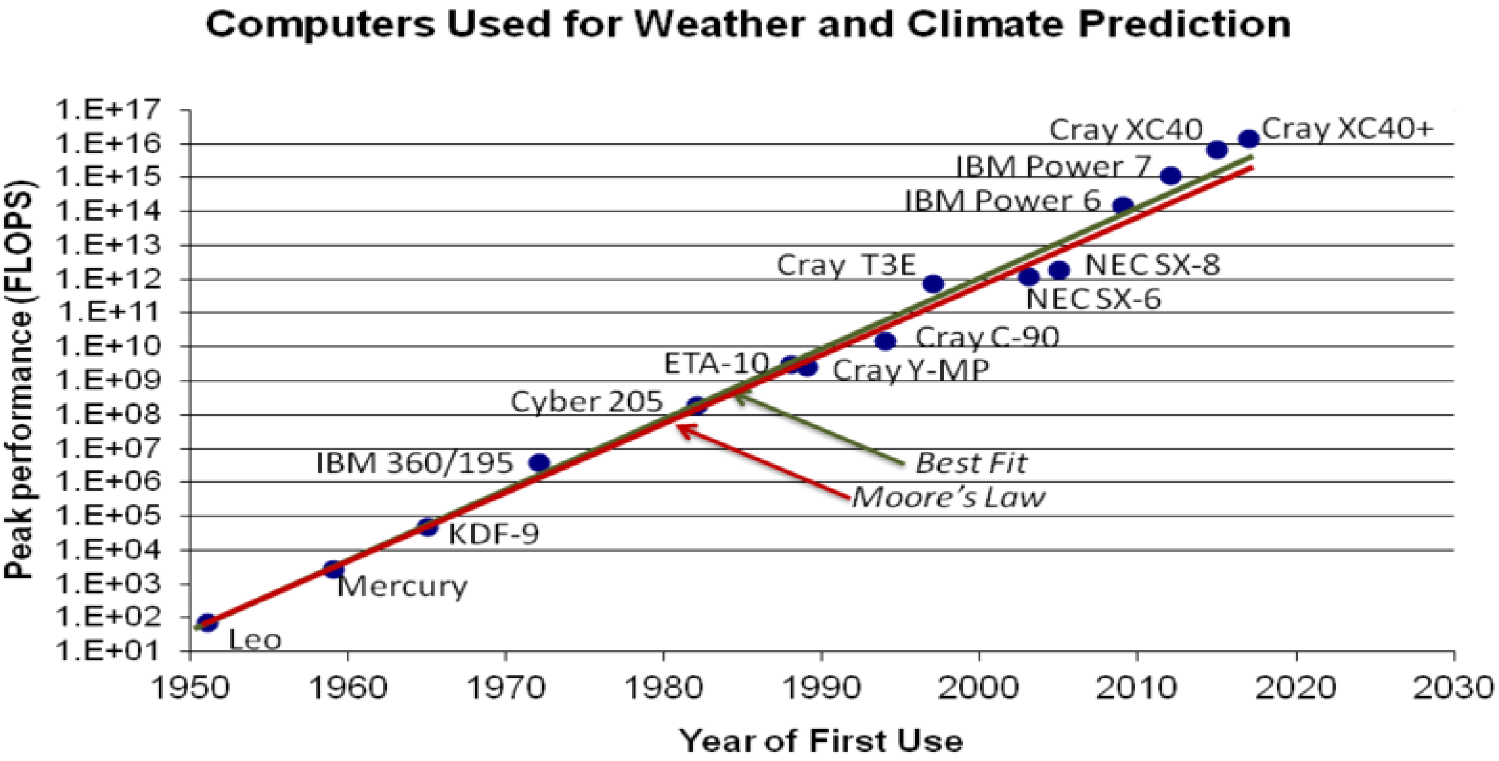
“It is of great concern to us that these scientific advances in weather forecasting and the associated public benefits are ready and waiting but are being held back by insufficient supercomputing capacity. We consider that a step-change in supercomputing capacity is required in the UK.”

House of Commons Science and Technology Committee Report on ‘Science in the Met Office’, March 2012

➔£97m investment in Met Office HPC

Cray® XC40™

Met Office


Computers Used for Weather and Climate Prediction
Peak performance (FLOPS)
1.E+17
1.E+16
1.E+15
1.E+14
1.E+13
1.E+12
1.E+11
1.E+10
1.E+09
1.E+08
1.E+07
1.E+06
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1950
1960
1970
1980
1990
2000
2010
2020
2030
Year of First Use
Leo
Mercury
KDF-9
IBM 360/195
Cyber 205
ETA-10
Cray Y-MP
Cray C-90
Cray T3E
NEC SX-6
NEC SX-8
IBM Power 6
IBM Power 7
Cray XC40
Cray XC40+
Best Fit
Moore's Law

Phasing

- Phase 1a – two systems of 4 cabinets to replace Power 7s.
  - 2x 3PB and 1x 6PB Lustre storage
- Phase 1b – Both systems extended by ~13 cabinets
  - Total performance > 6x Power 7
- Phase 1c – 1 new cluster in new IT Hall in early 2017 with own storage
  - Total performance > 15x Power 7

Where are we now?

Met Office

www.metoffice.gov.uk

System is now live!

- Producing operational forecasts from Tuesday 25 August
- Like-for-like computing capacity compared to the previous IBM

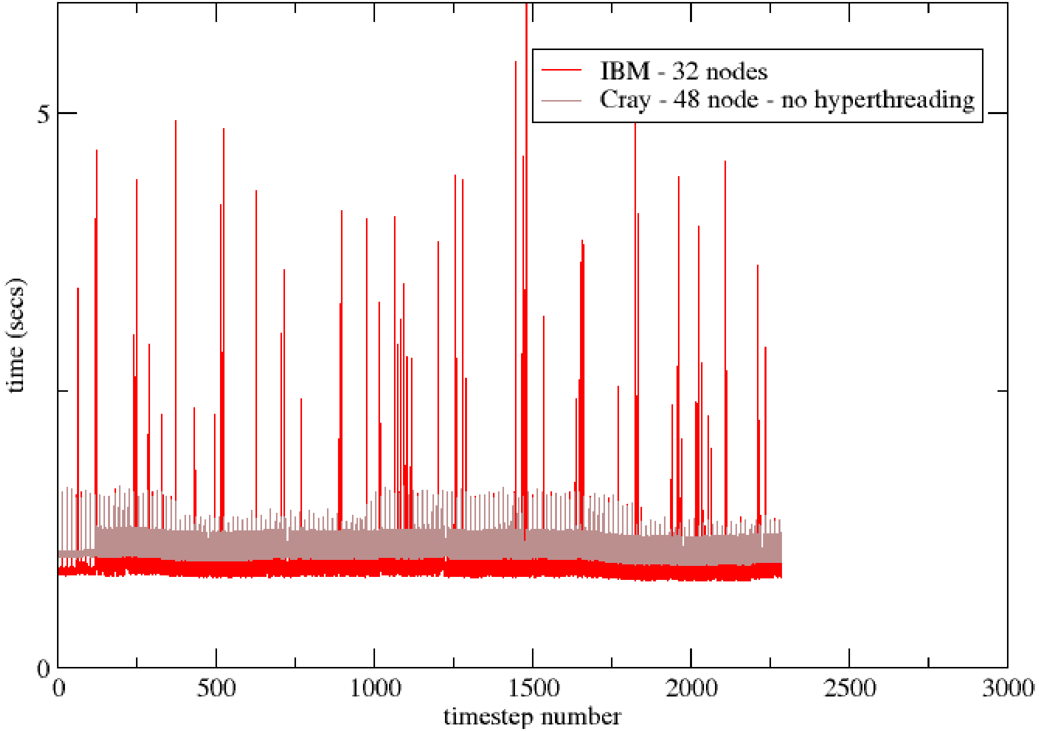
Porting

IBM to Cray

© Crown copyright

Model Timestep Comparison

Courtesy of Andy Malcolm



Lustre Concerns

- Lustre file systems can have problems with metadata
- We have been careful: -
    - avoid unnecessary metadata access
    - used striping of files
- Work done on the UM: -
    - Lustre API integration
    - Fortran IO statements
- Actually not caused many problems in practice

Code changes

- Several code changes have been made to UM
  - added compiler directives (e.g. !dir$ ivdep)
  - added ACTION='READ‘ to OPEN() statements
  - loop indices re-ordering
  - Adding options to only write output on certain processors
  - Removing unnecessary explicit synchronisations

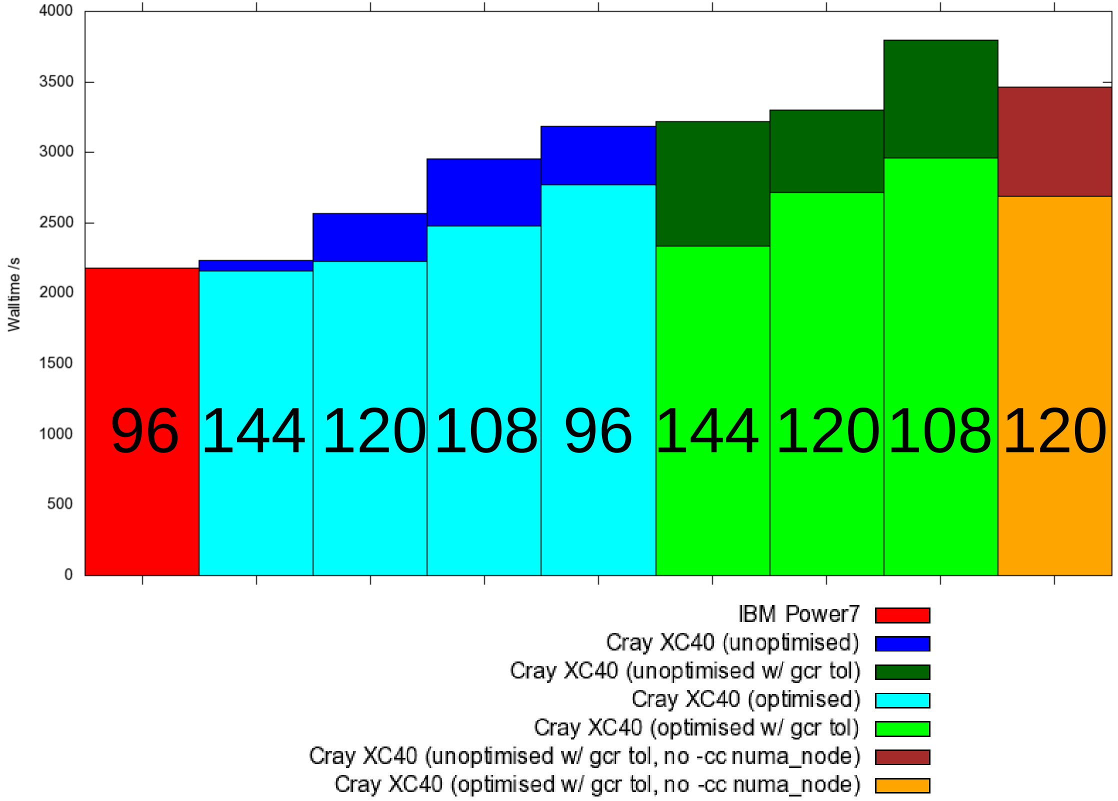
Code changes

- Several code changes have been made to UM
  - Speed up endianism conversion
  - Improve halo exchanges
  - Remove unnecessary INQUIRE() statements
  - Tune blocking sizes
  - extend OpenMP coverage

Compiler options

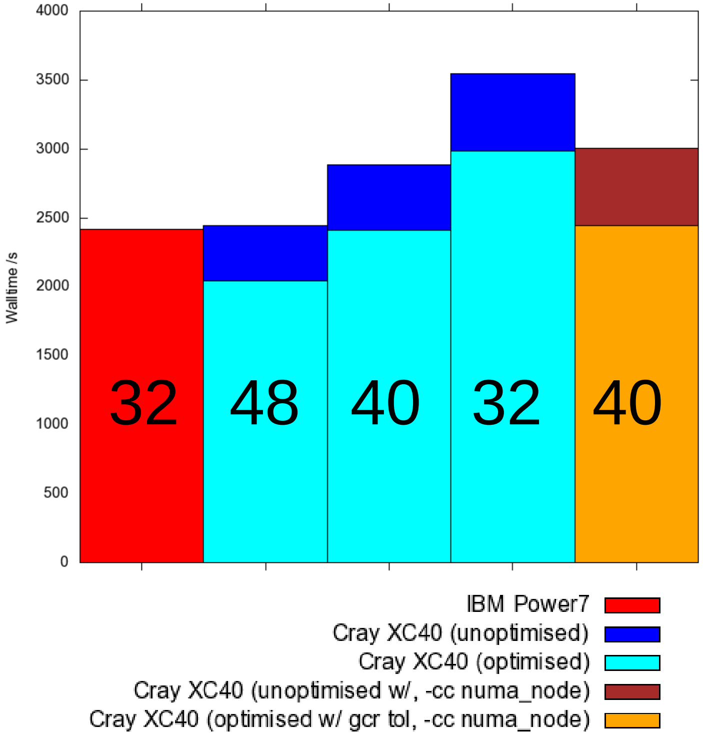
- Have done work to tune compiler optimisation flags
- ...But, bit-comparison must be maintained
- Started off conservatively; now try to be aggressive and override where code breaks or fails to bit-reproduce.
- looked at lots of flags
- main speed-up (and also problems!) from: -
    - -hvector
    - -hfp
    - -hflex_mp

Effect of Optimisations – N768



Met Office

Effect of Optimisations - UKV



Effect of Optimisations

- With the original configuration & optimisation:
  - For N768, 144 nodes can be reduced to 120 (~ 13% faster)
  - For UKV, 48 nodes can be reduced to 40 (~16% faster)
  - now slower with -cc numa_node
- With decreased tolerance:
  - Just decreasing tolerance increases runtime by ~30%
  - After optimisation, 108 node faster than 144 node before (~22% faster)

Effect of Optimisations

- The compiler option changes:
  - typically give ~4-10% improvement (configuration dependant)
  - One example of 10yr climate run
    2.2km resolution running on 16 nodes
    6620 sec per dump before changes improved to 5636 sec

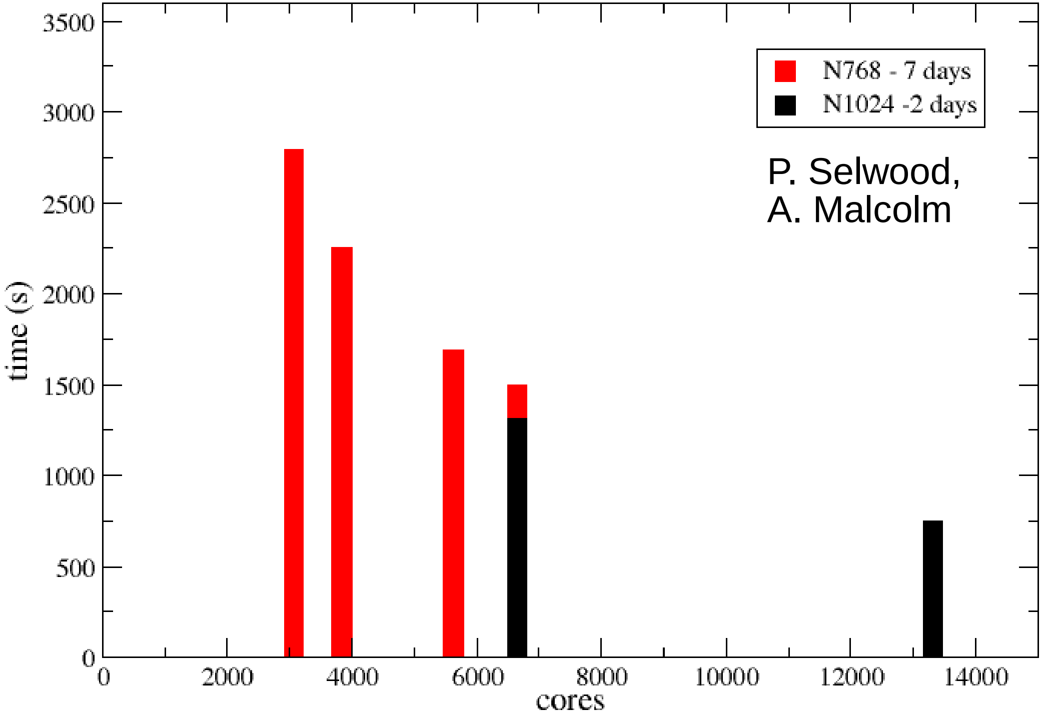
    ~14% improvement = 28,000 node hrs
    = 2 days running in one hall (current machine)

UM scaling on Cray XC40

Met Office HPC 3 phases: 1a 2x20K: 1b 2x100K: 1c 250K

Global Models
N768 ~ 17km res. 150M grid-points (operational)
N1024 ~ 12km res. 267M grid-points
MPI + 2 OMP threads
Cray compiler, IO server
Different solver config

Intel Haswell
16-core dual socket Xeon
2.6GHz



P. Selwood,
A. Malcolm

Met Office

The future

www.metoffice.gov.uk

IBM Switch-Off

- Machine switch off on 17 September 2015

Phase 1b

- Goes live Spring 2016

Phase 1c

- Goes live Spring 2017

Met Office

Phase 1c – New IT Hall

- Exeter Science Park
- Modern IT Facility
- 5.5 MVA, upgradeable
- Collaboration space

Met Office

More optimisation work

- More work to tune compiler optimisation flags
- Further extension of OpenMP
- More block size tuning
- etc...

Conclusions

Met Office

Conclusions

- New Cray XC40 machine is stable
- Now running operationally
- Slightly longer run times on same node count vs. IBM**
- ...But, optimisation work has (and continues to) narrow the difference
- Cray XC40 scales better
- IBM now being switched off
- We look forward to the next phases of expansion

Met Office

Thank You
Questions?