IFS migrates from IBM to Cray
'CPU, Comms and I/O'

Deborah Salmond & Peter Towers
Research Department Computing Department

Thanks to Sylvie Malardel, Philippe Marguinaud, Alan Geer & John Hague and many others …

ECMWF

CCA and CCB -
Cray XC30

2*19 Cabinets
with 2*3465
Compute Nodes

Nov 2013 -

ECMWF

C2A and C2B – IBM Power7
2*11 Frames with 2*768 Compute Nodes
June 2011 – Sept 2014

Comparison of IBM P7 and Cray XC30

	IBM	Cray
Processor	IBM Power7	Intel IvyBridge
Clock Speed	3.8 GHz	2.7 GHz
Switch	IBM HFI	Cray Aries
Nodes	768 *2	3465 *2
Cores per Node	32	24
Cores	24576 *2	83160 *2
Peak (Tflops)	754 *2	1796 *2
Memory per Node (GB)	64	64
Compiler	IBM XLF	Cray CCE
OS	AIX	CLE
Parallel File system	gpfs	lustre

Statistics for IFS 10-day Forecast

- Spectral truncation TL1279
- Horizontal Grid-Points = 2×10^6
- 16km grid-spacing
- Vertical levels =137
- Timestep = 600 Seconds

- Floating-point ops = 12×10^{15}
- MPI Communications = 150 TB
- SL Halo-width=18

- 1 Million lines of Fortran + C
- Shared with Météo-France
- Bit-reproducible
- Vector length = NPROMA
- Hybrid MPI and OpenMP

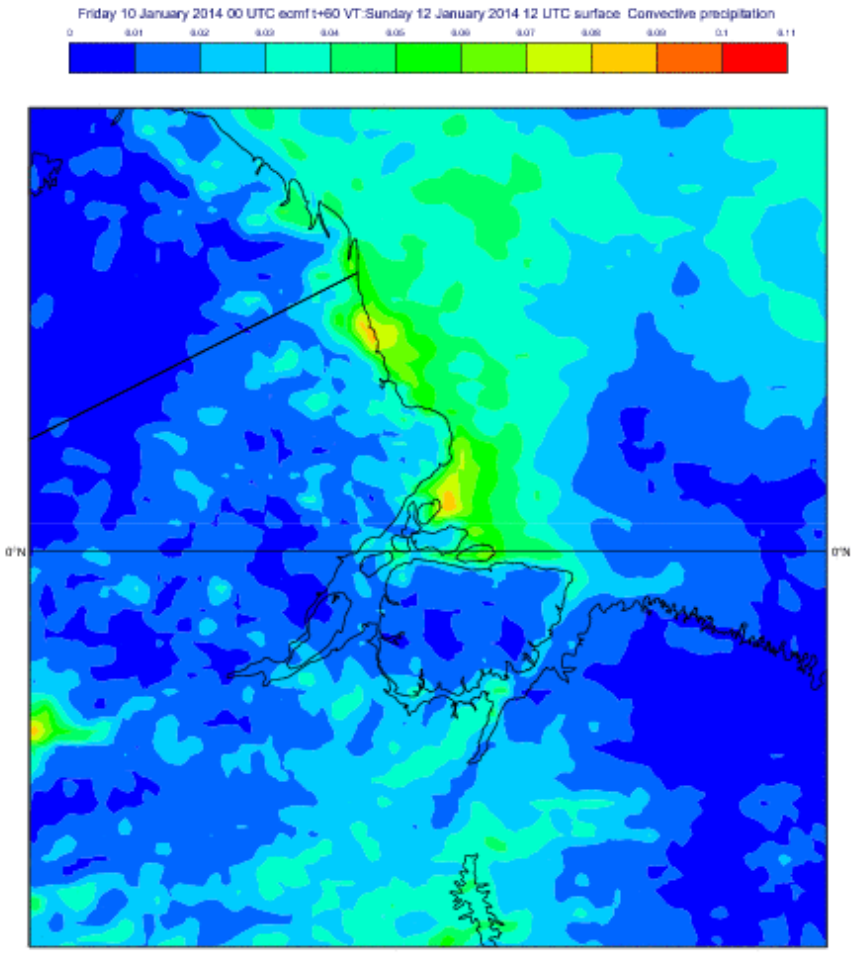
- Elapsed time ~3000 secs
- 4 Tflops

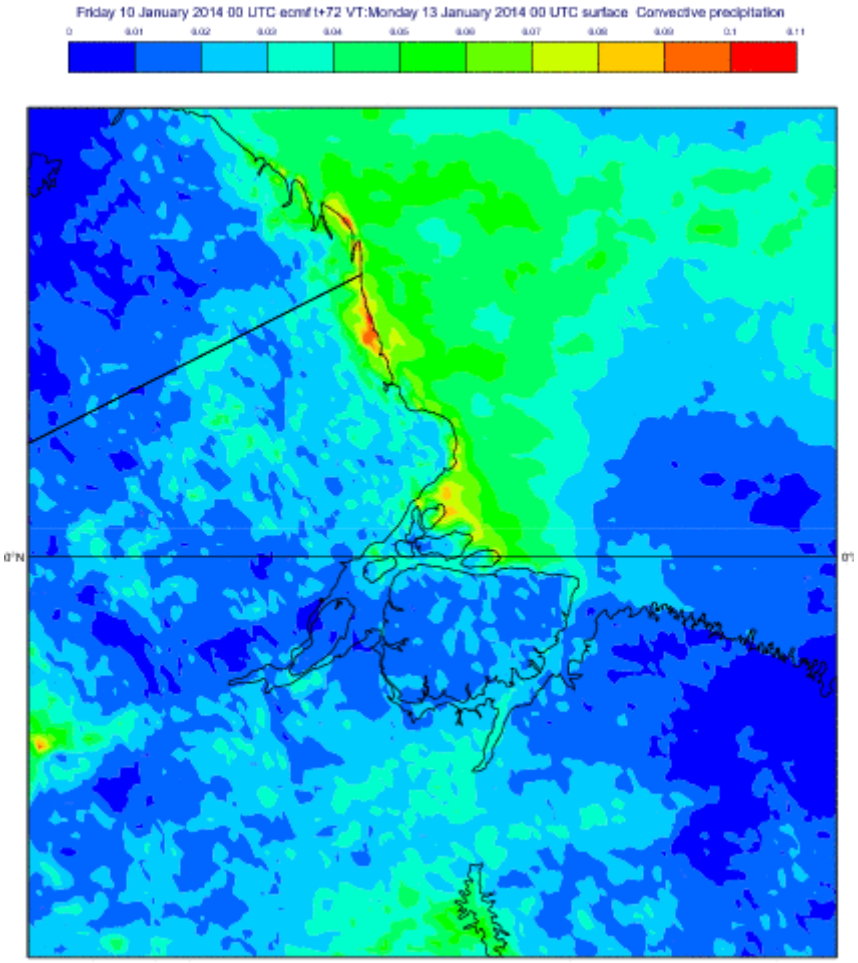
- IBM
  60 Nodes = 1920 Cores (+SMT)
  480 MPI x 8 OMP
  6.8% Peak

- Cray
  100 Nodes = 2400 Cores (+HT)
  400 MPI x 12 OMP
  7.7% Peak

* For RD config without full operational I/O

TL1279 (16km) and TC1279 (8km) Convective precipitation (accumulated over first 3 days of FC from 10 Jan 2014) in the Amazon delta





Thanks to Sylvie Malardel

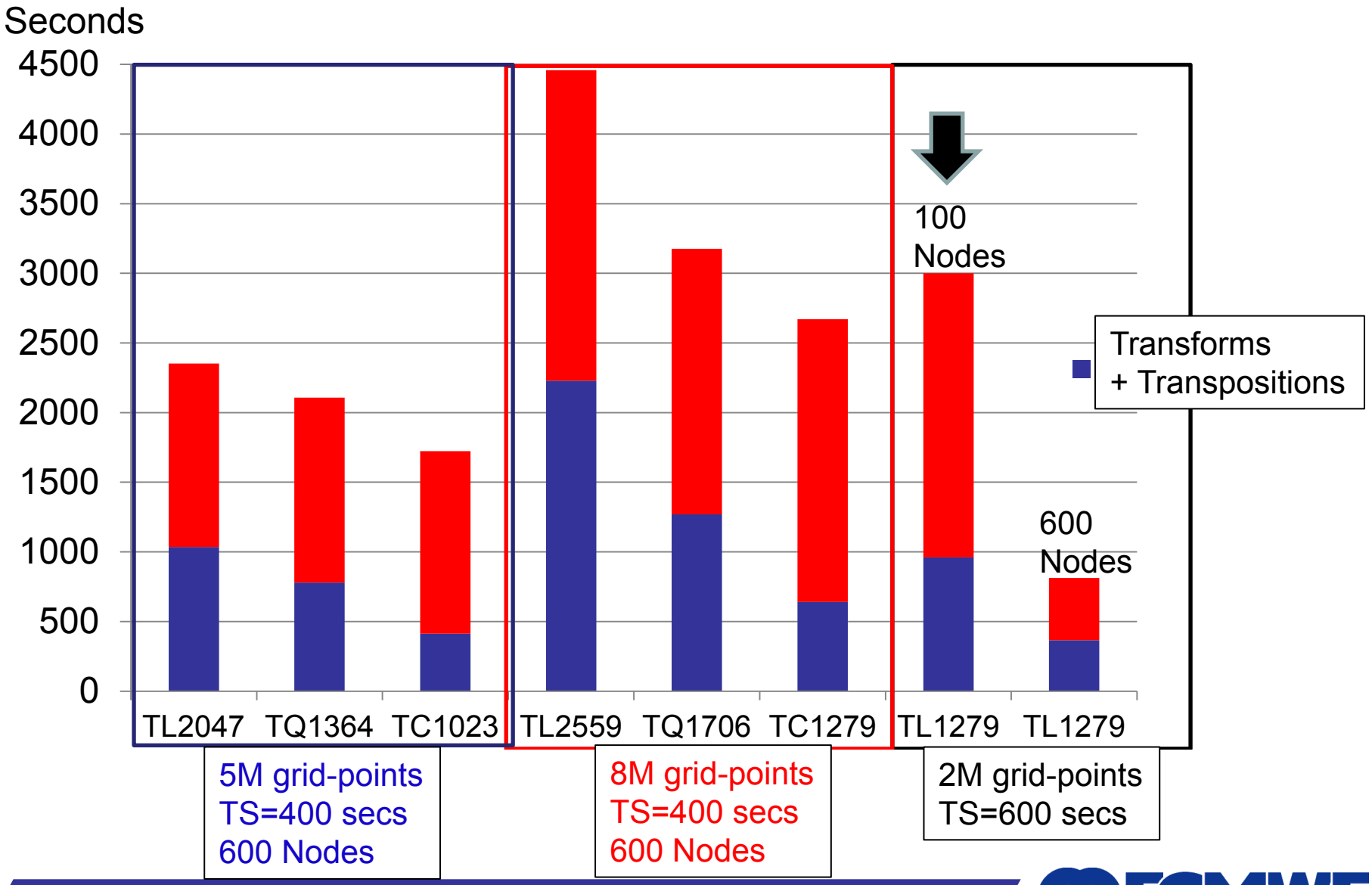
ECMWF

Choices for Higher resolution upgrade in 2015

Different Wavenumbers – Grid-point matches

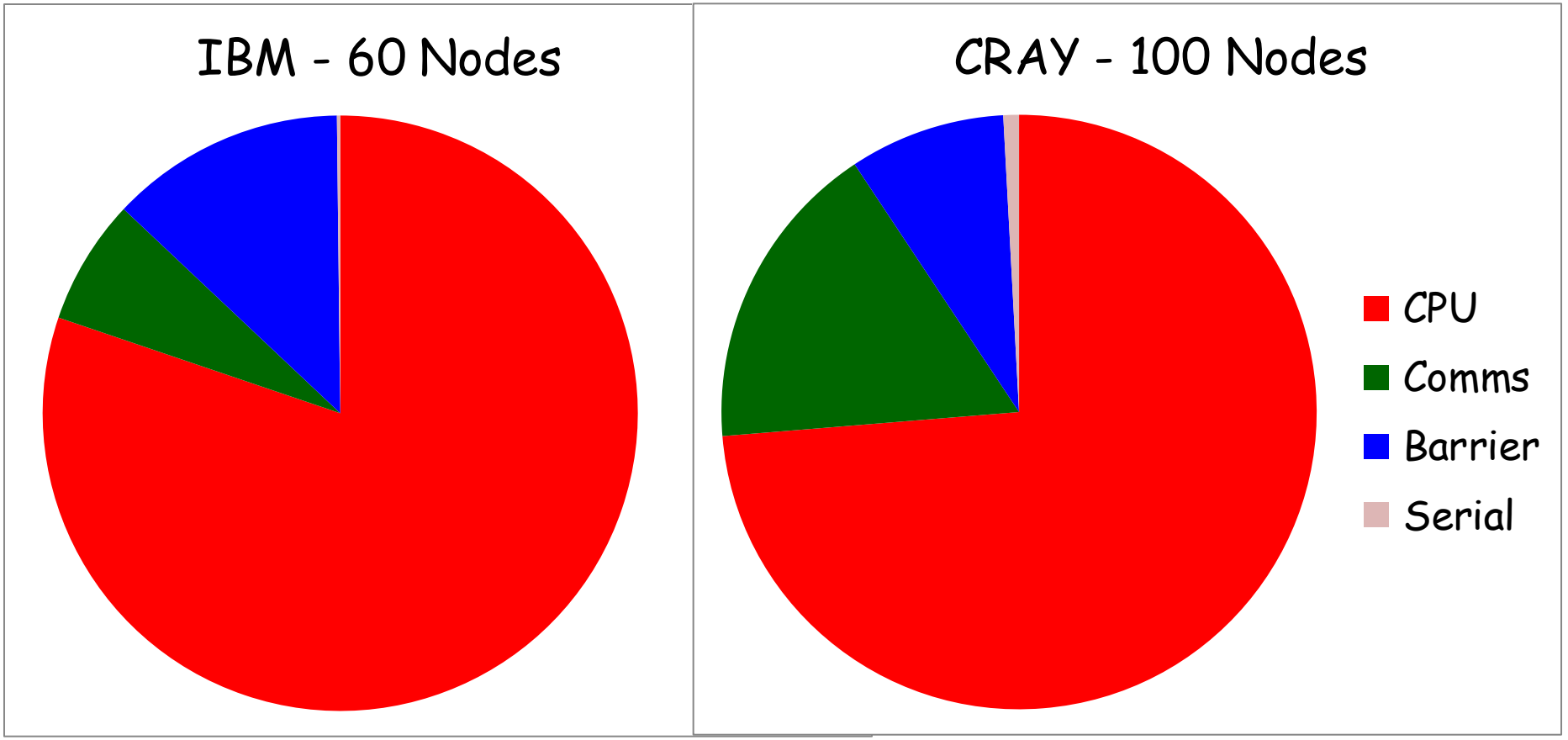
Horizontal grid-points	Linear	Quadratic	Cubic
2140702 (16km)	TL1279		
5447118 (10km)	TL2047	TQ1364	TC1023
8505906 (8km)	TL2559	TQ1706	TC1279

Costs of Different Resolutions



ECMWF

Comparison between IBM and Cray for IFS T1279 10-day Forecast

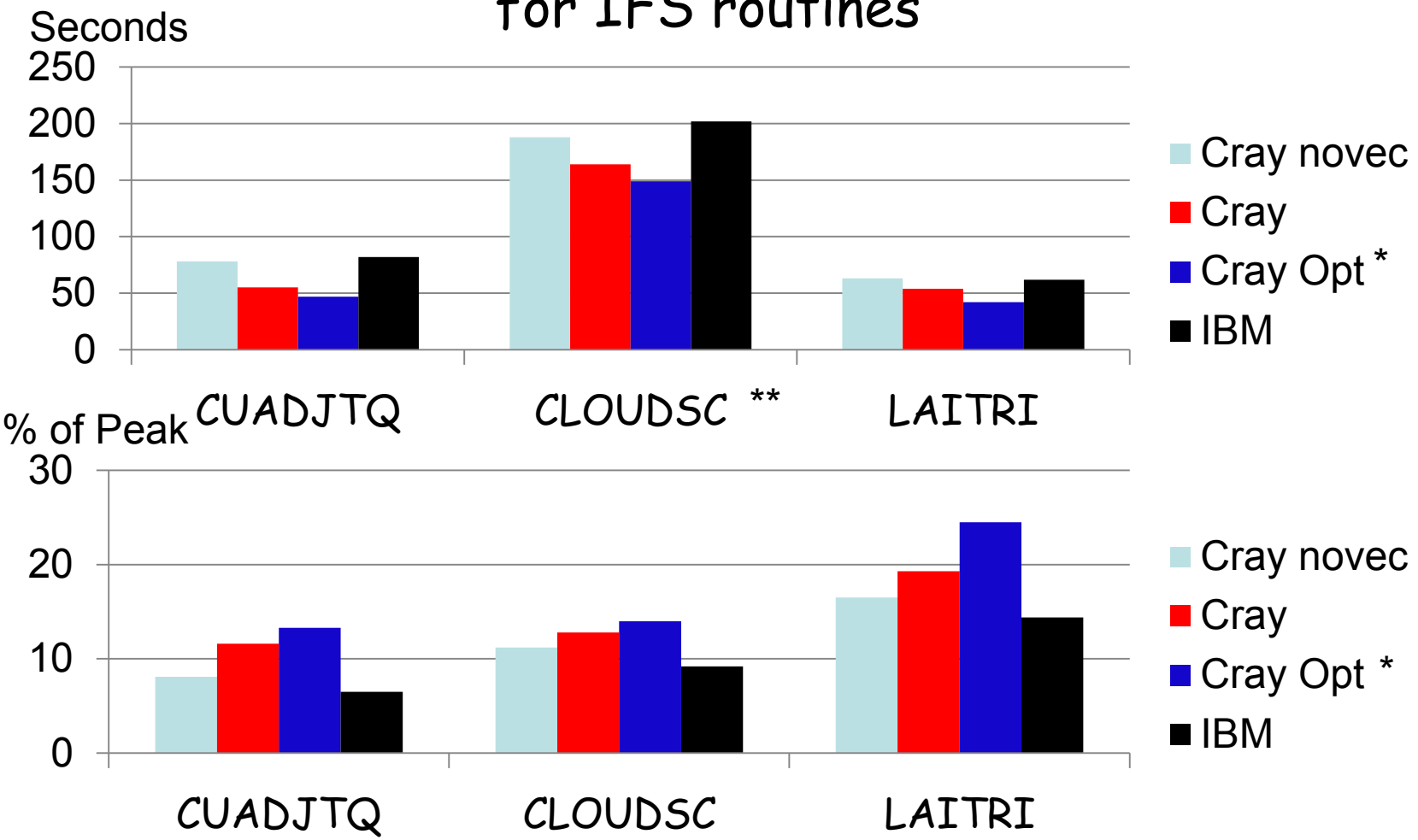


CCE compiler able to vectorise IFS source → Compute relatively faster

Aries has fewer hub chips per node than HFI → Comms relatively slower

Cray has light-weight kernel on application nodes → Less jitter

CPU Comparison: Cray (100 Nodes) and IBM (60 Nodes) for IFS routines

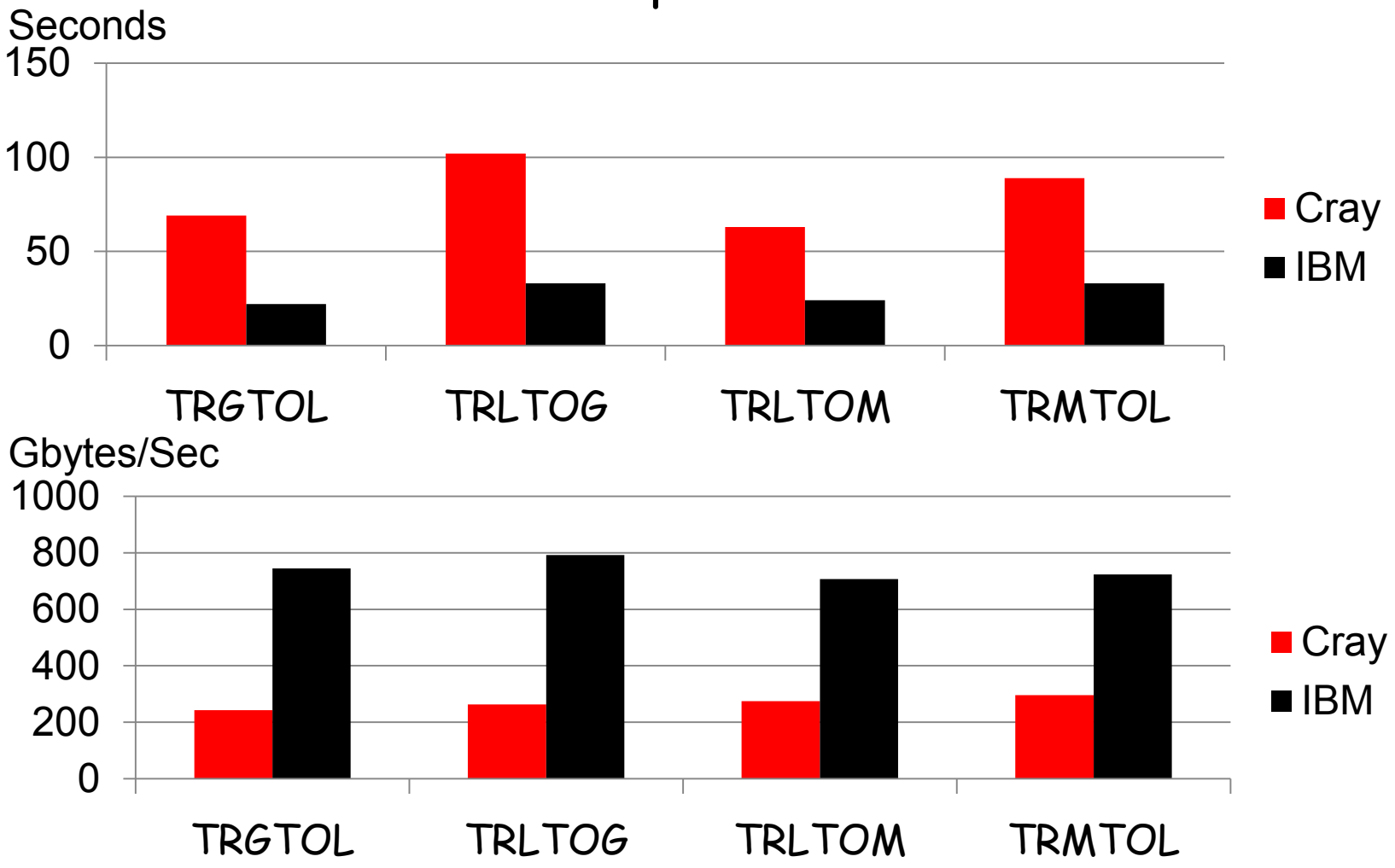


* Opt by improving vectorisation see John Hague's talk

** For more on CLOUDSC see Sami Saarinen's talk

ECMWF

Comms Comparison: Cray (100 Nodes) and IBM (60 Nodes) for IFS transposition routines



ECMWF

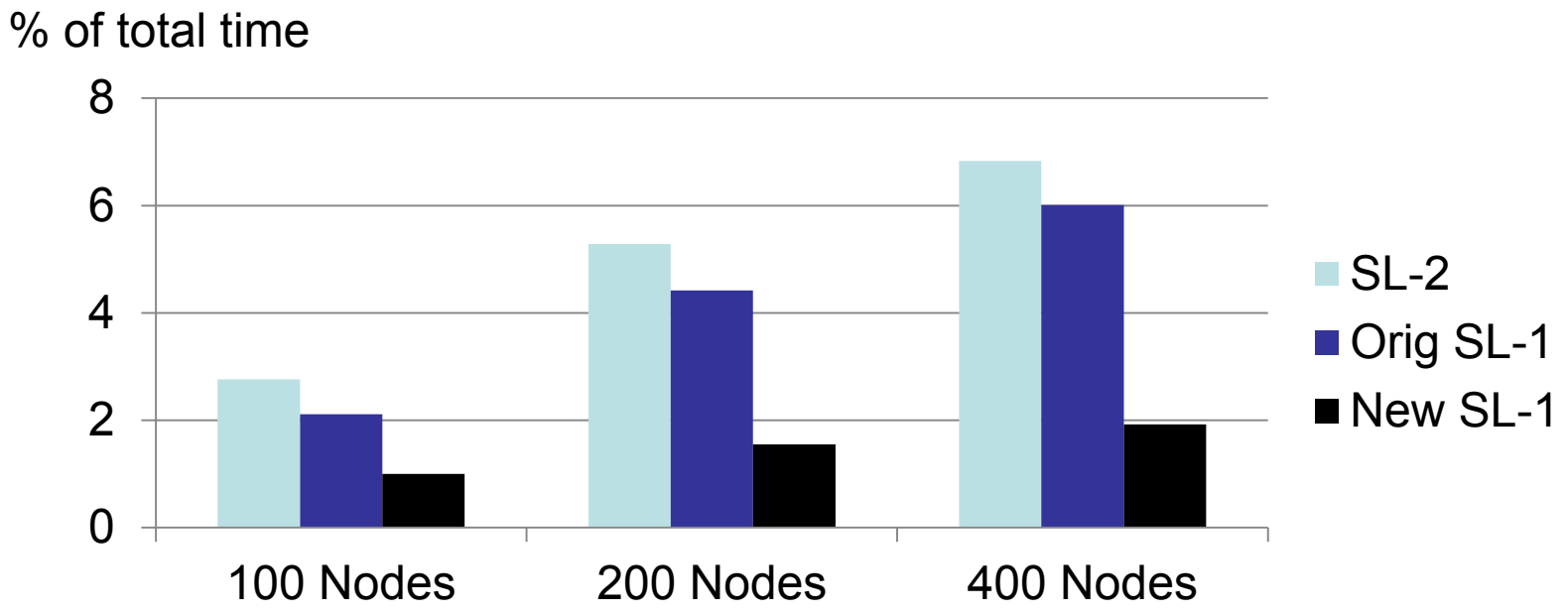
Comms developments: Overlap MPI with CPU

- Investigation of overlapping communications in direct and inverse Legendre transform by Philippe Marguinaud (Météo-France)
- MPI_alltoallv → MPI_Ialltoallv in transpositions
- Gives improvements for current and future resolutions

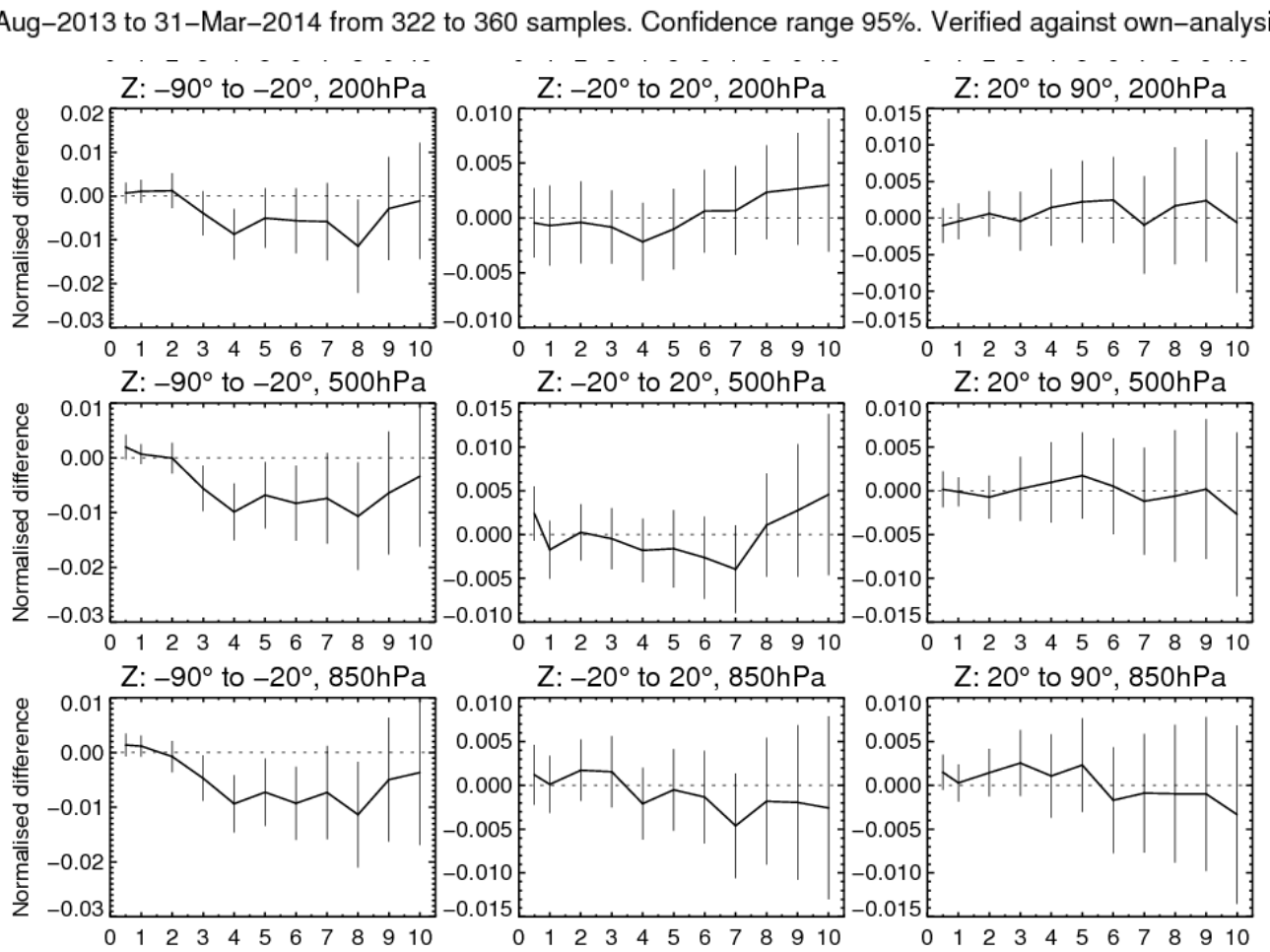
	T2047 (1024 nodes)	T1198 (64 nodes)
LT-INV: Orig	180 secs	121 secs
LT-INV: New	153 secs	113 secs

Comms developments: More on-demand SL-Comms using MPI

- 'SL-Comms part 2' already on-demand
- 'SL-Comms part 1' made on-demand to reduce volume of communications



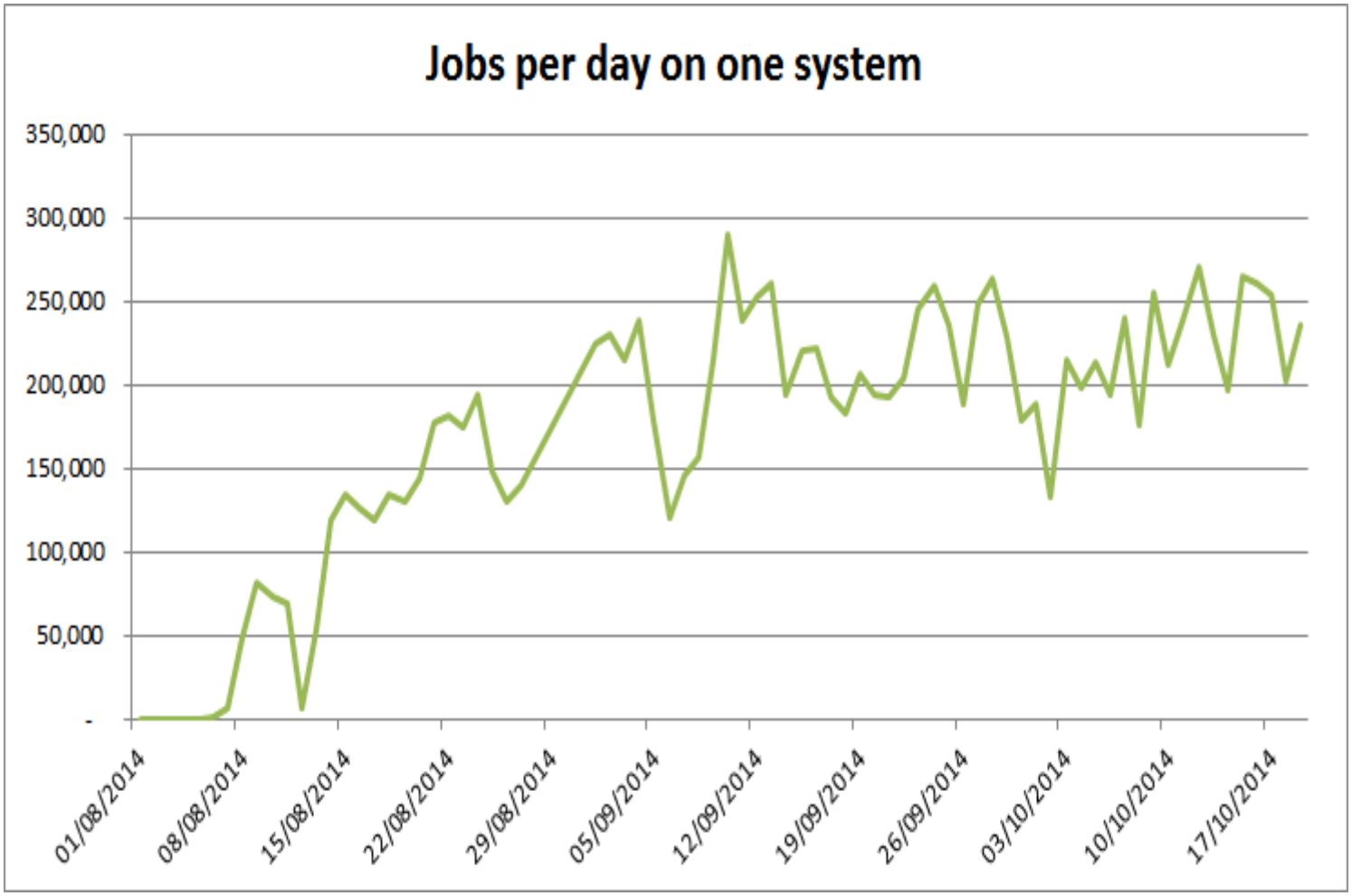
Improvement to scores with fixes to cce compiler difference of RMS error between runs with 8.2.7 & 8.2.0



Thanks to Alan Geer

Migration Timeline

Delivery of first cluster	*6 November 2013*
First user access	*6 December 2013*
Acceptance tests on first cluster	*6 February 2014*
Switch off of first cluster of old HPC	*8 April 2014*
Commissioning of second Cray cluster	9 April 2014
Acceptance tests on second cluster	1 August 2014
First operational forecast from Cray	17 September 2014
Final acceptance of Cray systems	30 September 2014


Jobs per day on one system
350,000
300,000
250,000
200,000
150,000
100,000
50,000
-
01/08/2014
08/08/2014
15/08/2014
22/08/2014
29/08/2014
05/09/2014
12/09/2014
19/09/2014
26/09/2014
03/10/2014
10/10/2014
17/10/2014

Operations on Cray

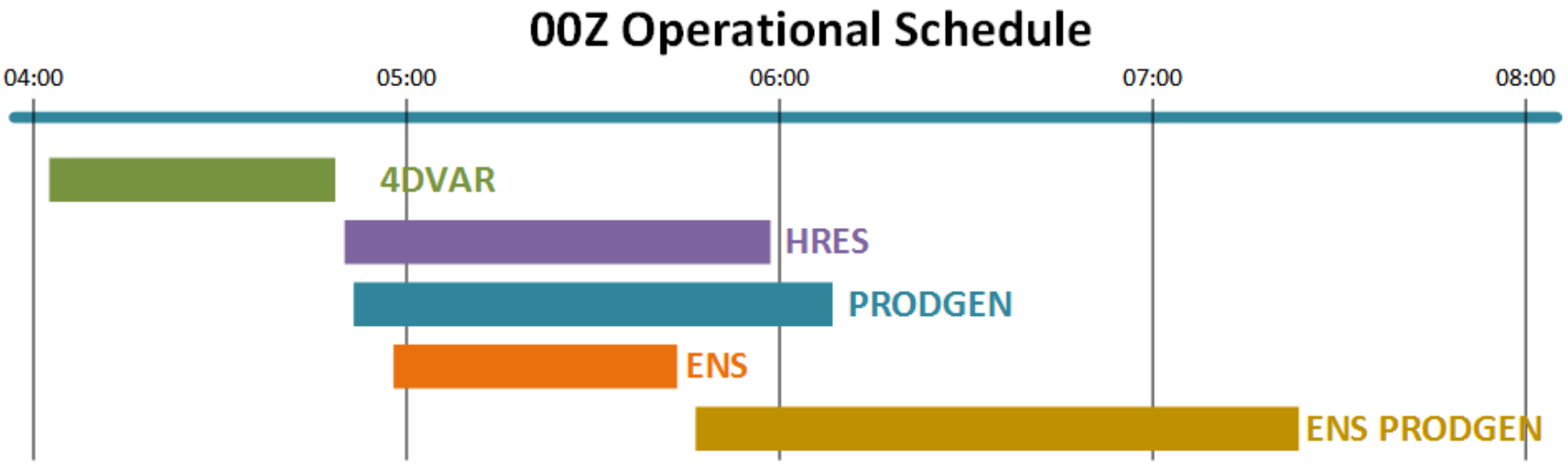
- Operational from mid September
- Many thanks to Cray staff
  - A huge effort by a large team of people
  - Both local and in the USA

Operations on Cray

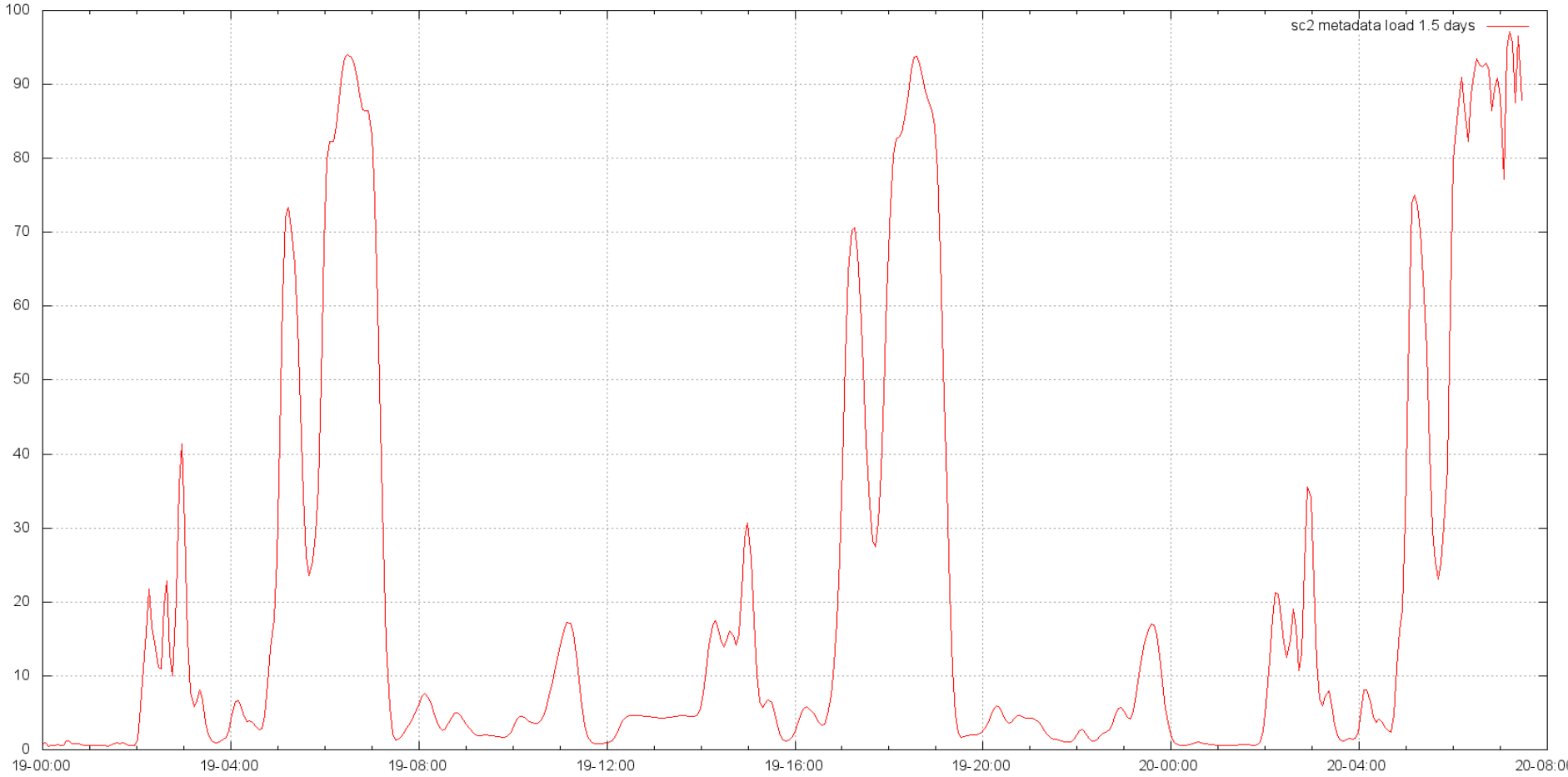
- Operational from mid September
- Many thanks to Cray staff
  - A huge effort by a large team of people
  - Both local and in the USA
- But……

We have a Tuning Challenge

- Lustre is NOT GPFS
  - Different performance characteristics
  - Seeing delays due to IO jitter
- Need to streamline
  - Both workflow and IO load
- Tuning efforts started
- Cray to provide additional expertise

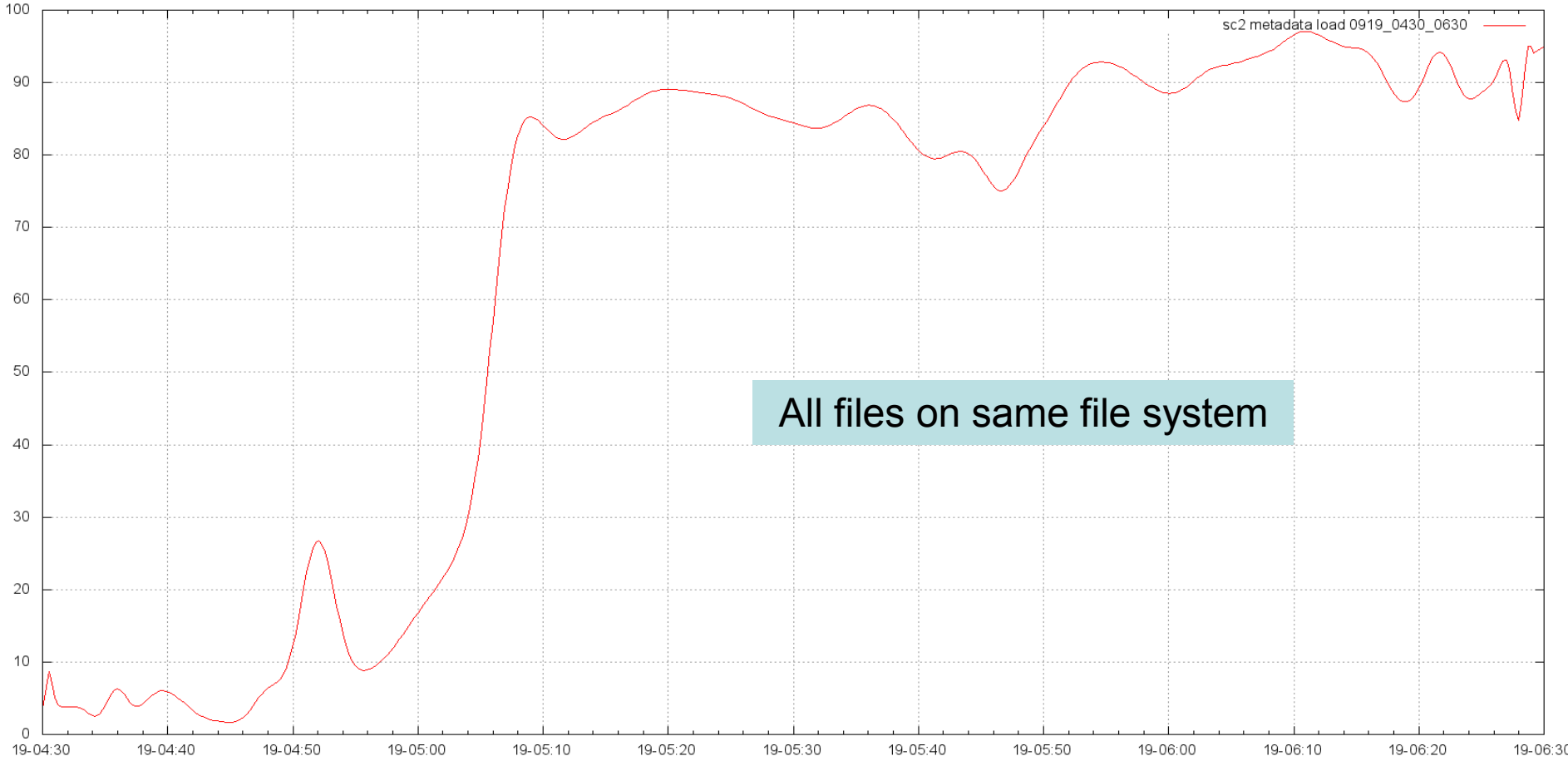

00Z Operational Schedule
04:00
05:00
06:00
07:00
08:00
4DVAR
HRES
PRODGEN
ENS
ENS PRODGEN

Operational File System on CCB: CPU Load on Meta Data Server on Oct 19



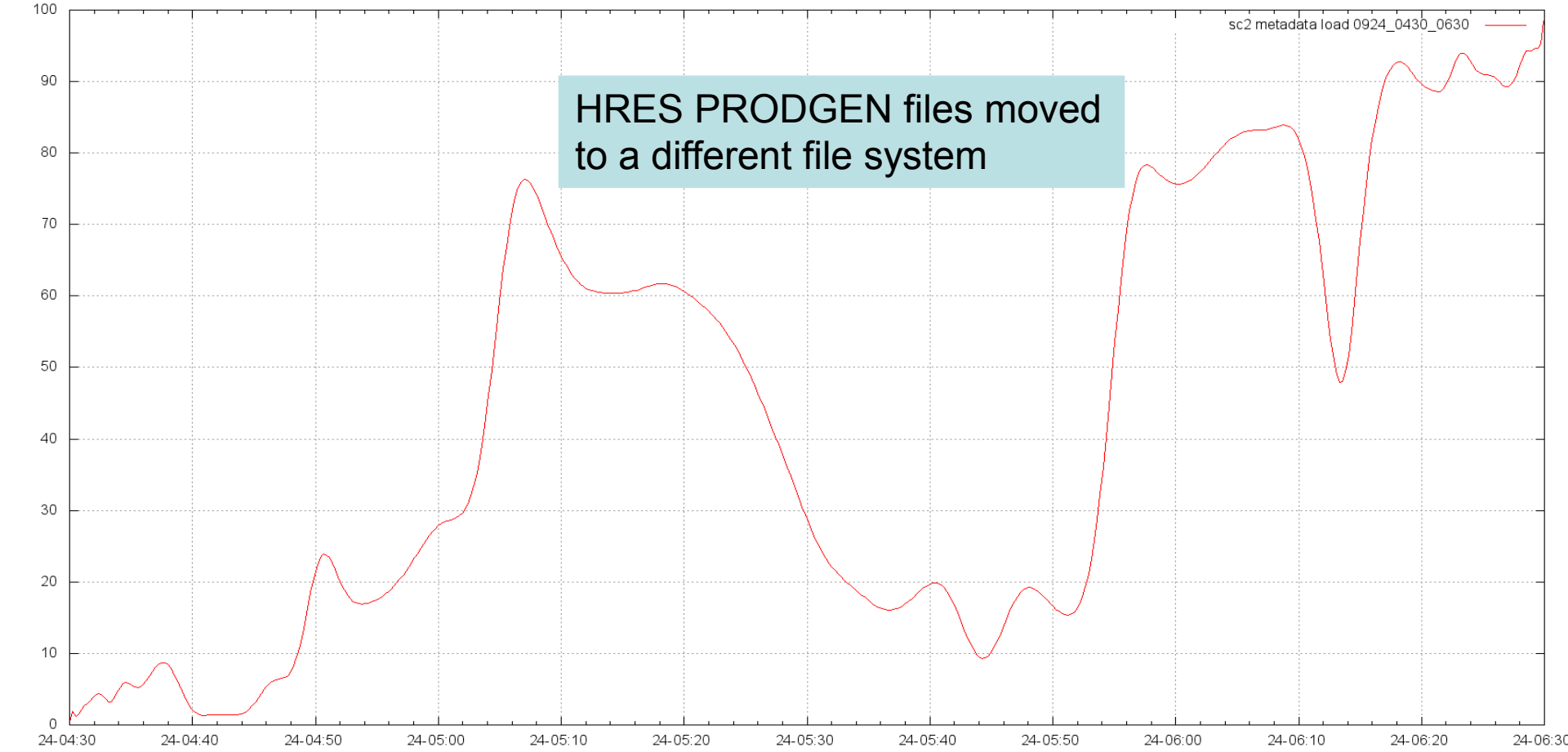
ECMWF

Operational File System on CCB: CPU Load on Meta Data Server on Sept 19



ECMWF

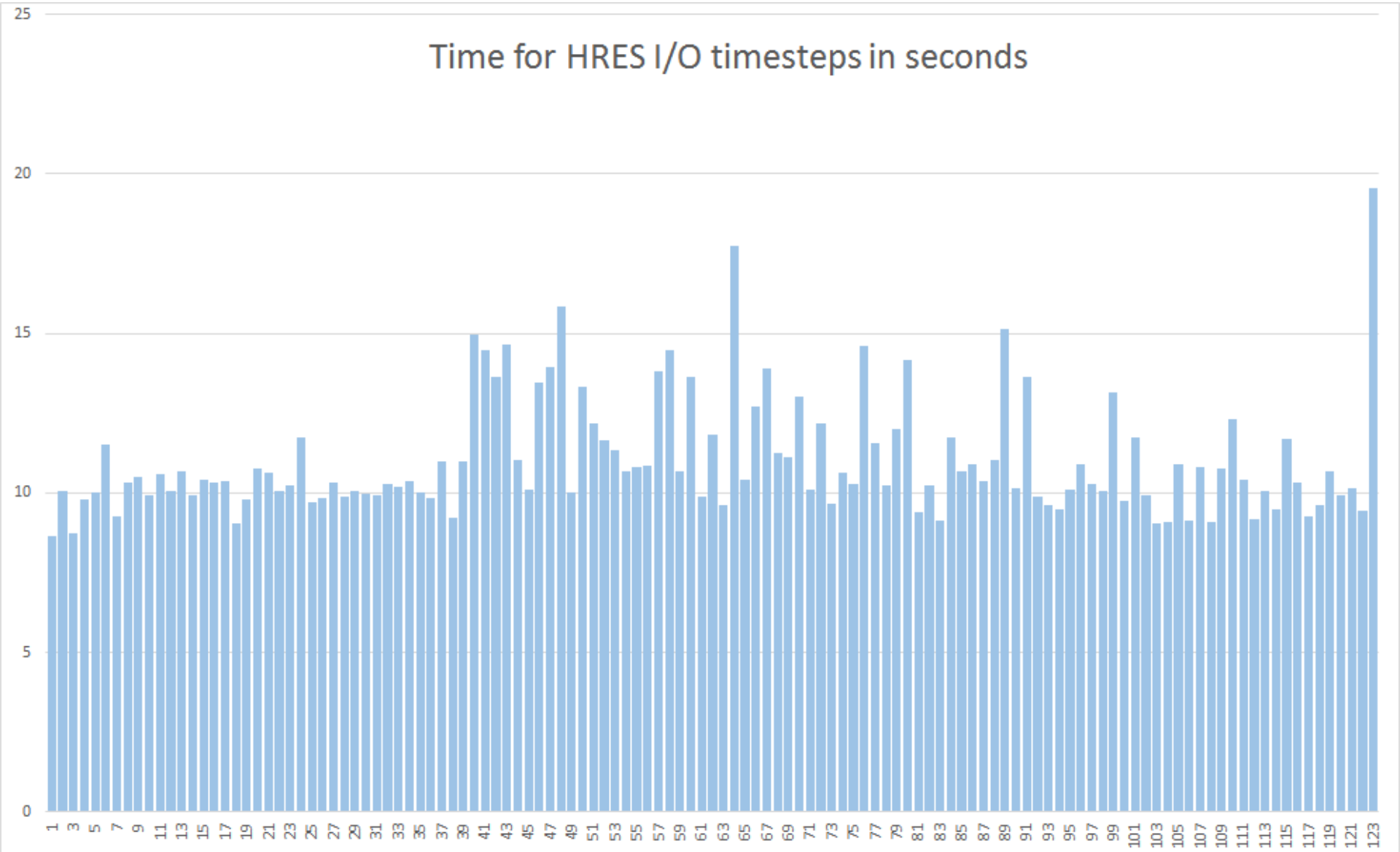
Operational File System on CCB: CPU Load on Meta Data Server on Sept 24



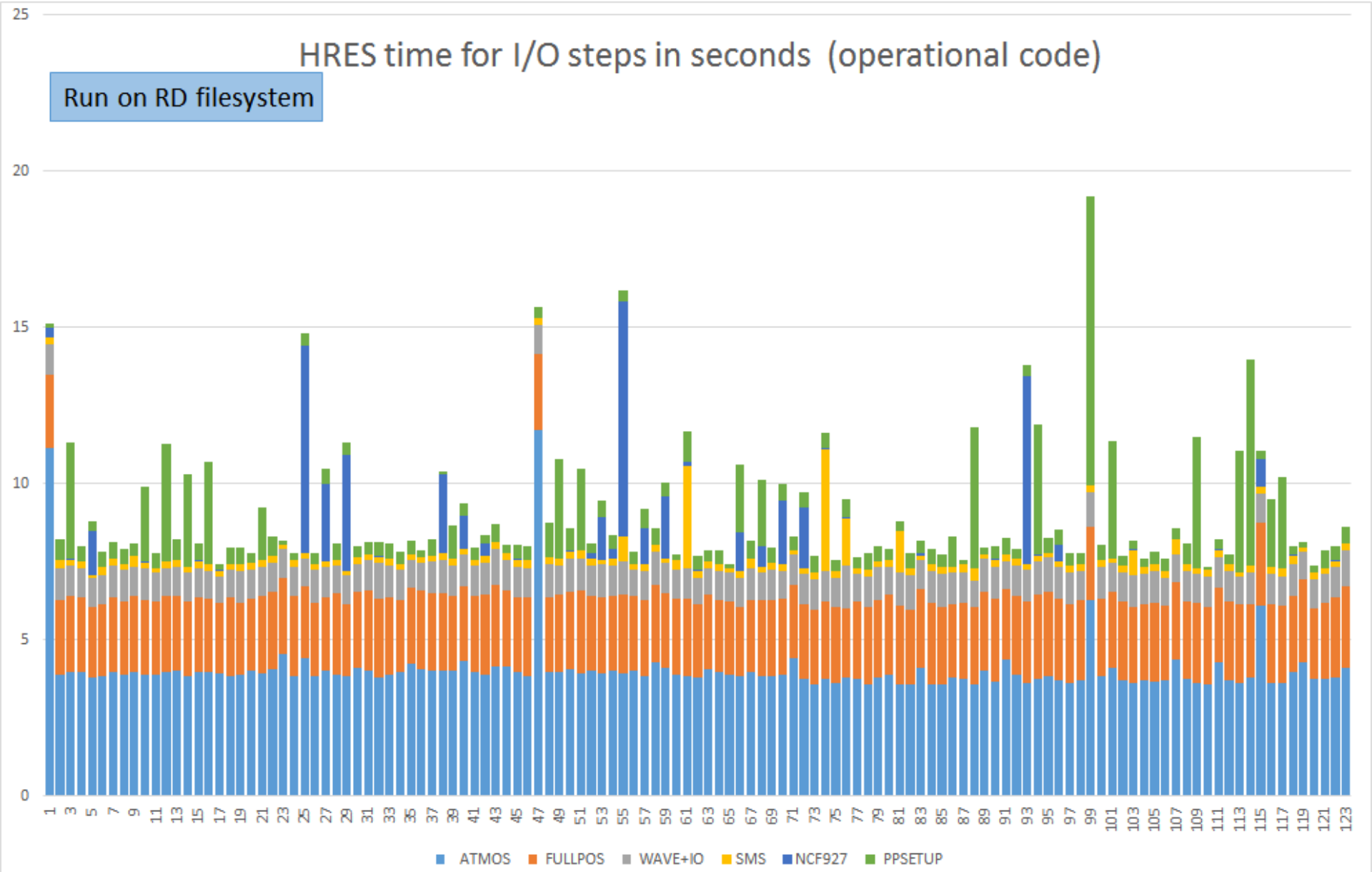
ECMWF

A Look at HRES IO

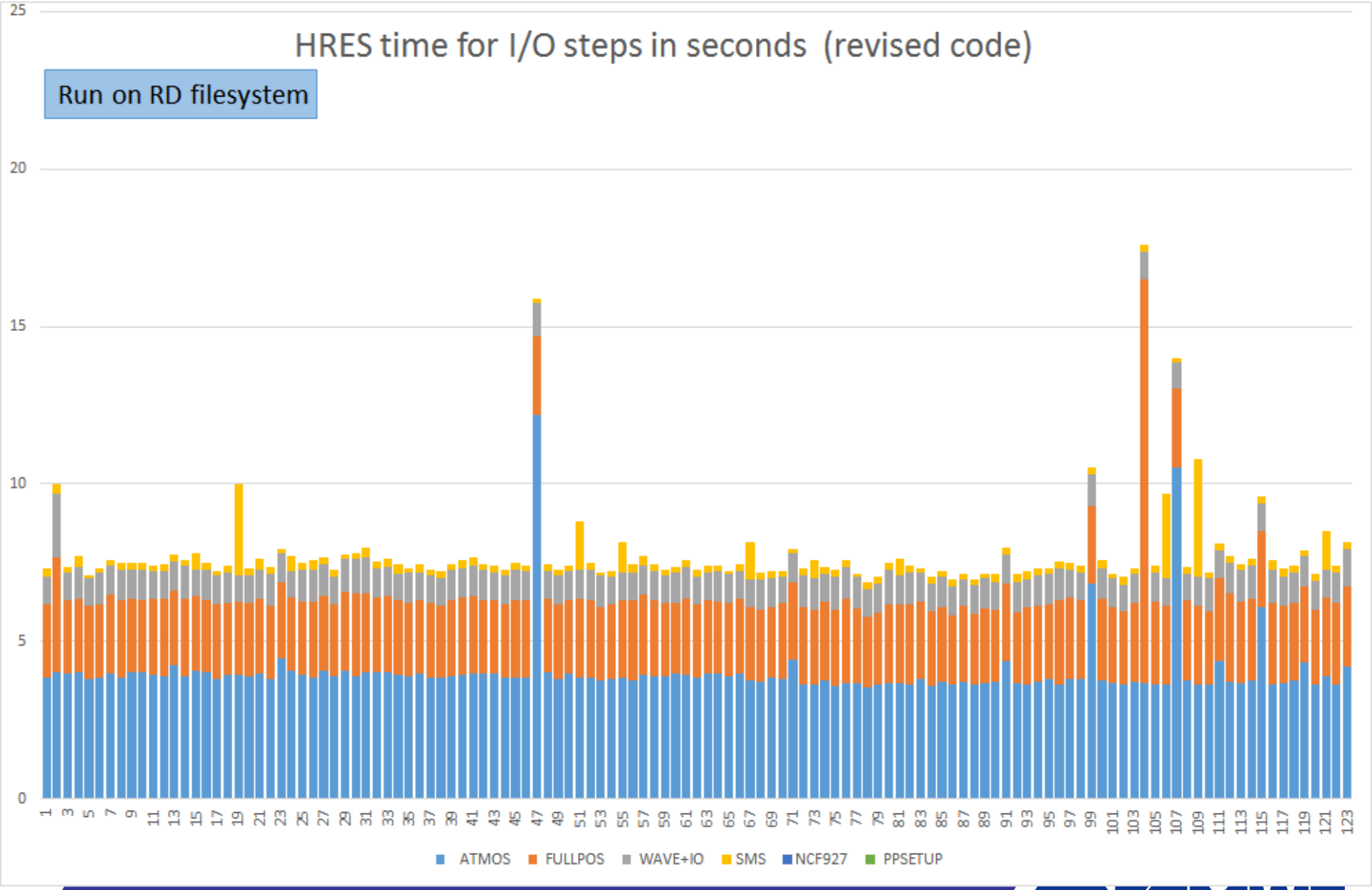
- 1.2TB total over 125 Post Processing IO steps
- Output to 120 files
- Bulk IO goes well (200+MB/s per file)
  - some delays due to jitter
- But we have IO jitter elsewhere
- Gets worse on a heavily loaded file system
- Tracked down to other IO operations every PP step
  - Open/read/close a file called dirlist by every task
  - Open/read/close 2 name list files by every task
  - Open/write/close a file called NCF927 by the master task


Time for HRES I/O timesteps in seconds
25
20
15
10
5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 109 111 113 115 117 119 121 123

ECMWF


HRES time for I/O steps in seconds (operational code)
Run on RD filesystem
25
20
15
10
5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 109 111 113 115 117 119 121 123
ATMOS
FULLPOS
WAVE+IO
SMS
NCF927
PPSETUP

ECMWF


HRES time for I/O steps in seconds (revised code)
Run on RD filesystem
25
20
15
10
5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 109 111 113 115 117 119 121 123
ATMOS
FULLPOS
WAVE+IO
SMS
NCF927
PPSETUP

ECMWF

IO that needs to be removed

HRES – 180K sets of open/read or write/close

ENS - 1.2M sets of open/read or write/close

HRES PRODGEN – 930K temporary files

- 2.8 M sets of open/read or write/close

ENS PRODGEN - 880K temporary files

- 2.6M sets of open/read or write/close

Lessons Learned

- Ever increasing complexity
- More resources required
  - To meet applications and systems challenges
- Must allow more time for performance testing of the operational suite