Performance of IFS on ECMWF's new HPCF

- HPCF configurations
- T1279 L91 10-day forecast on P7 & P6
- P7 compared with P6
  - CPU
  - Comms
  - Jitter
- Scalability of T2047 L137
- Latest T7999 tests

Deborah Salmond and Peter Towers

ECMWF

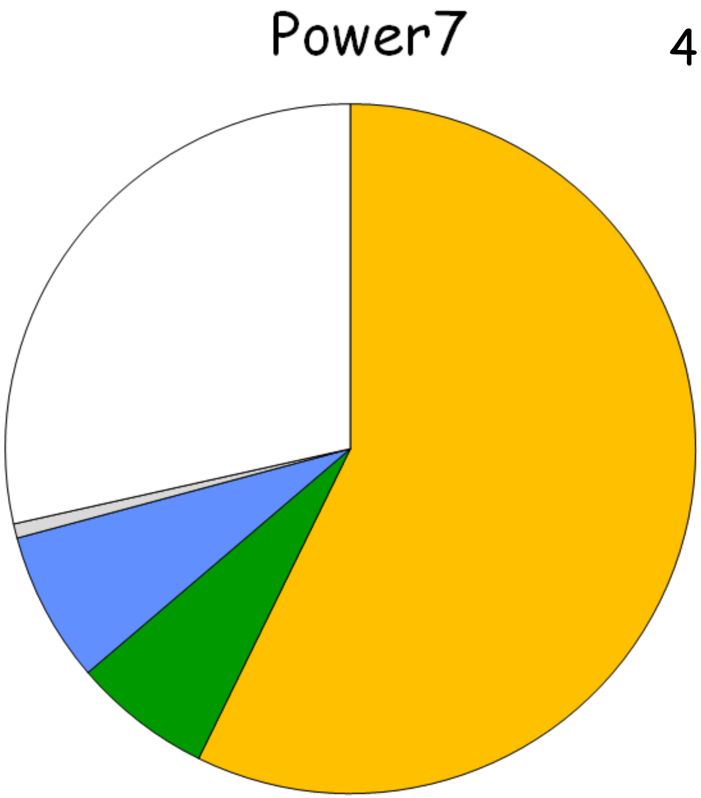
C2a – Power7 (11 Frames – 24k cores)

C1a – Power6 (24 Frames - 9k cores)

Power6 → Power7

	c1a & c1b	c2a & c2b
Contract phase	Phase 1	Phase 2
Processor	Power6	Power7
Clock	4.7 GHz	3.8 GHz
Peak Gflops /Core	18.8	2 * 15.2 (incl VSX)
Application nodes / cluster	262	732
Cores / cluster	8384	23424
Cores / node	32	32
SMT threads/core	2	2
Switch	IB – 8 links per node	HFI – 31 links per node

IFS T1279-L91: 10-day Forecast: CY38R1
48 Nodes: 384 MPI tasks * 8 OpenMP threads



48 Nodes = 1536 Cores = 3072 SMT threads
= no. of nodes for ECMWF operational T1279 forecast and 4D-Var

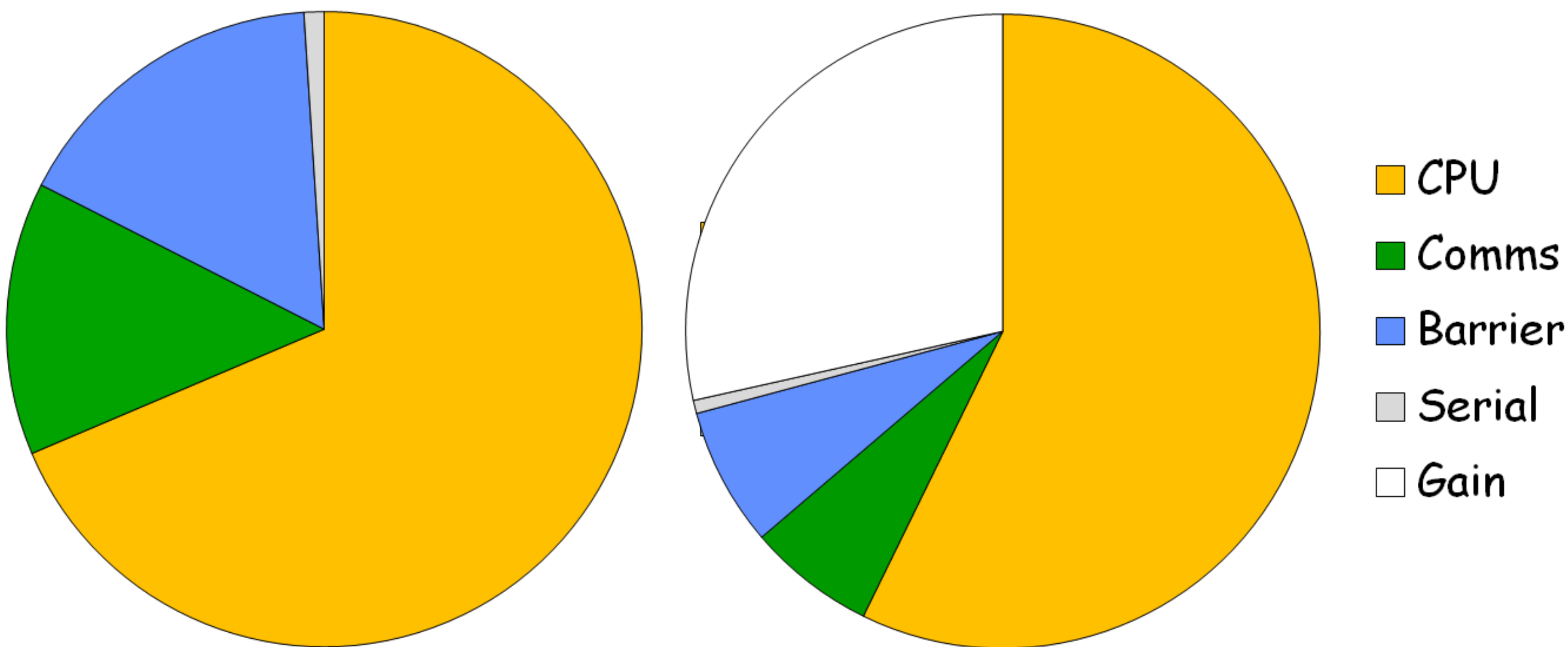
- CPU = time in OpenMP loops
- Comms = time in MPI communications
- Barrier = time at MPI barrier
- Serial = time not in OpenMP loops
- Gain = time on P6-time on P7

Notes on Barrier:
- Extra barriers inserted to get timings
- Barrier is a measure of Load imbalance + Jitter

Totals: 6.8 Pflop, 652 GB of Memory, 93 TB of Comms

IFS T1279-L91: 10-day Forecast: CY38R1
48 Nodes: 384 MPI tasks * 8 OpenMP threads

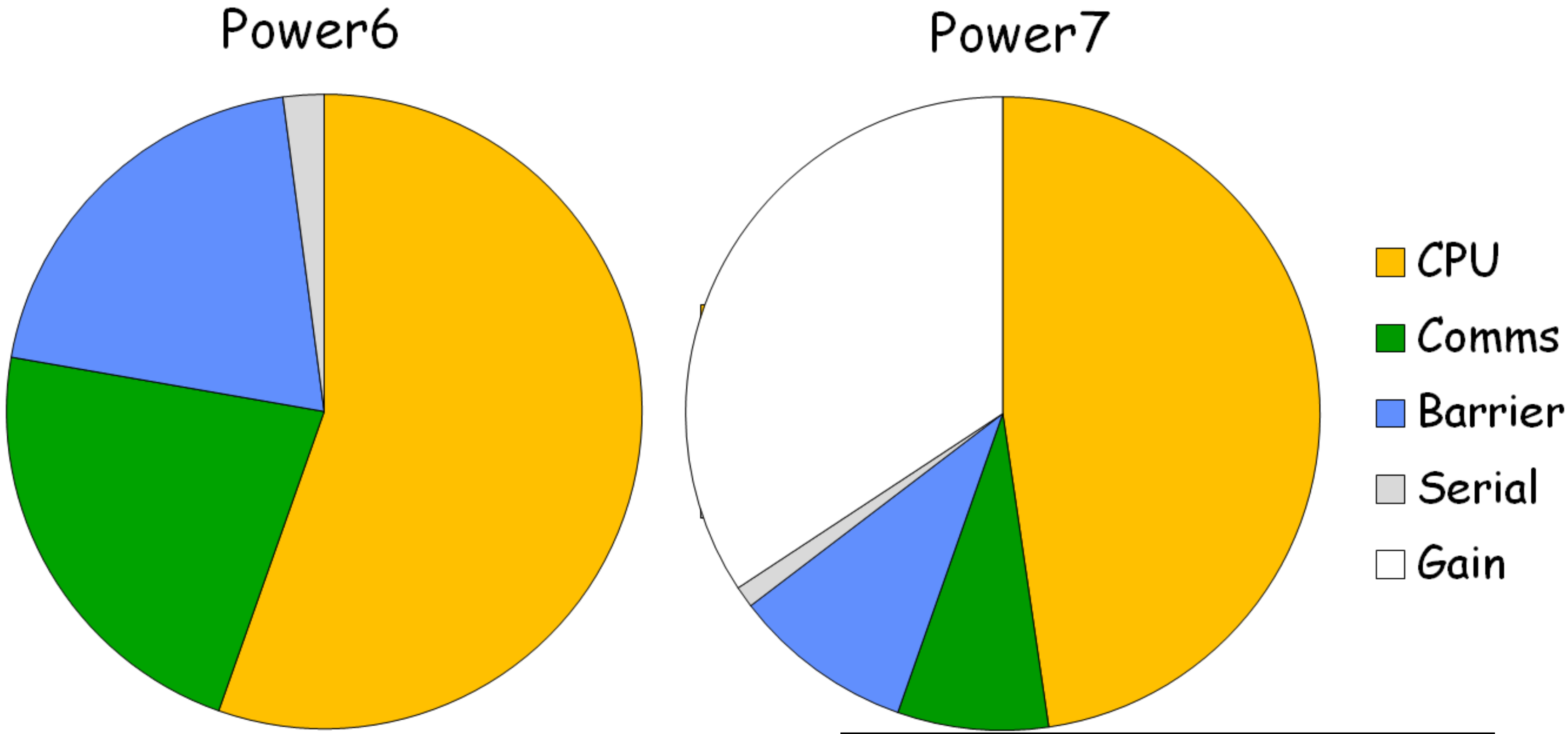
Power6 Power7



Power6	Power7
2813 seconds 2.4 Tflops (8.3% peak)	2007 seconds 3.4 Tflops (7.3% peak*)

* incl VSX

IFS T1279-L91: 10-day Forecast: CY38R1
144 Nodes: 576 MPI tasks * 16 OpenMP threads



1246 seconds
5.5 Tflops
2.25 speed-up from 48 nodes

809 seconds
8.5 Tflops
2.5 speed-up from 48 nodes

CPU on P7(oo) vs P6(not-oo)

Mflops per thread (runs with SMT)

ROUTINE	P6 (Mflops)	P7 (Mflops)	Ratio
CLOUDSC 'Many IF tests'	741	1577	2.1
LASCAW	138	615	4.4
LAITRI	1222	2439	2.0
LTDIR 'Matrix multiply'	5392	7837	1.4
TOTAL	780	1108	1.4

P6 peak Mflops per thread = 9400
P7 peak Mflops per thread = 15200 (incl. VSX)
= 7600 (if VSX not used)

ECMWF

Comms on P7(HFI) vs P6(IB)

Total Comms rates

ROUTINE	P6 (GB/s)	P7 (GB/s)	Ratio
SLCOMM2A 'Fat Halo'	268	724	2.7
TRGTOL 'Transpose local'	189	475	2.5
TRLTOM 'Transpose non-local'	189	244	1.3
TOTAL	179	382	2.1

HFI Switch

- Very fast communications between (8) nodes in a drawer
- Fast comms between (32) nodes in a super-node

On Node

- Use Shared memory for comms on-node

Jitter on P7 and P6

Jitter is random delays in tasks coming from a wide variety of causes

- Jitter test code:

  Each Task does 1000 repetitions of following sequence:

Barrier

Short CPU routine (0.4 msec)
time and save

Barrier

Long CPU routine (40 msec)
time and save

Barrier

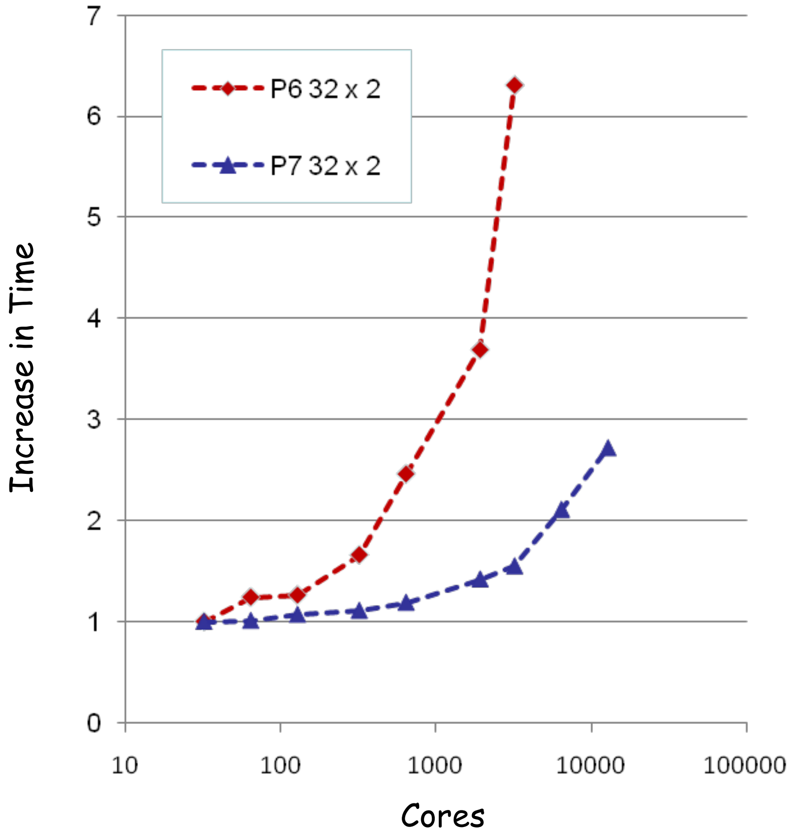
FATHALO routine
time and save

Barrier

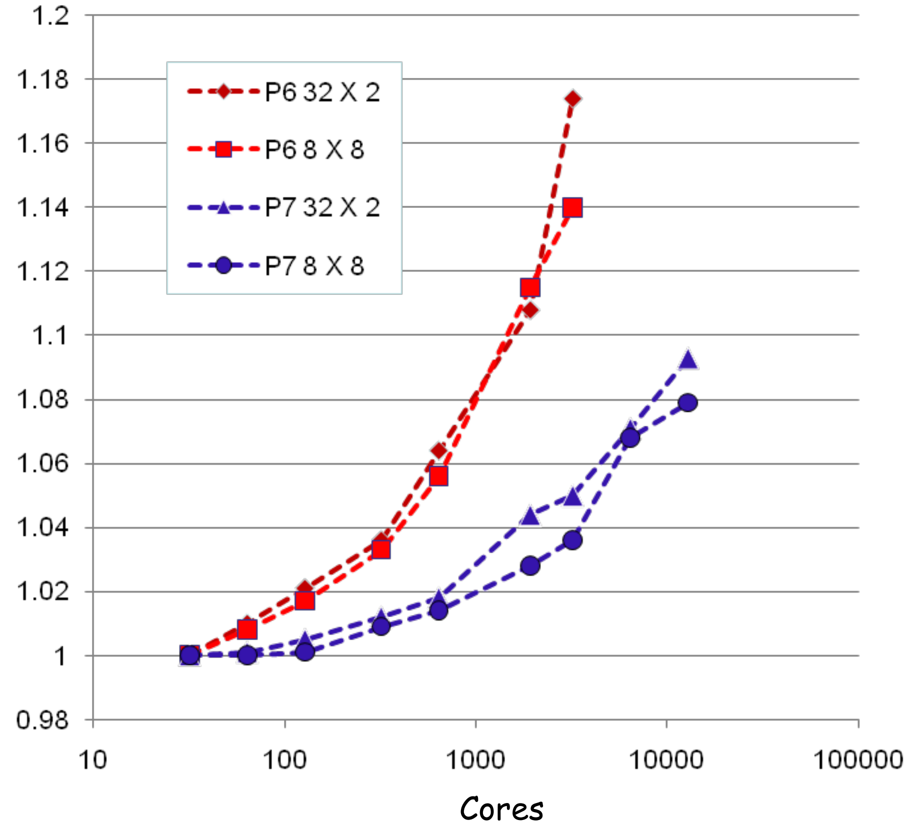
etc, etc

Slides from John Hague

Jitter measurements on P6 and P7

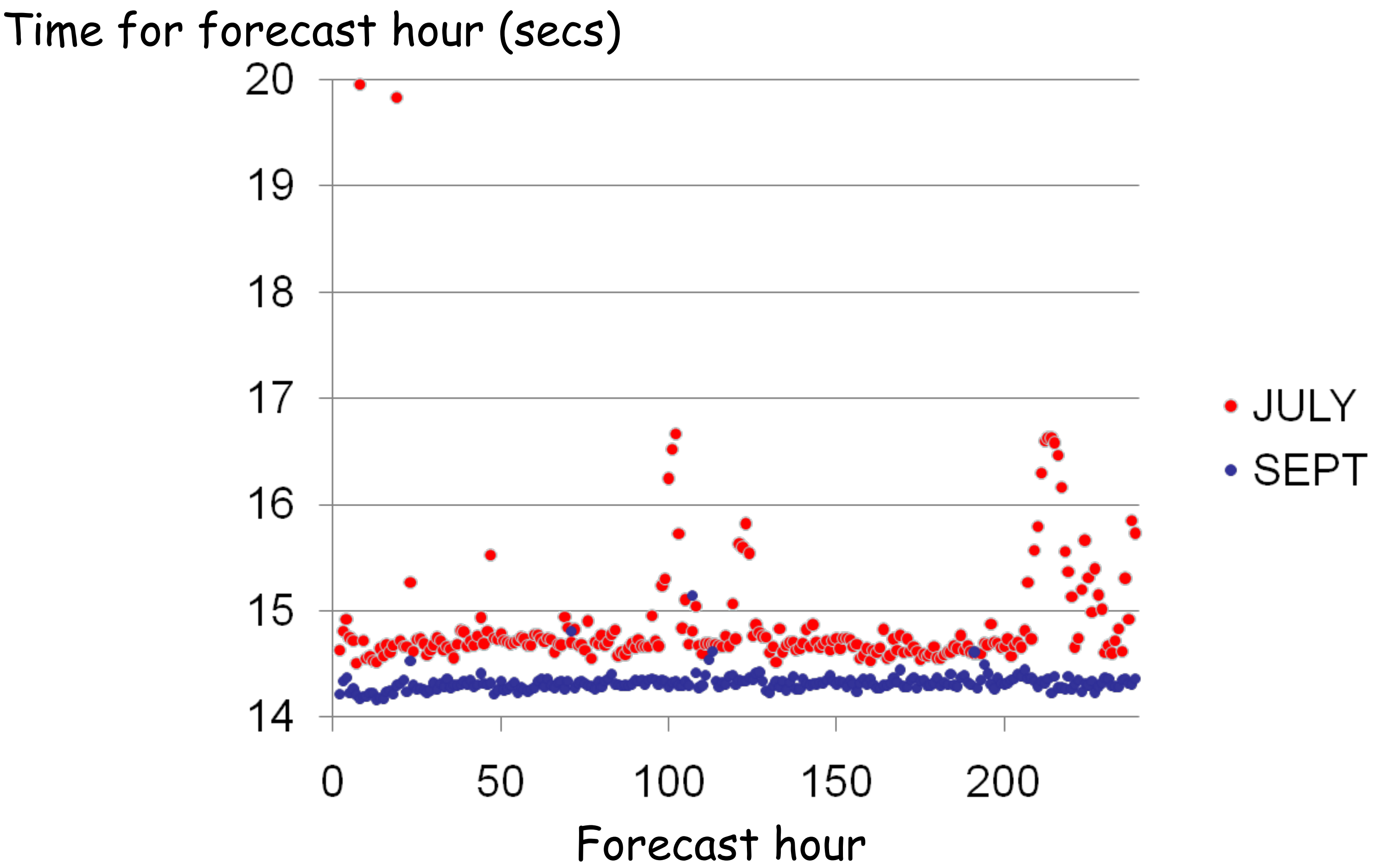


Short CPU between barriers (0.4 msec)



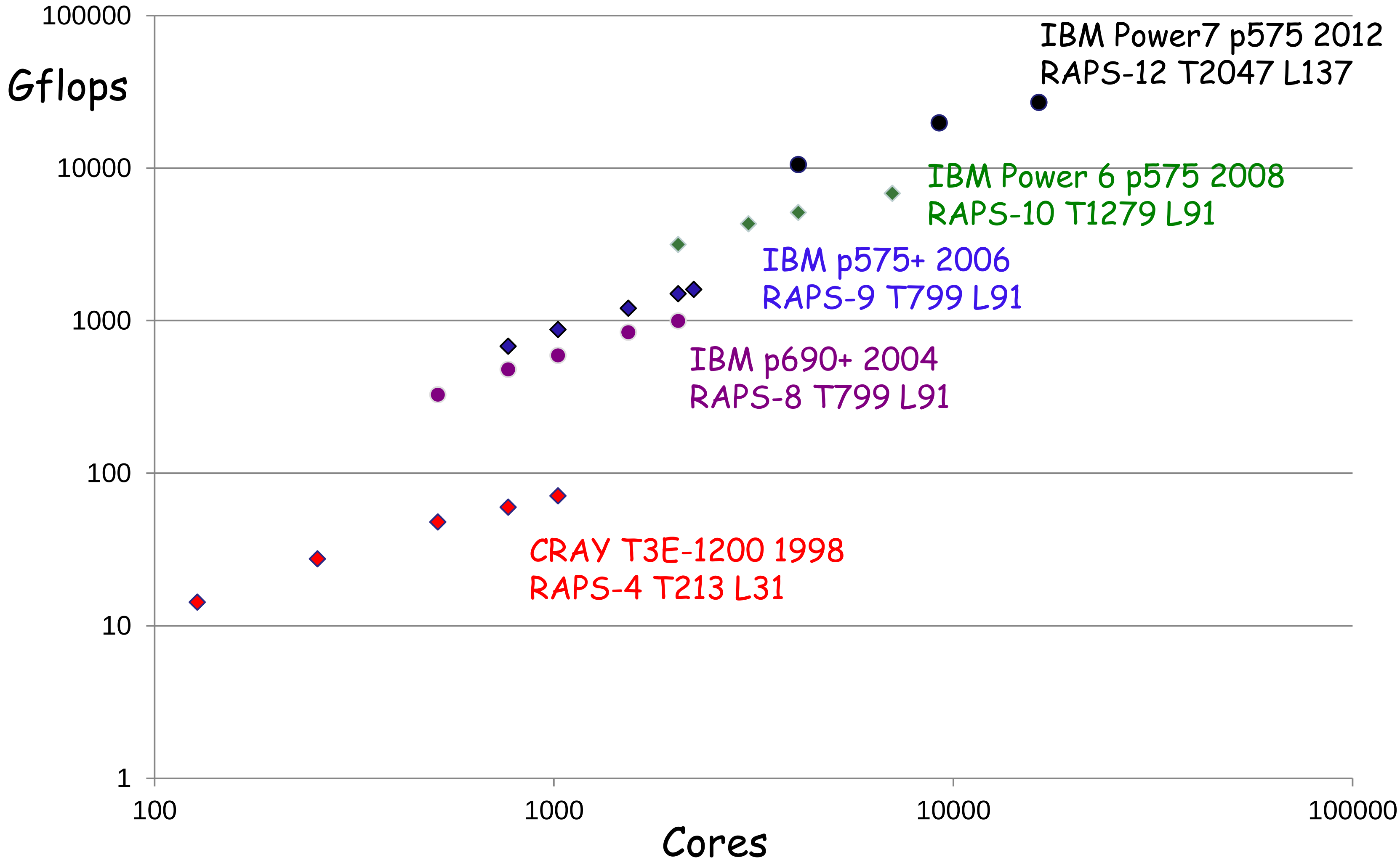
Long CPU between barriers (40 msec)

Effect of Jitter on IFS on c2a



SEPT = Correct environment Variables + Work of John Lewars

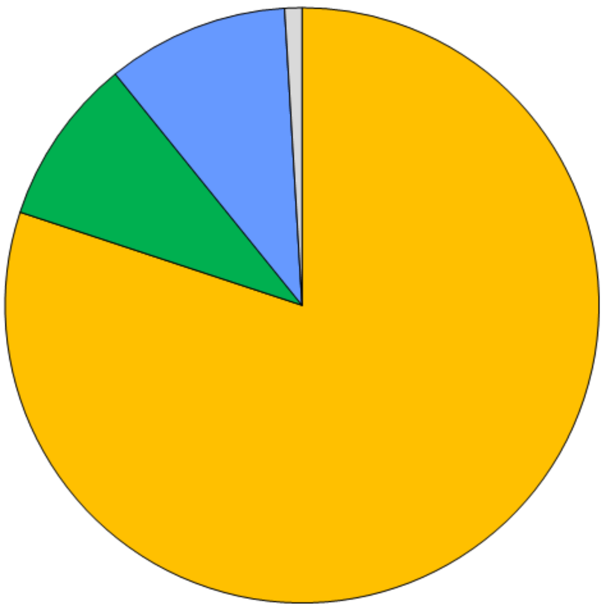
History of IFS Forecast scalability



ECMWF

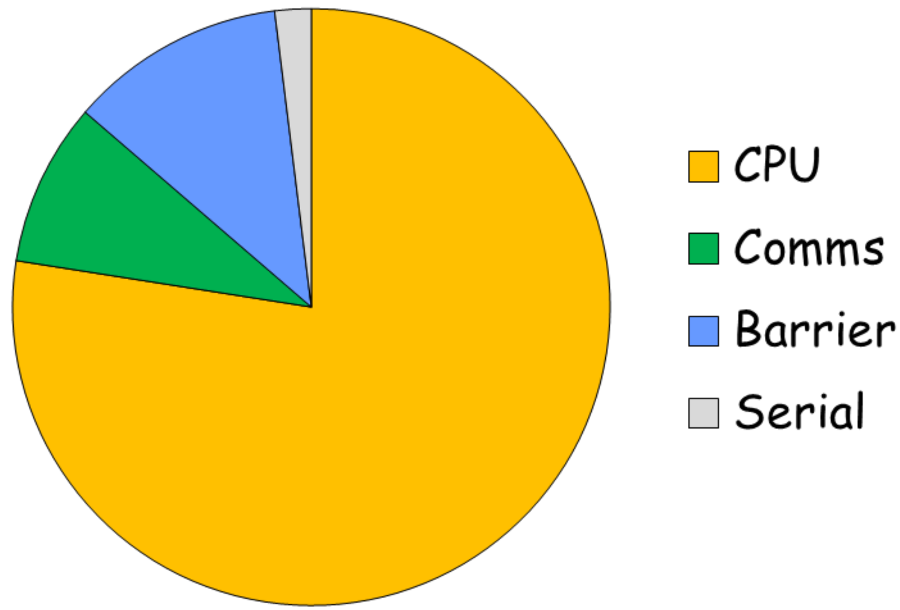
IFS 10-day forecasts: Comparison

T1279-L91 on 48 Power7 Nodes



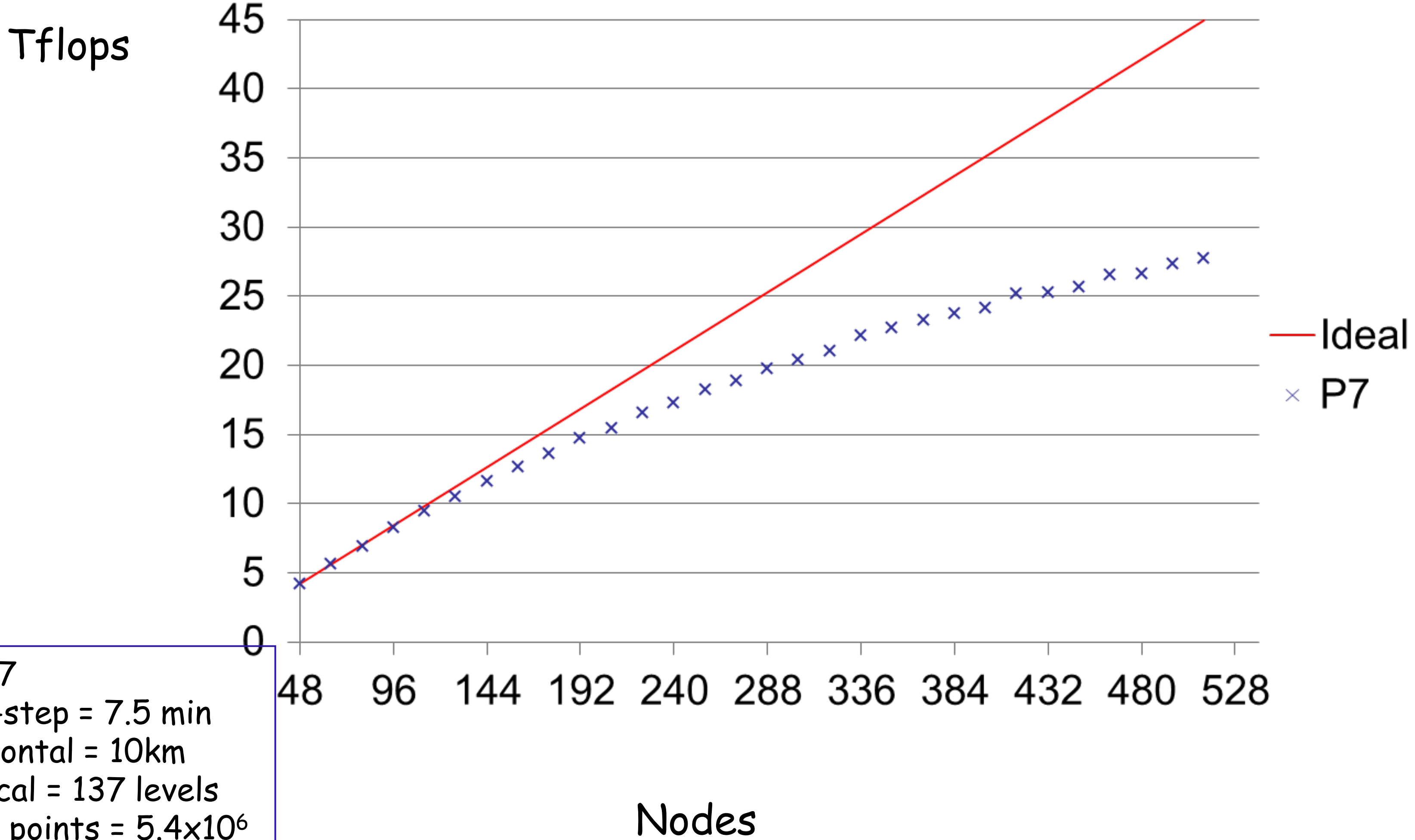
384 MPI tasks * 8 OMP threads
TOTALS: 6.8 Pflop
652 GB of Memory
93 TB of Comms
3.4 Tflops (2007 secs)

T2047-L137 on 144 Power7 Nodes



1152 MPI tasks * 8 OMP threads
TOTALS: 39.6 Pflop
3326 GB of Memory
455 TB of Comms
11.9 Tflops (3326 secs)

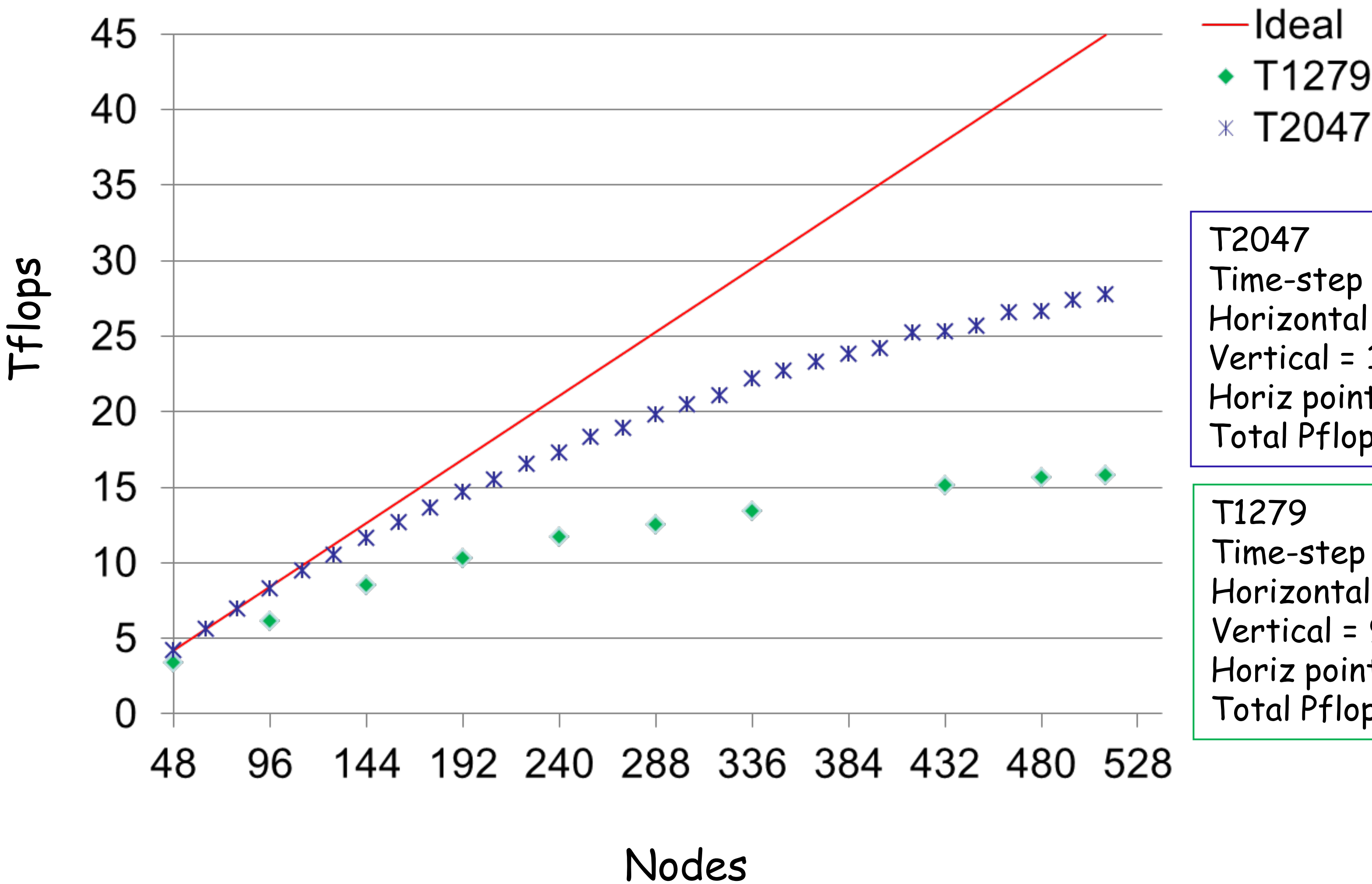
T2047 10-day Forecast runs up to 30k threads



T2047
Time-step = 7.5 min
Horizontal = 10km
Vertical = 137 levels
Horiz points = 5.4×10^6
Total Pflops = 39.6

ECMWF

T1279 compared with T2047



T2047
Time-step = 7.5 min
Horizontal = 10km
Vertical = 137 levels
Horiz points = 5.4×10^6
Total Pflops = 39.6

T1279
Time-step = 10 min
Horizontal = 16km
Vertical = 91 levels
Horiz points = 2.1×10^6
Total Pflops = 6.8

IFS model: current and planned resolutions

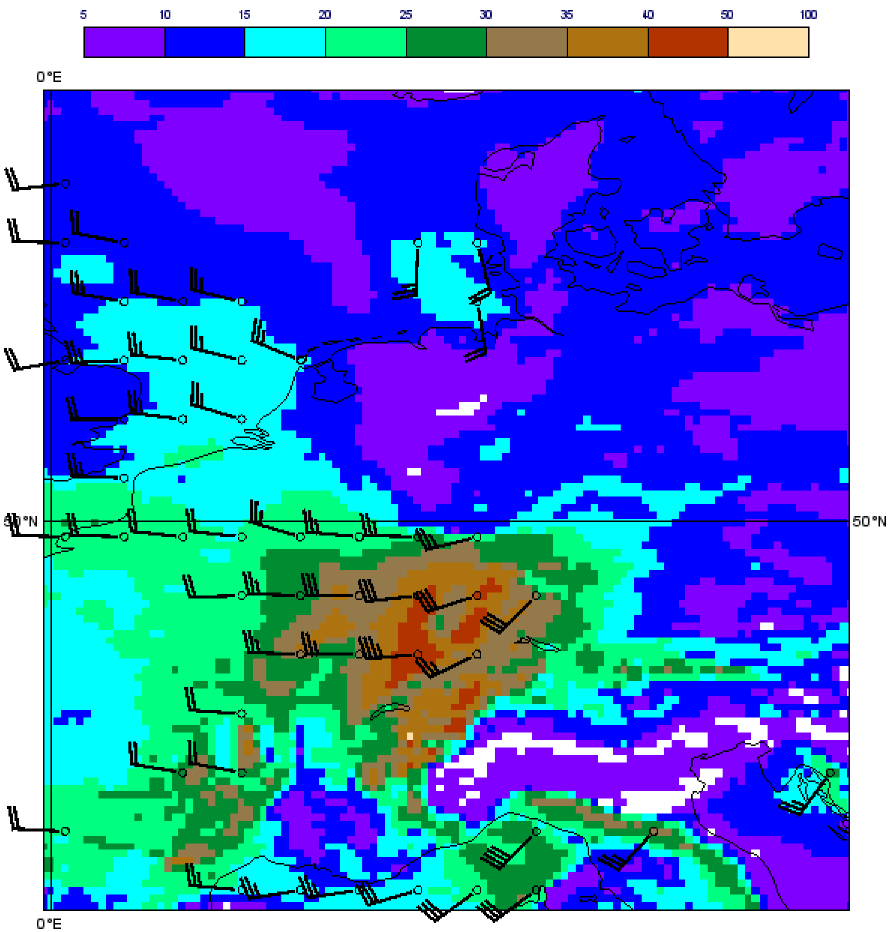
IFS model resolution	Envisaged Operational Implementation	Grid point spacing (km)	Time-step (seconds)	Estimated number of cores*
T1279 H	2010 (L91) 2012 (L137)	16	600	1100 1600
T2047 H	2014-2015	10	450	6K
T3999 NH	2020-2021	5	240	80K
T7999 NH	2025-2026	2.5	30-120	1-4M

***Rough estimate for the number of ‘Power7’ equivalent cores needed to achieve a 10 day model forecast in under 1 hour (~240 FD/D), system size would normally be 10 times this number.**

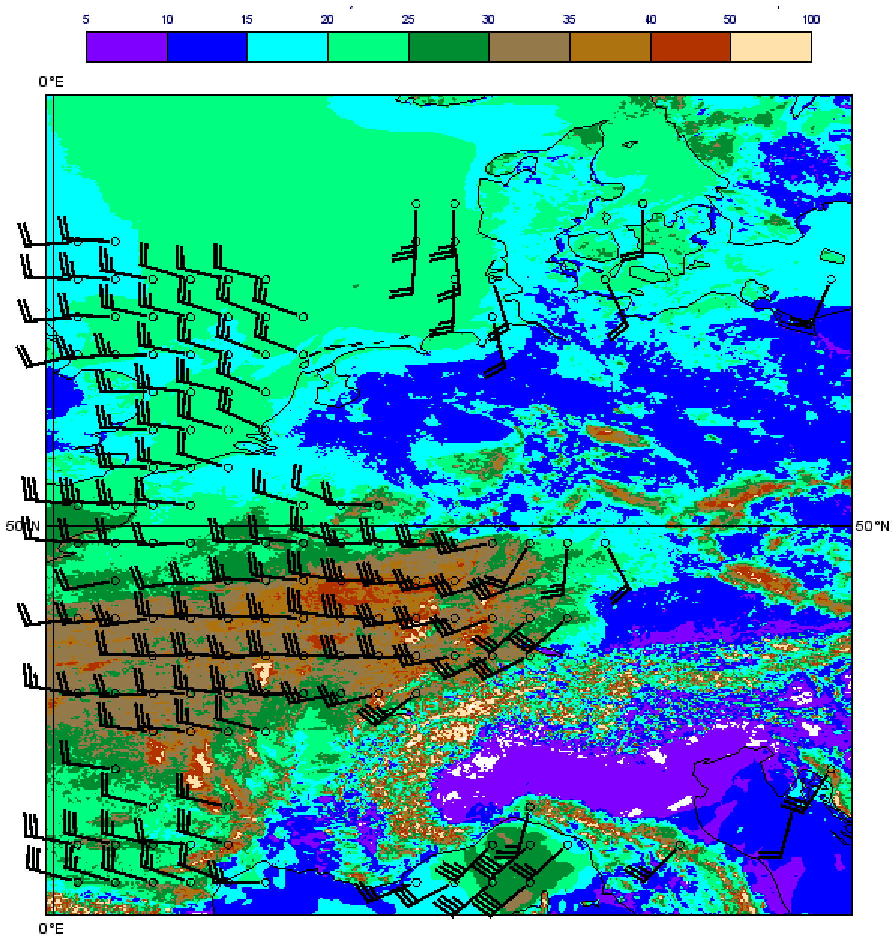
H = Hydrostatic Dynamics
NH = Non-Hydrostatic Dynamics

Slides from George Mozdzynski / Nils Wedi

T7999 results: Lothar "Christmas storm"



T1279 (~16km)



T7999 (~2.5km)

ECMWF

Conclusions

- IFS performs well on Power7 with speed-up of ~1.4 from Power6 for T1279 10-day forecast on 48 nodes.

- Jitter effects have been much reduced on Power7 compared with Power6

- Availability of new machine allowed first ever T7999 (2.5km) runs to be attempted