IFS performance on the new IBM Power6 systems at ECMWF

Deborah Salmond

Plan for Talk

- HPC systems at ECMWF
  - 2 x IBM p5 575+ clusters (Power5+)
  - New IBM p6 575 cluster (Power6)

- Preliminary performance measurements for IFS Cycle 35r1 - ECMWF's operational weather forecasting model - on Power6 compared with Power5+

- Power6 system is available to all ECMWF internal users for research experiments

Power5+ → Power6

hpce & hpcf	c1a & c1b
IBM p5 575+	IBM p6 575
Power5+ 1.9 GHz + SMT Peak 7.6 Gflops per core	Power6 4.7 GHz + SMT Peak 18.8 Gflops per core
2480 cores per cluster	7936 cores per cluster
16 cores per node	32 cores per node
Federation Switch	QLogic Silverstorm InfiniBand Switch

Power5+ → Power6

	Power5+	Power6	Increase
Clock	1.9 GHz	4.7 GHz	2.5 x
Cores per cluster	2480	7936	3.2 x
Compute nodes per cluster	155	248	1.6 x
Cores per node	16 (32 SMT)	32 (64 SMT)	2 x
Memory per node	32 Gbytes	64 Gbytes	1 x (per core)
L2 cache per core	0.9 Mbytes	4 Mbytes	4 x

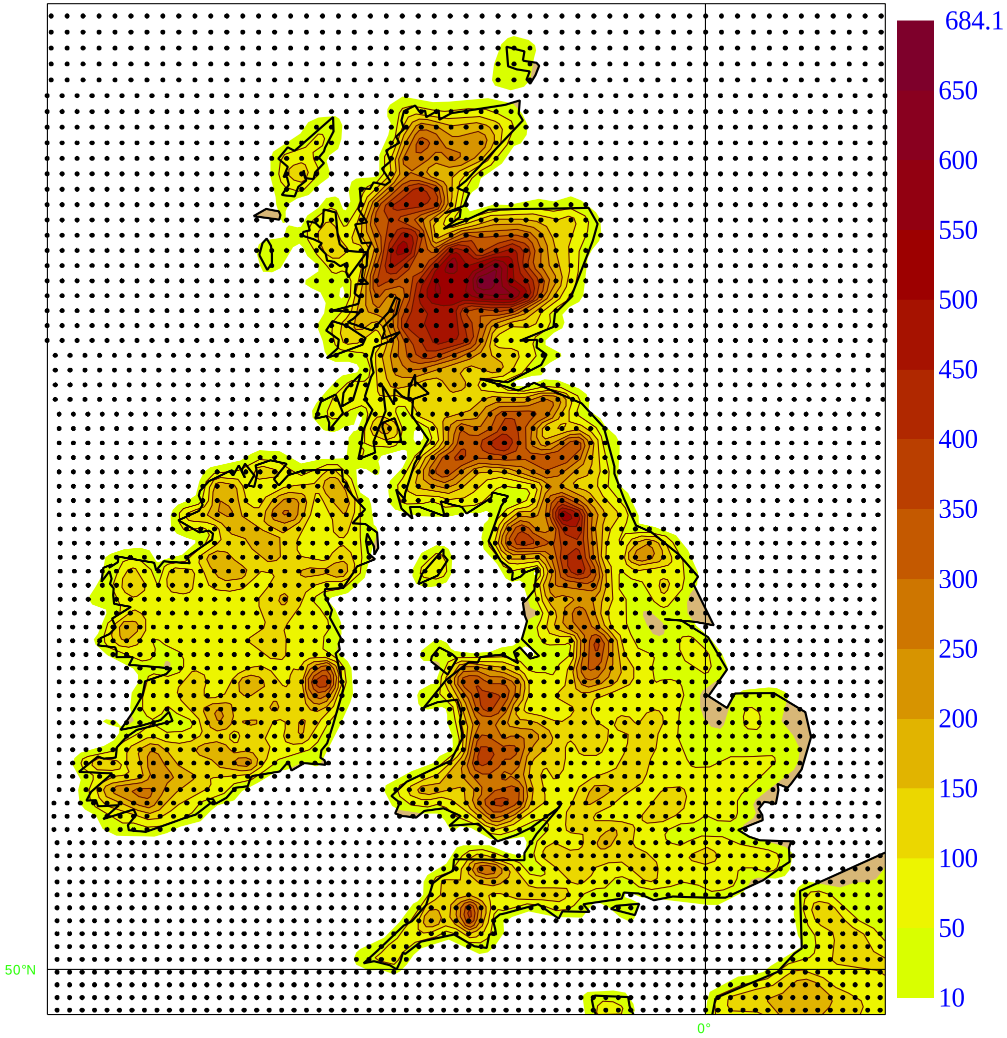
ECMWF's next operational resolution: T1279 L91

Horizontal grid-spacing = ~16km

Number of Horizontal gridpoints = 2,140,704

Timestep = 450 secs

Flops for 10-day forecast = $8*10^{15}$



ECMWF

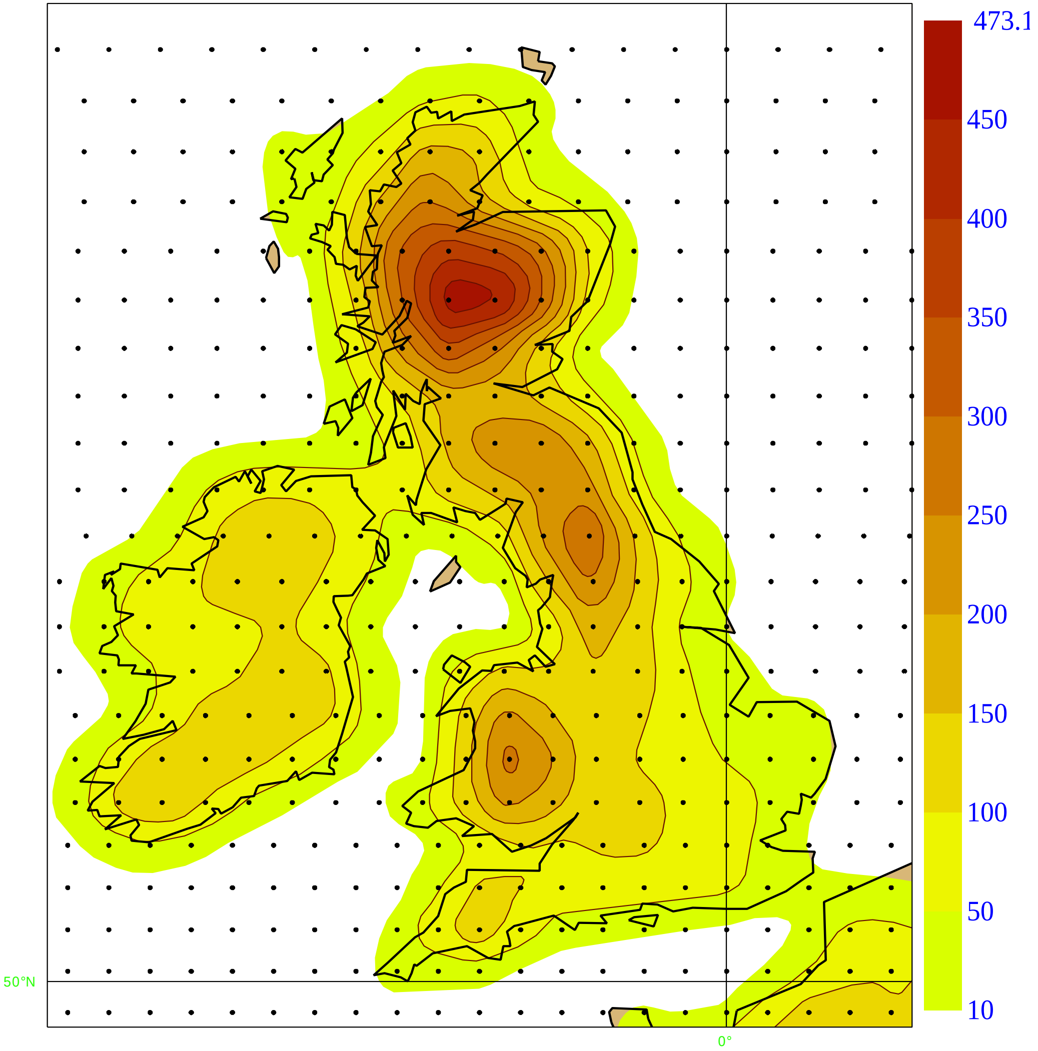
ECMWF's next operational resolution for EPS: T399

Horizontal grid-spacing = ~50km

Number of Horizontal gridpoints = 213,988

Timestep = 1800 secs

Flops for 51 member EPS = $5*10^{15}$



IFS T1279 L91 forecast (35r1) on P6 compared with P5+ – same number of cores

	Wall time	Number of cores	Tflops	% of Peak	Speed-up
P5+	10332	640* (40 nodes)	0.77	15.9	1
P6	6610	640 (20 nodes)	1.21	9.9	1.56

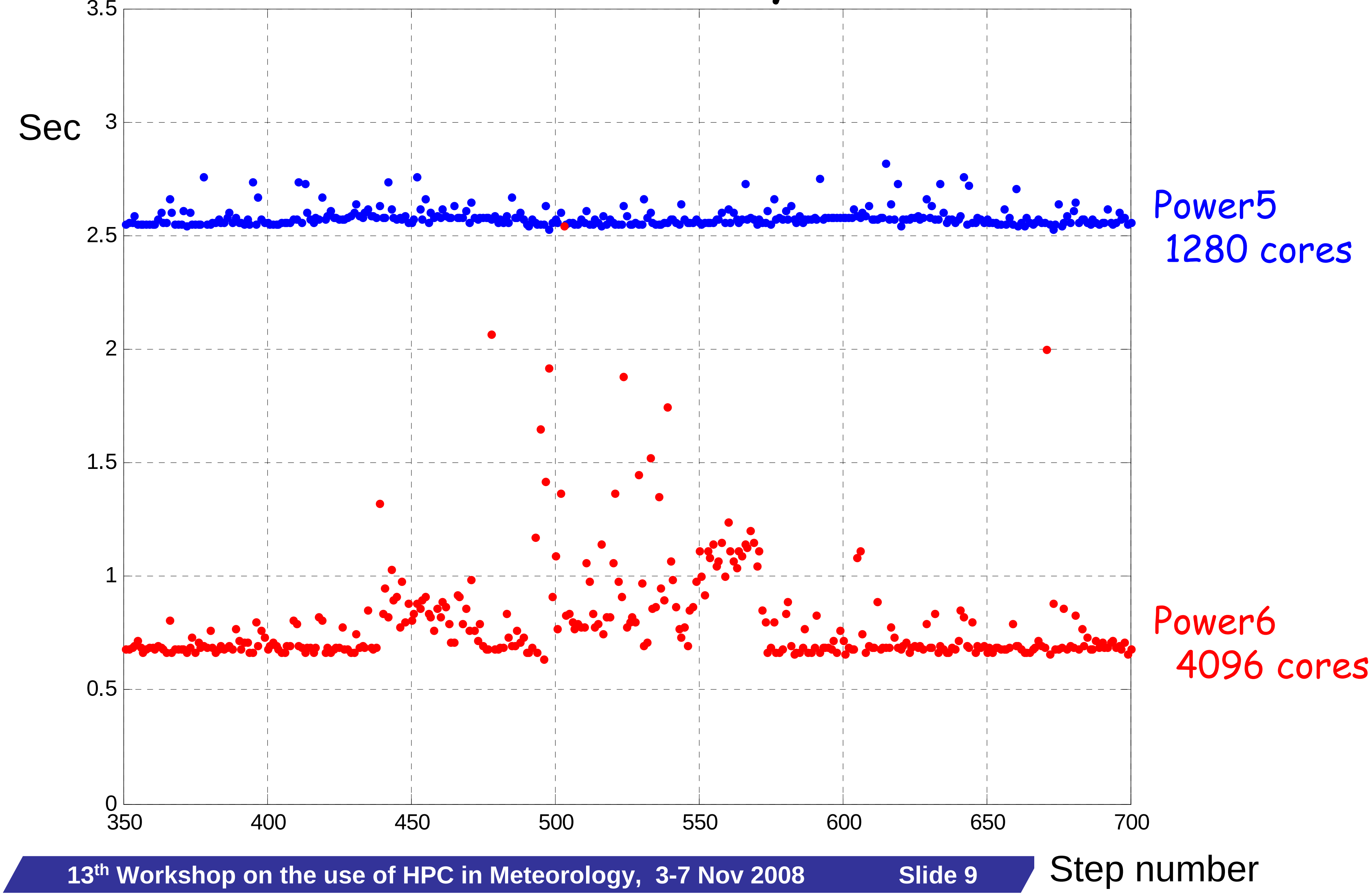
* 160 MPI tasks and 8 OpenMP threads – using SMT

IFS T1279 L91 forecast (35r1) on P6 compared with P5+ - 3.2 * number of cores

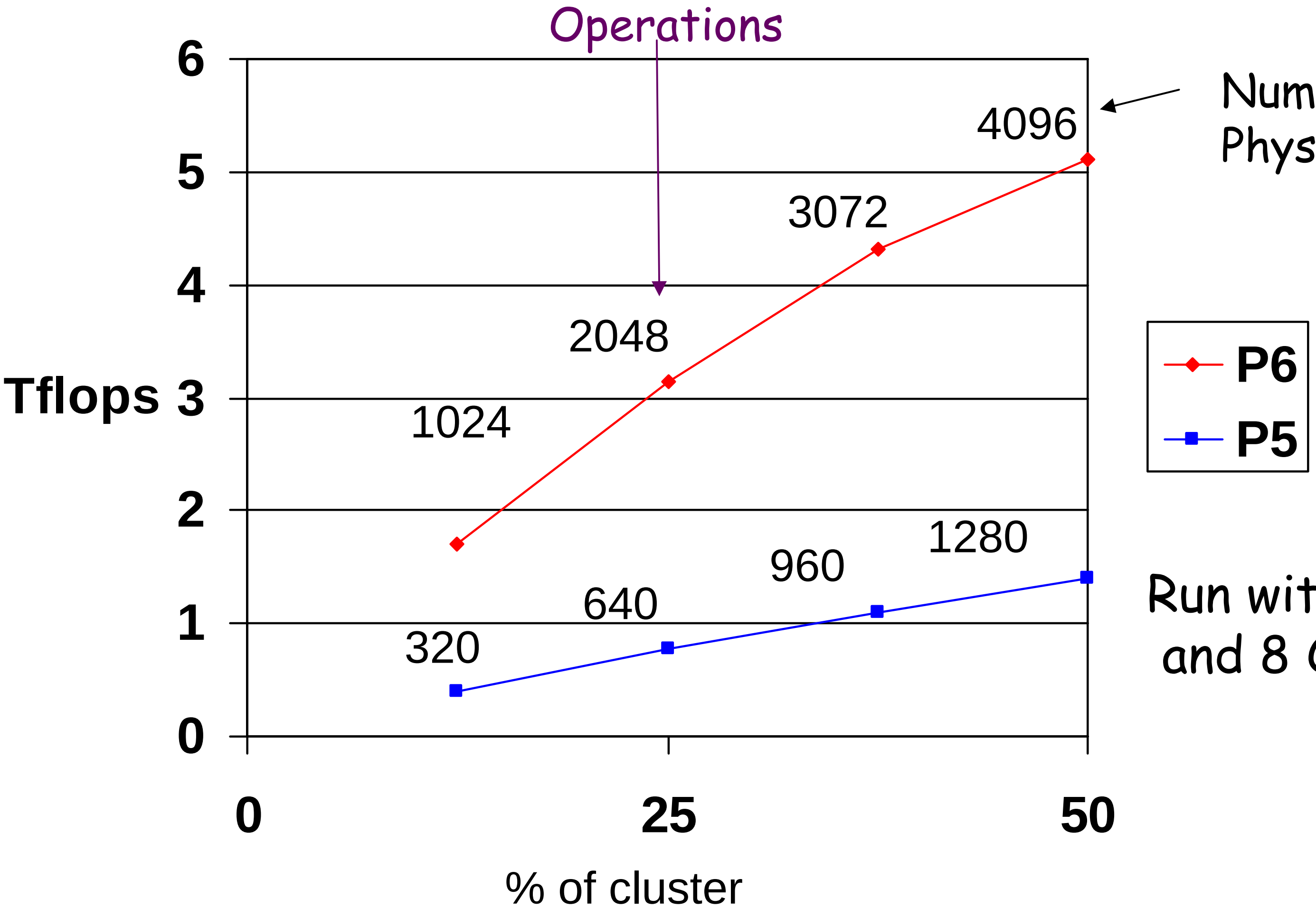
	Wall time	Number of cores	Tflops	% of Peak	Speed-up
P5+	10332	640 (40 nodes)	0.77	15.9	1
P6	2541*	2048 (64 nodes)	3.15	8.2	4.06

*Some performance problems
E.g. load imbalance from ‘jitter’

Variable time-step times
- now mostly fixed



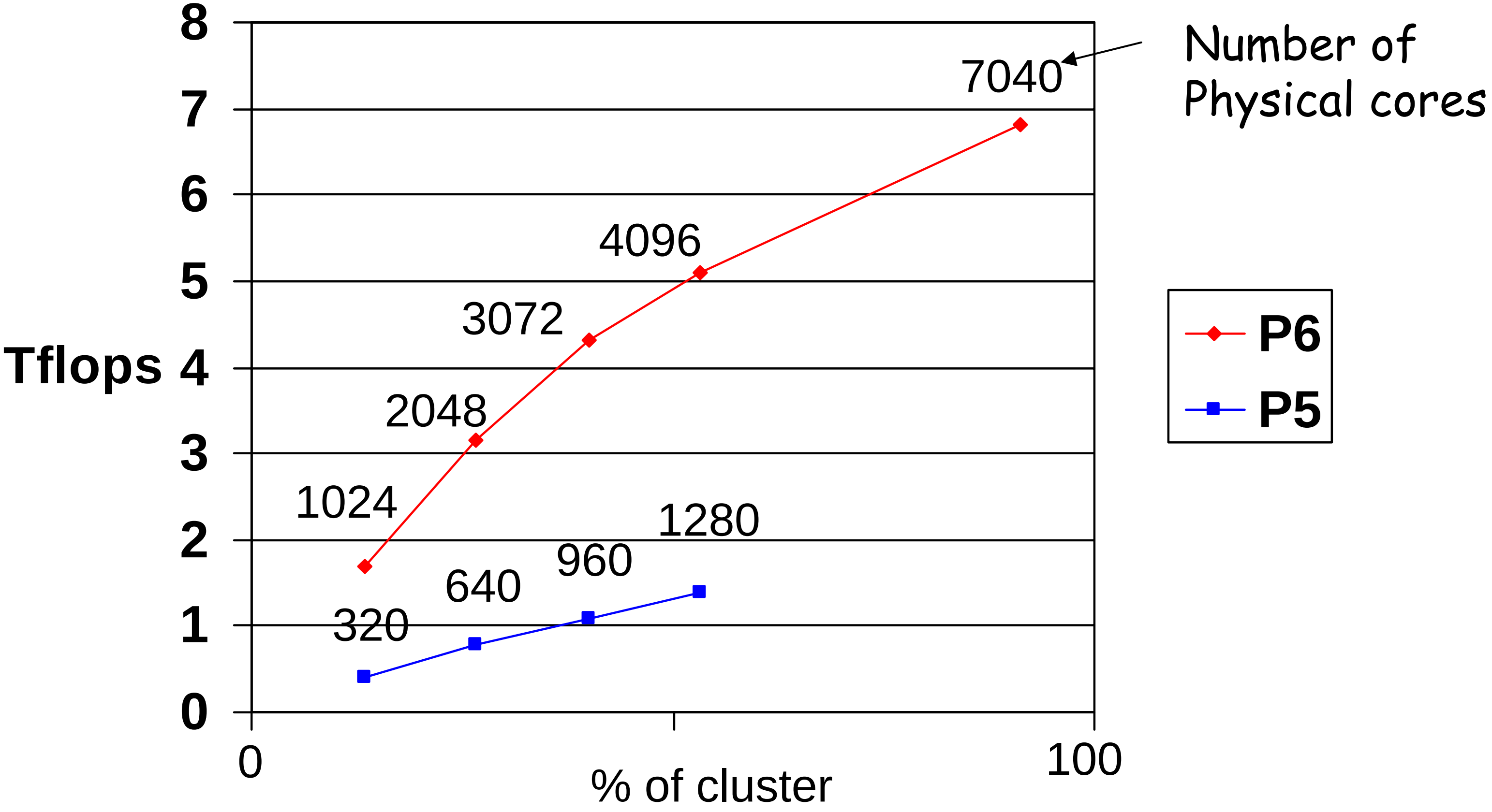
IFS T1279 10-day forecast on Power6



Number of Physical cores

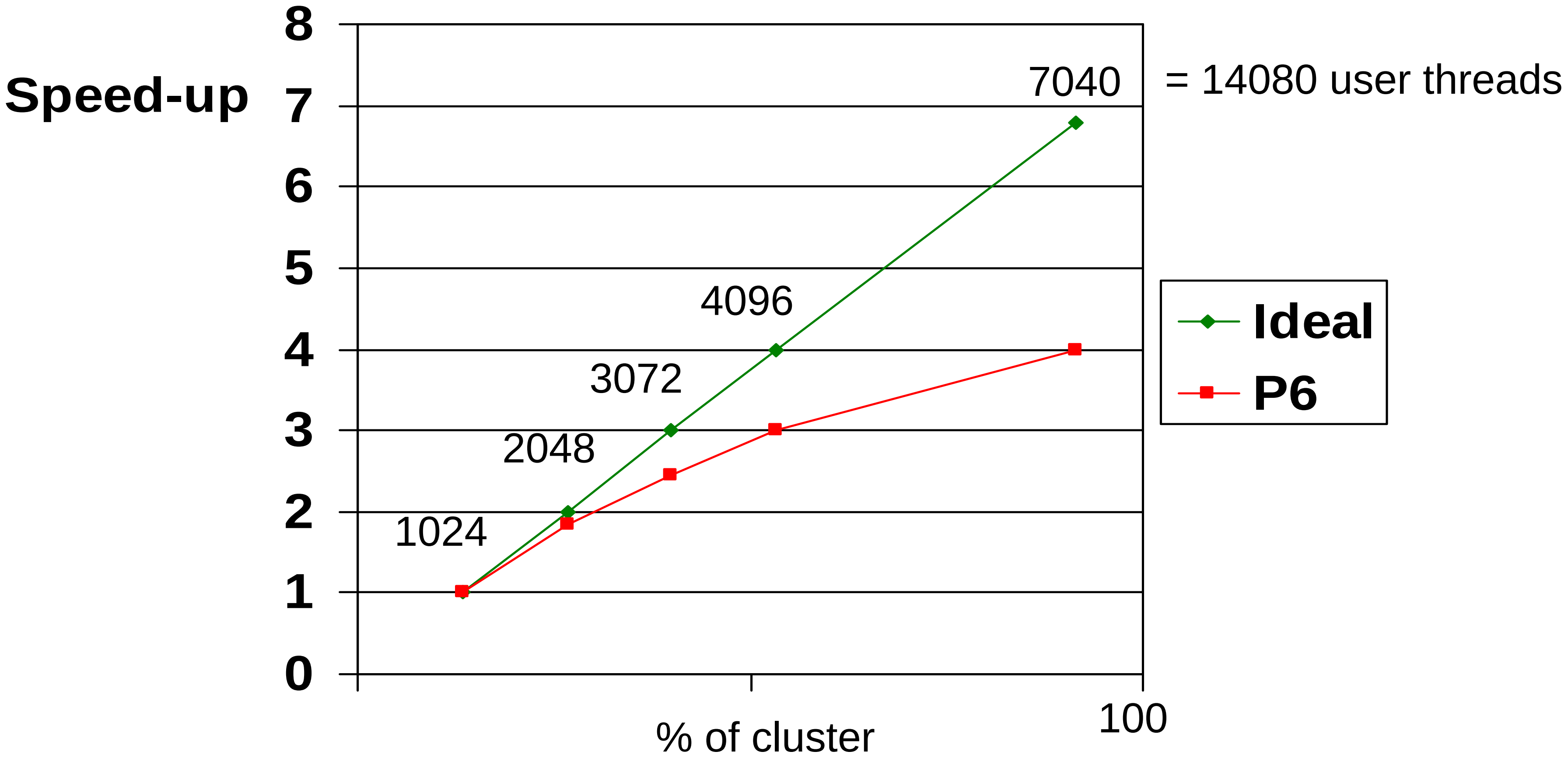
Run with SMT and 8 OpenMP threads

IFS T1279 10-day forecast on Power6 - scalability to whole cluster



Number of Physical cores

IFS T1279 10-day forecast on Power6 - speed-up curve



DrHook on Power5+ and Power6 for T1279 forecast

P5

% Tot	Ave	Min	Max	Ave	Min	Max	
	TIME			MFLOPS/Logical core			
6.64	85.6	73.7	97.7	645.4	518.0	733.0	CUADJTQ
6.72	86.6	80.5	94.6	362.3	352.0	374.0	CLOUDSC
6.49	83.6	76.8	87.6	2606.6	2342.0	2941.0	MXMAOP
2.00	25.7	7.4	50.8	197.5	71.0	232.0	CLOUDVAR
1.89	24.3	2.5	33.3	0.0	0.0	0.0	>MPL-TRLTOG_COMMS
2.28	29.3	26.5	32.9	290.6	266.0	311.0	VDFEXCU
2.33	30.0	28.1	32.0	2318.5	2268.0	2365.0	VERINT
1.96	25.2	21.6	32.0	989.4	852.0	1113.0	LAITQM
2.32	29.9	27.8	31.2	473.1	457.0	506.0	VDFMAIN
2.25	29.0	27.9	30.0	278.6	267.0	292.0	SRTM_SPCVRT_MCICA
2.22	28.5	27.2	29.4	0.0	0.0	0.0	>MPL-TRMTOL_COMMS

P6

% Tot	Ave	Min	Max	Ave	Min	Max	
	TIME			MFLOPS/Logical core			
7.23	54.7	51.4	59.4	573.5	551.2	595.6	CLOUDSC
5.42	41.0	35.1	44.6	5314.9	5124.3	5776.4	MXMAOP
3.95	29.9	26.4	33.5	1847.7	1446.0	2137.7	CUADJTQ
1.91	14.5	3.6	29.3	350.0	145.9	402.2	CLOUDVAR
2.64	19.9	18.1	21.2	0.0	0.0	0.0	>MPL-TRMTOL_COMMS
2.34	17.7	17.2	18.6	456.4	433.0	470.9	SRTM_SPCVRT_MCICA
2.31	17.5	16.9	18.5	808.3	751.7	853.3	VDFMAIN
1.79	13.5	3.8	17.8	0.0	0.0	0.0	>MPL-TRLTOG_COMMS
2.25	17.0	16.7	17.5	1466.6	1101.9	2035.2	LAITQM

Speed-up per core - Computation Power5+ → Power6

- Compute speed-up
  - Clock cycle = 2.47 x

Routine	Description	Gflops/core on Power6	Speed-up P5 → P6
CUADJTQ	Math functions	3.6	1.9
CLOUDSC	IF tests	1.1	1.4
MXMAOP	DGEMM call	10.6	2.2
SRTM	IF tests	0.9	1.5
LAITQM	Indirect addressing	2.9	1.5

Speed-up per core - Communications Power5+ → Power6

- Communications speed-up
  - 8 IB links per node on Power6
  - 2 Federation links per node on Power5+
  - Increase in aggregate Bandwidth per core = 2 x

Routine	Description	Speed-up P5 → P6
TRLTOG	Transposition Fourier to Grid-Point	1.89
TRMTOL	Transposition Spectral to Fourier	1.52
SLCOMM1	Semi-Lagrangian Halo	1.44

Power5 compared with Power6

- Very similar for users ☺
- Re-compile –qarch=pwr6
- No 'out-of-order execution' on Power6
  - so SMT is more advantageous
- Some constructs relatively slower on Power6
  - Floating point compare & branch
  - Store followed by load on same address

HPCF at ECMWF 1978-2011

CPUs

Sustained Mflops

1 CPU

40 Mflops

1978 CRAY1 — Vector

1984 CRAY XMP-2
1986 CRAY XMP-4
1990 CRAY YMP-8
1992 CRAY C90-16

Vector - Shared Memory Parallel

1996 Fujitsu VPP700
2000 Fujitsu VPP5000

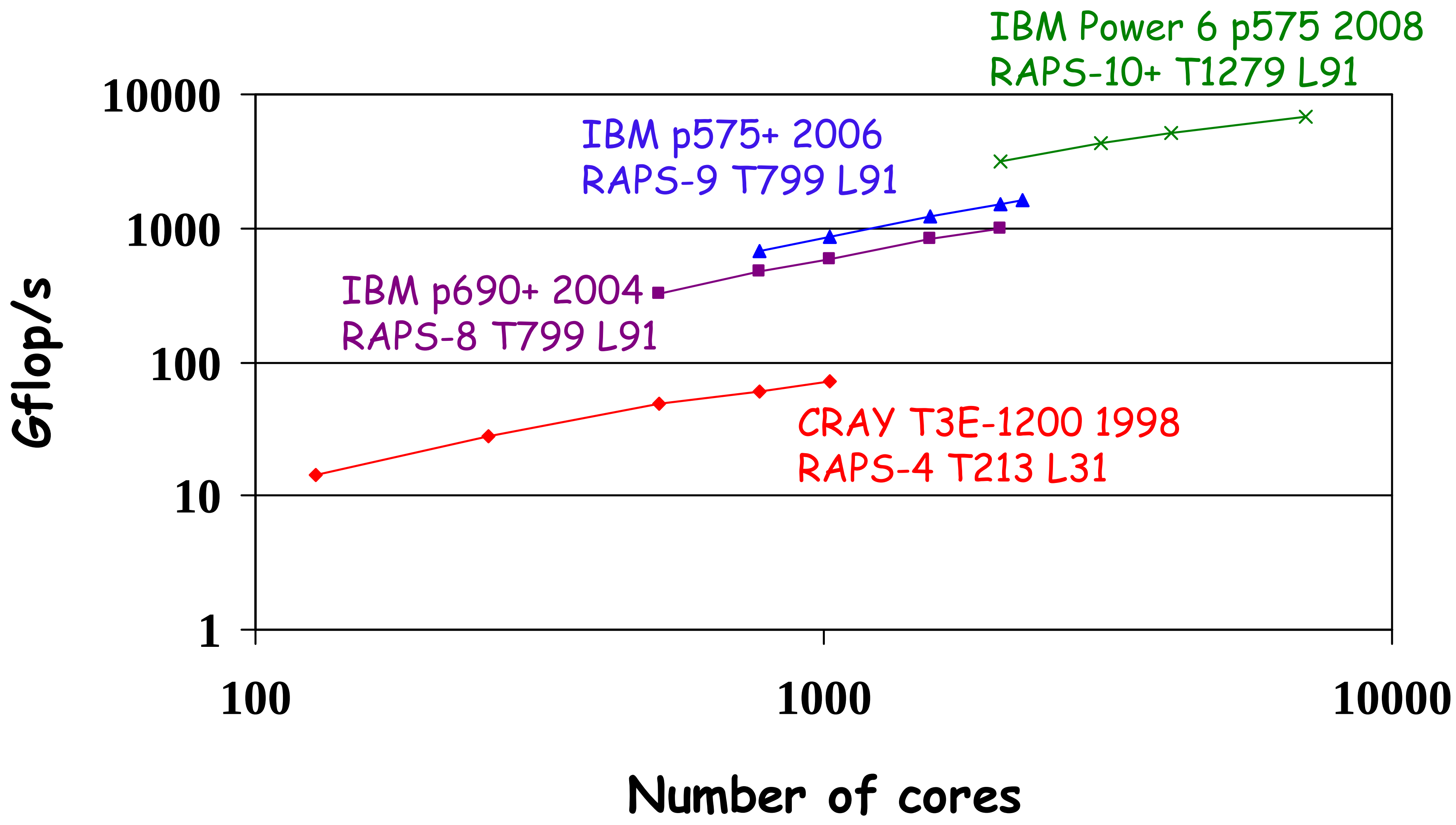
Vector - MPI

2002 IBM Power4
2004 IBM Power4+
2*2480 2006 IBM Power5+
2*7936 2008 IBM Power6

Scalar - MPI+OpenMP
-> SMT

4 Tflops
~20Tflops

History of IFS scalability


IBM Power 6 p575 2008
RAPS-10+ T1279 L91
IBM p575+ 2006
RAPS-9 T799 L91
IBM p690+ 2004
RAPS-8 T799 L91
CRAY T3E-1200 1998
RAPS-4 T213 L31
Gflop/s
Number of cores