Computational Efficiency of the ECMWF Forecasting System

Deborah Salmond

Research Department

Agenda

- Current HPC systems at ECMWF
- Operational configurations
- Acceptance tests for Phase 4 – hpce & hpcf
- Scalability with resolution T799 → T1279
- RAPS9 benchmarks on hpce & hpcf

Current HPC systems at ECMWF

- **2 IBM p575+ clusters – hpce & hpcf**
- **Dual-core 1.9GHz Power5+ processors**
  → 7.6 Gflop/s peak per core
- **140 Nodes per cluster**
- **16 PEs per shared memory Node (Note: PE = core)**
- **2240 PEs per cluster**
- **SMT → 4480 threads per cluster (2 threads per PE)**
- **32 Gbytes memory per node**

Phase3 → Phase4

hpcc & hpcd	hpce & hpcf
IBM p690+	IBM p575+
Power4++ 1.9 GHz Peak 7.6 Gflops per PE Sustained ~.5 Gflops per PE	Power5+ 1.9 GHz --> with SMT Peak 7.6 Gflops per PE Sustained ~1 Gflops per PE
2176 PEs per cluster	2240 PEs per cluster
32 PEs per node	16 PEs per node ---> 3*Mem BW per PE

Same Federation Switch

Current operational schedule on hpce

- **Run twice per day: 00Z and 12Z**
- **12Z run → Start 14:00**
- **4D-Var data assimilation — T799 / T95 / T255 L91**
  - **14:15 – 15:30 run on 24 Nodes – 75 mins (00dc)**
  - **16:15 – 16:55 run on 24 Nodes – 40 mins (12)**
- **10-day forecast — T799 L91**
  - **16:55 – 18:25 run on 24 Nodes – 90 mins**
- **EPS — 50*{T399 L62 10-day + T255 L62 6-day}**
  - **16:55 – 18:56 50*{ 3 nodes – 45 mins + 1 node – 20 mins }**
- **→ Finish 19:07**

Workload on hpce & hpcf

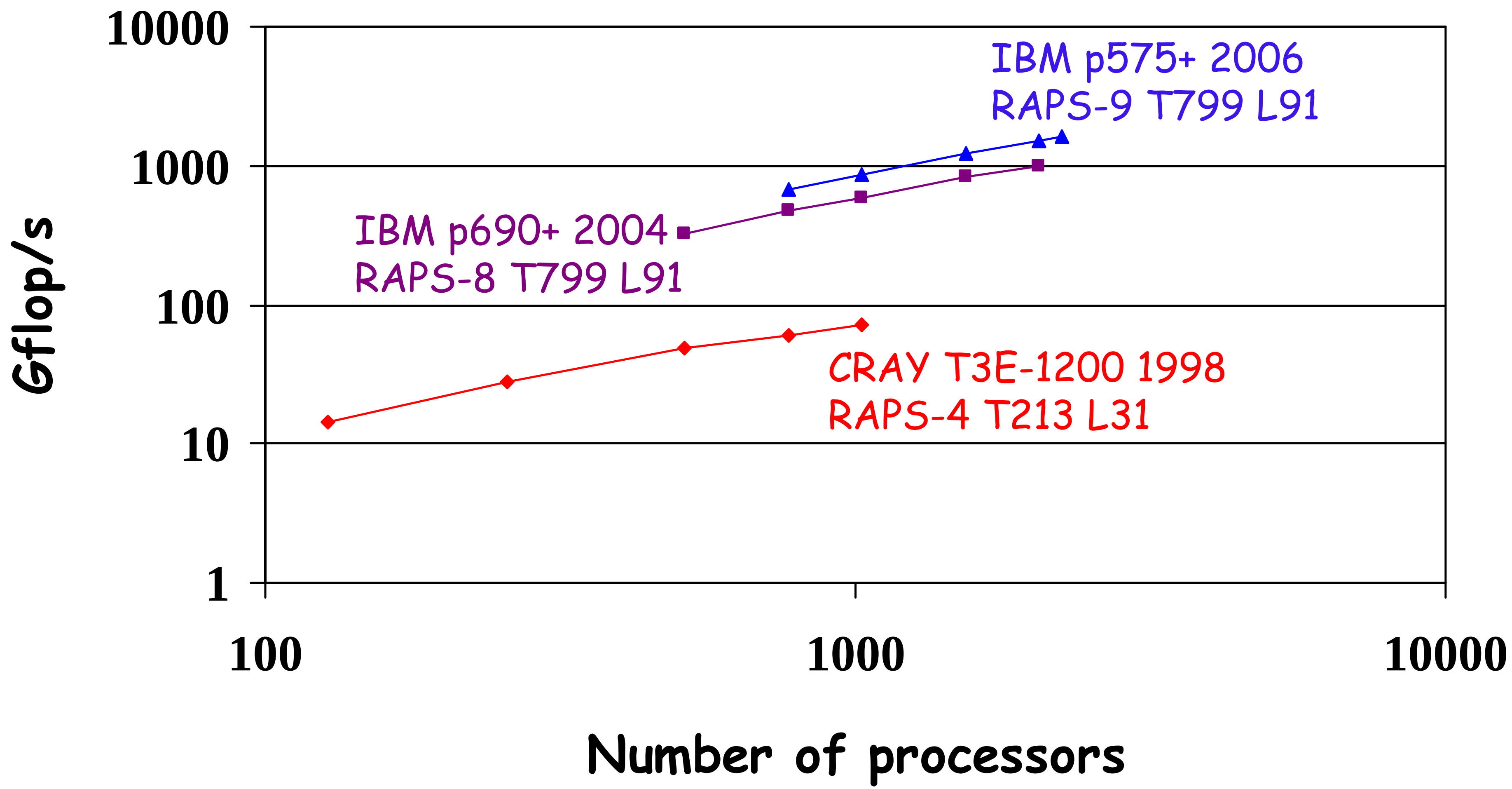
- Operations: 61% EPS : 22% 4D-Var : 17 % Forecast

- Research experiments – mostly 4D-Var

- Member States work = 25% of total time
  - – including ‘BC suite’

- 35000 jobs per cluster per day
  - – peak of about 50 jobs submitted per sec
  - – 12000 parallel jobs per cluster per day mostly using SMT

History of RAPS benchmark



Benchmark tests for Phase 4 (RAPS8)

- **4D-Var – 2 copies per cluster**

  Phase 4 = 1340 Seconds → Speed-up = 1.55 → 1.37 Tflop/s 8.1% peak

- **T799 10-day forecast – 2 copies per cluster**

  Phase 4 = 1471 Seconds → Speed-up = 1.72 → 2.43 Tflop/s 14.3% peak

- **T399 EPS forecasts – 47 copies on 141 nodes**

  Phase 4 = 1554 Seconds → Speed-up = 1.77 → 2.66 Tflop/s 15.6% peak

Notes: All times – first start to last finish

Speed-up – Phase 4 (hpce) compared with Phase 3 (hpcd)

Tflop/s – Aggregate Tflop/s per cluster & Peak is 17 Tflop/s

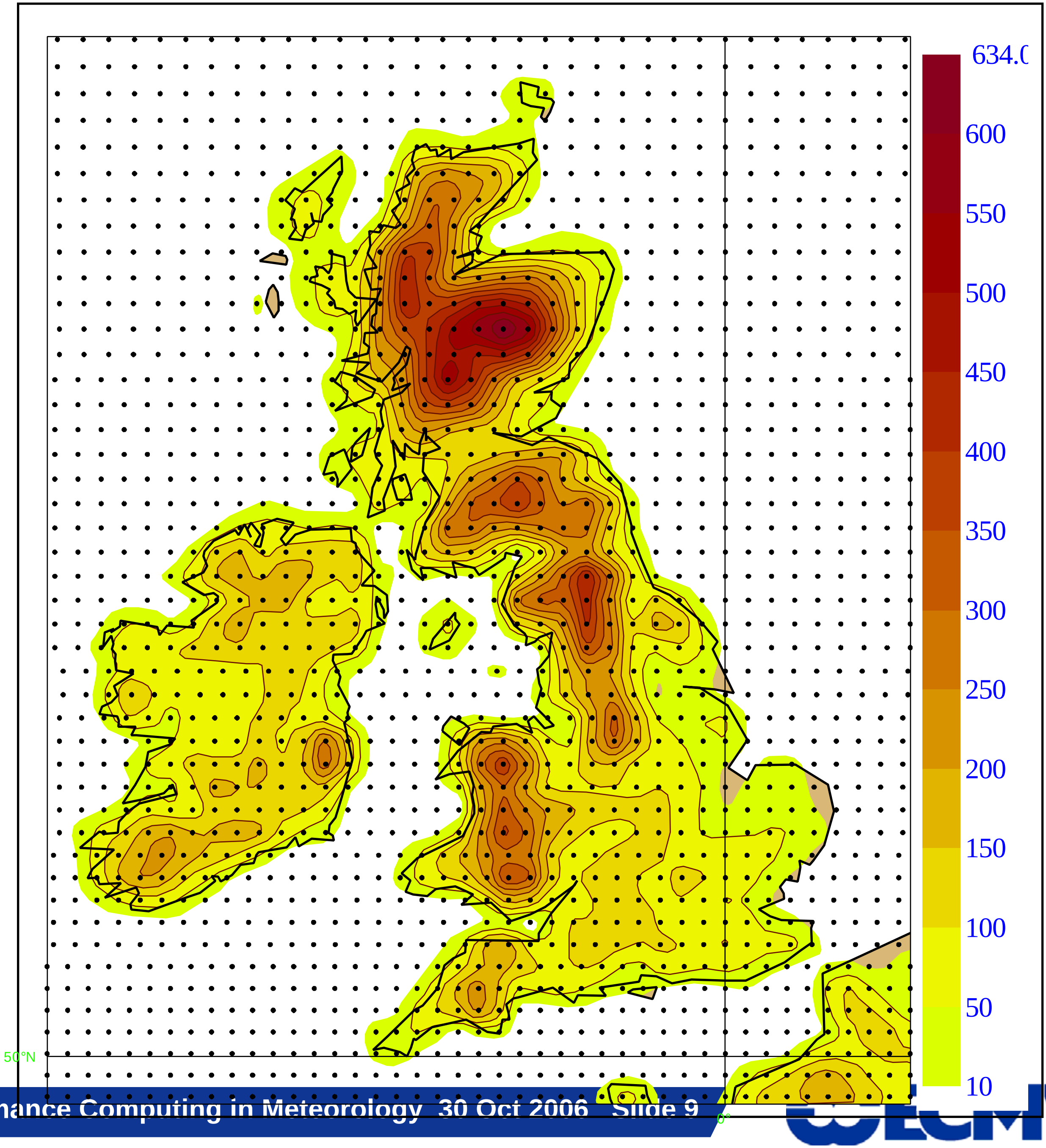
T799

25km

NGPTOT = 843,490

TSTEP = 720 secs

Flops for 10-day forecast = $1.615 * 10^{15}$



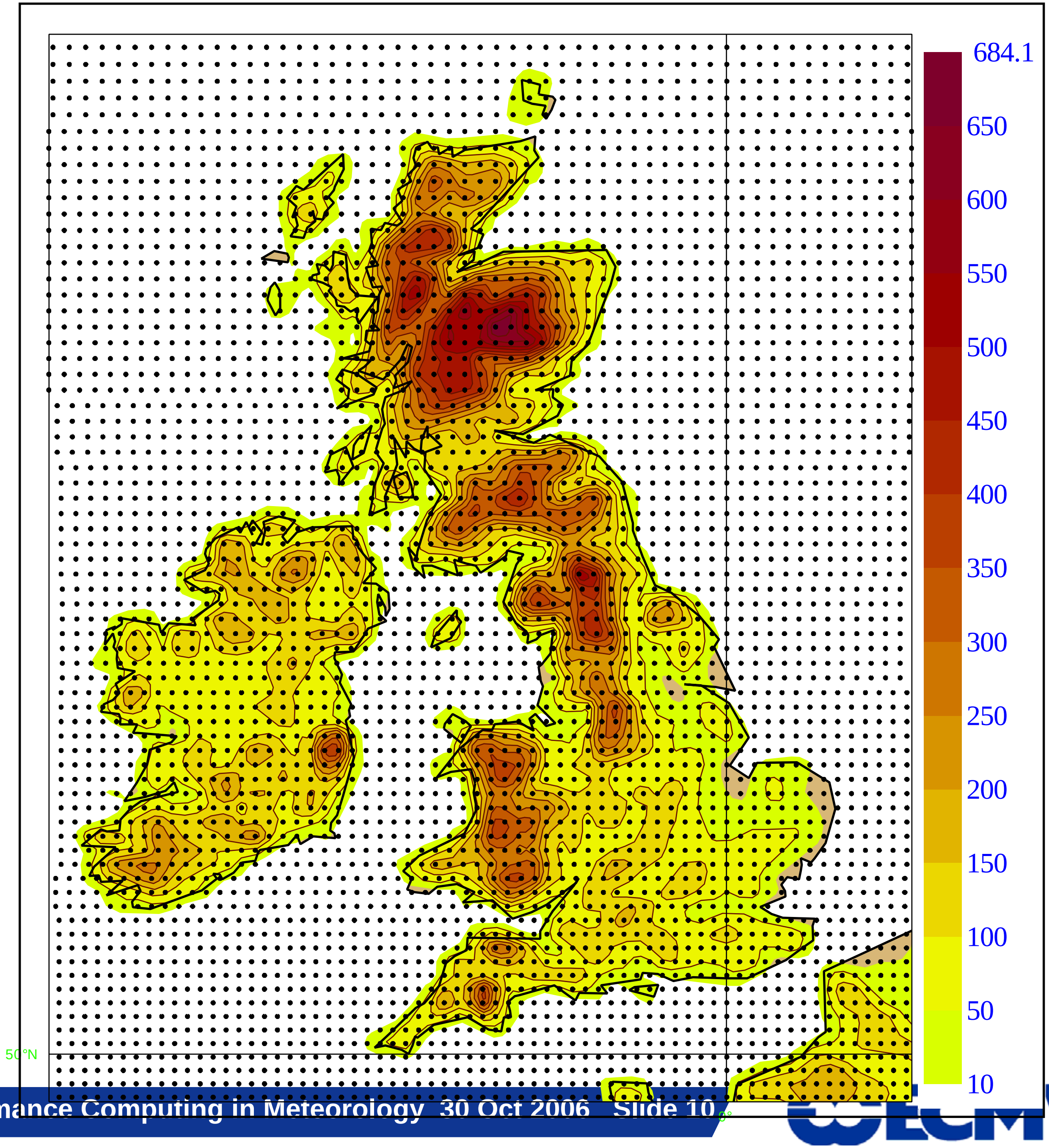
T1279

16km

NGPTOT = 2,140,704

TSTEP = 450 secs

Flops for 10-day forecast = $7.207*10^{15}$



ECMWF

T399

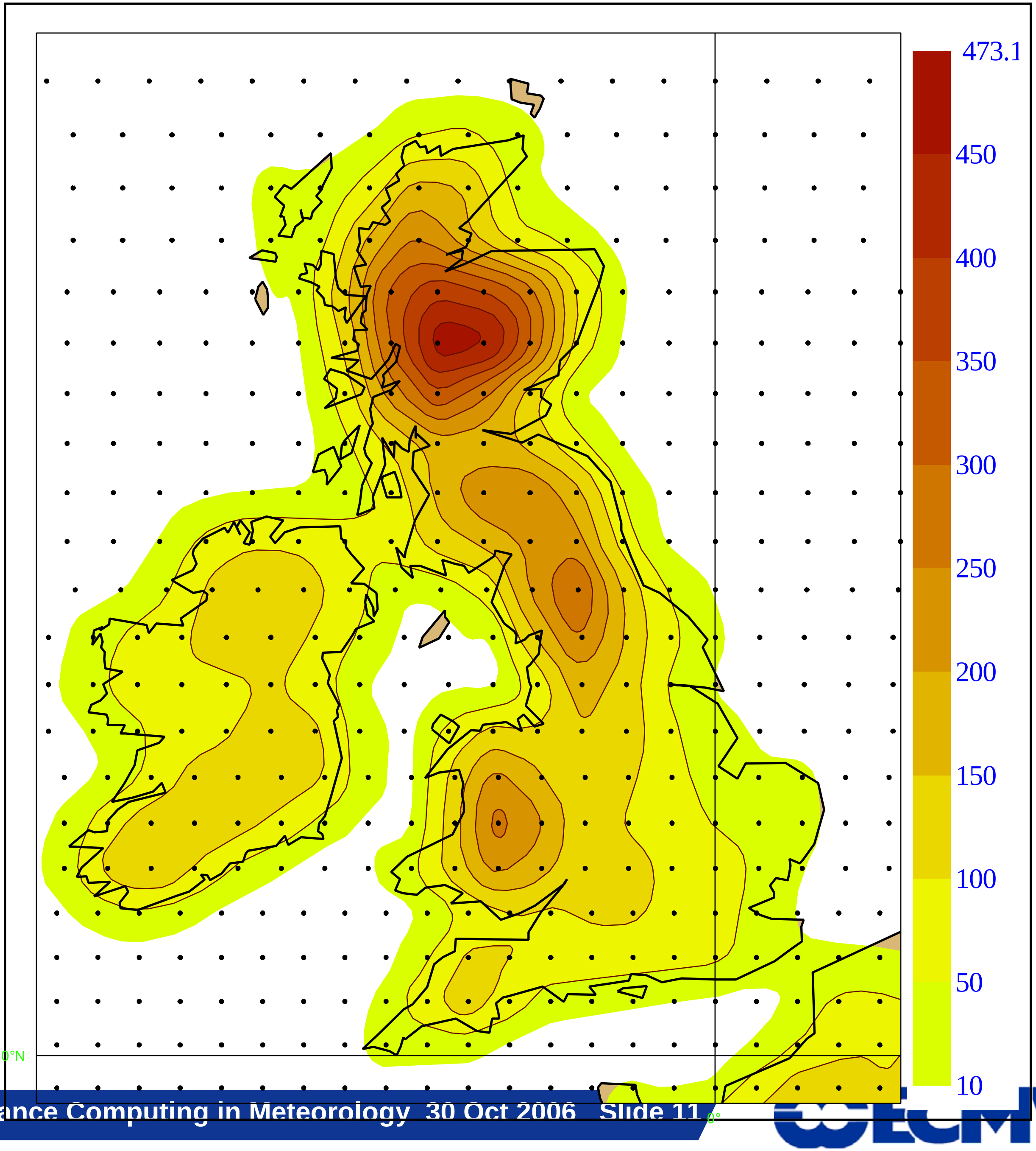
50km

NGPTOT = 213,988

TSTEP = 1800 secs

Flops for 10-day forecast = $0.1013*10^{15}$

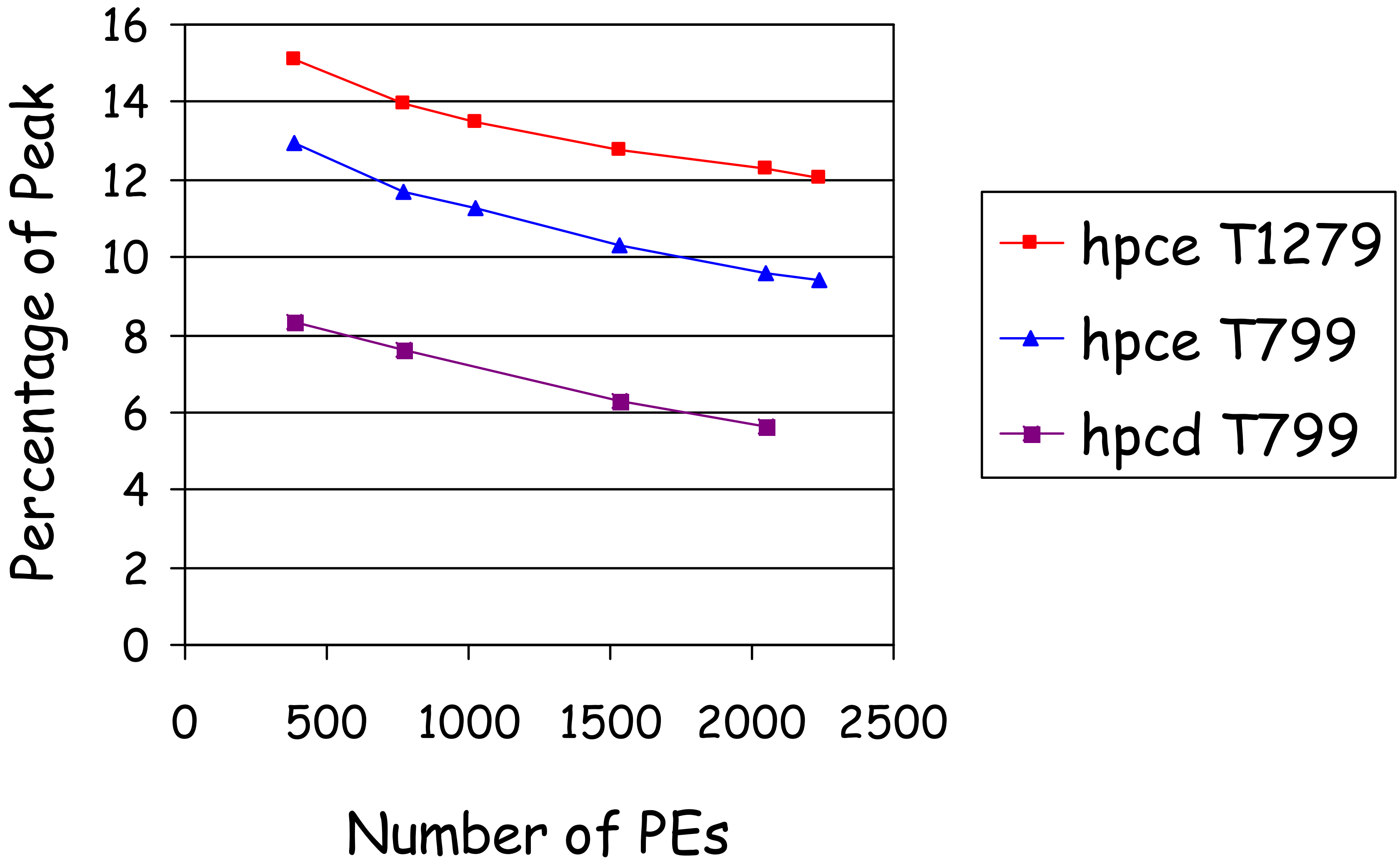
EPS=50*



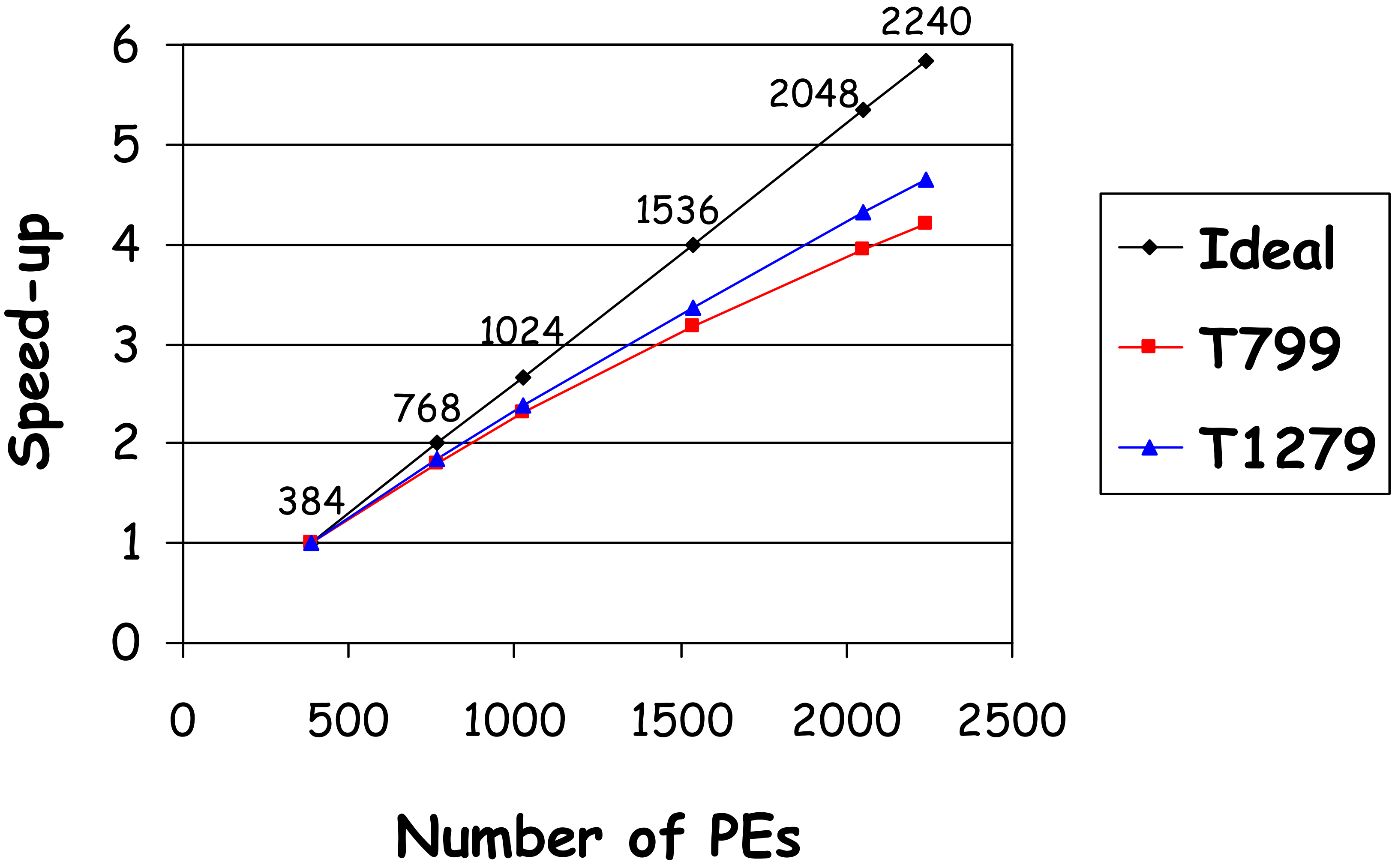
Comparison of Resolutions

Resolution	T1279 L91	T799 L91	T399 L62
Grid spacing	16km	25km	50km
Number of grid-points	2,140,704	843,490	213,988
Time-step	450 secs	720 secs	1800 secs
Flops for 10-day forecast	$7.207*10^{15}$	$1.615*10^{15}$	$0.1013*10^{15}$ → EPS * 50

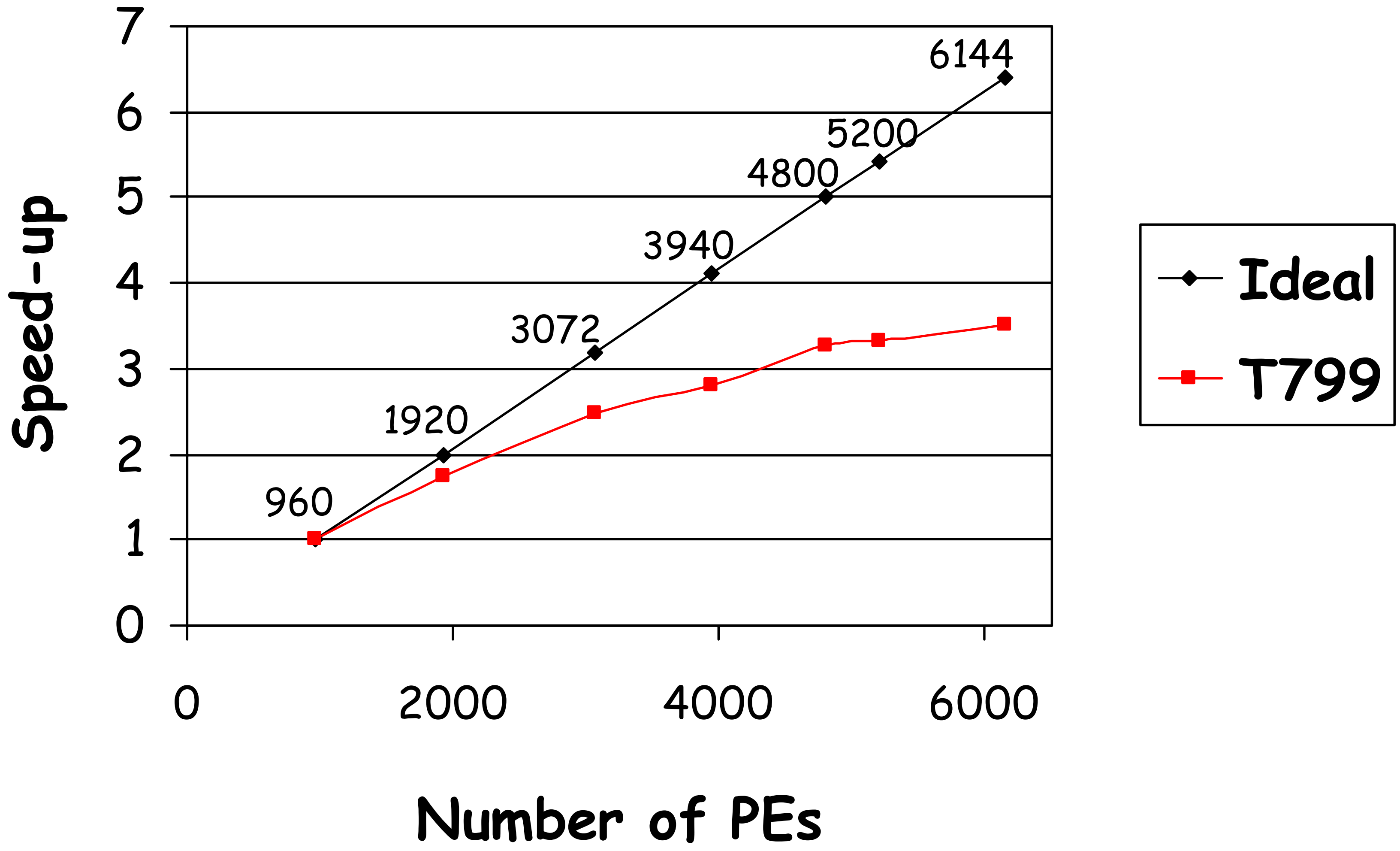
RAPS9 – 10-day T799 L91 Forecast



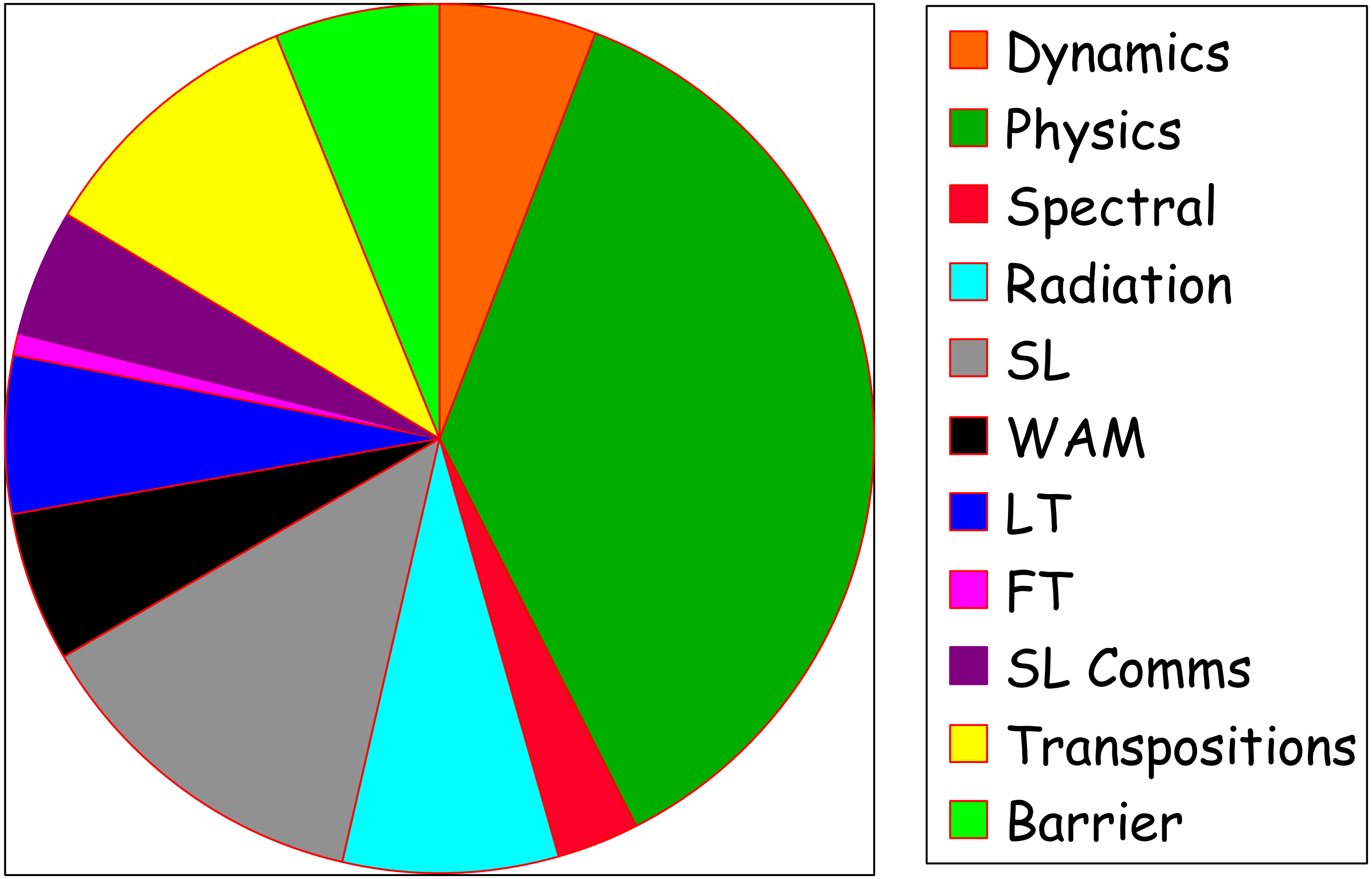
RAPS9 – T799 L91 10-day Forecast on hpce



RAPS9 – T799 L91 10-day Forecast on Cray XT3 at ORNL

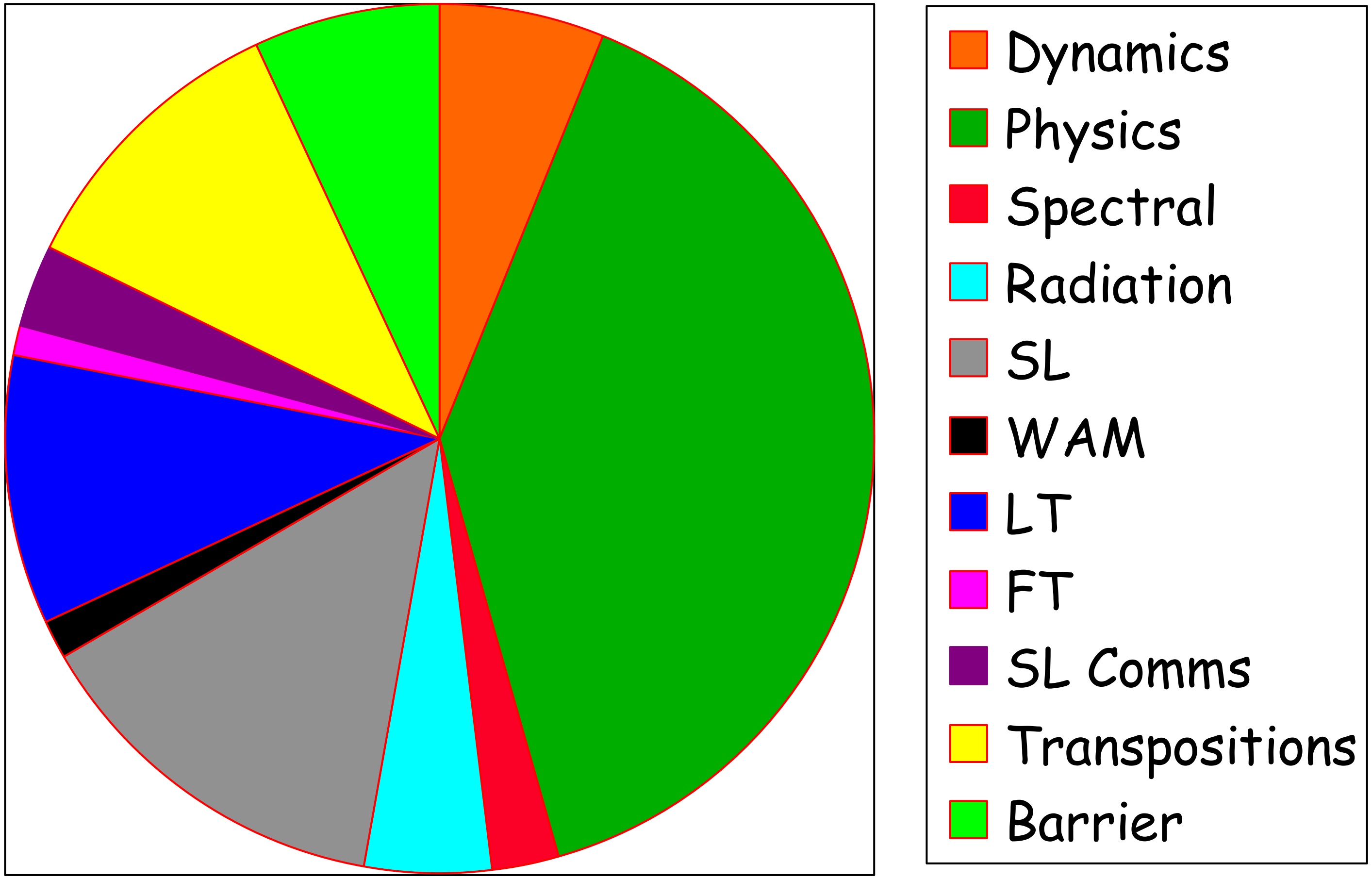


RAPS9 – T799 L91 10-day Forecast – 96 Nodes



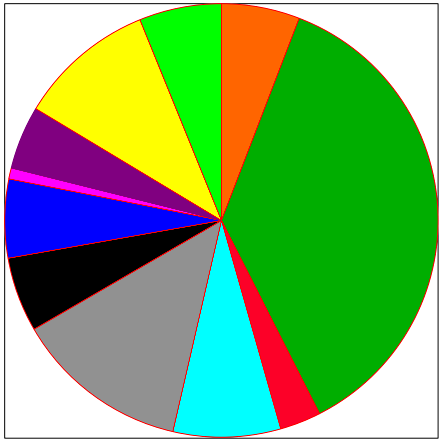
ECMWF

RAPS9 – T1279 L91 10-day Forecast – 96 nodes



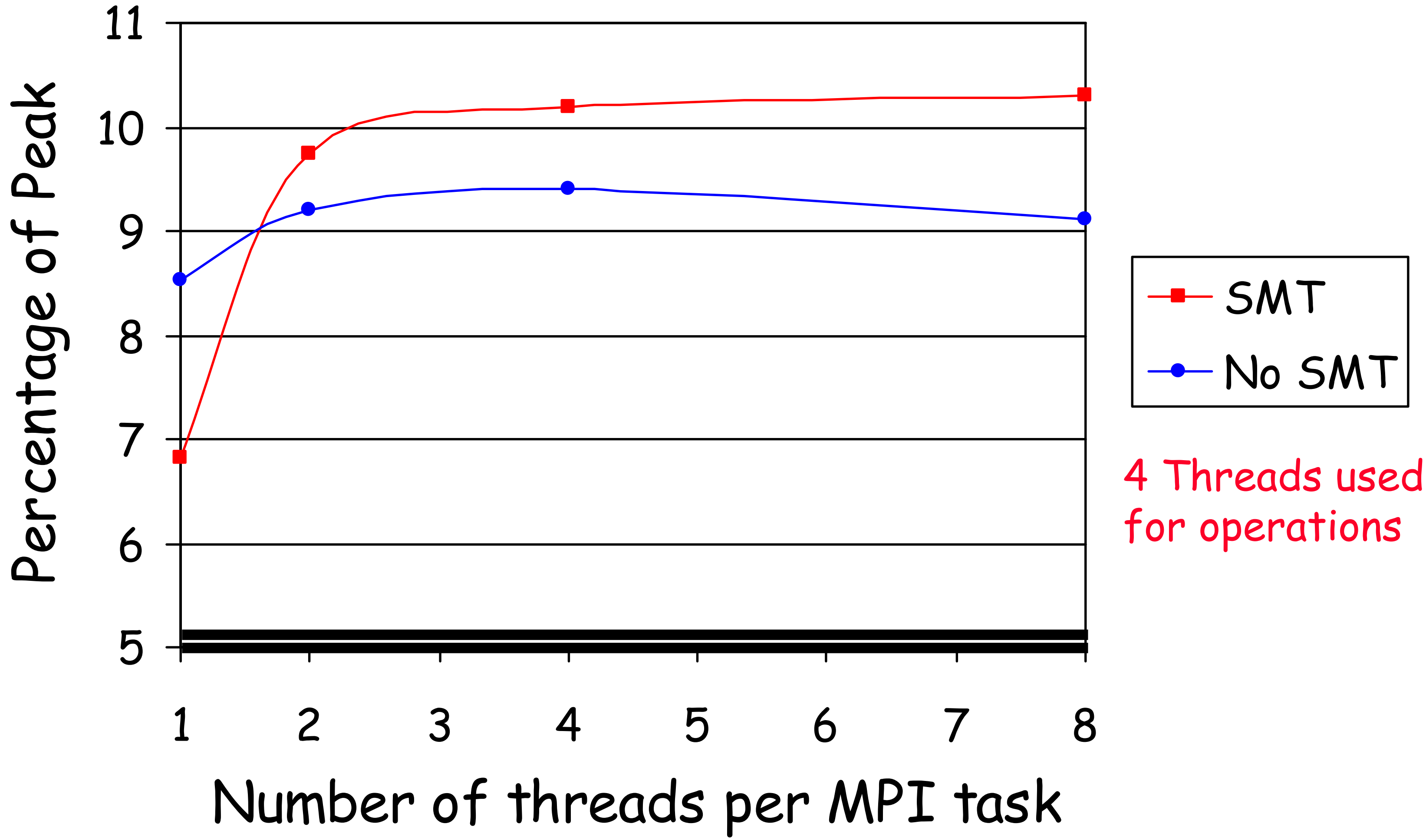
Mflop/s per Subroutine from Dr.Hook

→ T799 L91 forecast
run on 128 PEs



Subroutine	Mflop/s per PE hpcd	Mflop/s per PE hpce (no SMT)	Mflop/s per PE hpce (SMT)
CLOUDSC	533	559	758
MXMAOP	3267	5270	4700
LAITQM	1072	987	1350

RAPS9 - T799 L91 10-day forecast – OpenMP threads / MPI task on 96 Nodes



4 Threads used for operations

RAPS9 - 10-day forecasts
→ Message passing communications on hpce

Resolution	Nodes MPI x OMP	WALL (secs)	%Comms (barrier)	Tflop/s	% of peak
T799 L91	24 Nodes 96 x 8	4253	8.0%	0.38	13.0%
T1279 L91	96 Nodes 384 x 8	4836	11.5%	1.61	12.8%
T799 L91	140 Nodes 560 x 8	995	18.9%	1.60	9.4%
T1279 L91	140 Nodes 560 x 8	3506	13.8%	2.05	12.1%

RAPS9 - T799/T95/T255 L91 4D-Var run on hpce - 16 Nodes, 128 MPI tasks & 4 Threads

Step	Resolution	Wall (secs)	%Peak	Flops $*10^{15}$
Traj-0	T799	643	7.3%	0.091
Min-0	T95	422	4.3%	0.036
Traj-1	T799	509	8.9%	0.088
Min-1	T255	3070	12.1%	0.721
Traj-2	T799	688	6.6%	0.089
Total		5334	9.9%	1.025

Summary

- **Phase 4 is 1.5 - 2.0 times faster than Phase 3 for IFS**
- **SMT works well with MPI + OpenMP to give higher percentage of peak**
- **IFS scales well with resolution**
- **Benchmark should contain as much as possible of operational system**