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  
  \( \rightarrow 7.6 \text{ Gflop/s peak per core} \)
- 140 Nodes per cluster
- 16 PEs per shared memory Node (Note: PE = core)
- 2240 PEs per cluster
- SMT \( \rightarrow 4480 \text{ threads per cluster (2 threads per PE)} \)
- 32 Gbytes memory per node
<table>
<thead>
<tr>
<th>Phase3</th>
<th>Phase4</th>
</tr>
</thead>
<tbody>
<tr>
<td>hpcc &amp; hpcd</td>
<td>hpce &amp; hpcf</td>
</tr>
<tr>
<td>IBM p690+</td>
<td>IBM p575+</td>
</tr>
<tr>
<td>Power4++ 1.9 GHz</td>
<td>Power5+ 1.9 GHz -- with SMT</td>
</tr>
<tr>
<td>Peak 7.6 Gflops per PE</td>
<td>Peak 7.6 Gflops per PE</td>
</tr>
<tr>
<td>Sustained ~.5 Gflops per PE</td>
<td>Sustained ~1 Gflops per PE</td>
</tr>
<tr>
<td>2176 PEs per cluster</td>
<td>2240 PEs per cluster</td>
</tr>
<tr>
<td>32 PEs per node</td>
<td>16 PEs per node</td>
</tr>
<tr>
<td>--- &gt; 3*Mem BW per PE</td>
<td></td>
</tr>
<tr>
<td>Same Federation Switch</td>
<td></td>
</tr>
</tbody>
</table>
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

![Graph showing the history of RAPS benchmark](graph.png)

- IBM p690+ 2004
- RAPS-8 T799 L91
- IBM p575+ 2006
- RAPS-9 T799 L91
- CRAY T3E-1200 1998
- RAPS-4 T213 L31
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

25 km
NGPTOT = 843,490
TSTEP = 720 secs
Flops for 10-day forecast = 1.615 \times 10^{15}
16km
NGPTOT = 2,140,704
TSTEP = 450 secs
Flops for 10-day forecast = $7.207 \times 10^{15}$
T399

50km

NGPTOT = 213,988

TSTEP = 1800 secs

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

EPS=50*
## Comparison of Resolutions

<table>
<thead>
<tr>
<th>Resolution</th>
<th>T1279 L91</th>
<th>T799 L91</th>
<th>T399 L62</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid spacing</td>
<td>16km</td>
<td>25km</td>
<td>50km</td>
</tr>
<tr>
<td>Number of grid-points</td>
<td>2,140,704</td>
<td>843,490</td>
<td>213,988</td>
</tr>
<tr>
<td>Time-step</td>
<td>450 secs</td>
<td>720 secs</td>
<td>1800 secs</td>
</tr>
<tr>
<td>Flops for 10-day forecast</td>
<td>$7.207 \times 10^{15}$</td>
<td>$1.615 \times 10^{15}$</td>
<td>$0.1013 \times 10^{15}$ → EPS * 50</td>
</tr>
</tbody>
</table>
RAPS9 – 10-day T799 L91 Forecast

![Graph showing the percentage of peak performance over the number of PEs for different configurations: hpce T1279, hpce T799, and hpcd T799.](image)

Use of High Performance Computing in Meteorology 30 Oct 2006 Slide 13
RAPS9 – T799 L91 10-day Forecast on hpce

![Graph showing speed-up versus number of PEs for different configurations of RAPS9 models. The graph includes lines for Ideal, T799, and T1279, indicating speed-up values for different numbers of PEs.](image)

- **Number of PEs:** 384, 768, 1024, 1536, 2048, 2240
- **Speed-up:** Ideal, T799, T1279

Use of High Performance Computing in Meteorology  30 Oct 2006  Slide 14

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

![Graph showing speed-up vs number of PEs]

- **Ideal**
- **T799**

Use of High Performance Computing in Meteorology 30 Oct 2006 Slide 15
Mflop/s per Subroutine from Dr. Hook

→ T799 L91 forecast run on 128 PEs

<table>
<thead>
<tr>
<th>Subroutine</th>
<th>Mflop/s per PE hpcd</th>
<th>Mflop/s per PE hpce (no SMT)</th>
<th>Mflop/s per PE hpce (SMT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLoudsc</td>
<td>533</td>
<td>559</td>
<td>758</td>
</tr>
<tr>
<td>MXMAOP</td>
<td>3267</td>
<td>5270</td>
<td>4700</td>
</tr>
<tr>
<td>LAITQM</td>
<td>1072</td>
<td>987</td>
<td>1350</td>
</tr>
</tbody>
</table>
RAPS9 - T799 L91 10-day forecast
– OpenMP threads / MPI task on 96 Nodes

Number of threads per MPI task

Percentage of Peak

4 Threads used for operations

Use of High Performance Computing in Meteorology 30 Oct 2006 Slide 19
### RAPS9 - 10-day forecasts

Message passing communications on hpce

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Nodes</th>
<th>WALL (secs)</th>
<th>%Comms (barrier)</th>
<th>Tflop/s</th>
<th>% of peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>T799 L91</td>
<td>24 Nodes 96 x 8</td>
<td>4253</td>
<td>8.0%</td>
<td>0.38</td>
<td>13.0%</td>
</tr>
<tr>
<td>T1279 L91</td>
<td>96 Nodes 384 x 8</td>
<td>4836</td>
<td>11.5%</td>
<td>1.61</td>
<td>12.8%</td>
</tr>
<tr>
<td>T799 L91</td>
<td>140 Nodes 560 x 8</td>
<td>995</td>
<td>18.9%</td>
<td>1.60</td>
<td>9.4%</td>
</tr>
<tr>
<td>T1279 L91</td>
<td>140 Nodes 560 x 8</td>
<td>3506</td>
<td>13.8%</td>
<td>2.05</td>
<td>12.1%</td>
</tr>
</tbody>
</table>
### RAPS9 - T799/T95/T255 L91 4D-Var run on hpce
- 16 Nodes, 128 MPI tasks & 4 Threads

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