Preparing NWP-Models for Tera-Computing

Ulrich Schättler, Elisabeth Krenzien, Michael Czajkowski

Deutscher Wetterdienst
Contents

• T3E → IBM: Late Lament
• Upgrade of NWP-System at DWD
• Optimizations for Boundary Exchange and I/O
• Environment for the NWP-System: Now / Future
• LM_RAPS_3.0
• Conclusions
T3E → IBM: Late Lament

• If profiling various applications on a big IBM-System, which MPI-Routine do you expect to take the most time?
• It is: MPI_BARRIER!
• Because most codes have been developed on a T3E:
  – Barriers did take almost no time
  – MPI_SEND, MPI_RECV was not really blocking and a barrier could be helpful to ensure correct program execution!
• Is IBM-communication really so bad?
Timings for LM on T3E / IBM

<table>
<thead>
<tr>
<th>Timings in Seconds</th>
<th>T3E</th>
<th>IBM (pwr3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyn. Computations</td>
<td>1146.58</td>
<td>1105.09</td>
</tr>
<tr>
<td>Communications</td>
<td>324.90</td>
<td>568.19</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>187.09</td>
<td>373.63</td>
</tr>
<tr>
<td>Phys. Computations</td>
<td>708.80</td>
<td>712.22</td>
</tr>
<tr>
<td>Communications</td>
<td>30.84</td>
<td>60.05</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>200.41</td>
<td>241.18</td>
</tr>
<tr>
<td>I/O</td>
<td>576.05</td>
<td>358.05</td>
</tr>
<tr>
<td># Processors</td>
<td>484</td>
<td>160</td>
</tr>
</tbody>
</table>
Timings for LM on T3E / IBM

- Today’s operational domain size: $325 \times 325 \times 35$
- 48 hour forecast (should be finished within 1 hour)
- Timings: „Not so good, but acceptable“
- But what happens, if $O(1000)$ processors are used?
Upgrade of NWP-System

- New model components
  - Cloud ice scheme (GME / LM)
  - Multi-layer soil model (GME / LM in QI 2005)
  - Sea-ice model (GME)
  - Prognostic Precipitation (LM)
  - 2 time level Runge-Kutta numerical core
    - 3rd order in time; 5th order (horizontal) in space
    - at the moment tested for very high resolution runs
Upgrade of NWP-System

- In Preparation
  - 3D Var Physical Space Assimilation System (GME)
  - Assimilation of radar data (latent heat nudging; LM)
  - 3D Turbulence scheme (LM)
  - Graupel scheme (LM)
  - Parameterization of shallow convection (LM)
  - Lake-Model (LM)
Upgrade of NWP-System

• Local Model (LM):
  – The LM is used and further developed within the Consortium for Small Scale Modeling (COSMO)
  – Aim is to run LM with a very high resolution (≤ 3 km): LMK (LM Kürzestfrist)
  – But coming up next: Running the LM with 7 km resolution over the whole of Europe (LME)

• Global Model (GME):
  – GME is now run with a resolution of about 40 km and 40 vertical layers
LMK

- 2.8 km grid spacing
- 421×461 grid points
- 50 vertical layers
- 30 s time step
- 2 time level Runge-Kutta
- continuous upgrade by new components
- operational in 2006
LME

- Still 7 km grid spacing
- 665×657 grid points
- 40 vertical layers
- 40 s time step
- 3 time level Leapfrog
- (or Runge-Kutta: 72 s)
- operational in 2005 (QII)
# LME: First Timings

<table>
<thead>
<tr>
<th>Category</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyn. Computations</td>
<td>367.51</td>
</tr>
<tr>
<td>Communications</td>
<td>85.30</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>130.37</td>
</tr>
<tr>
<td>Phys. Computations</td>
<td>188.16</td>
</tr>
<tr>
<td>Communications</td>
<td>7.31</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>31.76</td>
</tr>
<tr>
<td>Input</td>
<td>126.62</td>
</tr>
<tr>
<td>Output</td>
<td>258.70</td>
</tr>
<tr>
<td>Total Time for the Job</td>
<td>1285.46</td>
</tr>
</tbody>
</table>
LME: First Timings

• 12 hour forecast on 900+1 procs (should be ≤ 900s)
• Boundary Exchange with MPI_ISEND, MPI_WAIT on the sender and MPI_RECV on receiver side
• Explicit buffering of data
• Extra processor for I/O (asynchronous IO); but realization with blocking communication
• Still using all that barriers!
LM Time Step with Trace Analyzer
Output Step
Input Step
Optimizations

- Boundary Exchange with
  - `MPI_ISEND`, `MPI_WAIT` and `MPI_RECV`
  - `MPI_Irecv`, `MPI_WAIT` and `MPI_SEND`
  - `MPI_SENDRECV`

- Implicit buffering of data by using `MPI_DATATYPES`
- Non-blocking communication for extra I/O processor
- Try to do a look-ahead reading
- There is really no need for barriers!
Optimized LM Time Step
Optimized Output Step
Optimized Input Step
Optimized Input Step - 2
# LME: Old vs. Optimized Timings

<table>
<thead>
<tr>
<th></th>
<th>Old</th>
<th>Comm+ Output</th>
<th>+ Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyn. Computations</td>
<td>367.51</td>
<td>349.29</td>
<td>349.65</td>
</tr>
<tr>
<td>Communications</td>
<td>85.30</td>
<td>138.96</td>
<td>221.58</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>130.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phys. Computations</td>
<td>188.16</td>
<td>187.36</td>
<td>187.50</td>
</tr>
<tr>
<td>Communications</td>
<td>7.31</td>
<td>25.40</td>
<td>26.09</td>
</tr>
<tr>
<td>Barrier waiting</td>
<td>31.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input</td>
<td>126.62</td>
<td>135.29</td>
<td>80.70</td>
</tr>
<tr>
<td>Output</td>
<td>258.70</td>
<td>142.21</td>
<td>142.19</td>
</tr>
<tr>
<td>Total Time for the Job</td>
<td>1285.46</td>
<td>1066.35</td>
<td>1095.01</td>
</tr>
</tbody>
</table>
Profiling with MPI_Trace (IBM)

325 × 325 × 35

<table>
<thead>
<tr>
<th>Function</th>
<th>Old Version</th>
<th>Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI_ALLREDUCE</td>
<td>0.24</td>
<td>3.71</td>
</tr>
<tr>
<td>MPI_BARRIER</td>
<td>30.08</td>
<td>0.01</td>
</tr>
<tr>
<td>MPI_ISEND</td>
<td>0.40</td>
<td>0.00</td>
</tr>
<tr>
<td>MPI_RECV</td>
<td>8.07</td>
<td>3.18</td>
</tr>
<tr>
<td>MPI_SEND</td>
<td>2.46</td>
<td>0.00</td>
</tr>
<tr>
<td>MPI_SENDRECV</td>
<td>-</td>
<td>21.49</td>
</tr>
<tr>
<td>MPI_WAIT</td>
<td>0.77</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Tools for Performance Analysis

- Trace Analyzer (VAMPIR)
  - good to detect problems in communication
  - helps to understand your code

- MPI Trace
  - to detect hot-spots in the communication
  - helps to understand MPI Performance

- HPMCOUNT (Hardware Performance Monitor)
  - produces a lot of data (at least on IBM)
  - results are not easy to understand
NWP Environment: IT Structure 2004
IBM SP 9076 550 (NH II, 375 MHz)

120 Nodes:
- Production (cos4)
- Development (cos5, cos6)
  - Login 4
  - Compute 84
  - GPFS 32

System software:
- AIX, PSSP, GPFS, LoadLeveler, f90, MPI, OpenMP
- C, JAVA, UNICORE

Applications:
- Assimilation, GME, GM2LM, LM, GSM, LSM, MSM, Trajektorien, LPDM RLM, csobank to das3/4, ecfs to das1, Projects with external partners

Filesystems:
- Production
  - 4344 GB SSA, GPFS
- Development
  - 1629 GB SSA, GPFS
- TEMP (gtmp)
  - Filesystems
    - 1086 GB SSA, GPFS

Internal network: colony switch, cisco router

Gigabit Ethernet Switches

- rus3 AIX
- rus4 AIX
- das3 AIX
- das4 AIX
- das1 IRIX

Testsysten ixt5
- 4 nodes
- Filesystems
  - 280 GB SSA, GPFS

Oracle

AMASS, DataMgr
Future Plans for DMRZ

2005: Application for funds
2006: Invitation to tender
2008: Start of operation in a new building

How does „forecast-process“ look like in 2008?

Performance enhancement by 8-10 (ca. 30 TeraFlop/s peak performance)

Investigate possible use of Linux-Clusters
LM_RAPS_3.0

- To reflect the recent model changes, a new RAPS-Benchmark has been released
  - Changes in communication for boundary exchange are included
  - Changes for asynchronous IO are not included
- Benchmark is available for vendors (and interested people), if an „Agreement“ is signed (⇒ change in the RAPS structure)
Conclusions

• A programmer’s work is never done

• Getting good communication performance on the IBM is not impossible, but takes time

• Optimizations may be machine dependent (offer choice of selection)

• Next: How to optimize the computations (Flop/s)?
Thank You