Linux Clusters at EARS

what is coming next?

Jure Jerman
Meteorological Office
Environmental Agency of Slovenia (EARS)

Outline

- Linux Cluster at Environmental Agency of Slovenia
- Operational experiences
- Future requirements for limited area modeling
- Needed ingredients for future system?

Tuba – current cluster system

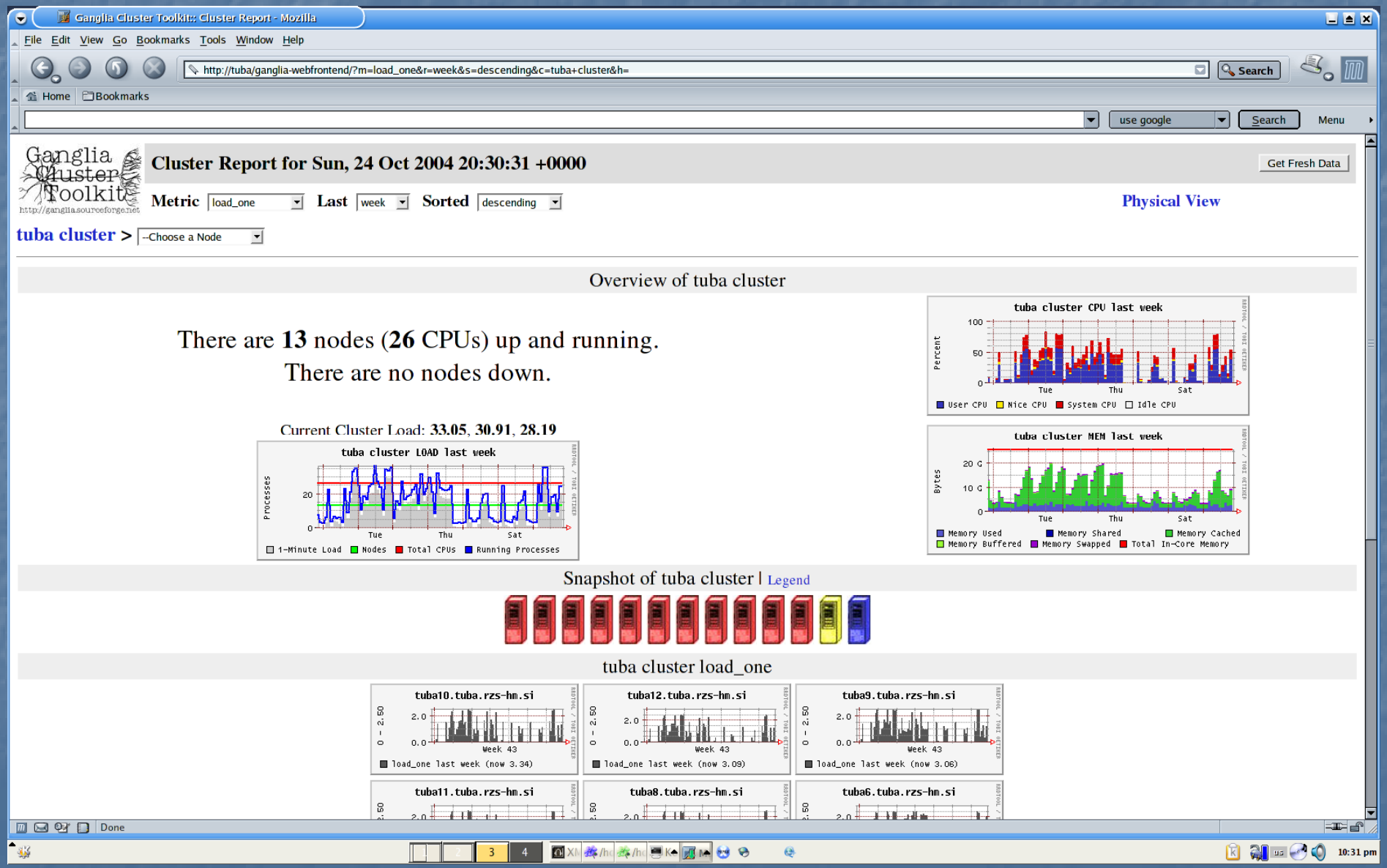
- Installed 2 years ago, presented at 10th ECMWF HPC workshop:
- Hardware:
  - 13 Compute Nodes,
  - 1 Master Node, Dual Xeon 2.4 Ghz,
  - 28 GB memory
  - Gigabit Ethernet
- New: 4 TB IDE2SCSI disk array, xfs filesystem

Tuba software

Open source, whenever possible

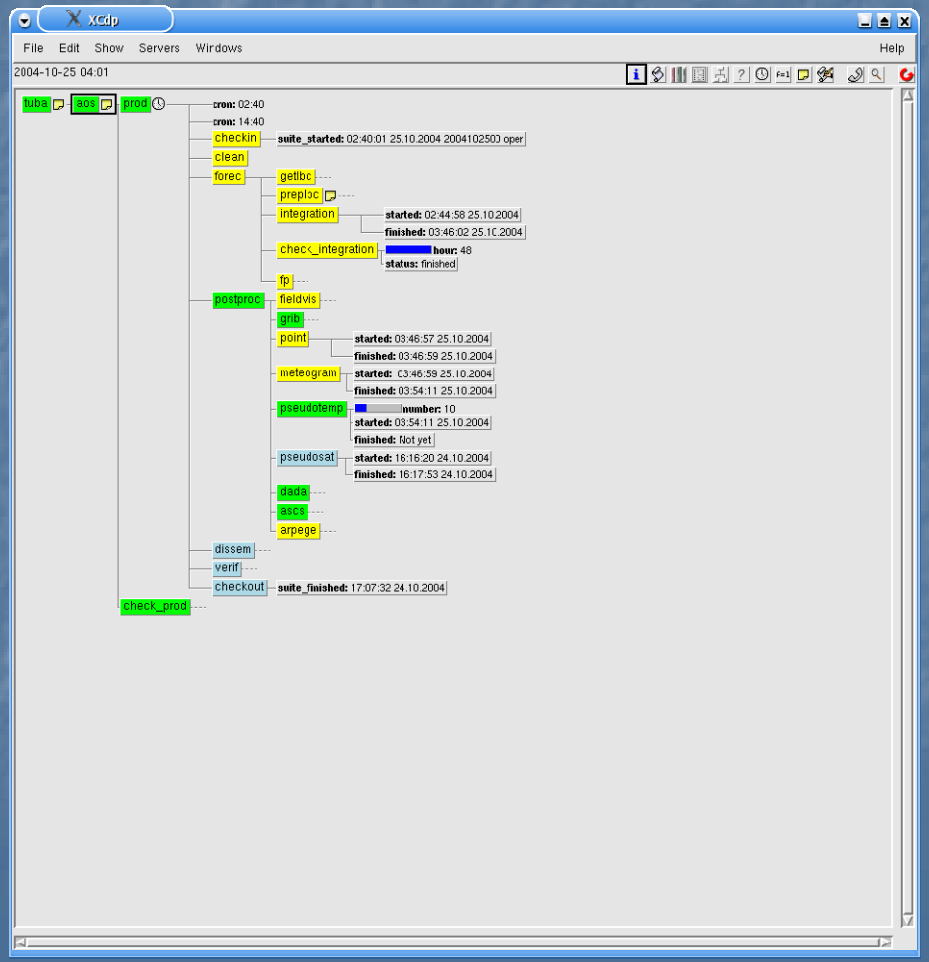
- Cluster management software:
- OS: RH Linux + SCore (5.4) (www.pccluster.org)
- Mature parallel environment
  - Lower latency MPI implementation
  - Transparent to user
  - Gang scheduling
  - Pre-empting
  - Checkpointing
  - Parallel shell
  - Automatic fault recovery (hardware of Score)
  - FIFO scheduler
  - Capability of integration with OpenPBS and SGE
- Intel compilers

Ganglia - Cluster Health monitoring

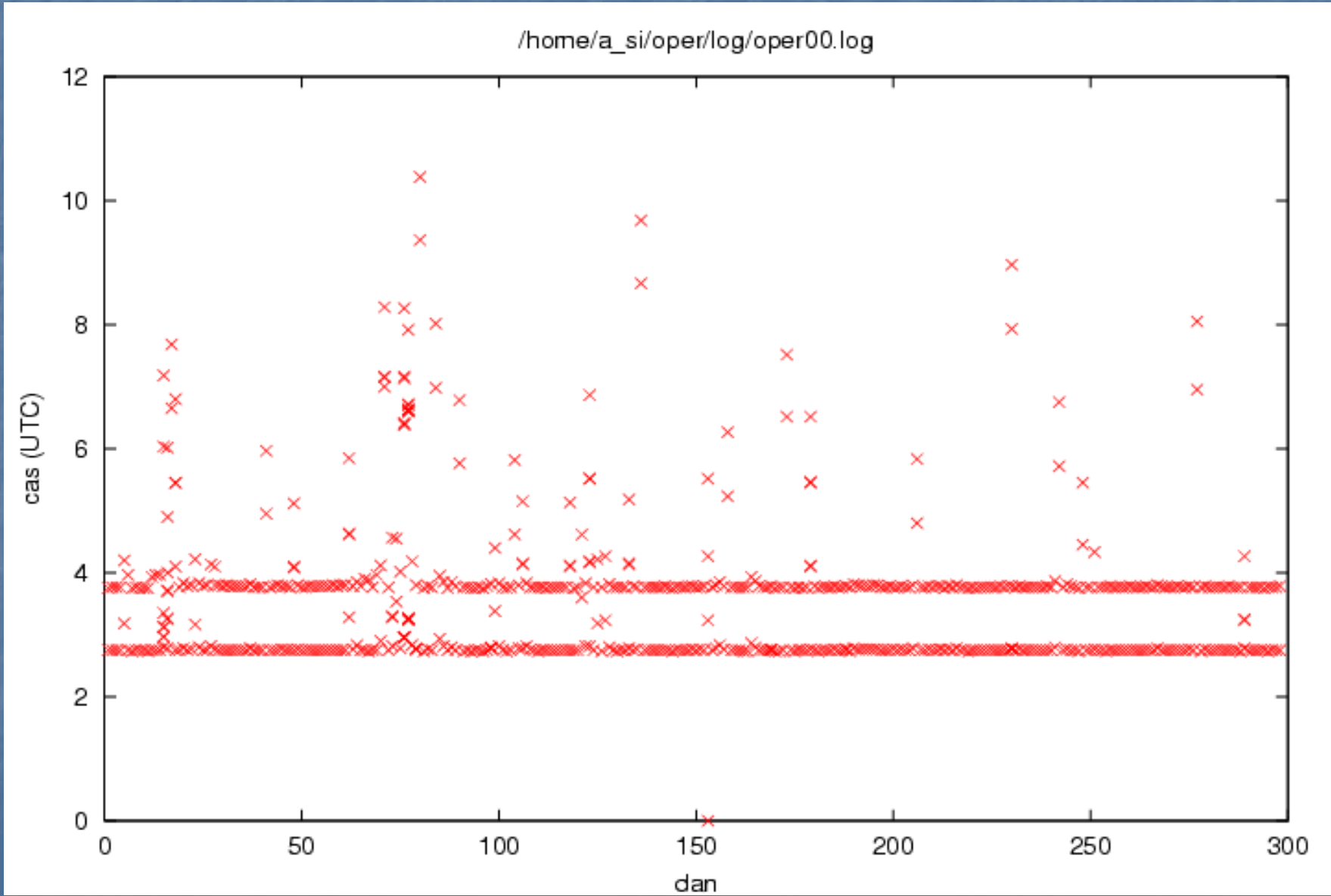


Operational experiences

- In production for two years
- Unmonitored suite
- NO hardware related problems so far!
- Some problems with SCore (mainly related to buffers in MPI)
- NFS related problems
- ECMWF's SMS, solves majority of problems



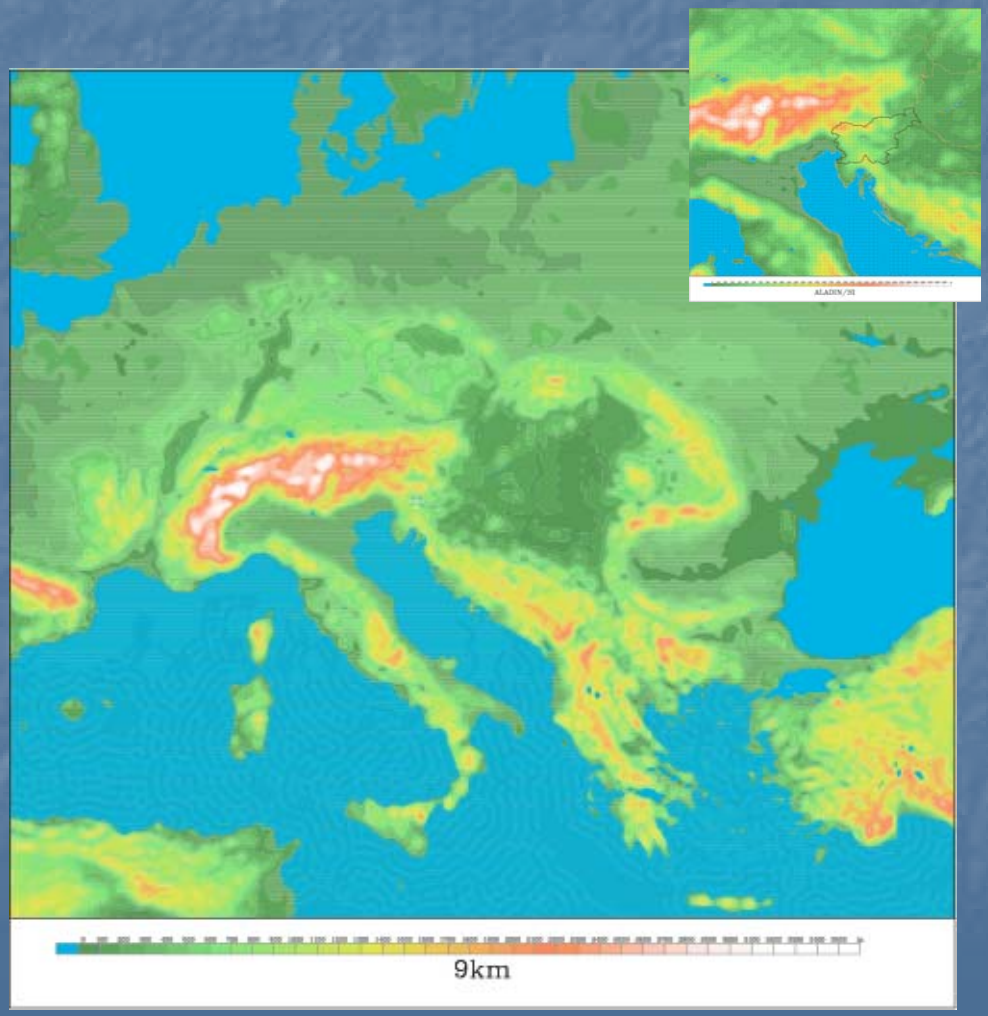
Reliability



New operational setup

ALADIN model

- 290x240x37 domain
- 9.3 km resolution
- 48h integration
- 55 min



Optimizations

Not everything in a hardware

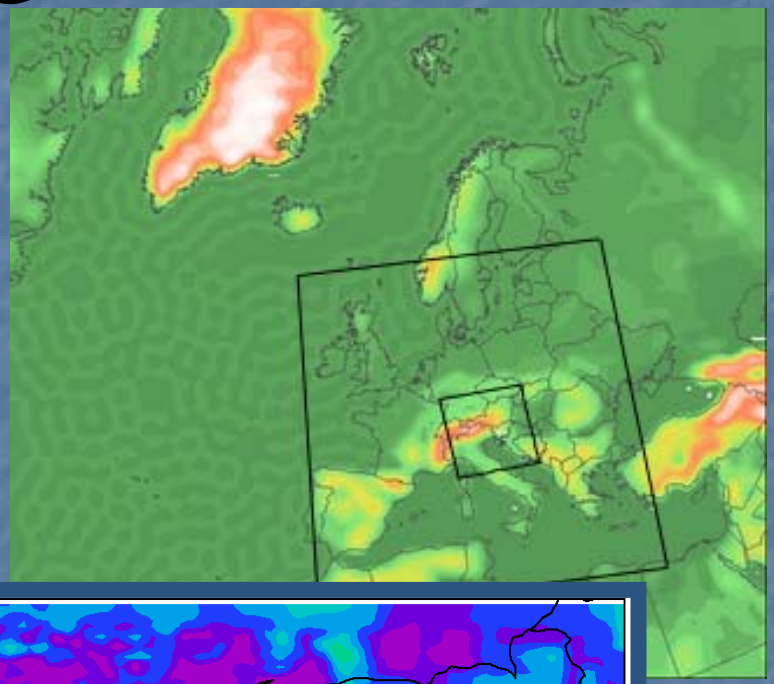
Code optimizations

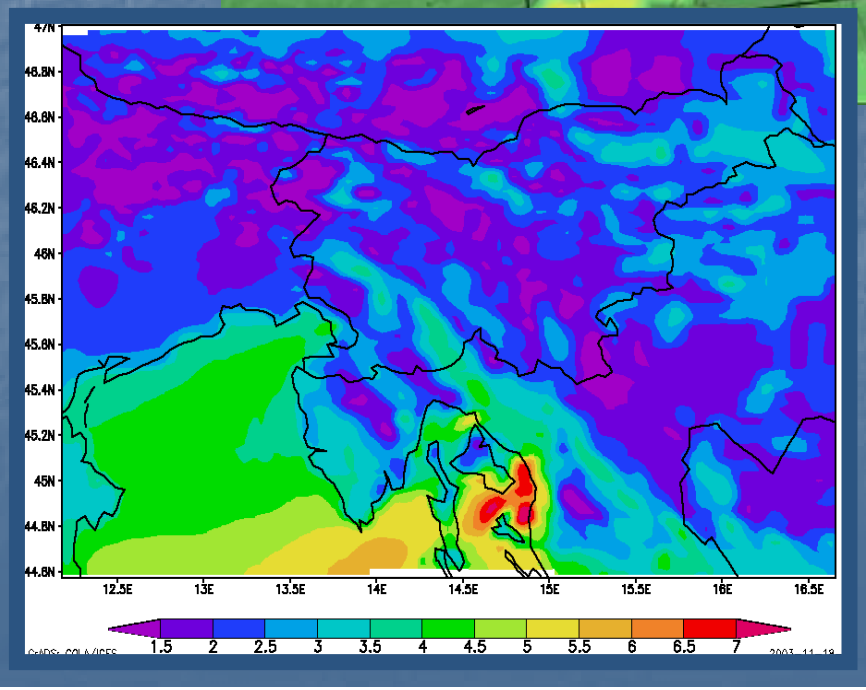
- B-Level parallelization (up two 20 % at greater number of processors)
- Load balancing of grid point computations (depending on the number of processors)

- Parameter tuning
  - NPROMA cash tuning
  - MPI message size
- Improvement in compilers (Lahey –> Intel 8.1 20 – 25 %)
- Still to work on: hyperthreading in combination with OpenMP

Non operational use

- Downscaling of ERA-40 reanalysis with ALADIN model
  - Estimation of wind energy potential over Slovenia
  - Multiple nesting of target computational domain into ERA-40 data
  - 10 years period, 8 years / month
- Other research jobs
  - Radar latent heat nudging
  - Spectral coupling
- Coexistence with operational suite





Foreseen developments in limited area modeling

- Currently ALADIN 9 km
- 2008-2009 Arome, 2.5 km (Meteo France project): ALADIN NH solver + Meso NH physics
- "Grey zone" problem
- Smooth convergence with Arome trough ALARO
- Expensive, 3 x per grid point
- Target Arome: ~100 x – 200 x more expensive

How to get there (if?)

Linux commodity cluster at EARS?

- First upgrade at the end of 2005
- 4-5 times the current system (if possible, below 64 processors)
- Tests going on with:
  - New processors: AMD Opteron, Intel Itanium-2
  - Interconnection: Infinyband, Quadrics?
- Compilers: PathScale (AMD Opteron)
- Crucial: Parallel file system (TerraGrid)

How to stay at the open side of the fence?

- Linux and other OpenSource projects are evolving
- Great number of more and more complex software projects
- Specific (operational) requirements in meteorology
- Space for system integrators
- Price/performance gap between commodity and brand name systems is getting smaller when the size of system is growing
- Pioneer time of Beowulf clusters seems to be over
- Importance of extensive test of all cluster components

Conclusions

- Positive experiences with small commodity Linux cluster, great price/performance ratio
- Our present type of development of new cluster works for small cluster, might work for medium sized and doesn't for big systems
- Future are Linux clusters, but branded