

Developments concerning parallel computing with the Telemac System

Jacek A. Jankowski

Abteilung Wasserbau im Binnenbereich
Bundesanstalt für Wasserbau
Karlsruhe



Telemac@BAW

- BAW provides scientific expertise for waterways maintenance and development:
 - physical and
 - numerical modelling
- Telemac applied in coastal and inland departments:
 - Ca. 10 engineers using Telemac
 - A co-development agreement with EDF



Parallel computing

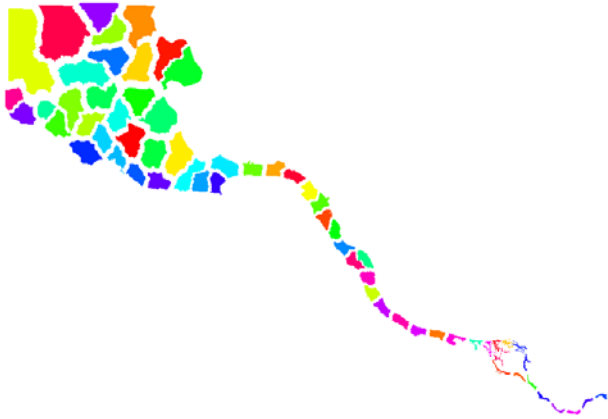
- Larger models + higher resolution
- Consequence:
 - parallel computing
 - advanced visualisation
 - grid computing
 - Linux clusters



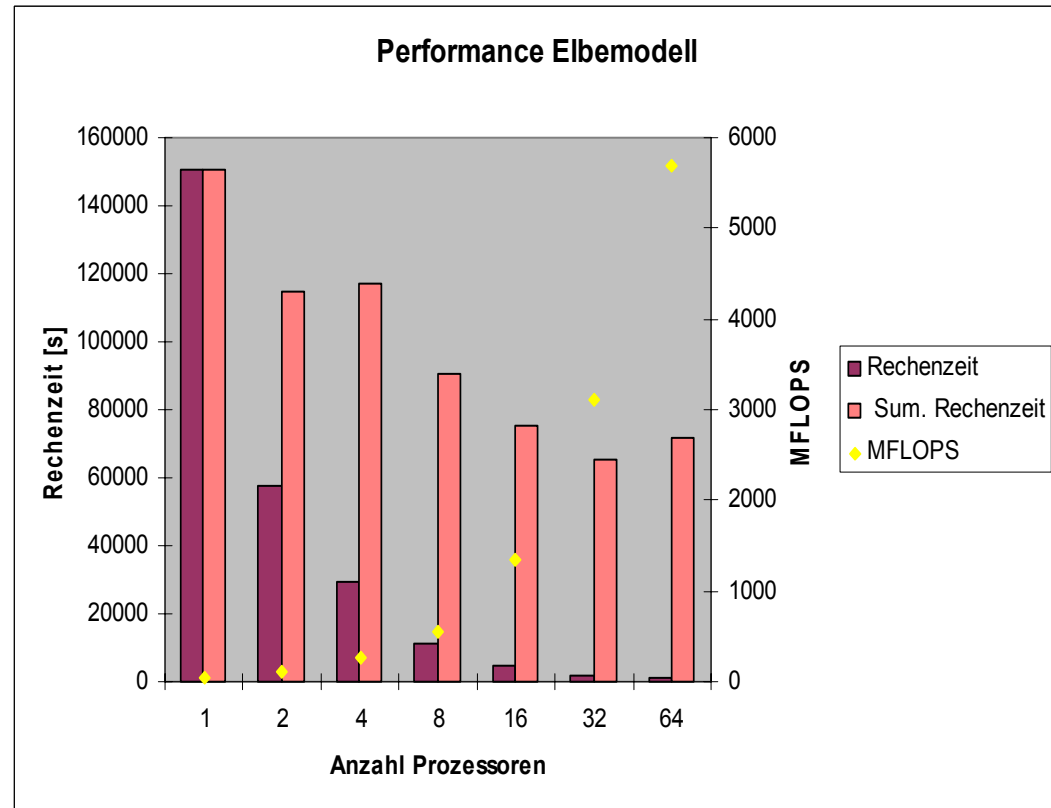
- BAW Hamburg: **2 SGI parallel computers** (256+32 Processors)
- BAW Karlsruhe: **2 SGI parallel computers** (32+8 Processors)
2 Linux cluster (16+32 Processors), 2 visualisation rooms
- DWD Offenbach: **IBM parallel computer** (1024+512 Processors)



Elbe model performance



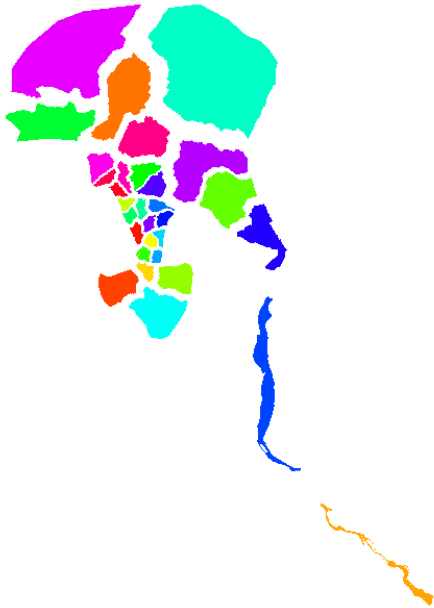
99192 nodes
190314 elements



Telemac (FEM, MPI): Computation:Reality = 1:42 by 32 proc.
Trim (FDM-25m, OpenMP): Computation:Reality = 1:20 by 16 proc.



Jade-Weser performance



150412 nodes
296466 elements



Telemac (FEM, MPI):

Computation:Reality = 1:25 by 32 proc.

UnTrim (FV/DM, OpenMP):

Computation:Reality = 1:10 by 16 proc.



Overall parallel performance

SGI Origin (2000, 3000 series),
comparable to IBM RS/6000 SP

- Ca. 100 Mflops / MIPS processor (400-600 MHz)
- Secondary (L2) cache of 8MB - optimally < 8000-10000 elements / processor
- Low network traffic by a good job placement
- Domain decomposition + MPI was a very good choice for computationally intensive Telemac-2D



Contents

Numerous smaller or larger improvements
delivered by BAW Karlsruhe, Hamburg,
University of Hannover, LNHE, SOGREAH, ...

- in the Telemac System *software environment* and
- in Telemac-2D *code* itself

**made mainly in order to improve the parallel
computing ease and efficiency**



Checkpointing & Restart

- **Checkpointing** describes saving a running application *upon a random request* in such a way, that a...
- ...**Restart** from the saved state is possible
 - system-level checkpointing
 - user-level checkpointing
 - application-level checkpointing



C&R: Ideas

- Send the Unix user-signal `SIGUSR1` to all running Telemac processes
- Telemac traps the signal...
- ...triggers saving the present computation state...
- ...and stops “gracefully”
- An additional system-specific library `system` presently tested for Linux, IRIX, AIX



C&R: Usage

- User calls a script `signal_cas754321.bat`
- When restarting:

GRAPHIC PRINTOUT PERIOD = 100

LISTING PRINTOUT PERIOD = 10

/ checkpointing was for the time step number 507:

NUMBER OF FIRST TIME STEP FOR GRAPHIC PRINTOUTS=93

NUMBER OF FIRST TIME STEP FOR LISTING PRINTOUTS = 3



Returning exit codes

- Introduced in Telemac, Partel and Gretel
- A correct run:
 - The exit code 0 must be guaranteed
- If a “controlled” or “uncontrolled” error occurs:
 - returning a non-zero exit code
- The exit codes interpreted by software environment and passed to the operating system



Work directory

- Most larger institutions have a computer network with specialised servers
 - Secure **file servers** for project data
 - **Compute servers** with temporary file systems
- The user must have the possibility to decide where the work directory of Telemac should be made

Call: `telemac2d -t parameterfile /work`



Killing softly

- User calls a script `delete_7654321.bat`
- It sends `SIGINT` to interrupt MPI processes...
and if it does not succeed...
- ... `SIGTERM` to terminate all processes
- ...`SIGKILL` to kill all processes

- Important for jobs using Network File System (NFS)
- Typical approach in most batch systems



Order in files

- Software environment and system:
`stdout` & `stderr`
- Telemac output redirected to log-files
(from separate processors)
- `partel` & `gretel` output always saved
- Error messages sent to files



Various other improvements

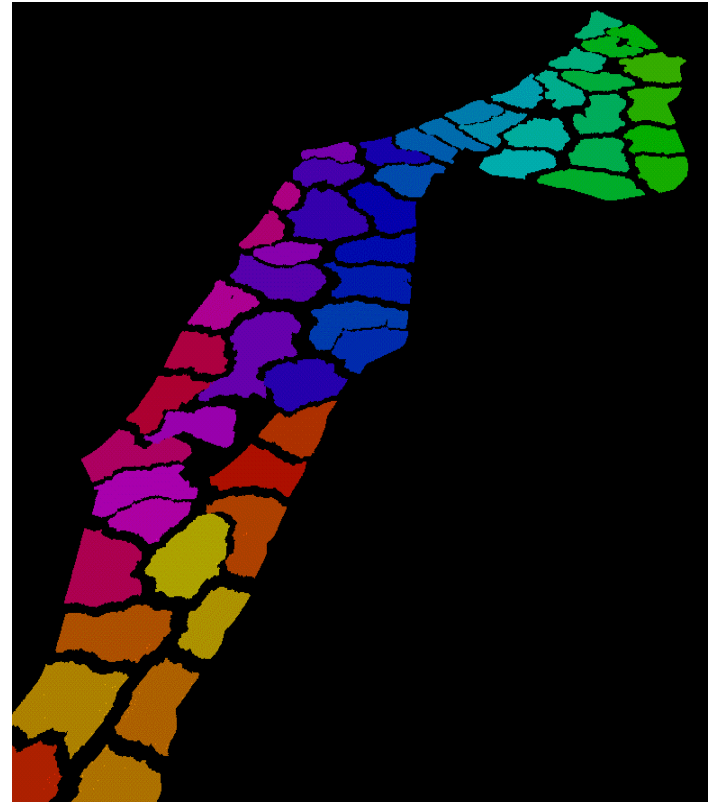
- Output synchronisation when variable time steps
- A default parallel executable
- A global `mpi_telemac.conf` file
- Compilation first, then partitioning!
- Compilation&link-only run:

```
telemac2d -cl cas
```



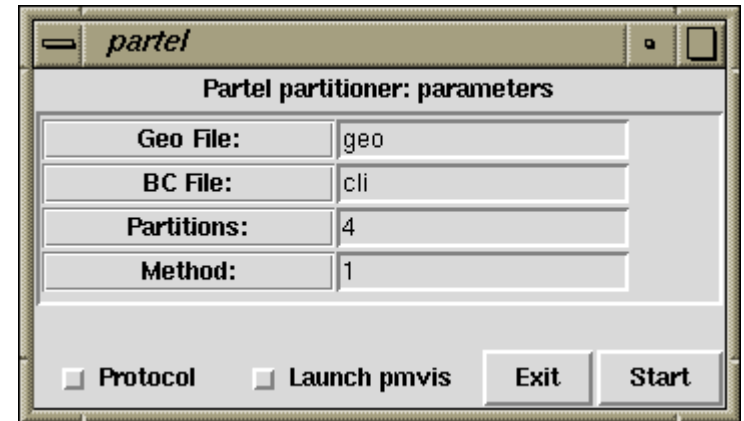
”Boundaryless” subdomains

- Difficulties when subdomains had no access to the global domain boundary
- Problem analysed (mainly boundary treatment) and solved by University of Hannover



partel

A *Metis*-based partitioning program (replacing **hanse1**)

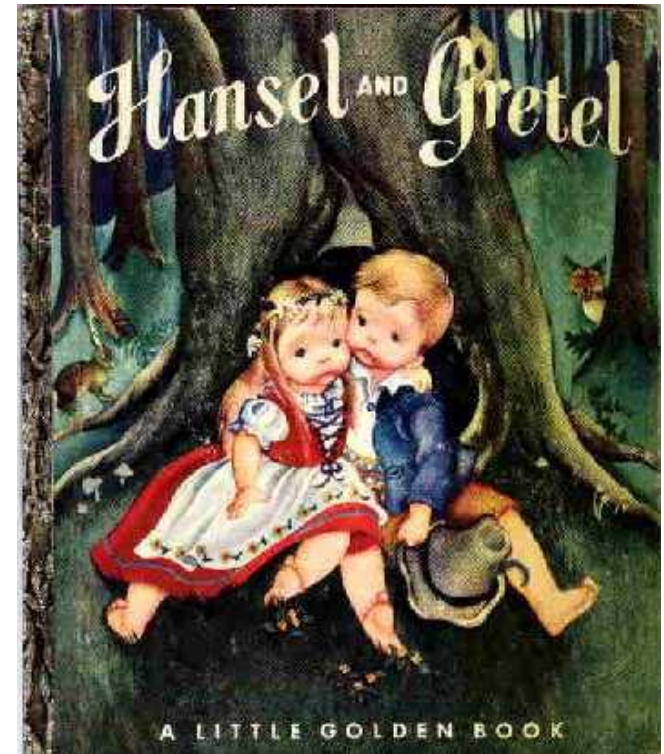


- Improved:
 - Treatment of islands
 - Treatment of difficult mesh topologies
 - Up to 1000 processors (beware!)
 - A few minor bugs removed
- Generating error codes

gretel

The “trailer” program `gretel`

- Merging of result files for a larger number of
 - partitions and
 - time steps savedproved to be very slow
- Improved: reading and merging all files simultaneously (beware!)



Grid computing

- Emerging as a premier usage mode of large-scale systems
- Revolution in the approach to high-performance computing resources – available anytime, anywhere?
- **Already 33% of BAW-Karlsruhe CPU-time usage originates from grid (distributed) computing**



Unicore



- *Uniform Interface to Computing Resources*
- Makes distributed computing and data resources available in a secure way over internet & intranet
- **Client:** a graphical interface for operations with jobs
- BAW supports *grid middleware* development



Job Preparation

- WesXan_IF3 [16:10:14 08/25/2003]
- WesXan_DWD [10:15:37 07/23/2003]
 - Import_DWD
 - Compile_DWD
 - Run_DWD
 - Export_DWD

Job Monitoring

- BAW-KA
 - onyx2 <NJS>
 - origin <NJS>
- DWD
 - globusSite(wren) <NJS>
 - ibm_sp(cos5)_ext <NJS>
 - WesXan_DWD [08:05:59 10/02/2003]
 - Import_DWD
 - Compile_DWD
 - Run_DWD
 - Export_DWD
- Gate Europe
 - SUPRENUM <NJS>
 - Zuse_Z1 <NJS>
- Gate USA

Standard Output Standard Error Details Log

Starting execution: telemac2d.bat

```

*** ONLY COMPILATION AND LINKING REQUIRED ***

- FORTRAN FILE                : princi.f

*** COMPILATION ***

xlf -c -qzerosize -qmaxmem=-1 -qspillsize=32704 -O2 -I/uhome/bawja

*** LIBRARIES ***

- /uhome/bawjanko/TELEMAC/telemac//telemac2d/tel2d_v5p3/ibm/telemac2dv5p
- /uhome/bawjanko/TELEMAC/telemac//bief/bief_v5p3/ibm/biefv5p3.a
- /uhome/bawjanko/TELEMAC/telemac//damocles/damo_v5p3/ibm/damov5p3.a
- /uhome/bawjanko/TELEMAC/telemac//parallel/parallel_v5p3/ibm/parallelv5
- /uhome/bawjanko/TELEMAC/telemac//special/special_v5p3/ibm/specialv5p3.

*** LINKING ***

*** COMPILATION AND LINKING FINISHED ***

```

Execution finished: telemac2d.bat

No compilation/linking/file errors detected.
 No execution errors detected.
 Working directory '/gtmp/unicadm2_4.0/ospace_b1692e6b/cas34568_tmp' clea
 Returning exit status 0

```

=====
Telemac System V5P3 delivered from LNHE DER EDF on 7th March 2003
Perl scripts modified by jaj for BAW Wed Mar 19 17:07:07 MET 2003
=====
...stopping.

```

Save

Job Preparation

- WesXan_IF3 [16:10:14 08/25/20...]
 - Import
 - Compile_Run
 - IfThenElse
 - Then
 - Export_OK
 - Else
 - If_tmp_exists
 - Then
 - tar_tmp
 - Export_tmpdir
 - Else
 - not_found
 - If_logs_exist
 - Then
 - Export_logs
 - Else
 - not_found
 - WesXan_DWD [10:15:37 07/23...]
 - Import_DWD

Job Group

UNICORE Site

- BAW-KA
- DWD
- Gate Europe
- Gate USA

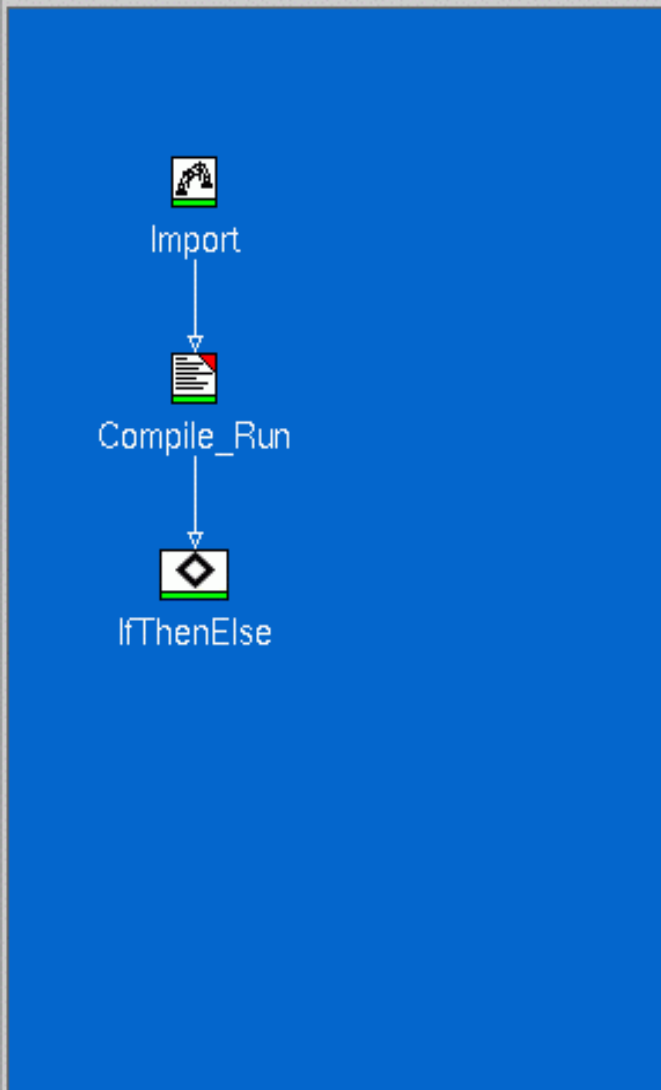
Virtual Site

- onyx2 <NJS>
- origin <NJS>

Name

Dependencies Resources Spec

Task Dependencies



Unicore Forum

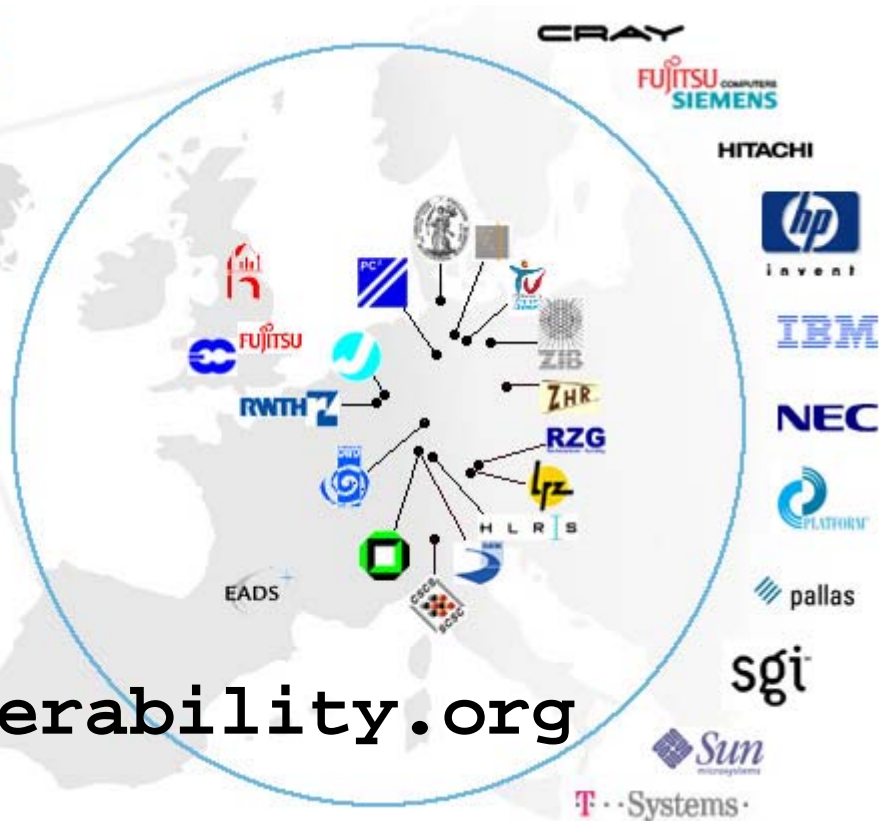
UNICORE
FORUM

www.unicore.org

www.globus.org

www.grid-interoperability.org

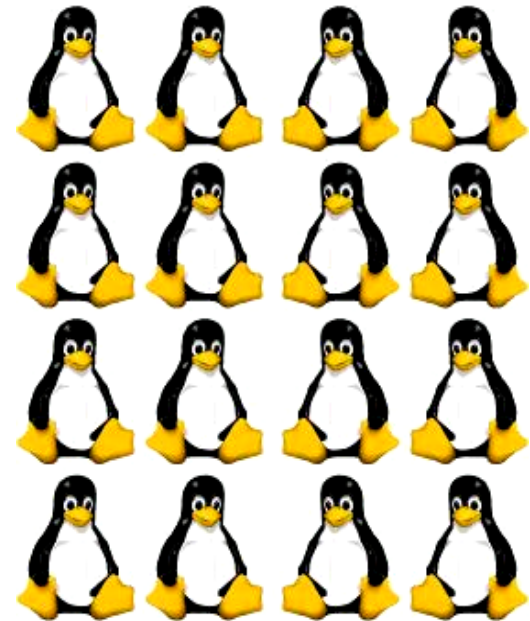
enter



Linux clusters

BAW Karlsruhe uses presently
2 Linux Clusters with Telemac:

- 16x Pentium III 800MHz,
NAGWare f95, MPICH, LSF
- 16x2 Xeon 2.8GHz, Intel
Fortran Compiler (7.1),
SCore+MPICH, PBS



To do...

- **Make the parallel execution a rule**, not an exception!
- Simplify decomposition, merging...
- The **error detection and treatment** in the present Telemac software should be changed in order to:
 - allow correct automatical reactions of the system (or grid software) in the case of an error
 - improve the error description passed to the human users



Conclusions

- The *informal support* works perfectly - most fresh developments for parallel computing have been shared between *HPC Telemac users* almost instantly
- Improvements from the field of high performance computing should find their way into the standard Telemac System



A new Telemac user



Jonathan
Kopmann

23rd September

54cm

3700g

...etc.

