Cluster Design in the Earth Sciences - TETHYS

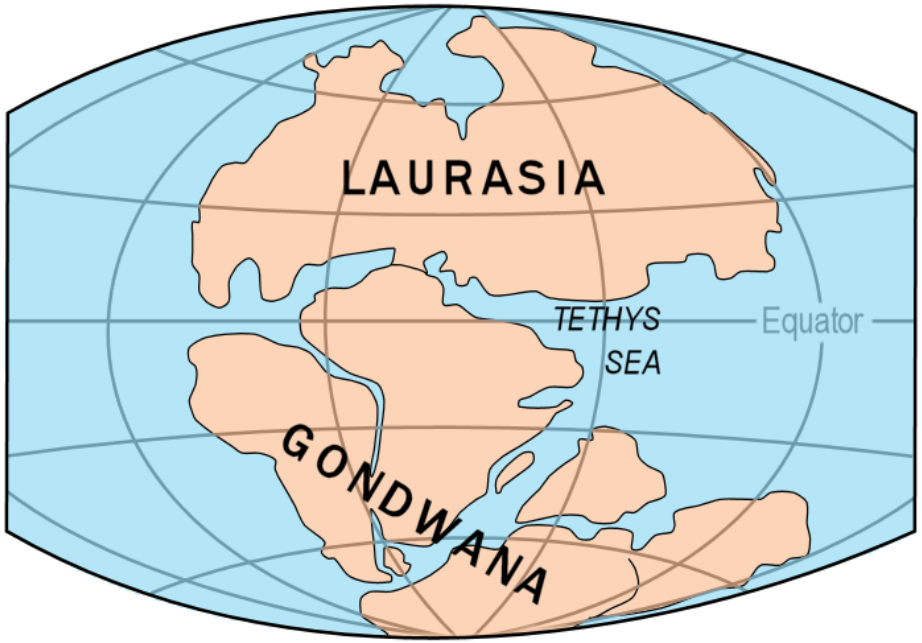
J. Oeser, H.-P. Bunge, M. Mohr

September 13, 2006

Department of Earth and Environmental Sciences
Geophysics
Ludwig-Maximilians-Universität München
Theresienstr. 41/IV
80333 Munich

TETHYS

TEc**T**onic
High Performance
Simulator



TRIASSIC
200 million years ago

Break-up of Pangaea with the Tethys Ocean is an iconic tectonic event.

Outline

- Topical Compute Clusters?
- TERRA – Mantle Convection
- Cluster – Design and Hardware
- TERRA – Performance Tests
- TERRA and SPECFEM3D
- Conclusions & Outlook

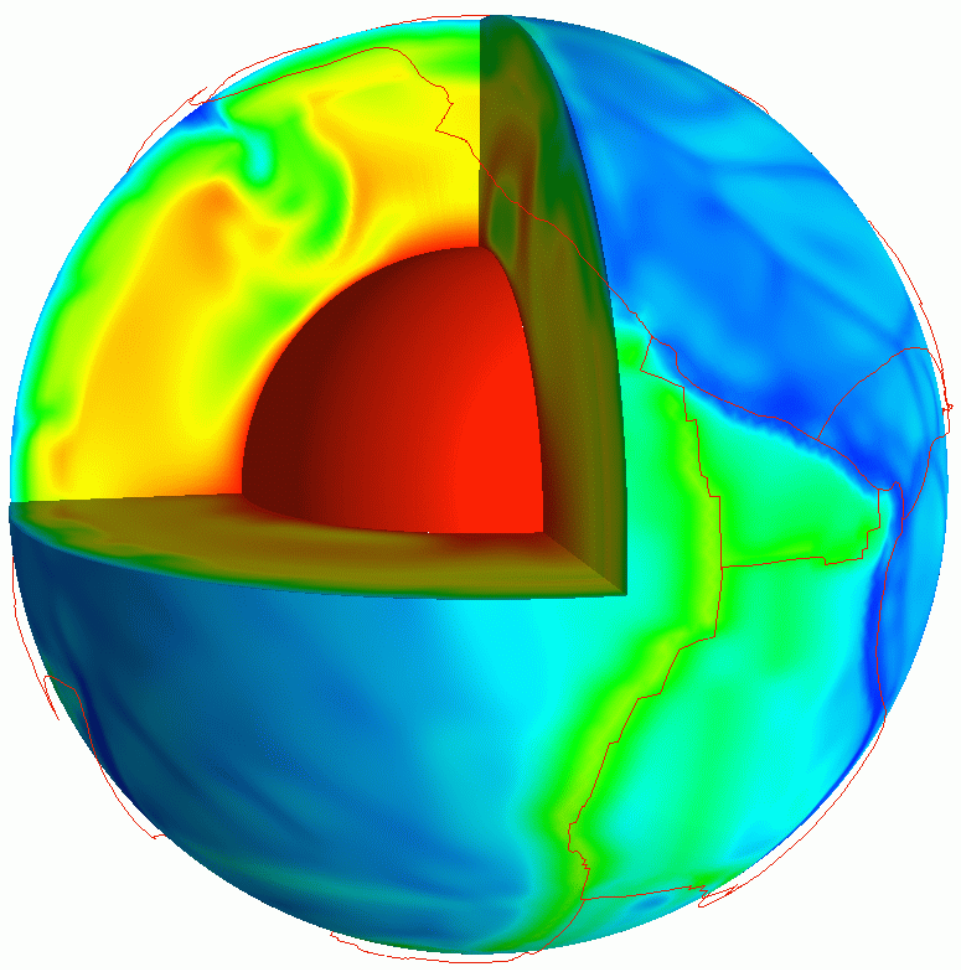
Topical Compute Clusters?

- scientific computing is increasingly important
  - models with 1000 1000 1000 grid points now feasible,
    implying scale-length resolution over 3 orders of magnitude
  - storms in the atmosphere, crust in the earth, eruptive conduits in volcanoes
  - model sensitivities now guide observations and experiments
- there is a growing need for **capacity** computing
  - 2-4 million node hours, dedicated and permanent usage of 200-500 processors every year
- computing platforms perform best when optimised for key applications
- topical computers are best run by scientific communities

Key Applications

- three key applications
  - TERRA – mantle convection
  - SPECFEM3D – seismic wave propagation
  - b3md – rupture/hazard modelling

TERRA – Mantle Convection



- earth consists of nested regions: crust, mantle, core
- convection is driven by primordial and radioactive heat
- solid state convection (creep) overturns mantle $\approx$ every 100 - 200 million years

convection drives large-scale geological activity
(plate tectonics, continental drift)

Mathematical-Physical Model

As a flow process mantle convection can be described by the compressible Navier-Stokes equations in combination with an energy equation.

Fortunately this can be simplified:

- assume quasi-static flow field and drop time-dependency from momentum equations
- small flow velocities allow to drop non-linear convection terms from momentum equations
- inertial and coriolis forces can be neglected
- assume that material is basically incompressible and use Boussinesq approximation

$\longrightarrow$ generalised Stokes problem coupled to time-dependent energy equation

Simplified PDE System

Conservation of mass:

$$\operatorname{div} u = 0$$

Conservation of momentum:

$$\operatorname{div}\left[\nu\left(\operatorname{grad} u + (\operatorname{grad} u)^T\right)\right] - \operatorname{grad} p + \varrho_0 \alpha \left(T - T_0\right) g = 0$$

Conservation of energy:

$$\varrho_0 \, c_p \left(\frac{\partial T}{\partial t} + u \cdot \operatorname{grad} T\right) - \operatorname{div}\left(\kappa \operatorname{grad} T\right) - \varrho_0 H = 0$$

u: velocity , p: pressure, T: temperature, α: coeff. of thermal expansion, c_p: specific heat at constant pressure, H: rate of internal heat production per unit volume, g: gravitational acceleration, κ: thermal diffusivity, ϱ_0: density, ν: kinematic viscosity

Algorithmic Core of TERRA

- time-dependent energy equation is integrated (forward) in time using a modified Euler scheme

- each time step requires two evaluations of the velocity field u via a generalised Stokes problem

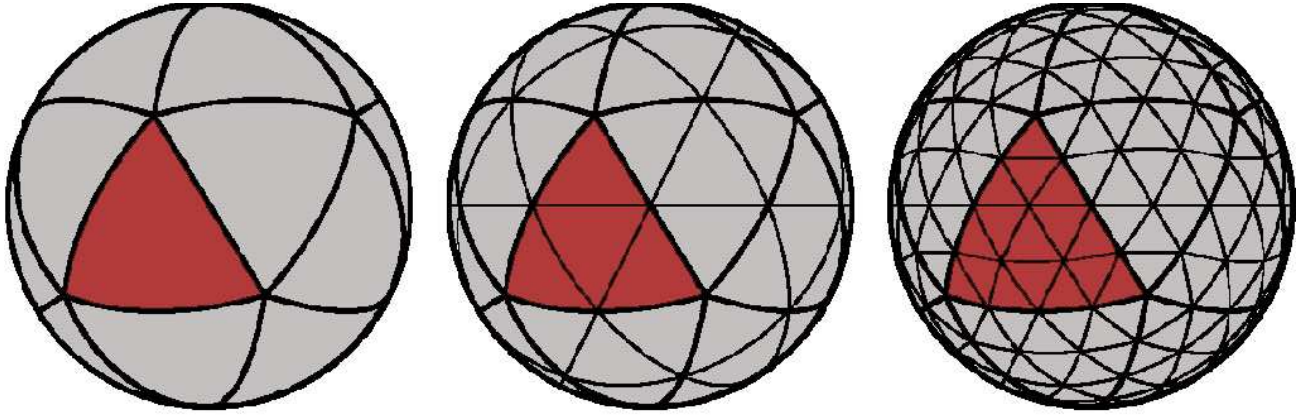
$$\text{div}\left[\nu\left(\text{grad}\, u + (\text{grad}\, u)^T\right)\right] - \text{grad}\, p = \varrho_0 \alpha \left(T_0 - T(t_i)\right) g$$

$$\text{div}\, u = 0$$

- this is done with a pressure-correction type scheme employing a multigrid method for the inner iteration

Discretisation and Parallelisation

- a surface grid is generated by mapping an icosahedron onto the sphere and successively refining it $\longrightarrow$ spherical triangles



- surface grid is radially extended down to the mantle-core boundary
- discretisation of PDE on the grid is performed with Finite Elements
- parallelisation via domain decomposition and explicit message passing

TETHYS – Design of Key Components

Topical computing allows us to choose the key components to best suit the applications under the consideration of price-performance.
What are the key components?

- CPU
  - dual better value than quad CPU nodes
  - single better value than dual core CPU nodes
- network interconnect
  - ethernet better value infiniband network connections
- memory requirements
  - driven by the key applications (for us 128 GB)

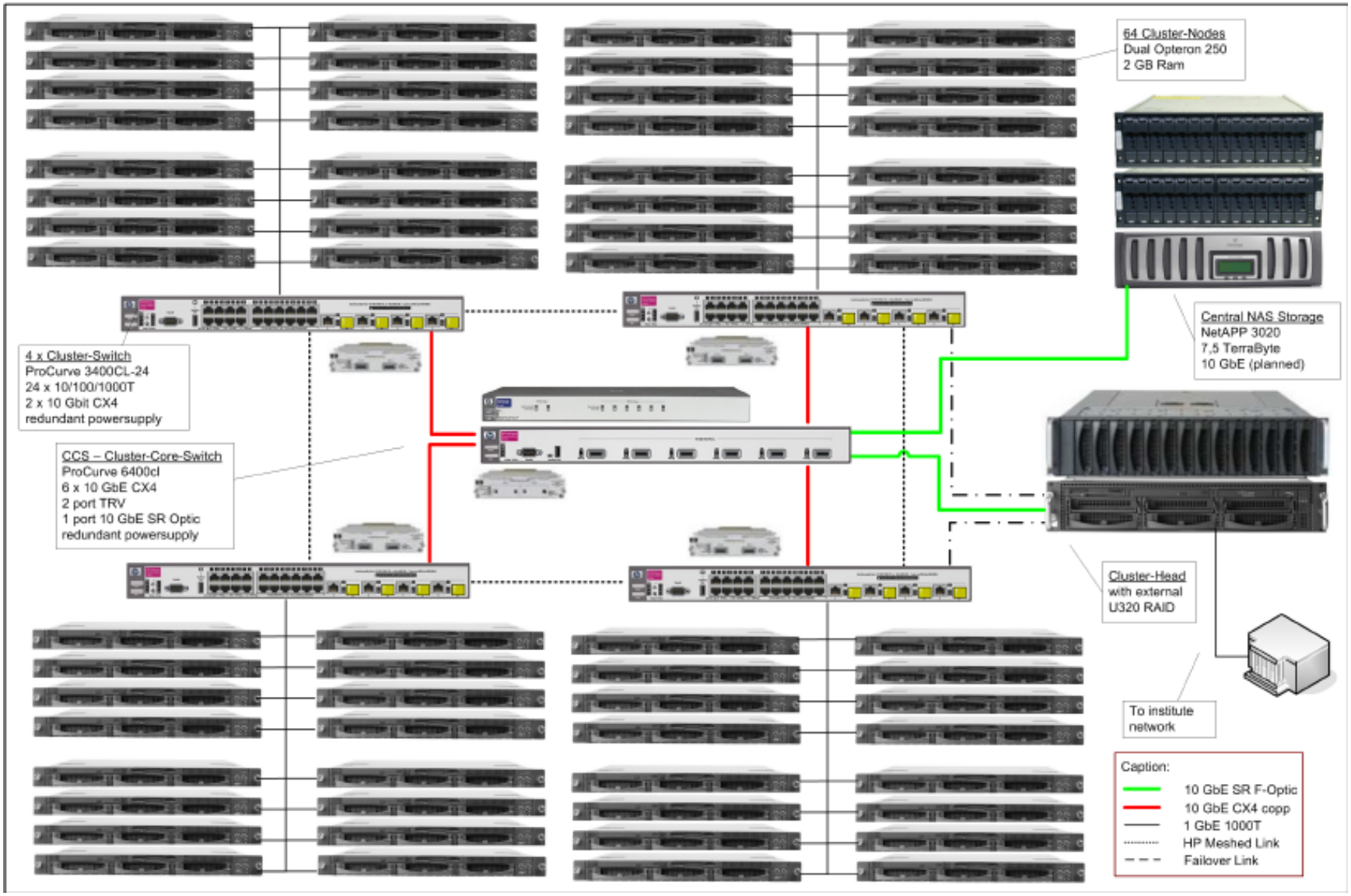
TETHYS – Hardware Specifications

- 1 head node
  - 1 INTEL XEON CPU (3.0 GHz, single core)
  - 2 GB RAM
  - 2 TB storage subsystem
  - 10 GBit and 1 GBit ethernet ports
  - high availability design (redundant power supplies, hard drives and network ports)
- 64 compute nodes
  - 2 AMD Opteron 250 CPUs (2.4 GHz, single core)
  - 1 GB RAM per CPU
  - 2 ethernet ports (1 GBit)
  - 160 GB hard drive

TETHYS – Hardware Specifications (cont.)

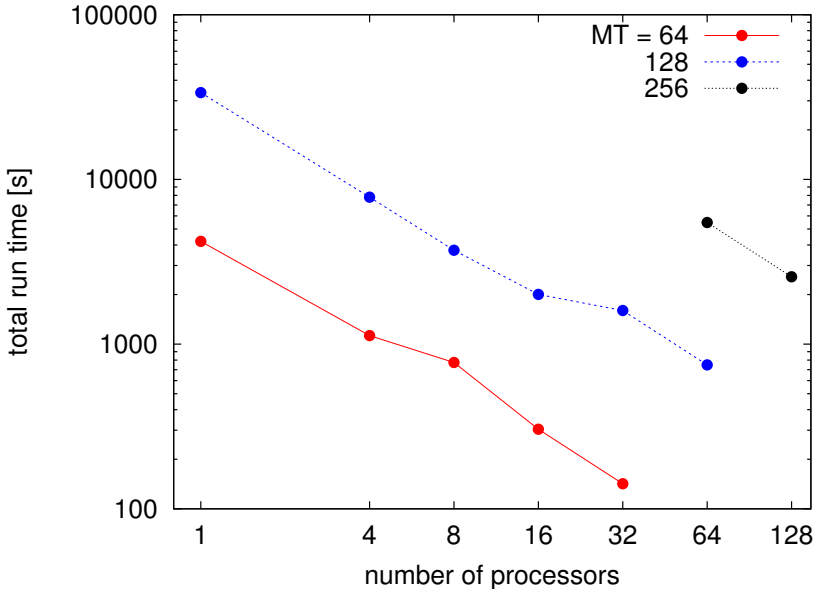
- 5 network switches
  - 1 HP ProCurve 6400cl (6 x 10 GBit ports) – cluster core switch
  - 4 HP ProCurve 3400cl (24 x 1 GBit ports) – cluster node switch
- operation system
  - Debian GNU/Linux 3.1 Sarge (AMD64 port)
  - FAI for installation of compute nodes
  - 8 TB parallel filesystem – PVFS2
  - queueing system – SGE or PBS Pro?

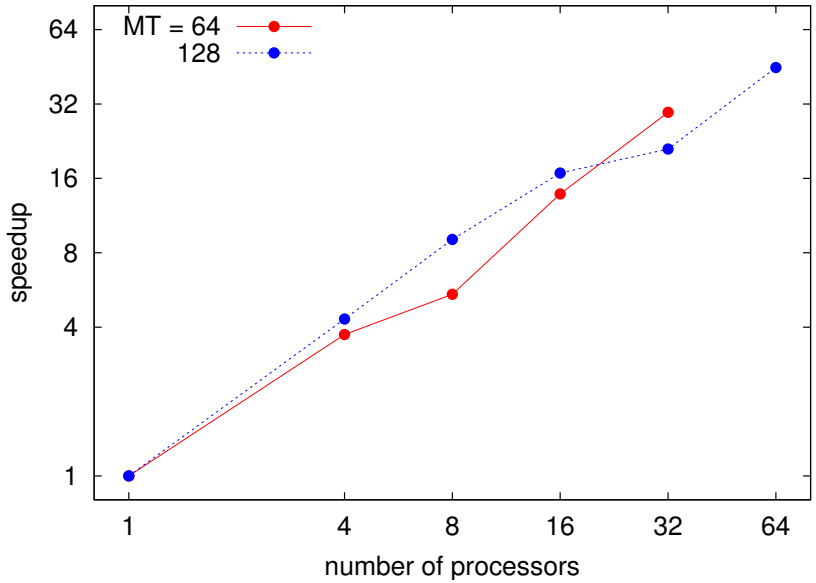
TETHYS – Cluster Topology



TERRA – Performance Tests

Good scalability shown by initial benchmark results.

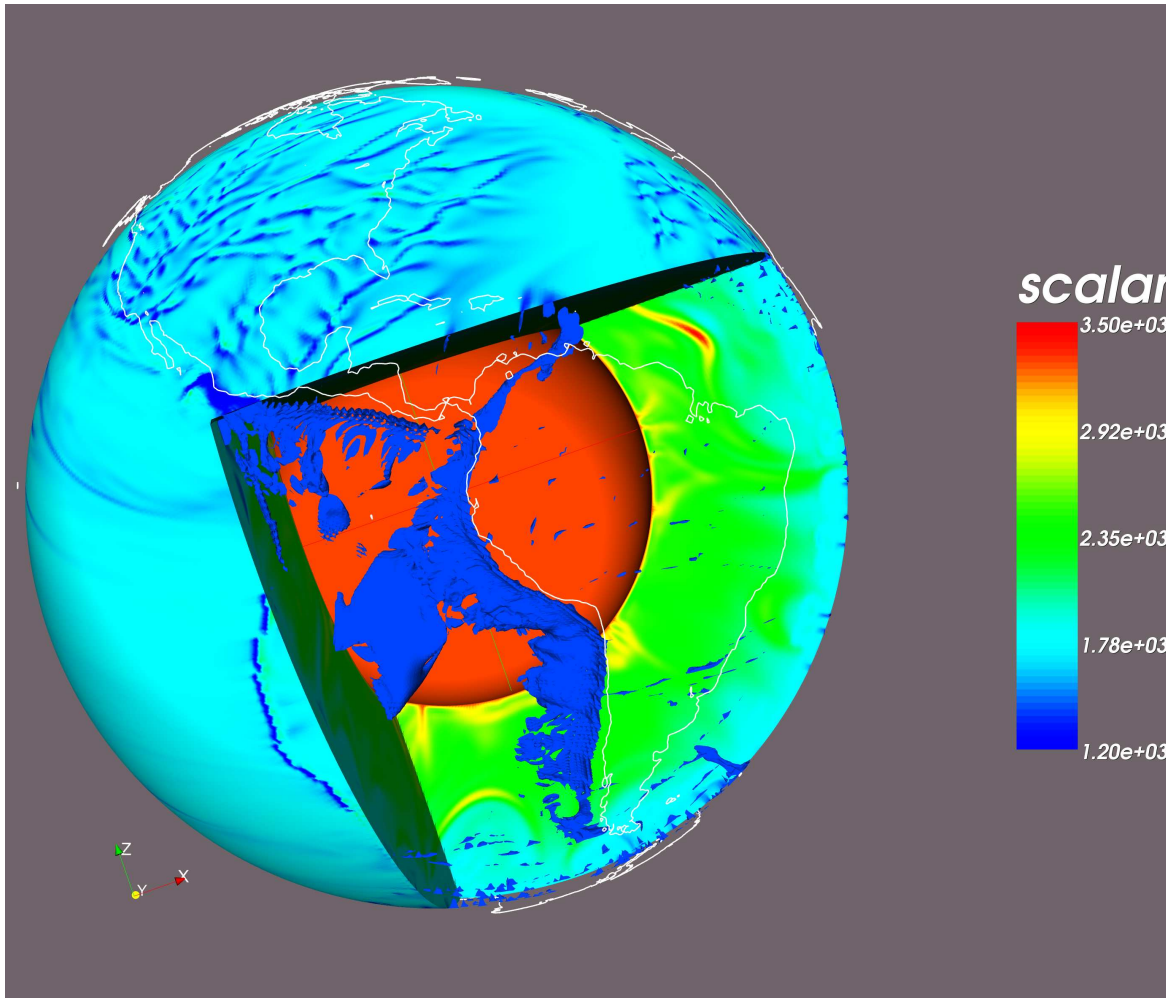




(MT = 64, 128, 256 ↔ resolution 100, 50, 25 km ↔ 1, 10, 85 mio. grid points)

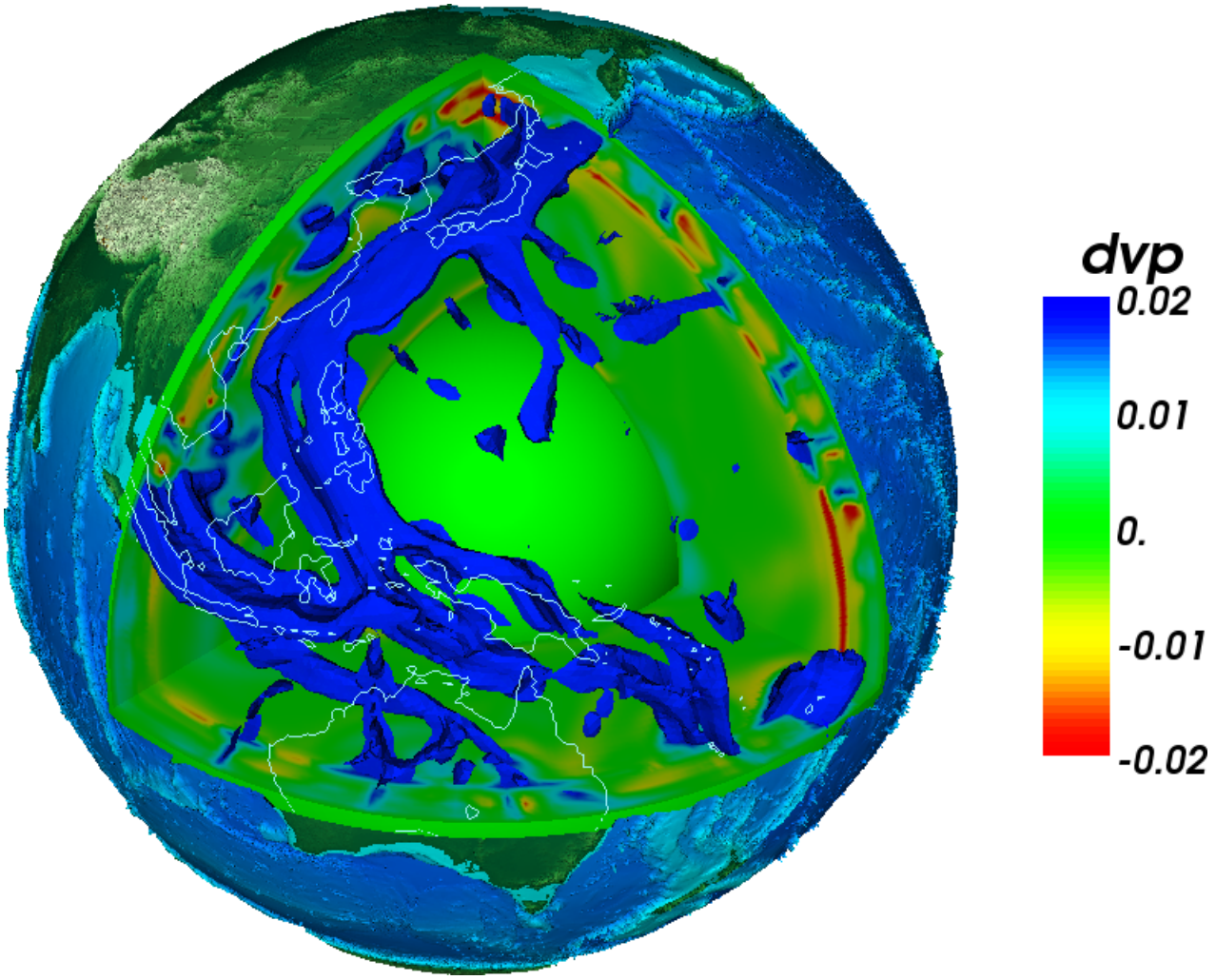
for 500 time-steps we obtain a run-time of 2002 s (mt=128 case on 16 processes) and 2564 s (mt=256 case on 128 processes), which both lead to the same workload per process

TERRA



circulation in earth interior, temperature denoted by colour (cold=blue, red=hot), isosurface shows subducting plate beneath South America

SPECFEM3D



synthetic seismic velocity structure predicted from mantle circulation modelling (blue=fast, red=slow)

Conclusions & Outlook

- large-scale geophysical modelling cluster is now feasible – TETHYS
- departmental supercomputer can efficiently perform geosciences simulations
- calculations for global earth modelling studies
- aggregate system performance of 200 Gflops
- cost-efficient Beowulf clusters viable part of modelling infrastructure in geosciences

Acknowledgements

- German Ministry of Education and Research (BMBF)
- Free State of Bavaria
- Microstaxx GmbH
- High-Performance Group of Fujitsu-Siemens Computers

APPENDIX

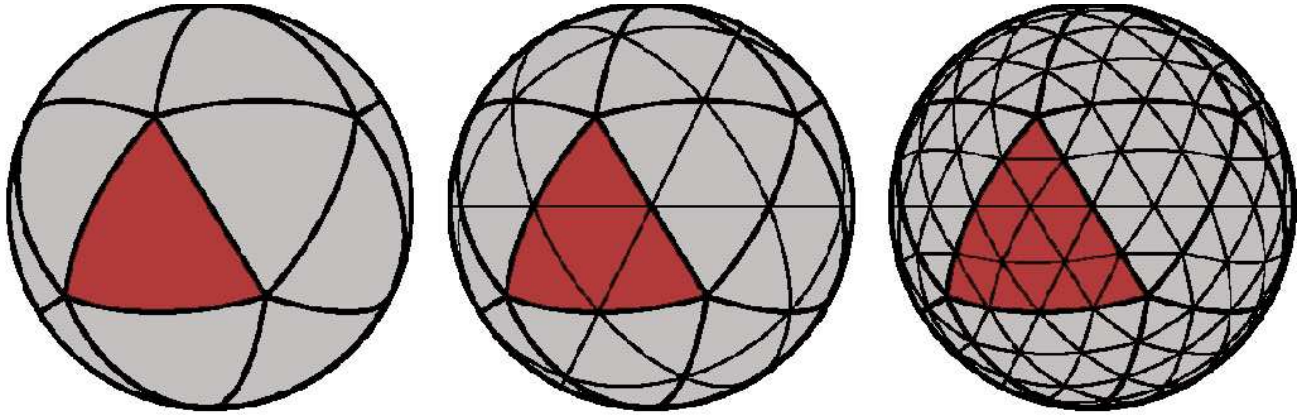
Space Discretisation I

(follows Baumgardner & Frederickson, 1985)

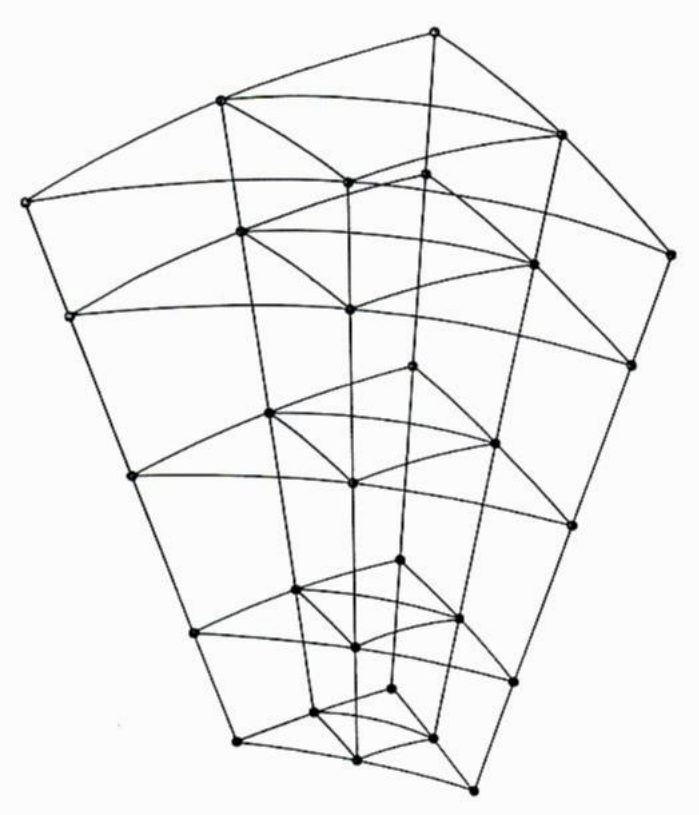
- a surface grid is generated by mapping an icosahedron onto the sphere
  $\longrightarrow$ spherical triangles

- surface grid is successively refined $\longrightarrow$ factor four in each step

- surface grid is radially extended down to the mantle-core boundary



Space Discretisation II



(figure: Karpik et al. 1991)

- the grid generation process naturally leads to a nested grid hierarchy
- characteristics of surface grid:

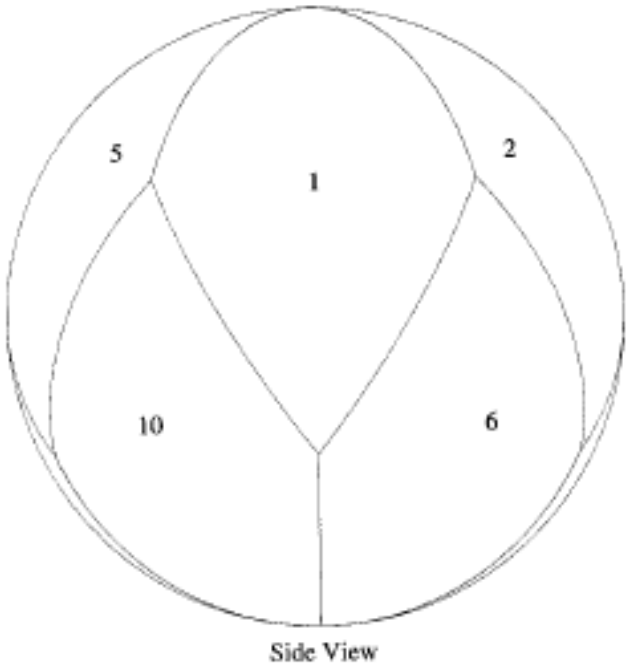
# triangles	$20\,n^2$
# nodes	$10\,n^2+2$
# arcs	$30\,n^2$

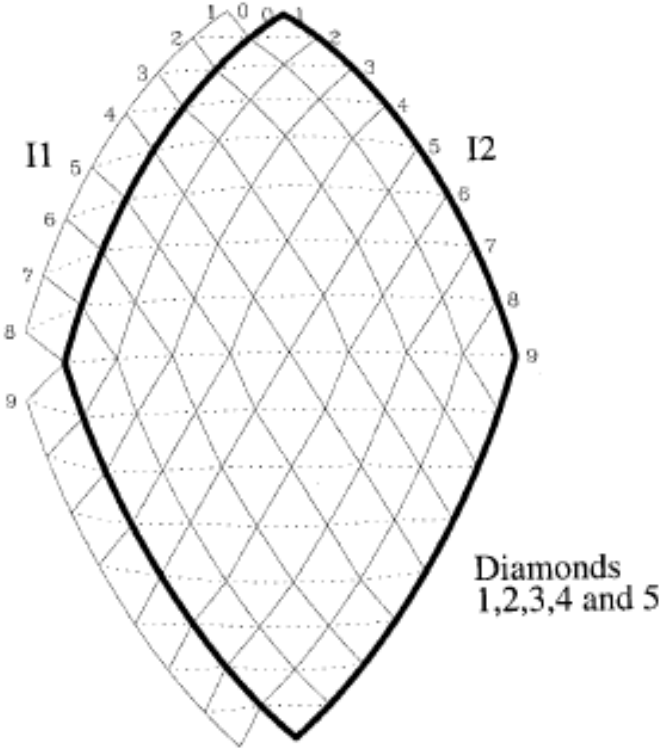
(with $n=2^k$, k the level of refinement level)

- radial resolution typically used is $n/2$ layers

Space Discretisation III

pair-wise combination of two base-triangles leads to
10 diamonds with a logically rectangular grid





(figures: Stuhne et al. 1996)

Finite-Element Spaces

For discretisation we need Finite-Element spaces

$V_0^h \subset H_0^1(\Omega)$ and $S_0^h \subset L_0^2(\Omega)$

- TERRA uses the same type of Ansatz-functions for both (scalar) velocity components (u_x, u_y and u_z) and the pressure p.
- Baumgardner & Frederickson (1986) generalised piecewise linear Finite Elements for spherical triangles
- Extension to 3D by cross-product with 1D piecewise linear Finite-Elements

Linear System

The mixed Finite-Element discretisation leads to a linear system

$$\begin{pmatrix} A & -G \\ -G^T & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$$

with

- the discrete viscous operator A
- the discrete gradient operator G
- the discrete right-hand side $f \approx \varrho_0 \alpha \left(T_0 - T\right) g$

$\longrightarrow$ saddle point problem

Schur Complement

consider a single block-GauSS step applied to the problem

$$\begin{pmatrix} A & -G \\ -G^T & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$$

$$\wr$$

$$\begin{pmatrix} A & -G \\ 0 & -G^T A^{-1} G \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ G^T A^{-1} f \end{pmatrix}$$

for A s.p.d. the matrix $S := G^T A^{-1} G$ is (at least) s.p.s.d

Pressure-Correction Type Scheme

Basically this can be seen as CG applied to solve $Sp = -G^T A^{-1} f$ with some modifications:

Setup:

- Given initial guess $p^{(0)}$ solve $Au^{(0)} - Gp^{(0)} = f$ for $u^{(0)}$ (momentum eqn.)
- Initial residual $r^{(0)} = G^T A^{-1} f + G^T A^{-1} G p^{(0)} = G^T u^{(0)}$

Loop:

- Given search direction $s^{(i)}$ solve $Av^{(i)} = Gs^{(i)}$ for $v^{(i)}$
- Use $v^{(i)}$
  - as auxilliary vector in CG
  - to perform update $u^{(i)} = u^{(i-1)} + \alpha v^{(i)}$

Multigrid

We solve the problem $Av^{(i)} = Gs^{(i)}$ iteratively using multigrid.

Characteristics:

- V-cycle
- Jacobi-type radial-line smoothing
- Operator-dependent transfers for tensor-valued stencils (Yang & Baumgardner, 2000)
- Galerkin coarse grid approximation
- Coarse grid agglomeration (for parallel MG)