# The Petaflops Challenge for NWP and Climate Research

**Geerd-Rüdiger Hoffmann** [1]

**Ulrich Trottenberg** [2]

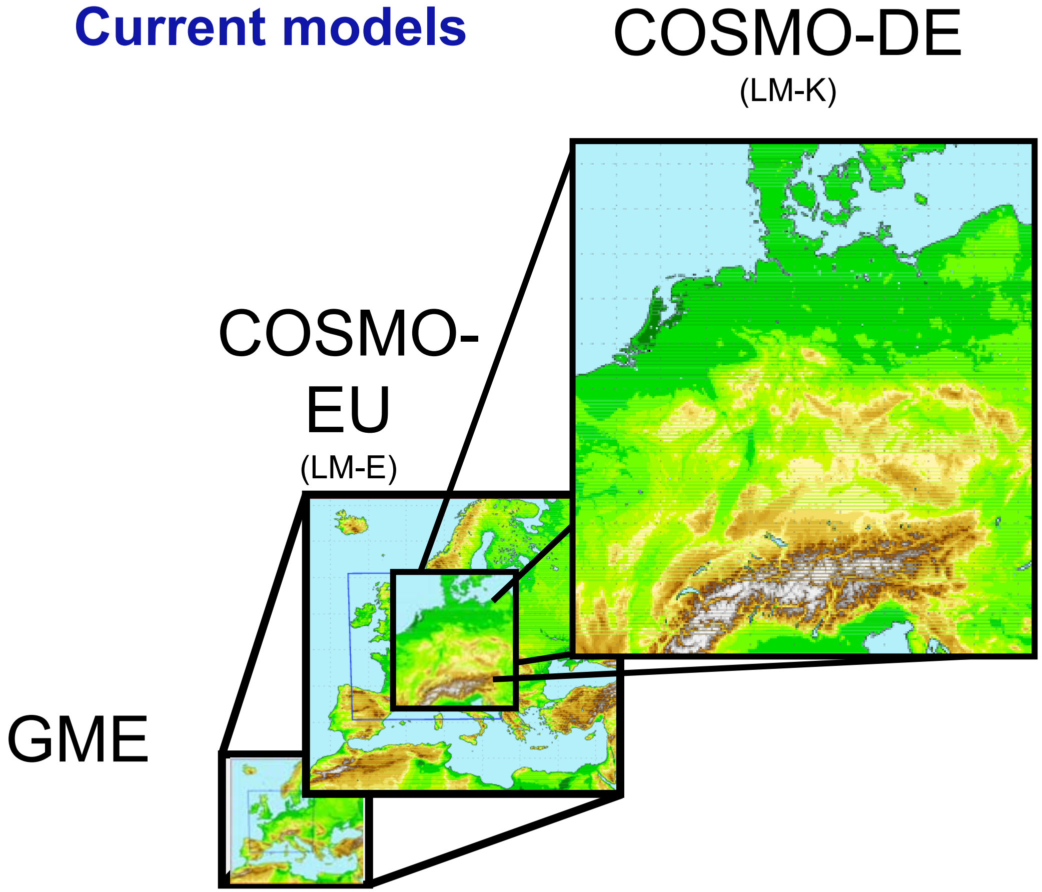
[1] Deutscher Wetterdienst (DWD), Offenbach, Germany
[2] Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, Sankt Augustin, Germany

# New challenges

- Meteorological requirements
- Computing requirements
- Archives
- Petaflops challenges
- Algorithms
- Outlook

# Meteorological requirements

## Current models

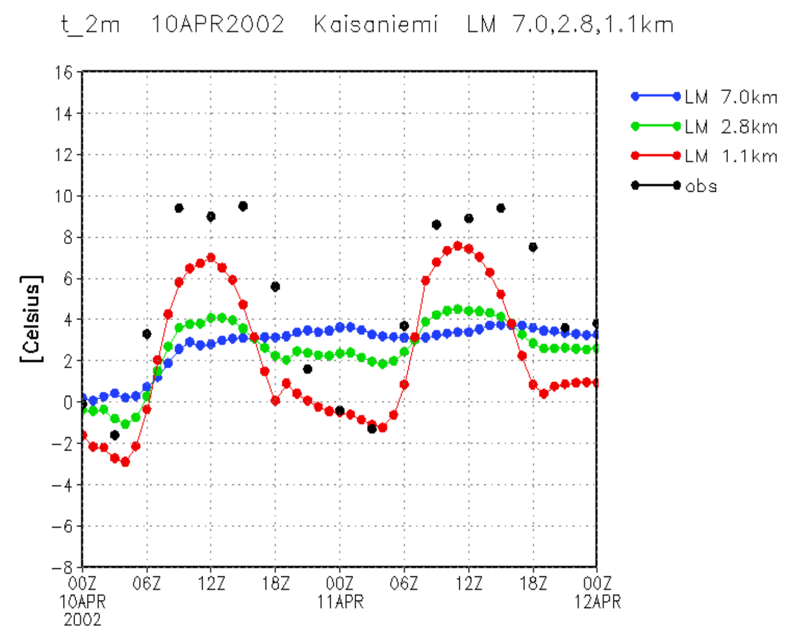


- convection-resolving
- Model Configuration
  - Grid Spacing: $\Delta x \approx 2.8$ km
  - 50 vertical levels
  - $\Delta t = 25$ s
- Boundary conditions
  - Interpolated COSMO-EU forecasts
- Data assimilation
  - Same as COSMO-EU
  - Including Latent Heat Nudging for Radar Reflectivities
- Cloud microphysics include graupel, snow and rain
- very short-range forecasts up to 21 hours
- operational at DWD since April 2007 (21 hours forecast, started every 3 hours)

# **Downscaling** COSMO-EU: 7 - 2.8 -1.1km

Example: 2m temperature

## EU-projectFUMAPEX



## Evaluation with measurements for increased resolution:

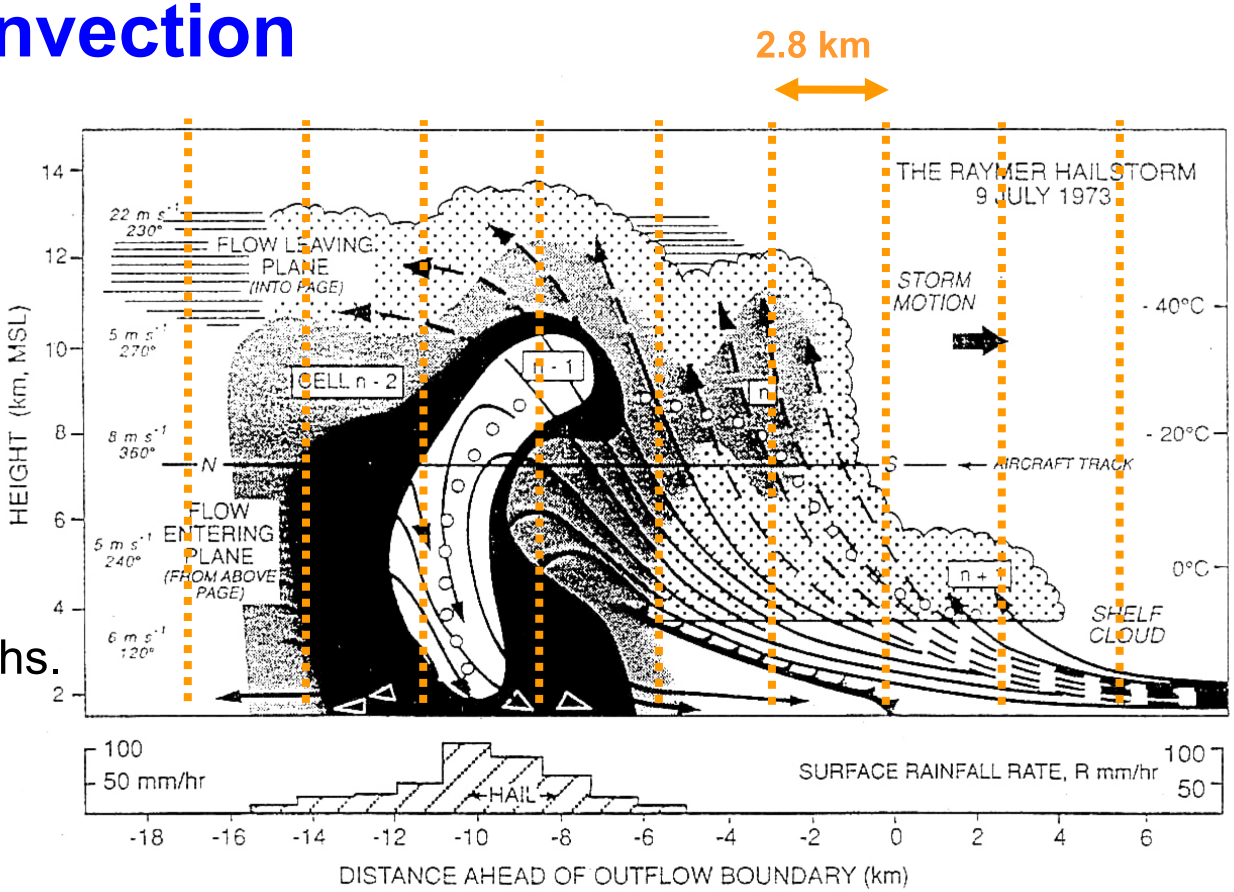
- improved local wind systems in mountains and along coast
- improvement through better land-sea-mask near coast (soil and roughness parameters)
- overall improvement
- nesessity of scale-adapted parametrisations!

© Fay, DWD

Fraunhofer Institute Algorithms and Scientific Computing

# Deep moist convection

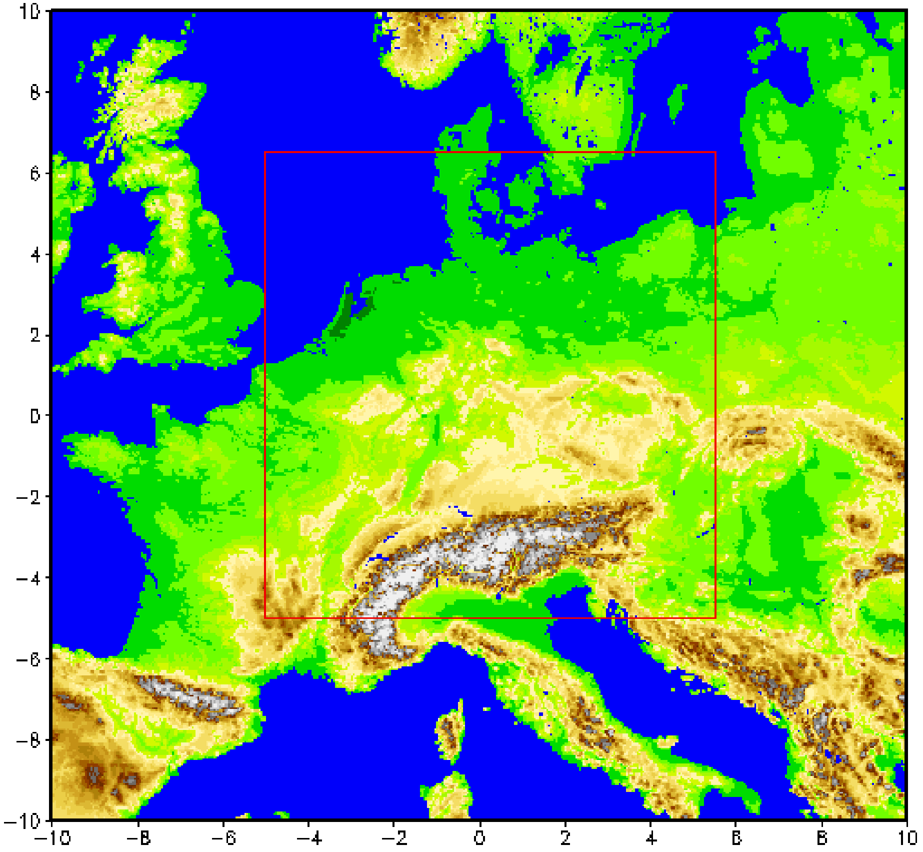
Schematic model from a Colorado storm case study (Raymer Hailstorm)



Effective resolution: approx. 5 to 7 grid widths.

from: R. A. Houze, Jr.: Cloud Dynamics
International Geophysics Series Vol. 53

# Area of future high resolution limited area model



- 2000×2000 × 100 grid points at a distance of one km
- up to 80 3D variables
- ~256 GB variables per time step

© Majewski

# Current computing requirements

**NEC SX-8R**

**7 nodes**

**56 processors**

**1.97 TFlop/s peak speed**

| | GME | COSMO-EU | COSMO-DE |
|---|---|---|---|
| grid spacing (km) | 40 | 7 | 2.8 |
| number of layers | 40 | 40 | 50 |
| number of grid points (Mill.) | 15.0 | 17.5 | 9.7 |
| forecast range (h) | 174 | 78 | 18 |
| time step (s) | 133 | 40 | 30 |
| number of time steps | 4698 | 7020 | 2160 |
| Flop per GP and time step | 4500 | 6000 | 9500 |
| wallclock time (min) | 112 | 62 | 20 |
| Flop per forecast ($10^{12}$) | 317 | 737 | 199 |
| computation speed (GFlop/s) | 47 | 198 | 166 |
| number of processors used | 16 | 32 | 32 |

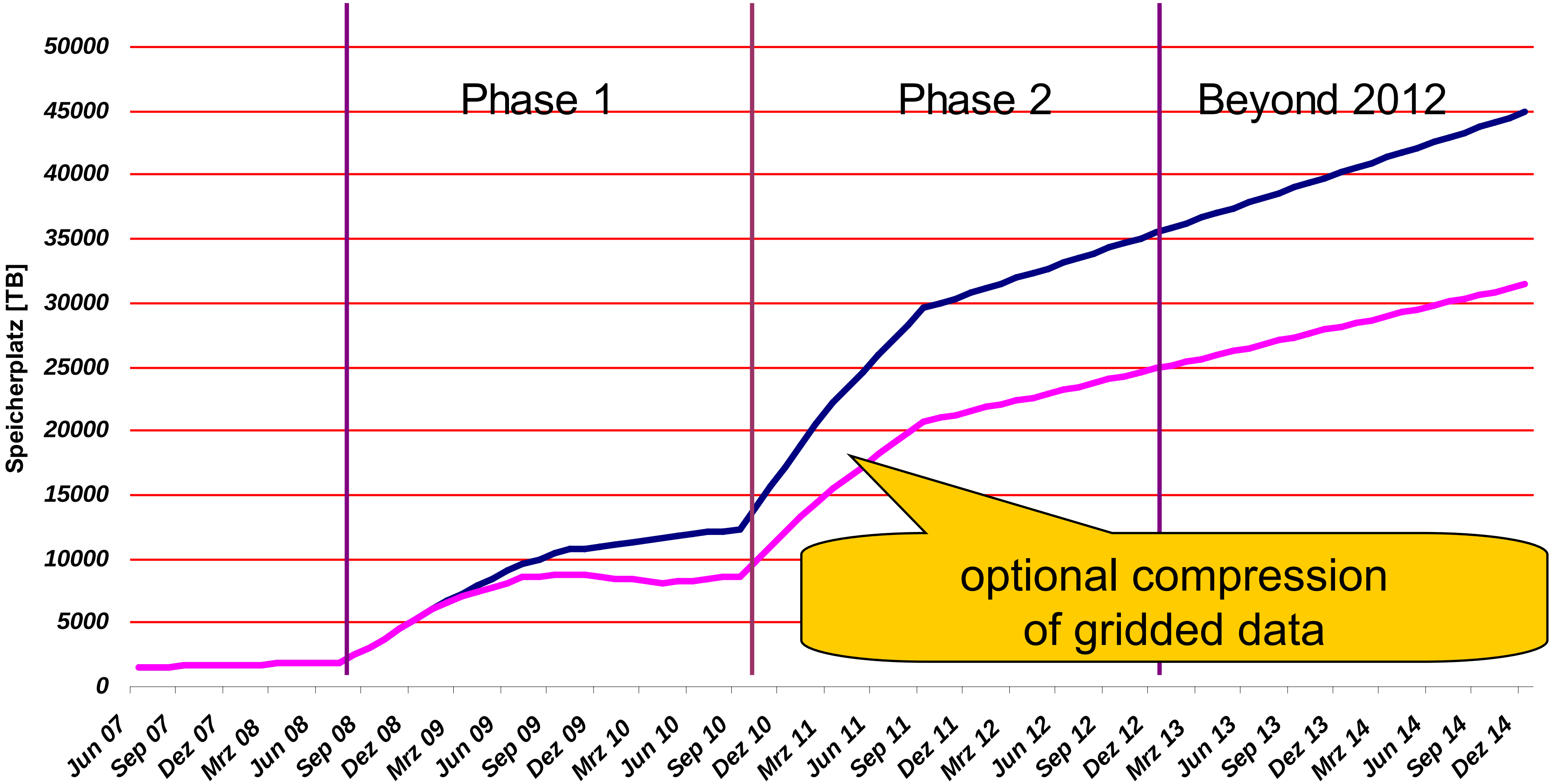
Flop: Floating point operation

© Majewski, DWD

# Future computing requirements

- Future limited area models for Germany will have a horizontal resolution of about 1 km, about 100 layers and will have about 400 million grid-points
- Enhanced physical parametrisation, including Aerosols, will result in up to 80 3D variables per grid point
- The resulting computing capacity for a weather forecast lies in the order of 90 TFlops/s sustained performance
- At present, the sustained performance of a scalar system with O(1000) processors is about 10% of peak performance for NWP
- Assuming that the efficiency in the increase of the number of processors necessary is 90%, the system will have to peak at about 1 PFlops/s
- If an EPS system with half the resolution is to be implemented, the system will have to peak at about 5 PFlops/s

# Mass Storage 2008-2012



optional compression
of gridded data

# GRIBzip – Compression for GRIB Data

- GRIBzip features
  - Loss-free compression of GRIB data
  - Specialized for GRIB data fields (2D and 3D)
  - Fast uncompression (40 MB/s)
  - Licensed SW, free read/uncompress (like pdf)
- Benefits
  - 2-3 x reduction in data volume, saves cost for storage media
    - DWD saved 140 TB space on magnetic tape in 2007/08, i.e. 30.000 €
  - Less data transfer, faster data exchange
  - Proven technology, in operational use at DWD since October 2007
  - Supported software, continuous development

# Petaflops challenges

## Hardware around 2010: scenario 1

Multi-core (≤ 12) processors („sockets“)

- Clock frequency in the order of 2 – 4 GHz (not more)
- At least 4 parallel floating point operations per clock
- Maximum performance per socket up to 192 GFlops/s
- 8 sockets per board, i.e. 1.5 TFlops/s per board
- Memora bandwidth only scales up to 4 GB/s per core, i.e. about ¼ B per Flop
- Memory size up to about 2 GB per core
- In order to achieve > 5 PFlops/s, in total about 316,800 cores are needed, i.e. 26,400 sockets in 3,300 boards.
- Power consumption would be around 20 MW (current technology?)

# Petaflops challenges

## Hardware around 2010: scenario 2

Heterogeneous systems consisting of variety of specialized processors

- Roadrunner with AMD and IBM Cell processors
- Cray XT5$_h$ with AMD and Cray vector processors
- Japanese Petaflop project with Fujitsu RISC and NEC vector processors
- PRACE prototype at BSC with IBM Power6 and IBM Cell processors
- …

# Petaflops challenges

## Hardware around 2010: scenario 3

Specialized processors

- Processors with SIMD instruction set
- Vector processors like NEC SX-9
- IBM BlueGene
- Nvidia nForce
- Broadcom HT
- FPA‘s
- …

# Petaflops challenges

## Software: Scenario 1

- Parallelisation of models across more then 60,000 cores with parallel efficieny of at least 90%
  - Depending on algorithms used the interconnect requirements become extreme in terms of latency and bandwidth
- Achieve at least 10% of peak performance with only ¼ B per Flop memory bandwidth
  - The choice of algorithms becomes crucial
- Parallel I/O with total bandwidth of about 20 GB/S average, assuming 10 s model time step with write-ups at every 15 min. model time
  - Depending on the I/O strategy, the interconnect features become essential

# Petaflops challenges

## Software: Scenario 2 and 3

- There is no relevant experience with the new computer architectures in terms of reliability and useability
- The programs will have to be partially re-written to make optimal use of the specialized processors, including possibly applying different algorithms
- New programming languages might have to be used
- There might be a lack of relevant programming experience depending on the different processor types

# Petaflops challenges

## Operations

- The cost of the systems might be prohibitive, except for very specialized computing centres, e.g. DOE or DOD installations
- Necessary infrastructure for the new systems might not be readily available: electricity suplly, cooling, space …
- Operating systems may not scale to the large number of processors, e.g. jitter
- The MTBF of the size of systems to be used may be smaller than the run-time of the individual jobs

# Petaflops challenges

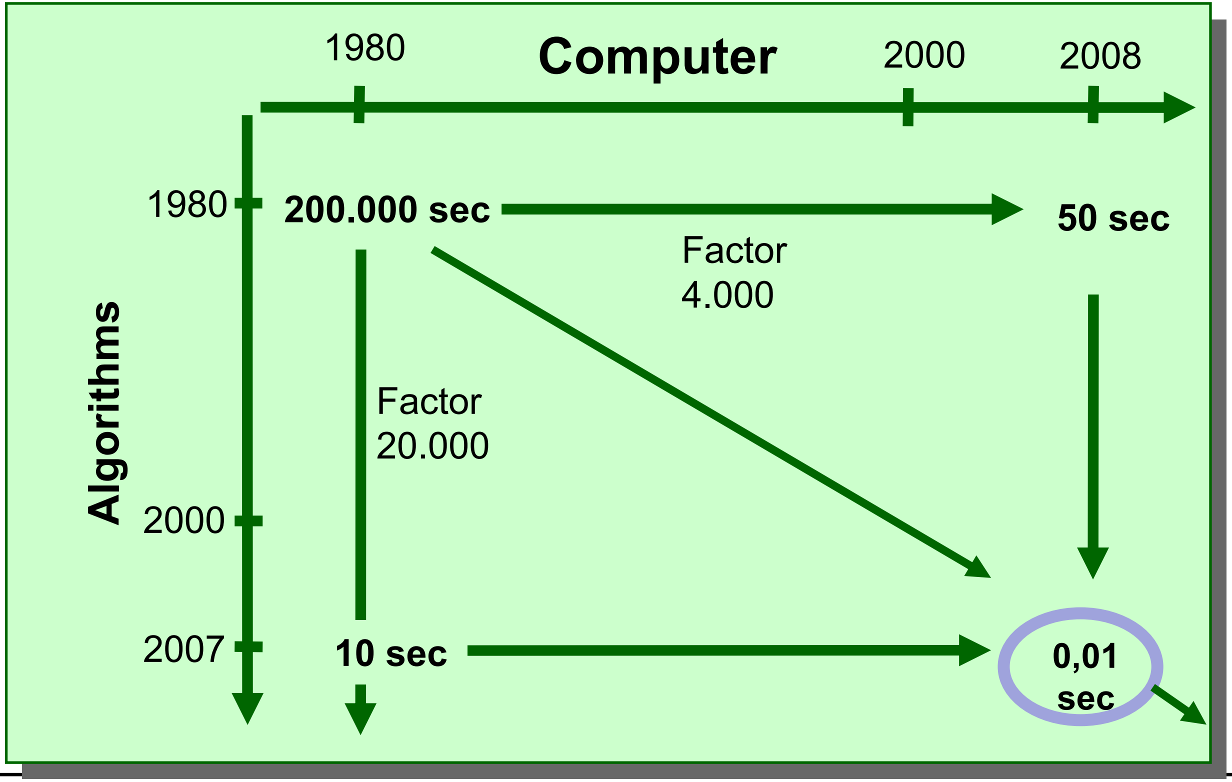
## Summary

In order to answer some of the questions, it is mandatory that systems of relevant performance are made widely available as soon as possible for application testing and tuning, like the PRACE prototypes..

# Algorithms

# Algorithms: On The Road to Petaflop Systems

## Algorithms versus hardware



Helmholtz like equation

to be solved in each time step

# Modelling and Computation

The Phanomenon (weather, climate,…)

⬇ Modelling

The Mathematical model

⬇ Discretization (discretization parameter h)

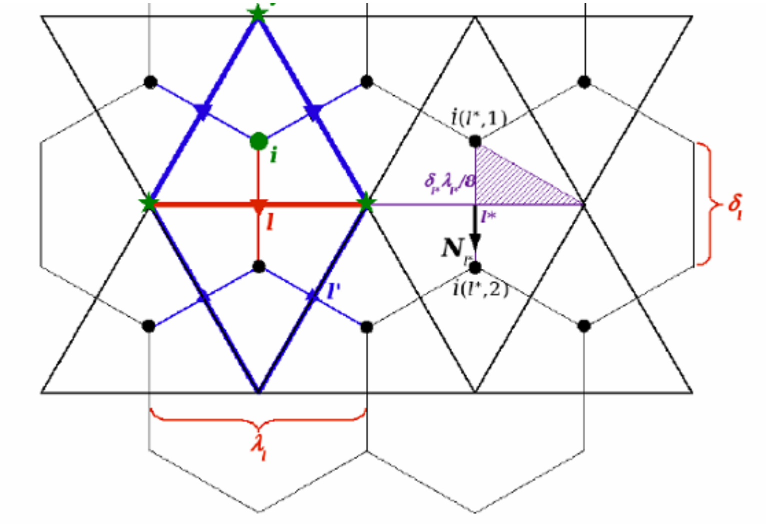
The Discrete mathematical model

⬇ Design of algorithm

The software system

⬇ Data, implementation

Computation, visualisation



$$\mathrm{div}(\vec{\chi})_i = \frac{1}{A_i} \sum_{l \in \mathcal{E}(i)} \chi_l \vec{N}_l \cdot \vec{n}_{i,l}\, \lambda_l$$

$$\mathrm{rot}(\vec{\chi})_v = \frac{1}{A_v} \sum_{l \in \mathcal{E}(v)} \chi_l \vec{N}_l \cdot \vec{t}_{v,l}\, \delta_l$$

$$\left(\nabla \psi \cdot \vec{N}\right)_l = \frac{\psi_{i(l.2)} - \psi_{i(l.1)}}{\delta_l}$$

# Prediction Uncertainty – Sources of Errors

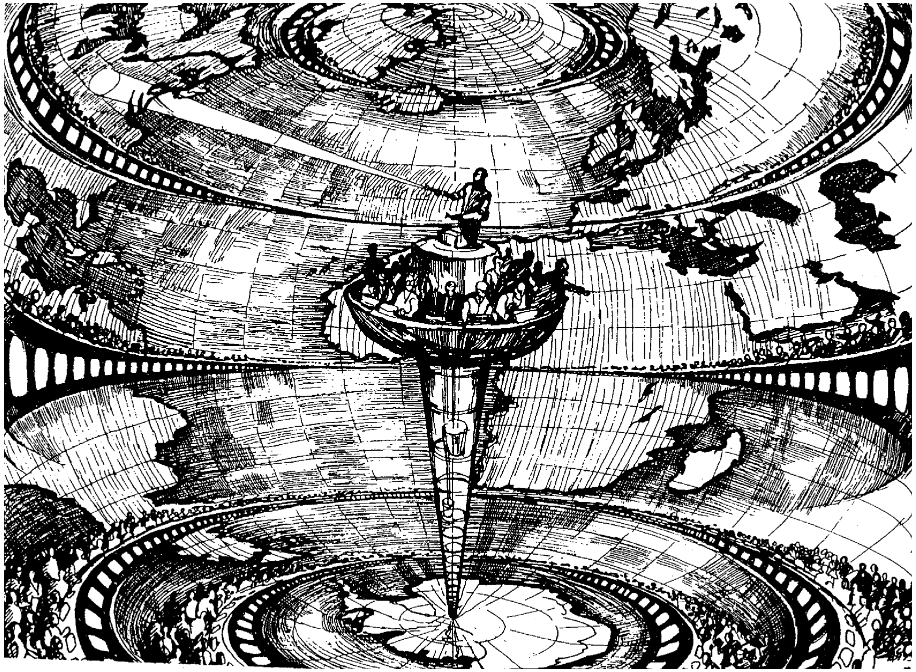
- Model error (in charge: meterology)
- **Discretisation error** (in charge: mathematics)
- Data error (in charge: technology)

- Chaos instability (in charge: reality)

Isolate, get information about the discretization errors by studying $h \to 0$

# Algorithmic Challenges: Parallelism

Efficient use of 12 x 26400 = 316.800 cores (2012) or more



**partitioning**

**load balancing**

**local communication**

**global communication**

**fault tolerance**

---

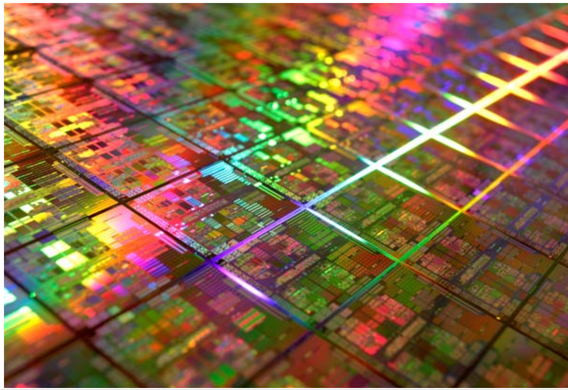
**ensemble calculations …**

# Architectures for the Next Supercomputers

**General purpose: Multicore/Manycore Processors**

Special purpose: Make use of heterogeneons components, e.g.:

- GPGPU – “General Purpose” Graphics Processing Units
- FPGA Accelerators (Field Programmable Gate Arrays)
- Co-Processors
- Cell Processors



# Programming Environments

- **OpenMP**
- **MPI**

- OS native Threads (pThread, MS Windows Threads, …)
- Remote DMA Libraries (shmem, …)

- nVidia's CUDA
- AMD's Brook+

- OpenCL

- Rapidmind
- Cilk+

# Hardware/Software Challenges

**Addressing a Zoo of Hardware Architectures: Complications**

Limited Main Memory Bandwidth

Many cores share the same physical memory

Limited Bus System bandwidth

Communication with Coprocessors costly

Different levels of Latency

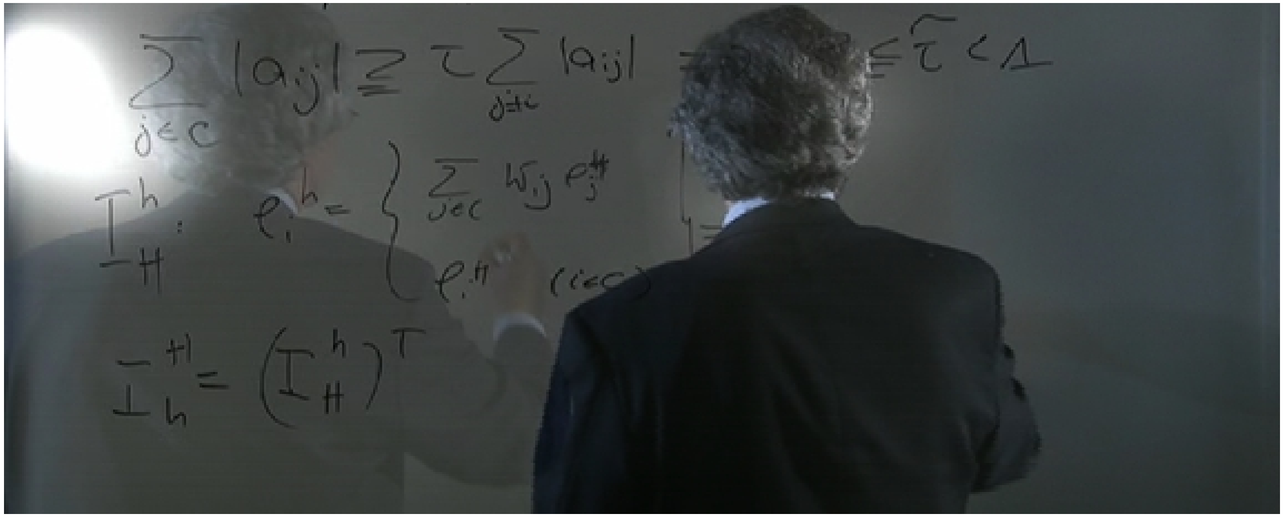
Communication synchronization difficult or costly (consider All-to-All)

Load Balancing among normal Processing or Vector units

Consider Multicore CPUs, many Sockets SMPs, GPGPU systems

# Project Proposal PeAKliM: Goals

- Meeting the increasing demand of compute power
- Answering, what kind of architecture is best suited for meteorology
- Prepare algorithms already now for future systems



# PeAKliM: Partners

- Max Planck Institute for Meteorology (MPI-Met)

Max-Planck-Institut für Meteorologie
Max Planck Institute for Meteorology

- Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

SCAI
Fraunhofer Institut
Algorithmen und Wissenschaftliches Rechnen

- German Climate Compute Center (DKRZ)

DKRZ

- Deutscher Wetterdienst (DWD)

Deutscher Wetterdienst
DWD

# Development of models

## 3D-atmosperic model ICON

Non-hydrostatic model with static local zooming

Hybrid parallelisation with MPI and OpenMP

Operational use as NWP model planned at DWD

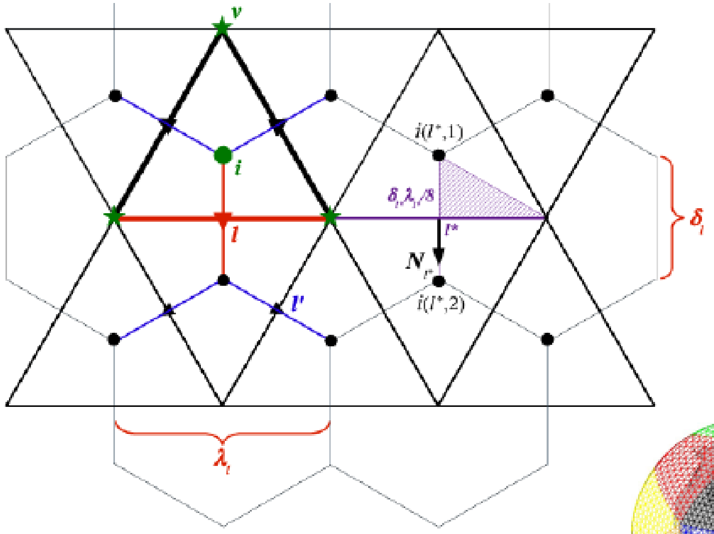
Atmosperic part of an earth system model to be used at MPI-M

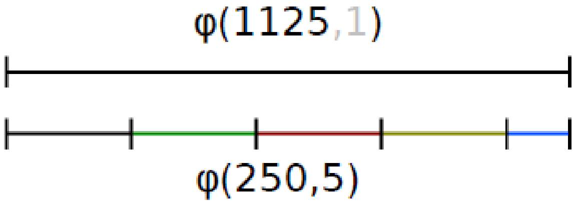
Modular approach: depending on application area different physical components (radiation, cloud micro physics, convection,…

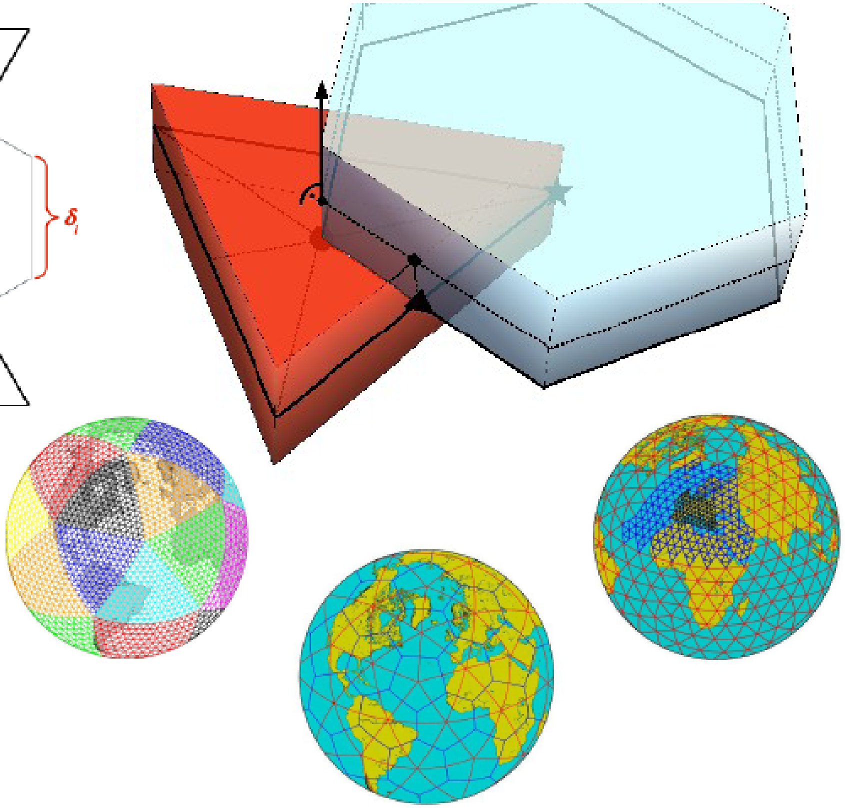
Programming to start in 2009

© Majewsk, DWD

# ICON Grid

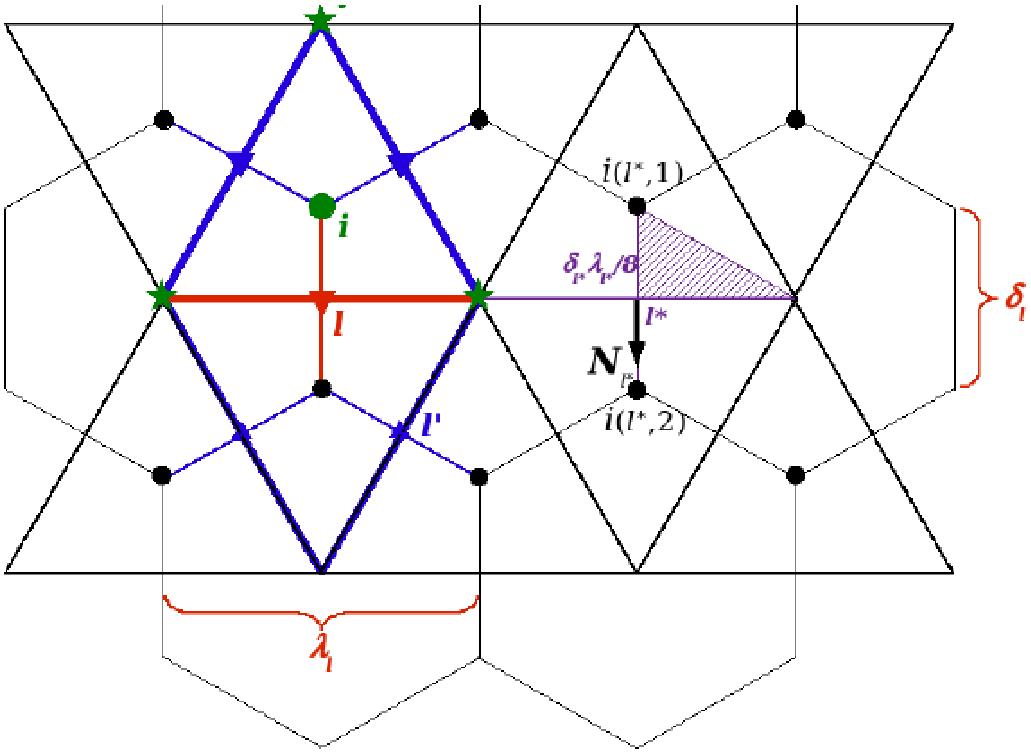






© Förstner

# ICON operators



© Förstner

$$\mathrm{div}(\vec{\chi})_i = \frac{1}{A_i}\sum_{l\in\mathcal{E}(i)} \chi_l \vec{N}_l \cdot \vec{n}_{i,l}\, \lambda_l$$

$$\mathrm{rot}(\vec{\chi})_v = \frac{1}{A_v}\sum_{l\in\mathcal{E}(v)} \chi_l \vec{N}_l \cdot \vec{t}_{v,l}\, \delta_l$$

$$\left(\nabla\psi\cdot\vec{N}\right)_l = \frac{\psi_{i(l,2)} - \psi_{i(l,1)}}{\delta_l}$$

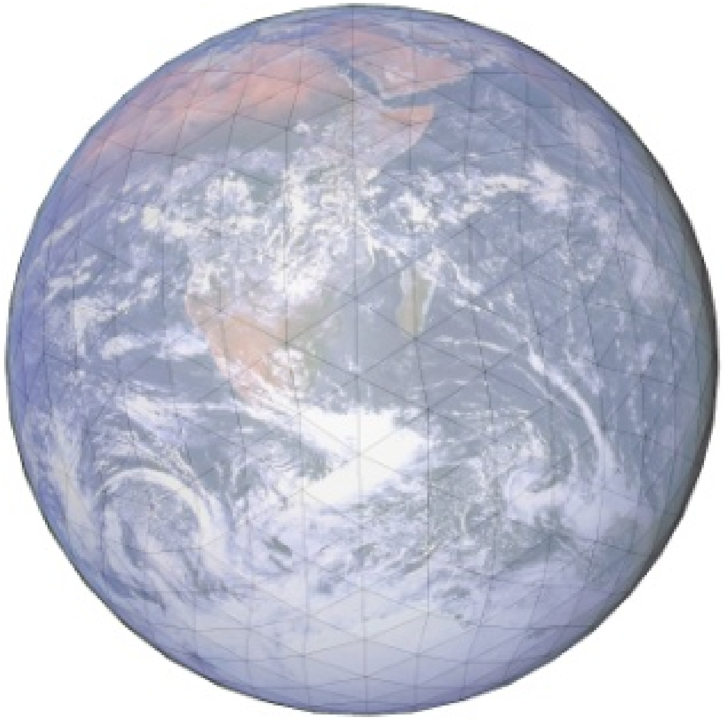
# Benchmark Kernels

- Identification of benchmark kernels relevant for Petaflop systems, addressing
  - Kernels from ICON and COSMO
  - Memory bandwidth
  - IO bandwidth + latency
  - Communication bandwidth + latency
- Optimization of benchmark kernels
  - Focusing on Hardware independence
- Performance measurements
  - Including Hardware dependent optimization

# Algorithmic + Numerical Challenges

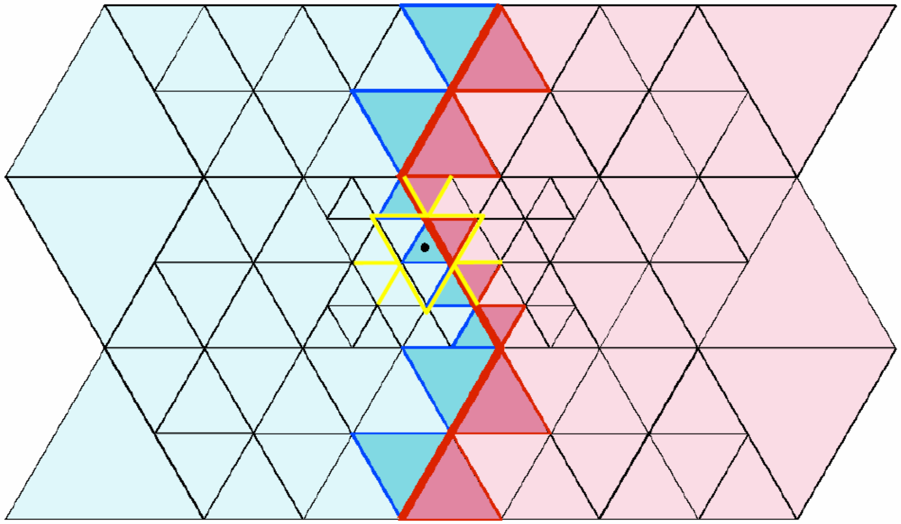
- Choice of algorithms for Petaflop Architectures
- High number of processors requires highly scalable algorithms
- Main issue: Solvers (horizontally explicit, vertically implicit or 3D implicit ?)
- **Multigrid techniques ?**

# Algorithmic Challenges: Solvers



- Solvers for (linear) systems of equations
  4 M x 100 gridpoints
- SOR, Krylov, GM-Res, Multigrid, AMG, …

# Algorithmic Challenges: Communication overhead



**Local:**

Boundary-volume effect ☺
    for purely grid based approaches

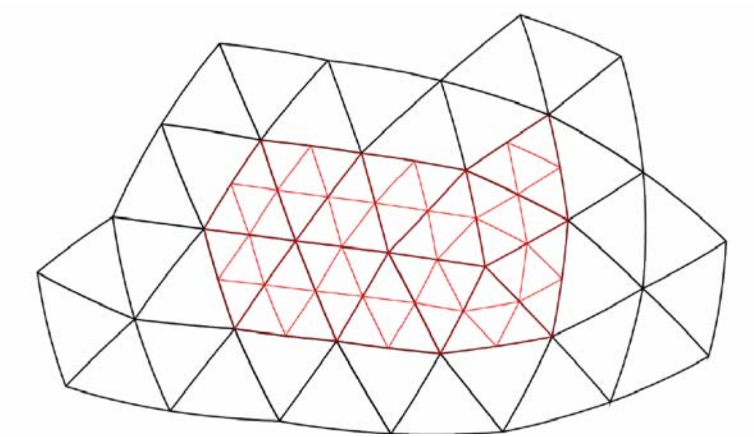
Granularity
    fine ~ 1.000 grid points / core ☺
    coarse ~ 10 grid points / core ☹

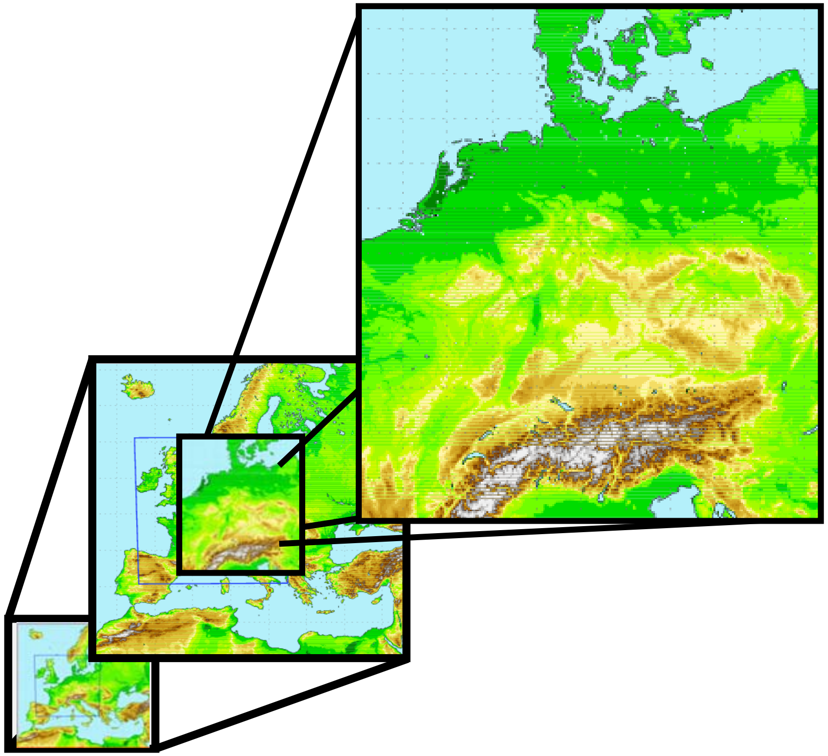
**Global:**

Global communication ☹
    (for FFT type algorithmic components)

# Algorithmic Challenge: Local Refinement

- Requires redistribution





- coarse → fine → coarse

?

Multigrid

# Algorithmic Challenges: Load balancing

**Weather (physics, clouds, etc.)**

↓

**will lead to load inbalance,**

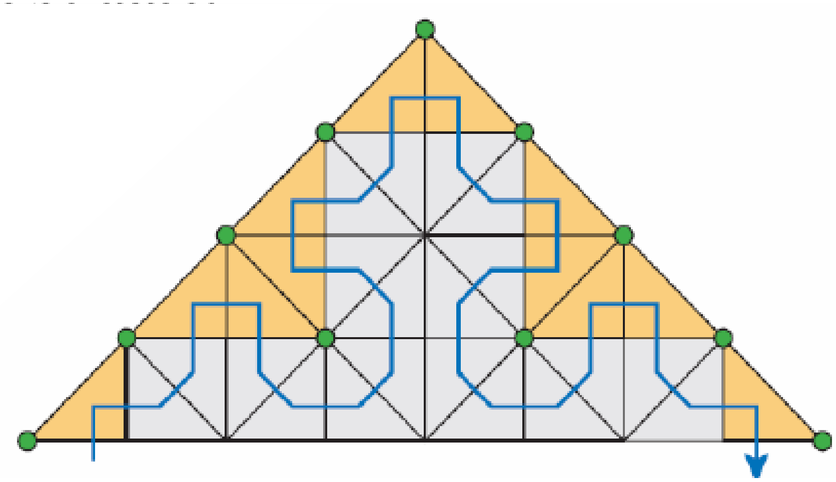
**even if volume-boundary effect is maintained.**

# Algorithmic Challenges: Load Balancing

- “Domain decomposition” by multilevel partitioning algorithms
  - Challenge: choice of algorithm with low imbalance at runtime
- Detection of load imbalance at runtime
  - Criteria for too much imbalance
  - Computation of new redistribution
- New redistribution techniques, e.g. space filling curves

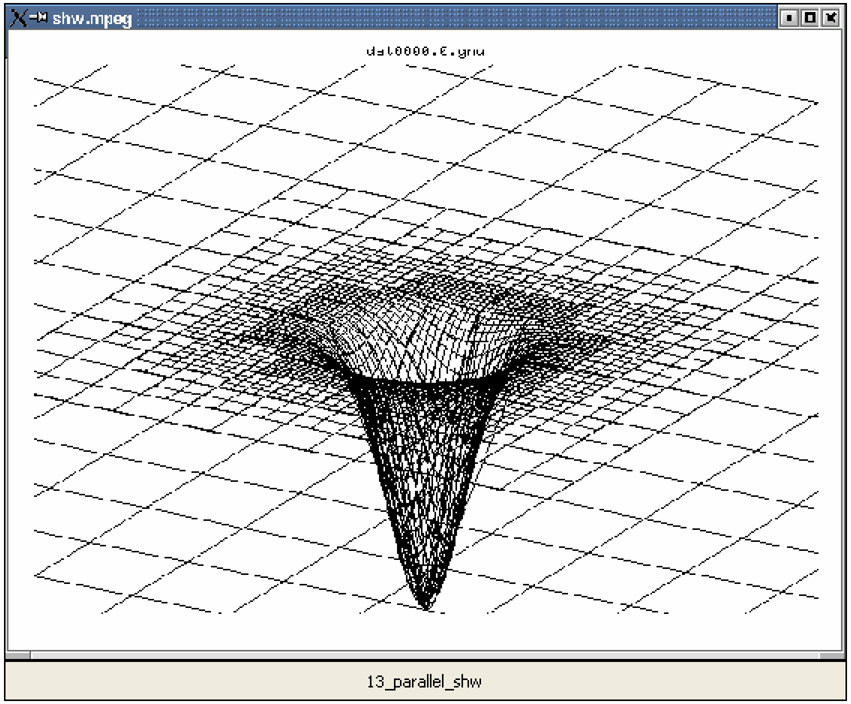
---

- Fault Tolerance



# Algorithmic Vision: Dynamic Local Refinement

## Adapt the Grid to Weather Phenomena dynamically



# Algorithmic Vision (in Climate Prediction)

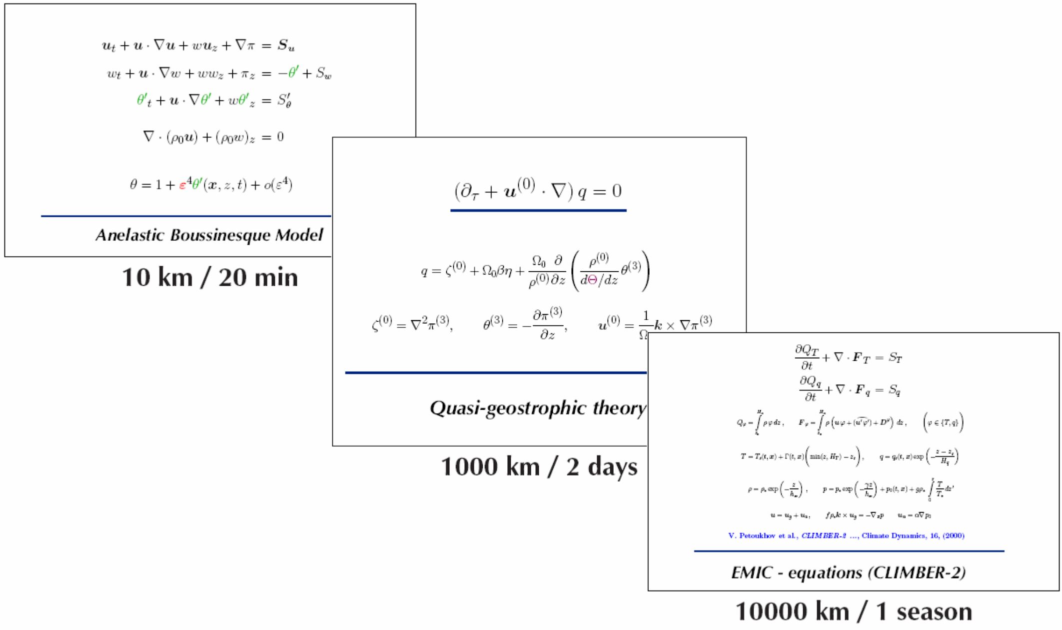
Get more information of discretisation error

by h → 0 studies:

Fix mathematical climate model
and let h tend to 0
(globally, locally, dynamically)

Identify the **numerical** error
(discretisation, grid resolution)

in the overall inaccuracy



$$u_t + u\cdot\nabla u + w u_z + \nabla\pi = S_u$$
$$w_t + u\cdot\nabla w + w w_z + \pi_z = -\theta' + S_w$$
$$\theta'_t + u\cdot\nabla\theta' + w\theta'_z = S'_\theta$$
$$\nabla\cdot(\rho_0 u) + (\rho_0 w)_z = 0$$
$$\theta = 1 + \varepsilon^4\theta'(x,z,t) + o(\varepsilon^4)$$

*Anelastic Boussinesque Model*

10 km / 20 min

$$(\partial_\tau + u^{(0)}\cdot\nabla)\,q = 0$$
$$q = \zeta^{(0)} + \Omega_0\beta\eta + \frac{\Omega_0}{\rho^{(0)}}\frac{\partial}{\partial z}\left(\frac{\rho^{(0)}}{d\Theta/dz}\theta^{(3)}\right)$$
$$\zeta^{(0)} = \nabla^2\pi^{(3)}, \qquad \theta^{(3)} = -\frac{\partial\pi^{(3)}}{\partial z}, \qquad u^{(0)} = \frac{1}{\Omega}k\times\nabla\pi^{(3)}$$

*Quasi-geostrophic theory*

1000 km / 2 days

$$\frac{\partial Q_T}{\partial t} + \nabla\cdot F_T = S_T$$
$$\frac{\partial Q_q}{\partial t} + \nabla\cdot F_q = S_q$$
$$Q_\varphi = \int_{z_s}^{H_a}\rho\varphi\,dz, \qquad F_\varphi = \int_{z_s}^{H_a}\rho\left(u\varphi + (\overline{u'\varphi'}) + D^\varphi\right)dz, \qquad \left(\varphi\in\{T,q\}\right)$$
$$T = T_s(t,x) + \Gamma(t,x)\Big(\min(z,H_T) - z_s\Big), \qquad q = q_s(t,x)\exp\left(-\frac{z-z_s}{H_q}\right)$$
$$\rho = \rho_s\exp\left(-\frac{z}{h_*}\right), \qquad p = p_s\exp\left(-\frac{\gamma z}{h_*}\right) + p_0(t,x) + g\rho_s\int_0^z\frac{T}{T_*}dz'$$
$$u = u_g + u_a, \qquad f\rho_s k\times u_g = -\nabla_z p \qquad u_a = \alpha\nabla p_0$$

V. Petoukhov et al., *CLIMBER-2* ..., Climate Dynamics, 16, (2000)

*EMIC - equations (CLIMBER-2)*

10000 km / 1 season

**Hierarchy of models**

Source R. Klein

# Outlook

- FhG / MPI project PeAKliM proposed for 2009
- Cooperation with European PRACE project
  - DWD is member of PROSPECT and Gauss Alliance e.V.
- Cooperation with American PERCS project
  - NCAR and DWD signed cooperation agreement in September 2008
- Cooperation with Japanese petaflop initiative
  - Visits by Dr. Watanabe (RIKEN) and Prof. Kobayashi (University Tohoku)