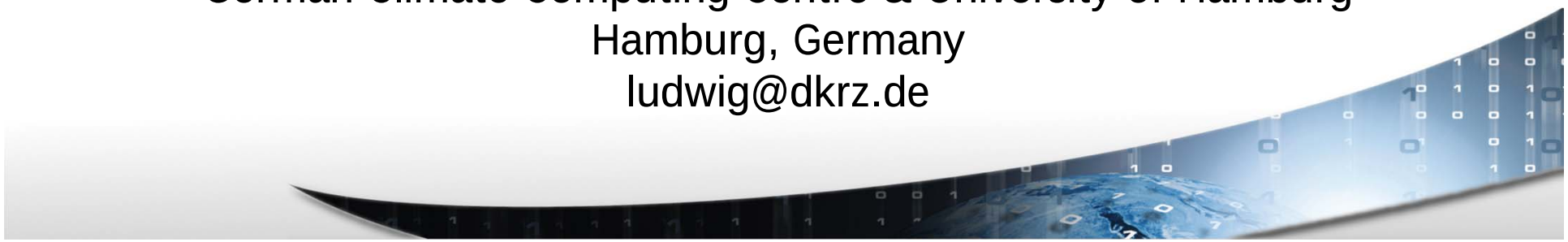
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

Tools for Earth System Modeling

Prof. Dr. Thomas Ludwig

German Climate Computing Centre & University of Hamburg
Hamburg, Germany
ludwig@dkrz.de



Talk Abstract

Earth system modeling is one of the grand challenges for super computers. Not only does it require massive computational power, it also challenges I/O systems of disks and tape drives. The whole workflow from early program development phases to data analysis asks for efficient tool support.

The talk will concentrate on the different phases of knowledge gaining with earth system modeling. We will discuss the usual issues like debugging and performance analysis. However, there are also more exotic requirements like bitwise reproducibility and application integrated checkpointing. With postprocessing we will look at tools for numerical and visual analytics of the data. Data volumes are tremendous, thus I/O performance plays an important role in the game.

The talk will discuss tool support during the data life cycle that is defined by the scientists´ workflows. We will analyse who will need what types of tools during which phases and what is available on the market to support this HPC-based research. With Exascale getting closer we have to review our requirements and define new priorities for extreme-scale tools.

Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

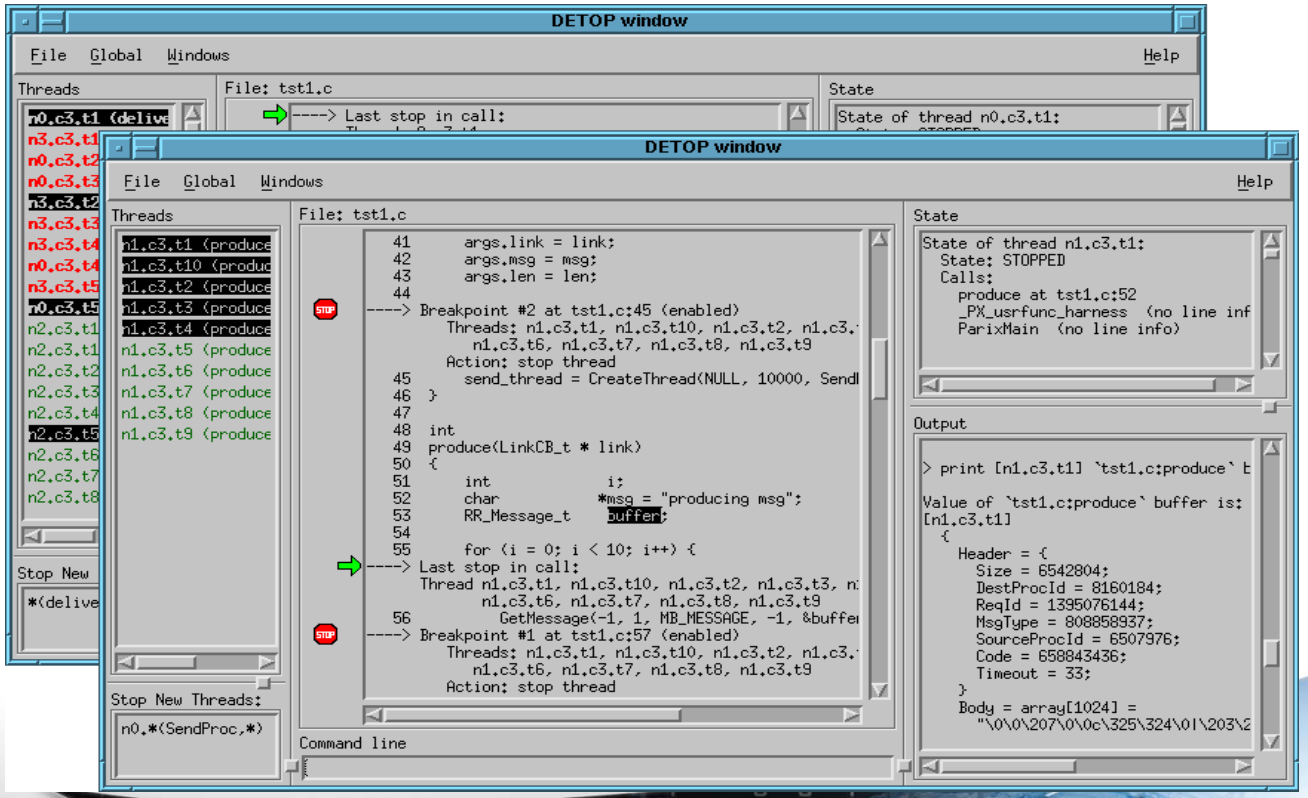
Outline

- **TL´s Past with Tools**
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

Early 90s

Tools@TUM: Tools for Parallel Systems (TOPSYS)

On-line: debugger, performance analyzer, visualizer, load balancer, checkpointer, computational steering



Late 90s

Ludwig / Wismüller

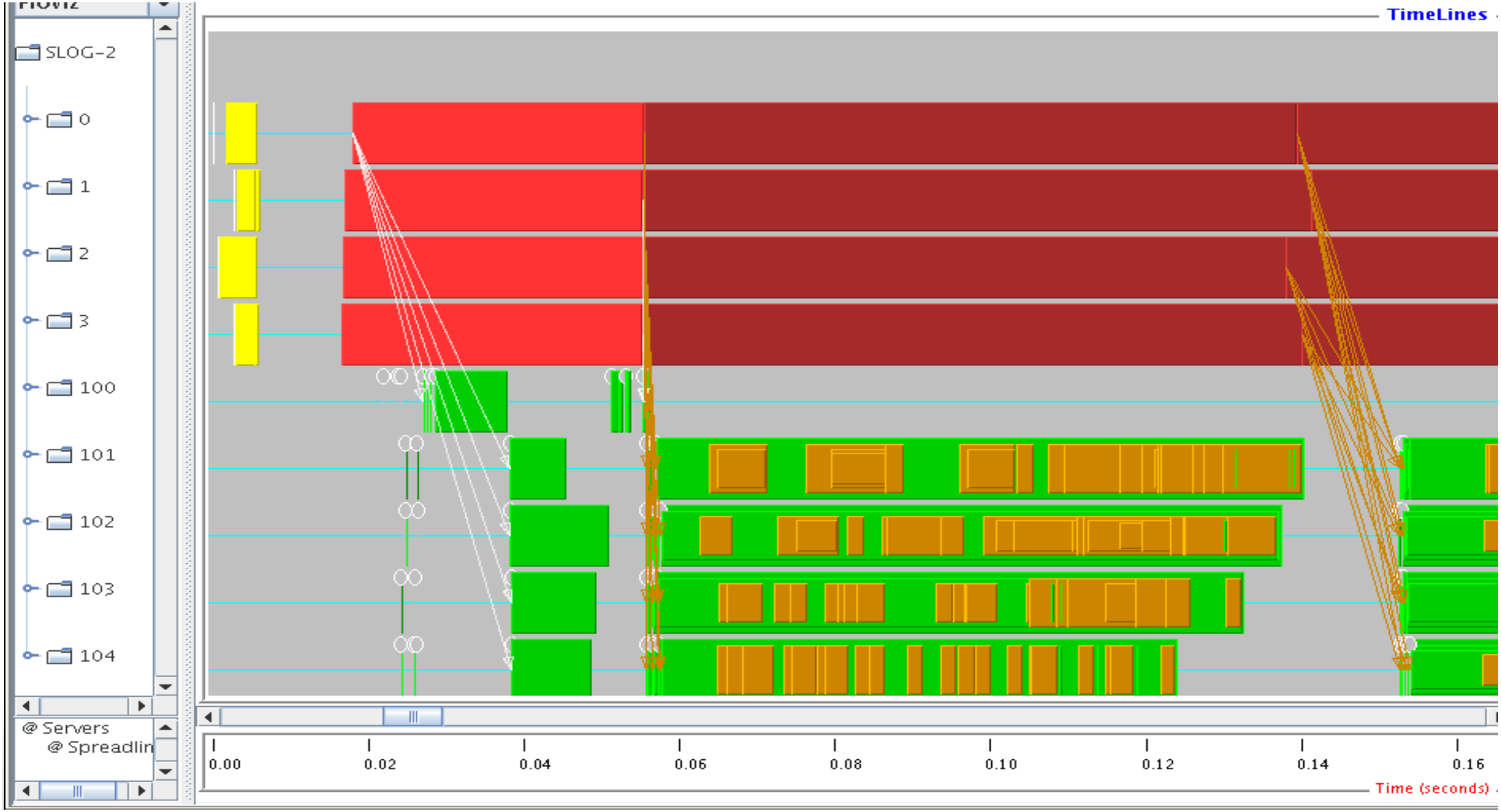
- OMIS – On-line monitoring interface specification
- OCM – OMIS compliant monitoring system

Cooperations

- Apart: Working Group on Automatic Performance Analysis – Resources and Tools

Early 2000

PIOviz:Trace-based tools for I/O server evaluation



Outline

- TL´s Past with Tools
- **TL´s Presence with DKRZ**
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

DKRZ – Partner for Climate Science

Maximum Compute Performance.
Mature Data Management. Competent Service.

Founded in 1987 – 25 years of HPC service

Operated as a limited non-profit company

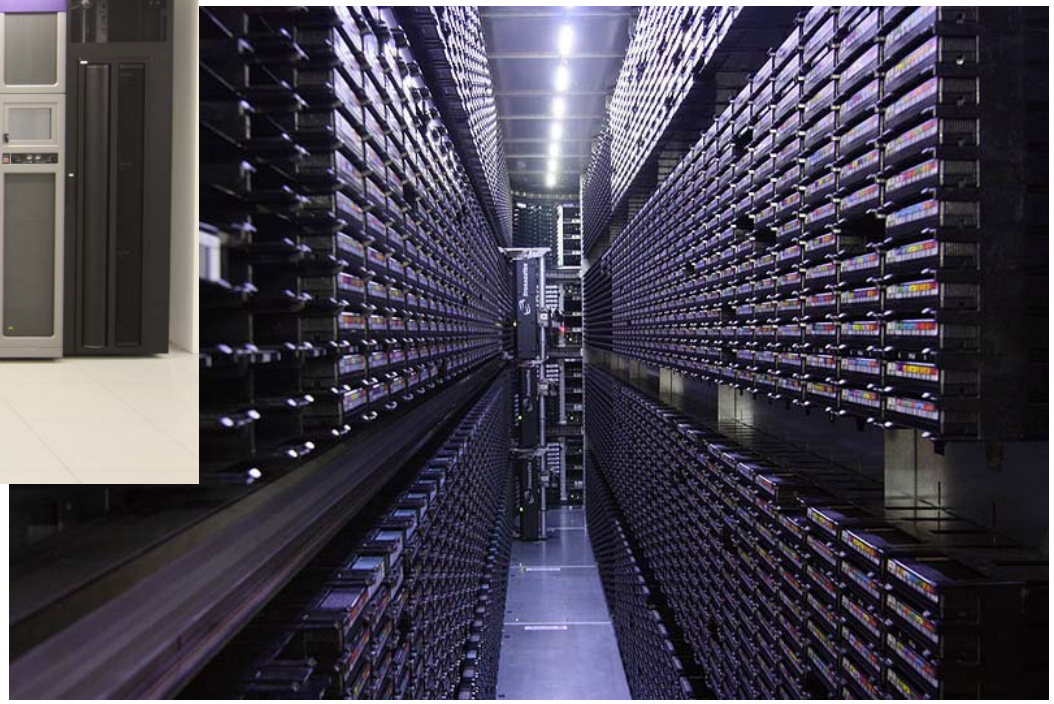
70 staff, 10 in research group

Details about DKRZ and climate research at booth #329

Computer

- IBM Power6
- Rank 368 in TOP500/Jun13
- 8,064 cores, 115 TFLOPS Linpack
- 6PB disks

Tape Library



- 90 tape drives
- 100 PB storage capacity
- HPSS HSM system

“Normal” Tools

DKRZ actively uses tools and teaches tool usage

Together with VI-HPS

- Basic profiling and hardware counter usage
- Score-P
- Cube
- Vampir
- Scalasca
- DDT

More tools for data and workflow management and visualization

However, tool usage is complicated

Future Computer and Storage

Next generation climate computer

- 1-3 PFlop/s
- 45 PByte on disk
- An estimated ½ EByte as tape library capacity

Storage: 30-50% of investment and energy costs

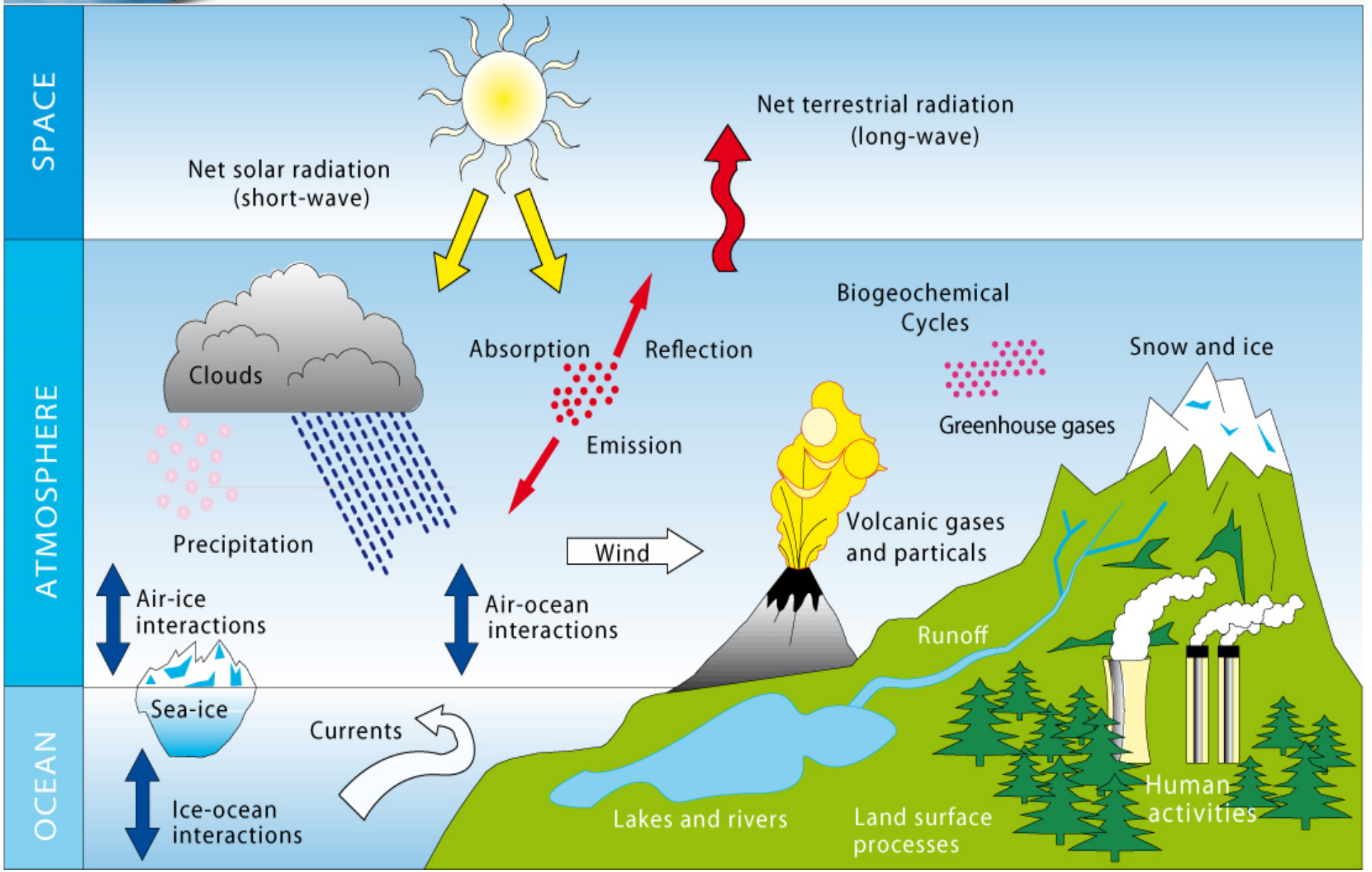
- Currently it is more like 10%

Storage 50% of the overall complexity ?

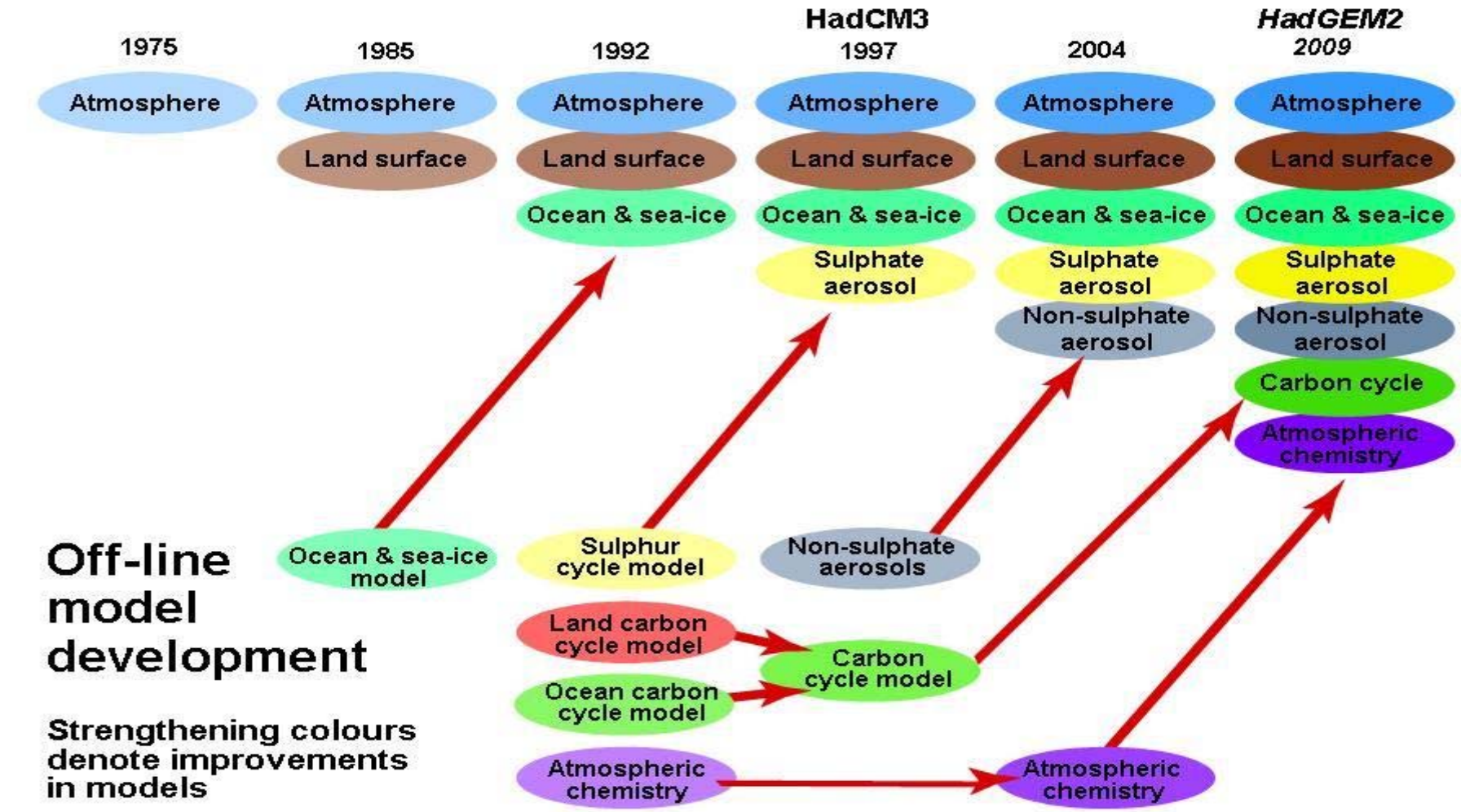
Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- **Climate Science Issues**
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

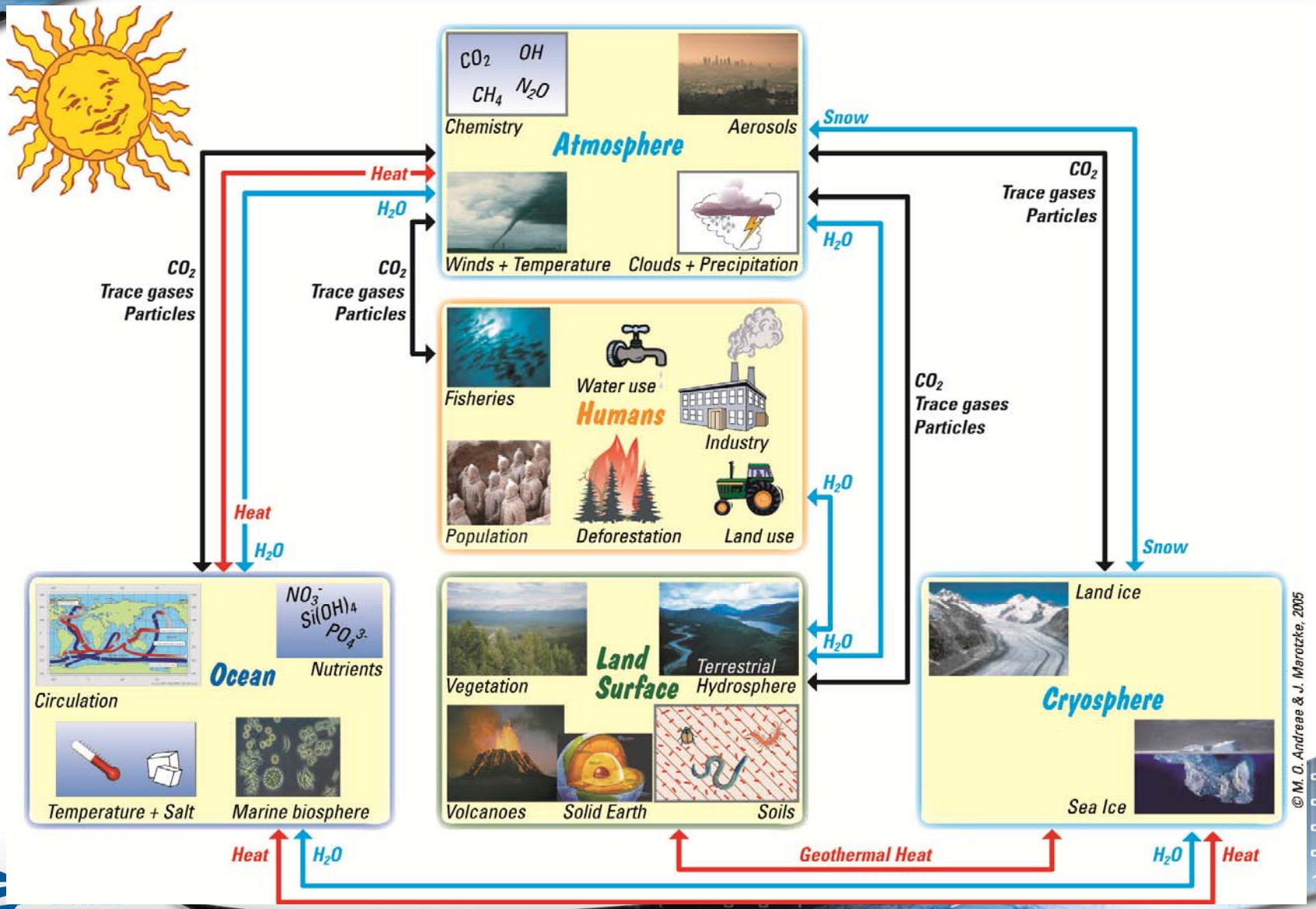
Climate Modeling



Model Components



Model Components...



Program Complexity

Mostly: **Multiple** Program, Multiple Data (MPMD)

E.g. MPI-ESM (Max-Planck-Institute Earth System Model)

- Atmospheric model: ECHAM (192 MPI processes)
  - Stand-alone as MPI/OpenMP program
- Ocean model: MPI-OM (63 MPI processes)
- Model coupler OASIS3 (1 MPI process)

How to debug? How to tune?

Program Complexity...

E.g. IRO-2 (Ice Forecast and Routing Optimization)

- Atmospheric model: METRAS (26 OpenMP threads)
- Sea ice model: MESIM (currently include in METRAS)
- Ocean model: HAMSOM (5 MPI processes)
- Model coupler: OASIS (1 MPI process)

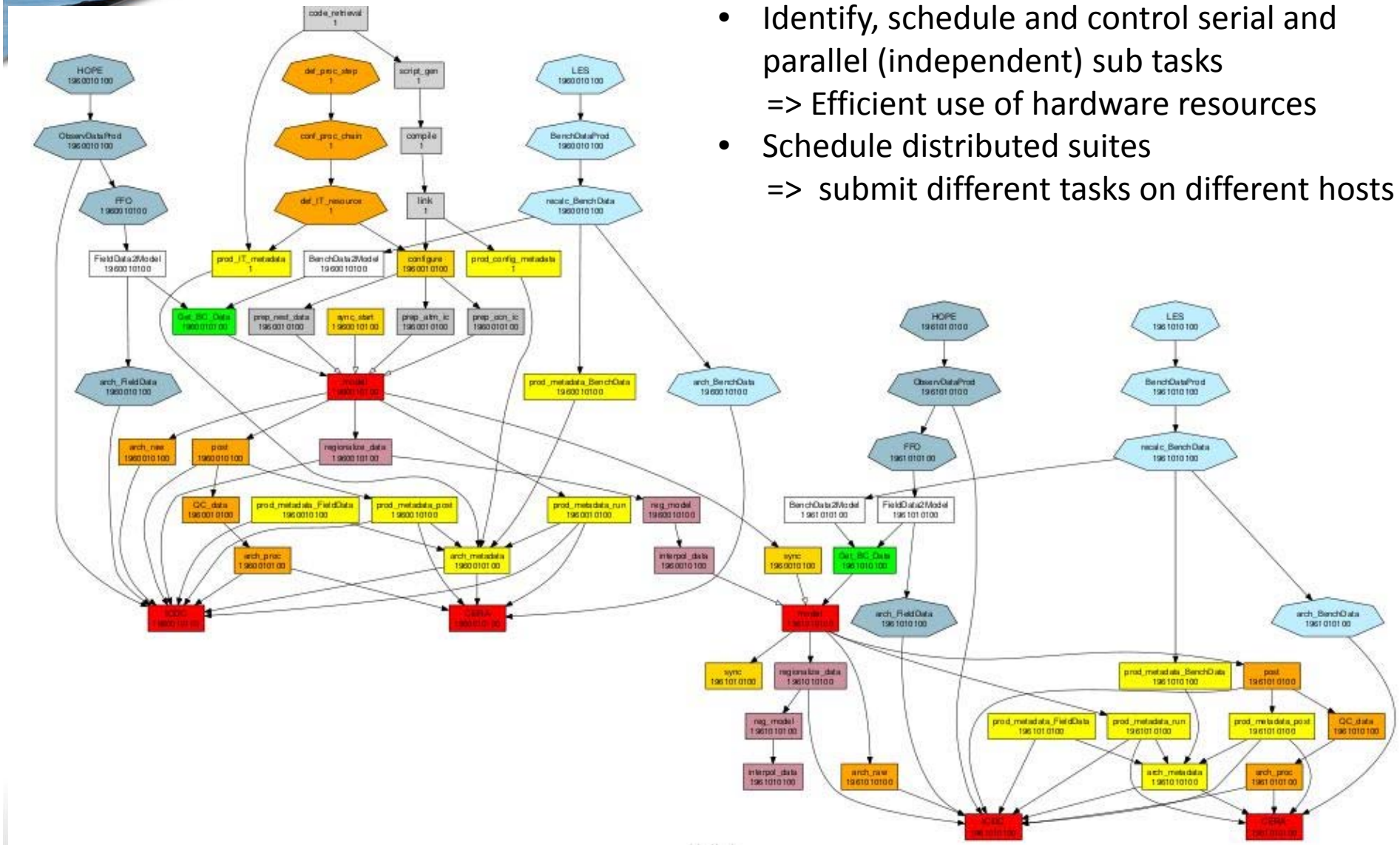
Workflow Complexity

Requirements for workflow management

- Schedule the individual steps of the process chain
- Be platforms independent
- Enable monitoring of processes
- Support testing and quality checking (QC)
- Ease failure handling
- Enable restart / controlled repetition of an experiment
- Deliver / produce provenance data

We need tools! E.g. Cylc, a meta-scheduler

Workflow Complexity...



- Identify, schedule and control serial and parallel (independent) sub tasks
  => Efficient use of hardware resources
- Schedule distributed suites
  => submit different tasks on different hosts

Software Engineering

Non-standard development process

Comprehensive ESM

- 500-1000 PY development effort
- 100.000s of LOC
- Moving target
  - Software is a dialog between scientists

Checkpointing

Checkpointing is mandatory

- Climate models are very long running applications

Good:

- Integrated application checkpointing good for resilience aspects with Exascale machines

Bad:

- Increases demands for I/O performance

I/O

Earth system science is extremely I/O-intensive

- High data volumes because of global models
- Long term storage required to validate results
- Uses own formats: netCDF, GRIB, no MPI-IO

Mostly, data must be kept in the center because of their size – move program to data

Tools: Measure program I/O, disk I/O, tape I/O

I/O

DKRZ (115 TFlop/s, 26 TByte main memory)
produces an estimated data transfer mem<->disk

- 5-10 GB/s (430-860 TB/day)
- ca. 100 TB/day are saved for further inspection
- ca. 20 TB/day are archived to tape

Next generation climate computer at DKRZ

- Main memory: > x10
- Disk space: x8
- Tape space: x5-x10

Performance

Overall performance for climate applications

- 5% to 8% of nominal system peak compute performance
- Only a fraction of nominal system peak I/O performance

Main reasons

- Algorithms, code structure, data intensiveness

Outline

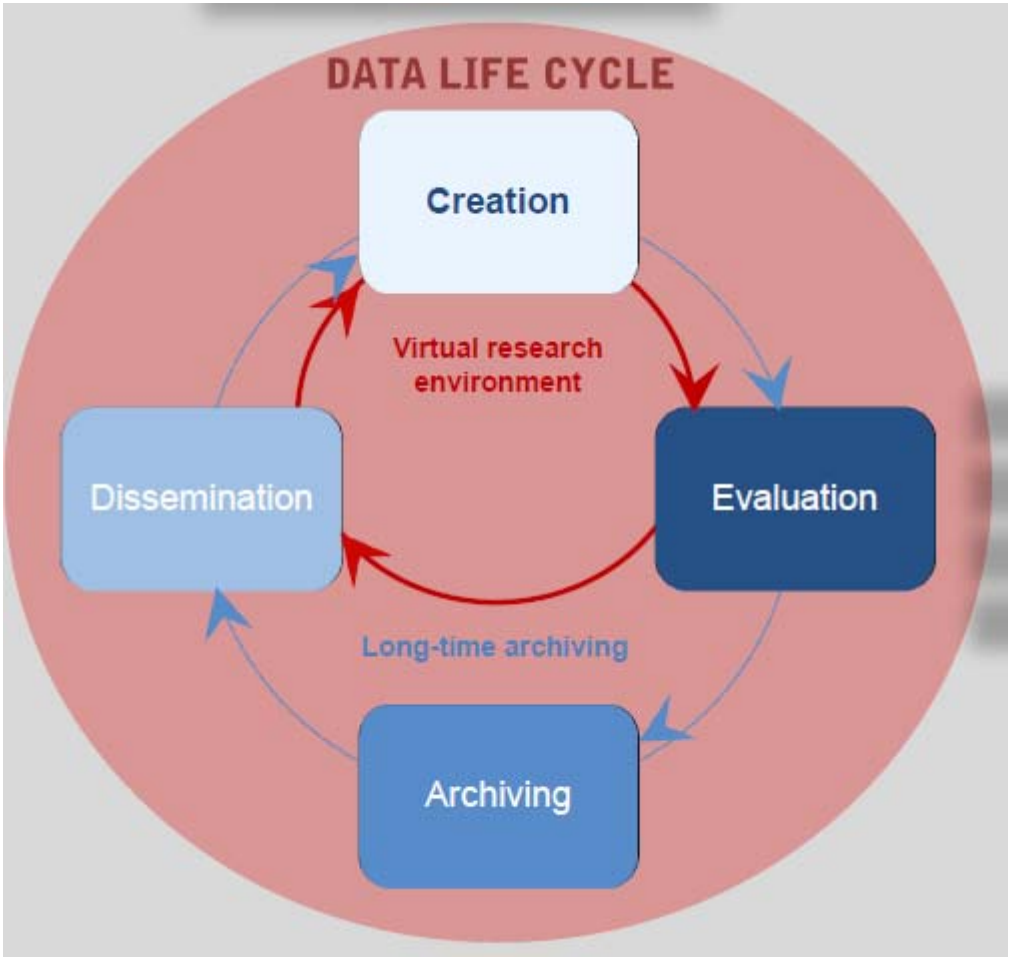
- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- **Data Life Cycle**
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

Data Life Cycle

Insight gaining is complicated...

- Create digital born data with climate models
- Evaluate data (numerical, visual)
- Archive data for future usage
- Disseminate data for further research

Data Life Cycle...



DKRZ distinguishes two layers:

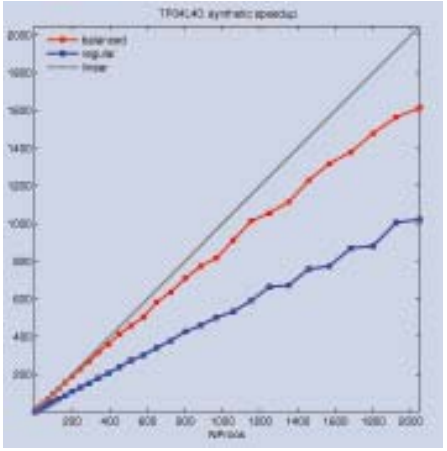
a) Virtual research environments integrates community-based scientific research

b) Long-term archiving supports interdisciplinary data utilization

Data Creation

climate modeling HLRE II optimisation

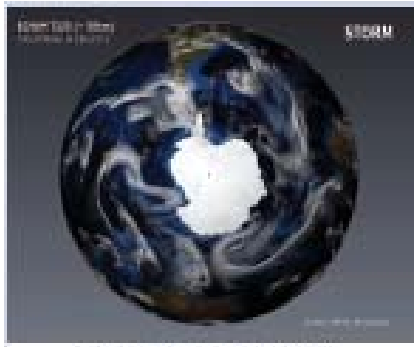
community agreed model calculations (consortial runs)

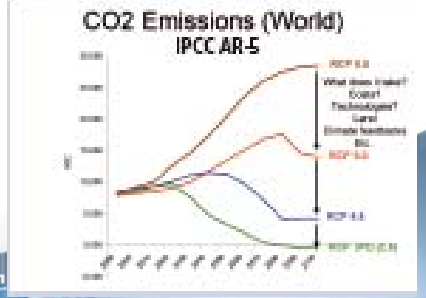


DATA LIFE CYCLE

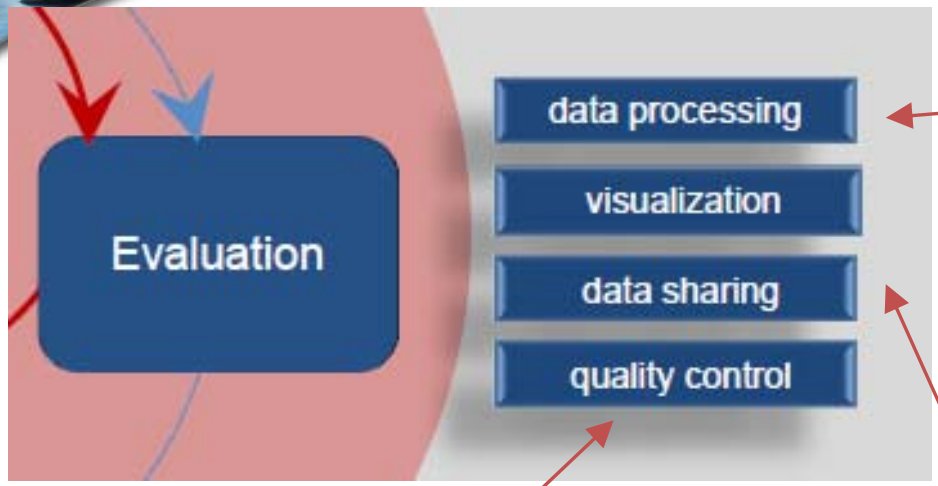
Creation

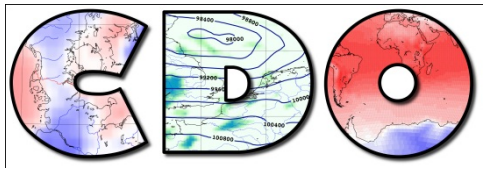
Virtual research

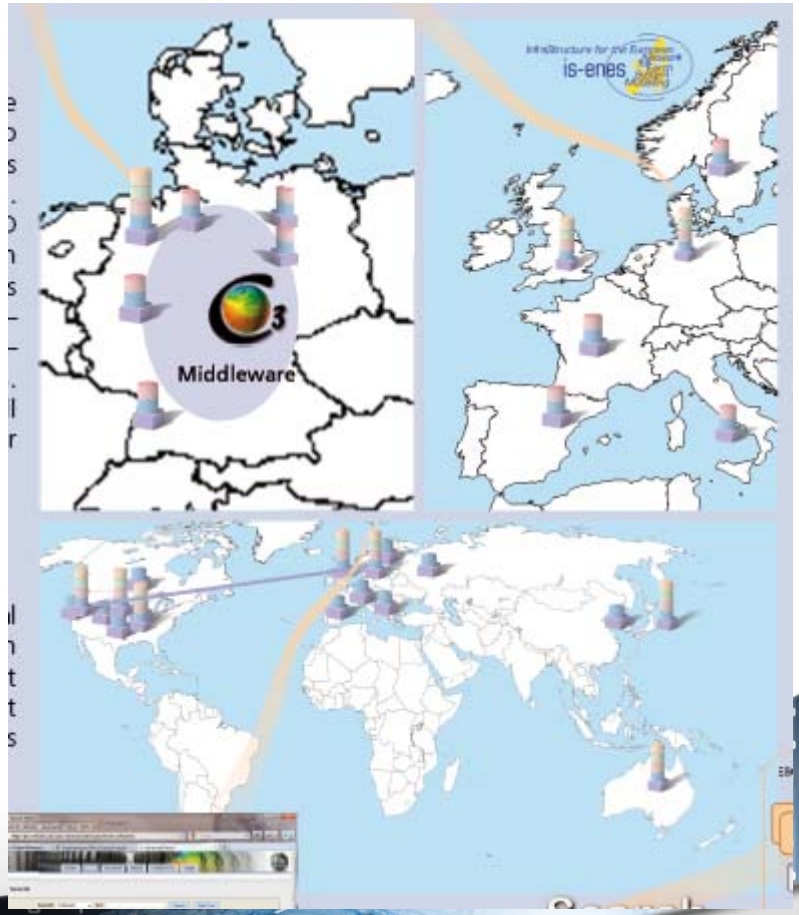


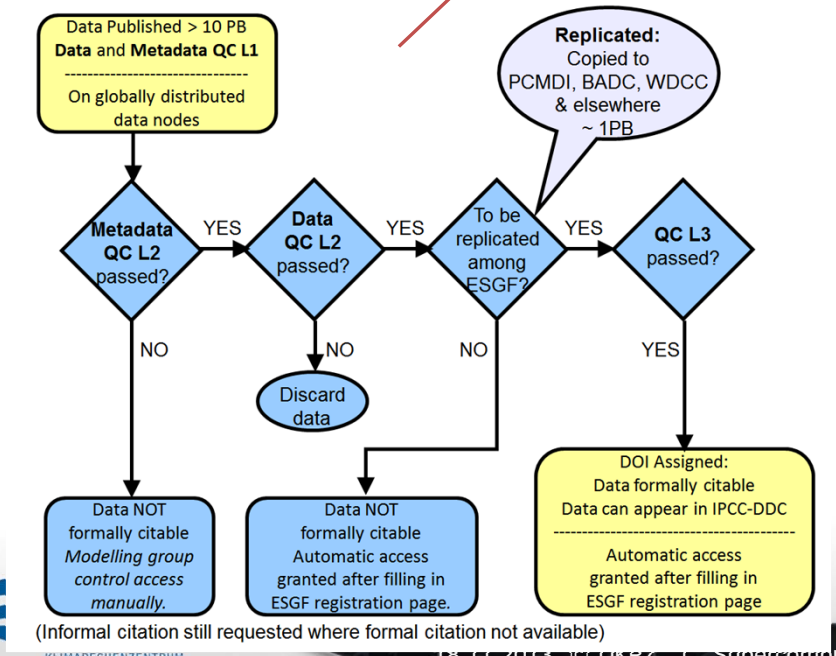


Data Evaluation

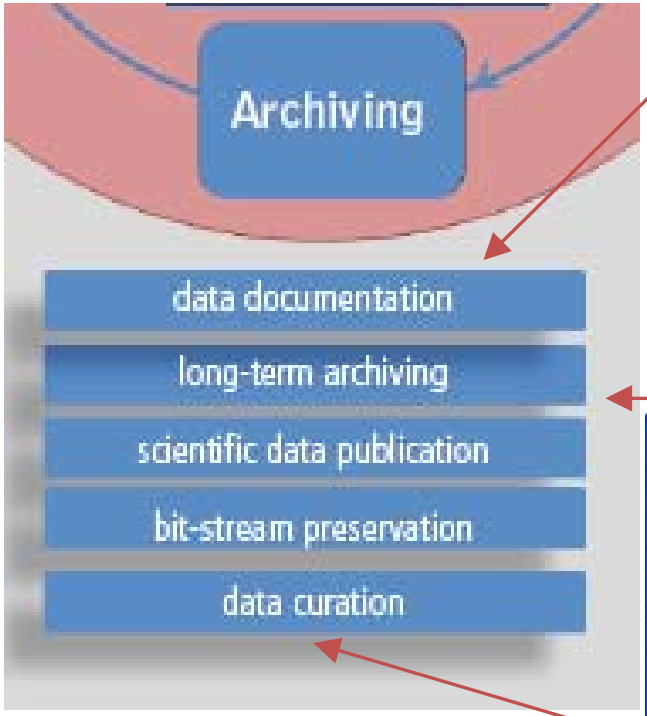


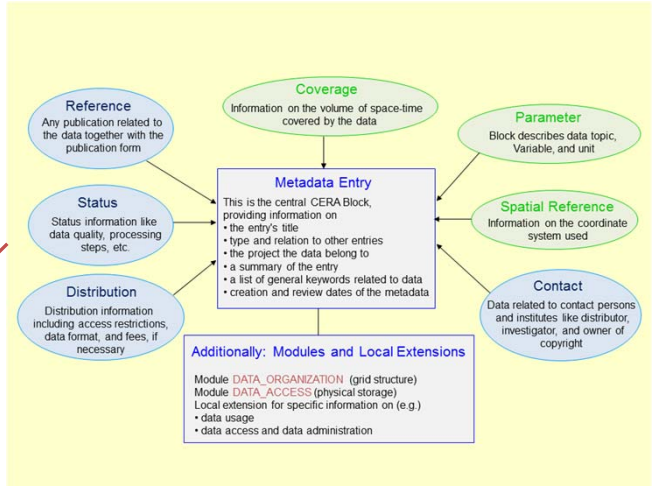


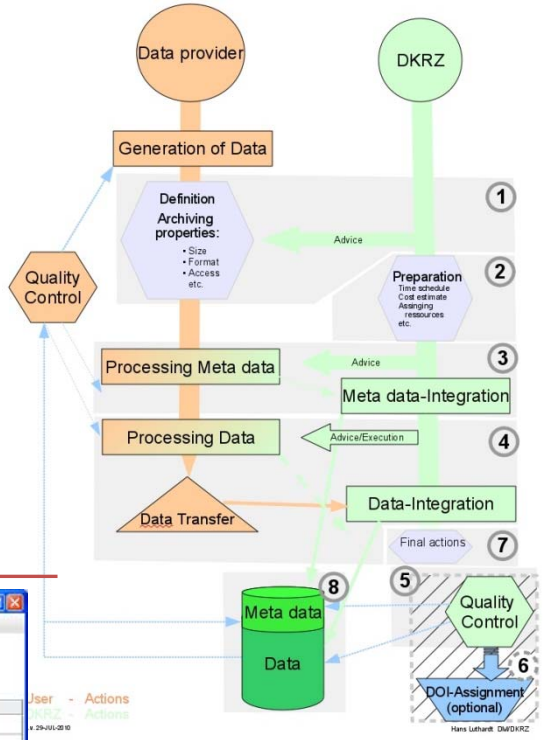


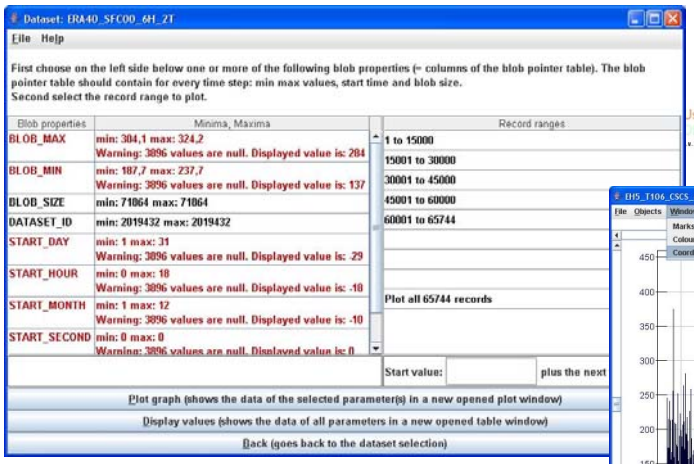


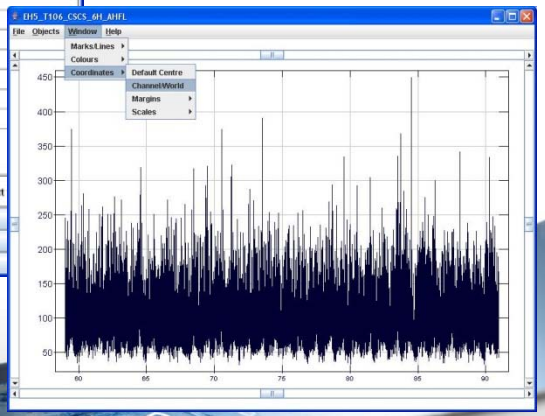
Data Archiving



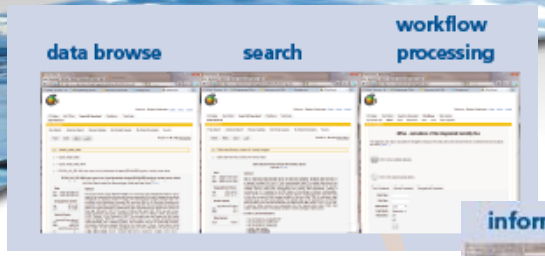


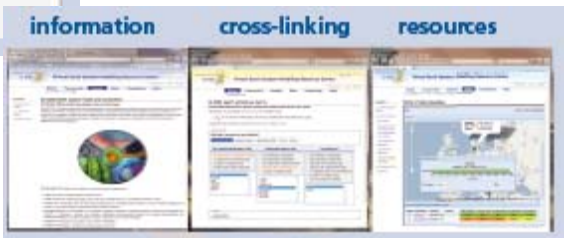


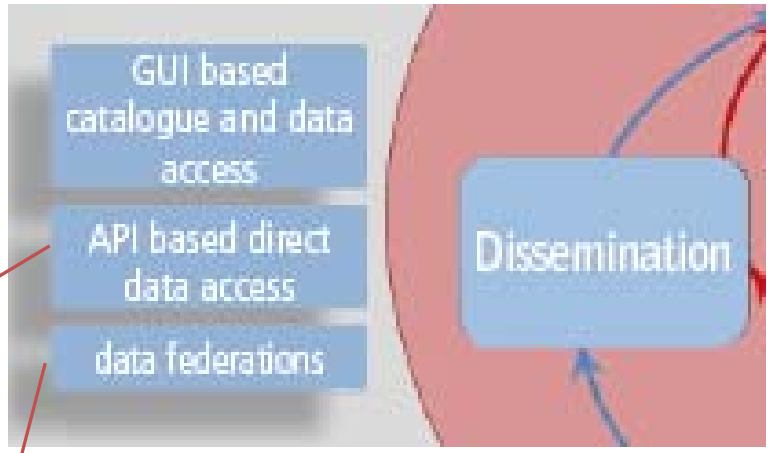


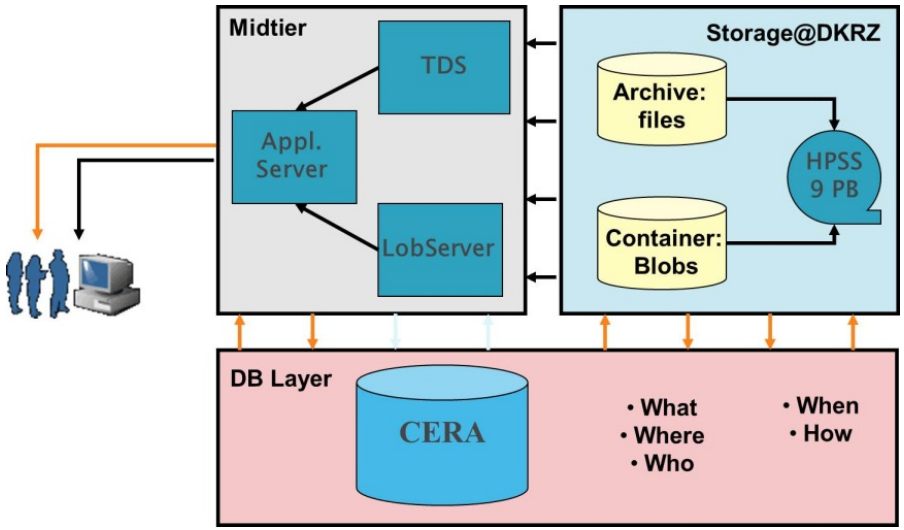


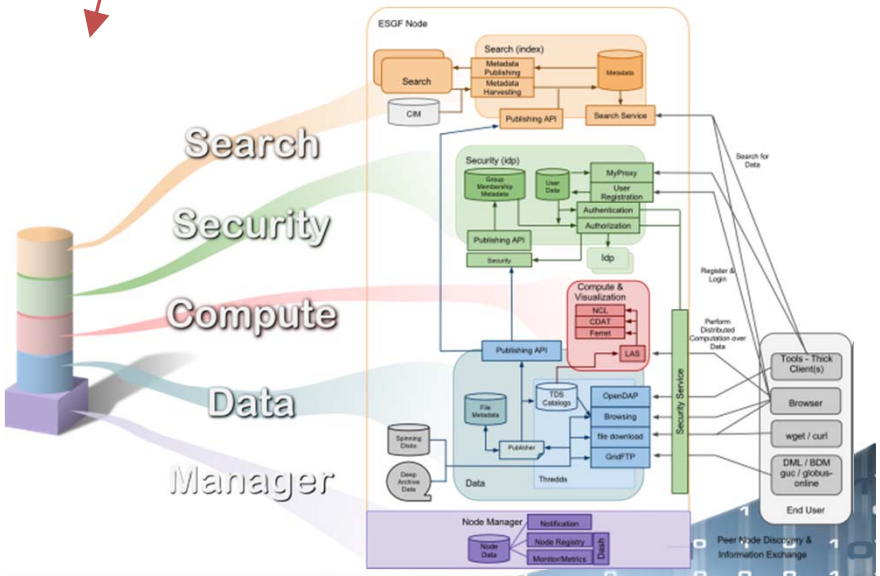
Dissemination











We need tools for the whole process !

A First Summary

Earth system modeling

- Has a complex scientific insight gaining workflow
- Has complex program and data structures

Tools

- Are needed for program development and tuning
- Are especially important for I/O tuning
- Are needed for workflow management
- Are crucial for data management

Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- **Exascale Challenge**
- Climate Science Challenges
- Wish List / Requirements
- Conclusion

Views: DKRZ vs. DOE

	DKRZ 2012	x 10.000	Exa
Linpack	110 TFLOPS	1 EFLOPS	1 EFLOPS
Main memory	26 TB	260 PB	32-64 PB
Disk space	6 PB	60 EB	0.5-1 EB
Tape library	100 PB	1 ZB	?
Memory-to-disk	30 GB/s	300 TB/s	60 TB/s
Disk-to-tape	3 GB/s	30 TB/s	?
Application-to-disk	**too slow**	**too slow**	**toooo slowww**

Exascale Problems

- Scalability of programs
- Resilience
- Energy consumption

- But also
  - Data I/O
  - Visualization of huge data volumes
  - Management of huge workflows

Extremely High Costs

Already now, with a 1/10 of a PFlop/s machine

- €2 M for electricity
- €0.5 M for tapes
- TCO is €16 M

E.g. IPCC contributions cost €1 M for electricity

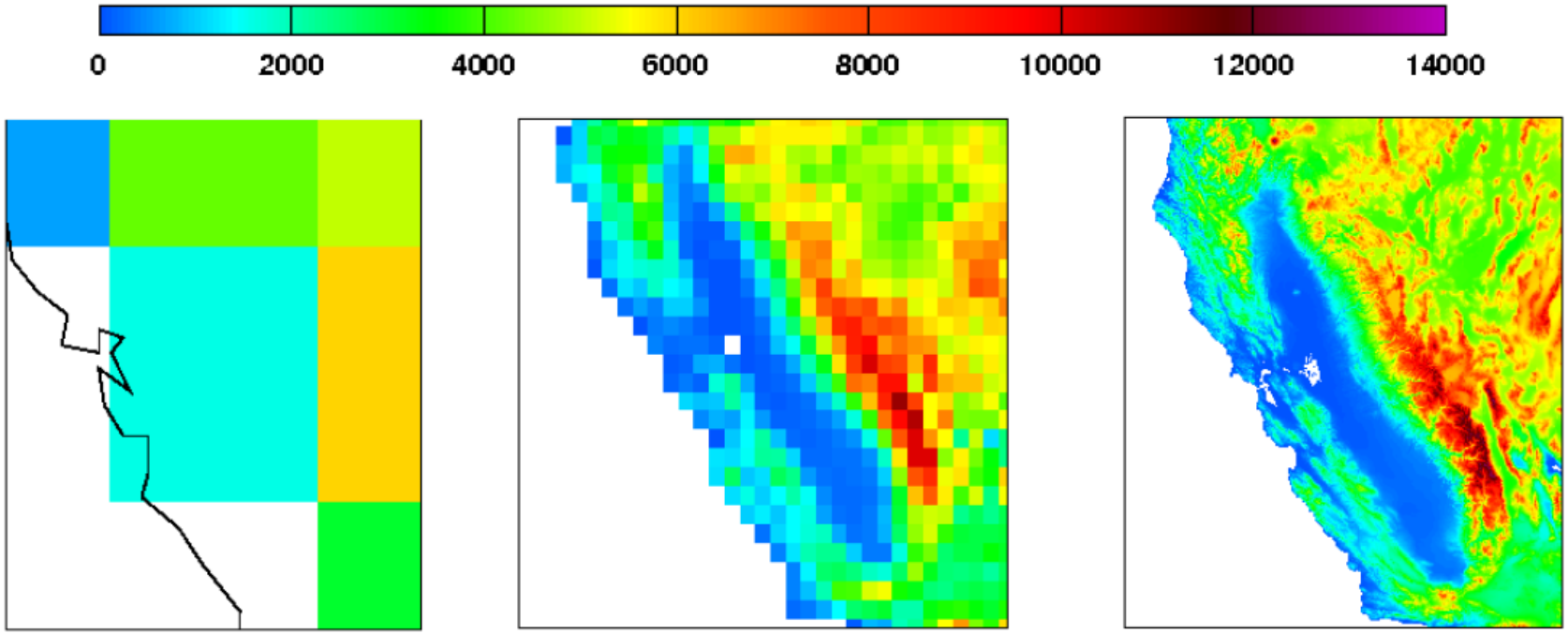
Program errors cost us 3 Cent/corehour for power

With 5% CPU time for finding errors this amount to €110.000 – enough for one more HPC specialist

Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- **Climate Science Challenges**
- Wish List / Requirements
- Conclusion

Cloud Computing “IS” Us



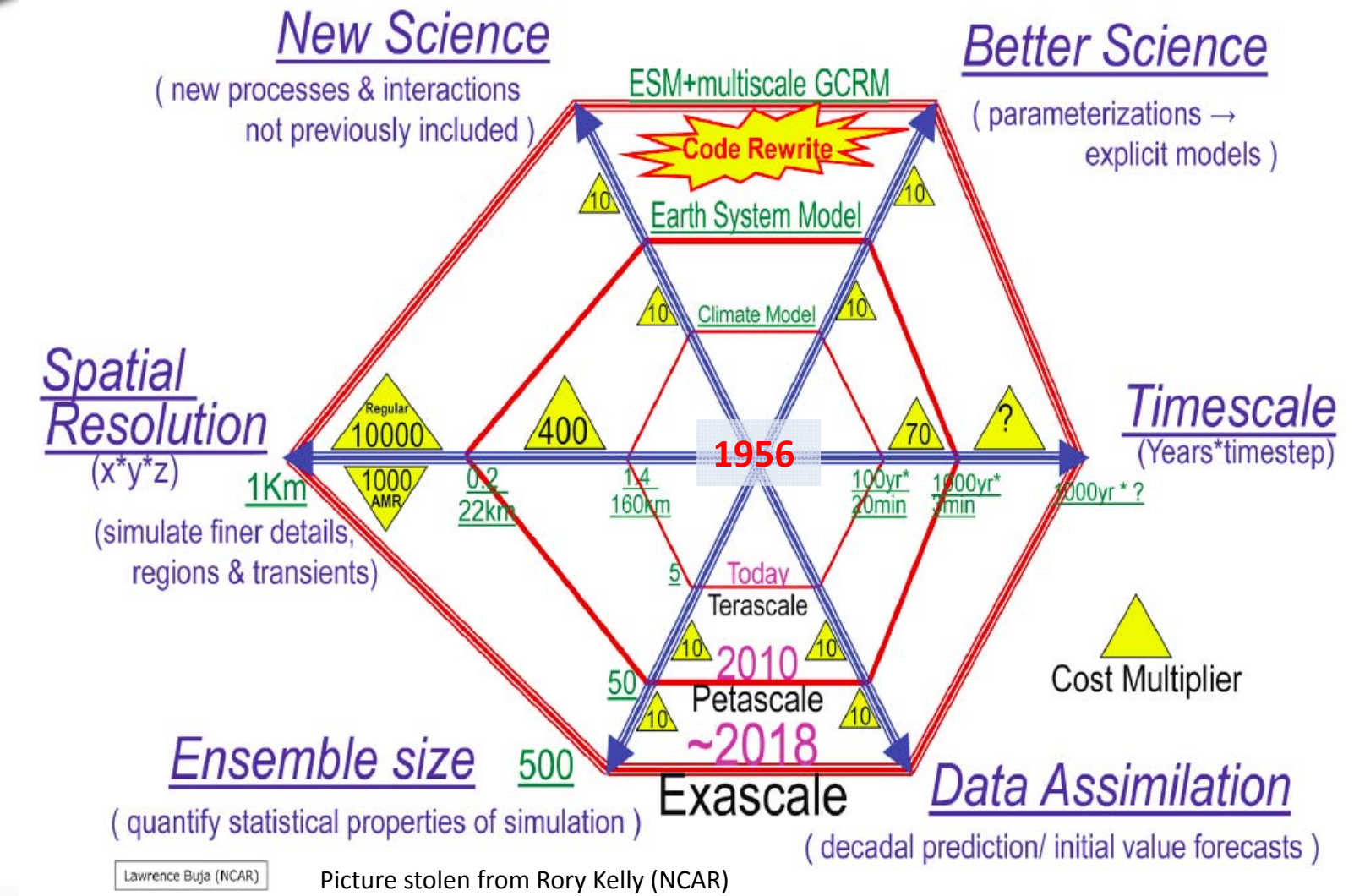
200km
Typical resolution of IPCC AR4 models

25km
Upper limit of climate models with cloud parameterizations

1km
Cloud system resolving models are a transformational change

GCRM – Global Cloud Resolving Model

The Big Bang



Picture stolen from Rory Kelly (NCAR)

The Big Data Bang

Climate Model Intercomparison Project (CMIP)

- CMIP5 finished 2013
  - 1.8 PB for 59,000 data sets stored in 4.3 Mio Files in 23 data nodes
  - CMIP5 data is about 50 times CMIP3
- We expect 100 PB for CMIP6 in 2020
  - Out of ½-1 EB raw model data

CMIP5 is only *one* big community project

Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- **Wish List / Requirements**
- Conclusion

Application Developer

Would love not to be forced to use tools

Otherwise some special tools for

- Automated evaluation of scalability
  - Add more nodes – what happens to performance?
- Computational steering
  - Quickly find problems with numerical solutions

Library Developer

In particular for I/O

- Measure and evaluate I/O performance
  - Application level view
  - System level view

Resource Provider

- Quickly identify low performers
  - wrt. compute performance
  - wrt. energy efficiency
- Decision basis for next generation machine
  - Resource usage profile
  - Scientific workflows
- Evaluate costs: per user/program/publication
  - Monetary aspects
  - Carbon footprint

Outline

- TL´s Past with Tools
- TL´s Presence with DKRZ
- Climate Science Issues
- Data Life Cycle
- Exascale Challenge
- Climate Science Challenges
- Wish List / Requirements
- **Conclusion**

Variety

We need tools for different users

- Scientist: maximize scientific productivity
- Support staff: optimize machine usage and help scientist
- Manager: decide on future resources

We need tools for all phases of insight gaining

- Model data generation
- Data visualization
- Data storage
- Data dissemination

New Tools

For earth system modeling (only?)

- Performance analysis of workflows

  How to map complex multi program structures onto complex machine structures?

- Cost analysis of workflows

  What is the overall resources usage and how can it be optimized?

People !

Invest in training of people (scientists, support staff, managers) for optimal team building

Will increase scientific productivity

Many thanks to Panos Adamidis, Hendryk Bockelmann, René Redler, Hannes Thiemann, Stephan Kindermann, and Kerstin Fieg **for discussions and ideas**