DKRZ

HPC-Infrastructure for Earth System Research

Joachim Biercamp
Panagiotis Adamidis
Stephan Kindermann
Mathis Rosenhauer

DKRZ

Outline

- Past
- Present
- Future

Outline

- **Past** *"Setting the stage"*
- **Present**
- **Future**

Outline

- **Past** ***"Setting the stage"***
- **Present** ***"Transition to new facilities"***
- **Future**

Outline

- **Past** ***"Setting the stage"***
- **Present** ***"Transition to new facilities"***
- **Future** ***"Challenges and plans"***

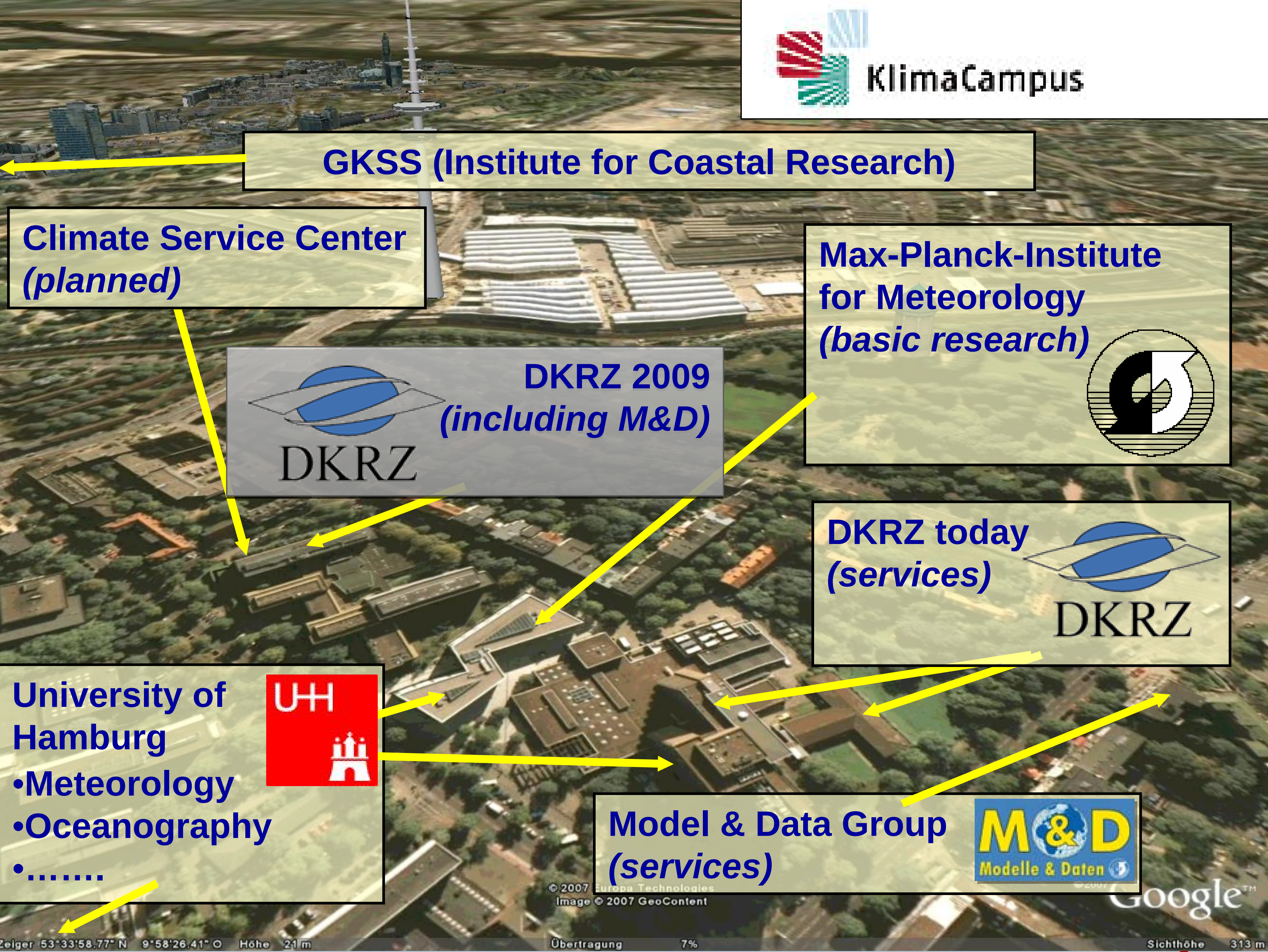
DKRZ: German Climate Computing Center

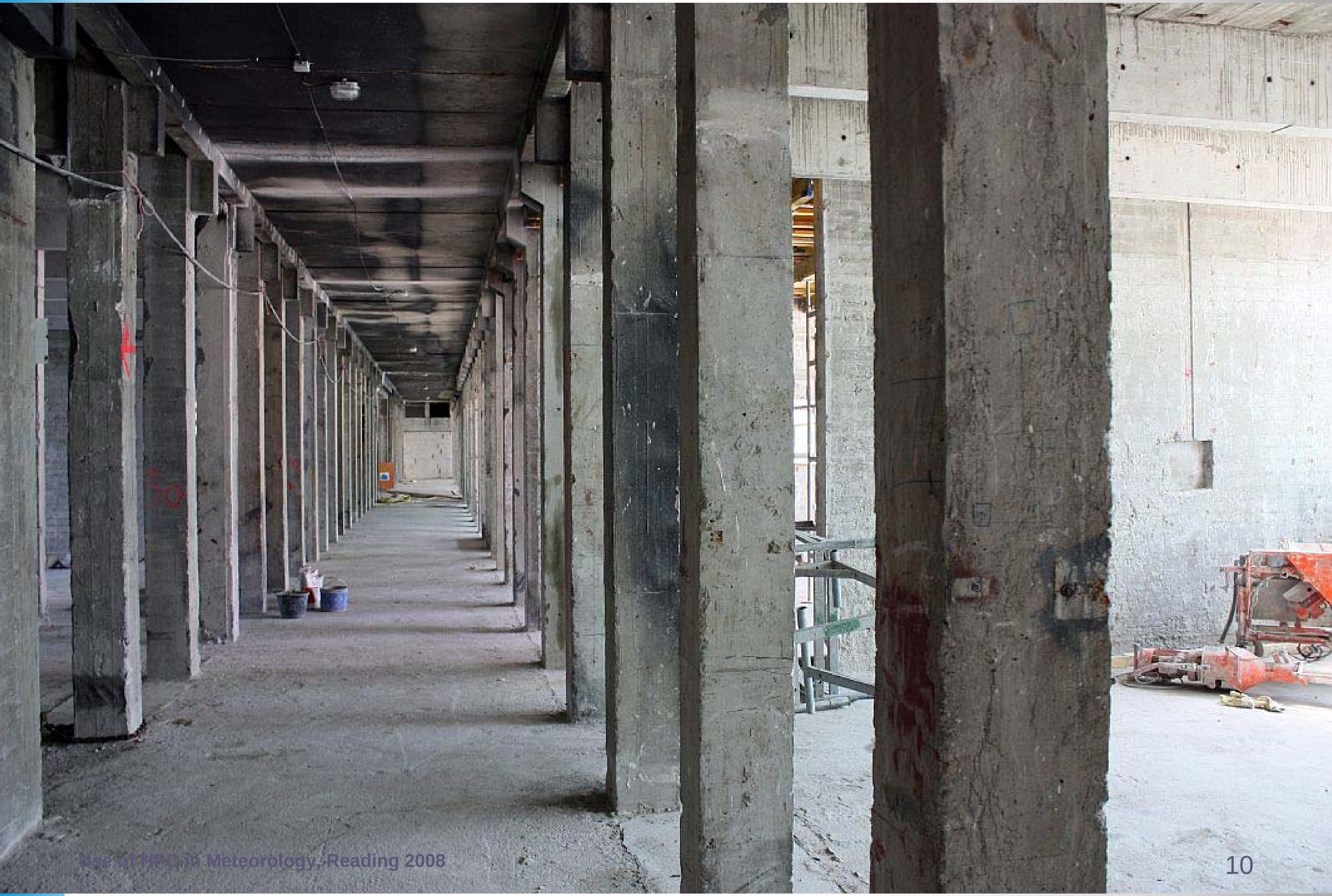
- DKRZ is a national facility
- Our mission is …

… to provide state-of-the-art supercomputing, data service and other associated services to the (German) scientific community to conduct top of the line Earth System and Climate Modelling.

The Hamburg “ClimateCampus”

- *Integration of different functions*
- *Provision of decentralized services to the community.*

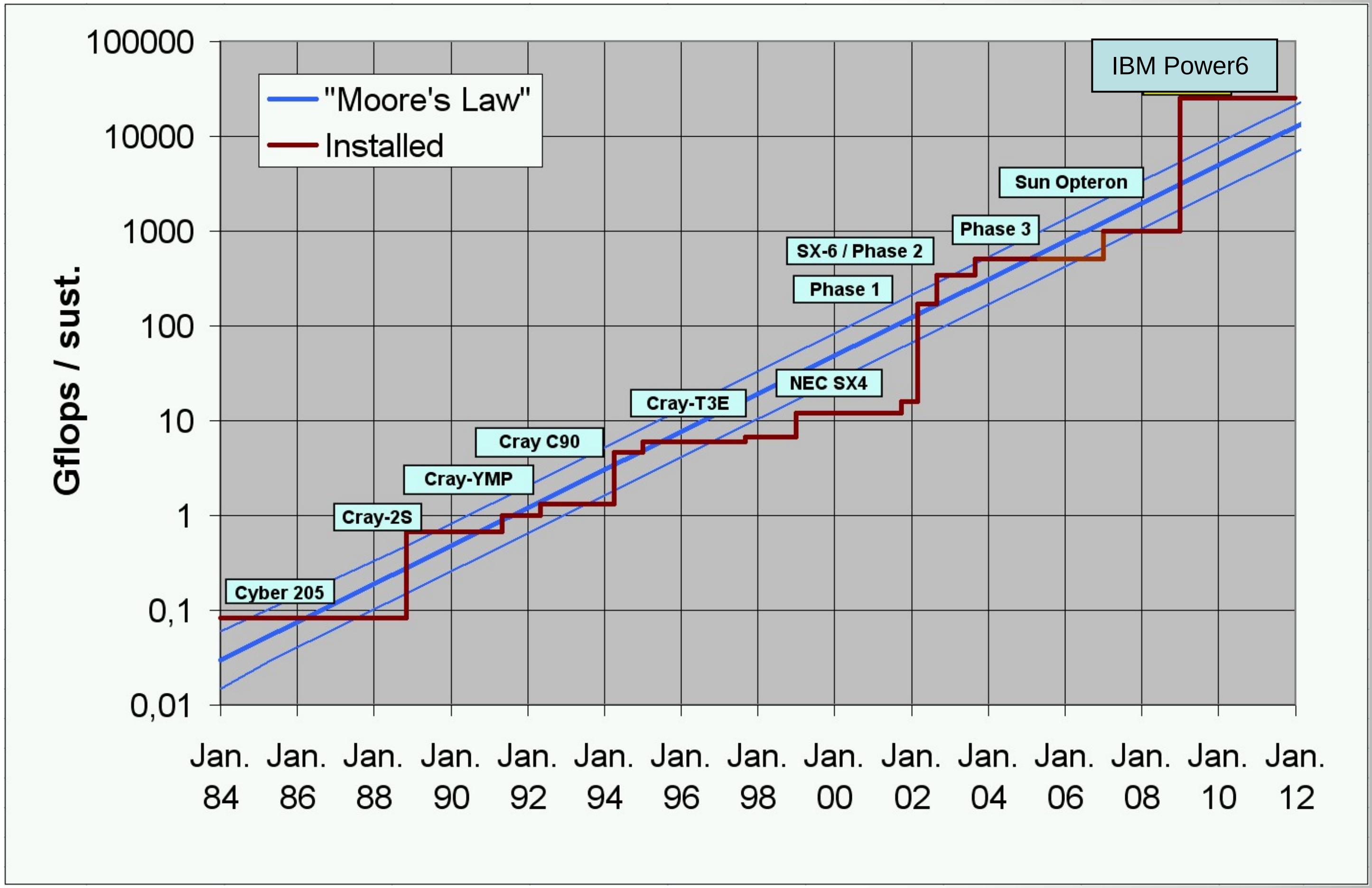

KlimaCampus
GKSS (Institute for Coastal Research)
Climate Service Center
(planned)
Max-Planck-Institute
for Meteorology
(basic research)
DKRZ 2009
(including M&D)
DKRZ
DKRZ today
(services)
DKRZ
University of
Hamburg
•Meteorology
•Oceanography
•…….
Model & Data Group
(services)
M&D
Modelle & Daten
© 2007 Europa Technologies
Image © 2007 GeoContent
Google
Zeiger 53°33'58.77" N 9°58'26.41" O Höhe 21 m
Übertragung 7%
Sichthöhe 313 m



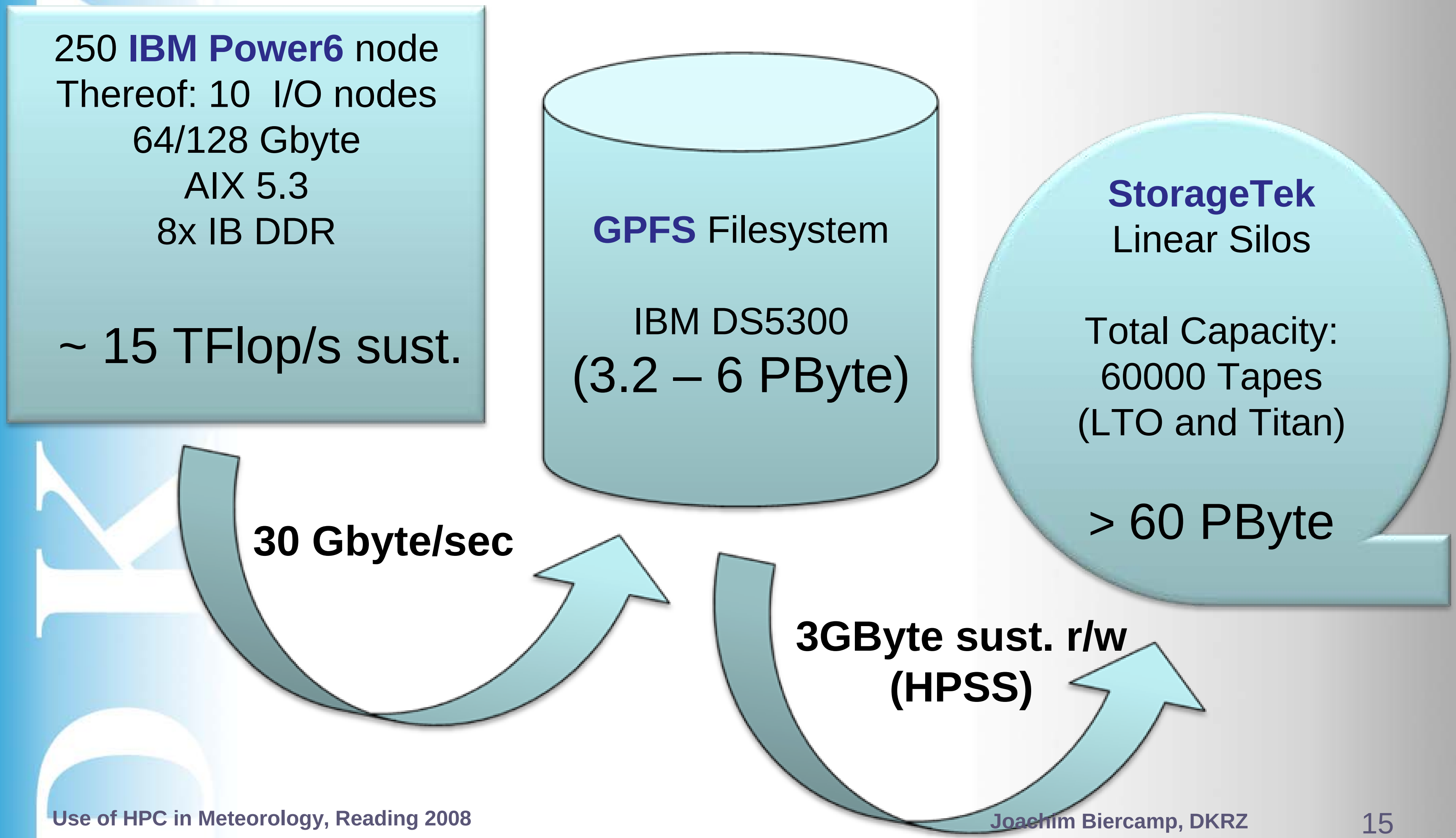

ENTZEL-CE

DKRZ, Spring 2009

Evolution of HPC at DKRZ



The new computing and data system ("HLRE2")



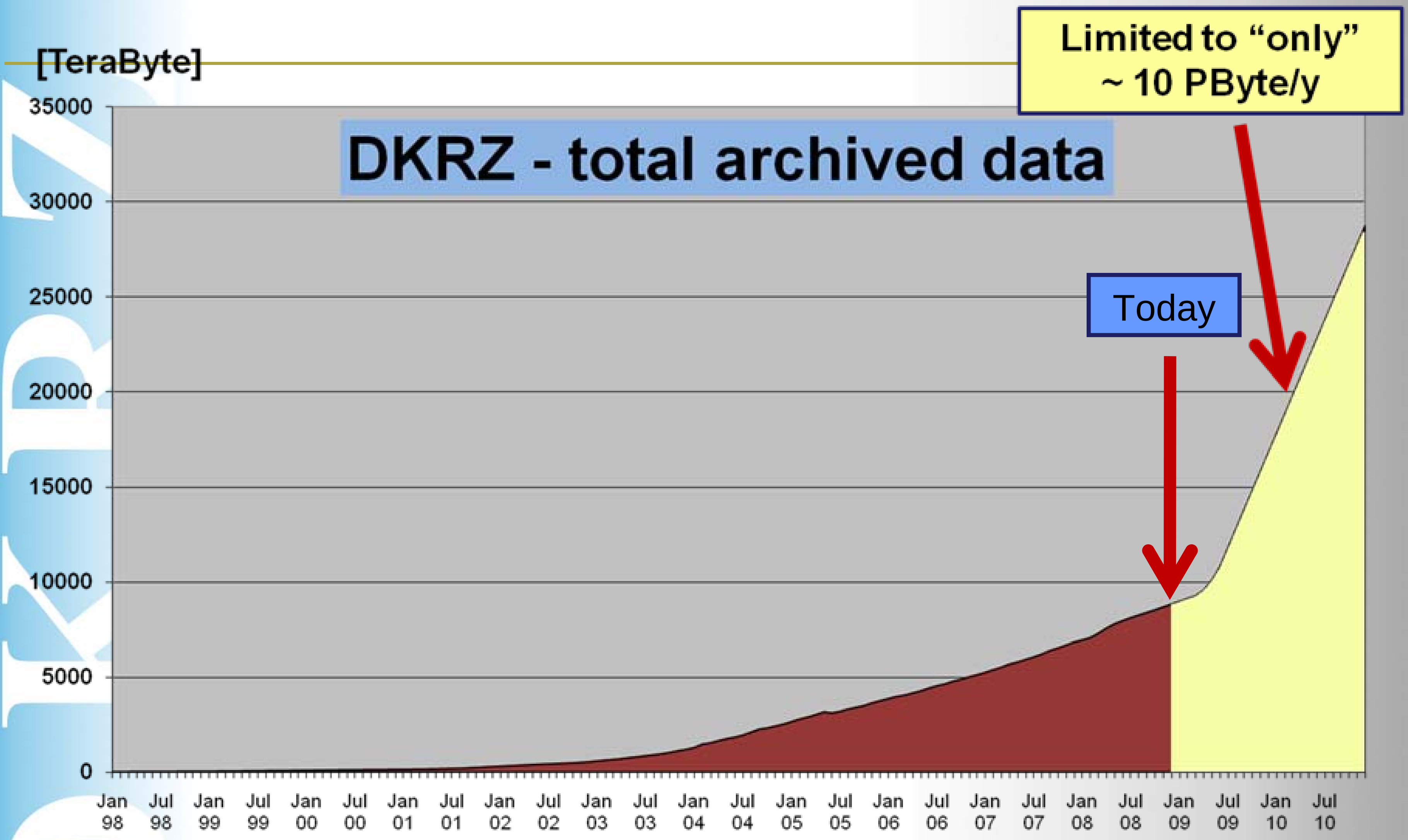
Challenges for HLRE2 and beyond

- **data**
  - annotation
  - mining
  - re-use
  - curation
- **power**
  - costs
  - supply
  - cooling
- **scaling**
  - thousands of cores
  - multi-core architecture

Efficiency

(Efficient) data handling

Data growth



[TeraByte]
Limited to "only" ~ 10 PByte/y
DKRZ - total archived data
Today
35000
30000
25000
20000
15000
10000
5000
0
Jan 98
Jul 98
Jan 99
Jul 99
Jan 00
Jul 00
Jan 01
Jul 01
Jan 02
Jul 02
Jan 03
Jul 03
Jan 04
Jul 04
Jan 05
Jul 05
Jan 06
Jul 06
Jan 07
Jul 07
Jan 08
Jul 08
Jan 09
Jul 09
Jan 10
Jul 10

Meeting Data Management Challenges @ DKRZ

(1) Limiting the increase in data volume

- *Compute power increase: factor of ~60*
- *Increase in storage capacity: from 1 PByte/y -> 10 PByte/y*
  - Media costs
  - Bandwidth of data server and tape devices

- *Incentive to reduce data volume before archiving*
  - As many disks as we can effort (6 PByte)
    ("Once on tape, forever on tape !")
  - Data is owned by PI, not individual user (e.g. student)
  - Quota
    - PI needs to apply for storage capacity
      => review by scientific steering committee)
    - Different quota for
      a. "project" data (limited life time)
      b. "Persistent" data (Publications, documented data, CERA database, IPCC process, ….)

(2) Storage and archiving of Model output

- Model I/O → GPFS
  - *some projects underway to address parallel, optimized I/O*

- GPFS → HSM (HPSS)
  - *raw data*
  - *chunks of (processed) time-series or time slices*
    - *catalogued in metadata database (CERA), to enable fast access to individual chunks/slices without the need to access and filter huge raw data files*
    - *chunks/slices are grouped and stored in* ***container-files*** *(own development, replacement of oracle blob storage)*
    - *simple transaction support for container-fill process*

(3) Becoming an IPCC AR5 data node

– Uniform AA infrastructure

→ integrate OpenID Authentication and SAML based Attribute Service Callouts into local data access infrastructure

– Metadata based discovery

→ publish Metadata for discovery in other IPCC portals (Infrastructure for this is in place, format is under discussion)

(efficient) porting and scaling

DKRZ

Typical challenges with codes coming from the NEC SX-6 vector architecture

- fitting into cache through loop exchanges and blocking is paramount.
- vector tolerates stride n memory access where n isn't power of 2 - *scalar does not!*
- No more intricate IF statements inside loops. *Compiler and processor don't help out with vector compression and mask registers*

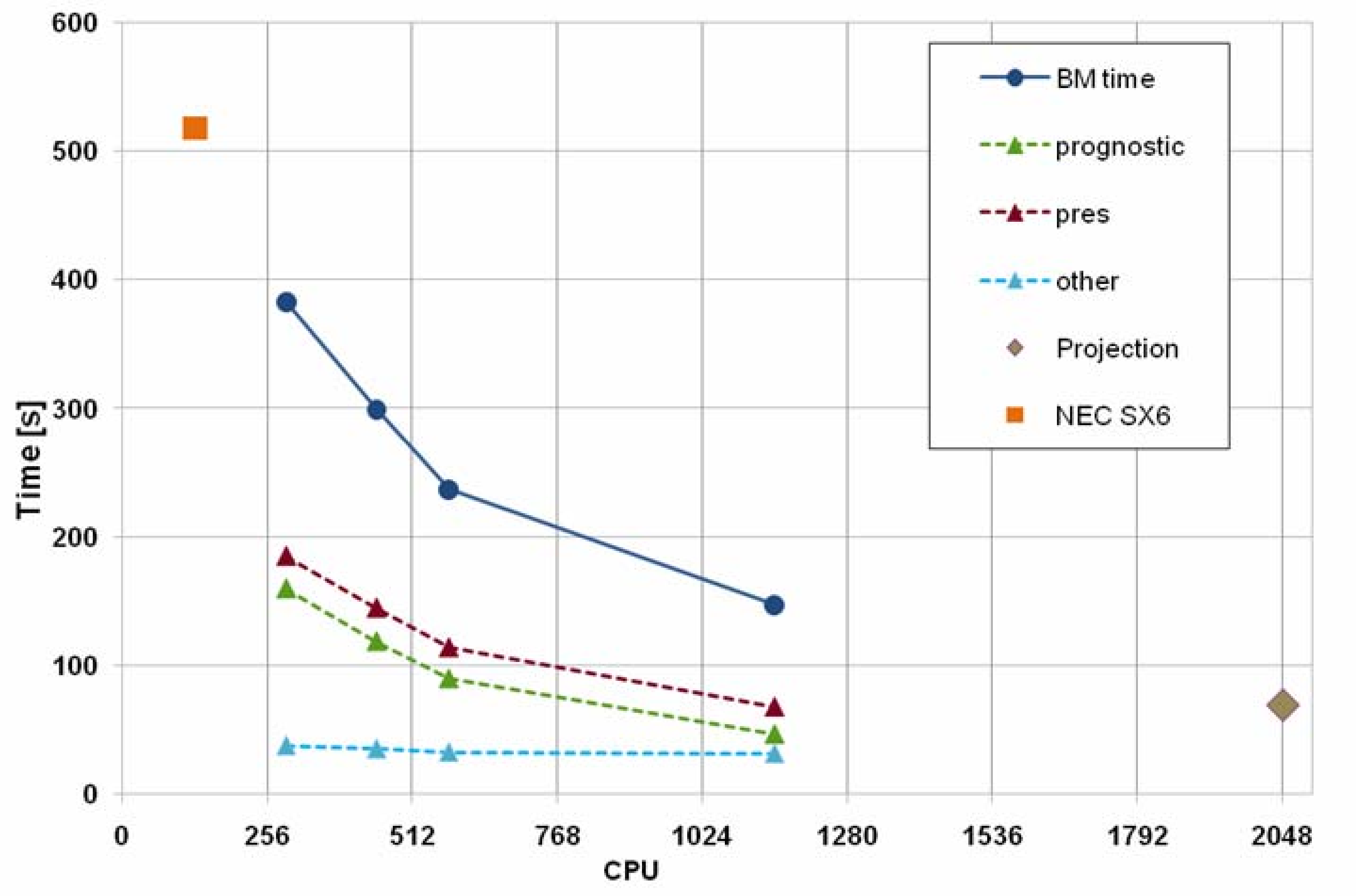
Some IBM-specific challenges we are tackling with their help

- Coupled ESMs love MPI2 process spawning and fancy inter-communicator tricks.
  IBM's Parallel Environment? Not so much.

- NEC had very easy to use and yet thorough performance analysis tools.
  IBM's tools - while powerful - give us plenty opportunity for user training.

- Even though AIX is much more widespread than SUPER-UX
  We still need to compile and support a wide range of open source software our users have come to be accustomed to.

First results (Here: PALM Large Eddy Simulation)



Plans and Projects: Next Steps

- Setting up **COSMOS** for IPCC AR5
  - talk by Luis Kornblüh on Wednesday
- **STORM**
  - *ambitious (reference) experiment -> to be run in 2009 / 2010*
  - coupled IPCC type ocean atmosphere model with (very) high resolution
    - ECHAM5 T255 (or higher ?!)
    - MPI-OM (tp 01: tripolar 0.1 deg)
    - OASIS 4
  - Use 1000+ cores per run

COSMOS/STORM: Code Optimization

Methods:

Cache Blocking Techniques

Computation vs. Communication: Expanding Halos

- More Local Operations
- Less Communication

Load Balancing: Land/Sea Points

Effort (MPI-OM ocean model only):

Treat ~40 subroutines

0.5 – 1 person years

Expected efficiency gain: 5% -> 10% -> 15% (of peak)

Later: Try CG Method with appropriate Preconditioner

Faster Convergence => Less Iterations

Less Communication

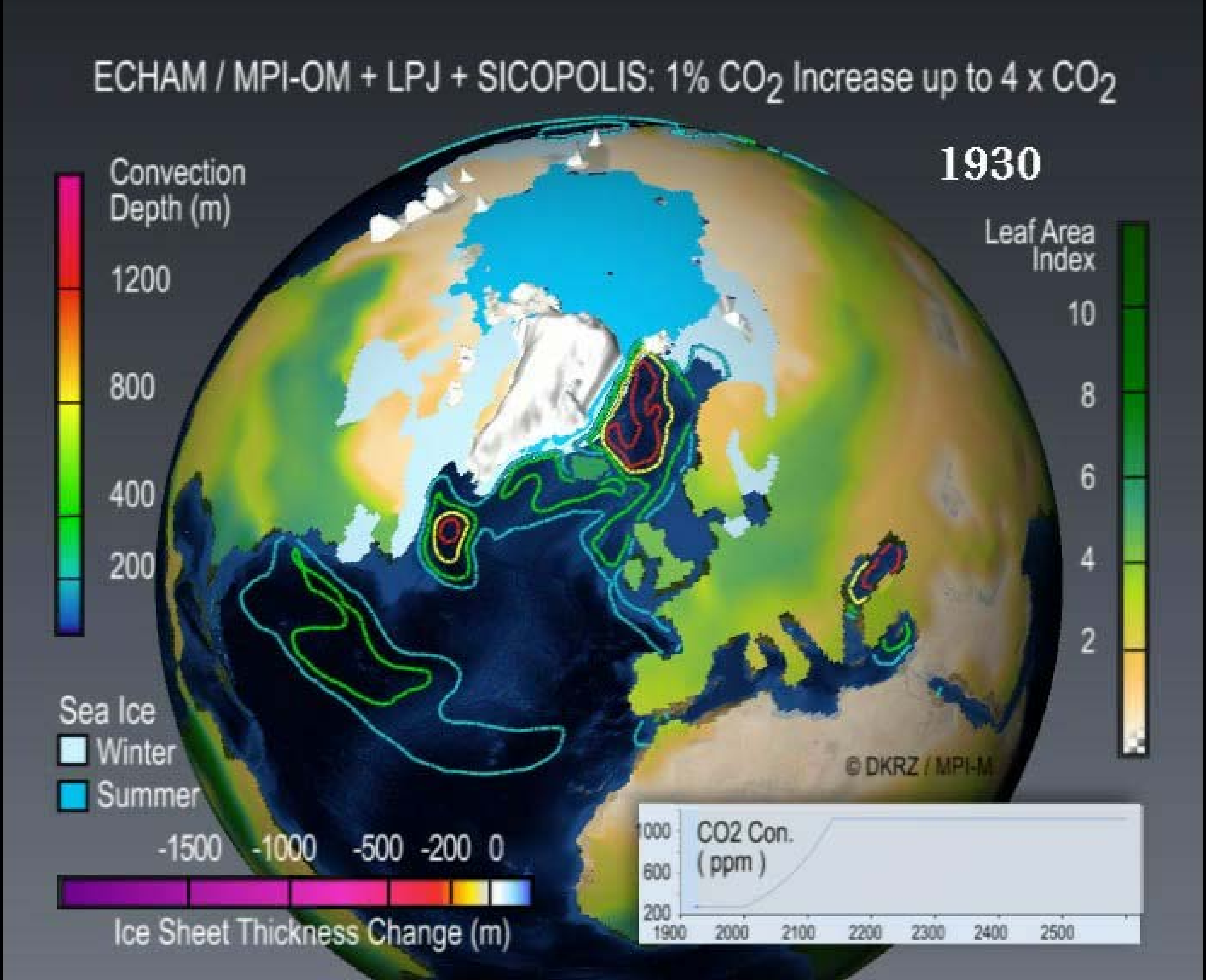
Plans and Projects: Going Petascale

- **ScalES** (**Sca**lable **E**arth **S**ystem Models; BMBF funded, Start Jan 09?)
  - Identify bottlenecks which inhibit scaling
    - I/O
    - Communication Network
    - Memory Bandwidth
    - Idle Processor Times
  - Address in COSMOS ESM (-> prototype for general procedure)
    - Parallel I/O
    - Load balancing
    - Coupler
    - Efficient use of state-of-the-art architectures
    - Dissemination and sustainability

Plans and Projects: Going Petascale

- **IS-ENES** (**I**nfra**s**tructure for **E**NES (**E**uropean **N**etwork for **E**arth **S**ystem Modelling))
  - FP7, expected start Jan 2009
  - JRAs to tackle scalability and portability
  - NAs to pave the path to DEISA and PrACE
  - Coaching and support
- **PeaKLIM** (**P**etaflop-**A**rchitekturen in **Kli**maforschung und **M**eteorologie)
  - Preparing for the petaflops challenge
  - Cooperation of Climate research (MPI-M), computer science (FHG-SCAI) and service providers (DWD and DKRZ)

    -> *Thursday: Talk by G–R. Hoffmann, U. Trottenberg*


ECHAM / MPI-OM + LPJ + SICOPOLIS: 1% CO2 Increase up to 4 x CO2
1930
Convection Depth (m)
1200
800
400
200
Leaf Area Index
10
8
6
4
2
Sea Ice
Winter
Summer
-1500 -1000 -500 -200 0
Ice Sheet Thickness Change (m)
© DKRZ / MPI-M
CO2 Con. (ppm)
1000
600
200
1900 2000 2100 2200 2300 2400 2500

Thank You !

Joachim Biercamp (biercamp@dkrz.de)