Disks, Tapes, and Climate Data

A Cost Perspective

59th HPC User Forum, Munich, Germany

Prof. Dr. Thomas Ludwig
German Climate Computing Center (DKRZ)
University of Hamburg, Department for Computer Science (UHH/FBI)

Overview

- 28 Years of DKRZ Systems
- Data Life Cycle Services
- Coupled Model Intercomparison Project (CMIP)
- Procurement Considerations
- Cost Reductions
- Exapolations

25 Years of DKRZ (1987-2012)

First computer in 1987

- Control Data Cyber-205
  - 1 processor, 200 MFLOPS, 32 MB main memory
  - 2.5 GB hard drive, 100 GB tape library

"Blizzard" system 2009-2015

- IBM Power6
  - 8,500 processor cores, 158 TFLOPS, 20 TB main memory
  - 6 PB hard drives, 100 PB tape library

factor 1,000,000 in all components

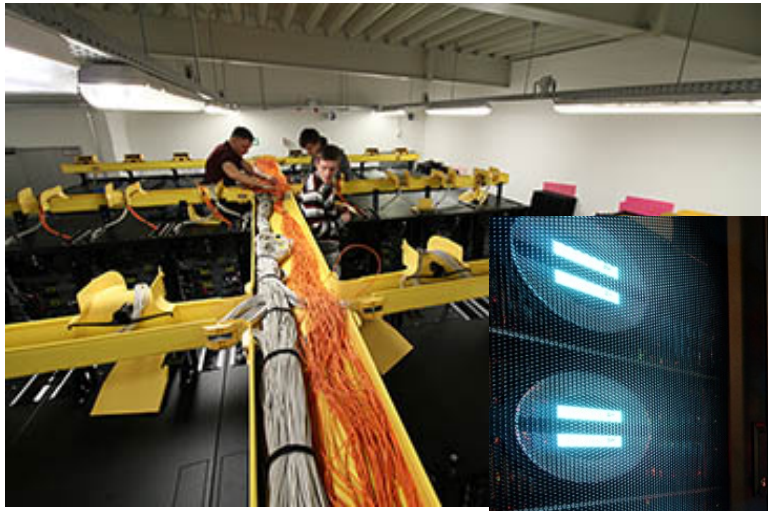
Data Generation with “Blizzard”

115 TFLOPS Linpack, 20 TByte main memory

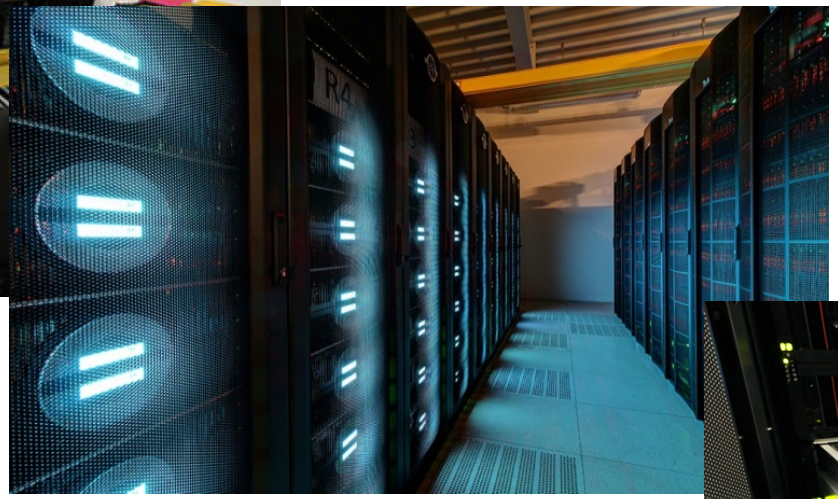
produces an estimated data transfer mem↔disk

- 5-10 GB/s (430-860 TB/day)
  - 20-40 times the complete main memory
- ca. 100 TB/day are saved on disk for further inspection
- ca. 20 TB/day are archived to tape

July 1st: “Mistral” put into Operation



bullx B700 DLC with
dual Intel E5-2680 Haswell@2.5 GHz

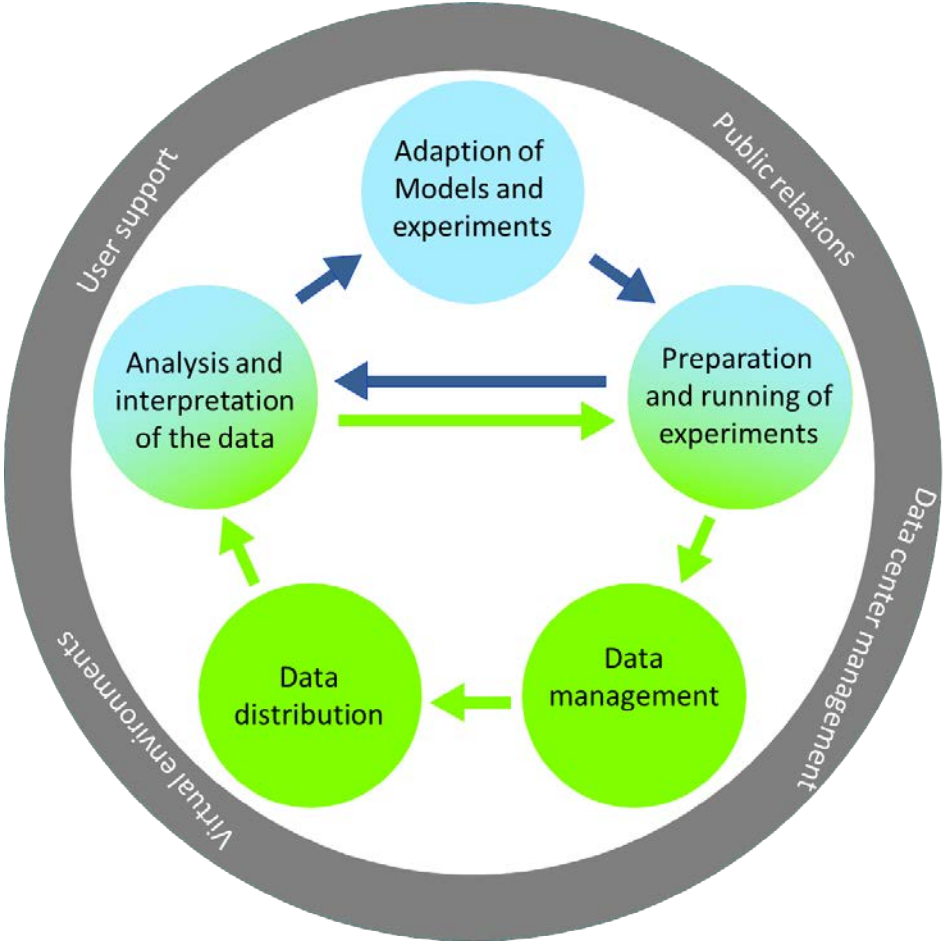


FatTree with FDR-14 Infiniband
3 Mellanox SX6536 core 648-port switches

From “Blizzard” to “Mistral”

Measure	2009	2015	Factor
Performance (no accelerators)	150 TFLOPS	3 PFLOPS	**20x**
Nodes	264	3,000	12x
Main memory	20 TB	200+ TB	10x
Hard disk capacity	6 PB	50 PB	**9x**
Throughput memory to disk	30 GB/s	400 GB/s	13x
Tape library capacity (2015, 2020)	120 PB	390 PB	3x
Throughput disk to tape	10 GB/s	20 GB/s	2x
Power consumption	1.6 MW	1.4 MW	0.9x
Investment costs	€ 30M	€ 35M	1.2x

DKRZ Service Structure



Basic workflows:

Climate Model Development

Climate Model Data Production

CMIP

CMIP5 – Coupled Model Intercomparison Project

- Provides key input for the IPCC report (5th AR, 2013)
  - Intergovernmental Panel on Climate Change
- ~20 modeling centers around the world (DKRZ being one of the biggest)
- Produces 10s of PBytes of output data from ~60 experiments (“digital born data”)

Data are produced without knowing all applications beforehand and these data are stored and archived for interdisciplinary utilization by yet unknown researchers

CMIP5 Summary

- Status CMIP5 data archive (June 2013)
  - 1.8 PB for 59,000 data sets stored in 4.3 Mio Files in 23 ESGF data nodes
  - CMIP5 data is about 50 times CMIP3

- Costs of CMIP5 at DKRZ
  - 20 M corehours in 2010/2011 = 1/3 annual capacity with IBM
  - Share of investments costs: € 1.6M
  - Share of electricity costs: € 0.6M
  - Share of tape costs: € 0.1M
  - Additional service staff: € 1.0 M

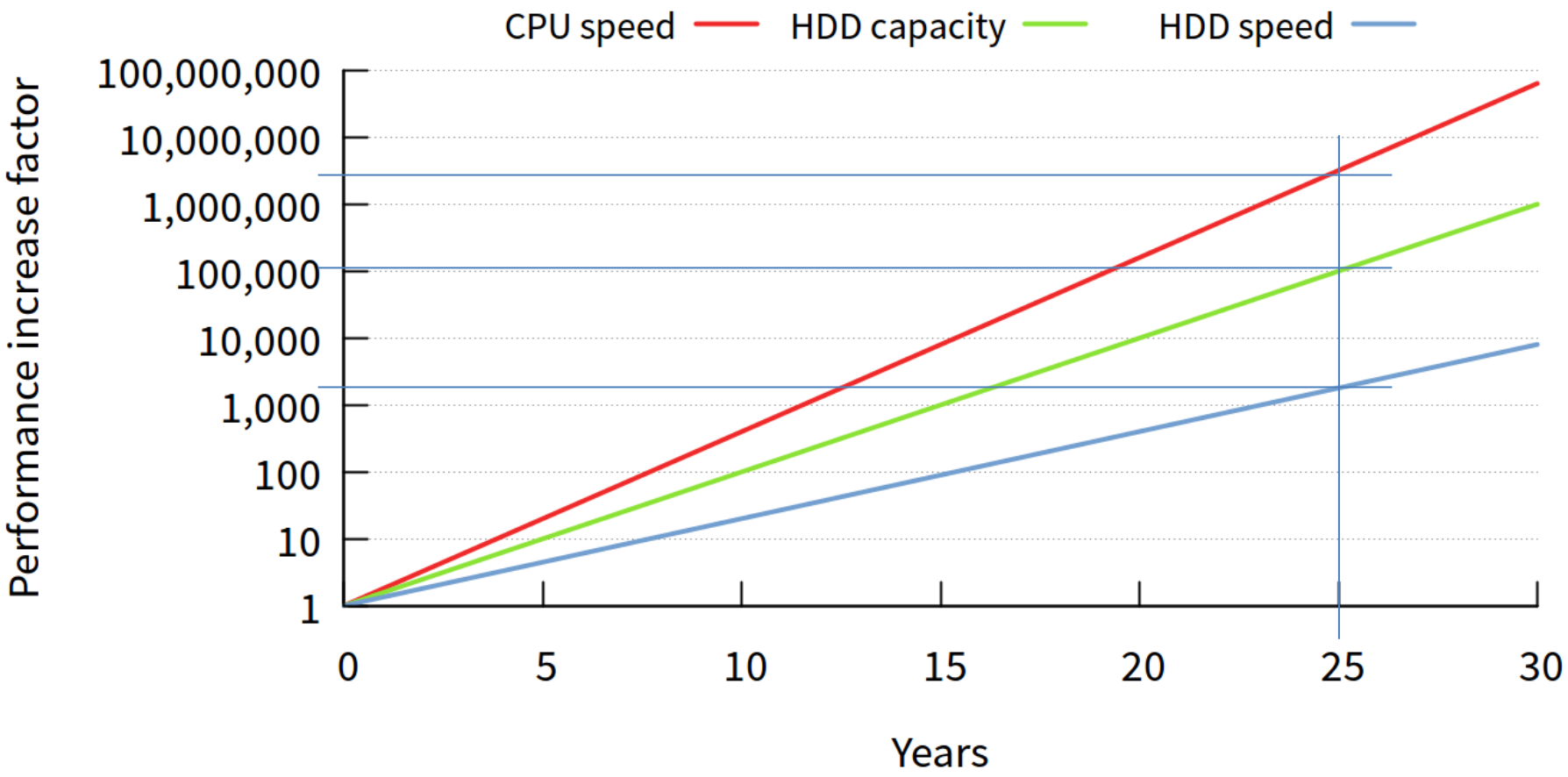
CMIP6 Data Volume Estimate

- Extrapolation to CMIP6 (2017-2019)
  - CMIP6 has a more complex experiment structure than CMIP5.
  - Expectations: more models, finer spatial resolution and larger ensembles
  - Factor of 20: 36 PB in 86 Mio Files
    - Potential DKRZ share: 3 PB on disk, 20 PB on tape
  - Factor of 50: 90 PB in 215 Mio Files
  - More accurate numbers in October 2015

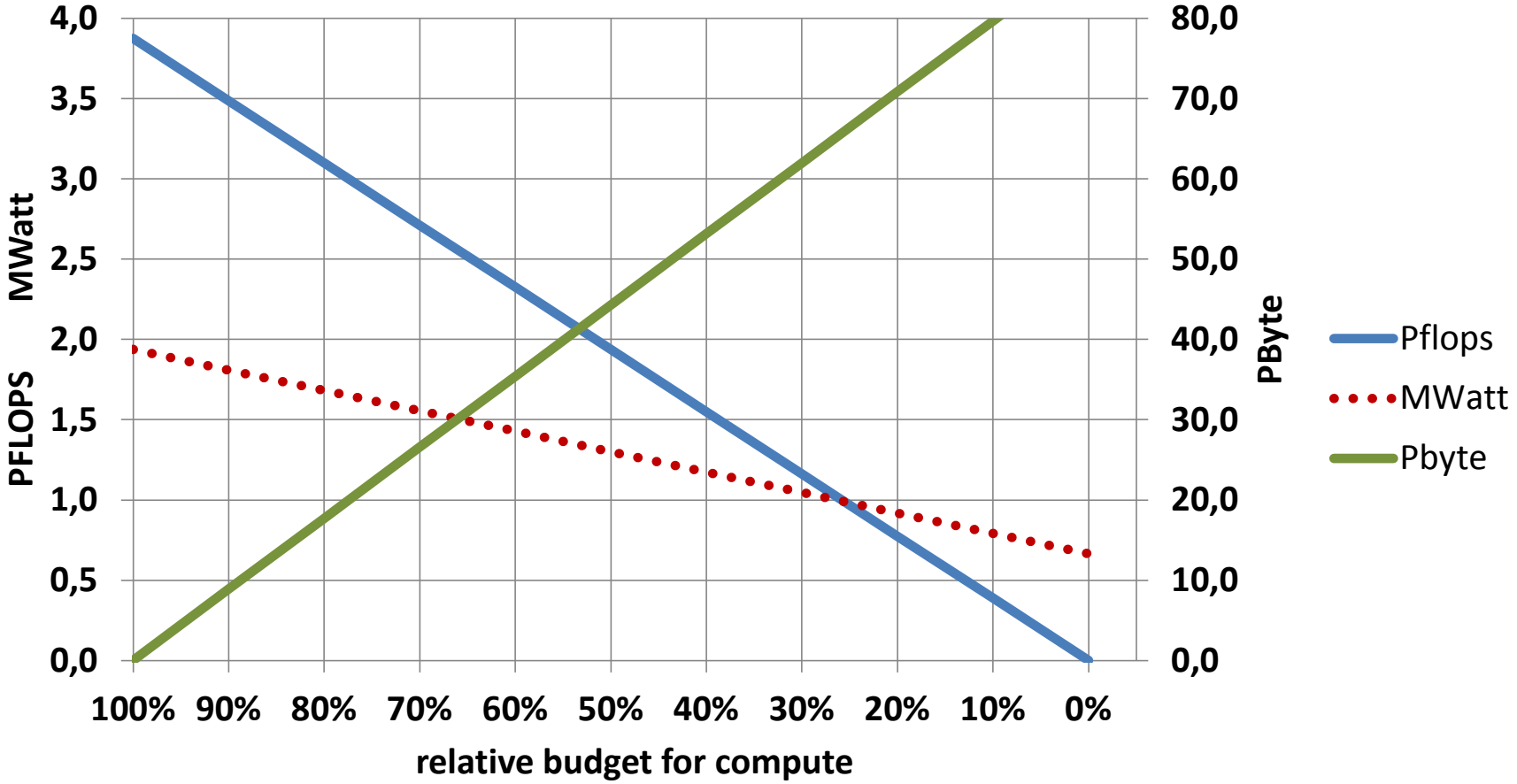
Planning Issues

- Procurement issues
  - How to distribute invest money onto compute and I/O?
- Operational issues
  - How much energy consumption?
  - How much tape consumption?

Technology Gap between Compute and I/O (>2 decades)



Investment into Compute and I/O (fall 2012 for Mistral)



Investment and Energy Costs for I/O

Measure	2009	2015	Factor
Performance	150 TFLOPS	3 PFLOPS	**20x**
Hard disk capacity	6 PB	50 PB	**9x**
Relative energy consumption of I/O	10%	20-25%	>2x
Relative investment costs for I/O	x%	y%	<2x?

From “Mistral” to “dkrz2020” (planning starts 2016)

Measure	2015	2020	Factor
Performance (no accelerators?)	3 PFLOPS	60 PFLOPS	**20x**
Main memory	200+ TB	2 PB	10x
Hard disk capacity	50 PB	500 PB	**10x**
Throughput memory to disk	400 GB/s	5 TB/s	13x
Tape library capacity (2020, 2025)	390 PB	1 EB	3x
Throughput disk to tape	20 GB/s	40 GB/s	2x
Power consumption	1.4 MW	1.4 MW	1x
Investment costs	€ 35M	€ 40M	1.15x

Cost Reductions

Dominating factors

- Energy consumption of disks
- Costs for tapes for mid and long term archival

Analysis

See paper and slides

“Exascale Storage Systems – An Analytical Study of Expenses”

In: Supercomputer Frontiers and Innovations, Vol. 1, Nr. 1, 2014
http://superfri.org/superfri/article/view/20

Data Reduction Techniques

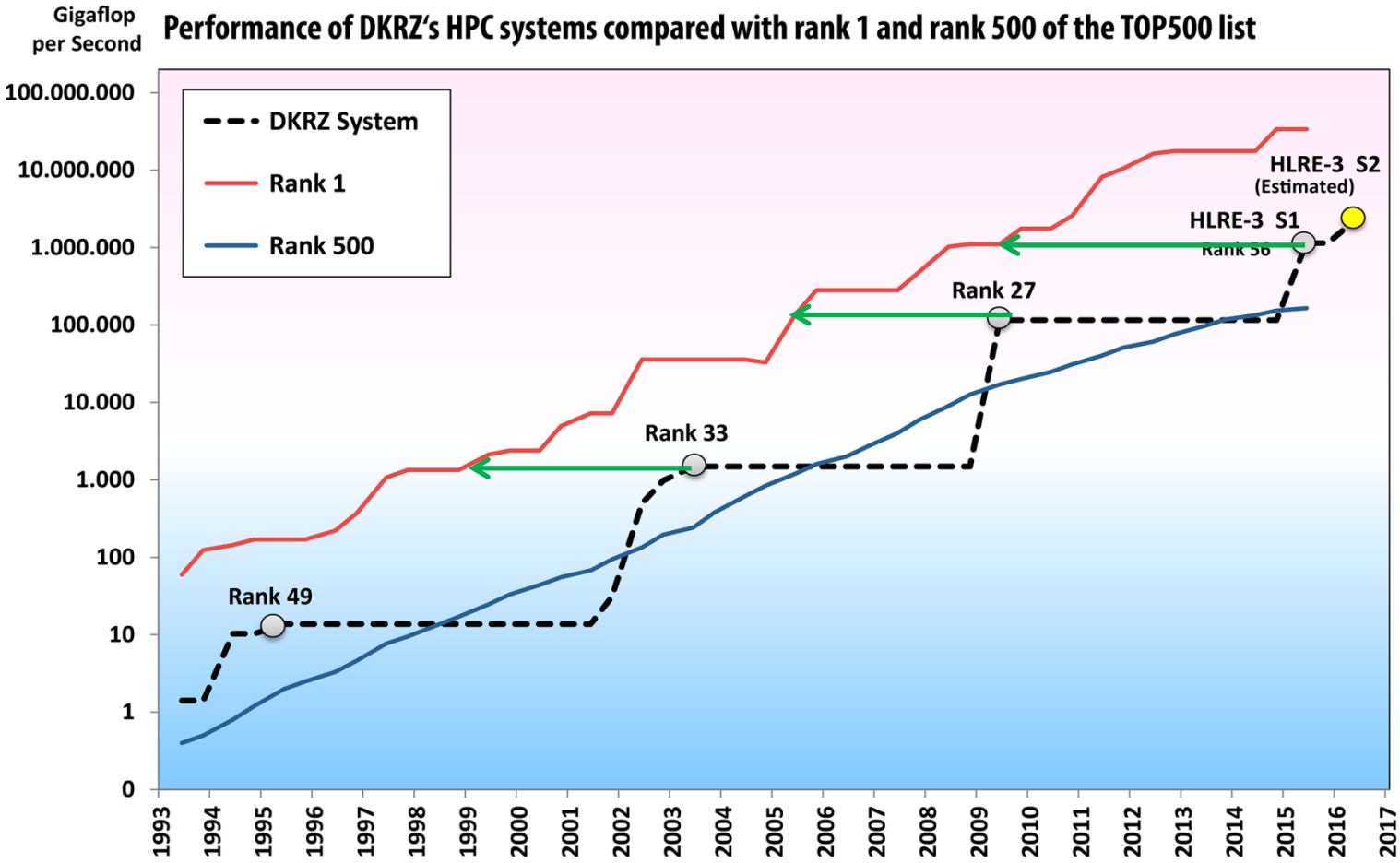
We compare

- Re-computation
  - Only for a very low number of accesses
- Deduplication
  - Not promising but can identify inefficient use of storage
- Compression (client and server side)
  - Can help to significantly reduce TCO
- User education
  - Most promising – most difficult

Extrapolations

- DKRZ´s role in the world
- DKRZ´s plan for Exascale

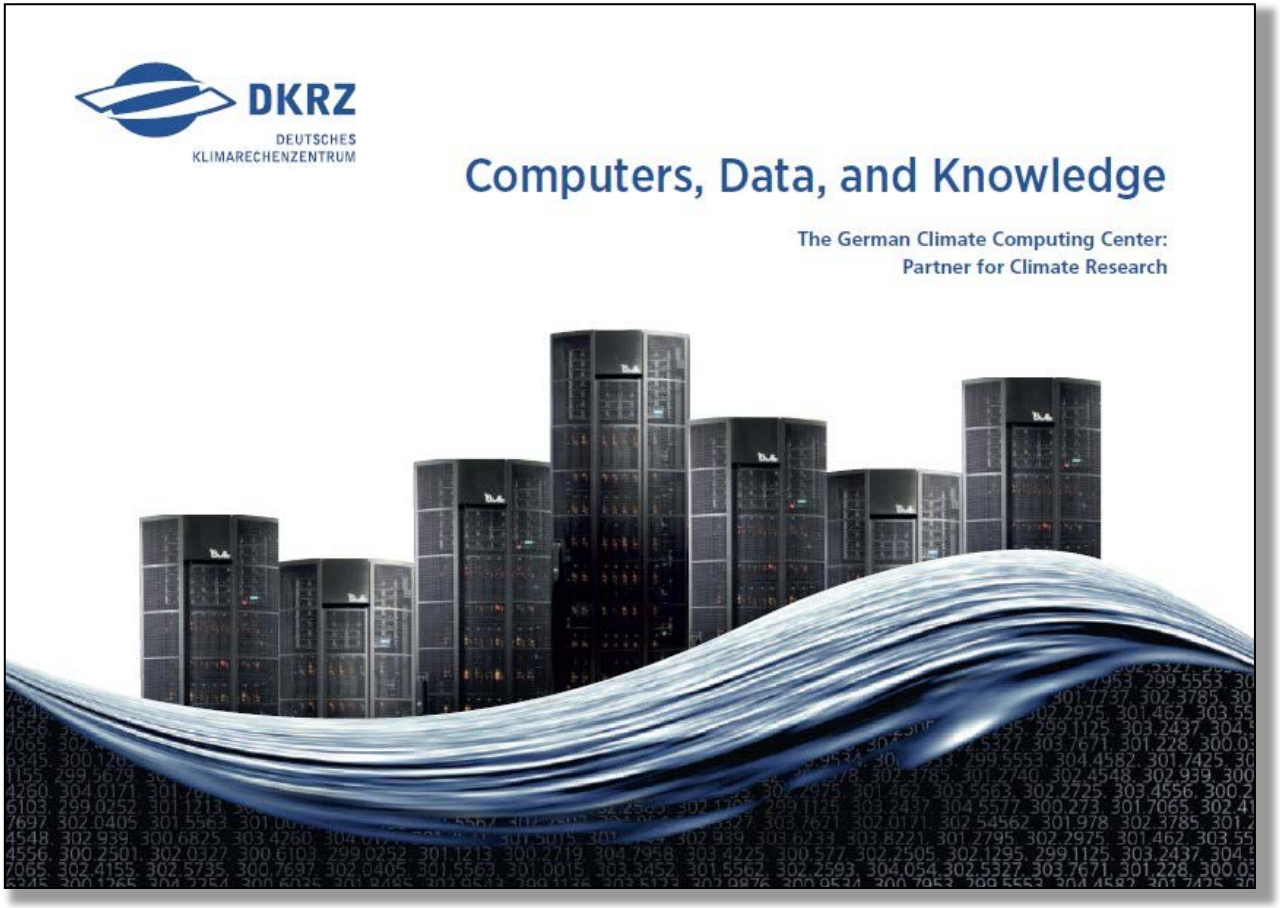
DKRZ in the TOP500 List



From “dkrz2020” to “dkrz-exa”

Measure	2020	2025	Factor
Performance (with accelerators?)	60 PFLOPS	**1.2 EFLOPS**	**20x**
Main memory	2 PB	**20 PB**	10x
Hard disk capacity	500 PB	**5 EB**	**10x**
Throughput memory to disk	5 TB/s	**65 TB/s**	13x
Tape library capacity (2025, 2030)	1 EB	**3 EB**	3x
Throughput disk to tape	40 GB/s	**80 GB/s**	2x
Power consumption	1.4 MW	**1.4 MW**	1x
Investment costs	€ 40M	**€ 48M**	1.2x

Status October 2015



New brochure at

www.dkrz.de

about us

media center

publications