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
- Exapolutions
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 1\textsuperscript{st}: “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”

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

![Graph showing the technology gap between CPU speed, HDD capacity, and HDD speed over time.](image_url)
Investment into Compute and I/O (fall 2012 for Mistral)
Investment and Energy Costs for I/O

<table>
<thead>
<tr>
<th>Measure</th>
<th>2009</th>
<th>2015</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>150 TFLOPS</td>
<td>3 PFLOPS</td>
<td>20x</td>
</tr>
<tr>
<td>Hard disk capacity</td>
<td>6 PB</td>
<td>50 PB</td>
<td>9x</td>
</tr>
<tr>
<td>Relative energy consumption of I/O</td>
<td>10%</td>
<td>20-25%</td>
<td>&gt;2x</td>
</tr>
<tr>
<td>Relative investment costs for I/O</td>
<td>x%</td>
<td>y%</td>
<td>&lt;2x?</td>
</tr>
</tbody>
</table>
From “Mistral” to “dkrz2020” (planning starts 2016)

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

Performance of DKRZ’s HPC systems compared with rank 1 and rank 500 of the TOP500 list
From “dkrz2020” to “dkrz-exa”

<table>
<thead>
<tr>
<th>Measure</th>
<th>2020</th>
<th>2025</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (with accelerators?)</td>
<td>60 PFLOPS</td>
<td>1.2 EFLOPS</td>
<td>20x</td>
</tr>
<tr>
<td>Main memory</td>
<td>2 PB</td>
<td>20 PB</td>
<td>10x</td>
</tr>
<tr>
<td>Hard disk capacity</td>
<td>500 PB</td>
<td>5 EB</td>
<td>10x</td>
</tr>
<tr>
<td>Throughput memory to disk</td>
<td>5 TB/s</td>
<td>65 TB/s</td>
<td>13x</td>
</tr>
<tr>
<td>Tape library capacity (2025, 2030)</td>
<td>1 EB</td>
<td>3 EB</td>
<td>3x</td>
</tr>
<tr>
<td>Throughput disk to tape</td>
<td>40 GB/s</td>
<td>80 GB/s</td>
<td>2x</td>
</tr>
<tr>
<td>Power consumption</td>
<td>1.4 MW</td>
<td>1.4 MW</td>
<td>1x</td>
</tr>
<tr>
<td>Investment costs</td>
<td>€ 40M</td>
<td>€ 48M</td>
<td>1.2x</td>
</tr>
</tbody>
</table>
New brochure at www.dkrz.de
about us
media center
publications