

Distributed Data Management at DKRZ

Wolfgang Sell

Deutsches Klimarechenzentrum GmbH

sell@dkrz.de

Table of Contents

- **DKRZ - a German HPC Center**
- **HPC Systemarchitecture suited for Earth System Modeling**
- **The HLRE Implementation at DKRZ**
- **Some Results**
- **HLRE2 (if time permits)**
- **Summary**

DKRZ - a German HPCC

- **Mission of DKRZ**
- **DKRZ and its Organization**
- **DKRZ Services**
- **Model and Data Services**

Mission of DKRZ

In 1987 DKRZ was founded with the Mission to

- ***Provide state-of-the-art supercomputing and data service to the German scientific community to conduct top of the line Earth System and Climate Modelling.***
- ***Provide associated services including high level visualization.***

DKRZ and its Organization (1)

Deutsches KlimaRechenZentrum = *DKRZ*
German Climate Computer Center

- organised under private law (GmbH) with 4 shareholders
- investments funded by federal government, operations funded by shareholders
- usage 50 % shareholders and 50 % community

DKRZ and its Organization (2)

DKRZ internal Structure

- **3 departments for**
 - systems and networks
 - visualisation and consulting
 - administration
- **20 staff in total**
- until restructuring end of 1999 a fourth department (now called Model & Data) supported climate model applications and climate data management

DKRZ Services

- operations center: **DKRZ**
 - technical organization of computational resources (compute-, data- and network-services, infrastructure)
 - advanced visualisation
 - assistance for parallel architectures (consulting and training)

Model & Data Services

competence center: *Model & Data*

- professional handling of community models
- specific scenario runs
- scientific data handling

Since Januar 2000 the Model & Data Group is external to DKRZ, administered by MPI for Meteorology, funded by BMBF

HPC Systemarchitecture suited for Earth System Modeling

- **IT Requirements for ESM**
- **Principal HPC System Configuration**
- **Links between Different Services**
- **The Data Problem**

General Requirements

- Earth System Modelling is both Compute and Data Intensive
- High Bandwidth needed between the Coupled Servers
- Scalability to be supported by Operating System (Linux)
- Record Level Access to Data with High Performance
- Support of Semantic Data Handling Mandatory

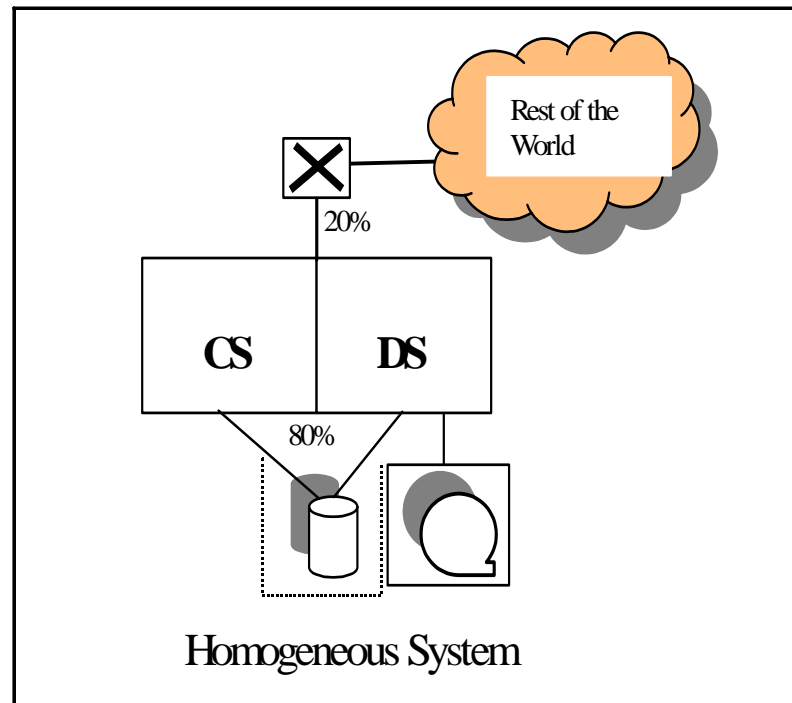
Development of Data Services

- **location oriented (1984 - 1991)**
 - **media administrated by users**
- **filename oriented (since 1992)**
 - **media transparent, HSM, unlimited disk space**
- **context oriented (since 1998)**
 - **metadata driven TOC, layered on top of HSM**
- **clustered distributed DS (since 2002)**

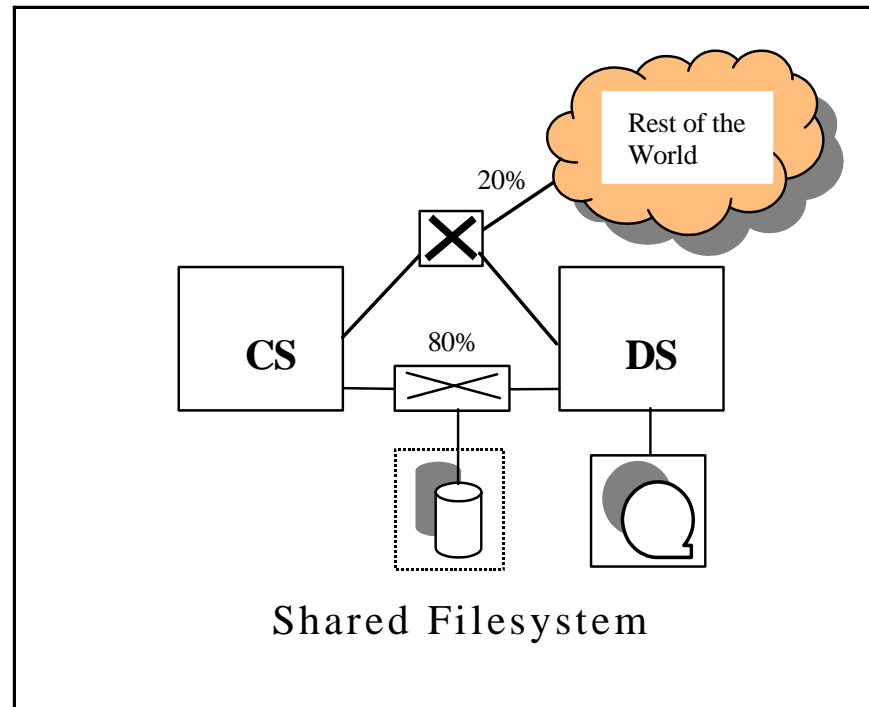
Final System architecture change

- **Data Server was until 2002 a monolithic system, hard to scale for higher loads**
- **clustered distributed DS**
 - **Clients organized into SMP cluster**
 - **DS also physically distributed**
 - **shared FS between CS/DS/VS**
 - **SAN architecture**
 - **easy to scale and adapt to increasing demands**

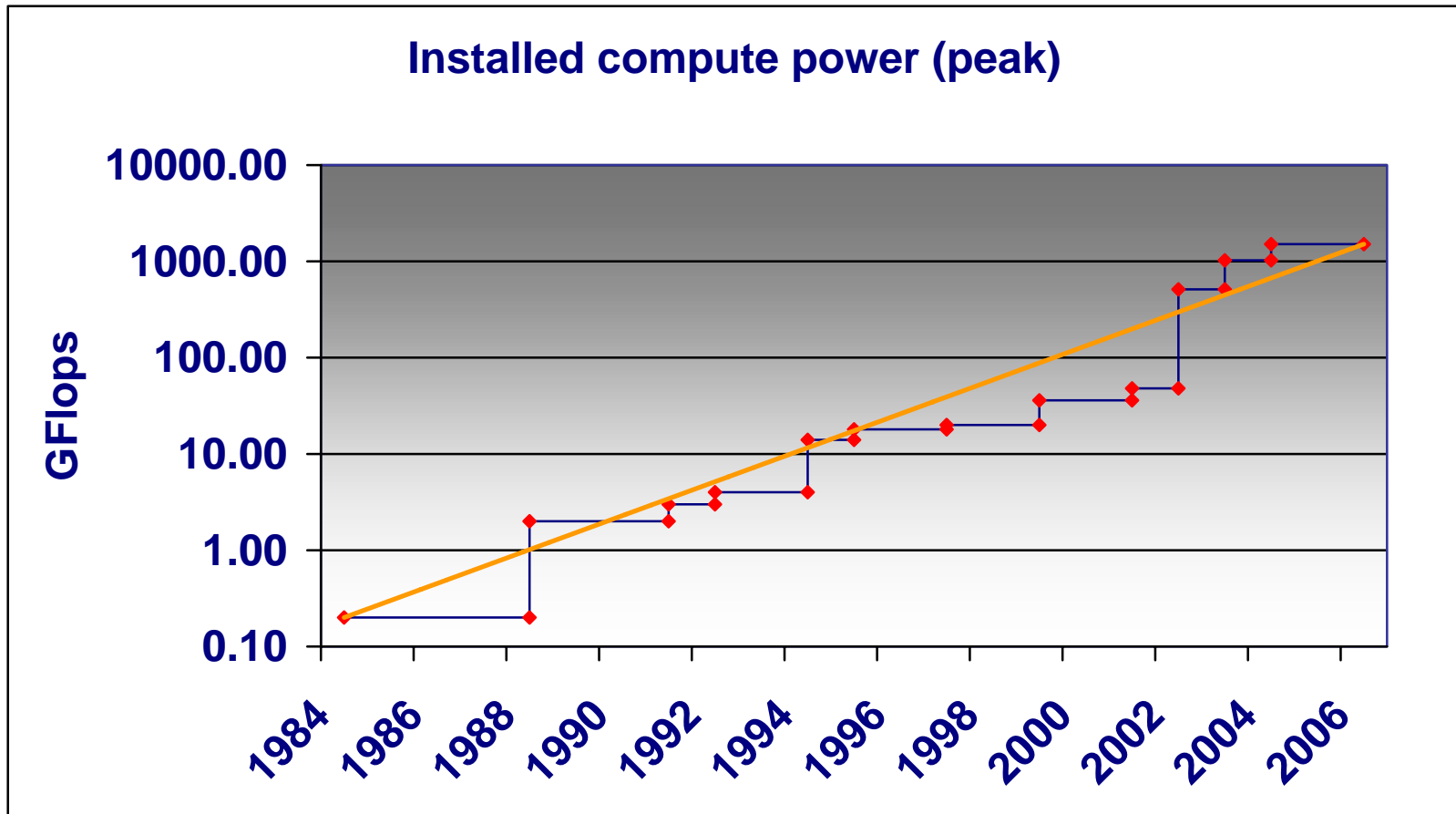
Variants of System Configuration (1)



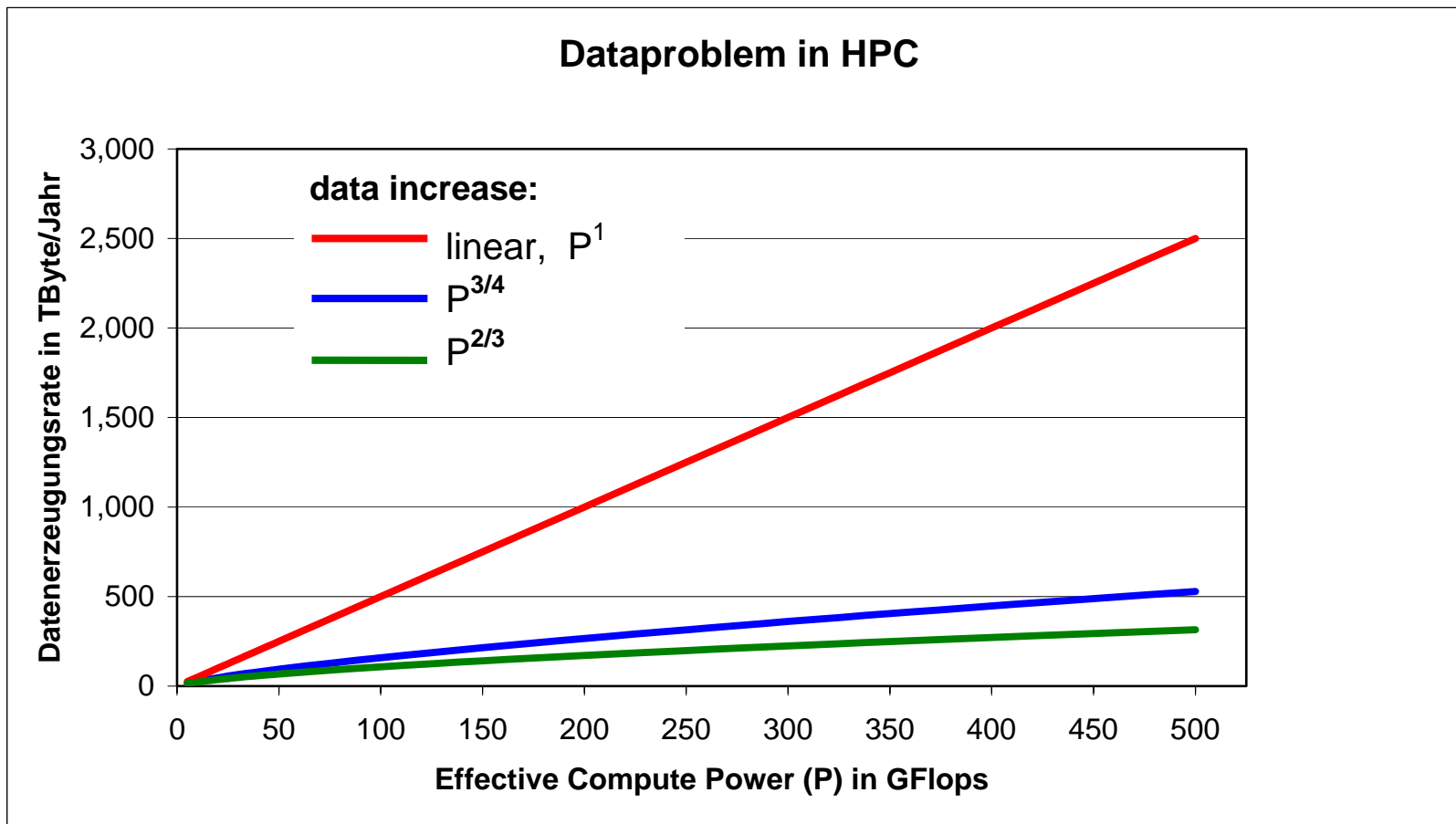
Variants of System Configuration (2)



Compute server power



Adaptation Problem for Data Server



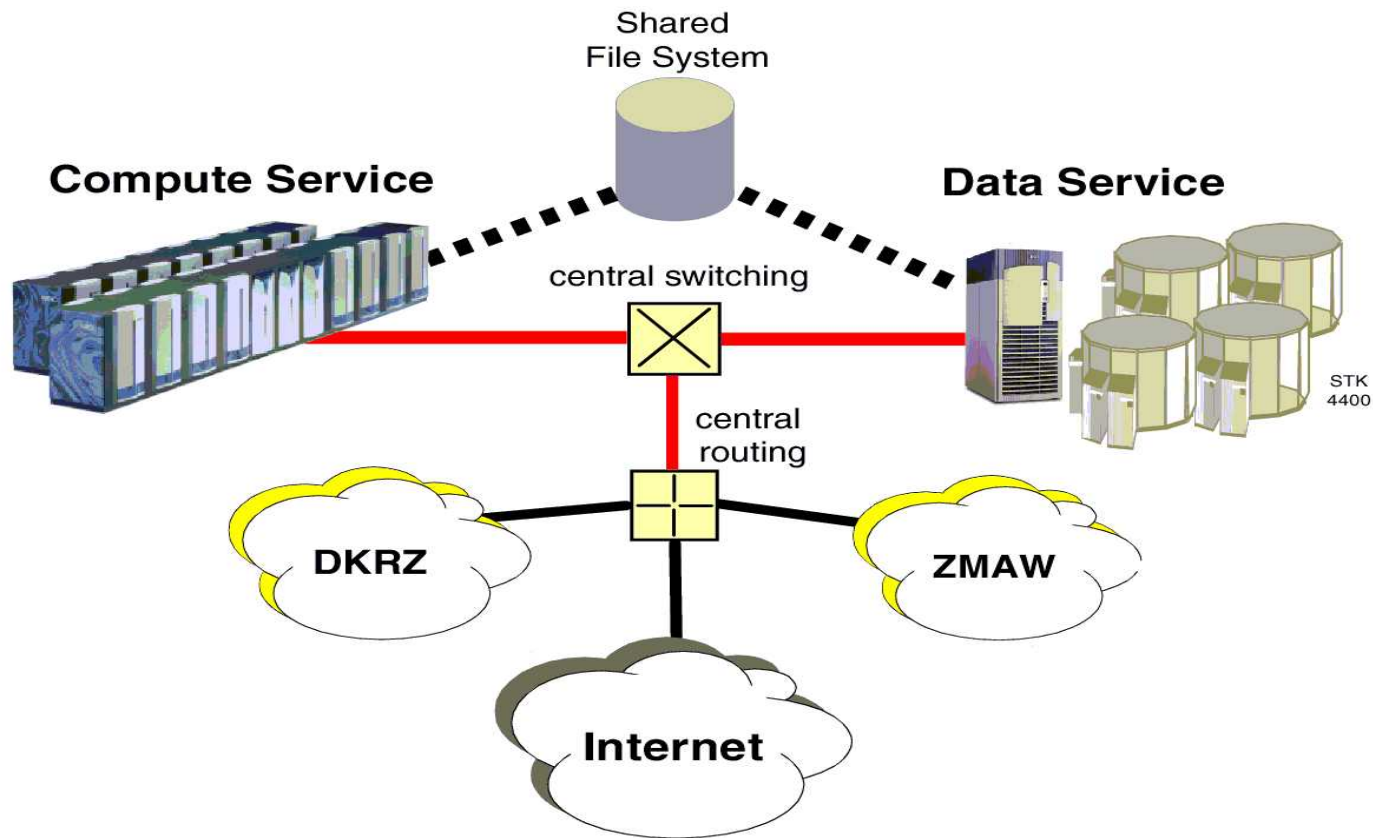
HLRE Implementation at DKRZ

Höchst**L**eistung**R**echnersystem für die **E**rdsystemforschung
= **HLRE**

**High Performance Computer System for Earth System
Research**

- **Principal HLRE System Configuration**
- **Requirements and Constraints**
- **Links between Different Services**
- **Option for Systemoperation**

Principal HLRE System Configuration

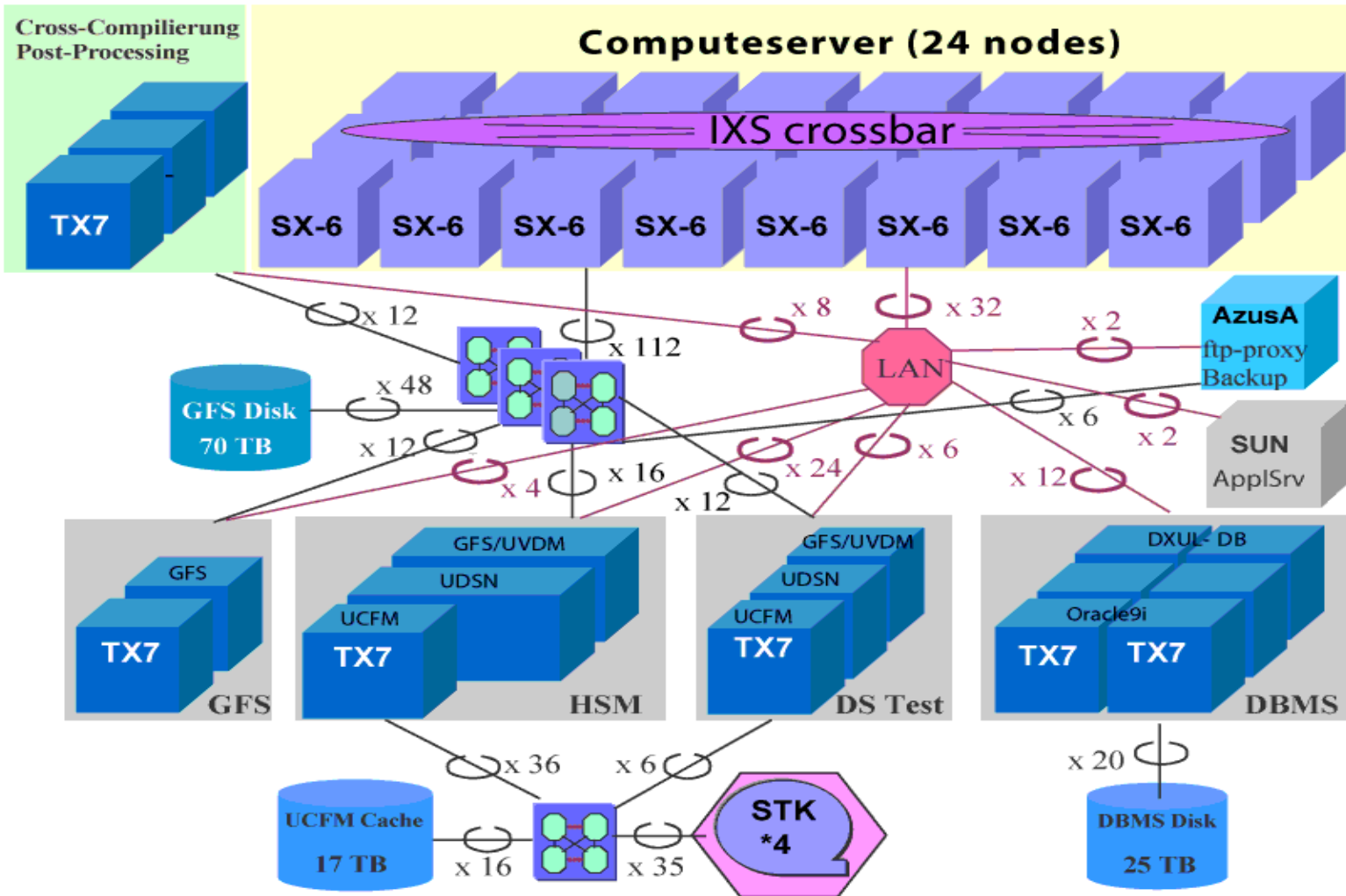


Current Hardware at DKRZ

- **24 SX-6 Nodes**
(192 Vector CPUs, 1,5 TByte CM and 1,5 Tflops peak)
- **IXS Crossbar switch**
(24 x 24, 2*8*24 GByte/s cross section bandwidth)
- **10 NEC AsAmA Nodes**
(132 Itanium-2, 1,0 and 1,5 GHz, Linux)
- **1 NEC Azusa**
(8 Itanium-1; 800 MHz; Linux)
- **6 STK Silos**
(total capacity ca. 6 / 30 PetaByte)
- **4 SUN Fire 4800 (Oracle Appl. Service)**

DKRZ Hardware (Phase 3)

Current Configuration



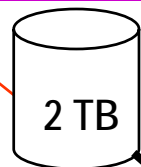
Filesystem Systematics

CS View

permanent

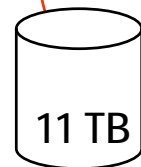
transient

\$HOME
all nodes
for ever
quota



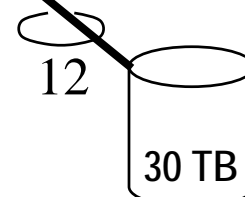
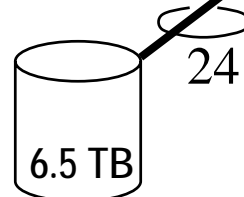
\$TMPDIR
node local
job temporal

\$UT
all nodes
for ever

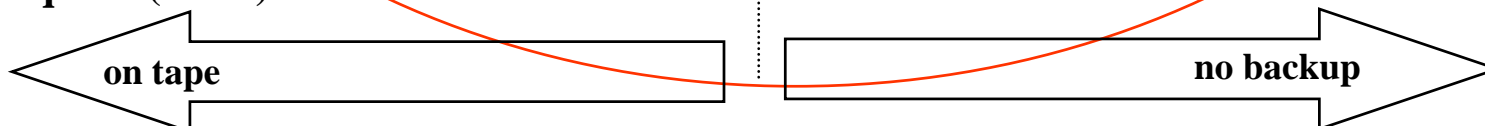


\$WRKSHR
all nodes
O(days)
quota

\$UTF
all nodes
1 year
quota (#files)



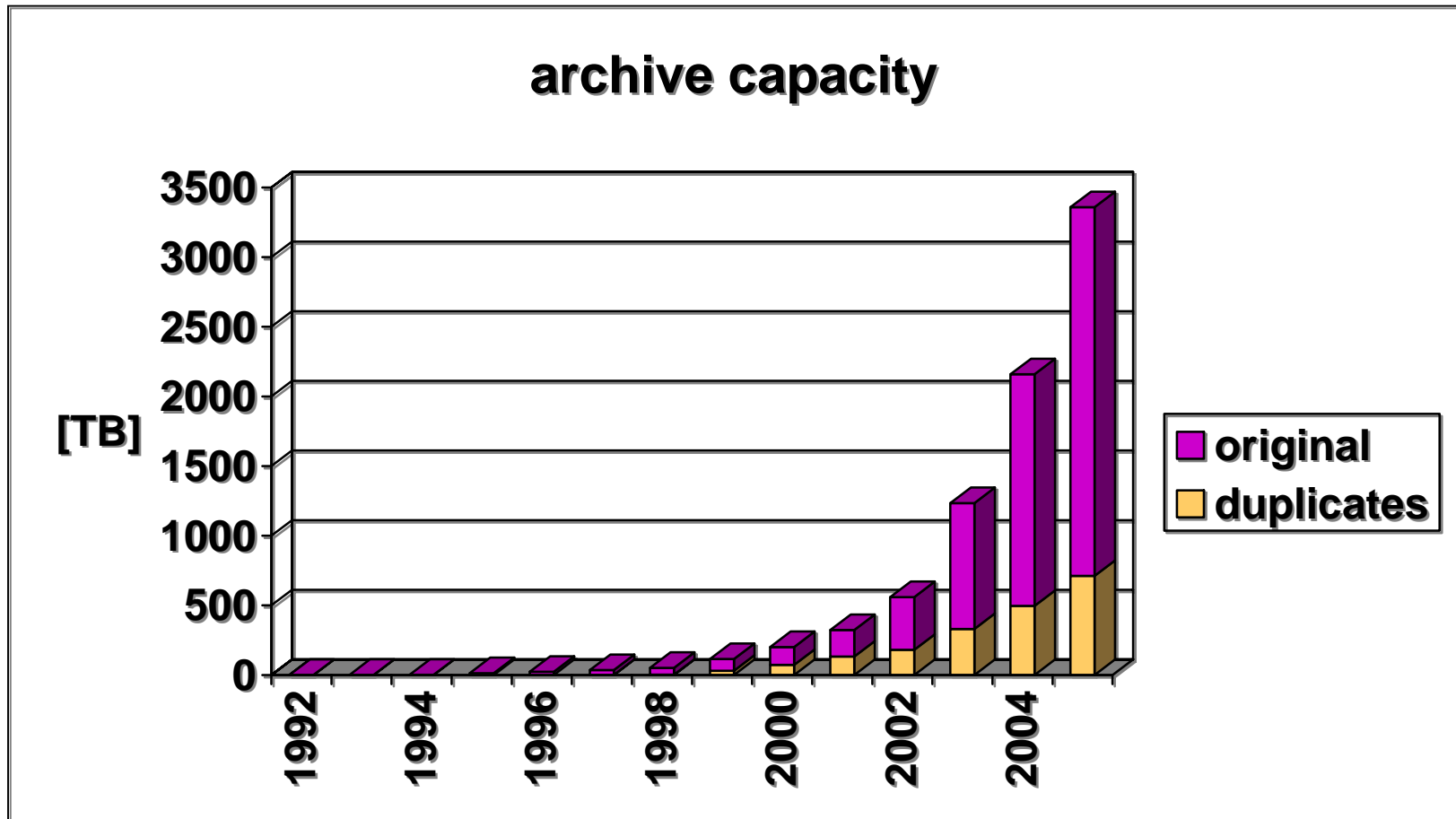
\$TMPSHR
all nodes
O(weeks)
quota



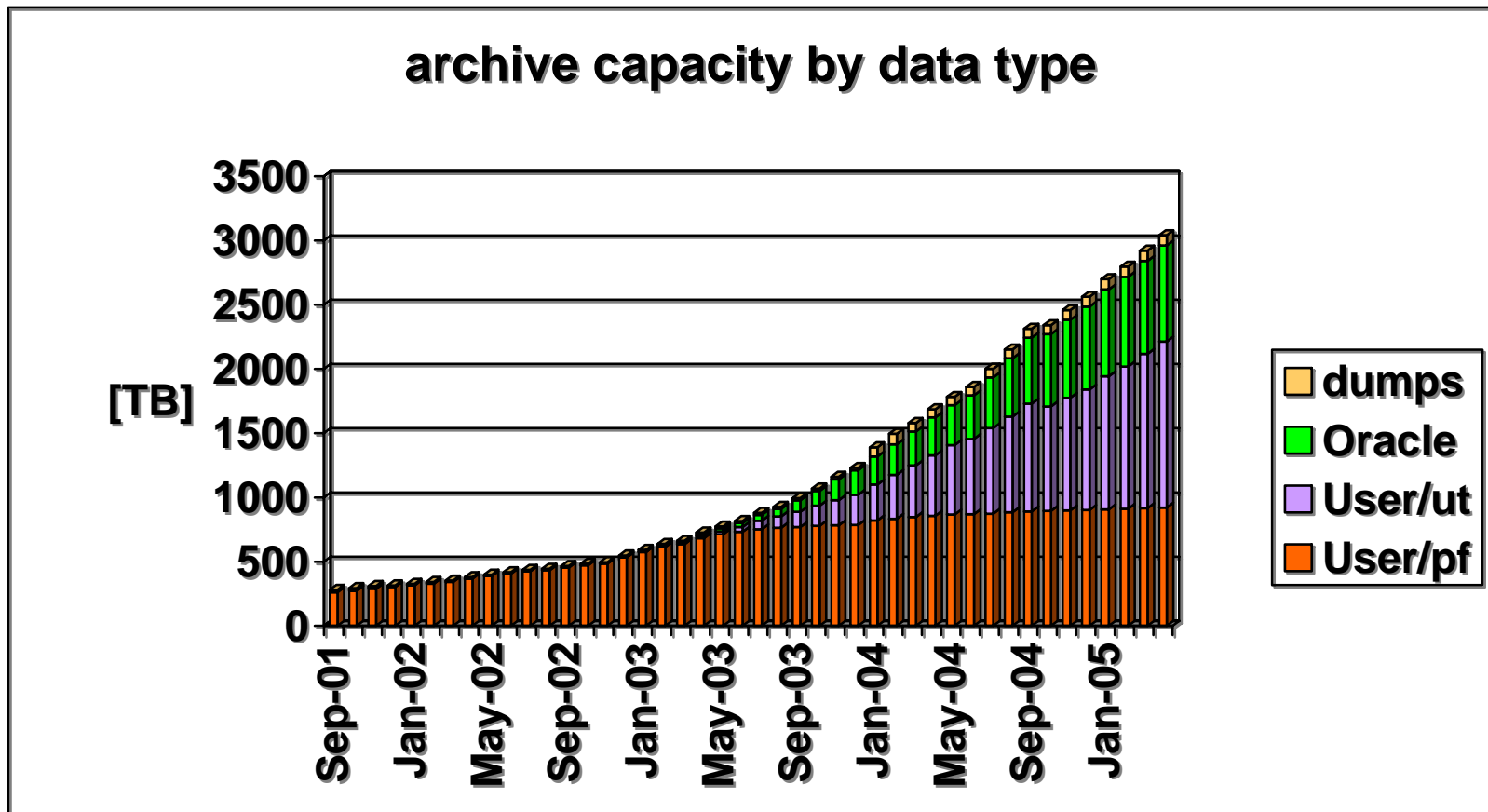
Some Results

- **Growth of the Data Archive**
- **Growth of Transferrate**
- **Lessons Learnt**

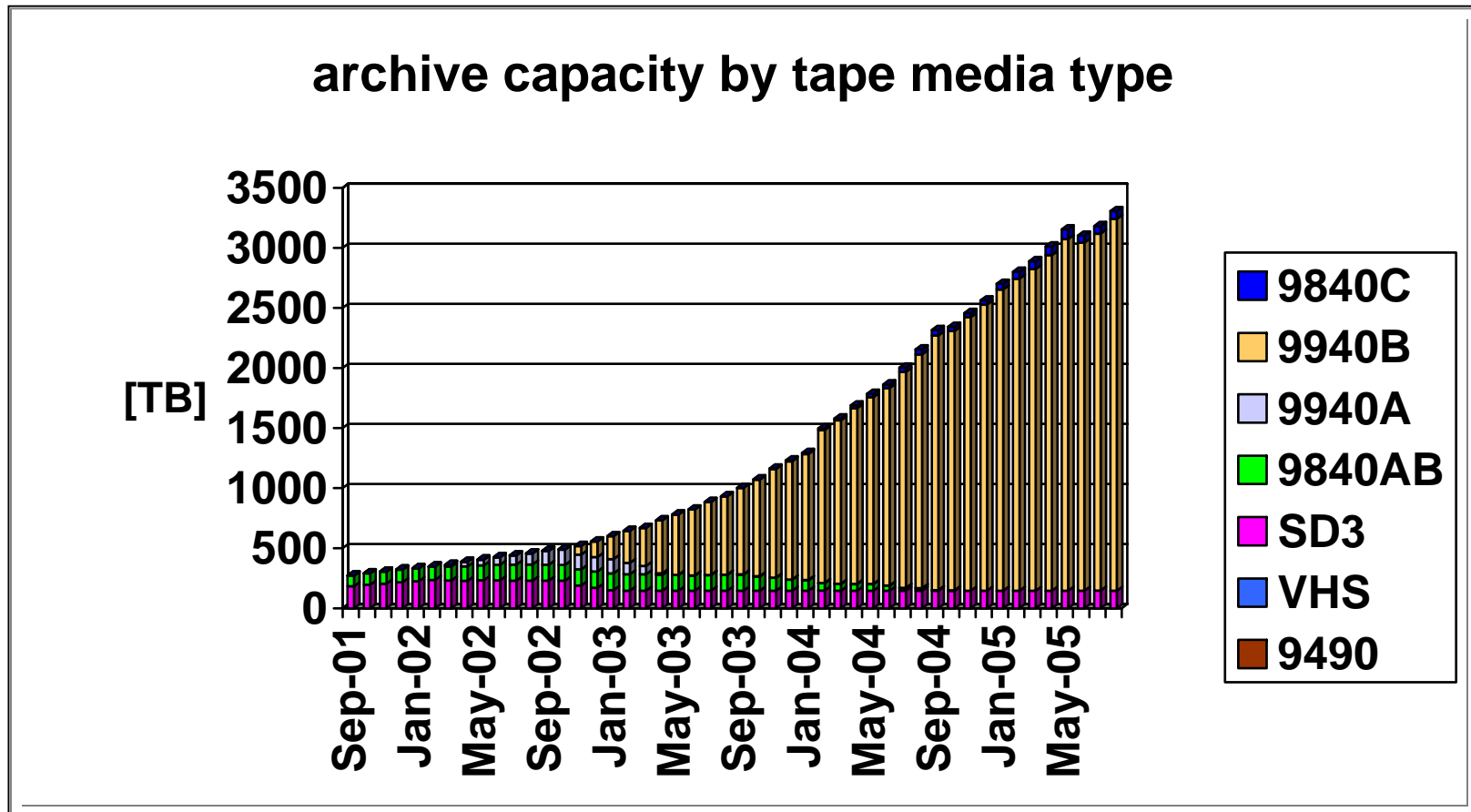
DS archive capacity (1)



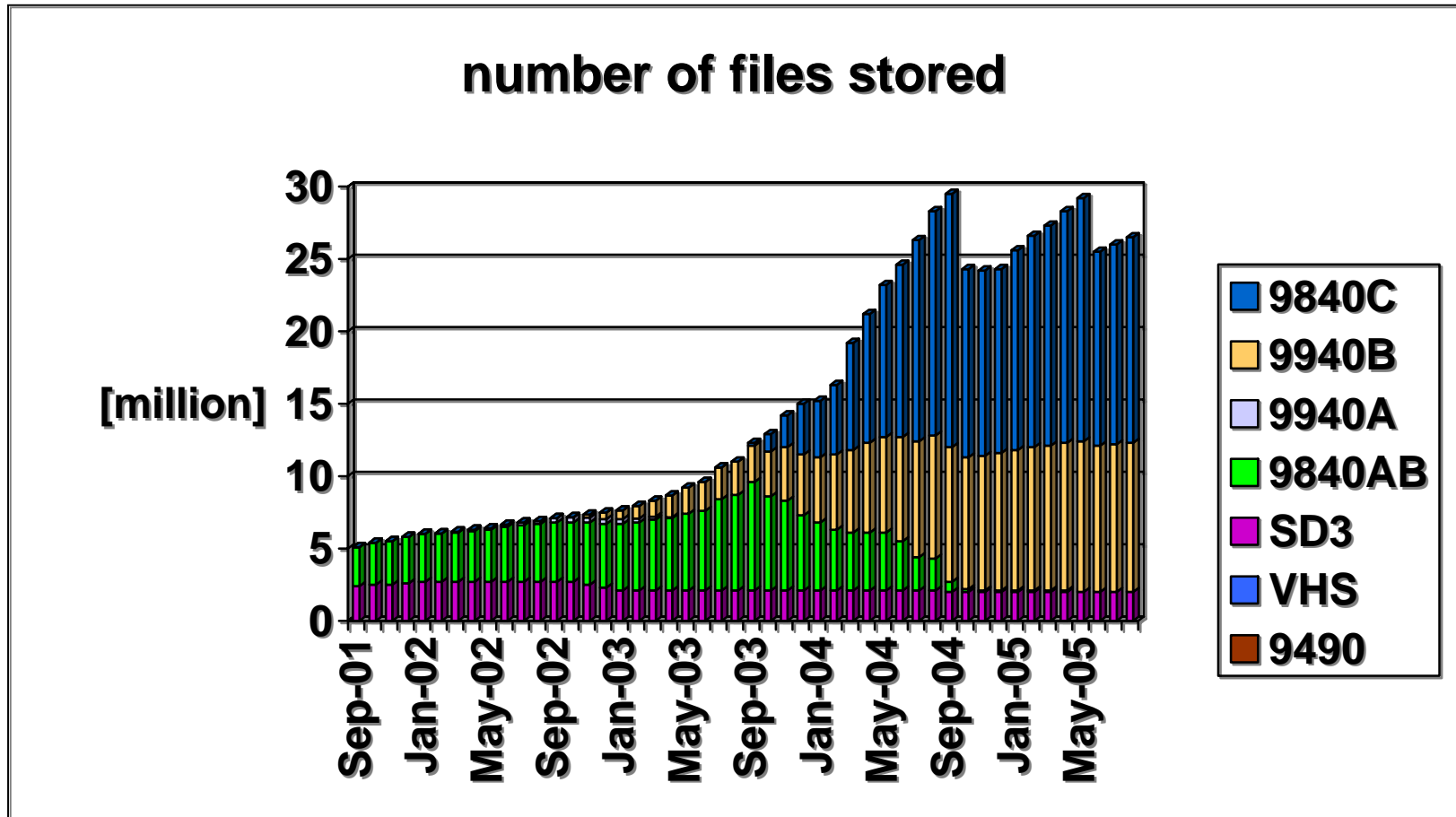
DS archive capacity (2001-2005)



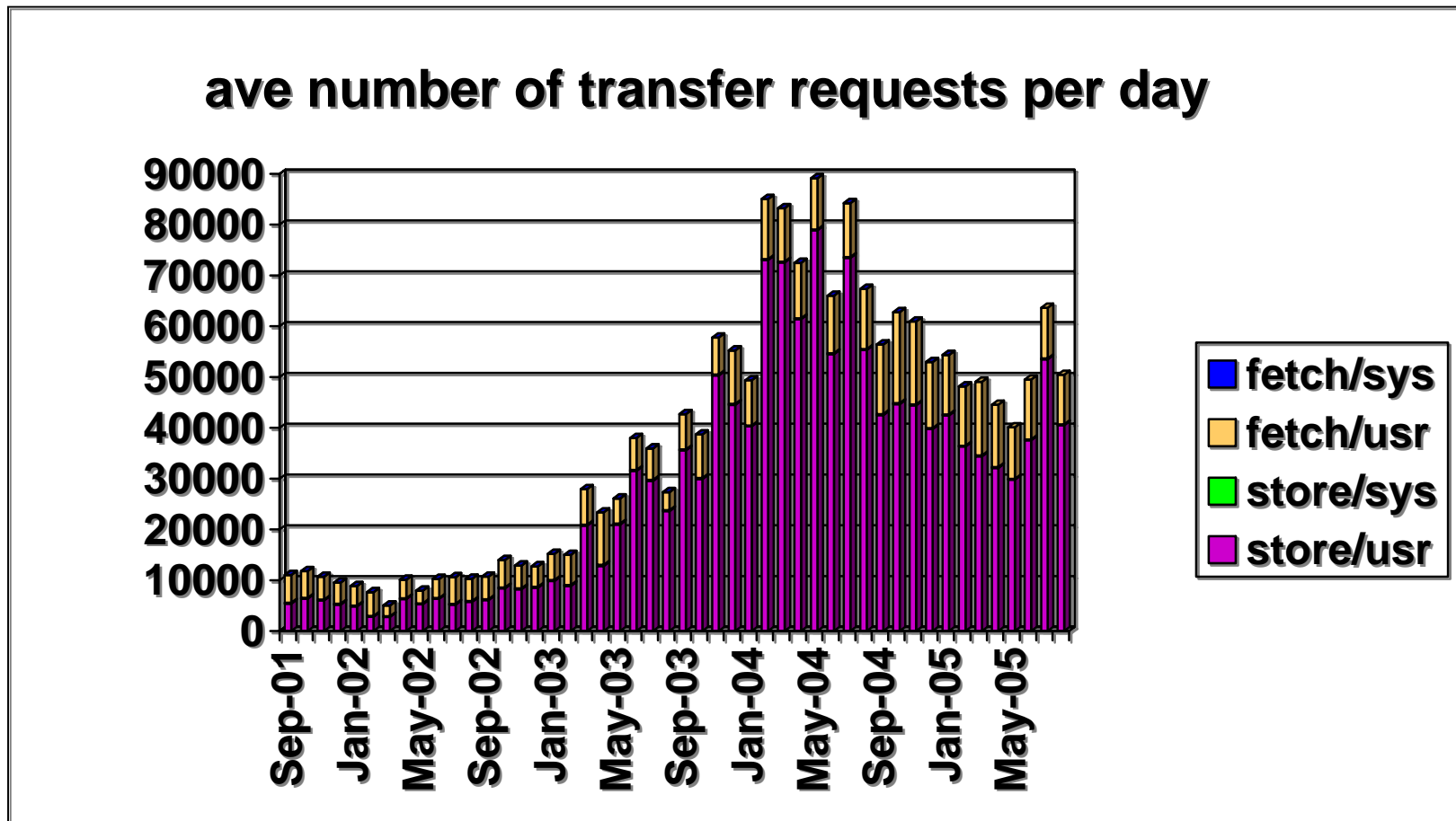
DS archive capacity (2001-2005)



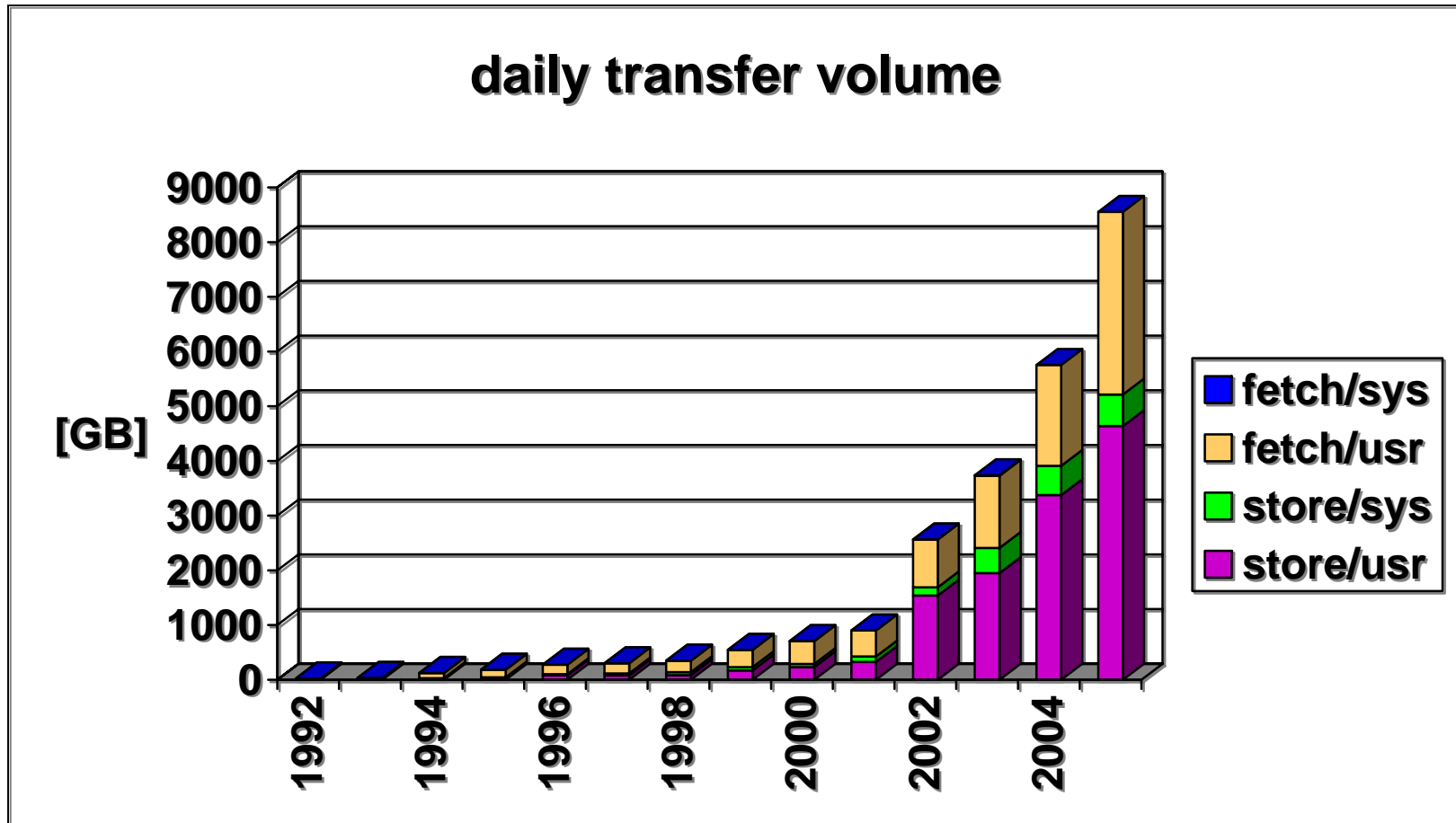
DS archive capacity (2001-2005)



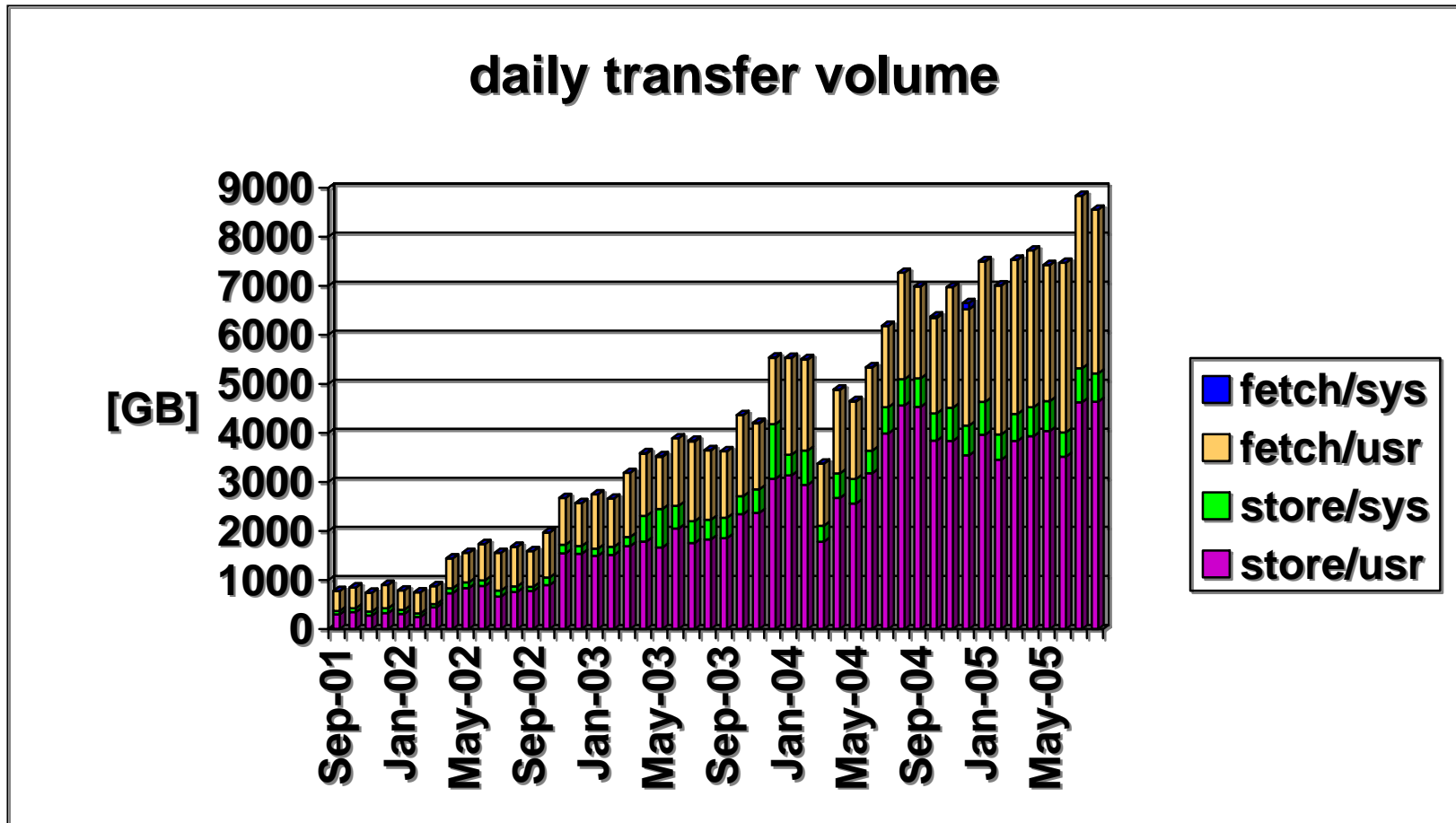
DS transfer requests (2001-2005)



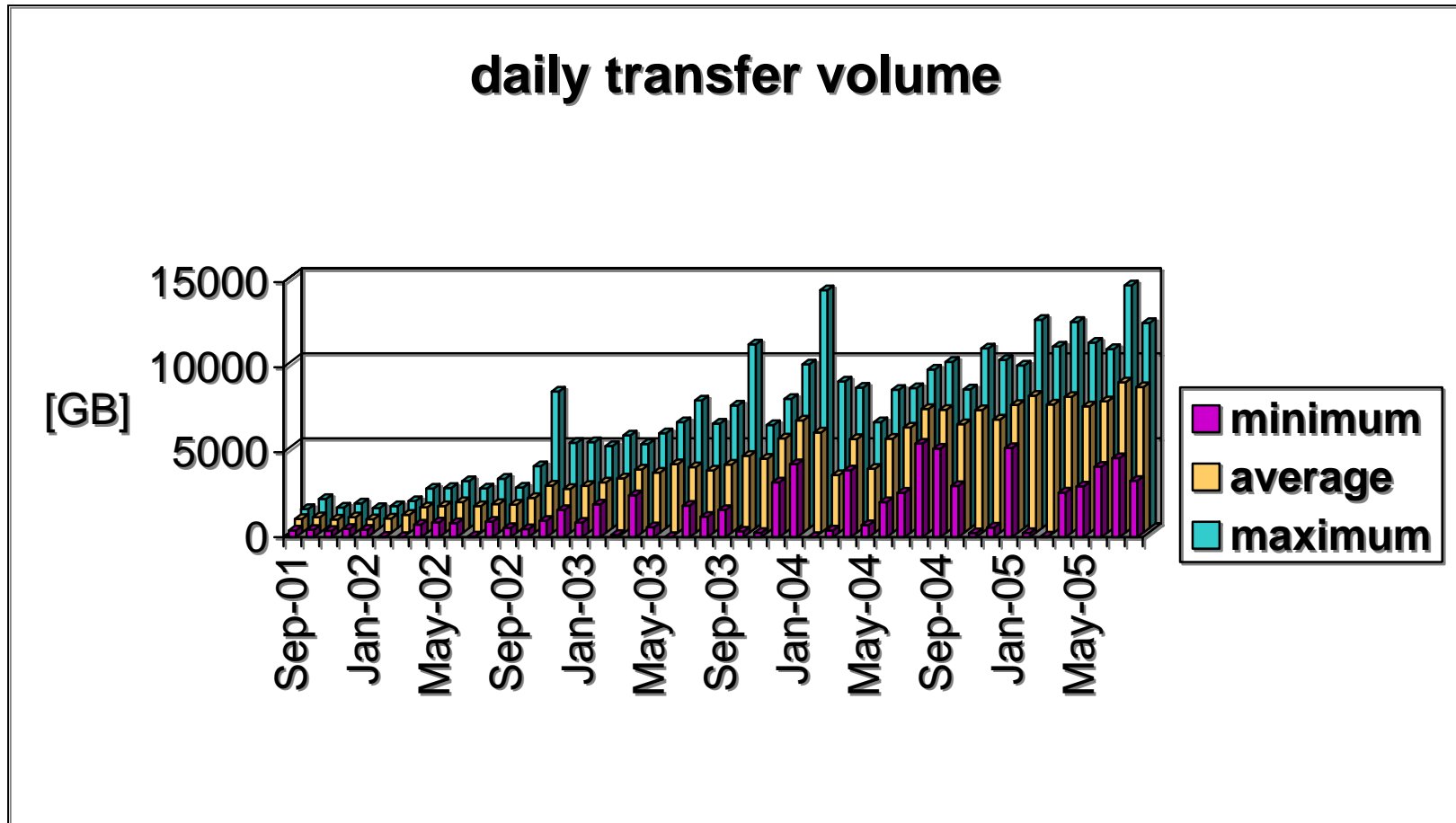
DS transfer rates (1)



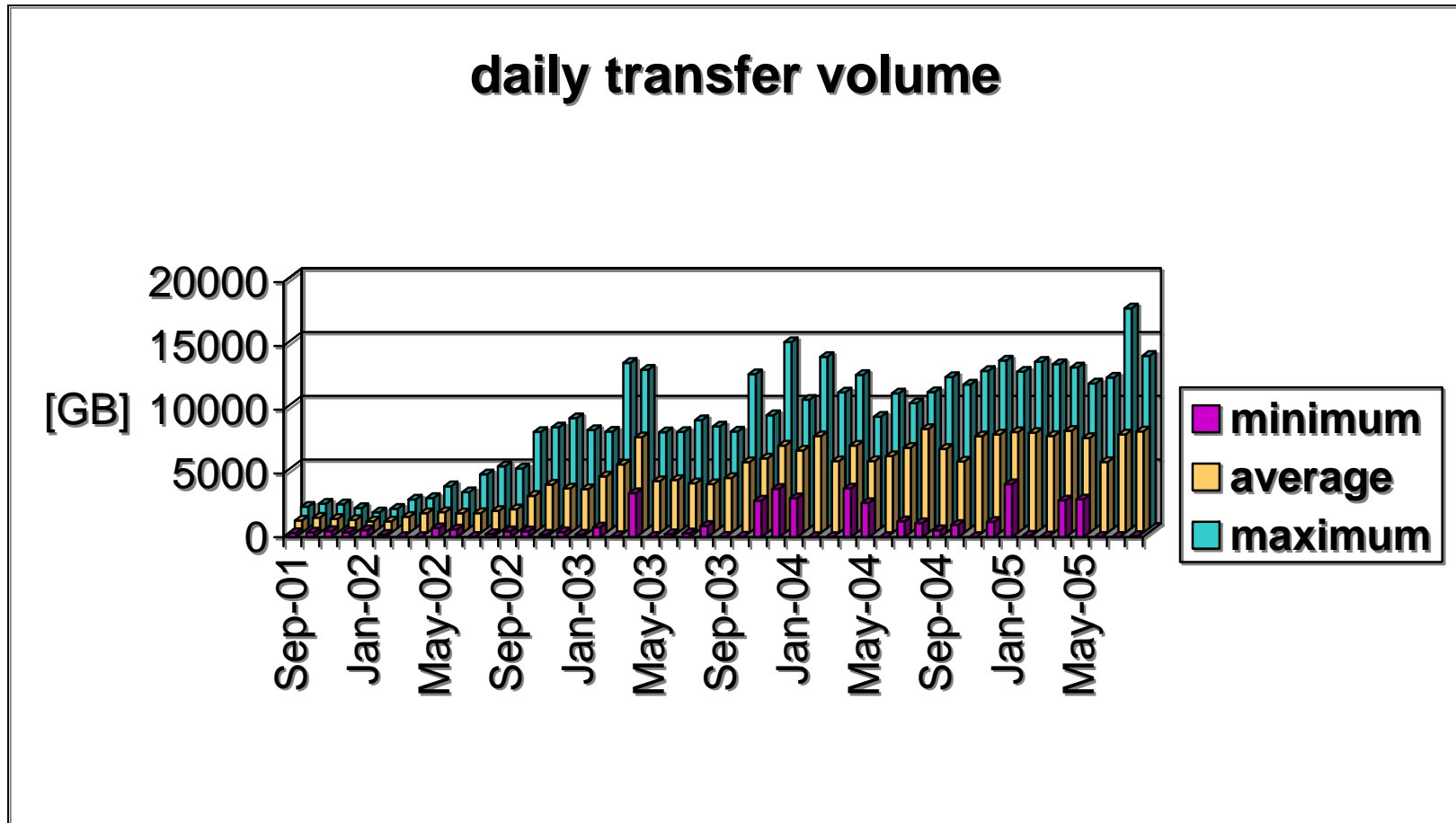
DS transfer rates (2001-2005)



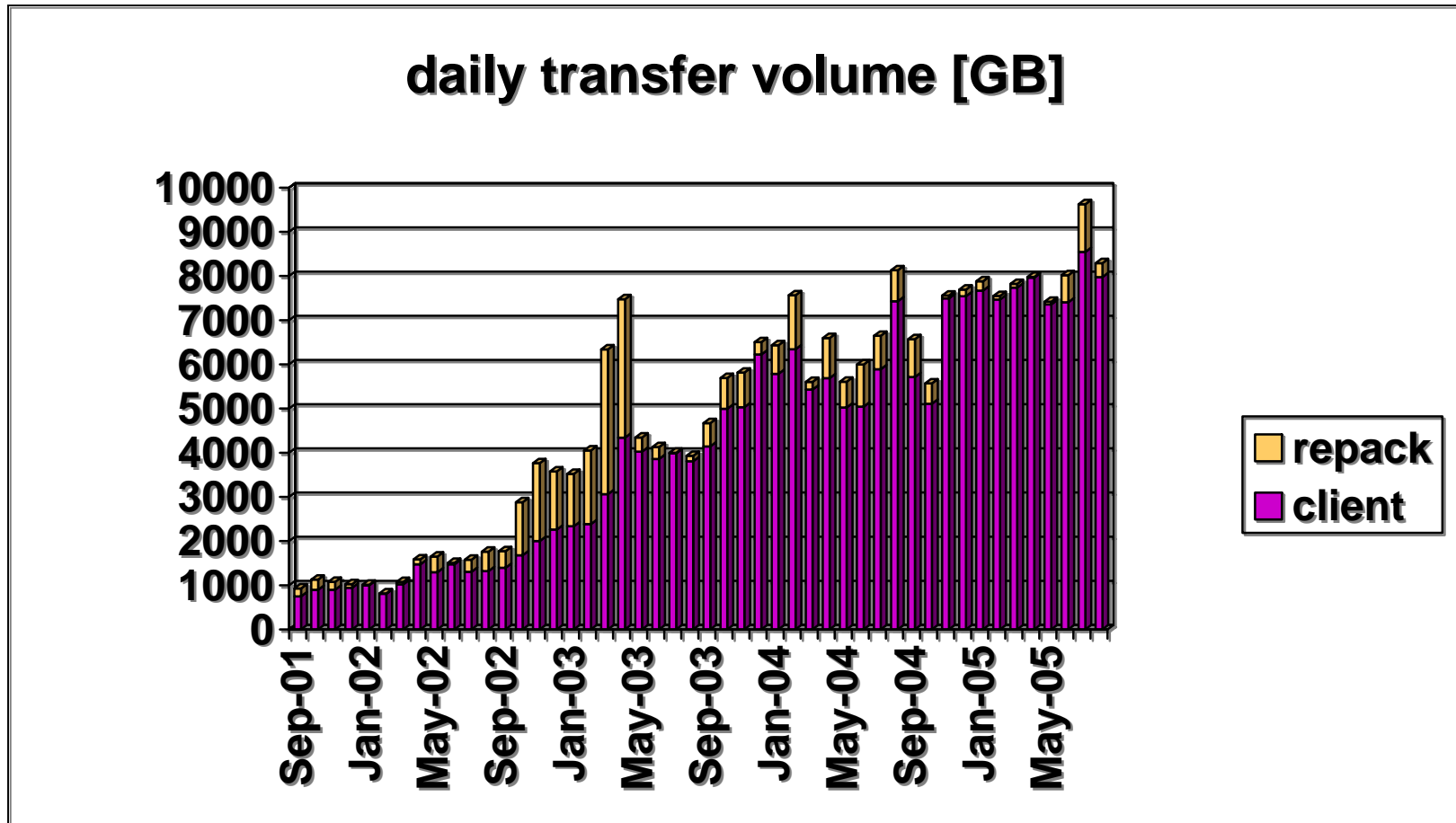
DS transfer rates (2001-2003)



Tape transfer rates (2001-2005)



Tape transfer rates (2001-2005)



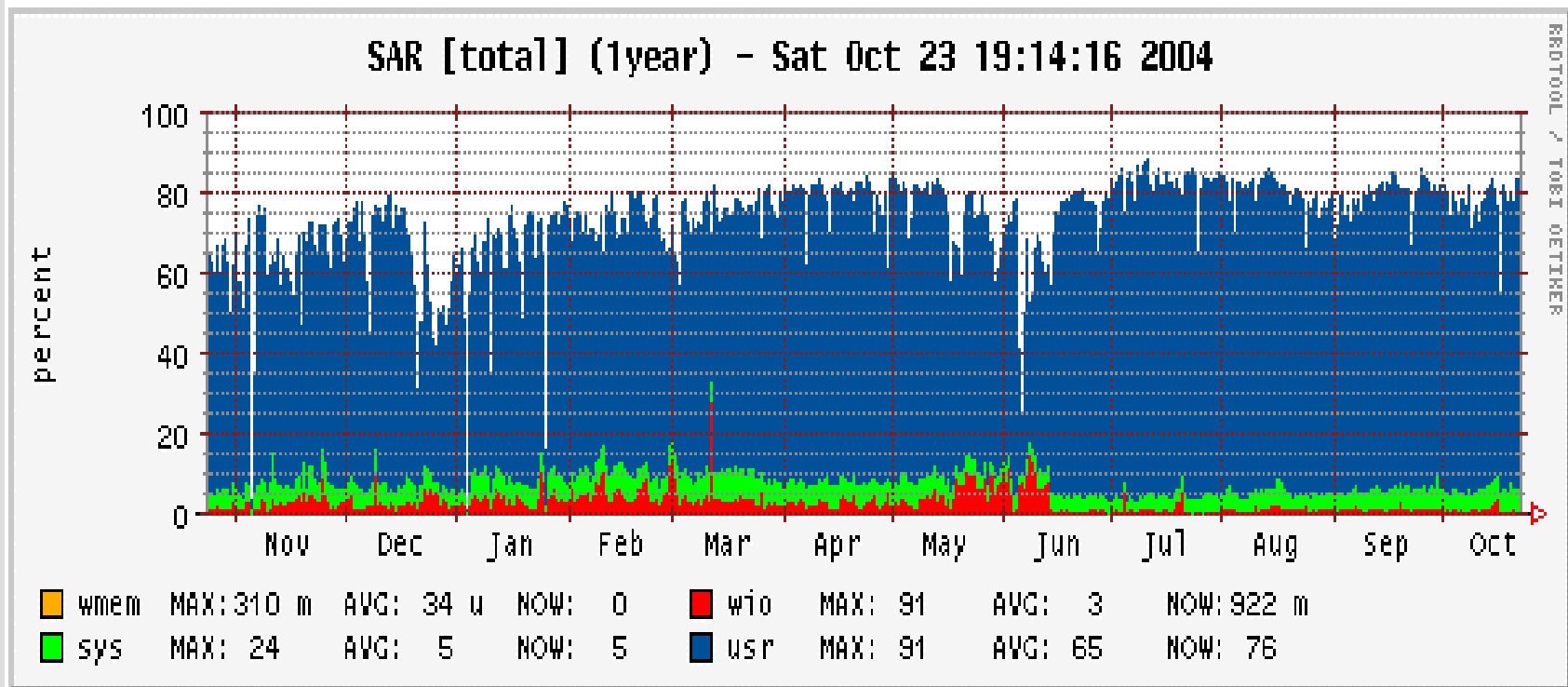
Some Lessons Learnt

- **Current Implementation of Non-Computing Services needs Significant Amount of Local Disk Space, e.g. HSM and DBMS need their Own Cache**
- **Lack of Standardisation for Shared Filesystems results in Dependence on Co-operativeness, e.g. Graphics Server Integration Pre/Post-Processing Servers from Different Vendors**
- **Fail-over Solutions needed in Complex Distributed Systems**

Some Lessons Learnt, cont.

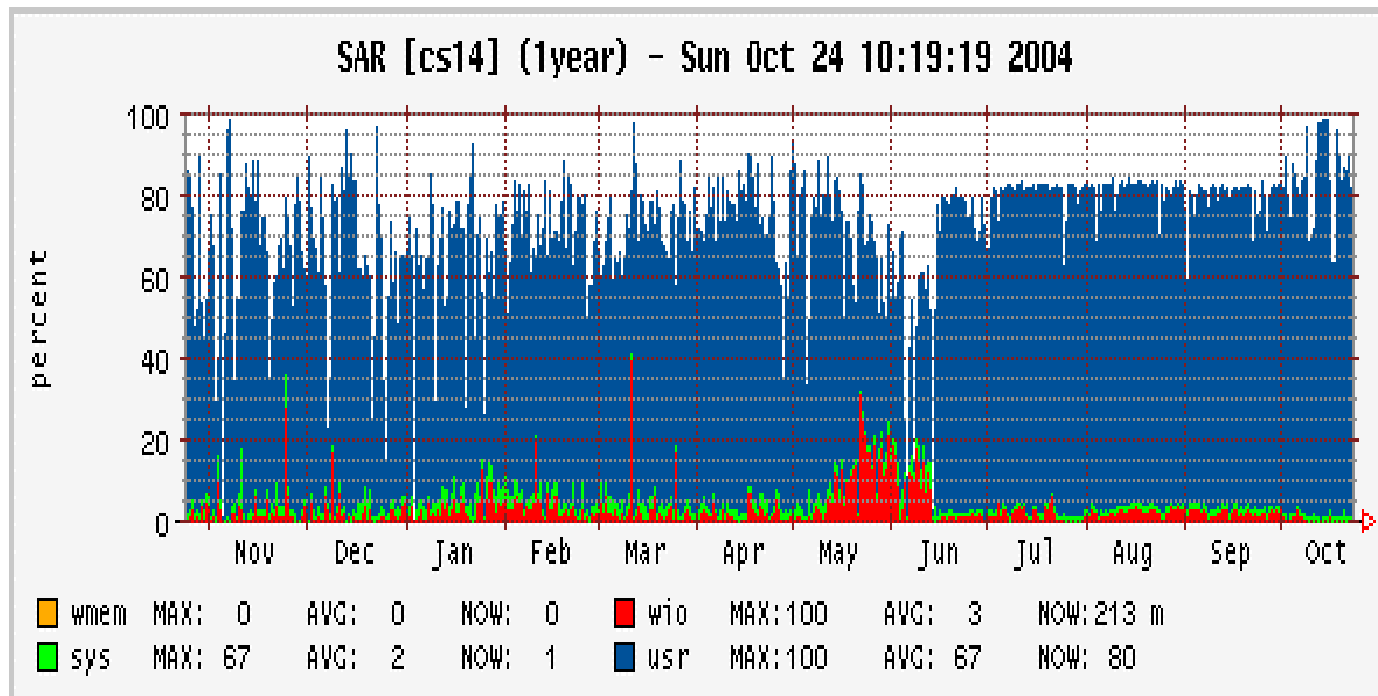
- **Server Scalability needed, but no Problem
Client Scalability may be a Problem, e.g
128 LUN Limitation for Linux 2.4**
- **Distributed Servers may Generate Intriguing
Dependencies, i.e. clearly Structured High Level
Services do not Guarantee Ease of Performant
Operation**

Effect of Client/Server Interaction



Invocation Period and Lifetime of Dirty Pages for kupdated

Effect of Client/Server Interaction



Invocation Period and Lifetime of Dirty Pages for kupdated

HLRE2

- **Time Schedule**
- **Basic Requirements**
- **Expected Compute Growth Path**
- **Demands for Data Services**
- **Observed Transferrates for HLRE**

(Personal View of the former DKRZ General Manager)

Reasons for Compute Intensity (and Need for More IT-Resources)

- **Spatial Model Resolution**
- **Complexity**
- **Reduction of Uncertainty**
- **Long Integration Periods**

- **Progress in Climate Modelling is Limited by Available Compute Power**

Resulting Expectations / Demands

Topic	Computing	Increase
Resolution:	300 -> 50 km	x 100
Complexity:	Clouds, Chemistry etc	x several
Uncertainty:	Ensemble Runs (30)	x 10
	Quantification	x 10 - 1000
Longer Integration Times		x 10

Source: John Mitchell, Hadley Centre

HLRE2 Time Frame

- Contract in May 2006
- Installation Phase 1 in Jan. 2007
- Installation Phase 2 in 1Q2009
- Decommissioning HLRE2 2011/12

(Since my successor as GM has not been determined the schedule will very likely be delayed)

Basic Requirements

- Fixed Budget for all Components
- Extended Warranty for 5 yrs Included in Budget
- General Contactor for Compute- and Data-Servers, Networks and Infrastructure
- Preferably Two Phased Installation
- Balanced System wrt Computing and Data, i.e. Compute Growth Path will depend on Data Growth Path

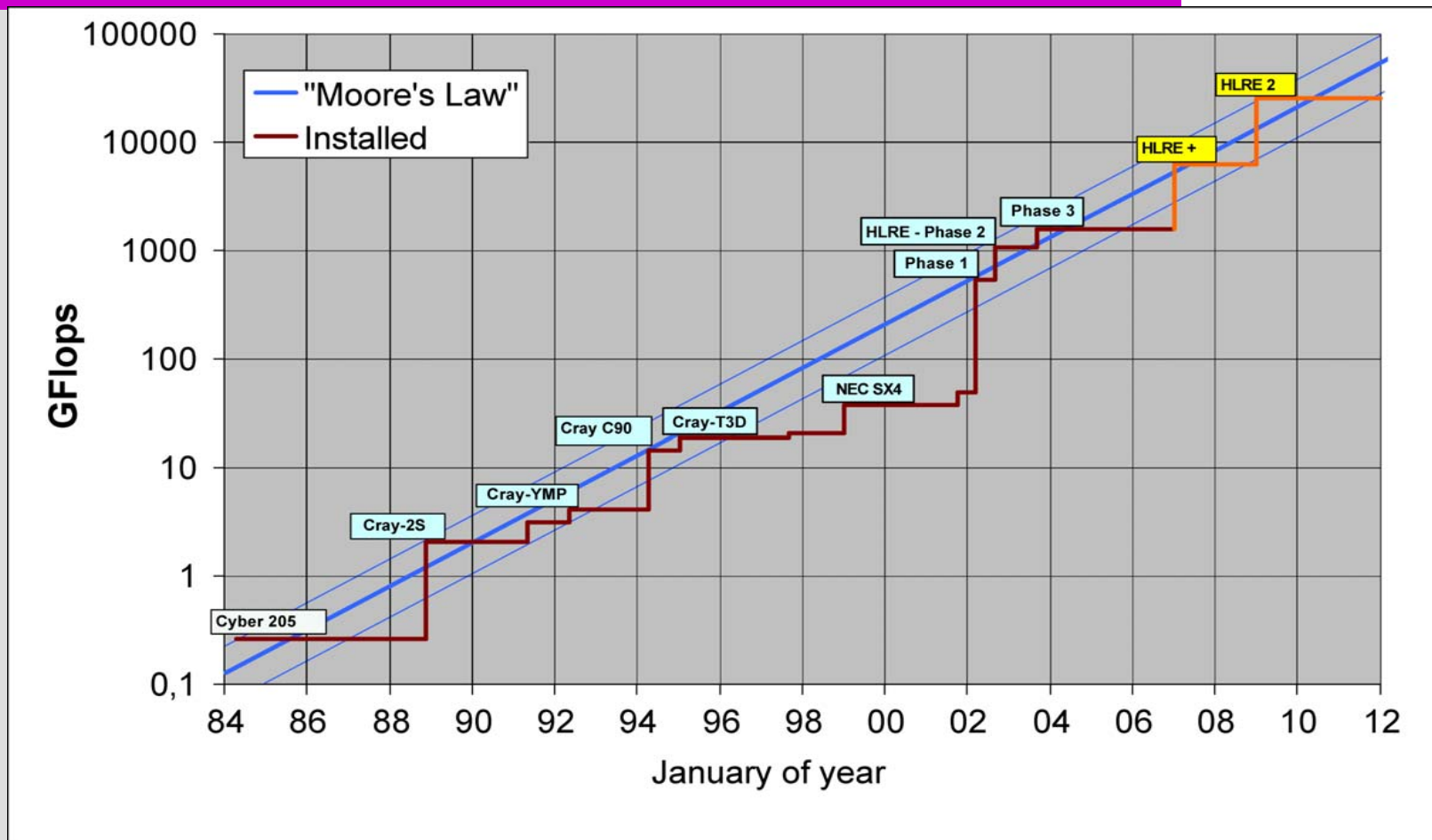
Basic Requirements, cont.

- Network and Infrastructure need to match Computing and Data Requirements
- Increase in Compute Power at least as shown in the following Graph (30 x current at least, pref. 40 x)
- Optional Service Contract Mandatory
- Technically Competent Project Manager at least Part Time On-Site needed

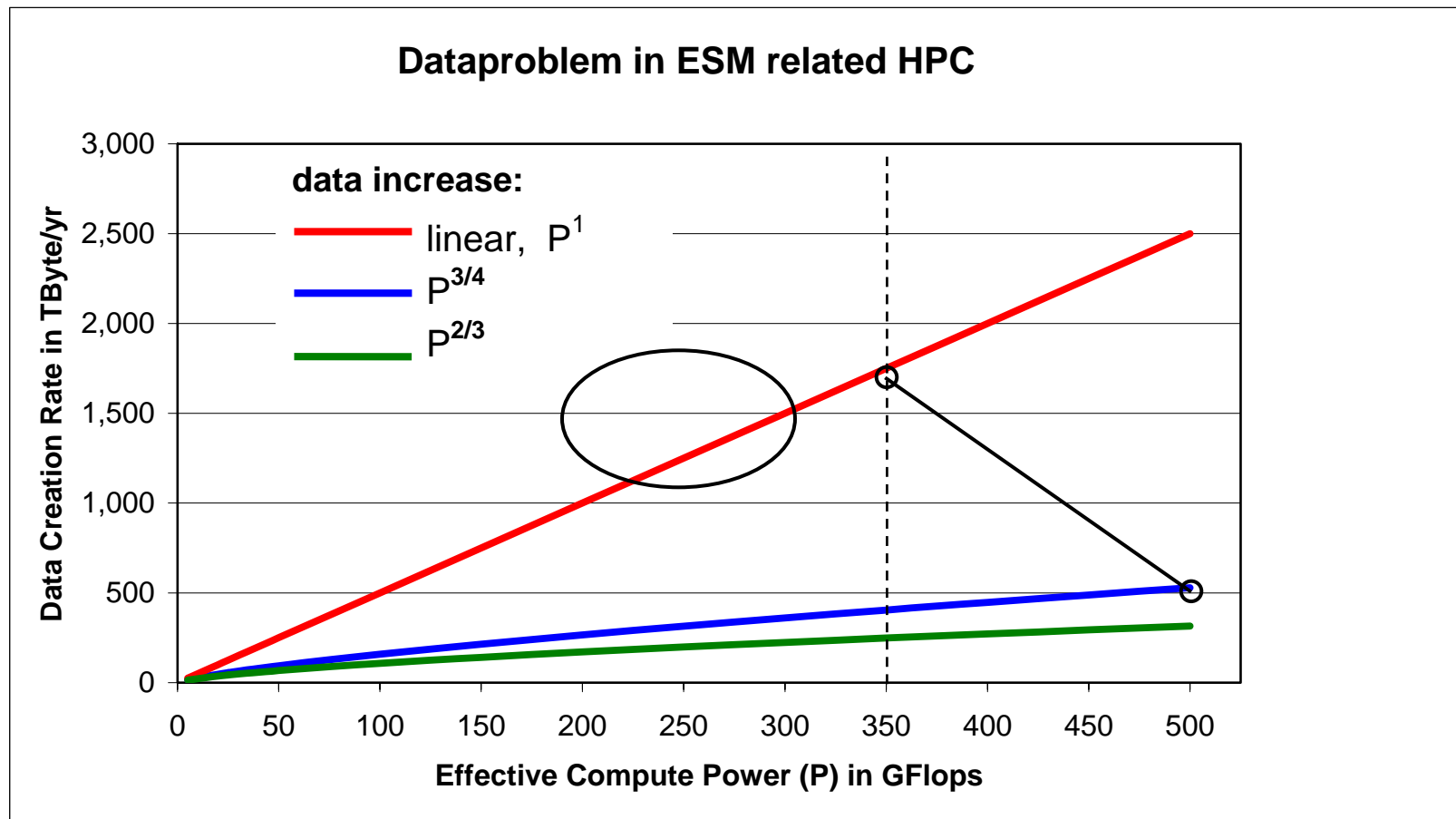
Additional Requirements

- DKRZ will request **Source Code** for all Major Software Components except for Compilers and DBMS (as for HLRE)
- In Order to make Efficient Use of the Source Code a Corresponding Documentation will be made **Mandatory (IMS-type docu)**
- The IMS Docu will have to be Submitted as Part of the Proposal

Evolution of Computing Power at DKRZ



Point of Operation in CS-DS-Space for HLRE



Rationale for Data Growth Path

- On the average for the HLRE Phase 3 the annual archive growth rate is 1.5 PByte/a
- The average annual archive increase rate will be set to 6 PByte/1 TFLOPS_{sust} for HLRE2
- 1 GByte/s nominal tape transfer rate can handle 3 PByte/a archive increase rate
- The DBMS will grow at about 30 % of the Archive

Minimal Growth Path for CS and DS

Year	Compute	Disk	Archive	DBMS
2006	0,25 TF	0,15 PB	6 PB	2 PB
2007	1 TF	0,6 PB	12 PB	3 PB
2008	1 TF	0,6 PB	18 PB	5 PB
2009	5–7,5 TF	3 PB	48 PB	12 PB
2010	5–7,5 TF	3 PB	78 PB	20 PB
2011	5–7,5 TF	4 PB	108 PB	28 PB

HSM Requirements and Strategy

- The HSM is considered as Seamless Extension of the Total File System
- GFS and HSM Components have to be laid out Redundantly and with Fail-Over
- **The HSM has to deal with 100 to 500 Mio Filename-Entries and a Total Storage Capacity of up to 150 PB**
(resulting procurement requirement)

DBMS Requirements and Strategy

- Consequently the HLRE2 DBMS must be Able to **Handle O(1 Billion) Entries**
- The HLRE2 **DBMS** will Grow in **Size up to 20 – 50 PB** Depending on the Data Intensity
- The HLRE2 DBMS must Provide Hooks that its Data can be Read by a yet unknown HLRE3 DBMS
- New (Reduced) Functionality Approach necessary because of costs?

Summary

- **DKRZ provides Computing Resources for Climate Research in Germany on a competitive international level**
- **The HLRE System Architecture is suited to cope with a compute- and data-intensive Usage Profile**
- **Shared Filesystems today are operational in Heterogenous System Environments**
- **Standardisation-Efforts for Shared Filesystems needed**

A photograph of a long row of server racks in a data center. The racks are arranged in a perspective that recedes into the distance. The text "Thank you for your attention !" is overlaid in a pink, italicized font across the middle of the image.

Thank you for your attention !