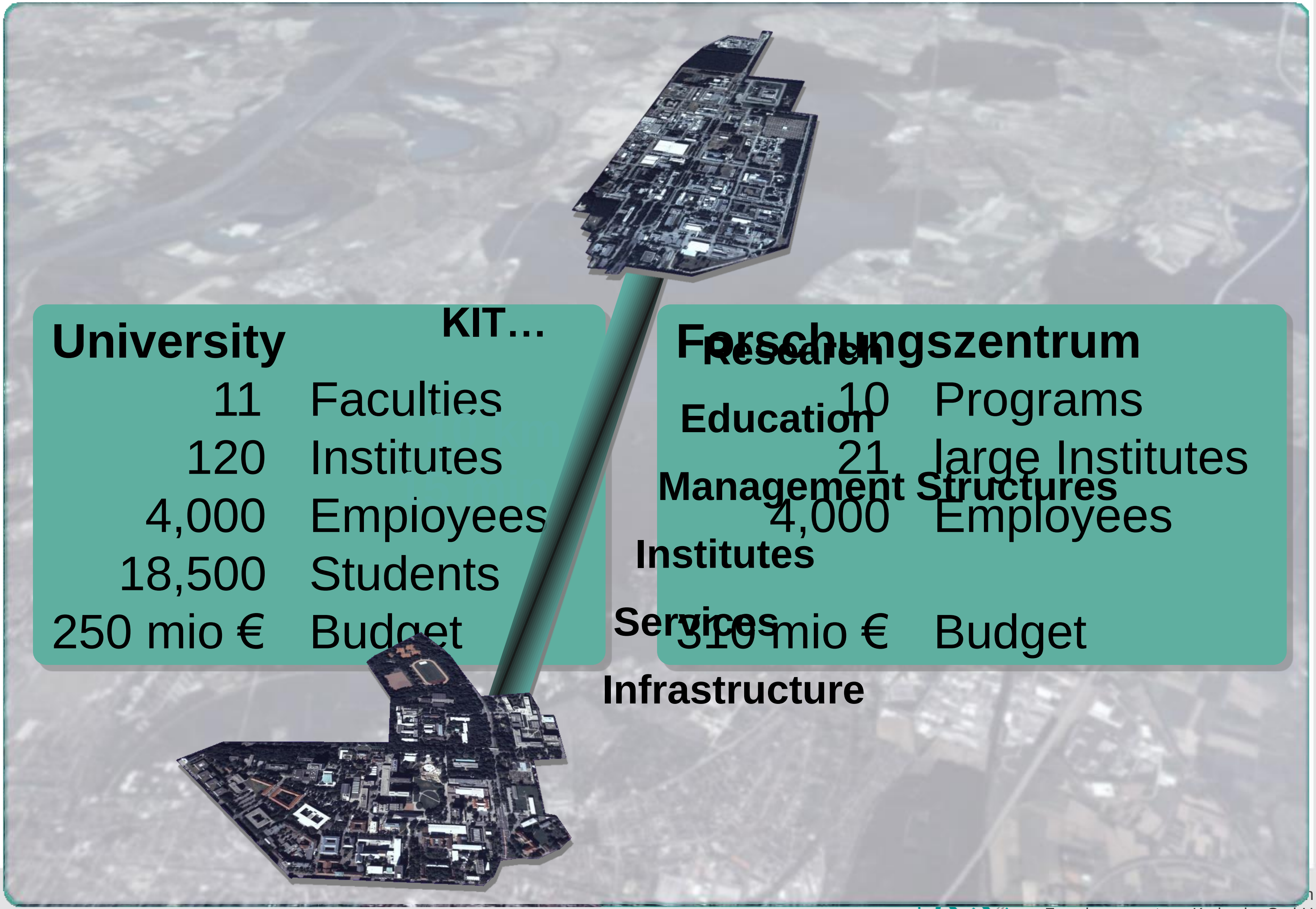
StorNextFS, a fast global Filesysteme in a heterogeneous Cluster Environment

Frank Schmitz

frank.schmitz@iwr.fzk.de

Karlsruhe Institute of Technology



University

11	Faculties
120	Institutes
4,000	Empioyees
18,500	Students
250 mio €	Budget

Forschungszentrum

10	Programs
21	large Institutes
4,000	Employees
310 mio €	Budget

KIT…

Research

Education

Management Structures

Institutes

Services

Infrastructure

Motivation for a global file system and the virtualisation

- Existing Linux- / Unix-Cluster / Windows environments and HPC systems like vector computers and InfiniBand-cluster
- New processors include virtualization technology (Vanderpool, Pacifica)
- Tasks not solvable with Linux (e.g. large excel sheets and other Microsoft based applications) → Windows needed
- Accessing data in a **global file system** solution from various operating systems and hardware platforms (like IBM, Intel, AMD, SUN, NEC)
- Testing six month (starting early 2006) in a heterogeneous SAN environment, we have found **StorNextFS** (SNFS) from Quantum/Adic as the best solution for KIT.

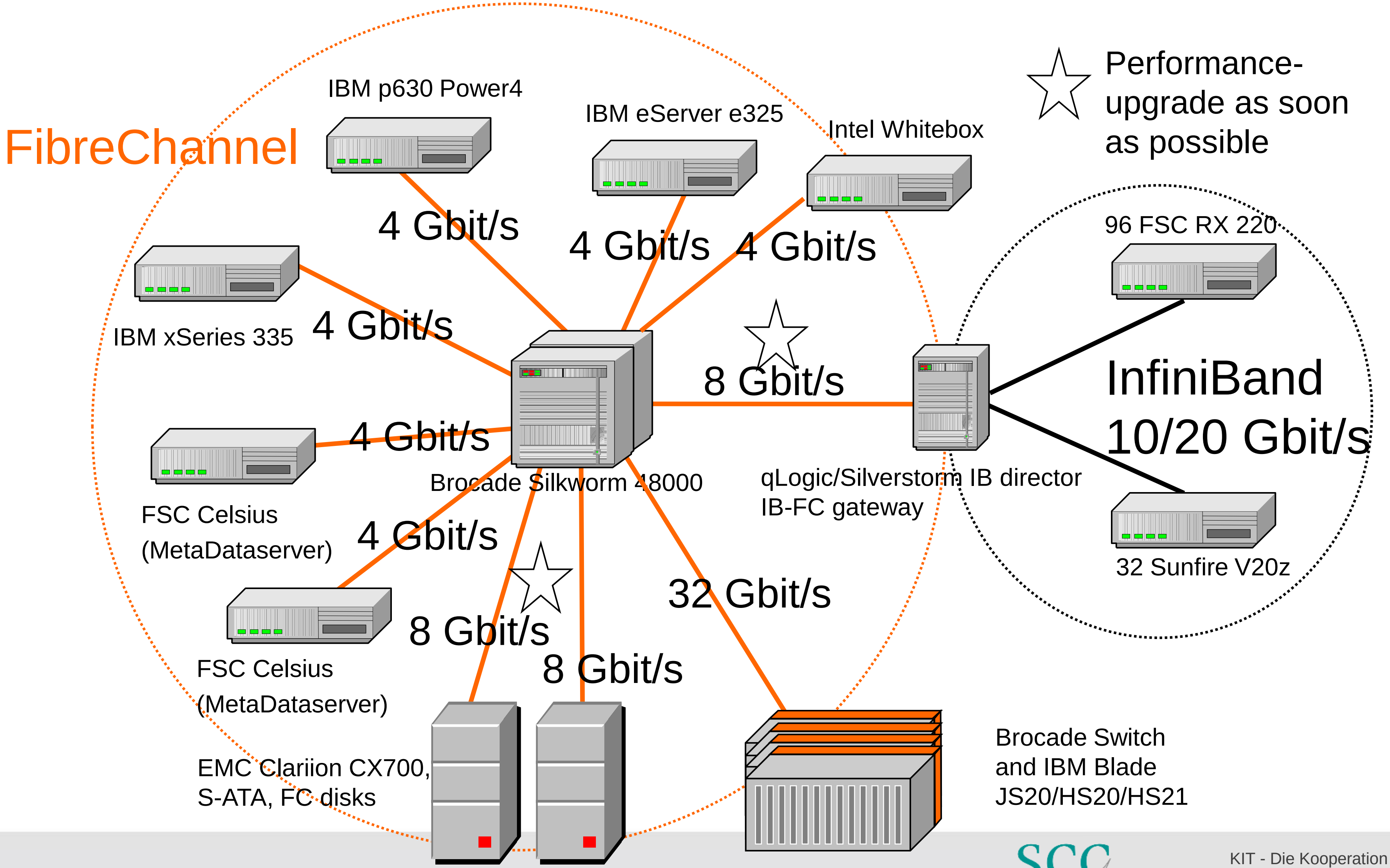
Ideas in 2006

- one global and fast file system to solve all needs (StorNextFS, SAN-FS, SAM-QFS, NEC GFS, PVFS, Sistina GFS, CXFS, Celerra High-Road,…),
- integration in low performance Grid file system or something like AFS
- InfiniBand, iSCSI, FC-SAN, gateway solutions
- first steps in 2004: gLite as the middleware layer, but …
- OGSA compliant Grid services → Globus ToolKit 4
- resource brokerage (TORQUE, LSF, CONDOR, LoadLeveler…)
- security → Kerberos 5 based solution

Since early summer 2006

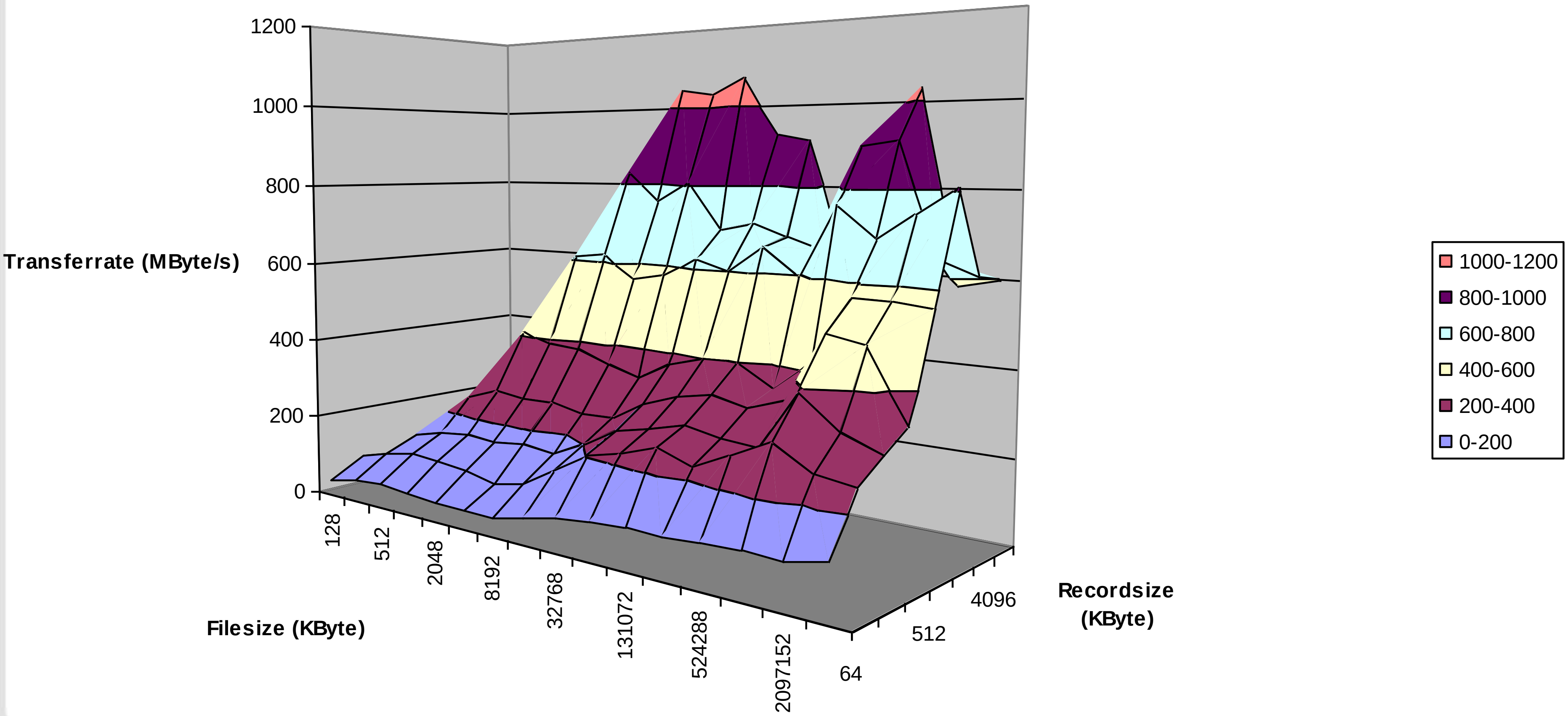
- Grid middleware GT4, gLite and Unicore are running in the CampusGrid (gLite and GT4 only Linux) environment
- GT4 will be available soon for the AIX and Super/UX operating systems
- integration in low performance Grid file system or something like AFS
- InfiniBand, iSCSI, FC-SAN, gateway solutions
- first steps in 2004: gLite as the middleware layer, but …
- OGSA compliant Grid services → Globus ToolKit 4
- resource brokerage (TORQUE, LSF, CONDOR, LoadLeveler…)
- security → Kerberos 5 based solution

The hardware structure for the new StorNextFS version 3.0 (redundant) in the CampusGrid environment



Performance of a small InfiniBand cluster using a FC-Gateway from qLogic

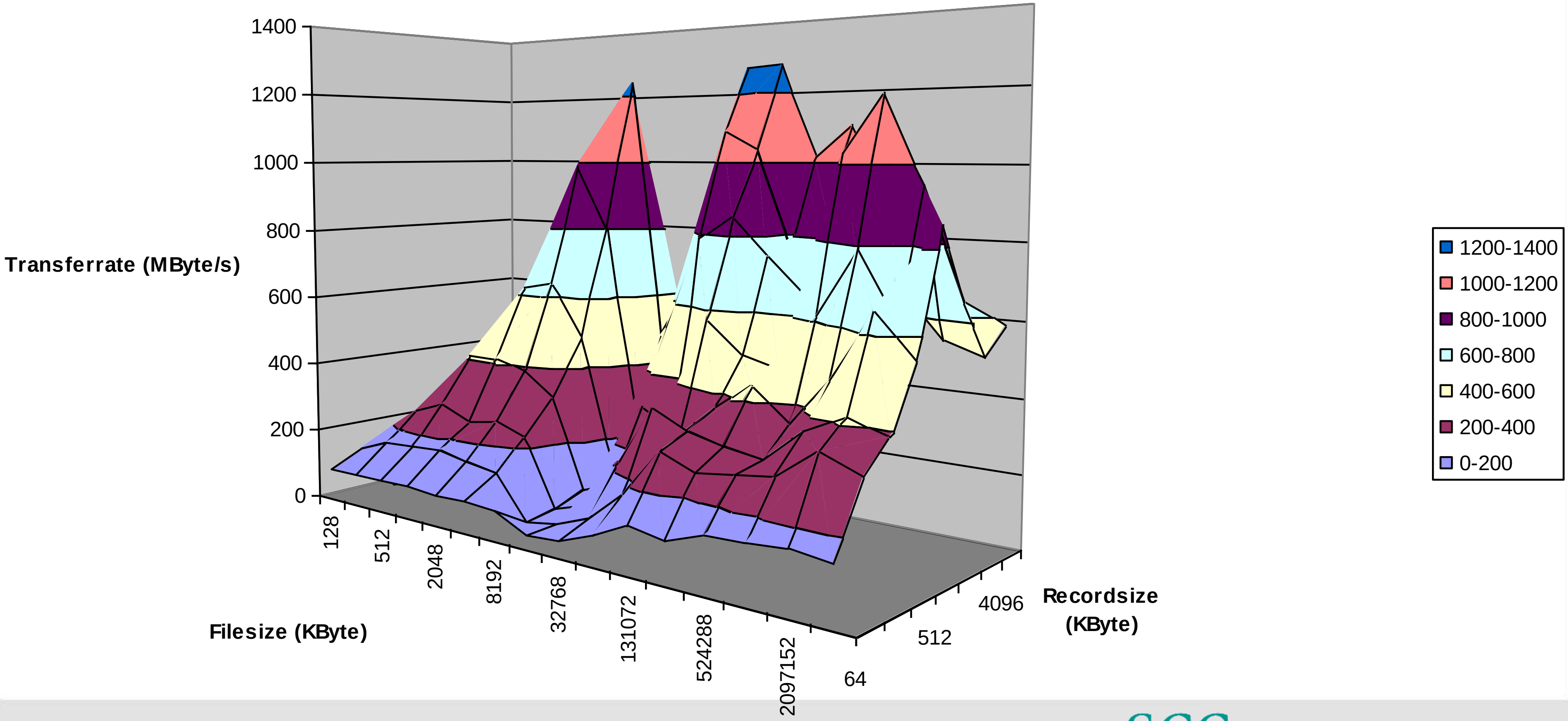
iozone (StorNextFS, write using fsync, 8 nodes, 8 processes)
The processes are started at the same time but have a run time diffenence of 1h 20m using a maximum of 5h 40m execution time (not only write)!



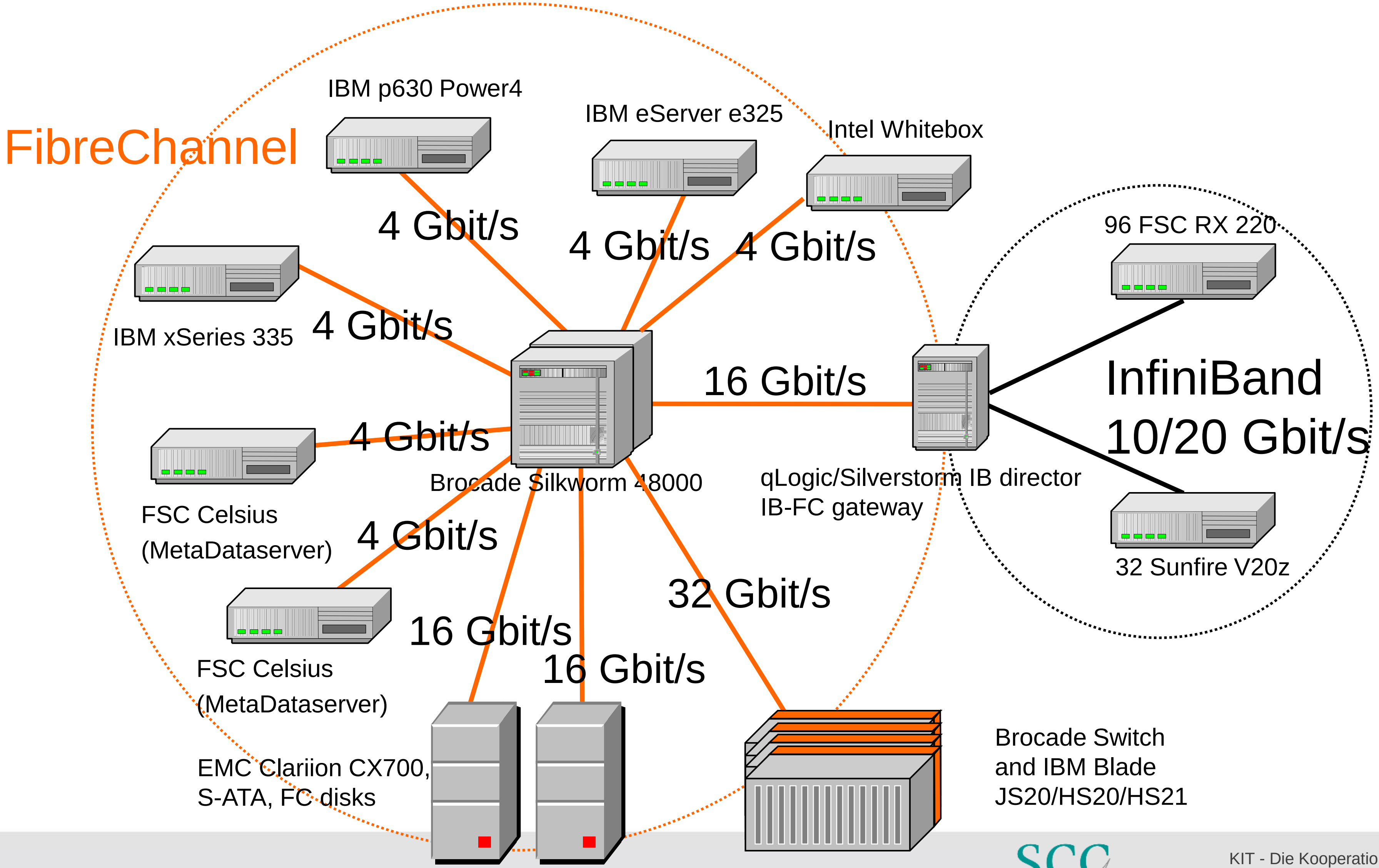
Performance of a small InfiniBand cluster using a FC-Gateway from qLogic

iozone (StorNextFS, write using fsync, 16 nodes, 16 processes)

The processes are started at the same time but have a run time diffenence of 1h using a maximum of 8h 20m execution time (not only write)!



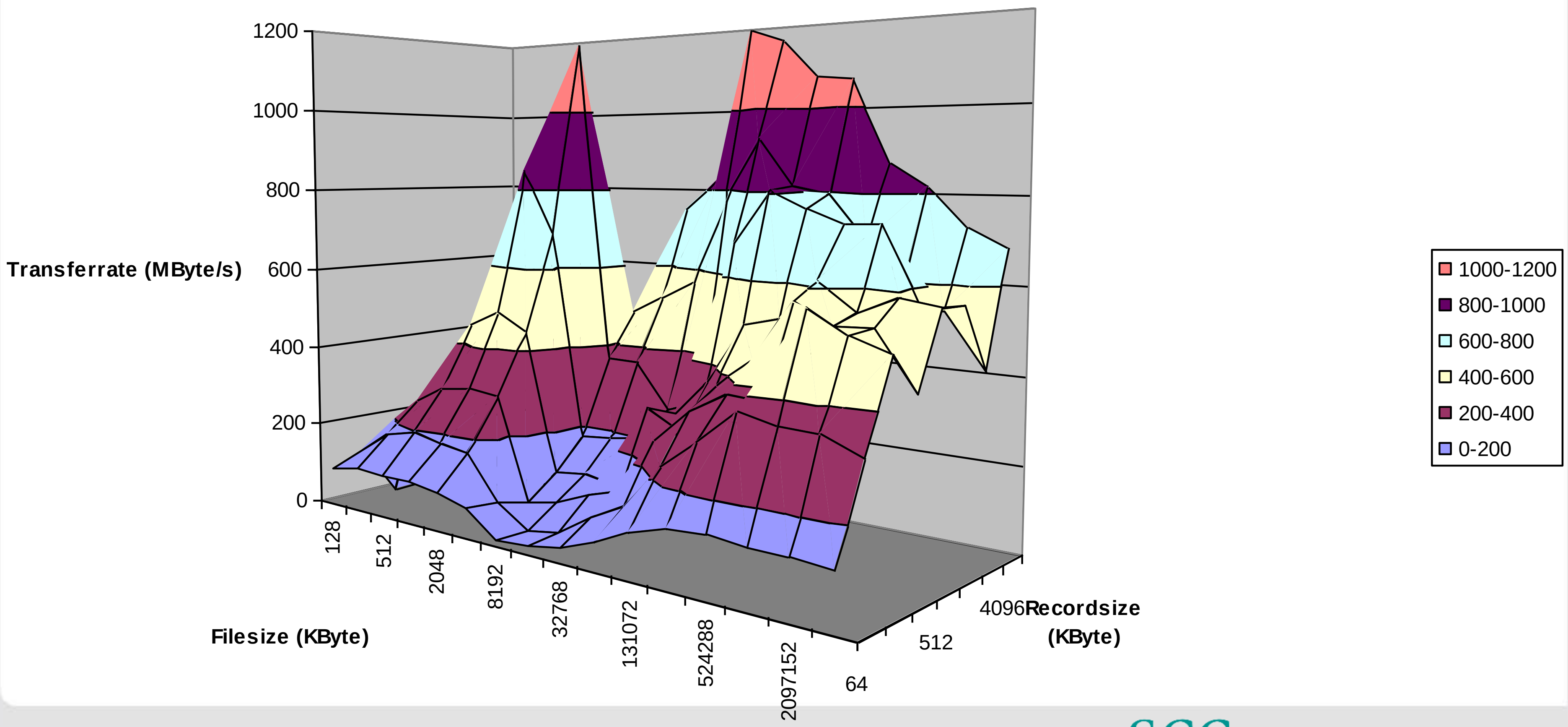
The hardware structure for the new StorNextFS version 3.0 (redundant) in the CampusGrid environment



Performance of a small InfiniBand cluster using a FC-Gateway from qLogic

iozone (StorNextFS, write using flush, 28 Nodes, 28 Prozesses)

The processes are started at the same time but have a run time diffenence of 2h using a maximum of 15h 30m execution time (not only write)!



Internals

- Limitations
  - A redundant Clariion controller running in secure mode is limited to 1.5 GByte/s for I/O
  - Because we are using two metadata-LUNs and four data-LUNs the limitation for writing on RAID-5 is 1.4 GByte/s (350 MByte each LUN)
  - InfiniBand DDR performance is 20 Gbit/s, the effective data rate is limited to 2 GByte/s because of the 8B/10B encoding schema (full duplex)
  - PCIe x8, it‘s limitation is 2 GByte/s
  - A single node using one HCA can achieve up to 250 MByte/s
- StorNextFS 3.0
  - One metadata-server could handle thousands of clients
  - Metadata will be send via Ethernet

Results

- Advantage for VMware virtualisation
  - No need to change existing Environments
  - offers more different Services
  - support for Windows- and Unix-programs including a unified file system solution
  - Using ADS as a user administration solution for KIT
- Disadvantage for virtualisation
  - maybe reduced Performance compared to native Installation
  - little overhead if CCN have to run all the time
- StorNextFS
  - performing very well, the bottlenecks are the EMC Clariion (only one system used for the benchmark) controller and the four FC data LUNs
  - the windows ADS integration is very well done! The windows performance for the version 2.7 was as assumed.
  - working in a VMware ESX Environment is no problem.
  - the EMC Clariion and the IB-FC gateway are the limiting factor!
  - Installation of StorNextFS is easy