Preview of a Novel Architecture for Large Scale Storage

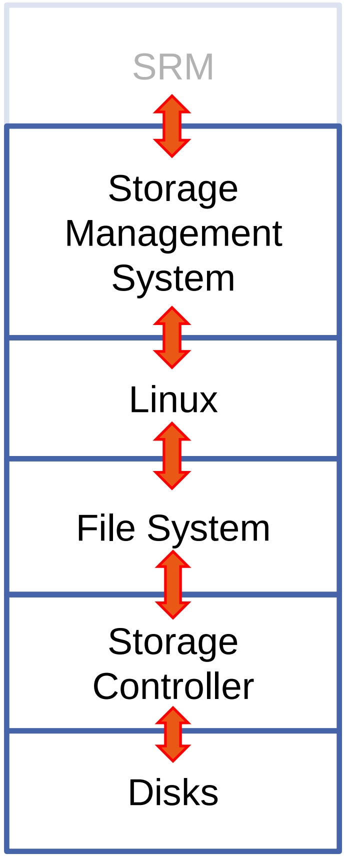
Andreas Petzold, Christoph-Erdmann Pfeiler, Jos van Wezel
Steinbuch Centre for Computing

STEINBUCH CENTRE FOR COMPUTING - SCC

Storage Management Systems at GridKa

- dCache for ATLAS, CMS, LHCb
  - 6 PB disk-only
  - 3 PB tape-buffers
  - 287 pools on 58 servers
  - Agnostic to underlying storage technology

- Scalla xrootd for ALICE
  - 2.7 PB disk-only/tape buffer
  - 15 servers
  - Agnostic to underlying storage technology



Current GridKa Disk Storage Technologies

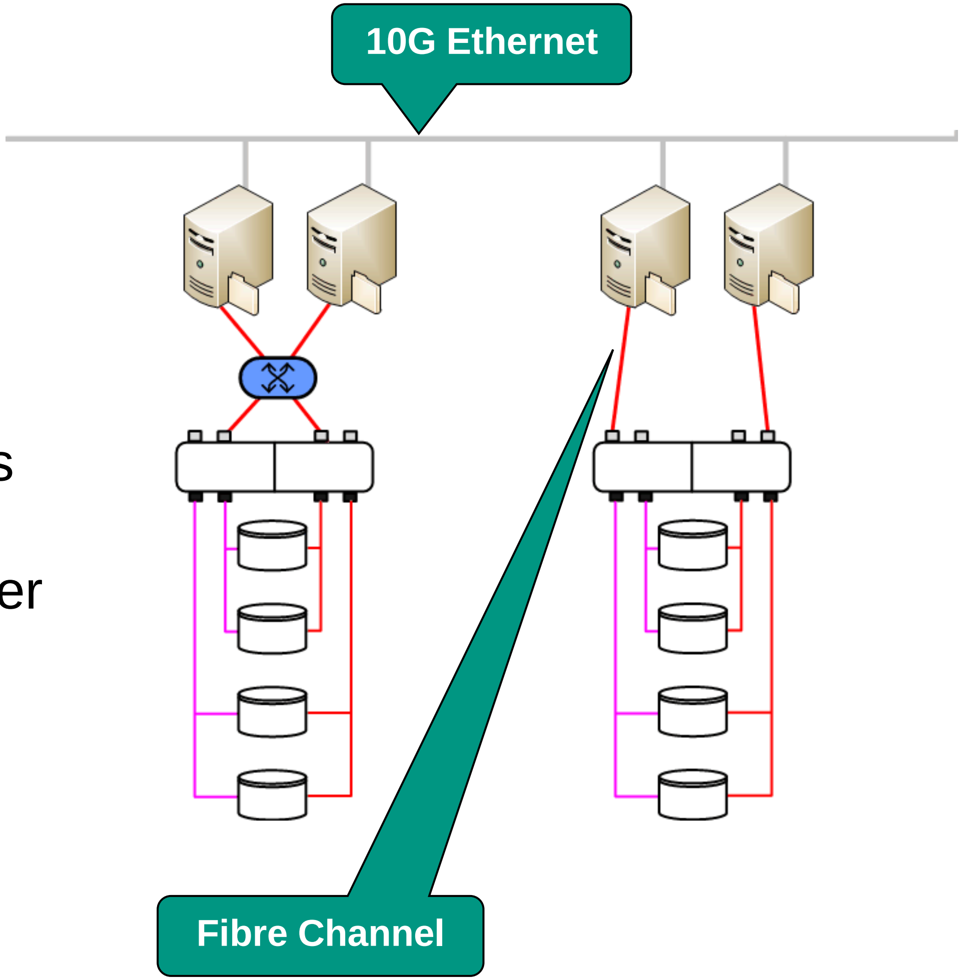
- 9 x DDN S2AA9900
  - 150 enclosures
  - 9000 disks
  - 796 LUNs
  - SAN Brocade DCX
- 1 x DDN SFA10K
  - 10 enclosures
  - 600 disks
- 1 x DDN SFA12K
  - 5 enclosures
  - 360 disks

Current GridKa Tape Storage Technologies

- 2 Oracle/Sun/STK SL8500
  - 2 x 10088 slots
  - 22 LTO5, 16 LTO4
- 1 IBM TS3500
  - 5800 slots
  - 24 LTO4
- 1 GRAU XL
  - 5376 slots
  - 16 LTO3, 8 LTO4

GridKa Storage Units

- Servers
  - connect directly to storage
  - or via SAN but not explicitly needed
- Cluster Filesystem (GPFS)
  - connects several servers
  - filesystem visible on all nodes
  - predictable IO throughput
  - nice storage virtualisation layer
- Currently evaluating alternative filesystems
  - XFS, ZFS, BTRFS, EXT4



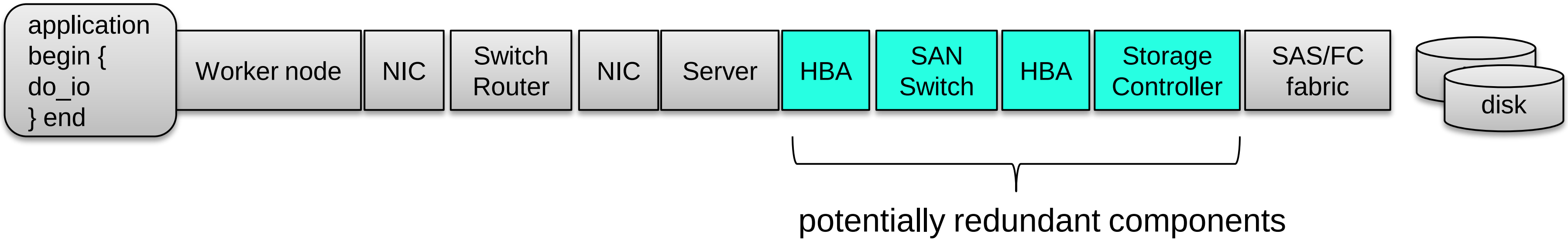
Novel Storage Solutions for GridKa

- Expect large resource increase in 2015
  - Chance to look at new solutions during LHC LS1
  - Simplification in operations and deployment required
- Solution 1: DataDirectNetworks SFA12K-**E**
  - Server VMs run embedded in storage controller

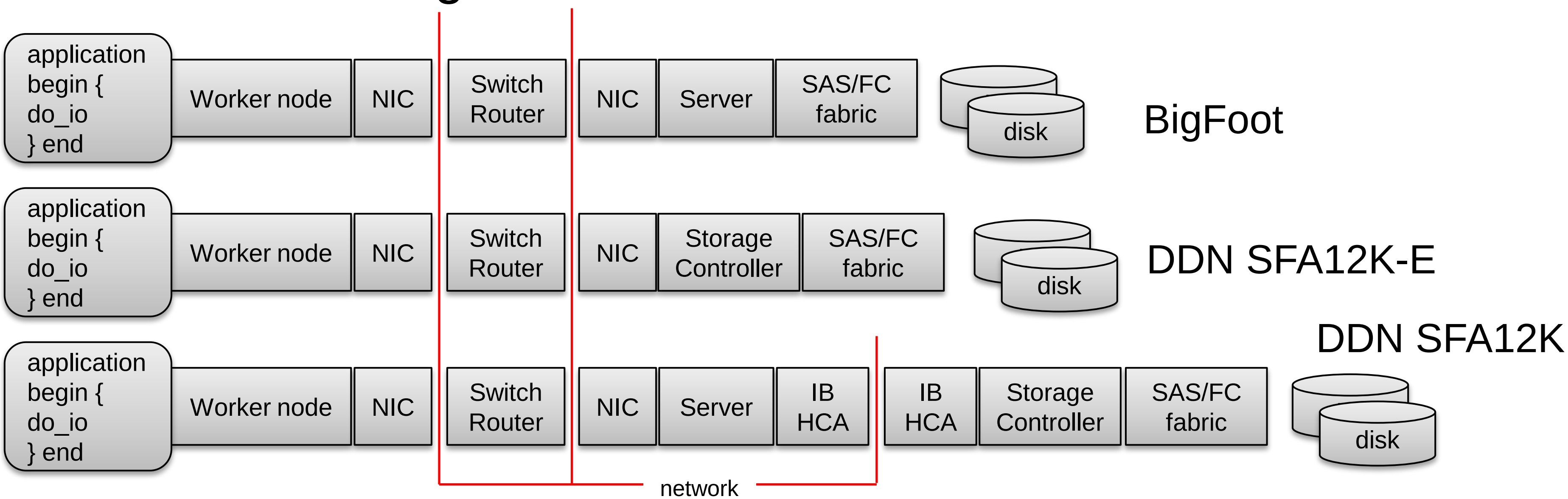
- Solution 2: Rausch Netzwerktechnik BigFoot
  - More conventional setup; server directly connected to local disks

Shortening Long IO Paths

Traditional storage at GridKa



Previewed storage at GridKa

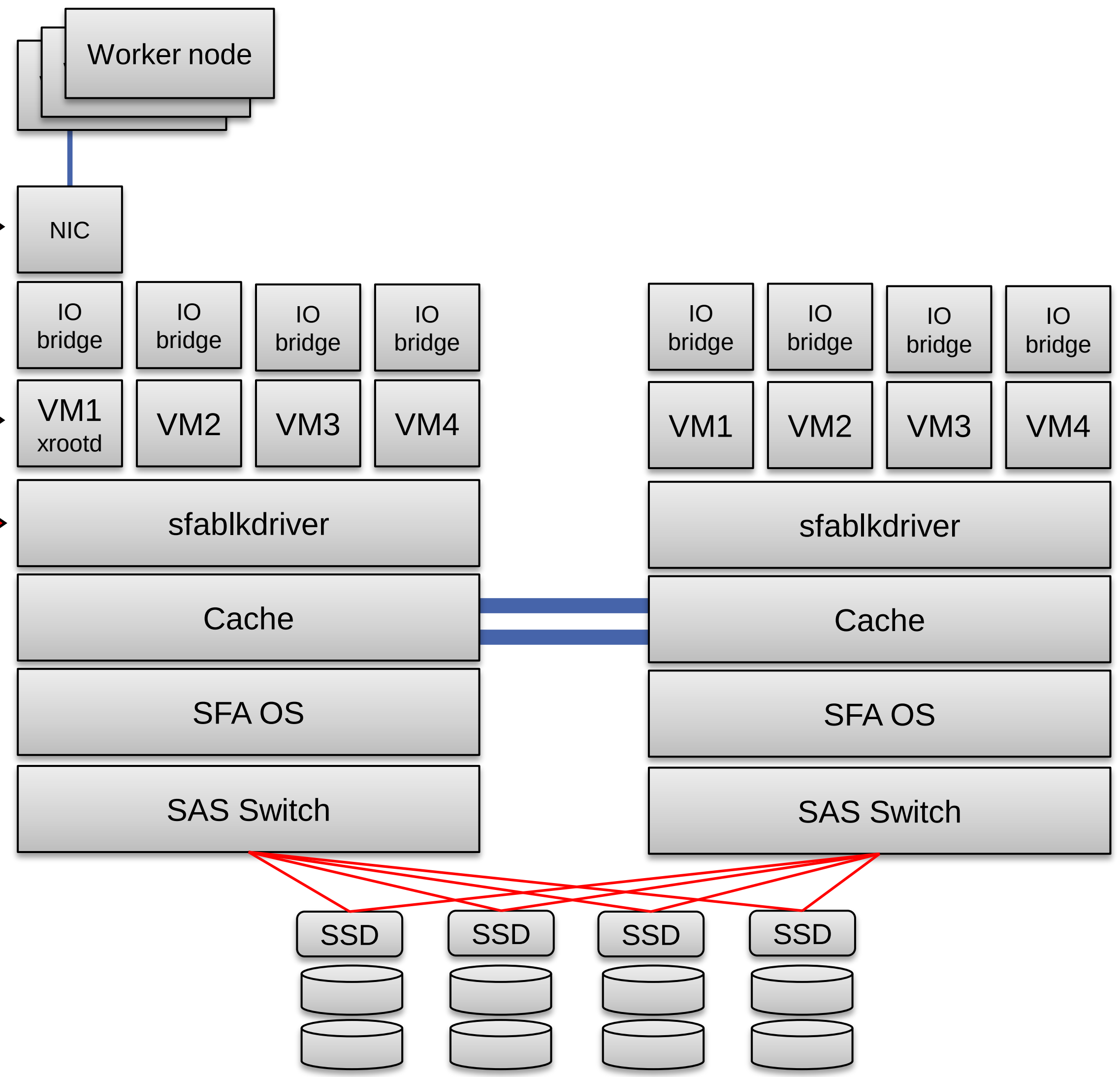


DDN SFA E Architecture

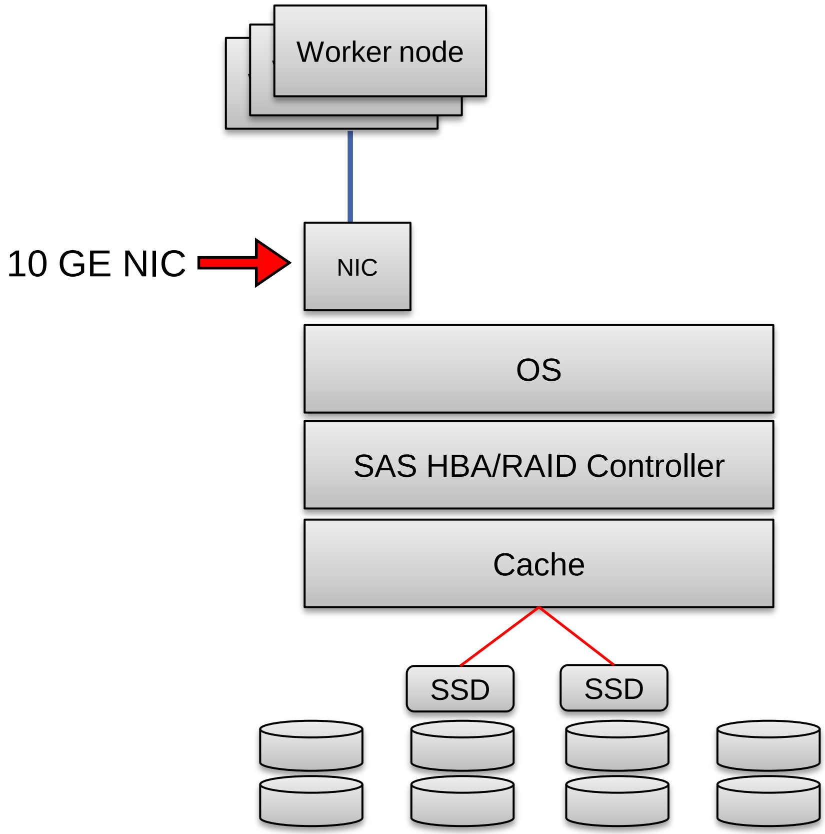
10 GE NIC dedicated to VM

8 VMs per SFA pair

DDN VM driver dma access to storage



BigFoot Architecture



Thoughts on Effect of SAN and Long Distance Latency

- IO is different for read and write
  - writes can be done asynchronously
    - but if you do them sync (like NFS) you have to wait for the ACK
  - workloads are read mostly
    - read are synchronous
    - reading huge files linearly defeats caches in servers
- SSD speeds
  - example: Enterprise MLC SSD, Violin memory, Fusion IO drive
  - number of IOPS: 250000/s (lower estimate) at 4 k/IO
  - $\frac{1\,s}{250000} = 4\,\mu s \rightarrow 4\,\mu s$ per IO
- SAN Latency
  - DCX director 2.1 µs → 4.2 µs round trip
  - not accounting HBA, fibre length (300 m = 1 µs)

Thoughts on Effect of SAN and Long Distance Latency

- Effect on high speed SSDs
  - $\frac{1}{2 * 2.1 \mu s + 4 \mu s} = 121951$ IOPS ***not*** 250000 IOPS
  - similar for disk controller caches when used for VMs or data servers

- Negligible impact of SAN for magnetic disks
  - $\frac{1}{2 * 2.1 \mu s + 5000\ \mu s} = 199$ IOPS

- Putting the SSD next to the server can possibly increase the number of IOPs
  - assumption will be tested in the coming weeks

Expected Benefits

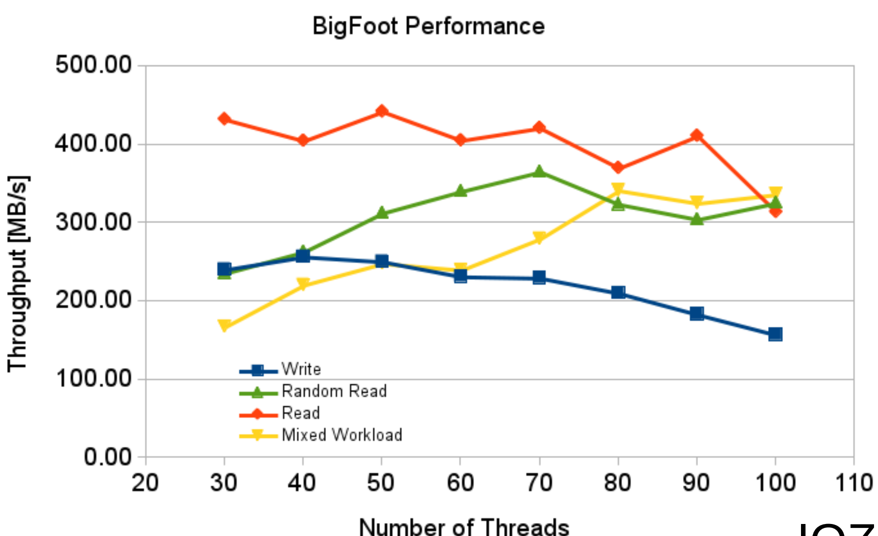
- Reduced Latency
  - HBA and SAN: 2.1 µs (e.g. Brocade DCX, blade to blade) storage controller
  - Improved IO rates
- Reduced power
  - Server and controller HBA, SAN Switch
  - ~300-400W = ~600Euro/server/year
- Reduced investment
  - 500-600 €/HBA, 400-600 €/switch port =1800-2400 €
- Improved MTBF
  - Less components

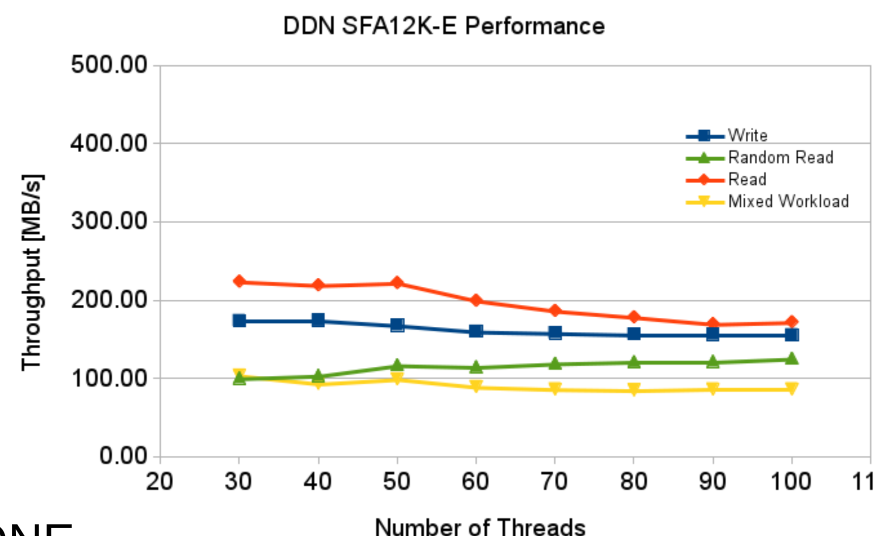
Possible Drawbacks

- Loss in flexibility
  - w/o SAN storage building blocks are larger
  - Limited server access to storage blocks
  - Storage systems are only connected via LAN
- VMs inside storage controller (DDN SFA12K-E)
  - Competition for resources
  - Limited number of VMs limits “server/TB” ratio
  - Loss of redundancy
- Simple server attached storage (BigFoot)
  - Limited by simple hardware controller
  - HW admin doesn’t scale to 100s of boxes
  - No redundancy

Glimpse at Performance

- Preliminary performance evaluation
  - IOZONE testing 30-100 parallel threads on XFS filesystem
  - Xrootd data server in VM, performance similar to IOZONE
  - Out-of-the-box settings, no tuning
  - Performance below expectations, reasons still to be understood
  - ZFS tested on BigFoot





IOZONE

Conclusions

- Storage extension at GridKa requires expensive upgrade of GridKa disk SAN – or novel storage solution
- Tight integration of server and storage looks promising
- Many possible benefits – further evaluation required
  - Less components
  - Less power consumption
  - Less complexity
- Performance needs to be understood together with vendors
- More tests with other vendors in near future


GridKa