

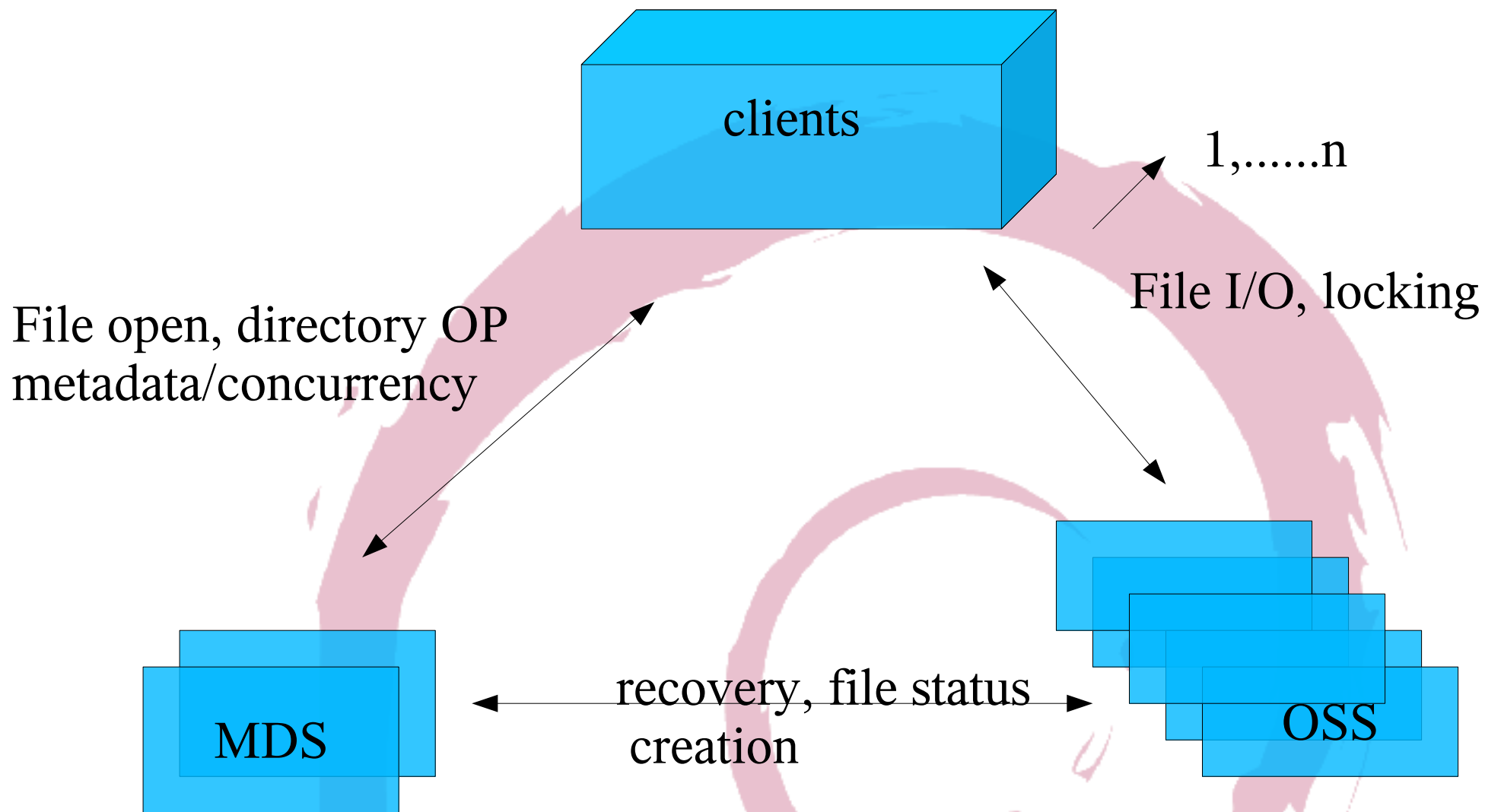
# Lustre Cluster at GSI

**Walter Schön, GSI**



# Topic

- **Architecture overview**
- **Lustre@GSI**
- **The “even cheaper” solution**
- **The MDS hardware/concept**
- **Outlook**



achieved meta data transaction rate: 15.000 ops/s  
aggregate I/O: > 130 Gbyte/s  
maximum file/file system size: 1,25 PB / >32 PB ( 1.6.x series)

# MDS – File Organization

- Meta data information/pointers are stored on the MDS
- No “mass data” are stored on the MDS

How much disk space does the MDS need?

- => Per file, one inode is used – independent from file size
- => lustre is efficient with “large files” (> 1 MB)
- => lustre is not efficient for small files e.g. 4k :-)

## (some) Lustre Features:

- **Fully POSIX compliant:** “general purpose” file system
- **File I/O of raw bandwidth:>90%** (experimentally proved)
- Capacity of FS is sum of storage targets
- Aggregate I/O bandwidth scale with the number of OSSs
- Fill balancing (configurable)
- Quota: user and group quota available ( ! in principle ! )
- Dynamically integration of new OSTs
- Controlled striping: FS default, recursively directory attribute, individual files at creation time

# Lustre Clusters: Storage Architecture

**GSI solution: even cheaper**

Failout solution: cheap

Failover/shared storage  
“standard”

**Enterprise class  
storage array and  
SAN fabric**

MDS

MDS

network: GigE,  
.....

clients

OSS+ 2 OST

OSS

OSS

OSS

OSS

OSS

OSS

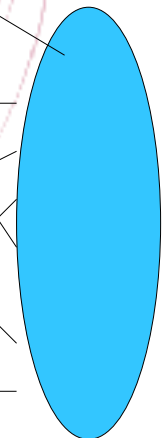
OSS

OST

OST

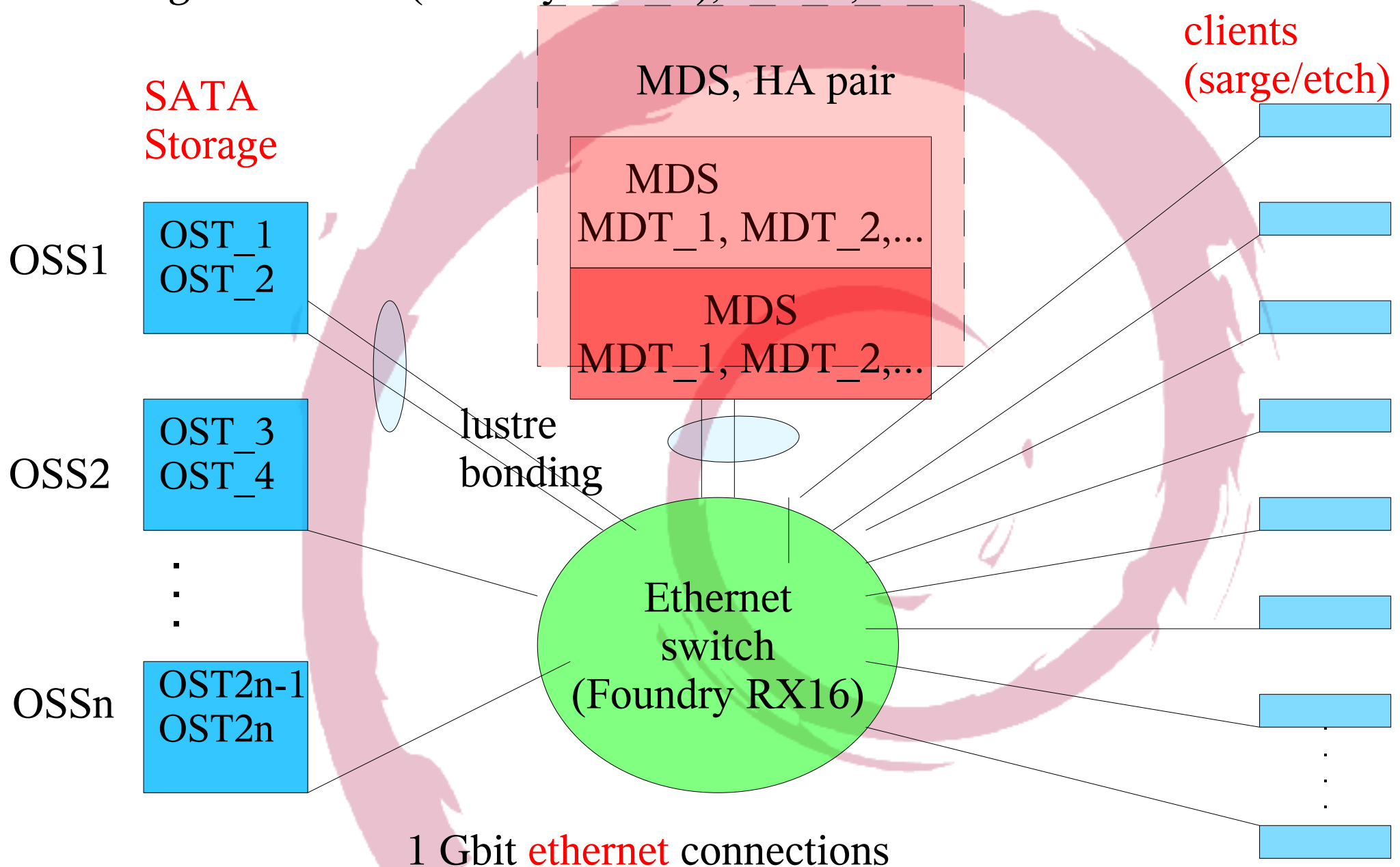
OST

OST



# lustre production/alpha cluster : architecture

running lustre 1.6.x (recently 1.6.4.3), debian, 2.6.22 Kernel



# The MDS Server

2 server in a HA configuration ( => talk K.Miers )

## hardware:

- 3HE Supermicro
- RAID 10 for MDT with 400 GB disk space, WD Raptor SATA disks
- 8 cores, 32 GB Ram
- 2 x GbE in lustre bonding (HA)
- 1 x GbE crossover for drbd
- 1 x GbE heartbeat

RAID 10 because of “small file” optimization of the db

## software:

- heartbeat-v2
- drbd



# Lustre@GSI:

**Test cluster**

testing new technology

**data file system, NFS**

40 file servers, nfs based

Alice Tier2

**Production cluster**

migration

FAIR computing  
GSI computing  
Theorie (Hydro)  
Alice Tier2

**“Alpha” cluster**

about 100 file servers, 0.5-0.7 PB

# The “Even Cheaper” Solution @ GSI:

hardware based on SATA storage and ethernet connections

OSS in “fail out mode”

default striping level: 1

default replication level: 1

number of MDS: ----- HA Pair  
number of OSSs : ----- 60 ->120  
number of OSTs : ----- 120 ->240  
lustre file systems: ----- 0.3 -> 0.7 PB, RAID 5  
number of client CPU's --- 1000 -> 1500  
aggregate I/O performance -- 6 -12 GB/s

cost (2007, including MDS, networking) : 660 Euro/TB

cost estimate with 24 slot servers/TB disks : 400 Euro/TB

3 RAID controllers, 4HE, 24 slots, 1 TB disks, 2x1GbE? 10 GbE?



## 3 HE server

- redundant power supplies
- LOM modul
- redundant fans
- excellent cooling of disks, memory, CPU
- 16 slot SATA, hot swap
- 14 slots for data
- 2 slots for RAID 1 system
- 2 SATA RAID controller
- 4/8 GB RAM
- Dual CPU Dual core
- 500 GB disks WD RAID ed.  
24x7 cert.,  
100% duty cycle cert.

5,6 TB per 3 HE RAID 5  
73 TB per rack

# Performance

- Lustre fills 1 GbE connection with **114 MB/s**
- Lustre scales linear with number of attached GbE connections to the OSSs  
=> for details and more numbers look at the GSI lustre talks at St. Louis
- Up to now: no derivation from linearity (I/O vs number of clients) discovered..  
( measurements with IOZONE in cluster mode )

## gStore:

- Data transfer between lustre cluster and data movers 114 MB/s per data mover ( connected via 1GbE)  
=> archiving of data, tape station

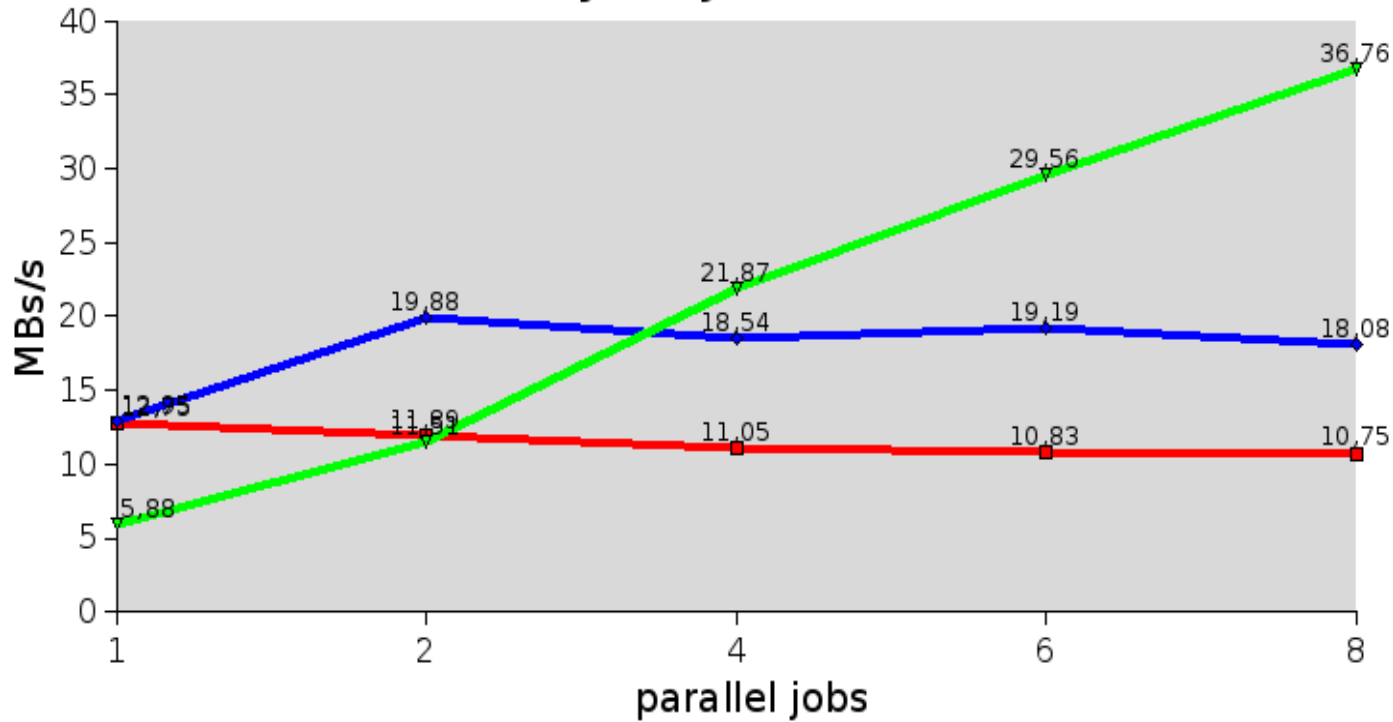
# HADES Lustre Test

- Comparison GSI data file system ( nfs based) - GSI lustre
- HADES Analysis ( I/O intense)
- Typically HADES data challenge with many jobs parallel

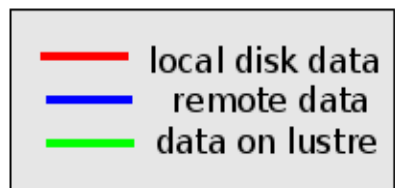
- GSI data file system (nfs): CPU load about 60% : **jobs I/O bound**
- GSI lustre; CPU load 99.x% : **jobs CPU bound**

.... measurement of the HADES group

# Aggregate Data Throughput Rate for Parallel Analysis Jobs ("train of tasks")



still a factor of 3 off from IOZONE



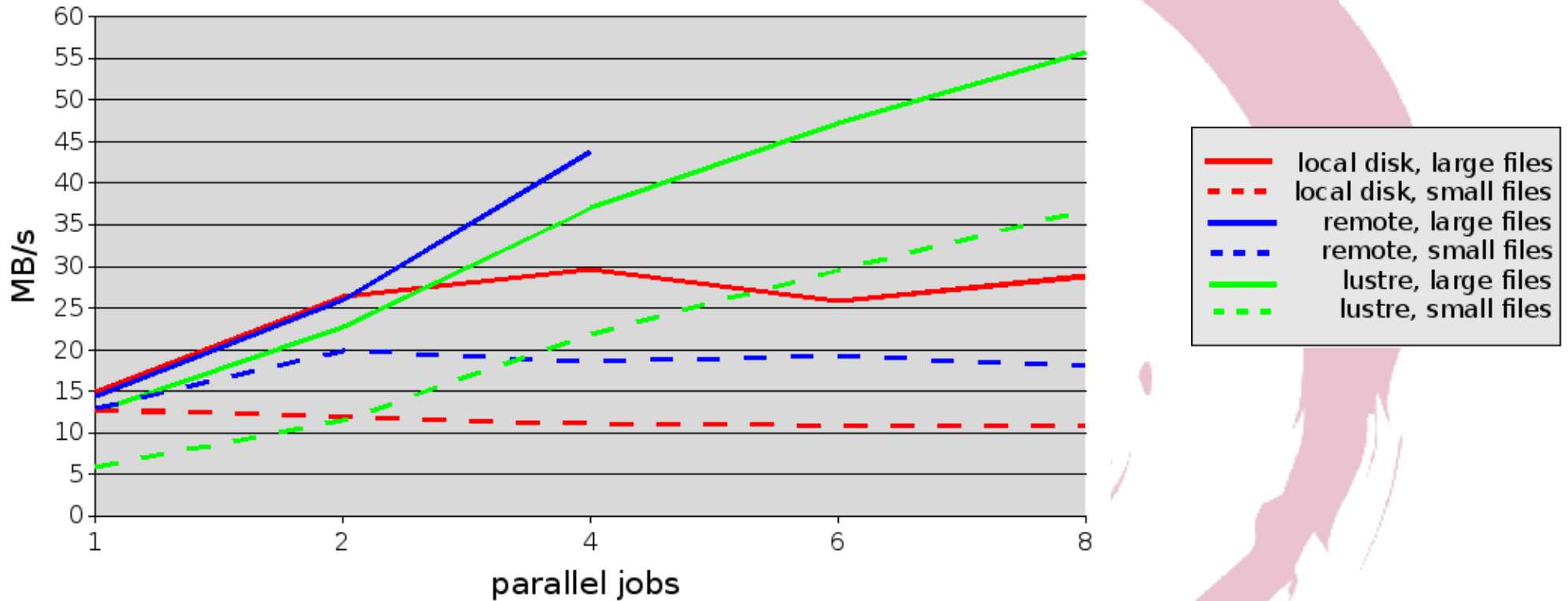
xrootd

lustre

measurements by our Alice group

# Test with ALICE analysis code

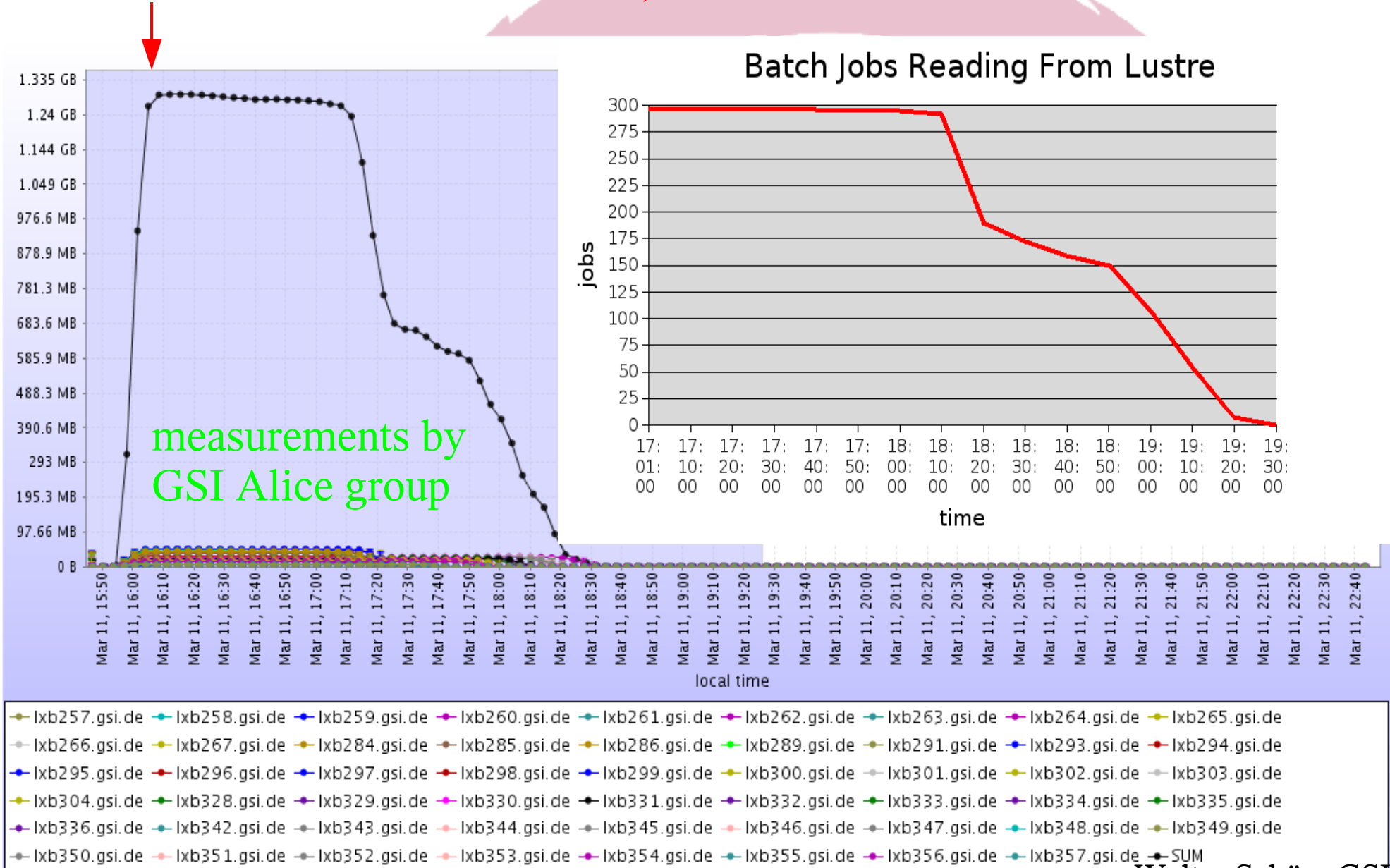
## Aggregate Data Throughput for Analysis Jobs



measurements by GSI Alice group

# More Measurements with the Alice Code ....

limit of network link of the clients in BG2  
(never reached in the time before lustre)





# “Look and Feel” , POSIX Compliance - User Test

- CBM: compiling huge amounts of code on 16 core boxes with “forking”  
=> ... very fast .... :-)
- : Looks like “normal” file system .....
- Thanks to POSIX compliance: No change of analysis code necessary
- Users “happy” ... :-)
- Fast access even during heavy load .....

# Performance - Conclusion

- GSI lustre cluster is only limited by number of GbE connections  
=> HEPIX talk in St. Louis
- For some parts of the cluster limitation is the  
client network connection @GSI => need to be improved
- Users happy ;-)

# Reliability of the Lustre Cluster

## “Regular” tests:

- Switching off one MDS
  - Established I/O still works
  - For short period no new files, no meta data information  
=> HA talk from K.Miers
  - Switching off both MDS
  - **Established I/O still works!**
  - No new files, no meta data information unless reboot of at least one MDS
- Destroying MDS db with nasty dd copy actions
  - Lustre “survives” => details talk K.Miers

# Real Life Tests of Lustre Reliability

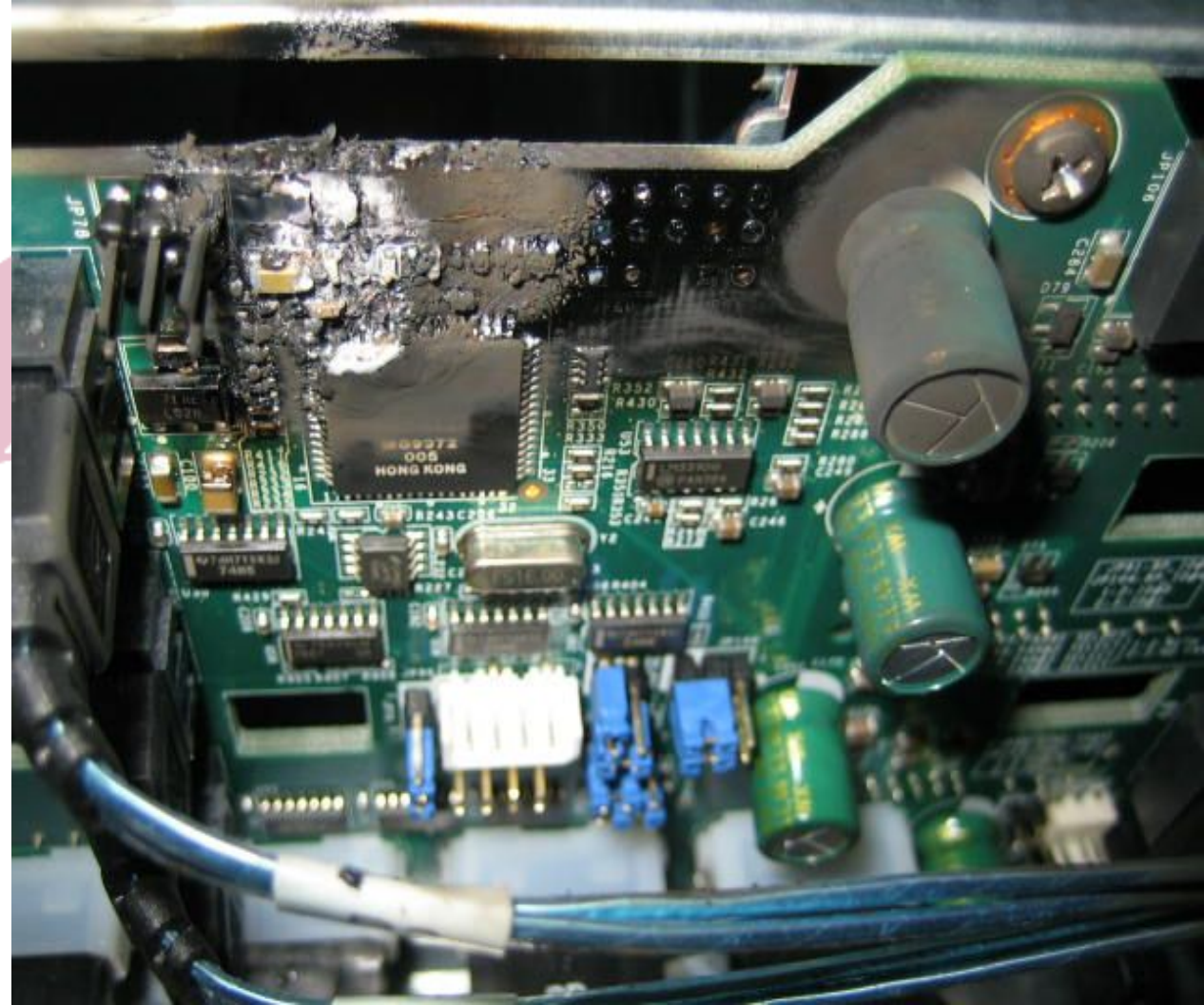
## Accident: Failure of a backbone switch

- **NFS:**
  - Lots of nfs stales in the batch farm which are connected to nfs file servers
  - Lots of client reboots necessary to get rid of the nfs stales
- **Lustre** ( test cluster )
  - Jobs “pending” during network problems
  - Jobs continue working after network connections o.k.
  - No manual interaction necessary

more real life tests...

Accident: smoking hardware

burned 16 slot OSS:  
loss of two OST's with  
about 6 TB data .....



Lustre:

- All jobs accessing the two OST's pending
- All other jobs continuing .....
- After switching the disks in a “spare” chassis:
  - All pending jobs continued working
  - We lost no single job

# Is Lustre the paradise?

Performance and Scalability – yes

However, some problems occurred especially with the early 1.6 series:

- Loss of data under “rare conditions” with patchless client :-) ..... solved
- Loss of data under very rare conditions still possible? ..... needs investigation
- WOM bug: If 32 Bit OS and OST with more than 2TB: WOM ..... solved
- Quota: Not working in 1.6.x ..... solved
- Quota not working for OSTs > sqrt(2) TB ..... solved
- Quota working, but not dynamically adjustable ..... solved
- Quota not working for > 4 TB ..... bug
- Root Squash not working ..... feature?
- Technical manual partly wrong for 1.6.x ..... much better
- Strange error codes ( Stone of Rosetta necessary )

our experience: lustre needs 6 months to go from raw to ripe in a series  
however: good cooperation with lustre developers

# Migration from NFS to lustre@gsi for the data file system

- We will copy data from existing nfs servers to lustre: server by server
- Each copied system will be closed for nfs, “lusterised” and dynamically integrated into lustre system (“assimilation” :-)
- With each integrated box, the size and I/O power of the collective lustre will rise .....
- Each group will get quota on the collective system according to the amount of disk space contributed
- The 100 TB offset of the “core lustre” enable us to deliver disk space immediately to the groups without the delay by banf(ing), ordering, mounting, installing, testing .....
- you just get the space immediately  
- and we book the money immediately

Fall 2008: about 0,7 PB in lustre should be reached....  
...if power and cooling available.. ;-)

# Lustre@GSI Outlook

- Next major release 1.8.x planned for autumn 2008
- RAID 5 over network in the next major release
- Kerberos
- Tests with cross site lustre
  - Tests with University Frankfurt

## Next year:

- HSM Module from CEA?
- ZFS ( available for lustre 2.x series? )
- Lustre as /home file system in failover mode with SAN switches