science + computing

A Bull Group Company

Lustre administration – and how it compares to its rivals

Daniel Kobras

science + computing ag

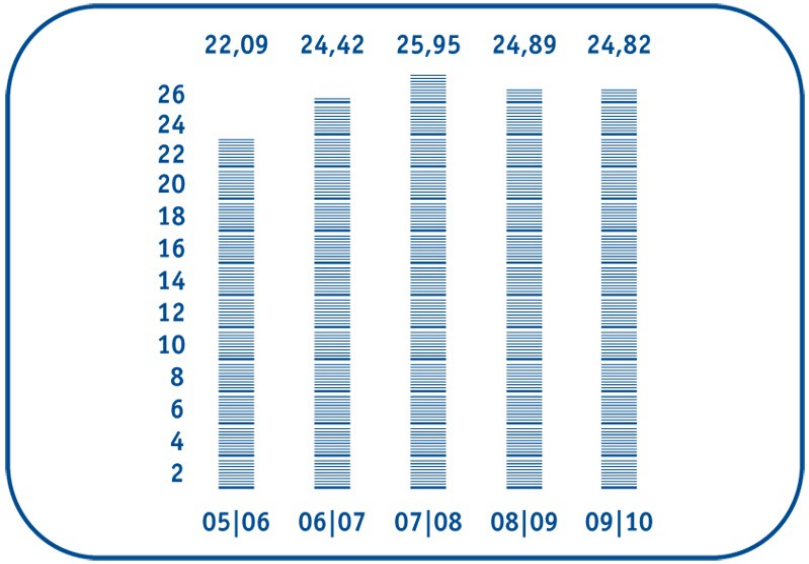
IT-Dienstleistungen und Software für anspruchsvolle Rechnernetze

Tübingen | München | Berlin | Düsseldorf

science+computing

science + computing
A Bull Group Company

Founded in	1989
Offices	Tuebingen
	Munich
	Berlin
	Duesseldorf
Employees	251
Shareholder	Bull S.A. (100%)
Turnover 09/10	24.8 Mio. Euro



Portfolio

IT Service for Complex Computing Environments
Complete solutions for Linux- and Windows-based **HPC**

scVENUS System management software for efficient administration of homogeneous and heterogeneous environments

Motivation

science + computing
A Bull Group Company

- with scalable storage, **performance** turns from a differentiator to a configurable item
- **administrative effort** becomes one of the main cost factors to consider when deciding between multiple implementations

Scalable Storage experience

science + computing
A Bull Group Company

Name	Use case	Type	Comment
Lustre	production	parallel FS	freely available (Linux)
IBM GPFS	production	parallel FS	license required (Linux, AIX)
IBM SoFS	production	parallel FS + scale-out NAS	GPFS + Samba CTDB (superseded by SONAS appliance)
HP X9000 (IBRIX)	production	scale-out NAS	global namespace
Oracle S7000	production	NAS	ZFS-based appliance
FhgFS	test	parallel FS	Linux
GlusterFS	test	parallel FS	freely available (Linux)
BlueArc Titan	deployment	scale-out NAS	HW accelerated appliance

Scalable Storage experience

science + computing
A Bull Group Company

Name	Use case	Type	Comment
Lustre	production	parallel FS	freely available (Linux)
IBM GPFS	production	parallel FS	license required (Linux, AIX)
IBM SoFS	production	parallel FS + scale-out NAS	GPFS + Samba CTDB (superseded by SONAS appliance)
HP X9000 (IBRIX)	production	scale-out NAS	global namespace
Oracle S7000	production	NAS	ZFS-based appliance
FhgFS	test	parallel FS	Linux
GlusterFS	test	parallel FS	freely available (Linux)
BlueArc Titan	deployment	scale-out NAS	HW accelerated appliance

Criteria
(incomplete, personal bias)

science + computing
A Bull Group Company

- **Configuration**
  How easily can I make my FS do what I want?
- **Transparency**
  How clearly does my FS tell me why it doesn't do what I want?
- **Storage Management**
  How does my FS reflect changes in my infrastructur?
- **Data protection**
  How does my FS help me to secure large amounts of data?

Configuration – Wish list

science + computing
A Bull Group Company

- unified configuration interface
- functionally oriented configuration commands
- central configuration
- traceable configuration
- configuration changes without downtime
- roll-back of configuration changes
- documentation

Configuration – GPFS

science + computing
A Bull Group Company

- comprehensive documentation
- configuration via custom set of commands (mm*)
- changes mostly possible in running system
- roll-out of changes via custom command set requires password-free root access between fs nodes

Configuration – Lustre

science + computing
A Bull Group Company

- comprehensive configuration possible
- comprehensive documentation
- configuration scattered across module options, mkfs/tunefs, Lustre-specific commands (lfs, lctl), or even implicit
- configuration options structured by subsystem (eg. OSS vs. OST vs. obdfilter) rather than function
- central configuration on MGS opaque
  - cannot (easily) read out current status
  - cannot roll back individual changes
- changes to network setup often require downtime

Configuration – Lustre example

science + computing
A Bull Group Company

Configure network interfaces of a Lustre server:

- options to kernel modules at LNET start time determine which interfaces are activated in which order
- list of interfaces is transmitted to MGS once at first start of the server
- clients receive server's network configuration from MGS upon start (mount)
- changes in server's network configuration become active locally, but aren't automatically forwarded to MGS or clients
- pushing changes to MGS requires wiping and replay of complete central configuration (--writeconf)

Transparency – Wish list

science + computing
A Bull Group Company

- instructive error messages
- fast and easy identification of malfunctioning components
- clear strategies for error recovery
- easy mapping of errors to affected users

Transparency – GPFS

- comprehensive troubleshooting guide
- terse error messages, impact not immediately obvious
- frequent strategy for error recovery: call support and keep fingers crossed

Transparency – GPFS example

science + computing
A Bull Group Company

- Error message on client

```
mmfs: Error=MMFS_FSSTRUCT, ID=0x94B1F045,
Tag=14402300:   Invalid disk data structure.
Error code 108.   Volume gpfs01
Sense Data … (hex dump)
```

  Which files are affected?

- Networking problem, potential data corruption

```
GPFS Deadman Switch timer [0] has expired;
IOs in progress: 0
```

Transparency – Lustre

science + computing
A Bull Group Company

- (mostly) open bug tracker
- constant stream of log messages
- not necessarily indicative of malfunction
- multitude of mostly similar messages
  -> syslog tends to combine messages, suppressing valuable information
- developer-friendly format of (most) error messages

Transparency – Lustre example

science + computing
A Bull Group Company

- typical message (MDS)

  LustreError:0:0:(ldlm_lockd.c:305:waiting_locks_callback()) ### lock callback timer expired after 101s: evicting client at 192.168.1.2@tcp ns: mds-lustre-MDT0000_UUID lock: ffff81010ca8dc00/0x2d5a67076b5b0e96 lrc: 3/0,0 mode: CR/CR res: 28424597/2754695384 bits 0x3 rrc: 2 type: IBT flags: 0x4000020 remote: 0x9b8763ea37421764 expref: 869 pid: 19255 timeout: 492121428

- typical message (client)

  Lustre: data-MDT0000-mdc-ffff81012037b900: Connection to service lustre-MDT0000 via nid 192.168.1.7@o2ib was lost; in progress operations using this service will wait for recovery to complete.
  LustreError: 167-0: This client was evicted by lustre-MDT0000; in progress operations using this service will fail.

  -> which files/users are affected?

Transparency – Lustre example

- typical message (OSS)

  LustreError: 21419:0:(ldlm_resource.c:719:ldlm_resource_add()) lvbo_init failed for resource **5719372**: rc -2

- problem with object on OST – which file is affected?

```
# debugfs -c -R "stat /O/0/d$((5719372 % 32))/5719372" \
   /dev/mpath/ost42
Inode: 12345   Type: regular    Mode: 0666    Flags: 0x80000
User: 31145    Group: 1337    Size: 4129115
(...)
Extended attributes stored in inode body:
  fid = "86 1e 23 00 00 00 00 00 ef 0a 29 81 00 00 00 00 00 64
 +12 00 00 00 00 00 00 00 00 00 00 00 00 00 " (32)
```

- affected file is inode 0x00231e86 on MDT

```
# debugfs -c -R "ncheck 0x00231e86" /dev/mpath/mdt01
2301574 /ROOT/home/user17/sim/nobelprize.dat
```

- alternatively: search complete filesystem for objid.

Storage Management – Wish list

science + computing
A Bull Group Company

- user-transparent migration of data to newly added server/from end-of-life'ed servers
- data replication
- support for different storage classes
- integration with archive systems/HSM

Storage Management – GPFS

science + computing
A Bull Group Company

- transparent migration of data between disks
- replication on GPFS level possible (separate configuration for data/metadata)
- replication level configuration per file
- management of several separate storage pools
- placement and migration policies
- Support for DMAPI (for TSM/HSM integration)

Storage Management – Lustre

science + computing
A Bull Group Company

- storage pools as groups of OSTs
- default pool assignment configurable per directory
- user can override pool assignment
- migration between OSTs only by copying
- new servers immediately become active (no burn-in testing possible)
- OST index of decommissioned servers is retained
- coming soon:
  - transparent migration
  - HSM support

Storage Management – Lustre example

science + computing
A Bull Group Company

- Pools:
  - central tools for storage management (with co-operative users), available since Lustre 1.8.0
  - but: cannot fsck MDT when using pools (Stand: Lustre 1.8.6)
- Migration:
  - possible by copying data
  - but: cannot lock down data
    - -> no central control over which data is still in use
    - -> on all clients: lsof | grep <datei>
      then: cp -p <datei> <datei>.new && \
      mv <datei>.new <datei>

Data protection – Wish list

science + computing
A Bull Group Company

- ACL support (Posix/NFSv4)
- strong authentication of
    - clients
    - users
- WAN capabilities (encryption, integrity checks, access control across domain boundaries)
- end-to-end checksums
- consistent backup of local data on each server
- snapshot functionality
- support for efficient backup on large filesystem, no full backups
- fast restore

Data protection – GPFS

science + computing
A Bull Group Company

- supports both Posix- and NFS4 ACLs
- mapping between ACL types (if possible)
- integration of several remote clusters, authenticated via key pairs
- no integrity protection via checksums
- efficient integration with TSM (mmbackup)
- backup/restore via multiple clients possible

Data protection – GPFS example

science + computing
A Bull Group Company

- „This is not supported“ phenomenon:

  mmbackup does not support file names containing quotation marks

Data protection – Lustre

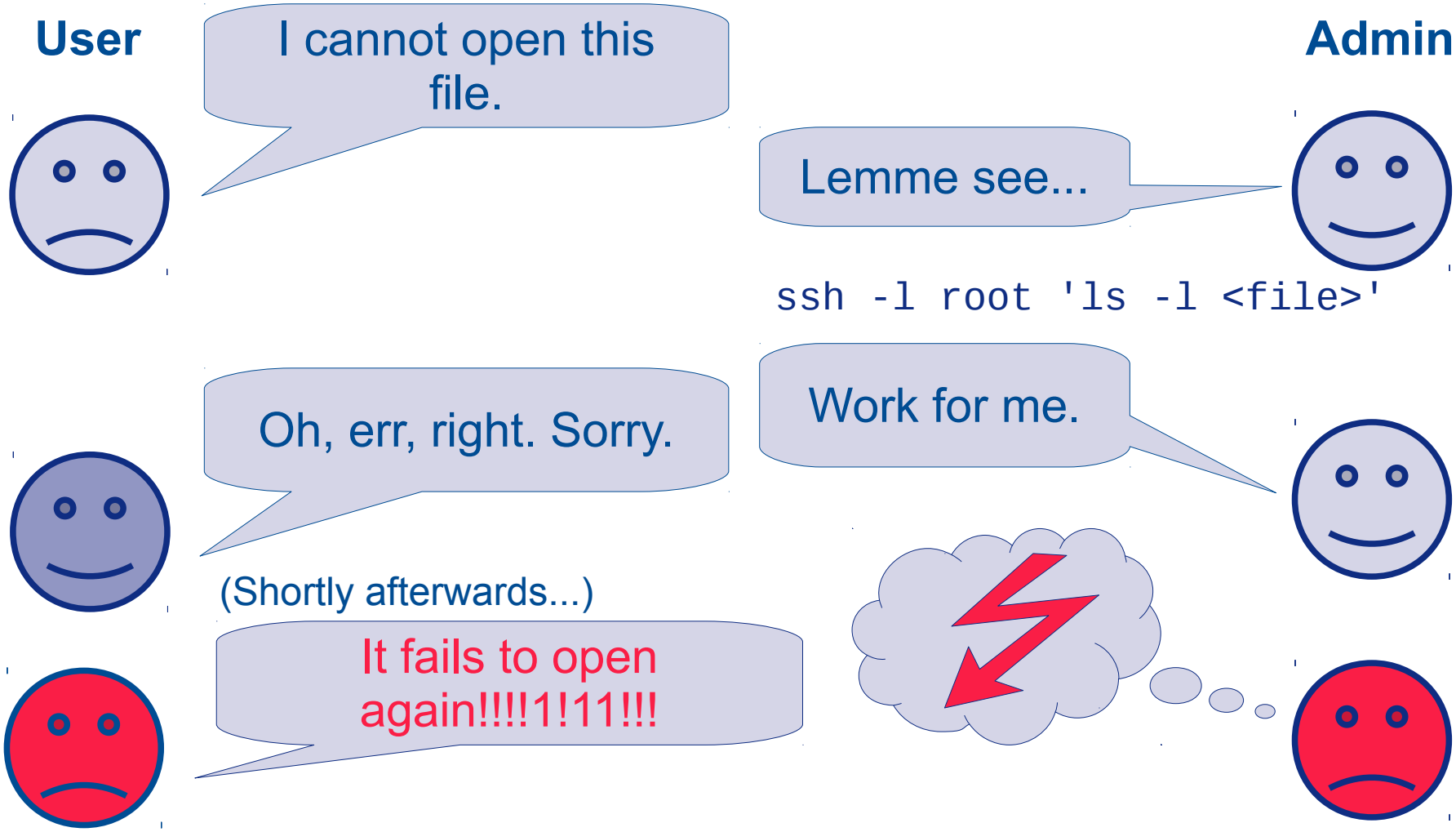
science + computing
A Bull Group Company

- Posix ACLs (currently 16 ACEs max.)
- access control on MDS
- no client authentication, only „world-wide“ export on Lustre level
- access control by UID, implicit client trust
- on-the-wire checksums
- server-side backups possible via local LVM snapshots, but not consistent across server (-> only useful on MDT)
- no snapshots on filesystem level
- backup/restore via (multiple) Lustre clients
- helper tool (e2scan) creates lists of changed files
- efficient implementation (changelogs) in Lustre 2.x

Data protection – Lustre example

science + computing
A Bull Group Company

Scenario: group mismatch between MDS and client



Data protection – Lustre example

science + computing
A Bull Group Company

- backup software capable of synthetic full backups is a must
- distribute load across several clients (subtrees) to increase backup/restore throughput
- staggered backup times to decrease MDS load
- without changelog feature, backup constrained by MDT load/performance

Conclusion

science + computing
A Bull Group Company

- **Lustre**
  - focus on users (**performance**), developers, but hardly on admins
  - tameable for the initiated (after steep learning curve)
  - open system, but admins constantly get to feel its **complexity**
  - most wanted: **GSSAPI** support, **transparent data migration**
- **GPFS**
  - more admin-friendly in general
  - closed, proprietary system may put you at the whim of support
  - shines when it comes to **data lifecycle**
- shortcomings can be alleviated with third-party tools (eg. RobinHood), and in-house extensions (eg. **`rbh-query`**)
- central storage driven by scalable filesystems still a net win in admin effort over scattered, stand-alone fileservers

science + computing

A Bull Group Company

Thank you!

Daniel Kobras
science + computing ag
www.science-computing.de
www.hpc-wissen.de
Telefon 07071 9457-0
info@science-computing.de