Bareos @ MPI SF

Dr. Stefan Vollmar
Head of IT Group
vollmar@sf.mpg.de

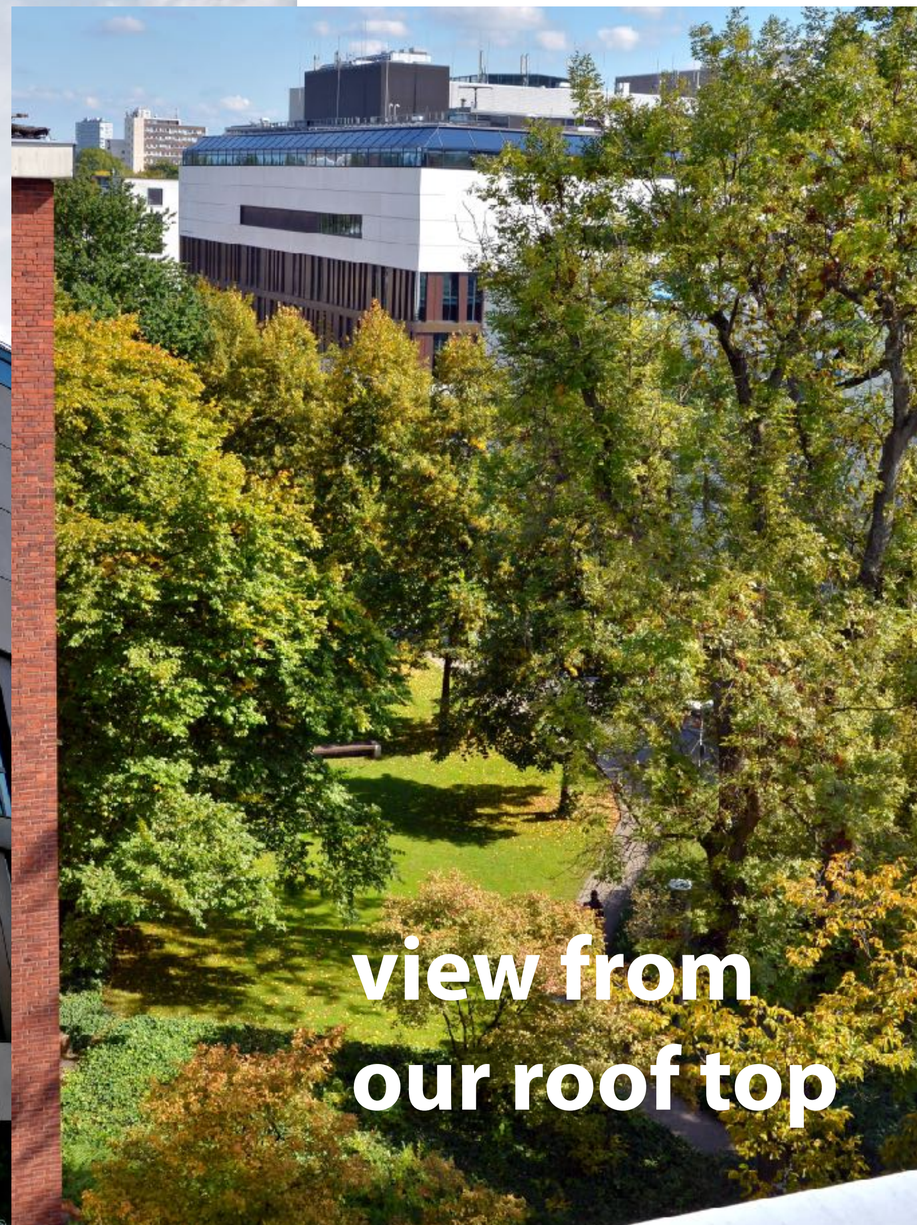
Max Planck Institute for
Metabolism Research
Cologne, Germany

MPI for Metabolism Research
Main Building

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

Neighbour Institute: MPI for Biology of Ageing

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne



view from our roof top

OpenSource @ MPI-SF

- Firefox
- Thunderbird
- LibreOffice/OpenOffice
- LaTeX

- GNU/Linux
- Samba
- Cendio ThinLinc

- VLC
- Inkscape, Gimp
- ImageMagick

- Bareos
- (BeeGFS)
- Grau OpenArchive

- OwnCloud/ Nextcloud
- LimeSurvey

- MySQL, PostgreSQL
- Apache, Nginx
- Python, PHP, Perl
- ZeroMQ
- LimeSurvey

- Ansible
- Slurm
- nagios/icinga

- mail server: dovecot
- RADIUS: freeRADIUS
- LDAP: OpenLDAP
- DNS: bind
- DHCP: isc-dhcp-server

- git
- Atom
- Emacs
- gcc

- FSL
- SPM
- FreeSurfer
- ImageJ/Fiji

Data (1)

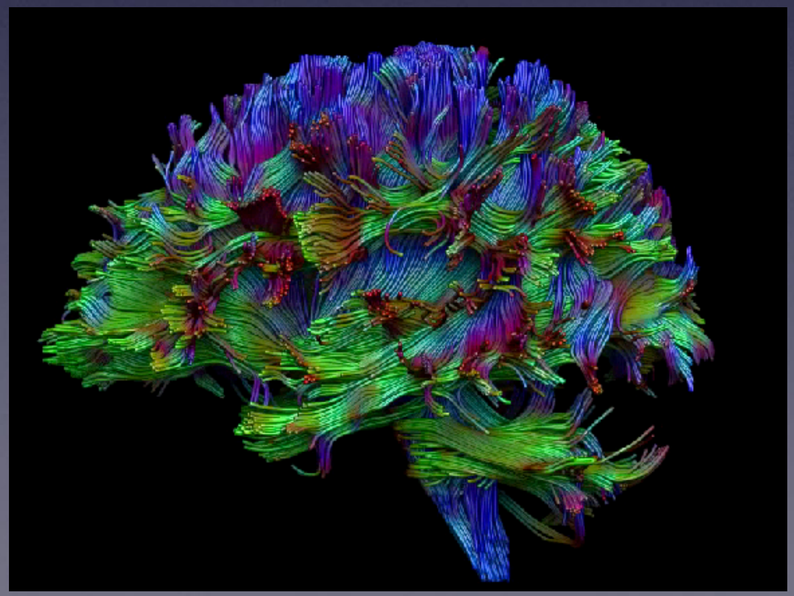
Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

MR Studies

Siemens MAGNETOM Prisma (3T)

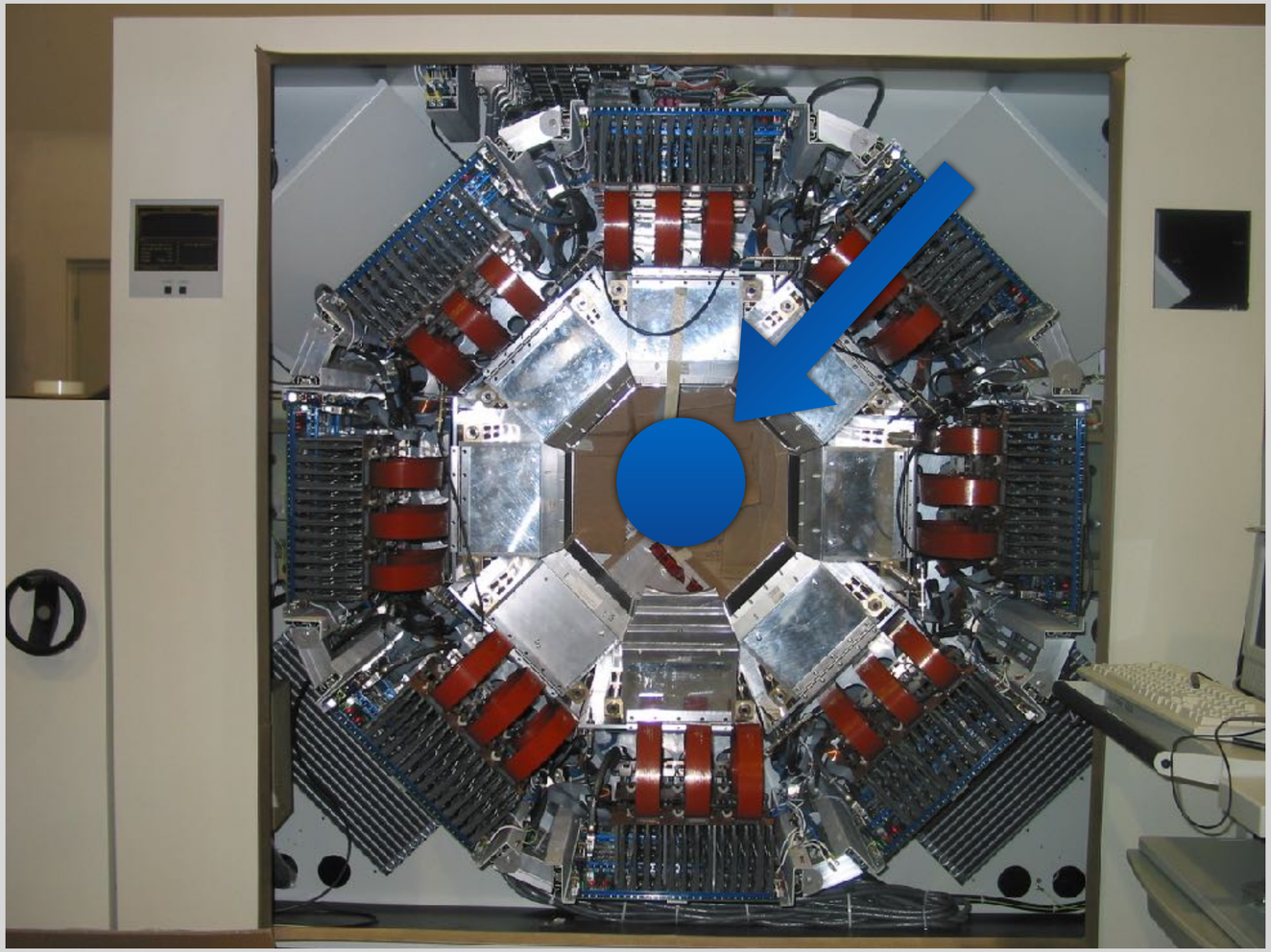
Diffusion MRI - Fibre Tracking

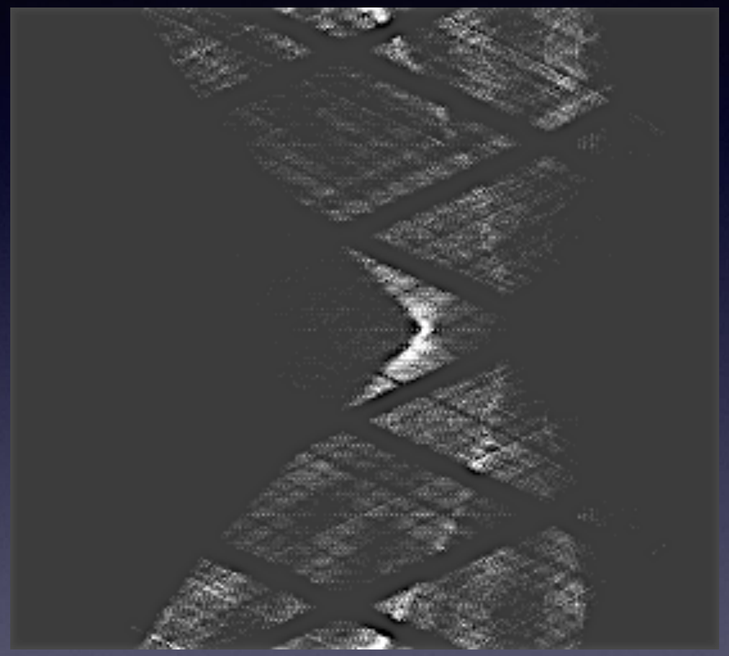
- "Nothing defines the function of a neuron better than its connections" Mesulam (2006)
- analyzing fiber structure (anatomical connectivity) in-vivo in human brains



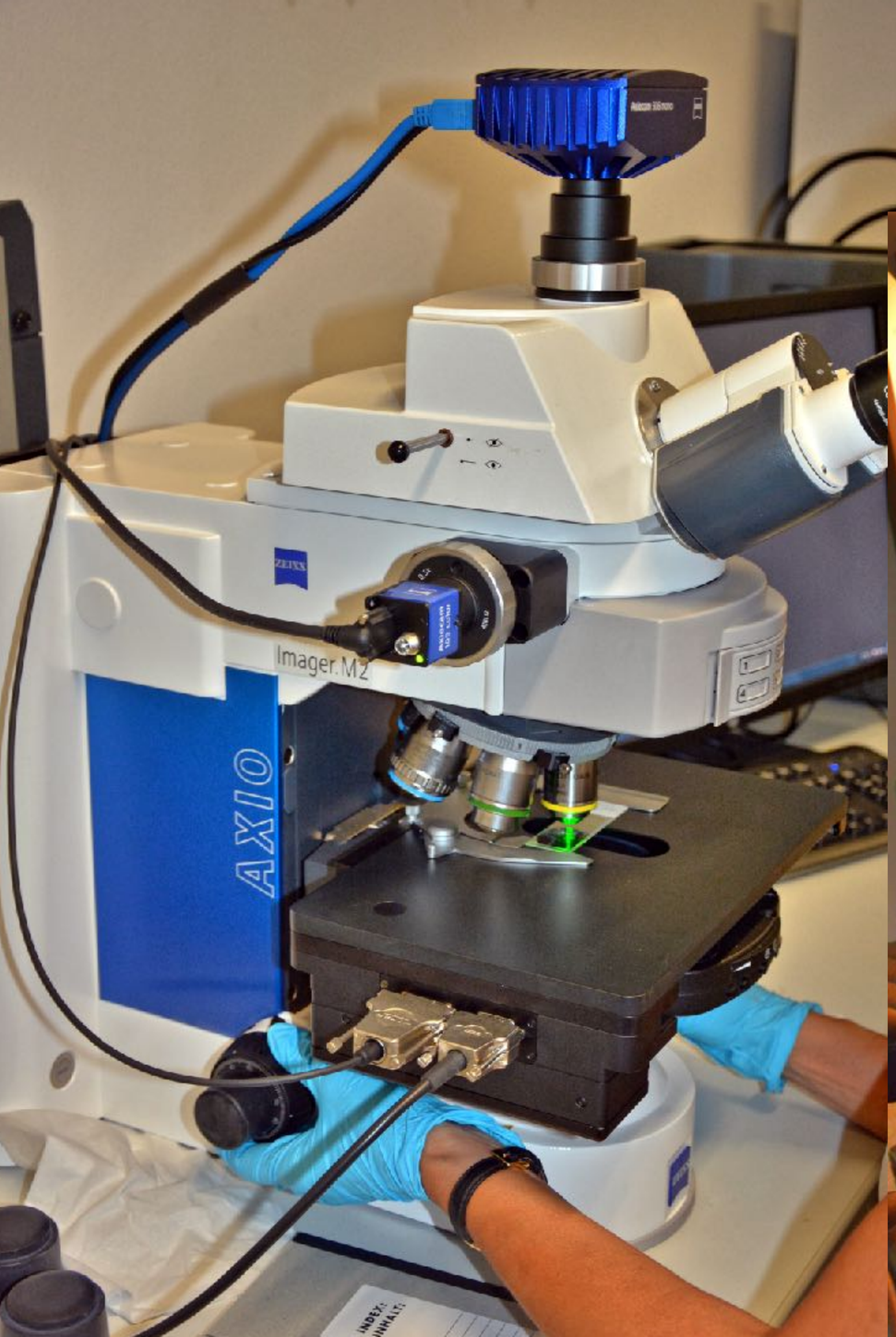
Data (2)

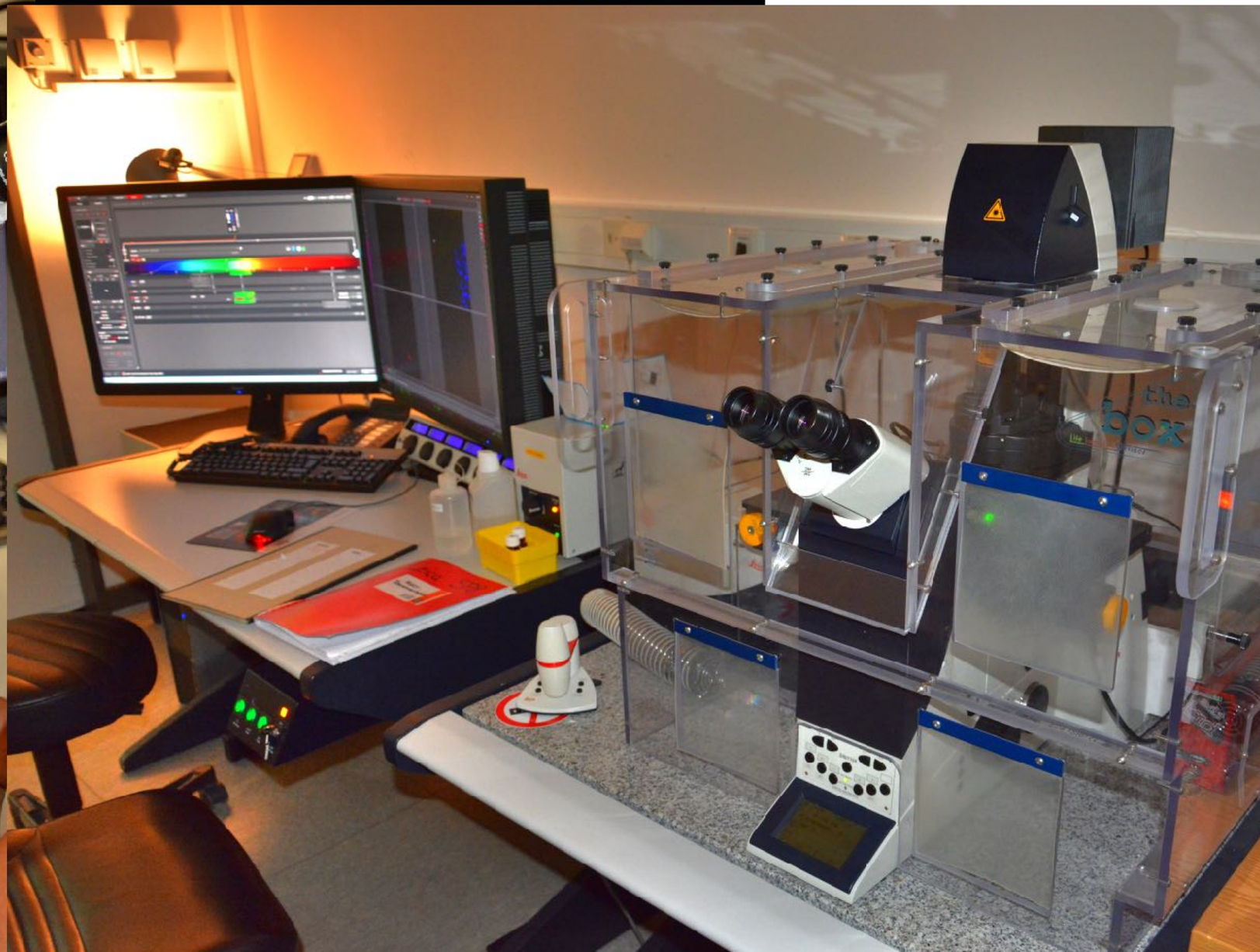
The High Resolution Research Tomograph





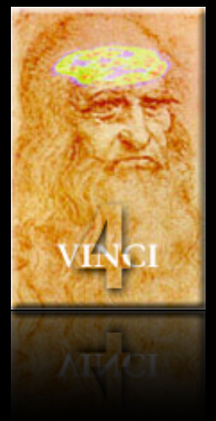
Data (3): Microscopy





VINCI Demo

V olume
I maging in
N eurological Research,
C o-Registration and ROIs
I ncluded



S. Vollmar, M. Sué, A. Hüsgen,
H. Endepols, M. Backes,
J. Čížek, K. Herholz

Max Planck Institute
for Metabolism Research

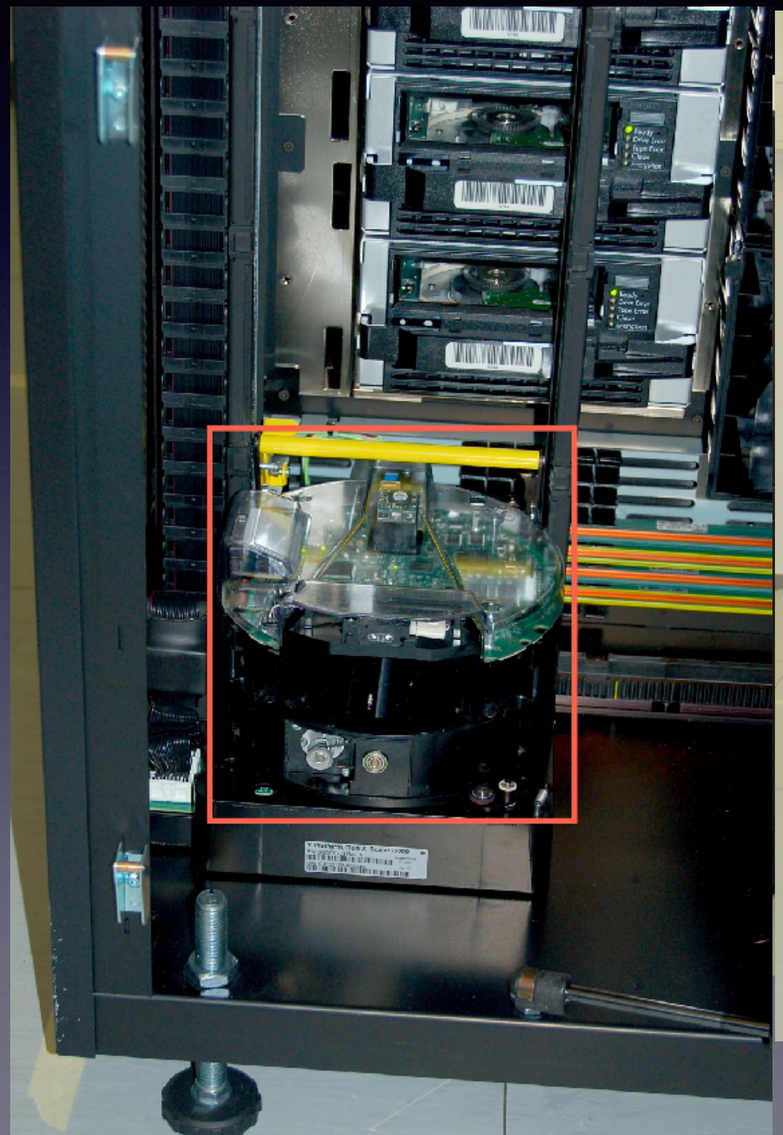
Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

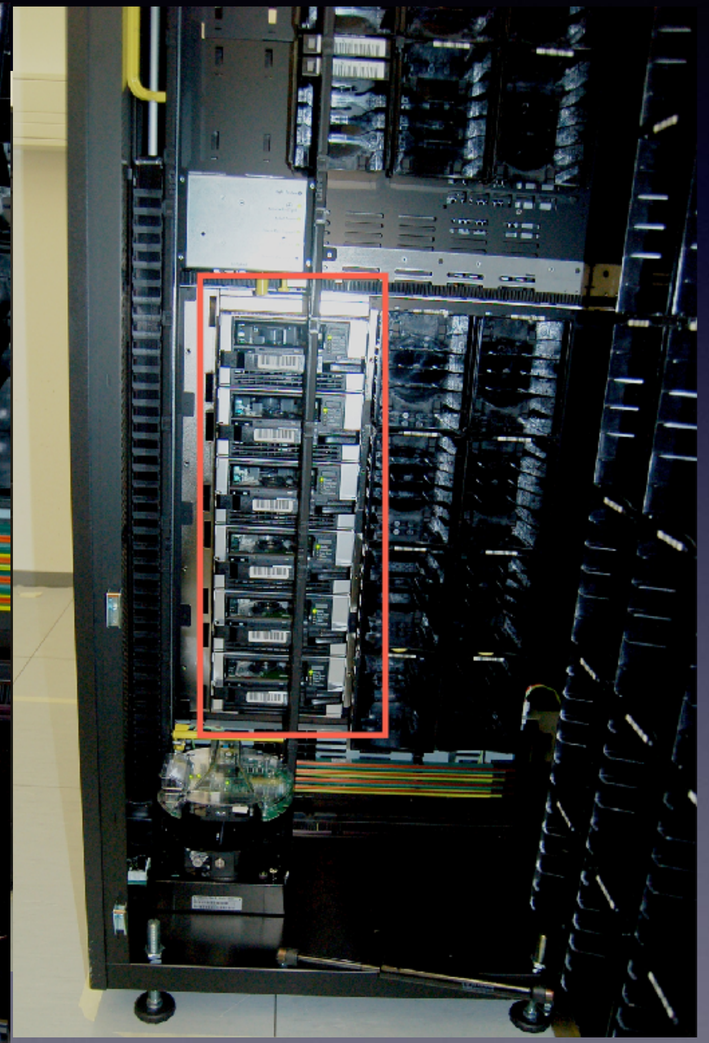
Storage and Archiving

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

LTO 6-WORM approx. 6 TB per Tape

1 MB = 1000000B = 10^6 bytes one CD ~ 700 MB
1 GB = 1000000000B = 10^9 bytes one DVD ~ 4.7 GB
1 TB = 1000000000000B = 10^{12} bytes one disk drive
1 PB = 1000000000000000B = 10^{15} bytes = 1000 terabytes





Grau OpenArchive and Bareos

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

"Historical Context"

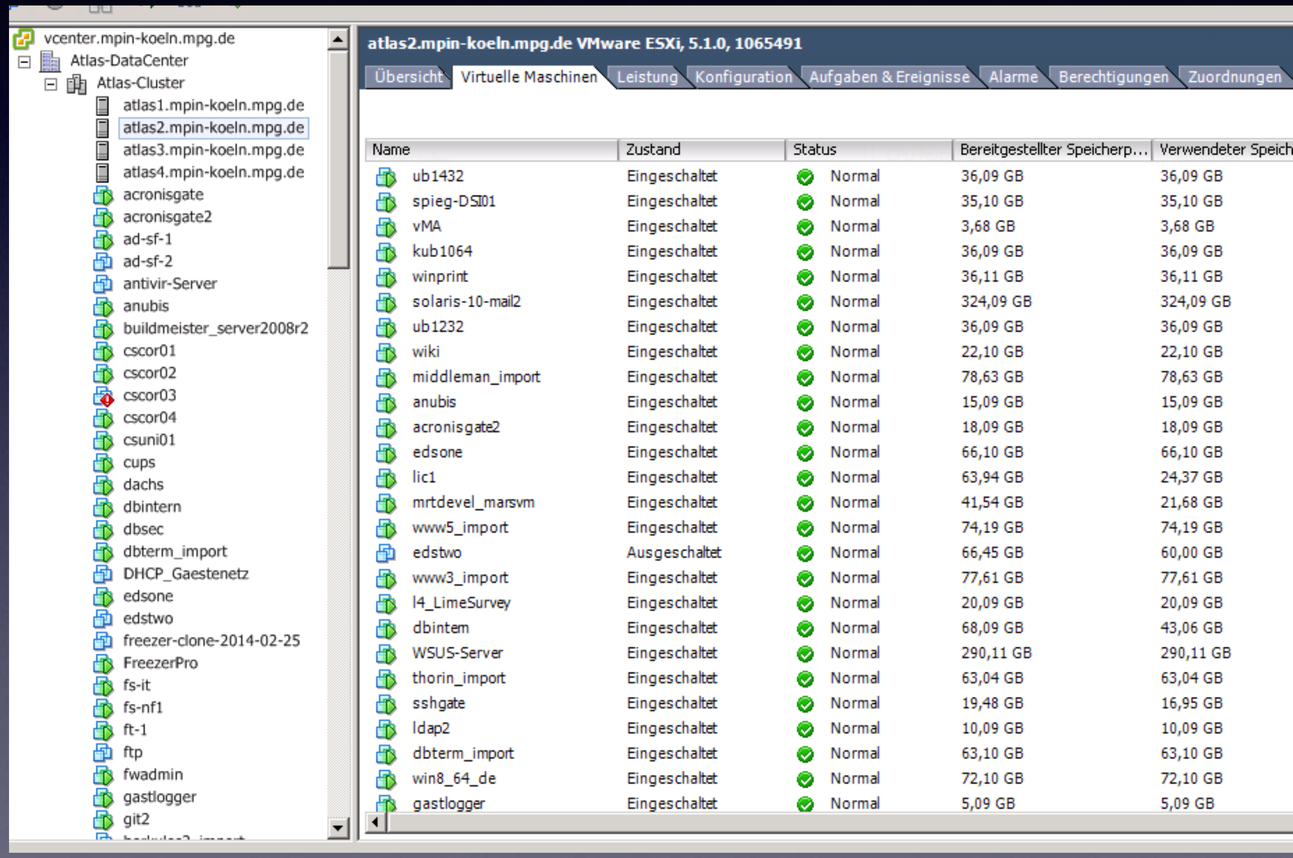
- Problem: users find it difficult to work efficiently with offline data (experiences with SAM-FS)
- new Grau OpenArchive HSM with 300 x LTO-6 in Quantum i6000-Library for offline storage
- But: GAM setup is not suitable for original SAM-FS concept - no shared pools, need fewer file systems
- Tried Bareos for backup with concept of D. Jahn (contac Datentechnik GmbH): backup-to-disk into a HSM file system
- Started with one virtual Bareos-Server for backup of smaller servers, added virtual network cards for backup in closed network segments. We use run-before and run-after jobs to enable/disable respective NICs.

Bareos - bread and butter

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

4 x HP DL585 G7
> 1 TB RAM
10-Gig-E networking
FC-SAN
approx. 64 cores each

approx. 100 virtual machines,
VMWare vSphere ESX cluster



Backup-to-Disk on HSM

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

- Files should have a sensible size limit (here: 5 GB)
- HSM needs suitable “MinFileAge”
- one job per virtual tape (file) to avoid writing earlier data of a particular file more than once

small but important database: full backup every Sunday

```
-rw-r-----  1 bareos bareos  2.8G Sep  3 02:34 labdb-50711
-rw-r-----  1 bareos bareos  114M Sep  4 22:31 labdb-50928
-rw-r-----  1 bareos bareos  164M Sep  5 22:31 labdb-50961
-rw-r-----  1 bareos bareos  164M Sep  6 22:31 labdb-51000
-rw-r-----  1 bareos bareos  106M Sep  7 22:31 labdb-51040
-rw-r-----  1 bareos bareos  164M Sep  8 22:31 labdb-51078
-rw-r-----  1 bareos bareos   86M Sep  9 22:30 labdb-51110
-rw-r-----  1 bareos bareos  2.9G Sep 10 02:34 labdb-51123
-rw-r-----  1 bareos bareos  111M Sep 11 22:30 labdb-51314
-rw-r-----  1 bareos bareos  106M Sep 12 22:30 labdb-51352
-rw-r-----  1 bareos bareos  164M Sep 13 22:31 labdb-51387
-rw-r-----  1 bareos bareos  164M Sep 14 22:31 labdb-51424
-rw-r-----  1 bareos bareos  164M Sep 15 22:30 labdb-51464
-rw-r-----  1 bareos bareos   89M Sep 16 22:30 labdb-51497
-rw-r-----  1 bareos bareos  114M Sep 18 22:31 labdb-51513
-rw-r-----  1 bareos bareos  164M Sep 19 22:31 labdb-51613
-rw-r-----  1 bareos bareos  164M Sep 20 22:31 labdb-51648
-rw-r-----  1 bareos bareos  164M Sep 21 22:30 labdb-51682
-rw-r-----  1 bareos bareos  165M Sep 22 22:30 labdb-51718
-rw-r-----  1 bareos bareos   86M Sep 23 22:30 labdb-51747
```

large project dir after full backup

```
-rw-r----- 1 bareos bareos 3113574226 Mar 20  2017 v1-kfo2-63953
-rw-r----- 1 bareos bareos 5368688709 Mar 20  2017 v1-kfo2-63947
-rw-r----- 1 bareos bareos        573 Mar 18  2017 v1-kfo2-63683
-rw-r----- 1 bareos bareos    4564090 Mar 17  2017 v1-kfo2-63461
-rw-r----- 1 bareos bareos        573 Mar 16  2017 v1-kfo2-63327
-rw-r----- 1 bareos bareos  358724677 Mar 15  2017 v1-kfo2-63245
-rw-r----- 1 bareos bareos        573 Mar 14  2017 v1-kfo2-63181
-rw-r----- 1 bareos bareos        573 Mar 13  2017 v1-kfo2-63104
-rw-r----- 1 bareos bareos        573 Mar 11  2017 v1-kfo2-63035
-rw-r----- 1 bareos bareos  148653493 Mar 10  2017 v1-kfo2-62986
-rw-r----- 1 bareos bareos        573 Mar  9  2017 v1-kfo2-62934
-rw-r----- 1 bareos bareos        573 Mar  8  2017 v1-kfo2-62879
-rw-r----- 1 bareos bareos   25588334 Mar  8  2017 v1-kfo2-62832
-rw-r----- 1 bareos bareos       1599 Mar  7  2017 v1-kfo2-62812
-rw-r----- 1 bareos bareos    8716459 Mar  6  2017 v1-kfo2-62722
-rw-r----- 1 bareos bareos 4465650813 Mar  5  2017 v1-kfo2-62588
-rw-r----- 1 bareos bareos 5368688723 Mar  5  2017 v1-kfo2-62587
-rw-r----- 1 bareos bareos 5368688682 Mar  5  2017 v1-kfo2-62586
-rw-r----- 1 bareos bareos 5368688716 Mar  5  2017 v1-kfo2-62585
-rw-r----- 1 bareos bareos 5368688722 Mar  5  2017 v1-kfo2-62584
-rw-r----- 1 bareos bareos 5368688706 Mar  5  2017 v1-kfo2-62583
-rw-r----- 1 bareos bareos 5368688686 Mar  5  2017 v1-kfo2-62582
-rw-r----- 1 bareos bareos 5368688718 Mar  5  2017 v1-kfo2-62581
-rw-r----- 1 bareos bareos 5368688763 Mar  5  2017 v1-kfo2-62580
-rw-r----- 1 bareos bareos 5368688767 Mar  5  2017 v1-kfo2-62579
-rw-r----- 1 bareos bareos 5368688757 Mar  5  2017 v1-kfo2-62578
-rw-r----- 1 bareos bareos 5368688740 Mar  5  2017 v1-kfo2-62577
-rw-r----- 1 bareos bareos 5368688747 Mar  5  2017 v1-kfo2-62576
-rw-r----- 1 bareos bareos 5368688739 Mar  5  2017 v1-kfo2-62575
-rw-r----- 1 bareos bareos 5368688737 Mar  5  2017 v1-kfo2-62574
-rw-r----- 1 bareos bareos 5368688758 Mar  5  2017 v1-kfo2-62573
-rw-r----- 1 bareos bareos 5368688771 Mar  5  2017 v1-kfo2-62572
```

Adding Clients with Templates


```
root@:conf.d# add-client.sh marvin
```

templates courtesy of J. Behrend, thanks!

add-client.sh

```
mkdir "/var/lib/bareos/storage/$1"
chown bareos "/var/lib/bareos/storage/$1"
cp _template-linux.dir.conf_ "$1.dir.conf"
cp _template-linux.sd.conf_ "$1.sd.conf"
sed -i "s/XXX/$1/g" $1.dir.conf
sed -i "s/XXX/$1/g" $1.sd.conf
```

template-linux.sd.conf

```
Device {
 Name = XXX
 Media Type = XXX
 Archive Device = /var/lib/bareos/storage/XXX
 LabelMedia = yes;       # lets Bareos label unlabeled media
 Random Access = yes;
 AutomaticMount = yes;  # when device opened, read it
 RemovableMedia = no;
 AlwaysOpen = no;
}
```

template-linux.dir.conf

```
Schedule {
 Name = "XXX-all"
 Run = Level=Full sun at 2:25
 Run = Level=Incremental mon-sat at 22:30
}
#
Job {
 Name = "XXX-all"
 Type = Backup
 Level = Incremental
 Client = XXX-fd
 FileSet= "XXX-all"
 Messages = Standard
 Storage = XXX
 Pool = XXX
 Accurate = true
 Schedule = XXX-all
}
#
Client {
 Name = XXX-fd
 Address = XXX
 Catalog = MyCatalog
...
Pool {
 Name = XXX
 Pool Type = Backup
 LabelFormat = "XXX-"
 Maximum Volume Jobs = 1
 Maximum Volume Bytes = 5G
 Recycle = no
}
```

```
root@esdb:/backup_postgres/autopostgresqlbackup# tree .
.
├── daily
│   ├── esdb
│   │   ├── esdb_2017-09-19_06h25m.Tuesday.sql.gz
│   │   ├── esdb_2017-09-20_06h25m.Wednesday.sql.gz
│   │   ├── esdb_2017-09-21_06h25m.Thursday.sql.gz
│   │   ├── esdb_2017-09-22_06h25m.Friday.sql.gz
…
│   │   ├── template1_2017-09-22_06h25m.Friday.sql.gz
│   │   ├── template1_2017-09-24_06h25m.Sunday.sql.gz
│   │   └── template1_2017-09-25_06h25m.Monday.sql.gz
│   └── test_esdb
│       ├── test_esdb_2017-09-19_06h25m.Tuesday.sql.gz
│       ├── test_esdb_2017-09-20_06h25m.Wednesday.sql.gz
│       ├── test_esdb_2017-09-21_06h25m.Thursday.sql.gz
│       ├── test_esdb_2017-09-22_06h25m.Friday.sql.gz
│       ├── test_esdb_2017-09-24_06h25m.Sunday.sql.gz
│       └── test_esdb_2017-09-25_06h25m.Monday.sql.gz
├── latest
│   ├── esdb_2017-09-25_06h25m.Monday.sql.gz
│   ├── postgres_2017-09-25_06h25m.Monday.sql.gz
│   ├── postgres_globals_2017-09-25_06h25m.Monday.sql.gz
│   ├── template1_2017-09-25_06h25m.Monday.sql.gz
│   └── test_esdb_2017-09-25_06h25m.Monday.sql.gz
├── monthly …
└── weekly
    ├── esdb
    │   ├── esdb_week.36.2017-09-09_06h25m.sql.gz
    │   ├── esdb_week.37.2017-09-16_06h25m.sql.gz
    │   └── esdb_week.38.2017-09-23_06h25m.sql.gz
    ├── postgres
    │   ├── postgres_week.36.2017-09-09_06h25m.sql.gz
    │   ├── postgres_week.37.2017-09-16_06h25m.sql.gz
```

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

Backup of small DB servers

- mostly (very) small MySQL/MariaDB or PostgreSQL databases
- Cronjob 1: create dump files (for each database) several times a day, store them locally
- Cronjob 2: run daily and make sure that older dumps are deleted as necessary
- Run Bareos several times daily
- new: use package autopostgresqlbackup

Why BeeGFS?

- had/have these file systems:
  - DataCore HP EVA/HP P2000
  - Sun SAM-FS (HSM)
  - Grau Archive Manager (HSM)
- conventional file systems
  - too slow
  - too small
  - too expensive
- compared parallel filesystems and found reasons why Lustre, Gluster and Ceph are not ideal for us
- BeeGFS not perfect but might get there: already very fast, very robust, very cost efficient, comparatively easy to install and maintain - but still needs: (more) redundancy, ACLs, quota enforcement
- Last but not least: commercial support and a very convincing configuration concept by ThincParQ with HP components

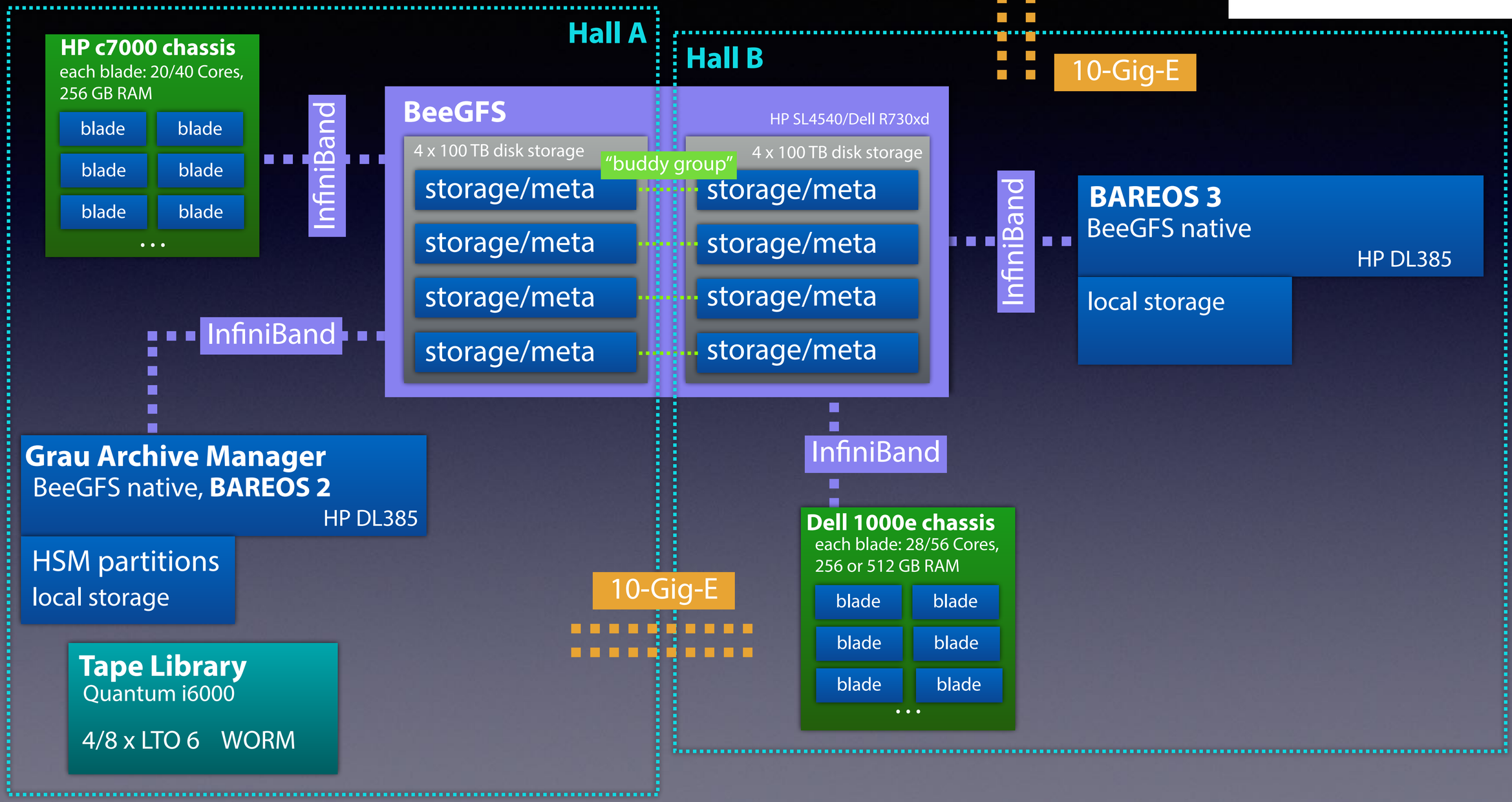
Why BeeGFS? beegfs-2015.03

- BeeGFS not perfect but might get there: already very fast, very robust, very cost efficient, comparatively easy to install and maintain - but still needs: redundant data chunks, ACLs, quota enforcement, redundant metadata servers, redundant management server

Why BeeGFS? beegfs-6.0

- BeeGFS not perfect but might get there: already very fast, very robust, very cost efficient, comparatively easy to install and maintain - but still needs: ~~redundant data chunks~~, ~~ACLs~~, ~~quota enforcement~~, ~~redundant metadata servers~~, redundant management server

BeeGFS HPC Environment



BeeGFS

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

BeeGFS Installation (1)

MPI-SF and MPI-Age 2014-2017

- HP SL4540 servers
  here: 2 servers in one chassis
  CentOS 7.3
- Infiniband QDR, FDR
- approx. 1 PB raw capacity for both Max Planck Institutes
- HP c7000 Blade Center
  Blades HP BL 460c 256 GB RAM
- Cluster management with Ansible
- SLURM for job management
- ThinLinc as SunRay replacement

HPC (2)

MPI-SF and MPI-Age 2014-2017

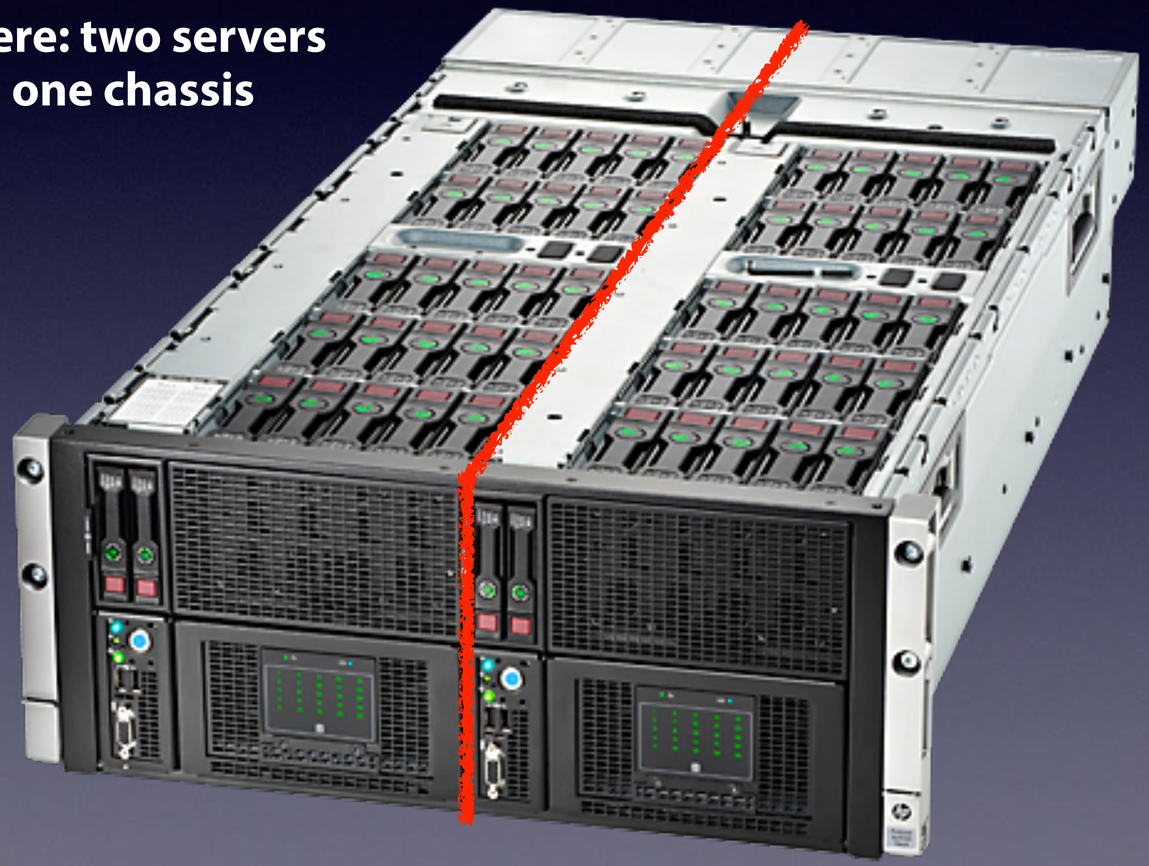
Dell PowerEdge M 1000e

- 15 blades
- 256/512 GB RAM
- 2 x 28 Cores each

BeeGFS Server (1)

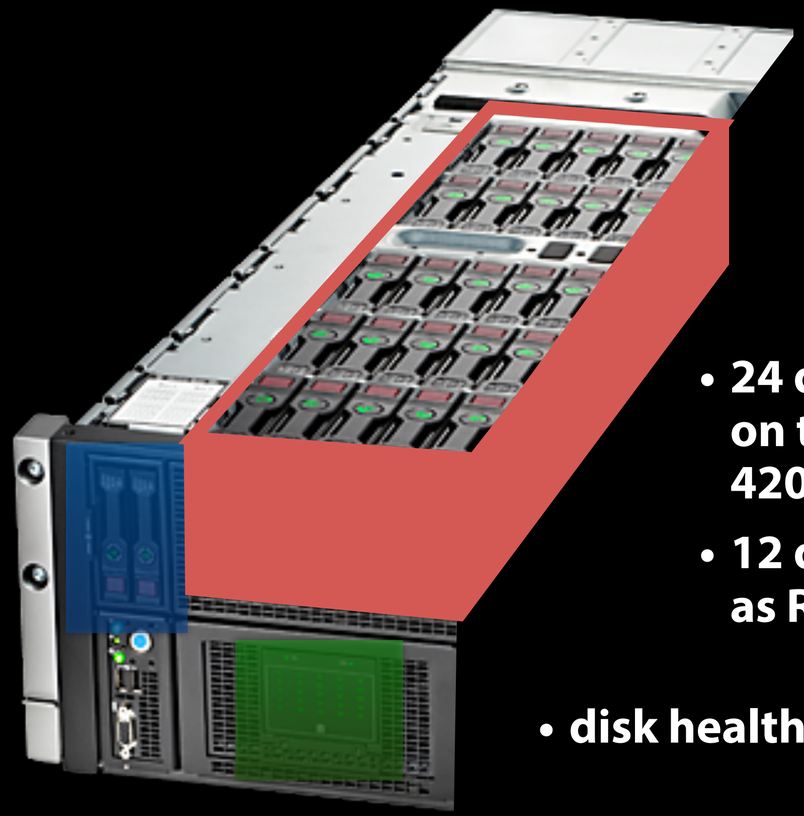
BeeGFS configuration HP SL4540

here: two servers
in one chassis



BeeGFS Server (2)

BeeGFS configuration HP SL4540



- SSD-RAID for Metadata and OS (800 GB)
- 24 disks, here: 4 TB each on two HP "Smart" Array 420i Controllers
- 12 disks per Controller as RAID6
- disk health on front panel

BeeGFS Performance


```
iozone -s16g -r1m -i0 -i1 -x -+n -t32

Children see throughput for 32 initial writers = 4907143.33 KB/sec
 Parent sees throughput for 32 initial writers = 4754264.38 KB/sec

[...]

Children see throughput for 32 readers = 6667562.31 KB/sec
 Parent sees throughput for 32 readers = 5905766.36 KB/sec
```

We see approx. 5 GB/s (writing) and 6 GB/s (reading) with *iozone*, if one configures a suitable number of threads (here: 32) in order to take advantage of Infiniband networking.
Benchmark for one blade and 4 BeeGFS nodes.

Git Filter for Bareos Config

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

```
[root@gam1 bareos]# more premove.sh
sed 's/.*Password.*/  Password = "<removed>"/'
[root@gam1 bareos]# git config filter.premove.smudge premove
[root@gam1 bareos]# git config filter.premove.clean premove
[root@gam1 bareos]# echo '* filter=premove' > .git/info/attributes
```

```
[root@gam1 bareos]# more bareos-dir.conf
...
Director {
  Name = gam1-dir
  QueryFile = "/usr/lib/bareos/scripts/query.sql"
  Maximum Concurrent Jobs = 10
  Password = "TeSTuEOklZ04/nMvOtV/UzeUElS"
  Messages = Daemon
...
```

on the institute's Github Server:

```
49  #
50  Director {
51    Name = gam1-dir
52    QueryFile = "/usr/lib/bareos/scrip
53    Maximum Concurrent Jobs = 10
54    Password = "<removed>"
55    Messages = Daemon
```

on the institute's Github Server:

changed compression type for vision

master

gam1 committed 30 seconds ago

Showing 1 changed file with 2 additions and 2 deletions.

4 conf.d/v1-vision_group.dir.conf

```
@@ -31,8 +31,8 @@ FileSet {
31  31    Include {
32  32      Options {
33  33        signature = SHA1
34      -     compression = LZ4
35      -     accurate = ms
    34  +     compression = GZIP
    35  +     accurate = mcs
36  36      }
37  37      File = /beegfs/v1/vision_group
38  38      # CAVEAT: check for other file syste
```

0 comments on commit 1c56464

Write Preview

Benchmark GZIP vs LZ4

161.2 GB total, approx 40,000 files in 2407 directories; incremental: approx 23 secs

total size	elapsed	written	%
uncompressed	42 mins 14 secs	161.2 GB	
LZ4	57 mins 37 secs	116.6 GB	27.7
GZIP	3 hours 16 mins 27 secs	93.86 GB	41.8

278 GB total, approx 0.9 million files in 29209 directories; incremental: approx 13 min

total size	elapsed	written	%
uncompressed	4 hours 51 mins 48 secs	297.2 GB	
LZ4	4 hours 25 mins 24 secs	265.4 GB	10.7

Unmounted Filesystems: problem and fix

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

Problem: after maintenance, a file system configured for Bareos backup might not be available, but Bareos

- **does not backup anything**
- **worse: it marks all previous files as deleted - next backup will be "full"; you can delete the corresponding Jobs and this will fix the database**

```
*delete jobid=21811
Jobid 21811 and associated records deleted from the catalog.
…
```

Permanent Fix with "run-before-job":

```
[root@gam1 conf.d]# more run-before-target-available.sh
#!/bin/bash
# expects absolute directory path in first argument
if [ ! -d "$1" ]; then
  # directory not available
  exit -1
fi
```

Unmounted Filesystems: fix

Configure RunBeforeJob in template:

```
Job {
  Name = "v1-XXX"
  Type = Backup
  …
  Pool = "v1-XXX"
  Accurate = true
  Schedule = "v1-XXX"
  RunBeforeJob = "/etc/bareos/conf.d/run-before-target-available.sh '/beegfs/v1/XXX'"
  Allow Duplicate Jobs = no
  …
```

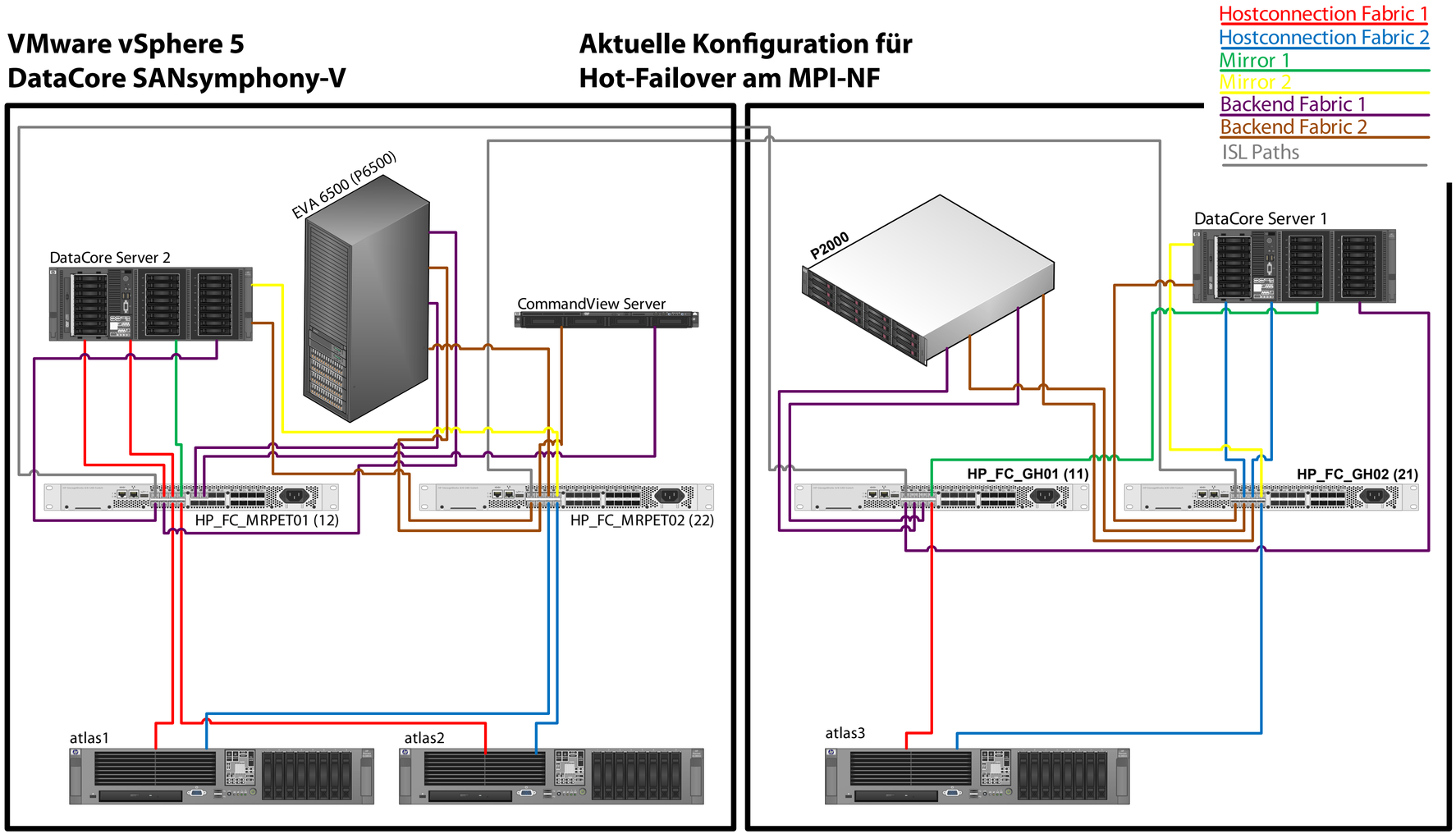
Job will fail (“die early” principle) if run-before-target-available.sh returns non-zero value:

```
25-Sep 10:58 gam1-dir: Console [default] from [127.0.0.1] cmdline run
25-Sep 10:58 gam1-dir JobId 8: shell command: run BeforeJob "/etc/bareos/conf.d/run-before-target-
available.sh 'v1-opt'"
25-Sep 10:58 gam1-dir JobId 8: Error: Runscript: BeforeJob returned non-zero status=255. ERR=Unknown
error during program execvp
```

Hot Failover (1): very complex…

VMware vSphere 5
DataCore SANsymphony-V

Aktuelle Konfiguration für
Hot-Failover am MPI-NF



MPI-NF: Kliniksgebäude PET/MR-Lab, Raum U08

MPI-NF: Hauptgebäude, Raum K04

More than “lukewarm” Failover (2)

Open Source Backup CONFERENCE
25 – 26 September, 2017 | Cologne

London Symphony Orchestra
Sir Simon Rattle, 2019-09-17
https://www.youtube.com/watch?v=73IcKxhM118

Principal Oboe: Olivier Stankiewicz
“XVIII Descent into Oboe Hell” 2:21:03

Hector Berlioz,
“The Damnation of Faust”

SilentBrick-Library Fast-LTA, München

SilentBrick-Library (2)

- complements Bareos concept: independent user-accessible “full” backup on read-only shares - “psychological factor”
- WORM-like file system
- additional building
- significant speedup with modified “parallel rsync” by W. Glick, https://wiki.ncsa.illinois.edu/display/~wglick/2013/11/01/Parallel+Rsync to profit from our parallel file system using 20+ threads for one transfer
- last but not least: backup of bricks with Bareos - avoid vendor lock-in