Clustered/Distributed File Systems

Petr Medonos, Lukáš Heřbolt

O nás

Petr Medonos

- Bc. ININ VŠCHT
- 5+ let v ETN
- RHCE, MongoDB for DBA
- databáze, performance, bezpečnost, ...
- realizace/dohled nad vetšinou současných projektů v ETN

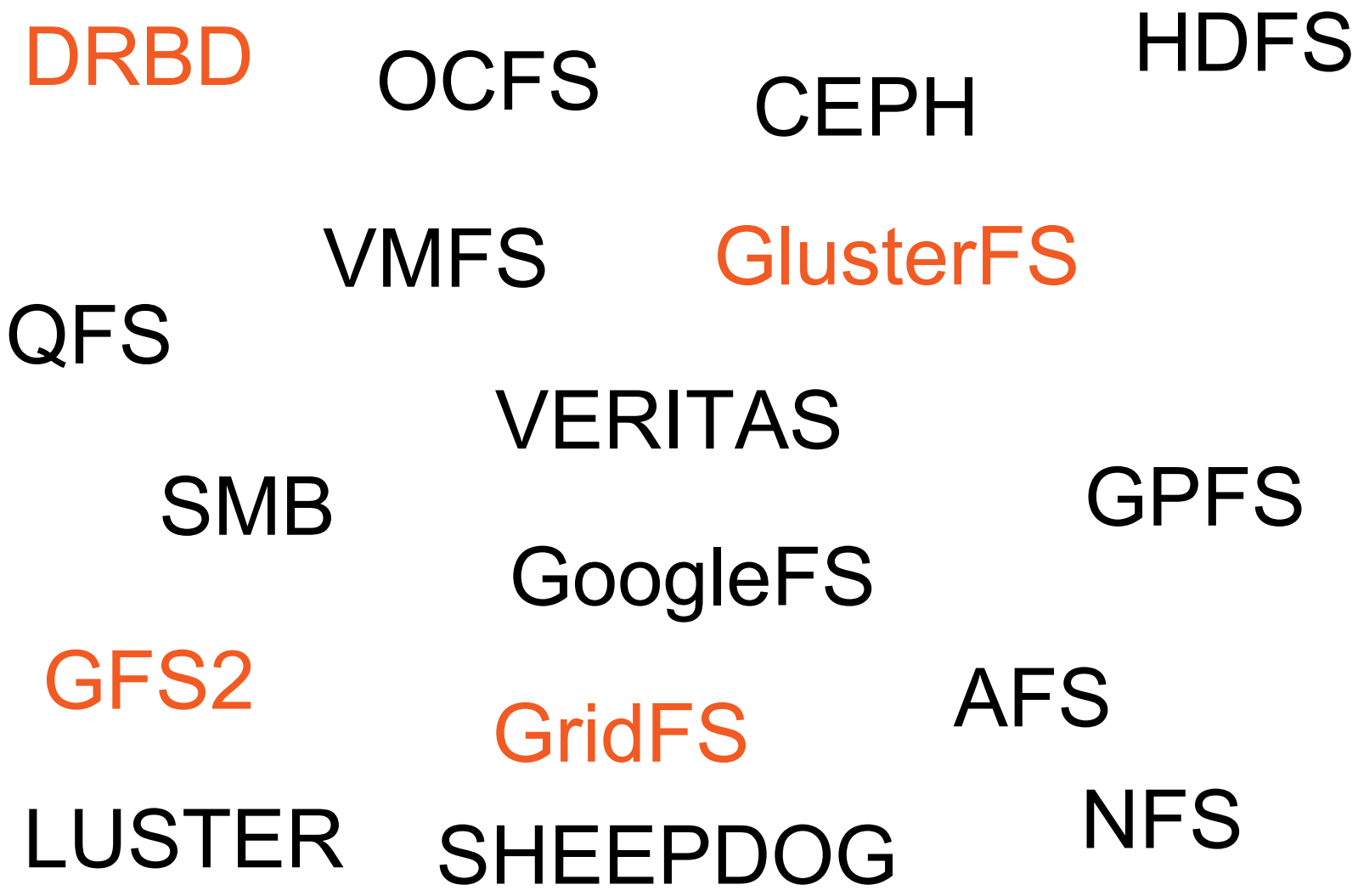
O nás

Lukáš Heřbolt

- Bc. FEL ČVUT
- 2 roky v ETN
- MongoDB for DBA
- Loadbalancing, cachování
- Realizace projektů
  - TO2 - (extravyhody.cz, firemnitelefony.cz), Fast, mojeallianz.cz, moje.partners.cz

Obsah

- **Možnosti**
- **Storage komponenty**
- **Clusterovaný vs. distribuovaný**
- **Block level (DRBD)**
- **FS**
  - GlusterFS
  - GFS2
  - GridFS


DRBD
OCFS
CEPH
HDFS
VMFS
GlusterFS
QFS
VERITAS
SMB
GPFS
GoogleFS
GFS2
AFS
GridFS
NFS
LUSTER
SHEEPDOG

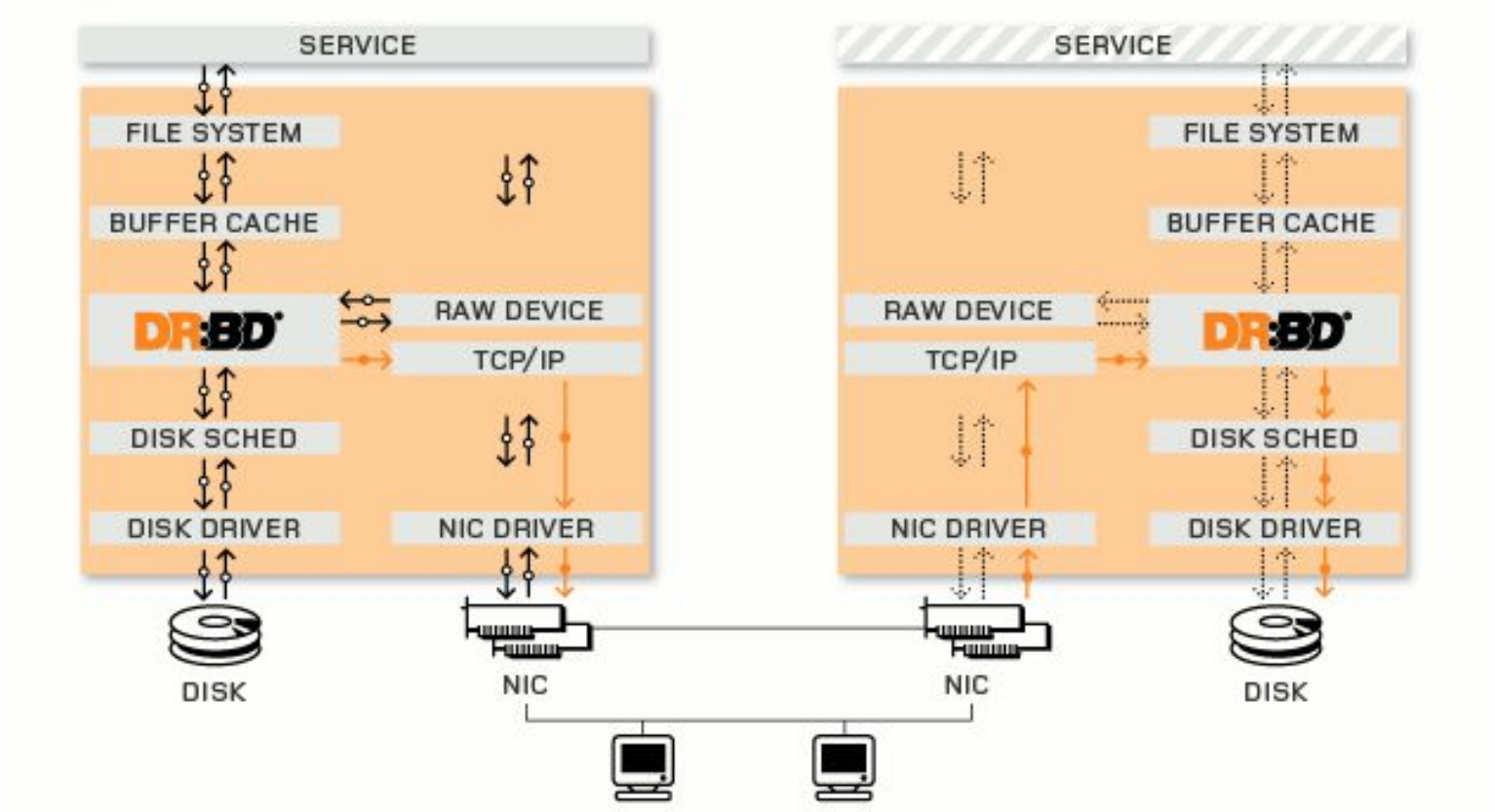
Clusterový vs. distribuovaný FS

- SAN vs. NAS
- block vs. file level
- cluster x parallel x distributed
- metadata
- split brain

DRBD

- replikovaný block device
- levná náhrada SAN
- od v. 8 umožnuje i dual-primary provoz
- synchronní i asynchonní replikace
- šifrování komunikace

DRBD


SERVICE
FILE SYSTEM
BUFFER CACHE
DRBD
RAW DEVICE
TCP/IP
DISK SCHED
DISK DRIVER
NIC DRIVER
DISK
NIC
SERVICE
FILE SYSTEM
BUFFER CACHE
RAW DEVICE
TCP/IP
DRBD
DISK SCHED
NIC DRIVER
DISK DRIVER
NIC
DISK

DRBD - konfigurace 1/2

```
common {
   startup {
      become-primary-on both;
   }}
resource r0 {
   net {
      protocol C;
      allow-two-primaries;
   }
   disk {
      on-io-error detach;
      resync-rate 700M;}
```

DRBD - konfigurace 2/2

```
on drbd-test1.local {
    device     /dev/drbd0;
    disk       /dev/sdb1;
    address    192.168.1.164:7789;
    meta-disk internal;
}
on drbd-test2.local {
    device     /dev/drbd0;
    disk       /dev/sdb1;
    address    192.168.1.207:7789;
    meta-disk internal;}}
```

DRBD - inicializace resourcu

```
drbdadm create-md <resource | all>
drbdadm up <resource | all>
drbdadm primary --force <resource>

cat /proc/drbd
```

DRBD - spit brain

victim:

```
drbdadm disconnect <resource>
drbdadm secondary <resource>
drbdadm -- --discard-my-data connect
<resource>
```

survivor:

```
drbdadm connect <resource>
```

victim:

```
drbdadm primary <resource>
```

GlusterFS

- distribuovaný POSIX FS
- nemá vlastní metadata server
- NAS
  - TCP/IP
  - infiniband
- snadno škálovatelný ~ PB
- XFS
- geo-replication

Základní pojmy

- peer
  - server, na kterém běží glusterd a sdílí volumy
- brick
  - zakladní stavební jednotka uložiště v GlusterFS - FS mountpoint
- translator
  - reší zakladní logiku mezi uložištěm a vysdílenými daty
- volume
  - spojené bricky prohnané přes translator

Elastic hashing algoritmus

- hash z cesty+názvu souboru
- přidělení logického disku podle hashe
- oddělení logického a fyzického uložiště
- přejmenování/přesouvání souboru

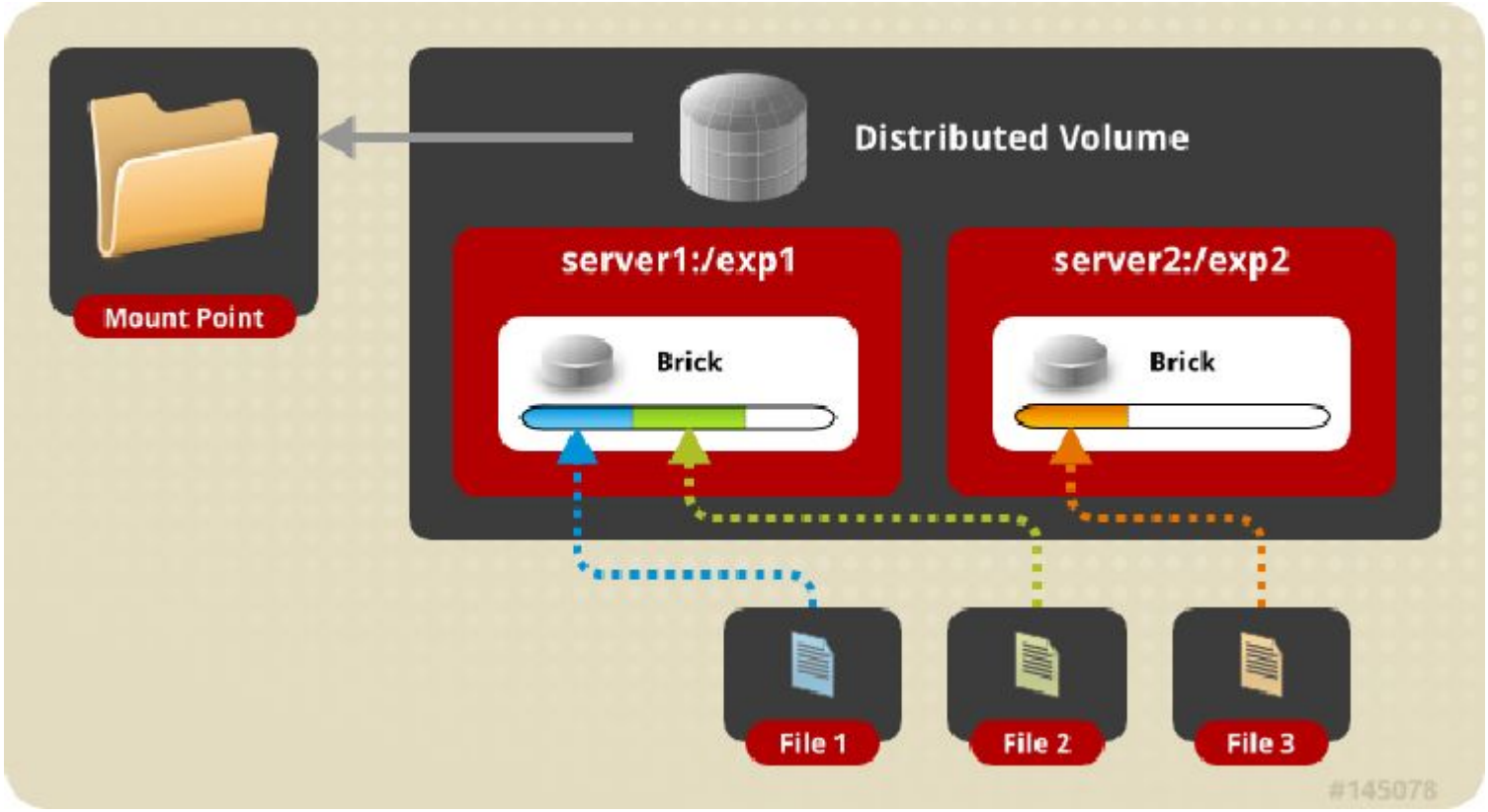
Přístup k datům

- nativní glusterfs client (FUSE)
- NFS
- SMB/CIFS

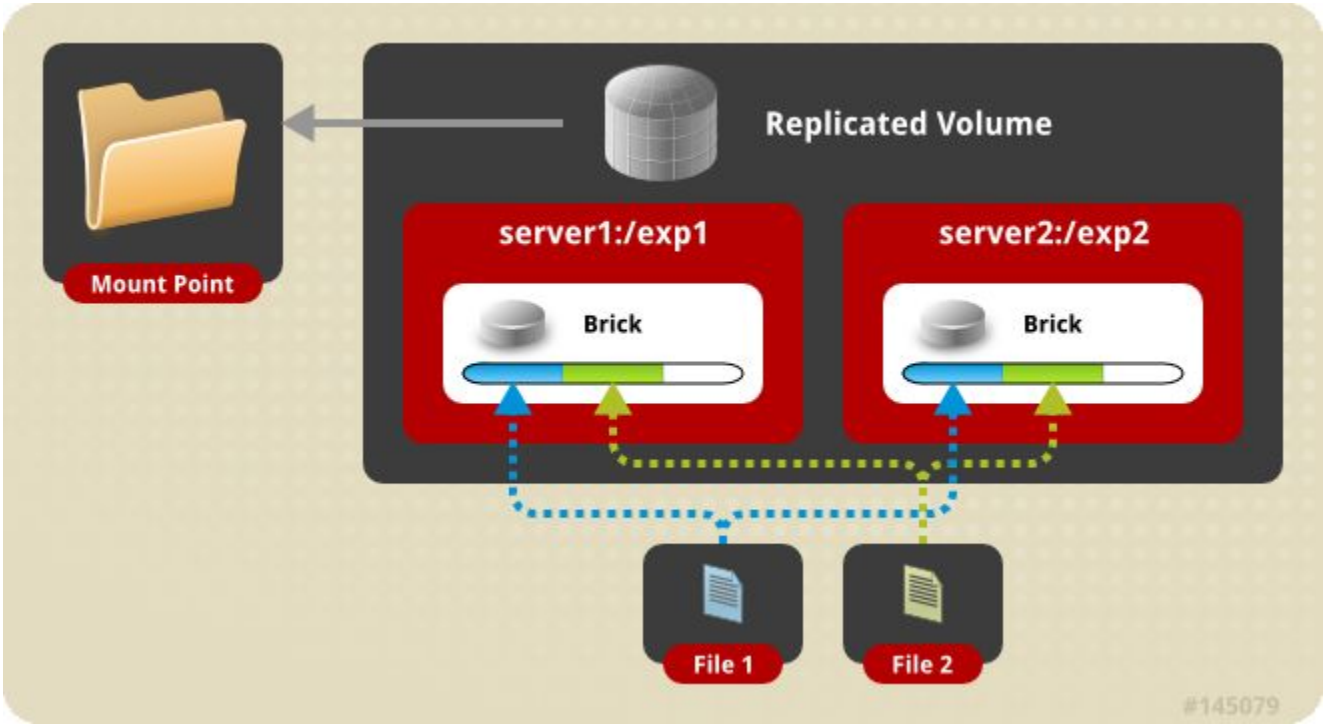
Redundance

- výchozí nastavení volumu - distributable!

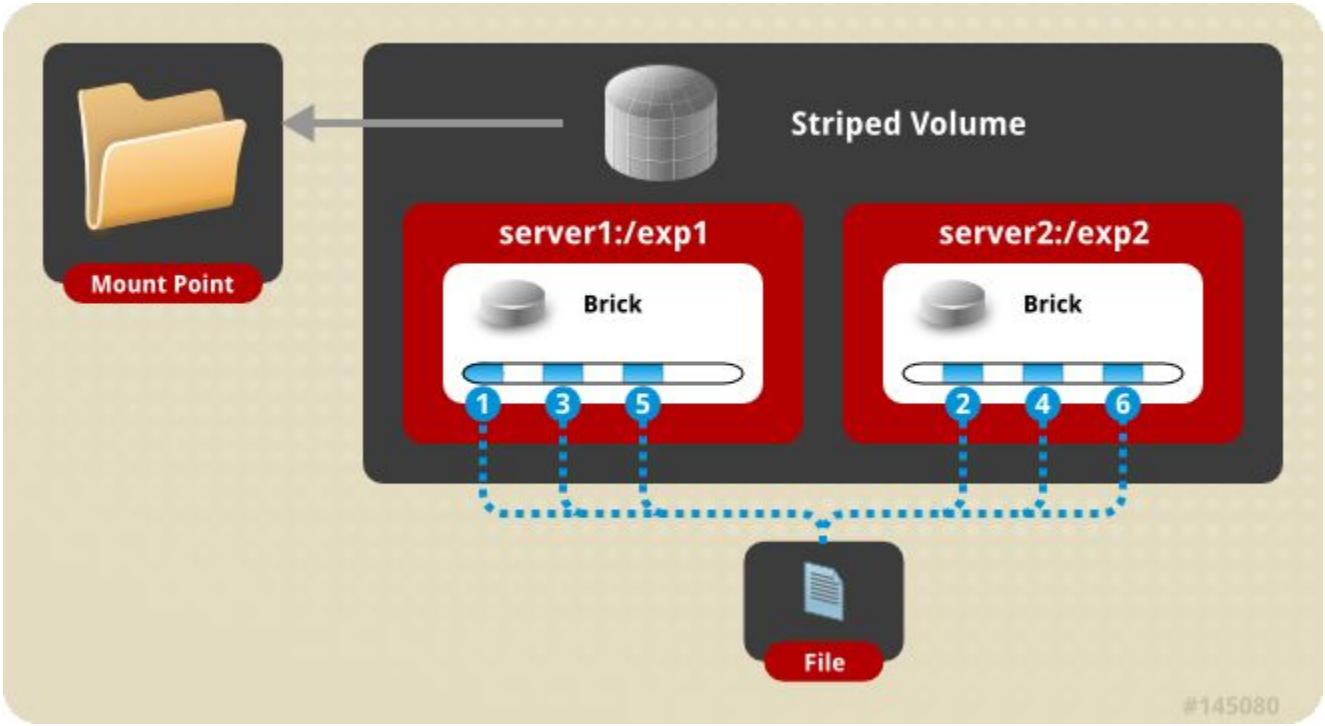
Distribuovaný volume



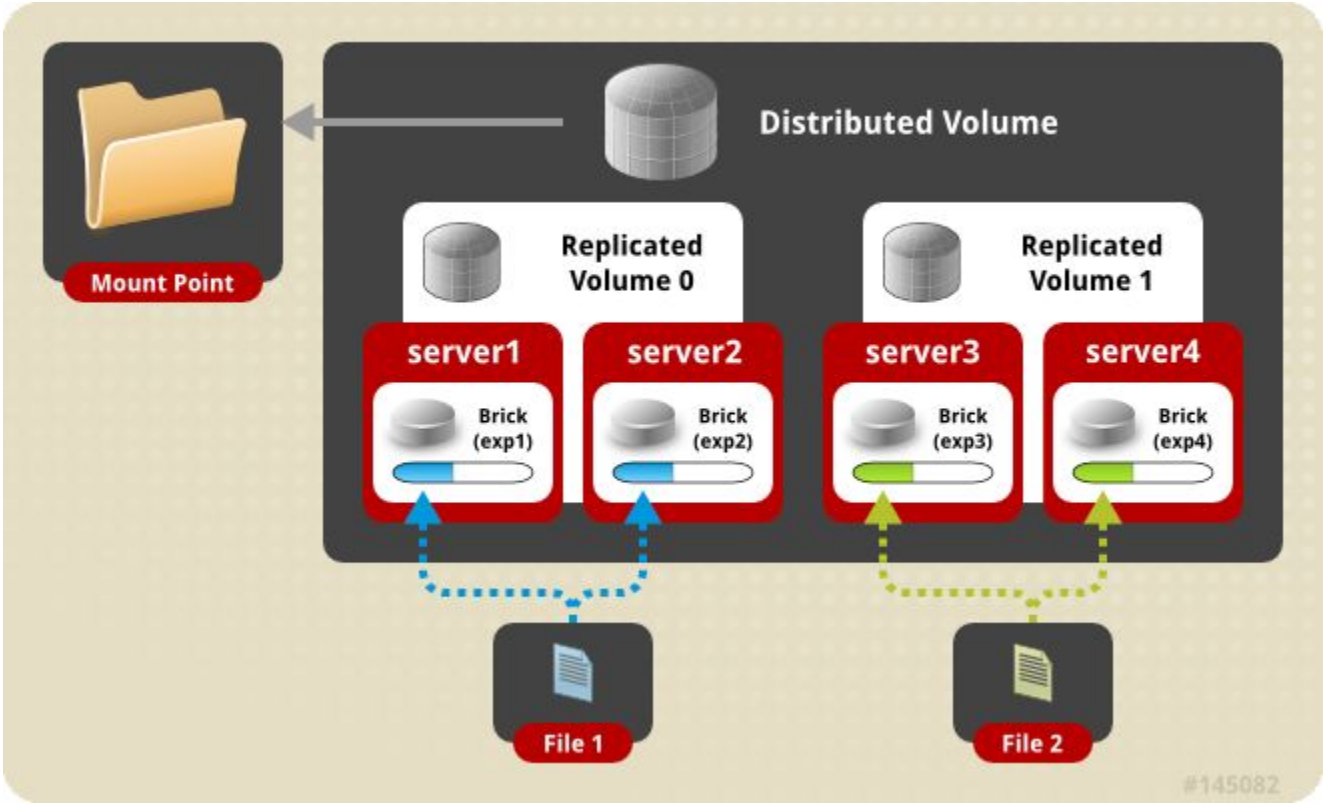
Replikovaný volume



Striped volume



Distribuovaný replikovaný volume



Možnosti volumů - summary

- **distributed**
- **replicated**
- striped
- distributed striped
- **distributed replicated**
- distributed striped replicated
- striped replicated

Geo-replication

- asynchronní
- master-slave
- rsync :)

Vytvoření bricku

```
[root@node1.local ~]# mkdir /brick1

[root@node2.local ~]# mkdir /brick2
```

Vytvoření replikovaného volumu

```
[root@node1.local ~]# gluster peer probe
 node2.local
[root@node1.local ~]# gluster peer status
[root@node1.local ~]# gluster volume create
 data replica 2 node1.local:/brick1
 node2.local:/brick2
[root@node1.local ~]# gluster volume start
 data
```

Pridani bricku

```
# gluster volume add-brick data
  c1-n1.local:/brick3 c1-n2.local:/brick4

# gluster volume rebalance data fix-layout
  start

# gluster volume rebalance data status
```

Volume shrink

```
# gluster volume remove-brick data
  c1-n1.local:/brick3 c1-n2.local:/brick4
  start

# gluster volume remove-brick data
  c1-n1.local:/brick3 c1-n2.local:/brick4
  status

# gluster volume remove-brick data
  c1-n1.local:/brick3 c1-n2.local:/brick4
  commit
```

Nastavení geo-replikace

```
# gluster peer probe c1-n1.local

/var/lib/glusterd/geo-replication/gsyncd.conf
  - remote_gsyncd =
  /usr/libexec/glusterfs/gsyncd

# ssh-keygen -f
  /var/lib/glusterd/geo-replication/secret.pe
  m
# ssh-copy-id -i
  /var/lib/glusterd/geo-replication/secret.pe
  m.pub root@c1-n1.local
```

Nastavení geo-replikace

```
# gluster volume geo-replication data
  ssh://c1-n1.local:/mnt/gluster-backup start

# gluster volume geo-replication  status
```

Rebalancing

- nutné vždy po přidání, odebrání bricku
- fix-layout
- fix-layout a migrace dat

Red Hat Cluster Suite (GFS2)

- **Kdy ho použít**
  - FC
  - FCoE
  - iSCSI
  - SAS
  - DRBD (master - master)
  - AoE
  - KVM
- **kolekce démonů pro běh clusteru a HA**
  - clvmd
  - dlm_controld
  - gfs_controld
  - rgmanager
  - ricci
  - cman
  - qdisk
  - fenced

Red Hat Cluster Suite

- **DLM**
  - distrinuted lock manager
  - správa zámků nad v CLVM a GFS2
    - cman
    - qdiskd
- **CLVM**
  - clustered Logical Volume Manager
  - distribuce konfigurace přes všechny nody v clusteru
- **GFS2**
  - Global File System 2
    - pozor na GPFS od IBM a GFS od Google
  - přímý konkureční přístup k sdílenému block device
  - journalování
  - online resize
  - nepodporuje SELinux contexty
  - řízení přístupu k datům přes DLM

Red Hat Cluster Suite

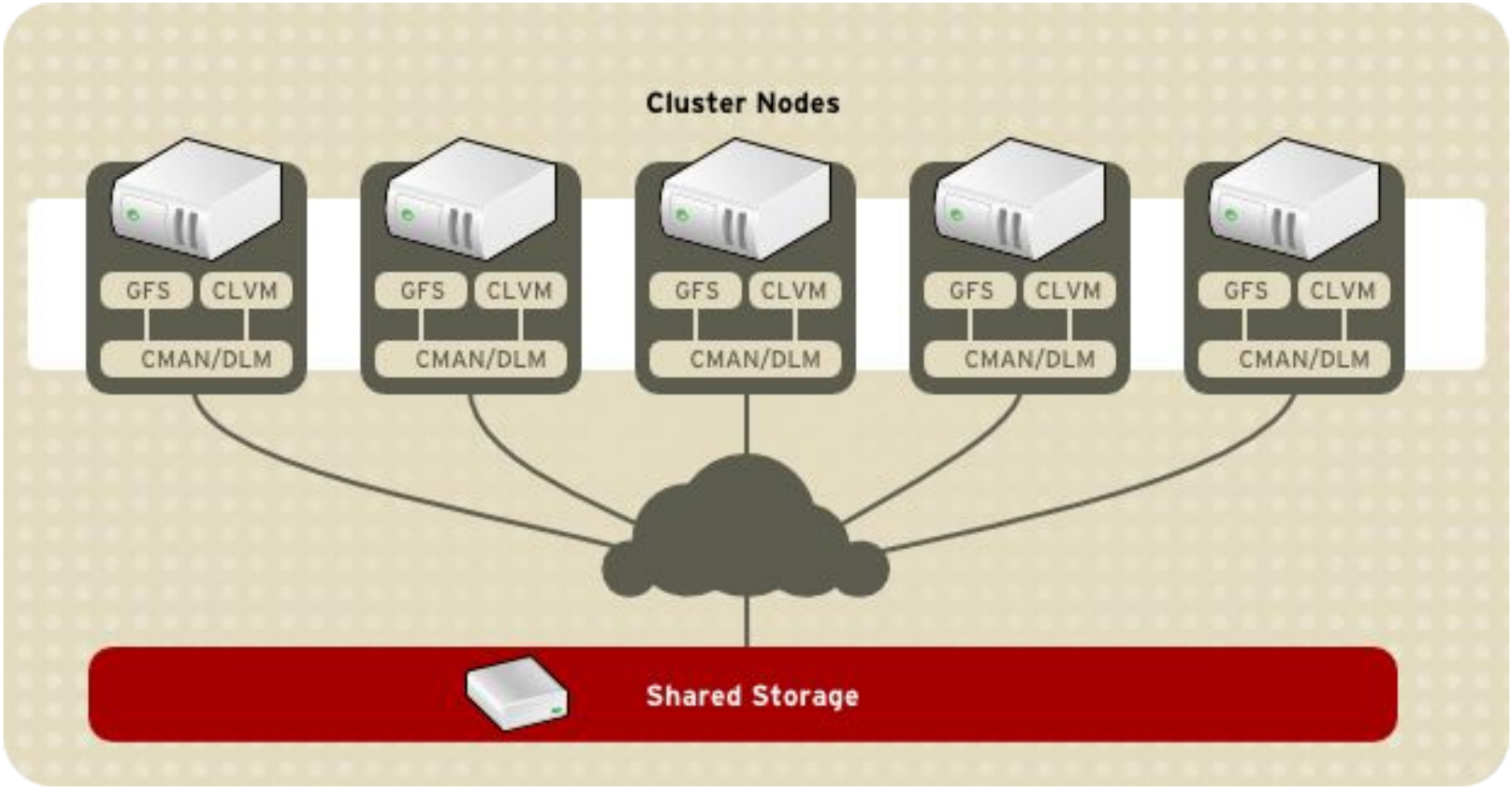
/sys/kernel/debug/dlm/clvmd_locks

LOCK	NL	CR	CW	PR	PW	EX
NL	YES	YES	YES	YES	YES	YES
CR	YES	YES	YES	YES	YES	NO
CW	YES	YES	YES	NO	NO	NO
PR	YES	YES	NO	YES	NO	NO
PW	YES	YES	NO	NO	NO	NO
EX	YES	NO	NO	NO	NO	NO

Red Hat Cluster Suit

- **GFS2 glocks**
  - 1 inode 2 zámky
    - iopen
    - inode glock
      - UN - NL
      - SH - PR
      - DF - CW
      - EX - EX

Red Hat Cluster Suite



Red Hat Cluster Suite

Nastavení

- **CLVM**
  - `vgcreate --clustered y`

    ```
    VG            #PV #LV #SN Attr   VSize   VFree
    shared-ebs-1   1   1   0 wz--nc 928,00g 89,80g
    ```
  - `/etc/lvm/lvm.conf`

    ```
    locking_type = 3
    ```
- **CMAN**
  - `/etc/cluster/cluster.conf`
  - `/usr/share/cluster/cluster.rng - xml schema`
- **GFS2**
  - `mkfs.gfs2 -p LockProtoName -t LockTableName -j NumberJournals BlockDevice`
    - LockProtoName - dlm_lock

Red Hat Cluster Suit

```
<?xml version="1.0"?>
   <cluster config_version="13" name="cluster1">
      <cman expected_votes="1" two_node="1"/> <!--2
nodes-->
      <clusternodes>
         <clusternode name="10.1.1.1" nodeid="1">
            <fence>
               <method name="single">
                  <device action="off" ipaddr="192.168.1.1"
                  name="fence_pub" port="hw-server-1"/>
               </method>
            </fence>
         </clusternode>
```

Red Hat Cluster Suit

```
<fencedevices>
    <fencedevice agent="fence_virsh" delay="5"
     identity_file="/root/.ssh/id_rsa" login="root"
     name="fence_pub"/>
</fencedevices>
<rm>
    <failoverdomains/>
    <resources/>
</rm>
```

Red Hat Cluster Suit

Na co si dát pozor

- **maximální počet nodů!**
  - omezeno na qdiskd na 16
- **pozor na čas**
  - TSC vs. HPET
- **výkon**
- **správný fencing**

GridFS - MongoDB

- **FS v DB**
- **alternativa k HDFS a jiným DBFS**
- **moduly pro**
  - nginx
  - httpd (apache)
- **připojení přes FUSE**
  - neumí adresáře
- **HA**
  - replika set
    - max 12 nodes 7 voting
- **horizontální škálování**
  - sharding

GridFS - MongoDB

```
[root@mongo ~]# mongofiles -d file put
  /root/passwd
[root@mongo ~]# mongo file
> show collections
fs.chunks
fs.files
system.indexes
>
```

GridFS - MongoFS

- **fs.chunks**
  - obsahuje data souborů
  - default velikost 256kb
    - změna přes driver v aplikaci při PUT

```
{
   “_id”: <ObjectID>,
   “files_id”: <string>,
   “n”: <num>,
   “data”: <binary>
}
```

- **fs.files**
  - “metadata” souboru

```
{
   “id”: <ObjectID>,
   “length”: <num>,
   “chunkSize”: <num>,
   “uploadDate”: <timestamp>,
   “md5”: <hash>
}
```


COME TO THE
DARK SIDE
WE HAVE COOKIES

Q&A

petr.medonos@etnetera.cz

lukas.herbolt@etnetera.cz