MetaCentrum

Souborové systémy v cloudu

Filip Hubík @ MetaCloud (hubik@ics.muni.cz)
Adam Tomek @ MetaCloud

Masarykova univerzita a CESNET

7.10.2014

CESNET

Obsah a cíle prezentace

Obsah

- Představení MetaCloud skupiny
- Definice problematiky
  - Cloud a úložiště
  - Náhled na souborové systémy
- Ceph podrobně
- Gluster podrobně
- Benchmarking
  - Testovací infrastruktura
  - Výsledky testů
- Závěr a otázky

Cíle přednášky

- Vytvořit základní představu o vhodnosti použití vybraných distr. soub. systémů v cloudovém prostředí
- Porovnat měřením jejich vlastnosti na testovací infrastruktuře a dopomoci tak k případnímu rozhodování o jejich nasazení

Kdo jsme

MetaCloud team

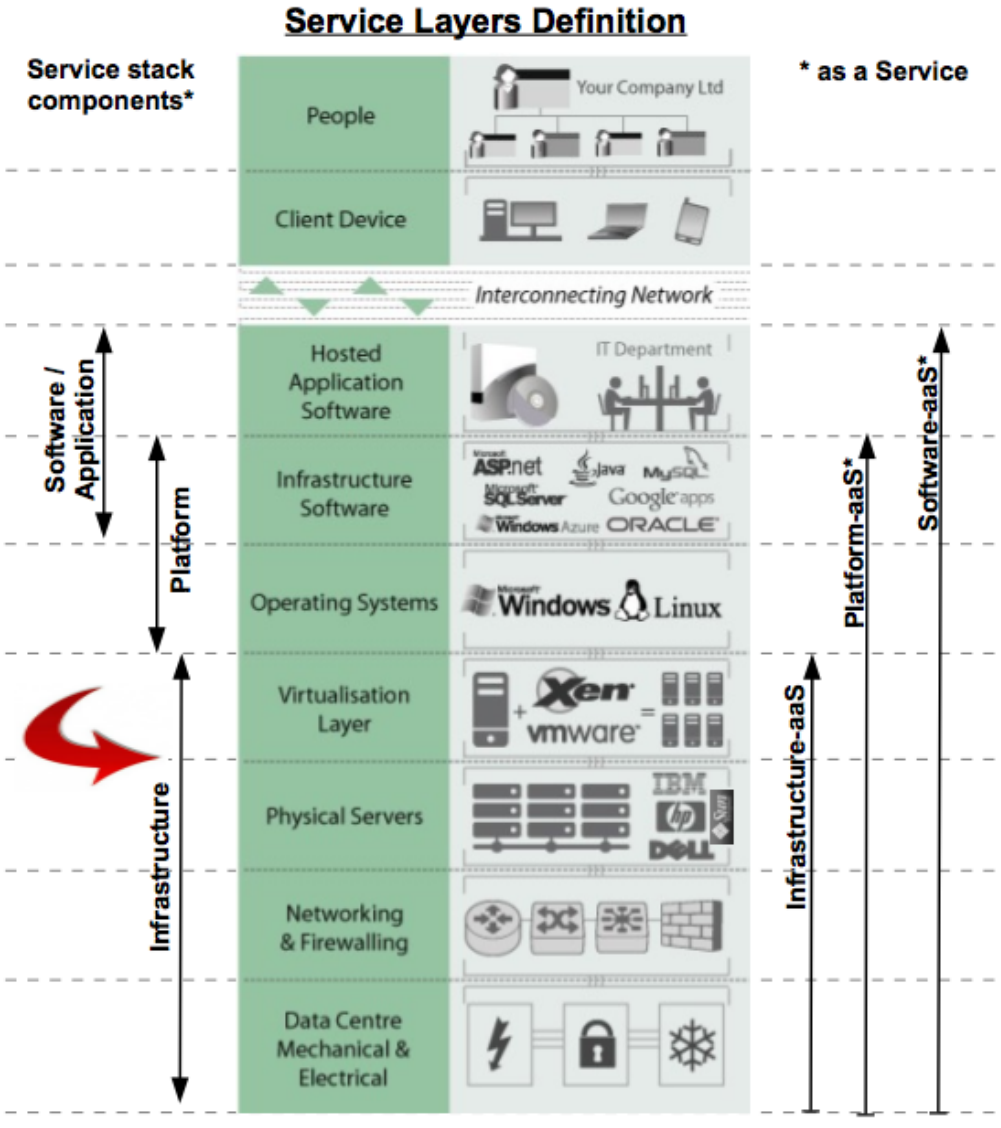
- Cloudové služby pro akademické sféry
- Spolupráce dvou iniciativ
  - CESNET – MetaCentrum – Xen
  - Masarykova univerzita – CERIT-SC – KVM
- Hybridní cloud
- Malá až střední velikost
- OpenNebula middleware
- Více jak 3 roky zkušeností
- Doprovodné aktivity (HA, support, . . .)
- Výzkumné projekty
- Kontakt: cloud@metacentrum.cz

Úvodem - architektura cloudu

Typy cloudů:

- Service (SaaS)
- Platform (PaaS)
- **Infrastructure (IaaS)**

* as a service [3]



Úvodem - skladba IaaS cloudu

Jádrem cloudu na IaaS úrovni je

- Middleware (uživ. rozhraní, scheduler, API, managment, . . .)
  - OpenNebula, Openstack, CloudStack, Nimbus
- Virtualizace HW (Xen, KVM, VMWare, . . .)
  - Množství uzlů
  - CPU
  - Pamět
  - Síťové prostředky (OpenFlow + Open vSwitch)
- **Úložiště obrazů a dat („Image storage“)**
  - Základní rozhodnutí ovlivňující dlouhodobý provoz cloudu
  - Různé formáty obrazů (QCOW, RDB, RAW, VDI, . . .)
  - Velikost obrazu alokována předem
  - **Problém - distribuce obrazů cloudem**
    - Kopírování (rsync, scp, ftp, . . .)
    - **Distribuovaný souborový systém**
    - Jiné přístupy (SAN/NAS magie, replikace, vzdálená úložiště & marketplace, . . .)

Typy souborových systémů

- Lokální – klasické FS, žádné sdílení dat mezi uzly
- Síťové – lokální + síťový protokol, klient/server 1:N model (NFS, FTP)
  - Distribuované – data jsou transparentně distribuována mezi více uzlů úložiště, M:N model (DFS, AFS)
    - **Paralelní – konkurentní I/O operace nad soubory, „striping“ (GPFS, Ceph, Gluster(!), MooseFS, HDFS, Lustre, . . .)**
- Symetrické – metadata umístěna na klientské i serverové části
- Asymetrické – metadata umístěna na klientské nebo serverové části
- Sdílené – sdílený přístup všech uzlů, zámky
  - Klastrované – symetrické, přímý přístup k blokovému zařízení, drahá kapacita, SAN (GFS, OCFS, . . .)
- „User-space“ – běží mimo kernel v uživatelském prostoru
- Objektové – data objekty + metadata + globalID, ne struktura, ploché, méně metadat
- Globální – každý uzel stejný pohled

Požadavky na FS v cloudu

- Každý host dokáže spustit každou VM
- Distribuovanost (levná a heterogenní infrastruktura)
- Lineární škálovatelnost kapacity a propustnosti
- Eliminace SPOF (metadata server)
- Konkurentní přístup k datům
- Vysoká dostupnost, tolerance chyb
- Ideálně paralelní I/O operace
- Správa přístupu (různé části úložiště různá ACL)
- Open-source
- Live migrace
- Snapshoty bez podpory obrazu

Požadavky na škálovatelnost

Kapacita

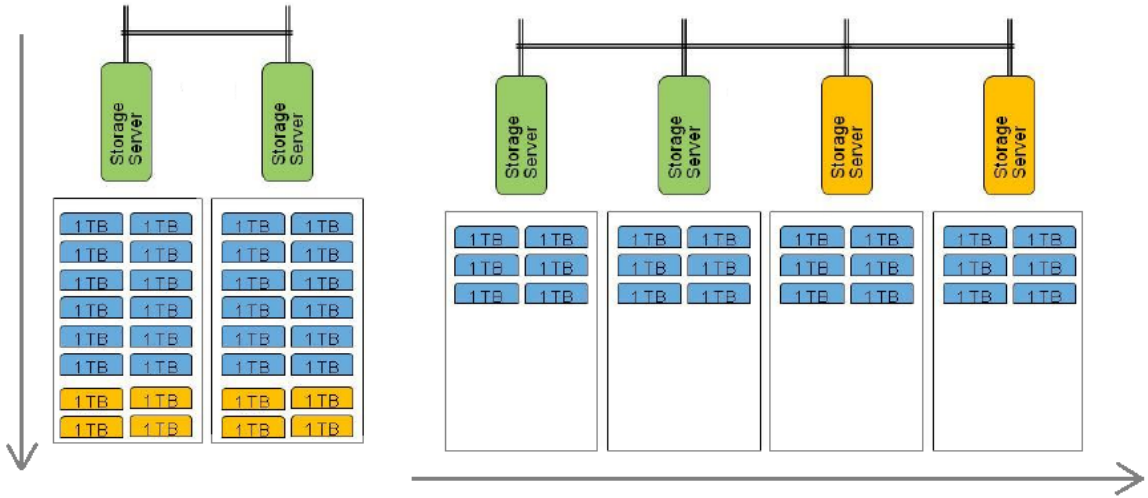
- Vertikální - počet disků na stroj (omezeno řadičem)

Propustnost

- Horizontální - počet strojů formujících úložiště

Problémy u distribuovaných FS

- Linearita - kapacita úložiště musí být rovna součtu kapacit jedn. strojů
- Síťová infrastuktura musí mít dostatečnou propustnost



Vhodné souborové systémy

Cíleny pro použití v cloudu

- **Ceph**
- **GlusterFS**

Další možné varianty

- GPFS - proprietární, problémy s Xen
- HDFS - metadata server SPOF, Copy-On-Write model
- MooseFS, XtreemFS - poměrně mladý, vyhodnocování
- Lustre - supercomputing
- ...

Souboj titánů

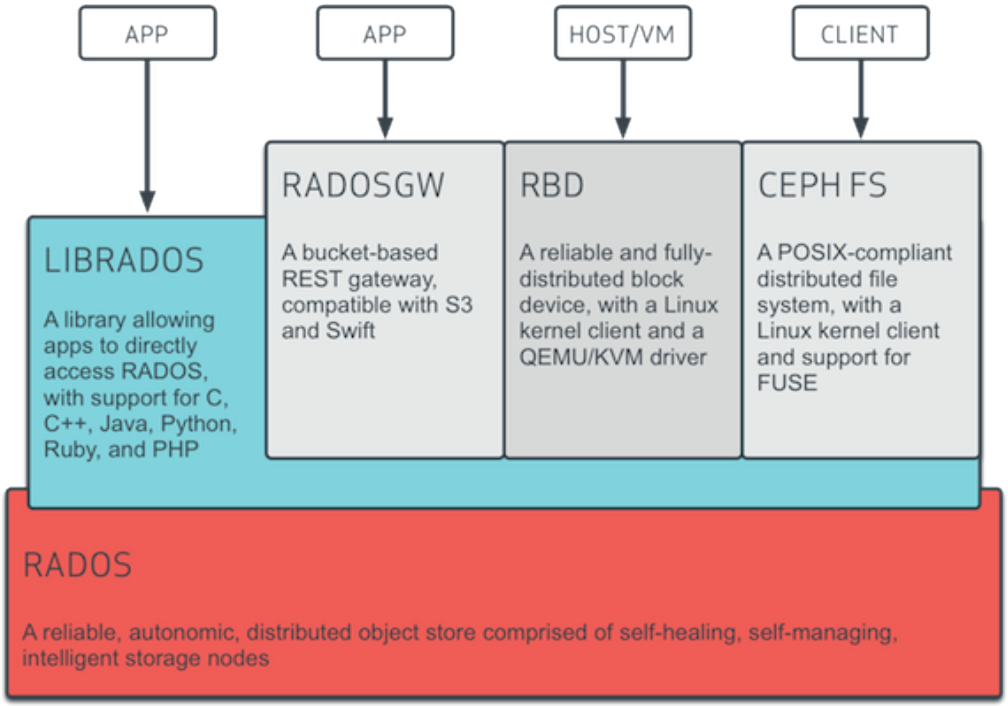
Ceph v produkci (Inktank, později Red Hat)

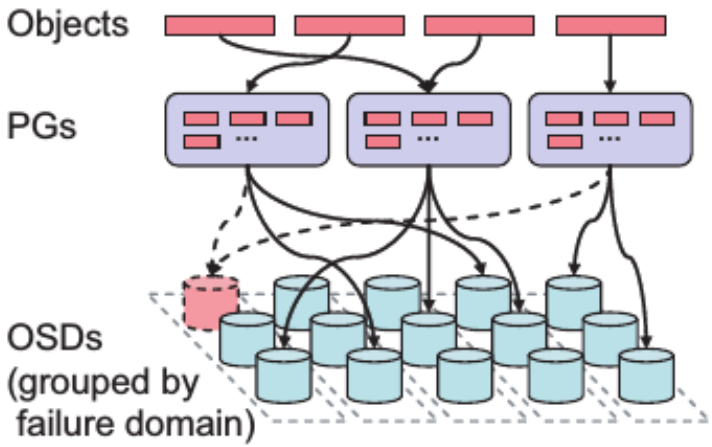
GlusterFS v produkci (Gluster, poté Red Hat)

Ceph

- Částečně POSIX distribuovaný soub. systém
- Nativně **objektový**, blokový (RBD), souborový (CephFS)
- Metadata oddělena i neoddělena
- CephFS není v produkční fázi (08-2014)
- „Rados block device“ – vhodné pro cloud bez FUSE
- Rozsáhlé možnosti konfigurace (v testech výchozí hodnoty)
- Komplexní ACL
- CRUSH algoritmus
  - Dynamický, load balancing, prokládání nativně, replikace
- Stavební kameny
  - OSD – CPU, paměť, síťové rozhraní, diskový prostor
  - Monitor – Informuje OSD o změnách topologie („cluster map“)
  - MDS – Metadata server (CephFS)
  - Klient – Jádro >= 3.9
- Není infiniband pouze TCP

Ceph – architektura

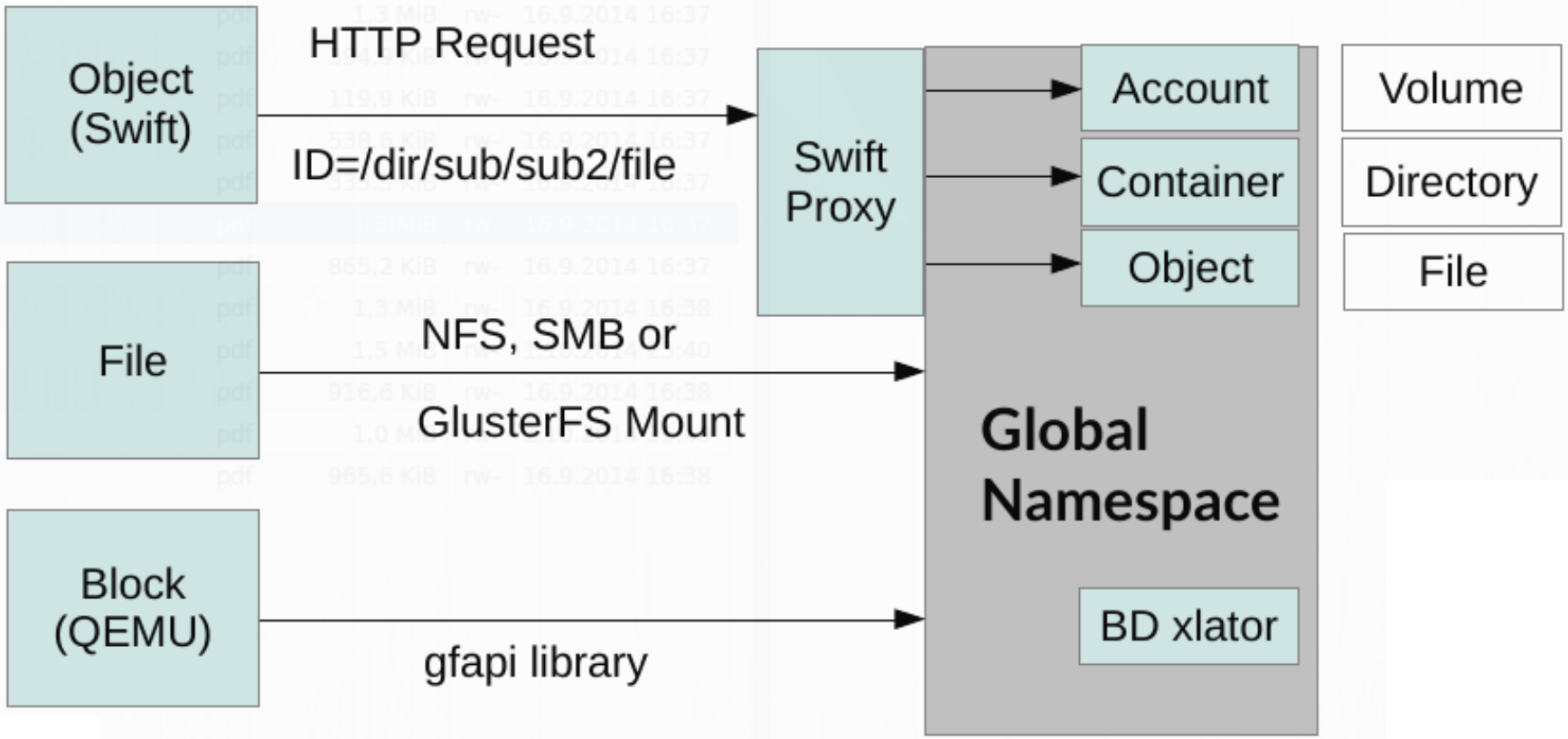




GlusterFS

- Distribuovaný POSIX FS primárně na souborové úrovni
- Není centralizovaný metadata server (elastický hash)
- Teoretická velikost až $72 * 10^6$ zettabyte (XFS subvol.)
- Integrace s QEMU (blokově bez FUSE, libgfapi - *drive file=gluster://server[:port]/volname/image[?transport=...]*)
- HA a samoopravné vlastnosti
- Souborový systém v otevřené formě, user-space model
- Georeplikace (asynchronní replikace)
- Pasivní load balancing, prokládání, replikace (synchr.)
- Dynamická práce se svazky
- Globální namespace
- RDMA, TCP
- Nejsou snapshoty! (potřeba použít QCOW)
- Jednoduchá konfigurace

Gluster – architektura a API



Gluster – terminologie a použití

Terminologie

- Brick – zákl. stavební kámen, úl. prostor zapojen do volume
- Volume (svazek) – log. org. souborů, globální namespace
- Klient – uzel připojující volume (i server)
- Server – uzel, na kterém leží brick
- Translator – kód prováděný nad daty (modulárnost)
  - Storage, Debug, Encryption, Scheduler, Protocol, ...
- Subvolume – brick, který už byl zpracován translatorem

Módy organizace dat ve svazku

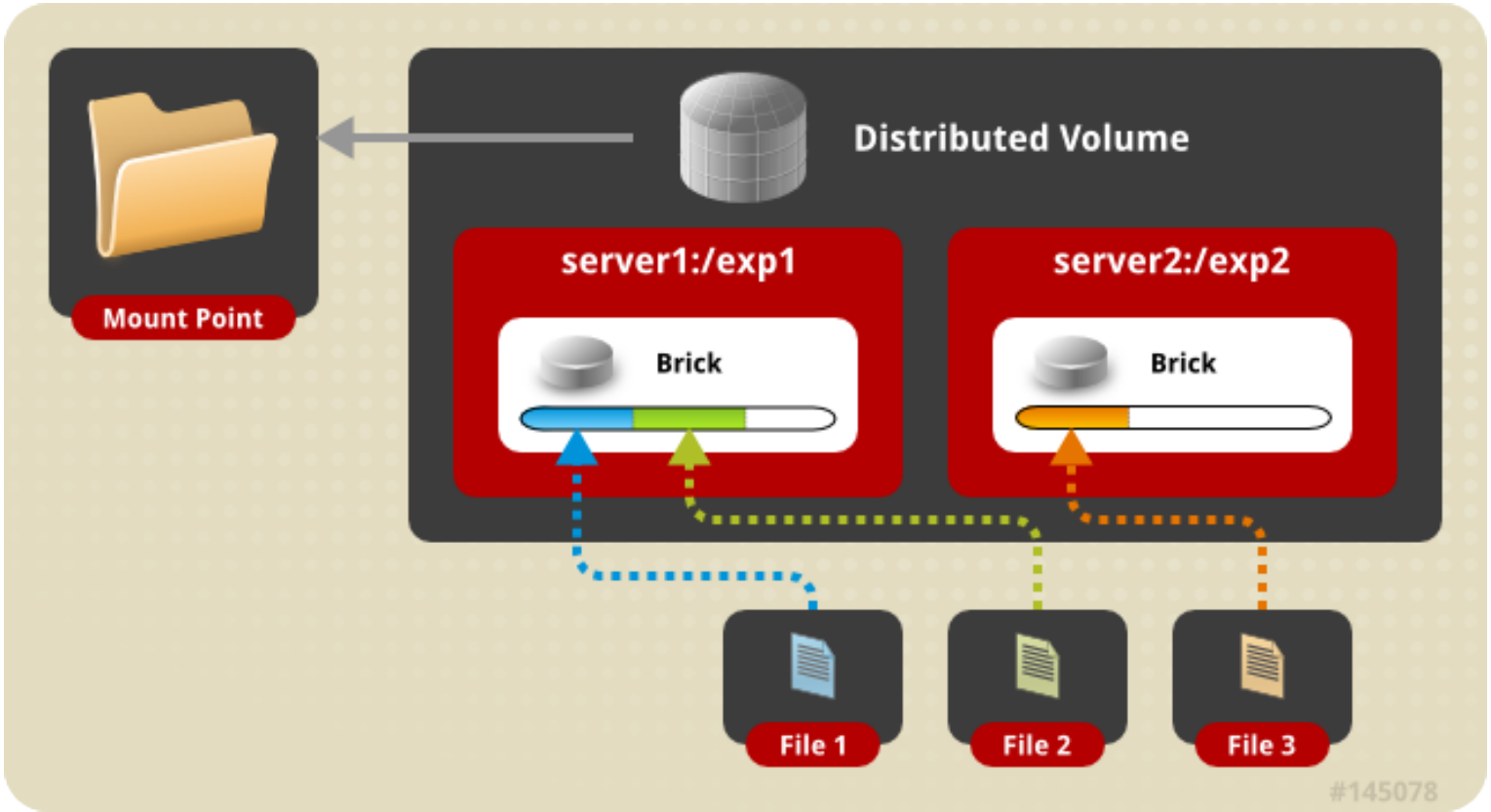
- Distribuce
- Replikace
- Prokládání

Příklad tvorby svazku

- *$ gluster volume create vol0 replica 2 stripe 2 server1:/data/gv0 server2:/data/gv0 server3:/data/gv0 server4:/data/gv0*

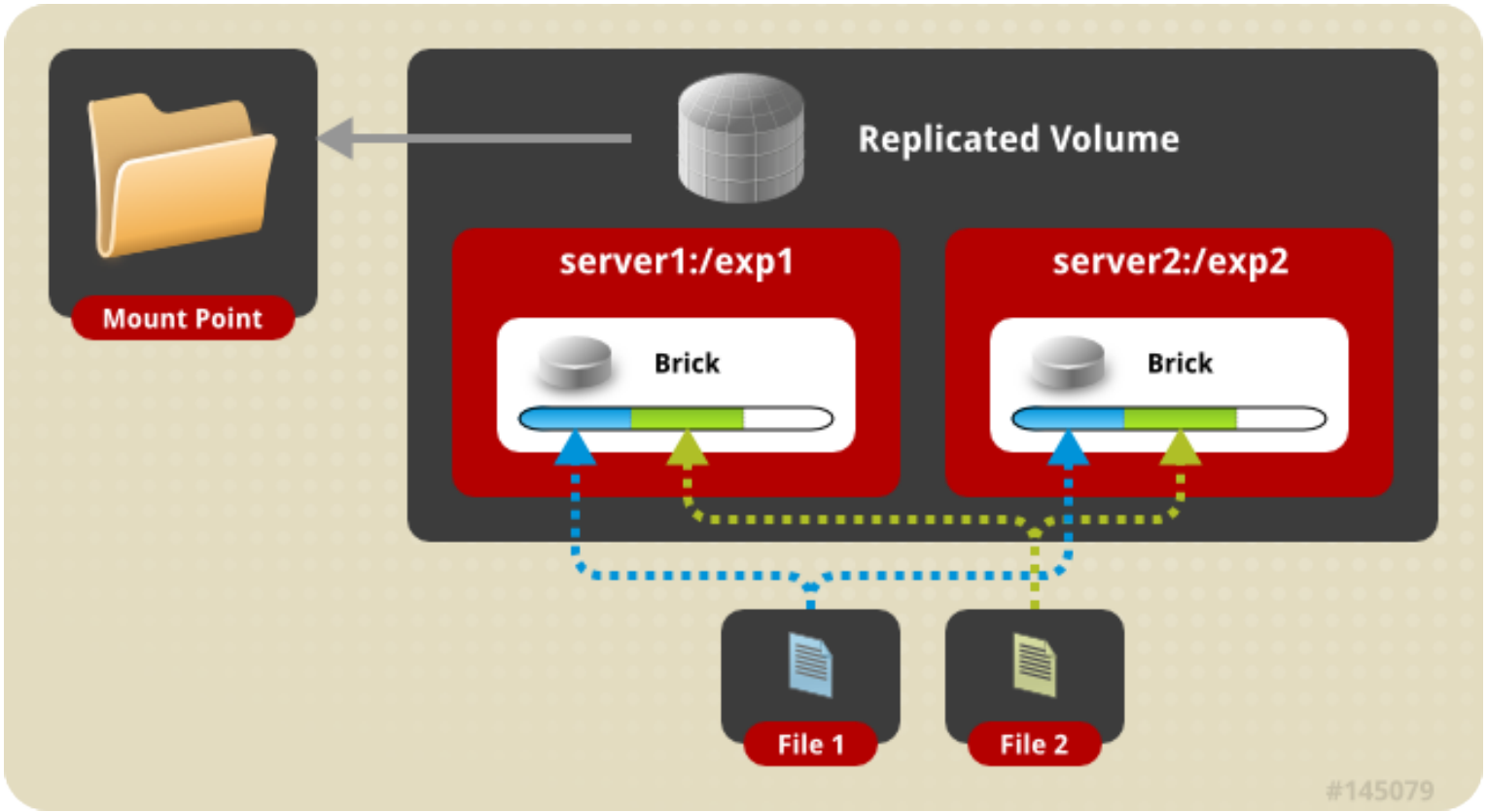
Distribuovaný mód

- Data distribuována na úrovni souborů
- Žádná redundance
- Obdoba RAID0 či JBOD
- Selhání uzlu znamená ztrátu dat



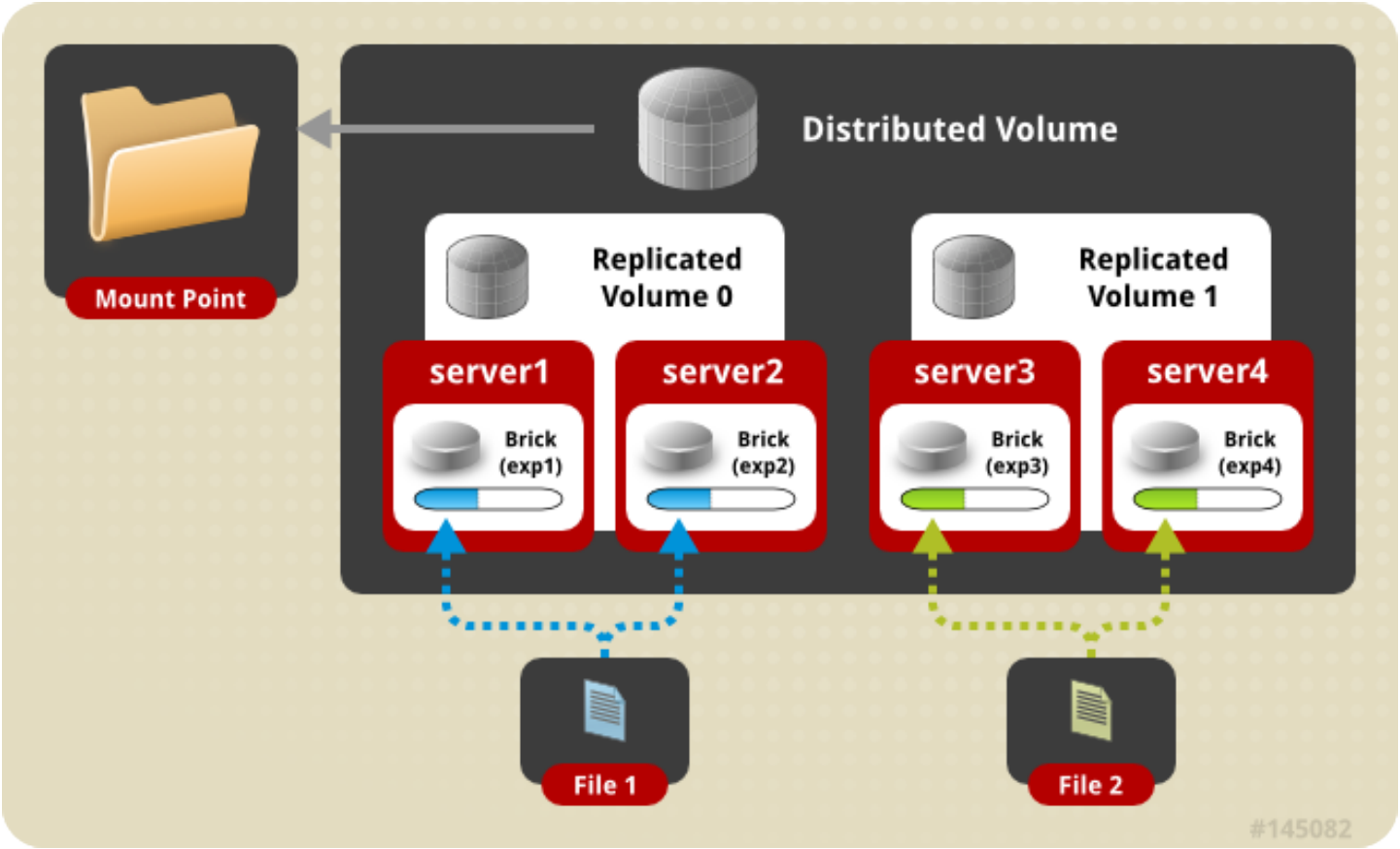
Replikace

- Redundance
- Selhání disku
  - Self-healing
  - Manuální intervence („Split-brain“ scénář)



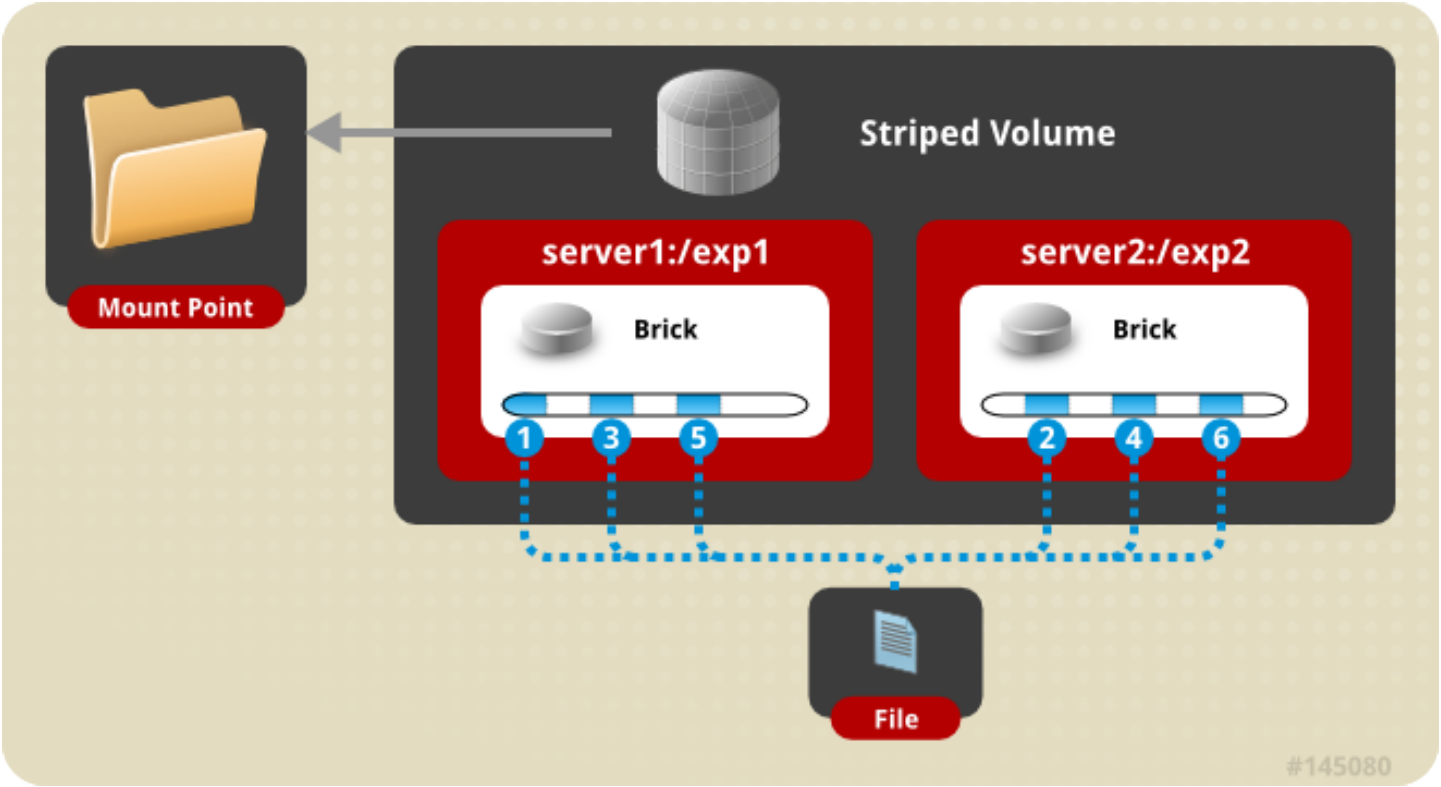
Distribuovaná replikace

- Koefient replikace r=2
- Počet bricků musí být nasobek r
- Redundance + propustnost při čtení



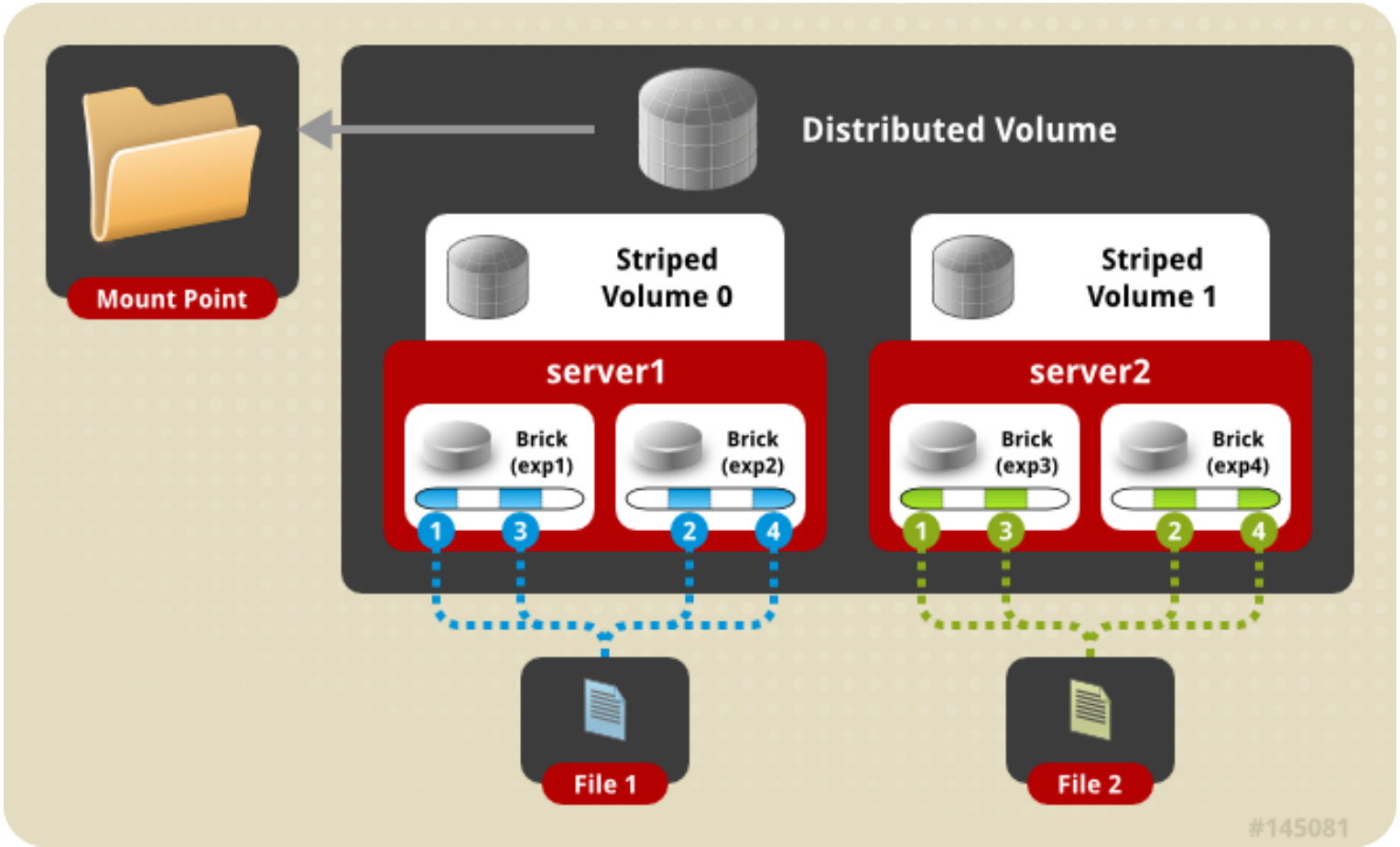
Prokládání („striping“) souborů

- Dělení na úrovni souborů (RAID0)
- Velmi velké soubory
- Vhodné pro vysoce konkurentní prostředí (DB, HPC)
- Počet bricků roven koeficientu prokládání



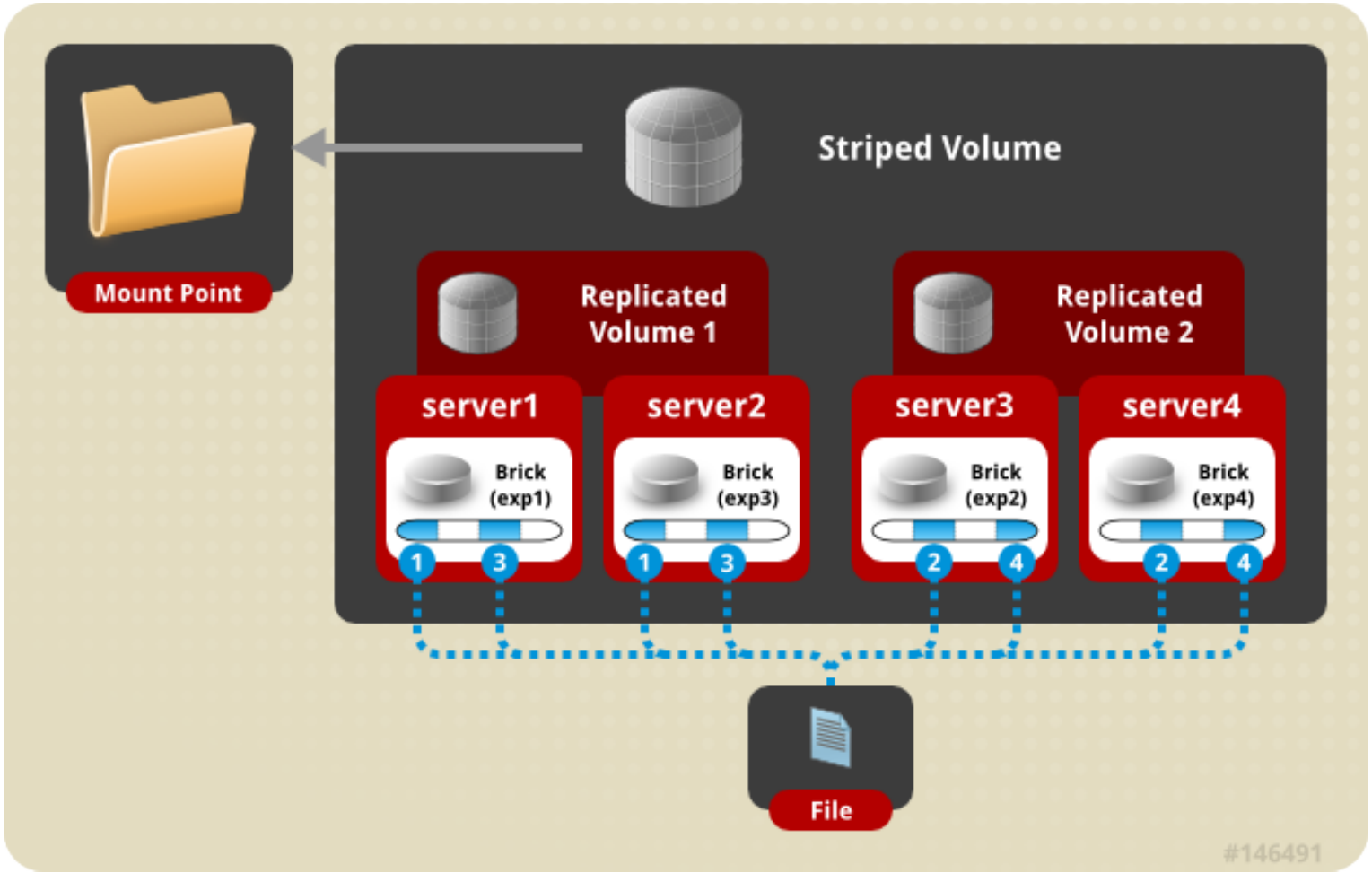
Distribuované prokládání

- Předchozí výhody + škálovatelnost
- Počet bricků násobkem koeficientu prokládání



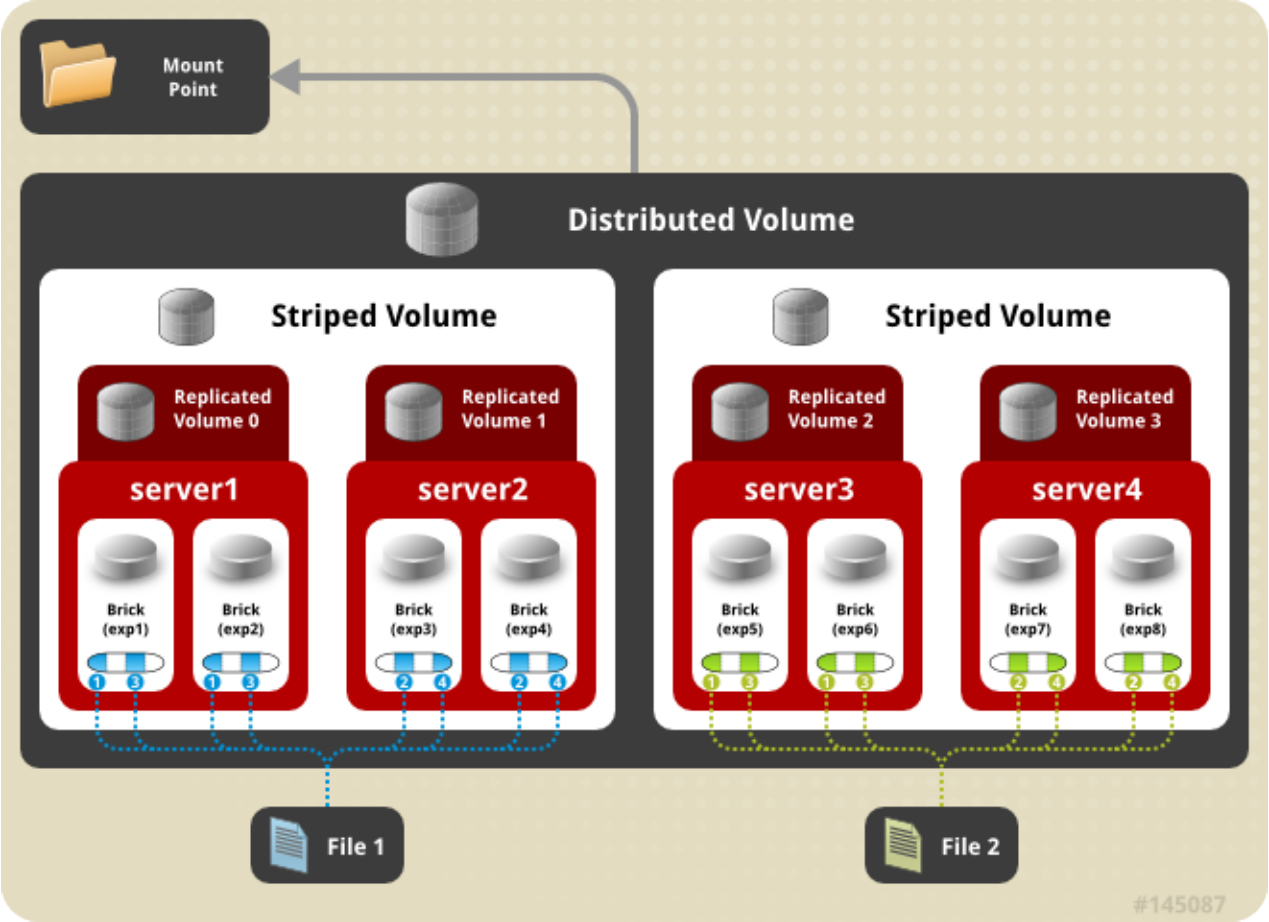
Prokládaná replikace

- Obdoba RAID1+0
- Velké soubory ve vysoce par. prostředí (MapReduce)



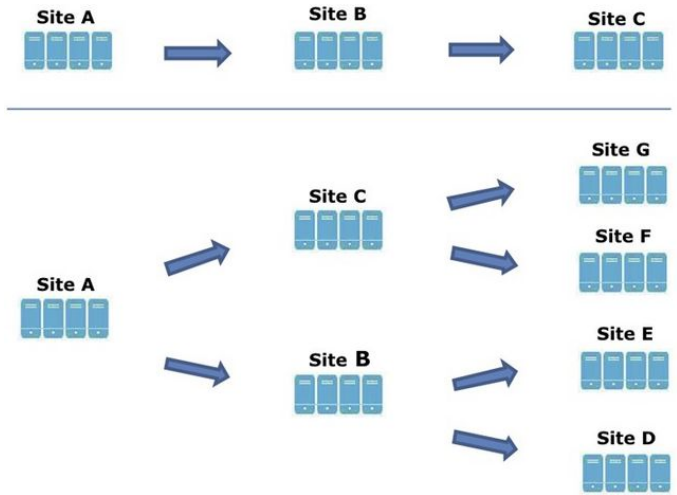
Distr. prokládaná replikace

- Výhody předchozího + škálovatelnost
- MapReduce
- Počet bricků násobkem násobků obou koeficientů



Gluster – georeplikace

- Asynchronní replikace (checkpointing)
- Inkrementální – rsync
- Master/slave model
- LAN, WAN
- Zálohování dat
- Čas musí být synchronizován na všech uzlech (NTP)
- V cloudu použití s QCOW real-time snapshoty
- Kaskádování (cíl = zdroj)



Testovaná architektura

Hardware

- 10 klientských uzlů, 4 serverové uzly stejného typu
- Intel Xeon CPU E5472 3.00GHz, 2GB RAM
- Disk
  - 250GB na uzel
  - hdparm – 192MB/s čtení
  - iozone – cca 190MB/s čtení i zápis
  - dd – 40MB/s zápis
- Odstranění virt. vrstvy (bez Xen či KVM)
- Gluster nutno připojovat přes FUSE
- Debian 7 Wheezy (jádra >= 3.9, 3.14-0.bpo.2-amd64)
- XFS (produkční u obou FS)
- IPoIB (Ceph neumí RDMA!) – 4X DDR (20 Gb/s)

Testovací konfigurace

Iozone

- V praxi 1-10 souběžných procesů na uzel (menší cloud)
  - **Souběžné vytížení úložiště ze strany klientů**
- Velikost obrazů 1-4GB na uzel
- Verze 3.397
- Distribuovaný režim (parametr -+m)
- Velikost záznamu (record size) 256kB

Ceph

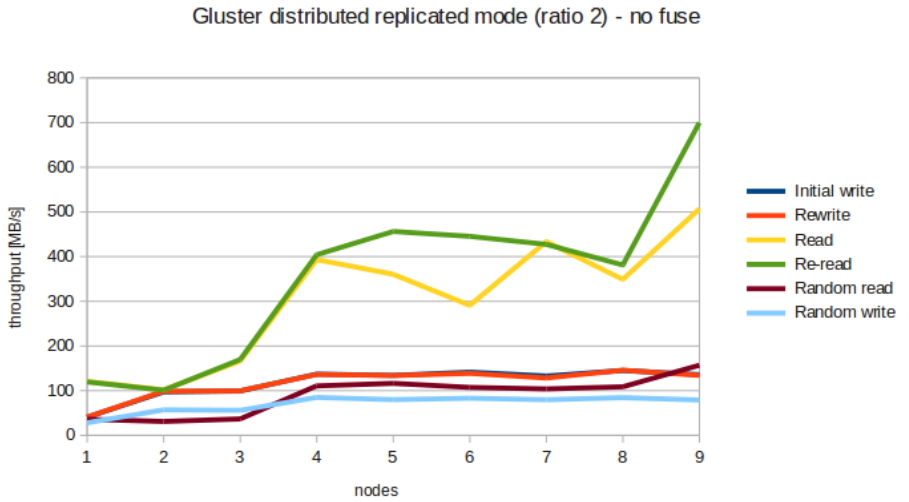
- Verze 0.80.4 (06-2014)
- Koeficient replikace 2, min repl. 1, prokládání výchozí
- pg_num & pgp_num = 256
- Obrazy připojeny přímo přes RBD blokové zařízení

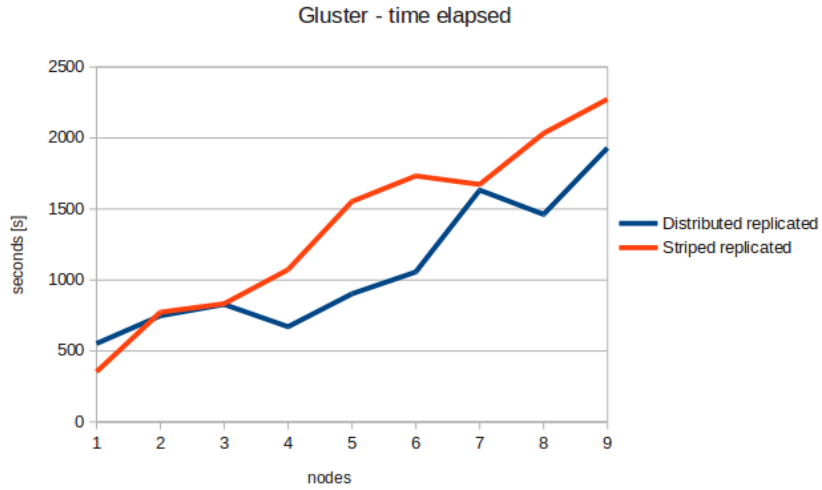
GlusterFS

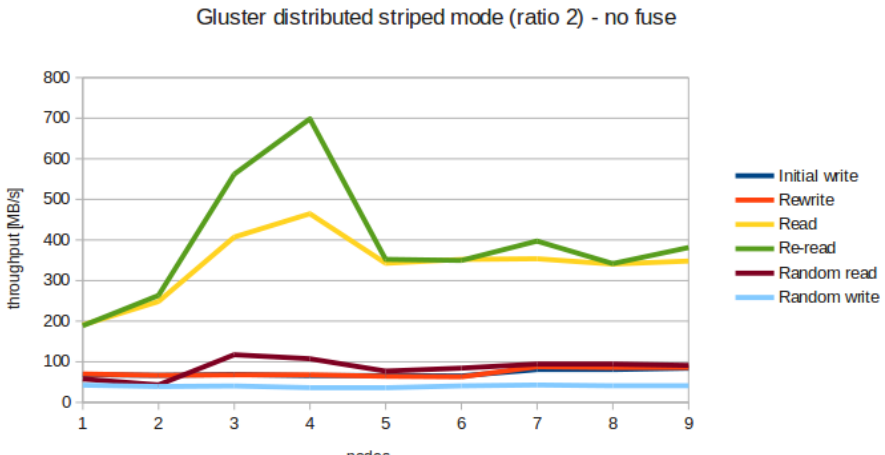
- Verze 3.5.2 (09-2014)
- Koeficient replikace 2, prokládání 2
- Obrazy připojeny přes FUSE (pokud není uvedeno jinak)

Gluster 1-9 uzlů, 1 proces/uzel

- DR mód ovlivněn replikací, DS pomalejší nelineární

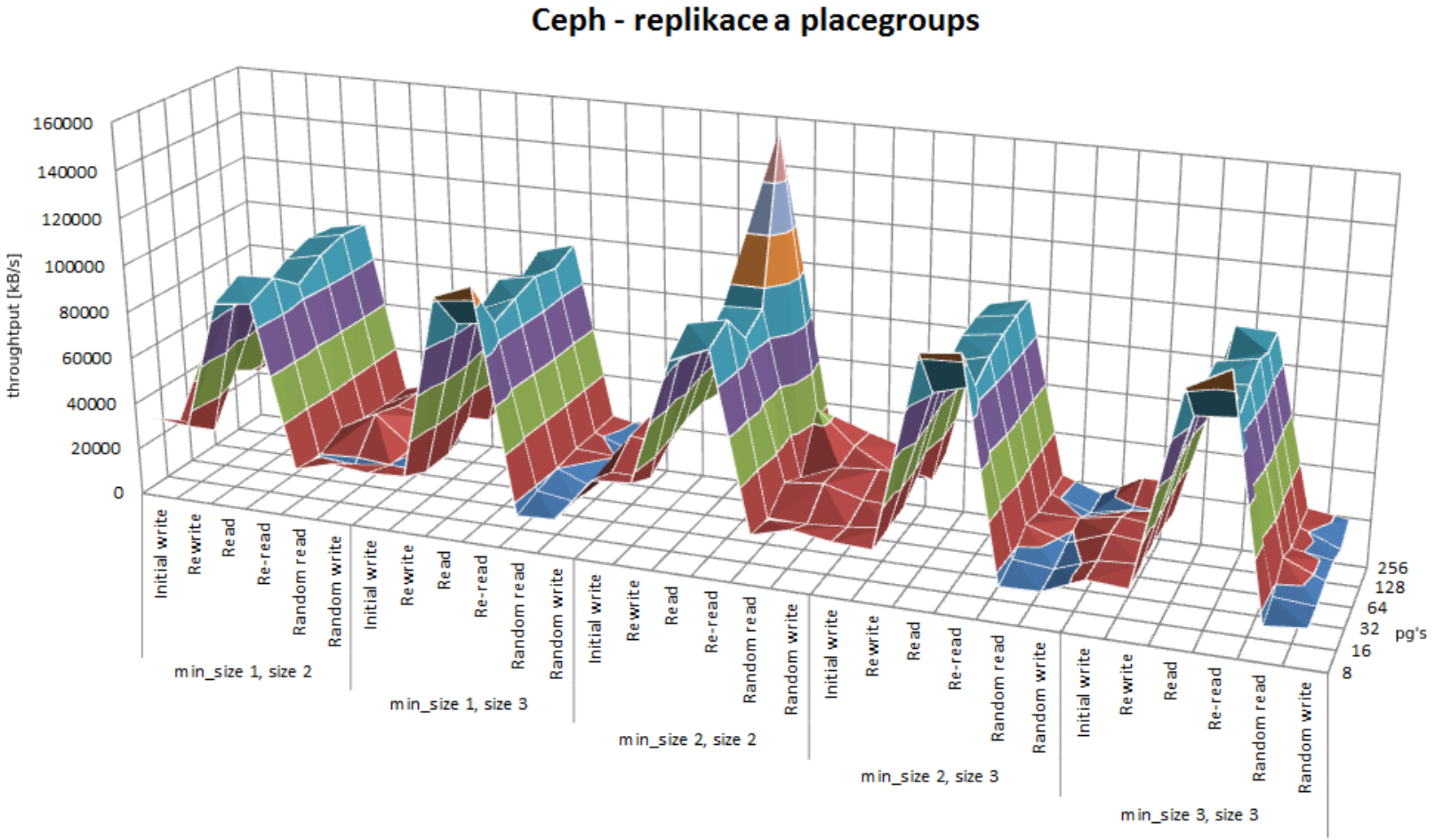






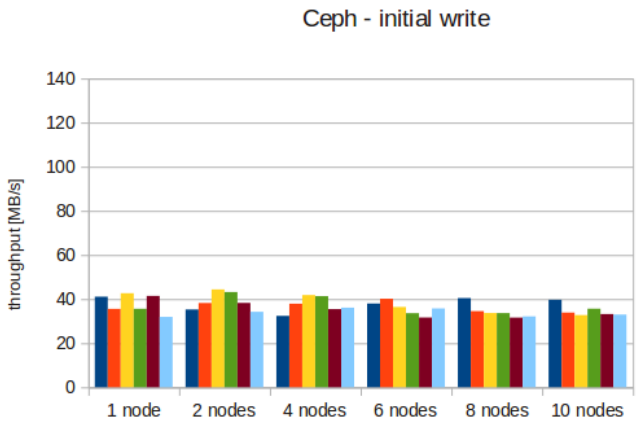
Meze propustnosti infrastruktury v MB/s	
Initial write	1496,30
Rewrite	1470,03
Read	1636,41
Re-read	1632,88
Random read	347,73
Random write	806,98

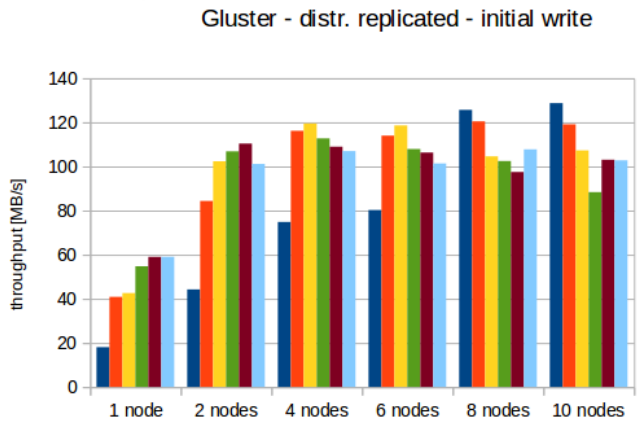
Ceph 1-10 uzlů, 4GB image

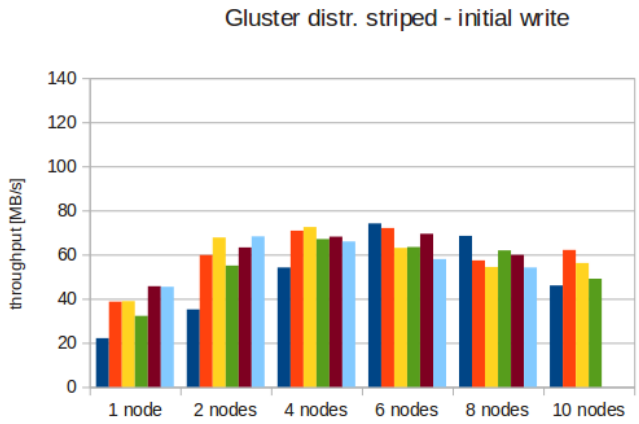


Initial write, 1-10 procesů/uzel

- Jak se v tom vyznat?
- Ceph - konst., Gluster DR 1 pr. lineární, další log.

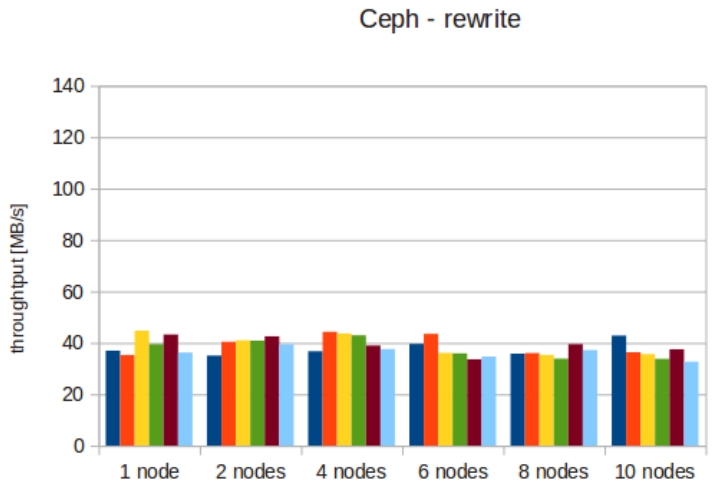


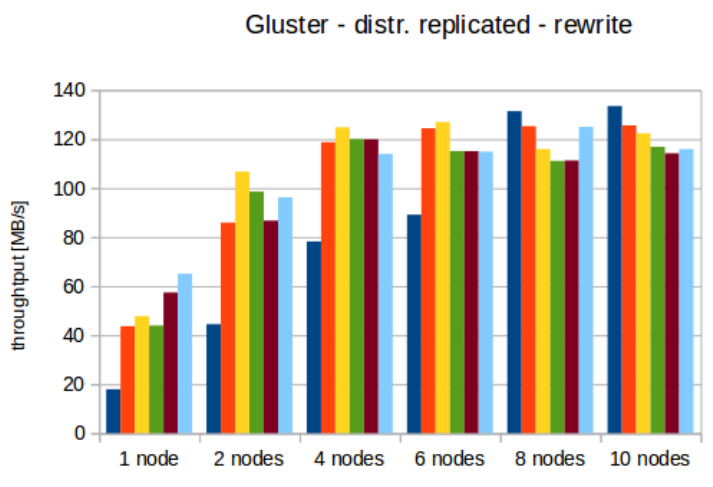


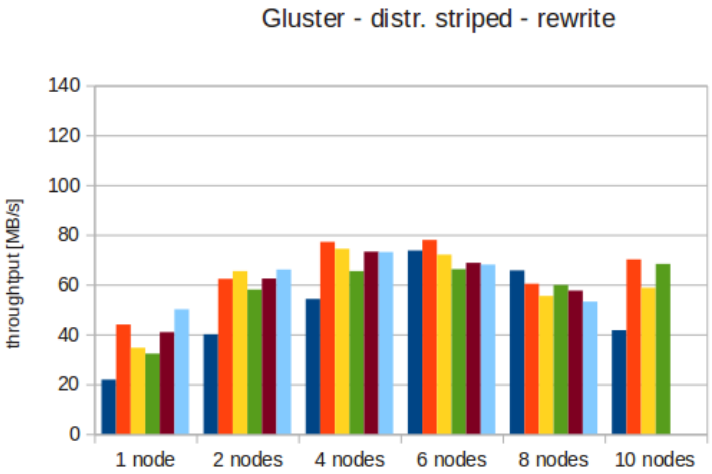


Rewrite, 1-10 procesů/uzel

- Obdobný stav

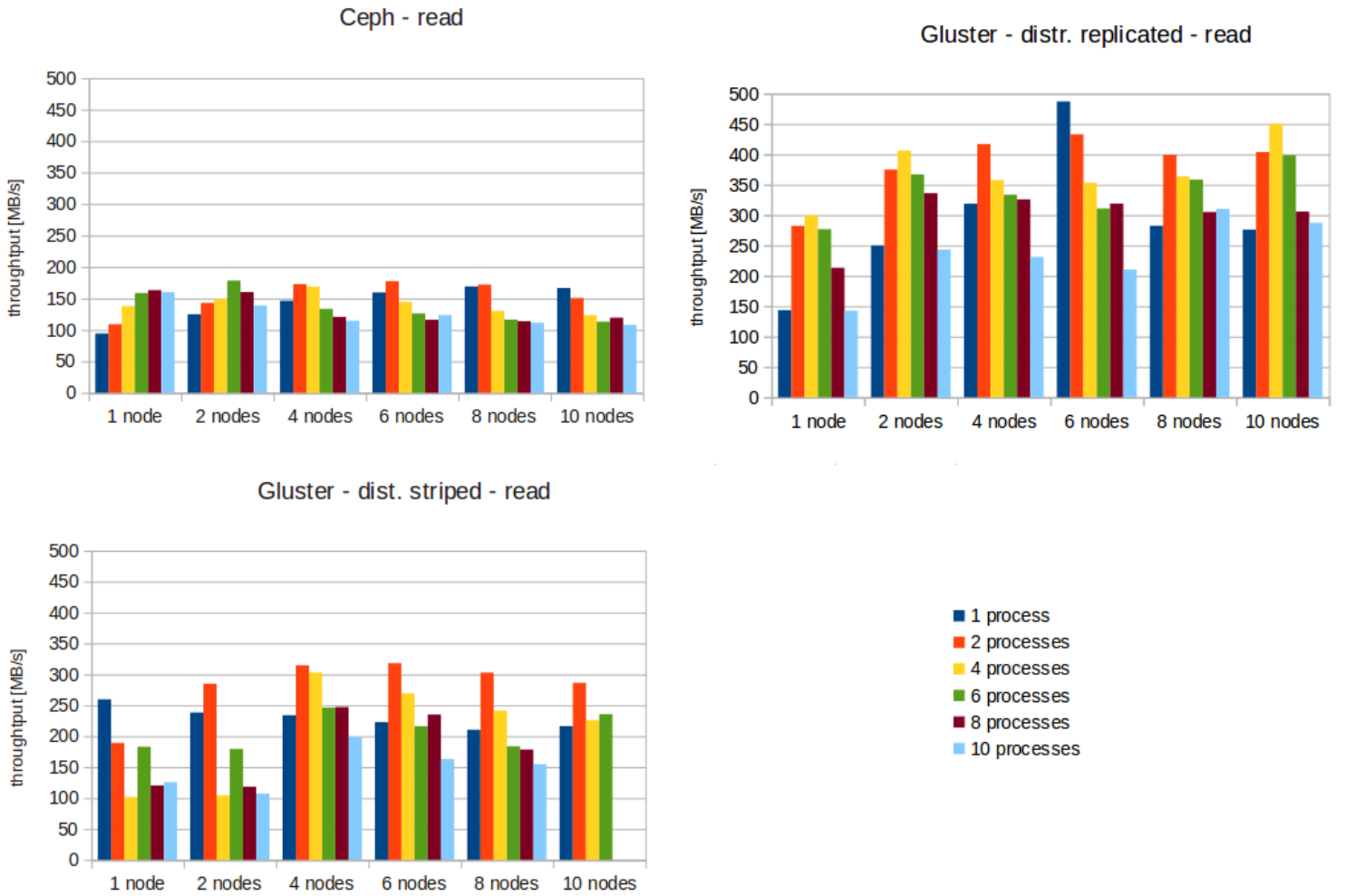




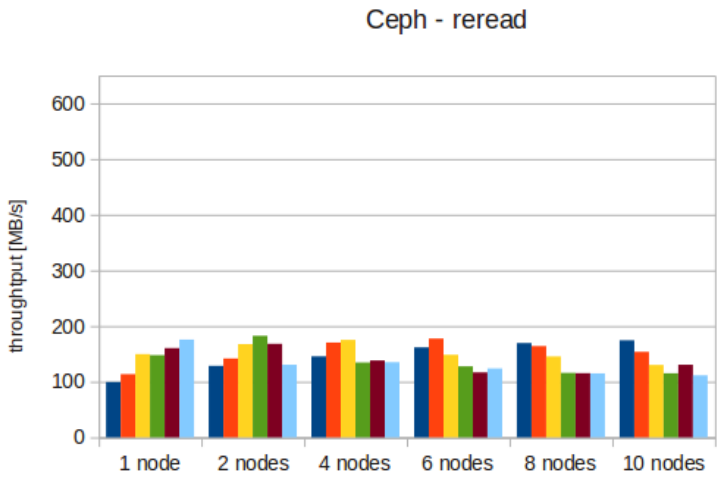


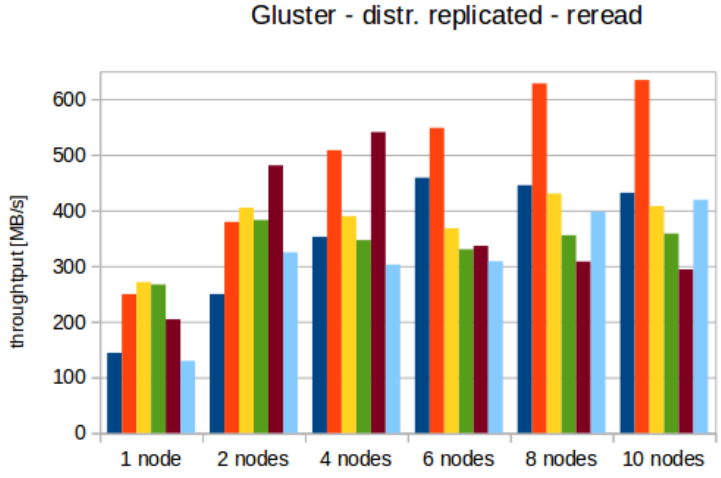
Read, 1-10 procesů/uzel

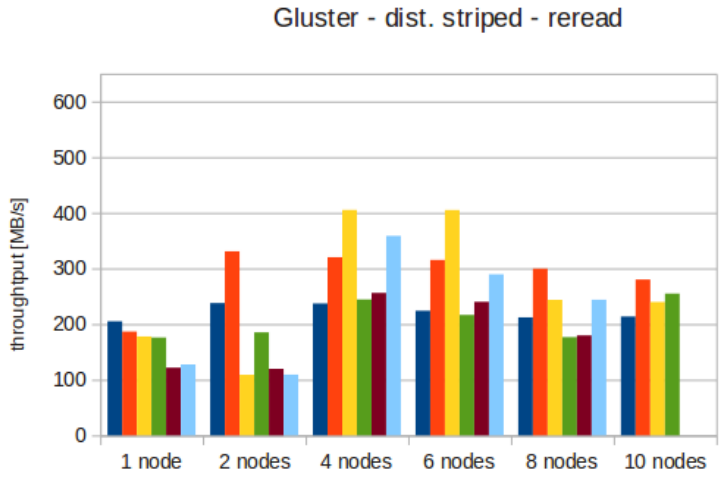
- 1 pr. DR 1-6 lineární, 7-10 klesá, Ceph pro 1 uzel taktéž



Re-read, 1-10 procesů/uzel

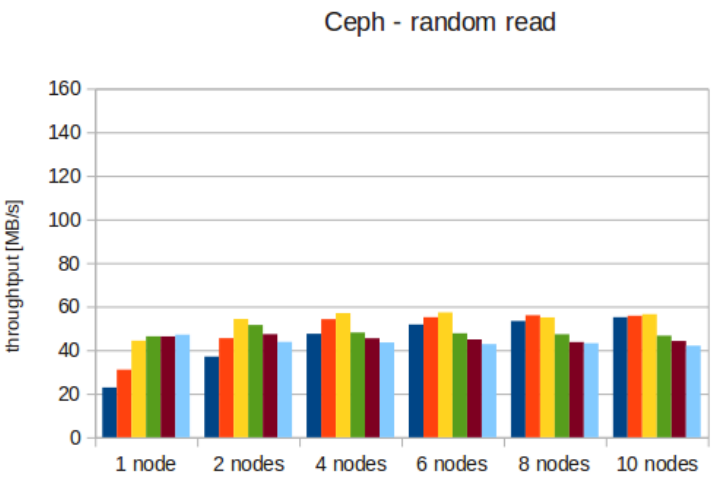


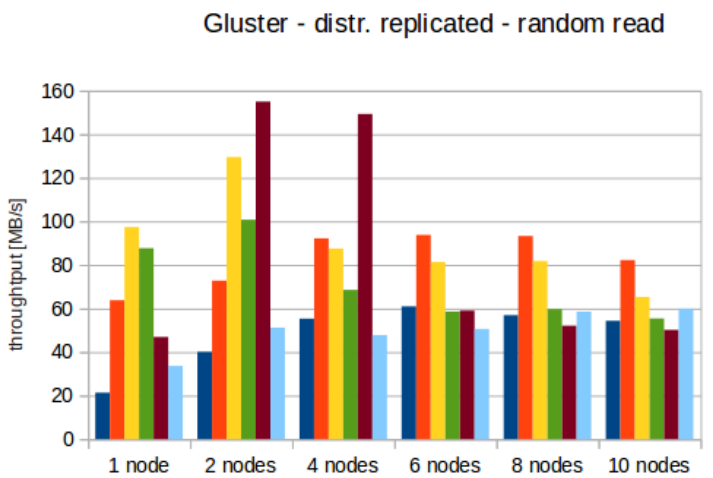


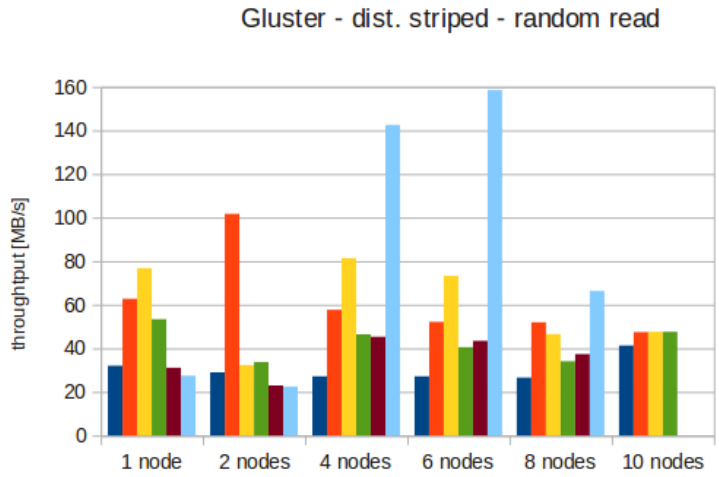


Random read, 1-10 proc./uzel

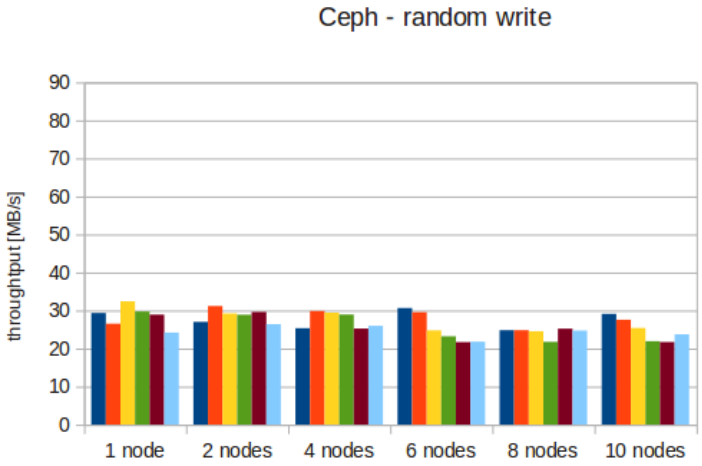
- Ceph někde rychlejší než DS

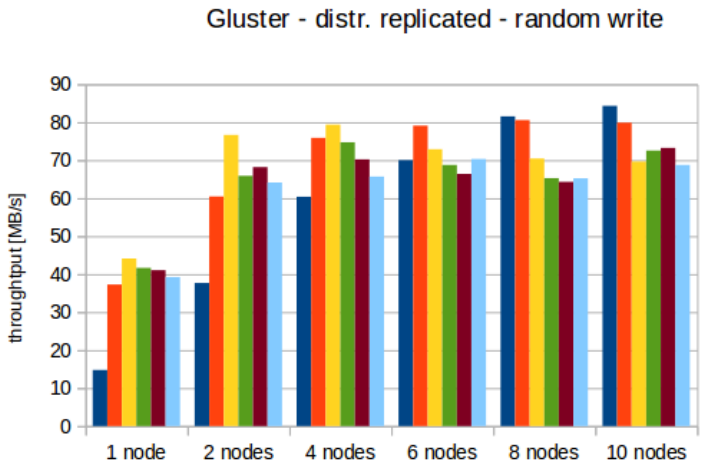


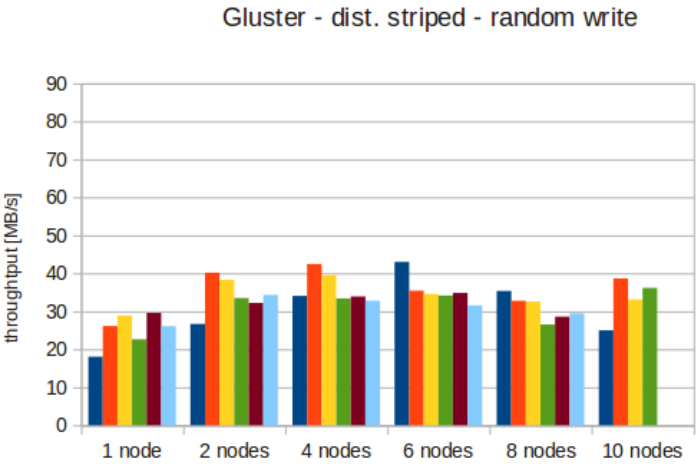




Random write, 1-10 proc./uzel







Výsledky měření - závěr

Ceph

- Kam se poděla škálovatelnost? Konstantní?
  - Nejspíš silně omezena RBD vrstvou na uzlu
- Pravděpodobně CRUSH - větší infrastruktury
- Gluster ukázal že to jde
- Nativní striping také hraje proti

Gluster

- Také omezen, ale snáší to lépe
- Škáluje lineárně či logaritmicky
- Ve stripingu propustnost klesá pro >=6 uzlů
- Ale pozor na čtení pro >= 8 uzlů
- Gluster padal ve stripingu pro 8 a 10 uzlů

Doporučení pro cloud

- Gluster v distribuovaném replikovaném módu

* Výsledky pouze simulují běh v cloudu

Další aktivity

- Bakalářská práce, další výzkum
- Testování Xen vs. KVM
  - Srovnání z pohledu vhodnosti pro cloudové využití
- Testování kompatibility Ceph a GlusterFS s virtualizační vrstvou (problém s GPFS)
- Hrátky s parametry Cephu i Glusteru
- Revalidace podivné škálovatelnosti Cephu

Konec

– Díky za pozornost –

Literatura

[1] Ceph offfical site – http://ceph.com
[2] Gluster offical site – http://www.gluster.org
[3] http://www.katescomment.com/iaas-paas-saas-definition
[4] Cloud Storage for the Modern Data Center, Copyright 2011, Gluster, Inc.
[5] http://www.inktank.com/partners/ceph-at-the-red-hat-summit/
[6] http://yourstory.com/2011/09/yourstory-exclusive-california-based-indian-entrepreneurs-powering-petabytes-of-cloud-storage-the-gluster-story/ [7] GlusterFS Cloud Storage, John Mark Walker
[8] Red Hat documentation – https://access.redhat.com/documentation