# Raspodijeljeni podatkovni sustav BeeGFS

Emir Imamagić
Sveučilišni Računski Centar (Srce)

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

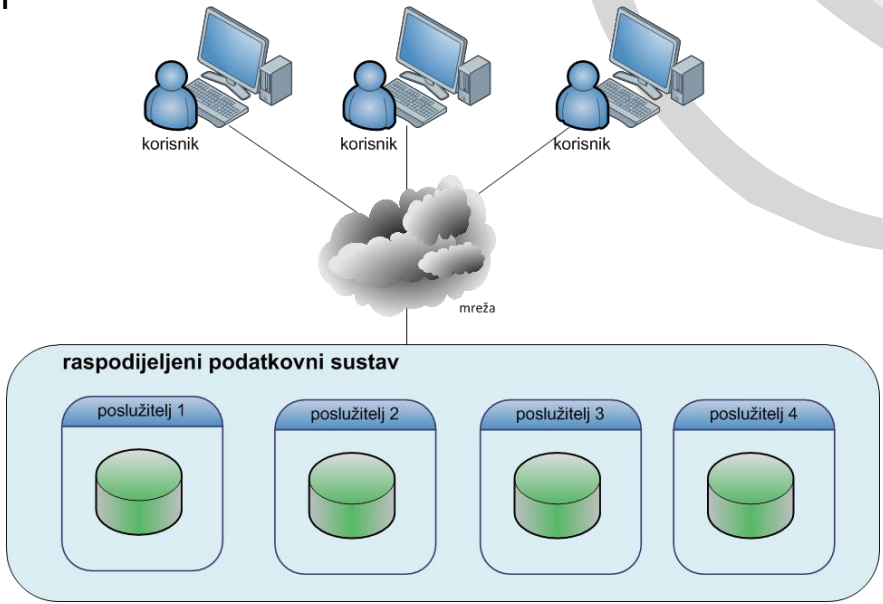
# Sadržaj

- Raspodijeljeni podatkovni sustavi
- BeeGFS
- Instalacija i konfiguracija
- Ostali raspodijeljeni podatkovni sustavi
- Raspodijeljeni podatkovni sustavi u Srcu
- Reference

# Raspodijeljeni podatkovni sustavi

# Uvod

- Spajanje više podatkovnih poslužitelja u jedinstven podatkovni prostor



# Uvod

- Prednosti
  - učinkovitiji pristup podacima – raspodjeljivanje opterećenja
  - proširivanje ili smanjivanje prostora prema potrebama
- Problemi
  - nedostupnost poslužitelja – potrebna replikacija podataka
  - konzistentnost podataka kod replikacije
  - mrežna povezanost
  - podržane platforme

# Vrste

- Mrežni datotečni sustavi
  - jedan podatkovni poslužitelj
  - primjeri: CIFS, NFS
- Paralelni datotečni sustavi
  - više podatkovnih poslužitelja
  - visoka učinkovitost
  - podatkovni poslužitelji na jednoj fizičkoj lokaciji
  - primjeri: GlusterFS, Lustre, **BeeGFS**, PNFS

# Vrste

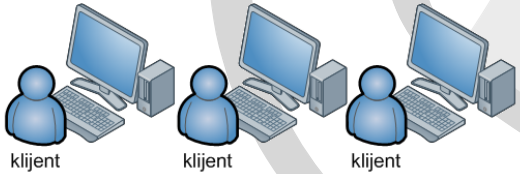
- Raspodijeljeni datotečni sustavi
  - veliki broj podatkovnih poslužitelja
  - podatkovni poslužitelji raspodijeljeni na više fizičkih lokacija
  - primjeri: Google FS, Hadoop FS, MogileFS

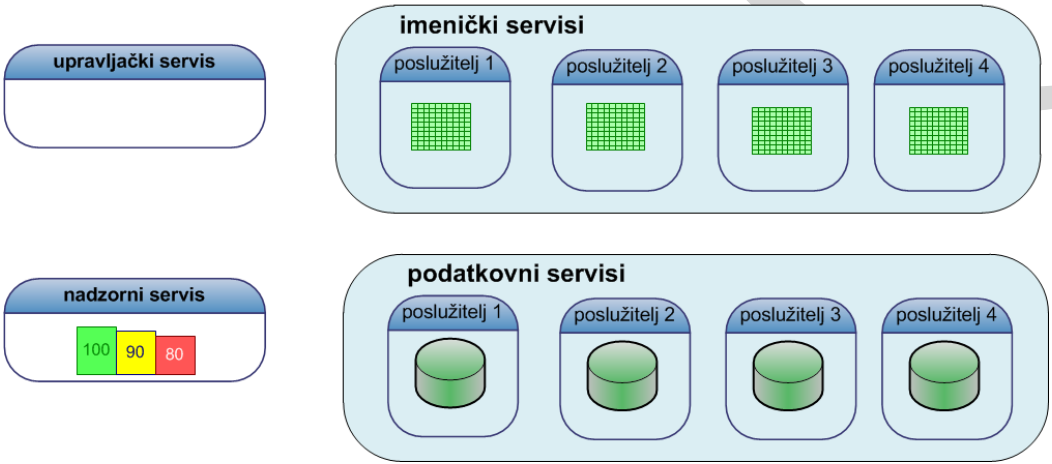
# BeeGFS

# BeeGFS

- Raspodijeljeni podatkovni sustav
- Razvijen od strane Fraunhofer Institute for Industrial Mathematics (ITWM)
  - inicijalno razvijen 2005. za potrebe HPC (engl. *High Performance Computing*) klastera
- Prethodno ime FhGFS
  - engl. *Fraunhofer Gesellschaft File System*
  - 2014. objavljeno novo ime i osnovan *spin-off* ThinkParQ
- Komercijalna programska potpora
  - dostupan je besplatno ili uz plaćanje podrške
  - neke funkcionalnosti zahtijevaju plaćanje podrške

# Arhitektura



# Komponente

- Upravljački servis (engl. *management*)
  - katalog svih servisa
  - omogućava dinamičko dodavanje i uklanjanje servisa
  - samo jedan za cijeli klaster
- Imenički servis (engl. *metadata*)
  - struktura datotečnog sustava
  - atributi i smještaj datoteka
  - proizvoljan broj u klasteru
  - raspodjela: jedan direktorij se sprema u jedan imenički servis

# Komponente

- Podatkovni servis (engl. *storage*)
  - pohrana podataka
  - proizvoljan broj u klasteru
  - moguće koristiti više particija (engl. *target*) po podatkovnom servisu
  - raspodjela: jedan podatak se sprema u jedan ili više podatkovnih servisa
- Nadzorni servis (engl. *admon*)
  - nadzire i prikuplja statistike ostalih servisa
  - grafički alat pisan u Javi
  - instalacija i upravljanje ostalim komponentama
  - opcionalan

# Komponente

- Klijenti
  - poseban kernel modul
  - ne zahtijeva izmjenu kernela

# Podržane platforme

- Operacijski sustavi
  - Red Hat Linux 5, 6 i 7
  - Suse Linux 11 i 12
  - Debian Linux 7 i 8
- Datotečni sustavi za pohranu podataka
  - XFS - preporučeni sustav za podatkovni servis
  - ext4 - preporučeni sustav za imenički servis
  - drugi datotečni sustavi sukladni specifikaciji POSIX (ext3, ext4, RaiserFS)
- Mrežne tehnologije
  - Ethernet (1Gb/s i 10Gb/s)
  - Infiniband (RDMA)

# Oblici spremanja datoteka

- Moguće podešavati na razini direktorija
- Dijeljenje datoteke (engl. *stripe*)
  - pogodno za velike datoteke
- Spremanje na više lokacija (engl. *mirror*)
  - omogućava visoku dostupnost datoteka
  - zahtjeva plaćenu podršku
- Spremanje podataka imeničkog servisa na više lokacija
  - eksperimentalna funkcionalnost
  - nije moguć automatski oporavak

# Napredne funkcionalnosti

- Odabir preferiranih imeničkih i podatkovnih servisa
- Korištenje više mrežnih sučelja
  - automatski prijelaz na drugu mrežu u slučaju ispada
  - postavljanje prioriteta mreža i automatski povratak na primarnu (engl. *failback*)
- Funkcionalnosti *extended attributes* i POSIX ACL
- Postavljanje korisničkih kvota
  - zahtjeva plaćenu podršku
- BeeOND (engl. *BeeGFS on demand*)
  - automatska uspostava klastera
  - korisno za računalne klastere i uspostavu testnih okolina

# Mane

- Nije u potpunosti implementirana visoka dostupnost
  - moguće ju ostvariti korištenjem dodatnih alata (npr. Red Hat Cluster Suite)
- Nadogradnje na glavne nove verzije zahtijevaju gašenje klastera
- Nije moguće balansirati podatke prilikom proširenja

# Instalacija i konfiguracija

# Instalacija

- Oblici instalacije
  - grafičko sučelje (admon) - ograničen skup mogućnosti
  - pomoću paketa - Red Hat, Suse i Debian
- Zadnja verzija je **2015.03**

# Instalacija

| Komponenta | Paketi | Servisi | Konfiguracijske datoteke |
|---|---|---|---|
| Upravljački servis | beegfs-mgmtd | beegfs-mgmtd | /etc/beegfs/beegfs-mgmtd.conf |
| Imenički servis | beegfs-meta | beegfs-meta | /etc/beegfs/beegfs-meta.conf |
| Podatkovni servis | beegfs-storage | beegfs-storage | /etc/beegfs/beegfs-storage.conf |
| Nadzorni servis | beegfs-admon | beegfs-admon | /etc/beegfs/beegfs-admon.conf |
| Klijent | beegfs-client<br>beegfs-helperd | beegfs-client<br>beegfs-helperd | /etc/beegfs/beegfs-client.conf<br>/etc/beegfs/beegfs-helperd.conf<br>/etc/beegfs/beegfs-mounts.conf |

# Konfiguracija - zajedničke opcije

| Opcija | Opis |
|---|---|
| sysMgmtdHost | Adresa upravljačkog servisa |
| storeAllowFirstRunInit | Da li je dozvoljeno inicijalizirati direktorij - postaviti u true prilikom instalacije, kasnije u false |
| connAuthFile | Datoteka koja sadrži dijeljeni ključ svih komponenata klastera |
| connInterfacesFile | Popis mrežnih sučelja koje servisi trebaju koristiti, lista je ujedno i prioritet pojedinih sučelja |
| connNetFilterFile | Popis mreža putem kojeg servisi mogu komunicirati s drugim servisima |
| tuneNumWorkers | Broj procesa koje će servis pokrenuti, 0 - broj procesorskih jezgri na stroju |

# Konfiguracija - upravljački servis

| Opcija | Opis |
|---|---|
| storeMgmtdDirectory | Direktorij u koji se spremaju podaci |
| sysAllowNewServers | Da li je dozvoljena prijava novih komponenata - postaviti u true prilikom instalacije, kasnije u false |
| sysAllowNewTargets | Da li je dozvoljena prijava novih particija - postaviti u true prilikom instalacije, kasnije u false |
| tune(Storage\|Meta)InodesLowLimit | Broj inodeova ispod kojeg servis prelazi u grupu *low* |
| tune(Storage\|Meta) InodesEmergencyLimit | Broj inodeova ispod kojeg servis prelazi u grupu *emergency* |
| tune(Storage\|Meta)SpaceLowLimit | Količina diska ispod kojeg servis prelazi u grupu *low* |
| tune(Storage\|Meta)SpaceEmergencyLimit | Količina diska ispod kojeg servis prelazi u grupu *emergency* |

# Konfiguracija - imenički servis

| Opcija | Opis |
|---|---|
| storeMetaDirectory | Direktorij u koji se spremaju podaci |
| storeUseExtendedAttribs | Podaci o direktorijima se spremaju u *extended attribute* - preporučeno uključiti zbog veće učinkovitosti servisa |
| storeClientXAttrs | Omogućiti funkcionalnost *extended attribute* - koristiti isključivo ukoliko postoji potreba jer smanjuje učinkovitost servisa |
| storeClientACLs | Omogućiti korištenje POSIX ACL (engl. *Access Lists*) - koristiti isključivo ukoliko postoji potreba jer smanjuje učinkovitost servisa; klijent mora koristiti kernel > 3.2. |

# Konfiguracija - podatkovni servis

| Opcija | Opis |
| --- | --- |
| storeStorageDirectory | Popis direktorija u koje se spremaju podaci |

# Konfiguracija - klijent

| Opcija | Opis |
|---|---|
| connCommRetrySecs | Broj sekundi nakon kojeg će pristup datoteci završiti greškom u slučaju ispada servisa, 0 - klijent će čekati beskonačno |
| connFallbackExpirationSecs | Broj sekundi nakon kojeg će klijent pokušati ponovno koristiti mrežno sučelje višeg prioriteta |
| tunePreferredMetaFile | Popis adresa imeničkih servisa koje će klijent primarno koristiti |
| tunePreferredStorageFile | Popis adresa podatkovnih servisa koje će klijent primarno koristiti |

# Konfiguracija - klijent

- Klijent zahtjeva dva servisa
  - beegfs-helperd - pomoćni servis za upravljanje sistemskim zapisima
  - beegfs-client - učitava kernel modul i montira BeeGFS sustave
- Popis BeeGFS sustava u datoteci /etc/beegfs/beegfs-mounts.conf
- Priprema kernel modula
  - automatski prilikom prvog pokretanja ili izmjene verzije BeeGFS-a ili kernela

# Pokretanje klastera

- Instalirati paketi
- Podesiti konfiguraciju svih servisa
- Pokrenuti upravljački servis
- Pokrenuti sve imeničke i podatkovne servise
- Provjeriti dostupnost svih servisa
- Pokrenuti montiranje na klijentima

# Upravljačke naredbe

- Potrebno instalirati paket beegfs-utils
- Provjera dostupnosti servisa

```
beegfs-check-servers
```

- Ispis svih aktivnih mrežnih veza sa servisima

```
beegfs-net
```

# Upravljačke naredbe

- Provjera iskorištenja direktorija svih servisa

```
beegfs-df
```

- Provjera stanja cjelokupnog BeeGFS datotečnog sustava

```
beegfs-fsck --checkfs
```

  - koristi se u slučaju problema u pristupanju datotečnom sustavu
  - moguće je izvođenje dok se sustav koristi (*online*)

# Upravljačke naredbe

- Glavna upravljačka naredba

```
beegfs-ctl
```

- Ispis servisa

```
beegfs-ctl --listnodes
```

- Ispis particija

```
beegfs-ctl --listtargets
```

# Upravljačke naredbe

- Uklanjanje servisa

```
beegfs-ctl --removenode
```

- Dohvat informacija o direktoriju

```
beegfs-ctl --getentryinfo /beegfs
Path:
Mount: /beegfs
EntryID: root
Metadata node: storage01 [ID: 18627]
Stripe pattern details:
+ Type: RAID0
+ Chunksize: 512K
+ Number of storage targets: desired: 1
```

# Upravljačke naredbe

- Postavljanje parametara direktorija (npr. *stripe*, *mirror*)

```
beegfs-ctl --setpattern --chunksize=1m --numtargets=4
/beegfs/mydir
```

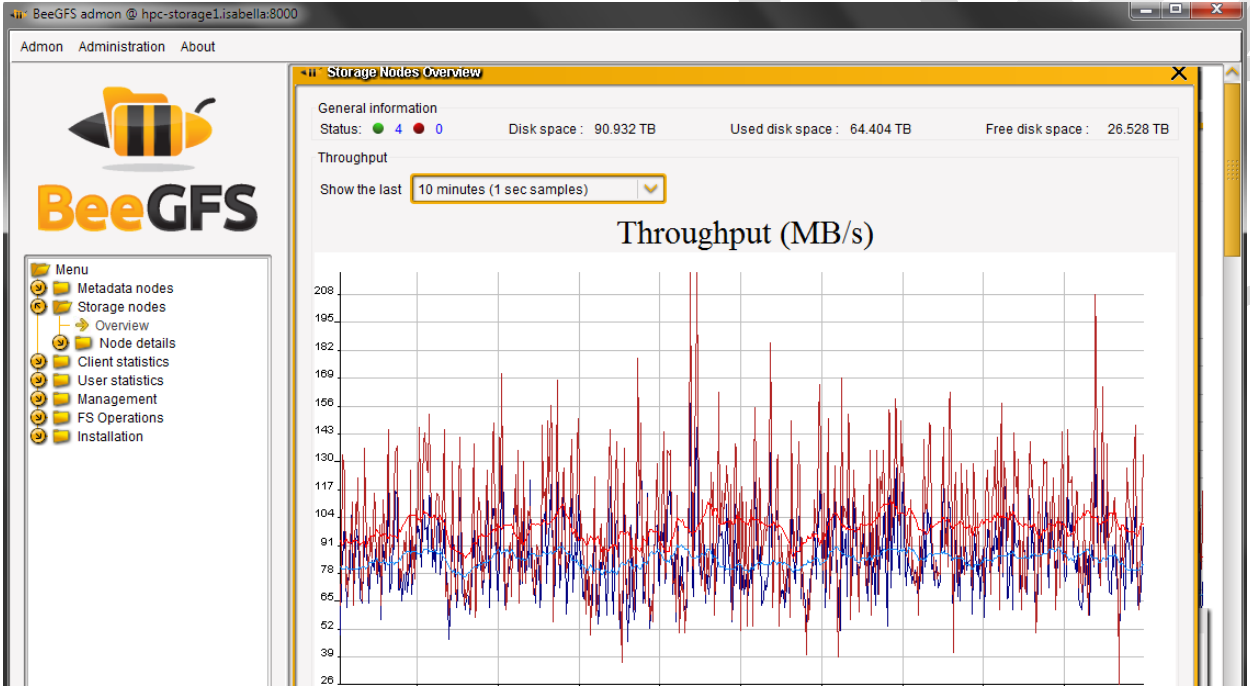
- Prikaz korištenja servisa

```
beegfs-ctl --serverstats
```

# Nadzorni servis

- Potrebno pokrenuti servis beegfs-admon
- Pristupiti web sučelju za dohvat Java grafičkog alata
- Početne lozinke
  - Administrator: admin
  - Information: information
- Ograničenje alata
  - moguć prikaz podataka do 5 dana u prošlosti
- Prikaz statistika
  - podatkovni servisi
  - imenički servisi
  - klijenti
  - korisnici

# Nadzorni servis



# Ostali raspodijeljeni podatkovni sustavi

# GlusterFS

- Paralelni datotečni sustav
  - namijenjen ostvarivanju velikih datotečnih sustava
  - izgradnja podatkovnih *clouda*
- Komponente
  - čvorovi
  - klijenti
- Prednosti
  - jednostavnost održavanja (nema imeničkih servisa)
  - podržani različiti oblici spajanja čvorova u podatkovne cjeline
- Mane
  - sporiji rad kroz modul FUSE (engl. *Filesystem in Userspace*)
  - neki oblici (*distributed*, *striped*) ne osiguravaju visoku dostupnost podataka

# Lustre

- Paralelni datotečni sustav
  - prvenstveno namijenjen HPC klasterima
- Komponente
  - *Metadata Server* (MDS) – imenički servis
  - *Object Storage System* (OSS) – podatkovni čvorovi
- Podržane mrežne tehnologije
  - Ethernet, Infiniband, Myrinet, Quadrics
- Mane
  - složena implementacija visoko dostupnog servisa MDS (Heartbeat ili RHEL Cluster Suite)
  - potrebno koristiti nestandardnu jezgru

# Ceph

- RedHat-ov raspodijeljeni sustav za podatkovni *cloud*
- Tri oblika spremanja podataka
  - objektno spremište
  - *block device* - za spremanje diskova virtualnih poslužitelja
  - standardni datotečni sustav
- Mane
  - datotečni sustav nije u potpunosti spreman za produkcijske sustave

# Hadoop File System

- Raspodijeljeni datotečni sustav
  - velika razina replikacije
  - pogodan za veliki broj zemljopisno raspodijeljenih čvorova
- Sastavni dio sustava Apache Hadoop
  - izvođenje aplikacija *Map-Reduce*
- Mane
  - implementacija u programskom jeziku Java

# MogileFS

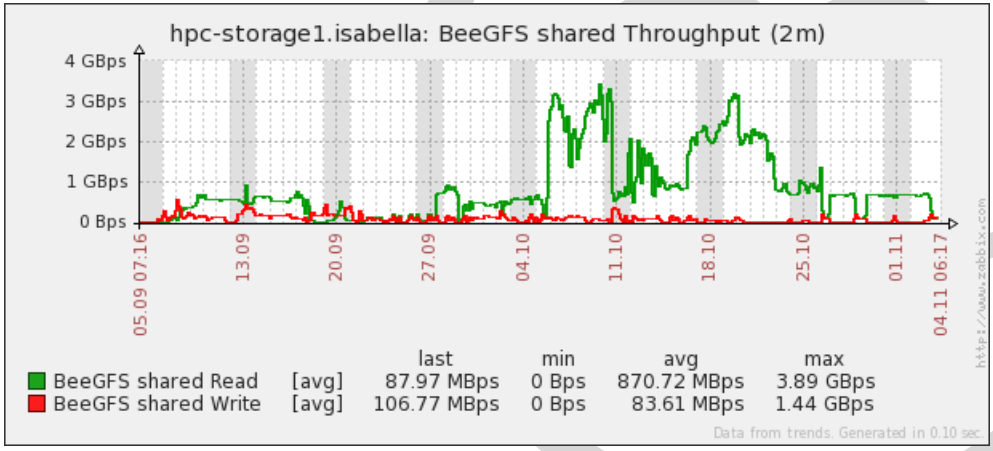
- Raspodijeljeni datotečni sustav
  - prvenstveno namijenjen web poslužiteljima
- Zasnovan na HTTP-u i WebDAV-u
- Mane
  - pristup podacima putem komandno linijskog sučelja

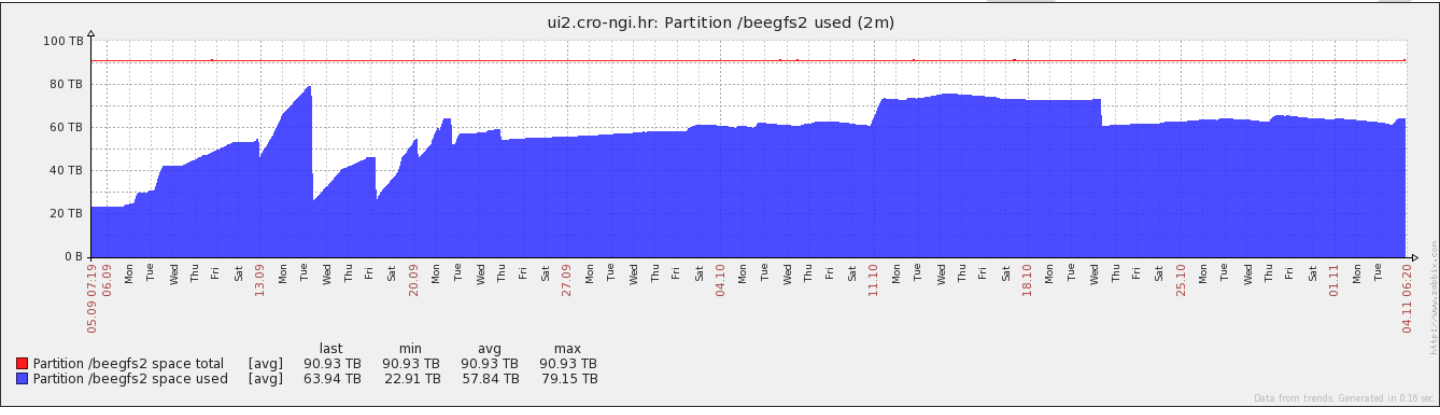
# Raspodijeljeni podatkovni sustavi u Srcu

# BeeGFS

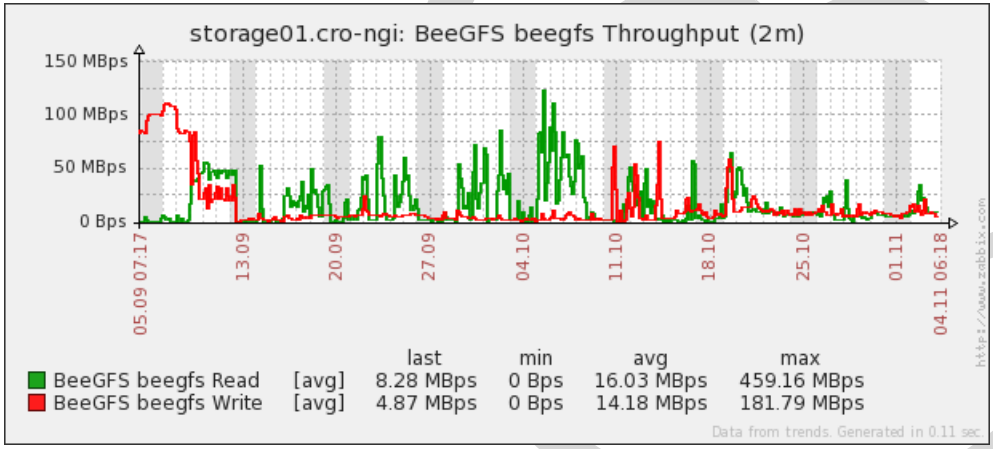
- Klaster Isabella i CRO NGI
- Korisnički direktorij (*home*)
  - ukupno 110 TB
  - 4 podatkovna poslužitelja
  - 1 imenički servis
  - iSCSI spremišni element
- Direktorij za pohranu privremenih datoteka (*scratch*)
  - učinkovit rad s velikim datotekama
  - ukupno 90 TB
  - 4 podatkovna poslužitelja
  - 4 imenička servisa

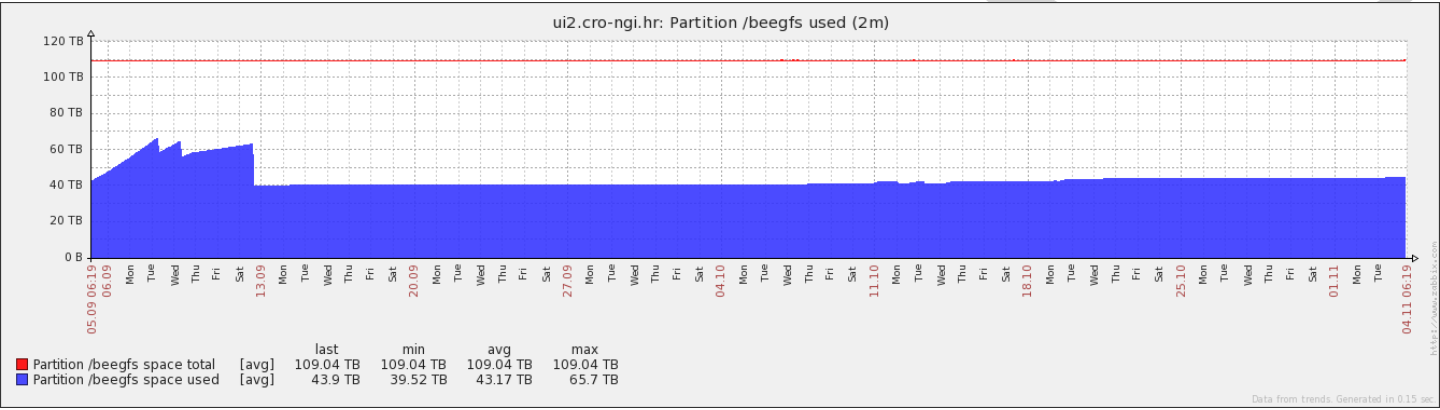
# BeeGFS


hpc-storage1.isabella: BeeGFS shared Throughput (2m)
4 GBps
3 GBps
2 GBps
1 GBps
0 Bps
05.09 07:16
13.09
20.09
27.09
04.10
11.10
18.10
25.10
01.11
04.11 06:17
last min avg max
BeeGFS shared Read [avg] 87.97 MBps 0 Bps 870.72 MBps 3.89 GBps
BeeGFS shared Write [avg] 106.77 MBps 0 Bps 83.61 MBps 1.44 GBps
Data from trends. Generated in 0.10 sec.


ui2.cro-ngi.hr: Partition /beegfs2 used (2m)
100 TB
80 TB
60 TB
40 TB
20 TB
0 B
last min avg max
Partition /beegfs2 space total [avg] 90.93 TB 90.93 TB 90.93 TB 90.93 TB
Partition /beegfs2 space used [avg] 63.94 TB 22.91 TB 57.84 TB 79.15 TB
Data from trends. Generated in 0.16 sec.

# BeeGFS


storage01.cro-ngi: BeeGFS beegfs Throughput (2m)
150 MBps
100 MBps
50 MBps
0 Bps
05.09 07:17
13.09
20.09
27.09
04.10
11.10
18.10
25.10
01.11
04.11 06:18
last min avg max
BeeGFS beegfs Read [avg] 8.28 MBps 0 Bps 16.03 MBps 459.16 MBps
BeeGFS beegfs Write [avg] 4.87 MBps 0 Bps 14.18 MBps 181.79 MBps
Data from trends. Generated in 0.11 sec.


ui2.cro-ngi.hr: Partition /beegfs used (2m)
120 TB
100 TB
80 TB
60 TB
40 TB
20 TB
0 B
last min avg max
Partition /beegfs space total [avg] 109.04 TB 109.04 TB 109.04 TB 109.04 TB
Partition /beegfs space used [avg] 43.9 TB 39.52 TB 43.17 TB 65.7 TB
Data from trends. Generated in 0.15 sec.

# BeeGFS

- Podatkovne usluge **Filesender** i **MojOblak**
  - usluge dijele isti datotečni sustav
- Dijeljeni prostor
  - 1 podatkovni i imenički poslužitelj
  - iSCSI spremišni element

# Sustavi korišteni u prošlosti

- GlusterFS
  - CRO NGI
  - podatkovne usluge Filesender, GSS, JKP
  - ukupno preko 200 TB prostora
  - napušten zbog slabije učinkovitosti od BeeGFS-a i ujednačavanja raspodijeljenih podatkovnih sustava
- Lustre
  - klaster Isabella
  - 6 poslužitelja (2 MDS, 2 OSS, 2 mosta između mreža Infiniband i Ethernet)
  - visoka dostupnost (Heartbeat)
  - napušten zbog složenosti održavanja, korištenja nestandardne jezgre i prestanka održavanja od strane Oraclea

# Sustavi korišteni u prošlosti

- GFS
  - klaster Isabella i CRO NGI
  - klasterski datotečni sustav
  - omogućava pristup istom logičkom disku na spremišnom sustavu s više poslužitelja
  - napušten zbog loše učinkovitosti
- NFS
  - klaster Isabella i CRO NGI
  - korisnički direktorij (*home*)
  - napušten zbog loše učinkovitosti pri velikom opterećenju
  - još uvijek se koristi na manjim klasterima

# Reference

- Službena web stranica BeeGFS
  - http://www.beegfs.com
- Korisnička dokumentacija BeeGFS
  - http://www.beegfs.com/content/documentation/
  - kvalitetni članci o testiranju učinkovitosti odabiru broja podatkovnih i imeničkih servisa
- Korisnički wiki
  - http://www.beegfs.com/wiki/TableOfContents
  - detaljne informacije o instalaciji i podešavanju svih komponenata sustava

# Hvala na pažnji!

Pitanja?

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

www.srce.unizg.hr



Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr/otvoreni-pristup

srce
otvoreni pristup