Költséghatékony adattárolás grid alapokon

Nagy Zsombor
Stefán Péter
Szalai Ferenc
Vitéz Gábor
NIIF/HUNGARNET

Vázlat

- A storage + grid technológiákról általában
- A kétrétegű modell
- Ami alul van - storage node kialakítása
- Speciális technológia: AoE
- Ami fölül van - GUG storage management
- Továbbfejlesztési elképzelések
- Összefoglalás

Grid + Storage

- A grid = elosztott számítási és adattárolási infrastruktúra.
- Milyen adatok vannak?
  - számítási folyamatok inputja,
  - átmeneti állományok,
  - számítások eredményeképp előállt nagyméretű adathalmaz (mammográfiai képek + feldolgozásuk).
- Óriási az adatigény: jelenleg 1 Pbyte csak CERN adattárolásra.

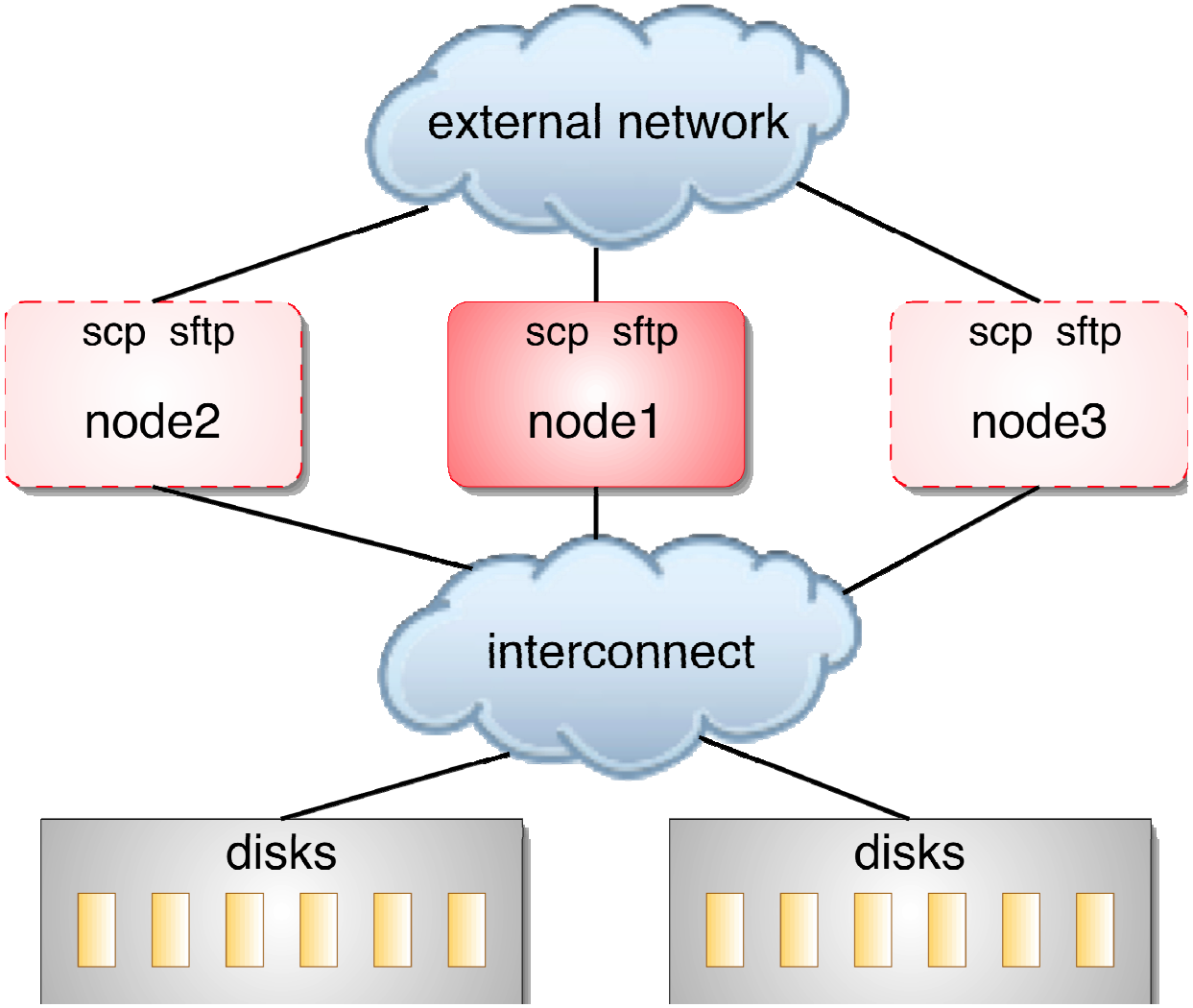
Storage felépítés

- Alapvetően kétfajta aspektusból foglalkozunk storage felépítéssel:
  - nagy méretű hely áll rendelkezésre csatolva egy (vagy több) file szerveren (storage node telepítése),
  - azok a szolgáltatások, amelyeken keresztül a felhasználók e lemezterületet akár grid-ből, akár közvetlenül elérhetik (grid storage management).
- E két réteg bár harmónikus egészet ad, független egymástól, és célszerű is külön-külön tárgyalni.

Storage csomópontok

- Olyan csomópontok, amelyek nagy adatkötetet kezelnek.
- Sokféleképp építhető fel:
  - optikai (pl. FC) vs. réz (pl. Ethernet) alapon,
  - HA vs. nem HA megoldások,
  - fürtözött, vagy nem fürtözött megoldások.
- Ilyen csomópontoknak kell nagy sávszélességgel, és kis késleltetéssel csatlakozni a gerinchálózathoz.

Storage csomópontok



ATA over Ethernet

- Egy lehetséges megoldás: (S)ATA over Ethernet.
- Olcsó megoldás: 250 eFt/Tbyte fajlagos költség.
- Teljesítményben: 50 Mbit/s ATA, 500 Mbit/s SATA.
- Volume managementtel bővíthető e teljesítmény.

NIIF grid storage csomópont

```
# ls -l /dev/etherd/e4.*
brw-rw---- 1 root disk 152, 40 Nov  4 11:29 /dev/etherd/e4.0
brw-rw---- 1 root disk 152, 41 Nov  4 11:29 /dev/etherd/e4.1
brw-rw---- 1 root disk 152, 42 Nov  4 11:29 /dev/etherd/e4.2
```

```
# cat /proc/mdstat
Personalities : [raid5] [raid6]
md2 : active raid6 etherd/e7.2[6] etherd/e6.2[5] etherd/e5.2[4] etherd/e4.2[3]
etherd/e3.2[2] etherd/e2.2[1] etherd/e1.2[0]
      1953556480 blocks level 6, 64k chunk, algorithm 2 [7/7] [UUUUUUU]
      [=========>.........] resync = 43.6% (170377004/390711296)
finish=1098.8min speed=3341K/sec
```

NIIF grid storage csomópont

- 7 x 10 x 400 Gbyte összkapacitás.
- AoE polcok, tartalék hardver elemekkel.
- 2 db 64-bites kiszolgálógép (jelenleg csak egy működik).
- RAID6 + LVM + XFS konfiguráció.
- HW failover a két kiszolgáló szerver között.
- Kb. 50 Mbyte/sec RW IO (Parallel ATA).

NIIF grid storage csomópont

- Jelenleg az NIIF Névtár szolgáltatásból autentikáljuk a felhasználókat.
- A szerver jól definiált szolgáltatásokon keresztül érhető el:
  - SCP-only, rsync, unison - grid témaszámmal,
  - FTPS - NIIF által kiadott X509 tanúsítvánnyal,
  - grid rendszerben (bővebben később).
- Tervezzük még:
  - GridFTP elérés.

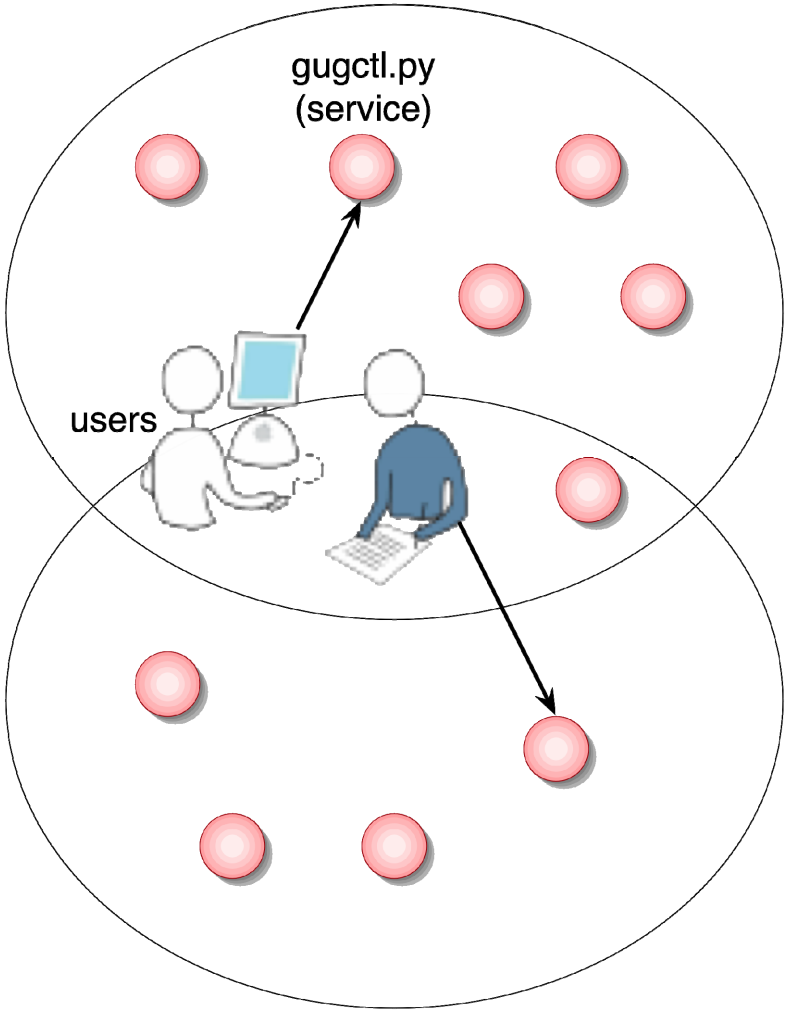
NIIF grid storage csomópont

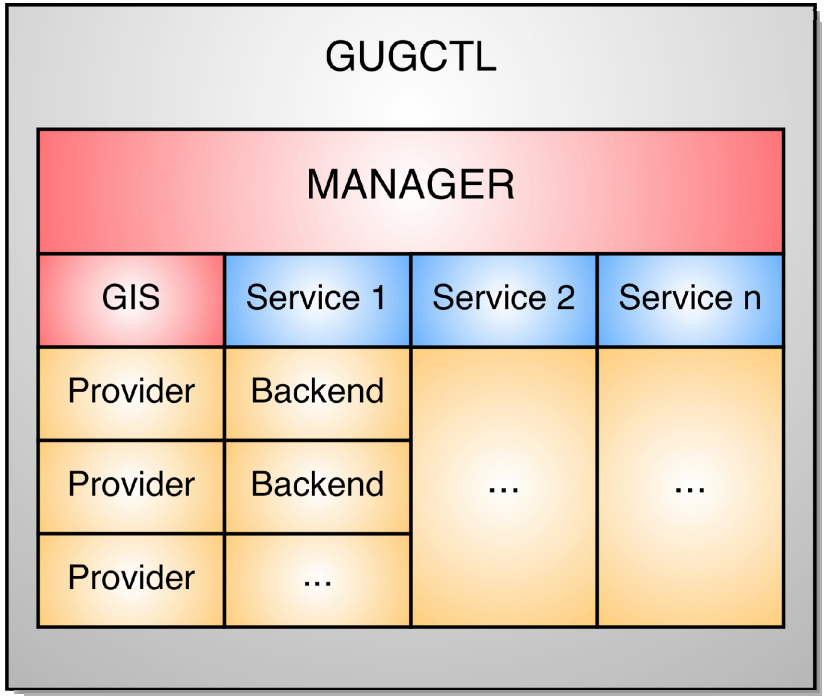
- Mire érdemes használni?
  - Szerver backup.
  - Nagy mennyiségű adat átmeneti tárolására (pl: molekula-dinamikai számítások outputja, később feldolgozandó adatok, logok).
  - Általános személyes adattárolásra (kivéve jogsertő tartalmak :).
- Mire NEM érdemes használni (egyenlőre)?
  - Állomány megosztásra több ember között.

Grid elérés

- Bár a storage csomópont elérhető grid-en kívül használt protokollok segítségével is, elsődleges feladata a grid jobok adatainak tárolása.
- Egy lépés a teljesen integrált infrastruktúra irányába.
- Elérhetőnek kell lennie a grid számára:
  - hálózati elérés (GRID VPN része),
  - grid kliens interfész része.
- Grid Underground (GUG) - a következő generációs ClusterGrid középréteg.

GUG architektúra

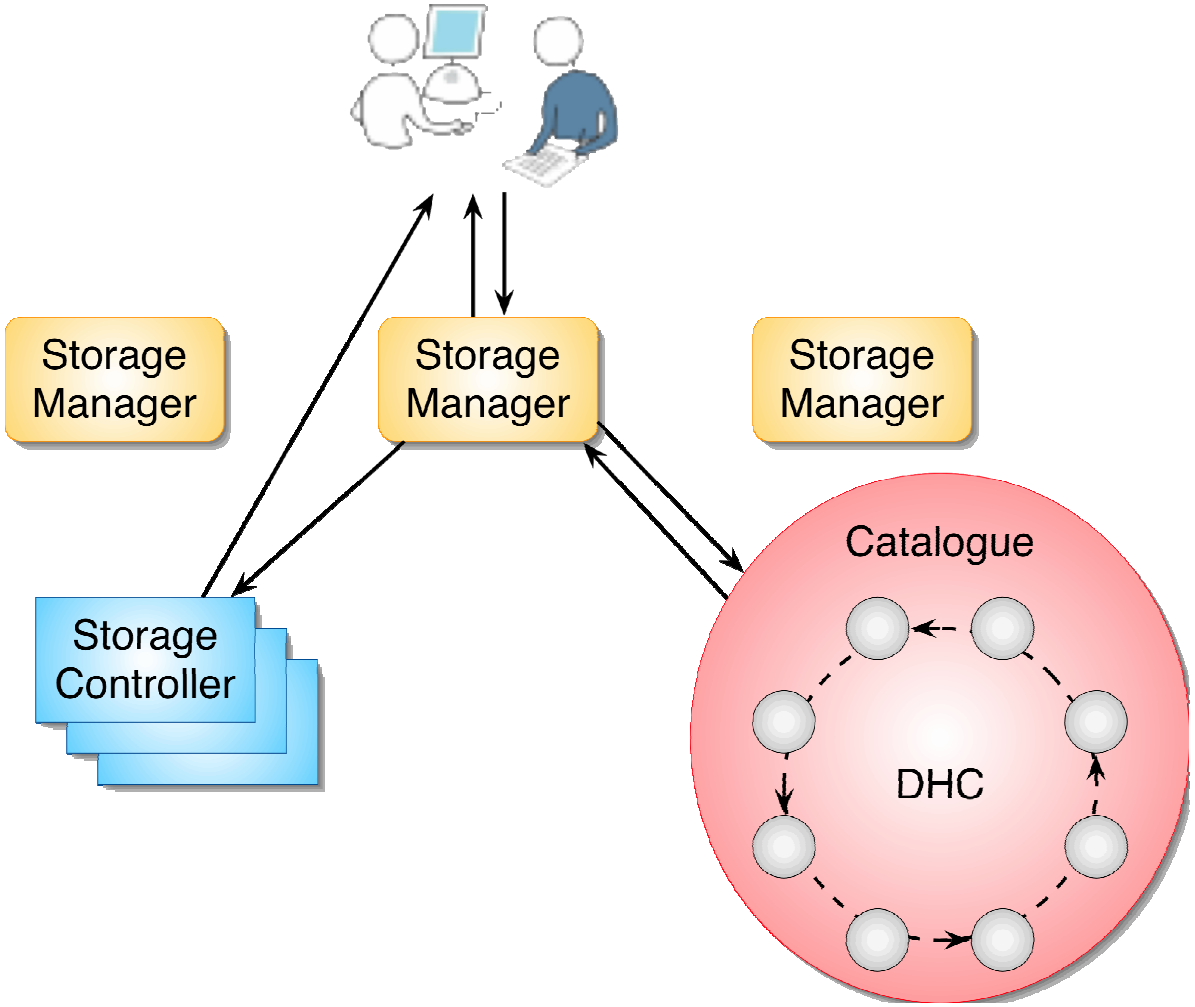




Storage szolgáltatás

- Storage szolgáltatás: olyan webszolgáltatások halmaza, amelyek segítenek a grid felhasználónak adatai automatikus menedzselésében.
- Független a storage node kialakításától, tetszőleges gép, amelyen van hely futtathatja a storage szolgáltatásokat.
- A storage szolgáltatás segít az adat elhelyezésében, menedzselésében, kezeli az elosztott környezetet, másolatokat készít az egyes adatállományokból (automatikus replika kezelés).

Storage szolgáltatás komponensei



Storage szolgálatás komponensei

- Hasonló architektúrát valósít meg, mint a Unix FS, csak teljesen elosztottan.
- Az elosztott csomópontok web szolgáltatás interfészen át kommunikálnak egymással.
- Felhasználói nézet a kliensen: egy nagy file rendszer a /grid alatt.
- Minden file rendelkezik egy globálisan egyedi GUID-val (olyan, mint az i-node).
- Minden GUID leképződik egy vagy több SURL-re (az állomány konkrét fizikai elérési helye)

Elosztott Hash Katalógus (DHC)

- Ő tárolja a filenév → GUID, illetve a GUID → Metaadat (pl: SURL) összerendeléseket.
- Redundáns gyűrű architektúra, nincs egyetlen sebezhető pontja.
- Kétféle bejegyzés lehet benne:
  - Könyvtár bejegyzés: GUID + file név,
  - Állomány bejegyzés: metaadat (méret, acl, SURL stb).

Elosztott Hash Katalógus (DHC)

- Példa: a /grid DHC bejegyzése.

```
'0' →
    <?xml version="1.0"?>
    <Directory>
        <Created>1129881054</Created>
        <Entry>
            <GUID>jufebez-945</GUID>
            <Name>home.txt</Name>
        </Entry>
    </Directory>
```

Elosztott Hash Katalógus (DHC)

- Példa: a /grid/home/ DHC bejegyzése.

```
'jufebez-945'
    <?xml version="1.0" ?>
    <Directory>
        <Created>1131204750</Created>
        <Entry>
            <GUID>labozuc-707</GUID>
            <Name>vitezg</Name>
        </Entry>
    </Directory>
```

Elosztott Hash Katalógus (DHC)

- Példa: /grid/home/{vitezg/, tls.py} DHC bejegyzése.

```
'jufebez-945' →
    <?xml version="1.0" ?>
    <Directory>
        <Created>1131204750</Created>
        <Entry>
            <GUID>labozuc-707</GUID>
            <Name>vitezg</Name>
        </Entry>
        <Entry>
            <GUID>vovoluv-014</GUID>
            <Name>tls.py</Name>
        </Entry>
    </Directory>
```

Elosztott Hash Katalógus (DHC)

- Példa: állomány DHC bejegyzése.

```
'vovoluv-014' =>
     <?xml version="1.0" ?>
     <File>
          <Created>1131204956</Created>
          <Size>796</Size>
          <SURL>srm://samu.ki.iif.hu:21111/StC?
                              ekatacu221</SURL>
     </File>
```

- TURL: http://samu.ki.iif.hu:21113/1be6779e-22f2-483b-98d8-8e825b676603.0

```
grid storage cp /grid/home/vitezg/tls.py .
```

Kliens felület

```
$ grid storage {POSIX parancs}

$ grid storage
Missing command.
usage:
  grid storage command [args]

commands:
  get   get a file from stroge (get <LFN> <localfilename>)
  mkdir create a directory on storage (mkdir <LFN>)
  mv    renames a file or directory in storage (mv <from>
                                               <to>)
  ls    list contents of a directory (ls <LFN>)
  put   put a file to storage (put <localfilename> <LFN>)
  rm    remove file from storage (rm <LFN>)
  cp    copy a file to/from/in storage (cp <from> <to>)
  rmdir removes an empty direcotry (rmdir <LFN>)
```

Összefoglalás

- Nagyon fontos, hogy nagy adatmennyiséget, egy arra alkalmas infrastruktúrán, robusztus módon tudjunk tárolni.
- Két dologról volt szó:
  - Hogyan lehet nagy köteteket kezelő ún. storage csomópontokat kialakítani?
  - Hogyan lehet ezeken az adatkezelést egységes rendszerbe foglalni?
- Fontos jövőbeli irányok:
  - a “felső kategória” tesztelése (Topspin eszközök),
  - a mostani rendszer mennyiségi és minőségi bővítése.

Köszönöm a figyelmet!

http://www.clustergrid.niif.hu
http://gug.grid.niif.hu