SZTAKI Felhő projekt

Ormos Pál
MTA SZTAKI
ormos.pal@sztaki.mta.hu

HBONE Workshop 2012

Miről lesz szó?

- Előzmények
- Feladatok
- Összefoglalás

Előzmények

- Labor szintű felhő kísérletek az intézetben
- Élenjáró GRID kutatások
  - EDGI http://edgi-project.eu/
  - S-CUBE http://www.s-cube-network.eu/
  - Brein http://www.eu-brein.com
- EU-s Cloud projekt
  - SCI-BUS http://www.sci-bus.eu

A Felhő projekt

A Projekt célja a felhőkhöz, mint elosztott informatikai rendszerekhez kapcsolódó kutatások végzése. Intézeti felhő kifejlesztése és üzembe helyezése, ami lehetővé teszi az intézeti informatikai infrastruktúra jelentős korszerűsítését.

- 4 osztály összefogásával (ITAK, LPDS, DSD, ILAB)
- Akadémiai és intézeti támogatás
- 2 éves projekt
- kutatási és fejlesztési feladatok
- Az alfa rendszer 2012.10.24-én elindult

A Felhő projekt

- Kutatási és fejlesztési feladatok
  - Automatikus és elasztikus skálázhatóság
    - Cél, a meglevő SZTAKI szolgáltatások és a Hadoop technológia felhőbe migrálhatóságának és skálázhatóságának vizsgálata.
  - Adatintenzív felhő
  - Biztonság és Identity Management

Infrastruktúra

- 2 db Dell R415 frontend szerver
  - 32GB RAM, 3TB disk, AMD processzor
- 7 db Dell R815 node
  - 256GB RAM, 1TB disk, AMD processzor
- 2 db Dell 6248 PowerConnect switch
  - 48 db 10/100/1000 port + 4 TGB port (2 optika, 2 réz)
- 1 db Dell 3600i storage
  - 2 vezérlővel, 36TB disk
- 1 db Dell R510 Linux storage
  - 48 GB RAM, 36 TB disk, Xeon processzor

Az eszközöket közbeszerzésen keresztül vásároltuk. A szállító a Humansoft volt.

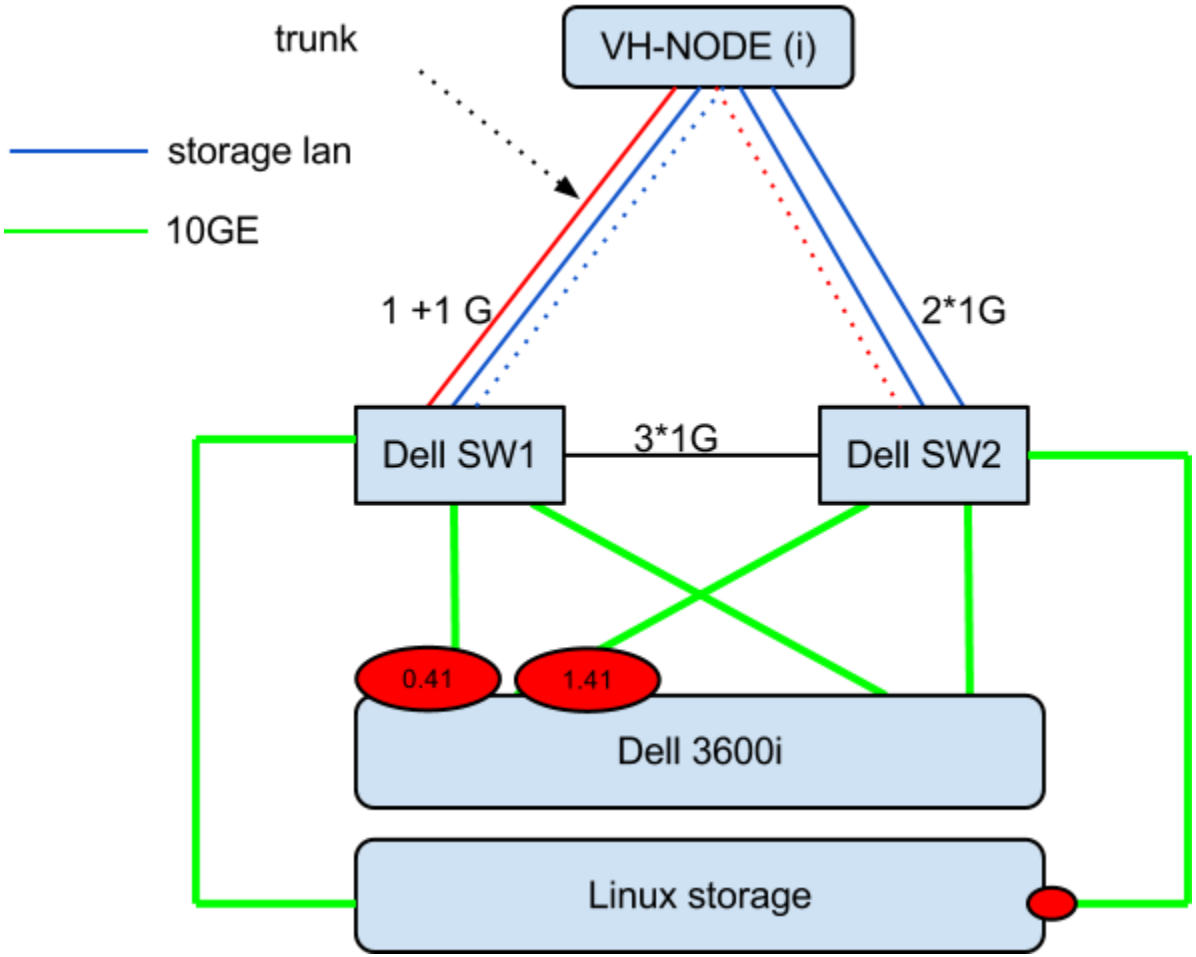
Hálózat

- IP cím tartományok, VLAN-ok tervezése, kialakítása megtörtént
- Különféle hálózati konfigurációk kipróbálása
  - Bonding
    - Hashing algoritmusok problémája
  - Több subnet-es iSCSI elérés az MD3600 felé, rdac optimális használata a multipath-ban (active és ghost portok)
    - Az MD3600i különböző switch-be kötött portjai különböző subnet-ben vannak, csakúgy mint a node-ok portjai. Így biztosítjuk, hogy a csomagoknak ne kelljen a switch-ek között utazni.
    - Az egy switchbe kötött két port közül az egyik aktív a másik passzív. Ezt a node-okon futó multipath daemon felismeri és csak az aktív utakat használja, a többi ghost állapotba kerül, ezekre csak hiba esetén kapcsol át. Az aktív utakat pedig round robin algoritmus szerint használja, így az összes aktív út ki van használva.

Hálózat

- Offload driver és bonding együttes működtetése, több MAC cím egy porton
  - A shared LVM és az iSCSI driverek másféle megoldást igényelnek, ha load-balancing-ot szeretnénk. Az iSCSI driver több TCP connection-t használ (image-enként egyet), viszont a targetnek egy IP címe van. Így ide az LACP bonding tűnt ideális megoldásnak. Az Dell storage-nak viszont több portja és IP címe van, ebből mindig kettő aktív, a másik kettőre hiba esetén kapcsol át. Gyárilag támogatja az rdac multipath megoldást, így ehhez az eszközhöz multipath-t kell használni.
  - A node-okban szereplő hálókártyák támogatják a bnx2i nevű iSCSI offload drivert. Ennek használatával egy másik interface jön létre a valódi fizikai interace mellett. Ennek a kettőnek külön MAC címe van, így lehetőség nyílt arra, hogy a kettőt (bonding, multipath) egymás mellett használjuk. Az offload interface-ek multipath-ban, az eredetiek pedig bondingban működnek.

Hálózat


trunk
VH-NODE (i)
storage lan
10GE
1 +1 G
2*1G
Dell SW1
3*1G
Dell SW2
0.41
1.41
Dell 3600i
Linux storage

Hálózat

- Az alfa szolgáltatáshoz megvalósított hálózati kialakítás még nem végleges
  - legalább háromszor át kellett kábelezni
  - legalább ennyiszer logikailag is módosítani kellett
- Discard és PAUSE framek problémája
  - Dell Powerconnect 6248 teljesítménye nem optimális

AAI

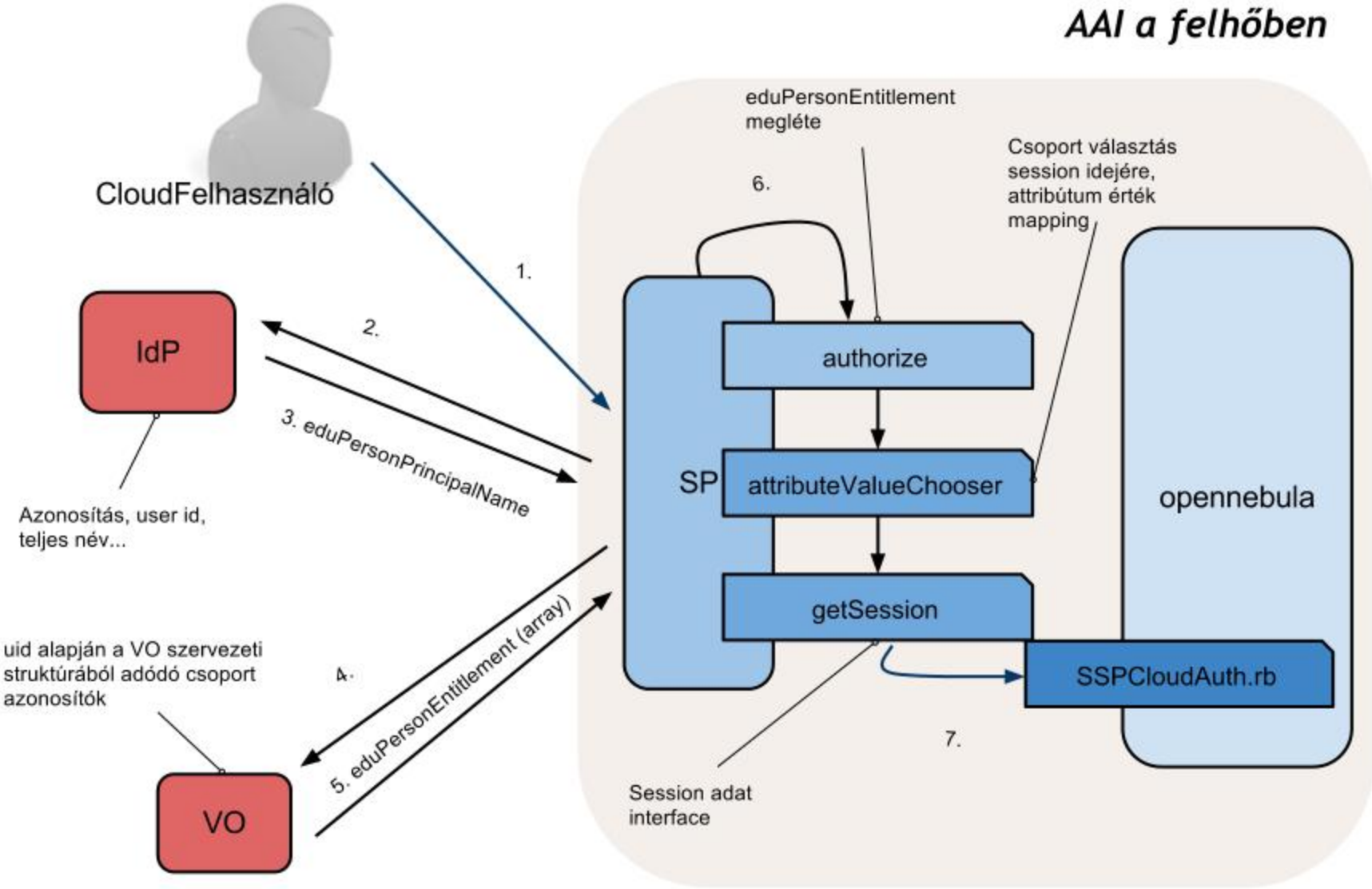
A cél, hogy az OpenNebula hozzáférés az intézeti SAML IdP-vel és Virtual Organization szoftverrel használható legyen

- Többféle variációt próbáltunk
- A végleges megoldás működése röviden:
  - a SimpleSAMLPHP-től egy létrejött session alapján - SessionID-val - lekérdezhetőek a bejelentkezés adatai JSON-ben
  - a lekérdezett JSON-ben szerepel minden szükséges adat egy felhasználó kezeléséhez (felhasználónév,csoportnév,autentikált-e?, authentikáció “erőssége”)
  - ha nem autentikált a SimpleSAMLPHP-ban, akkor nem engedélyezi a belépést
  - Az OpenNebula 3.8 megjelenése után patch készül amit beküldünk beolvasztásra
- Továbblépési lehetőség: csoportonként megkövetelhető erős authentikáció (két faktoros azonosítás, yubikey)

AAI

- Integrált a Virtual Organization szoftverrel, ráadásul több csoportba tartozás esetén van választó modul.
  - felhasználók, csoportok könnyebb kezelése VO-ban,
  - csoportkezelést a projekt menedzserei végzik, nem pedig rendszergazdák,
  - Sunstone-ban eredetileg a felhasználó nem tartozhat több csoporthoz, a megoldással bejelentkezéskor kiválasztható, hogy mely csoport jogosultságait akarja használni a felhasználó
  - a felhasználók, csoportok létrejönnek, ha szükség van rájuk
  - csoportváltozás esetén frissül a felhasználó csoportja

AAI a felhőben



iSCSI mérések

- Méréseket az alábbi eszközökkel végeztük:
  - vdbench (Dell eszköz)
    - Csak hiba meghatározására használtuk, hátránya, hogy csak egy úton tudjuk elérni az iSCSI diszkeket, multipath nem tesztelhető vele.
  - iometer
    - Ezzel végeztük a szisztematikus méréseket
    - Főbb tanulságok:
      - a teljes sebesség eléréséhez egy gépről is több workert kell indítani, valamint több MB nagyságrendű szegmensméretet kell beállítani.
      - Egy node a Dell storage-on olvasás esetén a három interface-t maximálisan ki tudja használni, írás esetén ennél kisebb , kb. 2-2,5 Gbps sebesség érhető el.

iSCSI mérések

- Iometer
  - a Linuxos storage-nál az 1 Gbps volt a maximális sebesség, mert egy felmountolt partició egy tcp connectiont használt, így a bonding korlátai miatt nem is lehetett e fölé menni.
  - Kis szegmensméret (4KB) esetén a linuxos storage volt a gyorsabb, a workerek számától függően 15-20 MBps (4500-5000 IOPS) sebességet produkált, az MD3600i esetén ez ~5 MBps (~1200 IOPS) volt.
  - A több node-ról egyszerre olvasás esetén kiderült, hogy a maximális sebesség az MD3600i esetén ~730 MBps.
- nagy fájl másolása
- Iozone

iSCSI mérések

- Összefoglalva:
  - 2G/node diszk sebességre lehet számítani a jelenlegi konfigurációval
- virtio/virtuális IDE overhead mérése
  - Olvasásnál nincs is olyan nagy különbség
  - Írásnál igen+CPU használata magasabb
- Image fájl/partíció
  - Csak a partíció 25-50%-al gyorsabb
  - Blokk mérettel +/- 15%

Cgroups-os IOLimit modul

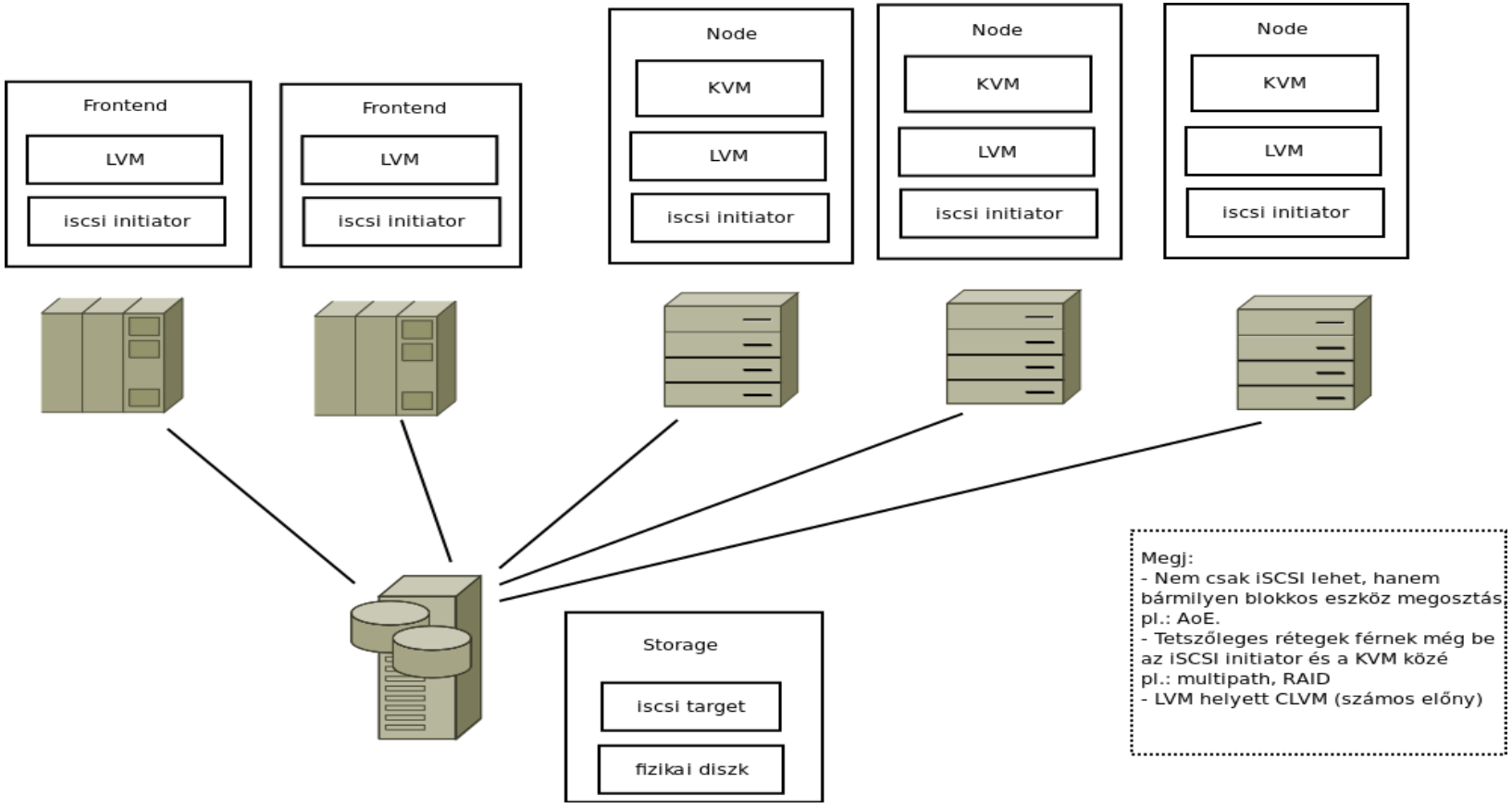
Ez az OpenNebula kiegészítés a Cgroups alrendszer segítségével a diszk hozzáférést szabályozza.

- A linux-ban elérhető képesség a control groups. Különféle erőforrásokra lehet limiteket mondani process group-onként
  - CPU/Mem/Disk IO
- A Shared LVM patch szükséges a helyes működéséhez
- Egy a Cgroups Block IO Controller alrendszerét használó hook-ról van szó, mely a virtuális gép RUNNING állapotába kerülésekor fut le
- az OpenNebula webfelületén limitek adhatók meg a virtuális gépre B/s-ben (lehetőség van IO/s-es megadásra is) a virtuális gép template-jében
- az ITAK cloud-on (Ubuntu) lett kifejlesztve, kipróbálva, jól működik
  - Az elkészült hook illeszthető Debian vagy RedHat alapú operációs rendszerekhez is
  - A Centos-ban más a cgroups elrendezése, adaptálni kell, de valószínűleg probléma nem lesz vele

Cgroups-os IOLimit modul

- részletes telepítési, konfigurációs és fejlesztői dokumentáció készült hozzá
- Későbbi tervek:
  - a Cgroups-ban a hálózati eszközök is megjelennek, így minden bizonnyal a hálózati sávszélességet is lehetne limitálni, de ezt nem próbáltuk még ki
  - teszt megtervezése, majd egy szemléletes
  - teszt bemutatása eredményekkel együtt

SharedLVM


Frontend
LVM
iscsi initiator
Frontend
LVM
iscsi initiator
Node
KVM
LVM
iscsi initiator
Node
KVM
LVM
iscsi initiator
Node
KVM
LVM
iscsi initiator
Storage
iscsi target
fizikai diszk
Megj:
- Nem csak iSCSI lehet, hanem bármilyen blokkos eszköz megosztás pl.: AoE.
- Tetszőleges rétegek férnek még be az iSCSI initiator és a KVM közé pl.: multipath, RAID
- LVM helyett CLVM (számos előny)

Shared LVM

- A shared LVM driver fejlesztése és tesztelése az ITAK cloud környezeten történt
  - A 3.4-es OpenNebula alatt
  - Kihasználtuk ehhez az NIIFI által biztosított storage infrastruktúrát
  - A már meglévő és az LPDS által továbbfejlesztett iSCSI driver azért nem volt megfelelő erre a célra

fejleszteni kellett egyet, ennek lett a neve shared LVM

- shared LVM patch: http://dev.opennebula.org/issues/1341
- hátra van még az illesztése a 3.8-as Opennebula alá Sztaki cloud-ba.

Üzemeltetés

- Felügyelet:
  - Mérések, riasztások
    - Icinga bevezetése
      - Jogosultság kezelés VO alapján → Gyufi bővebben
  - Support levelezési lista
  - Hibajegykezelés
- Konfiguráció menedzsment
- Hálózati hozzáférés szabályozása
  - Csak VPN-ből érhetők el a virtuális gépek
  - Különböző route-map szabályok

Összefoglalás

- Rengeteg jelenlegi és jövőbeni fejlesztési feladat
  - Beküldött patchek, fejlesztések az Opennebula számára
- Sok gond a hálózati konfigurációval
  - Teljesítmény problémák, sokszori átkábelezés
- Opennebula alapú felhő
  - VO kezelés saját fejlesztés, de megosztjuk a 3.8 bevezetése után a közösséggel
- Számos diszk alrendszer teszt
  - Shared lvm, cgroups, live migráció
- Diplomamunka és féléves feladatok
- Az alfa szolgáltatás 2012.10.24-én elindult
- A projektről bővebben : http://cloud.sztaki.hu/

Kérdések?

Köszönöm a figyelmet!

ormos.pal@sztaki.mta.hu