
# Pacemaker alapú HA szervervirtualizáció

Wágner Ferenc
KIFÜ, NIIF Program

Networkshop 2017

# Vázlat

# Vázlat

- Történet
  - Xen HA
  - Majdnem cloud
  - Private Cloud
- Mit is csináltunk?
  - Nem cloud
  - Erősségek

# Vázlat

- Történet
  - Xen HA
  - Majdnem cloud
  - Private Cloud
- Mit is csináltunk?
  - Nem cloud
  - Erősségek
- Beleszaladások

# Vázlat

- Történet
  - Xen HA
  - Majdnem cloud
  - Private Cloud
- Mit is csináltunk?
  - Nem cloud
  - Erősségek
- Beleszaladások
- Tervek

# Történet

# Xen HA

Vitéz Gábor kollégámtól vettem át 2006-ban:

- hardver: két pár Dell PE2650
- memória: 6 GB gépenként
- storage: AoE és FC, később iSCSI
- hálózat: Linux bridge + VLAN
- virtualizáció
  - Xen paravirtualizáció a CPU lehetőségei miatt
  - `xend` és `xm` eszközökkel
  - kernel/initrd kezelése nehézkes (PV GRUB nem hivatalos)
- VM konfig: független helyi fájlok (verziókövetéssel)
- erőforráskezelés: node szintű (Heartbeat 1)

```
xen1-ha 10.253.2.6 xenraid LVM::xenimages xendomains
```

**Elég jól működött**

ahhoz, hogy népszerűvé váljon

- fizikai vasakhoz képest
  könnyű kezelhetőség

**Elég jól működött**
ahhoz, hogy népszerűvé váljon

- fizikai vasakhoz képest könnyű kezelhetőség

**Hamar ki is nőttük,**
pedig többnyire 128-256 MB-os VM-eket használtunk, akár 15-öt egy hoston.

# 2013: majdnem cloud

Az NIIF Cloud projekt cloud célra dedikált hardvere:

4 darab

- 16 core
- 128 GB
- dupla 10 GE offload

szerver + iSCSI storage.

# 2013: majdnem cloud

Az NIIF Cloud projekt cloud célra dedikált hardvere:
4 darab

- 16 core
- 128 GB
- dupla 10 GE offload

szerver + iSCSI storage.

- Az én feladatom a cLVM storage réteg kialakítása volt (az Eternus iSCSI target nehézkes kezelésének kiváltására)

# 2013: majdnem cloud

Az NIIF Cloud projekt cloud célra dedikált hardvere:
4 darab

- 16 core
- 128 GB
- dupla 10 GE offload

szerver + iSCSI storage.

- Az én feladatom a cLVM storage réteg kialakítása volt (az Eternus iSCSI target nehézkes kezelésének kiváltására) ✓

# 2013: majdnem cloud

Az NIIF Cloud projekt cloud célra dedikált hardvere:
4 darab

- 16 core
- 128 GB
- dupla 10 GE offload

szerver + iSCSI storage.

- Az én feladatom a cLVM storage réteg kialakítása volt (az Eternus iSCSI target nehézkes kezelésének kiváltására) ✓
- a virtualizációs réteg (libvirt + KVM) is működött már ✓

# 2013: majdnem cloud

Az NIIF Cloud projekt cloud célra dedikált hardvere:
4 darab

- 16 core
- 128 GB
- dupla 10 GE offload

szerver + iSCSI storage.

- Az én feladatom a cLVM storage réteg kialakítása volt (az Eternus iSCSI target nehézkes kezelésének kiváltására) ✓
- a virtualizációs réteg (libvirt + KVM) is működött már ✓
- de a cloud middleware nem állt jól

amikor váratlanul megnyertük a Sulinetet.

# Menekülőjárat

Tulajdonképpen csak egy erőforráskezelő hiányzik!

# Menekülőjárat

Tulajdonképpen csak egy erőforráskezelő hiányzik!

$\Downarrow$

`apt-get install pacemaker`

# Menekülőjárat

Tulajdonképpen csak egy erőforráskezelő hiányzik!

$\Downarrow$

`apt-get install pacemaker`

- A Heartbeat projekt továbbfejlesztett erőforráskezelője
- Corosync már volt a cLVM (DLM) miatt
- Csak a VM RA scripteket kell megírni!

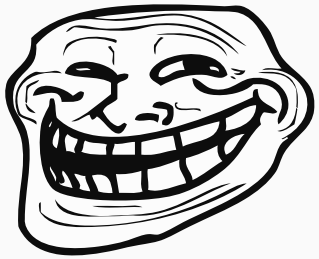
# Menekülőjárat

Tulajdonképpen csak egy erőforráskezelő hiányzik!

$\Downarrow$

`apt-get install pacemaker`

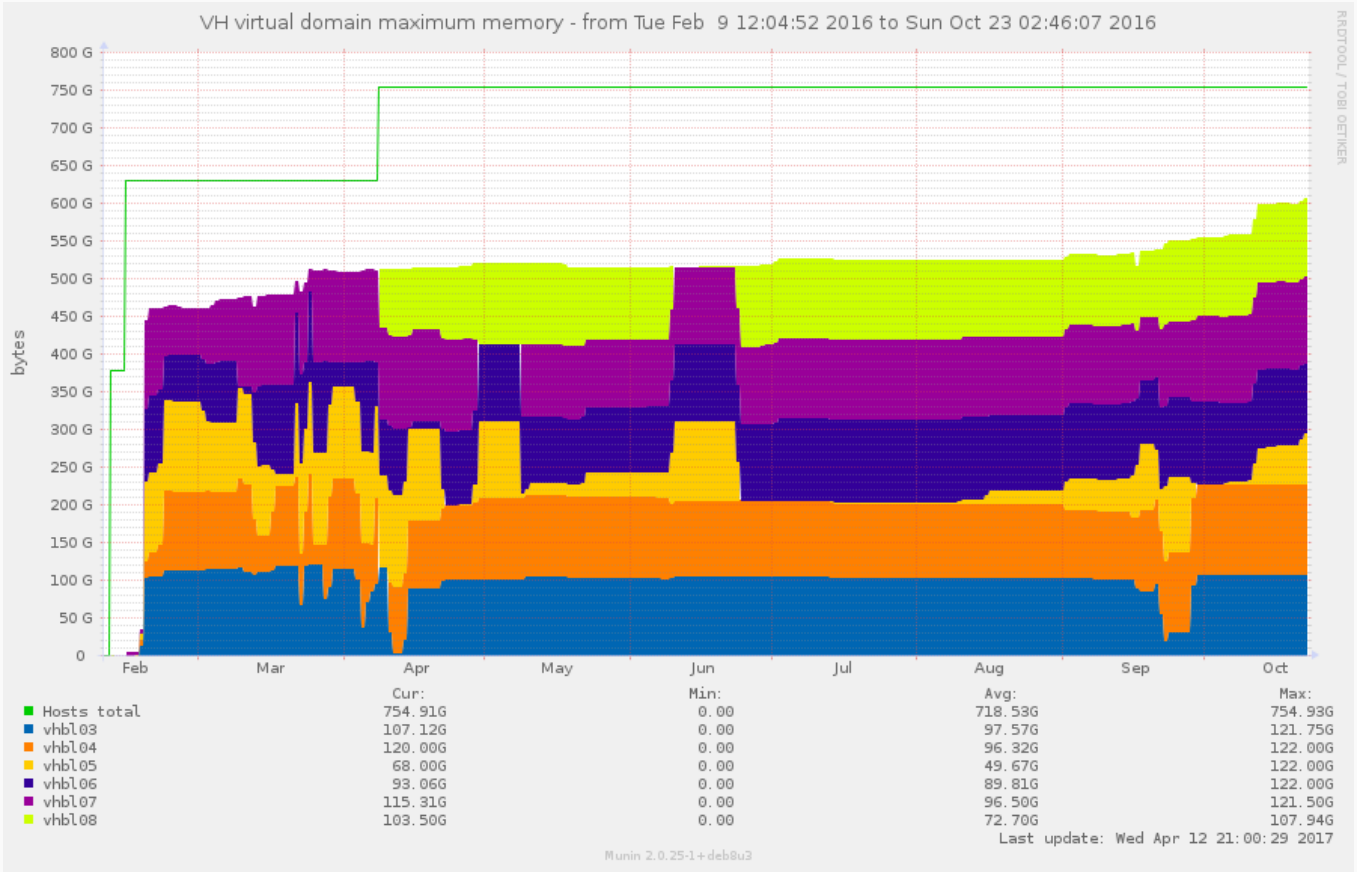
- A Heartbeat projekt továbbfejlesztett erőforráskezelője
- Corosync már volt a cLVM (DLM) miatt
- Csak a VM RA scripteket kell megírni!



# 2016: Private Cloud

Célzott bladecenter beszerzés

- alapvetően azonos specifikációjú gépek, csak több
- 10 GE Intel (`ixgbe`) helyett Emulex (`be2iscsi`)
- többlépcsős bővítés ($3 \rightarrow 5 \rightarrow 6$ node):



# Mit is csináltunk?

## Ez NEM cloud

- nincs felhasználói
  - storage/CPU pool
  - VM indítás
- nincsenek *tenantok*
- parancssoros felület (SSH)

```
$ sudo vm pws.ftp console
$ sudo vm elm     reset
$ sudo vm birch   ctrlaltdel
$ sudo vm dwdm    vnc
$ sudo vm niifidp sysrq c
$ wc -l /usr/sbin/vm
175 /usr/sbin/vm
```

# Hogyan működik?

```
# lvcreate -n gum -L 30g vm
# lvcreate -n gum -L 300g data /dev/mapper/data-ast-1
# mk_debian_installer --di=./stretch /dev/vm/gum
$ create_domxml --name gum --cpus 16 --mem 8g --ovs vh \
  --iface 150 --disk /dev/vm/gum /dev/data/gum >gum.xml
# domadd gum.xml
```

# Hogyan működik?

```
# lvcreate -n gum -L 30g vm
# lvcreate -n gum -L 300g data /dev/mapper/data-ast-1
# mk_debian_installer --di=./stretch /dev/vm/gum
$ create_domxml --name gum --cpus 16 --mem 8g --ovs vh \
  --iface 150 --disk /dev/vm/gum /dev/data/gum >gum.xml
# domadd gum.xml

$ wc -l [...]
  292 /usr/sbin/mk_debian_installer
  232 /usr/bin/create_domxml
  148 /usr/sbin/fixup_domxml
   72 /usr/sbin/domadd
```

```
<domain type="kvm">
  <name>gum</name>
  <uuid>auto</uuid>
  <memory unit="GiB">8</memory>
  <vcpu placement="static">16</vcpu>
  <os><type>hvm</type></os>
  <features><acpi/></features>
  <clock offset="utc"/>
  <devices>
    <emulator>/usr/bin/kvm</emulator>
    <disk type="block" device="disk">
      <driver name="qemu" type="raw" cache="none"/>
      <source dev="/dev/vm/gum"/>
      <target dev="vda"/>
    </disk>
    <disk type="block" device="disk">
      <driver name="qemu" type="raw" cache="none"/>
      <source dev="/dev/data/gum"/>
      <target dev="vdb"/>
    </disk>
    <interface type="bridge">
      <mac address="auto"/>
      <source bridge="vh"/>
      <virtualport type="openvswitch"/>
      <vlan><tag id="150"/></vlan>
      <target dev="gum.150"/>
      <model type="virtio"/>
    </interface>
    <console type="pty">
      <target type="serial"/>
    </console>
  </devices>
</domain>
```

gum.xml

```
<domain type="kvm">
  <name>gum</name>
  <uuid>auto</uuid>
  <memory unit="GiB">8</memory>
  <vcpu placement="static">16</vcpu>
  <os><type>hvm</type></os>
  <features><acpi/></features>
  <clock offset="utc"/>
  <devices>
    <emulator>/usr/bin/kvm</emulator>
    <disk type="block" device="disk">
      <driver name="qemu" type="raw" cache="none"/>
      <source dev="/dev/vm/gum"/>
      <target dev="vda"/>
    </disk>
    <disk type="block" device="disk">
      <driver name="qemu" type="raw" cache="none"/>
      <source dev="/dev/data/gum"/>
      <target dev="vdb"/>
    </disk>
    <interface type="bridge">
      <mac address="auto"/>
      <source bridge="vh"/>
      <virtualport type="openvswitch"/>
      <vlan><tag id="150"/></vlan>
      <target dev="gum.150"/>
      <model type="virtio"/>
    </interface>
    <console type="pty">
      <target type="serial"/>
    </console>
  </devices>
</domain>
```

gum.xml

# Ganeti alternatíva

Ezt akkor még nem tudtam, de kicsit utánanézve:

**Ganeti**

- KVM és
- DRBD

technológiákhoz
integrált cluster

# Ganeti alternatíva

Ezt akkor még nem tudtam, de kicsit utánanézve:

**Ganeti**

- KVM és
- DRBD

technológiákhoz integrált cluster

**Private Cloud**

- libvirt + Open vSwitch alapú
- storage-független
- Pacemaker-vezérelt

azaz standard technológiákra támaszkodó scriptgyűjtemény

# Rugalmas hálózati konfiguráció

- a Victor Hugo LAN (HBONE központ) része
- multicast, IPv6, VLAN trunking nem probléma

# Rugalmas kényszerek az erőforrások elhelyezésére

- társítás más erőforrással:

```
<rsc_colocation rsc="vm-bolha" score="-150" with-rsc="vm-cirkusz"/>
```

# Rugalmas kényszerek az erőforrások elhelyezésére

- társítás más erőforrással:

```
<rsc_colocation rsc="vm-bolha" score="-150" with-rsc="vm-cirkusz"/>
```

- csatolás például USB eszközhöz (node-hoz):

```
<rsc_location rsc="vm-mdsigner">
  <rule score="-INFINITY">
    <expression attribute="#uname" operation="ne" value="vhbl03"/>
  </rule>
</rsc_location>
```

# Rugalmas kényszerek az erőforrások elhelyezésére

- társítás más erőforrással:

```
<rsc_colocation rsc="vm-bolha" score="-150" with-rsc="vm-cirkusz"/>
```

- csatolás például USB eszközhöz (node-hoz):

```
<rsc_location rsc="vm-mdsigner">
  <rule score="-INFINITY">
    <expression attribute="#uname" operation="ne" value="vhbl03"/>
  </rule>
</rsc_location>
```

- kapacitás (tipikusan memória) függvényében

```
<primitive id="vm-fir" type="TransientDomain">
  <instance_attributes> [...] </instance_attributes>
  <utilization>
    <nvpair name="memoryMiB" value="1024"/>
  </utilization>
</primitive>
```

- Erőforrások priorizálására

```
# crm_resource --meta --resource=vm-elm --set-parameter=priority \
  --parameter-value=10
```

- Erőforrások priorizálására

```
# crm_resource --meta --resource=vm-elm --set-parameter=priority \
  --parameter-value=10
```

- sorrendezésére (függőségekre)
  - komoly adminisztrációs teher
  - a csökkenő párhuzamosság megéri (systemd fscks fail)

- Erőforrások priorizálására

```
# crm_resource --meta --resource=vm-elm --set-parameter=priority \
  --parameter-value=10
```

- sorrendezésére (függőségekre)
  - komoly adminisztrációs teher
  - a csökkenő párhuzamosság megéri (systemd fscks fail)
  - nagyon egyszerű konfiguráció:

```
<channel type="unix">
  <target type="virtio" name="startup_signal"/>
</channel>
```

- Erőforrások priorizálására

```
# crm_resource --meta --resource=vm-elm --set-parameter=priority \
  --parameter-value=10
```

- sorrendezésére (függőségekre)
  - komoly adminisztrációs teher
  - a csökkenő párhuzamosság megéri (systemd fscks fail)
  - nagyon egyszerű konfiguráció:

```
<channel type="unix">
  <target type="virtio" name="startup_signal"/>
</channel>
```

  - host kód:

```
socat UNIX-CONNECT:"$signal_socket" \
      SYSTEM:'read msg id && [ $msg = ready ] && echo ack $id'
```

- Erőforrások priorizálására

```
# crm_resource --meta --resource=vm-elm --set-parameter=priority \
  --parameter-value=10
```

- sorrendezésére (függőségekre)
  - komoly adminisztrációs teher
  - a csökkenő párhuzamosság megéri (systemd fscks fail)
  - nagyon egyszerű konfiguráció:

```
<channel type="unix">
  <target type="virtio" name="startup_signal"/>
</channel>
```

  - host kód:

```
socat UNIX-CONNECT:"$signal_socket" \
      SYSTEM:'read msg id && [ $msg = ready ] && echo ack $id'
```

  - és kliens kód:

```
port=/dev/virtio-ports/startup_signal
msg=foobar
reply="$(echo "ready $msg" | socat STDIO "$port")"
[ "($reply)" = "(ack $msg)" ]
```

# Teljesen szétválnak a felelősségek

- storage $\Rightarrow$ systemd + LVM
- hálózat $\Rightarrow$ Open vSwitch
- virtualizáció $\Rightarrow$ libvirt: KVM/Xen/LXC, akár rugalmasan
- magas rendelkezésre állás + erőforrás-kezelés $\Rightarrow$ Pacemaker
- felhasználói eszközök $\Rightarrow$ saját scriptek

Minden állapot a Pacemaker konfigurációjában van

# Teljesen szétválnak a felelősségek

- storage $\Rightarrow$ systemd + LVM
- hálózat $\Rightarrow$ Open vSwitch
- virtualizáció $\Rightarrow$ libvirt: KVM/Xen/LXC, akár rugalmasan
- magas rendelkezésre állás + erőforrás-kezelés $\Rightarrow$ Pacemaker
- felhasználói eszközök $\Rightarrow$ saját scriptek

Minden állapot a Pacemaker konfigurációjában van

$\Downarrow$

Homogén a cluster

```
$ wc -l [...]
  355 /usr/lib/ocf/resource.d/niif/TransientDomain
  557 /usr/lib/ocf/resource.d/heartbeat/VirtualDomain
```

# Beleszaladások

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` $\Rightarrow$ segfault

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` $\Rightarrow$ segfault
  - `destroy` $\Rightarrow$ segfault (ritkábban)

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` $\Rightarrow$ segfault
  - `destroy` $\Rightarrow$ segfault (ritkábban)
- libvirt 1.2 (jessie) hiba csak egy Windows VM-et érint

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` ⇒ segfault
  - `destroy` ⇒ segfault (ritkábban)
- libvirt 1.2 (jessie) hiba csak egy Windows VM-et érint
- jessie upgrade: nincs Pacemaker csomag, de support kell!
  - újraélesztettük a Debian-HA csapatot
    - systemd integráció
    - build system (hardening, reproducibility)
    - kisebb hibakezelési, dokumentációs, puffertúlírás javítások

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` ⇒ segfault
  - `destroy` ⇒ segfault (ritkábban)
- libvirt 1.2 (jessie) hiba csak egy Windows VM-et érint
- jessie upgrade: nincs Pacemaker csomag, de support kell!
  - újraélesztettük a Debian-HA csapatot
    - systemd integráció
    - build system (hardening, reproducibility)
    - kisebb hibakezelési, dokumentációs, puffertúlírás javítások
  - „újraélesztettük” DLM projektet is

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` ⇒ segfault
  - `destroy` ⇒ segfault (ritkábban)
- libvirt 1.2 (jessie) hiba csak egy Windows VM-et érint
- jessie upgrade: nincs Pacemaker csomag, de support kell!
  - újraélesztettük a Debian-HA csapatot
    - systemd integráció
    - build system (hardening, reproducibility)
    - kisebb hibakezelési, dokumentációs, puffertúlírás javítások
  - „újraélesztettük” DLM projektet is

```
  if (daemon_quit && list_empty(&lockspaces))
+         rv = 0;
          goto out;
```

# Beleszaladások

- libvirt 0.9 (wheezy) QEMU migrációs sávszélesség: 32 MB/s
- libvirt 0.9 (wheezy) hibák
  - aktív konzol mellett `destroy` ⇒ segfault
  - `destroy` ⇒ segfault (ritkábban)
- libvirt 1.2 (jessie) hiba csak egy Windows VM-et érint
- jessie upgrade: nincs Pacemaker csomag, de support kell!
  - újraélesztettük a Debian-HA csapatot
    - systemd integráció
    - build system (hardening, reproducibility)
    - kisebb hibakezelési, dokumentációs, puffertúlírás javítások
  - „újraélesztettük” DLM projektet is

```
  if (daemon_quit && list_empty(&lockspaces))
+         rv = 0;
          goto out;
```

  - cmirrord megjelent, de nem integrált

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát
- iSCSI boot setup

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát
- iSCSI boot setup
- bővítés: DX200 tiered storage

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát
- iSCSI boot setup
- bővítés: DX200 tiered storage
- DLM fencing

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát
- iSCSI boot setup
- bővítés: DX200 tiered storage
- DLM fencing
- iRMC firmware upgrade

- hardver upgrade: `be2iscsi` pár hetente kifagy
  - a STONITH kiválóan működik
  - Linux 4.6rc1 + új firmware javítja az Emulex hibát
  - de kitömi swapet (pls. bisect)
  - és migráció közben elrontja az órát
- iSCSI boot setup
- bővítés: DX200 tiered storage
- DLM fencing
- iRMC firmware upgrade
- Munin CPU plugin hiba

# Tervek

# Tervek

- folyamatos konzol log
- NUMA támogatás
- Ceph támogatás

Köszönöm a figyelmet!