Molnár Péter
Debreceni Egyetem Egyetemi és Nemzeti Könyvtár Informatika osztály

IT infrastruktúra egy modern egyetemi könyvtárban

Kivonat:

Egy mai, modern egyetemi könyvtár működésének alapját képezi a megbízható, nagy teljesítményű informatikai rendszer. A Debreceni Egyetem Egyetemi és Nemzeti Könyvtár a közelmúltban fejezte be informatikai infrastruktúrájának jelentős bővítését. Komoly előrelépés történt, mely már a kezdeti tervezéskor komoly szemléletváltozást igényelt: teljes átállást hajtottunk végre egy új szerverparkra. Törekedtünk a minden szempontból kiegyensúlyozott kiszolgálói infrastruktúra megteremtésére, mert lehetővé akartuk tenni az erőforrások hatékonyabb kihasználását, tartalék rendszerekre való gyors átállást és ezzel együtt a felhasználói elégedettség növelését.

Az előadás során áttekintést nyújtunk a követelmény-feltárástól az üzemeltetésig terjedő időszak fontosabb mérföldköveiről és kisebb-nagyobb buktatóiról. Bemutatjuk az elkészült infrastruktúra komponenseit, majd felvázoljuk a további fejlesztés lehetséges irányait. Összevetjük, mit nyújt a rendszer az elvárásokhoz viszonyítva és hogyan képes azt a tapasztalatok alapján megvalósítani. Megosztunk néhány gyakorlati tapasztalatot, mely segít egy hasonló rendszer felépítésében, fenntartásában.

Bemutatásra kerülnek az évek során összegyűlt követelmények és az azokra adott koncepcionális, s a megvalósítást érintő válaszok. Terveink szerint komoly szerepet szántunk egy több, azonos kiszolgálóból álló gépcsoportnak, valamint a hozzájuk kapcsolódó közös, osztott tárolónak, melyeken virtualizációs környezetet képzeltünk el. Több technológiát és szállító ajánlatát is alaposan megvizsgáltunk, míg végül összeállt a klaszterezett virtuális környezet és a SAN koncepciója. Követelményeink értékelése után Xen, Pacemaker alapokat építettünk Debian Lenny Linux-okkal és EMC storage-al. Az új rendszerben lehetőség nyílik bármely virtuális gép bármely fizikai gépen történő indítására, menet közbeni mozgatására. A klaszter tagjai jól definiálható szabályok alapján átvehetik egymás szerepét. Bizonytalan működés esetén önműködően kizárható a hibás node. Bár kész megoldások vásárolhatók kifejezetten ilyen célra, a mi rendszerünk előnye a nyílt forráskódú szoftverek előnyben részesítése és a könnyű átjárhatóság egyéb virtualizációs platformok felé.

Előadás:

Idén januárban fejeztünk be egy nagyarányú szerverpark bővítést könyvtárunkban. Gyakorlatilag a teljes kiszolgálói eszközbázist lecseréltük vagy jelentősen kibővítettük.

Ebben az előadásban a bevezetésre került klaszterünket mutatom be.

A kiinduló helyzet a teljesen hagyományos, szinte kizárólag fizikai szerverekre támaszkodó informatikai rendszer volt. A gépek több beszerzésből származtak, különböző paraméterekkel rendelkeztek. Jellemzően akkora erőforrással rendelkeztek, hogy egy nagyobb és néhány kisebb feladatot tudtunk rájuk bízni.

Az új eszközök beszerzése előtt sokat gondolkodtunk, hogy mit vásároljunk. Mi egy nagykönyvtár vagyunk: az informatika osztály több bevált szolgáltatást nyújt mind az olvasóknak, mind a könyvtárnak. Az egyik legnagyobb problémánk a szűkös memória kapacitás, a másik pedig a központi adattárolás kérdése volt. Számítási teljesítmény-igényünk folyamatosan növekszik, de még egy ideig bizonyosan lett volna felhasználható tartalékunk – de ezeket a tartalékokat csak a szolgáltatások összevonásával, a letisztultság feláldozása árán lettünk volna képesek kihasználni.

Leegyszerűsítve két út állt előttünk: az új kiszolgálók beszerzésekor a hagyományosabb és – felépítés szempontjából – egyszerűbb utat választjuk: erősebb gépeket vásárolunk több háttértárral

és „helyben” építkezünk. A másik út szerint alapjaitól újraszervezzük a vasakat és két új réteg kerül a felépítménybe: központosított adattárolás és virtualizáció.

Döntésünk meghozatala előtt egy listát állítottunk össze, ahova összeírtuk, mit szeretnénk. Jellemző volt, hogy:

- egy-egy vascsere – esetleg hiba – sok munkával járt (operációs rendszer és kiszolgáló szoftverek újratelepítése),
- az operációs rendszer naprakészen tartása nagy körültekintést igényelt, mert
- a szolgáltatások a gépek teljesítménye szerint csoportosultak (pl. web- és mentőszerver egy gépen, mert ott állt rendelkezésre a szükséges meretű háttértár)
- többször kellett erőforrás hiány miatt átszervezéseket végrehajtani.

További szempontok, melyeket figyelembe vettünk:

- igényeltük a minél magasabb szintű központosított rendszer-adminisztrációt,
- hardvererőforrások könnyű, gyors átszervezhetőségének lehetőségét,
- hiba esetén gyorsabb reagálást és helyreállítást (kézzel könnyen mozgatható virtuális gépek),
- fejlesztői és teszt környezet kialakítását (homokozót),
- ki kívántuk használni a második kiszolgálótermünk által nyújtott előnyöket (távoli mentések, tartalék kiszolgálók és hálózati útvonalak).

Miután összegyűjtöttük, nagyjából mit szeretnénk, elkezdtünk konzultálni több megoldásszállítóval. Röviden összefoglalva a tapasztaltakat elmondható, két típusú szállítóval találkoztunk: „mi tudjuk, mi kell Önöknek” és „egy új szerverszoba leszállítása pusztán logisztika”. Tapasztalatokat és tanácsokat vártunk; azonban végül levontuk a következtetést: elértük azt a üzemméretet, ahol a „dobozos” és „kulcsrakész” megoldások szinte egyáltalán nem használhatók. Legtöbbször felmerült olyan igény, kérés, mely személyre szabást, módosítást igényelt. Végül úgy döntöttünk, teljesen felvállaljuk a rendszerintegrátori szerepet.

Az elvárások tükrében elkerülhetetlenné vált tehát a szolgáltatások és a vas kontrasztosabb függetlenítése és adataink megbízható tárolása. A kiszolgáló számítógépenkénti helyi merevlemezekről és adattárolásról áttértünk SAN-ra, a fizikai gépekhez kötött szolgáltatásokról virtualizációra váltottunk. Gondolkodtunk, hogy Fibre Channel, iSCSI, esetleg NFS alapú központi adattárolási megoldást válasszunk-e. Úgy találtuk, hogy iSCSI és NFS alapon csak nagy nehézséggel tudunk megfelelni a felmerülő teljesítmény igénynek, így Fibre Channel-t választottunk. Felállítottunk a rendszerrel szembeni elvárásokat:

- storage és SAN alapú adattárolás,
- Fibre Channel kapcsolat a SAN elemei között,
- többféle Linux és Windows verziót támogató virtualizáció,
- több csomópontból álló kiszolgáló klaszter,
- hibás csomópont esetén könnyű átállás maradék csomópontokra,
- nyílt forrású virtualizációs és klaszter szoftver,
- lehetőség szerinti legnagyobb gyártói függetlenség (no vendor lock-in).

A virtualizációs megoldásnak támogatnia kell többféle operációs rendszert és viszonylag rugalmasnak kell lennie hardverigény szempontból. Ha egy számítógép meghibásodik, könnyen át lehessen csoportosítani a rajta futó szolgáltatásokat a maradék gépekre (3 darab összesen).

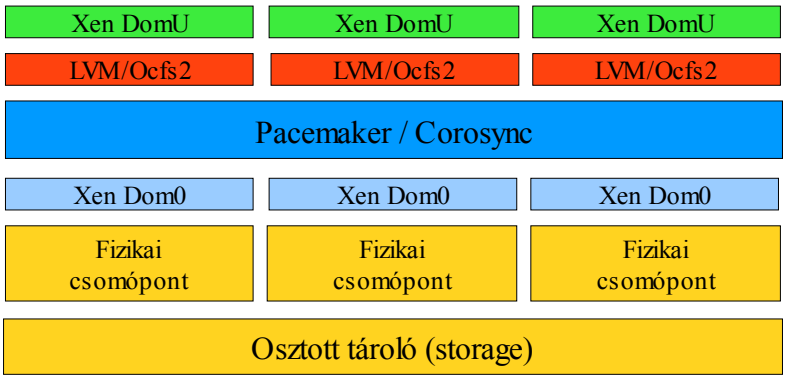
Végül az alábbi hardverelemek mellett döntöttünk:

- IBM x3650 M2 kiszolgálók (3 db)
- EMC Clariion CX4-240 storage (bruttó 39 TB)
- Brocade DS300B FC switch
- Cisco Sup32
- új rackszerkrény :)

A felhasznált szoftverek:

- Xen 3.2 (Qemu, GPLPV Windows meghajtók)
- Pacemaker/corosync
- Debian Linux 5 (Lenny)
- Ocfs2
- multipath-tools

Előnyben részesítjük a nyílt forrású megoldásokat, így állt össze a szoftvercsomag. A Pacemaker nyílt forrású klaszterszoftver. A Red Hat és a Novell, valamint a LinBit a fő fejlesztők. Képes hardver és alkalmazás szintű hibák kezelésére és helyreállítására. Szinte bármilyen erőforrást képes kezelni, ami szkriptekkel kezelhető, legyen az egy Apache vagy szerverek csoportja. Stratégiák állíthatók fel a gépek meghibásodása esetére. Képes az erőforrások indításának és leállításának sorrendjét kezelni. Szabályok alkalmazásával megadhatók, mely alkalmazások futhatnak/fussanak különböző vagy azonos csomóponton (http://www.clusterlabs.org/). Nálunk a Xen virtuális gépek és a fizikai csomopóntok kezelése a szoftver feladata. Az alábbi ábra az architekturális felépítést ábrázolja:



Tehát közös, osztott tárolón tartunk minden adatot. A gépek is innen indulnak. A Xen és Debian Linux Dom0 elindulása után azonnal indul a Pacemaker. A klaszterszoftver multicast üzenetekkel kommunikál a csomópontok között és igyekszik az elindított erőforrásokat elosztani a hardverek között. Először az ocfs2 klaszter állományrendszer indul el. Ezen tároljuk a virtuális gépek konfigurációs állományait, kerneleket. A szoftver hasznos képessége, hogy támogatja a *fencing*-et, azaz eltávolítja azt a fizikai csomópontot (és kikapcsolja), melyen egy erőforrás szabályos megállítására tett kísérlet kudarcot vall. Az eltávolítás az IBM gépek távoli menedzsment kártyája közreműködésével történik: a Pacemaker a bizonytalan gép menedzsment interfészére csatlakozik, majd kikapcsolja a gépet. A DomU-kat is a Pacemaker kezeli a Xen *auto* szkriptek helyett. Bekapcsoltuk a menet közbeni migrációs szolgáltatást a Xen-en, így a klaszter képes a virtuális gépeket igény szerint másik fizikai csomópontra mozgatni. Üzemszerű körülmények között ez valamilyen karbantartás alkalmával nagyon hasznos, hiszen a karbantartásra váró csomópontot készenléti állapotba (*standby*) helyezhetjük, a rajta futó DomU-k átköltöznek másik csomópontokra. A munka elvégzése után *online* állapotba helyezett csomópontra visszakerülhetnek az erőforrások.

Könnyűvé válik további csomópontok üzembe helyezése, hiszen elegendő a SAN és Ethernet kapcsolatokat beállítani, majd a Pacemaker telepítését követően azonnal költözhetnek menet közben az szolgáltatást nyújtó szoftverek.

A kiszolgálók rendelésekor olyan összeállításban kértük a gépeket, hogy tartalmazzon a SAN kétszeres összeköttetéseihez két darab Fibre Channel kártyát. Merevlemezt nem kértünk, mert a gépek a storage-ról indulnak. Az alapkiépítéshez képest kértünk mégegy processzort és 48 GiB memóriát. Így összesen 8 mag/16 szál és 54 GiB memória áll rendelkezésre gépenként. Három ilyen vasunk van jelenleg. A storage-ot egy fióknyi, 15 darab FC merevlemezzel kértük, ebből 5 darab 15000 fordulat/perc és 10 db 10000 fordulat/perc sebességgel pörög. További két fiókot kértünk 15-15 darab 5400 fordulat/perc sebességű SATA lemezekkel. Szívesen kipróbáltuk volna az SSD lemezeket, de erre anyagi okok miatt nem volt lehetőségünk.

A bemutatott komponensekből álló környezet teljesíti a követelménylistánkban felsorolt pontok mindegyikét: hibatűrő, gyors és rugalmasan konfigurálható rendszert kaptunk, mely képes több operációs rendszert párhuzamosan futtatni. Később egyszerűen bővíthető és nem alakult ki túlságosan nagymértékű vagy zavaró gyártói, beszállítói függés. Nagyrészt nyílt forrású komponensekből építkezünk. A kezdeti elvárásokhoz képest többletet hoz az önműködő erőforrás-kezelés és a fizikai csomópontok kényszerleállítása (*fencing*).

Az eddig eltelt idő alatt összegyűlt tapasztalatok alapján a klaszterünkkel teljesen elégedettek vagyunk. A gépek hozzák az elvárható teljesítményt és megbízhatóságot. A storage és a SAN elemei is gyorsak és üzembiztosak. A rendszer egésze olajozott működést mutat. Látványos sebességbeli javulás tapasztalható a jobb lemez alrendszernek köszönhetően. Szolgáltatásaink minősége annak köszönhetően javult, hogy több vassal gazdálkodunk, így finomabban elválaszthatók a futó szoftverek: ha egy processzus dől, akkor kevesebbet vihet magával.

Terveink között szerepel igény szerint újabb kiszolgáló számítógépek rendszerbe kapcsolása, további lemezdobozok beszerzése. A korszerűsítés jegyében szükségessé váló Xen és Debian upgrade tervezése megkezdődött. Elkezdtük egy olyan saját építésű storage terveinek kidolgozását, mely a SAN-ba Fibre Channel kapcsolaton keresztül illeszthető be. Elsődlegesen biztonsági mentések, és átmeneti adatok tárolására szánjuk. Vélekedésünk szerint egy iSCSI alapú megoldás helyett érdemes Fibre Channel-ben gondolkodni a teljesítményt szem előtt tartva.

Két érdekes nehézségbe ütköztünk a rendszer tervezése és telepítése során. A Xen 3.2-es verziója nem nyújt támogatást a gazda gép USB eszközeinek vendég általi használatához. Az usbip projekt (http://usbip.sourceforge.net/) lett a segítségünkre. Üzembe helyeztünk egy USB eszköz megosztásra használt gépet, melyről a Xen/Qemu-ban futó Windows Server számára tudunk ilyen csatolójú perifériát illeszteni. Másik érdekesség a Xen telepítésekor bukkant fel. Miután beállítottuk az Ethernet trönk kapcsolatot a hálózati kártyákon, csak a natív VLAN-ban volt elérhető a gép. Kiderült, hogy a Broadcom hálózati vezérlők firmware-e nem várt módon viselkedik és a Xen *bridge* szkriptjeit meg kellett változtatni.