A Beowulf fejlődése

A Beowulf-géptelepek második nemzedéke egyetlen térben elhelyezkedő folyamatokat, vékony rabszolga-csomópontokat és grafikus segédprogramokat kínál, továbbá az alkalmazkodóképessége és a kezelhetősége is jobb.

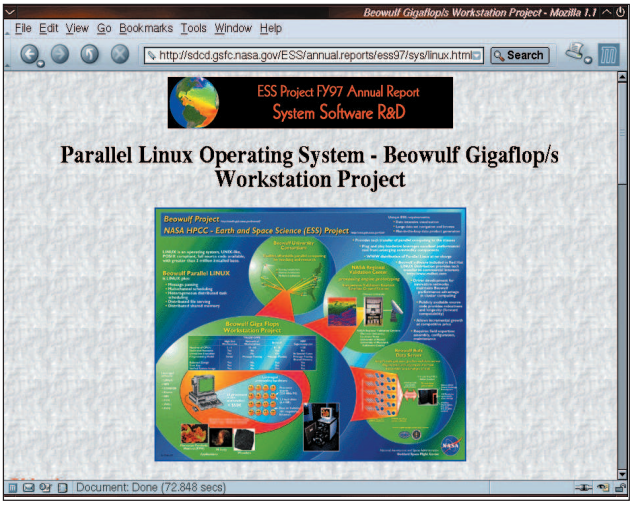
Képzeld el egy pillanatra, hogy gépkocsiddal betérsz egy benzinkúthoz, és azt mondod a kútkezelőnek: „Töltse tele, ellenőrizze az olajszintet és az ablaktörlőt, és adjon még 20 lóerőt, legyen szíves!” A benzinkutas nem lepődik meg a kérésen, hanem ezt válaszolja: „Összkerékmeghajtást is kér? Úgy hallottam, hóesés várható az éjjel.” Elgondolkozol egy másodpercre, majd beleegyezel, ugyanis az összkerékmeghajtás hasznos dolog.
Bárcsak ilyen könnyen alkalmazkodó autóink és Beowulf-géptelepeink lennének! A Beowulf 2 legfontosabb megkülönböztető sajátossága alkalmazkodóképessége – ha az igény növekszik, többletszámítási teljesítményt tud adni. A Beowulf alkalmazkodóvá válásának megértéséhez és értékeléséhez a először Beowulf 1-et kell megismernünk.

A Beowulf gyökerei

Mostanra már mindannyian tudjuk, hogy a Beowulf-géptelepek ötlete ***Donald Becker*** fejéből pattant ki, amikor a NASA Goddard intézetében dolgozott 1994-ben. A lényege az volt, hogy egyszerű számítógép-alkatrészekből felépített párhuzamos rendszer bizonyos feladatok esetén egy nagyságrendet javít az ár/teljesímény arányon. Az ötlet a valóságban is bevált, az első Beowulf-géptelep, a Wiglaf 1994 végén épült meg.
A Wiglafban 16 darab 66 MHz-es Intel 80486 processzor dolgozott, amelyeket később 100 MHz-es DX4-ekre cseréltek. A teljesítmény átlagosan 74 Mflops/s volt (74 millió lebegőpontos művelet másodpercenként). Három évvel később Becker és a CESDIS csapat elnyerte a tekintélyes Gordon Bell-díjat. A díjat azért a Pentium Pro processzorokat használó géptelepért adták, amelyet az 1996-os SuperComputing Conference-re építettek, és amely elérte a 2,1 Gflops/s (2,1 milliárd lebegőpontos művelet másodpercenként) teljesítményt. A Goddard-nál kifejlesztett programot ekkor már sok egyetemen és kutatóintézetben széles körben használták.

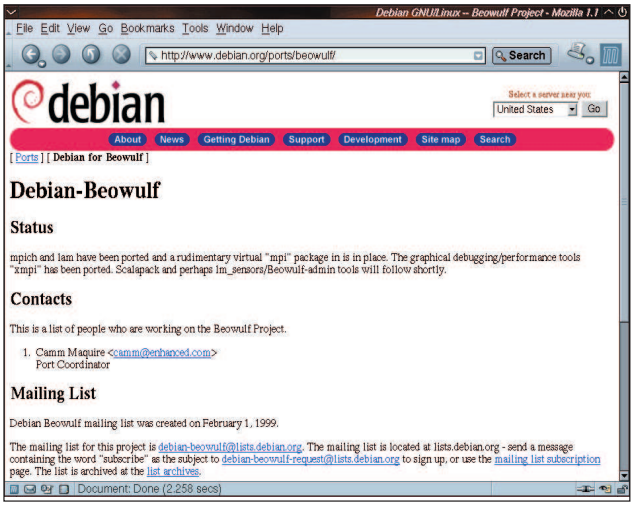
A Beowulfok első nemzedéke

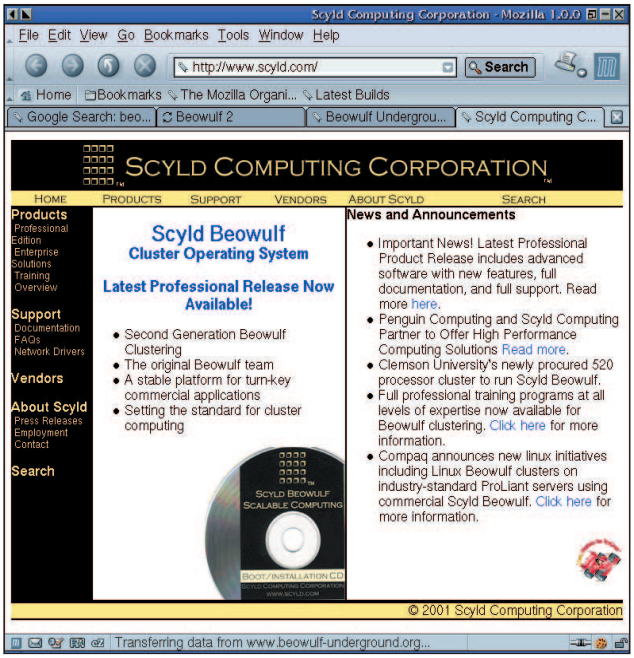
Az első nemzedékbe tartozó Beowulfok jellemző vonásai a következők voltak: közönséges számítógépek, nyílt forráskódú operációs rendszerek (például Linux vagy FreeBSD) és magánhálózaton elhelyezkedő számítócsomópontok. Ráadásul minden csomóponton teljes értékű operációs rendszer futott, és minden csomópontnak saját folyamattere volt.
Ezek az első nemzedékbeli Beowulfok az üzenettovábbításra külön programot használtak: vagy a PVM-et (párhuzamos virtuális gép), vagy az MPI-t (üzenettovábbító felület). Az adatcsere jellemző módja az üzeneteknek nagyteljesítményű géptelep számolócsomópontjai közti továbbítása volt.
Némi gond is akadt a Beowulfok első nemzedéke körül, amelyek legfőképpen abból következtek, hogy az új géptelepek kezelésére használt rendszerfelügyeleti eszközök fejlődése elmaradt a párhuzamos programokétól. Végtére is a Beowulf



nagyteljesítményű párhuzamos feladatokra lett kitalálva, és sokkal kevesebb figyelmet fordítottak a hordozható és megbízható rendszerfelügyeleti programok kifejlesztésére. A korai Beowulfokkal kapcsolatos nehézségek a következők voltak:

- Nehéz volt őket telepíteni: vagy a munkaigényes, minden csomópontot egyenként telepítő megközelítést választhattuk (ahol gyakran csúsztak be gépelési hibák), vagy a bonyolultabb, minden csomópontot egyszerre, a hálózaton keresztül telepítő módszert a PXE/TFTP/NFS/DHCP használatával – az összeset helyesen beállítani és egyszerre futtatni már önmagában is hőstettnek bizonyult.
- Telepítés után a Beowulfokat nehéz volt karbantartani. Gondoljunk csak egy közepesen nagy géptelepre, csomópontok tucatjaival vagy százaival. Mi történik, ha új Linux-rendszermag jelenik meg, mint például a 2.4-es az SMP-re kihegyezve? Ha a számolócsomópontokon az új rendszermagot szeretnéd futtatni, telepíteni kell a megfelelő helyre, majd közölni a LILO-val (vagy a kedvenc rendszerbetöltőddel) a változás tényét, mindezt több tucatszor vagy több százszor. A csomópontok frissítéséhez az `rsh` és az `rcp` programokat használták. Ezek a programok azt igénylik, hogy a számolócsomópontokon felhasználókat kezelő segédprogramok legyenek, továbbá biztonsági rések özönét nyitják meg.
- A géptelepet nehéz volt módosítani: a teljesítmény növelése új számolócsomópontok hozzáadásával csak a germán istenekhez történő buzgó fohászkodások közepette volt lehetséges. A csomópont hozzáadásához telepíteni kellett az operációs rendszert, frissíteni kellett a beállítóállományokat (sok kis trükkös fájlt), valamint a felhasználói területet a csomópontokon, és természetesen minden párhuzamos





számítást érintő kódot is, amely saját maga is beállítást igényel, elvégre használni is szeretnénk az új csomópontot, nem igaz?

- Az egész nem úgy nézett ki, mint egy számítógép. Sokkal inkább hasonlított független csomópontok halmazára, ahol mindegyik csomópont a maga dolgával törődött, és néha kellő ideig együttműködött a többiekkel, hogy egy párhuzamos számítási feladatot végezzen el.

Röviden, a Beowulf 1 sikeres volt a közönséges számítógépek teljesítményének kihasználásában, de még messze volt egy ipari erejű számítóeszköztől.
Az elmúlt egy évben a Rocks és az OSCAR géptelepes programterjesztései összegezték az eddigi Beowulf 1-fejlesztéseket (lásd a „Beowulf-tudatállapot” c. cikket a Linuxvilág 2002. júniusi és az „OSCAR-forradalom” c. cikket a Linuxvilág 2002. júliusi számában). De ha a beowulfos számítások bonyolultabbá válnak, a használatnak pedig egyszerűsödnie kell, akkor rendkívül sok Linux-programozásra van szükség. Itt lép a képbe a Beowulf 2, a Beowulf következő nemzedéke.

A Beowulf második nemzedéke

A második nemzedék újítása, hogy a leggyakoribb hibákat okozó összetevőket kiküszöbölték, az új felépítés sokkal egyszerűbb és megbízhatóbb lett, mint az első nemzedékben. ***Don Becker*** műszaki igazgató vezetésével a Scyld Computing Corporation, valamint pár ember az eredeti NASA Beowulf-csapatból olyan jelentős áttörést ért el a Beowulf-módszerben, mint amilyen maga a Beowulf megjelenése volt 1994-ben. Továbbra is közönséges számítógépeket és üzenetküldő programokat használnak a Beowulf 2-ben, de jelentős módosítások történtek a csomópontok telepítésének és a folyamattér elosztásának területén.

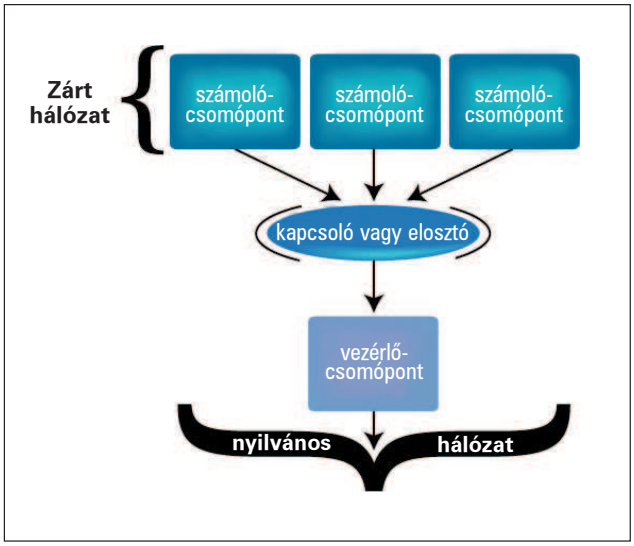
BProc

A Beowulf második nemzedékének lelke a BProc, ami az osztott Beowulf-folyamattér rövidítéséből kapta a nevét. A BProc fejlesztője ***Erik Arjan Hendriks***, a Los Alamos National Lab munkatársa. A BProc néhány rendszermag-módosításból és rendszerhívásból áll, egy folyamatnak egyik csomópontról a másikra történő áthelyezését ezek teszik lehetővé. A folyamat áthelyezése teljesen az alkalmazás ellenőrzése alatt áll – az alkalmazás saját maga határozhatja el, hogy másik csomópontra akar-e költözni, és ezt az `rfork` rendszerhívással kezdeményezheti. A folyamat a csatolt fájlkezelők nélkül költözik át, ami gyors és könnyű áttérést tesz lehetővé. Minden szükséges fájlt maga az alkalmazás nyit meg újra a célcsomóponton, minden hatalom az alkalmazás folyamatáé.
Természetesen értelmetlen volna egy folyamatot egy másik csomópontra áthelyezni, ha a távoli folyamatot nem lehet kezelni. A BProc ezt úgy oldja meg, hogy minden áthelyezett folyamathoz egy „szellemfolyamatot” helyez el a vezérlő csomópont folyamattáblájában. Ezek a szellemfolyamatok nem foglalnak memóriát a vezérlőcsomóponton, egyszerűen csak olyan bejegyzések, amelyek jelzéseket tudnak fogadni és képesek bizonyos műveleteket a távoli folyamat nevében végrehajtani. Például a szellemfolyamaton keresztül a távoli folyamat jelzéseket kaphat, beleértve a `SIGKILL` és `SIGSTOP` jelzéseket, továbbá gyermekfolyamatot is indíthat. Mivel a szellemfolyamatok a vezérlőcsomópont folyamattáblájában jelennek meg, a folyamatok állapotát kijelző segédeszközök az ismert módon működnek.
A BProc elegáns egyszerűsége messzire ható következményekkel jár. A legnyilvánvalóbb hatás, hogy a Beowulf-telep most látszólag egyetlen folyamattérrel rendelkezik, amelyet a vezérlőcsomópont irányít. Az egyetlen, az egész géptelepre kiterjedő folyamattér központosított felügyelettel egyetlen rendszer látomását kelti, mintha csak egyetlen számítógéppel lenne dolgunk. Ráadásul a BProc-nak nincs szüksége `rcp`-re és `rlogin`-ra a folyamatok kezeléséhez, hiszen a folyamatokat közvetlenül a vezérlőcsomópont vezérli. Ezeknek a parancsoknak a kiküszöbölése azt jelenti, hogy a számolócsomópontokon nincs szükség a felhasználók kezelésére, így jelentősen csökken az operációs rendszer mérete. A BProc futtatásához a számolócsomóponton mindössze néhány démon szükséges: a `bpslave` és a `sendstats`.

A Scyld megvalósítása

A Scyld teljesen átvette a BProc-ot, és olyan bővíthető géptelepeket építhetünk a segítségével, amely a számoló csomópontokon csak a BProc folyamat futásához elengedhetetlenül szükséges összetevőket hagyta meg. A végeredmény egy ultravékony számolócsomópont, amely a Linuxnak csak egy kis részét futtatja, épp annyit, amennyi a BProc-nak szükséges.

A BProc és az ultravékony Scyld-csomópontok előnyei összeadódnak, és ez nagy hatással van a géptelep kezelhetőségére. Két megkülönböztető jellemzője van a Scyld-terjesztésnek és a Beowulf 2 géptelepeknek. Az első, hogy géptelep bővíthető, egyszerűen lehet hozzáadni újabb csomópontokat. Mivel a csomópontok ultravékonyak, a telepítés annyiból áll csupán, hogy a rendszert a Scyld rendszermagjával el kell indítani,



Egy jellemző Beowulf-rendszer fizikai kialakítása

valamint a BProc vándorló folyamatait illetően fogadóképessé kell tenni. A második, hogy a vegyes változatok gondját kiküszöbölték. Az olyan géptelepeken, amelyekben a csomópontok teljes Linux-telepítéssel rendelkeznek, előfordulhat a változatok keveredése. Az idő előrehaladtával megeshet, hogy a programok frissítésekor egyes csomópontok nem működnek, akár frissítési hiba miatt, akár a programozó hibája miatt, ezért ahelyett, hogy minden csomóponton szigorúan ugyanaz a program futna, az egyes csomópontok között eltérések keletkeznek. Mivel a BProc futásához a csomópontokon csak nagyon kevés program futása szükséges, ezt a gondot gyakorlatilag elkerüljük.
Természetesen a folyamatok áthelyezésének képessége vékony csomópontokra önmagában semmit nem old meg. A Scyld a megoldás többi részét a különleges Scyld Beowulf-terjesztés részeként nyújtja, ami az alábbi sajátosságokkal bír:

- ***BeoMPI***: üzenettovábbító programkönyvtár, amely megfelel az MPI szabványnak, és az Argonne National Lab MPICH (MPI Chameleon) projektjének leszármazottja, amelyet a BProc programmal való együttműködés céljából továbbfejlesztettek.
- ***BeoSetup***: grafikus felület a számolócsomópontok BeoBoot rendszerindító lemezlenyomatainak elkészítéséhez.
- ***Beofdisk***: segédprogram a számítócsomópontok merevlemezének felosztásához.
- ***BeoStatus***: grafikus felület a géptelep állapotának figyeléséhez.

Nézzük meg, hogyan kell használni ezeket az eszközöket a Scyld Beowulf-géptelep felépítése közben!
Megveheted a Scyld Beowulf Professional Editiont (➲ http://www.scyld.com), amelyhez rendszerindításra alkalmas telepítő-CD, leírás és egyéves támogatás jár.
A Professional Edition látványos, és sok fejlett géptelepkezelő programot támogat, például a párhuzamos virtuális fájlrendszert (PVFS). A másik lehetőség a Scyld Basic Edition beszerzése, a CD a Linux Centralnál (➲ http://www.linuxcentral.com) 2,95 dollárba kerül. A Basic Editionből hiányzik néhány dolog, ami a Professional Editionben megtalálható: nincs leírás és támogatás. Mindkettőt használva építettem már géptelepet, nem volt semmi nehézség.
Fontos, hogy az *ábránkon* láthatóhoz hasonló Beowulf-elrendezést hozz létre, ez az általános Beowulf- (1 és 2) elrendezés. A vezérlőcsomópontban két hálókártya van, az egyik a nyilvános hálózat, a másik a számítócsomópontok zárt hálózata felé vezet. A Scyld Beowulf azt feltételezi, hogy a hálózatot úgy állítottad be, hogy az `eth0` vezet a nyilvános hálózathoz és az `eth1` a belső hálózathoz. A telepítés megkezdéséhez a Scyld CD-t helyezd a vezérlőcsomópont CD-meghajtójába, és kapcsold be a számítógépet.
A Scyld Beowulf telepítése gyakorlatilag megegyezik a Red Hat Linux telepítésével. A rendszerindító parancssorába írd be az `install` szót, ezzel indítható a vezérlőcsomópont telepítése. Ha megvárod, hogy a rendszerindítás magától folytatódjon, akkor alapértelmezés szerint a számolócsomópont telepítése indul el.
Lépkedj végig az egyszerű telepítési folyamaton, ahogyan a Red Hat Linux esetében tennéd. Ha először telepítesz géptelepet, azt javaslom (és a következőkben erről írok), hogy a Gnome-os telepítést válaszd a szöveges módú telepítő helyett. A Gnome-os telepítő választásával elérhetőek lesznek a menő grafikus Beo-segédprogramok, amelyek beépülnek a Gnome munkaasztali környezetbe, és géptelep telepítésének további részét jelentősen megkönnyítik.
Az `eth0` szokásos beállítása után jön az `eth1` beállítása a vezérlőcsomóponton és a számítócsomópontok IP-címeinek megadása. Ez a lépés jelenti az egyik kulcsfontosságú különbséget a Scyld és a Red Hat Linux telepítése között. A telepítőprogram kér egy IP-címet (pl.: 192.168.1.1) az `eth1` számára, és egy IP-címtartományt (pl.: 192.168.1.2 – 192.168.x) a számítócsomópontoknak. Ez elég egyszerű, de nem árt meggyőződni róla, hogy az IP-címtartomány elég nagy-e ahhoz, hogy minden számolócsomópontnak külön IP-címe legyen.
Hátravan még néhány lépés a telepítésből, például az X beállítása. Az egyszerűség kedvéért válaszd a grafikus bejelentkezést. Fejezd be a vezérlőcsomópont telepítését a rendszerindító lemez elkészítésével, távolítsd el a lemezeket a meghajtókból, és indítsd újra a vezérlőcsomópontot.
Lépj be rendszergazdaként. A Scyld által testreszabott Gnome-munkaasztal elindul, beleértve a BeoSetup és a BeoStatus programokat és a számolócsomópontok telepítésének leírását. A rendszerindításhoz minden számolócsomópontnak egy BeoBoot-lenyomat szükséges, amely lehet hajlékonylemezen vagy a Scyld CD-n. Jobban szeretem, ha minden csomóponthoz egy-egy hajlékonylemezt készítek, mint ha a CD-vel kell szaladgálni egyik géptől a másikig. A BeoBoot-lenyomatokat a BeoSetup segédprogram készíti el. A BeoSetup programban kattints a ***Node Floppy*** gombra, helyezz egy üres lemezt a meghajtóba, és a lemez elkészítéséhez kattints az ***OK*** gombra.
A folyamatot addig ismételd meg, amíg el nem készül a kellő számú lemez. Az indítólemezeket helyezd a számolócsomópontok meghajtójába, és kapcsold be a számítógépeket! Meglehetősen érdekes, ami ezután történik, bár a felhasználó nem láthatja (hacsak nem kötsz egy monitort a számolócsomópontra). Minden egyes számolócsomópont betölti a BeoBoot-le-

nyomatot, felismeri a hálózati eszközt, telepíti az eszközvezérlőket, és RARP-kéréseket küld ki. Ezekre a RARP-kérésekre a vezérlőcsomóponton figyelő Beoserv démon válaszol, IP-címet, rendszermagot és ramlemezt küld minden számítócsomópontnak. Külön nevet is adtak a folyamatnak, amely során a számolócsomópont feléleszti magát egy hajlékonylemezről betöltött minimális tudású rendszermaggal, amelyet ezután a végleges, sokkal bonyolultabb rendszermagra cserél le a vezérlőközpontból. A folyamat neve Two Kernel Monte. A számolócsomópont ezután újraindítja magát a végleges rendszermaggal, megismétli az eszközök felismerését és a RARP-kérést, majd felveszi a kapcsolatot a vezérlőcsomóponttal és a BProc részévé válik.
A Two Kernel Monte alatt a számolócsomópont hálózati kártyájának MAC-címe a ***BeoSetup Unknown Addresses*** (ismeretlen címek) ablakában helyezkedik el. A géptelephez úgy lehet őket hozzáadni, hogy a címeket kijelölöd és a középső ***Configured Nodes*** (beállított csomópontok) oszlopba húzod őket, majd megnyomod az ***Apply*** (alkalmaz) gombot. Miután a vezérlőcsomópont végzett a számolócsomópontok rendszerbe illesztésével, a csomópont mellett megjelenik az up címke. A csomópont állapota a BeoStatus programban is megjelenik.
A számolócsomópontok merevlemezét az alapértelmezett beállítás (***/etc/beowulf/fdisk***) szerint a következő két paranccsal lehet felosztani:

```
beofdisk -d
beofdisk -w
```

A -d kapcsolóval tudatjuk, hogy a ***/etc/beowulf/fdisk*** fájlban megadott alapértelmezett beállításokat használjuk, a -w végzi el a táblázatok kiírását a számolócsomópontokon. Ezután át kell írni a ***/etc/beowulf/fstab***} fájlt, hogy a csereterület (swap) és a / fájlrendszer az új lemezrészekre mutasson. Tedd megjegyzésbe a $RAMDISK sort a ***/etc/beowulf/fstab***-ban, ide volt befűzve a / fájlrendszer, mielőtt még a lemezt felosztottad volna. A következő két sorban add meg a csereterület és a / fájlrendszerek helyét a ***/dev/hda2*** és ***/dev/hda3*** eszközökön (a ***dev/hda1*** a rendszerindító lemezrésznek van fenntartva).
Ha a rendszert merevlemezről szeretnéd indítani, a Beoboot-lemezlenyomatot átírhatod a rendszerindító lemezrészre:

```
beoboot-install -a /dev/hda1
```

Ezután a ***/etc/beowulf/fstab*** fájlhoz hozzá kell adni egy sort:

```
/dev/hda1    beoboot    ext2    defaults    0 0
```

A változások érvényesítéséhez minden számolócsomópontot újra kell indítani:

```
bpctl -S all -s reboot
```

Ennél egyszerűbb nem is lehetne. A Beowulf 1-gyel ellentétben a Scyld Beowulf csak a vezérlőcsomópontra telepít teljes Linux-terjesztést. A számolócsomópontok merevlemezére semmi nem íródik a telepítés alatt, emiatt ezek ultravékonyak, könnyen karbantarthatók és gyorsan újraindíthatók lesznek.
A géptelep kipróbálásához futtathatod a nagyteljesítményű Linpack benchmark programot, amely része a terjesztésnek. A parancssorban add ki a linpack parancsot.
Kicsit látványosabb a Mandelbrot-halmaz kirajzoltatása, amelyet az mpi-mandel alkalmazással próbálhatunk ki. Ugyancsak része a terjesztésnek. Ha az mpi-mandel-t öt csomóponton szeretnéd elindítani, ezt írd a parancssorba:

```
NP=5 mpi-mandel
```

Mindent együttvéve az egyetlen folyamattér, a folyamatok gyors áthelyezésének lehetősége az alkalmazás segítségével, a vékony csomópontok és a grafikus segédprogramok, amelyekkel a Scyld géptelep felépíthető és figyelhető, olyan megoldást eredményeznek, amely a Beowulf 1-et teljességben, alkalmazkodóképességben és kezelhetőségben felülmúlja. Fenti kérdésünkre a válasz megerősítő, ehhez a géptelephez csakugyan további lóerőket lehet hozzáadni.

Köszönetnyilvánítás

A szerzők köszönetet mondanak ***Donald Becker***-nek, ***Tom Quinn***-nek és ***Rick Niles***-nek a Scyld Computing Corporationből, és ***Erik Arjan Hendriks***-nek a Los Alamos National Labból, akik türelmesen válaszoltak minden olyan kérdésükre, ami a Beowulf második nemzedékére vonatkozott.

Linux Journal 2002. augusztus, 100. szám

Glen Otero
PhD-fokozatot szerzett immunológiából és mikrobiológiából, továbbá a Linux Prophet nevű tanácsadó céget vezeti a kaliforniai San Diegóban.

Richard Ferri
az IBM Linux Technology Centerben vezetőprogramozó. Nyílt forrású linuxos géptelepeken dolgozik.

Kapcsolódó címek

Beowulf-háttéranyag ➲ http://www.beowulf.org
BProc ➲ http://www.sf.net/projects/bproc
A Gordon Bell-díj bejelentése:
➲ http://sdcd.gsfc.nasa.gov/DIV-NEWS/CESDIS.11_97.Becker.award.html
Thomas L. Sterling, John Salmon, Donald J. Becker és Daniel F. Savarese: How to Build a Beowulf. The MIT Press, 1999. ISBN 0-262-69218-X.

Egyéb Beowulffal foglalkozó oldalak:

➲ http://www.debian.org/ports/beowulf/
➲ http://sdcd.gsfc.nasa.gov/ESS/annual.reports/ess97/sys/linux.html
➲ http://freshmeat.net/projects/beowulf/?topic_id=136%2C141%2C143
➲ http://www.linuxjournal.com/article.php?sid=5710
➲ http://www.ibiblio.org/pub/Linux/docs/HOWTO/other-formats/html_single/Beowulf-HOWTO.html
➲ http://www.compaq.com/solutions/customsystems/hps/linux.html
➲ http://e-nef.com/linux/beowulf/