A másik gépem egy szuperszámítógép

Steve Jones 2002 novemberében kezdett el jegyzeteket készíteni a PBS-ről, MPI-ről és a Mosixról, majd 2003 júniusában már egy a TOP 500-as listán szereplő számítógéptelep vezetője volt. Jó példa ez arra, hogyan képes segíteni a Rocks-terjesztés egy telep vezetőjét a rendszer felállításában és üzemeltetésében.

Úgy két éve ***Mitch Davis*** (Stanfordi Egyetem műszaki tudományos vezérigazgatója) és ***Carnet Williams*** (a Stanfordi Egyetem műszaki tudományos igazgatója) egy igen komoly és nagy jelentőséggel bíró projekt kapcsán hívtak fel. A Stanfordi Jogi Egyetem hálózati műveletek osztályának vezetőjeként nagy örömömre szolgált együtt dolgozni Mitchel és Carnettel. (Mitch egyben a Jogi Kar dékánja is volt, Carnet pedig a főinformatikusi posztot töltött be ugyanitt.)

Ők meséltek nekem először ***Dr. Vijay Pande***-ről, a ***Folding@home*** projekt vezető kutatójáról, aki egy nagy géptelepet szeretett volna vásárolni. Megadták neki a nevemet, mivel olyan embert keresett, aki ezt a projektet képes sikerre vinni. Ösztönösen igent mondtam. Megbeszéltük a projekt részleteit és mielőtt leraktam volna a kagylót azért rákérdeztem: „Mekkora is lesz pontosan?" „300 kétprocesszoros csomópont" – válaszolták. „600 CPU ... Ez aztán tud tarolni" – gondoltam akkor.

Miközben Mitch, Carnet és Vijay a Dellel és az Intellel együtt a fürt megvásárlásáról tárgyaltak, én küldtem egy e-mailt Vijay Pande-nak, melyben leírtam, mivel tudnék segíteni neki a projekt hálózati és hardver részében. Egyben reményemet fejeztem ki, hogy az általa alkalmazni kívánt szoftvert a kivitelezés során jobban megismerhetem majd. Üzenetem utolsó sora a következő volt: „Valami nagynak szeretnék részese lenni ". Vijay azonnal válaszolt, és segítségemet örömmel fogadta. Megszerveztük első megbeszélésünket, mely során megvitattuk a projekt hatókörét.

Azon az első találkozón úgy tűnt, hogy a legtöbb dolog a levegőben lóg. Mindenki tudta, hogy jön a berendezés, de nem voltak valódi tervek. Vijay azt mondta, tisztában van vele, hogy a próbaüzemről és a használni kívánt fájlrendszerről dönteni kell. Emellett annak a lehetőségét is meg kellett fontolni, hogyan használhatjuk a már meglévő stanfordi szolgáltatásokat.

Vijay megemlítette a PBS, MPI és Mosix futtatásának lehetőségét is. Ezekről nagyon keveset tudtam, de feljegyzéseket készítettem és rákerestem a Google-lal a fentiekre, valamint a „beowulf" és „fürt" szavakra is. Belefutottam egy bemutatóba, ami a fürtépítésről, és egy Rocks nevű nyílt forráskódú szoftverről szólt. Utóbbit az NPACI szervezet készítette (➲ www.rocksclusters.org). Kitűnő bemutató volt, rengeteg kérdésemre azonnal választ adott. Ilyen volt többek között az az „egyszerű" probléma, hogy tulajdonképpen hogyan is kell összerakni egy ilyen fürtöt, hogyan kezeljük a csomópontokon futó szoftvereket, hogyan állítsuk be a mester-csomópontot és hogyan felügyelhetjük a többit. A bemutató gyakorlatilag megadta a vázát a mi tervezett fürtünk felépítésének. Kinyomtattam és magammal vittem a következő megbeszélésünkre. A készen kapott megoldás jó fogadtatásra talált.

1. kép Iceberg a Forsythe Adatközpontban

Folding@home az Icebergen

Az Iceberg ellenőrzi a Folding@home adatait. Ez egy elosztott számítási projekt, amelynek célja a fehérjék helyes és téves összekapcsolódásának, valamint az erre visszavezethető betegségek tanulmányozása. A megvalósításhoz önkéntesek szabad processzoridővel járulnak hozzá. Jelenleg körülbelül 80,000 processzor számolja az adatokat.
A Iceberget a Folding@home-mal kapcsolatos kutatások során kisebb feladatok szimulálására használom. A feladat itt egy olyan szimulációsorozat végrehajtását jelenti, amelyben egy fehérje egy előre meghatározott módon kapcsolódik össze egy másikkal. A kulcs egy olyan szkript megírása volt, amely leutánozza azt a folyamatot amelyben a Folding@home szétosztja a feladatot az ügyfeleknek. Egy futtatáshoz általában 10-20 CPU-t használok egyszerre.
Más, nagyobb projektekhez az Iceberget csak a kezdeti felosztáshoz használtam. Általában 10-50ns-os szimulációkat hajtunk végre 1ns-os darabokban. Az Iceberget az első 1ns-hoz használjuk, majd áttérünk a Folding@home-ra és ott folytatjuk a munkát. Új módszereinket gyorsan meg tudjuk ismételni az Iceberg által kínált kontrollált és stabil környezetben. Új projektek fejlesztésénél az Iceberget az eredmények ellenőrzésére használjuk, és amint biztosak vagyunk az új módszerben, ráeresztjük azt a 80,000 CPU-s elosztott számítógépre.

–Young Min Rhee a Folding@home projektből, folding.stanford.edu

Az Iceberg felállítása

Miközben ez a két találkozó zajlott, a Dell a Stanfordi Forsythe Adatközpontban a fürtöt szekrénybe szerelte és összekötötte a csomópontokat. Mindez hét napig tartott. Letöltöttem a Rocks 2.3 változatát, és a telepítettem a Rocks-zsargonban központnak nevezett gépre. Az egész valahogy elbűvölően egyszerű volt. A sikeren felbuzdulva harmadik megbeszélésünk után úgy döntöttünk, hogy a hardver és a hálózat felépítése mellett feladatköröm kibővült a szoftver kezelésével is. Biztos voltam benne, hogy a többi akadályt is sikerrel veszem majd, de utólag azt kell mondanom, hogy ezen a ponton még egyáltalán nem voltam tisztában a feladat valódi méreteivel. Valójában ugyanis minden idők legnagyobb Rocks fürtjét kezdtem el építeni.
Az első problémával akkor szembesültem, amikor megpróbáltam telepíteni egy számítási csomópontot. Erre való az `insert-ethers` nevű Rocks segédeszköz, amely a számítási csomópontok Ethernet MAC címeit megtalálja, IP címeket és gépneveket oszt ki nekik, majd ezt az információt – a PXE-t és a DHCP-t használva – egy megadott egyeztetési protokoll segítségével beilleszti egy adatbázisba. A csomópont beillesztését követően, az azon futó rendszer egy megfelelő Red Hat Kickstart állomány alapján épül fel és kerül beállításra, befejezve ezáltal a PXE rendszerbetöltési folyamatot.
Sajnos gondjaim adódtak a Dell PowerEdge 2650-ben található hálózati kártyával. Úgy tűnt a Rocks nem támogatja a Broadcom Ethernet vezérlőket. Elküldtem a problémámat a Rocks vitafórumára és a Dell-t is felhívtam támogatást kérve, és mivel arany fokozatú támogatási szerződésünk volt, nyitottam egy szervizjegyet.
A Rocks fejlesztői gyorsan készítettek számomra egy kísérleti változatot, amely tartalmazta a szükséges frissített eszközmeghajtókat. Ez megoldotta a problémát, és hamarosan viszontláttam a javaslataimat és észrevételeimet a Rocks 2.3.1 szerviz kiadásában.
Az utolsó gond amit a méretezésnél vettem észre, az volt, hogy nem voltam képes 511 aktív feladatnál többet futtatni. A felhasználóim sikítoztak a 100 kihasználatlan processzort látván. Ez a helyzet azért állhatott elő, mert az Icebergen futtatott feladatok nagy része rövid életű, egy-két processzoros folyamat volt. Miközben a Rocks fejlesztői csapatával dolgoztam együtt, megpróbáltunk előre meghatározott állandókat találni a Maui időzítő kódjában. Végül én találtam meg, és a Rocks csapatának útmutatásával, újrafordítottam és újraindítottam a Maui-t. A rendszer most már annyi aktív feladatot tud időzíteni, ahány szabad processzor van.

A TOP500-as futtatás

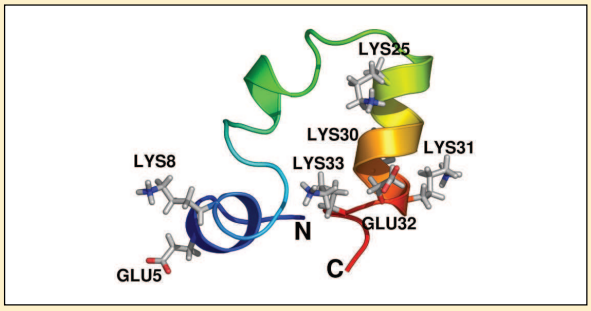
2002 végén, miután megoldottuk az utolsó hardver és szoftverproblémát is elhatároztuk, hogy az Iceberget feltesszük a TOP500 szuperszámítógép listára (➲ www.top500.org). A TOP500 egy félévente megrendezett verseny, amely a Linpack (egy lineáris algebrai feladatok megoldására használható csomag) tartós futtatásával mért teljesítmény alapján rangsorolja az 500 leggyorsabb gépet. A 2002 novemberi listán 97 hagyományos gépekből álló fürt szerepelt, így biztosak voltunk benne, hogy az Icebergnek van esélye a listára való felkerülésre.
A TOP500 listára való felkerülés azonban több munka volt, mint amire számítottunk. A Rocks egy előre lefordított Linpack állománnyal érkezik, amely képes ugyan jó teljesítményt elérni, de mi többet akartunk.
Kapcsolatba léptem a Dell Skálázható Rendszerek csoportjával. Tőlük a Linpack fürtre hangolásában kaptam segítséget. Dolgoztunk együtt például a Linpacknak a Goto BLAS könyvtárhoz való hozzáillesztésén (Basic Linear Algebr Subroutines; alapvető lineáris algebrai szubrutinok). Utóbbit ***Kazushige Goto*** írta (➲ www.cs.utexas.edu/users/flame/goto).
Ezen felül a Dell javaslatot tett egy kifinomultabb hálózati topológiára. A TOP500 futtatás előtt a 300 csomópont 16 db 100Mbit-es Ethernet kapcsolón osztozott (Dell PowerConnect 3024). Úgy találtuk, hogy a Linpacknak – mint megannyi más, erősen párhuzamosított alkalmazásnak – kifejezetten használ, ha egy jobb hálózati kapcsolatot építünk ki. (Magyarul kisebb késleltetési időt és/vagy nagyobb sávszélességet biztosítunk). A Dell kölcsönzött nekünk néhány Gigabites blokkolásmentes kapcsolót.
A fenti fejlesztések javították a teljesítményt, és 2003 júniusában bemutattuk eredményeinket a TOP500 listán. Az Iceberg a 319. helyen áll és érzésem szerint egy gyorsabb hálózati kapcsolattal akár előrébb is lehetne.

IN SILICO (számítógépes) biológia az Icebergen

Az Iceberg Dell szuperfürt létrehozása felért egy tudományos forradalommal. Tulajdonképpen évtizedek óta használunk szuperszámítógépes központokat, de mindig kényelmetlenül éreztük magunkat a meglehetősen szűkre szabott feldolgozási sorok, a kevés gépidő miatt, no meg azért, mert valahogy mindig nehézkesen lehetett csak a rendszert az igényeinkhez igazítani. Az elmúlt néhány évben felépítettük saját Linux fürtjeinket 50-100 CPU-val. A kész hardver azonban sajnos magasabb támogatási és adminisztrációs költségekhez vezet, arról nem is beszélve, hogy amíg cikket írunk, a fürt tétlenül áll. Az új Stanfordi Dell fürt ezzel szemben rendkívül költséghatékony megoldás kínál tudományos feladatokra. Megosztott erőforrásként csomópontok százai érhetőek el, ha mondjuk egyik éjszaka tesztelni szeretnénk egy új modellt, a költségünk viszont kizárólag az általunk használt átlagos csomópontszámmal arányos. Amellett, hogy a hardver teljes irányításunk alatt áll, az igénybe vehető erőforrások egy nagyságrenddel nagyobbak annál, mint amit valaha is használtunk a telepítés előtt vagy után. Ehhez képest az adminisztrációval hetente mindössze egy órát töltünk.

A számítógépes biológia fehérjekutatással kapcsolatos alkalmazásai általában rendkívül nagy számítási kapacitást igényelnek. Az egyik legáltalánosabb esetben az a feladat, hogy logikai összefüggést, feltűnő mintázatot találjunk a szekvenciacsaládok és a szerkezetek között. Hasonló, de kicsit más a biomolekulák belső mozgásainak szimulációja. A mintaegyeztetés nem nagy dolog, ha csak néhány tucat szekvencia áll rendelkezésünkre, a jelenlegi projektünkben azonban az cél, hogy a Shewanella baktérium fehérjéivel kapcsolatban tudjunk pontos előrejelzéseket adni. (Ez a baktériumfaj arról a képességéről híres, hogy radioaktív és mérgező hulladékot eszik). Ez pedig több százezer szekvenciát jelent, amelyeket páronként össze kell hasonlítanunk.

A helyi fürtünkön ez több heti alaposan megtervezett futtatást kívánt. Ugyanezt a jelenleg rendelkezésünkre álló eszközökkel egyetlen éjszaka alatt megtehetjük. Ennyi idő kell ahhoz, hogy kipróbáljunk egy új ötletet, másnap pedig már bemutathatjuk az eredményt.



1. ábra A Villin headpiece

A Villin headpiece önmagában egy nagyon kicsi, körülbelül 600 atomból álló protein. Ugyanakkor mindig víz veszi körül (vörös/fehér pálcikák), ami a kezelendő atomok számát körülbelül 10,000-re növeli. Minden egyes atom a legközelebbi 100-200 szomszédjával lép kölcsönhatásba. Ezeknek a hatásoknak az erősségét minden egyes lépésben ki kell számolni, majd ezt félmilliárdszor meg kell ismételni ahhoz, hogy a molekula mozgásának egy mikroszekundumnyi részletét pontosan leutánozzuk. Az 1. ábrához használt adatok az Iceberg 10 csomópontján két hetes futtatással készültek.

Az atomok mozgásának molekuladinamikai szimulációja szintén lenyűgöző feladat. Az ötlet nagyon egyszerű, ki kell számítani az atomok egymásra kifejtett erejét, majd Newton egyenlete alapján meg kell határozni az új pozíciókat egy megfelelően rövid idővel később (egy lépés általában 2 femtomásodperc, vagyis 2 × 10-15 másodperc). A szimuláció egy lépése ugyan gyors, egy biológiai reakció tanulmányozásához azonban lépések milliárdjai szükségesek . Ezért a kódot úgy optimalizáltuk, hogy kihasználhassuk a Pentium 4 Xeon processzorokon elérhető Intel Streaming SIMD Extensions utasításkészletet. Ezzel átlagosan kétszeres illetve négyszeres közötti gyorsulást érhetünk el. Ez tett elehetővé, hogy olyan méretű fehérjéket szimuláljunk több mint egy mikromásodpercnek megfelelő valós ideig, mint a Villin headpiece (1. ábra), s mindez csupán 2 hétbe telt a Iceberg 10 csomópontján. Valójában az optimalizált kód még egy önálló Dell/Intel Xeon processzoron is gyorsabb, mint a csúcskategóriás IBM Power4 vagy Alpha processzorokon, s mindezt tizedannyi költség mellett érjük el. Eddig ez volt messze a legjobb informatikai beruházásunk, így ha a jövőben bővítenünk kell, nem fogunk habozni.

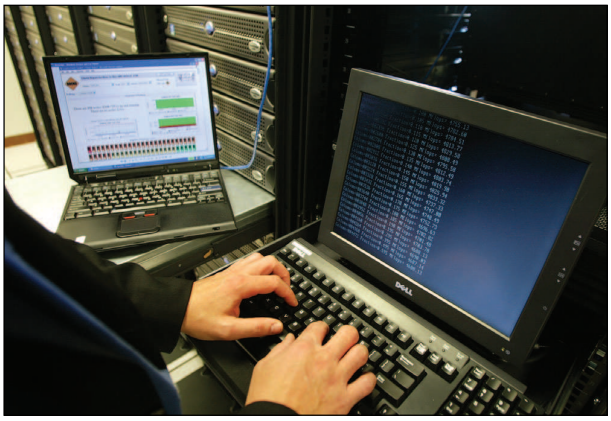
– Michael Levitt és Erik Lindahl a Szerkezeti biológia tanszékről, Stanfordi Egyetem Orvosi Kar

Az Iceberg új otthonába költözik

Az Iceberg állandó lakóhelyének a James H. Clark Központot szemeltük ki. Ez a Silicon Graphics és Netscape alapítójáról kapta a nevét, aki egyben a Bio-X Projekt alaptőkéjét szolgáltató mecénás is. 2003 augusztusában eljött a költözés ideje. A Forsythe Adatközpontból való elköltözés sok jó dolgot hozott számunkra, melyek közül az egyik legfontosabb a Rocks újratelepítése volt. Azért erőltettem ezt a dolgot, mert egy ekkora méretű fürt karbantartásának a stabil infrastruktúra a kulcsa. Ez teszi lehetővé a teljes tulajdonlási költség alacsonyan tartását. Amíg az Iceberg a Forsythe Adatközpontban volt arra a következtetésre jutottunk, hogy a hardveren és a szoftveren egyaránt van még mit javítani.

A költözés kapóra jött, hiszen az ezzel együtt járó állásidő lehetővé tette számunkra, hogy a módosításokat elvégezzük a fizikai modellen. Úgy döntöttünk, hogy a központi csomópontot továbbfejlesztjük, a felhasználói könyvtárakat pedig áthelyezzük egy másik, háttértárolóval rendelkező csomópontra.

A Rocks újra beváltotta a hozzá fűzött reményeket. Annyi volt csak a dolgunk, hogy az `insert-ethers` programban kiválasszuk a beillesztendő csomópont NAS készüléktípusát. Hogy teljes mértékben kiaknázhassuk az ebben található dupla Gigabit Ethernet hálózati kártyát, a kapcsolat-összefűzést használtuk. A központi csomóponton végzett néhány módosítás után, ami gyakorlatilag



2. kép Linpack futtatása a TOP500-ba való kerüléshez

3. kép Az Iceberg új otthonában
Balról jobbra: Vijay Pande, a Folding@home vezető kutatója; Steve Jones, az Iceberg tervezője; Erik Lindahl, posztdoktori ösztöndíjas; és Young Min Rhee, kutató

a felhasználók új készülékhez való hozzárendelését jelentette az előzőleg mentett adatok visszamásolása után újra működtünk.

A személyi szuperszámítógép filozófiája

Minden nagy teljesítményű számítógépfürt telepítésénél az elsődleges cél az, hogy azt a lehető leggyorsabban megbízhatóan működő állapotba hozzuk, és ez az állapot a hardver haláláig fenn is maradjon. A rendszert használó kutatók így biztosak lehetnek benne, hogy amikor egy problémával kapcsolatban nagyobb számítási teljesítményre van szükségük, az rendelkezésükre is áll. Egy ilyen méretű és kihasználtságú rendszert természetesen nem lehet csak a próba kedvéért módosítgatni.

Meglátásom szerint ezzel a koncepcióval tökéletesen összhangban van mindaz, amit az általunk választott standardizált fürt disztribúció biztosít. A rendszerfelügyelet ennek használatával kimerül a feldolgozási sorok illetve a fájlrendszerek telítettségének megfigyelésében. Szemmel kell tartani persze a különböző naplókat illetve magát a hardvert is, de ez tulajdonképpen természetes.

Kombinálva a Rocks és a Dell támogatási tervét a számítási pontokra ezüst, a központi csomópontra pedig arany fokozatú támogatást kaptunk, ami azt jelenti, hogy három évig nem lesz gondunk sem a teljesítményre, sem a meghibásodott alkatrészekre. A további finanszírozással kapcsolatban úgy gondoltuk, hogy a gépidő kiszámlázásával elő tudjuk teremteni azt az összeget, amellyel három éven belül lecserélhetjük a teljes rendszert. Erre aztán áttelepítjük a Rocks rendszert, és további három évig megint nem lesz problémánk.

Az Iceberg-II várhatóan ugyanannyi csomópontot tartalmaz majd, mint elődje, de a 2U helyett 1U magas gépek lesznek benne a helytakarékosság végett. Az elód különálló fürtként tovább működik majd, de egyre kevesebb csomóponttal, mert a meghibásodott elemeket már nem cseréjük benne.

Emellett tervezzük egy másik, 600 csomópontot számláló fürt beszerzését is, amit az elkövetkező 6-12 hónapban akarunk elindítani. Ez a fürt egy külső telephelyen lakik majd. Pillanatnyilag tárgyalunk egy kutatóintézettel, amely érdeklődött a lehetőség iránt. Ők megfelelő gépidőért cserébe speciális szolgáltatásokat (generált áram, szünetmentes táp, légkondicionálás) kínálnak.

Azt is tervezem, hogy az Iceberg mellett építek még egy fürtöt, kifejezetten kereskedelmi célokra. Szeretném, ha befejezése után nemcsak ez lenne a legnagyobb Rocks fürt, hanem bekerülne a TOP500 lista első 20 helyezettje közé.

Záró megjegyzések

Az Iceberg karbantartási költségeinek alacsonyan tartását alapvetően az teszi lehetővé, hogy körülötte éppen jó arányban vannak jelen az infrastruktúrával és a tényleges használattal foglalkozó emberek. Utóbbiak értelemszerűen azok a kutatók, akik számítási feladataikat futtatják a rendszeren. A műszaki csapat nemcsak karbantartja a rendszert, hanem lehetővé teszi annak az oktatásban való felhasználását, illetve azt is, hogy az érdeklődők megismerhessék az általa nyújtott szolgáltatásokat.

A feladatokat ésszerűen elosztottuk, így a tűzfal karbantartása, az ütemező beállítása, a felhasználói adatok vagy a hardver karbantartása egy-egy ember munkaidejének csak egy nagyon kis részét veszik el. A rendszer karbantartására fordított idő hetente átlagosan egy óra, ami talán a legfényesebb bizonyítéka a jó tervezésnek. Az sem elhanyagolható persze, hogy egy 203 csomópontot számláló fürt tulajdonlási költsége megegyezik egy 10 csomóponttal rendelkezőével.

Linux Journal 2003. november, 115. szám

Steve Jones (stevejones@stanford.edu)
Azért adta fel a Stanfordi Egyetem Jogi Karán betöltött állását, hogy egy internetszolgáltató biztonsági stratégájaként folytassa pályafutását. Jelenleg tanácsadóként a biztonsági rendszer kialakításáért felel. Jelenleg épp átköltözni készül Main államba, ahol egy oktatási intézmény stratégája lesz, illetve egy másik vállalatot is akar alapítani. Szabadidejében, egy 302 csomópontból álló, Icebergnek nevezett fürt rendszergazdája.