HPC, klastry, servery a Linux

HPC (High Performance Computing, vysoce výkonné výpočty) je hájemstvím superpočítačů a "velkých" serverů. Odborníci z univerzit v Mannheimu a v Tennesee každý půlrok vyhodnocují nejvýkonnější světové počítače a výsledky zveřejňují v seznamu Top500.

Seznam Top500 z listopadu 2002 potvrdil nástup open source klastrů - dva linuxové klastry s procesory Xeon se dostaly až mezi deset nejvýkonnějších superpočítačů světa. To už něco znamená; vítěz, monstrum Earth Simulator s 8192 vektorovými procesory firmy NEC, podává ustálený výkon 35,86 teraflopsu (trilionů floating-point operací za sekundu) - zhruba stejně, jako dává dohromady jedenáct jeho následovníků z posledního Top500, a dokonce všech 500 superpočítačů ze seznamu z listopadu 1998! Názor, že Linux je málo stupňovatelný a nevhodný pro nejvyšší výkony, už neplatí, což důrazně potvrdilo i ohlášení nových serverů SGI.

Altix

Servery SGI Altix řady 3000 s procesory Itanium 2 překvapily tím, že zvýšily stupňovatelnost linuxové platformy až do 64 procesorů a 512 GB vnitřní paměti v jediném systému (single-image system). Doposud byly takovéto parametry dostupné pouze při spojení několika systémů do klastru, tedy speciálním způsobem organizované počítačové sítě. Síťové propojení počítačů v klastru však je pro řadu velkých aplikací omezující, o něco složitější než u single-image systémů je i programování úloh.

Při benchmarkových testech 64procesorový Altix výrazně překonal konkurenční unixové servery ve výkonu v plovoucí řádové čárce a v datovém toku do/z paměti i v testech s reálnými vědecko-technickými aplikacemi (molekulární chemie, dynamika tekutin). Firma SGI uvedla i Altix Supercluster, tedy možnost spojovat několik systémů do tzv. superklastrů, které mohou mít až několik set procesorů pracujících s daty uloženými v jediné společné paměti.

Při zvyšování počtu procesorů od určité hranice nedochází ke zvyšování výkonu, protože režie systému převáží nad přidaným výkonem z více procesorů - tato hranice je tím vyšší, čím je architektura systému propracovanější. Již několik let existují servery SGI Origin, postavené stejně jako Altix na architektuře NUMAflex, která jim propůjčuje stupňovatelnost až do 2048 procesorů v single-image stroji (v praxi je největším 1024procesorový Origin 3800 v NASA Ames Research Centre). Jen nevýrazný výkon procesorů MIPS zabránil Originům v proniknutí do skutečné špičky, ale použití procesorů Itanium tento handicap odstraňuje. Dalšími velkými single-image servery v praktickém provozu (nejedná se o masivně paralelní či speciální superpočítače, to je jiná kategorie, hlavně cenová) jsou 128procesorové Fujitsu Siemens "Kaiser" PRIMEPOWER 2000 či 72(106)procesorové (34 pomocných procesorů) Sun StarCat 15000. Největším intelovským single-image serverem byl zatím 32procesorový Unisys ES7000, resp. NEC TX7.

Důležitý je i výkon procesorů. Ještě nedávno byly suverény riscové procesory (zejména Digital Equipment Alpha) a Intely se krčily skromně v pozadí. Dnes však Intel (resp.

i AMD) předstihl většinu konkurence a krok s ním drží jen znovuvzkříšená Alpha (dnes už od HP) a IBM Power4+. I když nelze vyloučit překvapivý pokrok či stagnaci některého výrobce (vzpomeňme, jak Itanium Merced zklamal vysoké ambice), dlouhodobý trend je zřejmý a hlavně vysoká efektivnost hromadné výroby Intelu a AMD přinesla ceny, kterým lze těžko konkurovat.

Nové servery SGI jsou připraveny i na příští dvě generace procesorů Intelu - Madison a Montecino. Pracují pod systémem Red Hat Linux 7.2, jehož schopnosti v oblasti stupňovatelnosti, práce s daty a využití výkonu byly beze změny v jádru zvýšeny extenzí SGI ProPack. Například díky výkonnému souborovému systému XFS (produkt SGI, nyní open source) dosáhl Altix I/O propustnosti přes 2 GB/s. Při vývoji extenzí společnost SGI oboustranně spolupracovala s komunitou open source, z pochopitelných důvodů jí však nevydala všechna svá zdokonalení. Proto sice můžete na Altixu provozovat standardní Red Hat a jeho aplikace, ale budete-li chtít špičkový výkon, musíte si pořídit ProPack. Výhodou Altixu proti klastrům i klasickým superpočítačům je snadné programování výkonných aplikací, protože nemá na vývoj aplikací zvláštní omezující požadavky a podporuje všechny známé paralelní programovací modely.

Obrovská operační paměť Altixu je přístupná kterémukoliv z procesorů, proto je vhodný především pro aplikace pracující s rozsáhlými daty (globální modelování počasí, bioinformatika, náročné simulace apod.). Velké datové objemy se tak mohou zpracovávat nesrovnatelně rychleji než při jinak nutném postupném načítání z disků. Altix však nabízí také spojení jednotlivých strojů propojením NUMAlink, které poskytuje unikátní možnost přístupu k jakékoliv oblasti paměti nejen procesorům v lokálním uzlu, ale i všem procesorům v ostatních uzlech superklastru. Toto řešení poskytuje až 200x rychlejší přístup k datům

v jiném uzlu než dnes nejrychlejší síťová propojení užívaná v linuxových klastrech. To umožní už dnes sestavit superklastry až se stovkami procesorů a s terabajty sdílené vnitřní paměti (na rozdíl od běžných klastrů, které jsou omezeny podstatně menšími objemy lokální paměti v uzlech a hlavně pomalým síťovým propojením mezi uzly).

Komerční servery

HPC je především vědeckou a akademickou oblastí - mnohem lukrativnější je uplatnění serverů v komerční sféře. V dnešním prosíťovaném světě pohánějí velké servery nejen internet, ale jsou skryty prakticky za vším "počítacím", od platební karty až k mobilu. Nový význam získaly s konsolidací - místo provozování řady aplikací na mnoha "menších" serverech pod různými platformami se všechny aplikace svěří jednomu "velkému železu". Konsolidace vede ke snadnější centrální správě dat i aplikací a tím ke snížení provozních a personálních nákladů. Efektivní využití "konsolidačního" serveru však vyžaduje podporu tzv. particií (partitions), tedy jakéhosi rozdělení serveru na několik nezávislých částí, z nichž každá provozuje svou instanci operačního systému.

Propracovaný systém particií už řadu let nabízejí systémy mainframe (hlavně IBM); u nich lze rozdělit i jediný procesor a jeho zdroje (paměť, I/O) na více logických (softwarově dynamicky spravovaných) particií. Na serverech zSeries a iSeries společnost IBM s velkým úspěchem u zákazníků zavedla i podporu linuxových particií. U unixových multiprocesorových serverů zpočátku existovaly pouze statické particie, rekonfigurovatelné jen po zastavení serveru. V prosinci 1999 však Sun zavedl i softwarově spravované particie (u Sunu zvané dynamic domains, u HP virtual partitions, ale podstata je obdobná), teprve později byly implementovány u dalších unixových dodavatelů (HP Superdome, IBM eSeries, SGI Origin, Fujitsu Siemens a další.). Dynamická rekonfigurace particií ve spolupráci s programy pro správu aplikací a optimalizaci výkonu a vytížení umožňuje pružnou rekonfiguraci a efektivní využití systému podle okamžitých požadavků.

Ošidná čísla benchmarků

S benchmarky je to jako se známkováním ve škole - lze říci, že jedničkář je chytřejší než trojkař, ale stoprocentně to neplatí ani v rámci jedné třídy, natož mezi dvěma školami a je řada nuancí, které se dají těžko vystihnout strohými čísly. Celkově se však dá říci, že jedničkáři jsou lepší žáci než pětkaři.

Nejpopulárnějším benchmarkem komerčních serverů je TCP-C, simulující on-line transakční zpracování (např. bankovních transakcí) nad distribuovanou databází. V tabulce výkonů TPC-C single-image systémů již dlouhou dobu vévodí Fujitsu Siemens PRIMEPOWER 2000, následovaný servery IBM pSeries 690 a HP 9000 Superdome. Je-li měřítkem cena za jednotku TPC výkonu, vítězí mezi výkonnými single-image servery Unisys ES7000 Orion s procesory Intel Xeon MP. Servery založené na standardních intelovských čipsetech (např. Dell) samozřejmě podávají lepší poměr cena/výkon, ale jejich celkový výkon je pro náročnější aplikace nepostačující. Proto se uplatňují zejména v klastrech, kde je však nevýhodou složitější software a relativně malá kapacita paměti dostupné v uzlech. Zahrneme-li do tabulky výkonu v TPC-C i klastrové systémy, dostane se do čela HP (Compaq) ProLiant DL760 se 32 uzly po 8 procesorech.

TPC-C je jen jedním z řady benchmarků. Některé zdůrazňují výkon procesorů či propustnost systému, jiné výkon v I/O, v podpoře Javy nebo webu. Skoro každá z těchto oblastí je doménou jiného výrobce - to odpovídá výše zmíněným, těžko rozlišitelným rozdílům mezi jedničkáři a trojkaři. Významné jsou zejména aplikační benchmarky (SAP, Oracle, Baan). Přehled výsledků benchmarkových testů (www.ideasinternational.com) nebo žebříček Top500 (www.top500.org) mohou něco naznačit, ale žádný benchmark nenahradí výsledky získané přímo v užívané aplikaci a v konkrétních podmínkách. Jisté však je, že do čela tabulek žádní "pětkaři" neproniknou.

Závěr

Platforma Intel/Linux má v oblasti výkonných systémů velké perspektivy nejen pro příznivý poměr ceny k výkonu, ale i z hlediska možnosti dosažení vysokého výkonu. Svědčí o tom například plány na HP Superdome s Itaniem, spolupráce Fujitsu Siemens s Intelem na vývoji linuxových HPC serverů i podpora Linuxu ze strany IBM a řady dalších "velkých IT hráčů".

Vrátíme-li se k serverům SGI Altix, v oblasti HPC patří k absolutní špičce. V komerční oblasti se zatím společnost SGI nepokoušela s konkurencí soupeřit. Tím, že používá standardní platformy Linux a Intel, se jí však otevírá přístup prakticky ke všem standardním aplikacím. Lze proto očekávat, že je to jedna z pravděpodobných oblastí růstu pro firmu SGI.

Josef Chládek