IBM p5 570+: Snabbhet - steg för steg

Notis

Gordon Haff
10 mars 2006

IBM System p5-570 är det första mellanklassystemet i IBM:s POWER5-serie. Det har i viss mån överskuggats av p5-590 och 595 Big Iron, den tunga klusternoden p5-575 och även de många aggressivt prissatta ingångssystem och lägre mellanklassystem som stärkt IBM pSeries ökande marknadsandelar inom små och medelstora företag och fjärranslutna kontor.[1] Modell 570 har inte fått så mycket uppmärksamhet, delvis eftersom dessa övriga system haft många starka sidor. Det stämmer också att mellanklassystem oftast inte är särskilt spännande. Oavsett i hur stor utsträckning de används och oavsett hur viktiga de är för systemleverantörernas intäkter och resultat, så är de ofta bara enklare versioner av större system.

Modell 570 är annorlunda. I en värld där skalbarhet, absoluta prestanda och modultänkande ofta utesluter varandra innehåller systemet förvånansvärt få kompromisser. Det kan fås i olika utföranden, allt från en processor (med dubbla kärnor/dual core) och hela vägen upp till ett system med 16 kärnor och en klockhastighet på över en miljon tpmC – en beräkningsvolym som verkade helt omöjlig även för de största servrarna bara för några år sedan. Och det är byggt av diskreta byggblock i ”lådform”, med kabelanslutningar. Modultänkandet gör att kunderna kan utöka sina system, betala stegvis och köpa ny kapacitet när det behövs i stället för alltsammans på en gång.

570 är kort sagt knappast bara ännu ett system i mellanklassen. Modulsystemet bygger på känd teknik inklusive IBM:s egna X3 chipset-baserade[2] xSeries-servrar – 570 är unik både vad gäller absoluta prestanda och den skalbarhet som kan fås fram ur en så komponentbaserad konstruktion.

Utbyggnad och linjär prestandaökning

Först tittar vi på skalbarheten i 570-systemet, d.v.s. hur väl systemet omvandlar fler processorer till extra prestanda. I en perfekt värld skulle dubbla antalet processorer också dubbla den programprestanda som användaren ser. Men detta perfekta tillstånd uppnås sällan i verkligheten. När ett system växer och belastningen ökar kan någon del av systemet (t.ex. en minnesbuss eller ett I/O-system) eller själva programvaran lätt bli en flaskhals – vilket hindrar de övriga delarna från att arbeta med full kapacitet. När systemets storlek ökar uppstår också fler kostnader för att hålla alla data koordinerade och konsekventa.

Skalbarhet kan vara ett problem i större moduluppbyggda serversystem genom att sätta samman två eller flera byggblock i SMP-servrar (Symmetric MultiProcessing) med externa kablar som i system 570. Eftersom sträckorna tenderar att vara längre

1 Se även våra engelska informationstexter ”POWERing the Performance Factory”, ”POWER5 Takes Off on pSeries”, ”IBM’s p5 Mothership Whizzes into Orbit” och ”IBM Does SMB”

2 Se ”Xeon Zips with X3.”

än i en enda låda och den exakta fysiska layouten mindre styrd tenderar ”kabelrören” att vara både mindre och långsammare när det gäller att skicka data och styra signaler över ett kabelsystem, både över backplane och crossbar. Kommunikationer från ett byggblock till ett annat, till exempel för att hämta vissa data från minnet i ett annat block, tar därför längre tid än att utföra samma åtgärd lokalt. Modulsystem är naturliga NUMA-arkitekturer (Non-Uniform Memory Access – Osymmetrisk minnesåtkomst).[3] I den första generationens NUMA-system från Data General och Sequent kunde det ta en processor mer än tio gånger längre tid att läsa data från ett minne som sitter i ett fjärranslutet byggblock än att läsa data från ett lokalt minne. Så stora skillnader innebär att vid ett betydande antal fjärråtkomster till minnet kommer minnets latencyvärde att bli ganska högt, vilket försämrar prestandan.[4] Med väl optimerade operativsystem (OS) och program kan dessa system fungera väl, men benchmarkvärdena visade på en akilleshäl i skalbarheten. Det mest uppenbara fallet var TPC-C – ett benchmarkvärde som i synnerhet omfattar minne och I/O-undersystem, och kräver omfattande kommunikation och fildelning över systemet – sådant som aldrig skalanpassats särskilt väl på dessa servrar ur första generationen.

p5-570 klarar sig *mycket* bättre än dessa tidigare system och överträffar till och med andra utmärkta modulsystem som t.ex. IBM:s egna processor-baserade X3-servrar i xSeries-sortimentet. I 570-systemet tar anrop till fjärrminnet bara cirka 25 till 50 procent längre tid än bästa kända lokala minnes-åtkomst. Det exakta värdet beror på antalet installerade byggblock och andra faktorer. Även fjärråtkomst går snabbt, mindre än 300 nano-sekunder. Detta är inte helt ”omedelbar” åtkomst (IBM undviker att använda NUMA-benämningen på 570-systemet) – men det är lika bra som eller bättre än många system som använder fasta inbyggda anslutningar och betydligt bättre än något annat modulsystem på marknaden, både relativt och absolut.

Benchmarkvärden för skalbarhet

Bevisen finns i siffervärdena. IBM genomförde en serie benchmarktester på POWER5-baserade 570-system, som gjorde det möjligt för oss att göra direkta jämförelser mellan olika systemstorlekar.[5] Vi tittar först på SPECjbb2000, ett test som huvudsakligen mäter processorkraften och förmågan att hantera trådar. Datagenomströmningen i modell p5-570 ökar i princip linjärt från två kärnor ända upp till 16. Det spelar heller ingen roll om operativ-systemet är IBM:s eget AIX 5L V5.3 UNIX-system eller Novell SUSE LINUX. Benchmarktestet körs aningen snabbare med AIX, men skalbarheten stämmer för båda operativsystemen.

Liknande resultat fås med benchmarkvärdet för SAP 2-tier Sales and Distribution Standard Application Benchmark (SAP SD, standardbenchmark för försäljning och distribution), som är en modell av en leveranskedja i ett företag. Detta är dels ett populärt benchmarkvärde och det används faktiskt för systemdimensionering, ett bevis på att det är en god beskrivning av belastningar i verkligheten. Även här: dubbla antalet kärnor, dubbla antalet användare.[6]

Även TPC-C, det klassiska *problembarnet* inom modulsystemkonstruktion, skalanpassas mycket väl på 570. TPC-C simulerar en användarpopulation som exekverar transaktioner mot en databas och därför kräver en stor mängd koordination över hela systemet för att säkerställa att varje transaktion fungerar med senaste tillgängliga data. Det gör det svårare att begränsa funktionerna helt och hållet till deras lokala byggblock – vilket kan göras enklare med tester som antingen koncentreras på minnes-avläsning eller som inte berör minnet särskilt mycket överhuvudtaget. Men även här kan 570-systemet skalanpassas väl. Resultaten är inte riktigt så noggranna och rena som hos övriga bench-markvärden – med tanke på att IBM genomfört några TPC-C-körningar med Oracle 10g och övriga med egna DB2 UDBV8.1 – men det övergripande mönstret visar fortfarande nästan linjär skalbarhet. Detta är ett bra resultat för alla system och det är

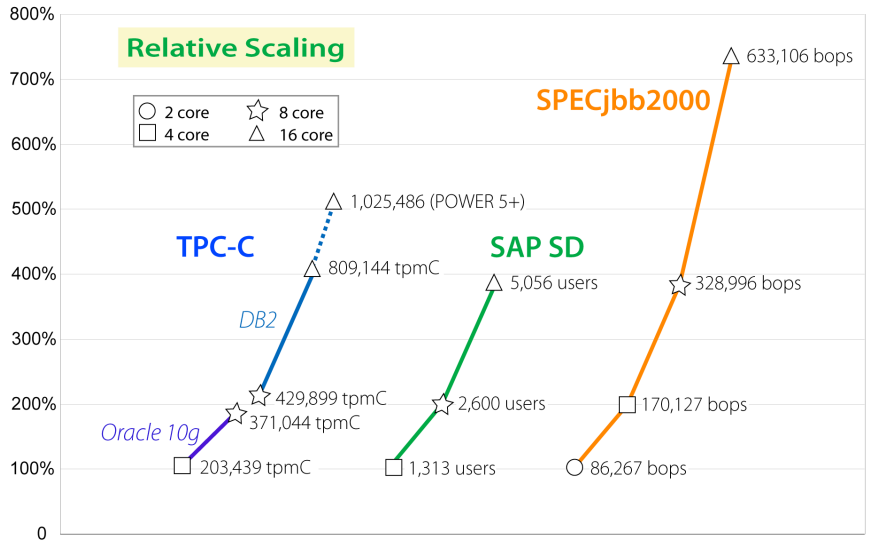
---

3 Nästan alla större serverkonstruktioner idag är NUMA-servrar i viss grad, oavsett om tillverkaren väljer att framhäva den egenskapen eller inte. Men termen associerades ursprungligen med modulkonstruktioner och används fortfarande oftast för den typen av system.

4 Se även dokumentet ”Latency Matters!”

5 IBM har inte släppt någon liknande uppsättning benchmarkvärden där nya POWER5+ använts, men det finns ingen större anledning att tro att skalbarhetsresultaten skulle vara annorlunda även om de absoluta värdena är något högre (ca 25 procent i genomsnitt).

6 Det vanligaste SAP SD-benchmarkresultatet är antalet simulerade SD-användare.



ganska anmärkningsvärt för ett moduluppbyggt serversystem.

Rena prestanda

Jämn skalanpassning indikerar inte heller någon kompromiss i övergripande systemprestanda. Med lanseringen av POWER5+ har topprestandan (med 16 kärnor) nu gått över en miljon tpmC (1 025 486 för att vara exakt). Till priset 4,42 USD/tpmC är pris/prestanda också det bästa bland topp 10-resultaten, om än med ganska liten marginal. Bland nuvarande produkter,[7] finns bara två som slår 570 i TPC-rankingen. Den ena är IBM:s p5-flaggskepp p5-595, som slår 570 mer eller mindre i förhållande till det högre antalet kärnor i benchmarkkonfigurationerna.[8] Den andra är en HP Superdome. Denna ger cirka 20 procent högre transaktionshastighet, men det krävs 64 Itanium 2 enkärniga processorer (single core) för att uppnå detta – fyra gånger fler än hos 570.[9]

TPC-C-resultaten på dessa nivåer påminner allt mindre om belastningar eller transaktionshastigheter som förekommer i verkligheten.[10] Kostnaden för att köra ett sådant benchmarkvärde är enorm, kräver en ofantlig specialiststab och utrustning för miljontals dollar. TPC-C fortsätter dock vara en allmän måttstock av olika anledningar: det genomgår oberoende granskning, det innefattar en priskomponent, det är vanligt förekommande och erbjuder många möjligheter till jämförelse med konkurrenter, och det tar hänsyn till många funktioner i systemkonstruktionen. Så även om det inte är ett perfekt mått – det finns det inget benchmarkvärde som är – så är TPC-C absolut ett relevant mått. Detta gör att 570-systemets senaste resultat blir mycket beaktansvärt.

Ett modulsystem dessutom

Det finns förstås många servrar före p5-570 som skalanpassats och fungerat väl. Det som verkligen är ovanligt är att p5-570 klarar detta samtidigt som det är en modulkonstruktion.

Grunden i en p5-570 är en 4U ”processorlåda” - som kan konfigureras som ett fristående system med upp till fyra kärnor på upp till två POWER5+-processorer. Varje processor är förpackad som en DCM-modul (Dual Chip Module) som kombinerar ett cachechip på 36 MB level 3 (L3) med processorchipset POWER5+, som innefattar två SMT-kärnor,[11] en 1,9 MB level 2 (L2)-cache, den integrerade minnesstyrenheten, katalogen för L3-cachen och den distribuerade switchlogikkretsen som ansluter till andra moduler. Jämfört med föregångaren POWER5 körs POWER5+ på högre klockhastigheter (upp till 2,2 GHz) och utrymmeskraven kan minskas med 37 procent vilket ger lägre effektförbrukning genom att växla från en 130 nm- till 90 nm-process. Den har

7 En IBM eServer pSeries 690 slår också 570 med en hårsmån, men det systemet har ersatts av p5-590/595.

8 Dessa är dock äldre POWER5-kärnor; IBM har ännu inte släppt POWER5+ till sina största servrar.

9 Förvisso kör HP Superdome Windows som operativsystem och SQL Server som databas, vilket kanske inte kan förväntas fungera lika bra som ett kommersiellt UNIX-system med Oracle eller DB2. Senare versioner av Microsofts™ OS och DBMS har dock förbättrats betydligt avseende skalbarhet, och deltavärdet bör vara relativt blygsamt. Se även den engelska texten ”IBM’s X3 Heads Into the Yukon.”

10 Se även texten ”TPC-C Passes Escape Velocity.”

11 I förhållande till andra SMT-konstruktioner har IBM satsat på både ett begränsat antal trådar (två trådar per kärna) och en avancerad prioritetsmekanism för att hantera dessa trådar. Se även den engelska texten ”IBM Takes the Smart Road to Multi-threading.”

också en mängd andra prestandaförbättringar, bland annat en minnesstyrenhet med högre prestanda, stöd för nya sidformat och dubbla antalet TLB-poster (Translation Lookaside Buffer), vilket används för att förbättra prestandan vid översättning av virtuella adresser.[12]

För att koppla samman flera processorlådor används en flexibel SMP-kabel som ansluter till det distribuerade switchnätverket för respektive POWER5+-processor. Det finns en unik uppsättning kablar för varje systemstorlek, så att de externa anslutningarna ska kunna optimeras på så många sätt som möjligt. SMP-bussen som går över dessa kablar sammanför logiskt alla block i en 570 till en enda SMP-server. Blocken är dessutom sammankopplade via en SP-buss som körs över en SP Flex-kabel som överför systemstyrsignaler som t.ex. JTAG, I2C och olika klockor.

Varje 4U-låda innehåller ett internt I/O-undersystem med sex PCI-X-platser (hot-plug), sex diskplatser (hot-swap), integrerade styrenheter och redundanta nätaggregat (hot-swap) och kylfläktar. De stöder också RIO2, IBM:s senaste generation av externa I/O-länkar, som stöder anslutning av ytterligare externa PCI-X- och lagringsplatser (utanför systemchassit).

Varje 570-låda är en helt fristående enhet och en fullutrustad server med fyra kärnor. Men den kan också utökas stegvis genom att helt enkelt lägga till ytterligare lådor och anslutningskablar.

Slutsats

Exemplet p5-570 visar tydligt att modulsystem numera kan konstrueras utan att prestanda behöver offras i någon större grad.

Det är ett tillägg till de övriga möjligheterna som systemet delar med andra alternativ i p5-serien. Capacity on Demand (CoD) för processor och minne gör att extra kärnor eller minnesenheter kan aktiveras permanent eller tillfälligt, antingen manuellt eller automatiskt, baserat på tillfälliga arbetslastkrav. 570 stöder också APV (Advanced POWER Virtualization) inklusive aktivering av firmware för mikropartitionering och VIOS (Virtual I/O Server), där flera partitioner kan dela samma uppsättning I/O-resurser och därmed minska deras kostnader och platsbehov över flera dynamiska logiska partitioner (DLPAR). Med APV kan en partition vara så liten som en tiondel av en POWER5+-kärna, och att dedicera en SCSI-styrenhet eller en nätverkanslutning till denna är ofta onödigt.

570 är moduluppbyggd, skalbar och ger höga prestanda – en vinnande kombination. POWER5+-processorn ger enastående prestanda i detta moduluppbyggda system. Prestandan skalanpassas i princip linjärt efter processorantalet. Samtidigt innebär modulsystemet att det blir fullt möjligt att starta i liten skala – även om man har planer på att bygga ut till ett stort system. Möjligheten att starta med köp av ett litet system och sedan bygga på efter behov, i princip utan prestandaförluster, ger en tydlig fingervisning om hur servrar kommer att konstrueras i framtiden, både stora servrar och mellanklasservrar.

---

[12] POWER5+ stöder också den nya QCM-modulen (Quad-Core Module) med fyra kärnor, en typ av multichipmodul i miniatyr. Den här enheten har två POWER5+-chip och två cache-chip på ett enda stycke keramiskt substrat, en lösning som är DCM-kompatibel. Den ger goda prestanda för pengarna, men till priset av lägre prestanda per kärna. IBM säljer huvudsakligen QCM-enheter p.g.a. pris-/prestandafördelar i enklare system, och erbjuder för närvarande inte dessa för 570-systemet. (En minimering av överlappandet mellan 570 och de bästa p5-systemen kan också ha påverkat IBM:s beslut.)

Through subscription research, advisory services, speaking engagements, strategic planning, product selection assistance, and custom research, Illuminata helps enterprises and service providers establish successful information technology.