Nahaufnahme: IBM Power

Timo Schöler

Die Power Prozessoren von IBM bilden die Basis für Hochleistungsworkstations und Superrechner. Ein Überblick über Technik und zugrunde liegende Philosophie jenseits von AMD und intel.

Bekannt als 'großer Bruder' des PowerPC waren die Power Chips von IBM stets in leistungsstarken Workstations und Servern der Linie RS/6000 zu finden; seit dem Power3 von 1997 implementiert diese Prozessorserie die 64 Bit PowerPC Architektur. Die 2001 erschienenen Power4/Power4+ Prozessoren führten das Dual Core-Prinzip ein, der Power5/Power5+ unterstützt nun Multithreading – dieser Beitrag soll die verwendeten Technologien wie Single- und Multithreading, Bussysteme und Energiesparmaßnahmen beschreiben.

Die Power5+ CPU baut auf dem mit Power4 eingeführten Dual Core-Prinzip auf und besteht ebenfalls aus zwei Kernen auf einem einzelnen Die, wobei jeder Kern darüber hinaus zwei Threads unterstützt – der Prozessor präsentiert sich dem Betriebssystem als Vierwege-CPU. Auf den 1,875 MByte großen L2 Cache können beide Kerne unabhängig zugreifen, durch Einteilung in drei jeweils zehnfach assoziativ verwaltete Partitionen werden Zugriffe möglichst einheitlich über den gesamten L2 Cache verteilt. Jeder Schreib- oder Lesezugriff auf den L2 Cache wird von 'Cache coherency engines' kontrolliert, deren Anzahl im Vergleich zum Vorgänger verdoppelt wurde. Die Anzahl der unterstützten Prozessoren konnte mit einem überarbeiteten Bussystem auf 64 verdoppelt werden und mit Simultaneous Multithreading (SMT) wurde die erreichbare Leistung weiter erhöht.

Der bei Power5+ bis zu 72 MByte große L3 Cache (Power5: 36 MByte) besteht ebenfalls aus drei Partitionen, wobei diese den jeweiligen Pendants im L2 Cache als 'Victim cache' zugewiesen sind – Daten werden nur in den L3 Cache geschrieben, wenn sie aus dem L2 Cache verdrängt werden, während eine Referenz zu Daten im L3 Cache dazu führt, daß sie in den L2 Cache geladen werden; nur modifizierte Daten werden in den Arbeitsspeicher geschrieben, unmodifizierte verfallen. Dank des in der Power5 CPU befindlichen L3 Verzeichnisses kann sehr zügig erkannt werden, ob Daten im Cache liegen oder nicht (sind die Daten nicht im Cache, spricht man von einem 'Cache miss'). Zugriffe auf den L3 Cache finden über zwei unidirektionale, 128-bit breite Busse statt, die mit halbem Prozessortakt betrieben werden. Der in die CPU integrierte Speichercontroller erspart – ähnlich dem L3 Verzeichnis – die sonst anfallenden Laufzeiten zu externen Controllern: Zugriffe des Speichercontrollers auf den Systemspeicher laufen über zwei unidirektionale Busse, für Schreibzugriffe 64 Bit, für Lesezugriffe 128 Bit breit, mit doppelter Taktfrequenz der eingesetzten Speichermodule.

Wie der Power4+ wurden die ersten Power5 Prozessoren in einem 130nm SOI (Silicon-On-Insulator)-Prozess hergestellt, wobei den 267mm^2 umfassenden 184 Millionen Transistoren des Power4+ ganze 276 Millionen Transistoren des Power5 gegenüberstehen, die mit 389mm^2 bei gleichem Produktionsprozeß entsprechend mehr Fläche auf dem Die einnehmen; beinahe noch beeindruckender ist die Anzahl der Pins eines MCM (Multichip Module) – insgesamt 5.370 Pins stellen die Verbindung zum

System her, 2.313 Pins für Datentransfer (hiervon 92% für SMP- Funktionen sowie L3- und Speicherbusse), 3.057 zur Energieversorgung. Mittlerweile wird der im Herbst 2005 präsentierte Power5+ im 90nm SOI-Prozeß hergestellt. Durch Umstellung auf 90nm mit dem Power5+ konnte die Chipfläche auf etwa 245mm² geschrumpft werden – mit verbesserten Energiesparfunktionen liegt die Leistungsaufnahme bei nur noch etwa 70 Watt. Im Februar 2006 präsentierte IBM die Midrange-Serie p5 570 und 575, in denen der Power5+ mit 2,2 GHz arbeitet.

Power5 Prozessormodule können, abhängig von der beabsichtigten Anzahl der Prozessoren im System sowie deren Topologie, in verschiedenen Bauformen auftreten: Als Multichip Modul (MCM), bestehend aus vier Power5 CPUs und vier L3 Cache Chips oder als Dual-Chip Modul (DCM), das einen Power5 Prozessor und einen L3 Cache Chip umfaßt, sowie seit dem Power5+ als Quad-Core-Modul (QCM), das aus zwei Power5+ CPUs – also vier Kernen – und zwei L3 Cache Chips besteht. Je nach Konfiguration können ein bis acht DCM (maximal 16 CPUs), bis zu vier QCM oder zwei bis acht MCM (maximal 64 CPUs) zusammengefaßt werden. Der Datenaustausch untereinander findet mittels eines auf jedem Prozessor befindlichen 'Fabric bus controller' (FBC) statt, insgesamt stellt sich das System als verteilter Switch dar – diese Funktionsweise ist allen Konfigurationen gemein. Systeme bis sechzehn Prozessoren können sowohl in MCM- als auch in DCM- oder QCM-Konfiguration auftreten, wobei High-End-Systeme in der Regel MCM einsetzen; zwei MCM können in diesen Systemen ein sogenanntes 'Book' formen, welches aus den beiden Multichip Modulen sowie Speicher und weiterer Logik besteht. Ein solches 'Book' bildet ein sechzehnfaches SMP-System, das sich dem Betriebssystem als 32fach präsentiert (zwei Threads pro Kern).

Die Verbindung jedes Power5 Chips zum L3 Cache, zum Speicher und Ein- /Ausgabe besteht aus separaten Paaren unidirektionaler Busse; während der 128 Bit breite Bus zum L3 Cache mit halber Prozessortaktfrequenz läuft – also mit steigender Taktfrequenz der CPU skaliert, operiert der als 'GX Bus' bezeichnete, jeweils 64 Bit breite Bus für Eingabe-/Ausgabeoperationen mit einem Drittel des CPU-Taktes. Die Power5 CPUs eines MCM untereinander sind ebenfalls über zwei Busse in Ringform verbunden, die ebenfalls unidirektional und mit vollem CPU-Takt betrieben werden.

Die Begrenzung von Power4-Systemen auf maximal 32 physikalische Prozessoren auf 16 Chips wurde durch tiefgreifende Änderungen der Systemstruktur aufgehoben – oberhalb von 32 CPUs könnte die Kommunikation zwischen den Prozessoren derart zunehmen, daß ein Leistungszuwachs nicht mehr garantiert werden konnte. Um diesen Engpaß zu umgehen, wurde im Power5 Design der L3 Cache enger an den Prozessor angebunden, er ist nun nicht mehr auf Seite des Hauptspeichers an die Fabric angebunden, sondern verfügt über einen eigenen Port – so können bereits beschriebene L2 Cache misses über den 72 oder 36 MByte großen L3 Cache abgefangen werden, ohne die Busse zwischen den Prozessoren zu belasten. Durch die engere Anbindung des L3 Cache im Vergleich zu Power4, der Integration des Speichercontrollers in den Prozessor sowie einem einzelnen Chip für den L3 Cache (statt zwei Chips für 32MByte im Power4) ergaben sich folgende Vorteile:

- Reduzierung der Latenz zu L3 Cache und Systemspeicher,

- erhöhte Transferleistung zu L3 Cache und Systemspeicher,

- erhöhte Zuverlässigkeit durch Reduzierung der Anzahl an Chips.

Ein extrem wichtiges Element zur Verbesserung der Skalierbarkeit von Power5-Systemen ist der 'Fabric bus controller' (FBC), der auf jedem Power5-Chip zu finden ist. Zum Arbeitsbereich des FBC gehören Koordination und Ansteuerung von L2 und L3 Cache, Anbindung zum Hauptspeichersubsystem sowie der Busse zwischen den einzelnen Prozessoren eines MCM und der Multichip Module untereinander. Der sogenannte 'Fabric bus' besteht aus drei unabhängigen Systemen, dem Adreßbus, dem Datenbus und dem Steuerbus, der als ein Zusatz zum Adreßbus verstanden werden kann.

Der Adreßbus zwischen Multichip Modulen ist 64 Bit breit und läuft mit halbem Prozessortakt, während er auf den MCM selbst 32 Bit Breit ist und mit vollem Takt läuft – der erreichbare Durchsatz ist gleich. Transfers auf dem Adreßbus sind ECC-geschützt, ein einzelner Transfer nimmt vier Taktschritte in Anspruch. Eine Transferoperation enthält eine Adresse (50 Bit) sowie weitere Angaben zum Transfertyp, die die Operation auf dem Bus eindeutig erscheinen lassen.

Während innerhalb der einzelnen MCM zwölf Punkt-zu-Punkt-Verbindungen als Adreßbus genutzt werden, baut die Kommunikation zwischen den MCM auf einer Ringstruktur auf. Adreßbusoperationen fließen (in einem 64 CPU-System) durch alle acht MCM, wobei jedes Modul die empfangenen Informationen sowohl an das nächste MCM, als auch an den eigenen, inneren Adreßbus weitergeben muß; sobald ein MCM eine selbst gestartete Adreßbusoperation auf dem Bus sieht, wird diese nicht weitergeleitet – der Ring ist geschlossen. Der Steuerbus ist von Struktur und Leistungsfähigkeit dem Adreßbus gleich, er läuft dagegen zu diesem etwas verzögert; auf ihm finden für Multiprozessorsysteme wichtige Operationen wie 'Cache snooping', 'snoop pushing' (Transfer eines Eintrages von dem Cache einer CPU in den Cache einer anderen, anfordernden CPU) und ähnliches statt.

Der Datenbus innerhalb der Multichip Module wird, wie der Adreßbus, mit vollem Prozessortakt betrieben, skaliert also vollständig mit steigender Taktfrequenz. Im Gegensatz zum Adreßbus, der innerhalb von MCM als Punkt-zu-Punkt-Bus aufgebaut ist, besteht der Datenbus aus zwei gegenläufigen, unidirektionalen Ringen. In Systemen mit mehreren MCM bauen sich weitere, modulübergreifende Ringe auf – in einem 64 Prozessor-System finden sich vier separate Ringe, die die einzelnen Power5 Prozessoren der MCM untereinander verbinden. Über den ebenfalls ECC-geschützten Datenbus finden ausschließlich Datentransfers von und zum Hauptspeicher, Abgleiche der Caches und ähnliches statt (zum Beispiel ein 'cast out' – ein modifizierter Eintrag im Cache wird in den Hauptspeicher oder rangniedrigeren Cache geschrieben; unmodifizierte Einträge verfallen). Im Unterschied zu Power4 Systemen wurde die Anzahl der Datenbusse innerhalb und zwischen den einzelnen Multichip Modulen verdoppelt, zusätzlich wurden weitere Ringe (vertical-node data busses) aufgebaut, um Latenzen innerhalb der 'Books' zu verringern.

Der im Prozessor selbst befindliche Speichercontroller greift über eine synchrone Speicherschnittstelle, die von SMI-Chips bereitgestellt wird (SMI – Synchronous Memory Interface), auf Speicher des Typs DDR oder DDR2 zu; mit Einführung des

Power5+ wurde die Taktfrequenz von angebundenem DDR2-Speicher von 266MHz auf 533MHz erhöht, die Verbindung zwischen SMI-Chips und dem Prozessor läuft weiterhin mit doppeltem Speichertakt. Die Schnittstelle zwischen einem Power5 Speichercontroller und den angebundenen SMI-Chips besteht aus drei Bussen, einem Addreß-/Steuerbus sowie je einem unidirektionalen Bus zum Lesen und Schreiben von Daten. Die SMI-Chips können vom Speichercontroller in Zweier- oder Vierergruppen angesprochen werden und sitzen in unmittelbarer Nähe zu den Speicherbausteinen (im DIMM-Format); sie stellen weiterhin Puffer zur Verfügung, um die unterschiedlichen Busbreiten zwischen Prozessor und SMI beziehungsweise SMI und Speicher auszugleichen. In einer Konfiguration mit vier SMI-Chips und einem Speichercontroller ist der Bus zum Schreiben von Daten 64 Bit breit, jeder SMI-Chip mit 16 Bit an diesen Punkt-zu-Punkt-Bus angebunden – in einer Konfiguration mit zwei SMI-Chips auf einen Speichercontroller wird nur die Hälfte der Bandbreite genutzt, 32 Bit liegen brach. Der mit 128 Bit doppelt so breite Bus zum Lesen von Daten verbindet demzufolge die einzelnen SMI mit jeweils 32 Bit Punkt- zu-Punkt, bei Verwendung von nur zwei SMI werden diese aber jeweils 64 Bit breit angeschlossen.

Neben der üblichen ECC-Technologie (Error Checking and Correction), die einzelne Bitfehler korrigiert und Fehler zweier oder mehrerer Bits erkennt, wird eine 'Memory scrubbing' genannte Technik verwendet: Der Speichercontroller liest im Hintergrund stetig Speicher aus, korrigiert auftretende Fehler (einzelner Bits) und schreibt den Inhalt zurück – hiermit soll verhindert werden, daß sich korrigierbare Einzelfehler auf nicht korrigierbare Fehler mehrerer Bits ausweiten können. Zudem steht pro Vierergruppe von Speichermodulen ein Modul als Reserve bereit, das den Ausfall eines Moduls transparent überbrücken kann, indem das System das fehlerhafte ausblendet und das Reservemodul einbindet – 'Chipkill' steht für diese Technik.

Energieumsatz und der Aufwand zur Kühlung heutiger Prozessoren waren ebenfalls wichtige Punkte des Power5-Designs; zusätzlich zu dem vor allem bei weiterhin schrumpfenden Chipstrukturen bekannten Problem der steigenden Leckströme weist die Power5 CPU im Verhältnis zum Vorgänger einen gesteigerten Anteil aktiv schaltender Elemente auf – der zu zahlende Preis für Multithreading. Auch die Anzahl der Busse wurde erhöht, so daß der Power5 Prozessor insgesamt bei gleichem Takt sowohl eine höhere Gesamtaufnahme als auch eine höhere Energiedichte aufweist – Power4+ und Power5 als 130nm-Chips angenommen. Um dem Abhilfe zu verschaffen, nutzt der Power5 intensiv fein regulierbare 'clock-gating' Mechanismen; hierbei werden Elemente vom Taktgeber getrennt, wenn abzusehen ist, daß sie im kommenden Taktzyklus nicht genutzt werden – wobei einzig das Schreiben nicht möglich ist, gespeicherte Werte können weiterhin ausgelesen werden. Während der Entwicklung der CPU wurden spezielle 'Power modeling tools' eingesetzt, um verschiedene Entwürfe der diese Funktionen steuernden Algorithmen mit typischen Anwendungsszenarien durchzurechnen und das Optimum für Teilbereiche oder die ganze CPU bestimmen zu können. Der Schwerpunkt lag hier auf der unbedingten Vermeidung von Einbußen der Rechenleistung sowie zeitkritischer Pfade auf dem Chip, die auf Abschaltungen durch clock-gating zurückzuführen wären. Die Schaltungen für die Steuerung des clock-gating an sich wurden bewußt möglichst einfach gehalten. Zum Schutz gegen thermische Überlastungen sind vierundzwanzig Temperatursensoren an strategischen Stellen verteilt. Bei Überschreiten des Grenzwertes an einem dieser Sensoren wird ein zweistufiges Verfahren gestartet, um die Situation zu normalisieren: In Stufe eins wird

die Schaltgeschwindigkeit der betroffenen Elemente herabgesetzt und mit ihr der Energieumsatz reduziert – dies wird durch wechselweises Schalten zwischen normalem Betriebszustand und 'stall condition', einem künstlichen Ausbremsen der überlasteten Elemente, erreicht. Wird durch die erste Stufe das Problem nicht innerhalb eines vorgegebenen Zeitraums behoben, greift Stufe zwei, die dann weitergehende Funktionen bremst (wie fetch, dispatch und completion) und damit die Ausführungseinheiten gleichsam lahmlegt, indem der Strom neuer Instruktionen nachläßt. Der Schutzmechanismus an sich benötigt keinerlei Unterstützung von Betriebssystem oder Service Prozessor, bietet jedoch einen sehr zeitnahen und flexiblen Ansatz zur Störungsvermeidung, ohne große Leistungseinbussen hervorzurufen.

Die bereits in den Vorgängermodellen zu findenden Elemente zur Steigerung der Zuverlässigkeit und Wartbarkeit des Systems (RAS – Reliability, Availability, Serviceability) – Erkennung von Fehlern, ihre Behebung oder Isolation vom Rest der Prozessors, sowie Ersatz der defekten Komponente im laufenden Betrieb – waren schon in Power4-Systemen sowohl in der CPU als auch in übrigen Komponenten (Arbeits- und Festplattenspeicher) implementiert. Im Entwurf des Power5 wurde die Absicht betrieben, die Gesamtstabilität auch unter Vermeidung von geplanten Wartungsfenstern weiter zu erhöhen – ein Großteil der Firmware-Upgrades kann ohne Reboot, im laufenden Betrieb stattfinden. Fehlerkorrektur mittels ECC wird nunmehr nicht nur in den Caches, sondern auch auf allen Bussen angewandt, Ein-Bit-Fehler automatisch korrigiert. Sollte der Fehler wiederholt und dauerhaft auftreten, kann eine Reparatur bis zum Wartungsfenster aufgeschoben werden. Ist der Fehler nicht zu beseitigen und stört den normalen Betrieb, kann das System heruntergefahren werden, das betroffene Element (zum Beispiel ein defektes 'Book') 'offline' geschaltet werden und das System startet neu – ohne Eingriff des Administrators.

Während die Leistungsfähigkeit moderner Prozessoren durch verschiedene Techniken wie größere Caches, Sprungvorhersage (Branch Prediction) und spekulative Ausführung von Befehlen (out-of-order execution) gesteigert werden konnte, erwies sich die niedrige Auslastung der einzelnen Prozessoreinheiten von durchschnittlich 25 Prozent als Hauptproblem; seit einigen Jahren versuchen diverse Hersteller dieses Problem hardwareseitig mit Hilfe von Multithreadingfunktionen zu lösen: Hierbei scheinen dem Betriebssystem auch bei Einzelprozessorsystemen (im Regelfall) zwei CPUs zur Verfügung zu stehen, die Auslastung der einzelnen Prozessoreinheiten läßt sich durch ausgeklügelte Mechanismen optimieren – im Idealfall steigt die Gesamtleistung. Mindestens drei Methoden Multithreading lassen sich unterscheiden:

- Coarse-grain Multithreading beschränkt sich auf einen einzelnen Thread zu jeder gegebenen Zeit und ist mit relativ geringem prozessorseitigen Mehraufwand machbar. Ergibt sich eine Situation, in der der laufende Thread stehenbleibt, etwa durch einen Cache miss – die zu verarbeitenden Daten sind nicht im Cache und müssen langwierig aus dem Hauptspeicher gelesen werden – schaltet der Prozessor auf den zweiten Thread um, Leerläufe der Ausführungseinheiten werden vermieden und der Gesamtdurchsatz steigt. Der minimale Mehraufwand ergibt sich durch den hohen Anteil der Elemente wie Register, die von beiden Threads genutzt werden können und nicht doppelt vorgehalten werden müssen.

- Fine-grain Multithreading hingegen schaltet Threads unabhängig von ihrer Laufzeitsituation wechselweise aktiv, auch obliegt die Ausführung der verschiedenen Threads allein dem Prozessor. Im Gegensatz zur Verfahrensweise des Coarse-grain Multithreading können kurze Leerläufe der Pipeline überbrückt werden, indem ein anderer Thread geschaltet wird. Die Effizienz nimmt mit der Anzahl der Threads zu, es können aber weiterhin Situationen auftreten, in denen die Ausführungseinheiten aufgrund eines auf Daten wartenden Threads nicht weiterarbeiten können; die dem Thread zugeteilte Rechenzeit bleibt ungenutzt.

- Simultaneous multithreading (SMT) unterhält ebenfalls mehrere Threads zur selben Zeit, kann aber tiefer in das Geschehen eingreifen: In jedem Moment – zu jedem einzelnen Takt – können Befehle verschiedener Threads gleichzeitig auf den Ausführungseinheiten abgearbeitet werden. Muß ein Thread auf Daten warten, werden die von ihm genutzten Einheiten so lange, bis der Thread weiterlaufen kann, zusätzlich den anderen Threads zur Verfügung gestellt. Die hierzu nötigen weiteren Logikelemente führen zu einer höheren Komplexität, der erreichte Leistungszuwachs rechtfertigt dies aber.

Der Power5 Prozessor bietet mit Simultaneous Multithreading die Ausführung zweier Threads pro Kern an, kann aber auch in einem Modus zur Ausführung nur eines einzelnen Threads betrieben werden – diesem stehen damit mehr Ausführungselemente zur Verfügung.

Durch die Erweiterung um symmetrisches Multithreading mußten Teile der CPU überarbeitet werden, andernfalls hätte die erwartete Mehrleistung nicht zuverlässig erbracht werden können: Mit Rücksicht auf die benötigte Chipfläche, vorgegebenen Grenzen bezüglich des elektrischen Bedarfs und der Umsetzung der Energie in Hitze wurden einzelne Elemente des Prozessors erweitert oder neu entworfen. Die L1 Caches, 64 KByte für Instruktionen und 32 KByte für Daten, blieben gleich groß, wurden jedoch statt zweifach im Power5 nun vierfach assoziativ ausgelegt; Einträge in den Caches können von beiden Threads genutzt werden. Die Instruction Pipeline des Power4 wurde identisch übernommen, ebenso konnten die Laufzeitcharakteristika (latencies, branch misprediction penalty, load-to-use latency) trotz höherer Komplexität einzelner Ausführungseinheiten (Queues) gleich bleiben.

Die Länge und Struktur einiger Queues wurde hinsichtlich der jeweils besten Balance zwischen Leistung, Energieumsatz, Komplexität und benötigter Fläche auf dem Die optimiert. Die 'Global Completion Table' (GCT) wurde überarbeitet, jeder Thread kann unabhängig alloziert werden: Instruktionen werden vor ihrem Weg durch die Pipeline gruppiert – hierdurch kann die Koordinierung mit weniger komplexen Einheiten erfolgen. Kontrollinformationen für jede in Ausführung befindliche Gruppe werden zu Ausführungsbeginn in der GCT abgelegt. Jeder Eintrag enthält die Adresse der ersten Instruktion der jeweiligen Gruppe; die Tabelle selbst folgt logisch der Programmfolge jedes Threads. Ist die Ausführung der Gruppe vollständig, wird dies im Eintrag der GCT vermerkt, nach Ausführung und bereits erfolgter Verarbeitung der Ergebnisse verfällt der Eintrag. Während die Einträge selbst der Programmfolge der Threads folgen, können sie beliebig zwischen beiden Threads umsortiert werden. Auch die Anzahl der physischen Register nahm im Verhältnis zur Power4 CPU zu (von 80 für Integer auf 120, von 72 FPU-Registern auf 120); jedem logischen Register wurde ein

weiteres Bit zugewiesen, aus dem die Zugehörigkeit hinsichtlich der laufenden Threads hervorgeht. In den ausführlichen Testszenarien, die IBM durchlaufen lies, um die optimalen Werte für Größe und Auslegung der einzelnen Einheiten zu finden, zeigte sich auch, daß nicht überall Änderungen notwendig waren: Die 'Branch Information Queue' (BIQ) wurde mit sechzehn Einträgen vom Vorgänger übernommen, da sie sich immer noch als ausreichend groß erwies. Die BIQ erhält Informationen über einzelne Verzweigungen vor der eigentlichen Ausführung der Instruktionen (zum Zeitpunkt 'instruction fetch'). Erweist sich der (spekulativ) durchgerechnete Zweig als gültig, verwirft die BIQ die dazugehörigen Einträge, andernfalls kann sie den Ausgangszustand wiederherstellen. Eine vergrößerte BIQ böte keine Vorteile – der Chip würde ohne Nutzen komplexer und größer.

In Systemen, die mehrere Threads verarbeiten, sollten alle Ausführungseinheiten möglichst ausgelastet sein, ohne jedoch den Programmfluß zu stören – leider kann ein Thread einen anderen Thread vollkommen blockieren. Um dies zu verhindern, verfügt der Prozessor über Funktionen, die stetig alle vorhandenen Ausführungseinheiten überwachen und gegebenenfalls eingreifen, übersteigt der Anspruch eines Threads einen vorbestimmten Grenzwert – die oben beschriebene Global Completion Table ist Teil dieser Maßnahmen, die IBM 'Dynamic resource- balancing logic' nennt. Ein Beispiel: Thread A benötigt zur Abarbeitung einige Daten, die weder im L1 Cache noch im L2 Cache stehen und fordert diese an; währenddessen sind die von Thread A belegten Ausführungseinheiten belegt, können jedoch nicht weiterarbeiten – Thread B kann aufgrund dieser Blockade zur Ausführung anstehende Befehlsgruppen nicht ausführen. Idealerweise bevor dieser Fall eintritt, wird Thread A so lange zurückgestellt und Thread B Priorität zugeteilt, bis Thread A die benötigten Daten erhalten hat und weiterarbeiten kann. Ähnliches gilt für von Threads beanspruchte Ressourcen – übersteigen diese einen Grenzwert, beispielsweise die Anzahl der Einträge in der Global Completion Table, wird eingegriffen.

Abhängig von der jeweiligen Laufzeitsituation kann der Prozessor einen von drei Mechanismen nutzen, um regulierend einzugreifen:

- Reduzierung der Priorität des Threads: Dieser hauptsächlich angewandte Mechanismus greift vor allem bei Threads, die zu viele Befehlsgruppen gleichzeitig fahren (und damit unter anderem die Global Completion Table füllen);

- Reduzierung des Dekodierens neuer Befehle des Threads findet vor allem bei Threads statt, die auf einen hohen Anteil an nicht im Cache vorhandenen Daten zurückgreifen müssen;

- Stoppen des Dekodierens neuer Befehle des Threads als letzte Maßnahme, vor allem bei Threads die in Befehlen festhängen, die sehr lange zur Ausführung brauchen (sync des Dateisystems zum Beispiel).

Jeder Power5 Kern kann beiden Threads individuelle Prioritäten auf einer von acht Ebenen zuteilen, wobei das Setzen der Priorität seitens der Software – sowohl durch das Betriebssystem als auch durch Applikationen – stattfindet, die Einhaltung wird dann in Hardware erzwungen. Es gibt verschiedene Szenarien, in denen eine ungleiche Priorität der Threads wünschenswert ist: Einerseits durch den Prozessor selbst gesetzt, um

Blockaden des Kerns durch einen einzelnen Thread zu verhindern, andererseits in konkreten Anwendungsfällen, in denen beispielsweise einer Echtzeitapplikation hohe Priorität, einem Daemon – etwa einem smtp-Server – eine niedrigere Priorität zugewiesen wird. Die Prioritäten werden als Level 0 bis Level 7 vergeben, wobei Level 0 einem ruhenden Thread entspricht. Eine besondere Bedeutung kommt Level 1 zu, der niedrigsten Priorität eines laufenden Threads: Sind beide Threads dem Level 1 zugewiesen, reduziert der Prozessor das Dekodieren neuer Befehle, der Kern arbeitet mit reduzierter Leistung und spart Energie.

Nicht alle Anwendungen profitieren von Multithreading, beispielsweise Programme mit extrem hohen Anforderungen an die Speicherbandbreite oder Code, dem durch die Leistungsfähigkeit einzelner Ausführungseinheiten ein oberes Limit gesetzt wird. Der Power5 Prozessor läßt sich hierzu auch einen Modus zur Ausführung eines einzelnen Threads (Single Thread, ST Mode) zu, hierbei werden diesem alle Register, Queues, die Global Completion Table und weitere Einheiten des Kerns komplett zur Verfügung gestellt – schon durch die große Anzahl an zur Verfügung stehenden physischen Registern zur Registerumbenennung (32 logischen Registern stehen 120 nutzbare physische Register zur Verfügung) führt die Power5 CPU einzelne Threads schneller aus als der Vorgänger mit gleicher Taktfrequenz. Ebenso wie die Zuweisung der Priorität kann zwischen Betrieb im Multithreading oder Singlethreading softwareseitig umgeschaltet werden.

IBM liefert mit dem Power5(+) einen sehr imposanten Prozessor; nicht erst seit der Vorstellung mit bis zu 2,2 GHz getakteter CPUs im Februar 2006 sind Power5 basierte System in weiten Bereichen das schnellste, was zu haben ist. Ein Beispiel sind die Tests der SPEC-Benchmarks für das IBM System p5-575 im Vergleich zu ebenfalls mit acht Prozessoren/Kernen bestückten Systemen (von den acht Doppelkernen der p5-575 ist jeweils ein Kern abgeschaltet, der aktive Kern greift somit auf knapp 1,9 MByte L2 Cache und 36 MByte L3 Cache zurück): Ergebnis 196 SPECint_rate_base2000 und 382 SPECfp_rate_base2000. Eine Siemens Primergy 650 mit SPARC 64V CPUs erreicht 133 SPECfp_rate_base2000, eine Sun V40z mit vier Dual-Core Opteron 880 kommt auf 140 SPECfp_rate_base2000.

Power im Internet

- SSH-Zugang auf einer Power5-Maschine im Rahmen des 'OpenPower Project': http://www-128.ibm.com/developerworks/linux/openpower/

- PowerPC Architecture Book: http://www-128.ibm.com/developerworks/eserver/articles/archguide.html

---

http://riscworks.net
eMail timo.schoeler@riscworks.net

http://timo-schoeler.de