Letztes Prozessorgeflüster

Von Ruhe und Bewegung

Zum ersten Mal in der 29-jährigen Geschichte der International Conference for High Performance Computing, Networking, Storage and Analysis (SC) wurde diese von einem Nichtamerikaner geleitet. Auf der SC17 gab es allerlei Neues von Intels Roadmap, von NECs Vektorprozessor, von ARM64-Servern und mehr.

Von Andreas Stiller

Bernd Mohr vom Jülich Supercomputing Centre (JSC) hatte die Ehre – und die Last – der erste internationale Chairman der SC zu sein. Drei Jahre dauerte die Vorbereitung, für die er vom JSC zu 50 Prozent freigestellt wurde. Für reichlich Gesprächsstoff sorgte diesmal Intel schon kurz vorher – nicht mit einem neuen Produkt, sondern mit einer Abkündigung. Der Xeon Phi in der bisherigen Form hat keine Zukunft mehr: Knights Hill wird es nicht mehr geben.

Das war eigentlich spätestens seit dem geänderten Vertrag für den Aurora-Supercomputer an den Argonne Labs schon klar, der statt eines mit Knights Hill bestückten Rechners für 2018 jetzt einen Exaflops-Rechner für 2021 mit neuer Architektur vorsieht. Der Abschied manifestierte sich auch auf Intels HPC Developer Conference am Wochenende vor der SC17. Zwar gab es hier noch Sessions zu Xeon Phi Knights Landing (KNL), aber in den gezeigten Roadmaps tauchte er meist nicht mehr auf. Gleichzeitig bloggte Trish Damkroger, Leiterin der Intel Data Center Group und General Manager der Technical Computing Initiative: „Ein Punkt über den wir [auf der SC17, Anm. d. Red.] reden werden, ist, einen der zukünftigen Intel-Xeon-Phi-Prozessoren (Codename Knights Hill) durch eine neue Plattform und neue Mikroarchitektur zu ersetzen, die speziell für Exascale designt ist."

Xeon Phi stirbt, Knights Mill (hier ein Prototyp auf Supermicro-Board) kommt aber noch – und zwar Ende des Jahres.

ICX-H mit HBM2

Darüber reden wollte Intel – aber nicht mit mir. Und so spitzte ich auch diesmal die Ohren im Convention Center in Denver, so wie schon vor 4 Jahren an gleicher Stelle, wo ich sehr zum Unwillen von Intels HPC-Chef Raj Hazra Details zum Xeon Phi Knights Landing aufschnappte und ausplauderte. Diesmal verriet Intel hinter den Kulissen die „Future direction of Intel Xeon und Intel Xeon Phi Processor and NVM solution". Und im geschwätzigen Convention Center war es abermals nicht schwierig, einiges davon mitzubekommen, wenn auch durch die rein mündlichen Übermittlungen in fremder Sprache ein paar kleinere Fehlerchen nicht ausblieben.

Demnach wird nicht nur, wie Trish im Blog aufführte, einer der zukünftigen Intel-Xeon-Phi-Prozessoren gecancelt, sondern die ganze Architektur-Linie. Es wird also auch der bereits in der Entwicklung befindliche Nachfolger Knights Peak gestrichen – allerdings nicht ersatzlos. Denn 2019 soll der Xeon Ice-Lake in 10nm+-Technologie herauskommen (ICX, nicht ISX wie ich online veröffentlicht hatte), und zwar nicht nur als normaler Scalable Xeon ICX-SP, sondern auch in der Ausführung ICX-H (Codename Knights Cove) mit 4 Stacks à 8 GByte High Bandwidth Memory (HBM2) mit bescheidenen 650 GByte/s Bandbreite.

Den Plattformnamen Whitley für ICX mit dem (bislang ungeklärten) Zusatz „with lewi" hatte Intel schon vor Wochen auf der Website verraten, ebenso wie den zukünftigen Serverprozessor Sapphire Rapids in der Tinsley-Plattform.

Den normalen ICX-SP – vermutlich ohne HBM2 – soll es mit bis zu 38 Kernen geben, mit 8 Kanälen DDR4-3200, deren Bandbreite bei zwei DIMMs pro Kanal allerdings auf 2933 sinkt. Hinzu kommen drei UPI-Links à 11,2 GBit/s, was dann insgesamt einen P+-Sockel mit 4184 Anschlüssen erfordert, 90 mehr als AMDs Epyc.

Die H-Version kommt mit 38 und 44 Kernen, wobei letzteres ein Multichip-Modul aus zwei Dies mit je 22 Kernen sein wird, intern verkoppelt über drei UPI-Links. Nach der ersten Performance-Einschätzung soll das Modul im Linpack etwa 40 Prozent schneller sein als ein ICX-SP mit 24 Kernen.

Zumindest der H, vermutlich aber auch der SP wird mit allen AVX512-Erweiterungen des Knights Landing versehen sein, sodass vorhandene Xeon-Phi-Software weiterläuft – nur muss man dann sicherlich neu optimieren. Hinzu kommt auch die AVX512-Erweiterung VNNI (Vector Neuronal Networks Instructions), womit er auch Nachfolger des für Deep Learning weiterhin vorgesehenen Knights Mill wird, dessen eigentlich geplanter Nachfolger Knights Bay ebenfalls gestrichen wurde. Daneben hörte man auch noch so Namen wie Ice Age und Knights Run für zukünftige Prozessoren.

Vektorrechner

Aber Intel bekommt jetzt Konkurrenz. So hat NEC den Vektorrechner SX-Aurora Tsubasa (Flügel) auf PCIe-Karte fertig. Er wird jetzt bemustert und soll ab Ende Februar 2018 auf den Markt kommen.

Der Prozessor hat acht Kerne mit einem gemeinsamen Cache von 16 MByte.

Jeder Kern speichert 64 Vektorregister mit jeweils 256 DP-Gleitkommawerten. 32 davon können pro Takt in den drei FMA-Einheiten verarbeitet werden, das macht 192 Flops/Takt/Core.

Er kommt zunächst in drei Ausführungen. Version 10A ist das (wassergekühlte) Flaggschiff mit 1,6 GHz Takt und 48 GByte HBM2. Damit schafft die Karte 2,45 TFlops in doppelter Genauigkeit bei einer Speicherbandbreite von 1,2 TByte/s.

Version 10B hat die gleiche Ausstattung, aber nur 1,4 GHz Takt. Die DP-Performance liegt bei 2,15 TFlops. Es gibt sie auch in einer luftgekühlten Ausführung. Version 10C mit aktiver oder passiver Luftkühlung hat ebenfalls 1,4 GHz Takt und 2,15 TFlops, aber nur 24 GByte HBM2 mit 0,75 TByte/s.

In der Stream-Performance ist die SX-Aurora Tsubasa 10B etwa fünfmal so schnell wie ein Skylake-System mit zweimal Xeon Gold 6148. Beim Linpack liegen beide etwa gleichauf, im Preis ebenfalls. Damit müsste er bei etwa 6000 US-Dollar liegen.

Anwendungen für Vektorrechner in C/C++ und vor allem in Fortran gibt es zuhauf, etwa bei den Klimaforschern – nur ist fraglich, ob sie mit dem begrenzten Speicher von 48 GByte werden auskommen können, denn zusätzlichen DDR4-Hauptspeicher bietet der Vektorrechner nicht. Braucht die Anwendung mehr als 48 GByte, muss sie träge über das PCIe-3.0-Interface auf den Hauptspeicher des Gastsystems zugreifen. Auch die Kommunikation zu weiteren Karten findet über PCIe 3.0 statt. Zu einem eigenen Link à la Nvlink oder zu dem im Oktober endgültig spezifizierten PCIe 4.0 konnte sich NEC leider nicht aufraffen.

ZettaScaler und Pezy-SC2

Da waren die japanischen Kollegen von ExaScaler mutiger. Ihr Pezy-SC2-Beschleuniger bietet bereits PCIe 4.0 – nur das aktuelle Gastsystem mit Xeon-D-1540 noch nicht. Und während NECs Vektorrechner das Deep Learning Trendformat fp16 nicht unterstützt, gehört auch dieses zum Repertoire von Pezy-SC2.

In den Liquid-Immersion-Cooling-Computern von ZettaScaler sorgten die Pezy-Systeme für große Aufmerksamkeit, konnten sie doch die ersten drei Plätze in der Green500-Liste erobern und kamen mit dem Gyoukou (Morgenlicht) des Forschungsinstituts RIKEN immerhin auf Platz vier der Top500-Liste der Supercomputer (siehe S. 20).

NEC will Skylake-Rackeinschübe mit bis zu 8 SX-Aurora-Tsubasa-Karten ab März 2018 anbieten.

In den Tanks mit sehr edlem Fluorinert von 3M (Fluorocarbon FC-43) stecken kompakte Xeon-D-Systeme mit jeweils acht Pezy-SC2-Karten. Jede dieser Karten besitzt sechs 64-bittige MIPS-Kerne sowie 2048 Processing Elements, von denen aber aus Ausbeutegründen nur 1984 freigeschaltet sind. Diese Rechenkerne arbeiten mit achtfachem Hyper-Threading. Bei FMA können sie 2 DP-, 4 SP- oder 8 FP16-Operationen pro Takt ausführen. Bei 1 GHz sind das rund 4 DP-TFlops bei 180 Watt. Zum Vergleich: Nvidias Tesla V100 (SMX) kommt auf 7,5 DP-TFLops bei 300 Watt. Volta ist also ein wenig energieeffizienter, aber ZettaScaler hat das sparsamere Gastsystem und das effizientere Kühlkonzept, sodass Nvidias DGX SaturnV Volta mit 15,1 GFlops/Watt von den (kleineren) ZettaScaler-Systemen mit 16,7 bis 17 GFlops/s knapp in der Green500-Disziplin übertroffen wird.

Der Pezy-SC2-Chip hat derzeit vier DDR4-3200-Speicherkanäle mit einer Gesamtbandbreite von 100 GByte/s. Doch im nächsten Jahr will ExaScaler das zusätzliche Wide-I/O-Speicherinterface mit 1024 Bit Breite fertig gestellt haben, mit dem dann über das „drahtlose“ Thru-Chip Interface (TCI) mit vier DDR4-Kanälen bis zu 2 TByte/s Bandbreite möglich ist – damit hängt man alle anderen HBM2-Stacks locker ab. Das von der Firma Thru-Chips des legendären Transmeta-Gründers Dave Ditzel und des japanischen Professors Tadahiro Kuroda entwickelte TC-Interface arbeitet mit induktiver Kopplung im Nahfeld. Es soll nicht nur schneller und platzsparender als die Stack-Technik mit Thru Silicon Vias (TSV) sein, sondern auch energieeffizienter.

Und die CPU-Konkurrenz?

Ja, auch die CPU-Konkurrenz kommt in Gang. HPE hat jetzt famose SPECfp-Werte für das Zweisockelsystem DL385 Gen 10 mit AMD Epyc 7601 gemeldet – und endlich, endlich wird auch die SPEC-CPU2017-Seite gefüllt. Danach liegt das Epyc-System in SPECfp2017_rate_base

Das Pezy-SC2-Modul in der Hand eines seiner Entwickler. Es leistet 4 TFlops in doppelter Genauigkeit.

Student Cluster Competition

Gleich 16 Teams traten diesmal zum Wettbewerb an. In einem Fotofinish gewannen die Studenten der Nanyang Technological University aus Singapur vor den hochfavorisierten Chinesen der Tsinghua University. Mit ihren acht Tesla-V100-Karten erzielten die Singapurer unter anderem einen fabelhaften Linpack-Wert von 51,77 TFlops, 40 Prozent mehr als der bisherige Rekord, den Studenten der FAU-Uni Erlangen-Nürnberg zur ISC17 erzielten. Das ist bei zugelassenen 3000 Watt in der Effizienz sogar geringfügig besser als beim Spitzenreiter der Green500. In Zukunft wird die Top500-Organisation auch eine zusätzliche Top50-Liste führen, die die Ergebnisse der Studentenwettbewerbe – zur amerikanischen SC, zur europäischen ISC und der asiatischen ASC – enthält. Das deutsche Team der Friedrich-Alexander- und der Technischen Universität München wurde übrigens Opfer eines Hackerangriffes von außen und verlor den Anschluss.

mit 257 Punkten knapp vor einem Cisco-System mit dem über doppelt so teuren Xeon Platinum 8180. Den Vorsprung hatten wir schon vor ein paar Monaten festgestellt. Das System soll im Dezember angeboten werden. Dell ist mit dem PowerEdge R745 für Epyc noch nicht so weit, zeigte ihn aber auf der SC16 als Prototyp „Sneek Peek“.

Das lang erwartete IBM-Power9-System mit Nvidia Tesla V100 soll ebenfalls ab 1. Dezember ausgeliefert werden, und zwar in der luftgekühlten Version mit vier V100 (SMX), im Februar soll das wassergekühlte System mit sechs V100 folgen.

So weit sind die neuen 64-bittigen ARM-Prozessoren für Server noch nicht – aber es gibt immerhin Prototypen. Qualcomm selbst war zwar nicht mit einem eigenen Stand auf der SC17, doch am ARM-Stand konnte man Qualcomms Centriq 2400 (Codename Falkor) im Centriq Open Compute Motherboard bewundern, der erste Serverchip in 10 nm – leider nicht in Betrieb. Anders als sein Cavium-Kollege beschränkt sich der 48-Kerner auf nur einen Sockel mit sechs Speicherkanälen. Partner aus der ersten Liga außer Microsoft mit dem Projekt Olympus kennt man derzeit noch nicht. Das sieht beim ThunderX2 von Cavium ganz anders aus, hier stehen sie bereits Schlange: Bull/Atos, Cray, Gigabyte, HPE und Penguin – alle zeigten ihre ersten ThunderX2-Prototyp-Systeme. Microsoft hat kurz vor der SC17 angekündigt, den ThunderX2 ebenfalls ins Projekt Olympus für Azure-Server aufzunehmen. Dieser ThunderX2 ist allerdings nicht der mit 56 Kernen, der ursprünglich mal von Cavium geplant war. Nach dem Aufkauf des ARMv8-Designs Vulcan von Broadwell hat der ehemalige Chefarchitekt des Xeon Phi, Avinash Sodani, das Beste aus zwei Welten zusammengerührt. Zunächst hat der Mischprozessor aber nur 32 Kerne.

Cray hat dabei den großen Vorteil, nicht nur mit der Cray XC50 mit ThunderX2 aufwarten zu können, sondern auch mit eigenen hochoptimierenden Compilern. Diese sollen bei einem Großteil der HPC-Applikationen gut 20 Prozent schneller sein als die GNU- oder LLVM-Compiler. Performance-Daten wollte noch keine Firma nennen, die Uni Bristol hat aber schon ein bisschen geplaudert und den aktuellen 32-kernigen Prototyp mit 2,5 GHz Takt im Cray-Scout-System gegen einen Broadwell Xeon E5-2695v4 (2,1 GHz) und Skylake Xeon Gold 6152 (2,1 GHz) im Single-Socket-System laufen lassen. Das sieht bei speicherintensiven Anwendungen nicht schlecht aus, im Stream dank der 8 Speicherkanäle sowieso (35 Prozent schneller als Skylake), aber auch bei OpenFoam und anderen. Bei rechenintensiven Anwendungen wie Gromacs obsiegen hingegen die Intel-Prozessoren, da ist schon der Broadwell um 17 Prozent schneller.

HPE hat die Apollo 70 im dichten Formfaktor fertig mit zwei Prozessoren und acht Speicherkanälen. Mit einem ähnlich aussehenden Bull-System Sequoia 1310 baut das Supercomputercentrum in Barcelona sein Mont-Blanc-Projekt Dibona auf, das 96 Prozessoren mit mindestens 3072 Kernen haben wird. Und sofern die Katalanen in Spanien und damit in der EU bleiben, können die europäischen Partner via PRACE auch daran teilhaben.

So weit, so gut, doch kurz nach der SC17 wurde Cavium von Marvell für 6 Milliarden Dollar aufgekauft – da muss man jetzt erst einmal abwarten, wie Marvell sich bei den Serverprozessoren positionieren will. Und vielleicht kauft ja Broadcom in einem Mega-Deal für geschätzt 130 Milliarden Dollar Qualcomm und hat dann wieder einen ARM64, nachdem man die Eigenentwicklung an Cavium abgestoßen hat – ein hübscher Ringtausch. Und wer weiß, wohin noch das XGene-3-Design von Applied Micro System gelangt, das Besitzer Macom mit dem gesamten Compute Business unlängst an eine Firma „Project Denver Holdings“ verhökert hat, die zur Carlyle Group gehört.

… und tschüss

Nicht nur vom Xeon Phi – dessen Pentium-artige Architektur übrigens zuerst hier im Prozessorgeflüster (c’t 15/2008) verraten wurde und nun kommen passenderweise Details zu seinem Ende – muss man sich verabschieden: Nach 34 Jahren c’t und über 500 Prozessorgeflüstereien und vielen Ausplaudereien entschwinde ich nun in den Ruhestand. Ich bedanke mich bei meinen treuen Lesern, vielleicht gibt es ja doch noch den einen oder anderen Artikel. Und wer das A20-Gate liebt, wird mich schon finden. *(as@ct.de)* c't