Prozessorgeflüster

Von Mode und neuen Kleidern

Ja, das ist Konkurrenzkampf! Mitten im Sommer führen Intel und AMD nahezu gleichzeitig ihre neuen Serverprozessorkollektionen vor. Prêt-à-porter ist aber noch nicht alles. Zukunftsmode sind auch noch die nächsten Xeon-Phi-Prozessoren, vor allem der Knights Hill, der vielleicht noch umdesignt wird.

Von Andreas Stiller

Früher waren für große Prozessor-events Frühjahr oder Herbst üblich und dann auch nur selten gleichzeitig. Nun kamen sie beide zur besten Ferienzeit. Wie gut, dass das österreichische Ötztal so ausgezeichnet mit WLAN versorgt ist, selbst in den Bussen und Wanderhütten. Nur für den Familienfrieden war das gelegentliche Remote-Benchmarking während des Urlaubs nicht übermäßig förderlich. Und dann seilte ich mich in Innsbruck auch noch einen Vormittag ab, um statt das goldene Dachl lieber das (idyllisch gelegene) Institut für Quantenoptik und Quanteninformation zu besuchen, dort wo die renommierten Professoren Rainer Blatt, Peter Zoller und Rudolf Grimm forschen und lehren.

Zum spannenden Quantenthema später mal mehr, jetzt stehen erst mal intensivere Benchmark-Orgien mit klassischen Programmen an, zwei Xeon-SP-Systeme und ein AMD-Epyc-7601-System wollen befeuert werden. Erste Ergebnisse dazu finden Sie auf Seite 60. Vielleicht schaffe ich es ja für die nächste c't-Ausgabe, die neue SPEC-Suite CPU2017 auf den Chips zum Fliegen zu bringen.

Lange pokerten die beiden Firmen auch mit den Preisen. Intel fühlte sich offenbar gezwungen, hier den ein oder anderen Abstrich insbesondere im Mittelfeld zu machen – da hatte man im Vorfeld von anderen ursprünglich geplanten Preisen gehört. Daran sieht man schon jetzt, wie nötig es ist, dass auf dem Servermarkt endlich mal wieder Konkurrenz herrscht. Im High-End-Desktop-Markt hat AMD mit dem 16-Kerner Threadripper ja schon mal klare Signale gesetzt. Da wird sich Intel für den Skylake-X wohl noch andere Preise einfallen lassen müssen.

So wirklich kaufen kann man AMDs Epycs allerdings im Unterschied zu den meisten neuen Xeons noch nicht. Auf Heise Preisvergleich findet man „noch kein Anbieter“, so gibts auch noch keine Straßenpreise. Größere Mengen im B2-Step hatte AMD wohl auch nicht eingeplant, wie man durchgetunnelten moderaten Wafer-Bestellungen bei Globalfoundries entnehmen kann. Den ein oder anderen Bug, der derzeit noch Workarounds etwa bei der I/O Memory Management Unit (IOMMU) nötig macht, möchte man wohl auch noch beseitigen, bevor die breite Marktauslieferung beginnt.

Intel könnte mit der Skalierbarkeit nicht nur auf die Hardware abzielen, sondern hat mittelfristig vielleicht für bestimmte SKUs auch Upgrades per Software im Sinn, so wie es IBM mit „Capacity on Demand“ für die Power-Linie schon lange vorsieht. Dann könnte man auf größeren Sockel-Support, mehr Speicher oder mehr L3-Cache upgraden oder mehr Kerne zubuchen, vielleicht auch so wie bei IBM nur zeitweise etwa am Monatsende.

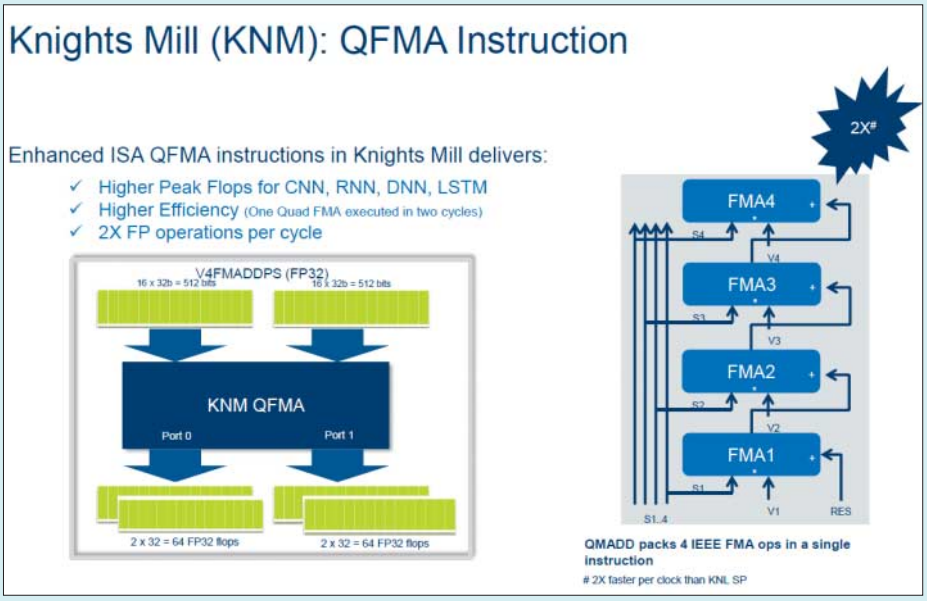

Bild: Intel

QFMA packt vier IEEE-FMA-Operationen in eine Instruktion und verdoppelt damit die FMA-Performance.

Vierfach … und mehr

Vergleiche zwischen Intels und AMDs neuen Server-Flaggschiffen sind natürlich recht schief, sind erstere doch mehr als doppelt so teuer. Da kann man beim AMD Epyc 7601 noch getrost eine Nvidia Tesla P100 zulegen, beim Zweisockelsystem dann gleich zwei – und bekommt Linpack-Werte in ganz anderen Regionen. Gegen die nackte Rechenpower einer P100 können jedenfalls auch 28 Kerne mit AVX512 nicht konkurrieren, da braucht man schon mindestens 72, so wie beim Xeon Phi Knights Landing. Dessen AVX512 sieht aber ein bisschen anders aus als das AVX512 der neuen Xeons. Gemeinsam sind nur die Grundfunktionen (AVX512 Foundation), dann teilt es sich auf, reziproke und exponentielle Funktionen hier,

Byte- und Word-Instruktionen da. Der Xeon SP kennt darüber hinaus auch AVX512VL, womit sich SSE und AVX2 auf 32 Register „aufblasen“ lassen. Nachfolger Cannon Lake soll dann zusätzlich FMA mit Integer sowie erweiterte Bit-Manipulationen bieten.

Der für Jahresende geplante Knights Mill wird neue Vector Neural Network Instructions (VNNI) hinzufügen für den Umgang mit fp16 bei interner 32-Bit-Rechengenauigkeit. Außerdem wird es auch Befehle für Quad Fused Multiply Add (QFMA), für Single Precision (4FMAPS) und für fp16 (QVNNI) geben.

Mit QFMA bekommt man gleich vier FMA-Operationen pro Instruktion, die dann aber zwei Takte zur Ausführung benötigen. Das verdoppelt immerhin bei SP die Performance, verglichen mit dem aktuellen FMA.

Für die bislang sehr teuren Xeon-Phi-Prozessoren hatte Intel Ende Juni die Preise zum Teil drastisch gesenkt, beim 72-Kerner sogar auf 3368 Dollar mehr als halbiert. Zwischenzeitlich hat Intel auch ein paar PCI-Express-Coprozessorkarten mit Xeon-Phi-Knights-Landing ins Angebot genommen – die Nachfrage hierfür hält sich aber in Grenzen.

Der Nachfolger Knights Hill in 10-nm-Technik mit 88 (vielleicht auch 90) Kernen dürfe sich weiter nach hinten verschieben – für den eigentlich damit fürs nächste Jahr geplanten Supercomputer Aurora kommt er definitiv zu spät. Aurora taucht im Unterschied zu den mit IBM Power9/Nvidia Volta bestückten Systemen Summit und Sierra in der Budget-Planung des DOE für 2018 gar nicht mehr auf, da wird mysteriös umschreibend nur von dem ALCF (Argonne Leadership Computing Facility) Upgrade Project gesprochen, das auf eine „Advanced Architecture“ verlagert werden soll, die besonders gut für Deep Learning geeignet ist und die irgendwann mit mehr als ein Exaflop Performance herauskommen soll. Sollte das etwa Knights Mill mit fp16-Exaflops sein?

Man hört jedenfalls davon, dass Knights Hill vielleicht noch umdesignt wird, um wettbewerbsfähiger zu sein. Bislang sollte die FMA-Performance wie beim Knights Mill mit QFMA, allerdings auch für DP verdoppelt werden. Mit nur 16 Kernen mehr und etwas höherem Takt hätte man sonst die geplante Verdreifachung der Performance nicht hingekriegt.

Nun, nach der Präsentation von Nvidias Tensor-Unit im 21-Milliarden-Transistor-Chip Volta, so raschelt es, sei Intel sehr beeindruckt und plane möglicherweise eine ähnliche Einheit für Knights Hill, dann natürlich nicht nur für fp16, sondern auch für Single und Double Precision. Aber Nvidia schläft nicht und der Volta-Nachfolger Ampere (ups, der Name ist ja noch geheim) könnte sich in einem Jahr vielleicht verstärkt wieder den doppeltgenauen Berechnungen samt passend aufgebohrter Tensor Unit widmen.

(as@ct.de) c't

AVX512

Kategorie	Beschreibung	CPUID 7	Xeon SP	Xeon Phi
AVX512F	AVX-512 Foundation	EBX:16	✓	✓
AVX512DQ	AVX- 512 Dword and Qword Instructions	EBX:17	✓	–
AVX512IFMA	Fused Multiply Add of Integers using 52-bit precision.	EBX:21	Cannon Lake	–
AVX512PF	AVX-512 Prefetch Instructions	EBX:26	✓	✓
AVX512ER	AVX-512 Exponential and Reciprocal Instructions	EBX:27	–	✓
AVX512CD	AVX-512 Conflict Detection Instructions	EBX:28	✓	✓
AVX512BW	AVX-512 Byte and Word Instructions (BW)	EBX:30	✓	–
AVX512VL	AVX-512 Vector Length	EBX:31	✓	–
AVX512VBMI	AVX-512 Extended Bit Manipulation	ECX:01	Cannon Lake	–
AVX512_VPOPCNTDQ	AVX-512 POPCOUNT Dword/Qword	ECX:14	–	Knights Mill
AVX512_4VNNIW	AVX-512 Vector Neural Network Instructions	EDX:02	–	Knights Mill
AVX512_4FMAPS	AVX-512 Fused Multiply Accumulation Packed Single precision	EDX:03	–	Knights Mill

✓ vorhanden – nicht vorhanden