Martin Fischer, Andreas Stiller

Die Superrechner

Intel, Nvidia und AMD stellen auf der SC12 gleichzeitig ihre neuen Hochleistungsrechenkarten vor

Nvidia setzte alles dran, bei den Supercomputern Nummer eins zu werden – und schaffte es mit Hilfe von Cray und 18 688 Tesla K20X-Karten. Newcomer Intel drängt jetzt mit dem Xeon Phi und unglaublicher Verve in diesen Markt – und lässt sich das einiges kosten. Doch auch AMD zeigt mit der FirePro S10000, dass sie mitspielen wollen.

Die Firma Intel musste erstmal reichlich Lehrgeld zahlen, hatte man doch ursprünglich vor, Nvidia und AMD auf dem Grafikmarkt herauszufordern. Das dafür gedachte Projekt Larrabee scheiterte. Unter anderem weil Intel erkennen musste, dass es nicht reicht, konkurrenzfähige Chips herauszubringen – man muss auch die Software-Schreiber und die Community an seiner Seite haben. In der Grafikwelt sind das vorrangig die Spiele-Entwickler. Die dachten nicht im Traum daran, ihre bewährten Pfade zu verlassen und auf Intels Programmierparadigmen zu wechseln. Außerdem hatte sich Intel bei Grafiktreibern noch nie mit Ruhm bekleckert.

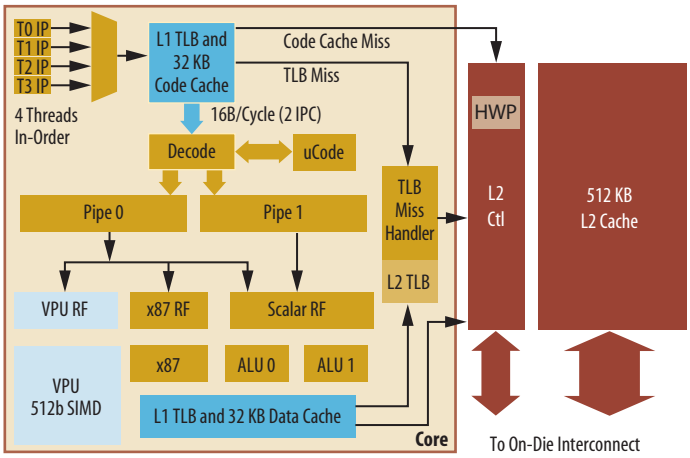
Ganz anders sieht die Sache im High Performance Computing (HPC) aus. Hier genießt Intels Softwareabteilung einen hervorragenden Ruf. Sie beschäftigt auch die bei weitem größte Compiler-Mannschaft, zu der auch die eingekauften DEC-Entwickler gehören, die schon vor langer Zeit solch schöne Dinge wie Auto-Vektorisierung und Auto-Parallelisierung erfunden hatten. Und mit dieser Qualität will Intel nun wuchern, wenn es darum geht, den Larrabee-Nachfolger Xeon Phi schmackhaft zu machen.

Die von Larrabee übernommene Grundidee besteht in einem großen Verbund kleiner x86-Kerne, ein jeder versehen mit einer leistungsfähigen Vektoreinheit. Jeder Kern besitzt einen L2-Cache von 512 KByte, der mit den anderen über einen Ringbus kommuniziert. Larrabee-Chefentwickler Doug Carmean fragte sich anfangs offenbar: Warum einen neuen Kern erfinden, wenn man einen effizienten Kleinkern doch in der Schublade hat – den ursprünglichen Pentium. Ein paar wichtige Updates musste der Pentium-Kern aber über sich ergehen lassen. Das machte die Anpassung von Linux etwas aufwendig, denn das rechnete nicht damit, dass es mal einen 64-bittigen Pentium mit Powermanagement-Funktionen und vierfachem Hyper-Threading, aber ohne I/O-Befehle geben würde.

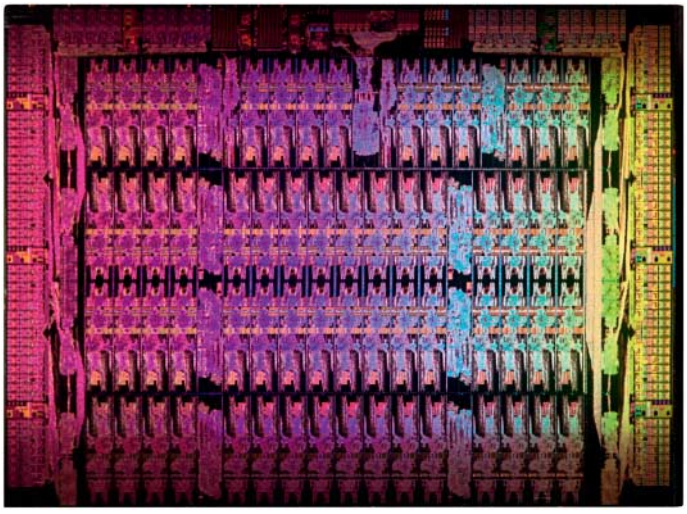
Der jetzt herausgekommene Xeon Phi ist eine Weiterentwicklung des Larrabee mit bis zu 62 x86-Kernen und ebenso vielen Vektoreinheiten mit 512 Bit Breite. Intel schwört Stein und Bein darauf, dass es keine 64 sind, von denen dann zwei nur als Reserve dienen. Als Beweis veröffentlichte der Chip-Gigant jetzt das Die-Plot, auf dem tatsächlich nur 62 Stück zu finden sind. Die nicht benötigten Grafikteile, die Larrabee noch hatte, wurden eliminiert. Zur Herstellung des Fünf-Milliarden-Transistor-Chips nimmt man nun auch nicht wie bei Larrabee einen alten Prozess, sondern den allerneusten mit 22-nm-Strukturen und Trigate-Transistoren. Nur fürs Interface hatte man das Update vergessen und so kommt der Xeon Phi noch mit dem antiken PCI Express 2.0 heraus – ein Umstand, der bei Intel hinter den Kulissen für heftige Debatten gesorgt hatte. Allerdings spezifiziert auch Nvidia den K20 vorsichtshalber nur für PCIe 2.0, weil es mit dem ein oder anderen Xeon-E5-Board Probleme gab. Nvidia unterstrich aber, dass die Hardware PCIe 3.0 unterstütze, das Grafikkarten-BIOS die Karten jedoch auf PCIe 2.0 festsetze. So stehe es den OEMs frei, für ihre Systeme K20-Karten mit „freigeschaltetem" PCIe 3.0 einzusetzen.

Xeon Phi kann man, anders als bisher GPUs, eigenständig betreiben. Ein kleines Embedded Linux läuft auf einem der Kerne und bindet die anderen als Co-Prozessoren ein, deren Arbeit man ganz einfach über den Befehl `top` sehen kann. Mit seinen bis zu 8 GByte Speicher läuft der Phi dann völlig unabhängig vom Hauptprozessor und holt sich die Daten wie jener direkt von der Festplatte oder vom Netz. Doch ein Phi kommt selten allein, im Cluster und bei Nutzung von MPI wäre heutzutage PCI Express 3.0 durchaus angesagt.

In manchen der Top500-Supercomputer befinden sich zum Teil noch Xeon-Phi-Prototypen mit anderen Kernzahlen und Taktfrequenzen, als sie in den ersten beiden Produktversionen herauskommen werden. So beherbergt der schnellste dieser Supercomputer, der Stampede an der Texas-Universität in Austin, derzeit rund 2000 Phis mit



Xeon Phi: eine Vielzahl kleiner Kerne nach Pentium-Vorbild, aber 64-bittig, mit vierfach Hyper-Threading, Powermanagement, L2-Cache und einer leistungsfähigen 512-Bit-Vektoreinheit.



Das Die des Xeon Phi wurde lange geheim gehalten: Man findet tatsächlich nur 62 Kerne.

Die Karte, die Xeon-Phi-Marketingleiterin Reinders in den Händen hält, ist eine Spezialausführung für den Stampede-Supercomputer. In den werden derzeit pro Tag etwa 80 Karten eingebaut.

Bild: Nvidia

Nvidias Tesla K20 will mit hoher Double-Precision-Leistung punkten und schluckt maximal 225 Watt. Ihre GK110-GPU besteht aus 7,1 Milliarden Transistoren.

60 Kernen und 1091 MHz. Als erstes Marktprodukt ist für Januar der Xeon Phi 5110P mit 1053 MHz vorgesehen. Später soll dann der Xeon Phi 3110 folgen, der nur 57 Kerne, weniger und langsameren Speicher, aber höheren Takt hat und auf eine TDP von 300 Watt kommt. Der Xeon Phi 5110P ist mit 225 Watt sparsamer und gleichauf mit Nvidias Tesla K20. Bei der theoretischen Spitzenleistung liegt er aber mit seinen 1,011 TFlops knapp zurück.

Die grüne Keule

Nvidia sieht Intels Phi zwar als die einzig wirkliche Konkurrenz im HPC-Beschleuniger-Markt, hat aber eigentlich mehr Respekt vor Ivy Bridge und dem kommenden Haswell-Prozessor. Aber mit den beiden schon erwähnten Tesla K20 und K20X fühlt sich Nvidia gut aufgestellt. Letztere gibt's ausschließlich für Server, erstere auch als aktiv gekühlte Varianten für Workstations.

Selbstbewusst bläst Nvidia dabei ins Horn der Rechenleistung: Selbst die kleinere Ausführung Tesla K20 schafft 1,17 TFlops bei doppelter Genauigkeit und überholt damit den Phi 5110P um rund 15 Prozent. Die K20X legt noch einen drauf und zeigt mit 1,31 TFlops, wo der Hammer hängt. Dabei braucht letztere nur 10 Watt mehr als die K20. Noch größer ist der Vorsprung bei einfacher Genauigkeit: Mit 3,52 beziehungsweise 3,95 TFlops pro Karte ziehen die Teslas der Intel-Konkurrenz davon.

Ihre Kraft schöpfen sie aus der neuen GK110-GPU, die Nvidia-Boss Jen-Hsun Huang im Mai 2012 auf der GPU Technology Conference vorstellte. Erstmals hat Nvidia eine neue Hochleistungs-GPU nicht im Consumer-Segment auf GeForce-Karten eingeführt, sondern im HPC-Markt. Das könnte durchaus mit Intels Phi zusammenhängen – Nvidia wollte Intel auf der SC12 nicht die alleinige Aufmerksamkeit überlassen und musste obendrein das Oak Ridge National Lab und Co. bedienen. Dass die Spieler-Gemeinde nun grummelt und endlich auch GeForce-Karten mit dem neuen Chip sehen will, erträgt Nvidia ob der sprudelnden Gewinne ganz gut. Aber für später plant man, all die GK110-Chips, bei denen einige Kerne nicht funktionieren, weit preiswerter für GeForce zu vermarkten – so eine lukrative Zweitverwertungsmöglichkeit hat Intel nicht.

Nvidias Ingenieure haben mit GK110 einen wirklich riesigen Chip geschaffen, der aus 7,1 Milliarden Transistoren besteht und je nach Ausführung 2496 (K20) oder 2688 Rechenkerne (K20X) beherbergt. Die laufen im Vergleich mit GeForce-Chips vergleichsweise konservativ mit 705 beziehungsweise 735 MHz. Das Nachsehen hat Nvidia allerdings bei der Speicheranbindung – denn 512 Bit bietet nur Intel. Entsprechend mau fällt auch die Bandbreite aus: Die K20 und ihr 5 GByte großer GDDR5-Speicher kommunizieren mit 208 GByte/s, die K20X bindet seine 6 GByte immerhin mit 250 GByte/s an. Sowohl die Speicher als auch die Caches sind ECC-geschützt. Bezüglich der Transferrate hatten sich einige im Vorfeld mehr erwartet. Laut Nvidias HPC-Manager Sumit Gupta soll sie aber ausreichen – wichtiger war für Nvidia die Einhaltung der TDP.

Doch die neuen Tesla-Karten sind nicht nur leistungsfähiger, sondern bringen dank GK110 auch zwei wesentliche neue Funktionen mit: Dynamic Parallelism und Hyper-Q. Sie setzen Version 5.0 der CUDA-Schnittstelle voraus und speziell angepassten Code. Gerade bei etablierten Supercomputing-Algorithmen dürfte die Adaption daher dauern.

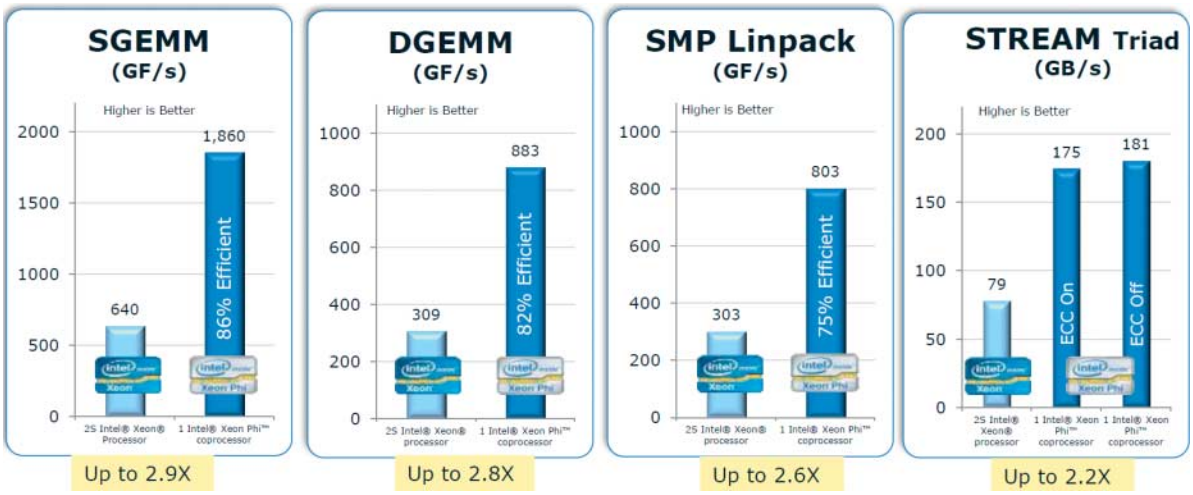


Bild: Intel

Performance-Werte einiger Klassiker auf dem Xeon Phi im Vergleich zu zwei Xeon-E5 – allerdings muss man das Kleingedruckte lesen, denn es handelt sich bei diesen Werten um einen schnelleren Prototypen (SE10P), der so nicht in den Handel kommt.

Dynamic Parallelism erlaubt es der GPU, in einem bereits laufenden Thread dynamisch neue Kernel zu erzeugen. Damit kann die GPU beispielsweise auch rekursive Funktionen ohne CPU-Hilfe abarbeiten. Um Leerlaufzeiten des Hauptprozessors zu vermeiden, verarbeitet die Tesla K20 via Hyper-Q gleichzeitig bis zu 32 MPI-Threads – die Fermi-Vorgänger schafften nur jeweils einen Prozess. Beispielsweise könnte ein 16-Kern-Prozessor also tatsächlich auch 16-MPI-Prozesse auf der GK110-GPU ausführen. Den Vorteil aus solch einer Konfiguration zeigte Nvidia anhand eines ausgewählten Code-Beispiels (CP2K), das 864 Wassermoleküle mit einem 16-Kern-Interlagos-System von AMD und einer Tesla K20 simuliert. Mit aktiviertem Hyper-Q lief der Code um den Faktor 2,5 schneller.

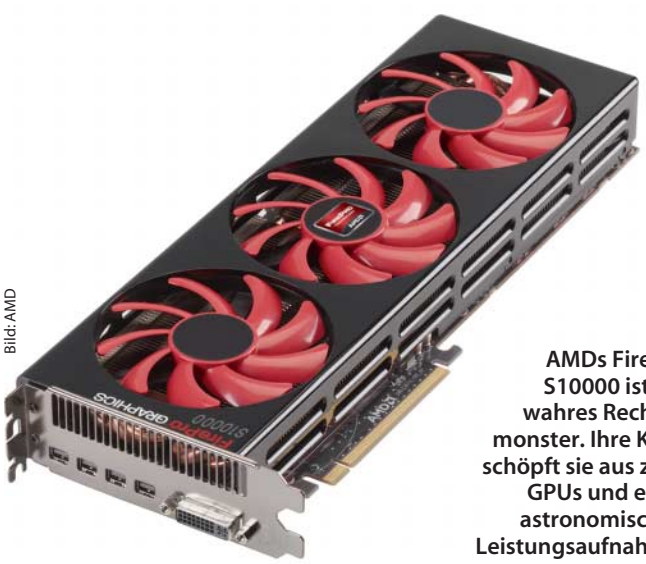


Bild: AMD

AMDs FirePro S10000 ist ein wahres Rechenmonster. Ihre Kraft schöpft sie aus zwei GPUs und einer astronomischen Leistungsaufnahme.

Effizienzhatz

Wie viel von der theoretischen Rechenleistung die Karten in der Praxis tatsächlich erreichen, ist noch etwas nebulös; beide Konkurrenten geben nicht genau spezifizierte Effizienzwerte von Linpack- zu Peakleistung inklusive Hauptprozessor an. Aus den Veröffentlichungen in der Top500-Liste kann man aber einiges ablesen.

Der Supercomputer Stampede bezieht derzeit nur rund 35 Prozent seiner Rechenleistung aus den Phi-Karten, den kann man damit nicht als Basis nehmen (und wenn, dann kämen schlechte Werte von unter 50 Prozent für den Phi alleine heraus). Intel hat einen eigenen Rechner in den Charts, dessen Phi-Karten eine Effizienz von 72 Prozent erreichen sollen. Aber dabei handelt es sich nicht um die Produktversion, sondern um einen 61-Kerner mit 1,1 GHz Takt. Der hat 1,073 TFlops Spitzenleistung und für diesen „Golden Release Candiate SE10P" gibt Intel 803 GFlops und damit 75 Prozent Effizienz an. Bei DGEMM soll er gar 883 GFlops erreichen. Für den Xeon Phi 5110P kann man den Linpack-Wert auf rund 756 und DGEMM auf 831 GFlops abschätzen. Und bei einfachgenauer Rechenleistung (SGEMM) liegt der K20 auch ohne „X" in ganz anderen Regionen.

Man findet in der Top500-Liste auch drei Supercomputer mit Xeon Phi 5110P, deren Effizienz zwischen 60 und 70 Prozent beträgt. Im gleichen Bereich rangieren auch die beiden K20X-Rechner, die auf 64 (Titan) beziehungsweise 70 Prozent (Todi) kommen, wobei die Coprozessoren hier jeweils gut 90 Prozent zur Gesamtrechenleistung beitragen. Gemeinsam mit den beiden Xeon-E5-Prozessoren kommen in den beiden NASA-Systemen, die mit Xeon Phi 5110P bestückt sind, 891 (Discovery) beziehungsweise 832 GFlops (Maia) pro Rechenknoten zusammen. Zum Vergleich: die Kepler-K20X-Karten im Supercomputer Titan, wo nur ein einziger Bulldozer-Prozessor pro Knoten mithilft, erreichen 941 GFlops/Knoten.

Nvidias eigenen Benchmarks zufolge soll eine einzelne K20X bei DGEMM-Berechnungen sogar 1,22 TFlops schaffen – das wäre dreimal so viel wie die M2090 (0,43 TFlops) und ein gewaltiger Effizienzsprung. Ein Nvidia-Mitarbeiter unterstrich dabei, dass dieser Wert tatsächlich allein von der GPU herrührt. Zum Preis der K20X äußerte sich Nvidia nicht. Die Workstation-Tesla K20 war bereits für 2950 Euro zuzüglich Mehrwertsteuer gelistet. PNY wird die Karte für 3090 Euro (ohne MwSt.) verkaufen. Sie sollte bereits Mitte November verfügbar sein, die Tesla K20X frühestens Ende November.

Xeon Phi 5110P wird jetzt an OEM-Partner ausgeliefert und soll ab dem 28. Januar 2013 für 2649 US-Dollar in den freien Handel kommen. Später im ersten Halbjahr 2013 soll dann Xeon Phi 3100 für unter 2000 US-Dollar folgen.

Und AMD?

Kurz vor Beginn der Supercomputing-Konferenz berief AMD eilig noch eine Telefonkonferenz ein, um die FirePro S10000 zu präsentieren – die Antwort auf Intels Phi und Nvidias K10 und K20. Sie unterstützt von Haus aus PCIe 3.0 und lässt sich auch zur Beschleunigung von Workstation-Grafik einsetzen. Im Unterschied zur Tesla lassen sich auch Displays anschließen und bis zu fünf Stück über vier Mini-DisplayPorts und einmal DVI gleichzeitig betreiben.

Die Rechenleistung der neuen AMD-Karte ist beachtlich: Knapp 6 TFlops bei einfacher und 1,48 TFlops bei doppelter Genauigkeit sind theoretisch drin – zur praktischen Effizienz macht AMD aber keine Angaben. Für die hohe Leistung müssen gleich zwei Tahiti-GPUs (aus je 4,31 Milliarden Transistoren) mit jeweils 1792 Rechenkernen auf der Karte schuften. Das treibt die Leistungsaufnahme aber in astronomische Höhen: 375 Watt soll die maximal betragen. In vielen Servern haben sich aber 225 Watt etabliert. Und das ist genau AMDs Problem: Im Vergleich mit Nvidias Tesla K10 ist die FirePro S10000 bei einfachgenauen Berechnungen 30 Prozent schneller – bei einer um zwei Drittel höheren Leistungsaufnahme. Bei doppelter Genauigkeit muss sie sich mit Nvidias Tesla K20 messen – dabei kommt ein ähnliches Verhältnis raus. Zum Preis des FirePro-Flaggschiffs machte AMD bis zum Redaktionsschluss noch keine Angaben. (as/mfi)

Varianten des Intel Xeon Phi

Bezeichnung	5110P	3100 (P und A)	SE10P/SE10X
Double-Precision-Performance (GFlops)	1011 GFlops	> 1000	1073
Kernzahl	60	57	61
Taktfrequenz	1053 MHz	nicht spez.	1100 MHz
GDDR5 Speichertransfers (GT/s)	5	5	5,5
Speicherbandbreite (GByte/s)	320	240	352
Speicherkapazität (GByte)	8	6	8
L2-Cache (MByte)	30	28,5	30,5
TDP (W)	225	300	300
Preis	**2649 US-$**	**<2000 US-$**	-

Neue Rechenkarten von AMD und Nvidia

Rechenkarte	Nvidia Tesla K20	Nvidia Tesla K20X	AMD FirePro S10000
GPU	GK110	GK110	2 × Tahiti
Fertigung	28 nm	28 nm	28 nm
Transistoren	7,1 Milliarden	7,1 Milliarden	2 × 4,31 Milliarden
Shader-Rechenkerne	2496	2688	2 × 1792
Rechengruppen	13 SMX	14 SMX	2 × 28 CUs
GPU-Taktfrequenz	705 MHz	735 MHz	825 MHz
Speicher	5 GByte GDDR5	6 GByte GDDR5	6 GByte GDDR5
Datentransferrate	208 GByte/s	250 GByte/s	480 GByte/s
ECC-Schutz	ja	ja	ja
Single-Precision-Leistung	3,52 TFlops	3,95 Tflops	5,91 TFlops
Double-Precision-Leistung	1,17 TFlops	1,31 Tflops	1,48 TFlops
Maximale Leistungsaufnahme	225 Watt	235 Watt	375 Watt
Ausführungen	Server, Workstations	Server	Server, Workstations
PCI Express	2.0[1]	2.0[1]	3.0

[1] Hardware beherrscht PCIe 3.0, aber im BIOS auf 2.0 festgesetzt