Prozessorgeflüster

Von vielen und von besonderen Kernen

Bisher schafft es niemand, Intels Vormachtstellung bei den Servern anzukratzen – jedenfalls wenn man US-Marktforschern glaubt. Doch an deren Sichtweise gibt es Zweifel und vielleicht kommt die Attacke auf das lukrative Server-Business aus einer ganz anderen Ecke.

Von Christof Windeck

Zehn Prozent Marktanteil: So lautete 2013 die Vorhersage von IHS iSuppli für Mikroserver im Jahr 2016. Diese Maschinen sollten zu einem erheblichen Anteil mit ARM- statt x86-Prozessoren bestückt sein. Daraus wurde aber nichts, stattdessen sinkt der Marktanteil der ARM-Systeme laut den Marktforschern von IDC schon wieder. Unverdrossen schicken APM und Cavium ihre jeweils zweiten Generationen von X-Gene- und ThunderX-SoCs ins Rennen (siehe S. 36). Laut Cavium-Manager Gopal Hegde sollen erste Muster des ThunderX2 freilich erst Anfang 2017 erscheinen.

Der streitbare Charlie Demerjian von SemiAccurate.com sieht die Marktanteile anders. Die US-Marktforscher Gartner und IDC seien blind für den chinesischen Markt, wo ARM-Server schon in erheblichen Stückzahlen liefen. Von Hannover aus können wir Server in China zwar auch nicht zählen, aber es fällt auf, dass AMD und Qualcomm für ihre ARM-Server-CPUs chinesische Partner gefunden haben. Diese wollen sich langfristig von US-Zulieferern unabhängiger machen und vielleicht auch Intels ominöser Management Engine aus dem Weg gehen, die in jedem Xeon-Chipsatz lauert. Diese ME, die außer für Fernwartung unter anderem auch für BIOS-Verdongelung und DRM zum Einsatz kommt, hält Intel nun schon seit zehn Jahren verschlossen wie eine Auster. AMD baut mit dem Platform Security Processor auf Basis von ARM TrustZone längst auch ein Trusted Execution Environment in APUs ein. Das ruft immer lautere Kritik hervor, etwa von deutschen Sicherheitsbehörden und aus der Open-Source-Szene. Dort trommelt man wegen der ME für „freiere“ Alternativen wie MIPS, ARM oder Power8.

Das Verteidigungsministerium der USA dürfte keine Probleme mit Intel ME und AMD PSP haben, sie stammen ja aus demselben Homeland. Für Waffensysteme verwendet man gerne Chips „Made in USA“ – wer weiß, was chinesische Zulieferer alles so einbauen. So hat das Department of Defense einen Zuliefervertrag mit Globalfoundries geschlossen: Der Chip-Auftragsfertiger betreibt nicht nur seine Fab 8 im Bundesstaat New York, sondern dort und in Vermont auch die 2015 zugekauften IBM-Fabs. Die hier einst gefertigten PowerPC-Prozessoren waren beim US-Militär lange beliebt. Dass Globalfoundries im Besitz des Emirats Abu Dhabi ist, scheint dabei kein Problem zu sein.

TPU für Deep Learning

Die US-amerikanischen Datenkraken Google und Facebook haben auch ein Problem mit Intel-Chips, aber eher wegen der hohen Preise. Die Cloud-Riesen liebäugeln mit ARM sowie OpenPOWER und Google kocht noch ein eigenes Süppchen, nämlich die selbst entwickelte Tensor Processing Unit (TPU) für Deep-Learning- und KI-Anwendungen.

Für ähnliche Zwecke, vor allem für die Erkennung von Sprache, offeriert nun die Firma KnuEgde ihren KnuPath „Hermosa“ mit 256 GFlops. Bisher hatte die vom ehemaligen NASA-Chef Dan Goldin 2005 als Intellisis in Austin gegründete Firma im Stillen fürs US-Militär Systeme zur Identifizierung von Sprechern entwickelt. Die Spracherkennung von KnuEdge können Firmen ausprobieren, nämlich in der Cloud-Anwendung knurld. Mit im KnuEdge-Boot sitzen zahlreiche Ex-AMD-Manager wie Pat Patla, der einst die Opteron-Sparte führte.

Chip-Auguren meinen, dass Spezialprozessoren wie Hermosa in Zukunft größere Marktanteile erobern – vielleicht nicht als separate Produkte, sondern als integrierte Hardware-Beschleuniger in anderen Prozessoren. Denn Allzweck-Prozessoren halten bei vielen Spezialaufgaben nicht mehr mit. Googles TPU soll typische Aufgaben in Suchmaschinen jedenfalls um ein Mehrfaches effizienter erledigen als x86-Chips, GPU-Rechenbeschleuniger oder sogar passend programmierte FPGAs. Das kennt man schon von Bitcoin-ASICs – die nebenbei bewiesen haben, dass selbst kleinere Firmen in endlicher Zeit Spezialbeschleuniger entwickeln und in Silizium gießen lassen können.

Das ist Intel nicht entgangen, dort hat man sich den FPGA-Spezialisten Altera einverleibt, um Kombiprozessoren zu entwickeln. Doch lassen sich die hohen Xeon-Preise halten, wenn es gar nicht mehr auf deren x86-Rechenleistung ankommt und Beschleuniger im Verbund mit anderen (ARM-)Rechenkernen billiger zu haben sind? Im Grunde hat ja auch AMD dieses Thema mit der Heterogeneous System Architecture (HSA) schon

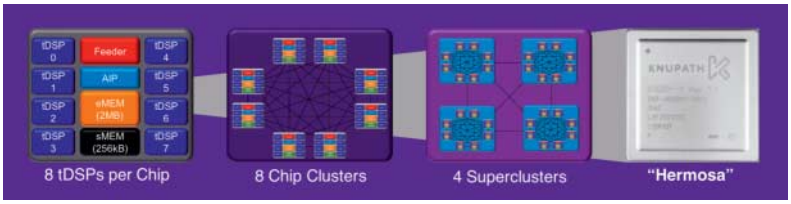
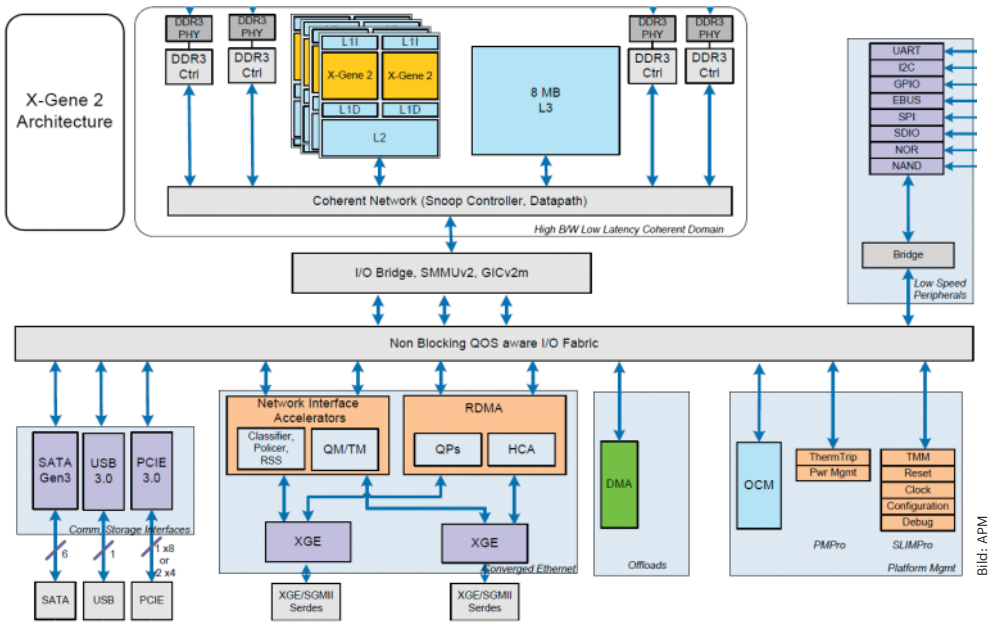


Bild: KnuPath

KnuPath Hermosa: ein KI-Beschleuniger mit 256 DSP-Kernen



Den X-Gene 2 hat APM subtil überarbeitet, jetzt enthält er etwa eine IOMMU.

vor Jahren besetzt, doch in letzter Zeit ist es darum recht still gewesen.

Absatzschwäche

Ziemlich ruhig in Bezug auf neue Hardware war es auch bei Apples WWDC – da hatte man mehr erhofft, etwa neue MacBooks. Doch vielleicht sieht Apple angesichts des weiter schrumpfenden PC-Markts keinen Anlass, sich mit Neuheiten zu beeilen, lieber verkauft man erst einmal die Lagerbestände ab. Das sehen angeblich die anderen PC-Hersteller auch so, weshalb Intels neue Kaby-Lake-Chips wohl erst später kommen als geplant. Bis auf HEVC- und VP9-Decoding auch mit HDR-Kontrasten bringen die Skylake-Nachfolger wenig Neues; die Desktop-Versionen starten möglicherweise erst Ende 2016. Ähnliches hört man auch von AMD Zen FX und Bristol Ridge für Desktops.

Früher als gedacht könnte hingegen DDR5-SDRAM kommen: Rambus-Manager Frank Ferro hat Semiengineering.org verraten, dass sich das zuständige JEDEC-Gremium bereits mit dem DDR4-Nachfolger beschäftigt. Er erwartet erste Produkte ab dem Jahr 2020. Wie üblich soll die neue Double-Data-Rate-DRAM-Generation viel höhere Datentransferraten bei gleichem oder besser noch niedrigerem Stromdurst liefern.

Das wäre auch für künftige Exaflops-Supercomputer wichtig. HPC-Experten rechnen detailliert aus, wie viele Picojoule (pJ) Energie pro Bit-Transfer nötig sind; bei aktueller Intel-Technik sind es etwa 5 pJ, um ein Bit aus einem DRAM-Chip bis in die Ausführungseinheit der CPU zu transportieren. Wenn man das auf ein Exaflops-System hochrechnet, summiert es sich zu mehreren Megawatt an Leistungsaufnahme – alleine für den Speicherzugriff.

Mehr Bits pro Watt versprechen High-Bandwidth Memory (HBM) auf Grafikkarten und Hybrid Memory Cube (HMC) für Intels Xeon Phi. Enge räumliche Kopplung zwischen CPU und Speicher vermeidet Stromvergeudung auf langen Leitungspfaden. Noch weiter geht der Active Memory Cube (AMC) von IBM, der selbst auch Rechenaufgaben übernimmt. Und im Vergleich zu aktuellen Computern will „The Machine“ von HPE die Speicher-Hierarchie gar auf den Kopf stellen.

Spezialrechner entfalten ihre Vorzüge allerdings erst mit speziell dafür geschriebenem Code – womit sich der Kreis zu den Rechenbeschleunigern schließt. Das sorgt für lange Übergangszeiten, wie auch der Bericht über Wetter-Computer in c’t 12/16 angerissen hat: Als erster Supercomputer mit GPU-Beschleunigern für Wetterprognosen ging 2015 Piz Kesch in Betrieb, erst acht Jahre nach der Vorstellung von Nvidia CUDA. Und bei der Umstellung ihrer Algorithmen hatten die Meteorologen trotzdem mit einigen Widrigkeiten zu kämpfen. Ähnliche Erfahrungen machte Johan de Gelas, der für Anandtech.com einen Server mit dem noch aktuellen ARM-SoC ThunderX getestet hat: Der lieferte zwar ansehnliche Resultate, aber erst nach langen Tüfteleien liefen Benchmarks vernünftig. Es liegt also noch viel Software-Arbeit vor jenen Firmen, die sich mit neuer Hardware ein Stück vom großen x86-Kuchen absäbeln wollen. *(ciw@ct.de)*