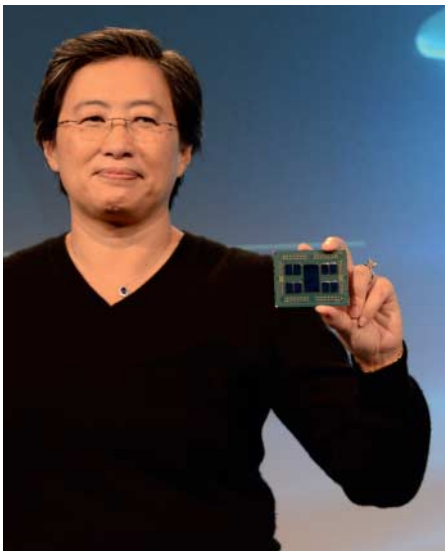
Zen 2 und Radeon Instinct für 2019

AMD setzt auf Chiplets und 7 Nanometer



Fertigung in 7-Nanometer-Technik, bis zu 64 Zen-2-Prozessorkerne mit verdoppeltem Rechendurchsatz und ein rasend schneller Beschleuniger für maschinelles Lernen – AMD will 2019 Vollgas geben, zuerst bei Servern.

Von Carsten Spille

Um sich von dem auf 29 Milliarden US-Dollar veranschlagten Umsatzpotenzial in Rechenzentren ein Stück zu sichern, will AMD mit der kommenden Prozessor- und Grafikchip-Generation voll angreifen. Dabei soll nicht nur topmoderne Herstellung mit 7 Nanometer (nm) feinen Strukturen helfen, sondern auch die Dekonstruktion des Prozessors. Auf der hauseigenen Veranstaltung „Next Horizon“ präsentierte AMD Anfang November die Pläne. Die CPUs sollen 2019 ausgeliefert werden, in größeren Stückzahlen aber wohl erst in der zweiten Jahreshälfte. Die Radeon Instinct MI mit 7-nm-Grafikchips will AMD noch 2018 verkaufen. Preise wurden nicht genannt.

7 nm, Chip-chen, mehr AVX

Anders als in den letzten Jahren kann AMD einen Fertigungsvorteil in die Waagschale werfen. Denn bei Intel halten die Schwierigkeiten mit der eigentlich schon vor Jahren fest eingeplanten 10-nm-Herstellung an, während die von AMD beauftragte Chipschmiede TSMC bereits Schaltkreise mit 7 nm feinen Strukturen liefert. Die sollen doppelt so viele Transistoren pro Quadratmillimeter packen wie die hauseigene 14-nm-Technik. AMD nutzt diesen Vorteil bei Zen 2 auf geschickte Weise und dekonstruiert den Prozessor code-namens „Rome“. Denn nur die eigentlichen Rechenkerne samt Cache und Infinity-Fabric-(IF-)Link zur Anbindung werden in 7 nm bei TSMC hergestellt, ein Compute-Cache-Die (CCD) gewissermaßen. Bis zu acht dieser kleinen Chip-chen (Chiplets) mit je acht CPU-Kernen sind via IF-Link mit einem zentralen I/O-Die verbunden, das Fertigungspartner Globalfoundries in 14 nm herstellt. Der beherbergt außer den acht Speichercontrollern und den PCI-Express Root-Hubs, die erstmals in der x86-Welt auch PCIe 4.0 beherrschen, auch den IF-Switch. Über ihn kommunizieren die Rechenkerne miteinander. AMD schlägt zwei Fliegen mit einer Klappe, da im I/O-Die auch die analogen Schaltungen sitzen, die schlechter als reine Logik oder Cache von der kleineren Fertigungsgeometrie profitieren würden. Zum genauen Aufbau schweigt sich der Hersteller noch aus. Auf Nachfrage gab AMD immerhin zu Protokoll, dass Programmierer nun mit deterministischen Latenzen beim Speicherzugriff rechnen können.

Auch die Innereien hat AMD bei der Zen-2-Architektur überarbeitet. Nicht nur hat man weitere, nicht genauer spezifizierte Härtungen gegen Spectre-Varianten ins Silizium geätzt, auch die internen Datenpfade, Caches und das Front End sind leistungsfähiger. Die Datenleitungen und AVX-Register sind mit 256 Bit nun doppelt so breit wie zuvor, sodass sich die Rechenleistung ebenfalls verdoppelt. In Verbindung mit den maximal 64 statt 32 Kernen ist es daher kein Wunder, dass ein Epyc-2-System nun an die Leistung eines Zwei-Sockel-Servers der ersten Epyc-Generation herankommt. In einem von AMD ausgewählten Benchmark mit dem Renderer C-Ray schlug ein Single-Sockel-„Rome“ den Zwei-Sockel-Vorgänger Naples und auch einen Dual-Xeon Platinum 8180M.

Maschinen lernen in 7 nm

Nicht nur Prozessoren, auch Grafikchips hat AMD überarbeitet und lässt diese in 7-nm-Technik fertigen. Sie sollen als Radeon Instinct MI50 und MI60, Beschleunigerkarten für Machine Learning, noch 2018 auf den Markt kommen. Die 13,2 Milliarden Transistoren quetscht AMD auf 331 Quadratmillimeter – zum Vergleich: Nvidias Tesla V100 ist 146 Prozent größer. Dank der kleineren Chipfläche kann AMD Nvidia vor allem preislich unter Druck setzen.

Die neue AMD-GPU basiert auf der Vega-GPU, hat aber einige Neuerungen bekommen. Statt zwei arbeiten nun vier Speichercontroller, was außer verdoppelter Transferrate von bis zu 1 TByte/s auch die zweifache Speichermenge von 32 GByte ermöglicht. Zudem sind nun sämtliche Caches und Register ECC-geschützt. Die 4096 Shader-Recheneinheiten schaffen bei doppelter Gleitkommagenauigkeit nun 50 Prozent des regulären FP32-Durchsatzes und toppen mit 7,4 TFlops Nvidias PCIe-Version der Tesla V100 knapp. Fürs Machine Learning hat AMD die Shader-Einheiten umgebaut, sodass sie nun auch INT8 (59 Tera-Ops/s) und INT4 (118 TOPS) beherrschen.

Für den Einsatz im Rechenzentrum sind die beiden Infinity-Fabric-Links spannend. Sie schaffen mit 100 GByte/s ein Unified Memory zwischen maximal vier Karten. Zwei dieser Vierergruppen können dann als Peer-to-Peer über PCIe aufeinander zugreifen. Damit das nicht zum Bremsklotz wird, hat AMD die neue Vega-GPU wie auch die kommenden „Rome“-Server-CPUs mit (abwärtskompatiblem) PCIe 4.0 ausgestattet. *(csp@ct.de)* c't

AMD lud den Autor zur Veranstaltung „Next Horizon“ nach San Francisco ein.