Prozessorgeflüster

Von Vektoren und Tensoren

Intel reist durch die Lande und predigt Vektorisierung und Google überrascht auf der Entwicklerkonferenz „I/O“ mit einer selbstentwickelten Tensor Processing Unit. Und von AMD gibt es auch was Neues.

Von Andreas Stiller und Christof Windeck

Der Juni hält bezüglich Prozessoren so allerhand bereit, beginnend gleich mit der Computex. Nicht umsonst hält hier Intels First Lady, die Chefin der Datacenter- und Server-Gruppe Diane Bryant, die Eröffnungsrede. Auch AMDs First Lady Lisa Su will spannende Sachen verkünden, etwa zur nächsten Grafikarchitektur Polaris. Zunächst aber gehts um den Launch von neuen Prozessoren und APUs, namentlich um Bristol Ridge. So heißt der letzte Mohikaner mit Bulldozer-Architektur, deren letzte Inkarnation als Excavator ins Rennen geht. Man hört von zwei und vier Excavator-Kernen mit einem Takt von bis zu 4 GHz im Boost. Ihnen zur Seite stehen bis zu 8 Compute Units der Gen3-Grafik. Die APU benötigt im Desktop-Bereich einen neuen 1331-poligen Sockel AM4, um unter anderem DDR4-Speicher zu unterstützen. Weitere Feinheiten (etwa über PCIe-Lanes) sind noch nicht bekannt. Im gleichen Sockel soll dann später auch der Zen-Prozessor Summit Ridge laufen.

Hier und da sind auch schon Benchmarkwerte von AMDs Referenzsystem Myrtle aufgetaucht, die, gemessen auf frühen Prototypen mit 2 GHz, aber noch nicht viel aussagen. Neben der Desktop-Version will AMD zwei Mobile-APUs herausbringen, eine mit dem neuen FP4-Sockel und eine für den alten, Carrizo-kompatiblen FP3.

Später im Juni erwartet man auch den offiziellen Stapellauf des reichlich verspäteten Xeon Phi Knights Landing (siehe S. 20) auf der ISC 2016 in Frankfurt. Für ihn hat Intel die Software-Infrastruktur von langer Hand vorbereitet, etwa mit OpenMP.

OpenMP-News

Ein Großteil der Neuerungen von OpenMP 4.5 hängt mit SIMD und der Vektorisierung auf unterschiedlicher Hardware zusammen. Für Intel trifft es sich gut, dass Ende April mit Dr. Michael Klemm ein Intel-Software-Ingenieur zum Chef des OpenMP Architecture Review Board gewählt wurde. Er tritt in die Fußstapfen von Michael Wong, der zuvor langjährig die OpenMP- und Transactional-Memory-Geschicke leitete und der im April von IBM zu Codeplay (nein, nicht zu Coldplay) wechselte – und mit dem ich zahlreiche interessante Gespräche führen konnte.

Sein Nachfolger Dr. Klemm „entstammt“ übrigens einer Hochburg des parallelen Programmierens in Deutschland, der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Von hier kam auch das „Team Kraut“, das 2013 und 2014 als einziges deutsches Team an der Student Cluster Competition auf den Supercomputer-Konferenzen in den USA mit gutem Erfolg teilgenommen hatte und das sich auch dieses Jahr für die SC16 in Salt Lake City wieder beworben hat. Interessierte junge Studenten und -innen der FAU (noch kein Bachelor) können möglicherweise noch mitmachen. Auch auf der ISC 2016 in Frankfurt wird es wieder eine Student Cluster Competition geben, mit zwölf Teams vor allem aus den USA und China, aber auch aus Estland und Katalonien. Die schwarzrotgoldenen Farben vertritt hier ein Team der Universität Hamburg.

Google kümmert sich ebenfalls intensiv um Studenten und lädt zum Google Summer of Code (GSoC) 2016 ein. Dabei geht es um die Mitarbeit an zahlreichen Open-Source-Projekten, für die Google mit einem Stipendium winkt. Vorher fand aber noch die hauseigene Entwicklerkonferenz I/O statt, diesmal nicht im Moscone Center in San Francisco, sondern im Open Air Amphitheater in Mountain View (siehe S. 28).

Prozessorbauer Google

Das Gerücht, Google gehe unter die Prozessorhersteller, wurde dort vom Chef der Google-Rechner-Infrastruktur, Urs Hölzle, bei strahlendem Sonnenschein bestätigt. Er verriet, dass man schon seit mehr als einem Jahr selbst entwickelte Tensor Processing Units (TPUs) in den eigenen Rechenzentren betreibe. Die TPUs kommen für die derzeit schwer im Trend liegenden KI-Anwendungen wie Machine Learning oder auch Deep Learning zum Einsatz, speziell als extrem effiziente Beschleuniger für die Open-Source-Library TensorFlow. Sie rechneten auch in der KI namens AlphaGo mit, die einen der besten menschlichen Go-Spieler Lee Sedol in den spektakulären Matches Anfang März bezwang.

Zu Details der TPUs schwiegen sich Hölzle und der Hardware-Entwickler Norm Jouppi, der dazu einen Blog-Beitrag verfasste, allerdings beharrlich aus. Hölzle verriet lediglich, dass man schon zwei TPU-Versionen von zwei unterschiedlichen Auftragsfertigern im Einsatz habe. Um welche Auftragsfertiger es sich handelt und welche Strukturbreiten zum Einsatz kamen, blieb geheim. Genauere Performance-Angaben oder relative Vergleiche zu Xeons oder Nvidia-Tesla-Karten gab es leider auch nicht, es hieß lediglich, dass die TPUs etwa zehnmal so energieeffizient bei Machine Learning seien wie herkömmliche Prozessoren.

Physisch sitzen die TPUs auf kompakten Steckkarten, die vermutlich per PCI Express mit normalen Servern verbunden sind. Sie kommen nicht bloß für AlphaGo, sondern auch schon seit gut einem Jahr produktiv zum Einsatz, etwa bei der Sortierung von Suchergebnissen. Dafür nutzt Google eine Technik namens RankBrain.

Auch bei Google StreetView helfen die TPUs und Jouppi verwies auch auf die Plattform Google Cloud Machine Learning, die Entwickler für eigene TensorFlow-Experimente nutzen können. In welchem Ausmaß daran TPUs mitrechnen, ist wiederum geheim.

Ob Googles TPU-Technik ein Rückschlag für Nvidia ist, bleibt folglich offen: Eigentlich kooperiert Google im OpenPower-Konsortium nämlich nicht nur mit IBM, sondern auch mit Nvidia, um Hardware für KI-Anwendungen an den Start zu bringen. Nvidia hebt diesbezüglich stets die Tesla-Karten heraus, aber vielleicht hat Google ganz andere Pläne. Schließlich hatte man sich schon 2010 die ehemaligen P. A.-Semi-Entwickler einverleibt, die nach der Übernahme von P. A. Semi durch Apple ihr eigenes Start-up Agnilux gegründet hatten. Und auch zu ARM-SoCs für Server hat sich Google mittlerweile öffentlich bekannt – nicht ohne hinzuzufügen, dass man bei OpenPower schon weiter sei.

Urs Hölzle teilte auch einen Seitenhieb auf die derzeit wieder im Aufwind fliegende FPGA-Technik aus – man weiß ja, dass sich Intel mit Altera gerade einen der größten FPGA-Spezialisten gekrallt hat. Jedenfalls sagte Hölzle, dass die eigenen TPU-ASICs wesentlich effizienter arbeiteten als eine TPU-Implementierung in FPGAs. *(as@ct.de)*

Bild: Google

Nicht CPU, nicht GPU, sondern TPU: Tensor Processing Unit. Mit solchen Clustern aus TPUs rechnet Google und hat damit den Go-Großmeister Lee Sedol geschlagen.