Seminar Hochleistungsrechner: Aktuelle Trends und Entwicklungen
Wintersemester 2016/2017

Aktuelle GPU-Generationen (NVidia Pascal, AMD Polaris)

Martin Sellmair
LMU

05.01.2017

Zusammenfassung

Diese Seminararbeit enthält eine Übersicht über die, vor kurzem erschienen, neuen Chipgenerationen für Nvidia und AMD Beschleunigerkarten: Pascal und Polaris. Es ist eine Beschreibung der wichtigsten Änderungen und Zusammenfassung der wichtigsten technischen Daten enthalten. Der Schwerpunkt liegt hierbei bei Pascal, da hier die größeren Änderungen vorgenommen wurden, und auch eine größere Vielfalt an GPUs herrscht. Ein Vergleich der Leistung und Energieeffizienz der neuen Chips untereinander wird ebenfalls gezogen. Dieser basiert vor allem auf den Ergebnissen bei der Berechnung von komplexen 3D-Szenen. Das Ende bildet ein kurzer Ausblick auf AMD Vega und folgende Architekturen.

1 Motivation:

Im letzten Jahrzehnt gewannen Beschleunigerkarten wie Grafikkarten, nicht nur für private Anwender und Firmen die im Bereich Grafikverarbeitung tätig sind massiv an Bedeutung, sondern wurden auch für den Einsatz in Supercomputern immer wichtiger. Heute sind GPUs bereits in den meisten führenden Supercomputern verbaut, um die Berechnungen immer komplexerer und umfassenderer Simulationen in verschiedensten Bereichen zu unterstützen.
Die GPUs werden für die Beschleunigung von Big Data Anwendungen genau so eingesetzt, wie auch für die Berechnung künstlicher Intelligenzen und für die Deep Learning Prozeduren Neuronaler Netze. Für Berechnungen dieser Art sind die auf hochgradig paralleles berechnen ausgelegten neusten Grafikkarten oftmals ideal. Die steigenden Anforderungen an HPC-Systeme machen es nötig, dass diese immer größere Leistung zur Verfügung stellen. Allerdings steigt bei der Verwendung bestehender Chiparchitekturen die benötigte Energie zum Betrieb noch größerer und leistungsfähigerer Supercomputer ebenfalls stark an. Eine neue Generation an Chips verspricht hier stets mehr relative Leistung für die aufgewandte Energie. Dies wird erreicht durch verkleinerte Strukturbreite und dadurch eine höher Anzahl an Transistoren, aber auch

durch Änderungen an der Chiparchitektur selbst.
Die beiden führenden Hersteller von GPUs (AMD und Nvidia) haben in den letzten Monaten die ersten Grafikkarten mit den neuen Architekturen mit Codenamen Polaris und Pascal auf den Markt gebracht.

1.1 Diese Seminararbeit

Das Ziel dieser Seminararbeit ist es, einen generellen Überblick über die Neuerungen der beiden Architekturen Pascal und Polaris zu geben. Da in beiden Fällen gegenüber den Vorgängern die Strukturbreite verkleinert wurde, wird zuerst ein Blick auf die neuen Herstellungsverfahren geworfen. Es folgt eine Betrachtung der Änderungen in den Chiparchitekturen selbst. Anschließend werden die Fortschritte im Speicherbereich beleuchtet, von verbesserte Komprimierung bis zu NVLink, Nvidias neuer Technik für Inter-GPU-Kommunikation. Abschließend werden Pascal und Polaris sowohl miteinander, als auch mit ihren jeweiligen Vorgängergenerationen, in Hinblick auf Performance und Energieeffizienz verglichen.
Die Quellenlage für fundierte wissenschaftliche Untersuchungen stellte sich während der Recherche als äußerst dürftig dar. Die Beiden behandelten Chipgenerationen sind erst seit so kurzer Zeit auf dem Markt, dass schlicht noch keine vernünftige wissenschaftlichen Arbeiten über ihrer Leistungsfähigkeit existieren. Gerade im Falle von AMDs neuen Karten ist dies besonders offensichtlich, da diese noch nicht einmal alle auf dem Markt sind. Es musste deshalb auch auf die technischen Angaben der beiden Hersteller zurückgegriffen werden.

1.2 Pascal

Wie alle Chiparchitekturen Nvidias der jüngeren Vergangenheit nach einem berühmten Physiker benannt, ist Pascal der Nachfolger zu Maxwell. Generell lassen sich die einzelnen Variationen der neuen Pascal Architektur in zwei Gruppen von Chips unterteilen:

- GP100: Der von der Tesla P100 GPU verwendete Chip
- GP102 bis GP107: Die von den GTX GPUs verwendeten Chips

GP100 Der GP100 Chip ist die explizit für General Purpose Computation on Graphics Processing Unit (GPGPU) ausgelegte Variante der Pascal Chips. Er wurde im Vergleich zu Maxwell speziell für den Einsatz in diesem Bereich verbessert. So wurde Beispielsweise für Deep Learning Verfahren eine Unterstützung für 16 Bit Fließkommaopertationen hinzugefügt. Diese Algorithmen benötigen keine sehr exakten Fließkommaoperationen, aber gewinnen stark durch die zusätzliche Leistung die 16-Bit Berechnungen ermöglichen. Des weiteren sinken durch die 16-Bit Datentypen die Speicheranforderungen pro einzelner Rechenoperation.[6]

GP102-GP107 Die Chips GP102 bis GP107 sind die neben GPGPU primär auch für die Berechnung von polygonaler 3D-Grafik ausgelegte Variante der Pascal Chips. Bei all diesen Karten handelt es sich im Vergleich zur Tesla P100 um Grafikkarten im klassischen Sinne. Wie schon bei den Maxwell-Karten reicht dabei die Bandbreite von High-End-Karten wie der Titan X, auf welcher der GP102 verbaut wurde, bis zu Einsteigerkarten wie der GTX 1050.
Der intendierte Einsatzzweck der Chips schwankt dementsprechend. Die Titan X am einen Ende des Spektrums ist, der Tesla P100 am ähnlichsten, primär für Workstations und GPGPU Aufgaben gedacht. Der Einsatzzweck der GTX 1050 am anderen Ende hingegen ist vor allem die Darstellung von anspruchsvoller 3D-Grafik in HD-Auflösung.

1.3 Polaris

Polaris ist die vierte Iteration der von AMD Graphics Core Next (GCN) genannten Chiparchitektur, die Ende 2016 von AMD neu veröffentlicht wurde. Es ist der Nachfolger von Fiji, dem leistungsstärksten Chip der dritten Generation GCN. Momentan existieren im wesentlichen zwei Varianten des Polaris Chips: Der schwächere Polaris 10, verbaut in RX460 und RX470, sowie der leistungs-

stärkere Polaris 11, welcher in der RX480 verbaut wird.

2 Herstellungsverfahren

Änderungen im Herstellungsverfahren spielen eine wichtige Rolle bei den Verbesserungen in Leistung und Energieeffizienz mit denen Pascal gegenüber Maxwell aufwarten kann. Wurde Maxwell, wie auch dessen Vorgängergeneration Kepler, noch in 28 Nanometer Strukturbreite hergestellt, so wird Pascal nun in TSMC's 16 Nanometer FinFET Verfahren gefertigt.
AMDs neue Polaris Chips dagegen werden (bei Global Foundries) im 14 Nanometer FinFET Verfahren gefertigt. Die angegebene Strukturbreite liegt damit etwas unter der von Nvidias Pascal. Es handelt sich um die bis dato kleinste Strukturbreite aller veröffentlichten Grafikkartenchips.
Allerdings ist zu beachten, dass diese Maße, genau wie die von Nvidia angegebenen 16nm, mit Vorsicht zu genießen sind, da sie oft Marketinghintergründe haben:

> "Mit realen Längen oder Maßen wie der Gate-Länge auf Chipebene haben Bezeichnungen wie 14 Nanometer schon seit Jahren nichts mehr zu tun. So sagte Intels William Holt, Leiter der Halbleiterfertigung, zu Broadwell: "Da ist wirklich nichts dran, was 14 Nanometer groß ist."14 Nanometer sind also nicht viel mehr als Marketing, wenn auch mit einem historischen Hintergrund."[8]

3 Chiparchitektur

Wie schon bei vorherigen in Tesla Karten verbaute GPUs ist auch der GP100 eine Aneinanderreihung von so genannten Graphic Processing Clustern (GPC). Ein voll ausgebauter GP100 Chip besteht in der obersten Ebene aus einer Anordnung von sechs GPCs. Diese enthalten insgesamt 30 Texture Processing Cluster (TCP) in denen jeweils 2 Streaming Multiprocessoren (SM) enthalten sind. Der gesamte Chip verfügt demnach über 60 SMs. Des weiteren existieren über insgesamt auf acht 512 Bit Memory Controller aufgeteilte 4096 Bit Speicheranbindung. Diese sind seitlich verbaut, da sich dort die HBM2 Speichermodule befinden. Der gesamte Aufbau des GP100 kann in schematischer Form in Abbildung 1 nachvollzogen werden.
Es ist anzumerken, dass selbst die Tesla P100 (als Nvidias stärkste Karte) aus produktionstechnischen Grnden nur über 56 der 60 möglichen Streaming Multiprocessoren verfügt, und daher zumindest aktuell noch keine Karte existiert, welche die volle theoretische Leistung des GP100 bringt. Prinzipiell jedoch kann jeder GPC des GP100 zehn Streaming Multiprocessoren enthalten.[6].

3.1 Streaming Multiprocessoren

Die Erste augenfällige Neuerung an den Pascal SMs ist deren erhöhte Taktfrequenz im Gegensatz zu Kepler und Maxwell. Ein Vergleich mit den älteren Teslamodellen zeigt den Unterschied: Die Tesla K40 (Kepler) rechnete mit einem Basistakt von 745 Mhz und einem maximalen Boosttakt von 875 MHz, die Tesla M40 (Maxwell) mit einem Basistakt von 948 Mhz und einem Boosttakt von 1114 MHz. Die P100 hingegen verfügt über einen deutlich schnelleren Basistakt von 1328 MHz und Boosttakt von 1480 MHz. Der Basistakt von Pascal liegt also bereits 200 MHz über dem Boost von Maxwell.[6]
Ein SM des GP100 ist jeweils mit 64 CUDA-Kernen und vier Textureinheiten bestückt. Diesen Aufbau veranschaulicht Abbildung 2

3.2 Simultane Multi-Projektion

Die Simulanous Multi-Projektion Einheit ist eine Neuheit in der in Pascal erstmals verbauten Polymorph Engine 4.0. Bei der Polymorph Engine handelt es sich um einen Teil des Streaming Multiprocessors der speziell für die Verarbeitung von Geometrieinformationen zuständig ist. Die neue SMP Einheit erweitert diese Engine um die Fähigkeit identische Geometrie aus Verschiedenen Winkeln oder von verschiedenen Positionen aus zu berechnen, ohne dass die Anwendung diese Instruktionen

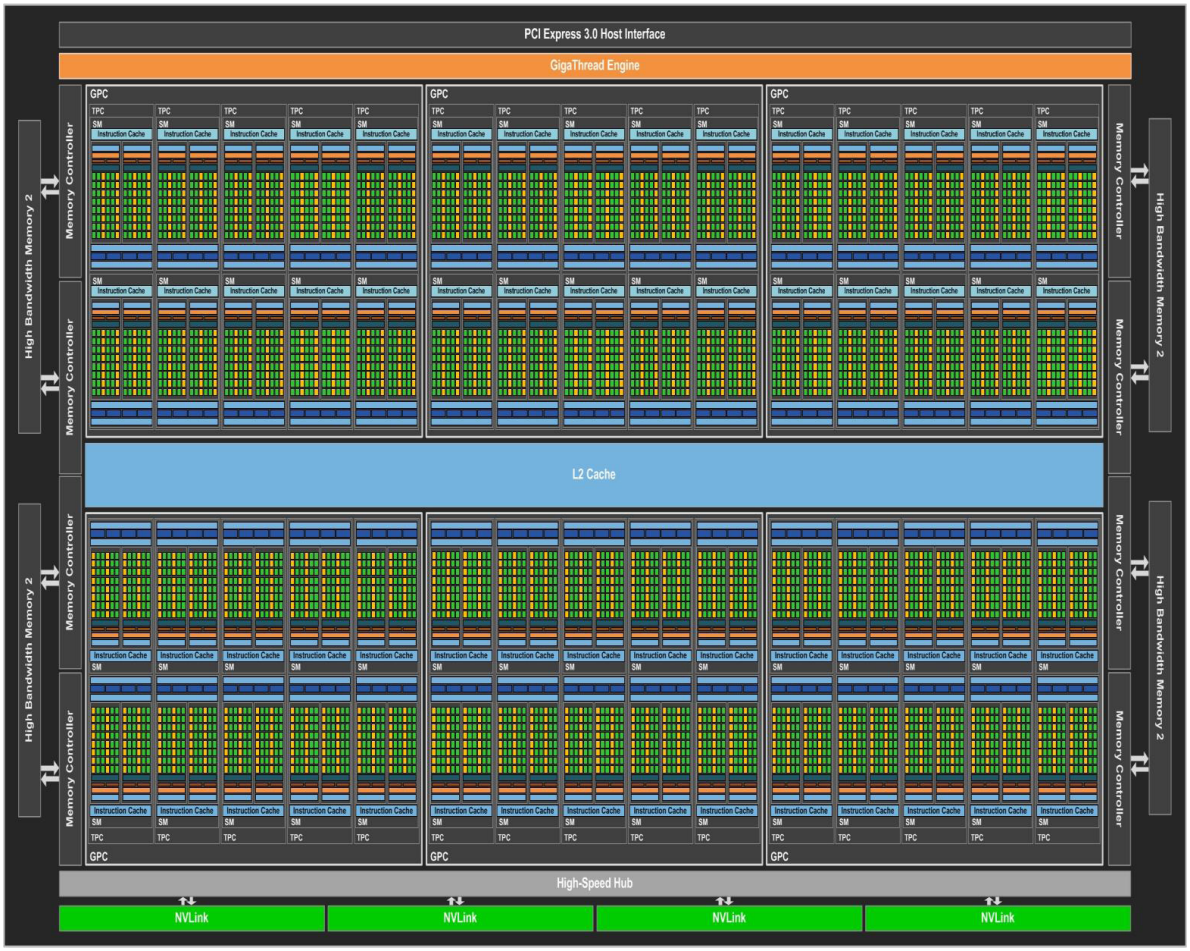

Abbildung 1: GP100 Architektur im Vollausbau mit allen 60 möglichen Streaming Multiprocessoren [6]

und die Geometrieinformationen mehrmals an die GPU senden muss. Die nötigen Berechnung für diese Änderungen in der Projektion sind hardwarebeschleunigt, und dementsprechend performanter als eine reine Softwarelösung.

Der hauptsächliche Einsatzzweck der SMP Einheit ist das berechnen von Virtual Reality Umgebungen, in welcher identische Geometrie für beide Augen separat (eben in der Position um den Abstand der Augen versetzt und mit leicht anderem Winkel) berechnet werden muss. Zwar beherrschten auch GPUs mit Maxwell Chips bereits einige der von der SMP Einheit zur Verfügung gestellten Funktionen, allerdings nur in sehr eingeschränktem Umfang, und weniger effizient.[5]

3.3 Präemptives Multitasking

Weitere für Pascal vorgenommene Verbesserungen betreffen den Scheduler der GPUs. Dieser wurde auf Präemptives Multitasking umgestellt, und erlaubt nun das Scheduling in der Granularität einzelner Instruktionen, und nicht mehr nur wie in Maxwell und Kepler für ganze Threads. Der Programmierer

Abbildung 2: GP100 Streaming Multiprocessor [6]

muss nun nicht mehr von Hand lang laufende Programmblöcke in einzelne Tasks aufteilen. Geschah dies bei den älteren GPUs nicht, hatte es das Blockieren des ganzen Systems bis zur Beendigung des Tasks zur Folge. In einer Situation in der ein Thread nur auf das eintreten eines anderen Ereignisses wartete, konnte die GPU nicht genutzt werden, obwohl sie eigentlich ihre volle Leistung für andre Aufgaben bereitstellen hätte können. Der mit Pascal verbesserte Kontextwechsel zwischen einzelnen Threads macht genau hier dem Programmierer das Leben leichter.

3.4 Asynchronous Compute Engine

Mit der Graphics Core Next Generation 4 Architektur verbessert AMD hauptsächlich die Tessellation[1]-Leistung der Polaris Karten. Sie war eine der größten Schwachstellen des originalen GCN Designs, und wurde massiv erhöht.[4]
Ein wichtiger und ebenfalls überarbeiteter Chipbestandteil sind die Asynchronous Compute Engines. Sie sind für die parallele Aufteilung der Rechenlast bestimmter Shader (z. B. Pixel und Texturshader) auf die einzelnen Compute Units (AMDs

[1] Ein Verfahren zur dynamischen Generiereung von Geometrie in 3D-Szenen relativ zur Kameradistanz

Äquivalent zu Nvidias CUDA Cores) zuständig. Den ACEs wurden zwei neue Funktionen hinzugefügt. Die Quick Response Queue und Compute Unit Reservation werden im Folgenden erläutert.

Quick Response Queue Die QRQ ermöglicht es dem Entwickler einzelnen Tasks die nicht Teil von Grafikberechnungen sind höhere Priorität zuzuweisen. Die ACEs werden diesen Tasks dann mehr Resourcen zuweisen und sie werden dadurch in der parallelen Verarbeitung bevorzugt und schneller beendet. Dies geschieht ohne dass der Command Processor die Berechnung anderer laufender Tasks komplett unterbricht. (Siehe Abbildung 3) Durch dieses Verfahren kann sichergestellt werden, dass bestimmte Berechnungen gegenüber der Berechnung der Grafik priorisiert sind, und stets so schnell wie möglich beendet werden, ohne die Grafikberechnung kurzeitig komplett aussetzen zu mssen.[1]

Compute Unit Reservation CUR erlaubt es, die zu Verfügung stehenden Compute Units in einzelne Gruppen einzuteilen. Auf diese Weise kann sichergestellt werden, dass für einen bestimmten Typ von Berechnung, wie etwa Pixelshader, eine Gruppe von CUs exklusiv reserviert sind. Sie stehen dann natürlich für andere Berechnungen nicht mehr zur Verfügung. Es kann so aber für den reservierten Berechnungstyp immer die ausreichende Leistung bereitgestellt werden, um diese Berechnungen schnellst möglich zu erledigen.[1] Abbildung 4 zeigt dieses am Beispiel von CUs die für Audioberechnungen reserviert sind.

3.5 Schaltkreisäderungen

AMD hatte es sich für Polaris auch zur Aufgabe gemacht, einige ihrer Techniken beim Design von CPU-Schaltkreisen für ihre GPUs zu adaptieren. Als wichtigste Änderungen sind hier vor allem drei Techniken zu nennen:

Adaptive Frequency and Voltage Scaling AVFS bezeichnet eine Echtzeitüberwachung der Temperatur und Spannung der Transistoren, wodurch Betriebsspannung und Frequenz dynamisch angepasste werden können. Dies erlaubt es zum Beispiel die anliegende Spannung dynamisch zu erhöhen, falls die, von der Einheit welche die Spannungsregulierung übernimmt, nominell zur Verfügung gestellte Spannung gar nicht wirklich bei den Transistoren anliegt.

Adaptive Clocking Hierbei handelt es sich um eine Technik die auf AVFS aufbaut, und für den Fall von spontan stark abfallender Spannung die Frequenz in sicheren Bereichen hält. Da diese Spannungsabfälle nur auftreten wenn zur selben Zeit fast alle CUs gleichzeitig zu rechnen beginnen, was nur selten vorkommt, ist es möglich die generelle Taktfrequenz zu um bist zu 140MHz zu erhöhen, weil durch das Adaptive Clocking für den Fall eines Spannungseinbruchs vorgesorgt ist. [1]

Multi-Bit Flip-Flop MBFF ist eine Optimierung der Taktung innerhalb des Chips. Anstatt in jedem der in den CUs verbauten Flip-Flops einen eigenen Takteingang einzubauen, werden die Flip-Flops zu Vierergruppen zusammengefasst, die gleichzeitig getaktet werden. Dies führt zu höherer Energieeffizienz:

> "AMD developed special quad-flops, where four flip-flops share a single stronger clock input [...] A single quad-flop takes about twice the energy compared to a normal flop, but performs the work of four flops reducing the load on the clock network by a factor of two." [1]

4 Speicher

Als erste Beschleunigerkarte von NVidia wird auf der Tesla P100 High Bandwidth Memory in Version 2 verbaut. Da es sich bei HBM2 um vertikal angeordnete Speichermodule handelt, können diese viel platzsparender verbaut werden als die zuvor verwendeten GDDR5 Module. Diese Entwicklung erlaubt es insgesamt mehr Speicher einzubauen, wel-

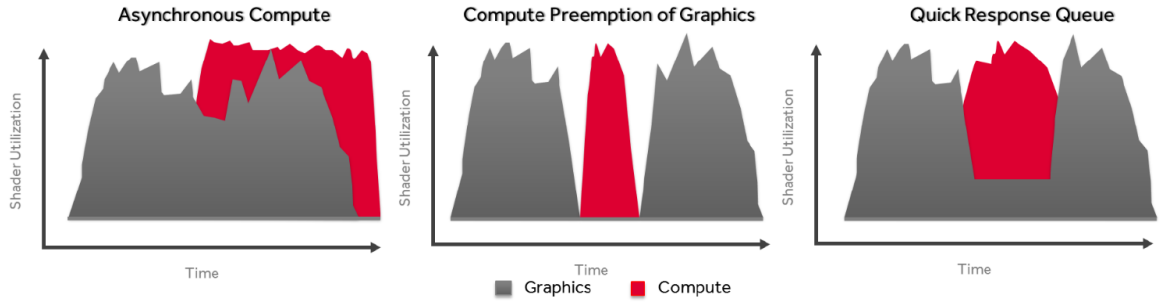

Abbildung 3: Schematisches Beispiel für die Quick Response Queue[1]

cher auch schneller angebunden ist als noch auf den Karten der Vorgängergenerationen. Die Tesla P100 verfügt daher über 16 Gigabyte Speicher, und Nvidia gibt zudem eine drei mal höhere Speicheranbindung als noch auf der Maxwell GM200 GPU an.[6] Im Vergleich zur P100 sind die dedizierten Grafikkarten deutlich zurückhaltender mit Speicher ausgestattet. Es kommt dort im Falle der Titan X und GTX 1080 (sowie bei der wohl in naher Zukunft zu erwartenden GTX 1080 TI) GDDR5X Speicher, und im Falle von GTX 1070, GTX 1060, und GTX 1050 (TI) nur GDDR5 zum Einsatz. GDDR5X zeichnet sich gegenüber dem ältern GDDR5-SGRAM vor allem durch bis zu doppelt so hohe Datentransferraten aus, und erreicht damit mit 900 GByte/s annähernd das Niveau von HBM2 (Theoretisches Maximum: 1 TByte/s).[10] Auch die bei GDDR5X gegenüber GDDR5 von 1.5 V auf 1.35 V sinkende Betriebsspannung sorgt dafür, dass dieser in Sachen Energieeffizienz zumindest in diesem Vergleich deutlich die Nase vorne hat.[2] HBM2 ist mit nur 1.2 V allerdings noch um ein gutes Stück besser.[9]

4.1 Cache

Die Leistung von atomaren Speicherzugriffen und die Leistung des Caches war bis vor kurzem bei GPUs noch ziemlich niedrig.[7] Polaris und Pascal setzen hingegen den bestehenden Trend hin zu größeren Caches und schnelleren atomaren Zugriffen in neuen GPUs fort, und erlauben es auch auf GPUs auf eine effiziente Art und Weise Algorithmen zu verwenden, die von diesen beiden Punkten stark profitieren. Das erleichtert auch das Portieren von bestehenden - eigentlich für CPUs optimierten - Code auf die GPU.[7]
Während etwa Fermi und Kepler GPUs noch über 64 KB an Speicher verfügten, der je nach Bedarf aufgeteilt, sowohl als shared memory und auch L1 Cache benutzt werden konnte, wurde diese Architektur beginnend mit Maxwell geändert. Die SM eines GP100 hat nun jeweils ein separates 64 KB shared memory Element und einen 64 KB L1 Cache. Diese Aufteilung führt effektiv dazu, dass für einzelne Programme keine Optimierung der Aufteilung des Speichers in shared memory und L1 Cache mehr vorgenommen werden muss, sondern für beides stets die vollen 64 KB zur Verfügung stehen.
Darüber hinaus verfügt der GP100 über 4096 KB L2 Cache und bietet damit einen sehr effizienten gemeinsamen Cache für die ganze GPU. Der L2 Cache des GK110 ist im Vergleich dazu nur 1536 KB groß, und der des GM200 ebenfalls geringere 3072 KB. Mit mehr direkt auf dem Chip verbautem Cache, verringert sich die Anzahl der Zugriffe direkt auf den Videospeicher, was die Leistung verbessert, den Energiebedarf senkt, und die benötigte Speicherbandbreite verringert.[6]
Auch die Architektur von Polaris zieht Vorteile aus der kleiner Strukturbreite, was in diesem Fall zu einer Verdopplung des verfügbaren L2 Caches führt.

[2] Bei fast gleichbleibender elektrischer Ladung

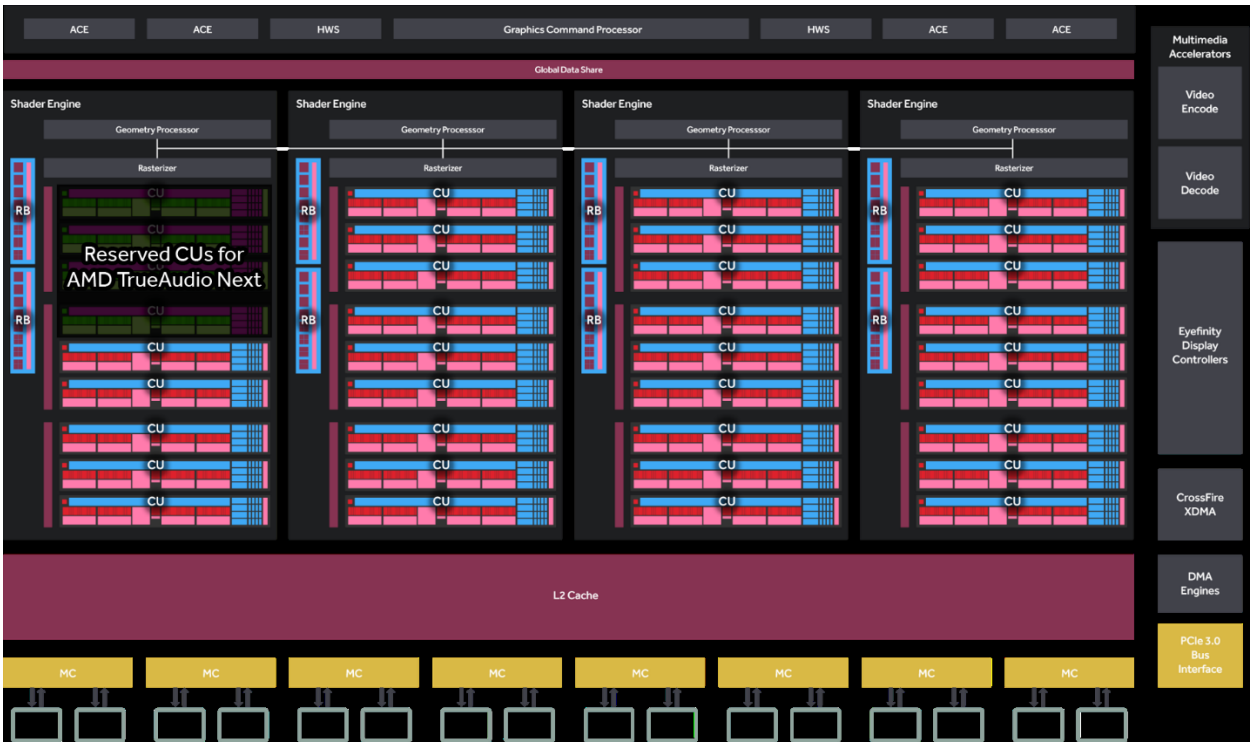

Abbildung 4: CUR reserviert einzelne CUs für Audiopipeline[1]

Zusammen mit der besseren Delta Color Compression, führt das zu einer Einsparung von bis zu 40 Prozent beim Strombedarf für Speicheroperationen, vergleicht man die RX 480 mit der älteren R9 290X GPU.[1]

4.2 Verbesserte Speicherkomprimierung

Die sogenannte Enhanced Memory Compression beschreibt die verlustfreie Komprimierung von Bildern in Pixelbuffern, um deren Speicherbedarf und Bandbreite bei ihrer Befüllung einzusparen. Sie wird schon seit längerer Zeit in Grafikkarten benutzt und Pascal und Polaris enthalten eine neuere und verbesserte Version. Die meisten Änderungen beziehen sich dabei auf eine Technik die Delta Color Compression genannt wird.
Hierfür werden Blöcken von Pixeln auf die farbliche Ähnlichkeit ihrer Pixel untersucht. Unterscheiden diese Pixel sich kaum farblich voneinander, wird für jeden Block nur ein Referenzpixel gespeichert, und für die verbleibenden Pixel nur das Delta zu diesem. Kann auf diese Weise der Speicherbedarf auf weniger als die Hälfte der Ursprungsmenge reduziert werden, gilt die Kompression als erfolgreich, und die Pixel werden, in auf diese Weise komprimierter Form, abgespeichert.[5]

Zusätzlich zu dieser schon in Maxwell und Fiji verwendeten 2:1 Technik, wurden in Pascal und Polaris noch eine 4:1 und 8:1 Kompression hinzugefügt. Die 4:1 Kompression funktioniert hierbei analog zur 2:1 Kompression. Der wesentliche Unterschied ist, dass in Bereichen eines Bildes mit sehr geringen Farbunterschieden Blöcke auf höchstens ein Viertel ihrer unkomprimierten Größe reduziert werden. Die 8:1 Kompression hingegen nimmt bereits 4:1 komprimierte Pixelblöcke, und führt dann auf die-

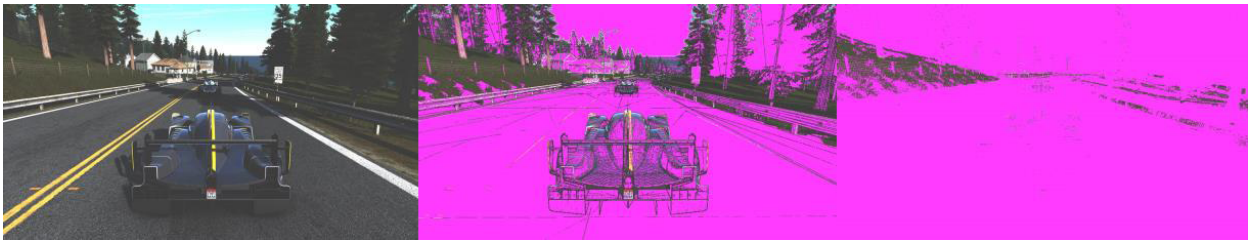

Abbildung 5: Delta Kompression: Originalbild, Maxwell und Pascal Kompression - Rosa Pixel wurden komprimiert in den Buffer geschrieben[5]

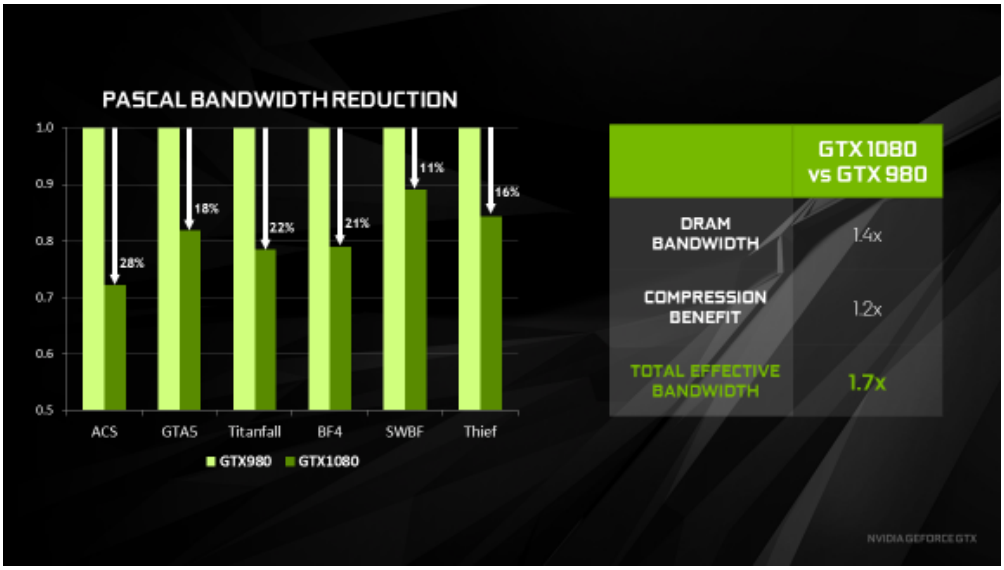

Abbildung 6: Vergleich der Speicherkomprimierung zwischen GTX 980 (Maxwell) und GTX 1080 (Pascal)[5]

sen Blöcken noch eine 2:1 Kompression durch. Es wird folglich noch ein zweiter Satz an Deltawerten gespeichert, eben genau der Unterschied vom Referenzpixel des ersten Blocks zum Referenzpixel des nächsten Blocks. Abbildung 5 veranschaulicht mittels eines Beispielframes aus Project Cars den Zugewinn an Pixeln, die durch die beiden neuen Komprimierungsverfahren zusätzlich in komprimierter Form abgespeichert werden können. Eine Übersicht über den Performancegewinn in Relation von GTX 1080 zur Maxwell Vorgängerkarte im gleichen Preissegment, der GTX 980, gibt Abbildung 6

4.3 Inter-GPU Kommunikation

Soll ein bestehendes Programm auf GPUs portiert werden, wird in parallelen Systemen mit mehreren Nodes, die jeweils eine oder mehr GPUs besitzen, die Kommunikation zwischen den einzelnen GPUs oft zum Flaschenhals. Der Grund hierfür ist, dass die Kommunikation zwischen CPU und GPU, genau wie auch die Kommunikation mittels MPI zwischen den CPUs, ziemlich langsam ist.[7] Arbeitet man mit einem solchen Setup, ist es wichtig, den Datentransfer zwischen den GPUs so niedrig wie möglich zu halten, um hohe Leistung im System zu erreichen.
Eine der signifikantesten Neuerungen die mit den Tesla P100 Karten Einzug halten, ist ein Ansatz, dieses Problem zu umgehen: Ein inter-GPU Kommunikationsinterface, das Nvidia NVLink nennt. Dieser Bus wird von Nvidia mit dem fünffachen maximalen Durchsatz von PCI Express Gen 3 x16 beworben.[6] Dadurch können hohe Datenmengen direkt von GPU zu GPU gesandt werden. Dieses Verhalten wiederum ist mehr und mehr wünschenswert, da nicht nur die Anzahl der Cores und die absolute Anzahl an Beschleunigerkarten in HPC-Systemen in letzter Zeit stark angestiegen ist, sondern auch die relative Anzahl an GPUs pro CPU sich erhöht hat:

> "Multiple groups of multi-GPU systems are being interconnected using InfiniBand and 100 Gb Ethernet to form much larger and more powerful systems. 2012s fastest supercomputer, the Titan located at Oak Ridge National Labs, deployed one GK110 GPU per CPU. Today, two or more GPUs are more commonly being paired per CPU as developers increasingly expose and leverage the available parallelism provided by GPUs in their applications." [6]

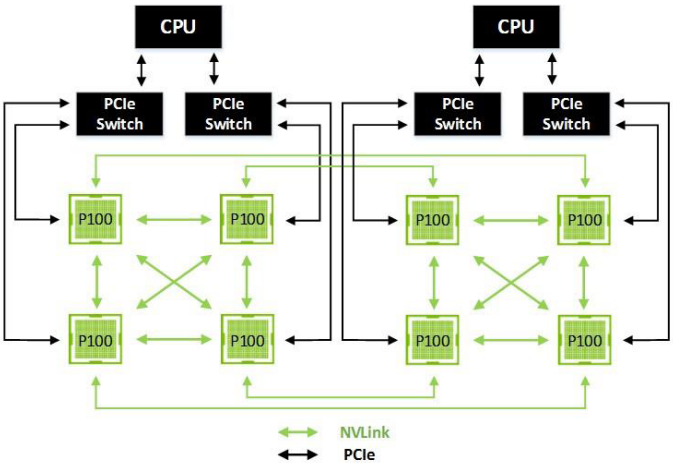

Abbildung 7: NVLink Setup mit acht Tesla P100[6]

Eine Entwicklung die dazu führt, dass die Bandbreite heutiger PCIe Anschlüsse in Systemen mit mehreren GPUs eine relevante Einschränkung darstellt. Um diese Einschränkung aufzuheben, stellt NVlink eine bidirectionale Bandbreite von 160 Gigabytes/Sekunde zur Verfügung.[6] Ein beispielhaftes Setup von zwei CPUs mit jeweils über zwei PCIe Switches angeschlossene P100 Einheiten ist in Abbildung 7 zu sehen.
Eine weitere wichtige Neuerung bei der Programmierung von Pascal GPUs ist die Einführung eines Gemeinsamen virtuellen Speichers für CPU und GPU. Bis zu 512 Terrabyte an gemeinsamen virtuellen können adressiert werden. Dies vereinfacht das Programmieren für GPU Anwendungen, und das portieren bestehender Anwendungen für die GPU erheblich, da sich der Programmierer nicht mehr explizit um das Speichermanagement zwischen CPU und GPU Speicher kümmern muss.

Will man Code kompatibel zu GPU-CPU Umgebungen halten, in denen eine direkte Kommunikationsmöglichkeit zwischen den GPUs nicht besteht, ist mit deutlichen Einschränkungen bei der Leistung, oder Erhöhtem Aufwand bei der Programmierung, zu rechnen.[7]

5 HPC Benchmark

In einer japanischen Studie wurden, für eine bereits bestehende Simulation von Erdbeben, die Leistung eines reines CPU basierten Systems, eines Systems mit Maxwell GPUs, und eines Systems mit Pascal GPUs miteinander verglichen. Die ursprüngliche Entwicklung der Simulation zielte auf einen Einsatz für den K Supercomputer ab. Neues Ziel war es nun, die Simulation auch in anderen, GPU-CPU heterogenen, Umgebungen zu testen. Für die Auslagerung der Berechnungen auf die GPUs wurde OpenACC verwendet, das es ähnlich wie MPI erlaubt, einzelne Programmteile mittels Annotations für die Berechnung auf der GPU zu markieren. Dies sollte den Portierungsaufwand so niedrig wie möglich halten, schöpfte allerdings auch nicht das volle Potential der GPU basierten Systeme aus. Für die Tests wurden acht Knoten des K Supercomputers einem Cluster aus acht Knoten mit jeweils einer K40 GPU und einem NVIDIA DGX-1 System mit ebenfalls acht P100 GPUs gegenübergestellt.
Insgesamt konnten die Japaner bei ihren Tests eine Beschleunigung um den Faktor 3.49 für das Kepler System, und 5.1 für das Pascal System feststellen. Es stellte sich heraus, dass die Implementierung mit OpenACC eine Verbesserung in etwa analog zu den angegebenen Verbesserungen bei der Hardware von Kepler zu Pascal erlaubt. Darüber hinaus wird betont, dass durch die höher Leistung für Berechnungen mit doppelter Fließkommapräzision die Pascal bietet, noch weiteres Verbesserungspotential vorhanden ist. Einer konservativen Schätzung zur Folge könnten weitere Teile der Simulation, bei einem analogen OpenACC Port auf eine Pascal GPU, um den Faktor acht beschleunigt werden, was den Kern der Simulation um mindestens den Faktor elf gegenüber dem reinen CPU System beschleunigen würde. Entsprechend positiv fällt in der Studie auch das Fazit gegenüber den Systemen mit GPUs aus:

> "From these results, it is evident that by using OpenACC and minimal algorithmic development for performance-sensitive kernels, high performance can be achieved in a GPU-CPU heterogeneous compute environment with low development costs. [7]

Ist man darüber hinaus gewillt mehr Aufwand in die Portierung zu stecken, und ein Programm speziell für die verwendete Pascal Hardware anzupassen, ist sicherlich ein nochmals deutlich vergrößertes Leistungspotential vorhanden.

6 Vergleich: Polaris 10 und GP 106

Für einen direkten Vergleich von Polaris und Pascal bietet sich im Moment wohl am ehesten ein Vergleich zwischen Polaris 10 und GP106 an. Polaris 10 ist AMDs momentan leistungsstärkster Polaris Chip, und der GP106 das Nvidia Modell das von den Spezifikationen am nächsten kommt. Auch preislich sind die Grafikkarten auf denen diese Chips verbaut werden (RX 480 und GTX 1060) sich momentan sehr ähnlich.

6.1 Performance

Im Vergleich zur RX 480 liefert die mit 6 GB Speicher bestückte Version der Geforce GTX 1060 in DirectX12-Spielen weniger Bilder pro Sekunde. Dafür liegt sie in DirectX11-Spielen meist vor der AMD-Karte.

> "Die GeForce GTX 1060 Founders Edition [ist beim Rendern unter DriectX11] in 1.920 * 1.080 im Durchschnitt sechs Prozent schneller als die AMD Radeon RX 480. Das anvisierte zweite Ziel, die GeForce GTX 980 zu erreichen, wird mit Ach und Krach geschafft."[3]

Betrachtet man einen Benchmark abseits der Berechnung von 3D-Grafik in Spielen, so ist allerdings beim GPU Computing die RX 480 mit 10-14 Prozent schneller. Bei diesen Berechnungen macht sich wohl vor allem der größere Speicher (8 GB gegen 6 GB) bei der RX 480 bemerkbar, da kaum ein Spiel heutzutage auch nur die 6 GB Videospeicher der GTX 1060 ausnutzt. Die Speicherkompression von Pascal funktioniert im Vergleich zur RX 480 hier auch schlechter als in Spielen, so dass die 2 GB Unterschied beim Speicher mehr ins Gewicht fallen.[3]

6.2 Energieeffizienz

Die GTX 1060 profitiert von der aktuellen Pascal-Architektur und besitzt eine hervorragende Energieeffizienz. Im Vergleich zur RX 480 braucht sie unter Last trotz durchschnittlich besserer Performance bei Berechnungen unter DirectX 11 deutlich weniger Strom:

> "Im Spielebetrieb kommt die GeForce GTX 1060 Founders Edition im Durchschnitt auf 188 Watt. Das sind 44 Watt weniger als die direkte Konkurrenz, die AMD Radeon RX 480 und zeigt zugleich die größte Stärke der Pascal-Architektur: die sehr gute Energieeffizienz, gegen die AMD auch mit der Polaris-Generation nichts entgegenstellen kann." [3]

Die deutlich verringerte Leistungsaufnahme lässt sich somit zumindest in den unteren Preissegmenten als größter Pluspunkt von Pascal gegenüber Polaris feststellen.

7 Fazit

Durch die Betrachtung der bereits veröffentlichen Beschleunigerkarten lässt sich feststellen, dass momentan ein Vergleich zwischen Pascal und Polars basierten GPUs nur im unteren Leistungssegment sinnvoll ist. Während Nvidias aktuelles Lineup zwar schon größtenteils komplett ist (mit Ausnahme der wohl zu erwartenden GTX 1080 TI), macht im HPC-Bereich ein Vergleich von Leistung und Effizienz erst zwischen Pascal und AMDs Anfang 2017 erscheinenden Vega Karten Sinn. Zu eindeutig sind die Unterschiedlichen Ausrichtungen bei den veröffentlichten Pascal und Polaris GPUs im Moment.

7.1 Zusammenfassung

Betrachtete GPUs Im Rahmen dieser Arbeit wurden die sich aktuell im Markt befindlichen GPUs mit den neuen Architekturen Polaris (AMD) und Pascal (Nvidia) betrachtet. Bei Pascal lag ein besonderer Fokus auf den Modellen Tesla P100 als Modell für HPC-Systeme und GTX 1080 als high-end Karte für 3D-Anwendungen. Der Fokus bei Polaris lag auf der Radeon RX 480 als derzeit schnellstes Polaris Modell.
Insgesamt lag der Schwerpunkt bei den Pascal Modellen, begründet darin, dass die Neuerungen bei den Nvidia Karten tiefgreifender sind als auf der Seite von AMD, auf welcher Polaris eher eine sanfte Weiterentwicklung darstellt, und einschneidendere Änderungen erst mit Vega zu erwarten sind.

Leistung und Effizienz Unterm Strich sind sowohl Pascal als auch Polaris deutliche Verbesserungen im Vergleich zu ihren Vorgängern. Die GPUs beider Architekturen liefern bis zum eineinhalbfachen an Leistung bei selbem Energiebedarf, vergleicht man sie mit Karten der selben Preisklasse aus der Vorgängergeneration. Auch der Vergleich zwischen Nvidia und AMD liefert im selben Preissegment sehr ähnliche Ergebnisse. Nur im HPC und generell im high-end Sektor ist Nvidia konkurrenzlos, schlicht aufgrund dessen, dass aktuell von AMD zu Titan X und Tesla P100 noch keine Konkurrenzkarten veröffentlicht wurden.
Der Hauptgrund für die Verbesserung gegenüber Maxwell und Fiji dürfte mehr noch als in Änderung an der Chiparchitektur in der Verkleinerung der Strukturbreiten, und der damit gestiegener Anzahl an Transistoren, zu finden sein. Sowohl für Nvidia als auch für AMD waren die beiden Vorgängergenerationen im Vergleich zu deren Vorgängern vor allem Updates der Architektur.

Beim Sprung zu Pascal und Polaris hingegen haben sich auch die Strukturbreiten wieder verkleinert, was sich schon bei den letzten Generationswechseln von Grafikkarten als der größte Motor für Leistungssprünge herausgestellt hat.

7.2 Ausblick

AMD Vega Über AMDs für 2017 geplante neue Architektur Vega sind noch relativ wenige feste Fakten bekannt. Es ist davon auszugehen, dass die Basis des Chips wie Polaris GCN Gen. 4 sein wird. Im Moment scheint es, dass Vega, wie auch Pascal es schon beherrscht, zwei 16 Bit Floatingpointoperationen gleichzeitig auf den FP32 Einheiten auszuführen können wird. Dies wird mit einer vermuteten 25 Teraflops FP16-Leistung begründet.[2] Das lässt auch Rückschlüsse auf die FP32 Taktrate zu:

> "Die 25 Teraflops FP16-Performance würden also automatisch 12,5 TFLOPs FP32-Leistung bedeuten. Das wären 45 Prozent mehr als die Fiji-GPU auf der Radeon R9 Fury X bietet." [2]

Für die angekündigte stärkste HPC Karte mit Vega Chip, die wohl den Namen MI25 tragen wird, wird eine Speicherbandbreite von 512 GB/s vermutet.[2]

Fernere Zukunft Für die Zeit nach Vega hat AMD für 2019 die GPU-Generation Navi angekündigt: Navi-GPUs sollen sich vor allem durch eine besonders hohe Skalierbarkeit profilieren und auf einer noch nicht angekündigten zukünftigen Speichertechnik basieren. Es ist also offenbar schon eine Nachfolgetechnologie von HBM2 geplant, welche wiederum selbst ja erst seit kurzem im Markt ist.
Die nächste GPU Generation von Nvidia wird Volta heißen und soll schon 2018 erscheinen. Laut Gerüchten sollen alle GPUs über HBM2 verfügen.
Bei beiden Firmen ist wohl zunächst keine weitere Verkleinerung der Strukturbreite zu erwarten, sondern vor allem Architekturoptimierungen.

8 Literatur

Literatur

[1] AMD. Dissecting the polaris architecture whitepaper, 2016.

[2] Wolfgang Andermahr. AMD nennt 12,5 TFLOPs, NCUs und 512 GB/s, 2016. https://www.computerbase.de/2016-12/amd-vega-details/.

[3] Wolfgang Andermahr. Drei Mal GP106 im Duell mit der Radeon RX 480, 2016. https://www.computerbase.de/2016-07/geforce-gtx-1060-test/3/.

[4] Wolfgang Andermahr. So viel Leistung bringt die 4. Generation GCN, 2016. https://www.computerbase.de/2016-08/amd-radeon-polaris-architektur-performance/.

[5] NVIDIA Corporation. Nvidia geforce gtx 1080 whitepaper, 2016.

[6] NVIDIA Corporation. Nvidia tesla p100 whitepaper, 2016.

[7] Kohei Fujita, Takuma Yamaguchi, Tsuyoshi Ichimura, Muneo Hori, and Lalith Maddegedara. Acceleration of element-by-element kernel in unstructured implicit low-order finite-element earthquake simulation using openacc on pascal gpus, 2016.

[8] Marc Sauter. Der 14-Nanometer-Schwindel, 2015. http://www.golem.de/news/fertigungstechnik-der-14-nanometer-schwindel-1502-112524.html.

[9] Anton Shilov. JEDEC Publishes HBM2 Specification as Samsung Begins Mass Production of Chips, 2016. http://www.anandtech.com/show/9969/jedec-publishes-hbm2-specification.

[10] Christof Windeck. Schneller im Vierviertelteltakt - Speicher für Grafikkarten: auf

GDDR5-SGRAM folgt GDDR5X, 2016. https://www.heise.de/ct/ausgabe/2016-13-Speicher-fuer-Grafikkarten-auf-GDDR5-SGRAM-folgt-GDDR5X-3228173.html.