Seminar Hochleistungsrechner: Aktuelle Trends und Entwicklungen
Winter Term 2015/2016

Aktuelle Speichertechnologien: HBM/HMC

Andreas Seibold
Technische Universität München

20.12.2015

Zusammenfassung

Seit den Anfängen der Computerentwicklung hat sich das Verhältnis der Geschwindigkeiten zwischen Speicher und Prozessor zugunsten der Prozessoren immer weiter vergrößert. Heutige Speichertechnologien müssen versuchen diesem Trend entgegenzuwirken und deutliche Zugewinne bei der Übertragungsrate vorzeigen. Zwei sehr ähnliche Technologien, HMC und HBM, setzen auf gestapelten DRAM, der mit Verbindungskanälen an einen Logikblock angeschlossen ist. Der Speicher wird dreidimensional aufgebaut und erhält eine breite Schnittstelle, sodass bei geringerem Energieverbrauch eine höhere Durchsatzrate erreicht werden kann.

1 Einleitung

Der anhaltende Trend der immer schneller werdenden Prozessoren im Vergleich zur eher langsam steigenden Datenrate des Hauptspeichers ist ein ernstes Problem. Bei Heimanwendungen, wie normalen Officearbeiten oder Internetsurfen fällt es nicht sehr ins Gewicht, jedoch beim High Performance Computing sehr wohl. Moderne Mehrkernprozessoren befinden sich 3 aus 4 Takten im Leerlauf. Die nutzbare Leistung liegt dabei bei oftmals bei weniger als 10% der Spitzenbelastung.[3]

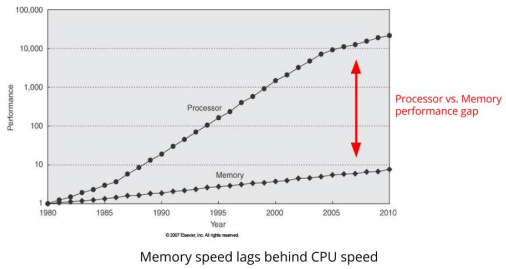


Abbildung 1: Die Entwicklung des Performanceunterschieds zwischen Prozessoren und Hauptspeicher von 1980 bis 2010. Quelle [1]

Diese sogenannte „Memory Wall“ gilt es zu durchbrechen, um neue Spitzenwerte im Supercomputingbereich zu erreichen. Die prognostizierten Speicherbandbreiten sind für 2018 bereits auf 200 bis 400 GB/s pro Rechnerknoten geschätzt[4], andere sprechen von 500GB/s[6]. Die maximale Übertragungsrate eines DDR3 mit einer Zugriffsbreite von 8 Byte ist mit 2GB/s jedoch sehr langsam und auch GDDR5 Speicher mit einer Zugriffsbreite von 32 Byte erreicht nur 32GB/s. Stapelbare Technologien, sogenannte 3D-Chips, kombinieren mehrere herkömmliche DRAM-Bausteine zu einem großen Speicherblock, der dann bis zu 256GB/s erzielen kann[4].

Die Stapelung wird mit Hilfe sogenannter Wege durch das Silikon, Through Silicon Via (TSV) ermöglicht. Dies wird in Abschnitt 2 erläutert. Die Sektion 3 erklärt den Aufbau der neuen Speichertechnologien High Bandwidth Memory (HBM) und Hybrid Memory Cube (HMC). Anschließend werden in Abschnitt 4 die beiden Speicherarten mit herkömmlichen Ansätzen verglichen. Abschließend wird in Sektion 5 ein Ausblick gegeben.

2 Through Silicon Via

Die 3D-Chip Technologie ist eine vergleichsweise neue Technologie und wird nach und nach weiterentwickelt sowie verbessert[6][5]. Die einzelnen DRAM-Blöcke werden an bestimmten Stellen in der Vertikalen mit einem leitenden Material durchzogen, die die Kanäle der Speicherblöcke jeweils mit dem Logikchip verbindet.

Dafür werden, wie in Abbildung 2 gezeigt, kleine Mikroerhebungen an die TSV angebracht, um eine leitende Verbindung herzustellen. Dafür muss aus dem Material ein Teil in Form eines Kanals entfernt werden und anschließend mit einer Schutzschicht ausgekleidet werden. Die Schutzschicht hat sowohl eine leitende als auch eine nicht leitende Schicht, so dass sie Verbindung in der Horizontalen in den DRAM-Block geführt werden kann. Danach wird das elektrisch leitende Füllmaterial in den Kanal eingefüllt und der Kontakt fertiggestellt[8].

3 High Bandwidth Memory und Hybrid Memory Cubes

Die beiden Speichertechnologien sind sich im generellen Aufbau sehr ähnlich und setzen beide auf gestapelte mit TSV verbundene DRAM-Blöcke, die mit einem Logikchip auf der untersten Ebene kommunizieren. Die Organisation der DRAM-Blöcke unterscheidet sich jedoch etwas.

Abbildung 3 zeigt einen High Bandwidth Memory Baustein, der mit insgesamt 1024 TSV

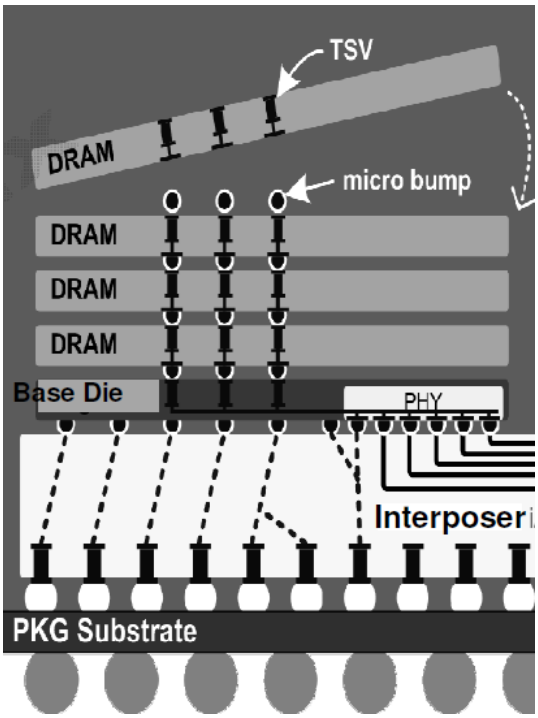


Abbildung 2: HBM mit Through Silicon Via. Quelle [4]

Ein-/Ausgabe Verbindungen ausgestattet ist. Diese sind alle mit dem Logikblock verbunden und werden von dort an die vier darüberliegenden DRAM-Bausteine aufgeteilt. Jeder Baustein ist in zwei Bereiche aufgeteilt, so dass insgesamt 8 Kanaäle mit jeweils 128 TSV-Verbindungen genutzt werden. Jeder Kanal hat unabhängige Adress- und Daten-TSV mit einer Point-to-Point-Verbindung, um eine Isolierung der Kanäle zueinander zu gewährleisten[6].

Hybrid Memory Cubes werden ebenfalls mit gestapelten DRAM-Blöcken und TSV aufgebaut. Die Organisation wird jedoch anders gemacht. Die Speicherbereiche werden in sogenannte Vaults unterteilt (vgl. Abbildung 4), die in der Vertikalen übereinanderliegende Partitionen der der einzelnen Ebenen zusammenfasst. Dabei ist jeder Vault funktional und operativ unabhängig. Zudem haben alle

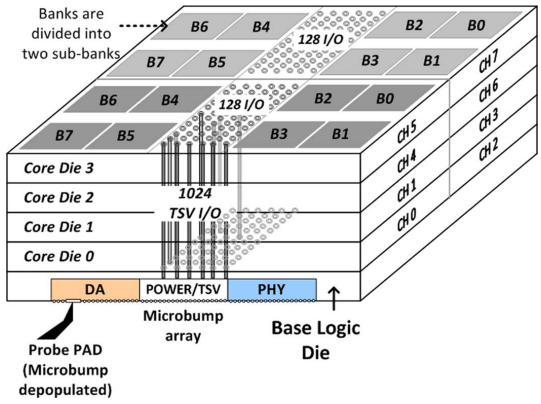


Abbildung 3: Aufbau eines gestapelten DRAM High Bandwidth Memory. Quelle [6]

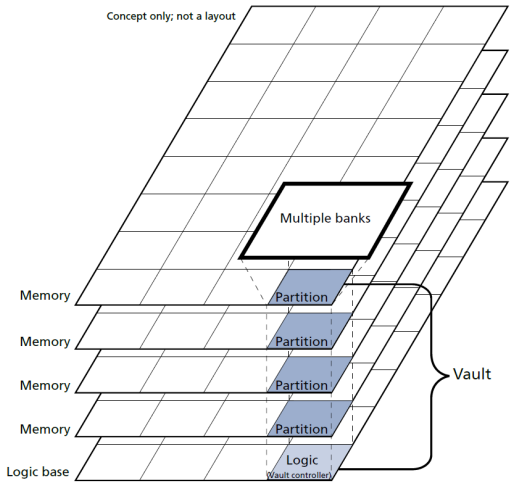


Abbildung 4: Aufbau eines Hybrid Memory Cubes. Quelle [2]

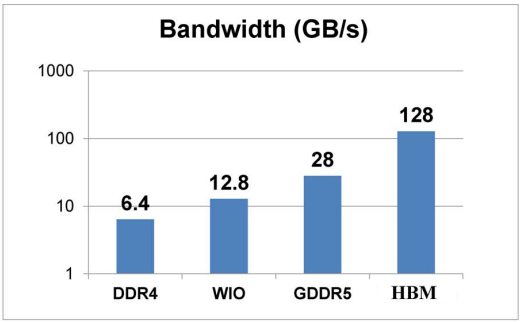
einen eigenen Speichercontroller (Vault-Controller) in der untersten Logikebene, der alle Speicherreferenzoperationen innerhalb des zugehörigen Vaults regelt.Timing-Anforderungen sowie Refresh-Operationen werden vom Vault-Controller übernommen, so dass diese Funkionalität aus dem Hostspeichercontroller verschwindet[2].

Weiterhin kann jeder Vault-Controller eine Puffer für Referenzen für den zugehörigen Speicher besitzen. So kann ein Controller Referenzen je nach Bedarf aus dem Puffer bearbeiten anstatt in der ankommenden Reihenfolge der Anfragen. Daher sind Antworten vo Vault-Operationen zurück zur seriellen Ein- und Ausgabe Out-Of-Order. Jedoch bleiben Anfragen einer externen seriellen Verbindung zum selben Vault in Reihenfolge[2].

Zudem lassen sich die einzelnen HMC-Bausteine durch den Aufbau des Logikchips, der wie ein Switch organisiert ist, miteinander verknüpfen[3]. Sie können als Kette organisiert werden oder als Sternmuster. Ebenfalls möglich ist ein Aufbau mit mehreren Hostchips[2].

4 Leistungsvergleich

Die Performanz der gestapelten Speicher zeigt sich in Abbildung 5. Mit höherer Speicherdichte und mehr Ein-/Ausgabeanbindungen als bei den anderen Technologien kann bei gleicher oder sogar niedrigeren Spannung (gegenüber GDDR5) ein deutlicher Durchsatzvorteil für HBM ausgemacht werden. Die logarithmische Skala ist hier zu beachten. Der geringere Energieverbrauch ergibt sich aus den deutlich kürzeren Wegen, die die Daten zurücklegen müssen, da die Speicherblöcke übereinander angeordnet sind[6]. Eine um bis zu 68 % bessere Energieeffizienz ist erreichbar[4][3].
Neue Versionen von HBM sollen bis zu 256GB/s Übertragungsrate erreichen[4], die einer Geschwindigkeit von 160GB/s von HMC gegenüberstehen[5]. Diese wird in der nächsten Version auf bis zu 320GB/s angehoben. Kombiniert man mehrere Chips, sind Werte von 512GB/s bis 1TB/s möglich, die sich bei aktuellen Technologien auf lediglich 336GB/s belaufen[7].



Type	DDR4	WIO	GDDR5	HBM
Total density	4Gb	2Gb	2Gb	**8Gb**
Total IO width	16	512	32	**1024**
Total # of banks	16	16	16	**64**
Total # of channels	1	4	1	**8**
Supply voltage	1.2v	1.2v	1.5v	**1.2v**

Abbildung 5: Bandbreite verschiedener Speichertechnologien. Quelle [6]

5 Ausblick

Die gebotene Leistung der neuen Technologien wird sicherlich dem High Performance Markt einen Schub geben und noch schnellere und weniger speicherlimitierte Großrechner ermöglichen. Welche der beiden Technologien, HBM und HMC, wo seinen Verwendungszweck findet, kann noch nicht abgeschätzt werden. High Bandwidth Memory wird jedoch bereits von AMD in der neuesten Grafikkartengeneration verwendet und von NVIDIA[7] ebenfalls für die kommenden Produkte fokussiert. Hier scheint die Entscheidung bereits gefallen zu sein. Da Grafikkarten seit einigen Jahren aus dem HPC Bereich nicht mehr wegzudenken sind, wird wohl HBM Speicher sehr sicher ein Teil zukünftiger Hochleistungssysteme sein.

Literatur

[1] Gocon-2014-10, 2014. http://dave.cheney.net/wp-content/uploads/2014/06/Gocon-2014-10.jpg.

[2] Hybrid Memory Cube Consortium. Hybrid Memory Cube Specification 2.0, 2014. http://www.hybridmemorycube.org/.

[3] Joe Jeddeloh and Brent Keeth. Hybrid memory cube new dram architecture increases density and performance. In *2012 Symposium on VLSI Technology (VLSIT)*, 2012.

[4] Joonyoung Kim and Younsu Kim. HBM: Memory Solution for Bandwidth-Hungry Processors, 2014. http://www.hotchips.org/wp-content/uploads/hc_archives/hc26/HC26-11-day1-epub/HC26.11-3-Technology-epub/HC26.11.310-HBM-Bandwidth-Kim-Hynix-Hot

[5] Mitsumasa Koyanagi. Heterogeneous 3d integration—technology enabler toward future super-chip. In *Electron Devices Meeting (IEDM), 2013 IEEE International*, pages 1–2. IEEE, 2013.

[6] Dong Uk Lee, Kyung Whan Kim, Kwan Weon Kim, Hongjung Kim, Ju Young Kim, Young Jun Park, Jae Hwan Kim, Dae Suk Kim, Heat Bit Park, Jin Wook Shin, et al. 25.2 a 1.2 v 8gb 8-channel 128gb/s high-bandwidth memory (hbm) stacked dram with effective microbump i/o test methods using 29nm process and tsv. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pages 432–433. IEEE, 2014.

[7] Mike O'Connor. Highlights of the High - Bandwidth Memory (HBM) Standard, 2014. http://www.cs.utah.edu/thememoryforum/mike.pdf.

[8] Makoto Motoyoshi. Through-silicon via (tsv). *Proceedings of the IEEE*, 97(1):43–48, 2009.