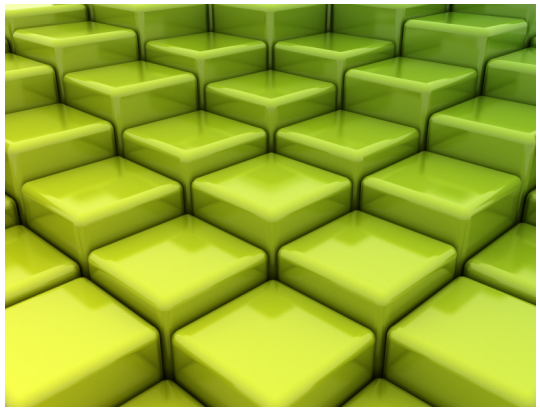
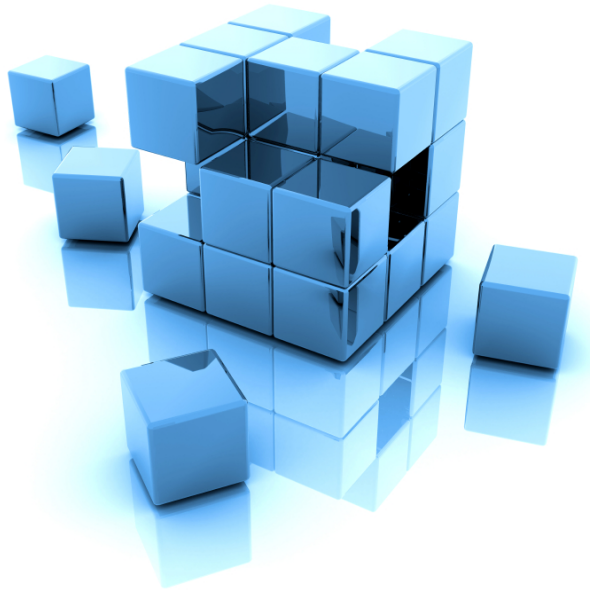
Stacked DRAM: The Hybrid Memory Cube



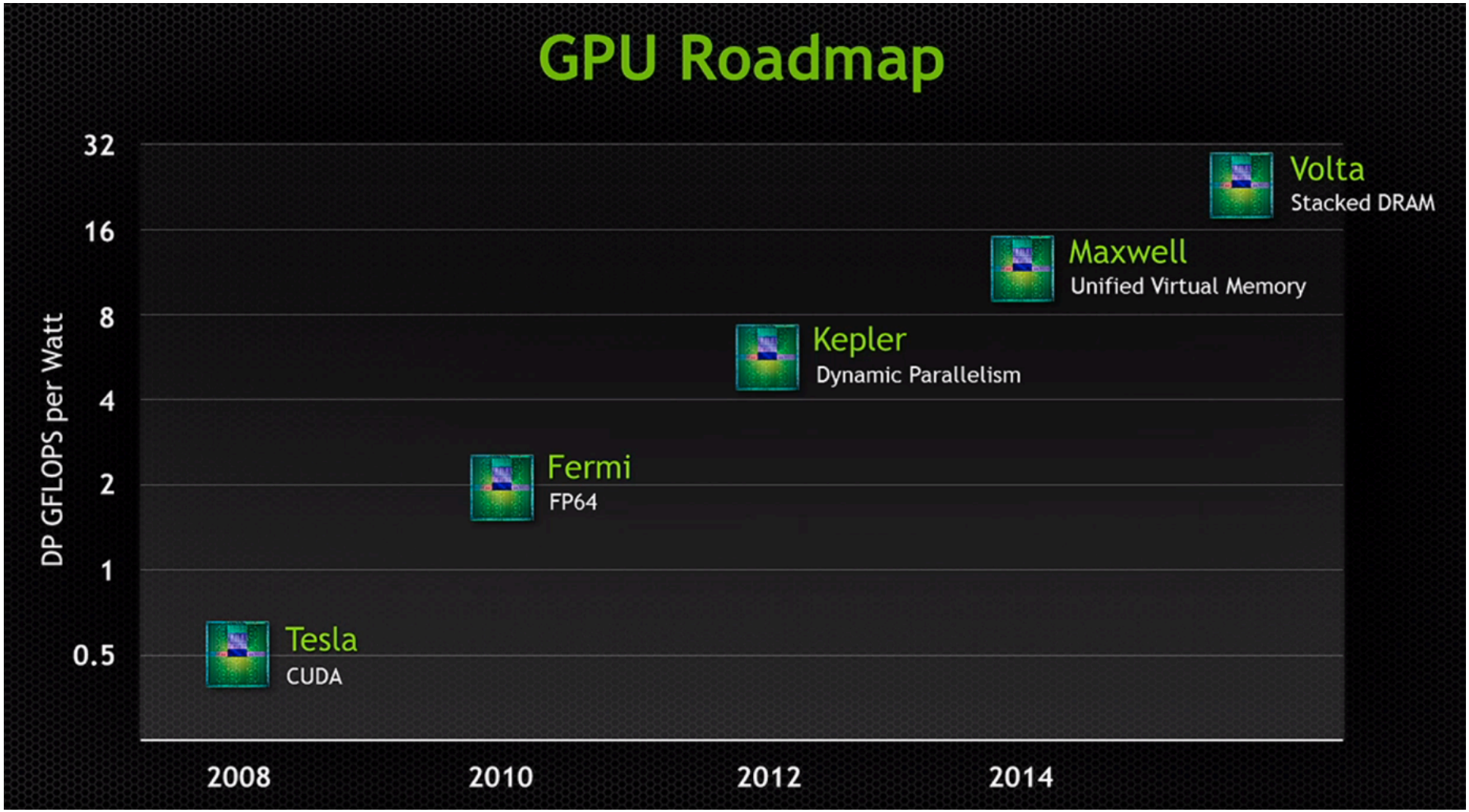
Manuel Ujaldon

Computer Architecture Department

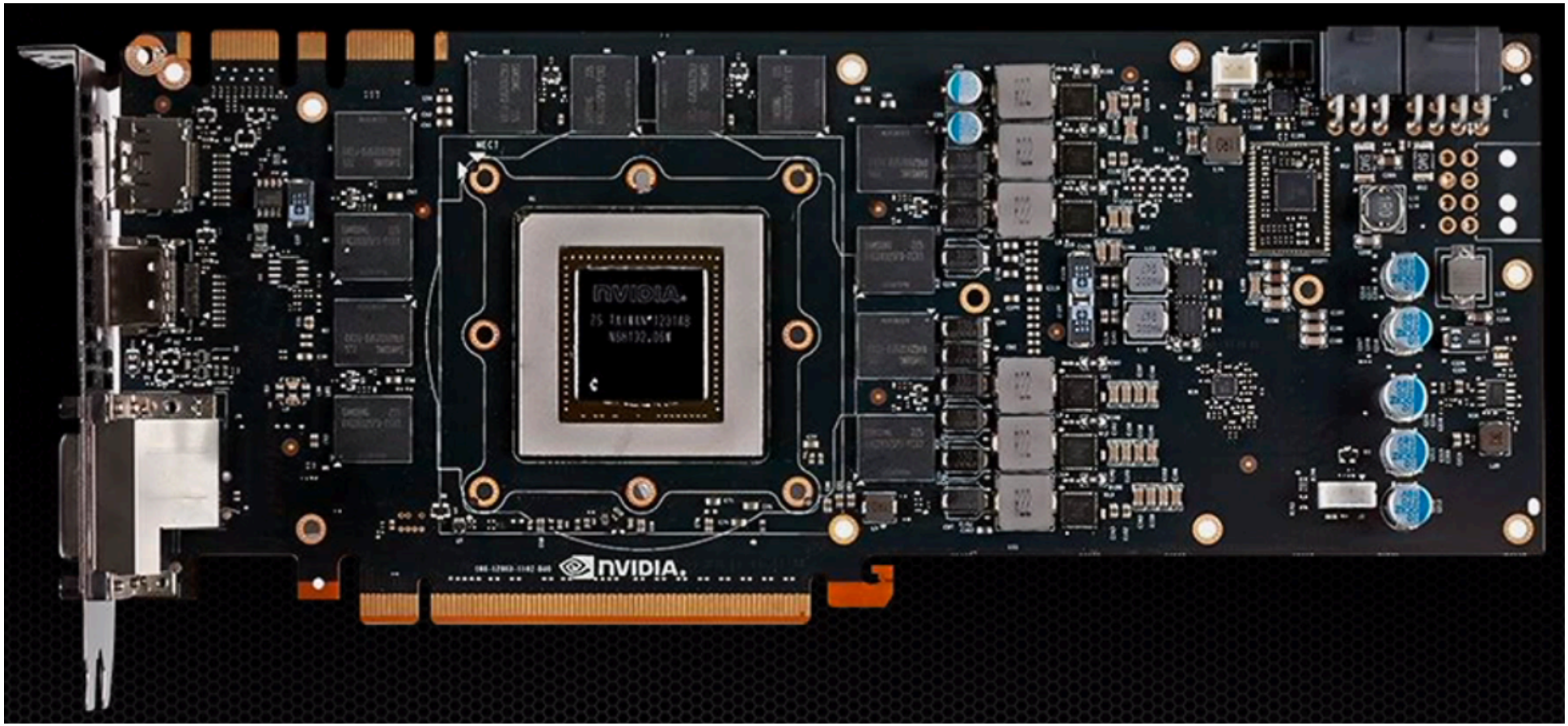
University of Malaga



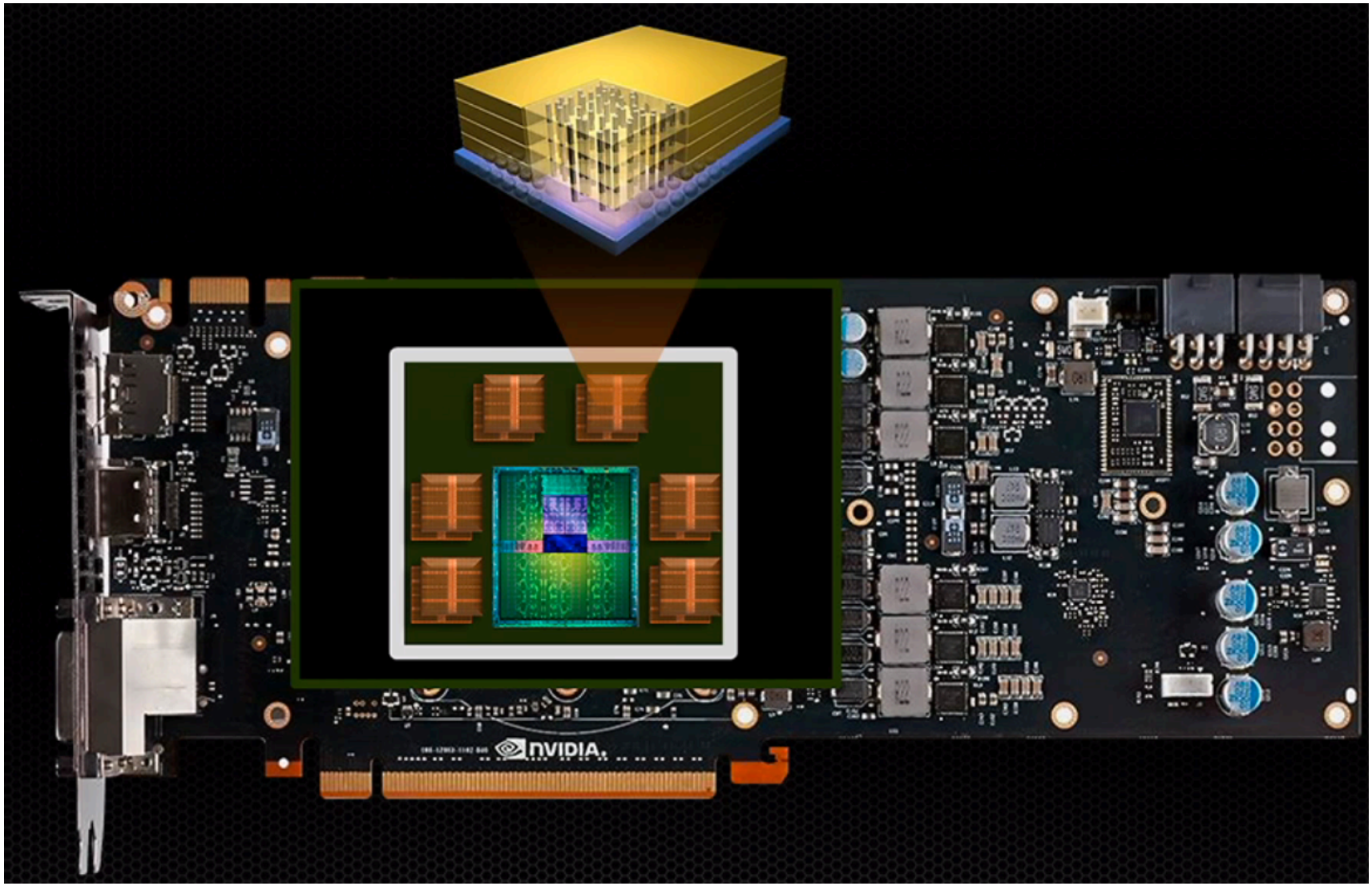
A look ahead through Nvidia's GPU roadmap



A 2013 graphics card: Kepler GPU with GDDR5 video memory



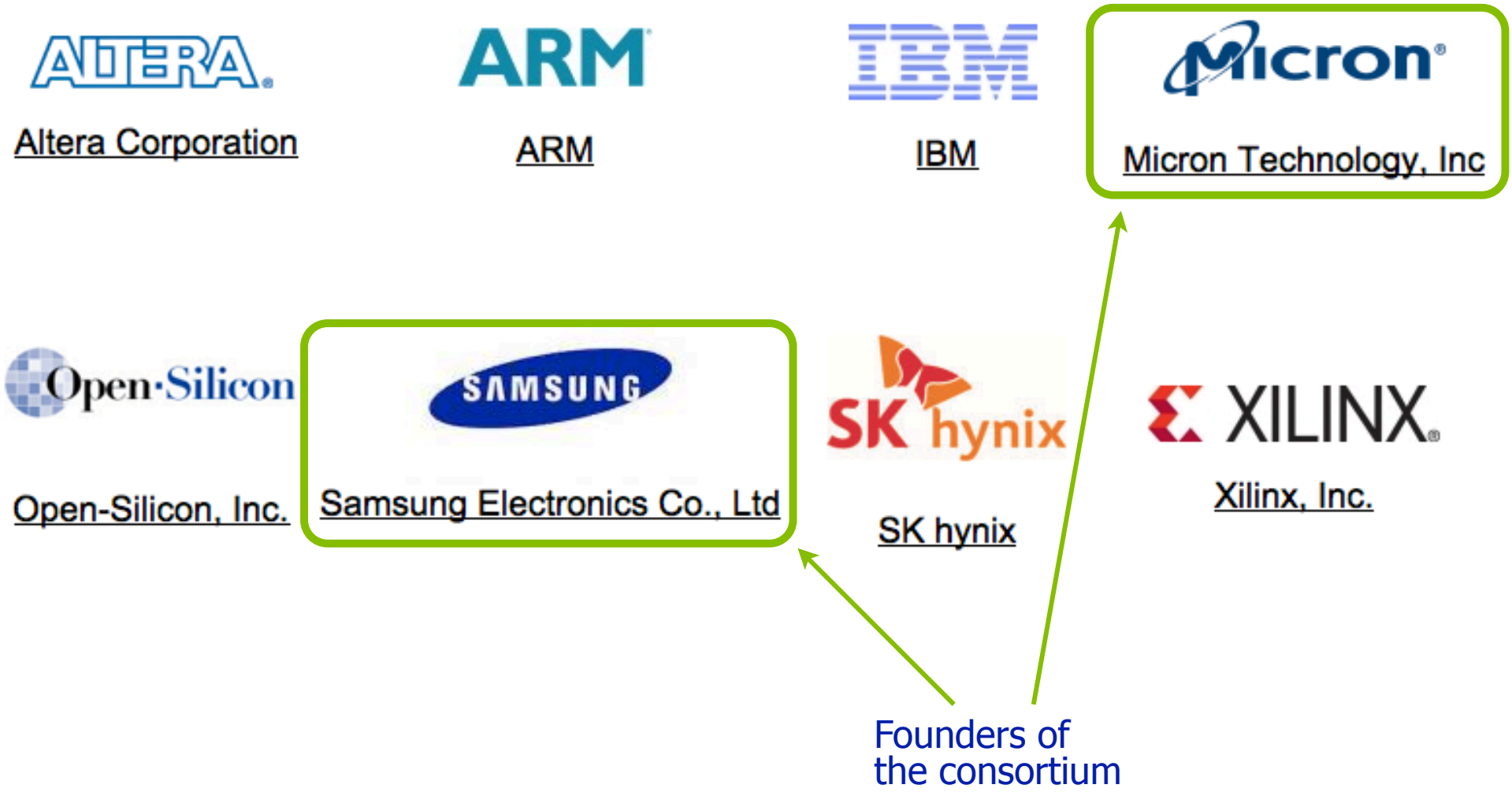
A 2017 graphics card: Volta GPU with Stacked DRAM



A promising Stacked DRAM development: The Hybrid Memory Cube Consortium (HMCC)

HMCC achievements and milestones	Date
First papers published about Stacked DRAM (based of research projects)	2005, 2006
First commercial announcement of the technology	February, 2011
HMC Consortium is launched by Micron Technologies and Samsung Electronics	October, 2011
Stacked DRAM announced for Volta GPU by Nvidia	March, 2013
Specification 1.0 available	April, 2013
Production samples	Second half of 2014 (estimated)
2.5 configuration available	End of 2014 (estimated)

Developer members of HMCC (as of May'13)



Broader adoption

- HMC was primarily oriented to HPC and networking, but it can also be useful for mobile and DDR-like technol.
- HMC is tightly coupled with CPUs, GPUs and ASICS in point-to-point configurations, where HMC performance is available for optical memory bandwidth.

Adopter Members:

- Accel, Ltd.
- Achronix Semiconductor Corporation
- ADATA Technology Co., Ltd.
- AIRBUS
- Altior
- Analog Bits
- APIC Corporation
- Arira Design
- Arnold & Richter Cine Technik
- Atria Logic, Inc.
- BroadPak Corporation
- Cadence Design Systems, Inc.
- Cascade Microtech
- Convey Computer Corporation
- Cray Inc.
- DAVE Srl
- Design Magnitude Inc.
- Dream Chip Technologies GmbH
- eSilicon Corporation
- Exablade Corporation
- EZchip Semiconductor Ltd.
- FirstPass Engineering
- FormFactor, Inc.
- Fujitsu Advanced Technologies Ltd.
- Galaxy Computer System Co., Ltd.
- GDA Technologies
- GLOBALFOUNDRIES
- GraphStream Incorporated
- Green Wave Systems Inc.
- HGST, a Western Digital Company
- HiSilicon Technologies Co., Ltd.
- HOY Technologies
- Huawei Technologies
- Industrial Technology Research Institute (ITRI)
- Infinera Corporation
- Inphi Corporation
- Integrated Device Technology
- Ircona
- ISI / Nallatech
- Juniper Networks
- KALRAY
- Kool Chip Inc.
- Korea Advanced Institute of Science and Technology
- Lawrence Livermore National Laboratory
- LeCroy Corporation
- Liquid Logic, LLC
- LogicLink Design, Inc.
- Lomonosov Moscow State University
- Luxtera, Inc.
- Marvell
- Mattozetta Technologies
- Maxeler Technologies Ltd.
- MediaTek
- Memoir Systems Inc.
- Mentor Graphics
- Miranda Technologies Partnership
- Mobiveil, Inc.
- Montage Technology, Inc.
- Napatech A/S
- National Instruments
- NEC Corporation
- Netronome
- New Global Technology
- Northwest Logic
- Obsidian Research
- OmniPhy
- Oregon Synthesis
- Perfcraft
- Pico Computing
- Renesas Electronics Corporation
- Science & Technology Innovations
- SEAKR Engineering
- SIMMTECH Co., Ltd.
- Somerset Technology Services, Inc.
- STMicroelectronics
- Suitcase TV Ltd.
- T-Platforms
- Tabula
- Tech-Trek Ltd.
- Technion - Israel Institute of Technology
- Teledyne LeCroy
- Teradyne, Inc
- The Regents of the University of California
- Tilera Corporation
- Tongji University
- TU Kaiserslautern, Lehrstuhl Entwurf Mikroelektronischer Systeme
- UC Irvine
- United Microelectronics Corporation
- University of Heidelberg ZITI (Center for Computer Engineering
- University of Rochester
- University of Southern California
- Winbond Electronics Corporation
- Woodward McCoach, Inc.
- ZTE Corporation

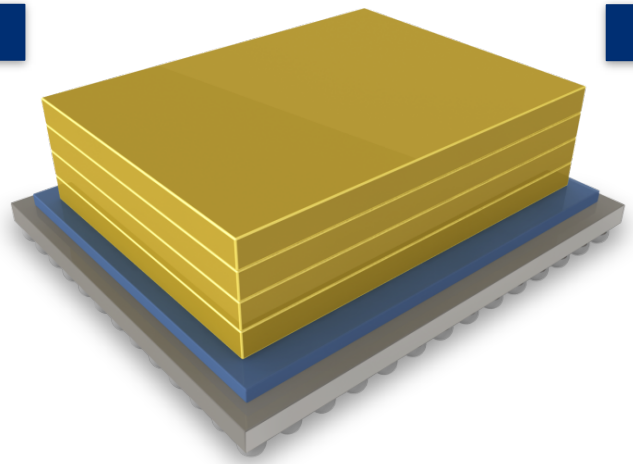
The Hybrid Memory Cube at a glance

Revolutionary Approach to Break Through the "Memory Wall"

- Evolutionary DRAM roadmaps hit limitations of bandwidth and power efficiency
- Micron introduces a new class of memory: Hybrid Memory Cube
- Unique combination of DRAMs on Logic

Key Features

- Micron-designed logic controller
- High speed link to CPU
- Massively parallel "Through Silicon Via" connection to DRAM



Unparalleled performance

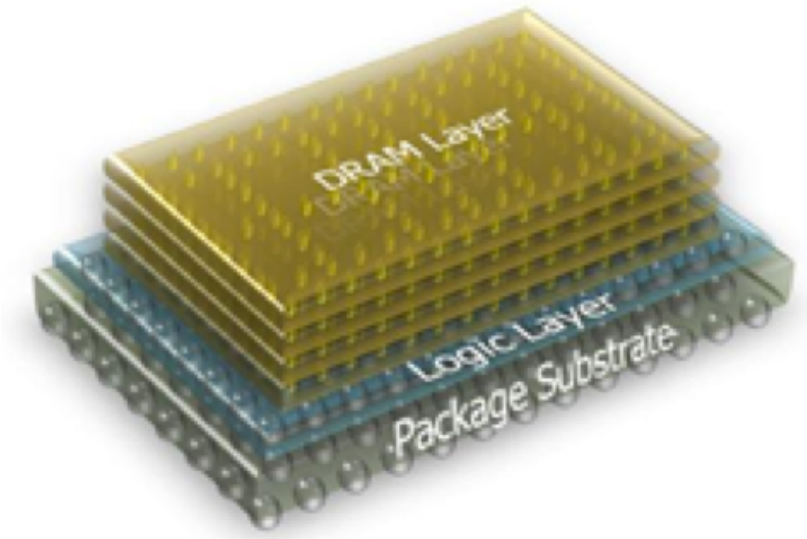
- Up to 15X the bandwidth of a DDR3 module
- 70% less energy usage per bit than existing technologies
- Occupying nearly 90% less space than today's RDIMMs

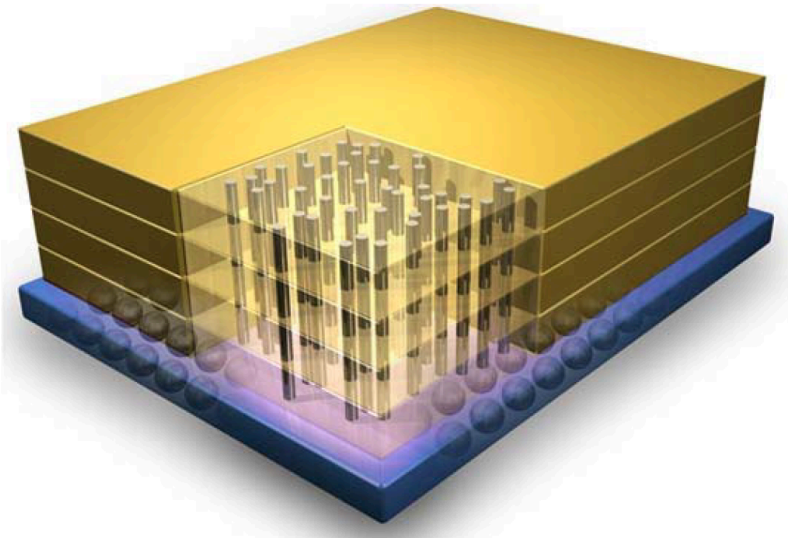
Full silicon prototypes in silicon
TODAY

Targeting high performance computing and networking, eventually migrating into computing and consumer

Architectural highlights

- Stacked DRAM is an abstracted memory management layer.
- The traditional DRAM core cell architecture is restructured to use memory vaults rather than arrays.
- A logic controller is placed at the base of the DRAM stack.
- The assembly is interconnected with through-silicon vias (TSVs) that go up and down the stack.
- The final step is advanced package assembly.

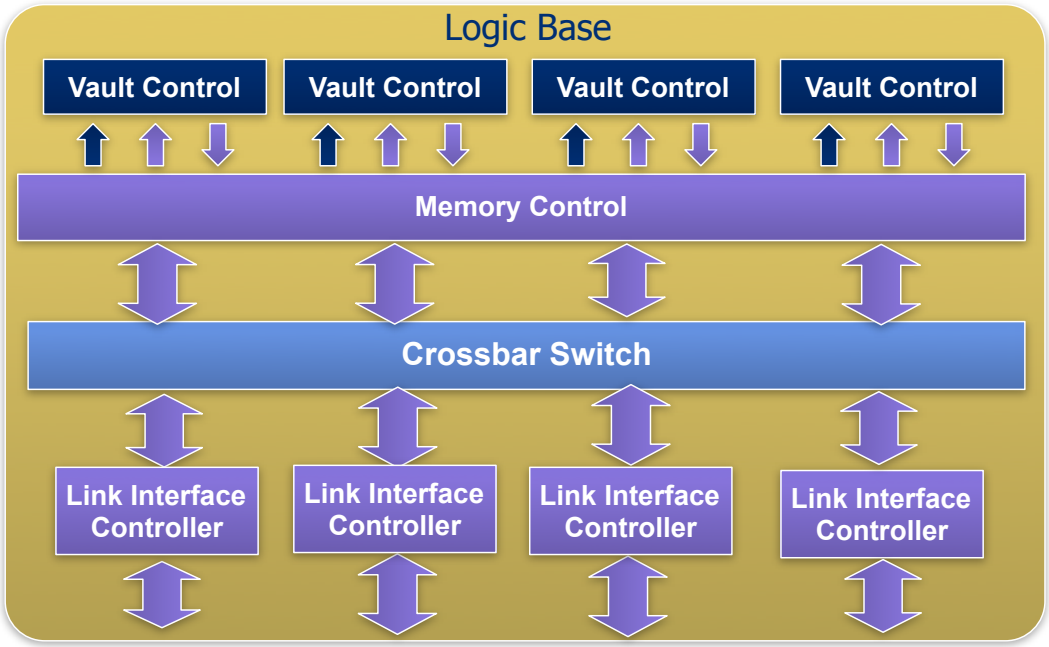




Architectural details

1. DRAM is partitioned into 16 parts like DDR3 and DDR4.
2. Common logic is extracted from all partitions.
3. DRAM is piled up in 4-high or 8-high configurations.
4. Common logic is re-inserted at the logic base die.
5. 16 vaults are built. Each consists of either 4 or 8 parts of each layer plus logic underneath, and can be thought of as individual channels in the regular architecture.
6. A high speed link connects DRAM and processor, with:
    1. Advanced switching.
    2. Optimized memory control.
    3. Simple interface.
    4. 16 transmits and receive lanes, each running at 10 GB/s.

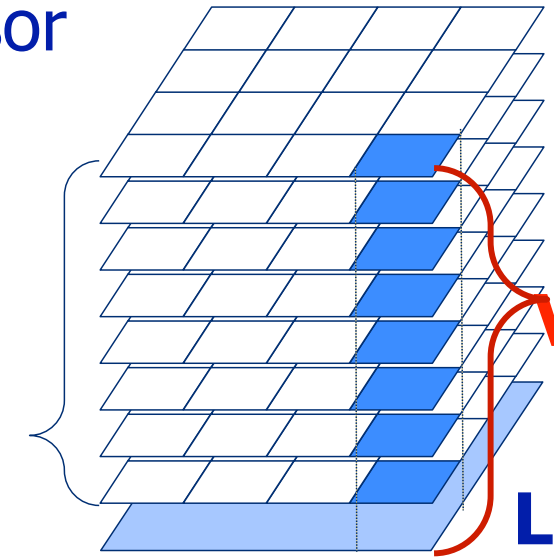
HMC Architecture



Add advanced switching, optimized memory control and simple interface to host processor(s)…

Processor Links

3DI & TSV Technology



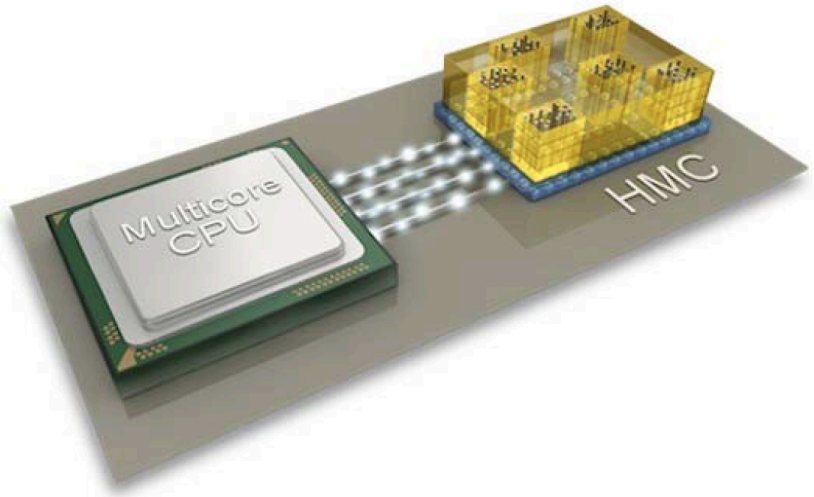
DRAM

Vault

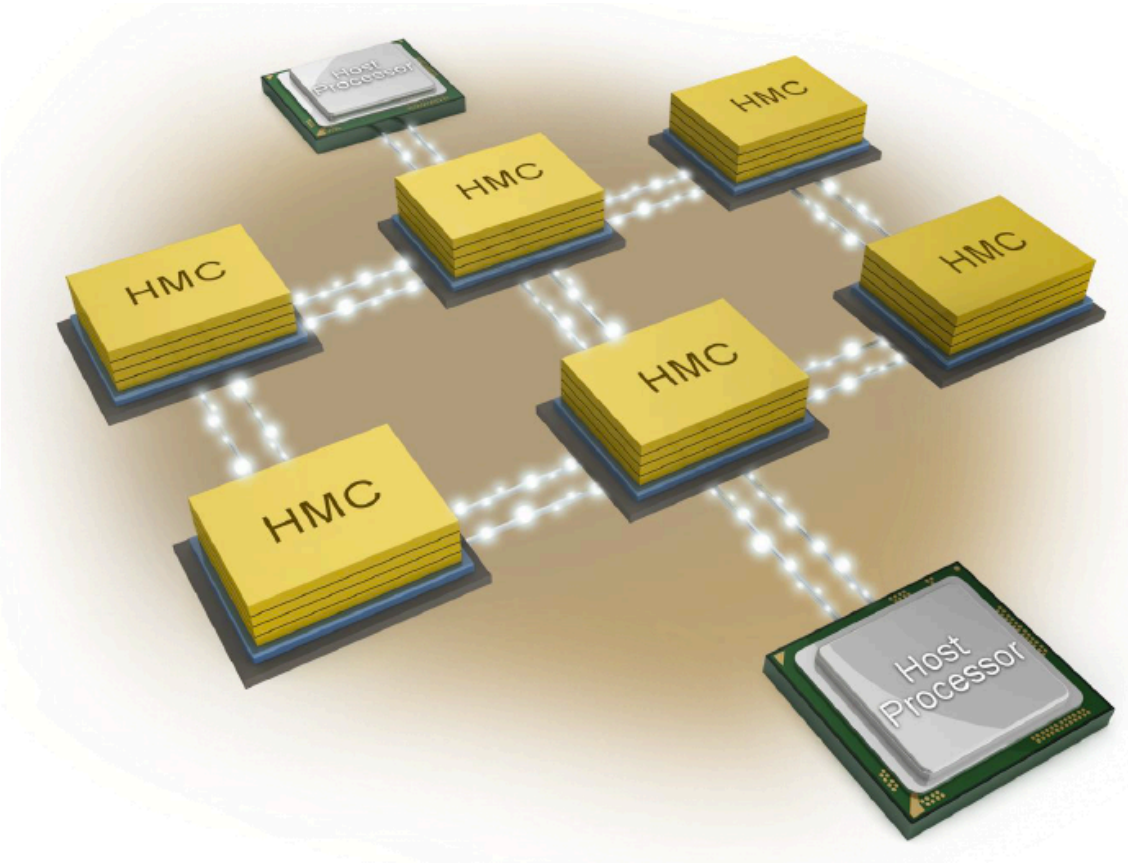
Logic Base

HMC supports stacked DRAM in two different flavours: Near memory and far memory

- Near memory:

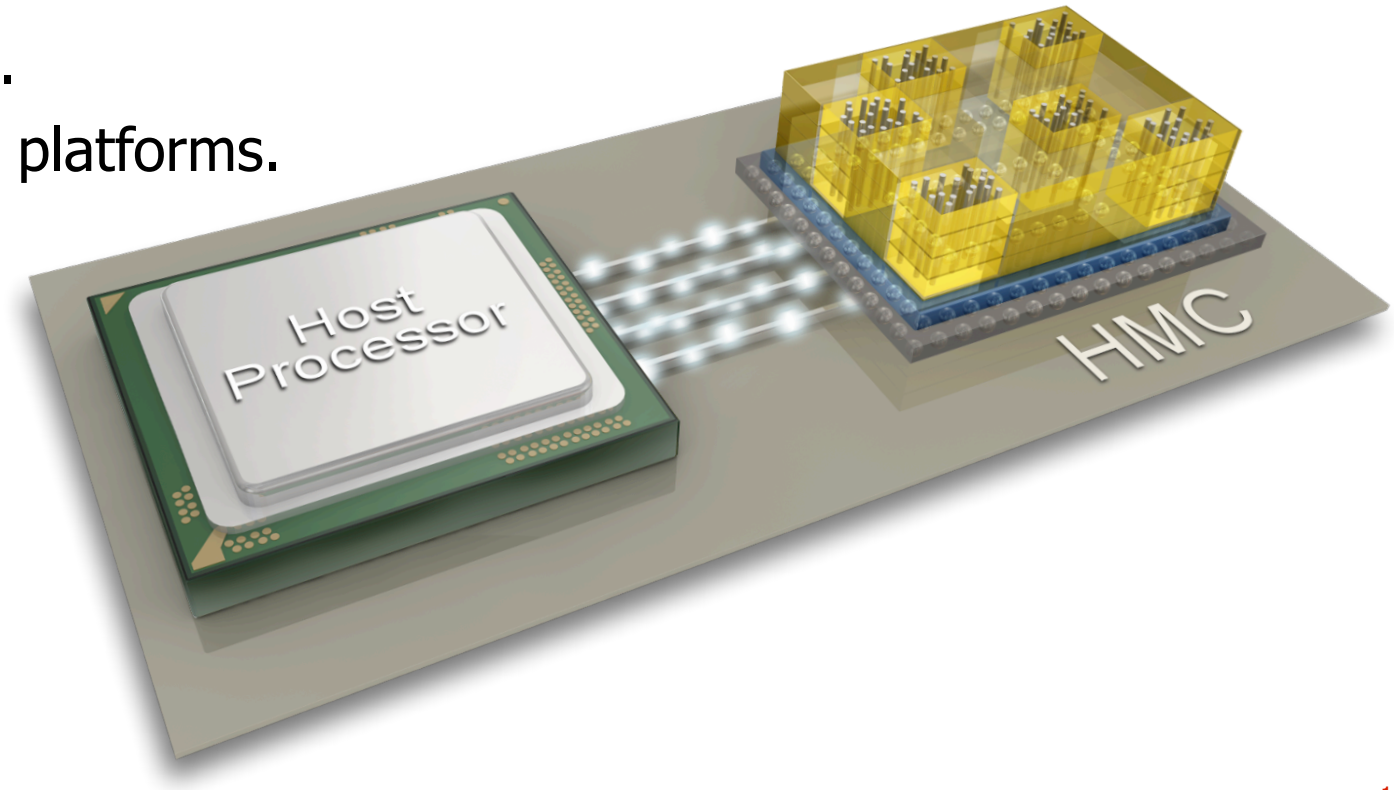


Far memory:



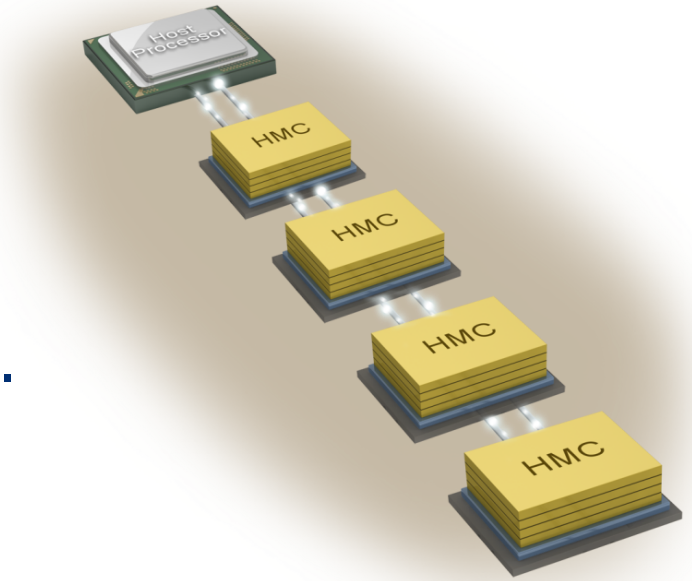
HMC near memory

- All links between CPU and HMC logic layer.
- Maximum bandwidth per GB. capacity.
- Target systems:
  - HPC and servers.
  - Hybrid CPU/GPU platforms.
  - Graphics.
  - Networking.
  - Test equipment.

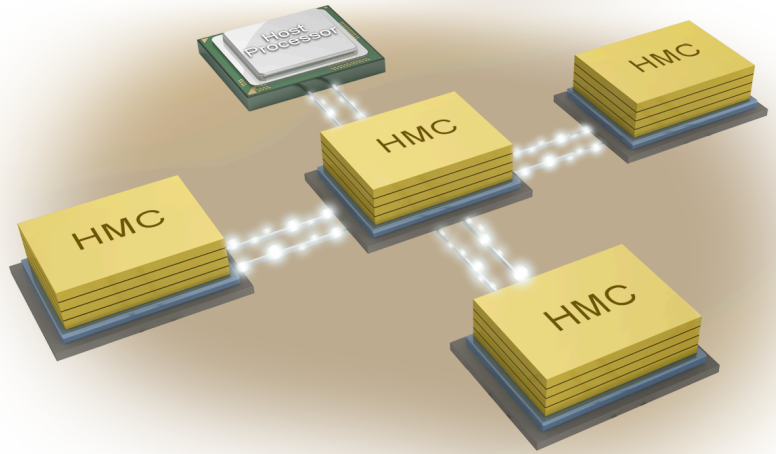


HMC far memory

- Far memory
  - Some HMC links connect to host, some to other cubes.
  - Scalable to meet system requirements.
  - Can be in module form or soldered-down.



- Future interfaces may include
  - Higher speed electrical (SERDES)
  - Optical
  - Whatever the best interface for the job!



A comparison in bandwidth with existing technologies

- On a CPU system (PC with a dual channel motherboard):
  - [2013] DDR3 @ 4 GHz (2x 2000 MHz): 64 Gbytes/s.
  - [2014] HMC 1.0 (first generation): 640 Gbytes/s.
  - [2015] HMC 2.0 (second generation): **898 Gbytes/s**.
  - A 2x improvement can be reached in a quad-channel motherboard.

- On a GPU system (384-bits wide graphics card):
  - GDDR5 @ 7 GHz: 336 Gbytes/s.
  - 12 chips 32-bits wide are soldered to the printed circuit board, where HMC 2.0 chips achieve **2688 Gbytes/s** (2.62 Tbytes/s).

Additional information available on the Web

- The Hybrid Memory Cube Consortium:
  - http://www.hybridmemorycube.org (specification 1.0 available as PDF).
- CUDA Education (presentations, exercises, tools, utilities):
  - http://developer.nvidia.com/cuda-education
- Keynotes and technical sessions from GTC'13:
  - http://www.gputechconf.com/gtcnew/on-demand-gtc.php
  - You will find more than 300 talks. Particularly recommended:
    - "Future directions for CUDA" by Mark Harris.
    - "Multi-GPU Programming" by Levi Barnes.
    - "Performance Optimization Programming Guidelines..." by Paulius Micikevicius.
    - "Performance Optimization Strategies for GPU-accel. Applications" by David Goodwin.
    - "Languages, Libraries and Development Tools for GPU Computing" by Will Ramey.
    - "Getting Started with OpenACC" by Jeff Larkin.
    - "Optimizing OpenACC Codes" by Peter Messmer.

Acknowledgements

- To the great Nvidia people, for sharing with me ideas, material, figures, presentations, ... In alphabetical order:
  - Bill Dally [2010-2011: Power consumption, Echelon and future designs].
  - Simon Green [2007-2009: CUDA pillars].
  - Sumit Gupta [2008-2009: Tesla hardware].
  - Mark Harris [2008, 2012: CUDA, OpenACC, Programming Languages, Libraries].
  - Wen-Mei Hwu [2009: Programming and performance tricks].
  - Stephen Jones [2012: Kepler].
  - David B. Kirk [2008-2009: Nvidia hardware].
  - David Luebke [2007-2008: Nvidia hardware].
  - Lars Nyland [2012: Kepler].
  - Edmondo Orlotti [2012: CUDA 5.0, OpenACC].
- ... just to name a few of those who contributed to my presentations.
- Also thanks to Scott Stevens and Susan Platt from Micron

Thanks for attending!

- You can always reach me in Spain at the Computer Architecture Department of the University of Malaga:
  - e-mail: ujaldon@uma.es
  - Phone: +34 952 13 28 24.
  - Web page: http://manuel.ujaldon.es (english/spanish versions available).