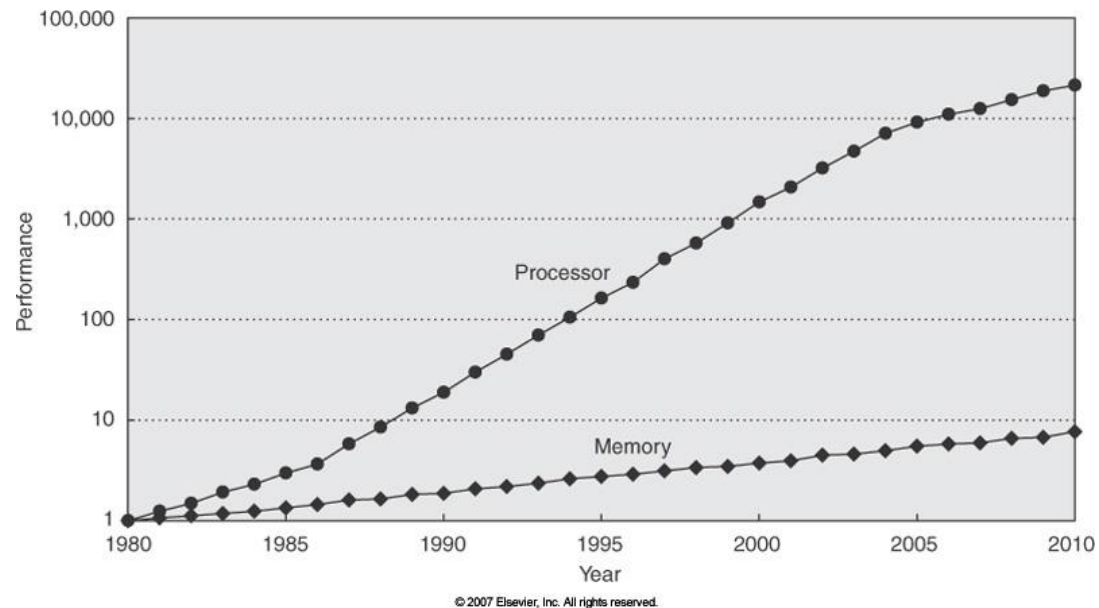


# Hybrid Memory Cube

Saptadeep Pal

# General perspective in 3D memories

## ? Why do we need 3D memory technology



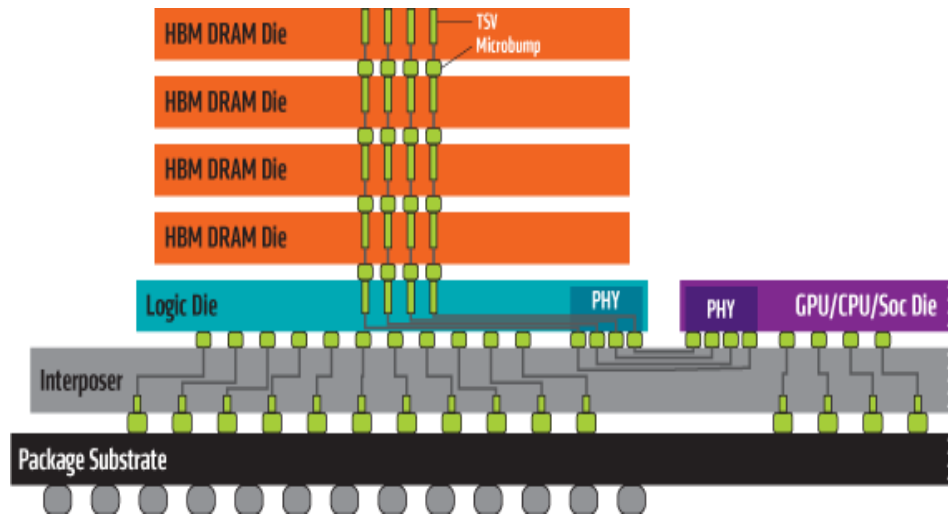
- Big performance gap between processor and memory

### 3D - memory:

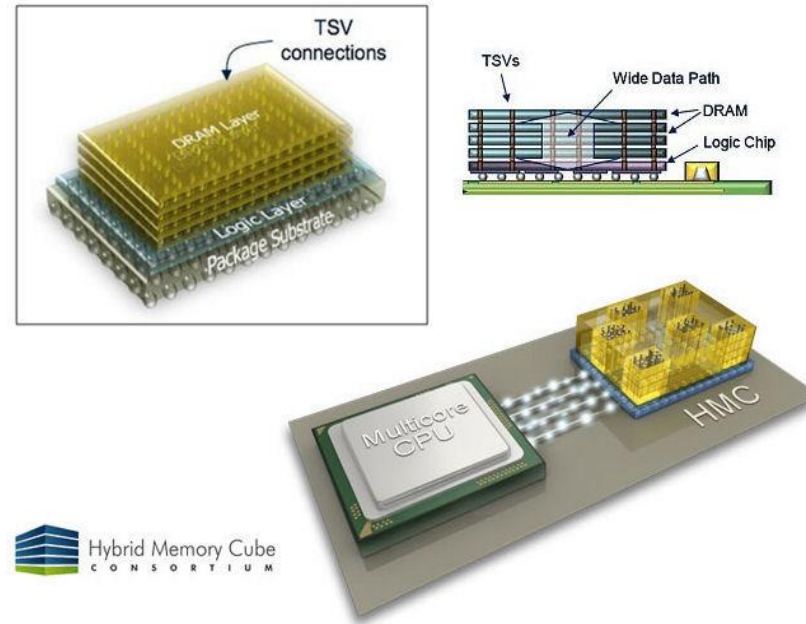
- Multiple layers of die stacked using TSVs
- Shorter memory access latency
- Higher achievable bandwidth

# New Developments in 3D Memory Technology

## I. High Bandwidth Memory (HBM)



## II. Hybrid Memory Cube (HMC)



## III. Wide-IO on Interposers

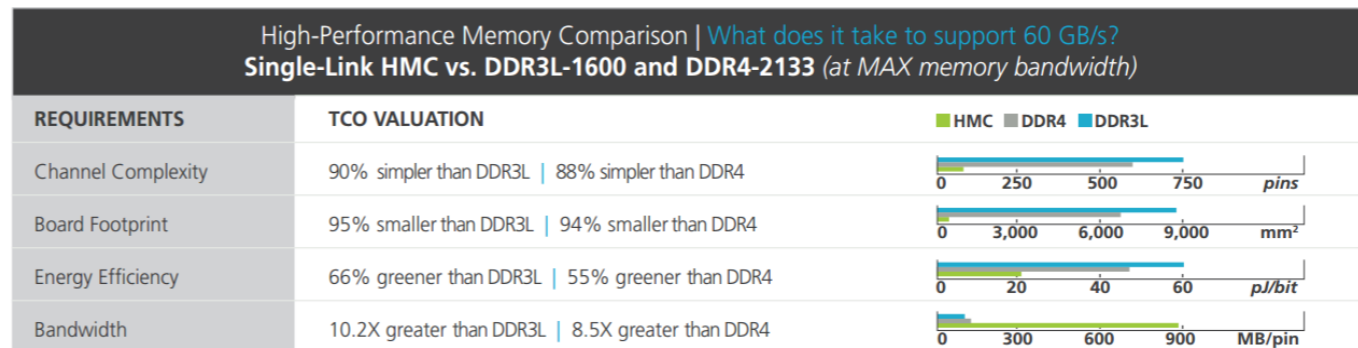
# HMC Advantage

High bandwidth with its scalability, power efficiency, PCB connectivity between host&DRAM, lower latency

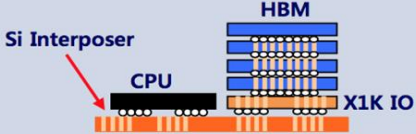
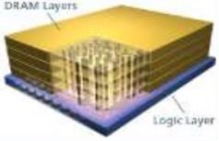
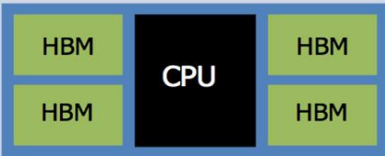
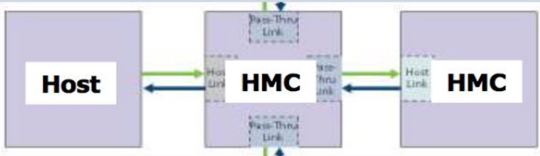
## HMC<sub>Gen1</sub>: Technology Comparison

Generation 1 ( 4 + 1 memory configuration)

Technology	VDD	IDD	BW GB/ s	Power (W)	mW/ GB/ s	pj/ bit	real pJ/ bit
SDRAM PC133 1GB Module	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB Module	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDRII-667 2GB Module	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB Module	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB Module	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC, 4 DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7



# HBM vs. HMC

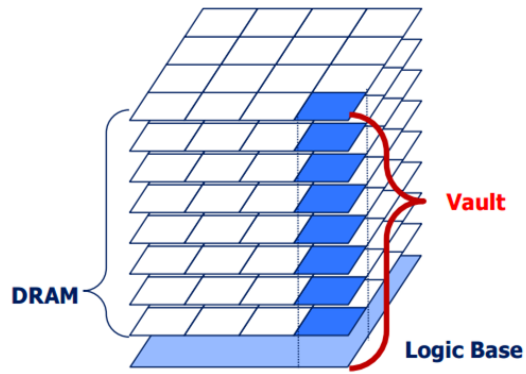
	HBM	HMC
		
<b>PKG type</b>	MPGA(Micro Pillar Grid Array)	BGA
<b>Logic function</b>	Buffer / Rerouting	Memory controller, SERDES
<b>CMD protocol</b>	Deterministic	Non-deterministic
<b>Max. bandwidth</b>	128~256GB/s	4link: ~160GB/s, 8link: ~320GB/s
<b>Power* / Chip size</b>	1X / 1X	1X(USR**) / 1.1X <small>**Ultra Short Reach</small>
<b>Capacity per cube</b>	2/4GB	2/4/8GB
<b># of bank</b>	~128banks (@4GB)	~512banks (@8GB)
<b>Capacity extension</b>		

Target Market:

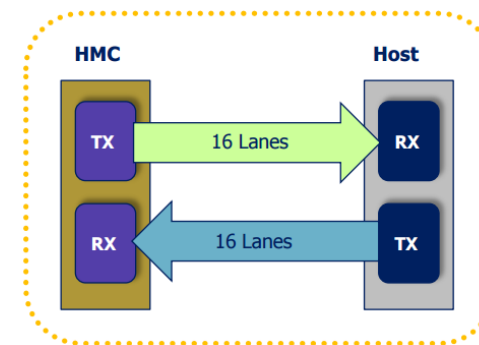
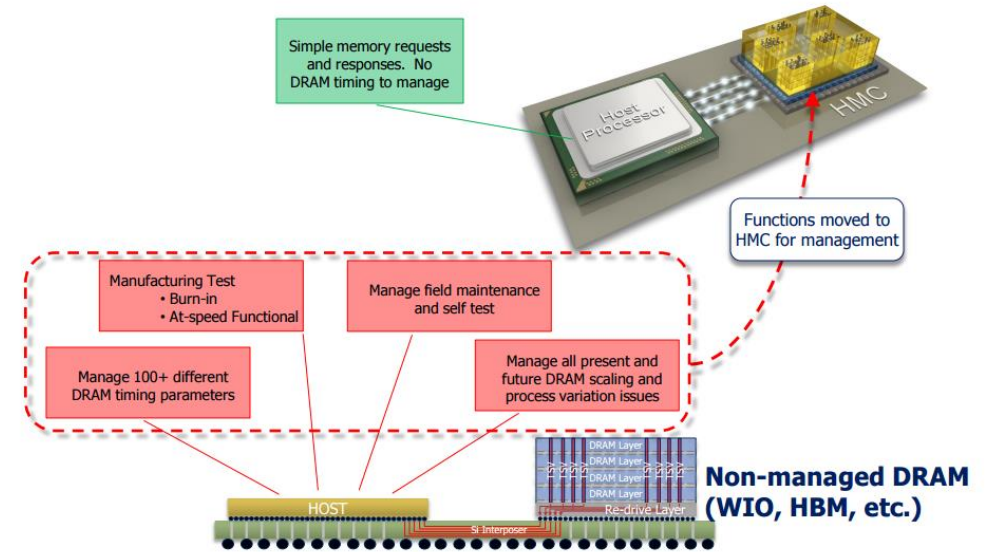
HBM: high-performance graphics accelerators and network devices

HMC: High end servers, high end enterprise

# HMC architecture



- Each vault has a memory controller (called a vault controller) which determines its own timing.
- All in-band communication across a link is packetized.
- Each vault controller determines its own timing requirement
- Refresh operations are controlled by the vault controller, eliminating this function from the host memory controller
- Responses from vault operations back to the external serial I/O links will be out of order. However, requests from a single external serial link to the same vault/bank address are executed in order
- There is no specific timing associated with memory requests. The vaults generally reorder their internal requests to optimize bandwidths and to reduce average latencies

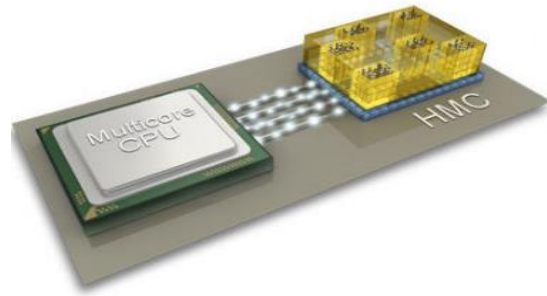


# Logic Base Architecture

The logic base manages multiple functions for the HMC

- All HMC I/O, implemented as multiple serialized, fully duplexed links
- Memory control for each vault; Data routing and buffering between I/O links and vaults
- Consolidated functions removed from the memory die to the controller
- Mode and configuration registers
- BIST for the memory and logic layer
- Test access port compliant to JTAG IEEE 1149.1-2001, 1149.6
- Some spare resources enabling field recovery from some internal hard faults

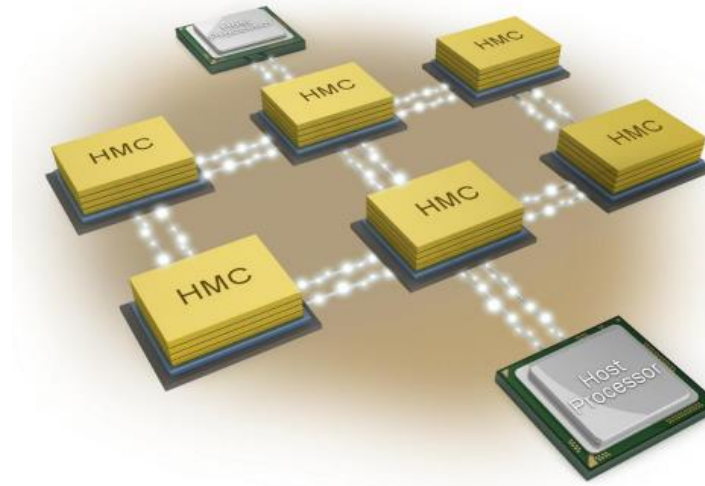
# HMC Near Memory and Far Memory



All links between HMC and Host CPU.

Maximum bandwidth per GB capacity

- HPC/Server – CPU/GPU
- Graphics
- Networking systems
- Test equipment



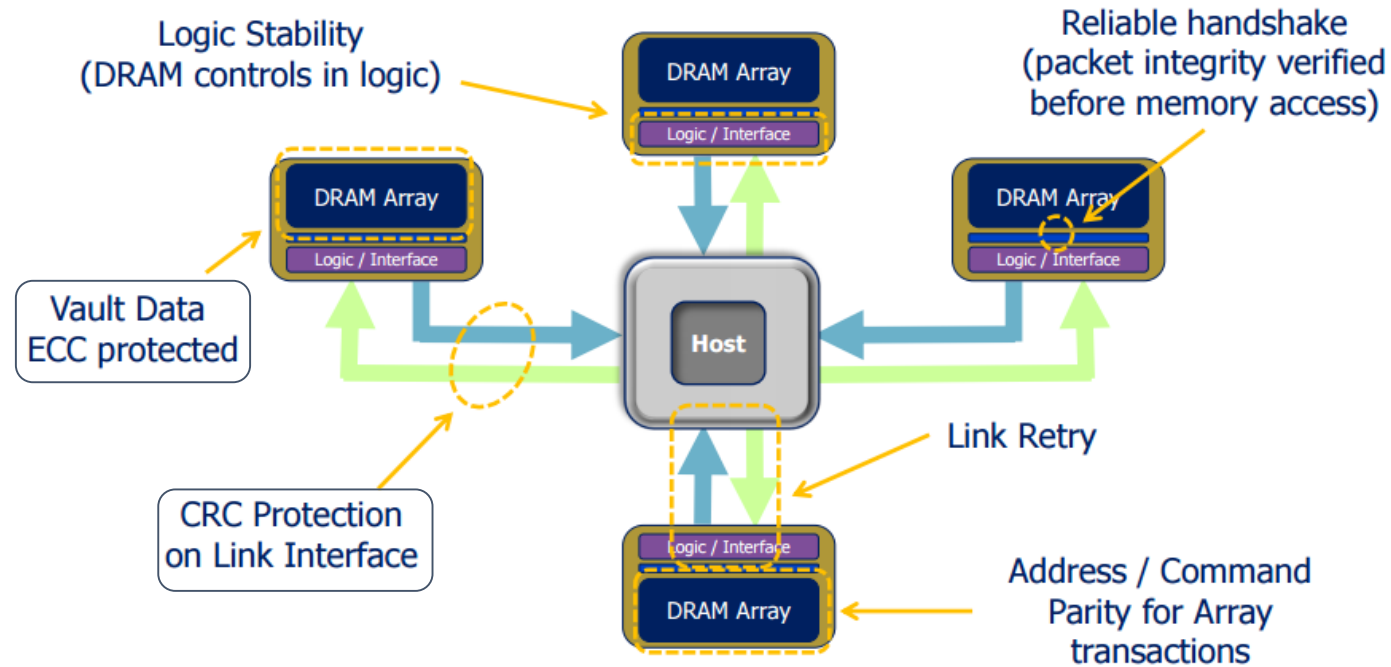
HMC links connect to host or other cubes

- Links form networks of cubes.
- Scalable to meet system requirements



# HMC Reliability

Built-In RAS features at a high level...



# HMC Products in Market now

- Xeon Phi
- Micron's short reach HMC SERDES PHY
- Xilinx Virtex-7 FPGAs support HMC 10