Architecture or Package Design, Which Comes First?

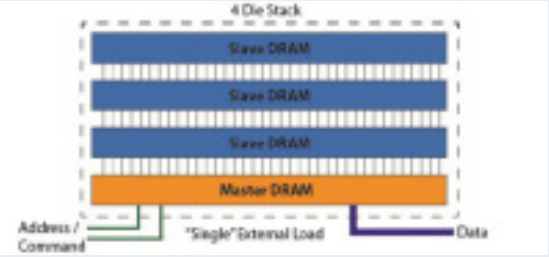
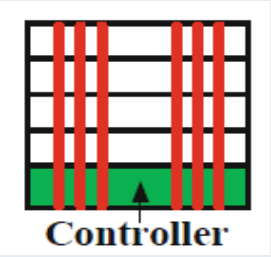
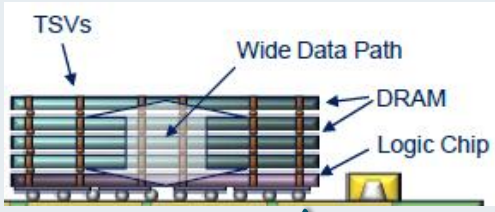
Craig Hampel
Rambus Inc.
Sunnyvale, CA, USA

Rambus

How does the availability of TSVd memory effect system architecture?

- Provide an introduction to applications for DRAM that utilize TSV and 3D technologies
- Discuss the options of TSVs being applied to increase bandwidth
- Discuss the architectural implications of these devices in building high bandwidth memory solution
- Summarize and compare high bandwidth memory systems

Applications of Through Silicon Via for DRAM

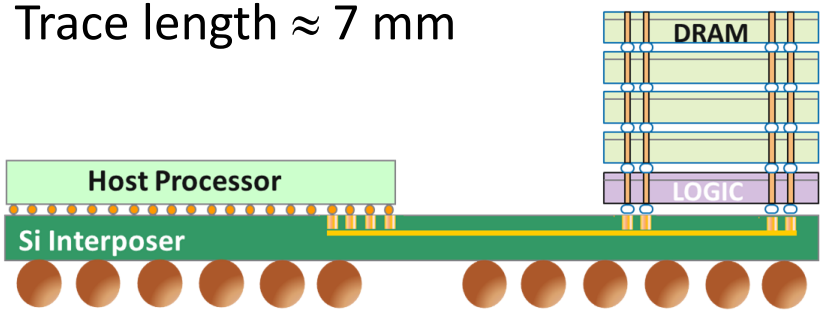
	Capacity	Power	Performance
Example	3DS DRAM	Wide I/O	HBM and HMC
			
Basic Idea	Utilize the 3D connection of TSVs within a DRAM stack to reduce load and Increase capacity and density DRAM to DRAM TSV	Utilize the reduced length and parasitics of TSVs to reduce power DRAM to Controller TSV	Utilize the high density of TSVs to increase bandwidth DRAM to Logic Device TSV
Application	Data center main memory DDDR4 DIMMs	Mobile devices	Graphics Networking HPC/Throughput Processing

This Talk

High Performance TSV Applications

HBM

Trace length ≈ 7 mm

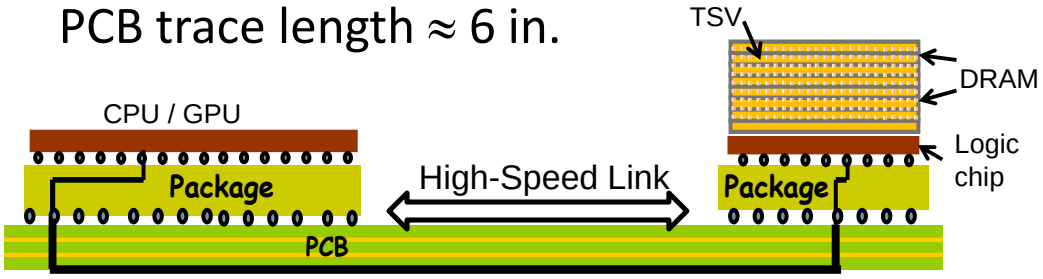


Silicon substrate interface
Wide 1024
Slower 2Gb/s data rate
128-256 GB/s
Bidirectional data

Fairly simple logic layer

HMC

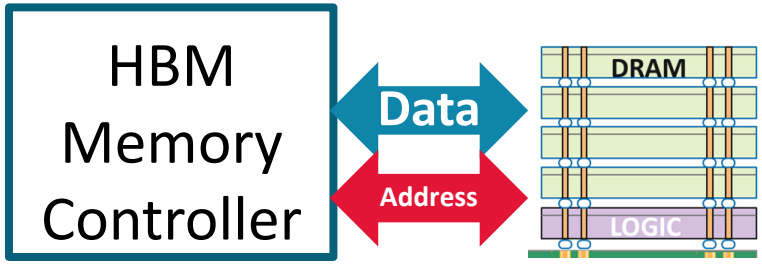
PCB trace length ≈ 6 in.



External PCB interface
32 x 4 or 2 I/Os per HMC
High speed serial interface
10-30Gb/s data rate
160-320 GB/s total
Unidirectional busses

Advanced features in logic layer

Distribution of Memory Functions



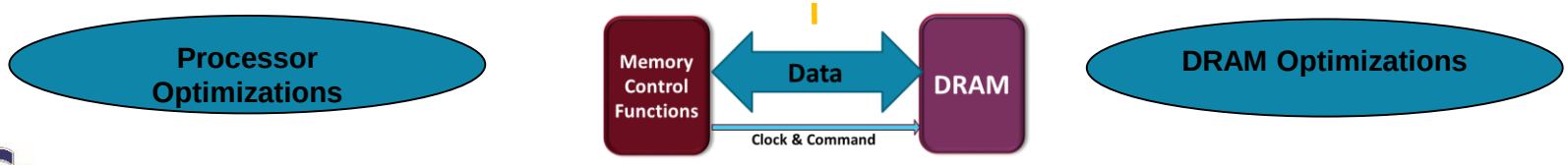
- In shorter channels all memory functions can be performed in controller
- DRAM side logic die is primarily used for manufacturability and shared functions like voltage regulation, BIST etc…

- Longer serial channels encourage packetization and more memory functions in DRAM side logic die

DRAM Management Historically - HBM

- Transaction scheduling
- DRAM even timing
- Transaction ordering
- DRAM state - page policy
- Request priorities
- Data formats – data scrambling
- Error detection/correction
- Bank/row/channel mapping
- In situ repair

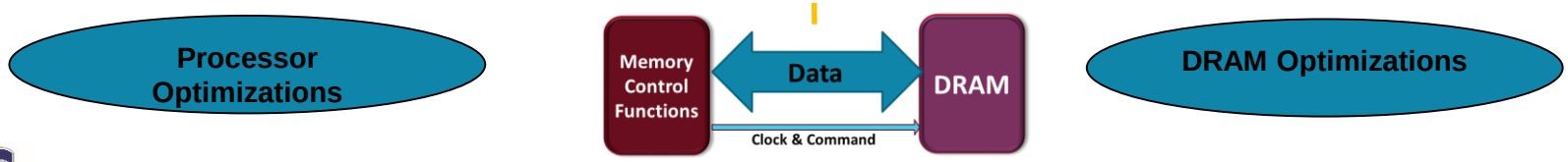
- Manufacturing time repair
- Minimum, “binned” and fully deterministic response times
- Undefined/non coherent data
- Self refresh



DRAM Management - HMC

- Data formats – data scrambling

- Transaction scheduling - nondeterministic
- DRAM event timing
- Transaction ordering
- DRAM state - page policy
- Request priorities
- Error detection/correction
- In situ repair
- Bank/row/channel mapping



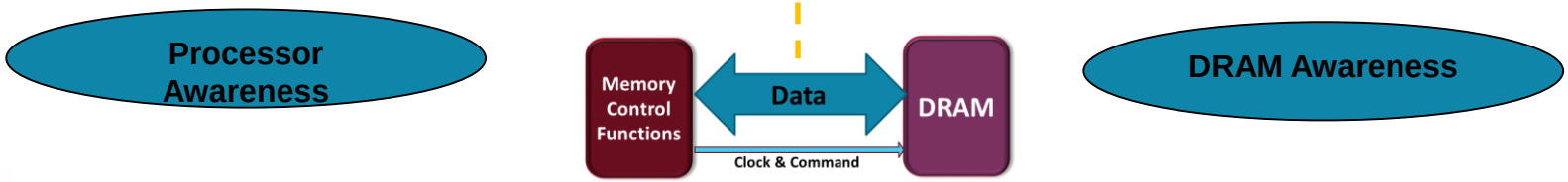
DRAM Functional Optimizations:

Execution and Application Aware

- Data type and priority
- Earliest awareness of address and type of access
- System level QOS requirements
- System level power management

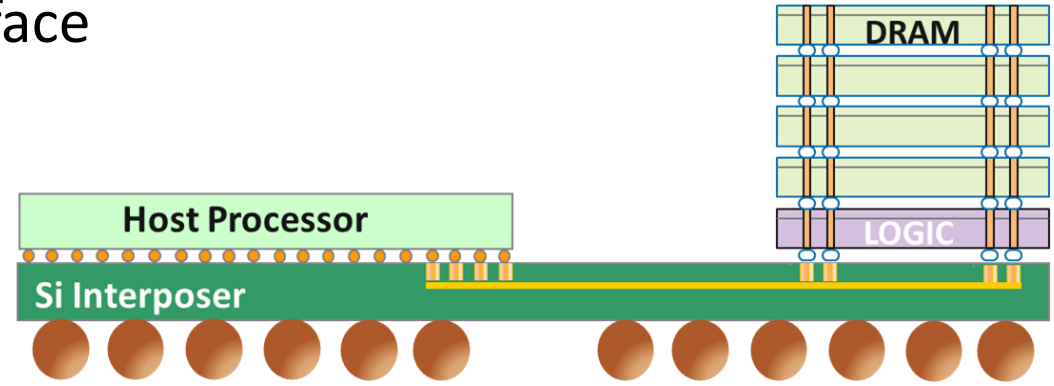
Memory Environment Aware

- Memory characteristics aware
- Earliest access to data and errors
- Memory environment aware
- Memory timing aware - dynamic
- Storage characteristics aware



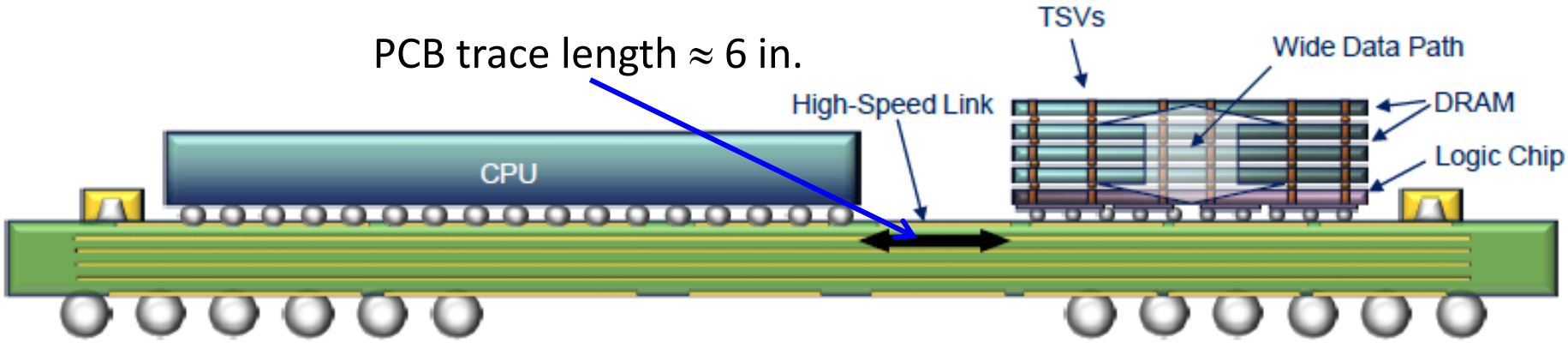
HBM Memory System

- Traditional DRAM interface
- Wide, slow, data bus
  - HBM: 1 Gbps
  - HBM2: 2 Gbps



- Signaling
  - Bidirectional data bus allows dynamic allocation
  - Control functions reside near host processor
- Co-packaged memory
  - Channel is short
  - Silicon substrate required with DRAM near processor
  - Substrate trace length : 7 mm

HMC Memory Systems



- Discrete independent full memory system
- Good for for HPC/Server (CPU) & Network (NPU) Applications
  - Ideal for steady state workloads with equal read-write ratio
  - HMC logic performs many of the functions typically in a CPU
  - 160 – 320 GB/s Memory BW

Comparison of HBM and HMC Memory Systems

Variables	HBM	HBM2	HMC	HMC2
VDD (V)	1.5	1.5	1.35	1.2
Max. Data Rate (Gbps)	1	2	15	30
Bus Width (bits)	1024		4 Links (16 TX/RX lanes per link)	
Max Stack Bandwidth (GB/s)	128	256	120	320
Signaling	Single ended		Differential	
Interface	Wide Parallel		Serial	
Channel overhead	Short		Long	
Format	In a Si Interposer		Stand alone as a complete package	
Control distribution in logic	Simple DRAM like		Advanced Transactional	

3D Memory System Summary

- TSVs provide a number of potential benefits
  - Increased capacity like 3DS
  - Reduced power like wide-I/O
  - Higher bandwidth in systems like HBM and HMC

- Different packaging decisions and allocation of functions in the logic layer make HMC and HBM very different