Highlights of the High-Bandwidth Memory (HBM) Standard

Mike O'Connor

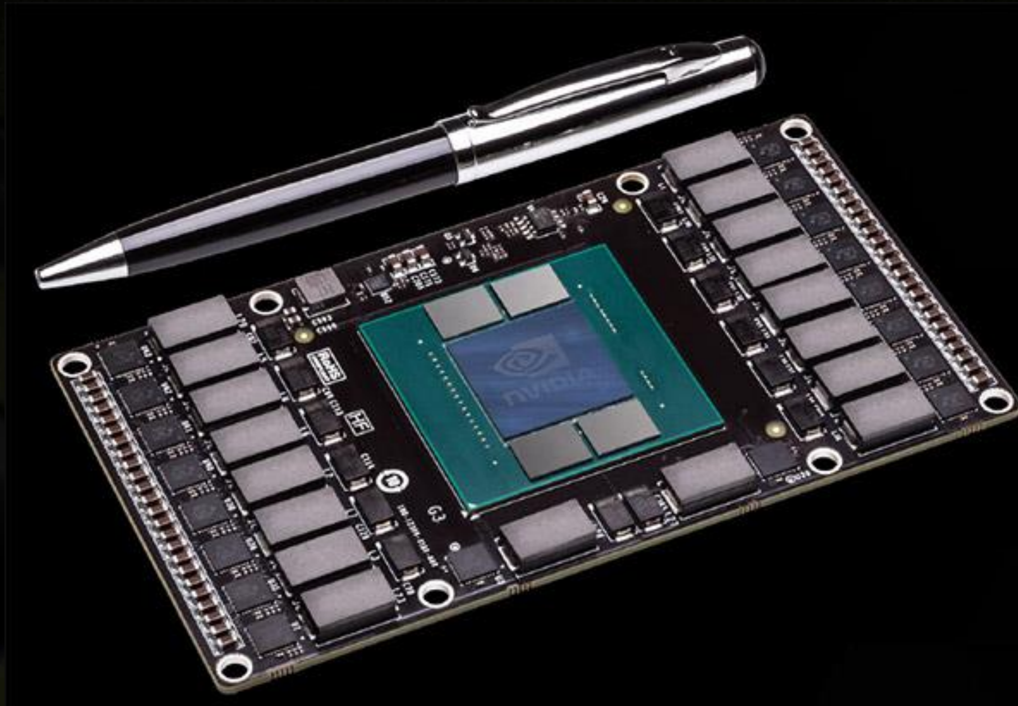
Sr. Research Scientist

What is High-Bandwidth Memory (HBM)?

- Memory standard designed for needs of future GPU and HPC systems:
  - Exploit very large number of signals available with die-stacking technologies for very high memory bandwidth
  - Reduce I/O energy costs
  - Enable higher fraction of peak bandwidth to be exploited by sophisticated memory controllers
  - Enable ECC/Resilience Features

- JEDEC standard JESD235, adopted Oct 2013.
  - Initial work on standard started in 2010

What is High-Bandwidth Memory (HBM)?



- **Enables systems with extremely high bandwidth requirements like future high-performance GPUs**

HBM Overview

- Standard defines an HBM stack
  - Bonding footprint
  - Interface Signaling
  - Commands & Protocol
  - Some optional features:
    - ECC support
    - Base-layer logic/redistribution/IO die
- Standard does *not* define
  - Internal architecture of the stack
  - Precise DRAM timing parameters

HBM Overview

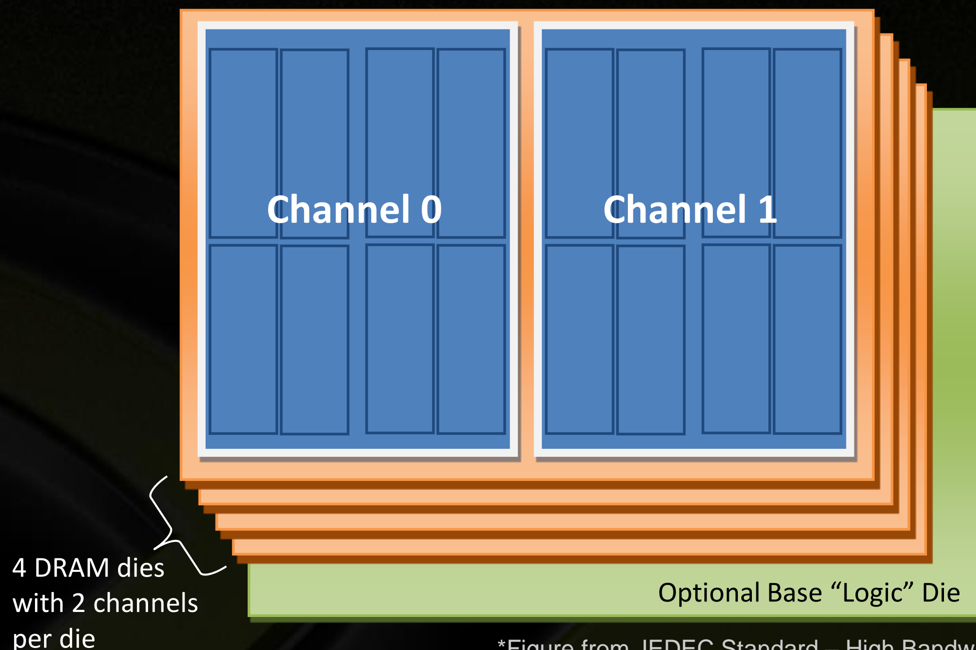


*Figure from JEDEC Standard – High Bandwidth Memory (HBM) DRAM, JESD 235, Oct. 2013

- **Each HBM stack provides 8 independent memory channels**
  - **These are completely independent memory interfaces**
    - **Independent clocks & timing**
    - **Independent commands**
    - **Independent memory arrays**
    - **In short, nothing one channel does affects another channel**

HBM Overview - Bandwidth

- Each channel provides a 128-bit data interface
  - Data rate of 1 to 2 Gbps per signal (500-1000 MHz DDR)
  - 16-32 GB/sec of bandwidth per channel
- 8 Channels per stack
  - 128-256 GB/sec of bandwidth per stack
- For comparison:
  - Highest-end GPU today (NVIDIA GeForce GTX TITAN Black)
    - 384b wide GDDR5 (12 x32 devices) @ 7 Gbps = 336 GB/s
  - Future possible GPU with 4 stacks of HBM
    - Four stacks of HBM @ 1-2 Gbps = 512 GB/s - 1 TB/s
      – cost

HBM Overview - Bandwidth

- Each channel provides … 128-bit … interface
  - Data r… of 1 to 2 Gb… 1000 M… DDR)
  - 16-32 G…
  - … Black)
  - …evices) … 7 Gbps = 336 GB/s
  - Future … ssible GPU with 4 stacks of HBM
    - Four stacks of HBM @ 1-2 Gbps = 512 GB/s - 1 TB/s
      – power cost

At lower overall DRAM system power.

~6-7 pJ/bit vs.

~18-22 pJ/bit for GDDR5

HBM Overview - Capacity

- Per-channel capacities supported from 1-32 Gbit
  - Stack capacity of 1 to 32GBytes
  - Near-term, at lower-end of range
    e.g. 4 high stack of 4Gb dies = 2GBytes/stack
- 8 or 16 banks per channel
  - 16 banks when > 4Gbit per channel (> 4GBytes/stack)
- Not including optional additional ECC bits
  - A stack providing ECC storage may have 12.5% more bits

HBM Channel Overview

- Each channel is similar to a standard DDR interface
  - Data interface is bi-directional
    - Still requires delay to “turn the bus around” between RD and WR
    - Burst-length of 2 (32B per access)
  - Requires traditional command sequences
    - Activates required to open rows before read/write
    - Precharges required before another activate
    - Traditional dram timings still exist (tRC, tRRD, tRP, tFAW, etc.) – but are entirely per-channel

HBM Channel Summary

Function	# of µBumps	Notes
Data	128	DDR, bi-directional
Column Command/Addr.	8	DDR
Row Command/Addr.	6	DDR
Data Bus Inversion	16	1 for every 8 Data bits, bi-directional
Data Mask/Check Bits	16	1 for every 8 Data bits, bi-directional
Strobes	16	Differential RD & WR strobes for every 32 Data bits
Clock	2	Differential Clock
Clock Enable	1	Enable low-power mode
Total	**193**	

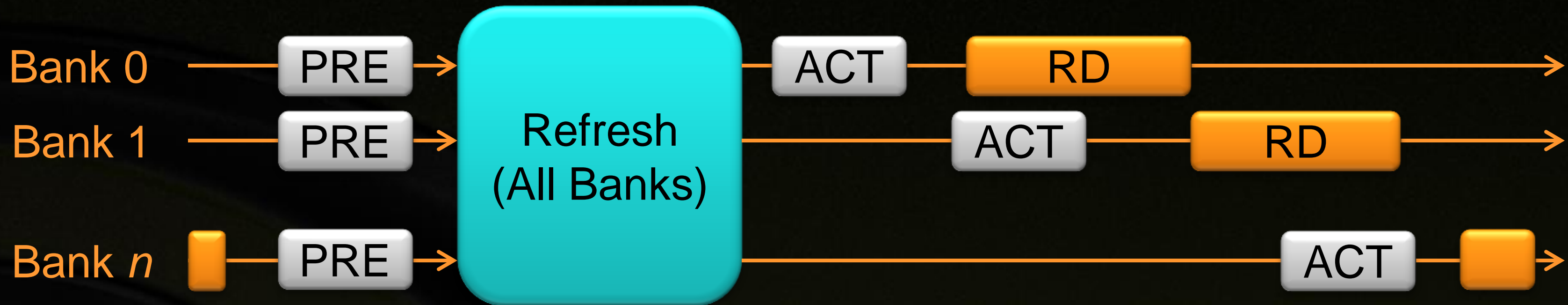
New: Split Command Interfaces

- 2 semi-independent command interfaces per channel
  - "Column Commands" – Read / Write
  - "Row Commands" – ACT / PRE / etc.
- Key reasons to provide separate row command i/f:
  - 100% col. cmd bandwidth to saturate the data bus w/ BL=2
  - Simplifies memory controller
  - Better performance (issue ACT earlier or not delay RD/WR)
- Still need to enforce usual ACT→RD/WR→PRE timings

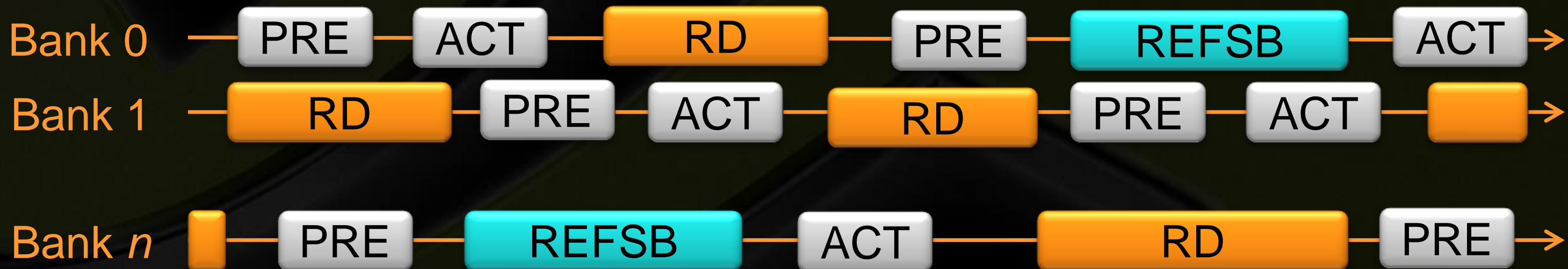
New: Single-Bank Refresh

- Current DRAMs require refresh operations
  - Refresh commands require all banks to be closed
  - ~ 1 refresh command every few µsec
  - Can consume 5-10% of potential bandwidth
  - Increasing overheads with larger devices
- Sophisticated DRAM controllers work hard to overlap ACT/PRE in one bank with traffic to other banks
  - Can manage the refresh similarly
  - Added "Refresh Single Bank" command
    - Like an ACT, but w/ internal per-bank row counter
    - Can be issued to any banks in any order
    - Memory controller responsible for ensuring all banks get enough refreshes each refresh period

New: Single-Bank Refresh



Traditional Precharge-All and Refresh-All



Arbitrary Single-Bank Refresh

New: RAS Support

- HBM standard supports ECC
  - Optional: Not all stacks required to support it
- ECC and non-ECC stacks use same interface
  - Key insight:
    Per-byte data mask signals and ECC not simultaneously useful
  - Data Mask Signals can carry ECC data
    - makes them bi-directional on HBM stacks that support ECC
- Parity check of all cmd/addr busses also supported

Other HBM Features

- HBM supports Temperature Compensated Self Refresh
  - Temperature dependent refresh rates with several temperature ranges (e.g. cool/standby, normal, extended, emergency)
  - Temperature sensor can be read by memory controller to adjust its refresh rates as well
- Data Bus Inversion coding to reduce number of simultaneously switching signals
  - No more than 4 of 9 (DQ[0..7], DBI) signals switch
  - DBI computation maintained across consecutive commands

Thank You

QUESTIONS?

moconnor@nvidia.com

BACKUP

Footprint

BACKUP

HBM Footprint

TEST PORT {DIRECT ACCESS}	Power Supply Region - Channels [f:e, b:a]	DWORD0_Channele	DWORD0_Channela
		DWORD0_Channelf	DWORD0_Channelb
		DWORD1_Channele	DWORD1_Channela
		DWORD1_Channelf	DWORD1_Channelb
		AWORD_Channele	AWORD_Channela
		AWORD_Channelf	AWORD_Channelb
		DWORD2_Channele	DWORD2_Channela
		DWORD2_Channelf	DWORD_Channelb
		DWORD3_Channele	DWORD3_Channela
		WDORD3_Channelf	DWORD3_Channelb
Depopulated Micropillar "NO BUMP" Area for optional probing		MIDSTACK	
TEST PORT {DIRECT ACCESS}	Power Supply Region - Channels [h:g, d:c]	DWORD0_Channelg	DWORD0_Channelc
		DWORD0_Channelh	DWORD0_Channeld
		DWORD1_Channelg	DWORD1_Channelc
		DWORD1_Channelh	DWORD1_Channeld
		AWORD_Channelg	AWORD_Channelc
		AWORD_Channelh	AWORD_Channeld
		DWORD2_Channelg	DWORD2_Channelc
		DWORD2_Channelh	DWORD_Channeld
		DWORD3_Channelg	DWORD3_Channelc
		WDORD3_Channelh	DWORD3_Channeld

*Figure from JEDEC Standard – High Bandwidth Memory (HBM) DRAM, JESD 235, Oct. 2013

HBM Footprint

148		D		D		D		D		D		D		D		D		D		D		D		D
149	D		D		D		D		D		D		D		D		D		D		D		D	
150		DQ h39		DQ h37		DR FU h2		DQ h35		DQ h33		DM h4		DQ d39		DQ d37		DR FU d2		DQ d35		DQ d33		DM d4
151	DBI h4		DQ h38		DQ h36		par h1		DQ h34		DQ h32		DBI d4		DQ d38		DQ d36		par d1		DQ d34		DQ d32	
152		DQ h47		DQ h45		wdq sh1 _c		DQ h43		DQ h41		DM h5		DQ d47		DQ d45		wdq sd1 _c		DQ d43		DQ d41		DM d5
153	DBI h5		DQ h46		DQ h44		wdq sh1 _t		DQ h42		DQ h40		DBI d5		DQ d46		DQ d44		wdq sd1 _t		DQ d42		DQ d40	
154		M		M		M		M		M		M		M		M		M		M		M		M
155	M		M		M		M		M		M		M		M		M		M		M		M	
156		DQ h55		DQ h53		rdqs h1_ c		DQ h51		DQ h49		DM h6		DQ d55		DQ d53		rdqs d1_ c		DQ d51		DQ d49		DM d6
157	DBI h6		DQ h54		DQ h52		rdqs h1_t		DQ h50		DQ h48		DBI d6		DQ d54		DQ d52		rdqs d1_t		DQ d50		DQ d48	
158		DQ h63		DQ h61		DR FU h3		DQ h59		DQ h57		DM h7		DQ d63		DQ d61		DR FU d3		DQ d59		DQ d57		DM d7
159	DBI h7		DQ h62		DQ h60		DE RR h1		DQ h58		DQ h56		DBI d7		DQ d62		DQ d60		DE RR d1		DQ d58		DQ d56	
160		D		D		D		D		D		D		D		D		D		D		D		D
161	D		D		D		D		D		D		D		D		D		D		D		D	
162		Cg7		Cg5		CK Eg		Cg3		Cg1		AR FU g0		Cc7		Cc5		CK Ec		Cc3		Cc1		AR FU c0
163	AR FU g2		Cg6		Cg4		AR FU g1		Cg2		Cg0		AR FU c2		Cc6		Cc4		AR FU c1		Cc2		Cc0	
164		AR FU g4		Rg5		CK g_c		Rg3		Rg1		AR FU g3		AR FU c4		Rc5		CK c_c		Rc3		Rc1		AR FU c3
165	AE RR g		AR FU g5		Rg4		CK g_t		Rg2		Rg0		AE RRc		AR FU c5		Rc4		CK c_t		Rc2		Rc0	
166		M		M		M		M		M		M		M		M		M		M		M		M
167	M		M		M		M		M		M		M		M		M		M		M		M	
168		Ch7		Ch5		CK Eh		Ch3		Ch1		AR FU h0		Cd7		Cd5		CK Ed		Cd3		Cd1		AR FU d0
169	AR FU h2		Ch6		Ch4		AR FU h1		Ch2		Ch0		AR FU d2		Cd6		Cd4		AR FU d1		Cd2		Cd0	
170		AR FU h4		Rh5		CK h_c		Rh3		Rh1		AR FU h3		AR FU d4		Rd5		CK d_c		Rd3		Rd1		AR FU d3
171	AE RR h		AR FU h5		Rh4		CK h_t		Rh2		Rh0		AE RR d		AR FU d5		Rd4		CK d_t		Rd2		Rd0	
172		D		D		D		D		D		D		D		D		D		D		D		D

Half of One channel Data i/f

Four channels Command i/f

*Figure from JEDEC Standard – High Bandwidth Memory (HBM) DRAM, JESD 235, Oct. 2013

Commands

BACKUP

Column Commands

Command	Clock	C[0:7]
Column NOP	Rising	CNOP / XXXXX
	Falling	XXXXXXX / Parity
Read	Rising	RD / Autoprecharge / Bank
	Falling	Column Address / Parity
Write	Rising	RD / Autoprecharge / Bank
	Falling	Column Address / Parity
Mode Register Set	Rising	MRS / Mode Reg
	Falling	Opcode

Row Commands

Command	Clock	R[0:5]
Row NOP	Rising	RNOP / XXX
	Falling	XXXXX / Parity
Activate	Rising	ACT / Bank
	Falling	Row Address[15:11] / Parity
	Rising	Row Address[10:5]
	Falling	Row Address[4:0] / Parity
Precharge	Rising	PRE / Bank
	Falling	XXXXX / Parity
Precharge All Banks	Rising	PREA / XXX
	Falling	XXXXX / Parity
Refresh (single bank)	Rising	REFSB / Bank
	Falling	XXXXX / Parity
Refresh (all banks)	Rising	REF / XXX
	Falling	XXXXX / Parity

RAS

BACKUP

HBM RAS Challenges

- Stacked Memory has some challenges with respect to RAS requirements
- Traditional DRAM DIMMs get only a subset of bits (e.g. 4) from each burst from a single DRAM device
- HBM gives you all the bits of a burst from a single row of a single bank of a single DRAM device
  - Good for power, but RAS-wise all our eggs are in one basket
    - Including the ECC bits
  - Need techniques to detect failures (e.g. row decode fault)
  - Need techniques to recover from failures (e.g. RAID-like schemes)