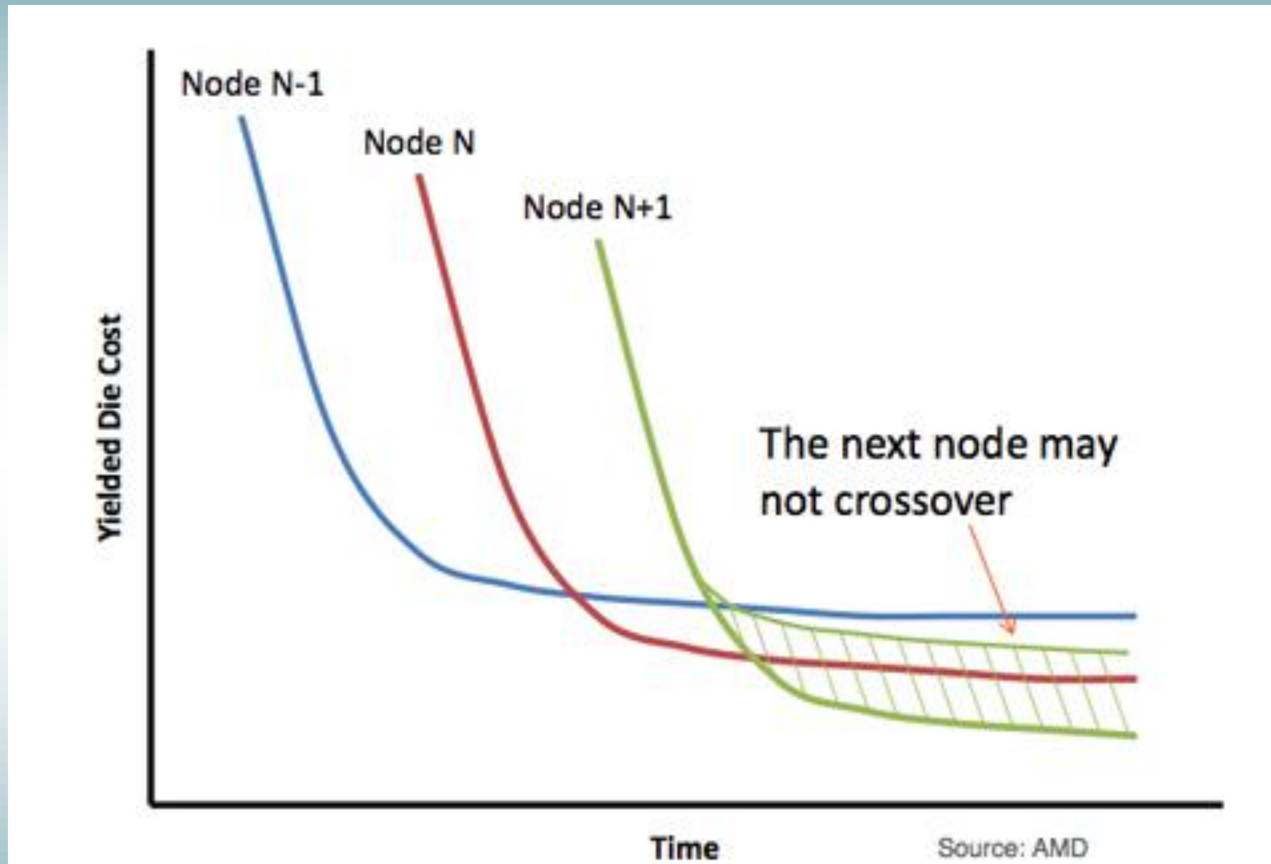
How 3D Memory is Changing Computing

DATE 2015 W05

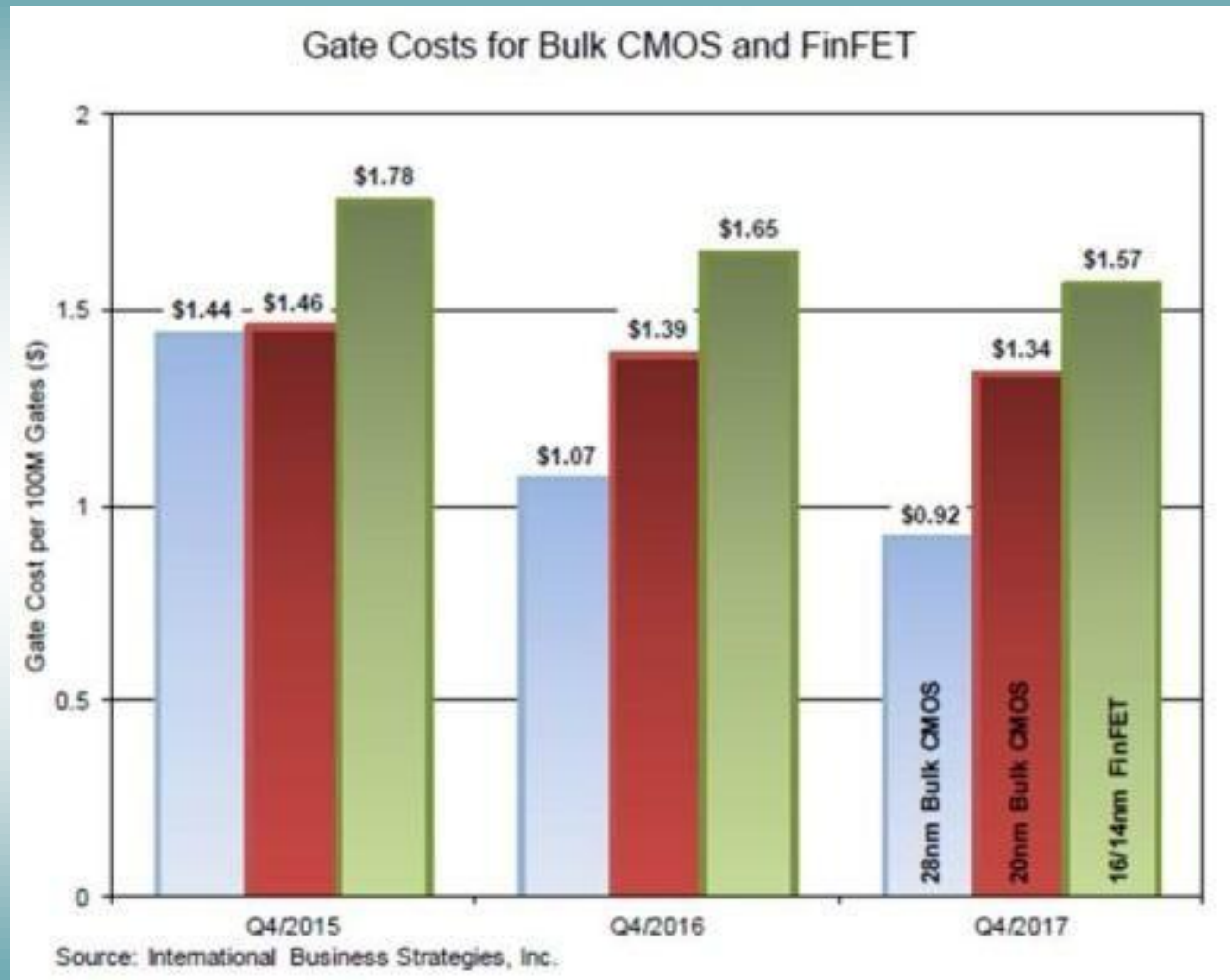
Robert Patti, CTO
rpatti@tezzaron.com

Why 3D? – Expiring Economics



AMD 2014 3D-ASIP

Why 3D? – Expiring Economics



The Ugly Truth

Sparse Matrix Operations
Particle Physics, Weapons Dev.
5.9% efficiency

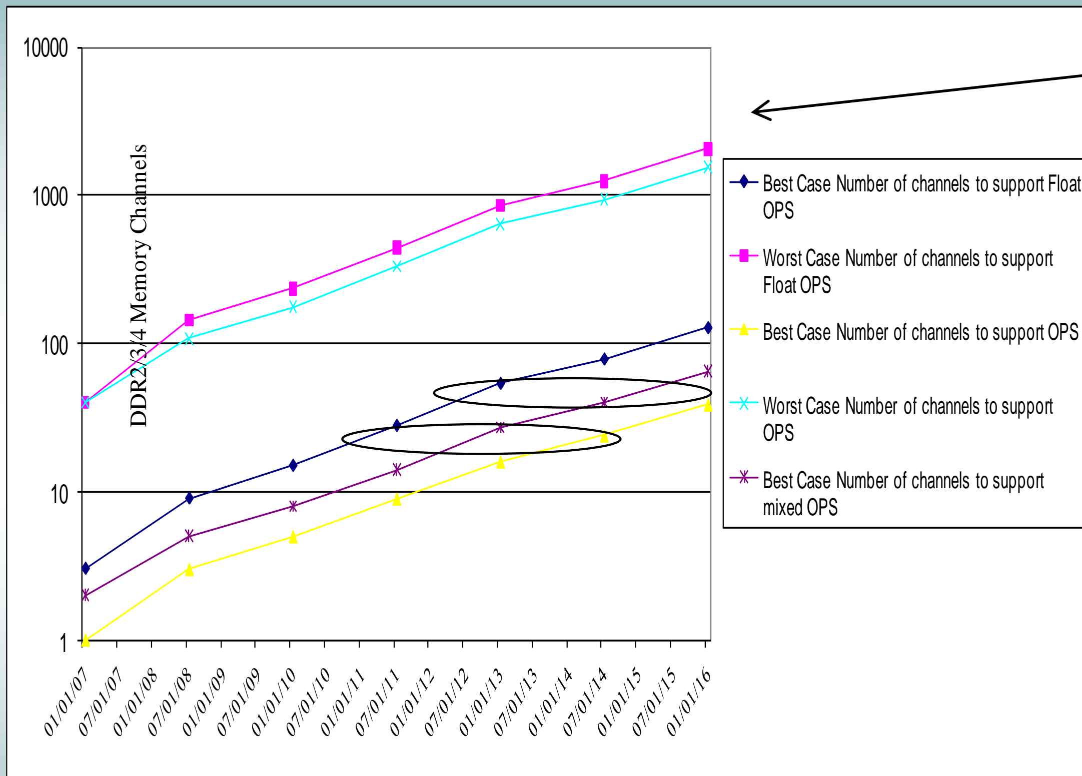
I/O BW to Processing Ratio
Radar, Sonar, Imaging Sensors
12% efficiency

Finite Element Analysis
Weather & Ocean Forecasting
7.1% efficiency

Memory Intensive Calculations
Cryptanalysis
< 3.0% efficiency

Large Matrix Manipulation
Engineering Design of Complex Structures
8.4% efficiency

More Bad News
The Bandwidth Knothole



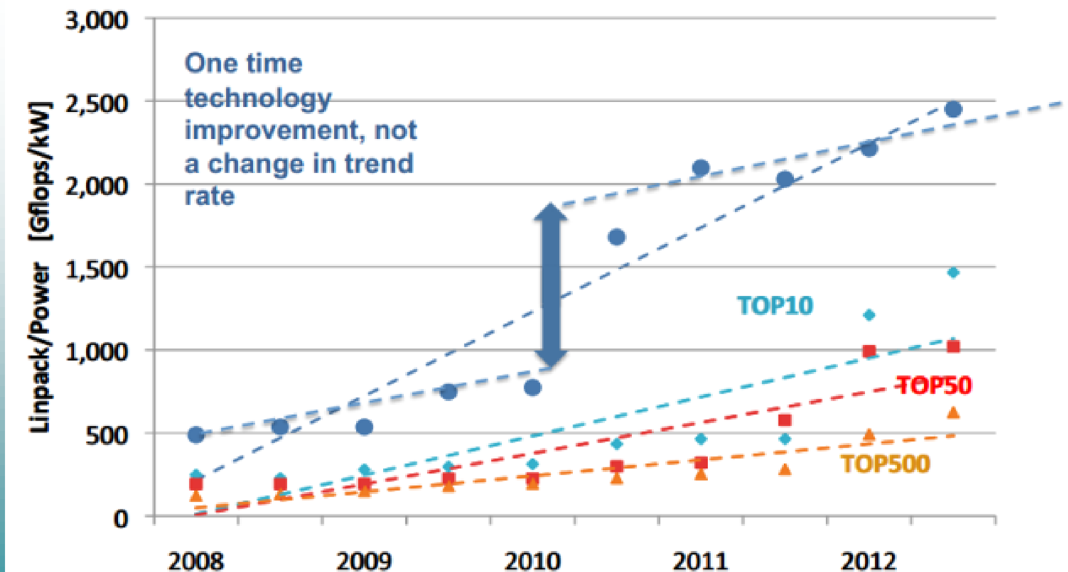
- To continue to increase CPU performance, exponential bandwidth growth required.
- More than 200 CPU cycles of delay to memory results in cycle for cycle CPU stalls.
- 16 to 64 Mbytes per thread required to hide CPU memory system accesses.
- Memory bound HPC performance 1/500th of CPU bound performance
- Memory I/O power is running away.

Need 50x bandwidth improvement.

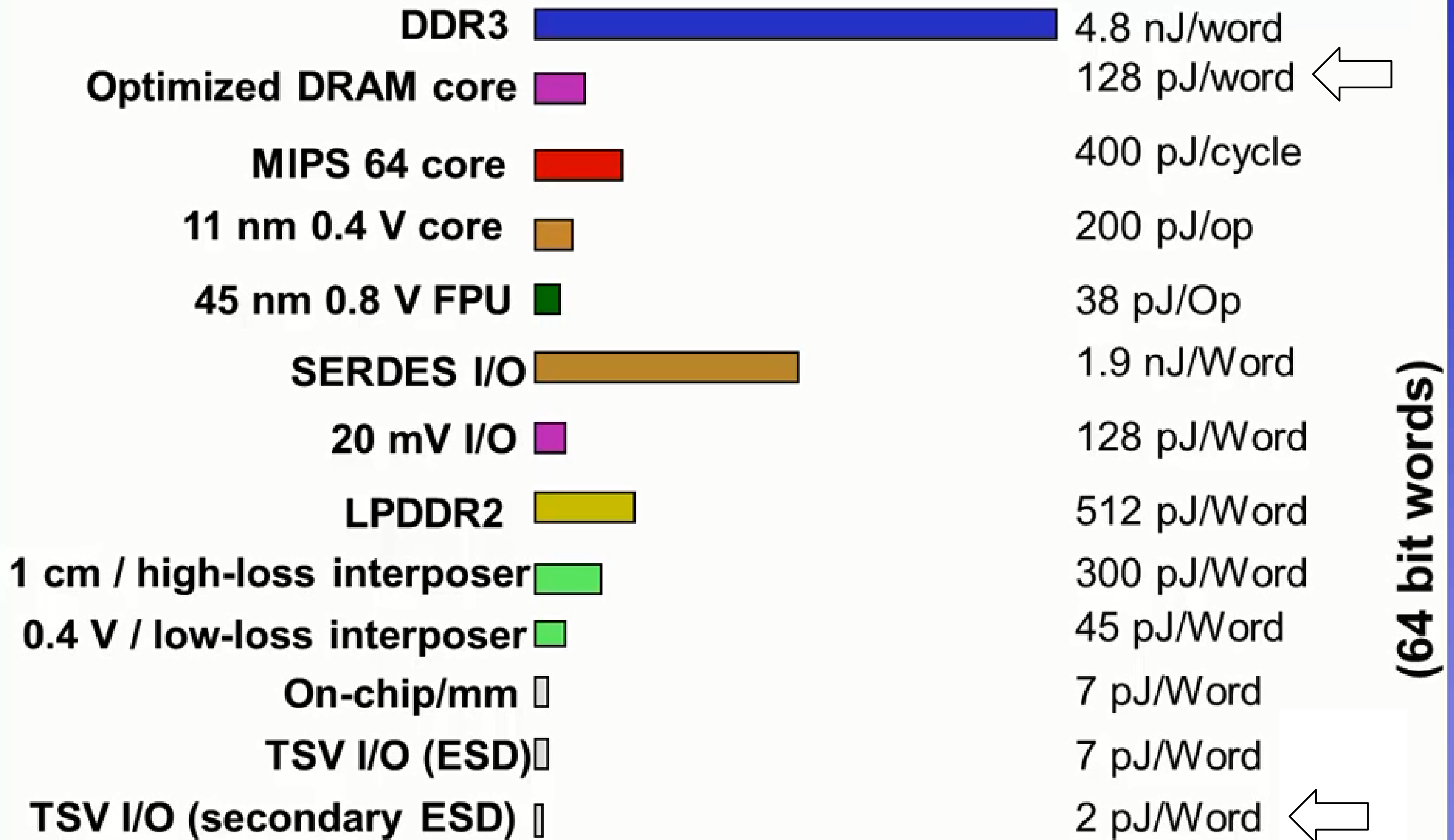
Need 10x better cost model than embedded memory.

HPC Power Limit

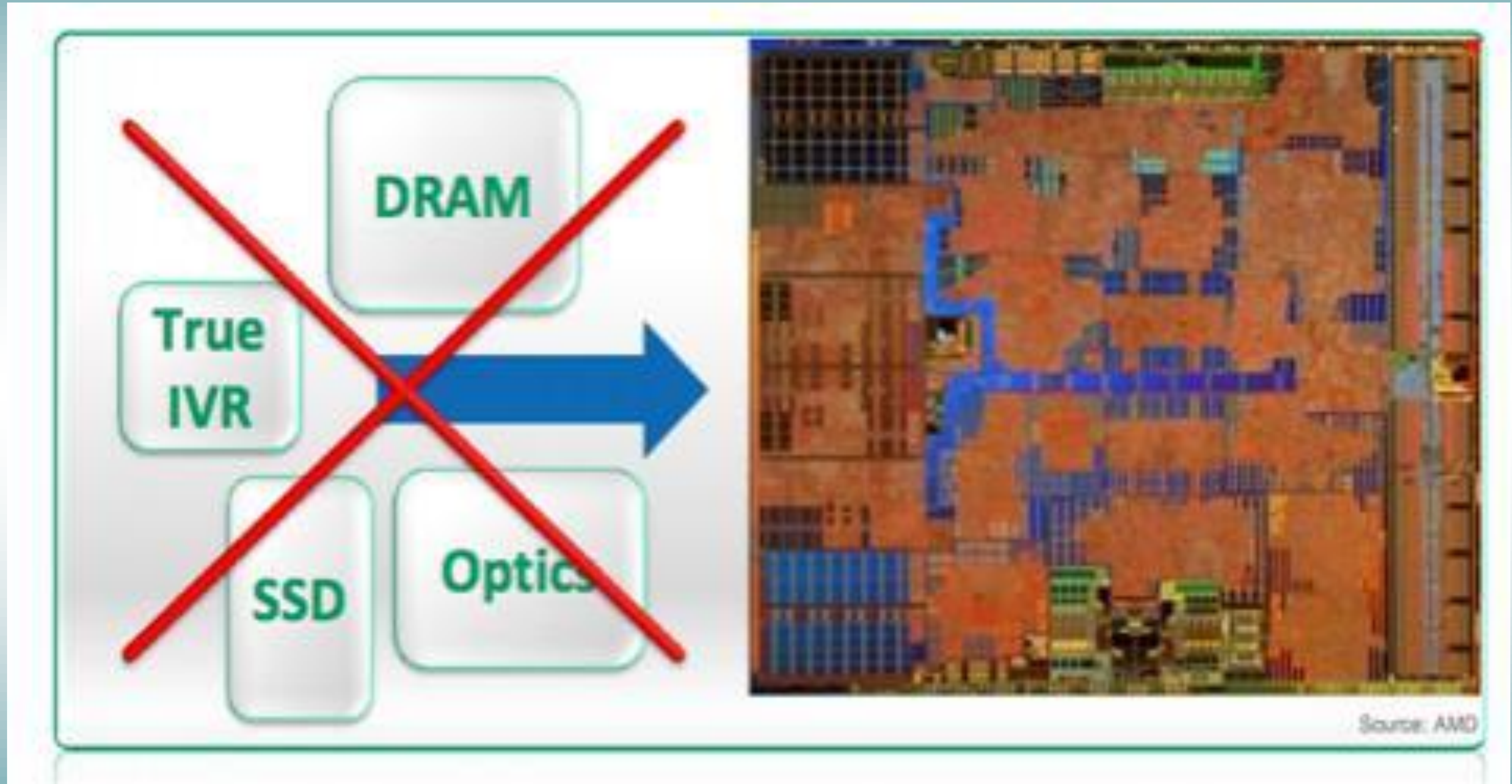
- Exascale
  - 1e18 FLOPs
  - ~10MW logic power budget
  - 10pJ/FLOP power goal --- 20pJ/FLOP needed



Still need 40x power improvement


Energy per Operation
DDR3
4.8 nJ/word
Optimized DRAM core
128 pJ/word
MIPS 64 core
400 pJ/cycle
11 nm 0.4 V core
200 pJ/op
45 nm 0.8 V FPU
38 pJ/Op
SERDES I/O
1.9 nJ/Word
20 mV I/O
128 pJ/Word
LPDDR2
512 pJ/Word
1 cm / high-loss interposer
300 pJ/Word
0.4 V / low-loss interposer
45 pJ/Word
On-chip/mm
7 pJ/Word
TSV I/O (ESD)
7 pJ/Word
TSV I/O (secondary ESD)
2 pJ/Word
(64 bit words)
Various Sources
Paul Franzon NCSU

On-Chip Memory, Flash, etc… Unfriendly Integration



AMD 2014 3D-ASIP

HPC Future Scaling

- Need to fix efficiency problem
  - Change the existing target of EPC; limits scope of usage
    - New architectures using close memory
      - 3D heterogeneous integration
  - Paradigm shift in machine nodal interconnect
    - Network fabric becomes memory (like) fabric
    - NUMA
- Need to fix bandwidth problem
  - 3D provides lots of bandwidth
- Need to fix power problem
  - 3D gives shorter wires, less long I/O, allows BoC processes

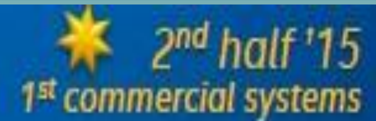
INDUSTRY DIRECTION

Intel's HPC – Heavily Dependent on New Memories

3+ TFLOPS[1]
In One Package
Parallel Performance & Density

New for Knights Landing

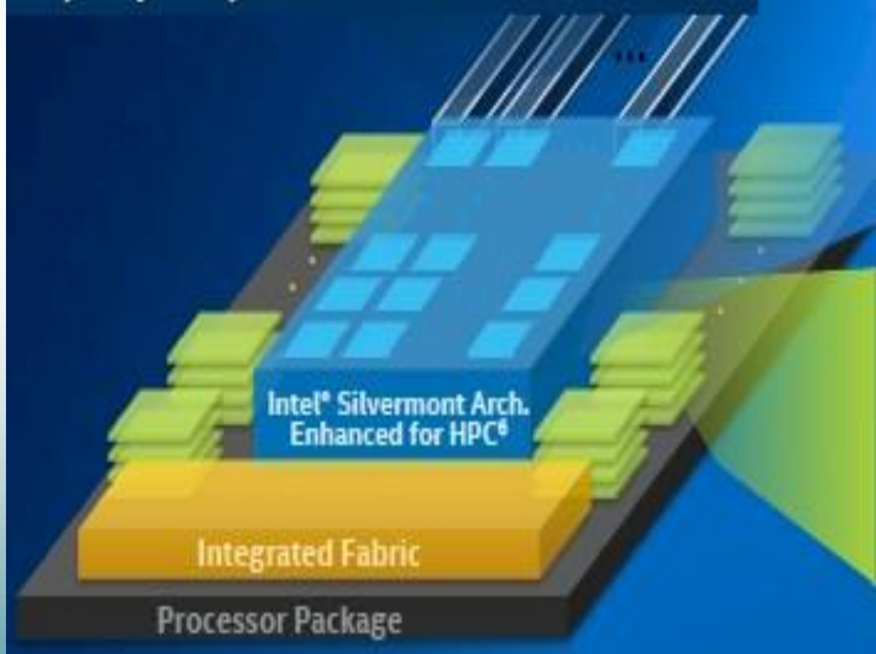
(Next Generation Intel® Xeon Phi™ Products)



Platform Memory: DDR4 Bandwidth and Capacity Comparable to Intel® Xeon® Processors

Compute: Intel® Silvermont Arch. (Intel® Atom™)[2]

- *Low-Power Cores with HPC Enhancements*[3]
- *3X Single Thread Performance*[4] vs Prior Gen.
- *Intel Xeon Processor Binary Compatible*[5]



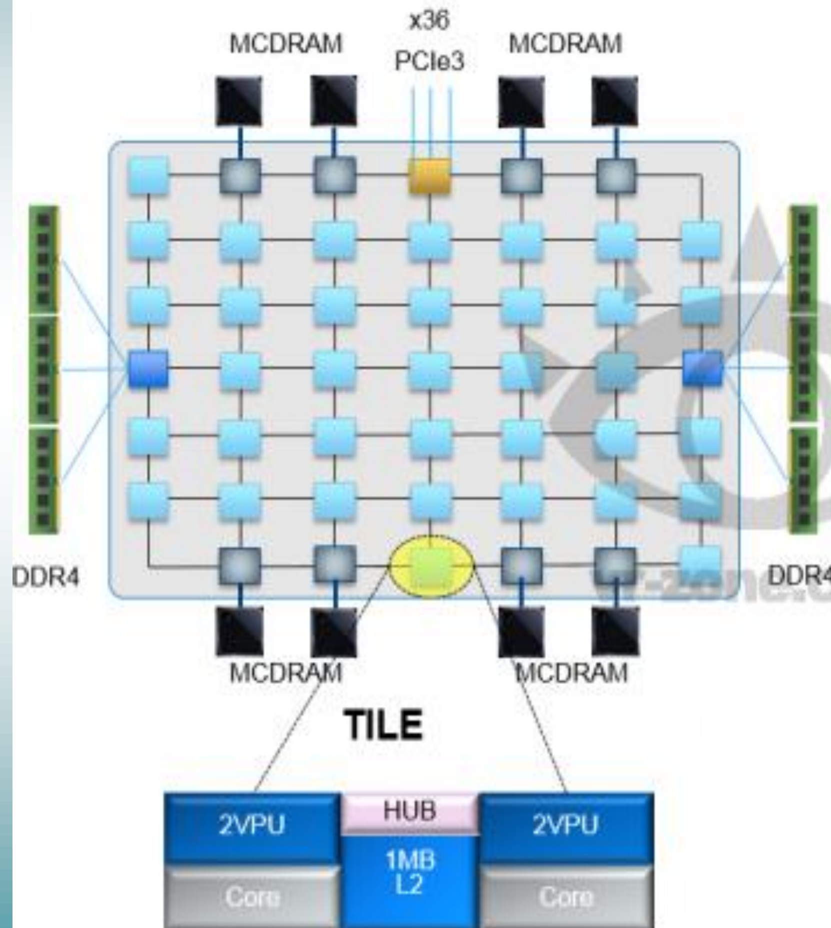
On-Package Memory: High Performance

- *up to* **16GB** *at launch*
- **1/3X** *the Space*[6]
- **5X** *Bandwidth vs DDR4*[7]
- **5X** *Power Efficiency*[6]

Jointly Developed with Micron Technology

Moving Forward: Memory is part of the Fabric

Knights Landing Processor Architecture



Up to 72 Intel Architecture cores based on Silvermont (Intel® Atom processor)

- Four threads/core
- Two 512b vector units/core
- Up to 3x single thread performance improvement over KNC generation

Full Intel® Xeon processor ISA compatibility through AVX-512 (except TSX)

6 channels of DDR4 2400 MHz -up to 384GB

36 lanes PCI Express* Gen 3

8/16GB of high-bandwidth on-package MCDRAM memory >500GB/sec

200W TDP

Memory, Switches, and Photonics

Intel® Omni Scale—The Next-Generation Fabric

- Designed for Next Generation HPC
- Host and Fabric Optimized for HPC
- Supports Entry to Extreme Scale
- End-to-End Solution

INTEGRATION

Intel® Omni Scale Fabric

Knights Landing

Intel® Omni Scale Fabric

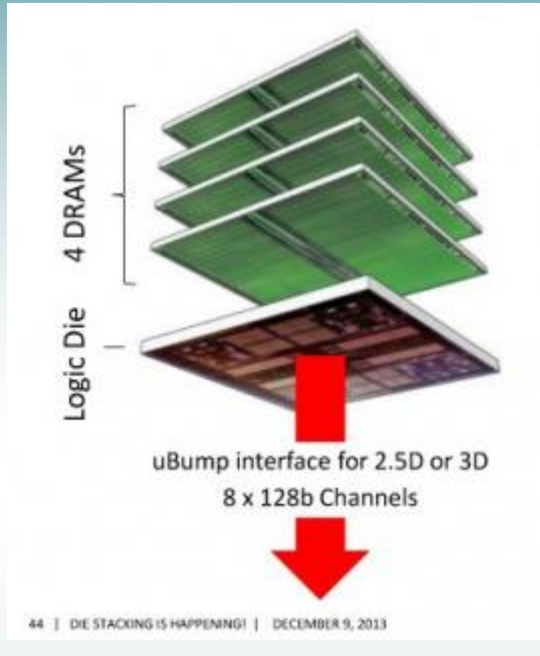
Future 14nm generation

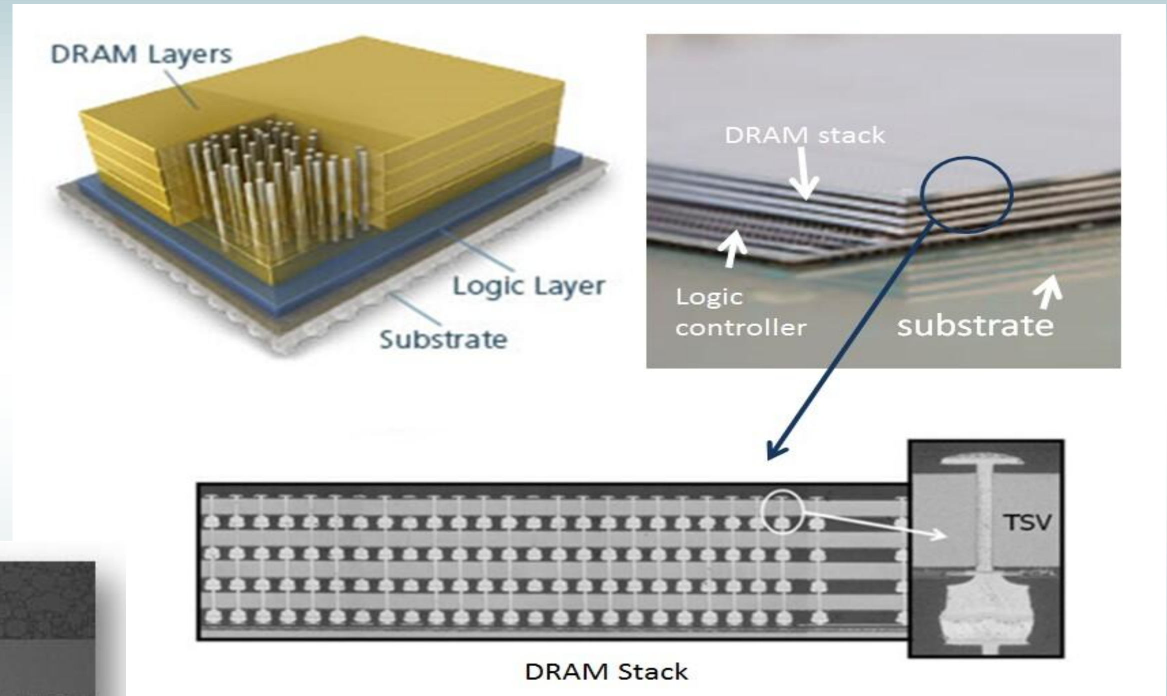
Coming in '15

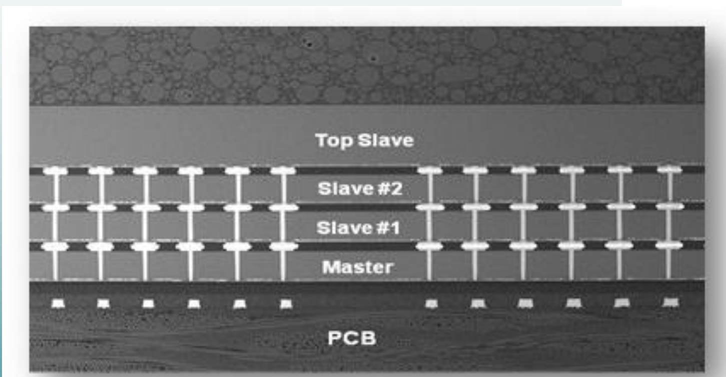
- √ PCIe Adapters
- √ Edge Switch
- √ Director Systems
- √ Intel Silicon Photonics
- √ Open Software Tools*

Intel® True Scale Fabric Upgrade Program *Helps Your Transition*

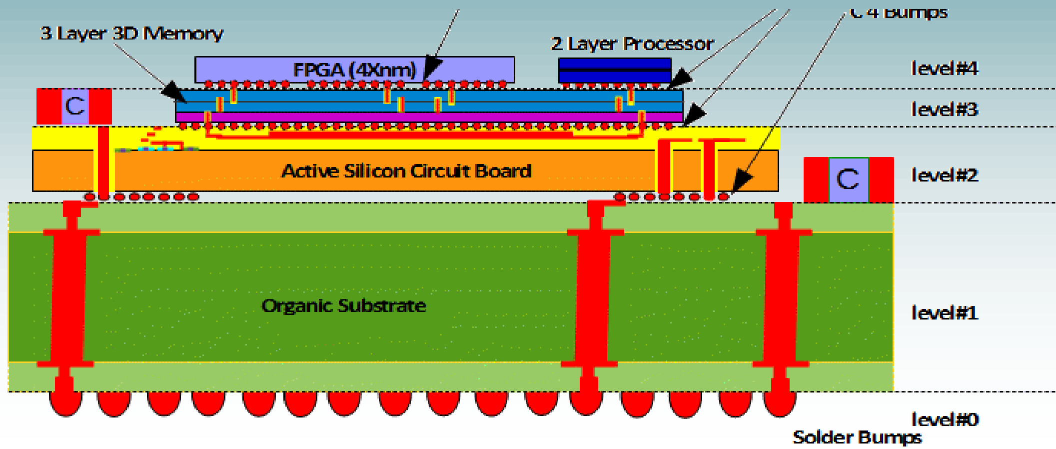
Industry Memories HBM/HMC/DDR4

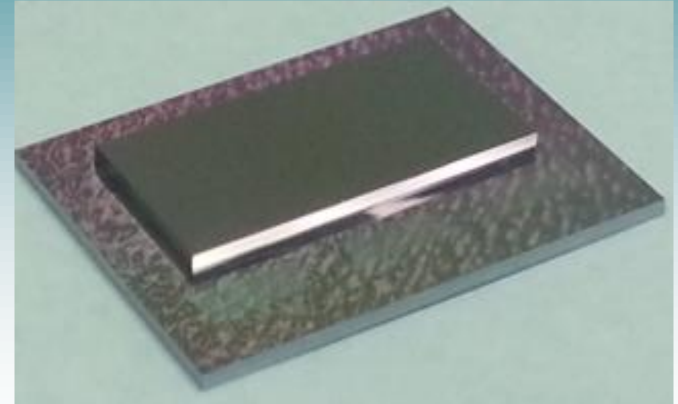


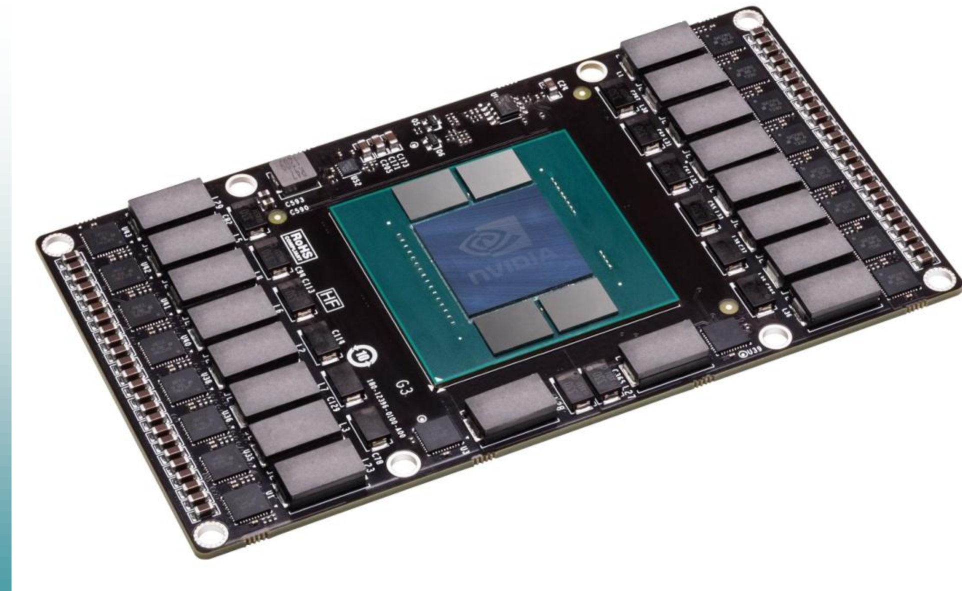




Advanced Packaging Drives Systems







Conventional RAM Architecture

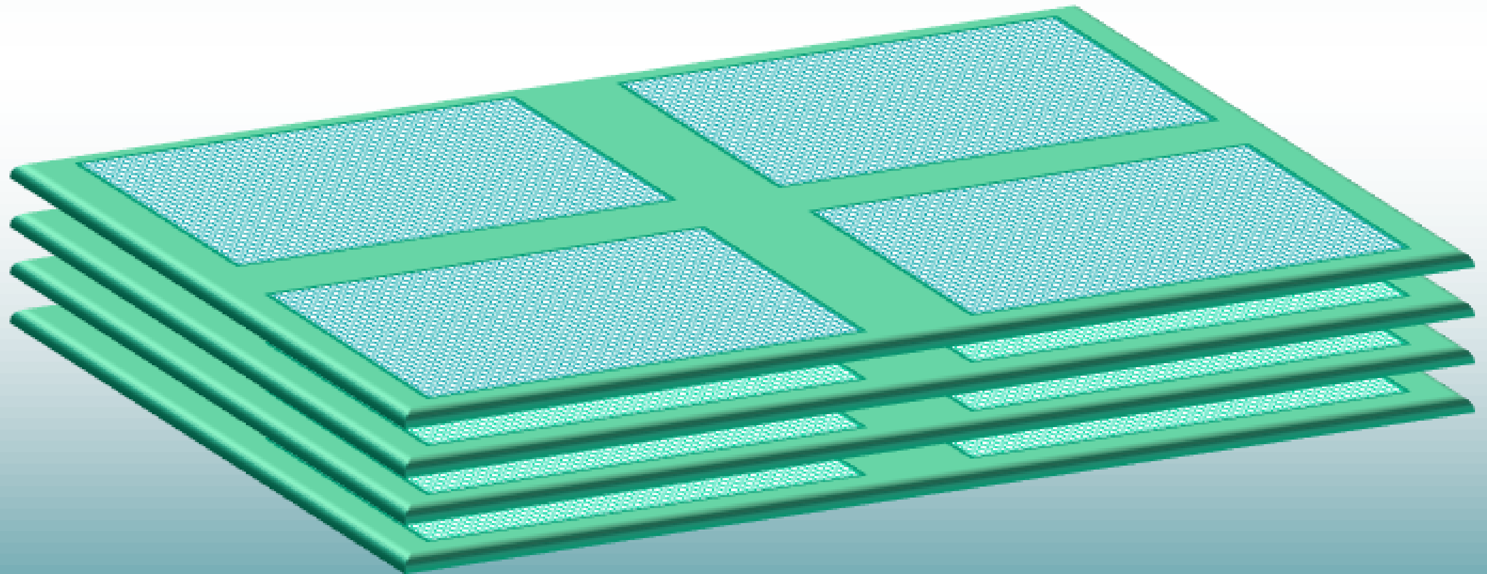
Memory Bits

Periphery

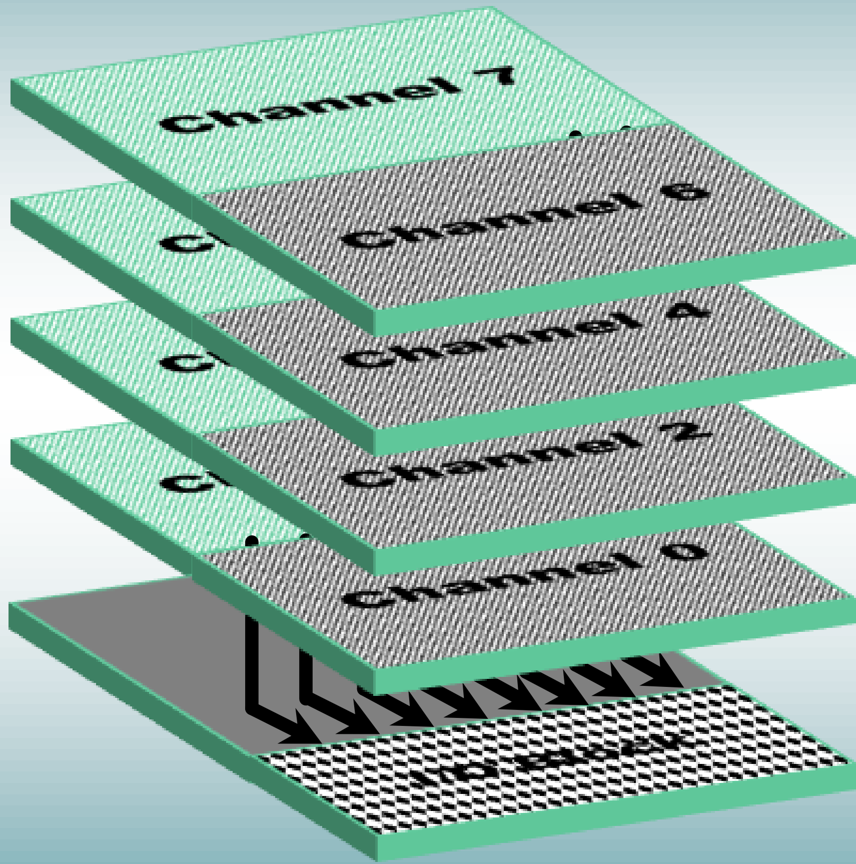
- Decoders
- Amps
- Drivers
- etc.

Conventional 3D *Packaging*

- Preserves traditional RAM problems
- Adds stacking costs



HBM Architecture

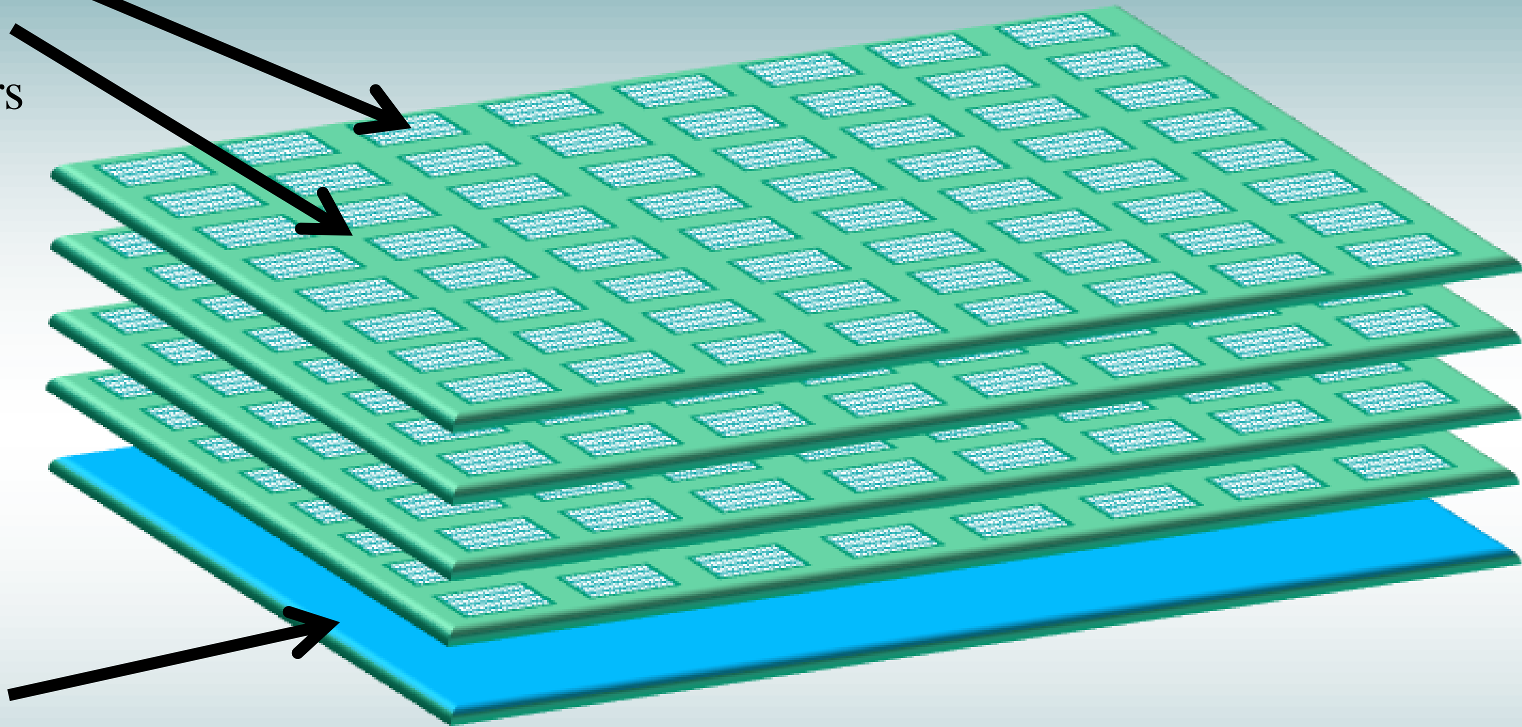


- 2 Channel / Die
- 4 or 8 Channel Stacks
- 128 bit CIO bus
- B2 or B4
- 2K or 4K Bytes / Page
- 1.2V I/O
- Staggered 96mm x 55mm Pitch, 25µ bump

HMC Architectural Approach

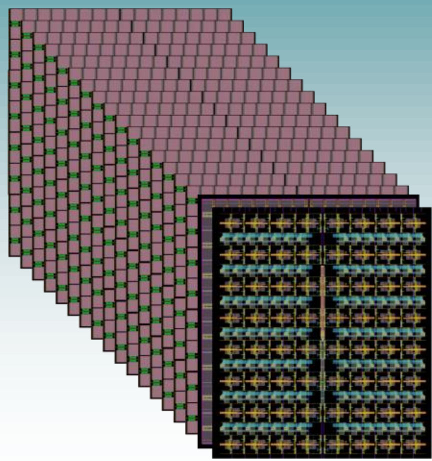
Memory Bits

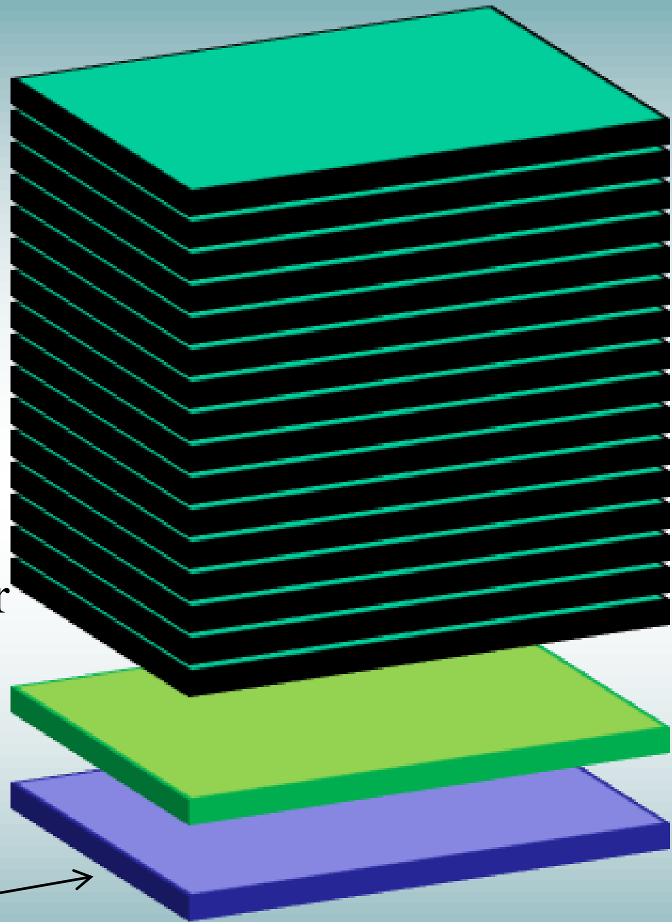
Periphery
- Decoders
- SAmps
- Drivers



High Speed SERDES I/O In Logic Process

Gen4 “Dis-Integrated” 3D Memory





DRAM layers
4xnm node

2 million vertical connections per layer per die

I/O layer contains: I/O, interface logic and R&R control CPU. 65nm node

Controller layer contains: sense amps, CAMs, row/column decodes and test engines. 40nm node

Better yielding than 2D equivalent!

New Memories

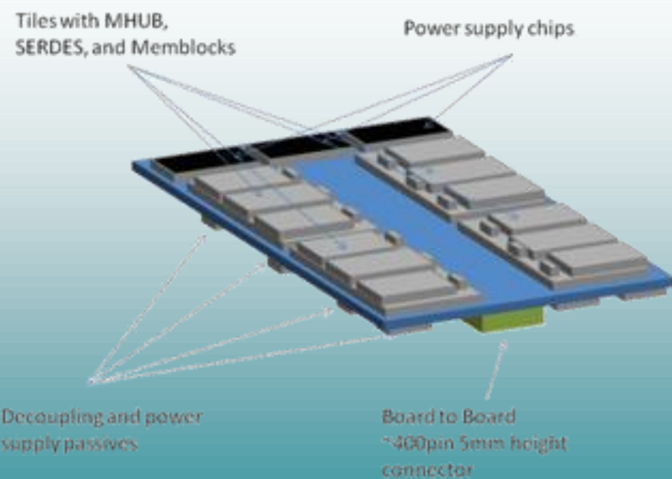
	DiRAM4	DiRAM4 CMOS	DiRAM4 Hub	HMC	HBM
Density	**64 Gb**	**64Gb**	**128 Gb**	16 Gb	8 Gb
	8 GB	**8 GB**	**16 GB**	2 GB	1 GB
Latency	**7 ns**	**7 ns**	**Variable**	Variable	33 ns
Min Ref	**64 bits**	**64 bits**	**256 bits**	256 bits	256 bits
Interface	**0.7 V**	**0.7 – 1.2 V**	**1.2 V**	SerDes	1.2 V
tRC	**15 ns**	**15 ns**	**15 ns**	50 ns	48 ns
BW	**16 Tb/s**	**4 Tb/s**	**1Tb/s**	1 Tb/s	1 Tb/s
	2 TB/s	**500 GB/s**	**128 GB/s**	128 GB/s	128 Gb/s
Channels	**256**	**64**	**1**	1*	8
Banks per Channel	**16**	**64**	**8192**	128	8

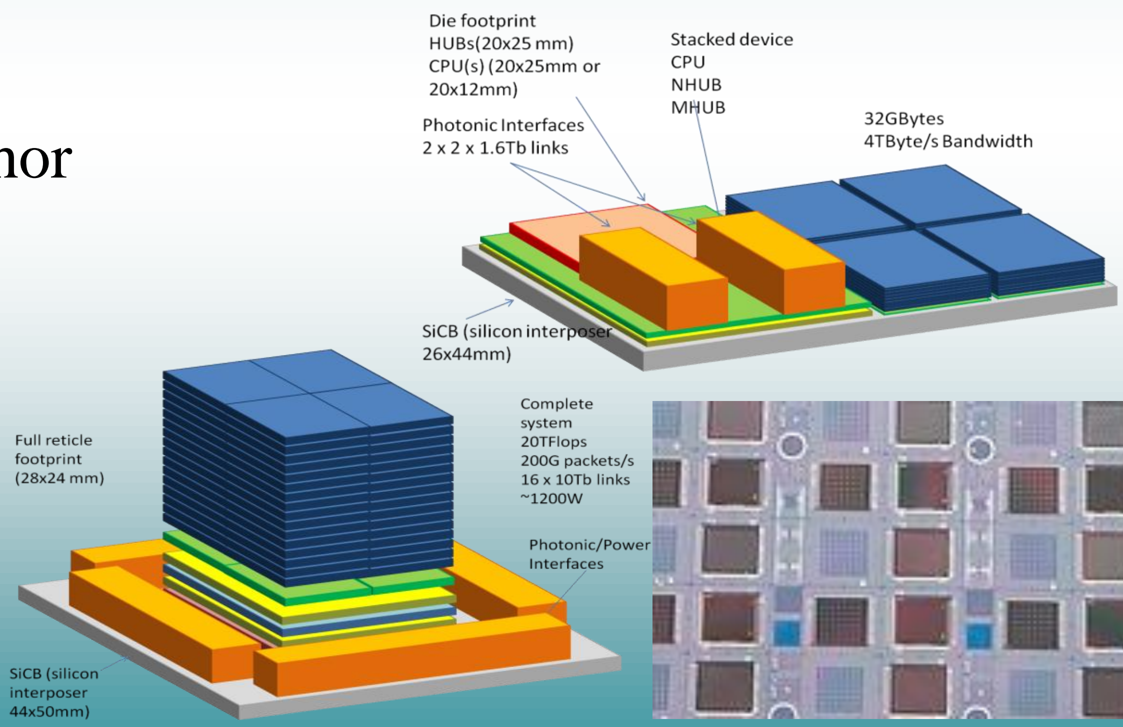
Next: NDP/ Processor in Memory

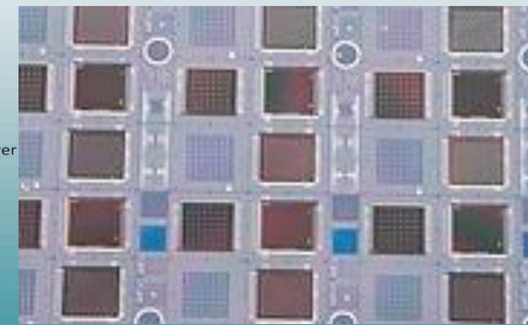
- 64 light weight cores
  - ~32GOPS + 1Tb/s external request traffic + 3Tb/s route through traffic
- 4 port
  - ~1Tb/s/port
- Built-in full crossbar
- 64Gb
- Packet based
  - Nodal addressing support
- Memory fabric is the machine fabric

Interposer Systems

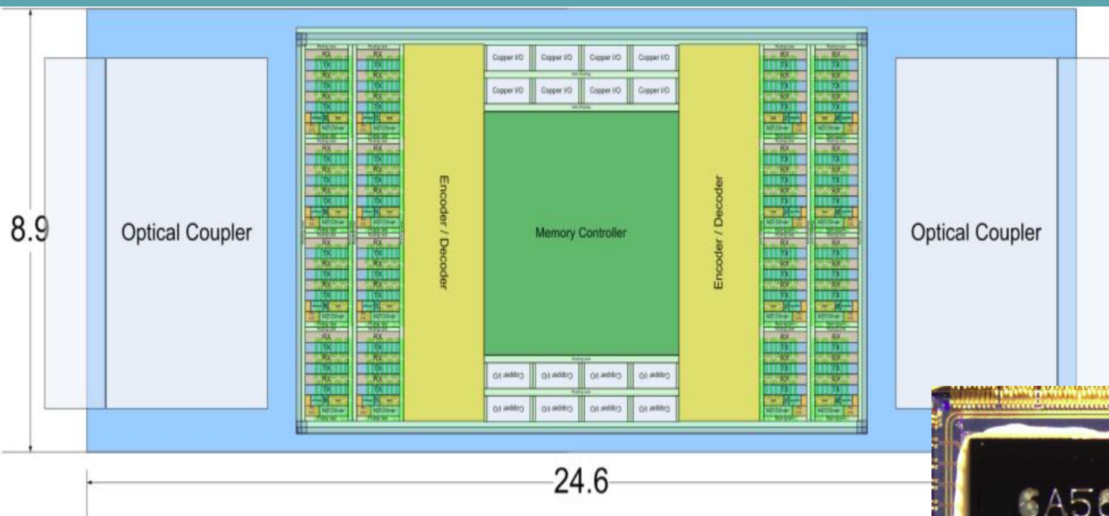
- SIP/SSIP
  - Power Conversion
  - Cooling
  - Photonics
- Optimization
  - Extending to power
- Mixed PCB/IC Metaphor

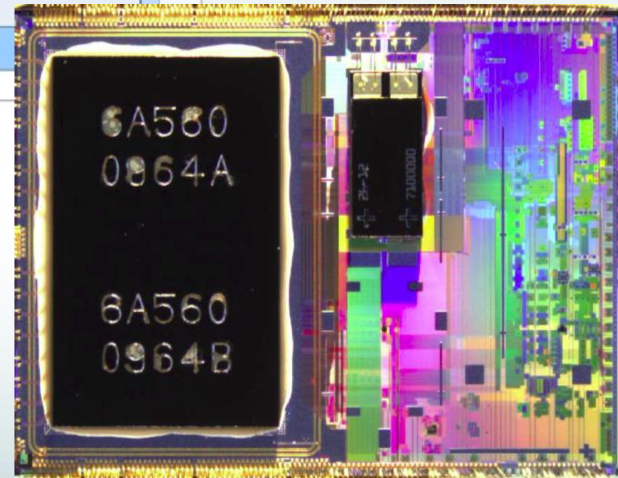


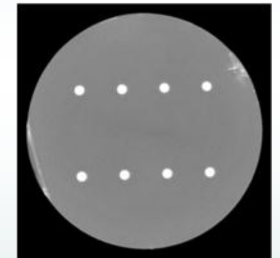




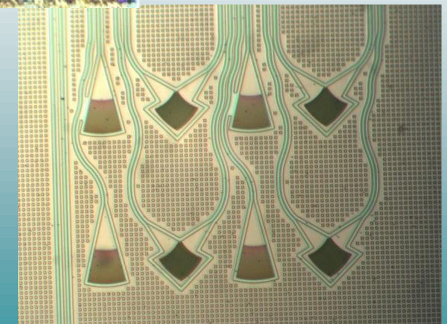
Luxtera Photonic Transceiver



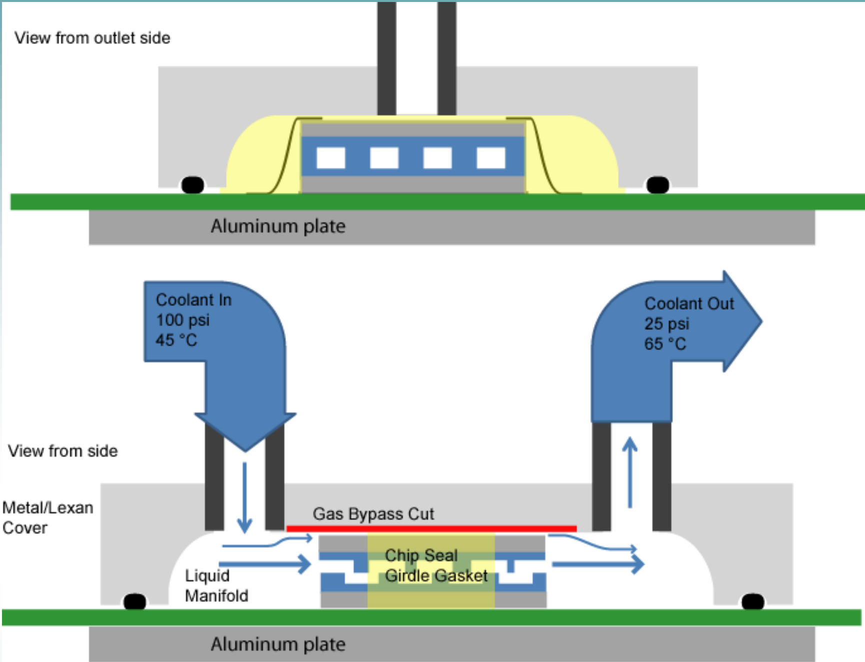


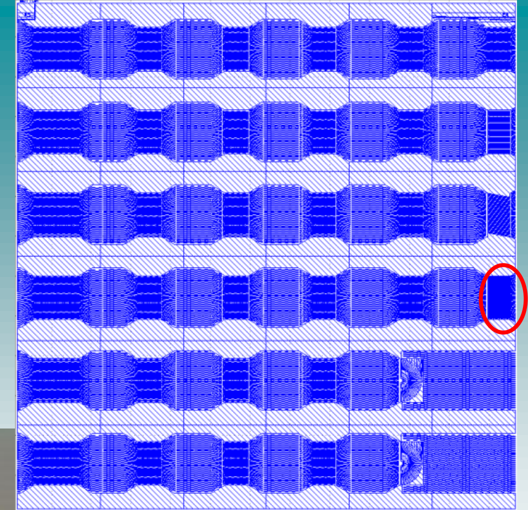


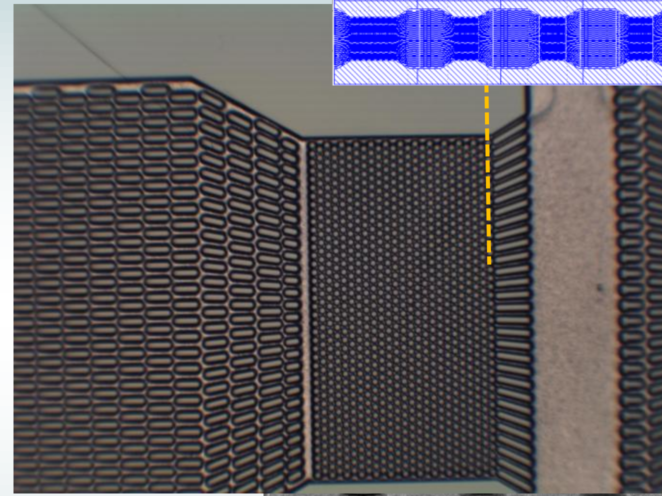
- 2.5pJ/bit
- 8 core fiber
- Logic die
- Smart self-calibrating self-tuning
- >1.6Tb/s payload

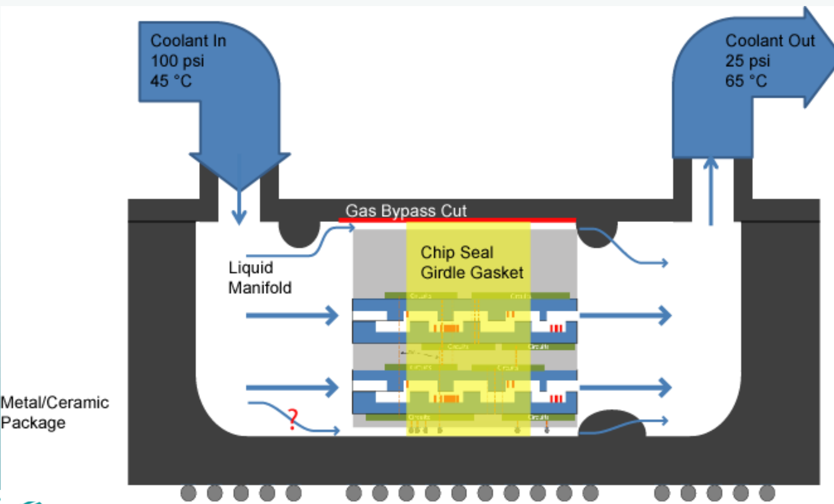


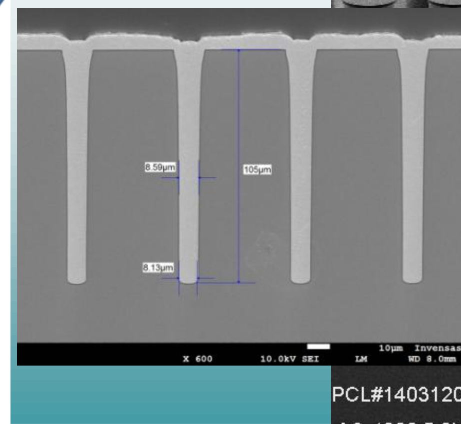
Cooling Block Illustrations

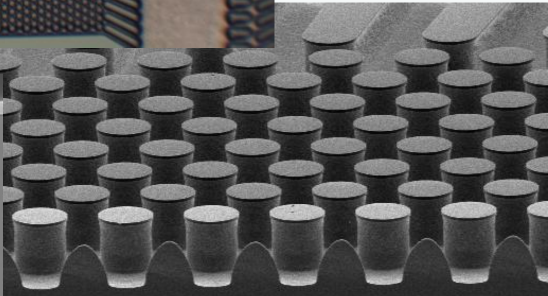






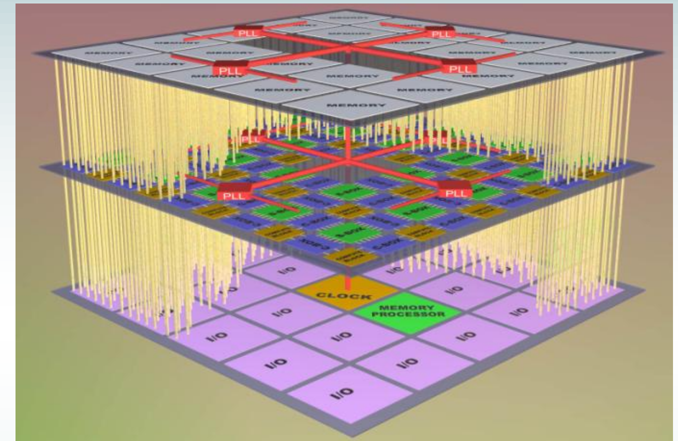






Summary

- 3D is required for new levels of performance
- 3D enables new applications
- 3D does what More Moore can't
- It's not just about new manufacturing technology
  - Its new design
  - Its new architecture



Sensors

Computing

MEMS

Communications