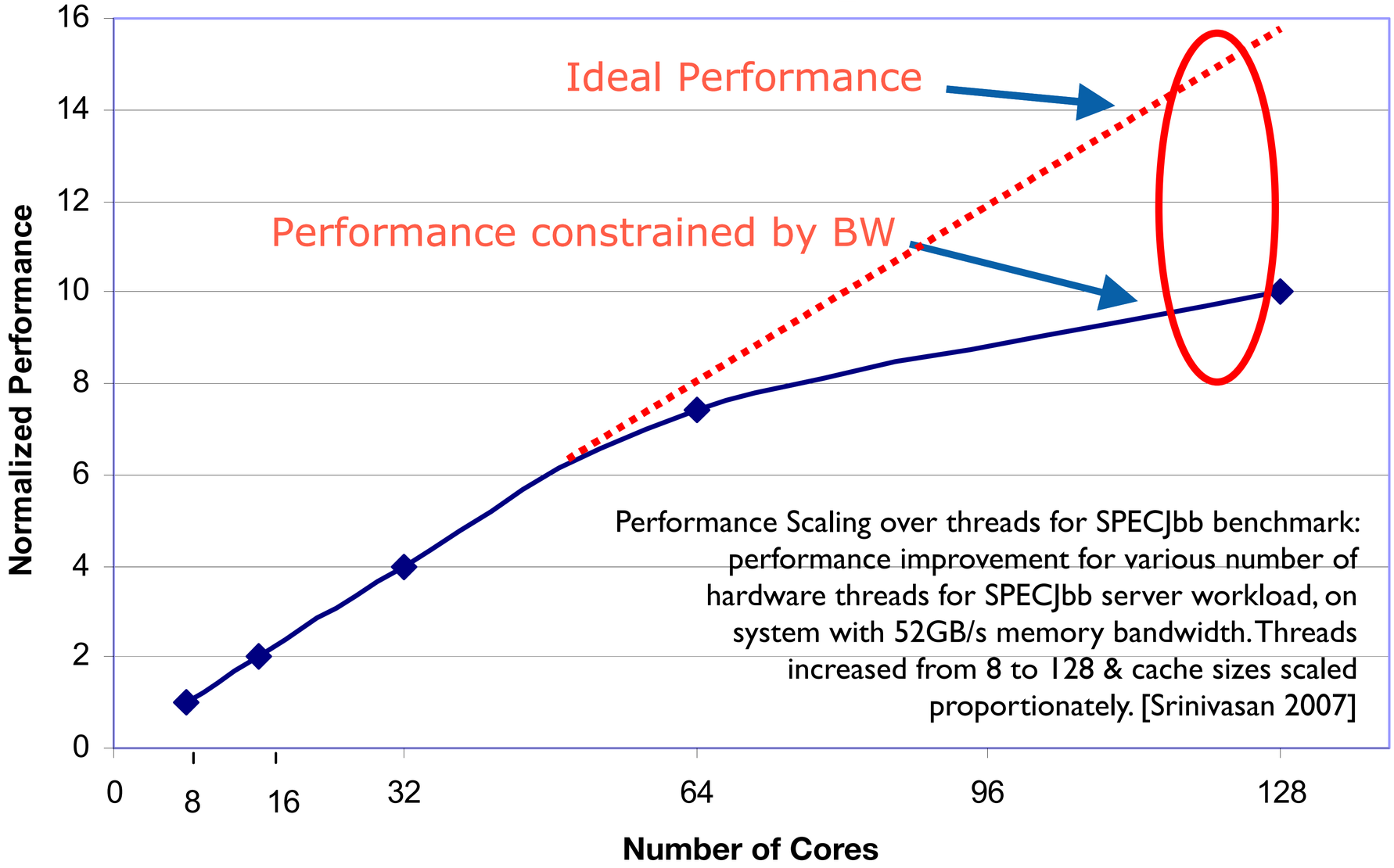
All Tomorrow's Memories

... for multicore

Bruce Jacob

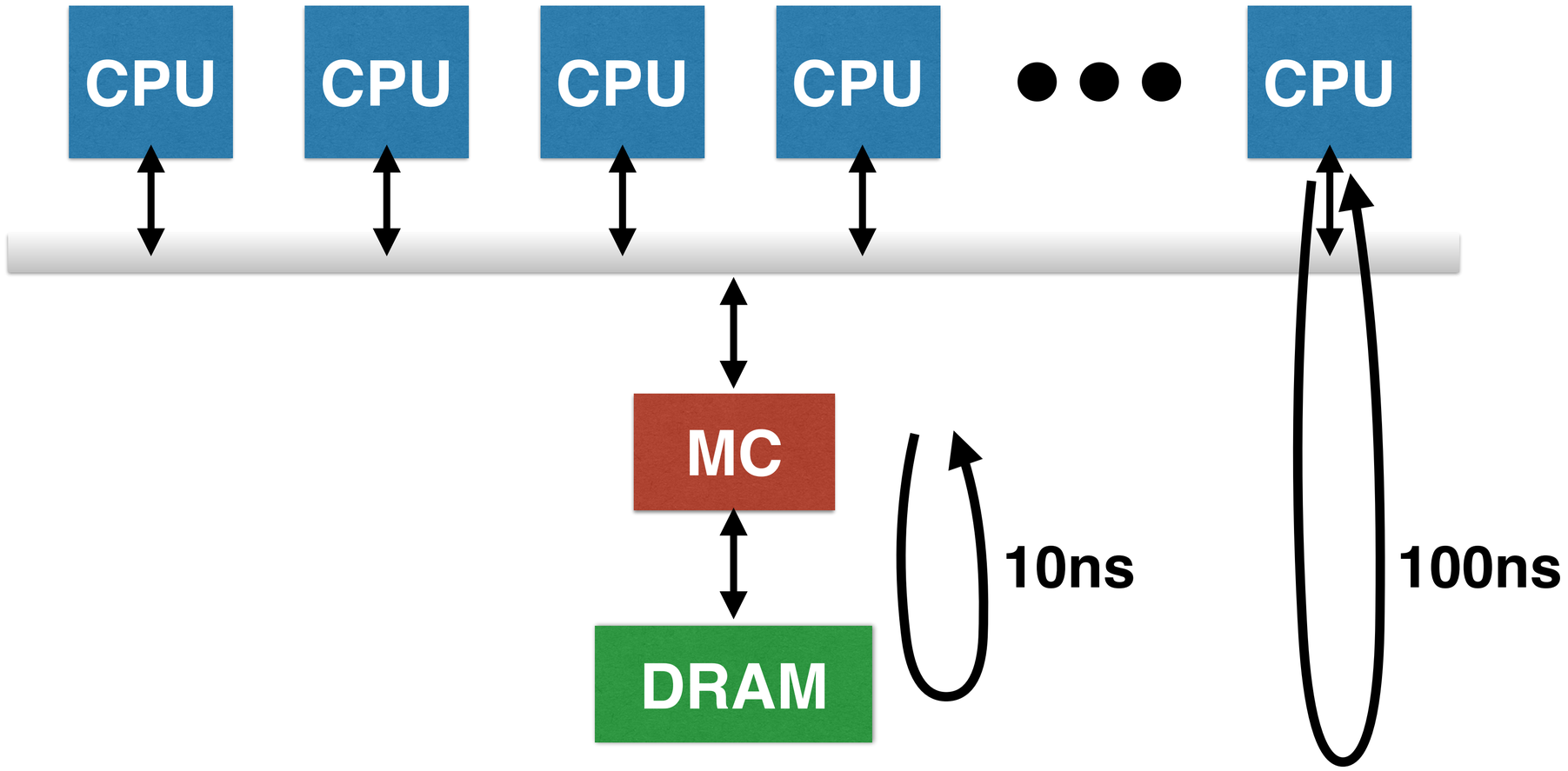
Keystone Professor
University of Maryland

Multicore Memory Bandwidth



Performance Scaling over threads for SPECJbb benchmark: performance improvement for various number of hardware threads for SPECJbb server workload, on system with 52GB/s memory bandwidth. Threads increased from 8 to 128 & cache sizes scaled proportionately. [Srinivasan 2007]

Multicore Memory Latency



Wish List [& Talk Outline]

Fine-Grained Access

Bandwidth

Capacity

Low Power

Nonvolatility

Wish List [& Talk Outline]

Fine-Grained Access

Bandwidth

Capacity

Low Power

Nonvolatility

] **DRAM - HBM/HMC***

* Things we did and/or are doing now (I'll cover in talk)

Wish List [& Talk Outline]

Fine-Grained Access

Bandwidth

] **DRAM - HBM/HMC***

Capacity

Low Power

Nonvolatility

] **Flash, 3DXP, ReRAM, PCM, etc - NVMM***

* Things we did and/or are doing now (I'll cover in talk)

Wish List [& Talk Outline]

Fine-Grained Access

Bandwidth

DRAM - HBM/HMC*

Capacity

Low Power

Nonvolatility

Flash, 3DXP, ReRAM, PCM, etc - NVMM*

HBNV*

* Things we did and/or are doing now (I'll cover in talk)

Wish List [& Talk Outline]

Fine-Grained Access

Bandwidth

DRAM - HBM/HMC*

Capacity

Low Power

Nonvolatility

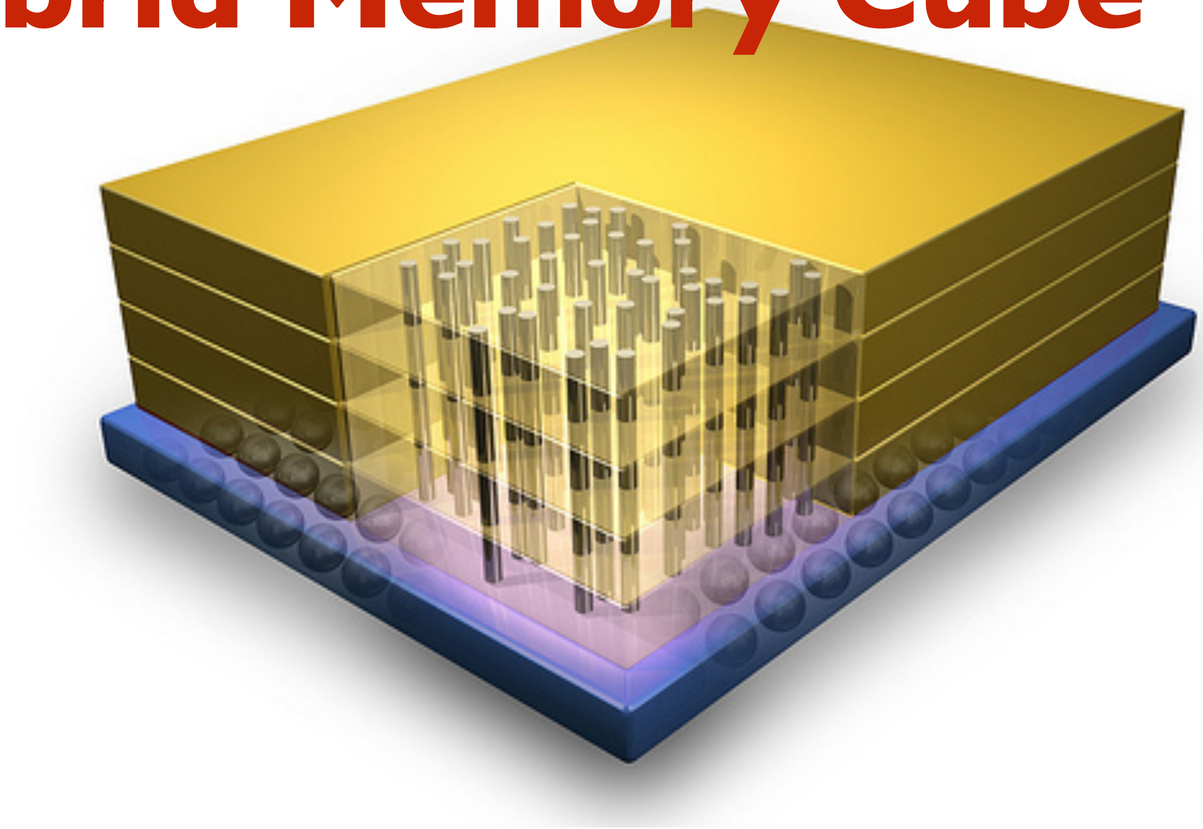
Flash, 3DXP, ReRAM, PCM, etc - NVMM*

HBNV*

Major implications for OS and applications

* Things we did and/or are doing now (I'll cover in talk)

Hybrid Memory Cube



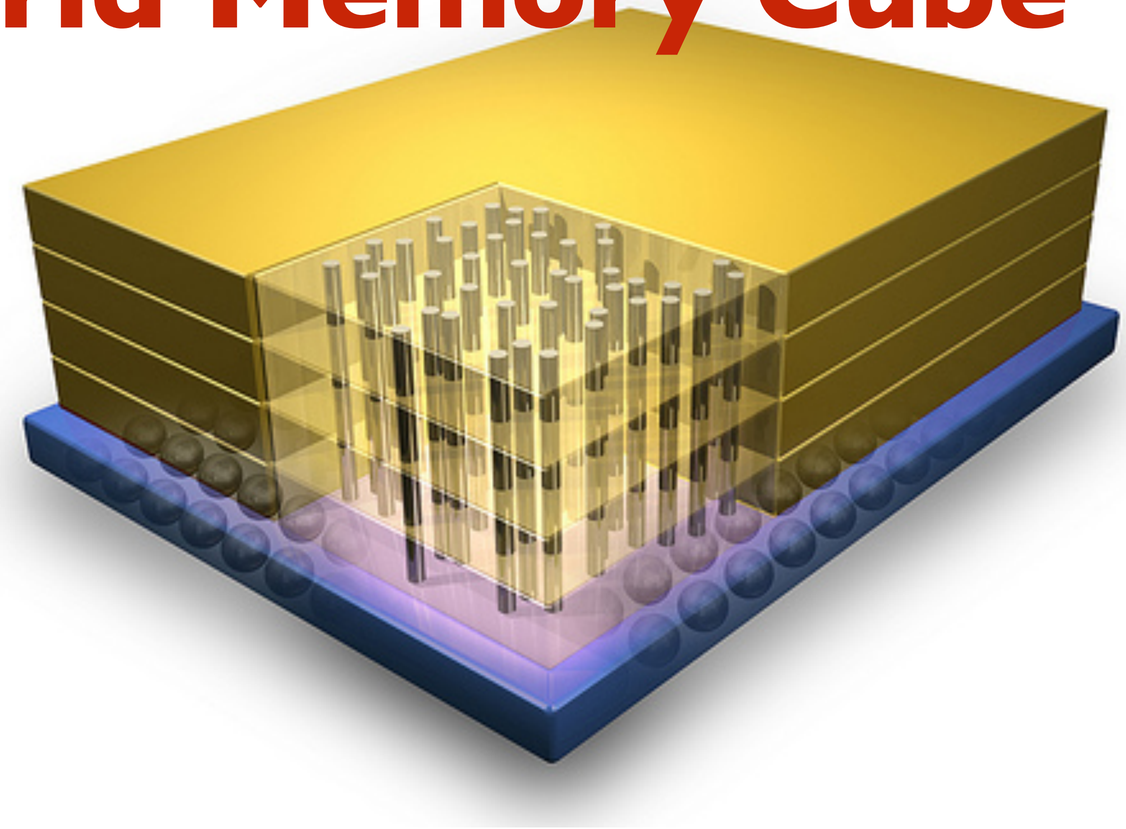
Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

Next gen is 160 GB/s per (640 total)

Max in-flight = 16 x 8 x 2..8 (256–1024)

Hybrid Memory Cube



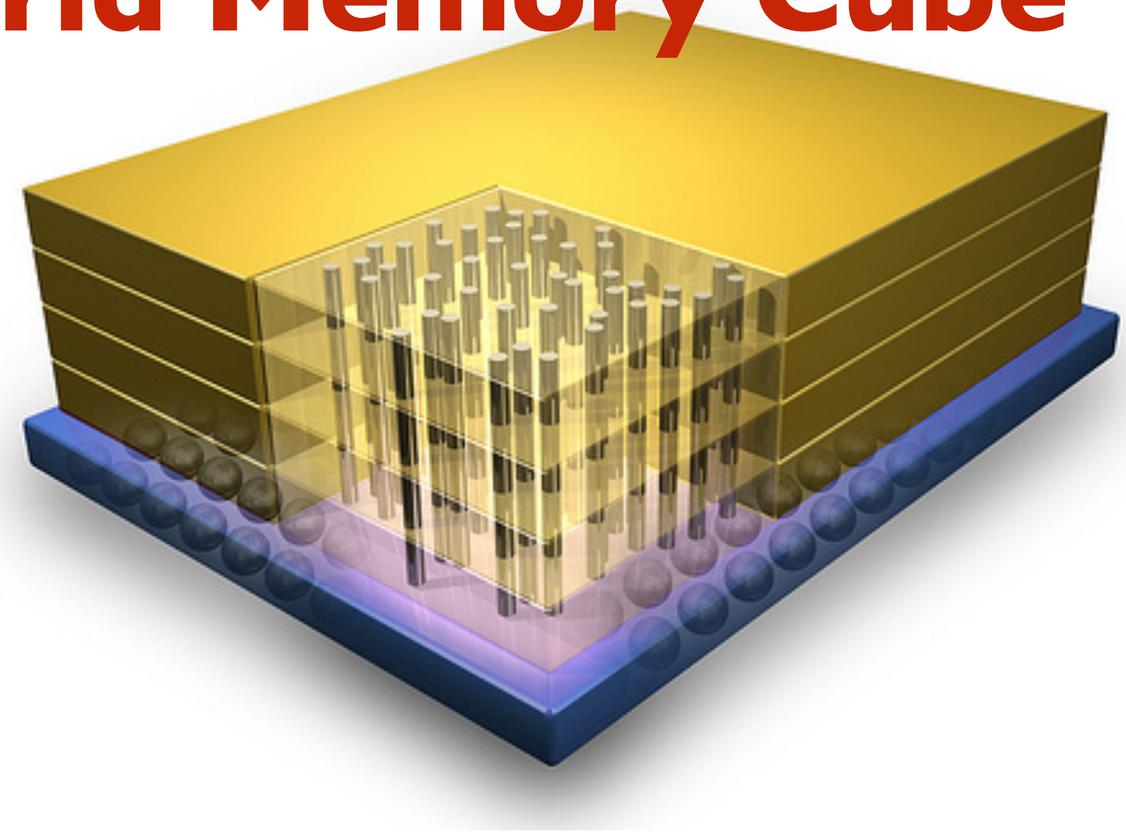
Off-chip: high speed SerDes and generic protocol

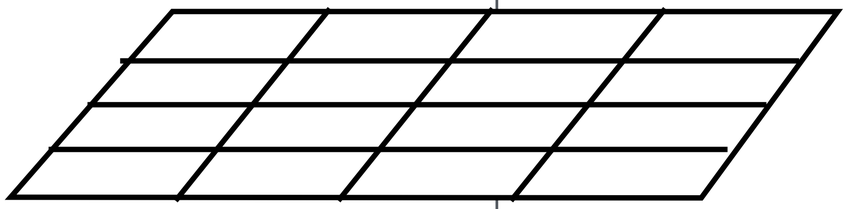
4 I/O Ports, up to 80 GB/s each

Next gen is 160 GB/s per (640 total)

Max in-flight = 16 x 8 x 2..8 (256–1024)

Hybrid Memory Cube





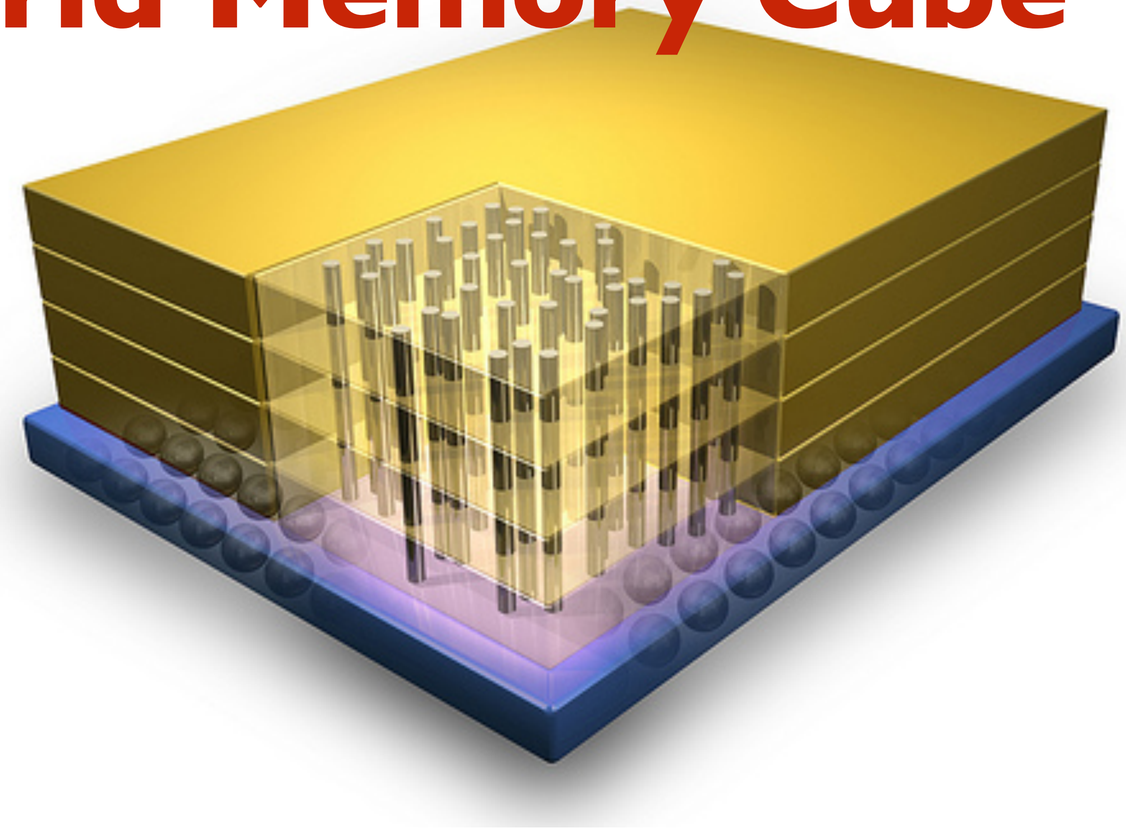
Off-chip: high speed SerDes and generic protocol

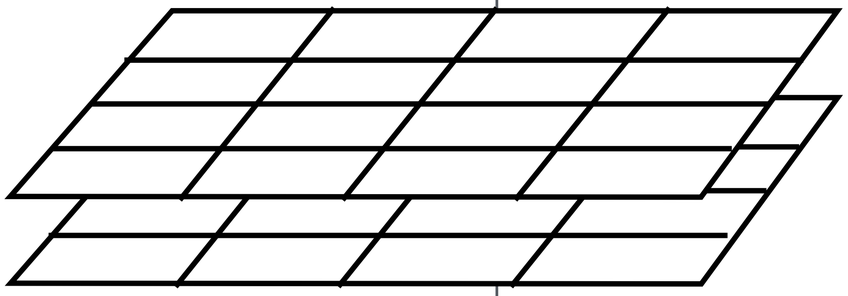
4 I/O Ports, up to 80 GB/s each

Next gen is 160 GB/s per (640 total)

Max in-flight = 16 x 8 x 2..8 (256–1024)

Hybrid Memory Cube





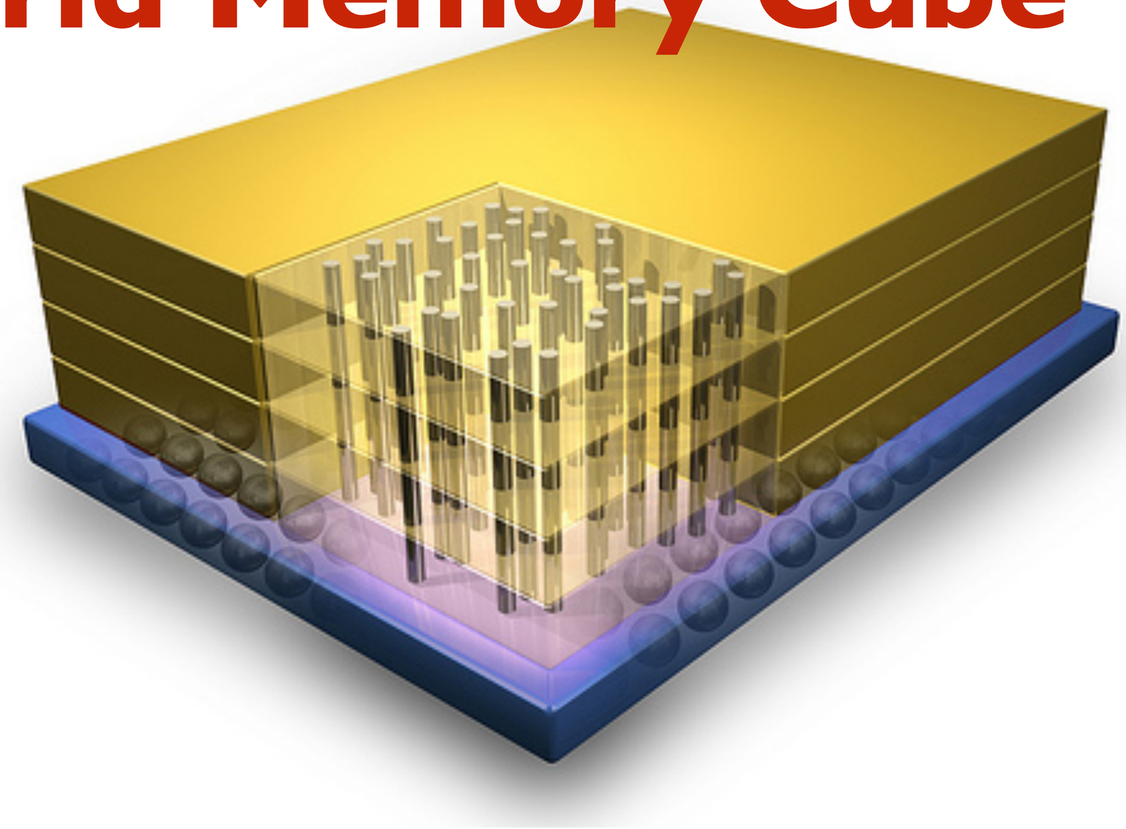
Off-chip: high speed SerDes and generic protocol

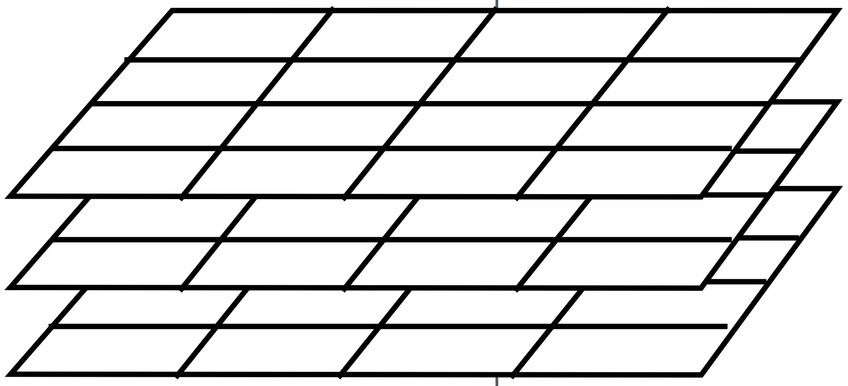
4 I/O Ports, up to 80 GB/s each

Next gen is **160** GB/s per (640 total)

Max in-flight = **16 x 8 x 2..8** (256–1024)

Hybrid Memory Cube





Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

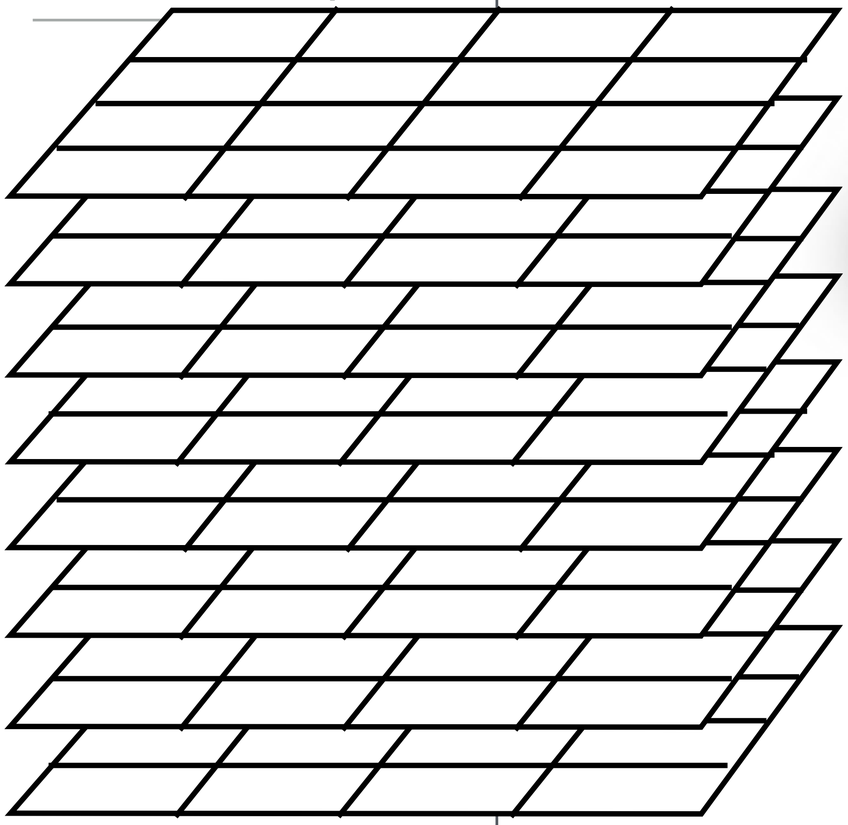
Next gen is **160** GB/s per (640 total)

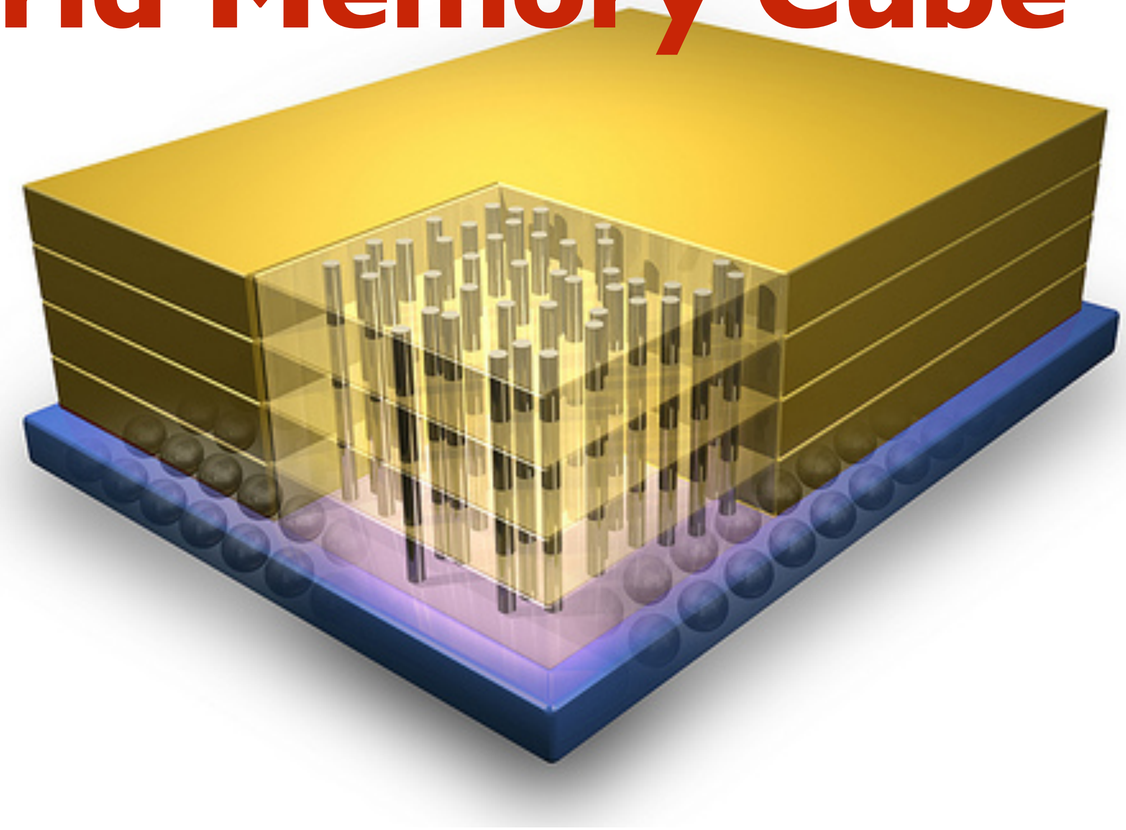
Max in-flight = **16 x 8 x 2..8** (256–1024)

Bruce Jacob

University of Maryland

Hybrid Memory Cube





Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

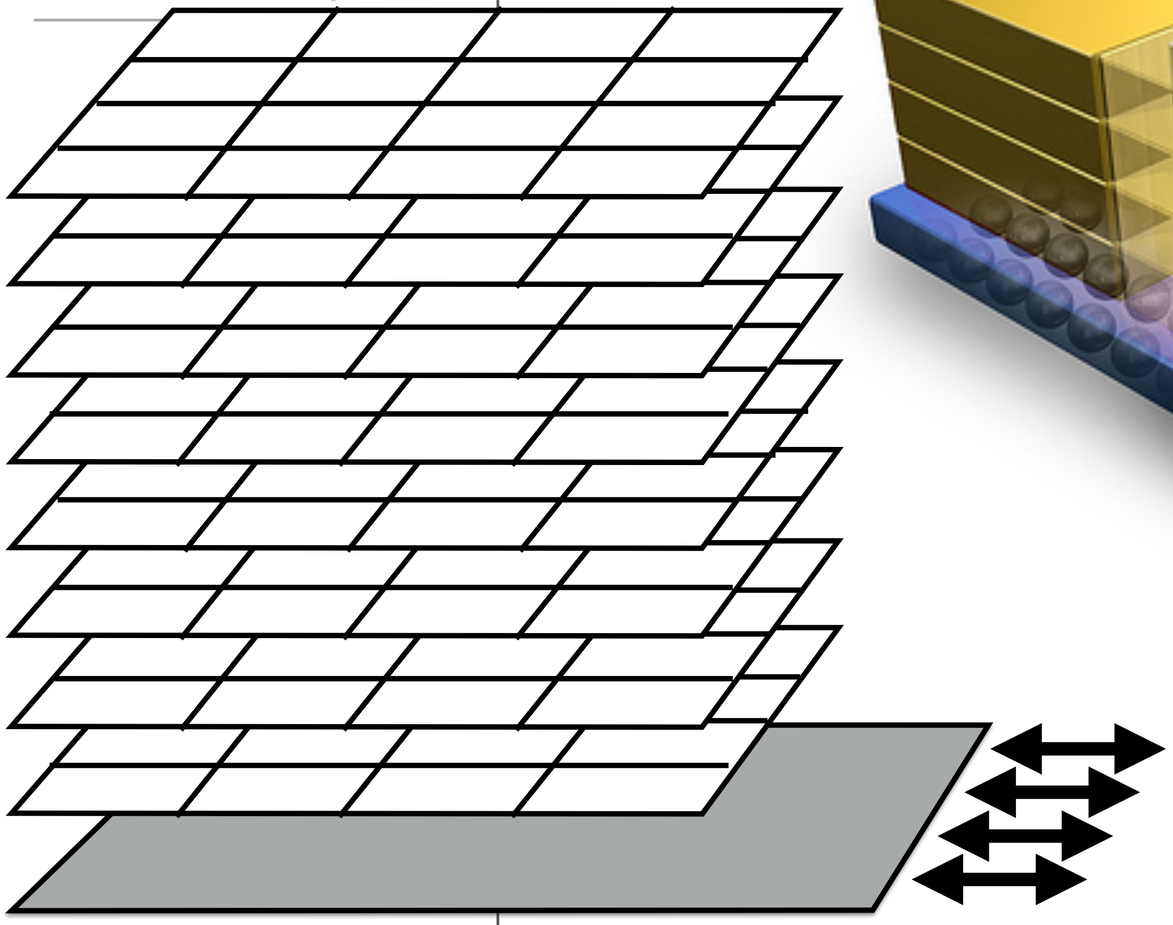
Next gen is 160 GB/s per (640 total)

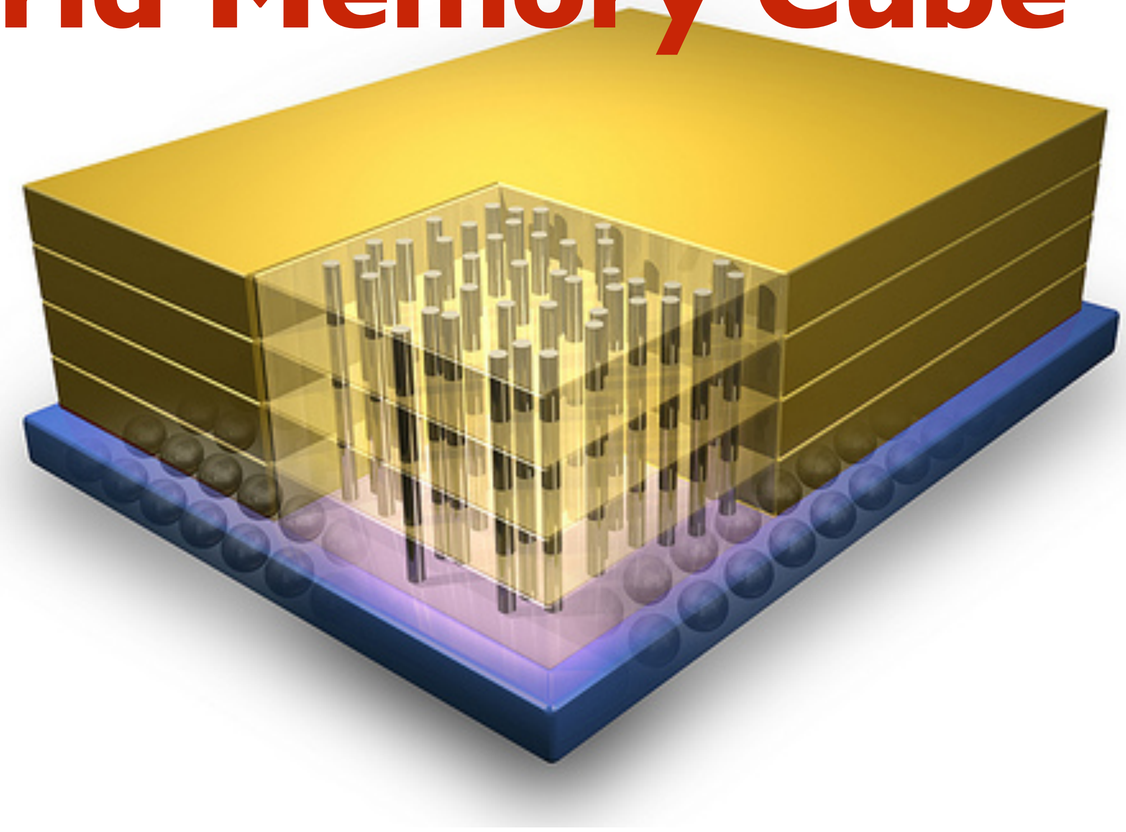
Max in-flight = 16 x 8 x 2..8 (256–1024)

Bruce Jacob

University of Maryland

Hybrid Memory Cube





Logic Base
(I/O & CTL)

Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

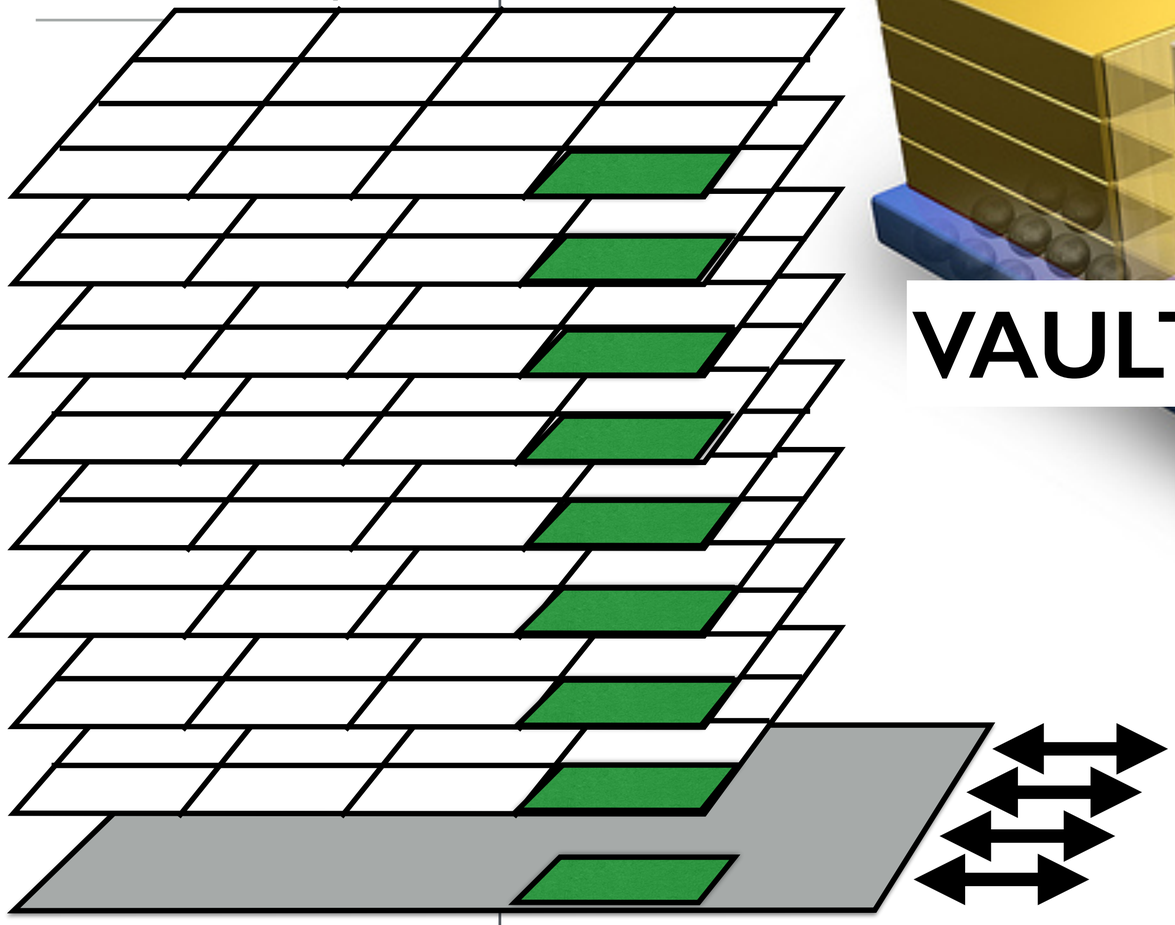
Next gen is **160** GB/s per (640 total)

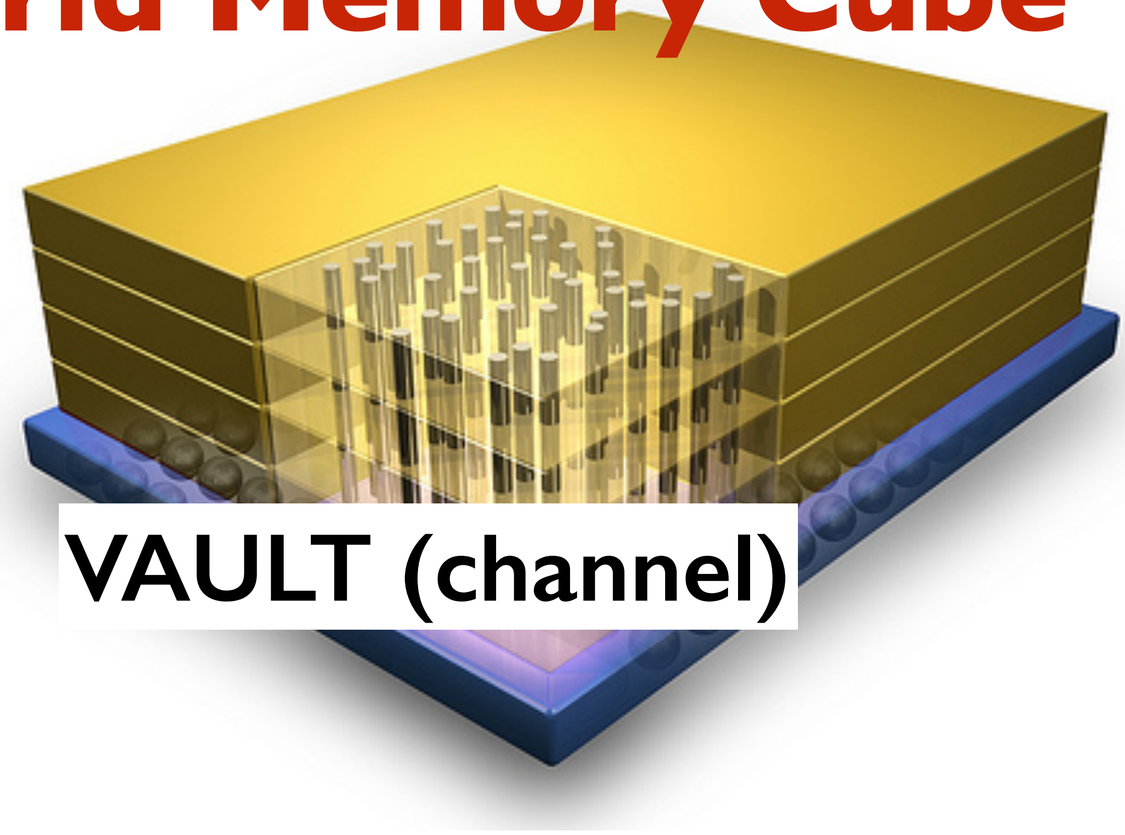
Max in-flight = **16 x 8 x 2..8** (256–1024)

Bruce Jacob

University of Maryland

Hybrid Memory Cube





Logic Base
(I/O & CTL)

Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

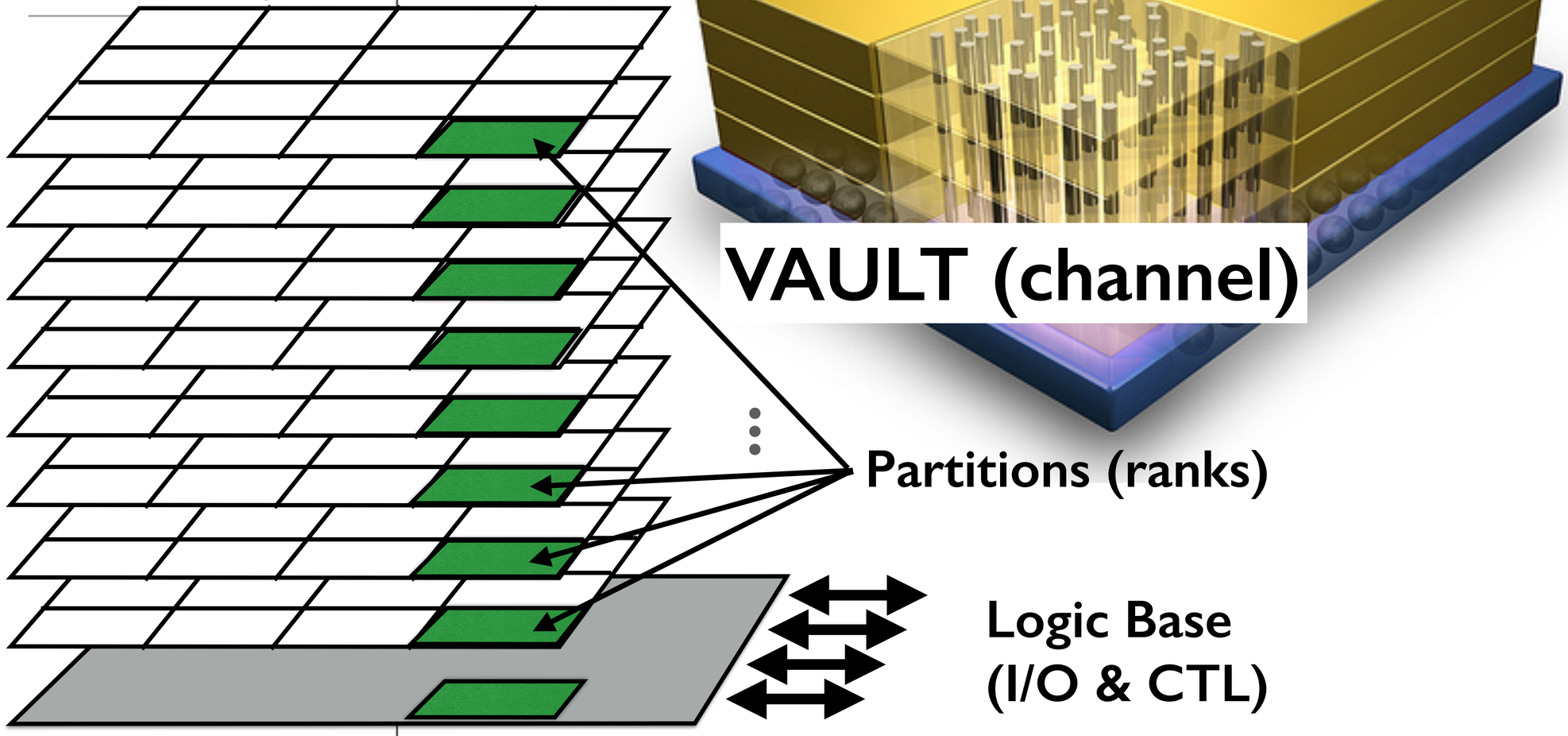
Next gen is **160** GB/s per (640 total)

Max in-flight = **16 x 8 x 2..8** (256–1024)

Bruce Jacob

University of Maryland

Hybrid Memory Cube



Off-chip: high speed SerDes and generic protocol

4 I/O Ports, up to 80 GB/s each

Next gen is **160** GB/s per (640 total)

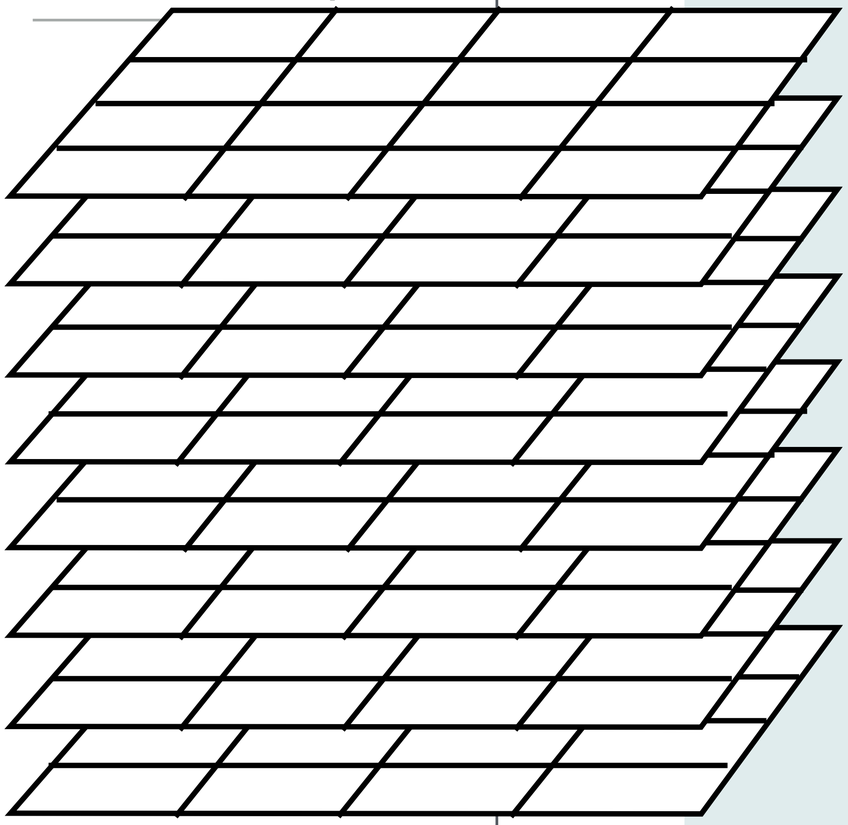
Max in-flight = **16 x 8 x 2..8** (256–1024)

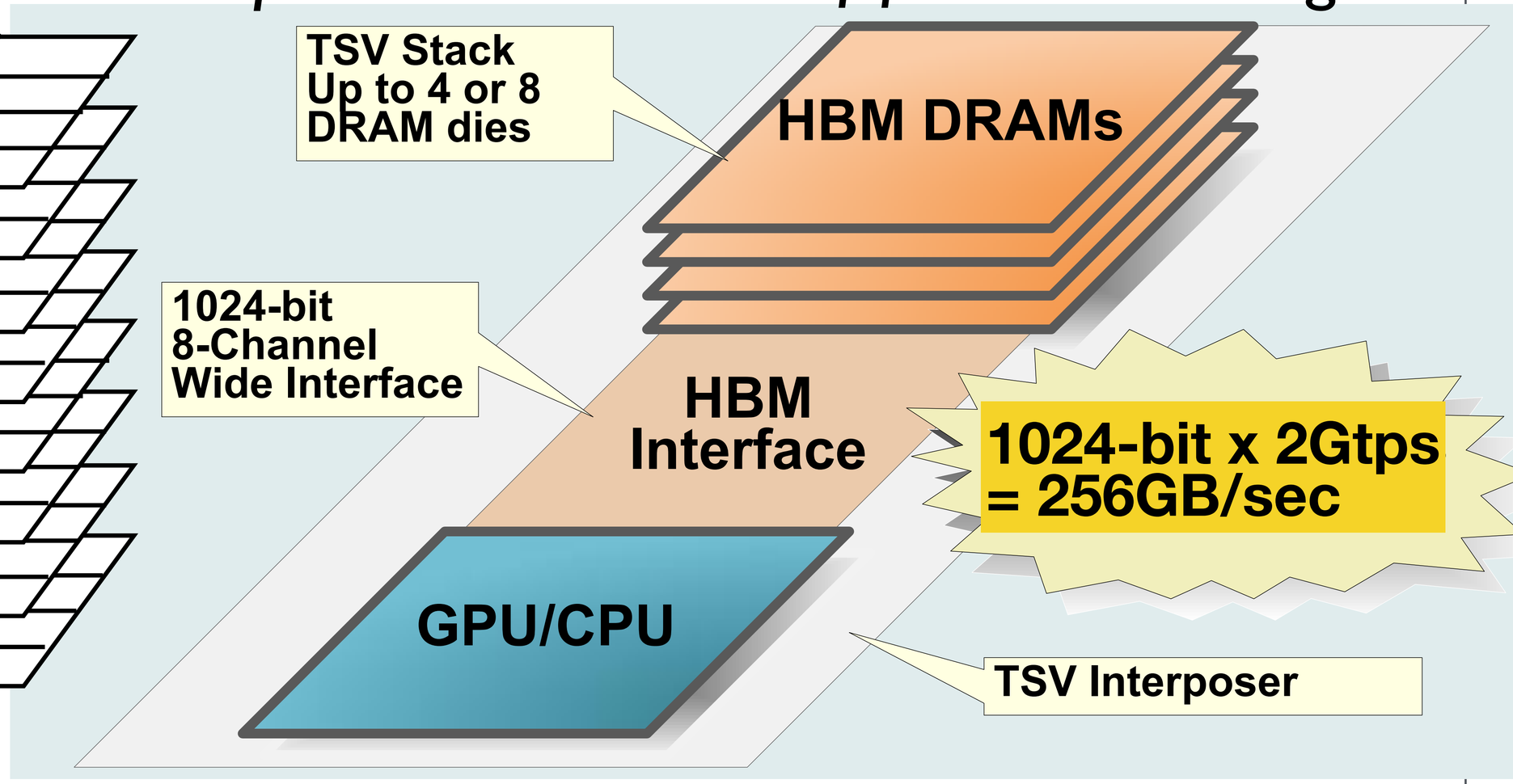
Bruce Jacob

University of Maryland

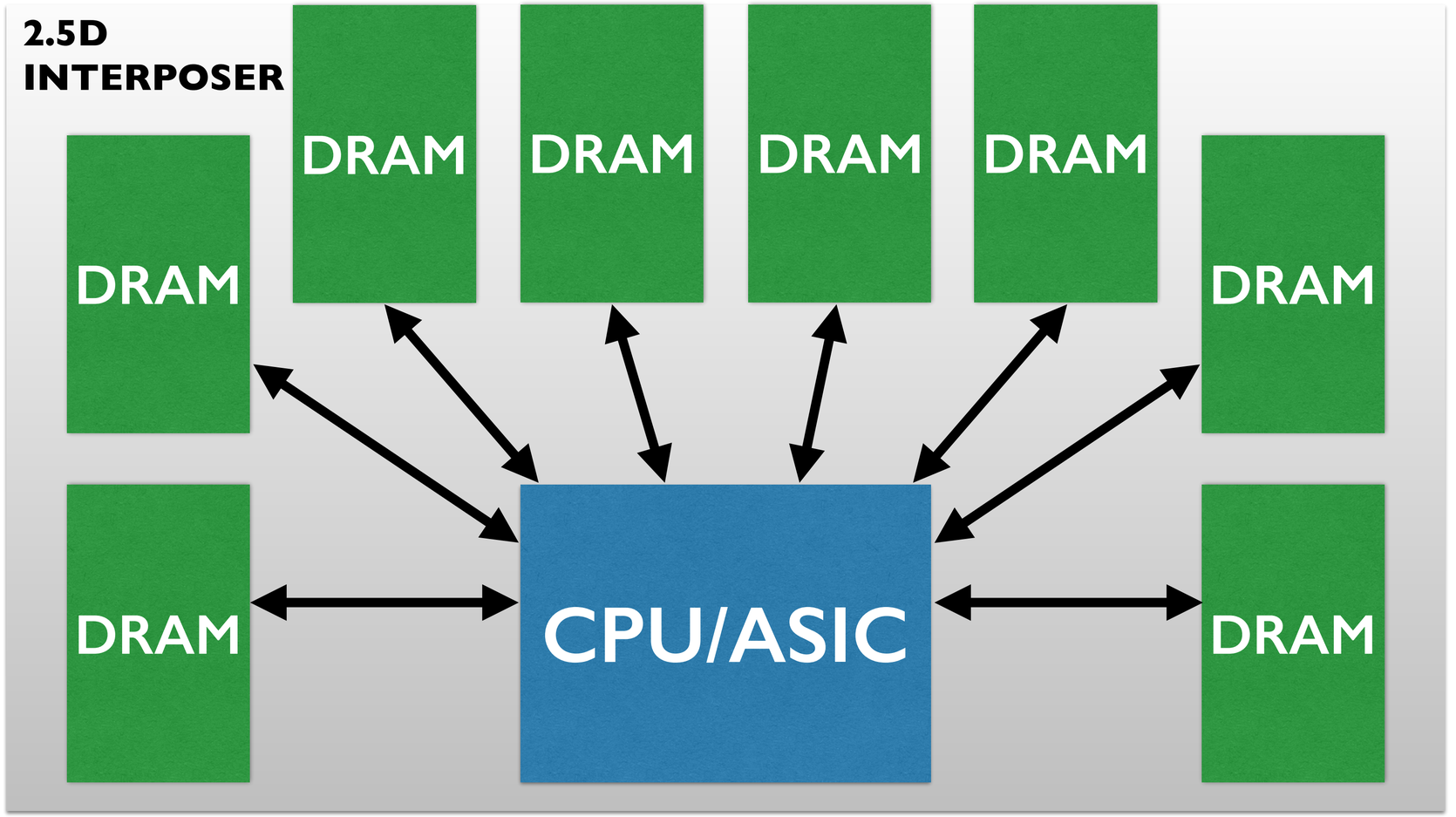
High Bandwidth Memory

Uses a simpler '2.5D' instead of full 3D stacking





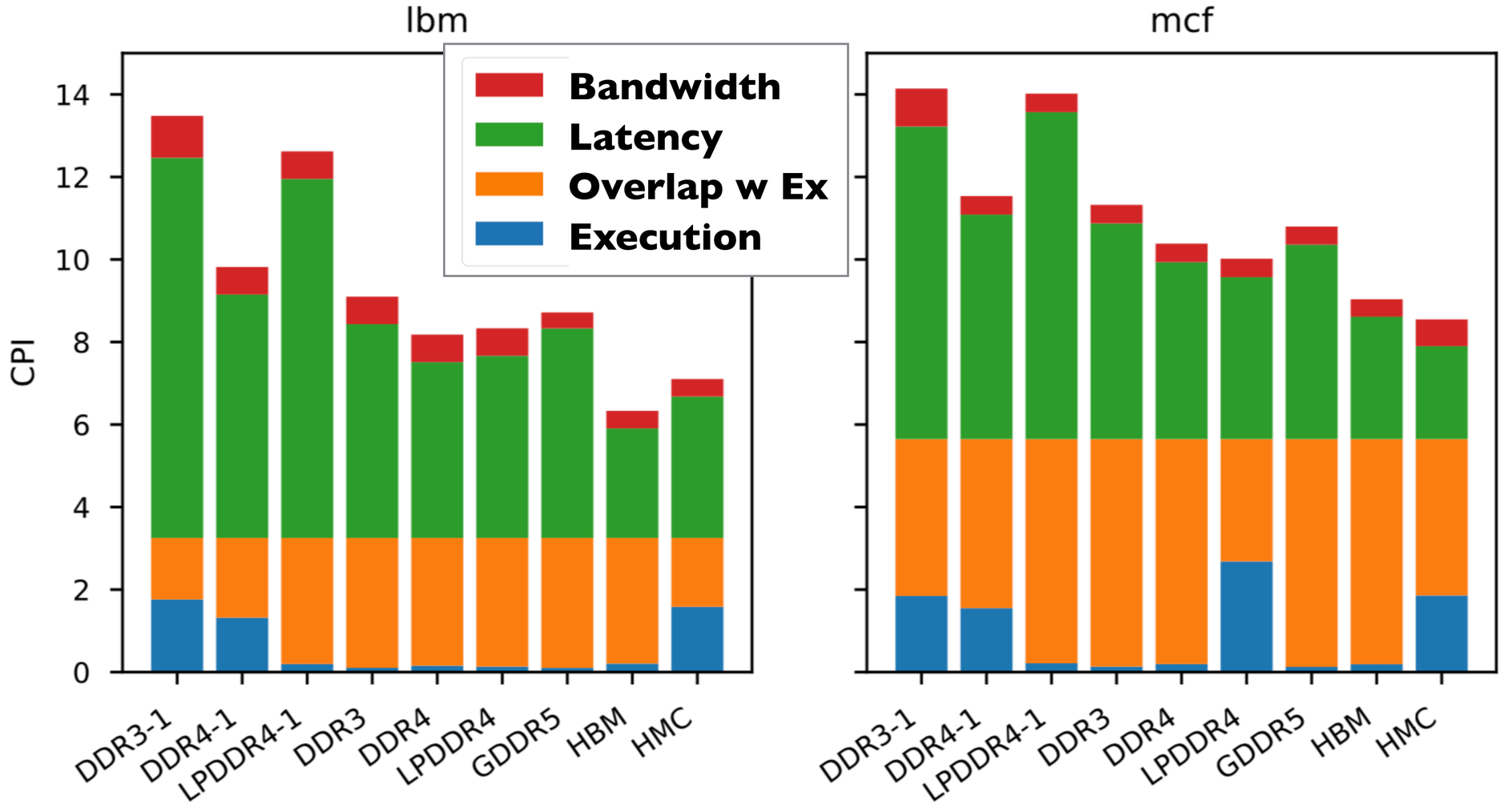
High Bandwidth Memory



Each Link is 128 Bits Wide: *1024 Total*

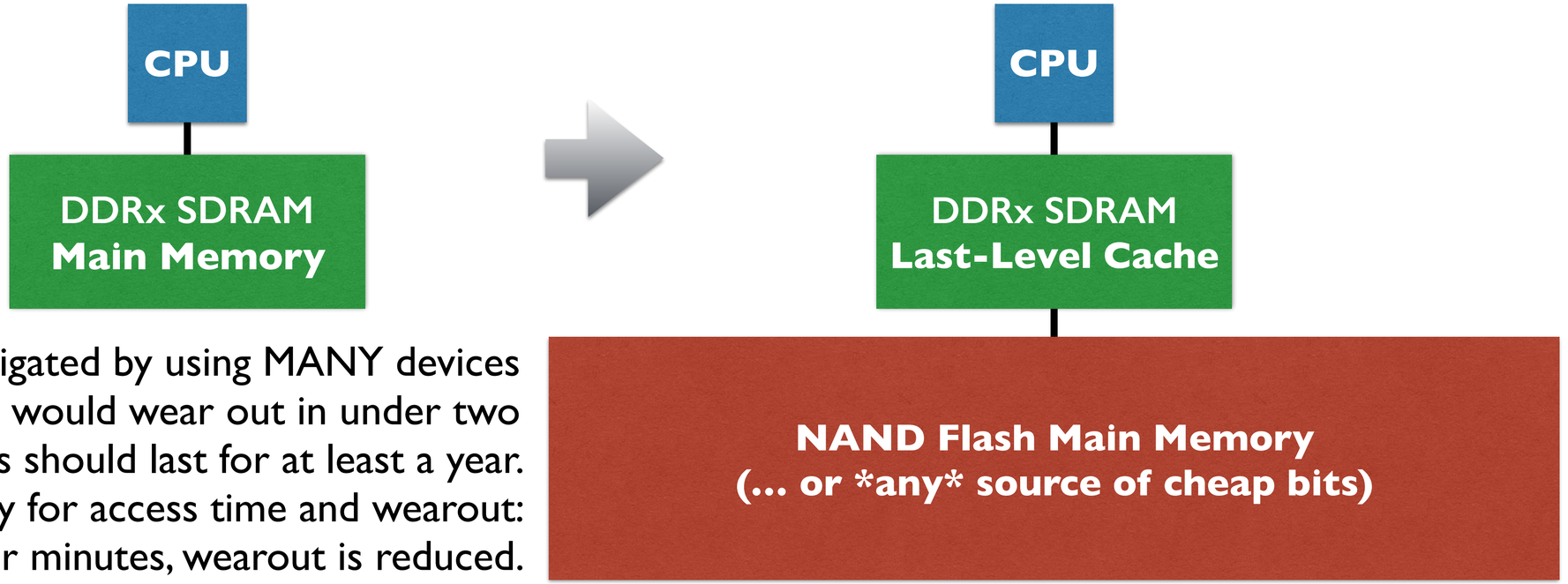
Performance Comparison

MEMSYS 2018



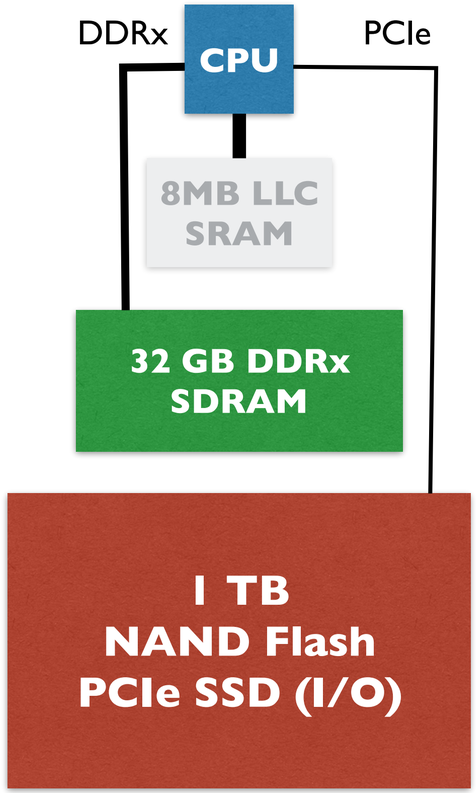
Non-Volatile Main Memory

	Cost for 10 GB	Size of 10 GB	Power for 10 GB	Power per GB/s
Off-Chip SRAM	$1,000	1 bucket	0.1–1 W	0.1 W
DDR4 SDRAM	$100	1 DIMM	1 W	0.1 W
NAND Flash	$10	<1 chip	0	0.1 W (?)
3D XPoint	$40	<1 chip	0	0.1 W (?)

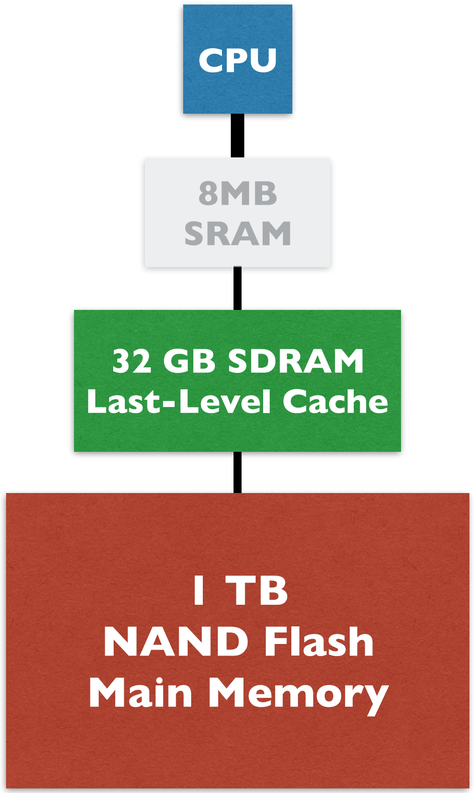


Note: wear-out mitigated by using MANY devices (thousands). A single device would wear out in under two days; therefore, 1000 devices should last for at least a year. Next, you can trade off longevity for access time and wearout: if the data need only last hours or minutes, wearout is reduced.

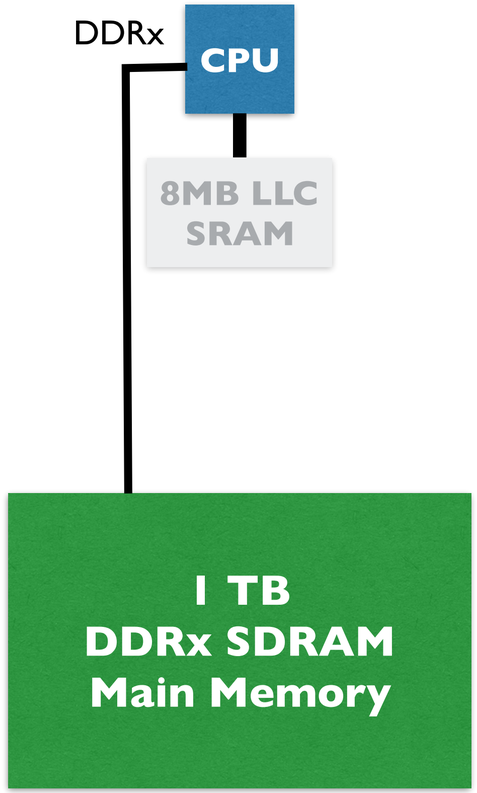
A Tale of 3 Memory Systems



SSD
$500 – 10W

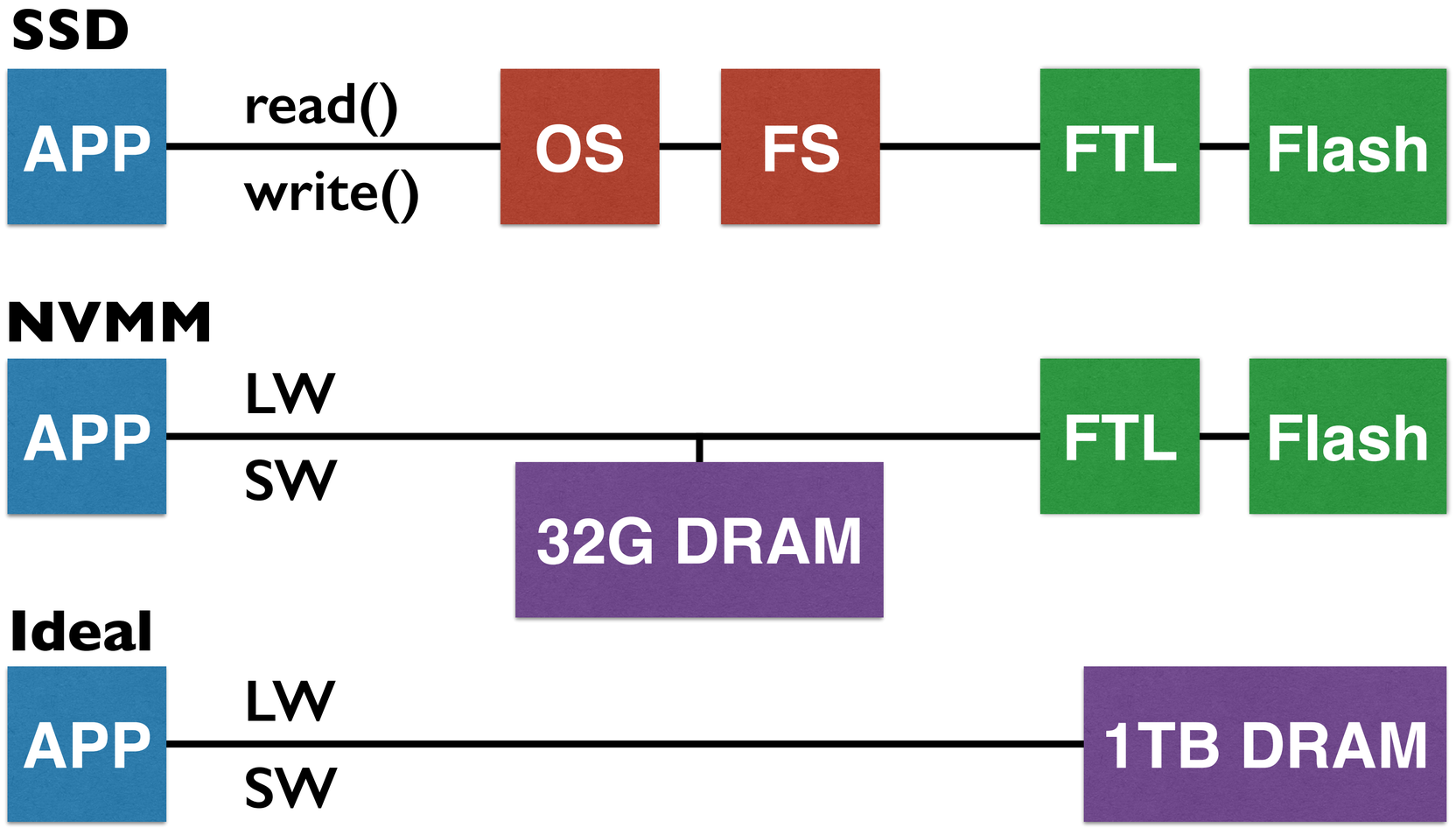


NVMM
$500 – 10s of W

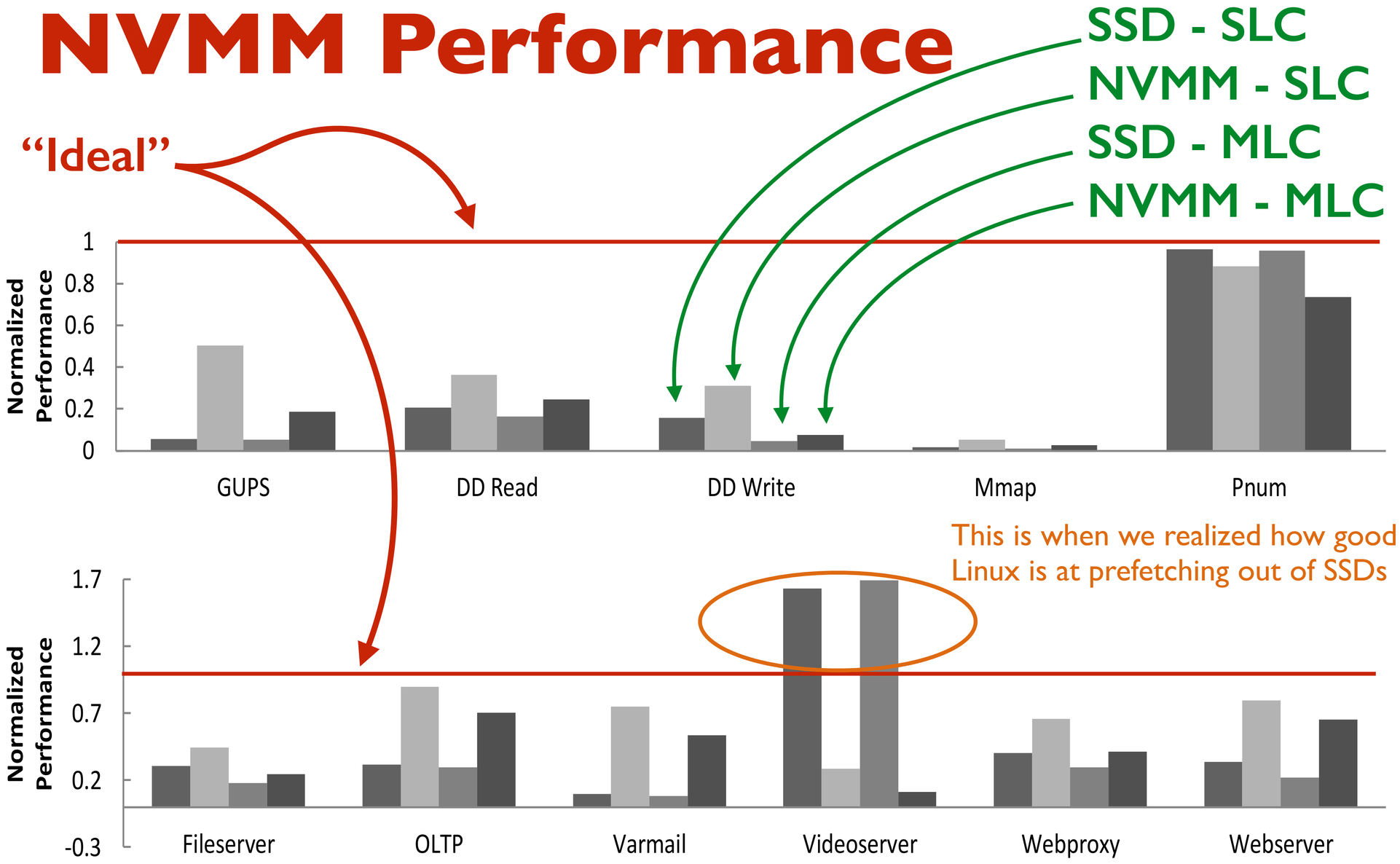


Ideal
$10,000 – 100W

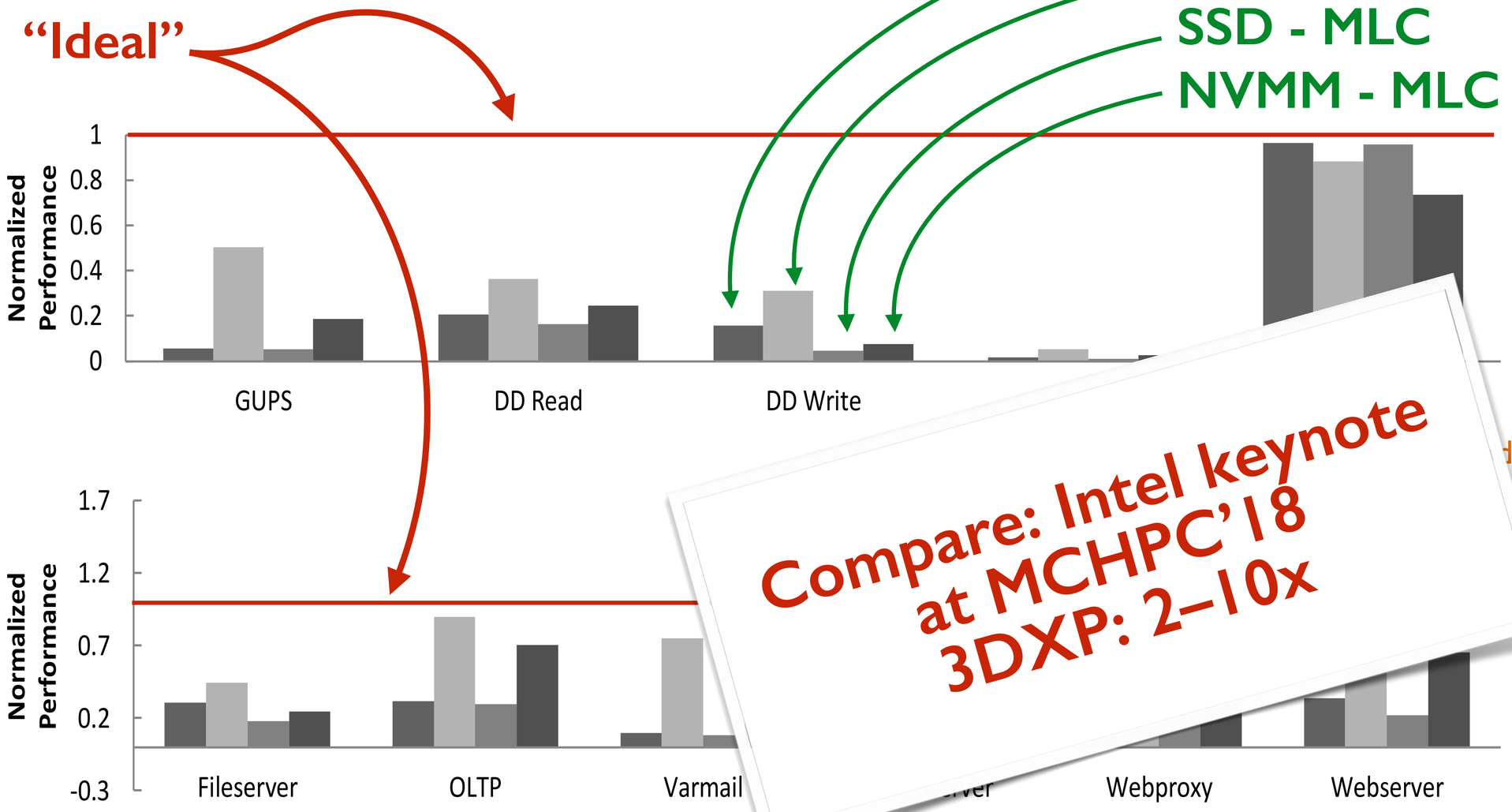
A Tale of 3 Memory Systems



NVMM Performance



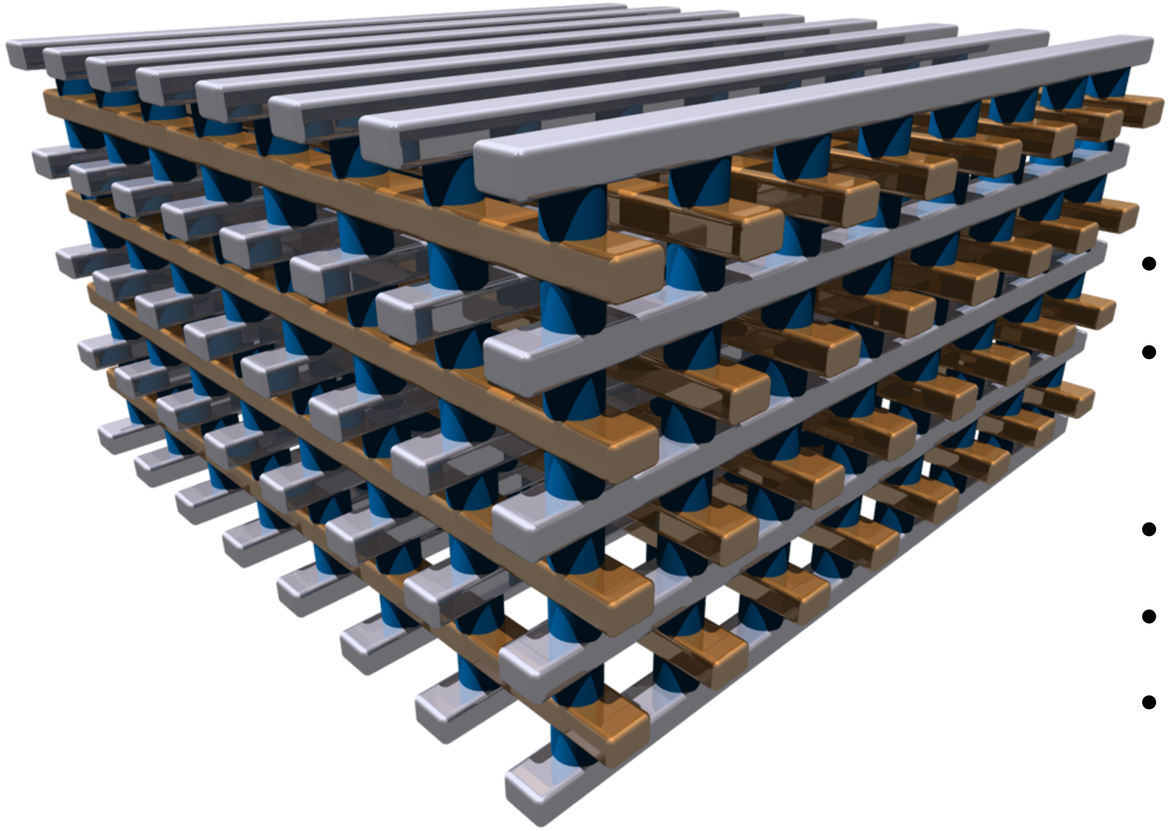
NVMM Performance

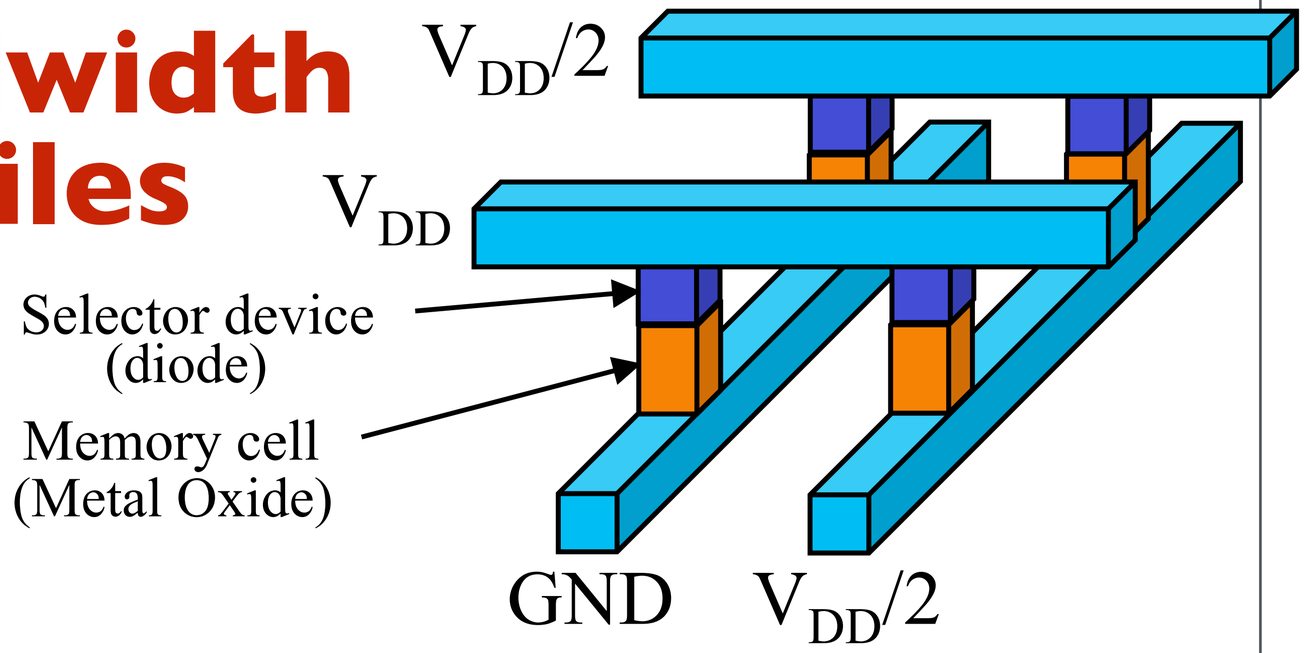


Compare: Intel keynote at MCHPC'18
3DXP: 2–10x

High Bandwidth Non Volatiles

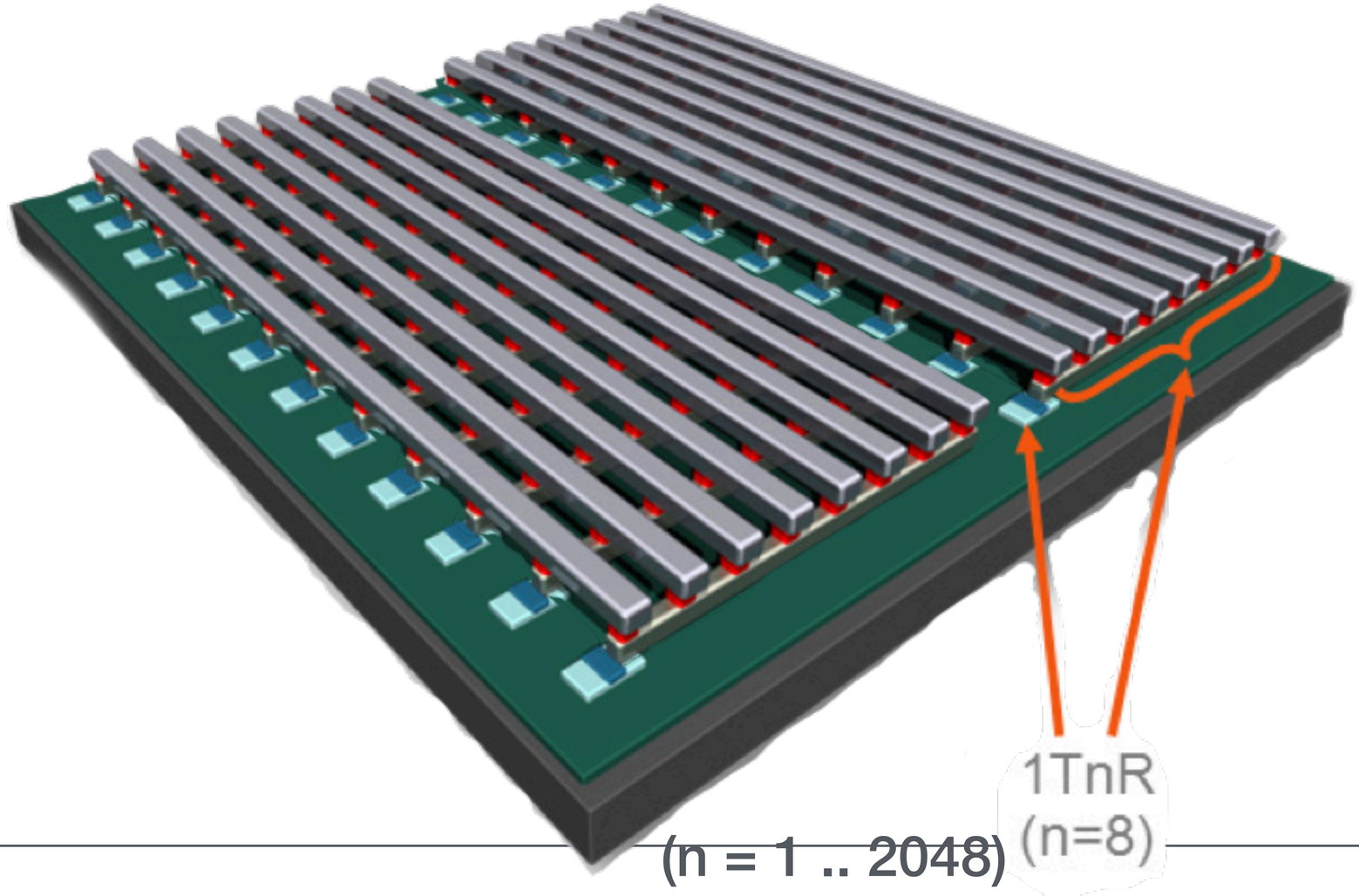
Crossbar ReRAM



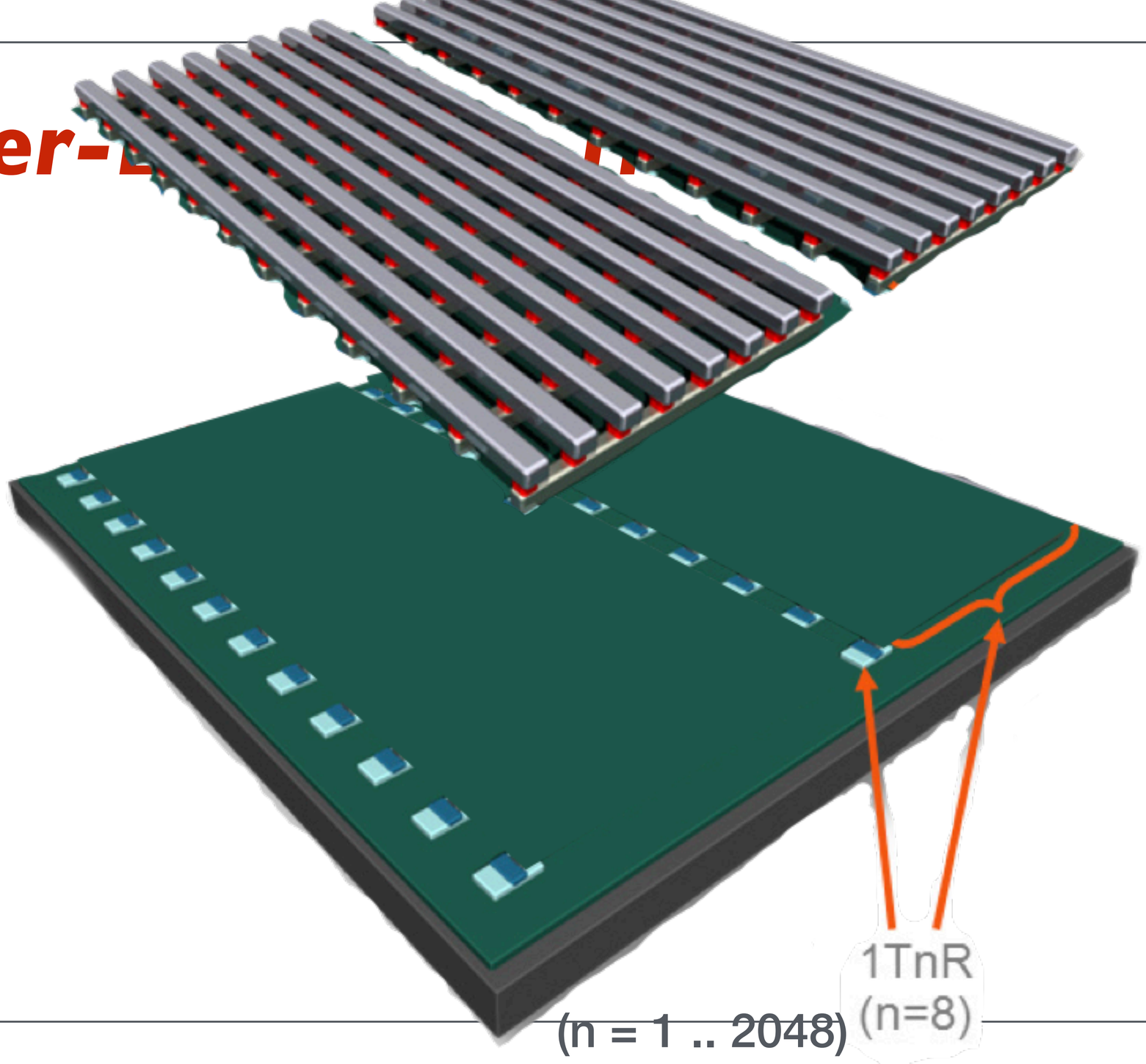


- Cells minimum area (no access transistor)
- **2-stack arrays** @ 16nm, 20 x 20 mm die: **64GB of ReRAM**
- **8-stack arrays => 256 GB of ReRAM**
- Stacks arbitrarily high
- No. Access. Transistor.

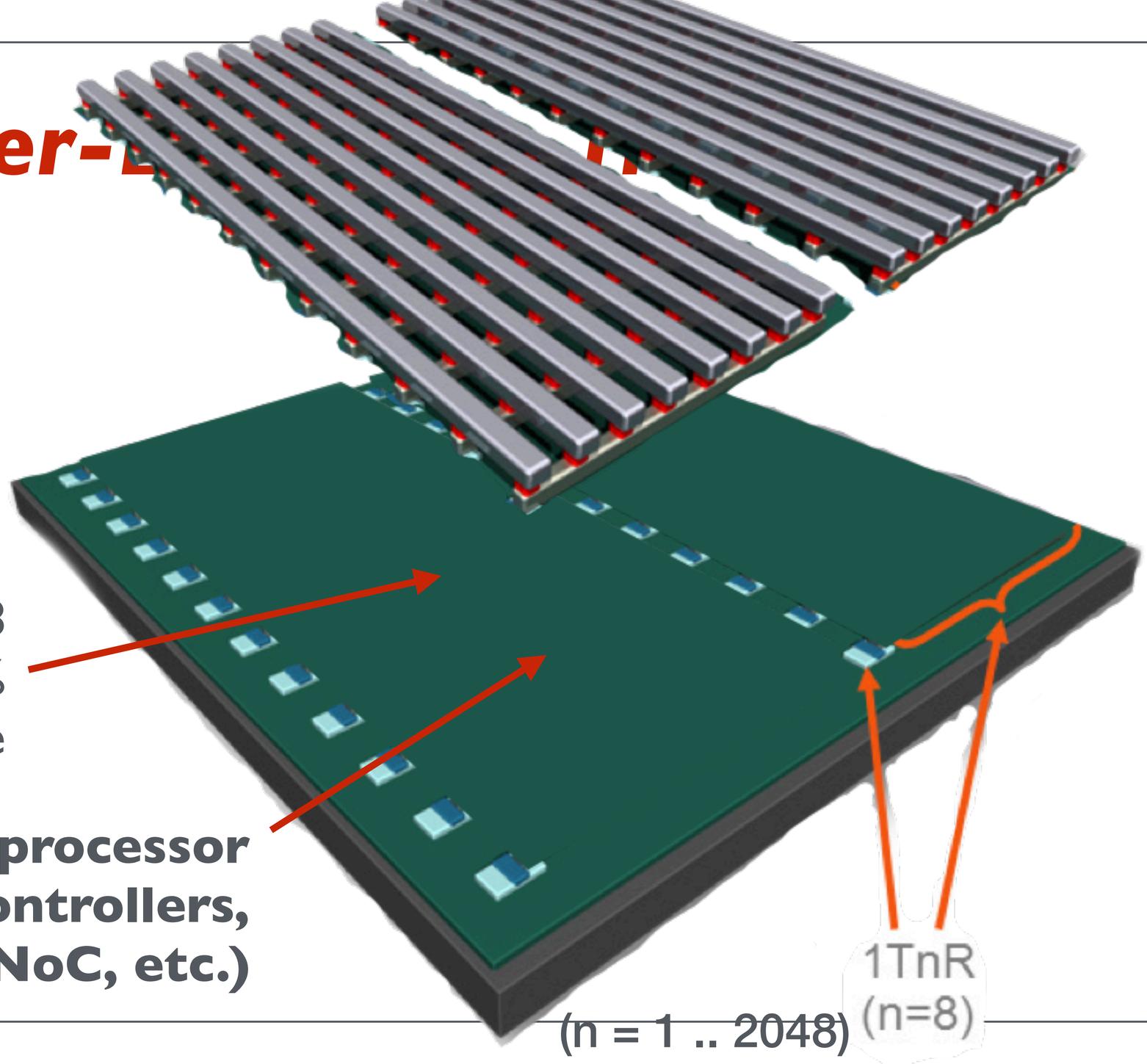
No Per-Bit Access Transistor



No Per-L



No Per-L

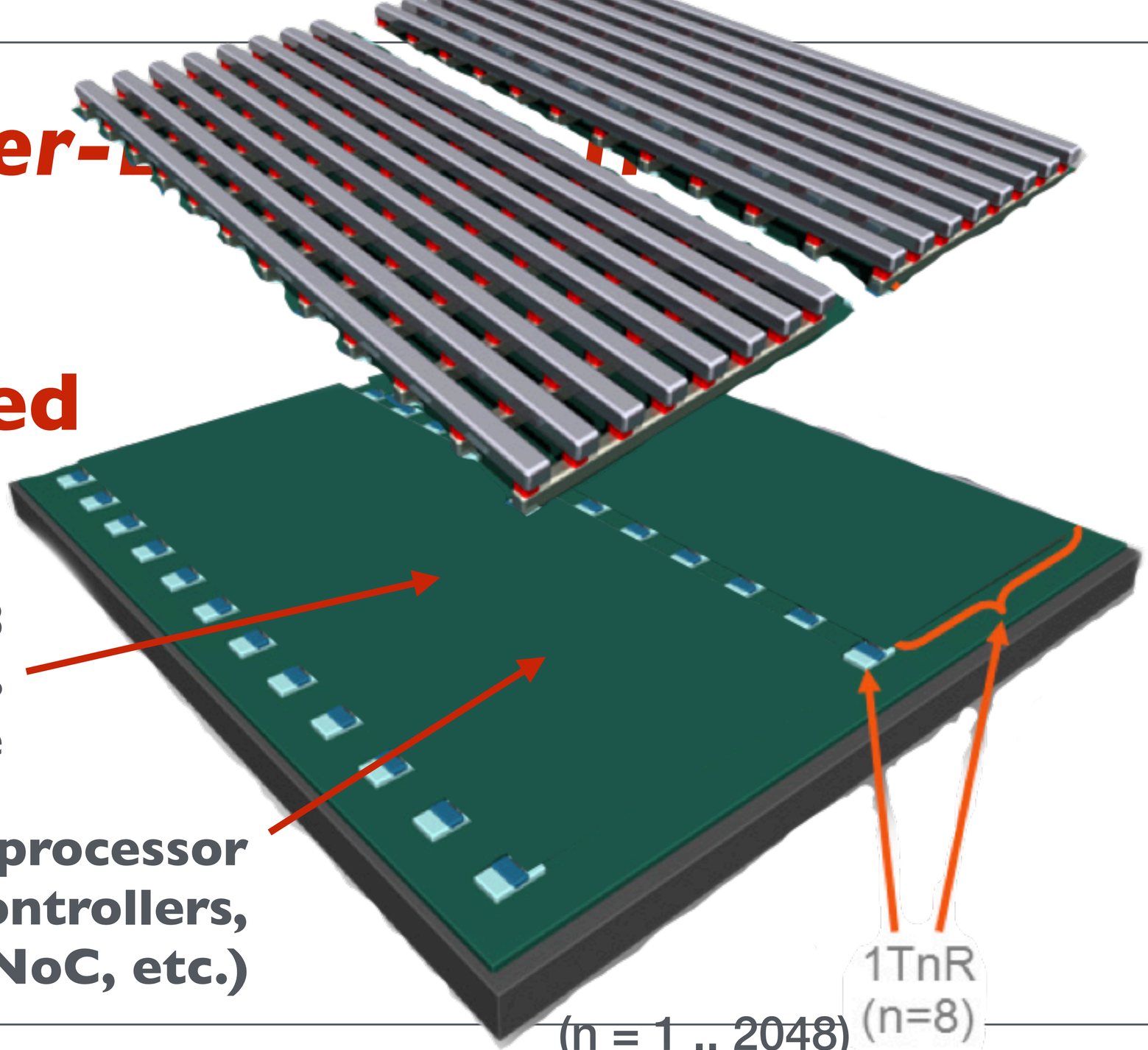


For n = 2048 area is ~75% white space

Use for processor (cores, controllers, routers, NoC, etc.)

No Per-L… …ll

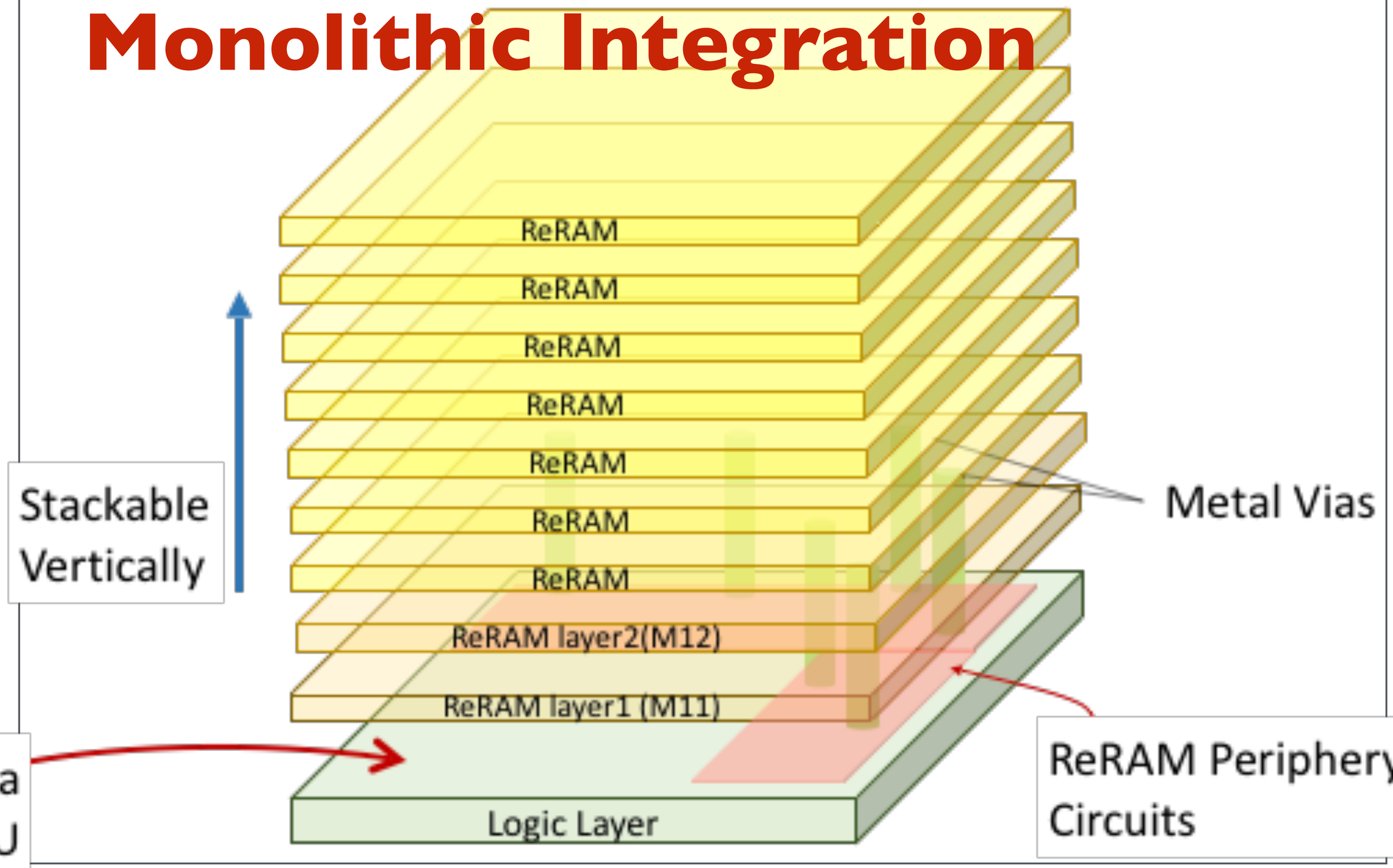
Monolithic
Not Die-Stacked



For n = 2048 area is ~75% white space

Use for processor (cores, controllers, routers, NoC, etc.)

Monolithic Integration



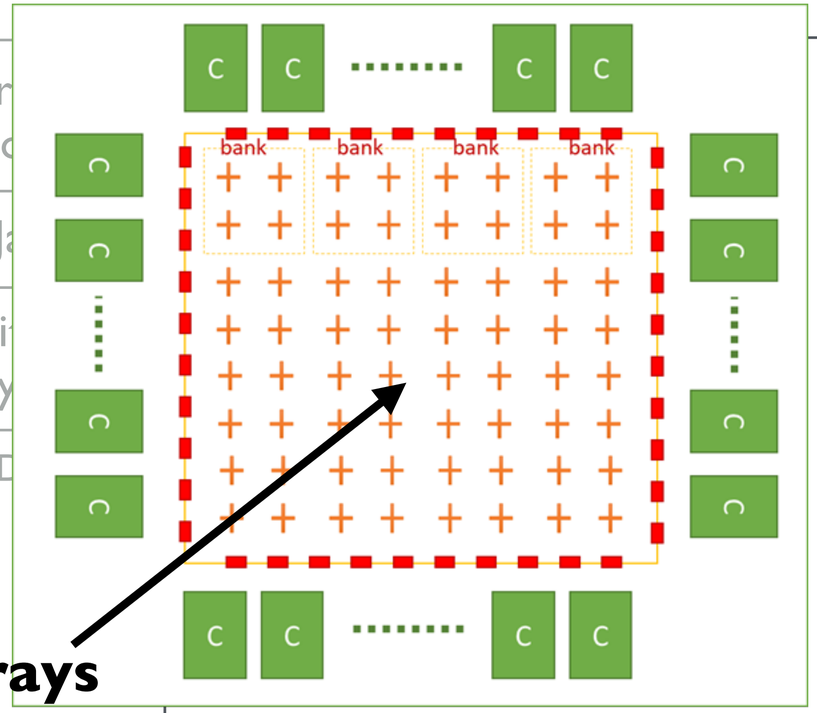
All Tomorr
Memc

Bruce Ja

Universi
Mary

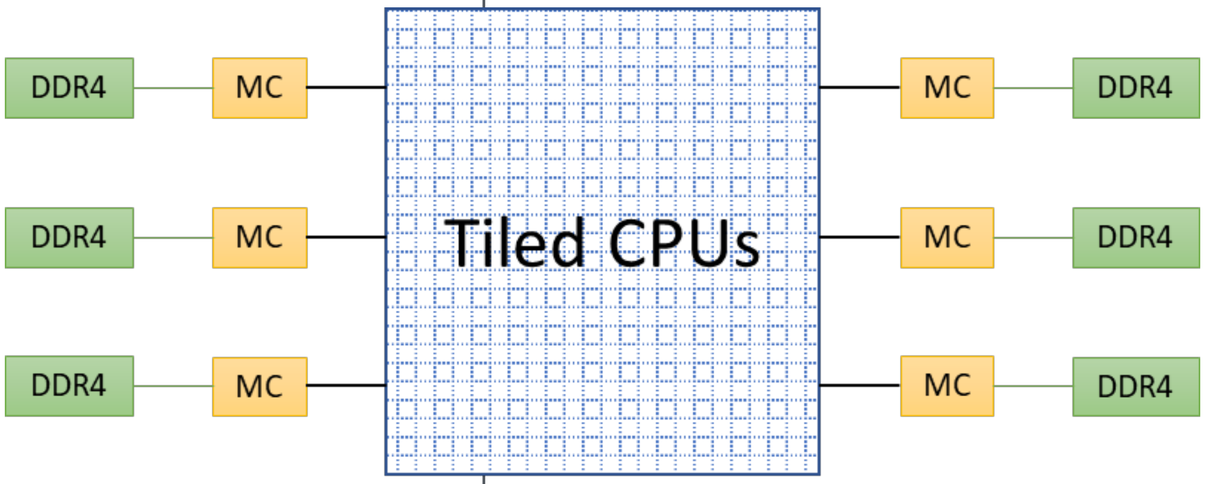
SLID

Monolithic Integration



ReRAM Arrays NoC-connected

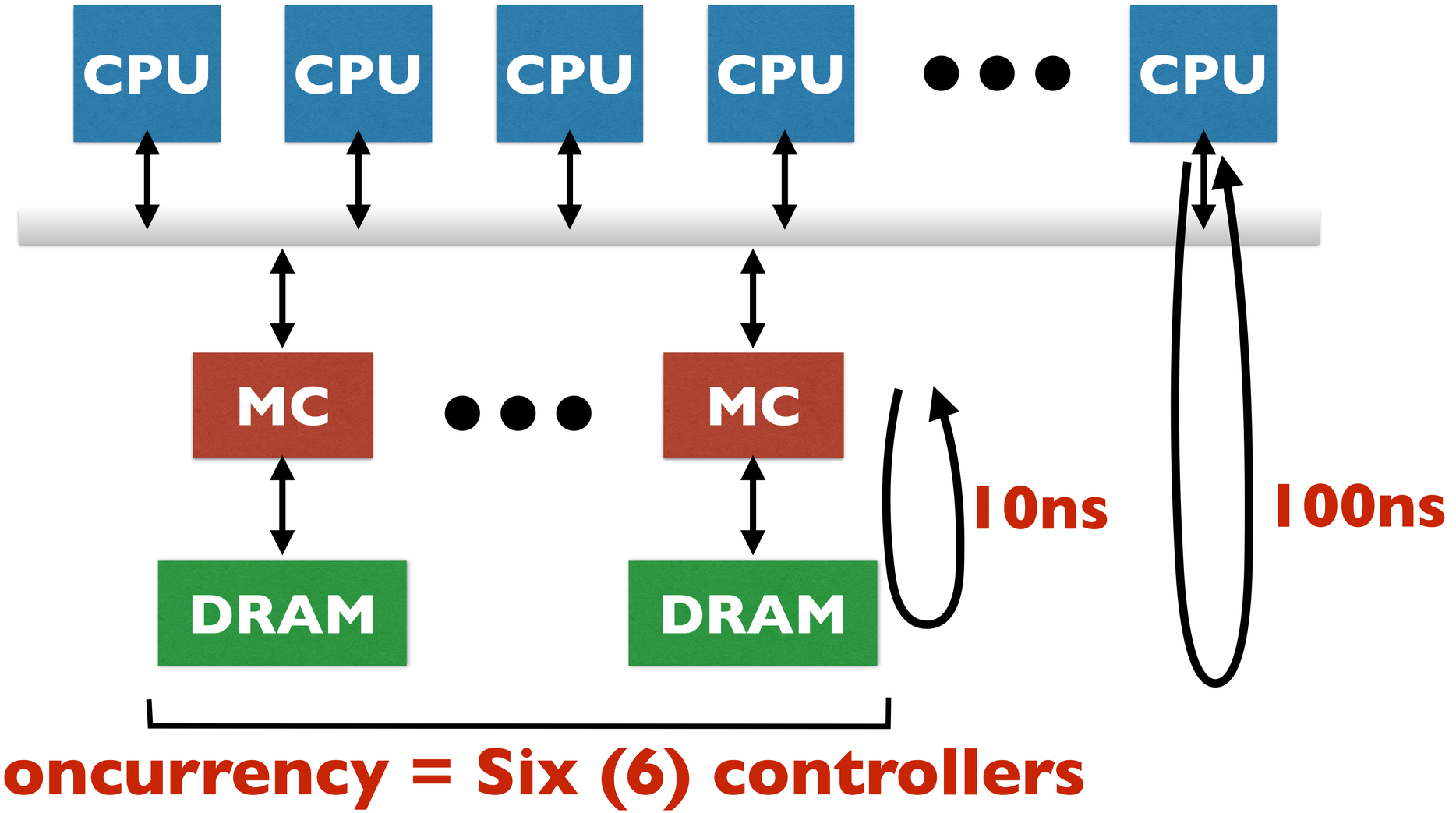
ReRAM Architecture



DRAM Architecture

	DRAM	ReRAM
# Mem Controllers	6	**1000**
L1 Cache size	32KB	32KB
L2 Cache size	1MB	1MB
Main Mem Parameters		
Mem Latency	~30ns DRAMsim3 simulated	SST Messier **Read: 200ns** **Write: 1us**
request_width (access granularity)	**64 Bytes** bus-width = 8B burst-length = 8	**8 Bytes** bus-width = 8B burst-length = 1
Topology	Mesh	Mesh

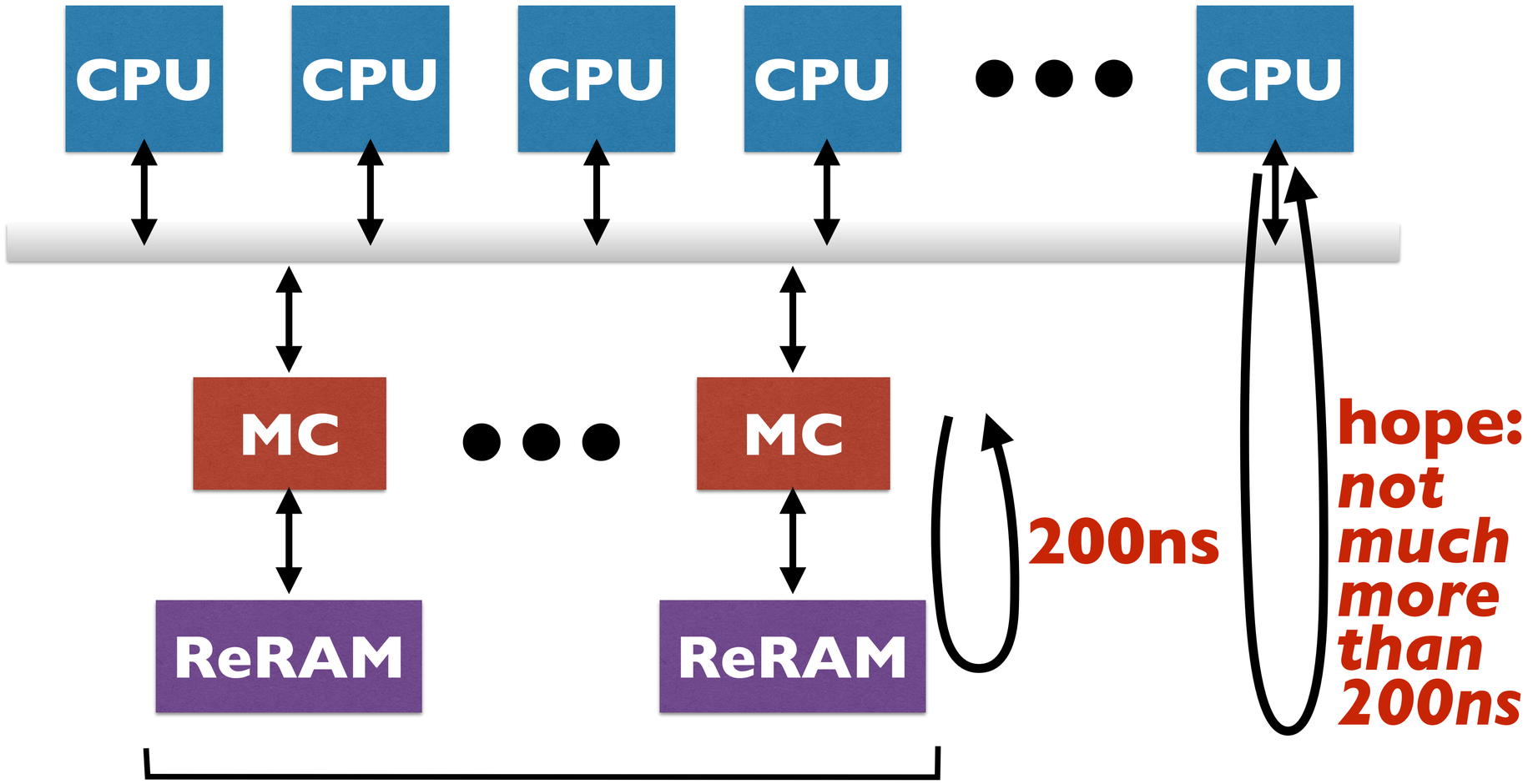
DRAM-Based System



cores

memory access points

Monolithic System

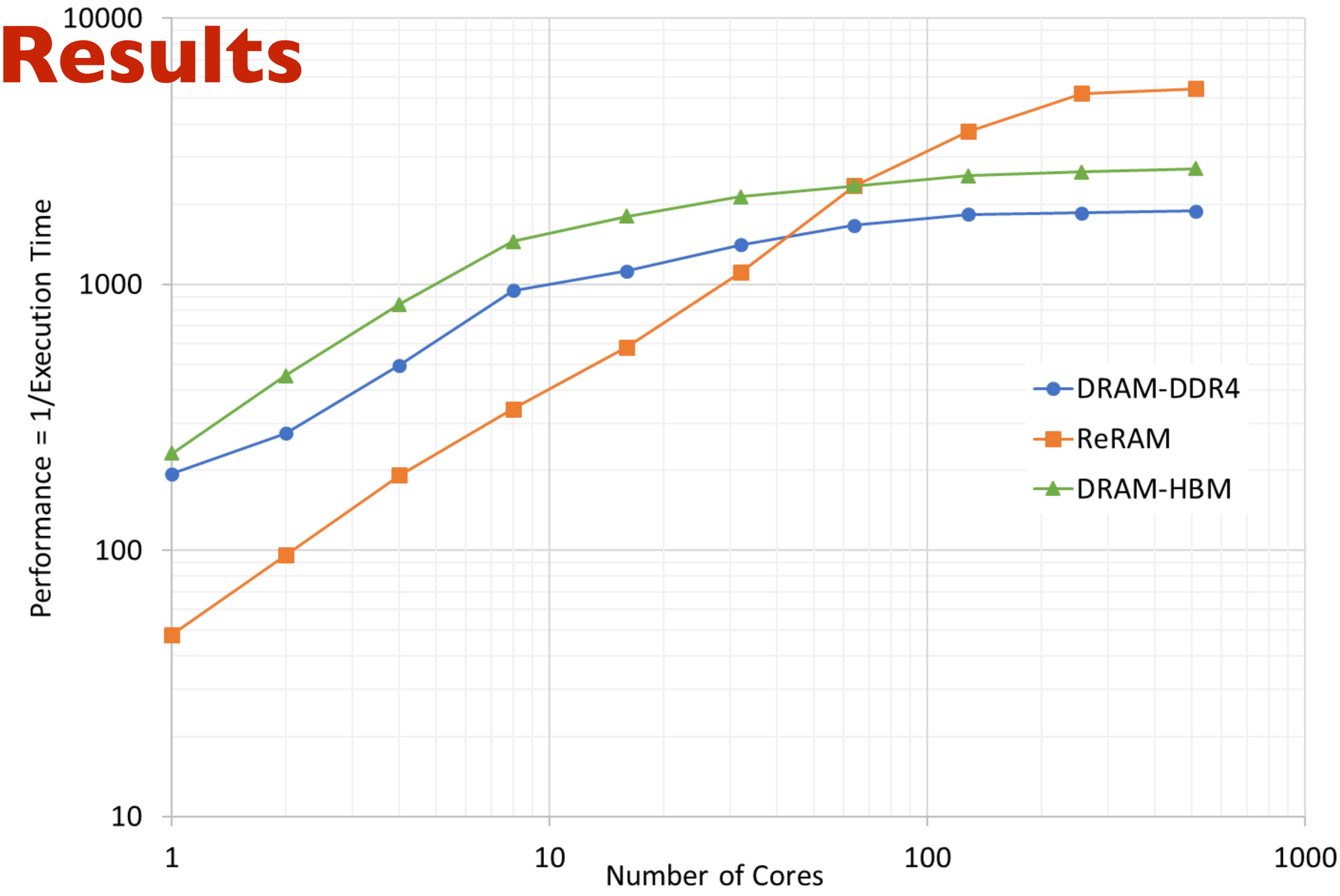


cores

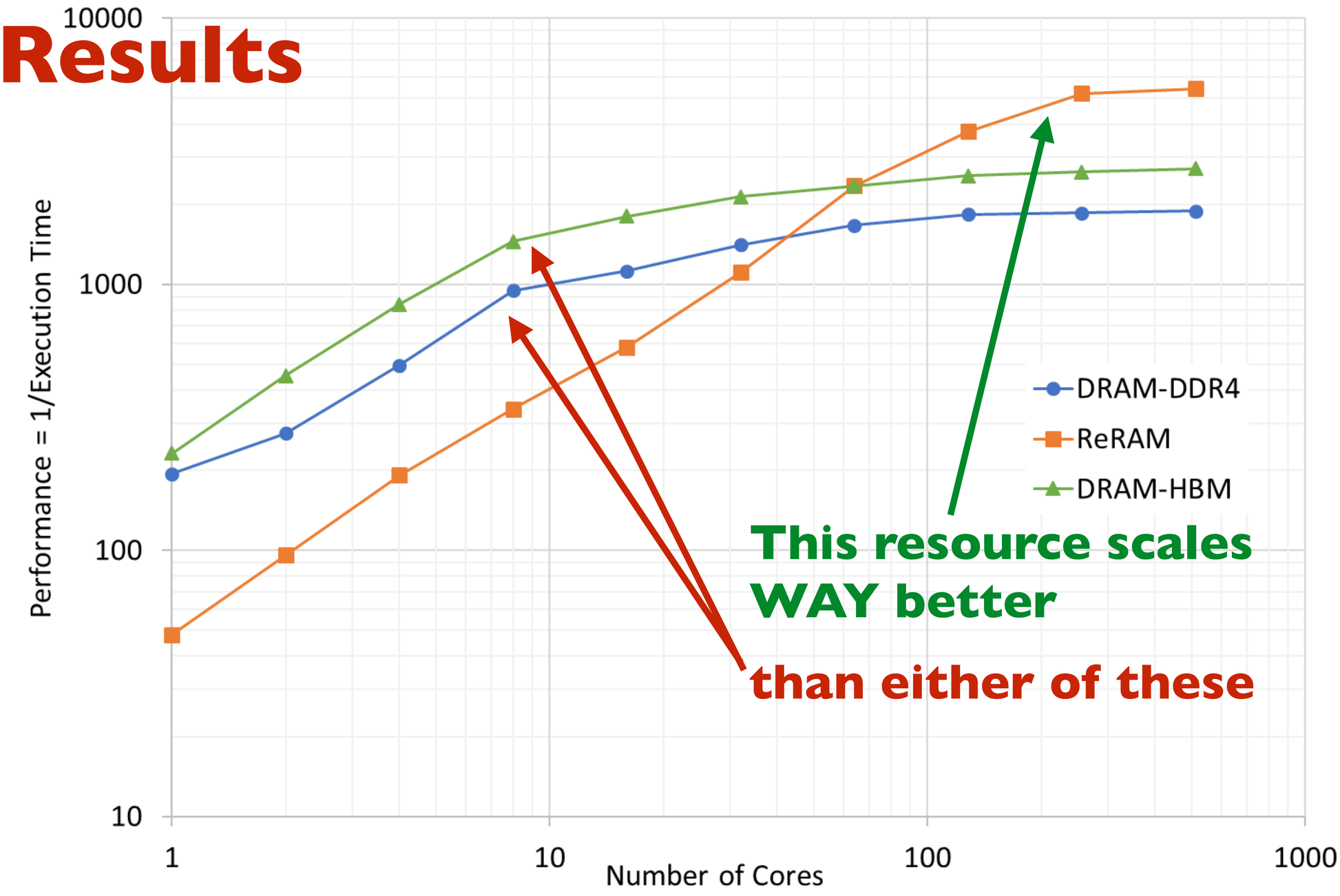
memory access points

Concurrency = 1000 controllers

Results


STREAM: DRAM vs ReRAM comparison
Performance = 1/Execution Time
10000
1000
100
10
1
10
100
1000
Number of Cores
DRAM-DDR4
ReRAM
DRAM-HBM

Results


STREAM: DRAM vs ReRAM comparison
Performance = 1/Execution Time
10000
1000
100
10
1
10
100
1000
Number of Cores
DRAM-DDR4
ReRAM
DRAM-HBM
This resource scales WAY better
than either of these

Bottom Line

The hardware is here now (mostly)

- 1–10TB main memories will be common
- TB/s off-chip bandwidths are here now
  The costs: power and capacity
- Lower-power solution: monolithic
  (~TB, 100s GB/s, 1000s concurrent ops)

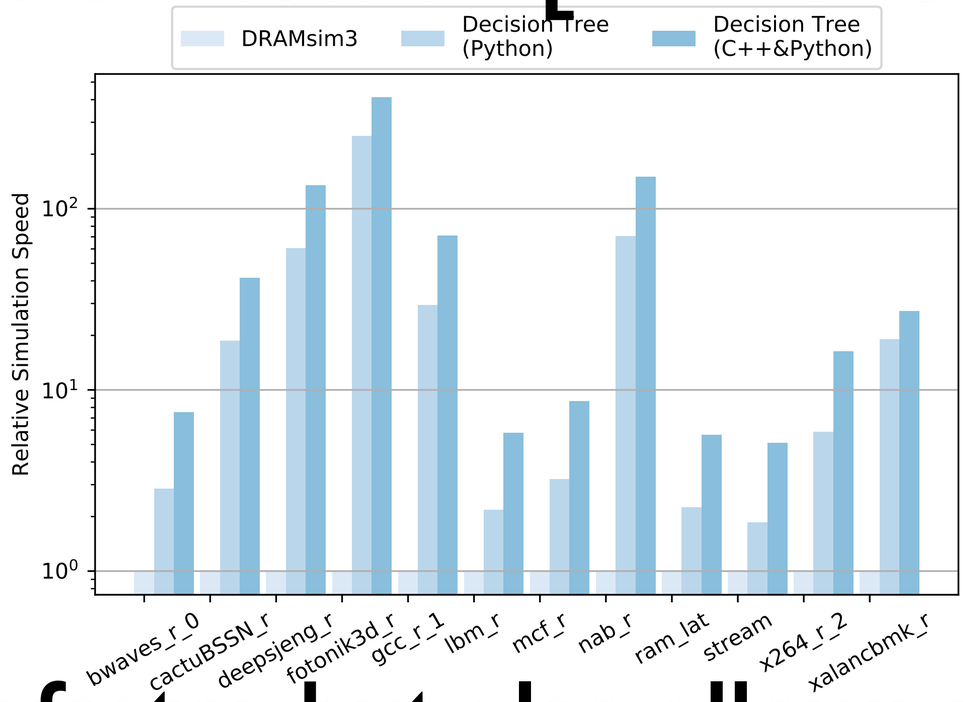
The software is waiting to happen

- Combined VM+FS subsystems
- Journaled main memory, etc.

Bottom Line

[next time] How to simulate it

- ML-Based simulation [MEMSYS 2019]

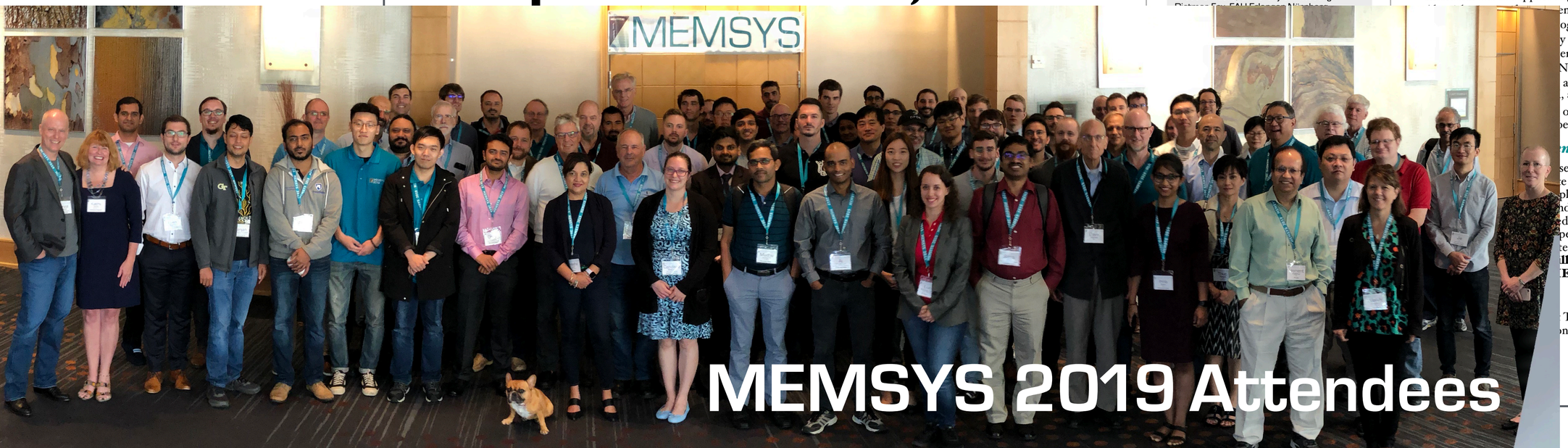


- Not only faster but also allows multiple simulations without need for sync'ing

Shameless Plug

www.memsys.io

Washington DC
Sep 28 – Oct 1, 2020



MEMSYS 2019 Attendees

Call For Papers www.memsys.io Call For Papers

MEMSYS 2020

The International Symposium on Memory Systems ❖ Sep 28–Oct 1, Washington DC

Important Dates

Submission: 29 May, 2019 (+7 days)*
Notification: 31 July, 2019
Camera-Ready: 14 August, 2019

...mission extension

...nats

...ers
...pers

...t, ACM 'sigconf'
...lind submission
16 pages long

...elists

Yitzhak Birk, Technion
Petar Radojkovic, BSC

Organizers

Bruce Jacob, University of Maryland
Kathy Smiley, Memory Systems

Ameen Akel, Micron
Abdel-Hameed Badawy, NMSU
Jonathan Beard, Arm
Bruce Childers, University of Pittsburgh

Memory-device manufacturing, memory-architecture design, and the use of memory technologies by application software all profoundly impact today's and tomorrow's computing systems, in terms of their performance, function, reliability, predictability, power dissipation, and cost. Existing memory technologies are seen as limiting in terms of power, capacity, and bandwidth. Emerging memory technologies offer the potential to overcome both technology- and design-related limitations to answer the requirements of many different applications. Our goal is to bring together researchers, practitioners, and others interested in this exciting and rapidly evolving field, to update each other on the latest state of the art, to exchange ideas, and to discuss future challenges. *Please visit memsys.io for more information.*

Areas of Interest

Previously unpublished papers containing significant novel ideas and technical results are solicited. Papers that focus on system, software, and architecture level concepts specifically memory-related, i.e. topics outside of traditional conference scopes, will be preferred over others (e.g., the desired focus is away from pipeline design, processor cache design, prefetching, data prediction, etc.). Symposium topics include, but are not limited to, the following:

- Memory-system design from both hardware and software perspectives
- Memory failure modes and mitigation strategies
- Memory and system security issues
- Memory for embedded and autonomous systems (e.g., automotive)
- Operating system design for hybrid/nonvolatile memories
- Technologies including flash, DRAM, STT-MRAM, 3DXP, etc.
- Memory-centric programming models, languages, optimization
- Compute-in-memory and compute-near-memory technologies
- Data-movement issues and mitigation techniques
- Interconnects to support large-scale data movement
- ...emory-management techniques
- ...ogies, their controllers, and novel uses
- ...y level across datacenter applications
- ...eration of large-memory machines
- ...NoSQL stores
- ... and memory technologies to support them, ..., and heterogeneous memories
- ... on topics *outside* the scope of traditional ...e preferred over others.

...ntations

...se interesting ideas that will spark ...e groups—to get applications ...ple, system architecture ...d circuits people to talk to ...d abstracts, position ...pers, and each accepted ...e presentation time
...ll be published in the ...EEE Xplore.

... The Westin ...n VA.

acm

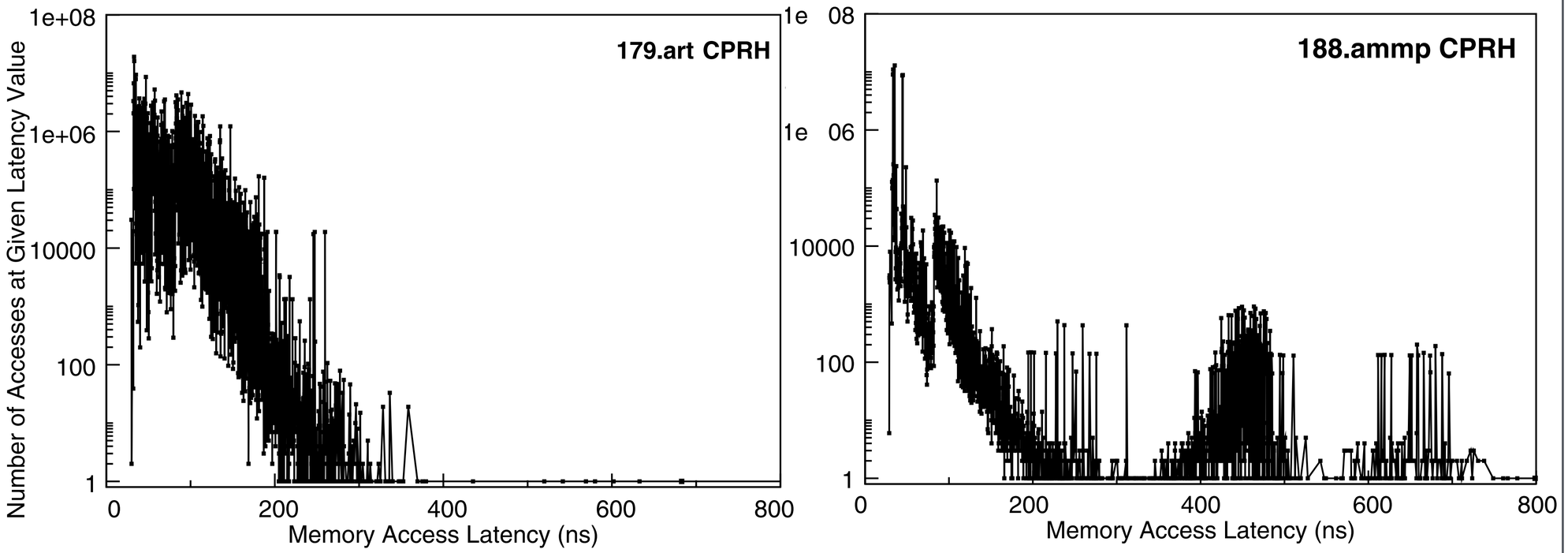
IEEE

Thank You!

Bruce Jacob

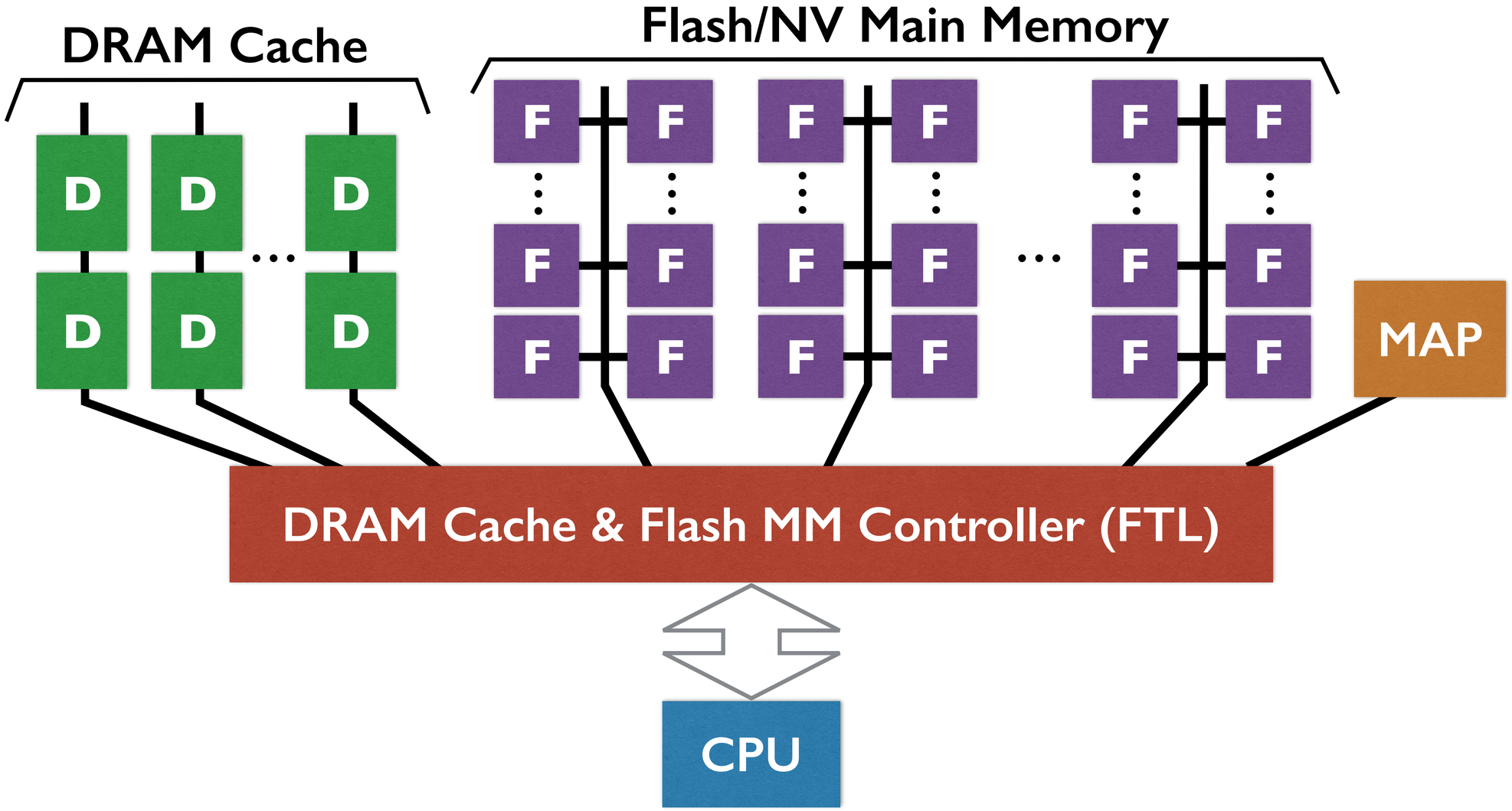
blj@umd.edu
www.ece.umd.edu/~blj

Perspective: DRAM Latency in Real Systems Is Quite Long



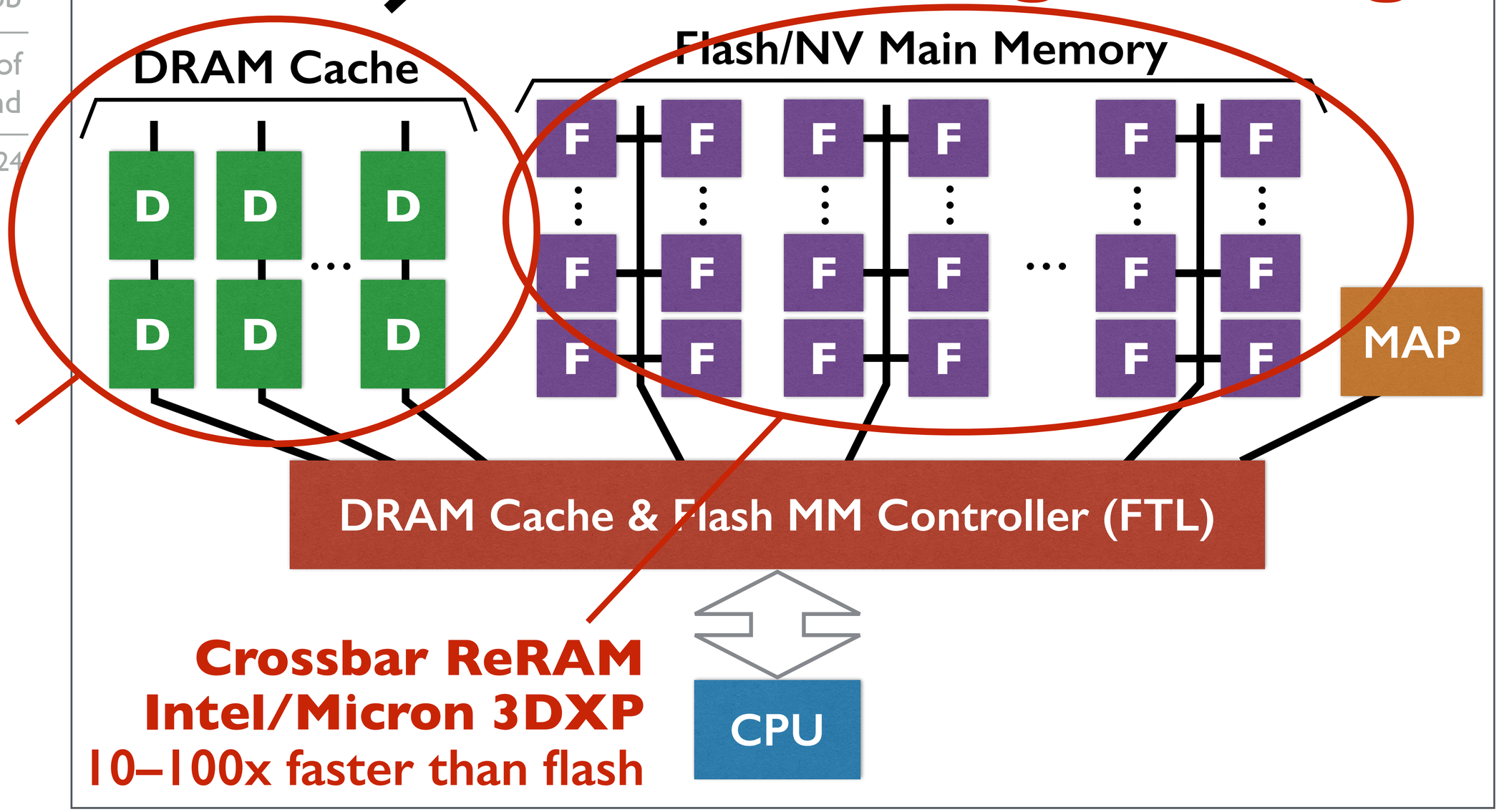
This is for single core. Multicore can be much, much worse.

Yeah, it's a lot of engineering



it ~~was~~

Yeah, ~~it's~~ a lot of engineering



HMC:
320GB/s
16 channels
HBM:
256GB/s
8 channels

Crossbar ReRAM
Intel/Micron 3DXP
10–100x faster than flash