All Tomorrow's Memories

Bruce Jacob

Keystone Professor
University of Maryland

Stealth Revolution

Stealth Revolution

Stealth Revolution

Fusion-ish SSD

Stealth Revolution

Fusion-ish SSD

Stealth Revolution

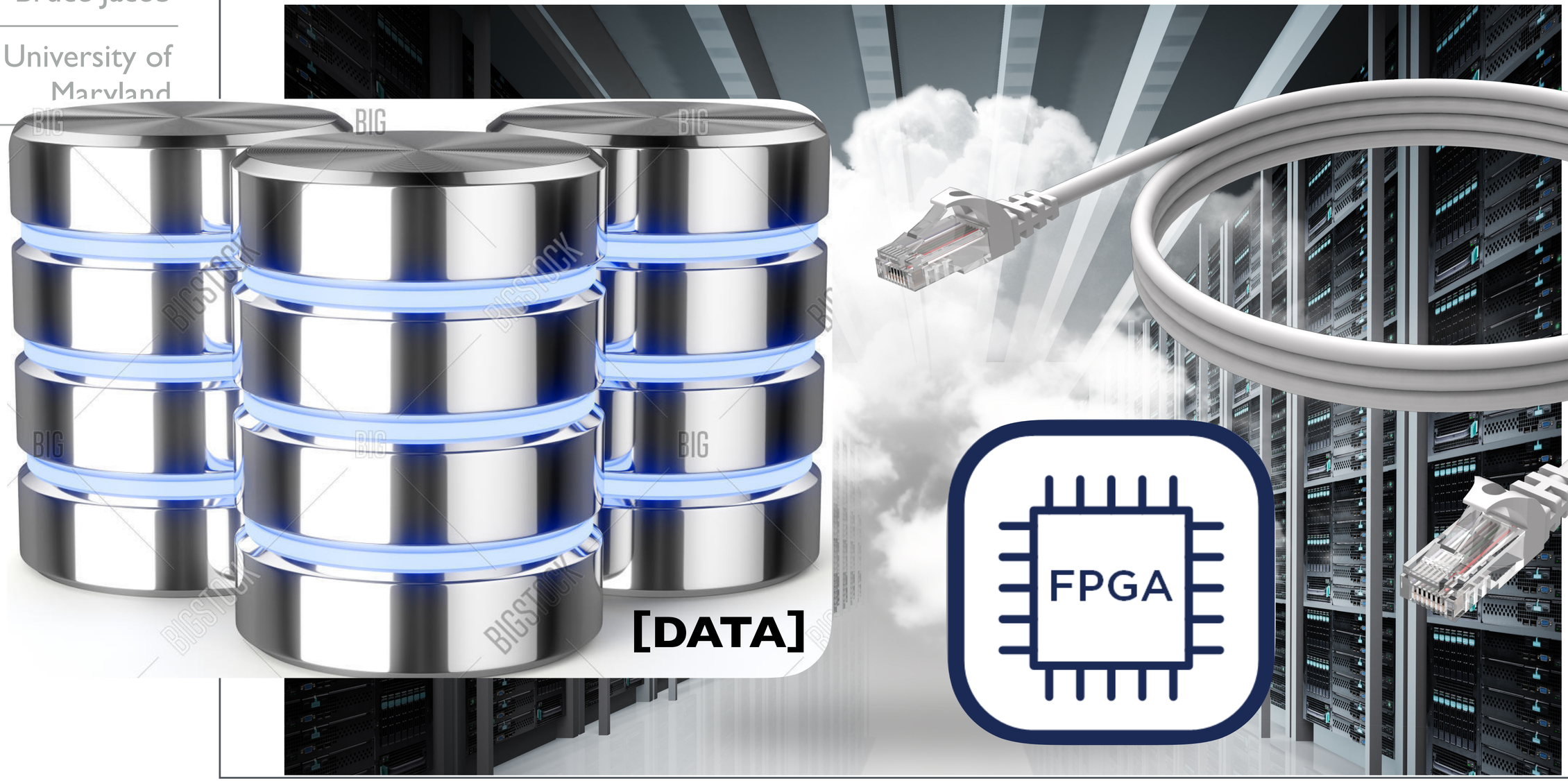
Stealth Revolution

Stealth Revolution

Bruce Jacob

University of Maryland

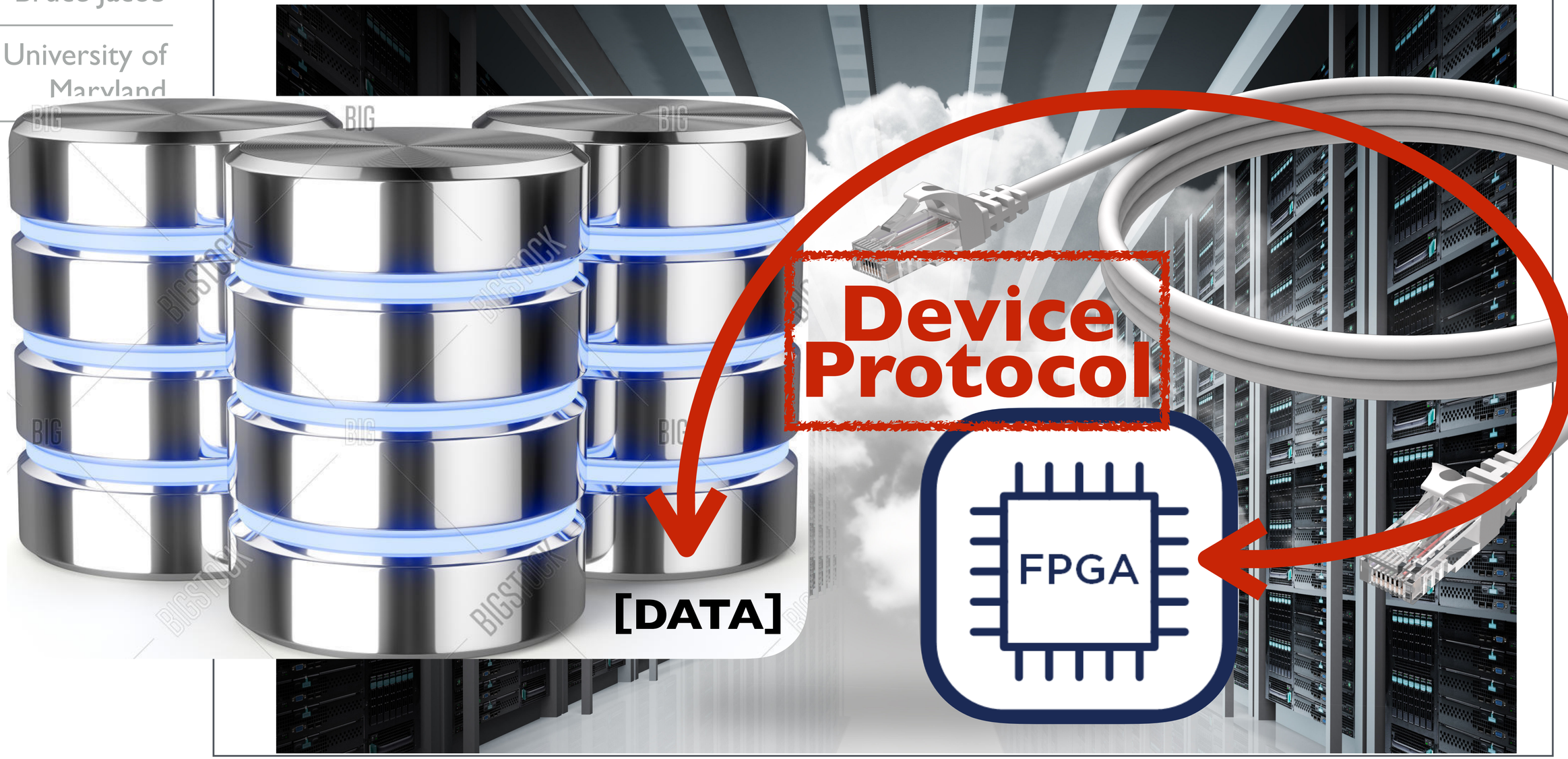
Stealth Revolution



Bruce Jacob

University of Maryland

Stealth Revolution



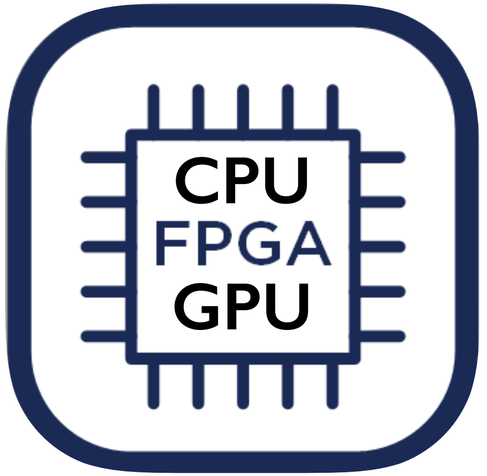
Stealth Revolution

Question: where is data?

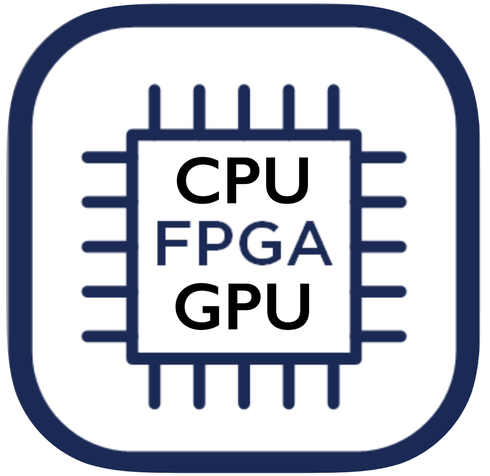
Answer: <data owner>

To data owner: read X

To requester: mem[X]



Stealth Revolution



Question: where is data?

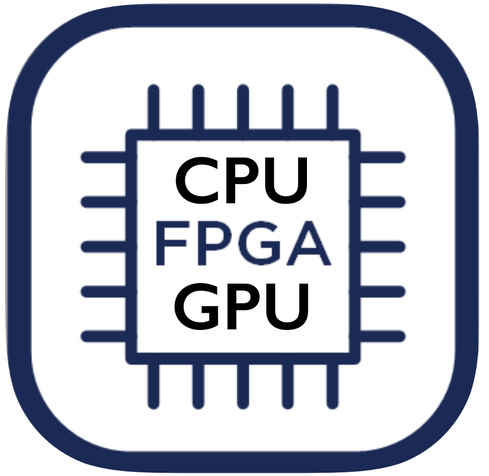
Answer: <data owner>

To data owner: read X

To requester: mem[X]

Major implications for OS and applications

Stealth Revolution



Question: where is data?

Answer: <data owner>

To data owner: read X

To requester: mem[X]

(esp. considering Fusion-io like capabilities)

Major implications for OS and applications

Background: Wish List

Fine-Grained Access

Bandwidth

DRAM - HBM/HMC*

Capacity

Low Power

Nonvolatility

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

HBNV*

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

Background: Wish List

Fine-Grained Access

Bandwidth

DRAM - HBM/HMC*

Capacity

Low Power

Nonvolatility

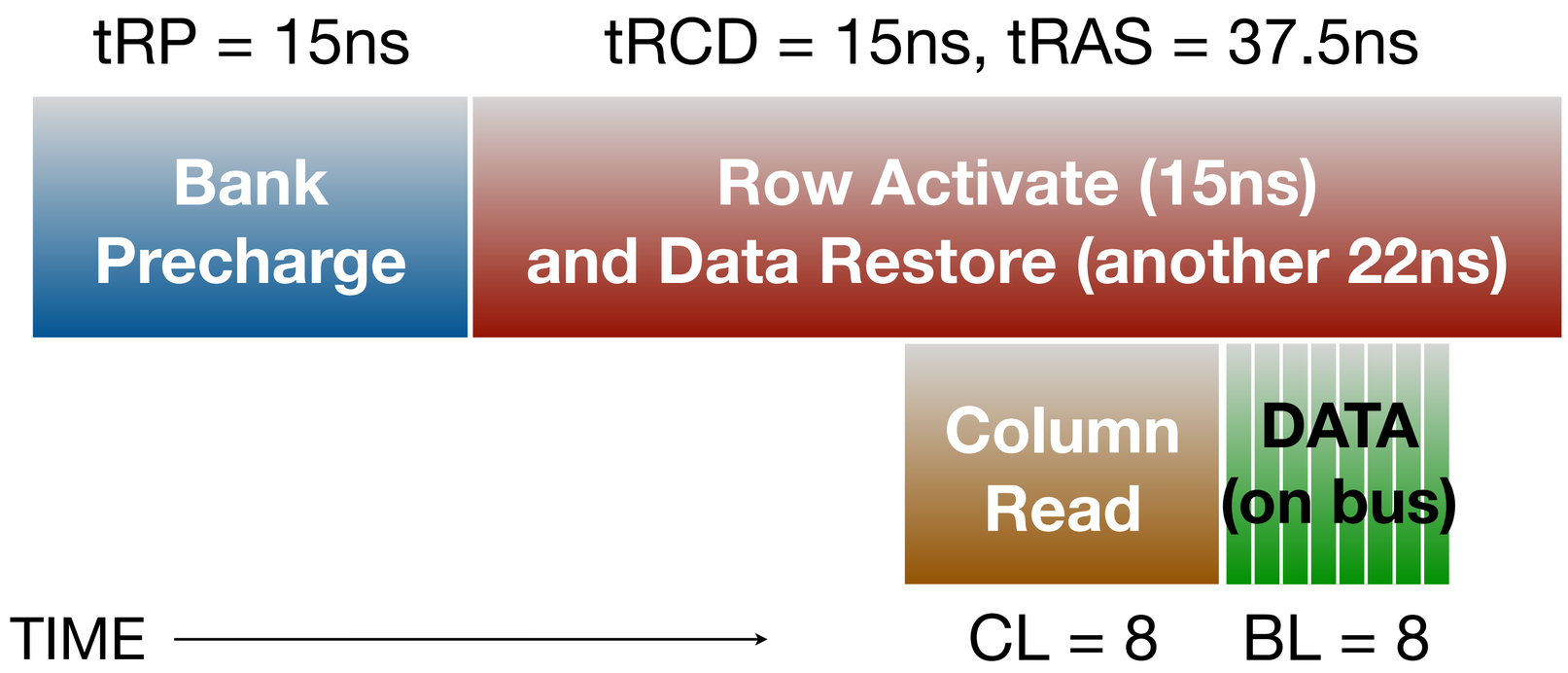
Flash, 3DXP, RRAM, PCM, etc - NVMM*

HBNV*

Major implications for OS* and applications

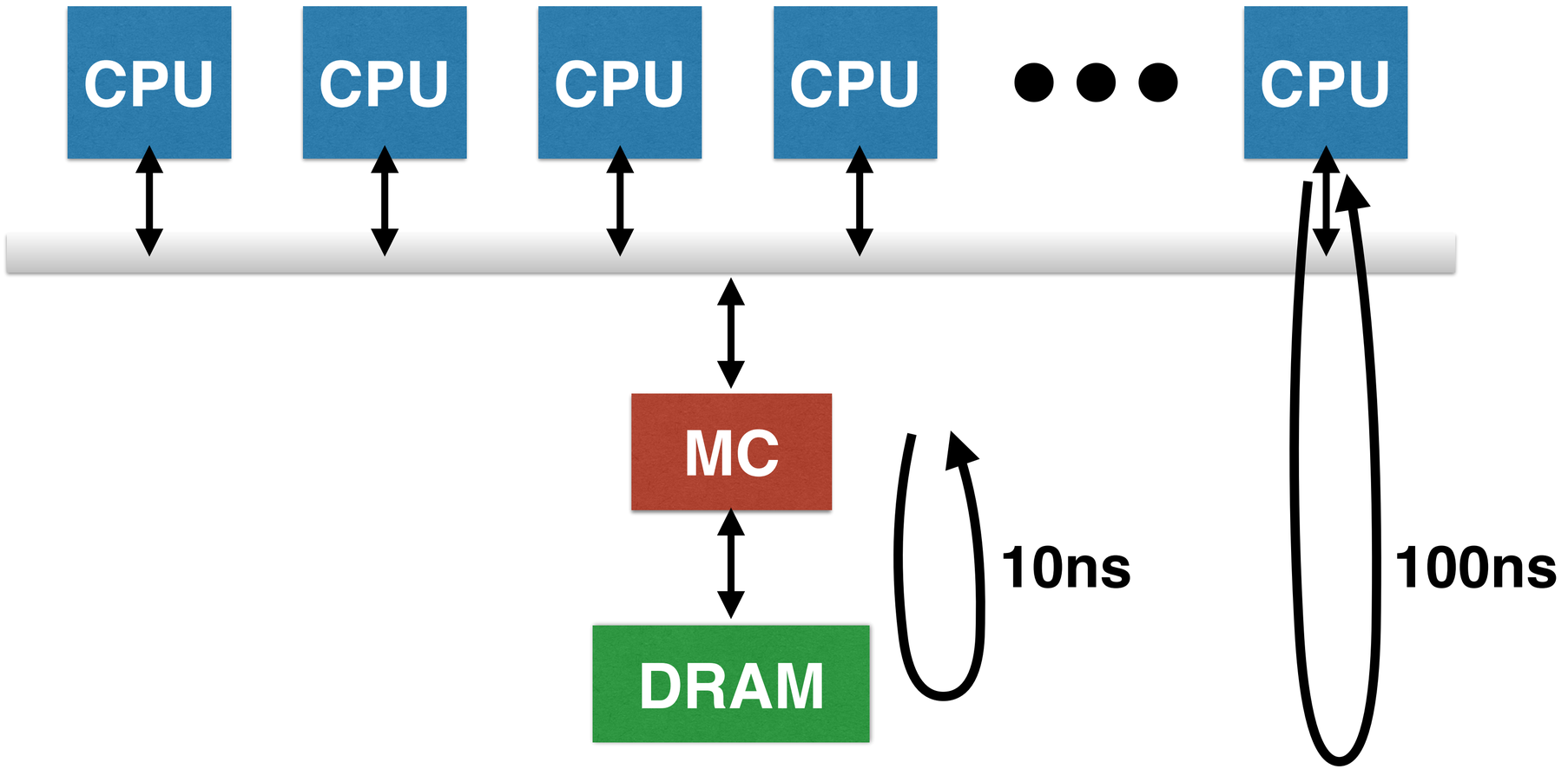
* Things we did and/or are doing now (I'll cover in talk)

Background: Memory Latency



Cost of access is high; requires **significant effort** to amortize this over the (increasingly short) payoff.

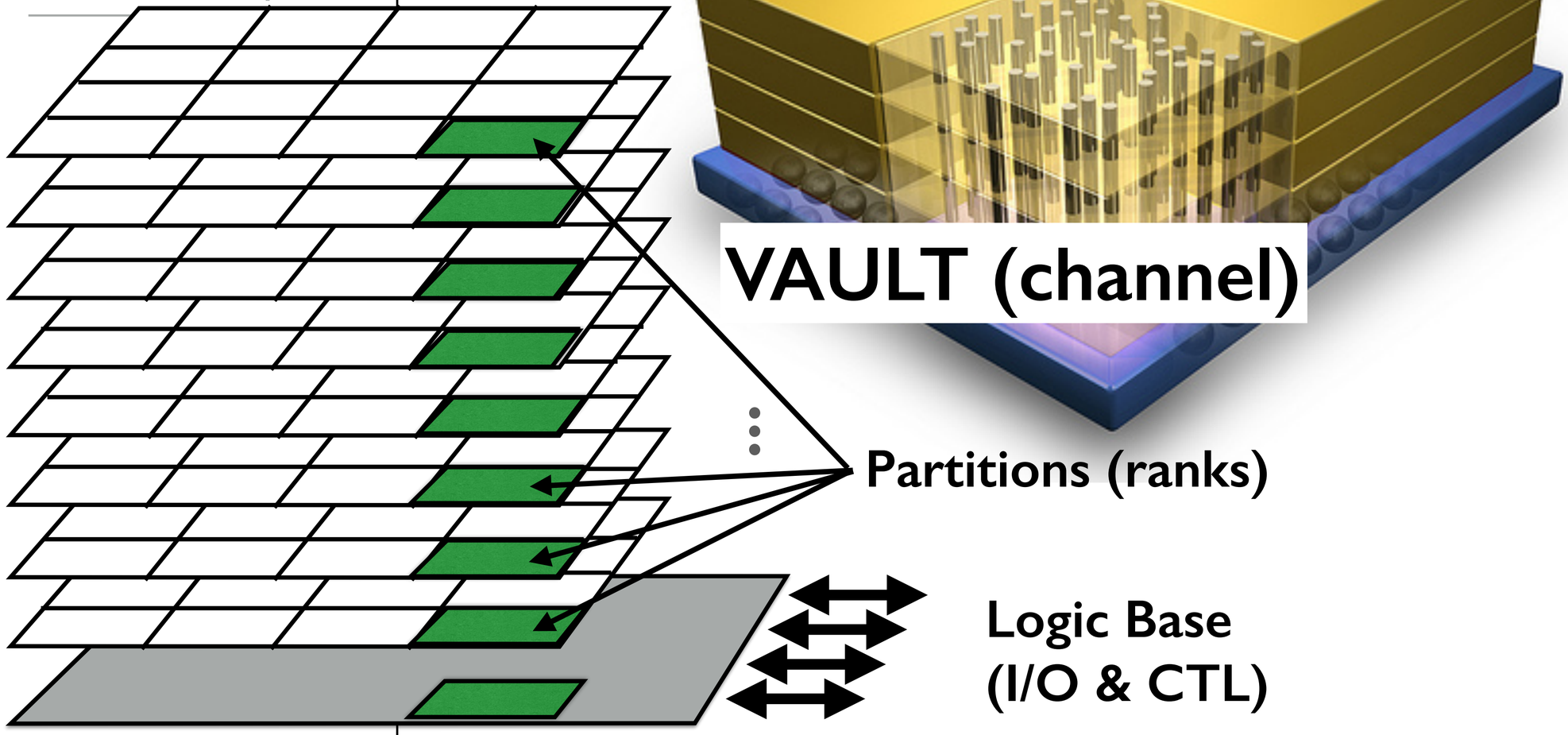
Background: Memory Latency



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)

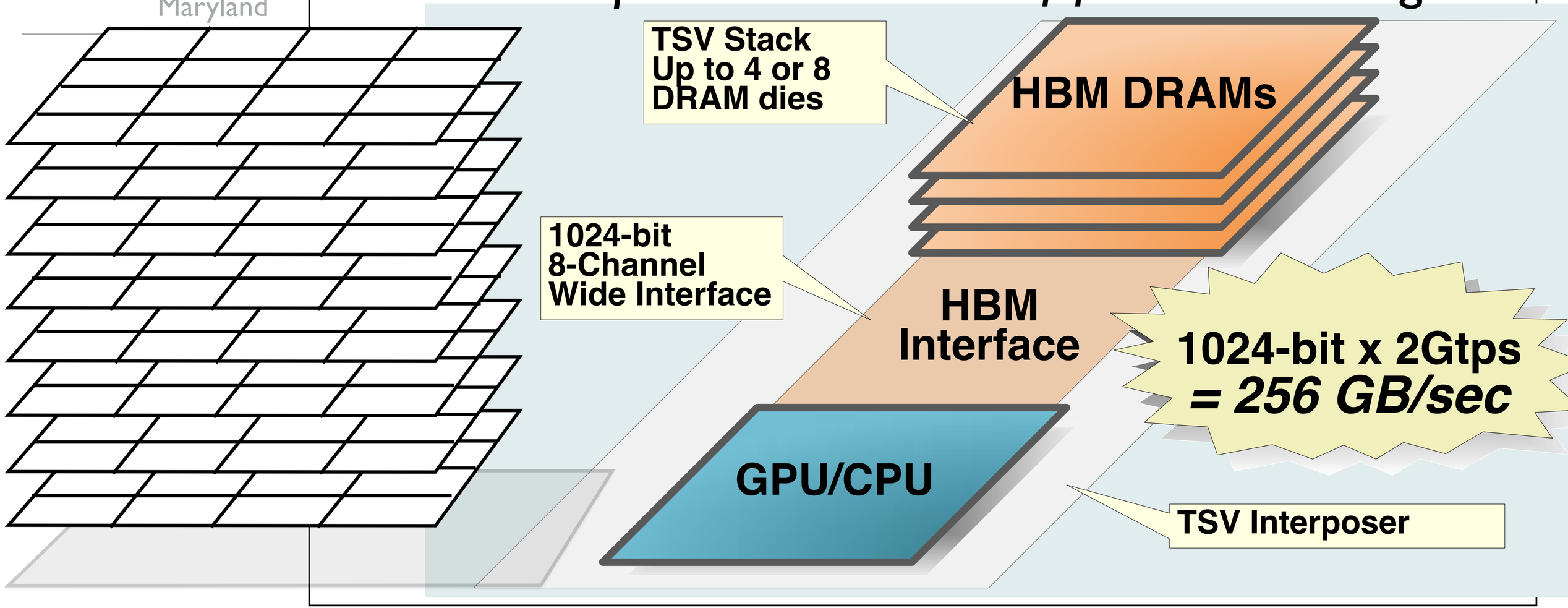
Total conc'y = **16 x 8 x 2..8** (256–1024)

Bruce Jacob

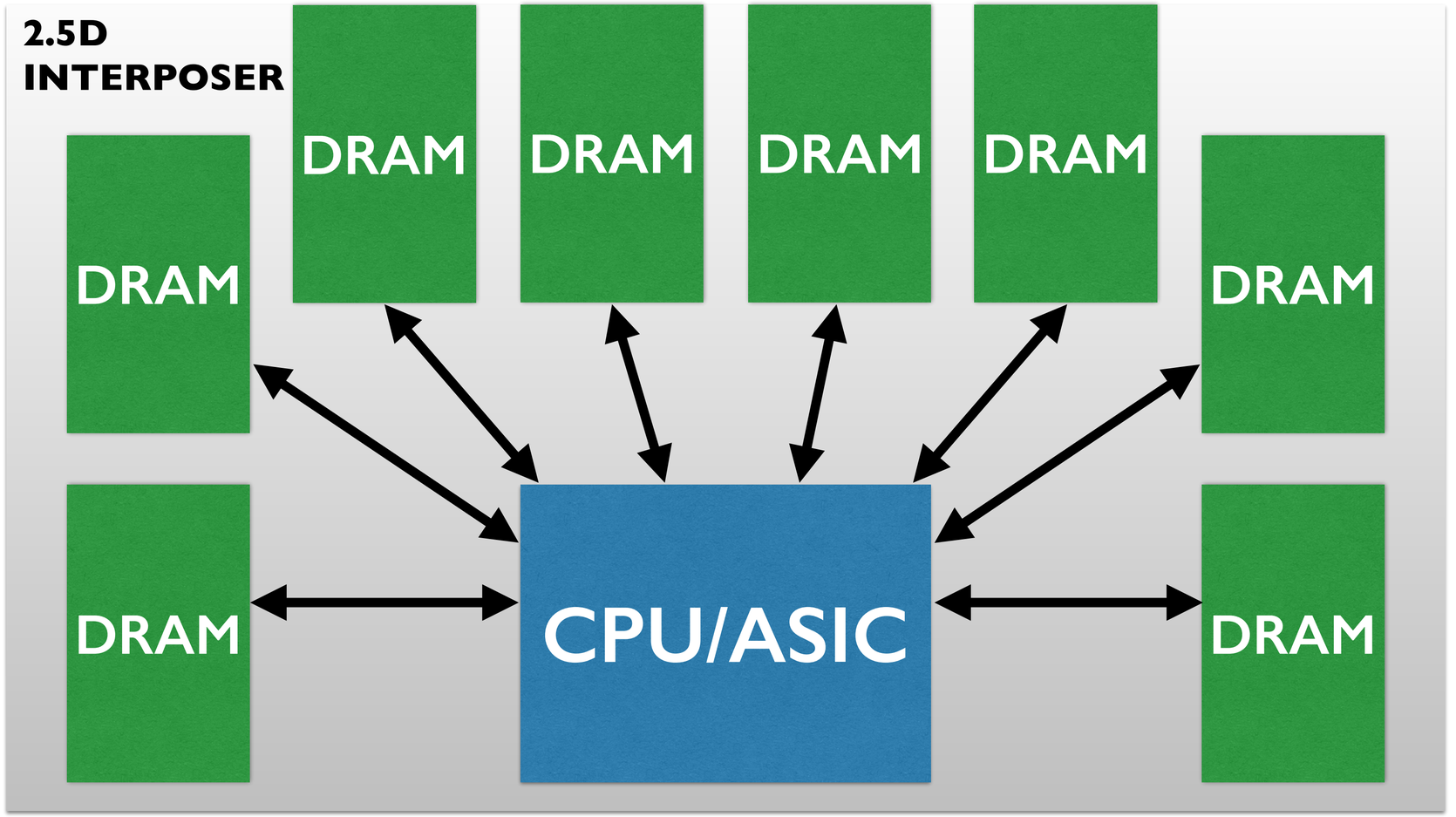
University of
Maryland

High Bandwidth Memory

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



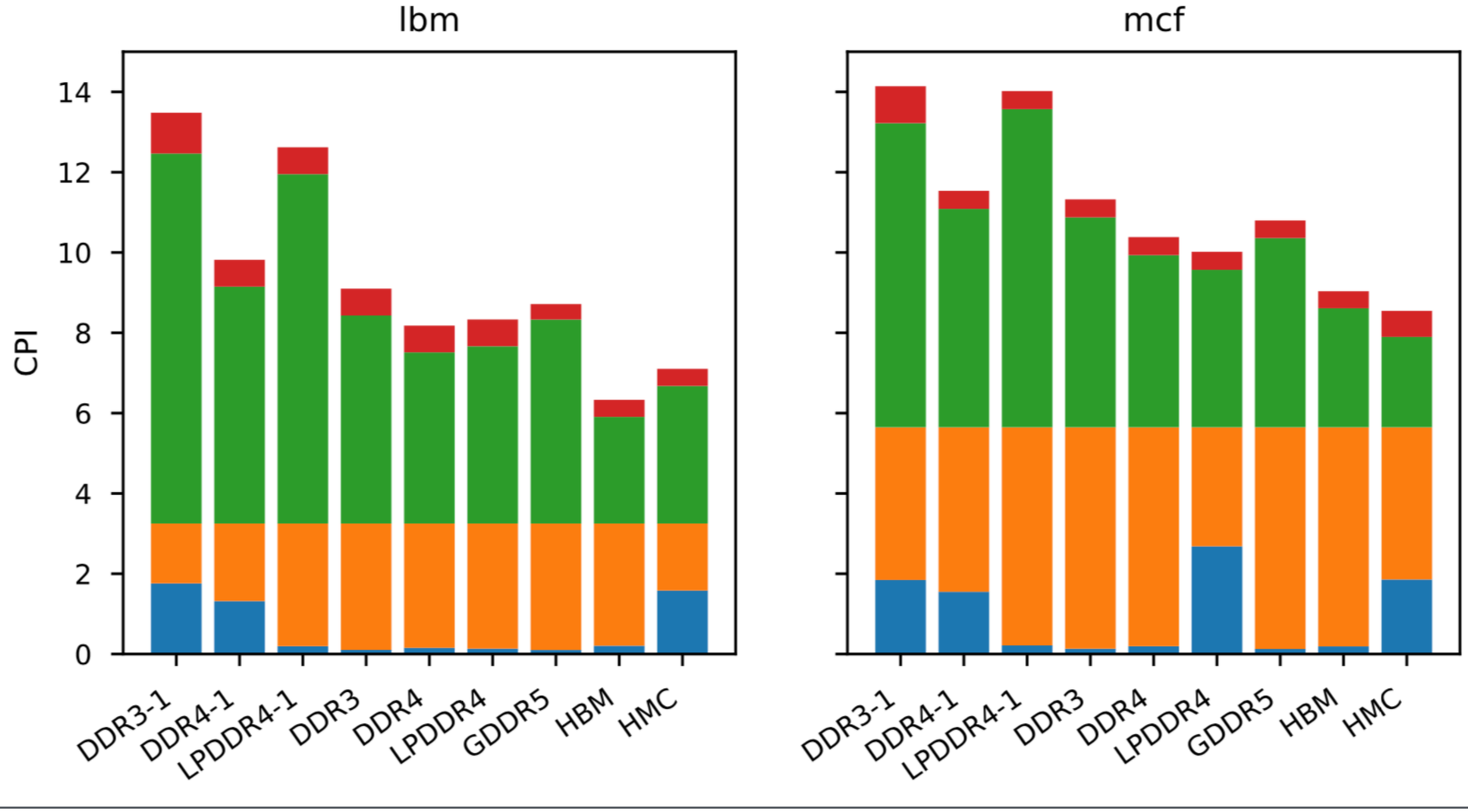
High Bandwidth Memory



Each Link is 128 Bits Wide: *1024 Total*

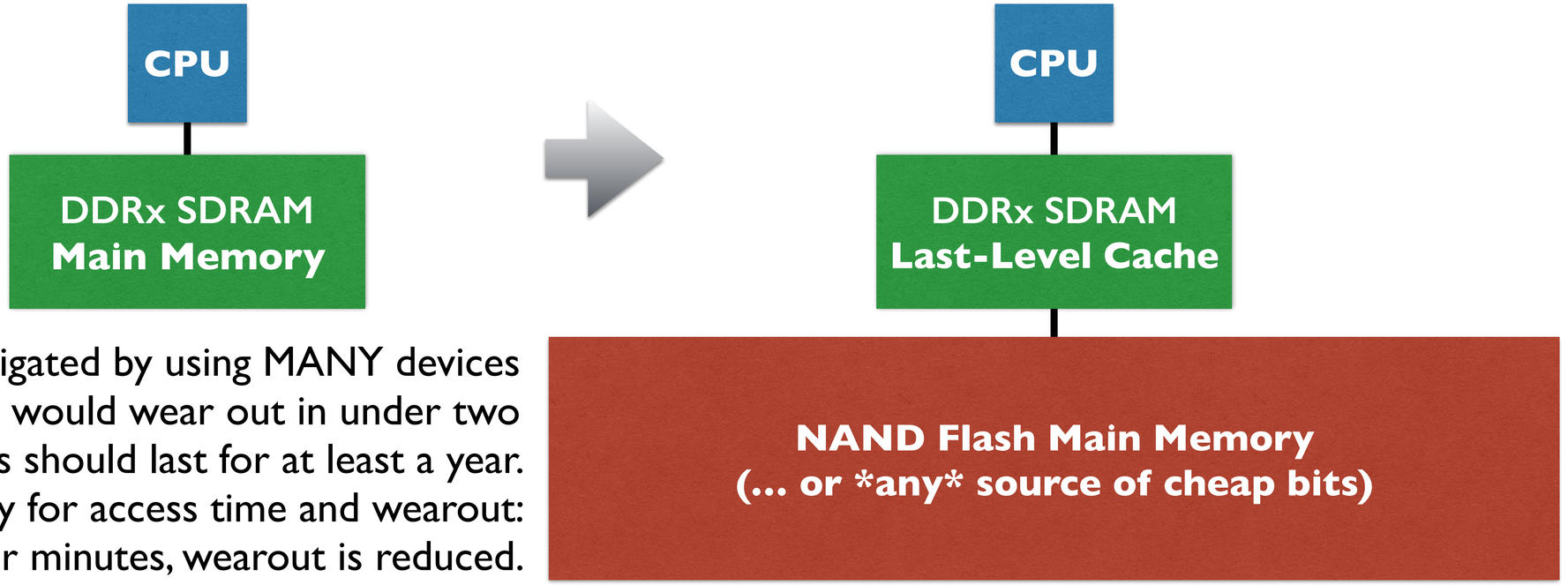
Performance Comparison

MEMSYS 2018



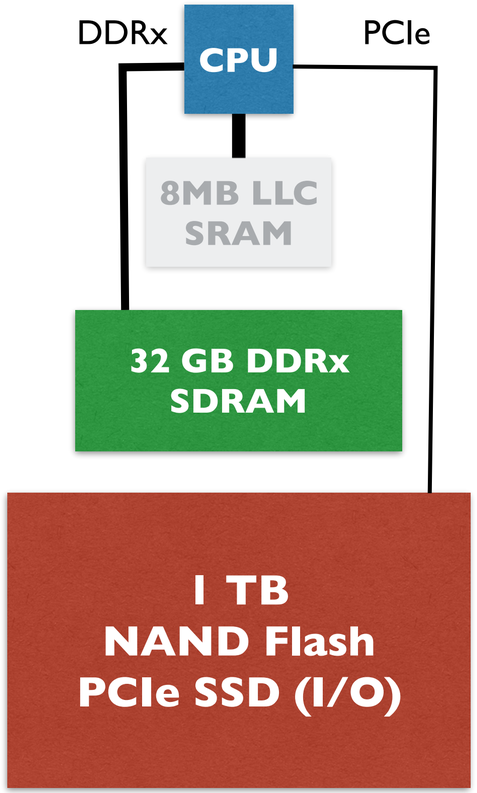
Non-Volatile Main Memory

	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 (?)

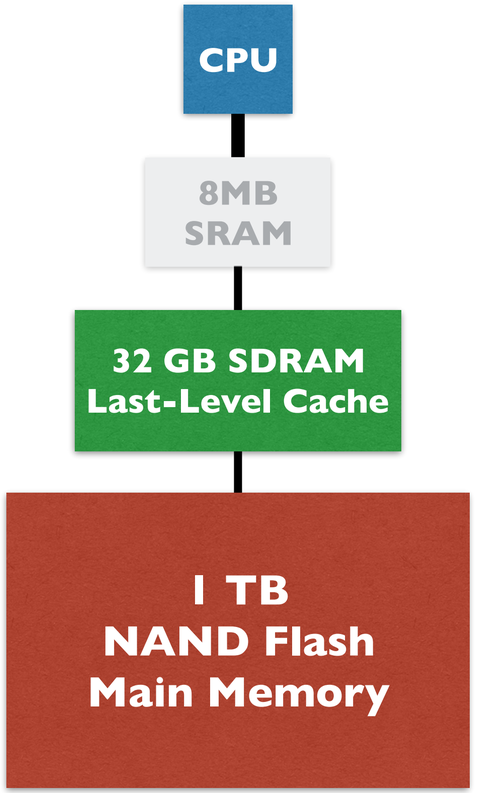


Note: wear-out mitigated by using MANY devices (thousands). A single device would wear out in under two days; therefore, 1000 devices should last for at least a year. Next, you can trade off longevity for access time and wearout: if the data need only last hours or minutes, wearout is reduced.

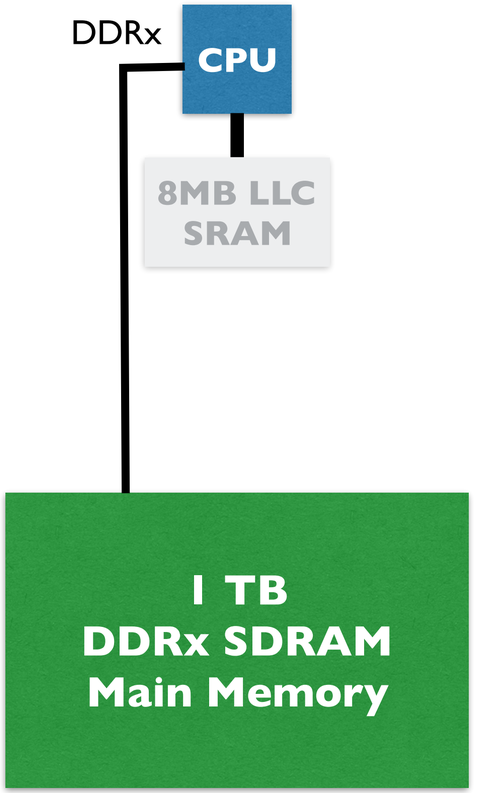
A Tale of 3 Memory Systems



SSD
$500 – 10W

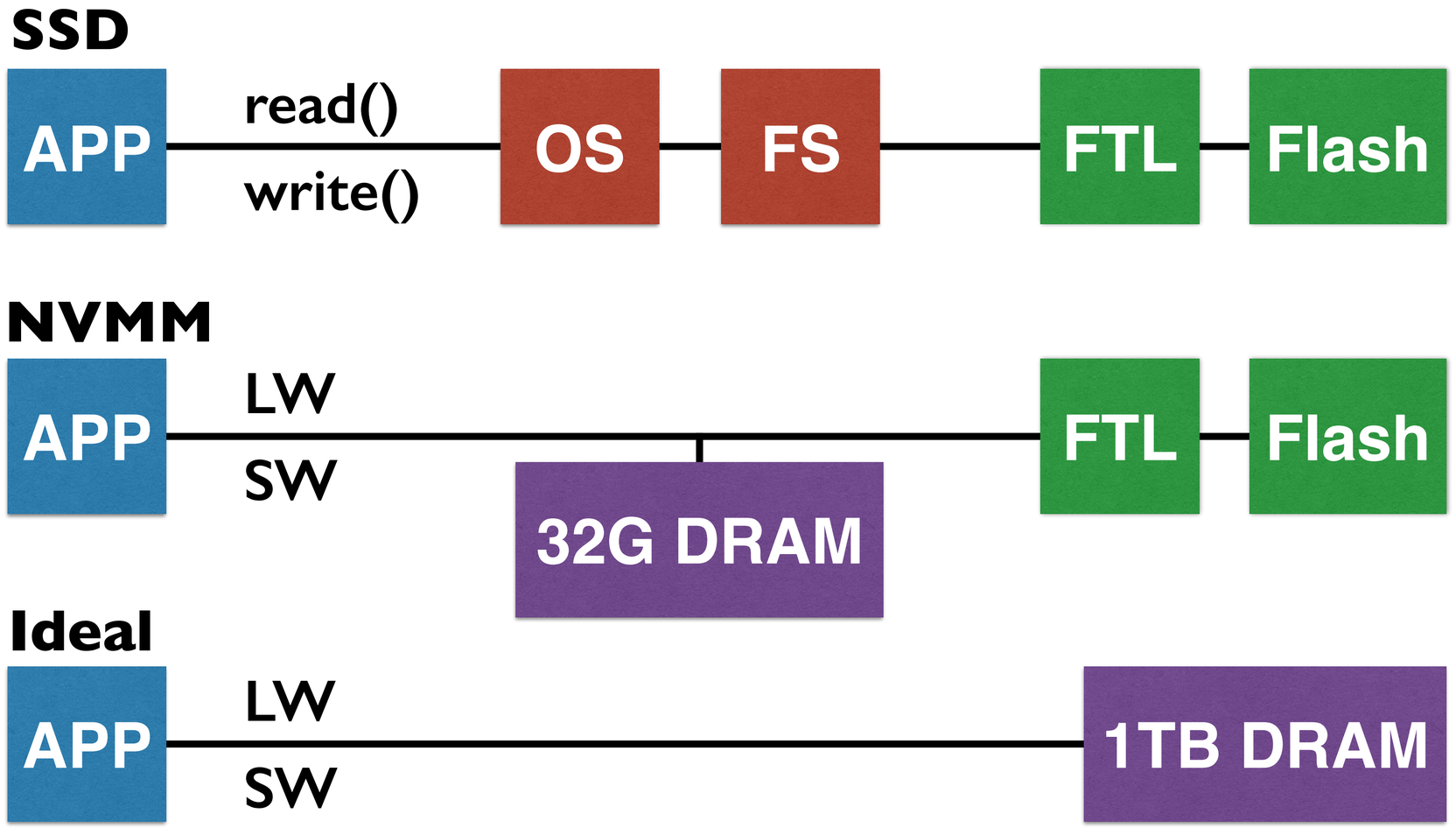


NVMM
$500 – 10s of W

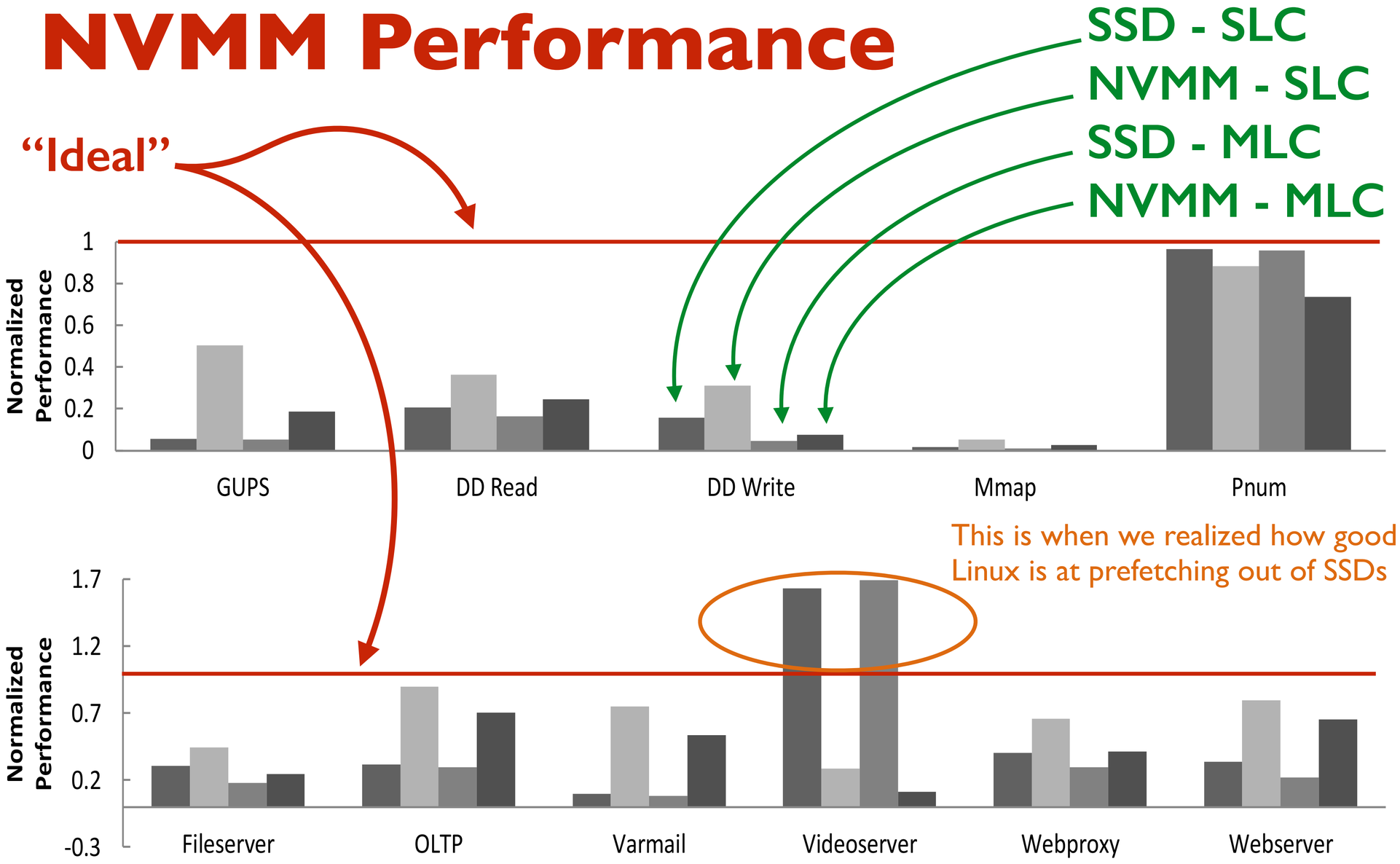


Ideal
$10,000 – 100W

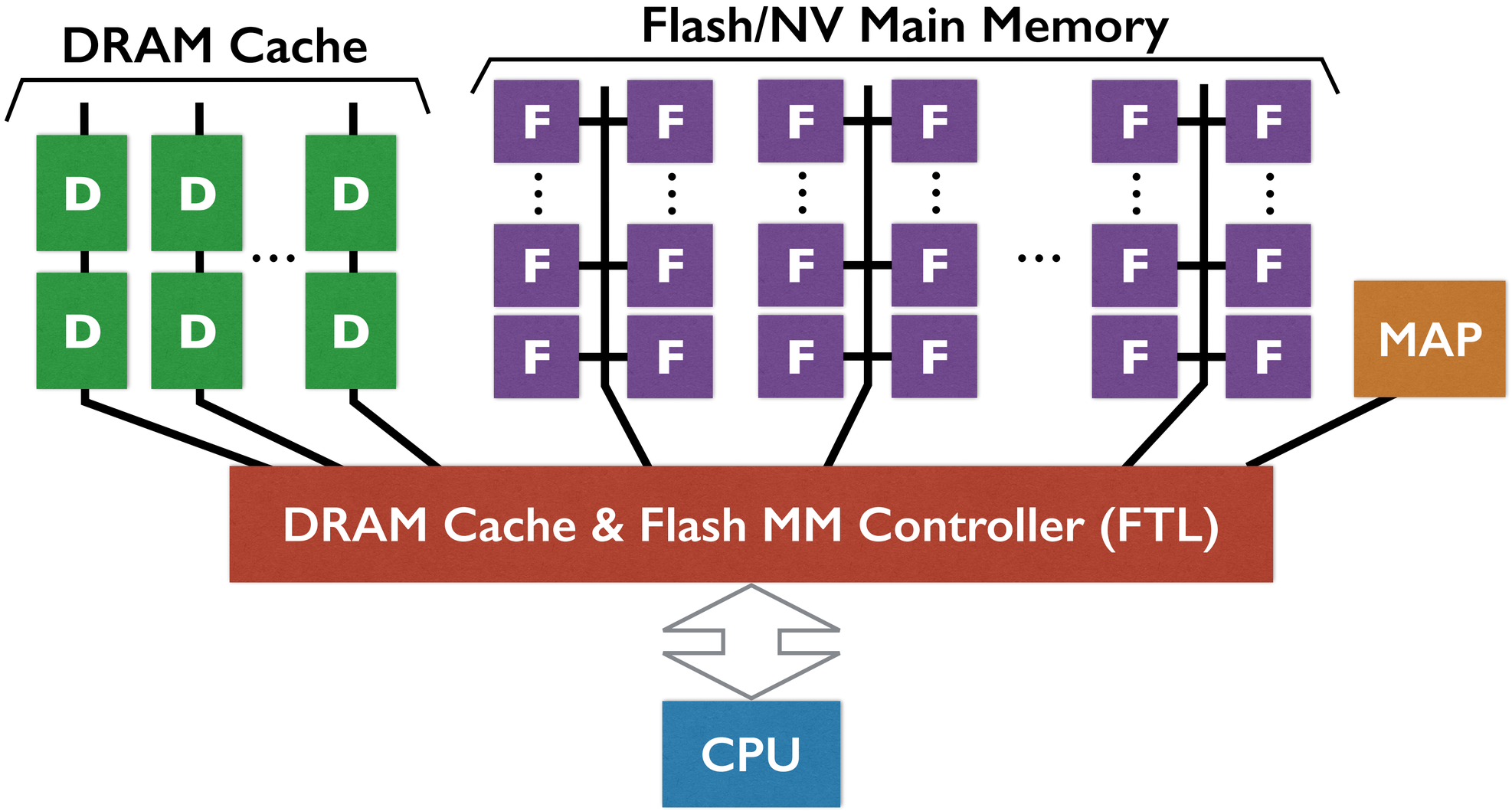
A Tale of 3 Memory Systems



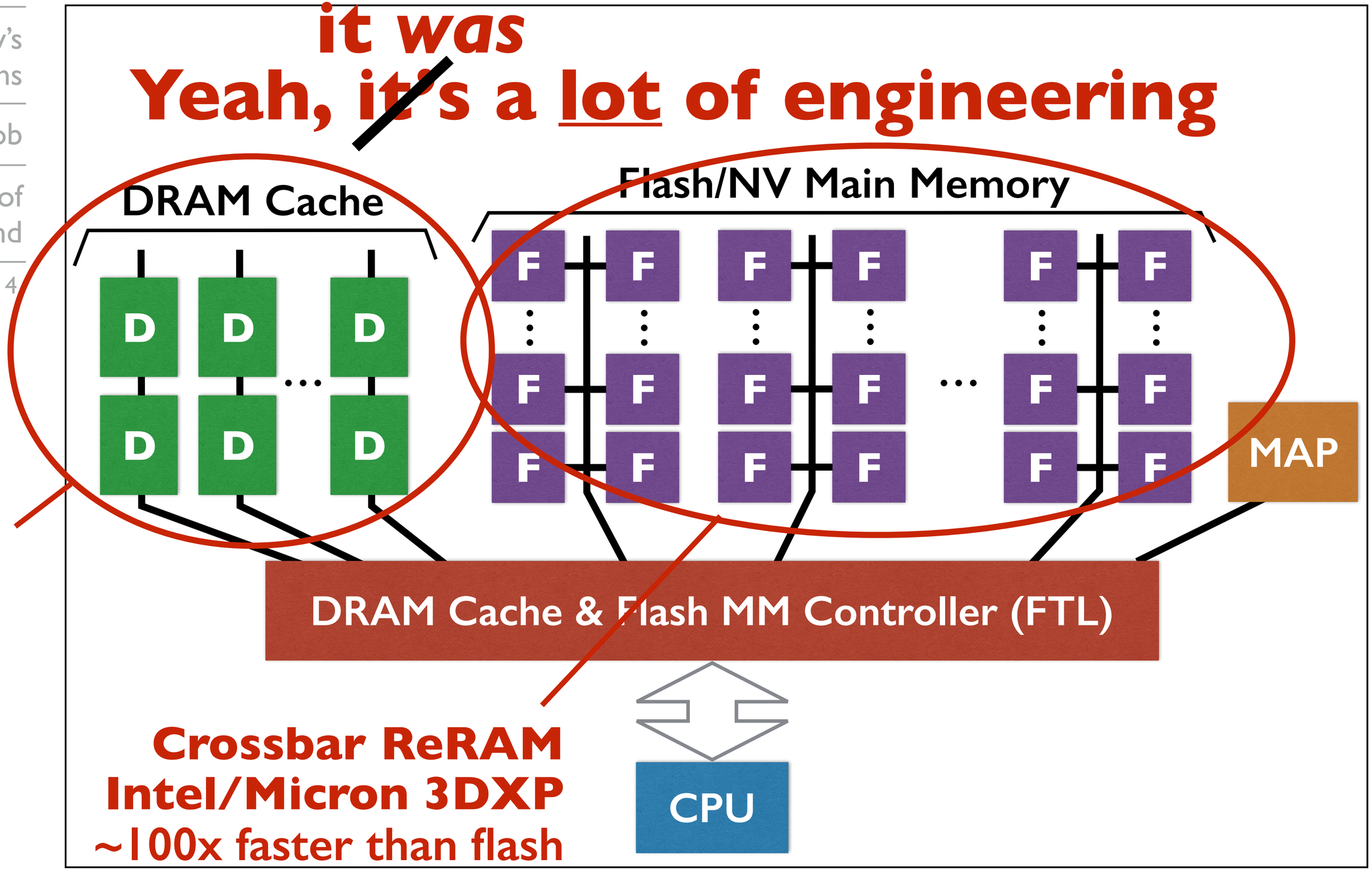
NVMM Performance



Yeah, it's a lot of engineering



it was
Yeah, it's a lot of engineering

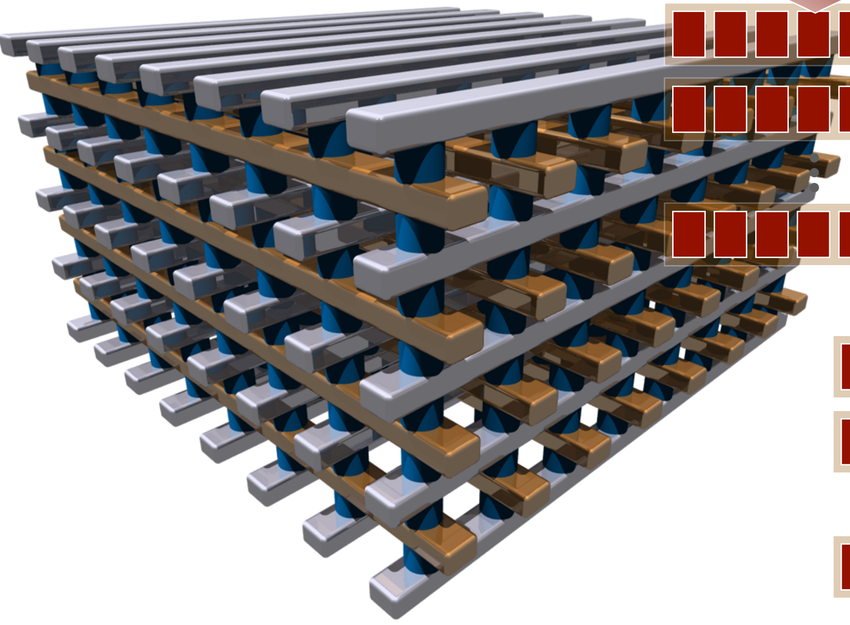


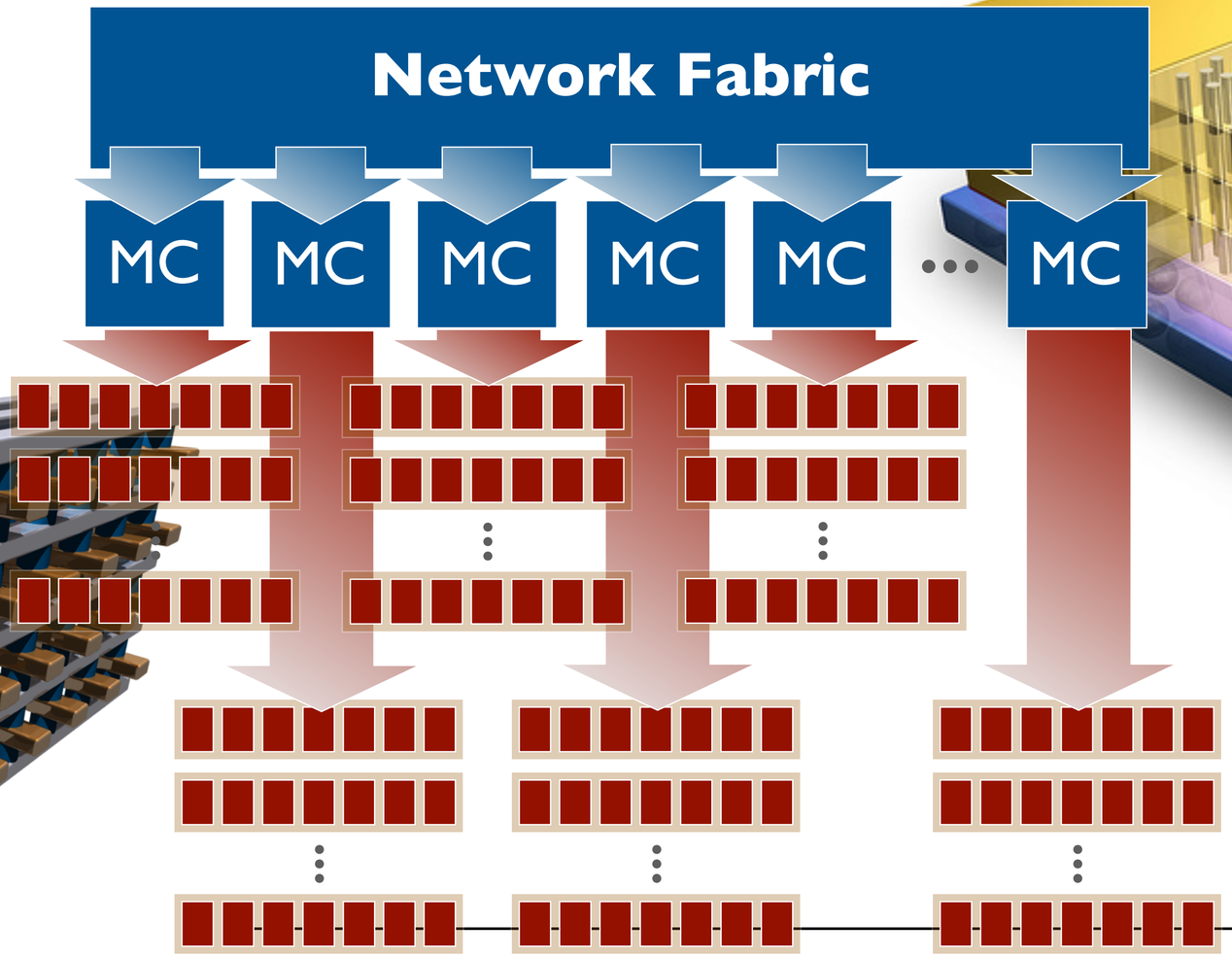
HMC:
320GB/s
16 channels
HBM:
256GB/s
8 channels

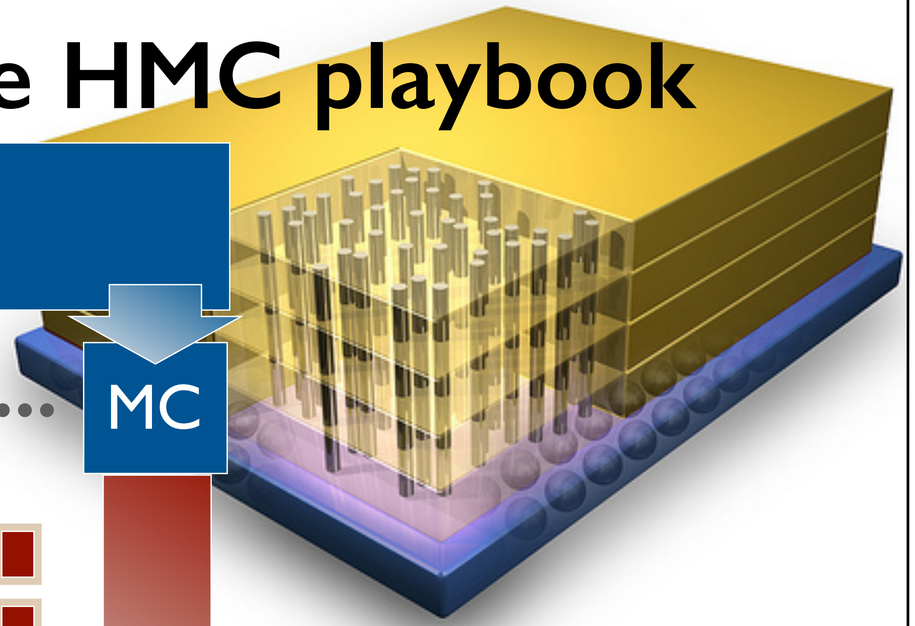
Crossbar ReRAM
Intel/Micron 3DXP
~100x faster than flash

High Bandwidth Non Volatiles

Borrow a page from the HMC playbook



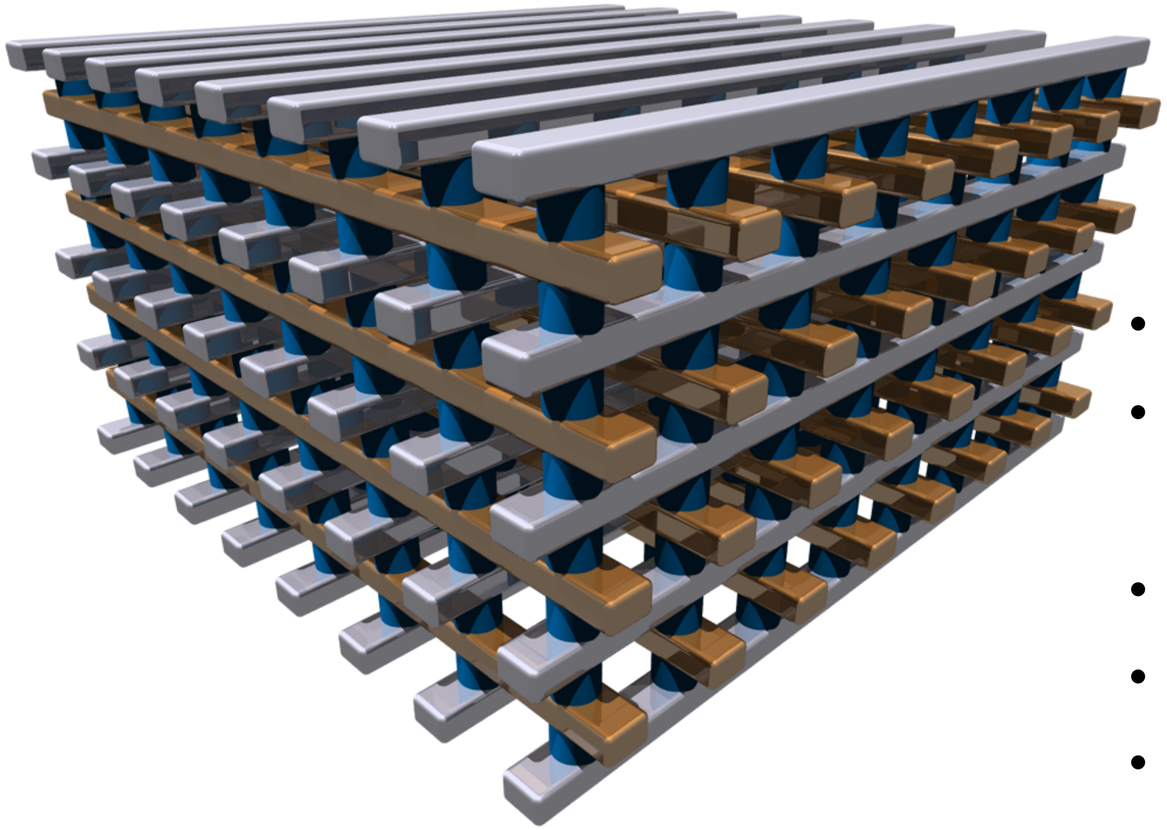


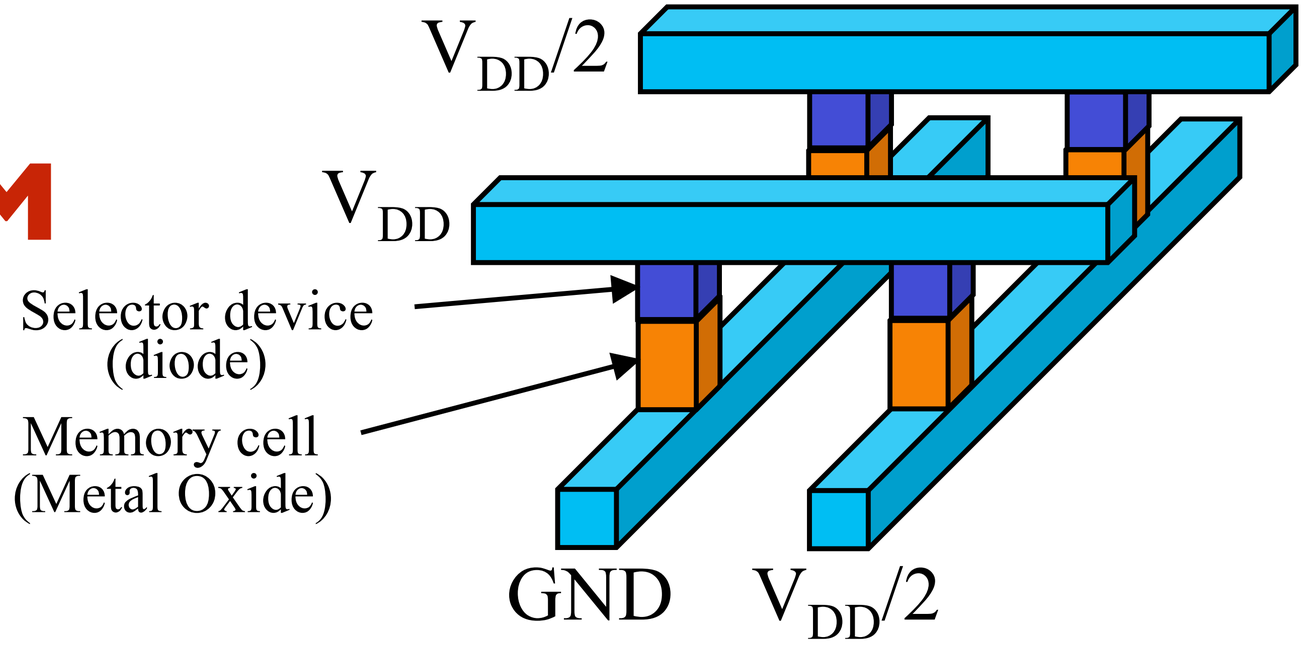


NV ReRAM: up to 1000ns expected*

***trade-offs?**

Crossbar 3D ReRAM



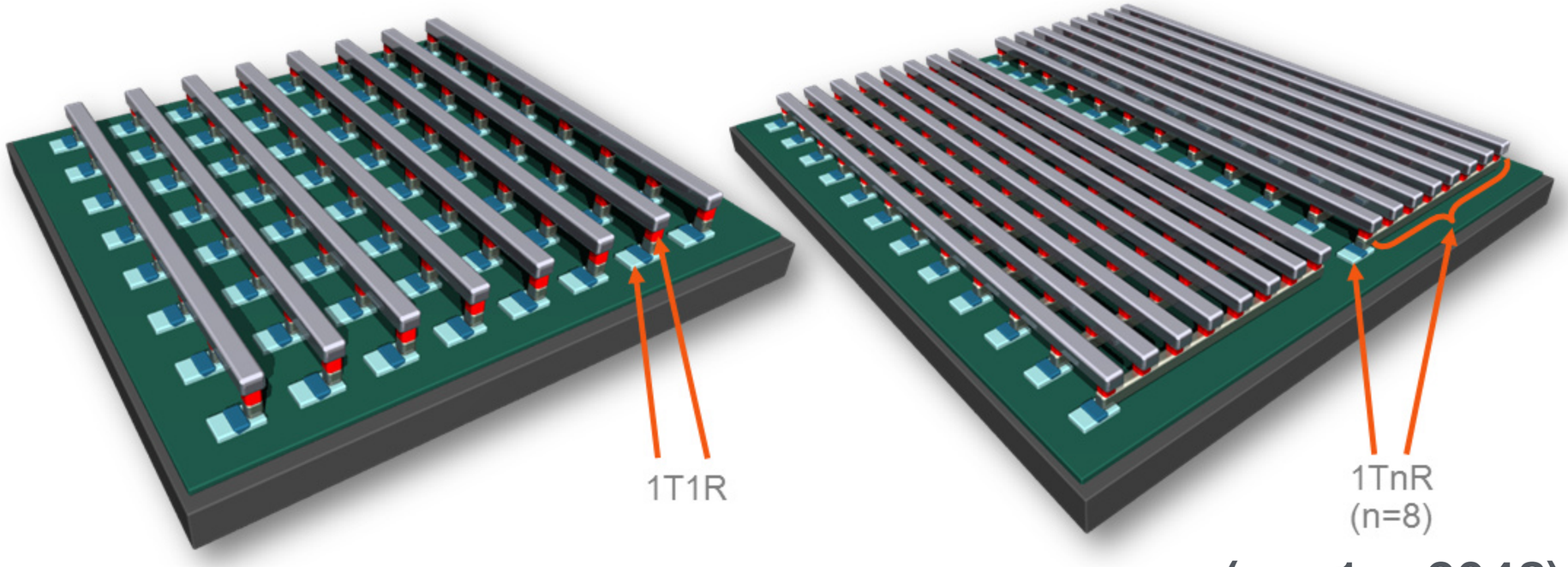


- Cells minimum area (no access transistor)
- **2-stack arrays** @ 16nm, 20 x 20 mm die: **64GB of ReRAM**
- **8-stack arrays => 256 GB of ReRAM**
- Stacks arbitrarily high
- No. Access. Transistor.

No Access Transistor

1T1R Memory Array
Low Latency, Low Density

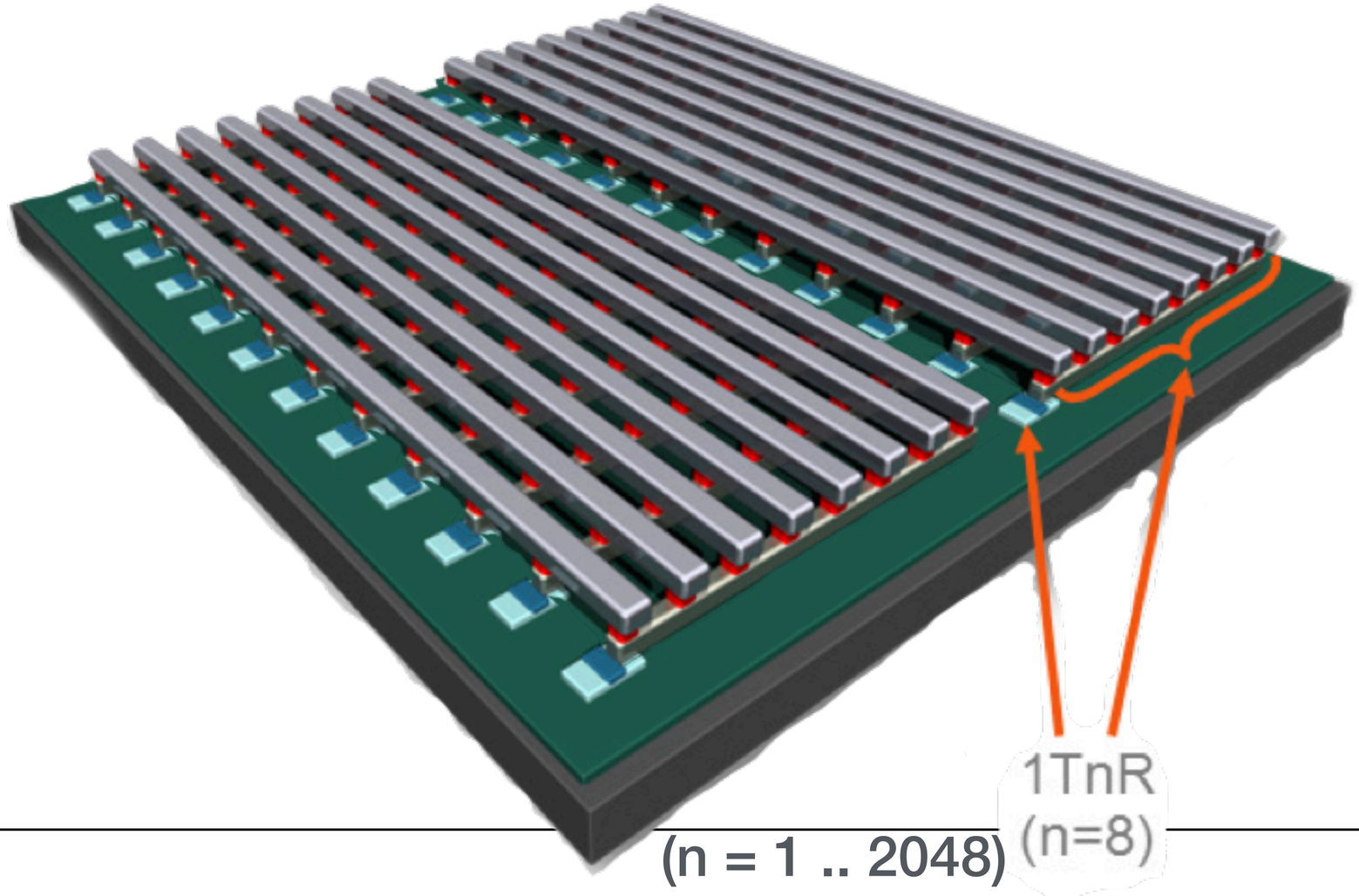
1TnR Memory Array
High Performance, High Density



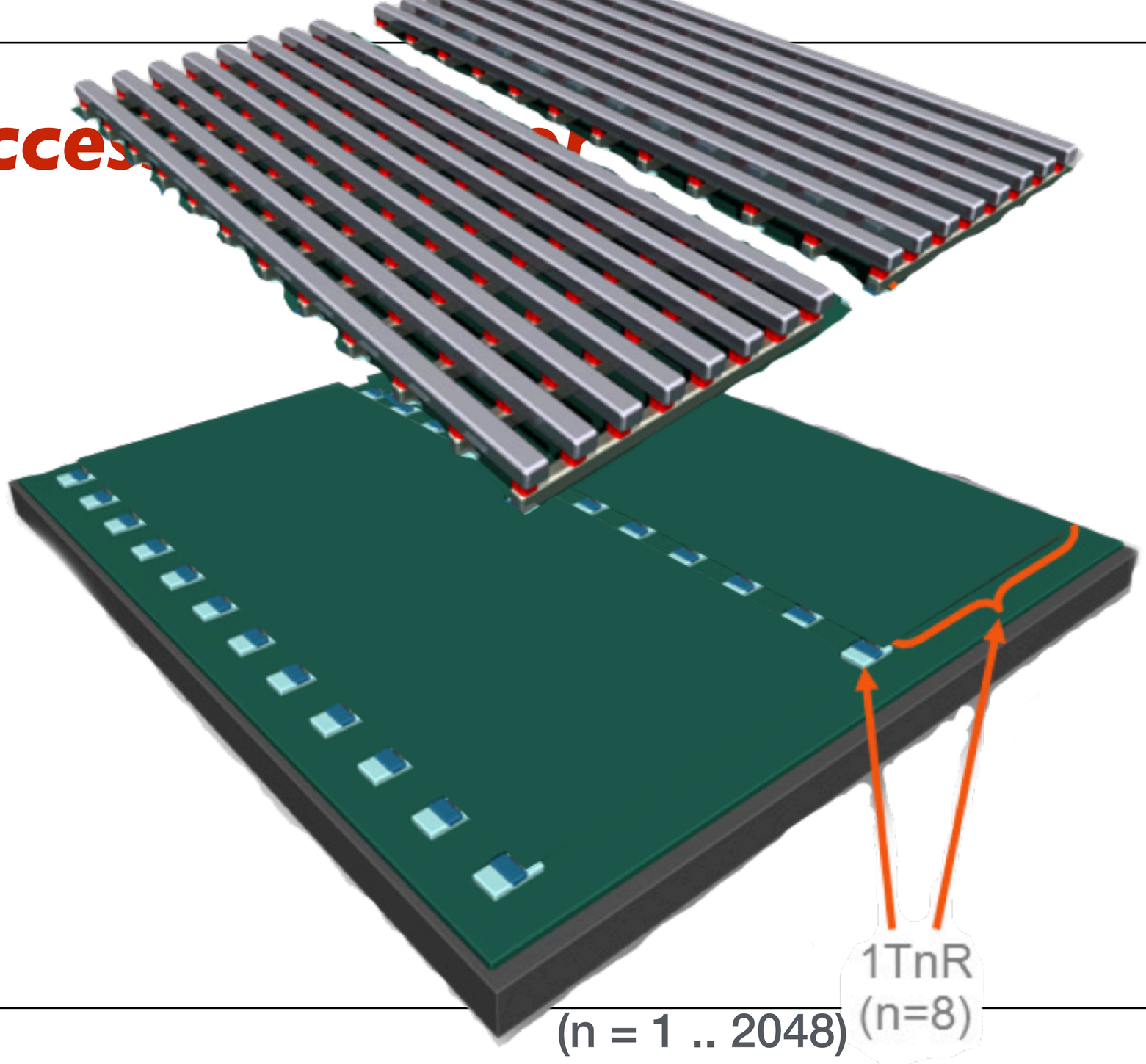
(n = 1 .. 2048)

Crossbar RRAM Technology

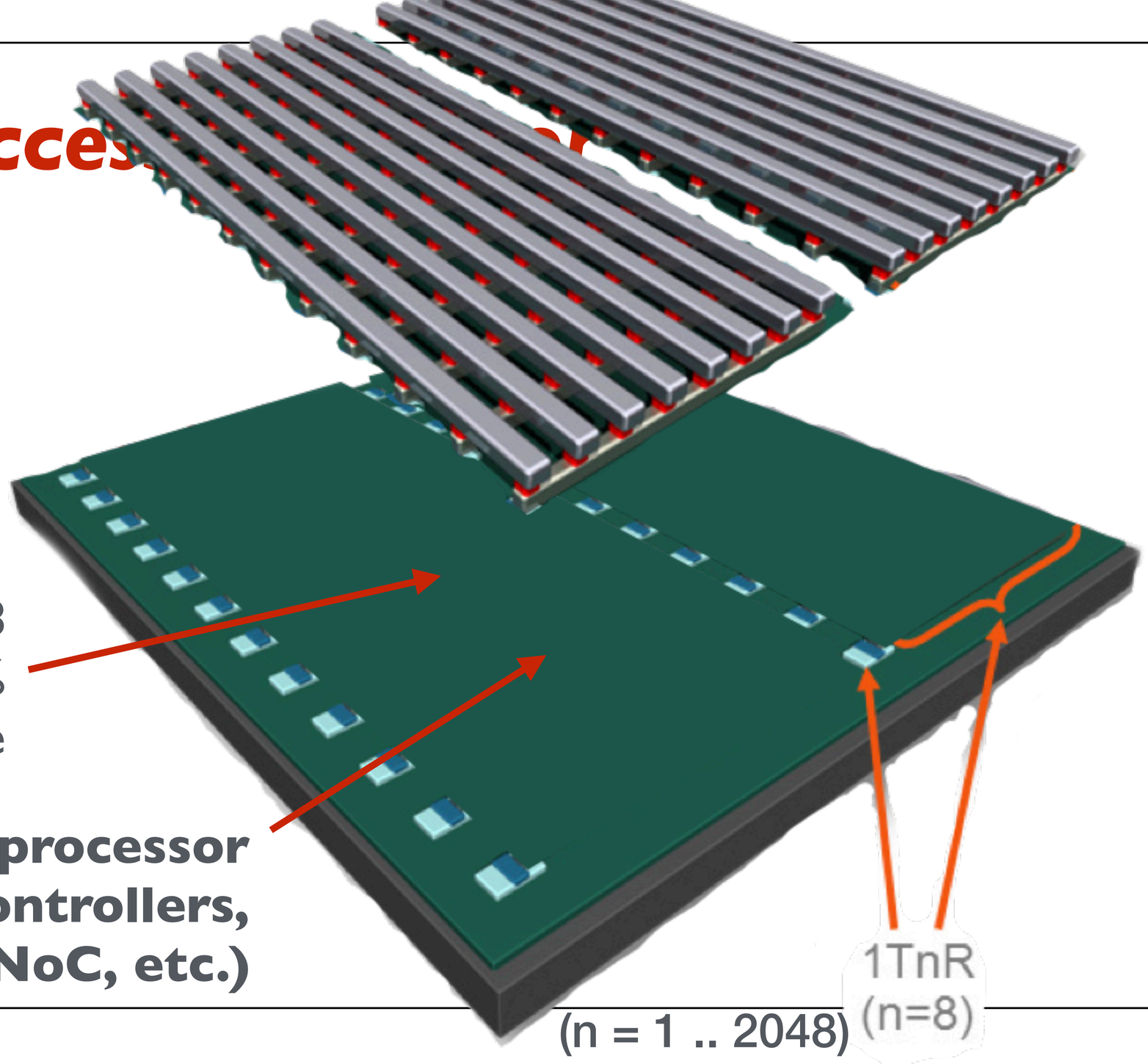
No Access Transistor



No Acces



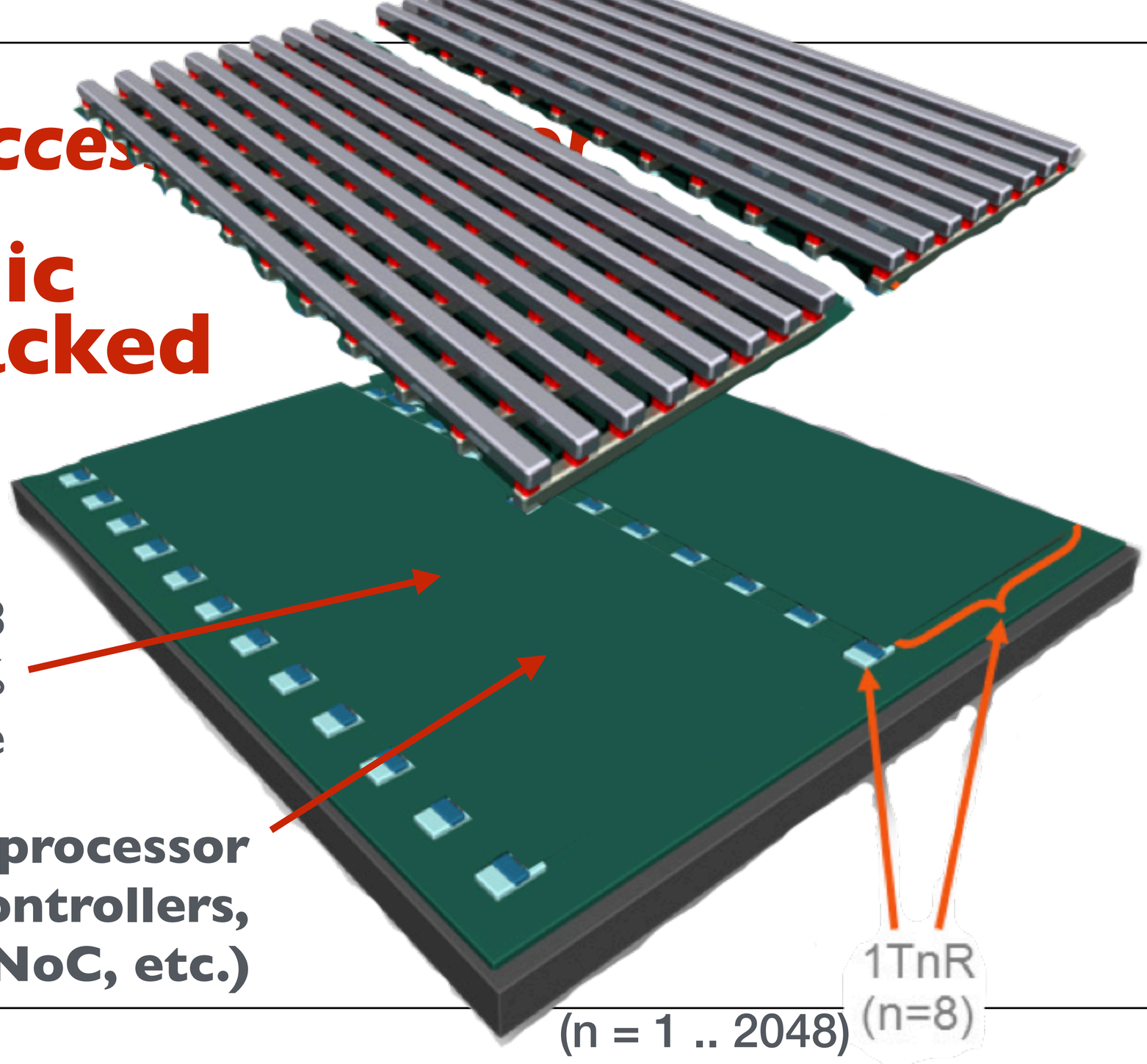
No Acces



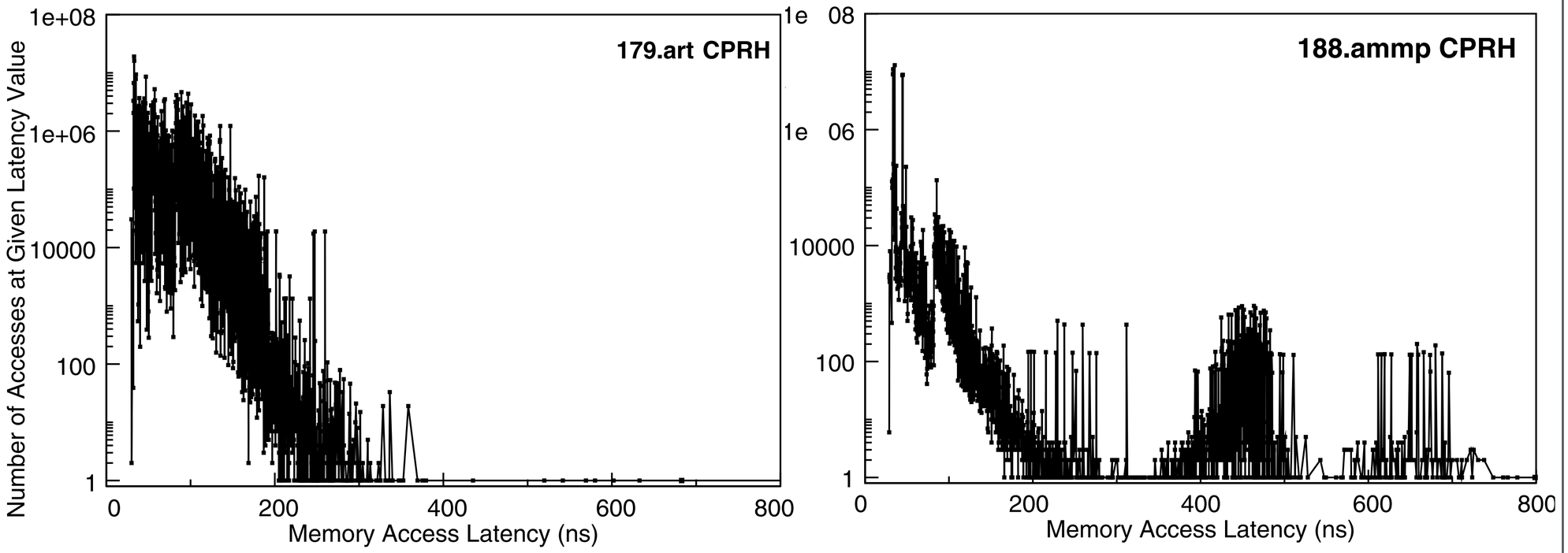
For n = 2048 area is ~75% white space

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

No Acces… Monolithic Not Die-Stacked



Recall: Real DRAM Latency Is Actually Quite Long (100ns)



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

Example Monolithic Numbers

~64 cores, ~256GB ReRAM, ~4k banks

Assume 200ns latency for 8-byte payload:

Bandwidth = 4k * 8 bytes / 200ns
= 4k * 40 MB/s
= 160 GB/s

e.g., 64 cores, each 4-way multithreaded,
each with 512-bit (8-longword) SIMD,
vectored & scatter-gather loads,
4-deep non-blocking => 8k

So what all does this enable?

HBM/HMC: hugely parallel systems (the duality of bandwidth and parallelism), streaming applications, 2x performance

NVMM: massive data sets, new OS paradigms such as merged VM+FS and journaled main memory (built-in checkpoint/restart)

HBNV: fine-grained operations on enormous sparse data sets

Datacenter & Cloud Issues

Distribution at storage-level interface simplifies application development

Potential for significant performance

RoCE appropriate for supercomputing? How about RoXX?

At what round-trip latency does this rival MPI as programming model?

Expect a shake-up soon.

Nonvolatility Issues

Unified VM+FS Subsystems (OS redesign)

➡ *By default, data in process address space temporary, garbage-collected at exit(); **permanentify** function to keep around*

➡ *Possible directions:*

- Persistent objects (e.g. Mneme, POMS) [failed only due to reliance on disk]
- Named regions

➡ *Journaled main memory w/ checkpointing*

Capacity Issues

Rethink Protection & Translation

➡ *TLB overhead is ~20%*

- So get rid of it already!
- BUT: need protection, authentication

➡ *Why not waste bits? Simplify both sharing and translation by eliminating much of VM*

➡ *OS/HW co-design needed: e.g., sharing via vaddr instead of paddr, language support?*

Recall: Nonvolatile main memories ~TB per node

Bottom Line

It's going to happen. :)

- Combined VM+FS subsystems
- Journaled main memory
- Persistent Object Store work from 80s
- OS: Simpler design, fewer potential bugs
- VM arguably a way better abstraction to distribute than the FS
- Monolithic = good for many applications

Bottom Line

It's going to happen. :)

- Combined VM+F
- Journal
- om 80s
- wer potential bugs
- way better abstraction
- bute than the FS
- Monolithic = good for many applications

… and the storage guys are showing us the way!

Shameless Plug

www.memsys.io

Washington DC Oct 2019

The IEEE International Symposium on Memory Systems 2018

Call For Papers www.memsys.io Call For Papers

MEMSYS 2018

The International Symposium on Memory Systems ❖ October 1–4, Washington DC

Important Dates

[sigconf], blind submission [no authors listed], up to 16 pages long

Organizers

Bruce Jacob, U. Maryland
Kathy Smiley, Memory Systems

Rajat Agarwal, Intel
Abdel-Hameed Badawy, NMSU

Jishen Zhao, UC San Diego

Memory-device manufacturing, memory-architecture design, and the use of memory technologies by application software all profoundly impact today's omputing systems, in terms of their performance, ty, predictability, power dissipation, and cost. Existing gies are seen as limiting in terms of power, capacity, and ging memory technologies offer the potential to overcome and design-related limitations to answer the requirements applications. Our goal is to bring together researchers, l others interested in this exciting and rapidly evolving ach other on the latest state of the art, to exchange ideas, ure challenges. *Visit memsys.io for more information.*

rest

lished papers containing significant novel ideas and re solicited. Papers focusing on system, software, and concepts, 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-system resilience, especially at large scale
- Memory and system security issues
- Operating system design for hybrid/nonvolatile memories

RAM, 3DXP, memristors, etc.
nguages, optimization
memory technologies
rdware, software, mitigation
rage/memory/accelerators
ment techniques
rdware and software,
plications
latacenter applications
e-memory machines

echnologies to support them,
l heterogeneous memories

side of traditional
ers.

ideas that
oups—to
eople,
ple and
ept extended
h papers, and each
accepted submission is given a 25-minute presentation time slot.
All accepted papers will be published in the ACM Digital Library.

Thank You!

Bruce Jacob

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