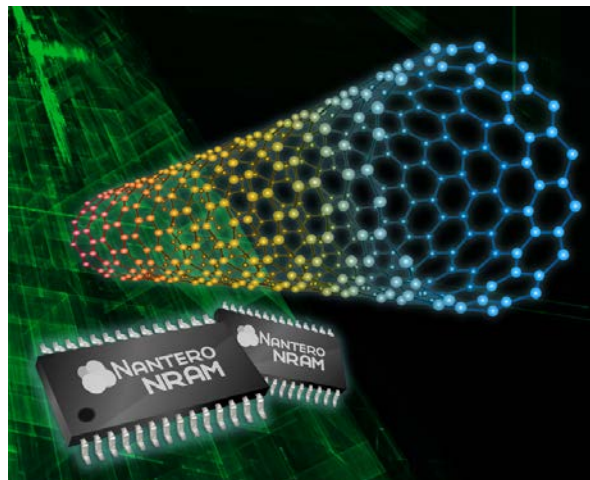


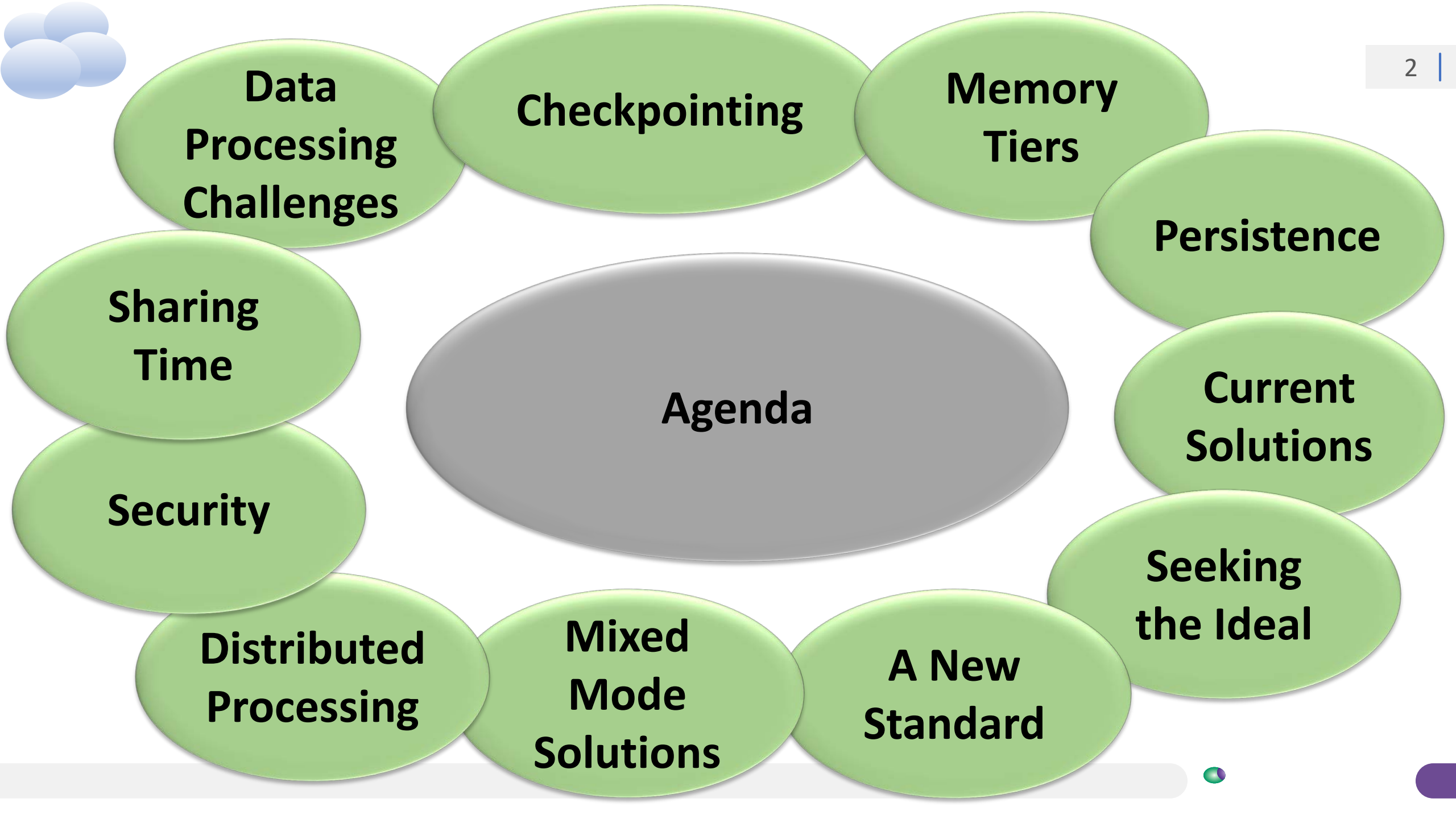
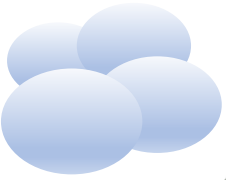


Expanding the World of Heterogenous Memory Hierarchies

The Evolving Non-Volatile Memory Story



Bill Gervasi
Principal Systems Architect



**Data
Processing
Challenges**

Checkpointing

**Memory
Tiers**

Persistence

**Current
Solutions**

**Seeking
the Ideal**

**A New
Standard**

**Mixed
Mode
Solutions**

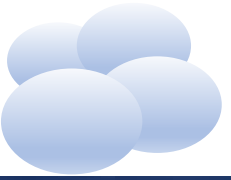
**Distributed
Processing**

Security

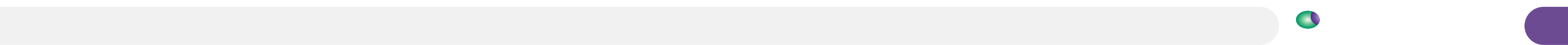
**Sharing
Time**

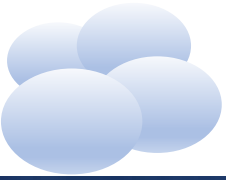
Agenda





Data processing is great



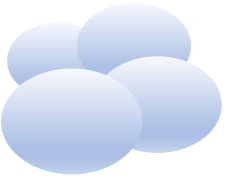


Data processing is great



Until something goes wrong





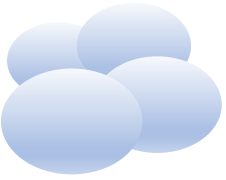
The Cost of Power Failure

According to Gartner, the average cost of IT downtime is **\$5,600** per minute. Because there are so many differences in how businesses operate, downtime, at the low end, can be as much as \$140,000 per hour, **\$300,000** per hour on average, and as much as \$540,000 per hour at the higher end. Jun 18, 2018

The 20 | The Cost of IT Downtime | The 20
<https://www.the20.com/blog/the-cost-of-it-downtime/>

Amazon.com Goes Down, Loses \$66,240 Per Minute





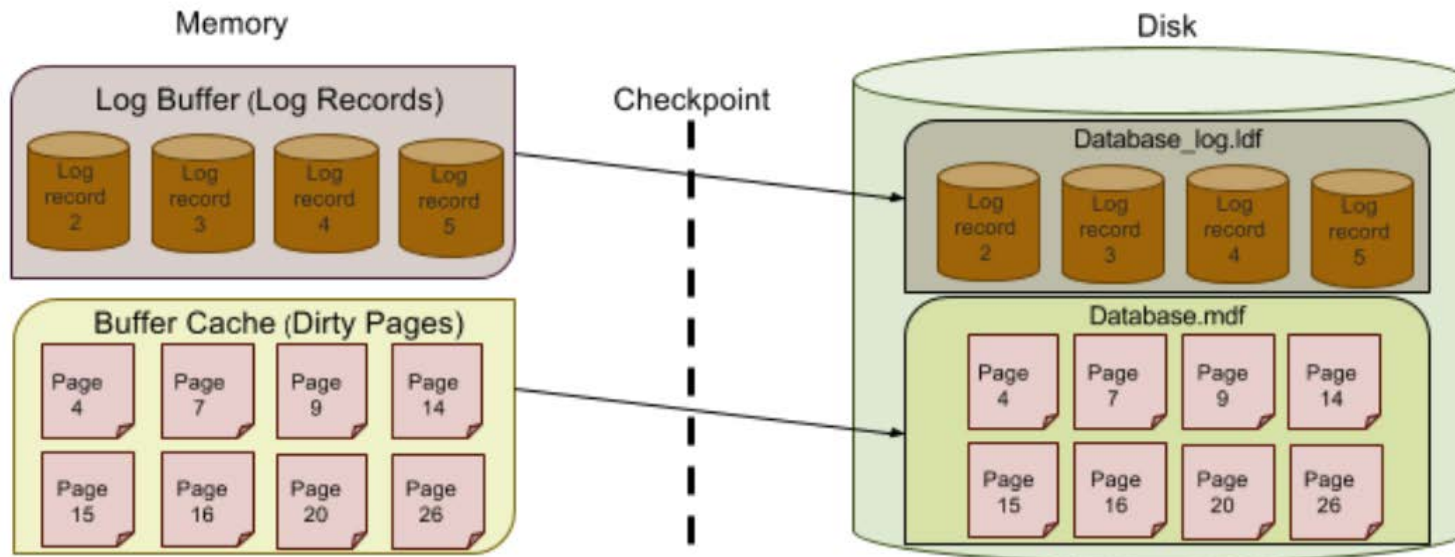
Checkpoint

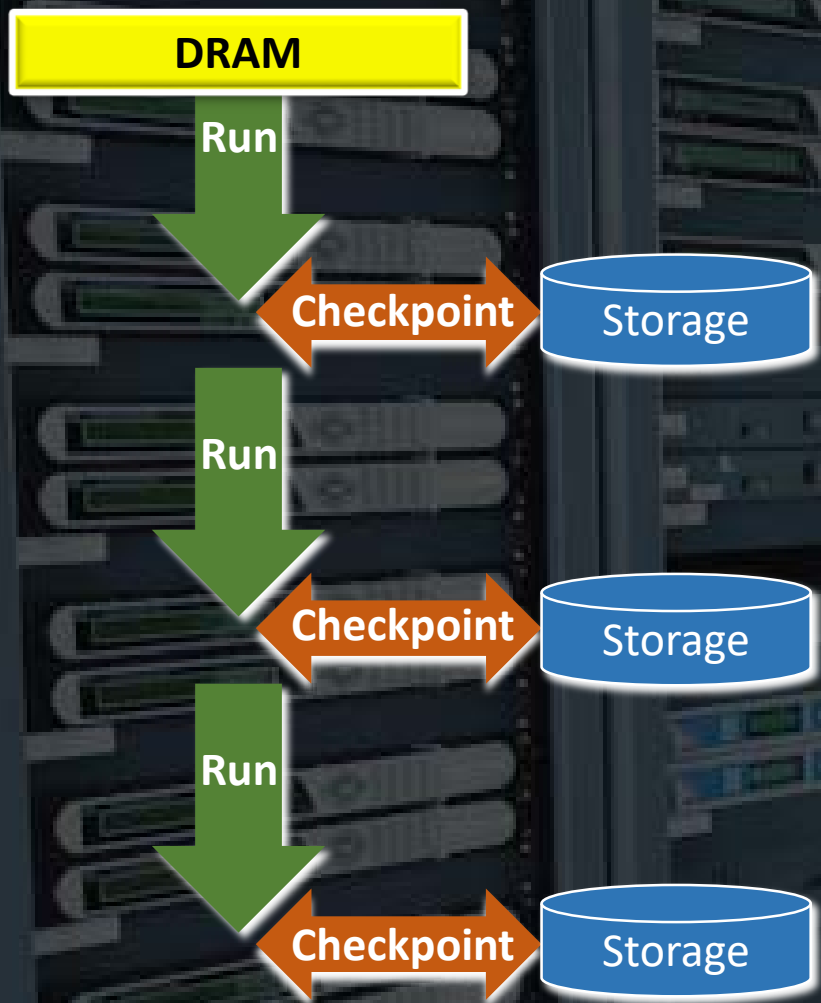
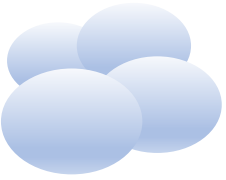
📅 November 12, 2015 👤 Alexandr Omelchenko 📁 Glossary

★★★★☆ [Total: 21 Average: 4.2/5]

Checkpoint is a process that writes current in-memory dirty pages (modified pages) and **transaction log** records to physical disk. In SQL Server checkpoints are used to reduce the time required for recovery in the event of system failure. **Checkpoint** is regularly issued for each database. The following set of operations starts when checkpoint occurs:

1. Log records from log buffer (including the last log record) are written to the disk.
2. All dirty data file pages (pages that have been modified since the last checkpoint or since they were read from disk) are written into the data file from the buffer cache.
3. Checkpoint **LSN** is recorded in the database boot page.



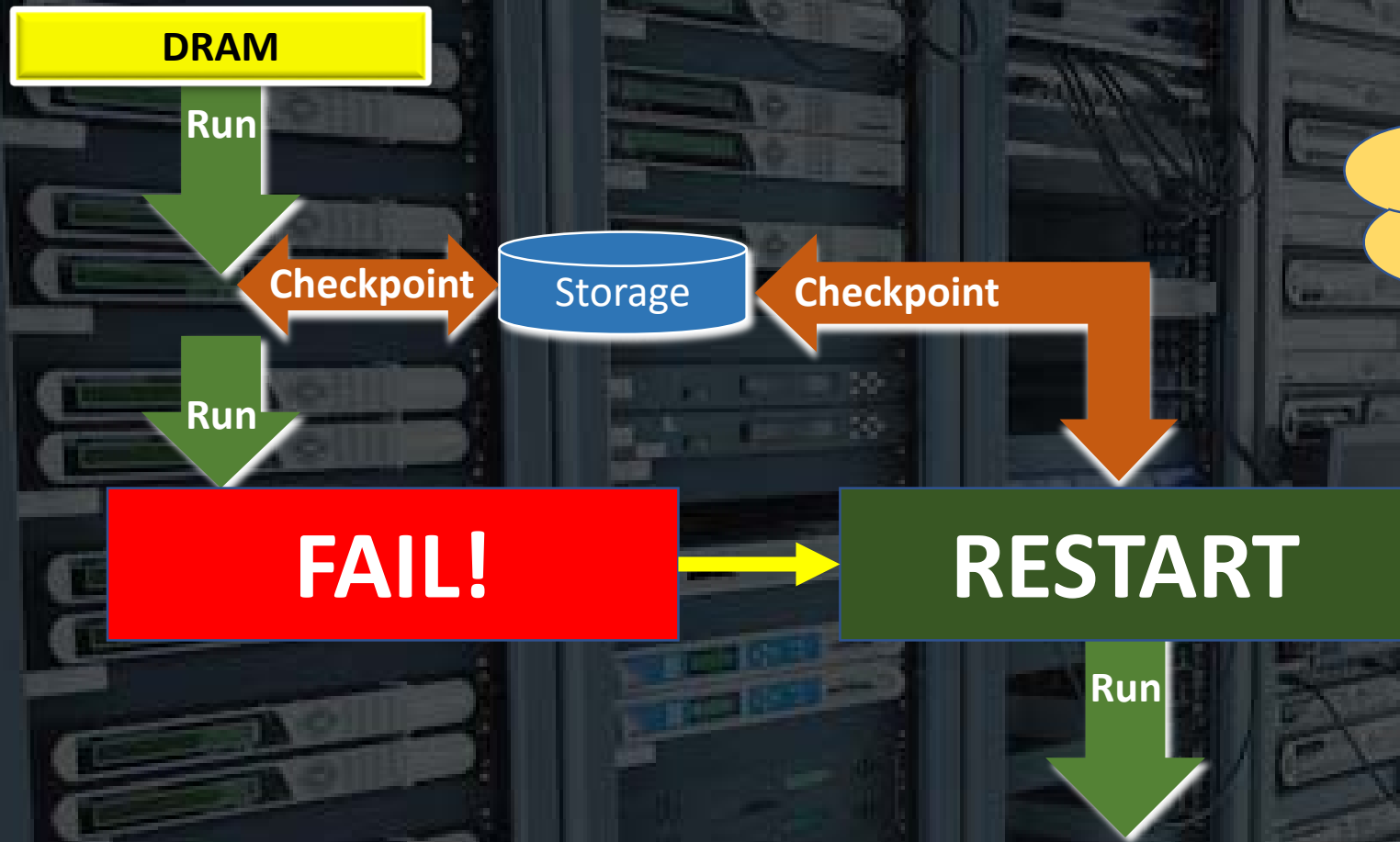


Checkpointing degrades performance

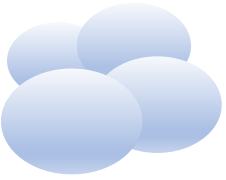
Checkpointing burns power

Checkpointing sucks





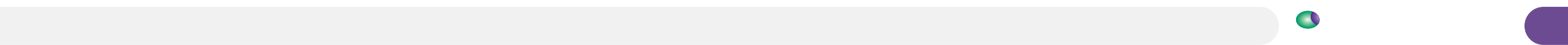
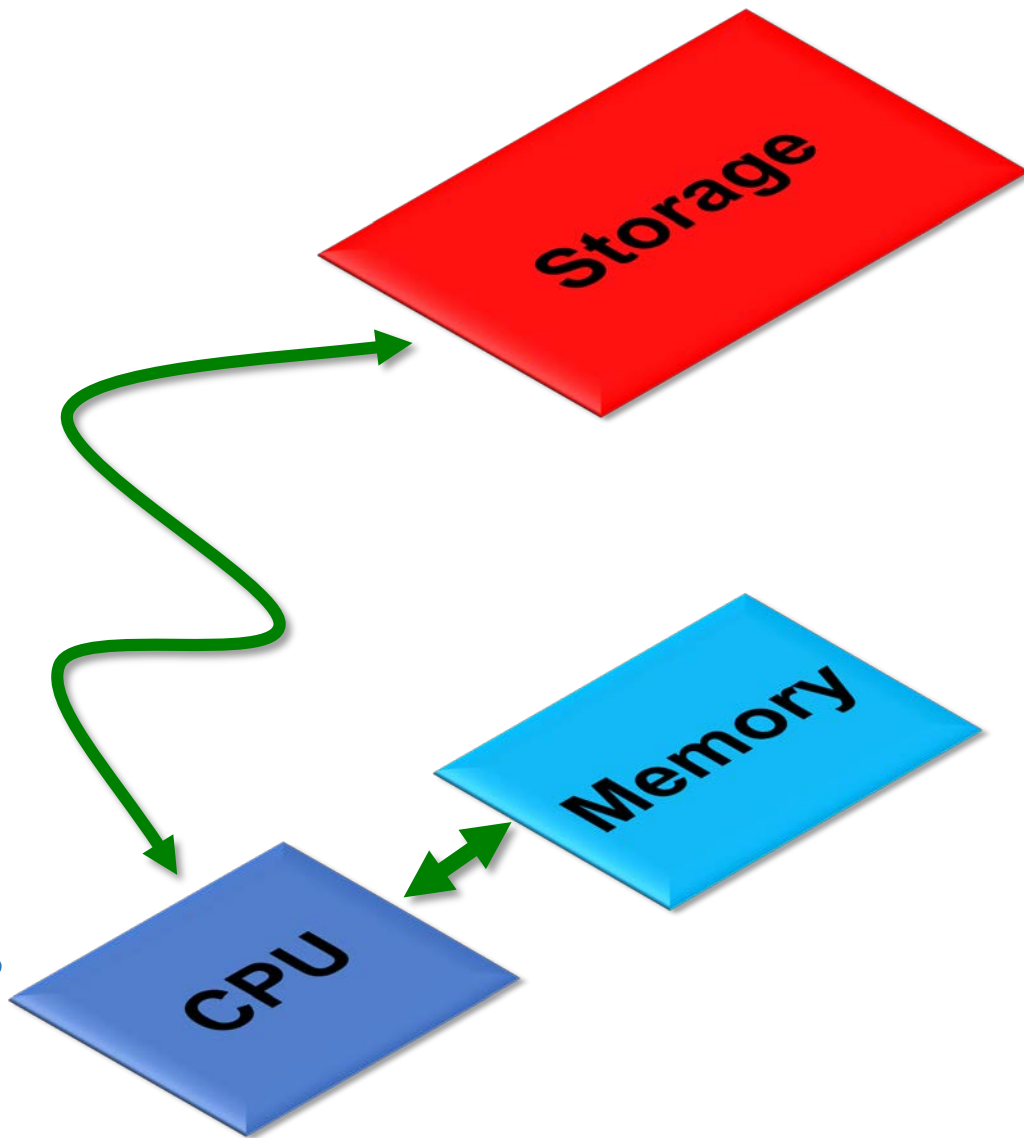
But checkpointing
avoids data loss from
failure

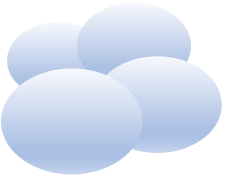


System failure is a key factor in server software design

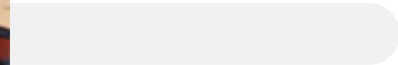
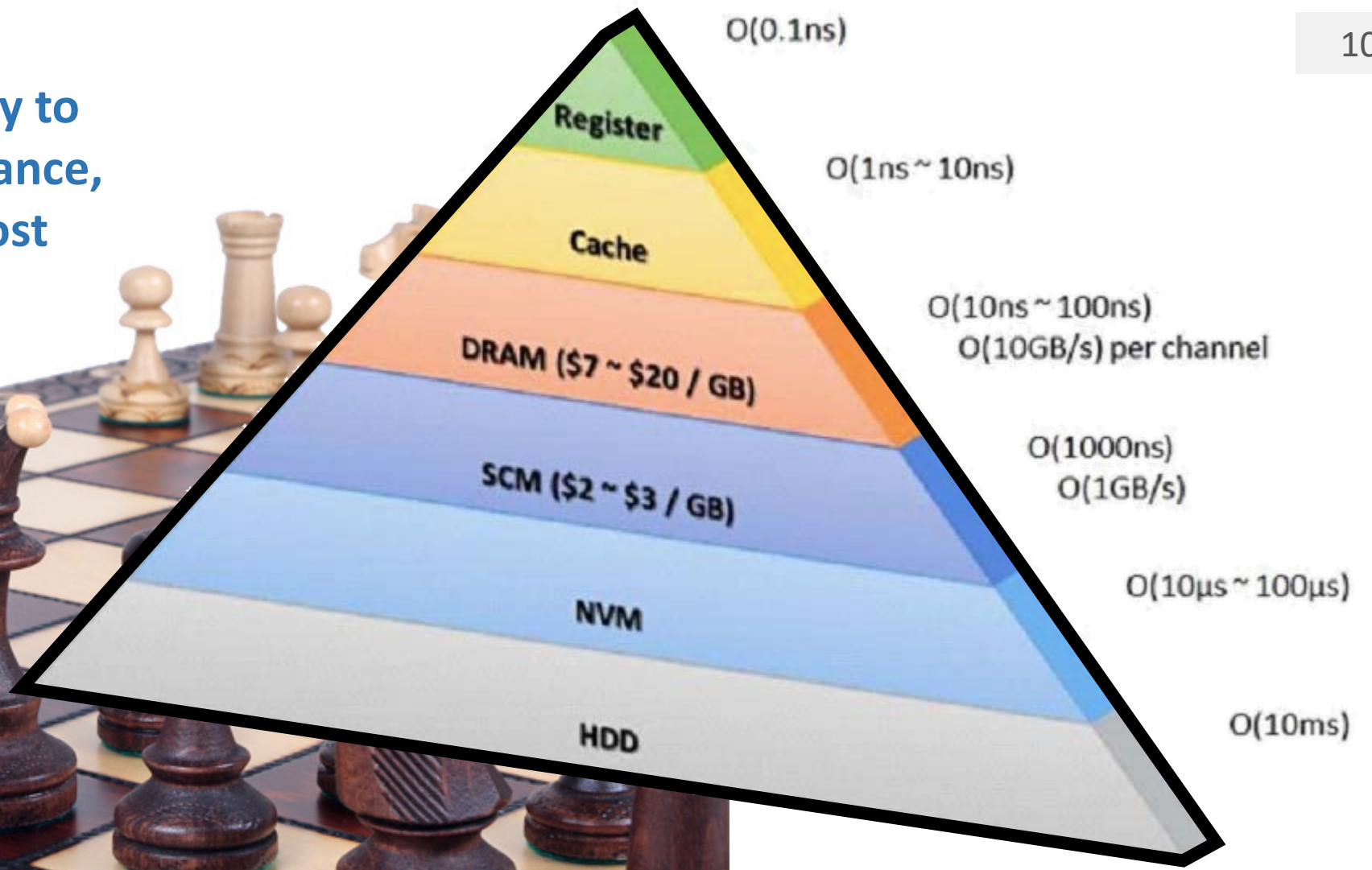
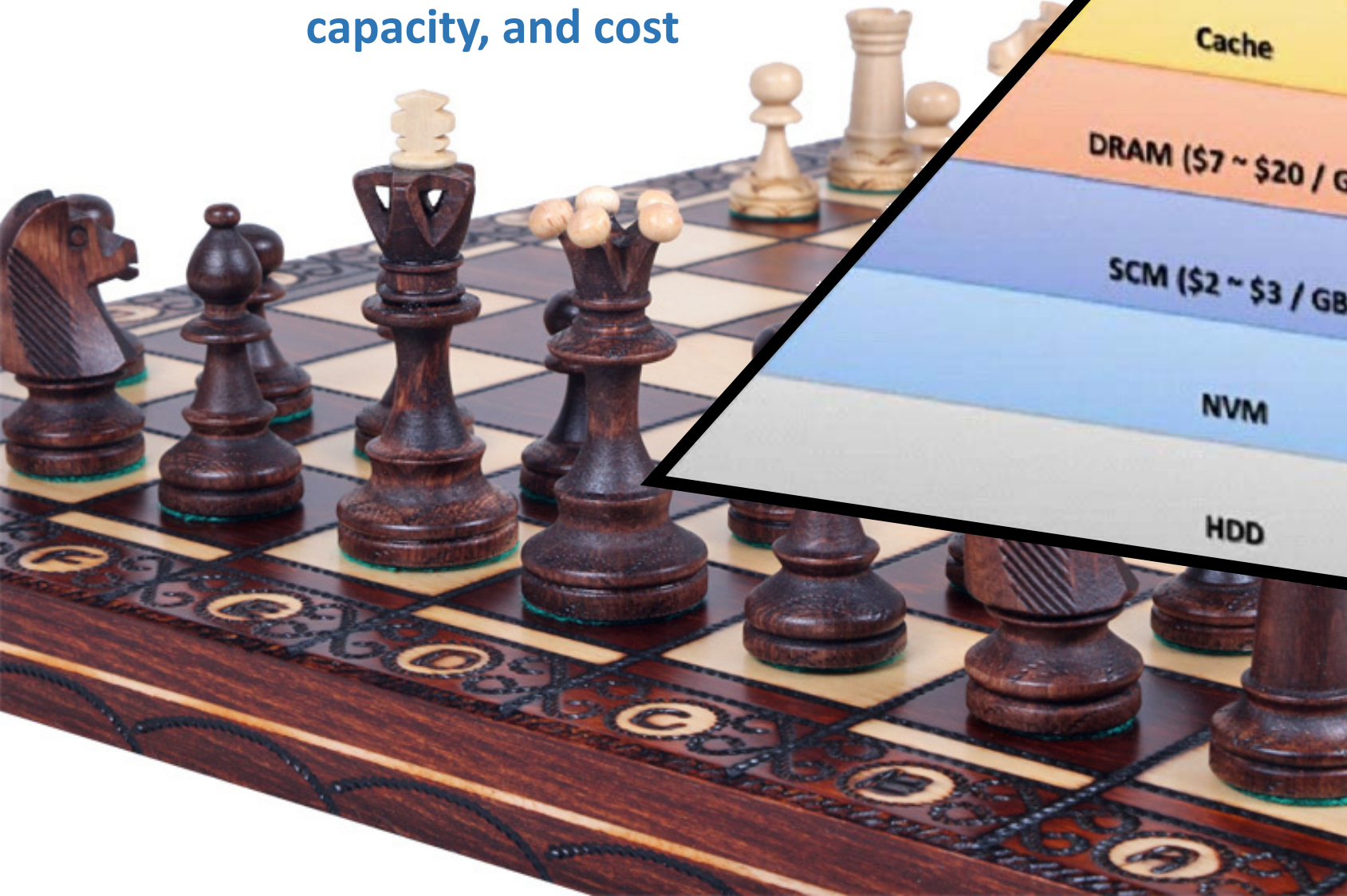
Data persistence is essential

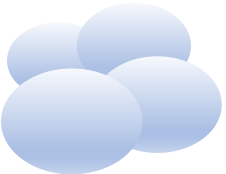
Storage access time impacts transaction granularity





The game we play to trade off performance, capacity, and cost



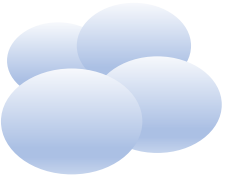


To reduce the penalties from checkpointing...

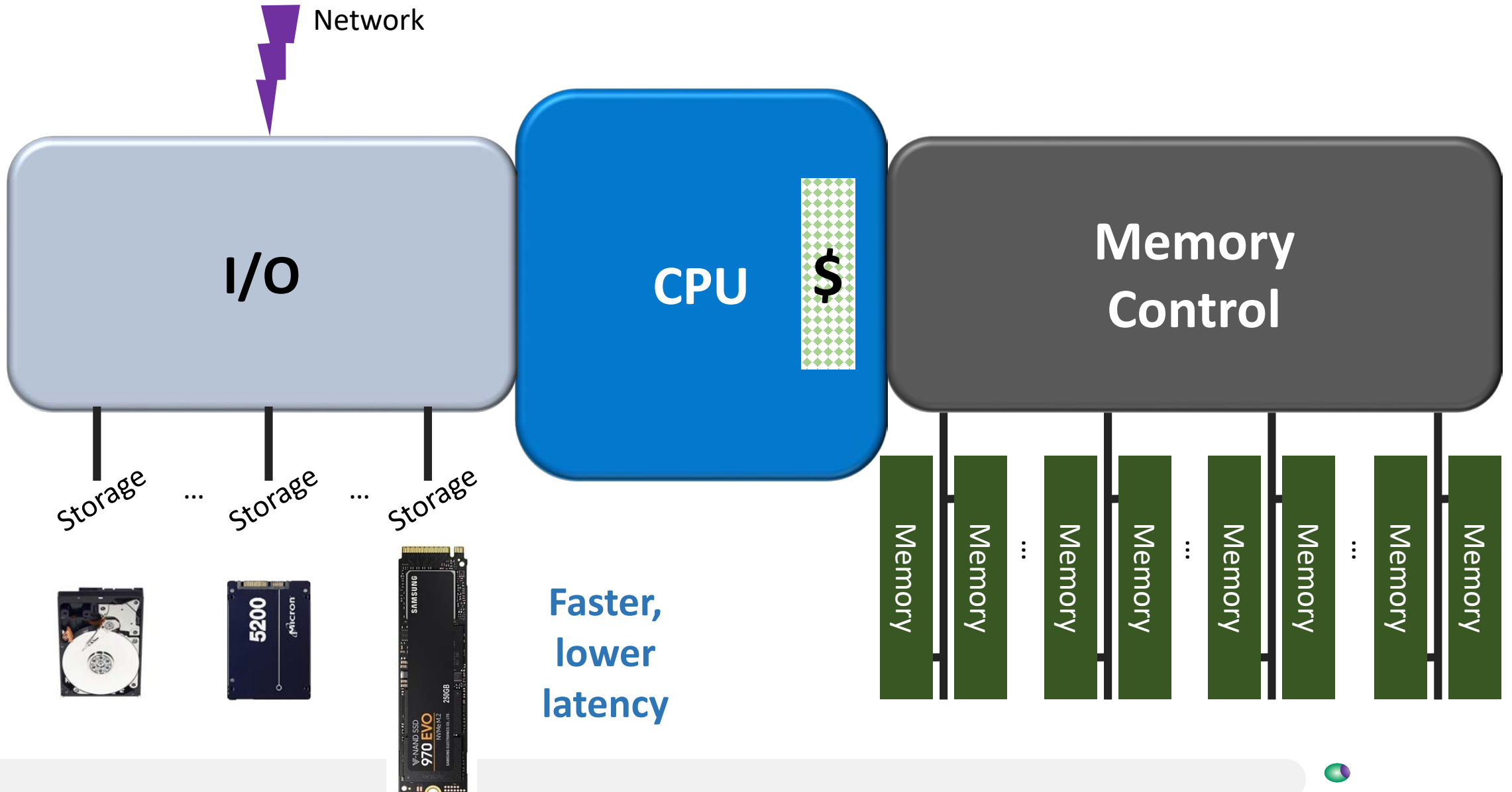


...move non-volatile storage closer to the CPU





Traditional Server Architecture Review

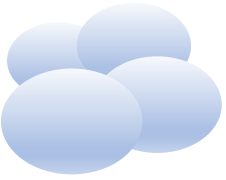




The Search for

THE HOLY GRAIL





When we no longer
fear power failure...

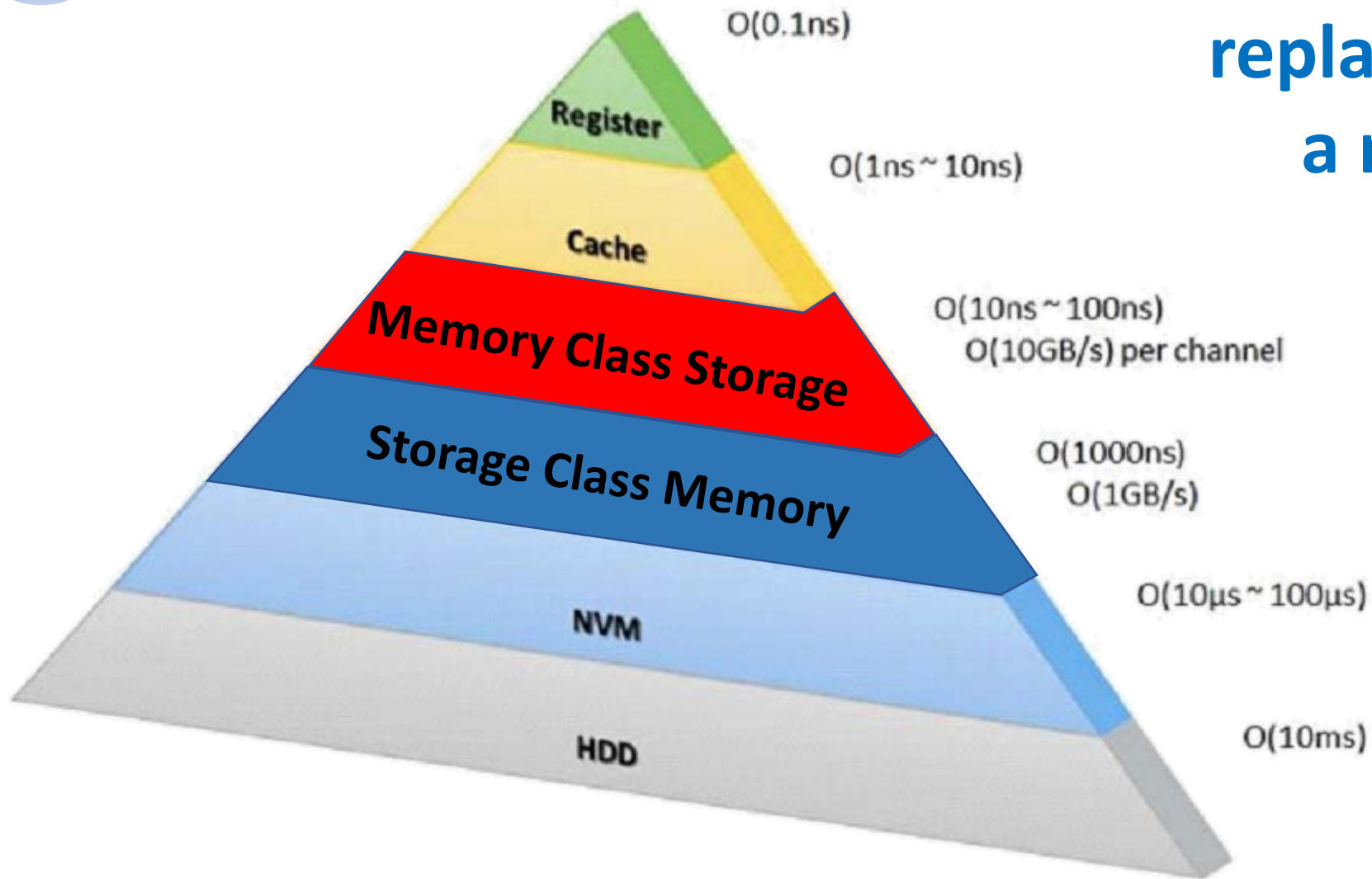


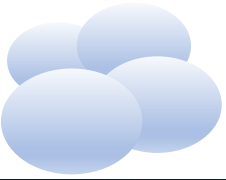
DATA PERSISTENCE



What if you could
replace DRAM with
a non-volatile
memory?

You'd call it
**Memory
Class
Storage**





When was the last time you read about a new volatile memory?

NRAM™

MRAM

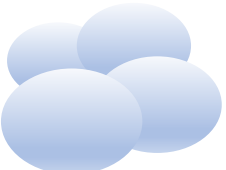
PCM

3DXP

ReRAM

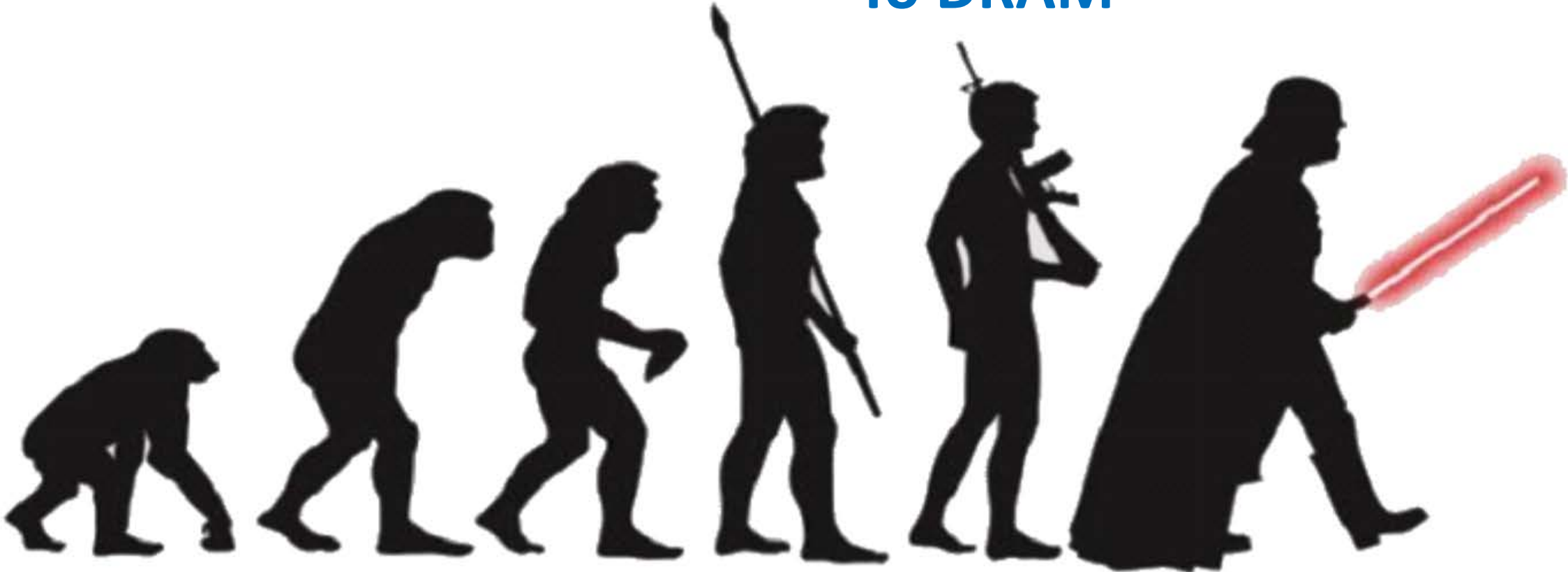
The non-volatile memory revolution is under way





To NVRAM

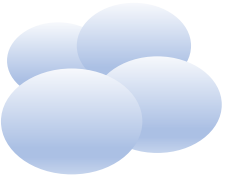
To DRAM



To core memory

From vacuum tubes



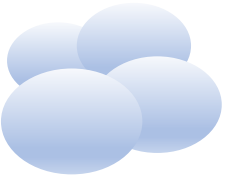


THIS is why the term “Persistent Memory” is **insufficient**

The industry must distinguish between **deterministic**
and **non-deterministic** persistent memory

Only “**Memory Class Storage**” is
fully deterministic **AND** persistent



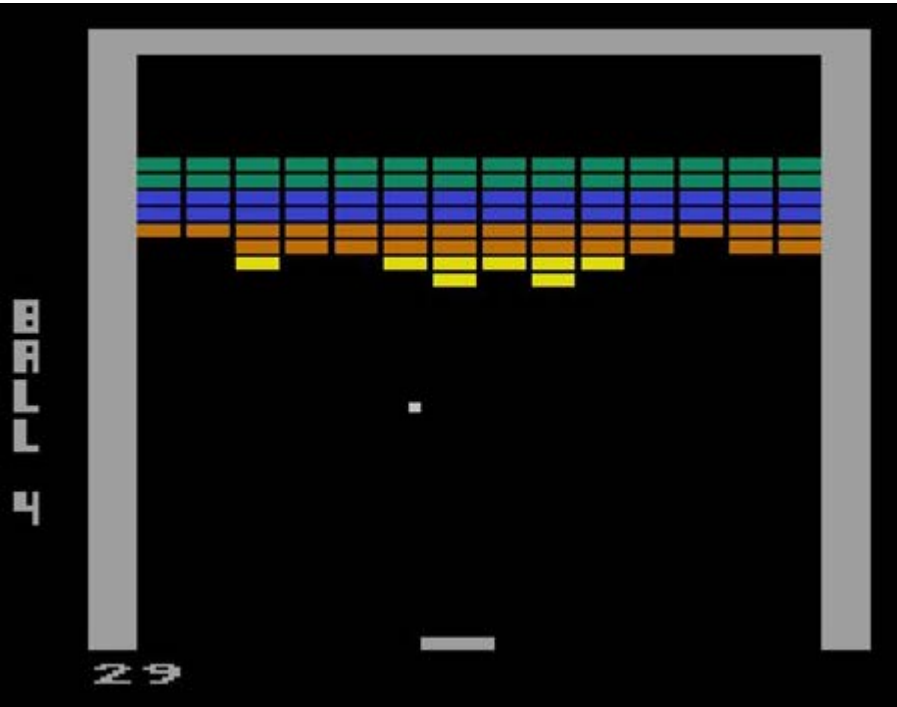
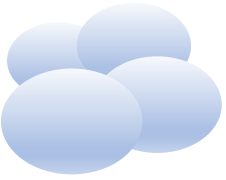


3DXpoint
MRAM
ReRAM



**Not all “persistence”
is created equal**





Flash Architecture	Layers of Cells	Bits per Cell	Number of Cell Voltage States	Cell Endurance ¹ (P/E Cycles)
Planar SLC	1	1	2	~100,000
Planar MLC	1	2	4	~3,000
Planar eMLC/iMLC/pSLC	1	1	2	~20,000
Planar TLC	1	3	8	<1,000
Vertical SLC	Varies, 64 typical	1	2	TBD ²
Vertical MLC	Varies, 64 typical	2	4	TBD ²

	375GB Intel DC P4800X	1.6TB Intel DC P3700	1.6TB Intel DC P3608	2.4TB Micron 9100 Max	2.7TB Mangstor MX6300
Endurance Per Usable GB	32.8 TB	27.35 TB	5.45 TB	2.73 TB	12.77 TB
Usable Capacity	375GB	1.6TB	1.6TB	2.4TB	2.7TB
Raw Capacity	448GB	2TB	2.3TB	4TB	4TB
Spare Area / %	73GB / 16.3%	400GB / 20%	700GB / 30.4%	1600GB / 40%	1300GB / 32.5%
Media Endurance Per Raw GB	27.4TB	21.9TB	3.8TB	1.6TB	8.63TB

“Write endurance” determines HOW persistent

Wear leveling needed if writes are limited



Application Class	Workload	Active Use (power on)	Retention Use (power off)	Functional Failure Rqmt (FFR)	UBER
Client	Client	40°C 8 hrs/day	30°C 1 year	≤3%	≤10 ⁻¹⁵
Enterprise	Enterprise	55°C 24hrs/day	40°C 3 months	≤3%	≤10 ⁻¹⁶

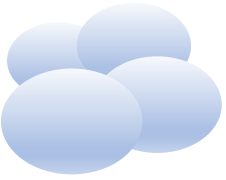
Temperature sensitivity impacts long term retention

Client

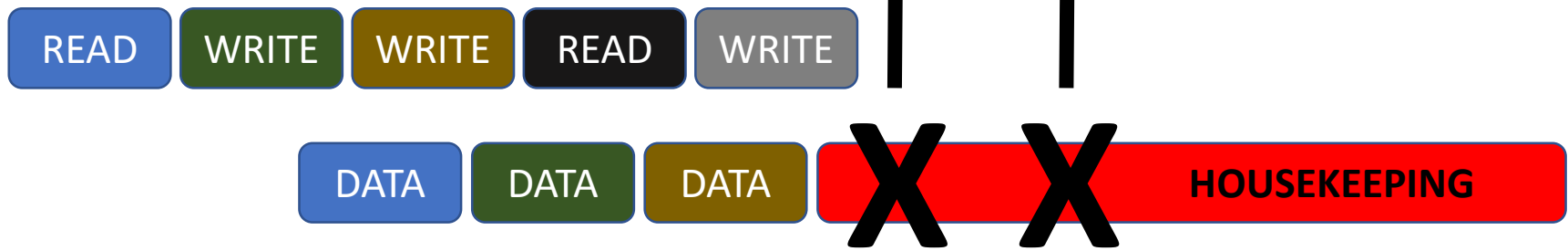
Power Off Temperature	55	1	1	2	2	3	5	8
	50	2	2	3	4	6	9	15
	45	4	4	5	7	10	17	27
	40	7	8	10	14	20	31	52
	35	14	16	20	26	38	61	101
	30	28	32	39	52	76	120	199
	25	58	65	79	105	155	244	404
	25	30	35	40	45	50	55	
Active temp								

Weeks of Data Retention





DRAM interface is deterministic
Data latency is FIXED



Any endurance limit breaks determinism





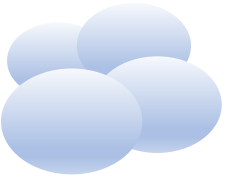
Memory Class Storage

Full DRAM Speed

No endurance limits

Fully deterministic



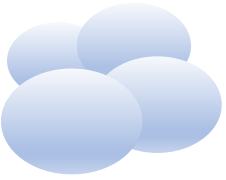


NVRAM

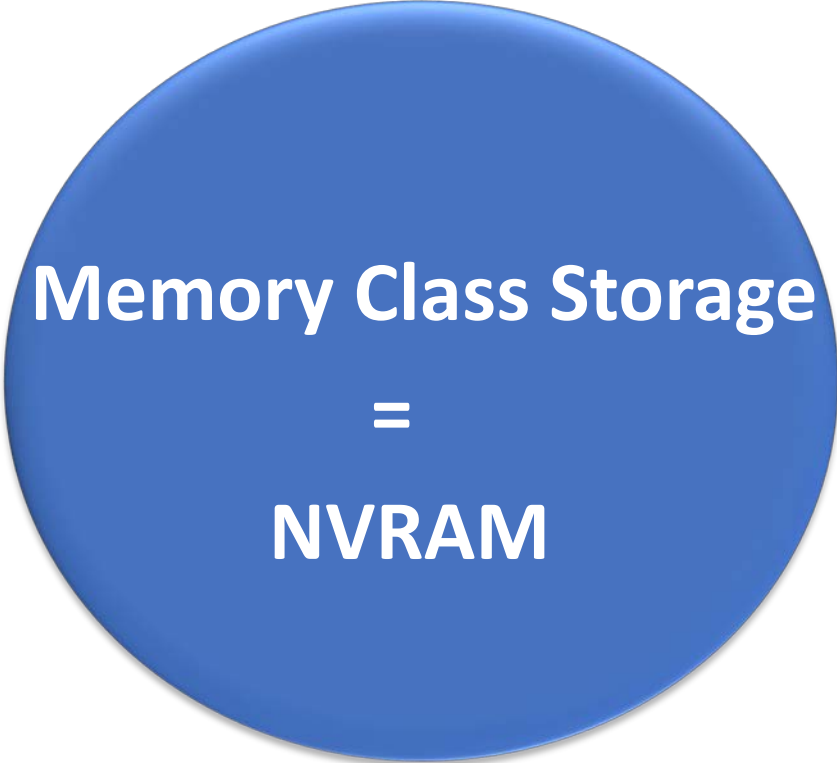
is a

Memory Class Storage

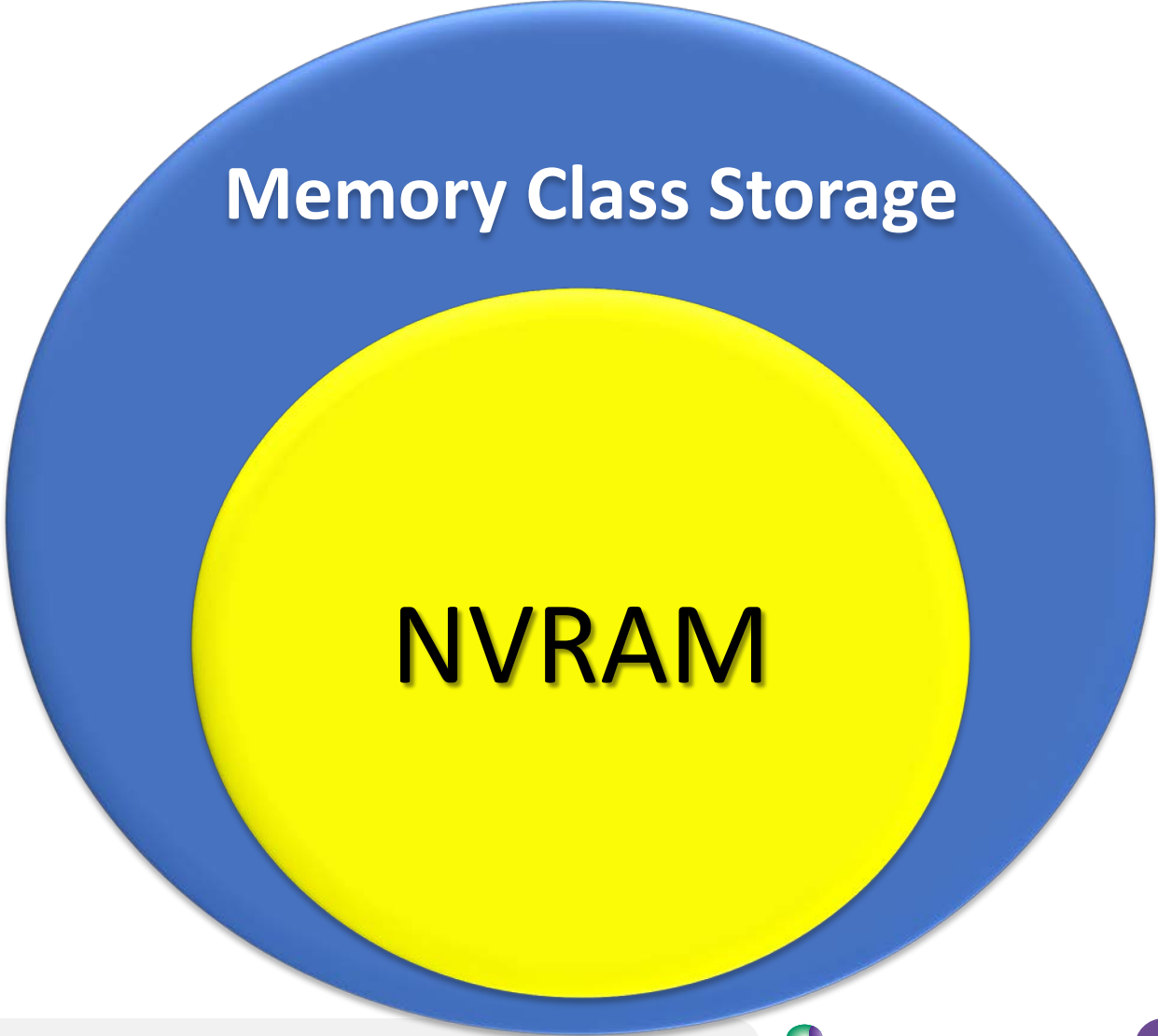


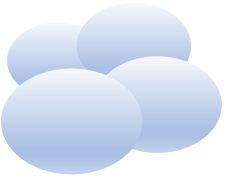


In the future?



For now...

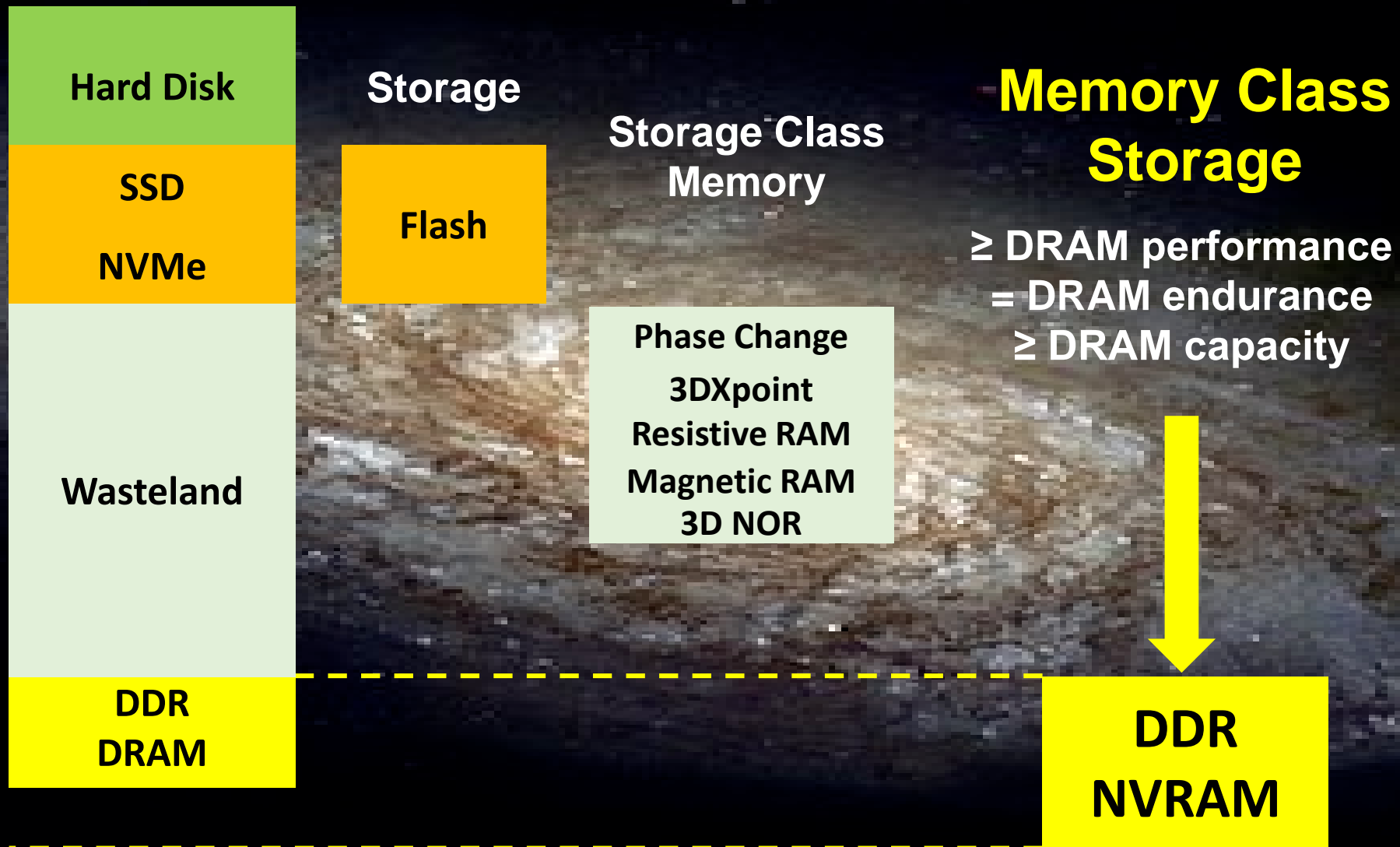
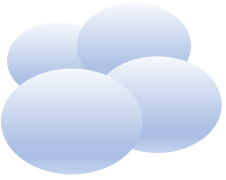


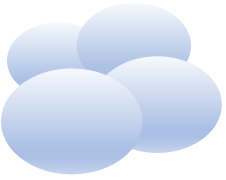


Storage Class Memory

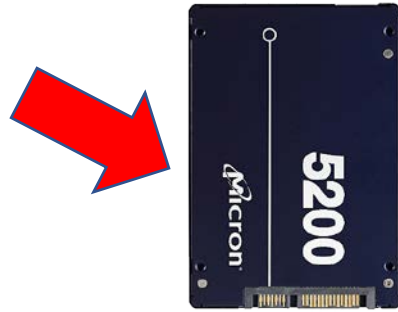
Is NOT a

Memory Class Storage





Non-Deterministic



Non-Deterministic



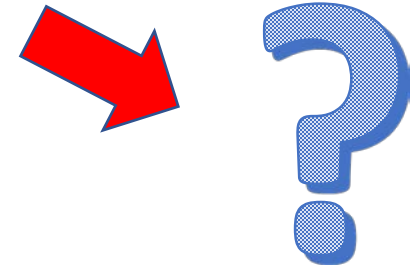
Non-Deterministic



Deterministic

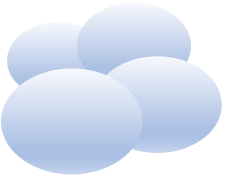


Deterministic

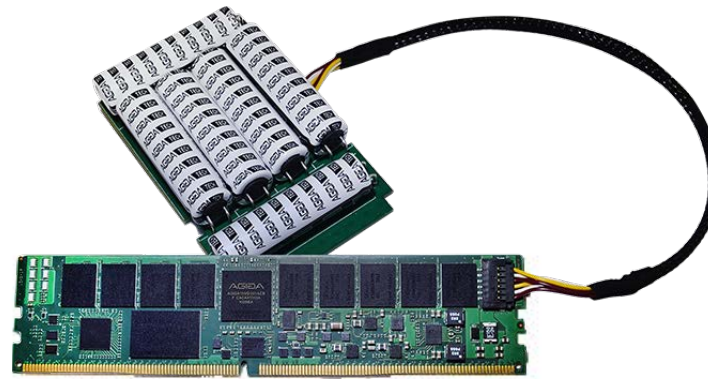


Deterministic





DRAM



NVDIMM-N



Optane

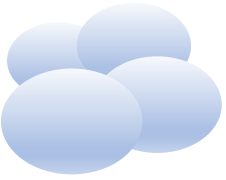


NVRAM Memory Class Storage



NVDIMM-P

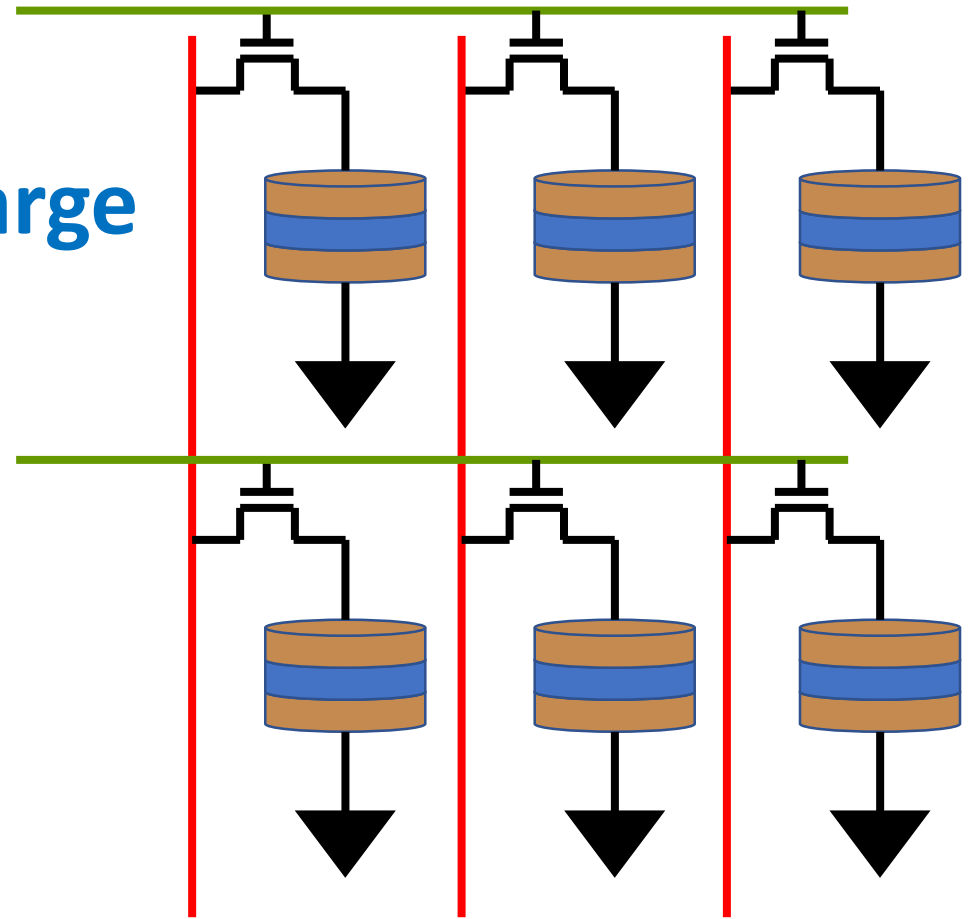
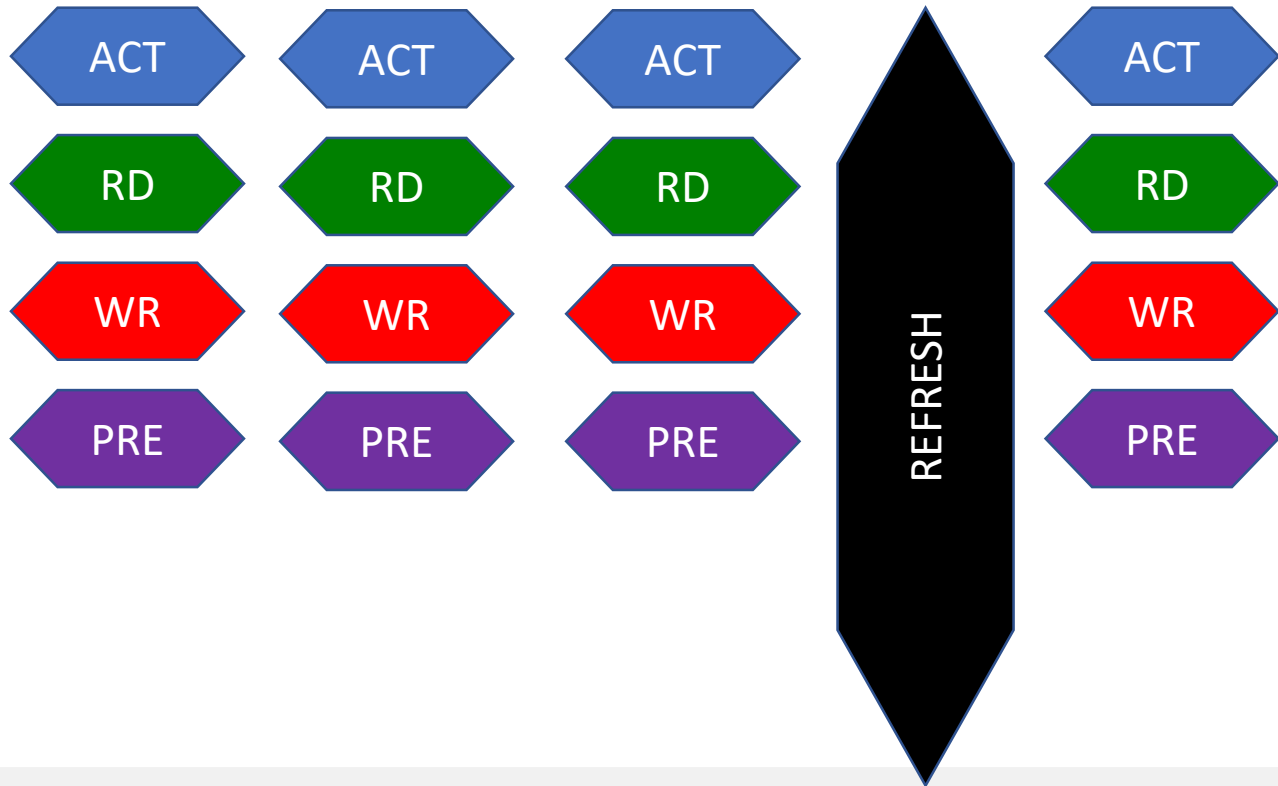




DRAM

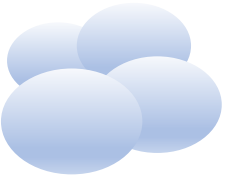
Itty bitty leaky capacitors lose charge

On power fail, you lose



Refresh time consumes up to 15% of bandwidth



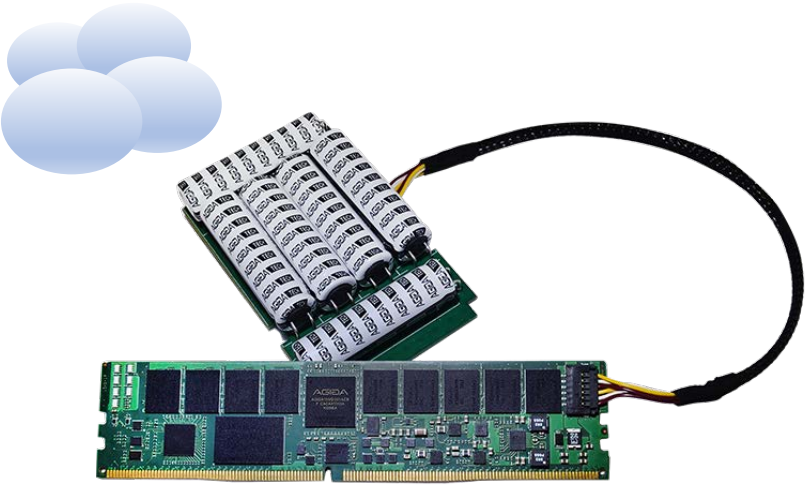


DRAM

Run

FAIL!





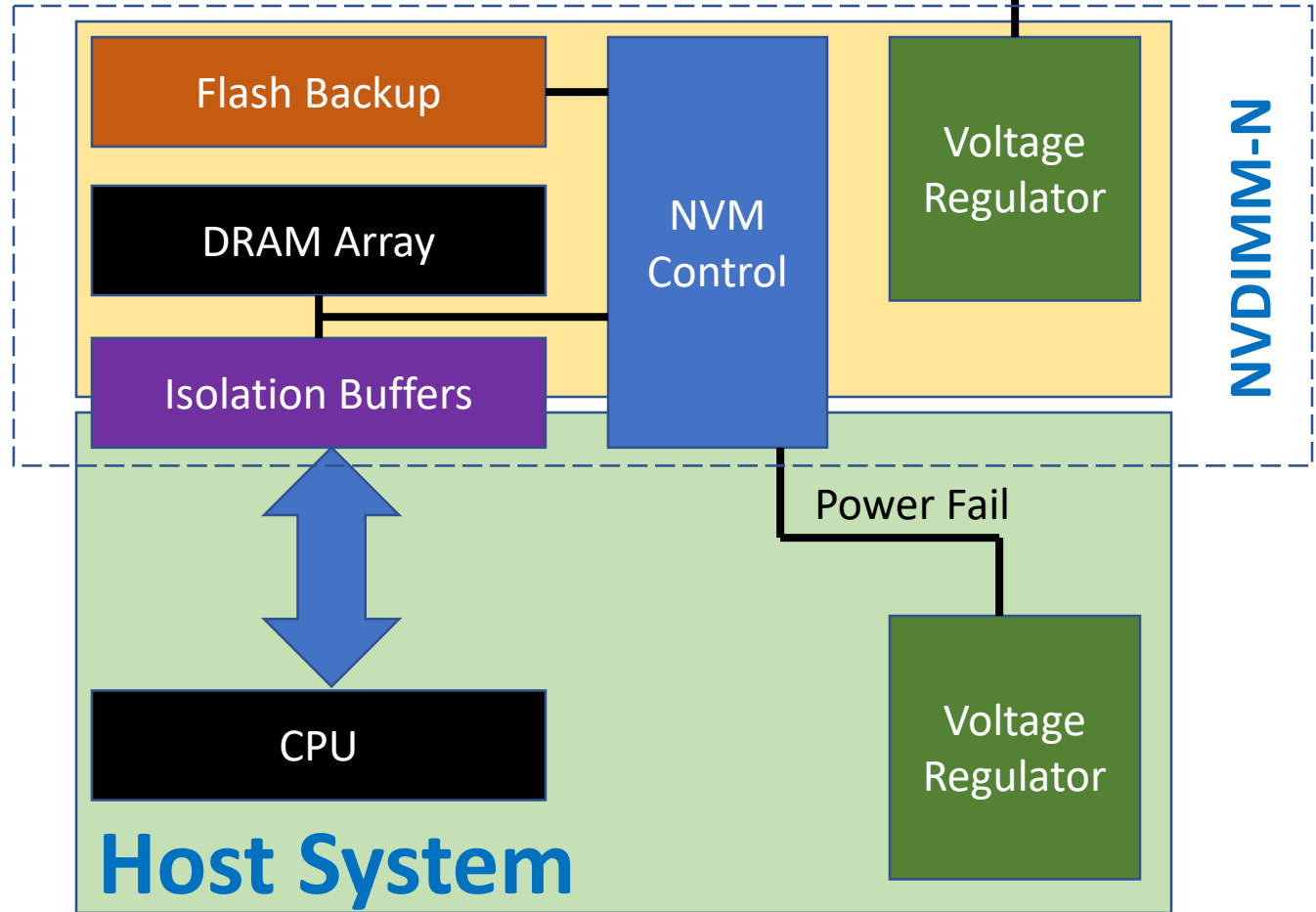
NVDIMM-N

Use DRAM normally

On Power Fail, copy to Flash

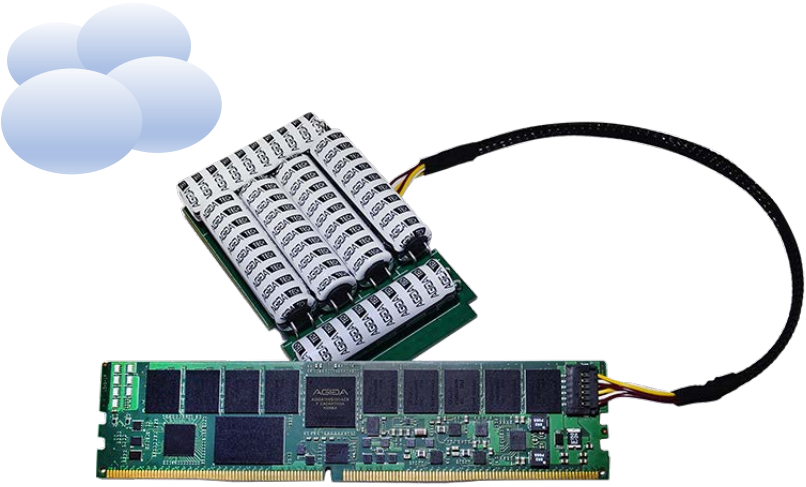
Power restored, copy to DRAM

Energy Source



NVDIMM-N

Host System



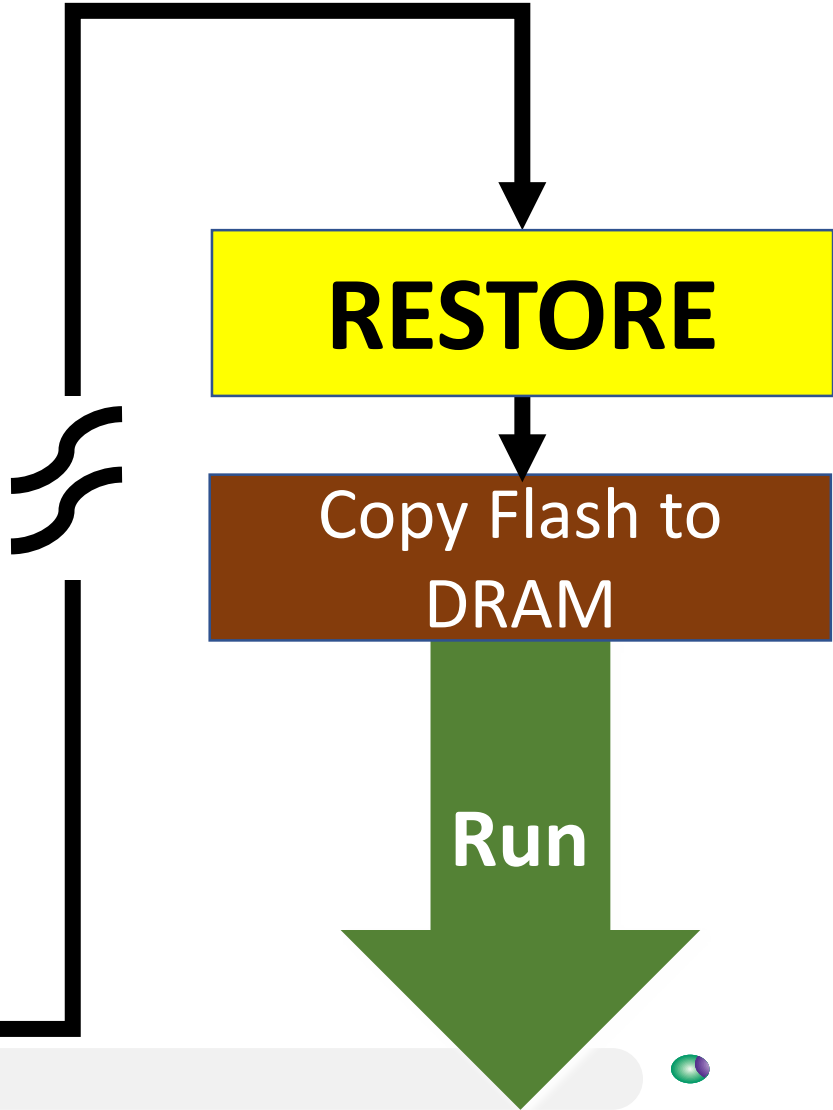
NVDIMM-N



FAIL!

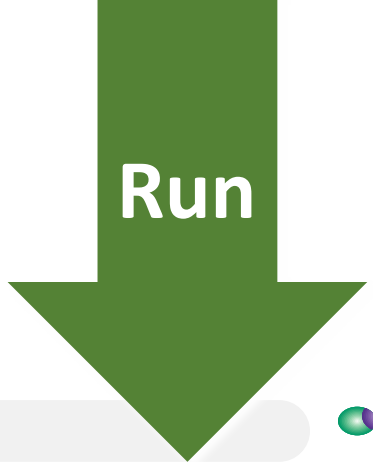
Switch to Battery Power

Copy DRAM to Flash

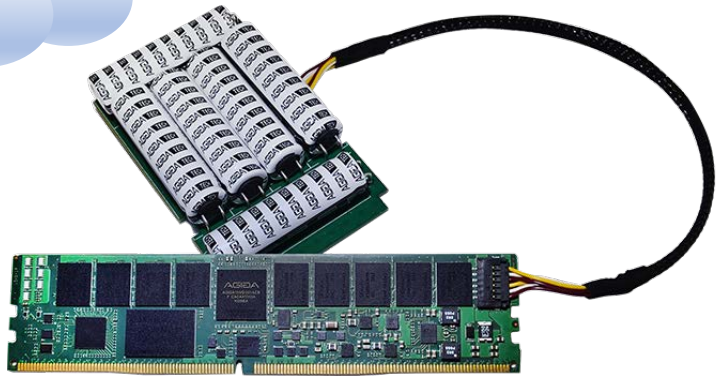


RESTORE

Copy Flash to DRAM



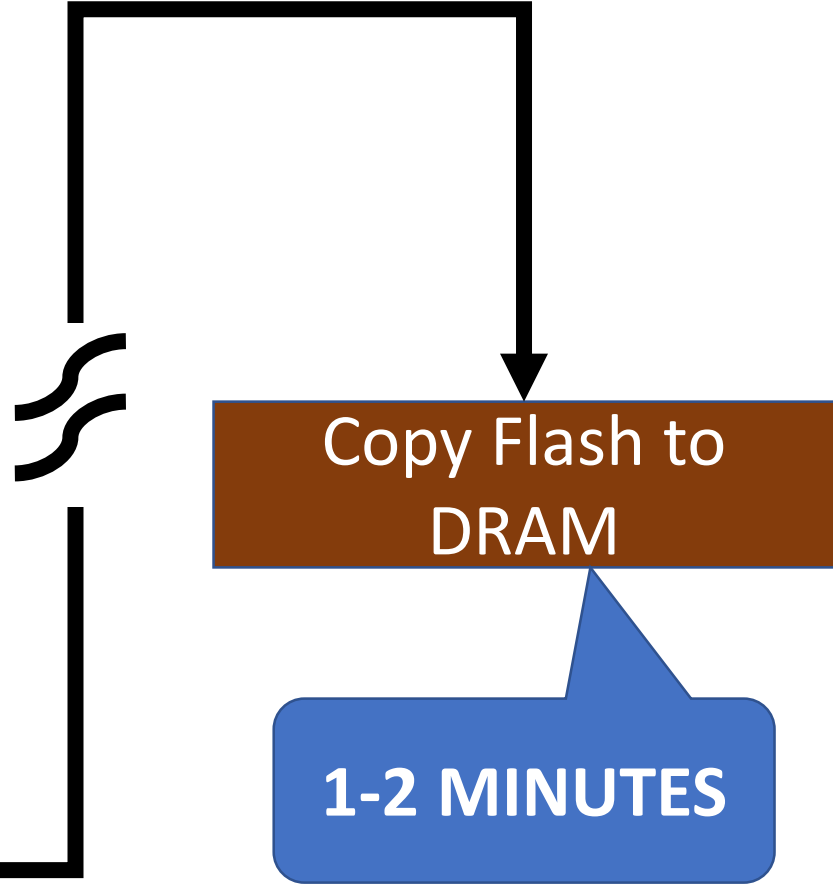
Amazon.com Goes Down, Loses \$66,240 Per Minute



NVDIMM-N

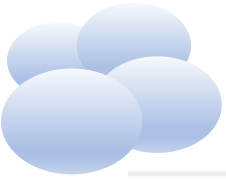
1-2 MINUTES

Copy DRAM to Flash



1-2 MINUTES

One power fail cycle pays for a LOT of protection



Faster than Flash!!!

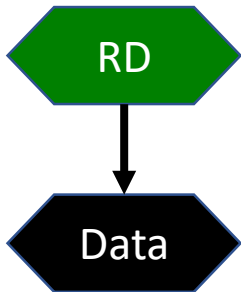
But vs DRAM? Meh

Decent capacity, though

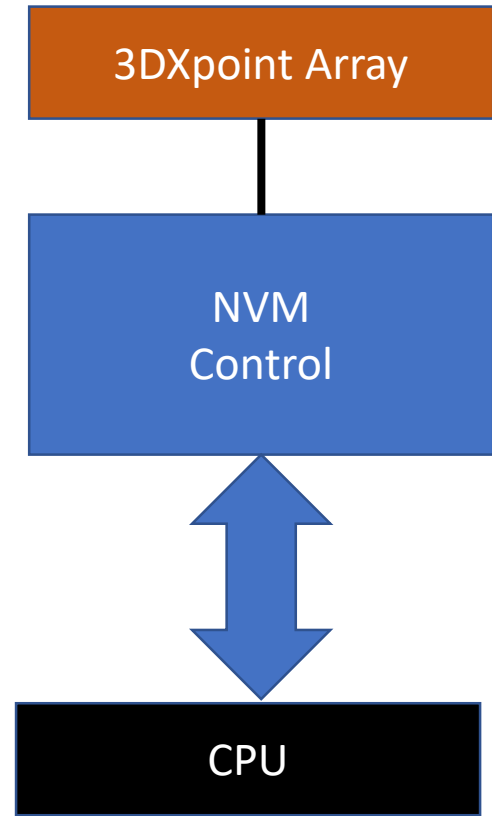
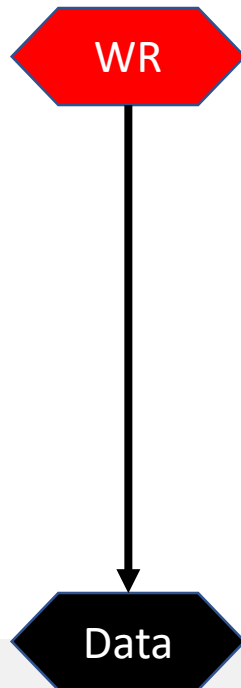


Optane

Reads are slow



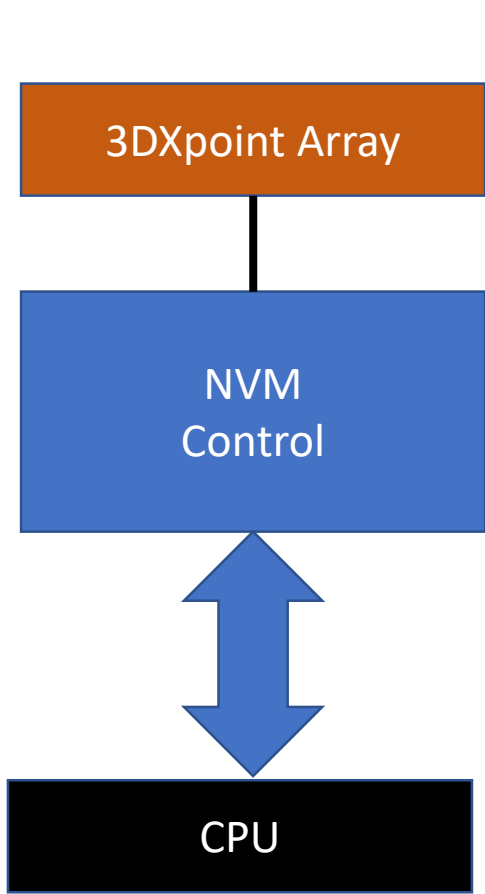
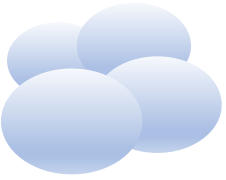
Writes are deathly slow



Host System

**Could be used as a very slow DRAM
but more common as expansion**



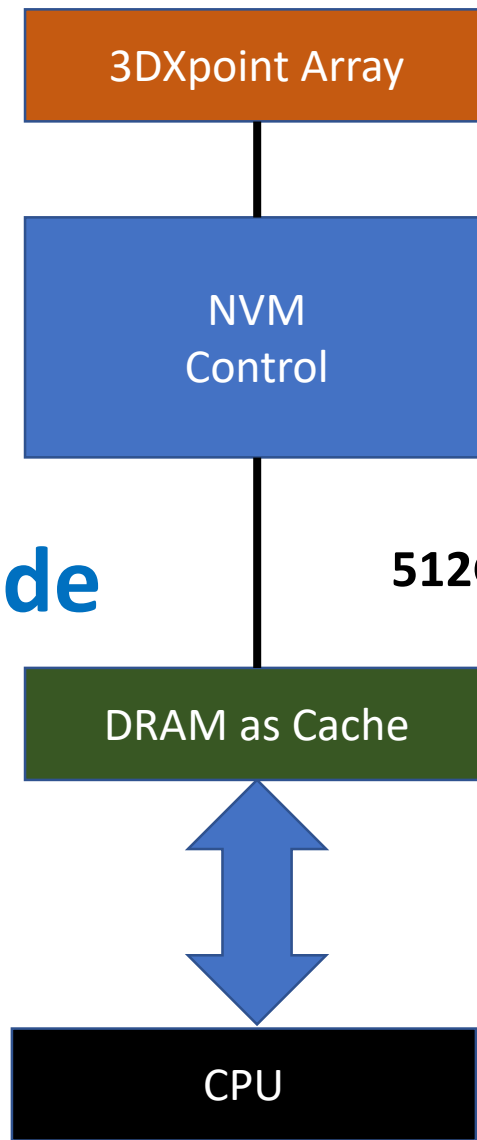


512GB = 512GB

App Direct

Memory Mode

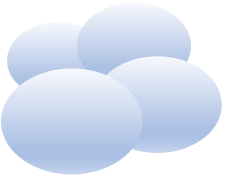
Optane



512GB + 64GB = 512GB

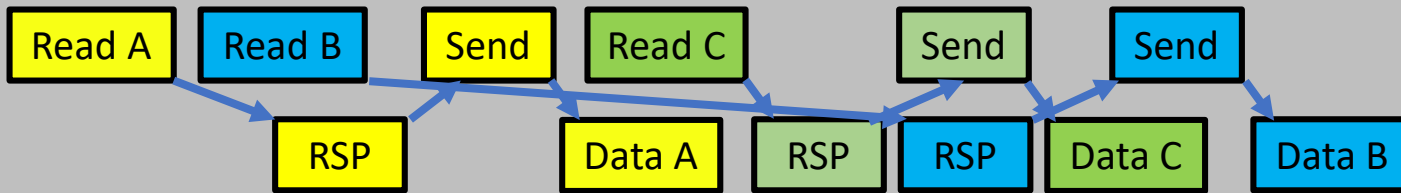
Host System





NVDIMM-P

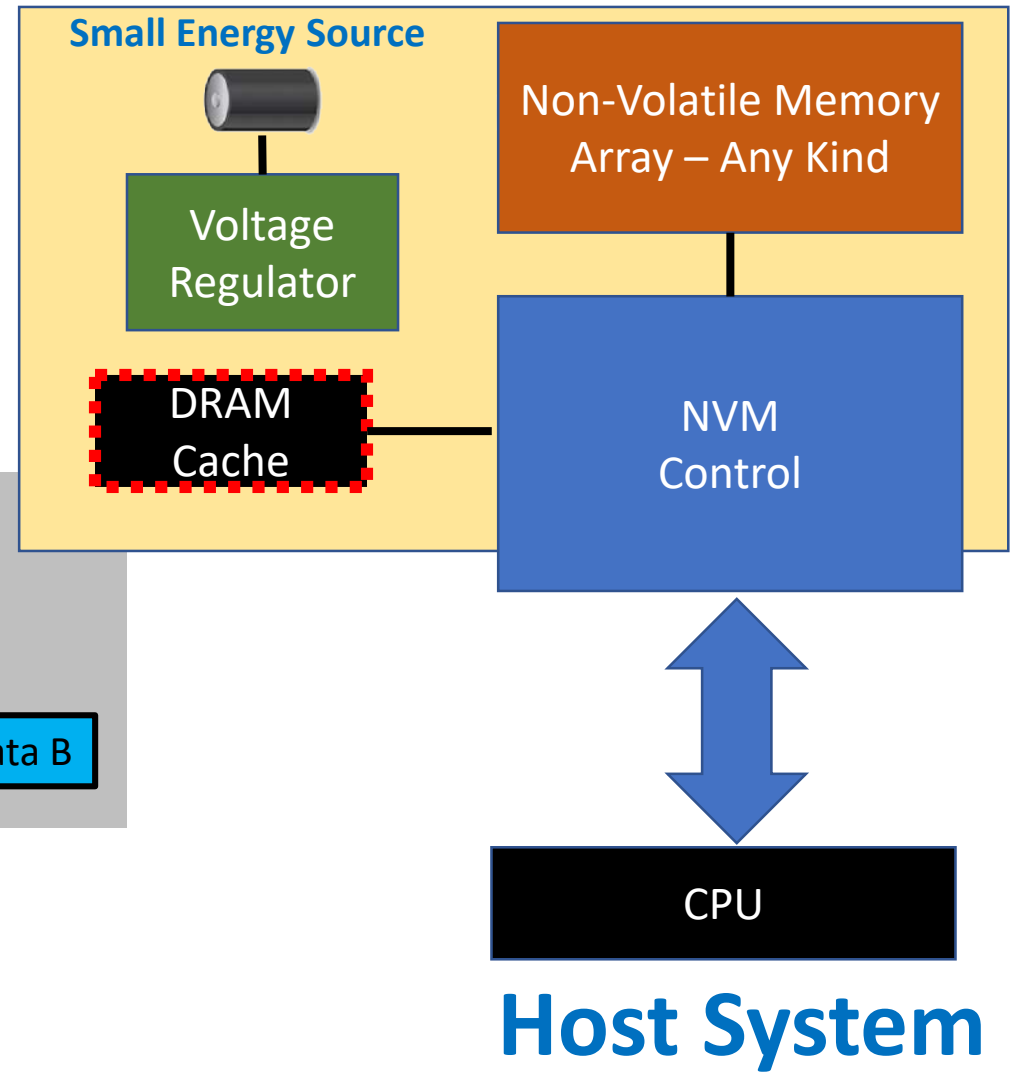
NVDIMM-P Protocol



New non-deterministic protocol

Not backward compatible with DDR

Requires NVDIMM-P aware CPU





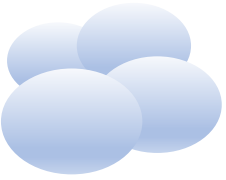
**Volatile Mode
No Persistence**

**Battery Backup
ala NVDIMM-N**

**NVDIMM-P
Persistence
Options**

**Explicit FLUSH
Command**

**Reduced
Energy,
Cacheless**



NVRAM



DRAM speed

Non-volatility

Unlimited write endurance

Wide temperature range

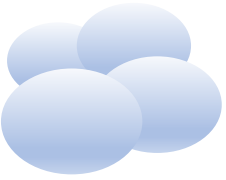
Scalable beyond DRAM

Flexible fabrication & application

Low power

Low cost





Drop in replacement for DRAM

Fully Deterministic

Permanently persistent

Always available

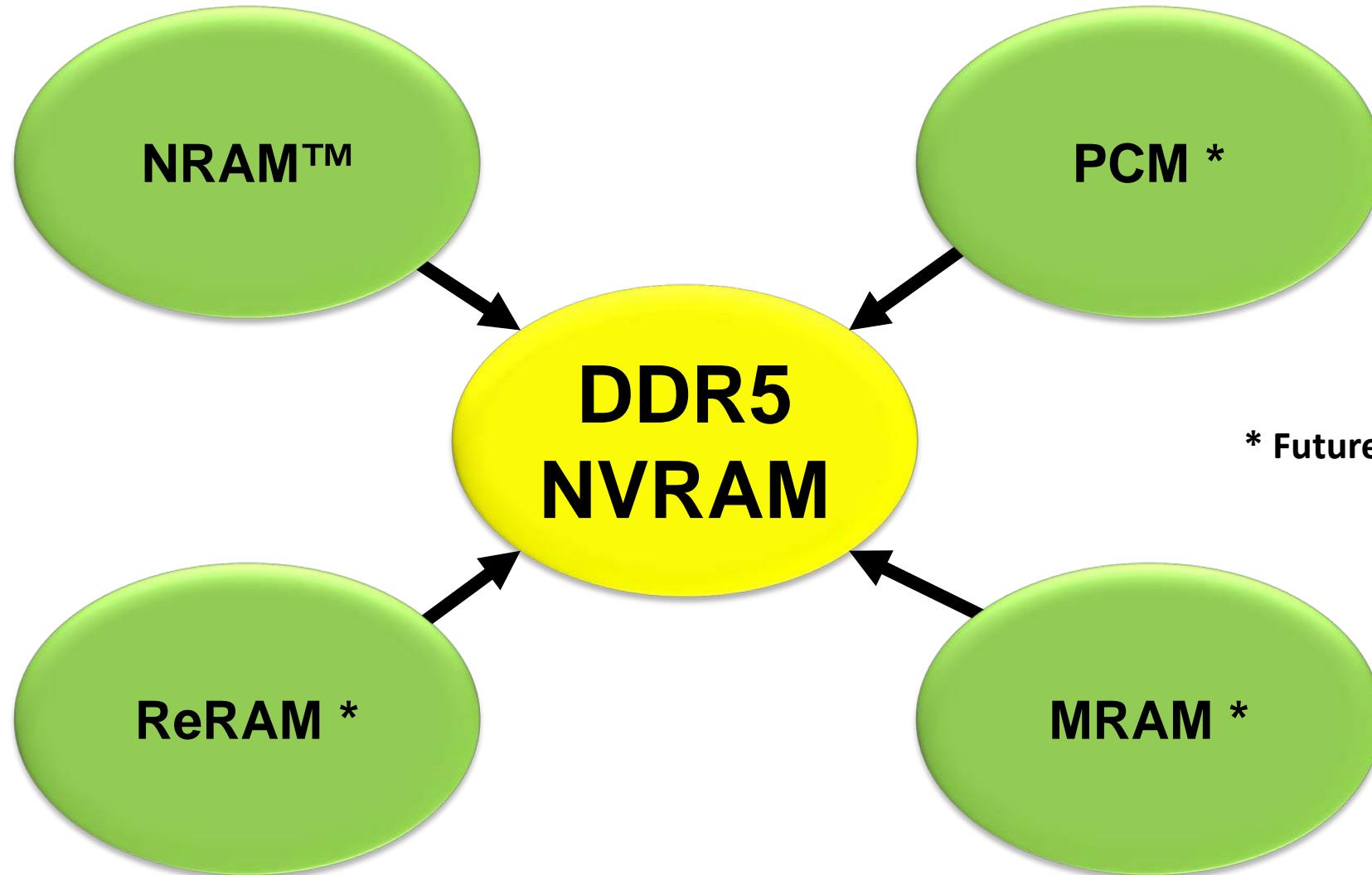
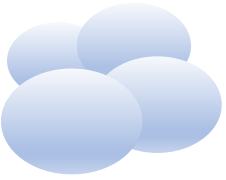


NVRAM Memory Class Storage

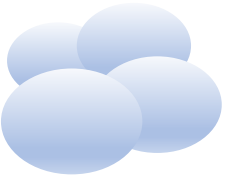


Host System





* Future generation devices



JEDEC STANDARD

DDR5 SDRAM

JESD79-5

JEDEC SOLID STATE TECHNOLOGY ASSOCIATION

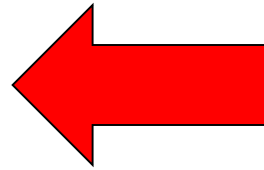


JEDEC STANDARD

DDR5 Non-Volatile Random Access Memory (DDR5 NVRAM)
Addendum #1 to JESD79-5

JESD79-5-1

JEDEC SOLID STATE TECHNOLOGY ASSOCIATION

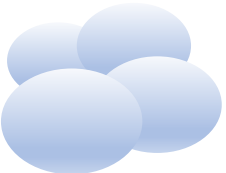


“DDR5 NVRAM is like a DRAM and...”

PROPOSED

PROPOSED





Comparing DRAM & NVRAM



No refresh is required

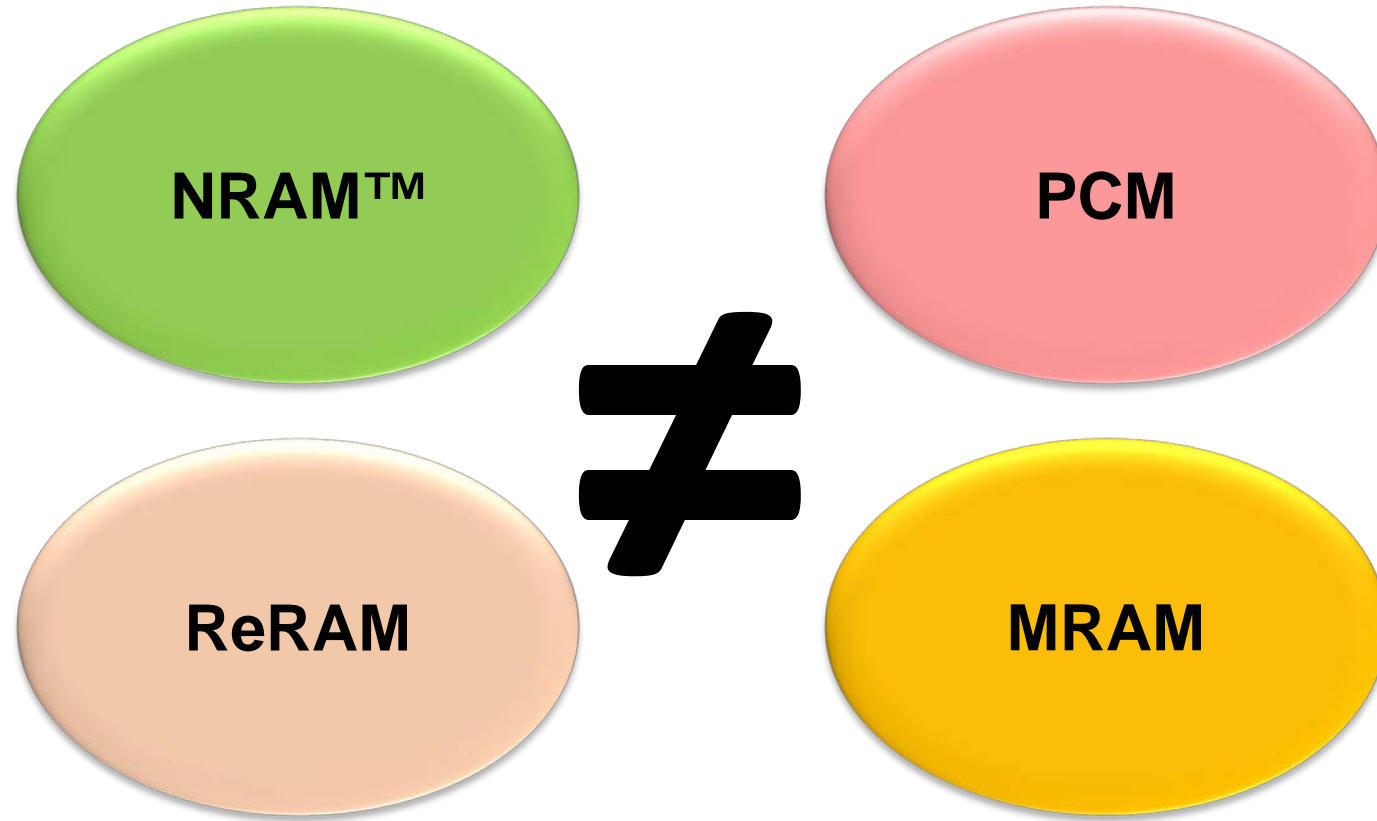
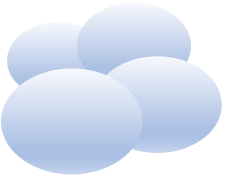
“Self refresh” can be power OFF

Some timing differences (but deterministic!)

Data persistence definitions

Greater per-die capacity





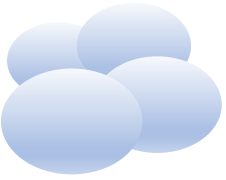
Timings

Precharge
requirement

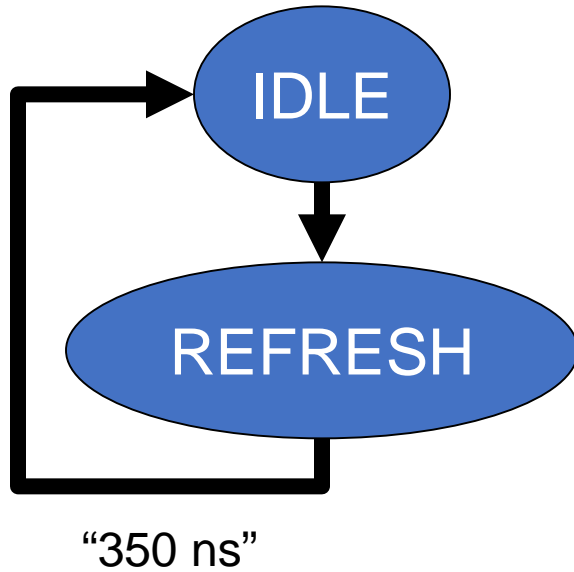
Persistence
definition

DDR5 NVRAM Specification brings coherence

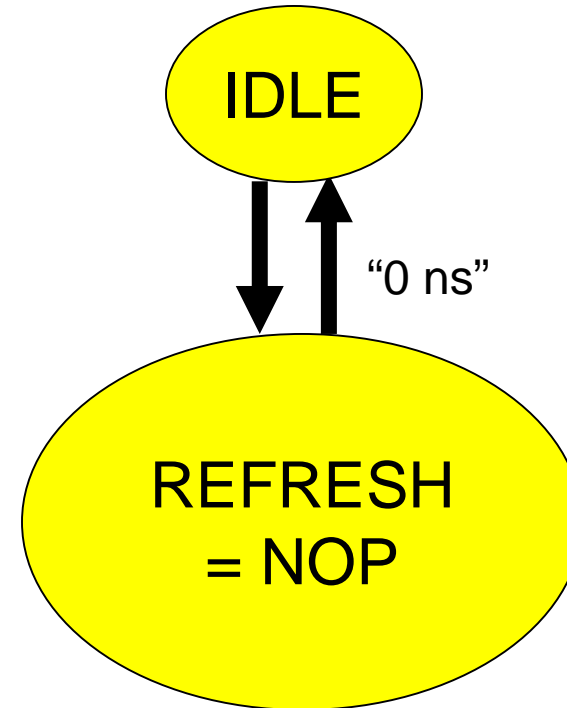




DRAM



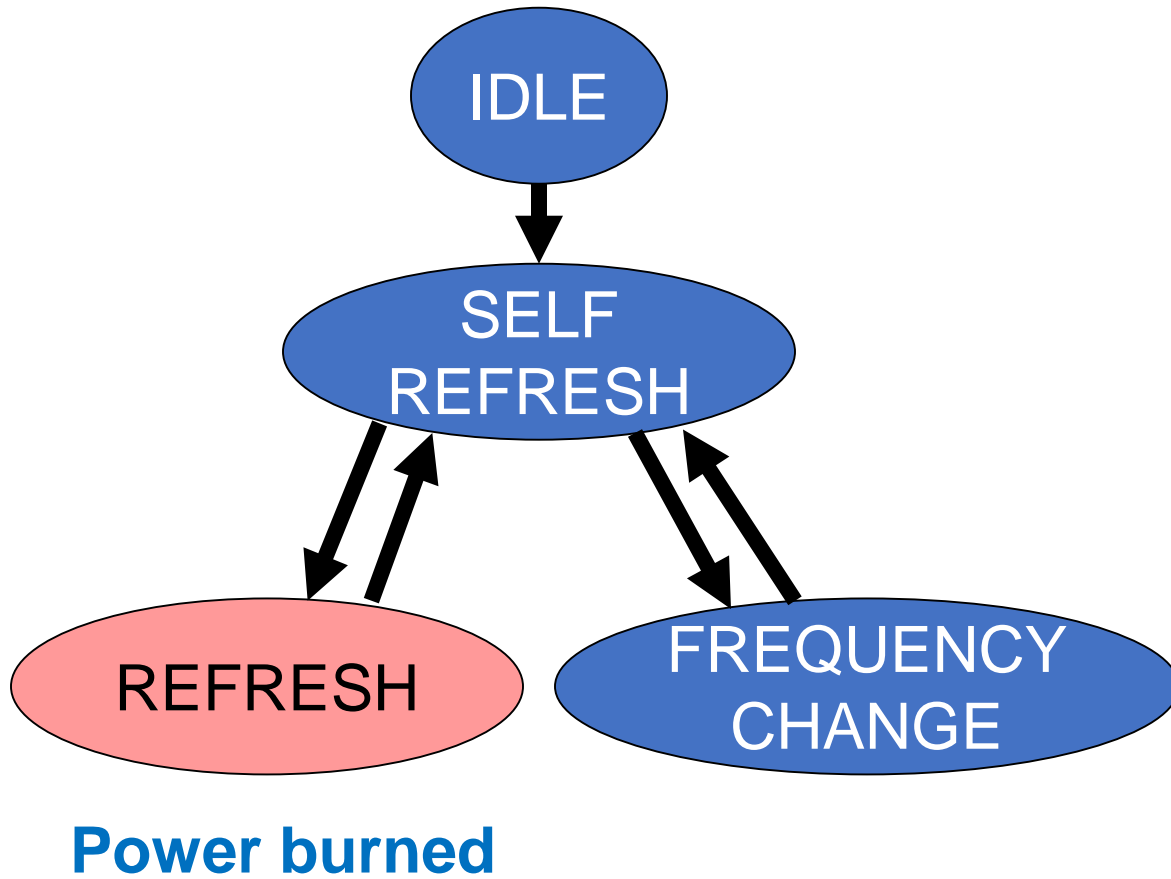
NVRAM



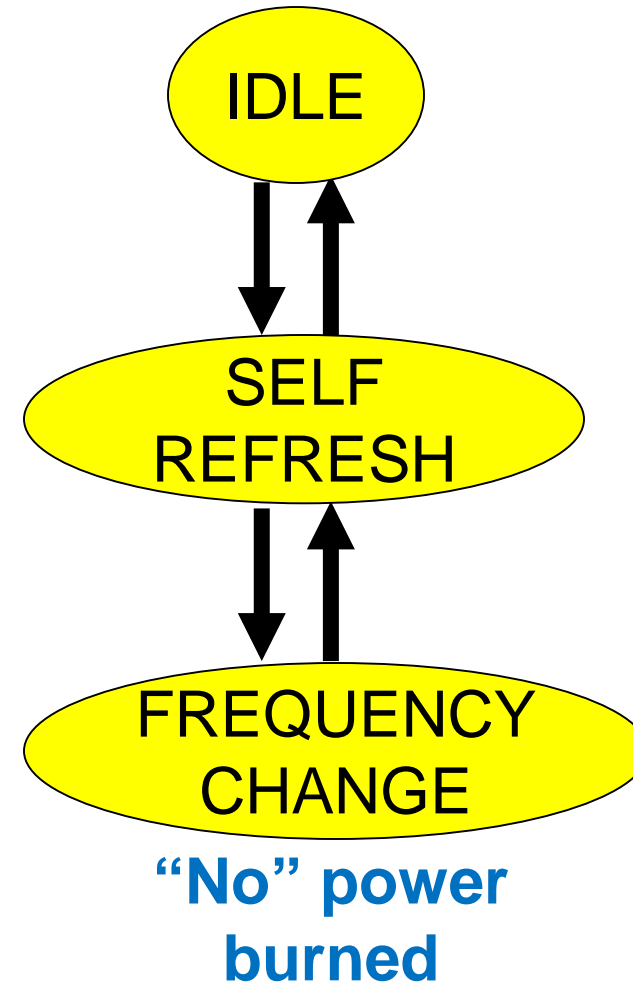
**Refresh command is not needed
Decoded as NOP for compatibility**

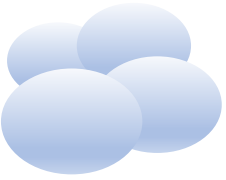


DRAM

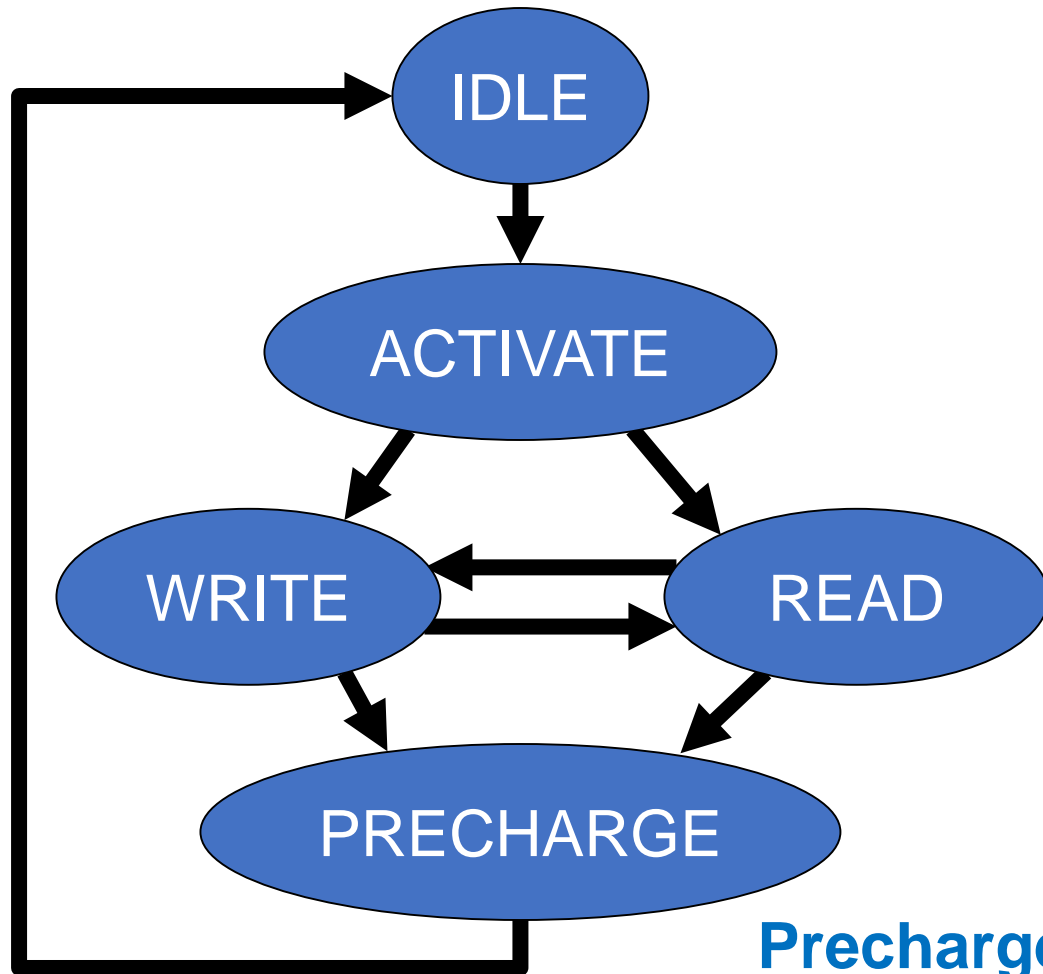


NVRAM

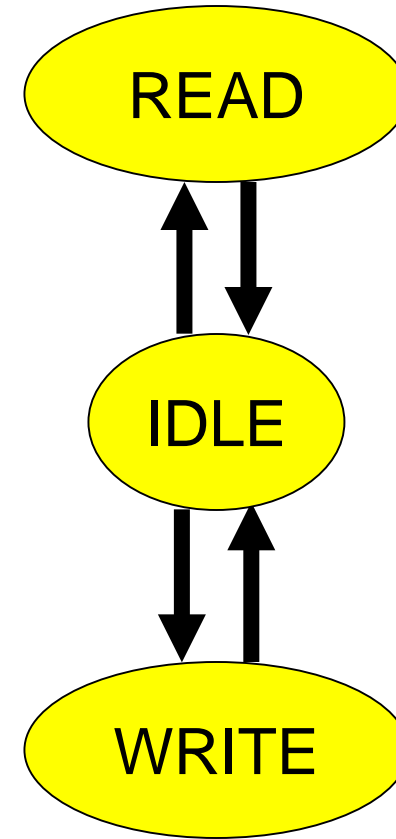




DRAM

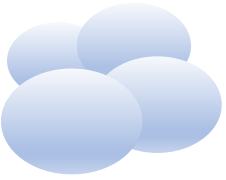


NVRAM



**Precharge command is not needed
Decoded as NOP for compatibility**





Persistence Definitions*

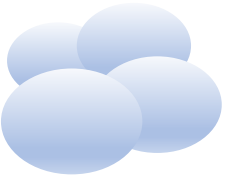
**Intrinsic:
Immediately
After
WRITE**

**Extrinsic:
After
FLUSH
Command**

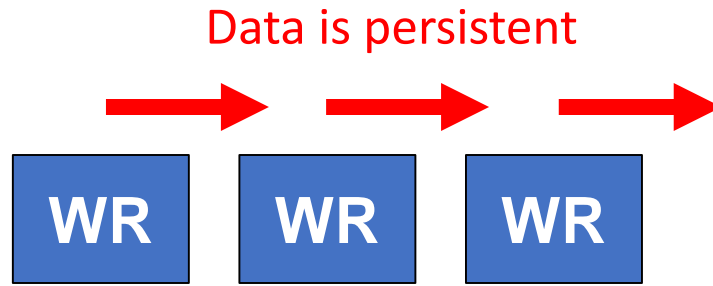
**Power Fail:
On
NVRAM
RESET**

* Discussions on-going

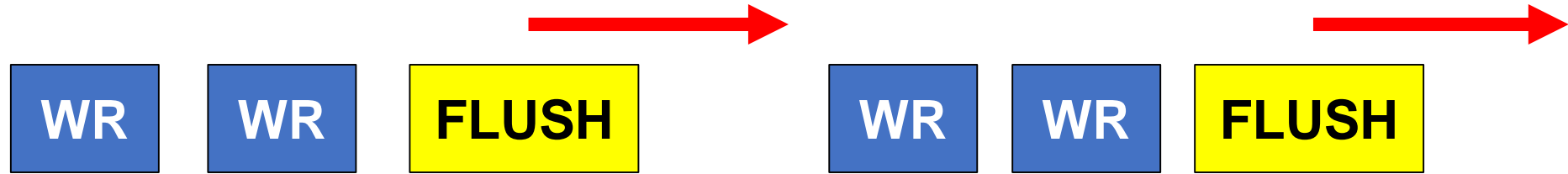




Intrinsic Persistence



Extrinsic Persistence



Power Fail Persistence



* Discussions on-going





DDR5 DRAM
is limited
to 32Gb per die

DDR5 NVRAM
enables up to
128Tb per die



DDR5 SDRAM

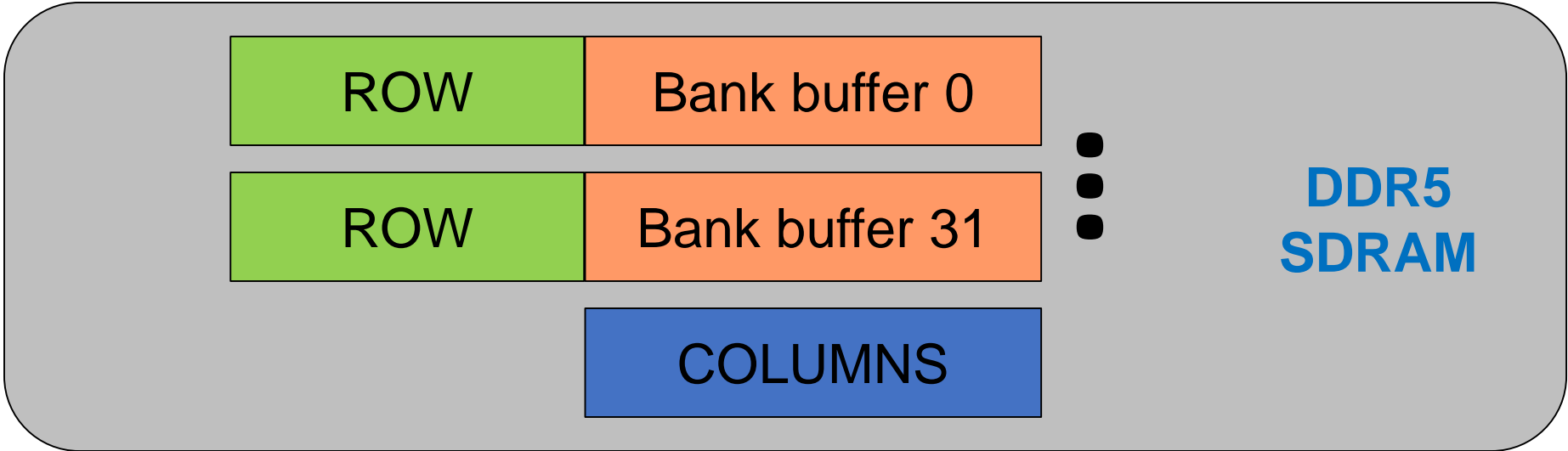
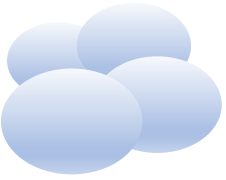


DDR5 NVRAM

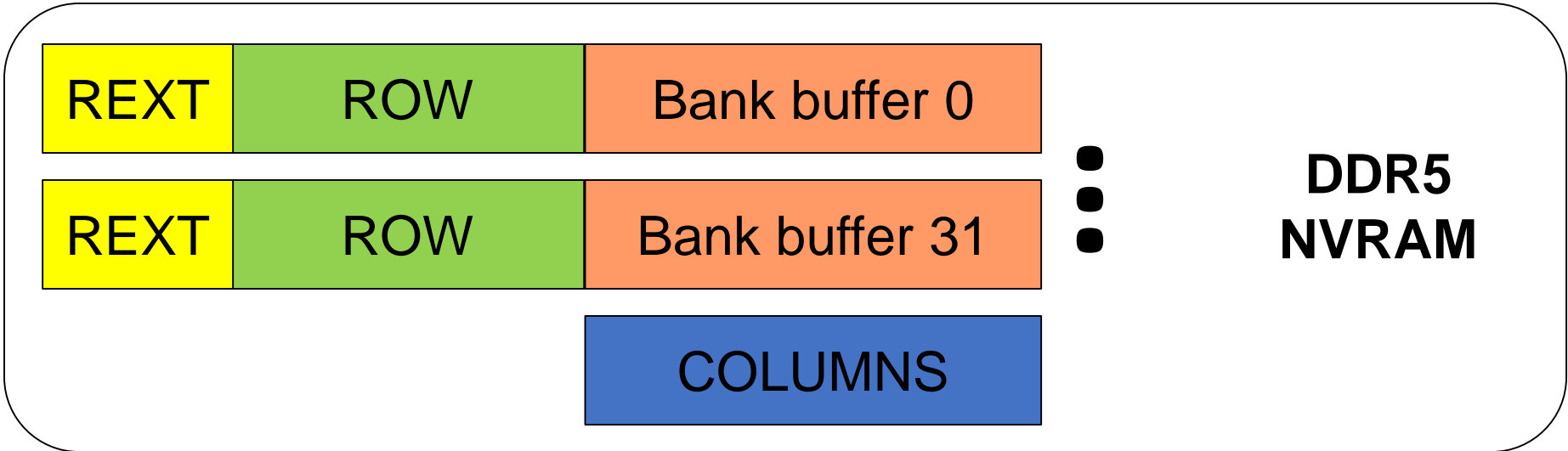


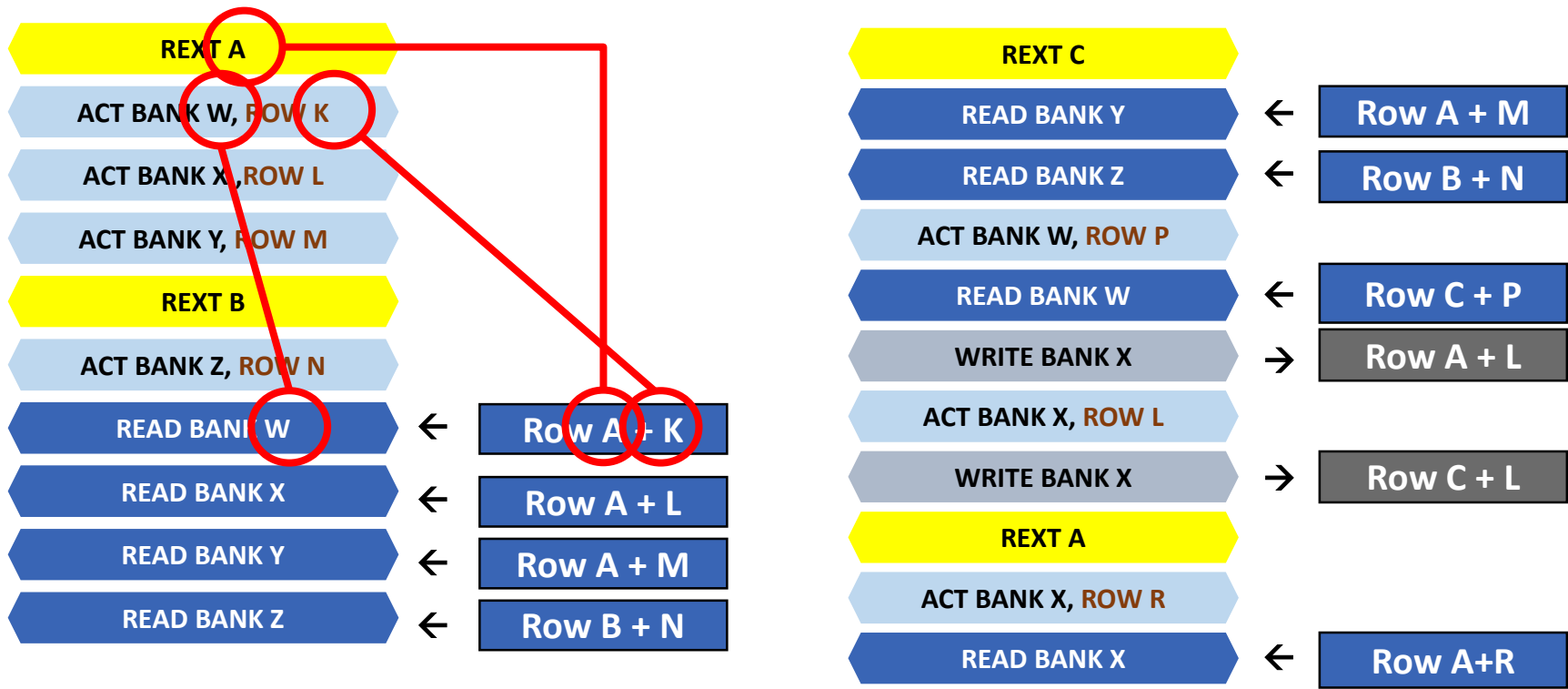
Row Extension adds up to 12 more bits of addressing

Backward compatible with DDR5 – Acts like REXT = 0 until needed



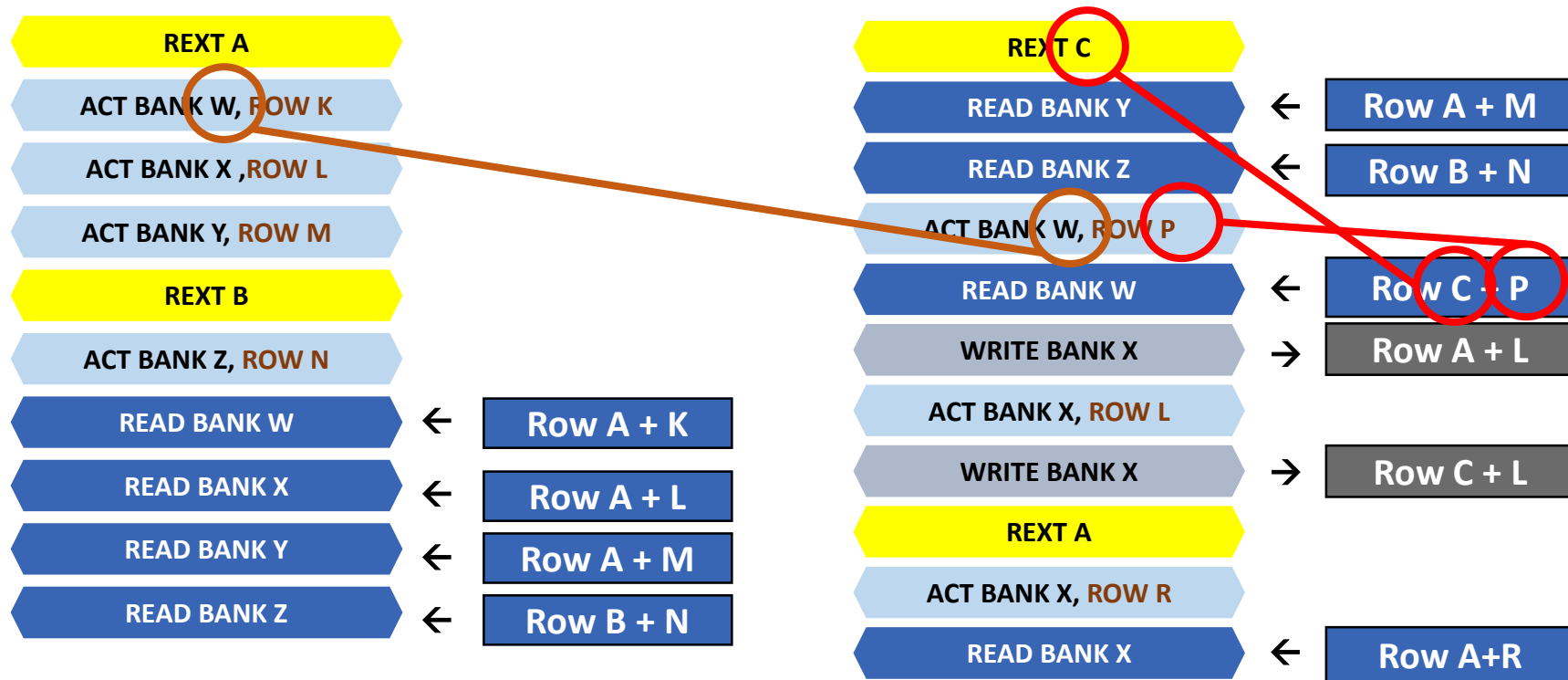
"ROW" includes bank group & bank...



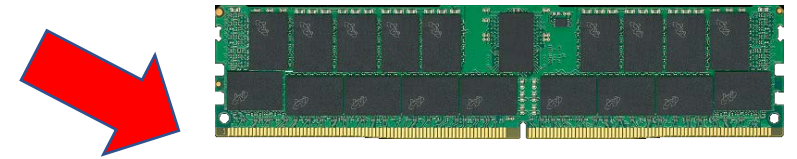
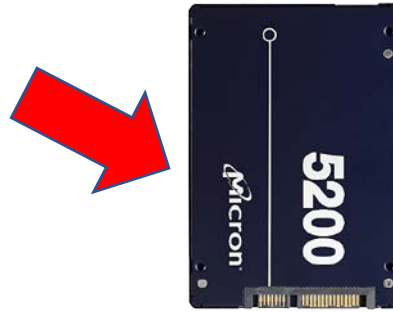
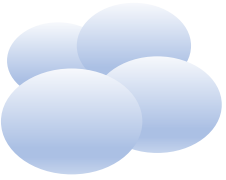


Row Extension Example



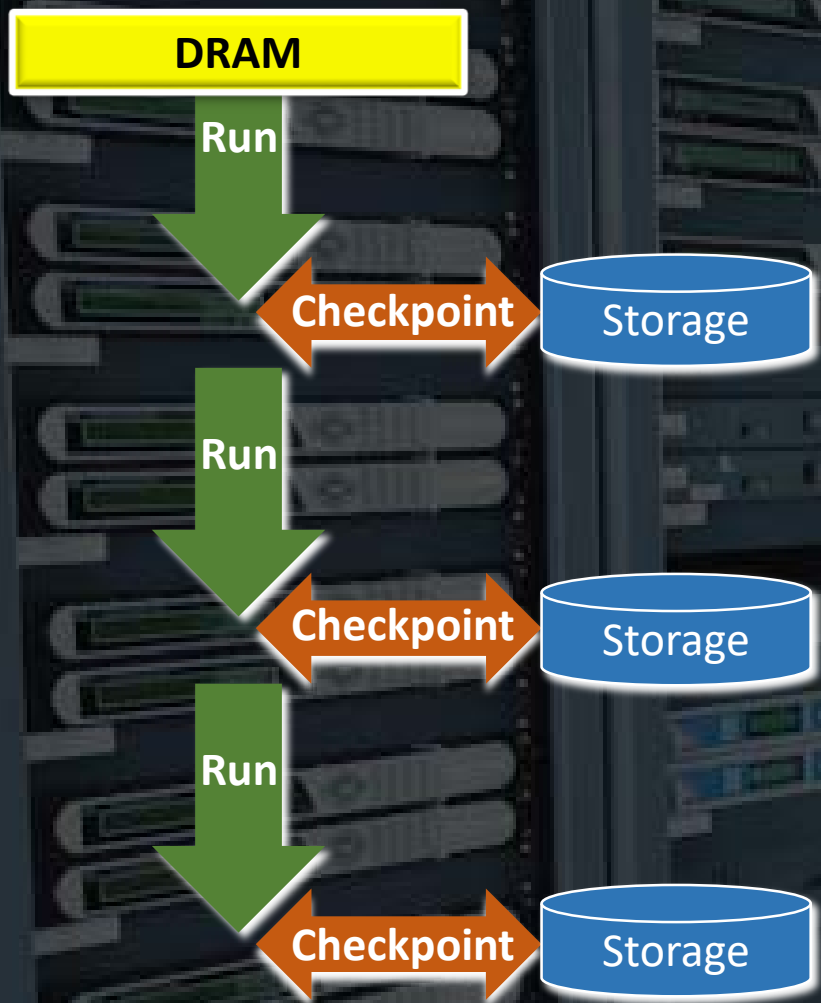
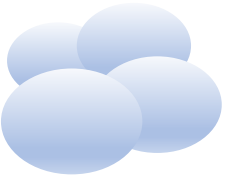


Row Extension Replacement Example



NVRAM Memory Class Storage



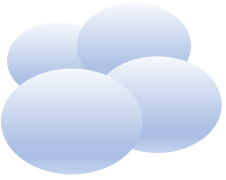


Checkpointing can be made to persistent memory

OR

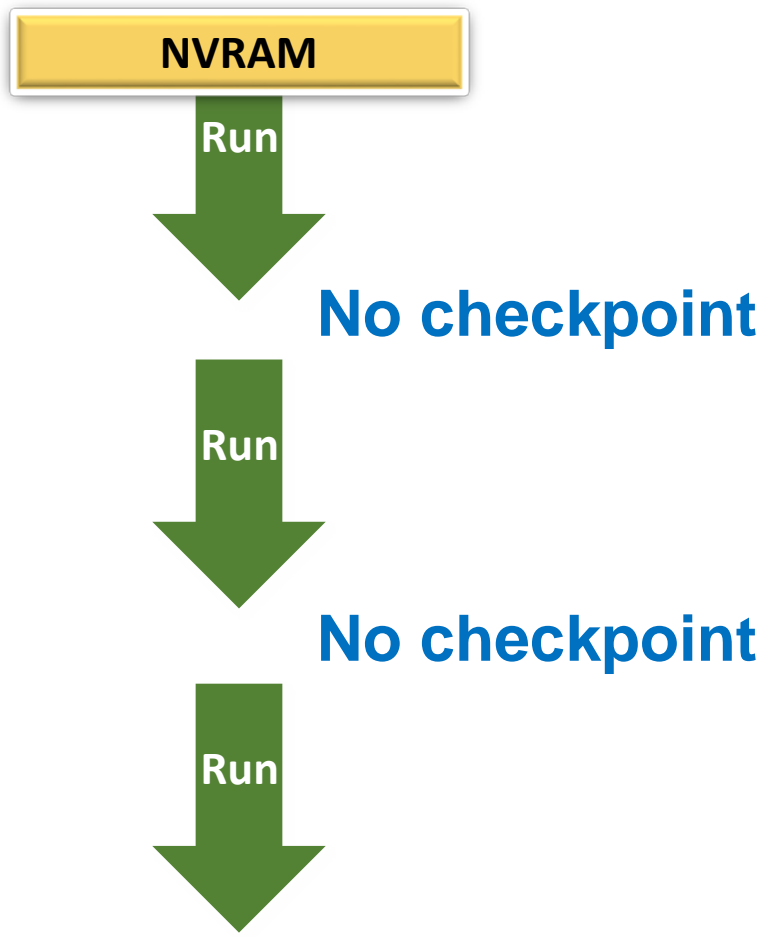
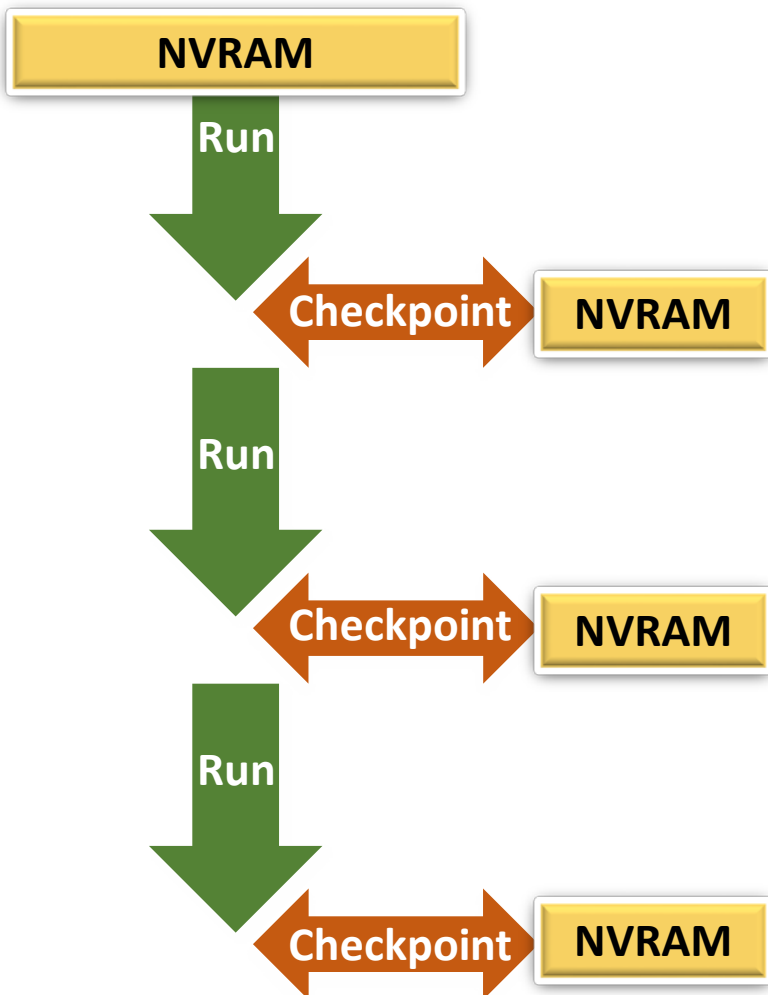
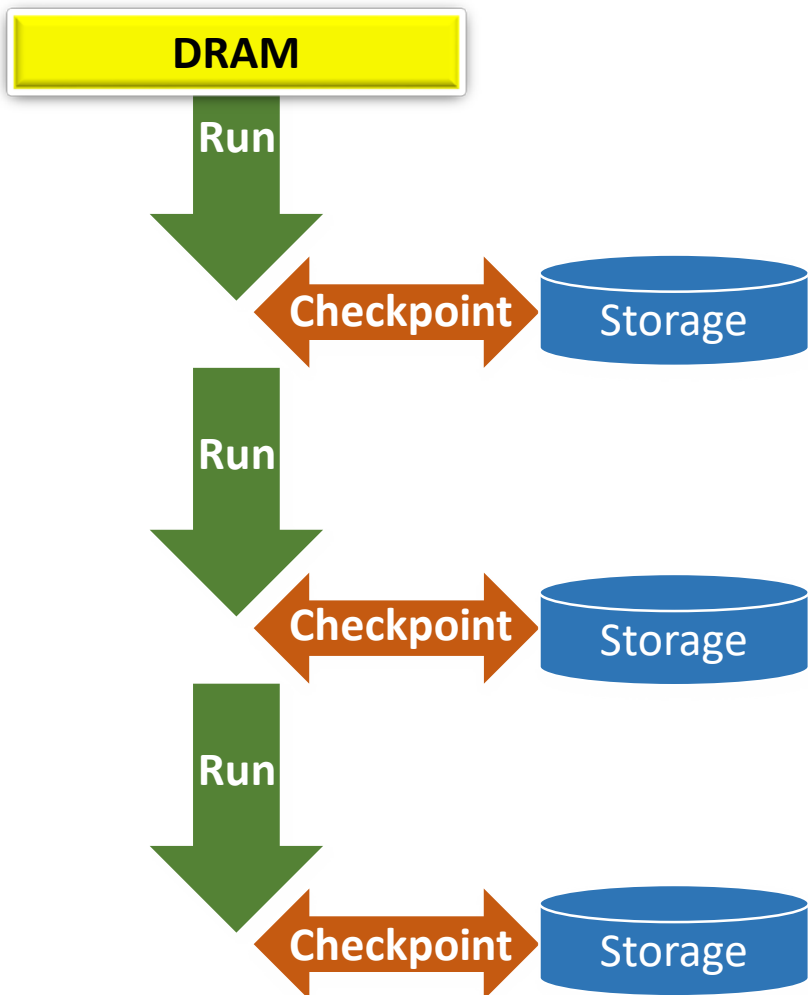
Checkpointing can be turned off completely

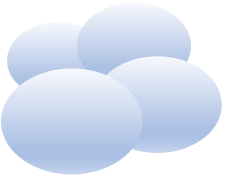




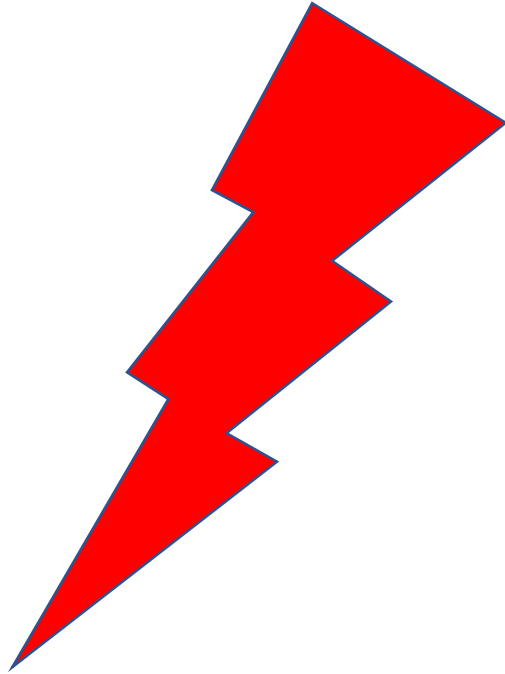
Phase 1

Phase 2





Keep in mind...



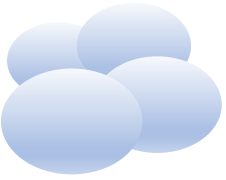
Power failure
is not the only
thing to fear



Checkpoints
may include
system failure

Knowing when
a task may
resume is
complicated





Remember Those Persistence Definitions

Immediately After WRITE

Tasks may be safe in nanoseconds

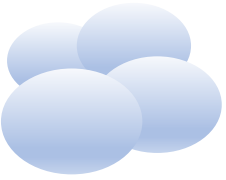
After FLUSH Command

Tasks may be safe in microseconds

On NVRAM RESET

Tasks may not be safe until system stability confirmed





Performance

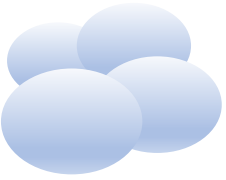
Persistence

Capacity

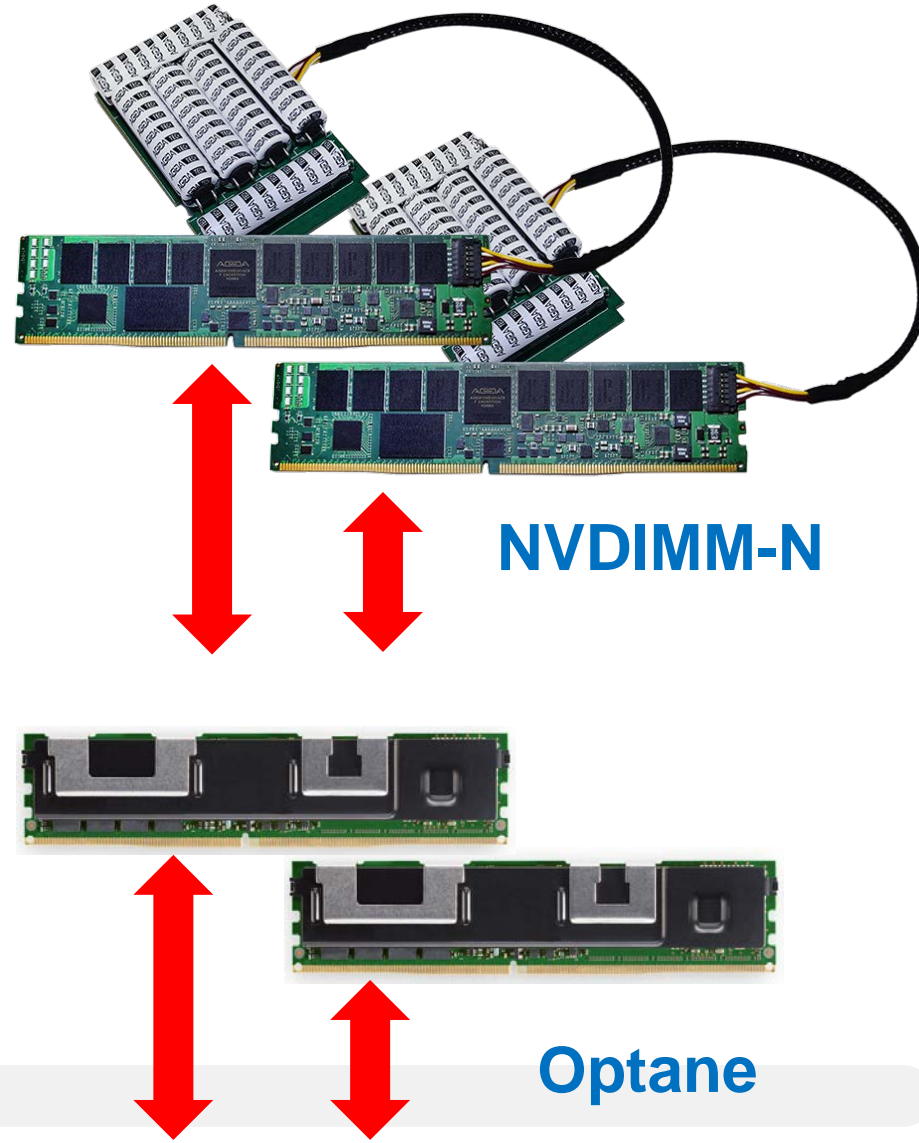
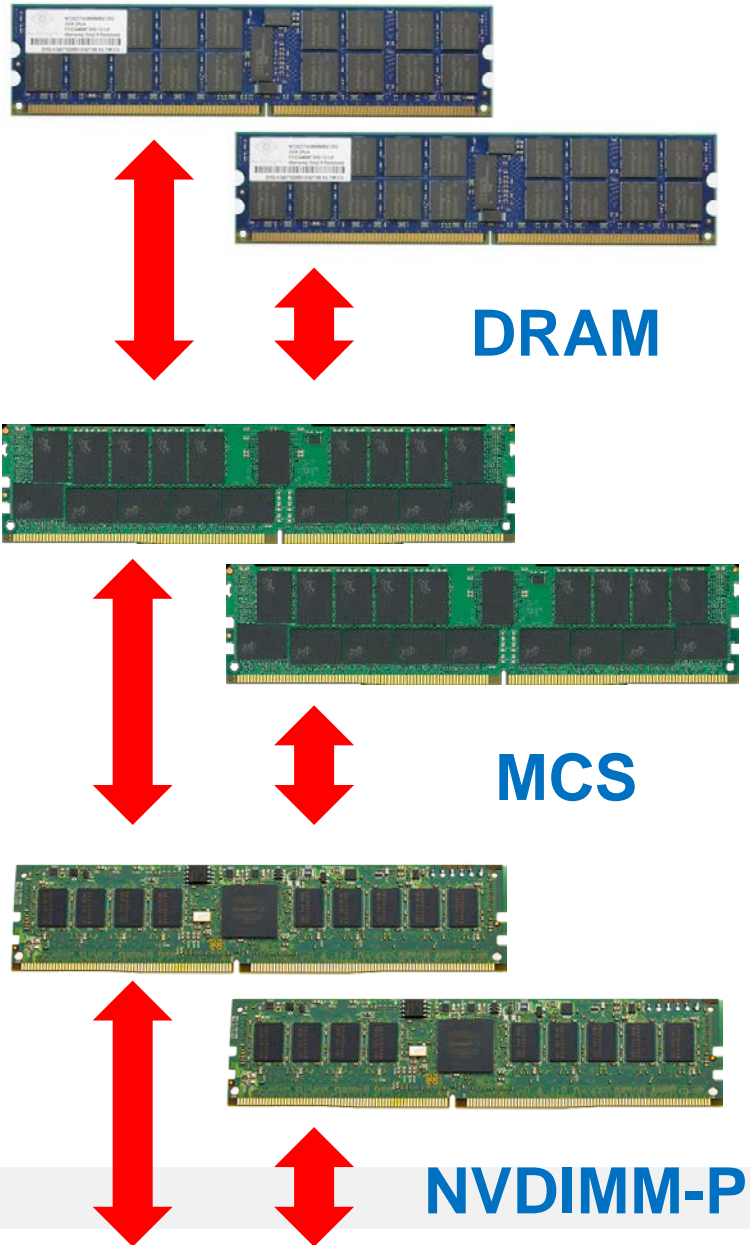


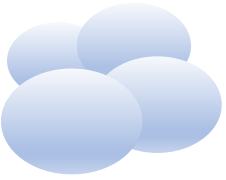
**System designers have
a lot of options to balance**



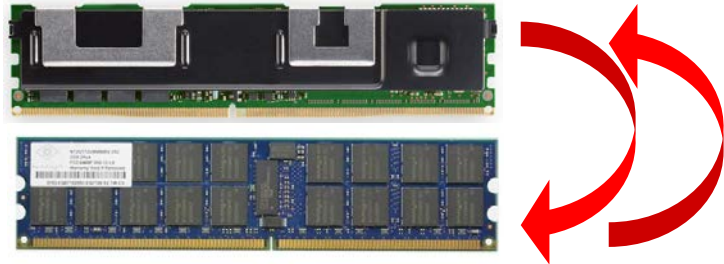


Homogenous Main Memory

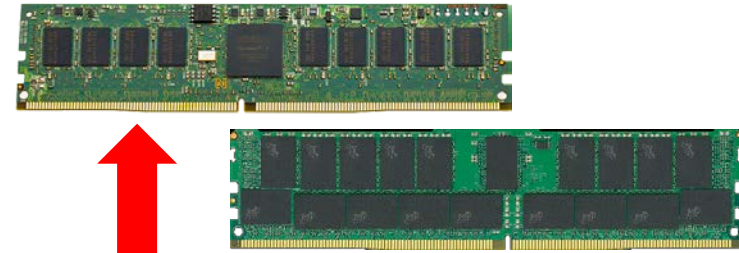




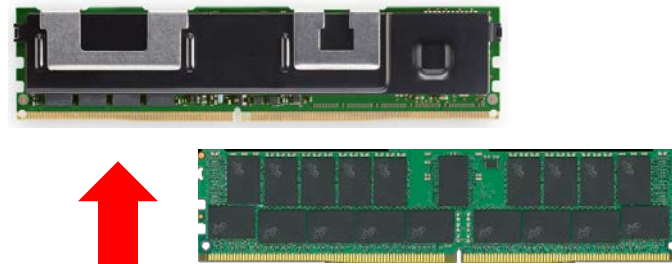
Heterogenous Main Memory



**DRAM +
Optane**

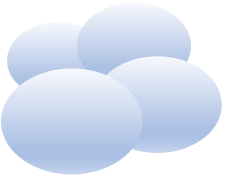


**MCS +
NVDIMM-P**



**MCS +
Optane**





When capacity meets persistence

512GB



NVDIMM-P



Optane

64GB

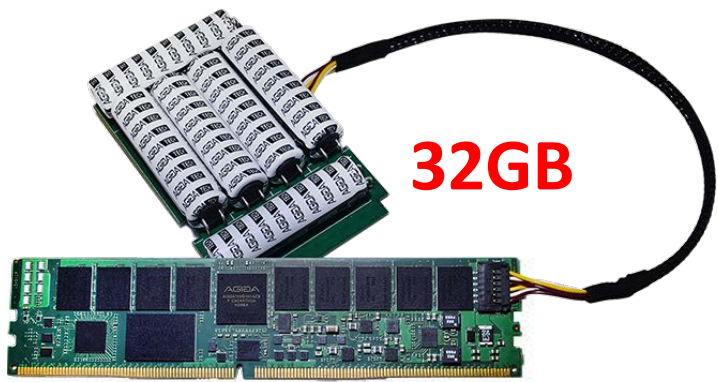


DRAM



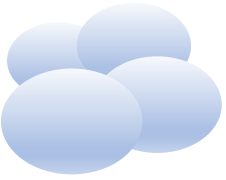
NVRAM
Memory Class Storage

32GB



NVDIMM-N



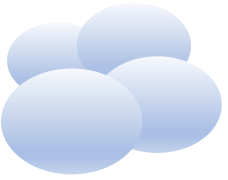


Homogenous

Main Memory Combinations

	Data Safe	Performance	Capacity
DRAM	No	Best	1.0 X
NVDIMM-N	Yes	Best	0.5 X
Optane	Yes	Worst	10 X
NVDIMM-P	Yes	Mid	10 X
MCS	Yes	Best+	1 X+



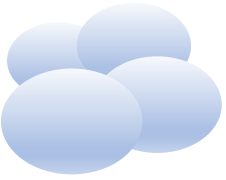


Heterogeneous

Main Memory Combinations

	Data Safe	Performance	Capacity
DRAM + Optane	No	High	6 X
DRAM + NVDIMM-P	No	High	6 X
MCS + Optane	Yes	High	6 X
MCS + NVDIMM-P	Yes	High	6 X





Homogenous

Main Memory Combinations

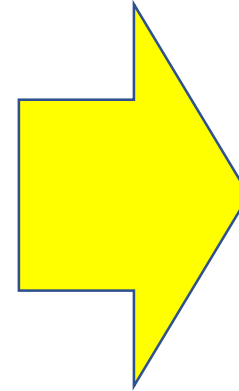


Software need not care

All functions take the
same time

Heterogeneous

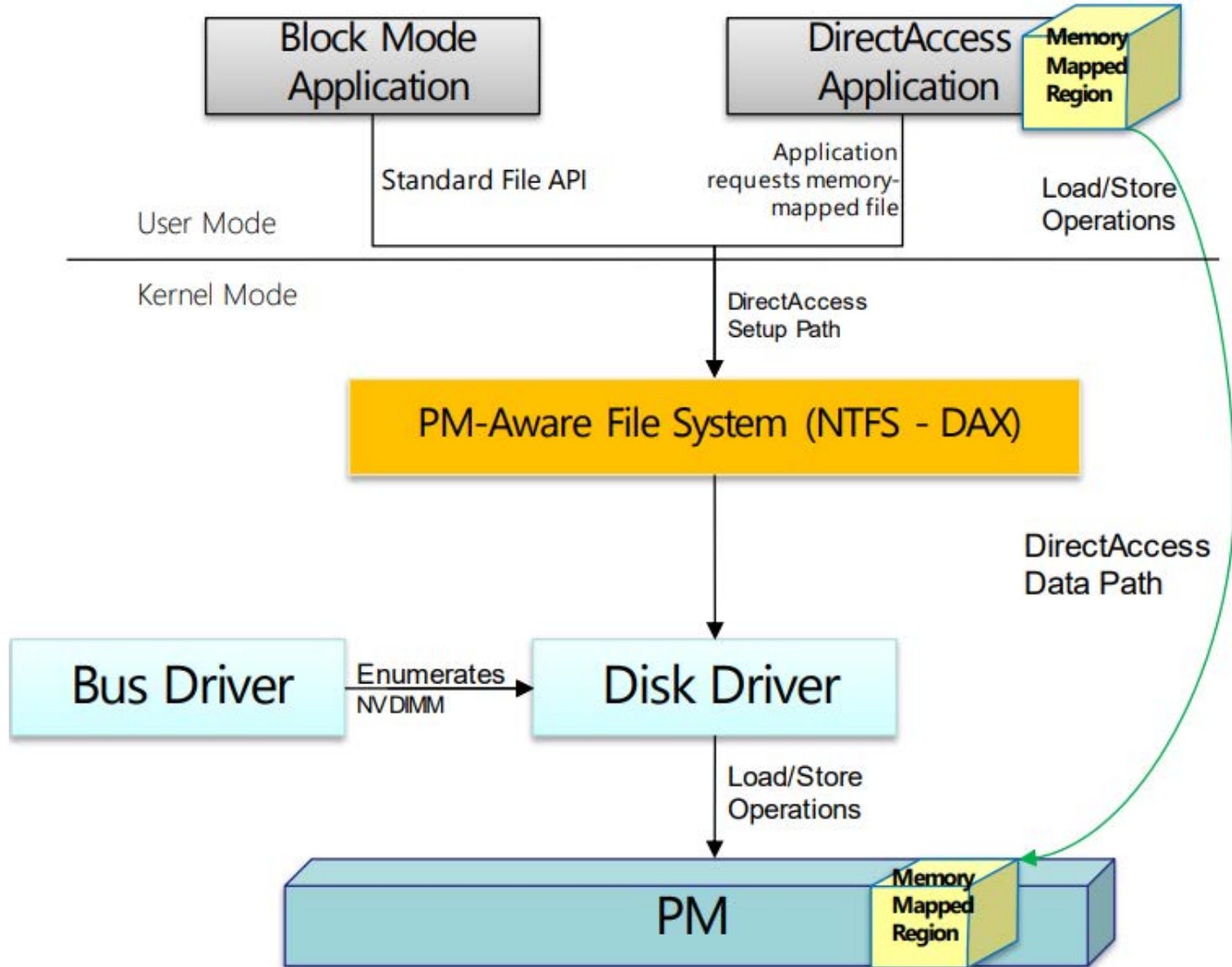
Main Memory Combinations



Software encouraged to
put critical functions
in faster memory

Often mount slower
memory as RAM drive



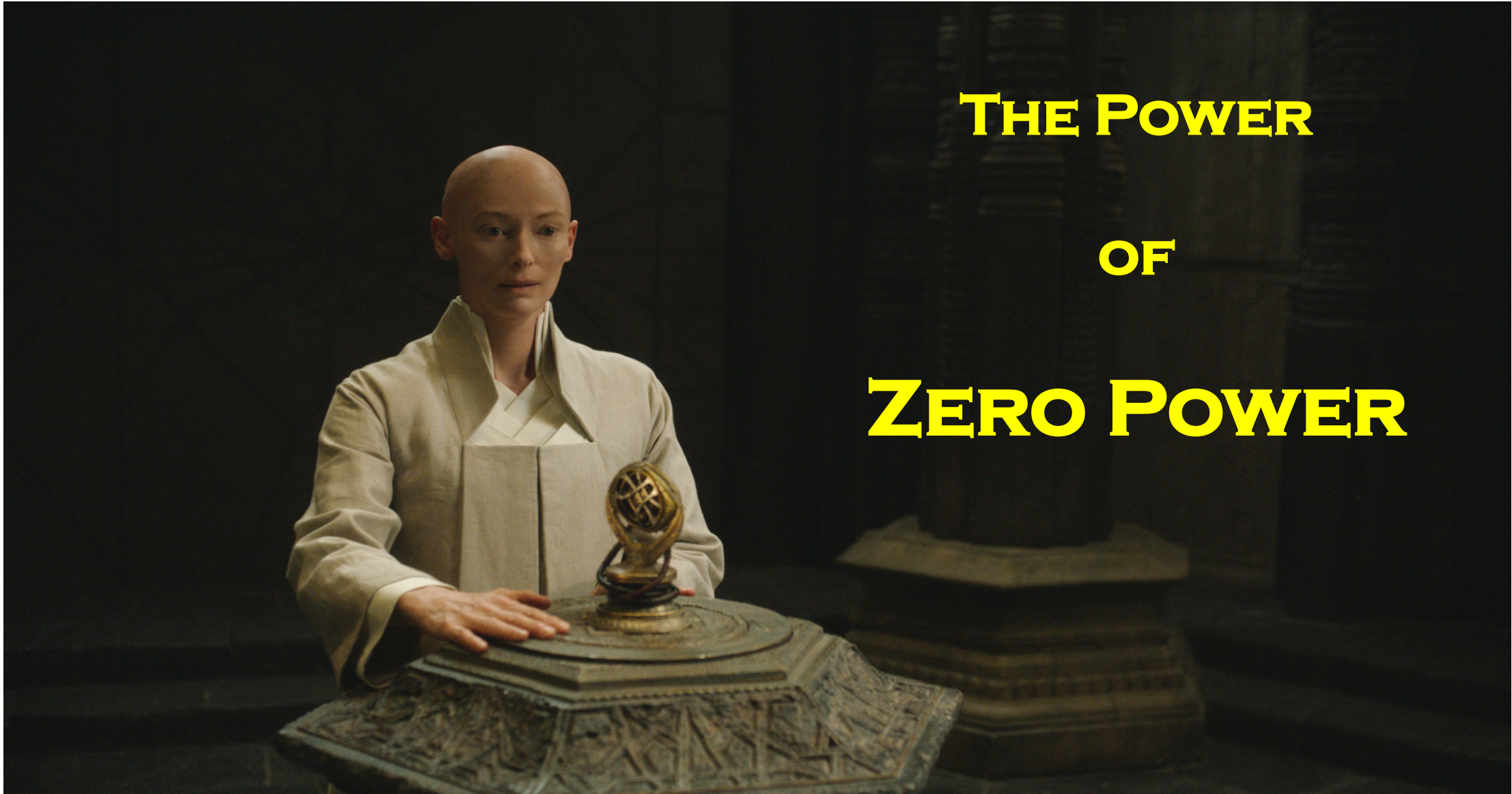
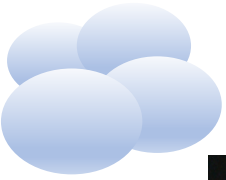


**Software support via DAX
assists in moving...**

from mounted drives...

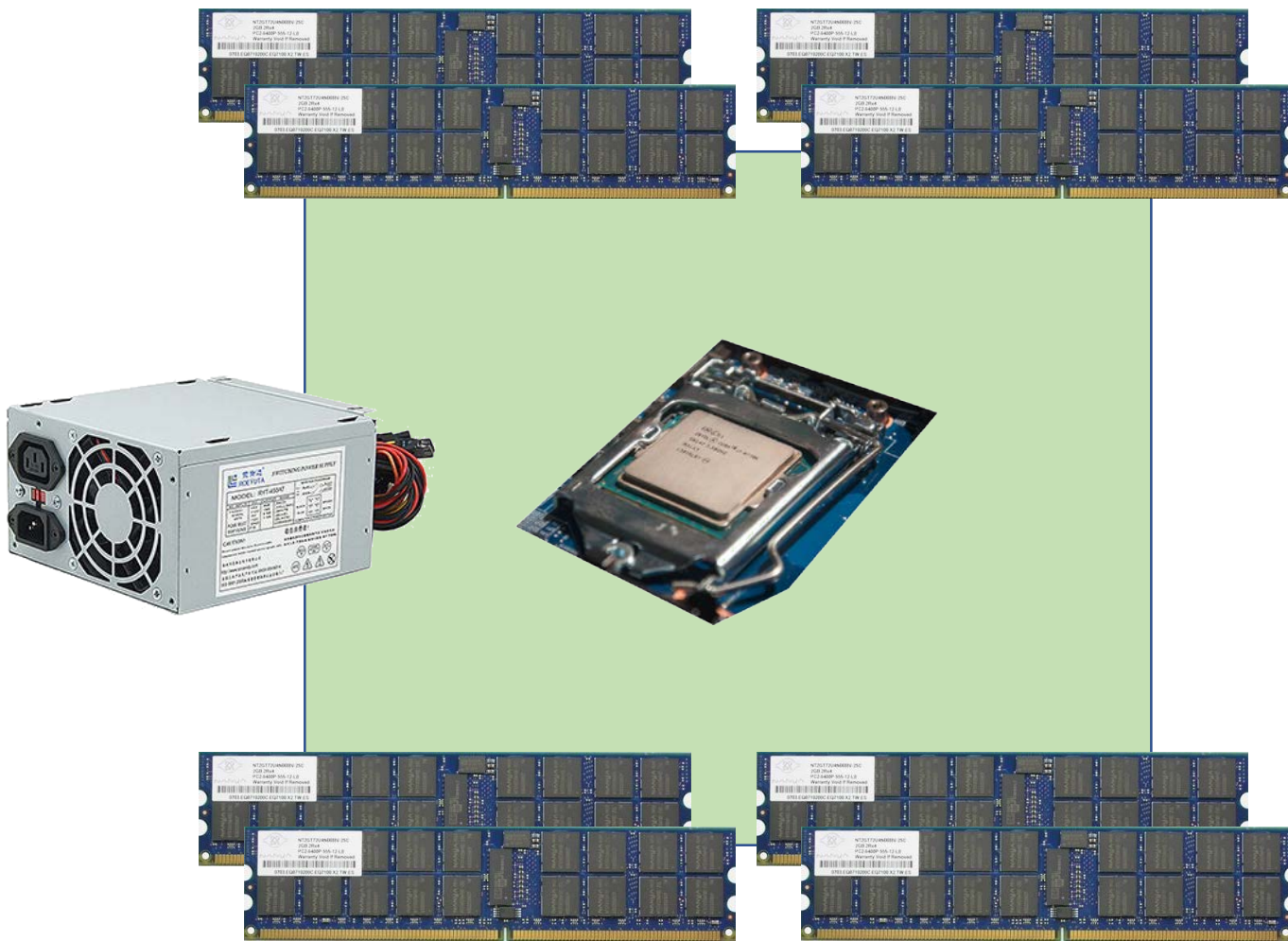
...to RAM drive...

...to direct access mode

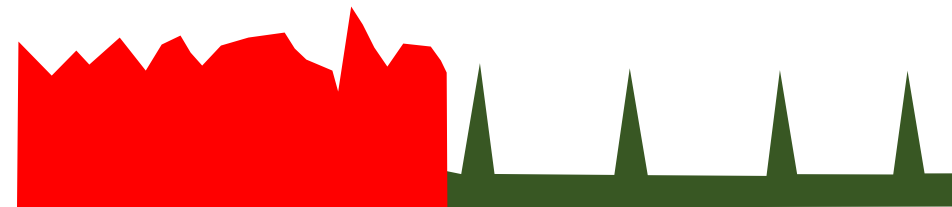


**THE POWER
OF
ZERO POWER**





Putting a Node to Sleep



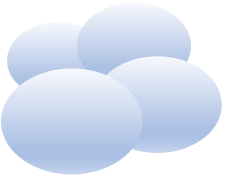
**Operating
Mode**

**Self Refresh
Mode**

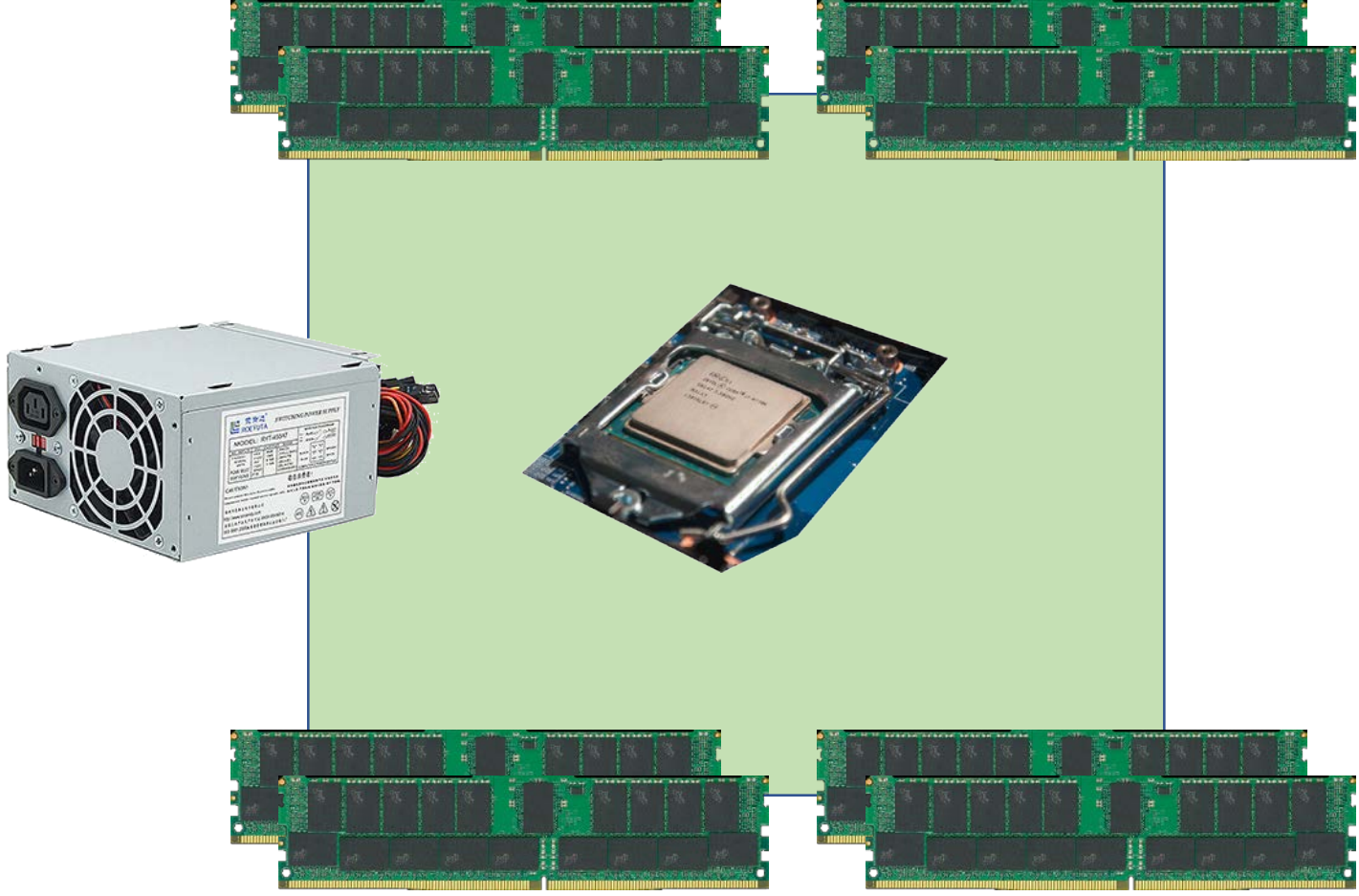
**Instant On means
power must stay alive**

**Refresh operations burn
significant power**





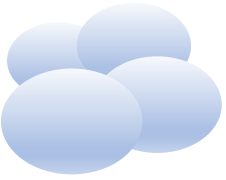
Memory Class Storage can be turned off entirely



Operating
Mode

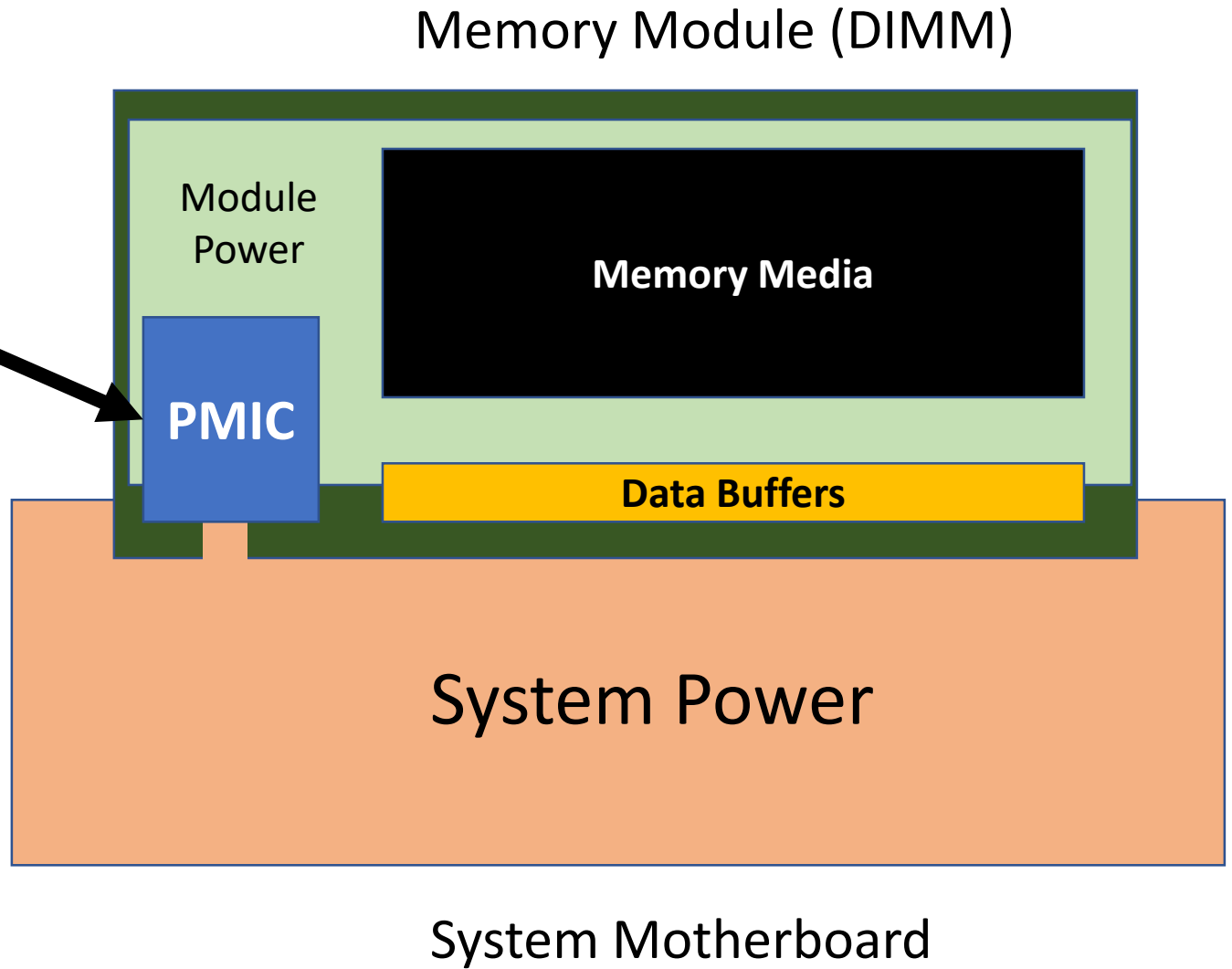
Power
Off

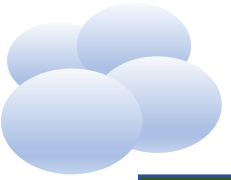




DDR5 memory modules have on-DIMM voltage regulation (PMIC)

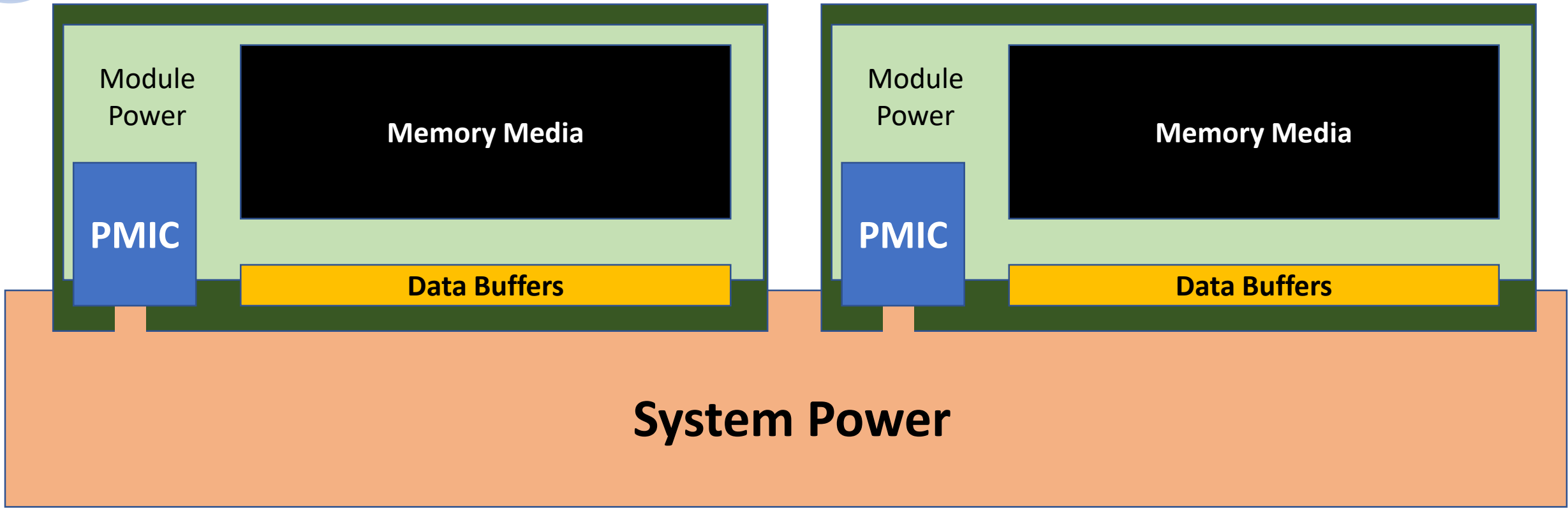
DIMM power may be shut off independently of system power





DIMM1

DIMM2



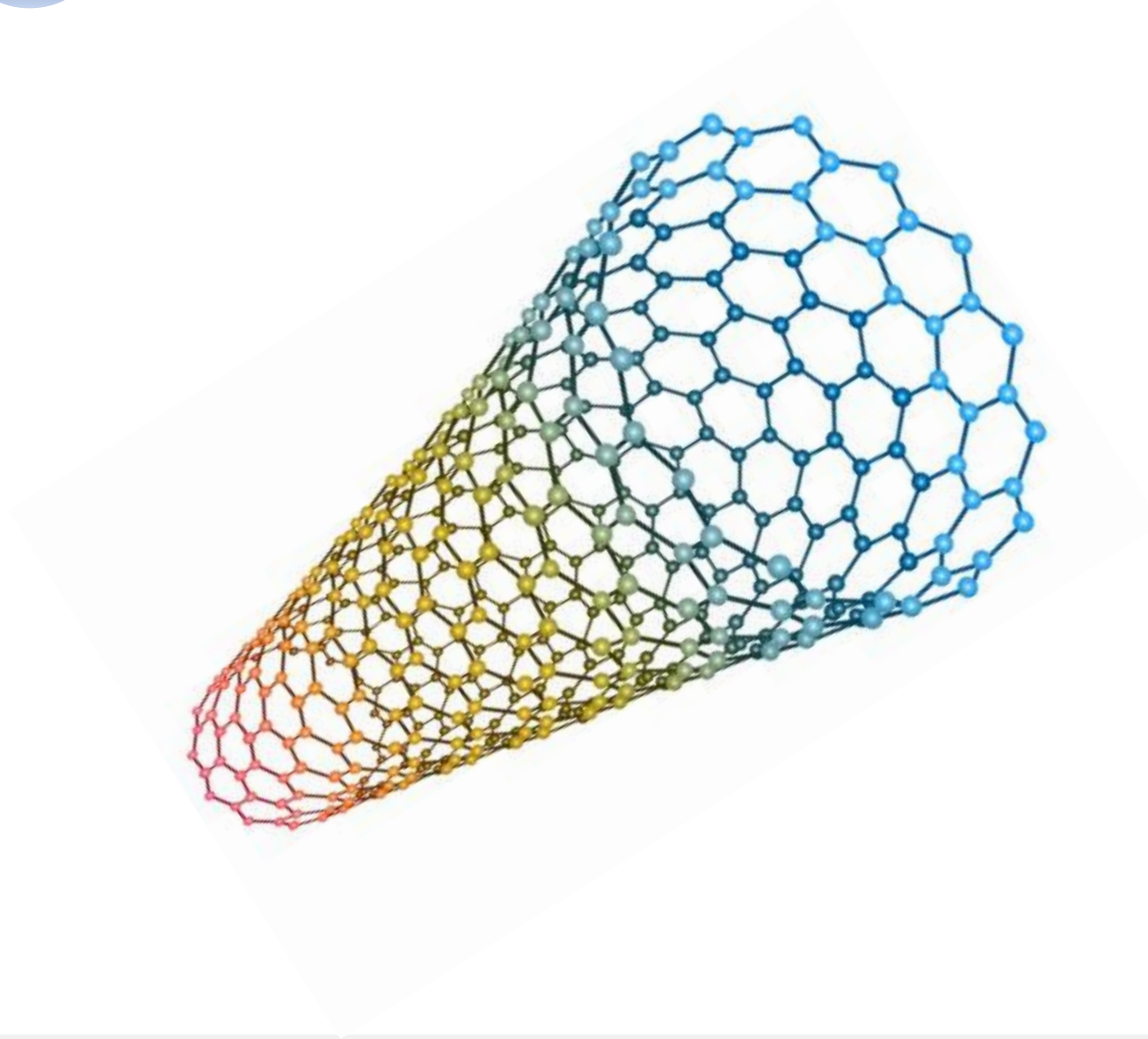
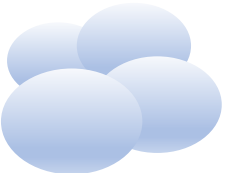
Multiple power management options

System power off; both DIMMs off

System power on & both DIMMs off

System power on & DIMM1 on, DIMM2 off

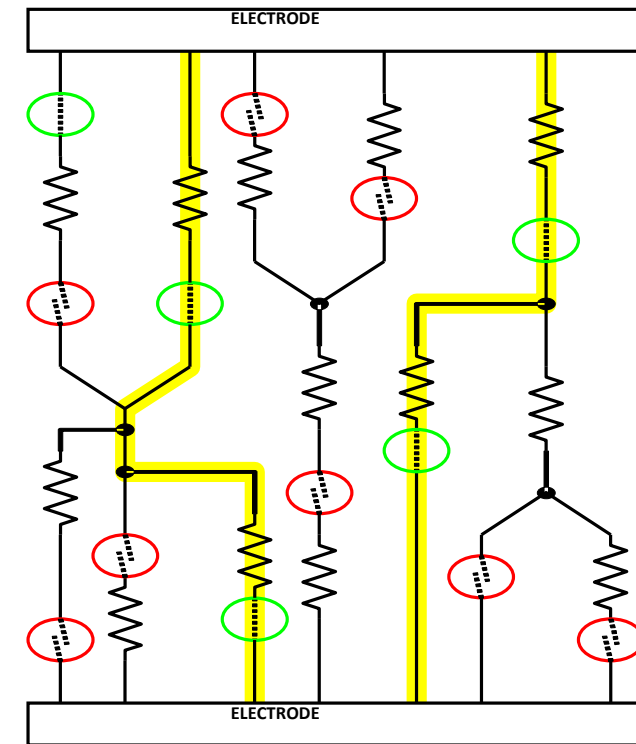
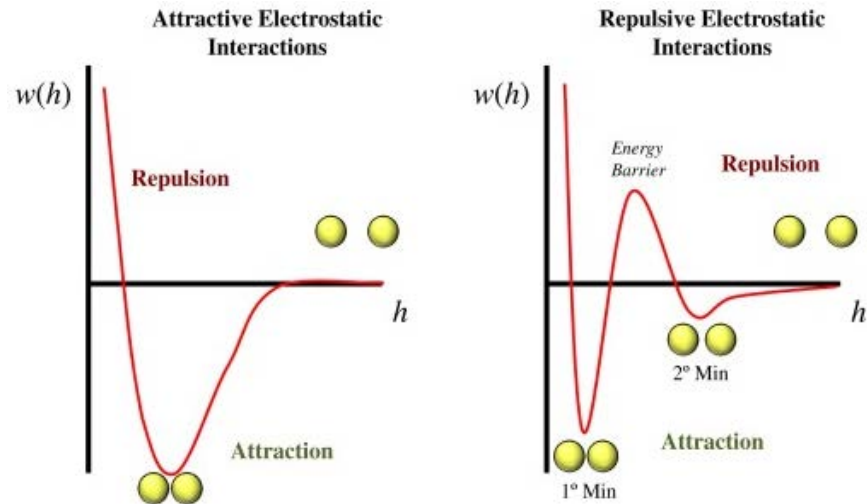
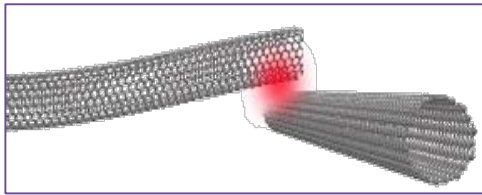




Nantero NRAM™

My favorite NVRAM



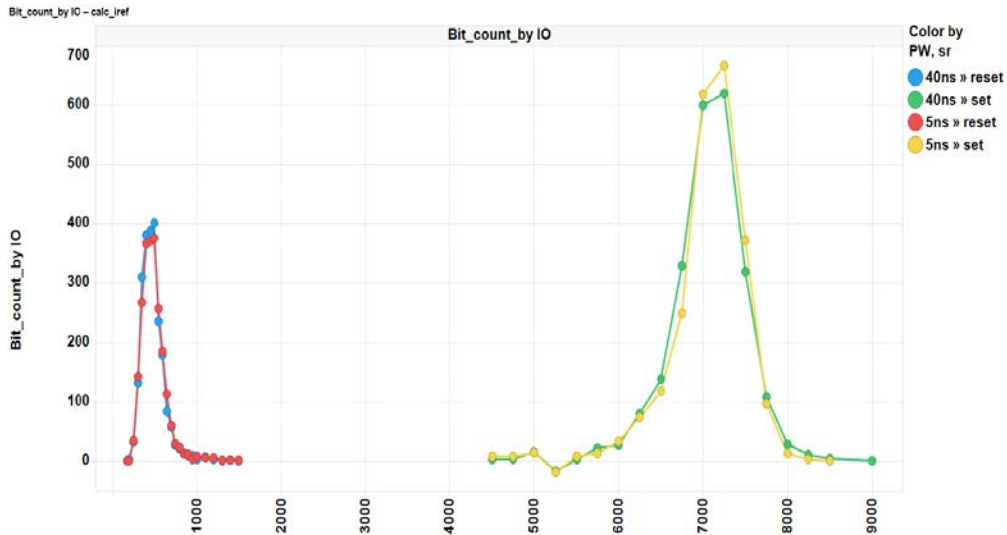


Van der Waals energy barrier keeps CNTs apart or together

Data retention >300 years @ 300 °C, >12,000 years @ 105 °C

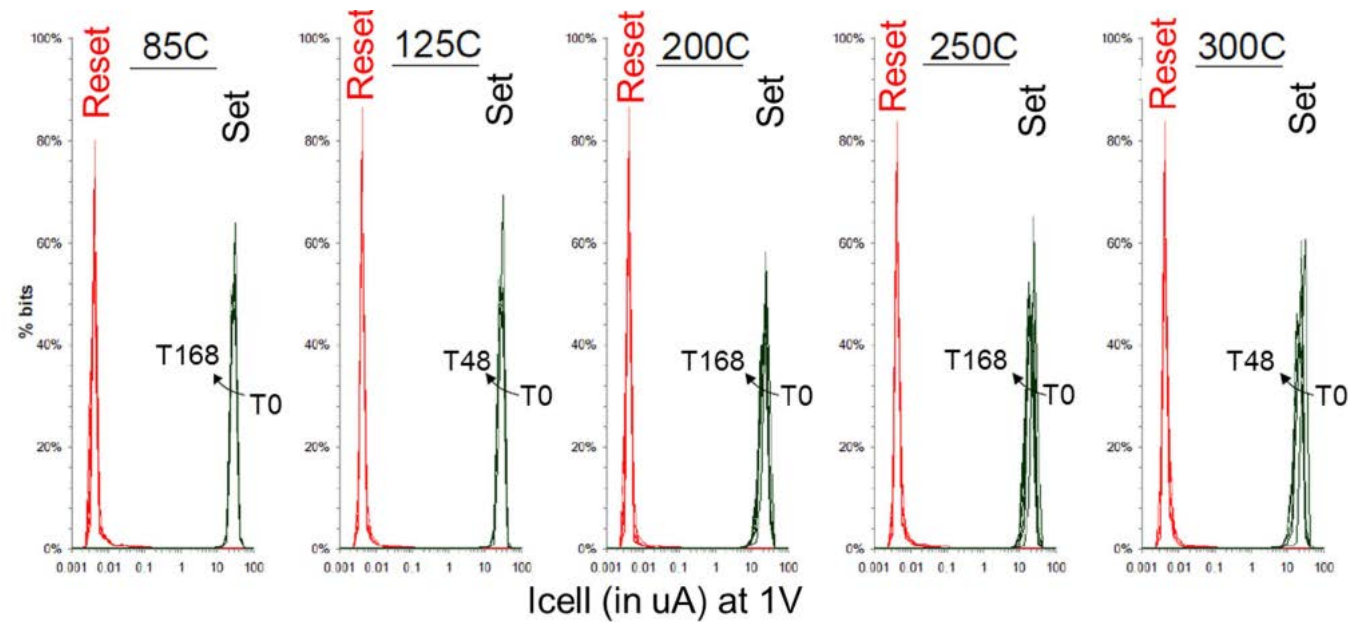
Stochastic array of hundreds nanotubes per each cell





**5 ns balanced
read/write performance**

No temperature sensitivity



**NRAM Data Retention
= 12,000 Years**



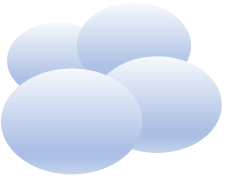
10,000 years ago

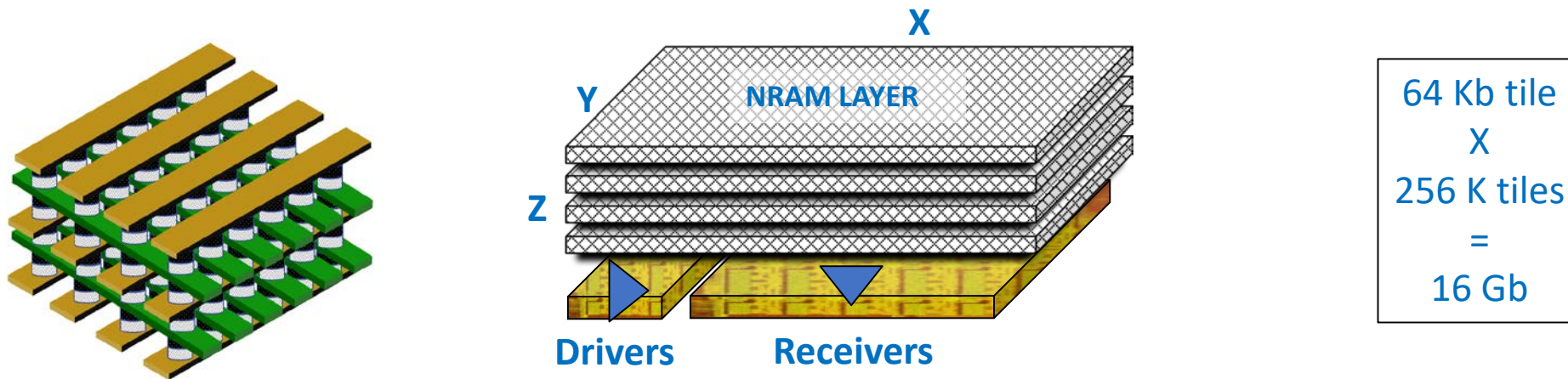


4,500 years ago



2,500 years ago



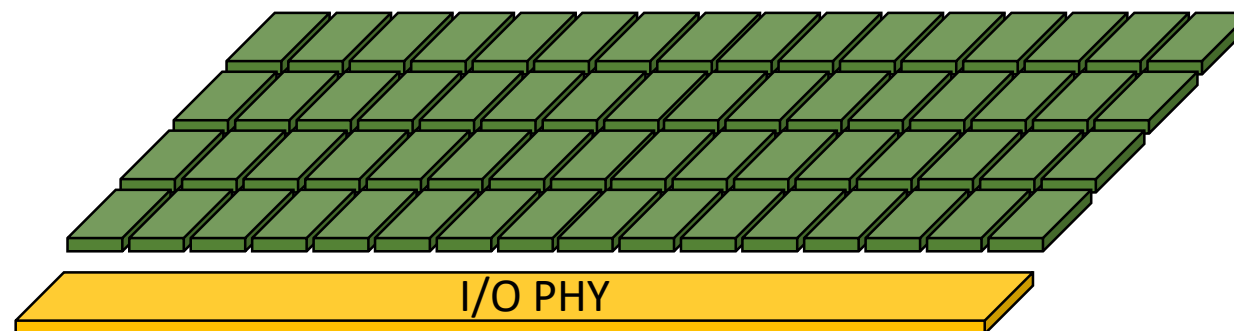


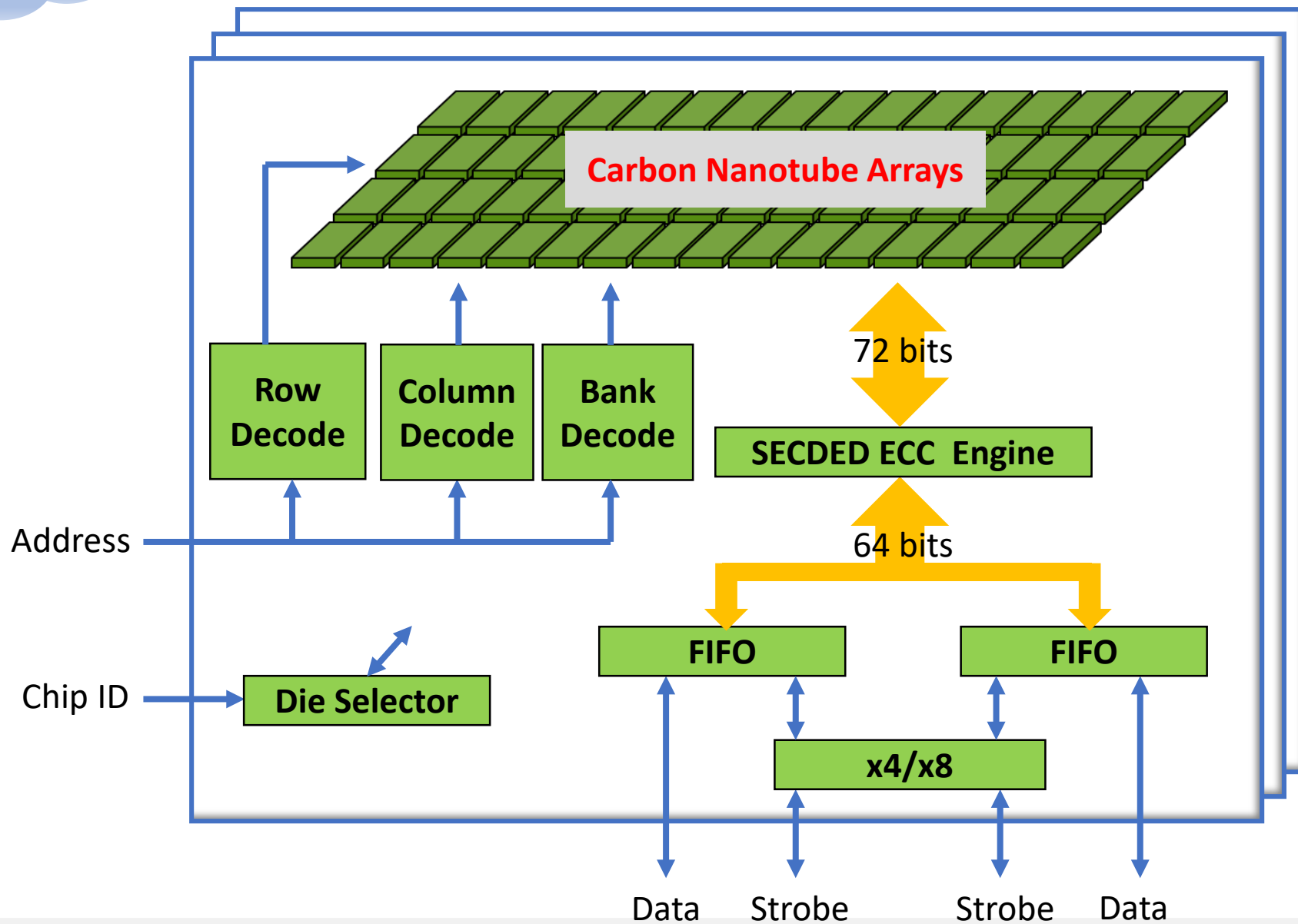
Array size tuned to the size of drivers & receivers

Chip-level timing is a function of bit line flight times

Replicate this “tile” as needed for device capacity

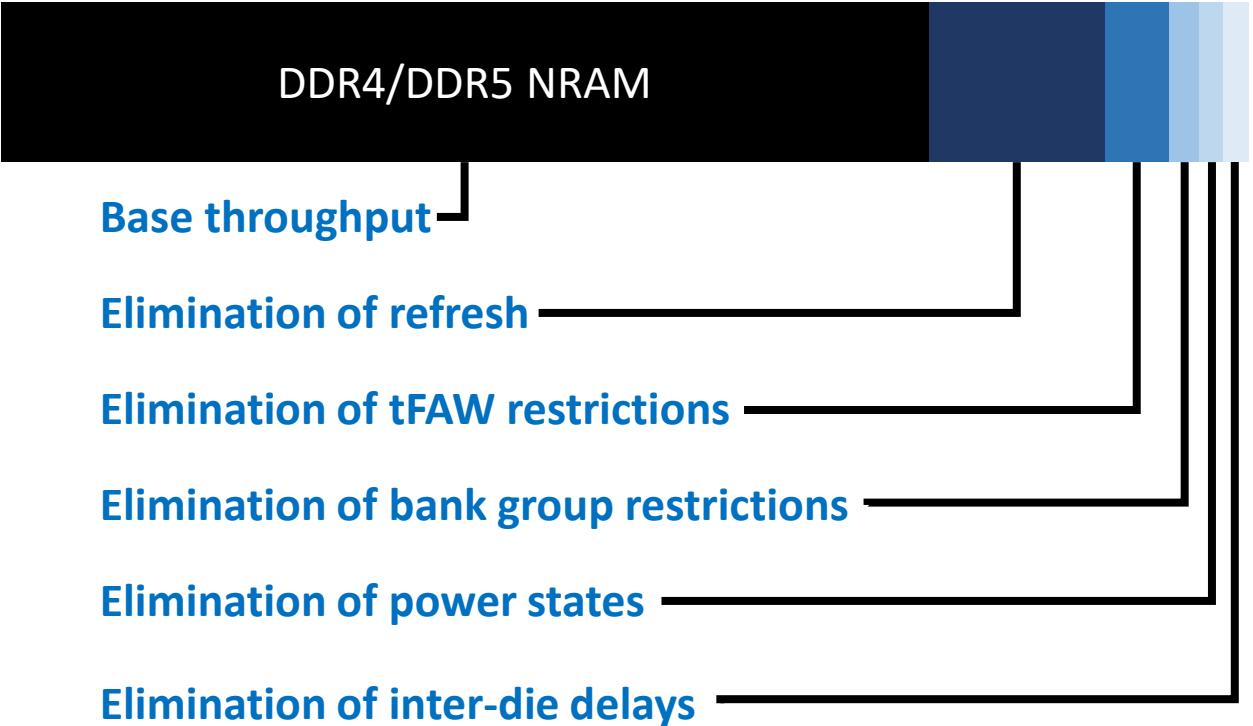
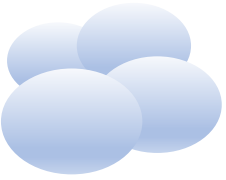
Add I/O drivers to emulate any PHY needed





**DDR4, DDR5
NRAM**

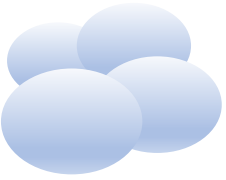




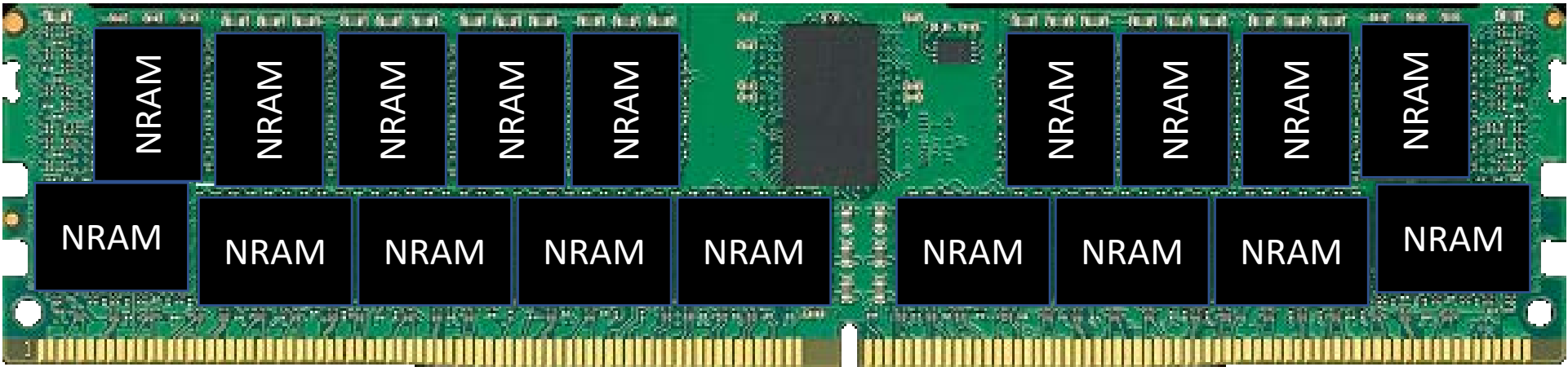
**Architectural improvements
improve data throughput
15% or greater at the same
clock frequency**

Bandwidth: larger is better





NVRAM Memory Class Storage

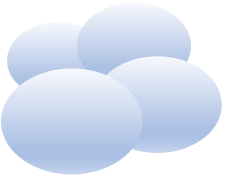


Plugs into an RDIMM slot

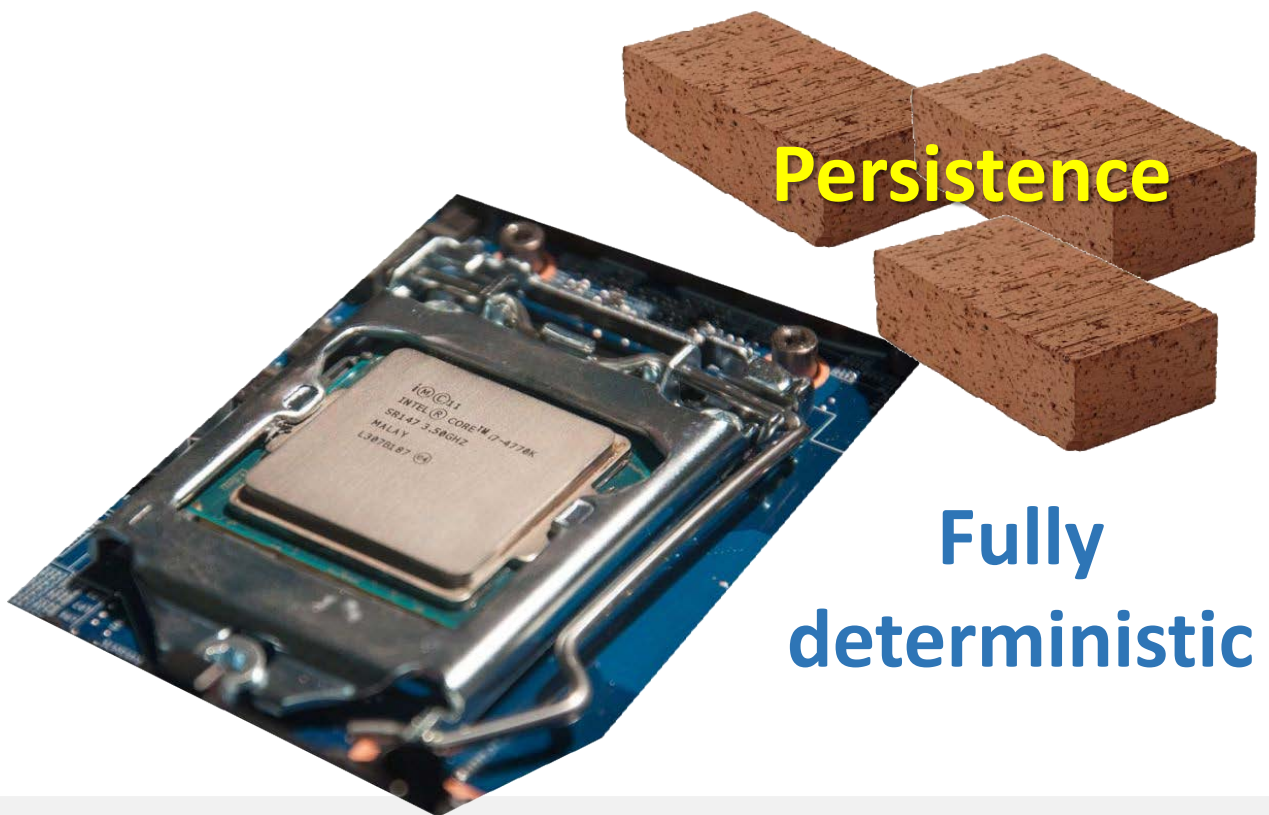
Appears to the CPU as DRAM

Memory controller may optionally be tuned for NVRAM





One less layer of marshmallows to deal with



Persistence

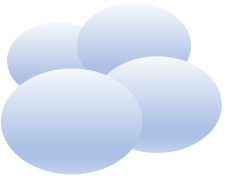
Fully deterministic

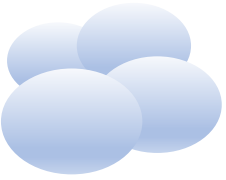


Persistence

Non-deterministic



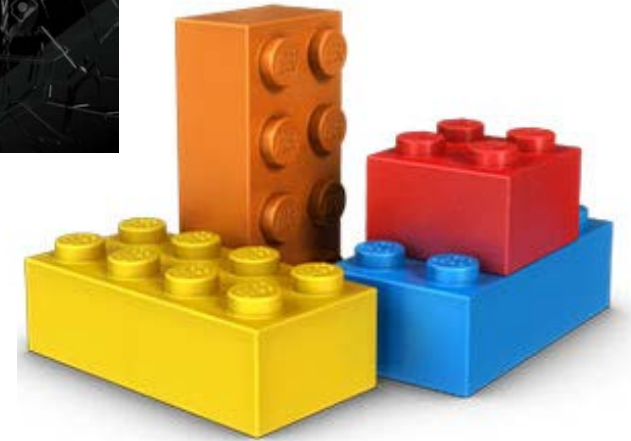




Would you rather...



Step on
broken glass?

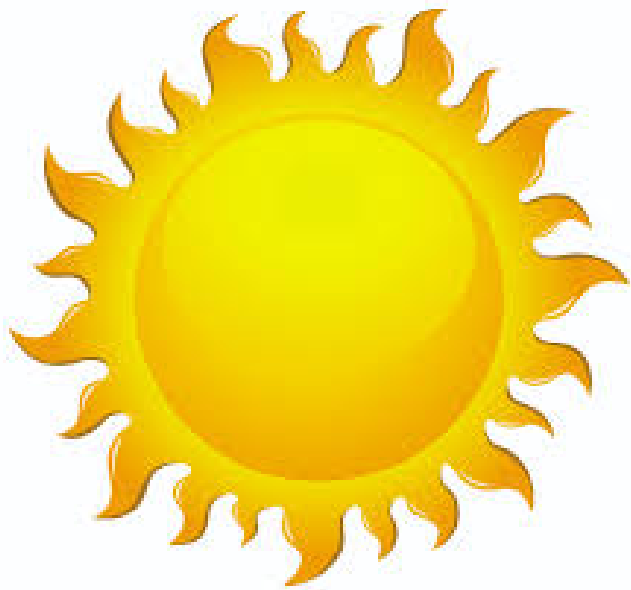


A LEGO?



Or some
jacks?





...about those energy stores...

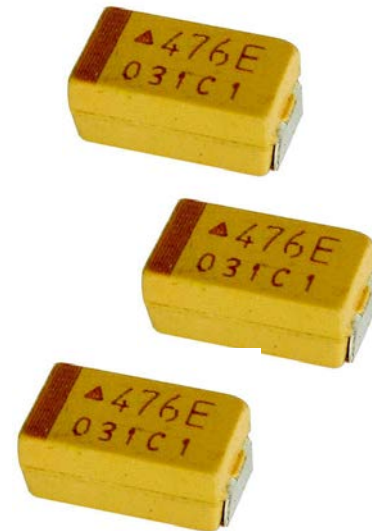
Batteries

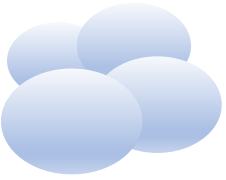


Supercapacitors



Tantalums (etc.)





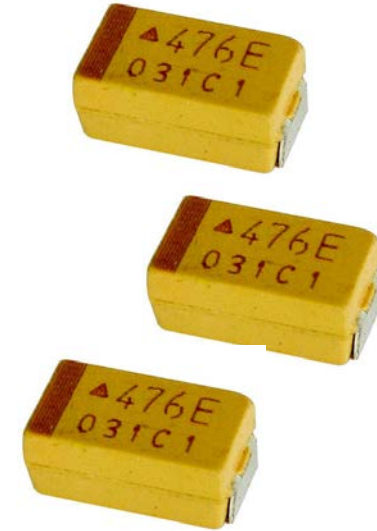
Batteries



Supercapacitors



Tantalums (etc.)



High capacity
High energy density
Low reliability

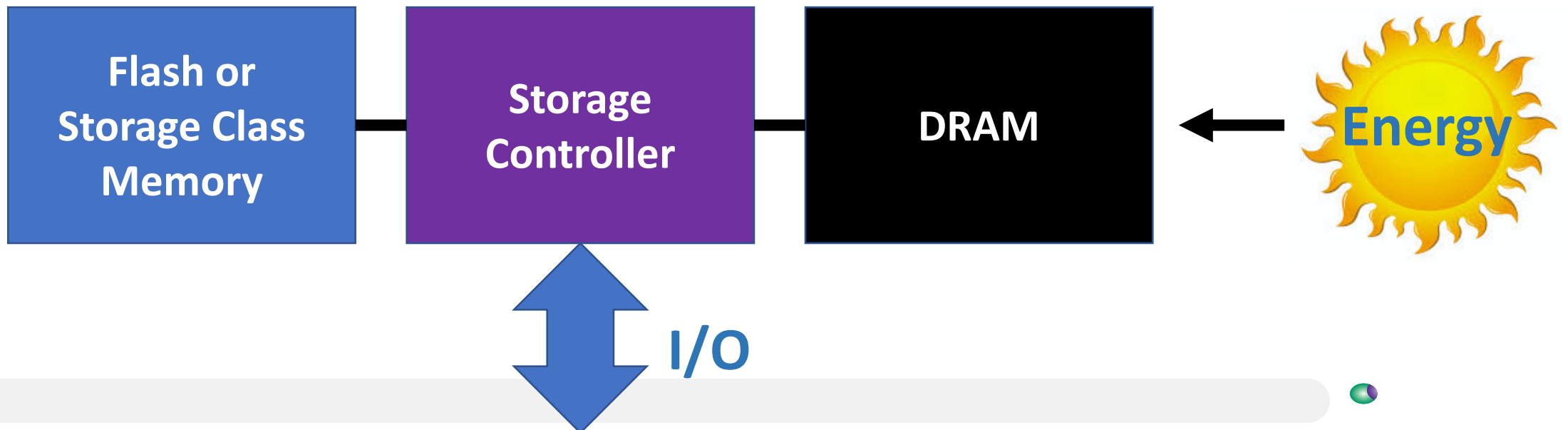
Medium capacity
Low energy density
Degrade over time

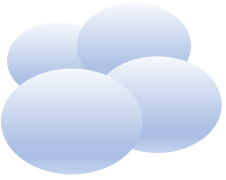
Low capacity
Low energy density
...but stable





Energy needed
for backup of
DRAM cache

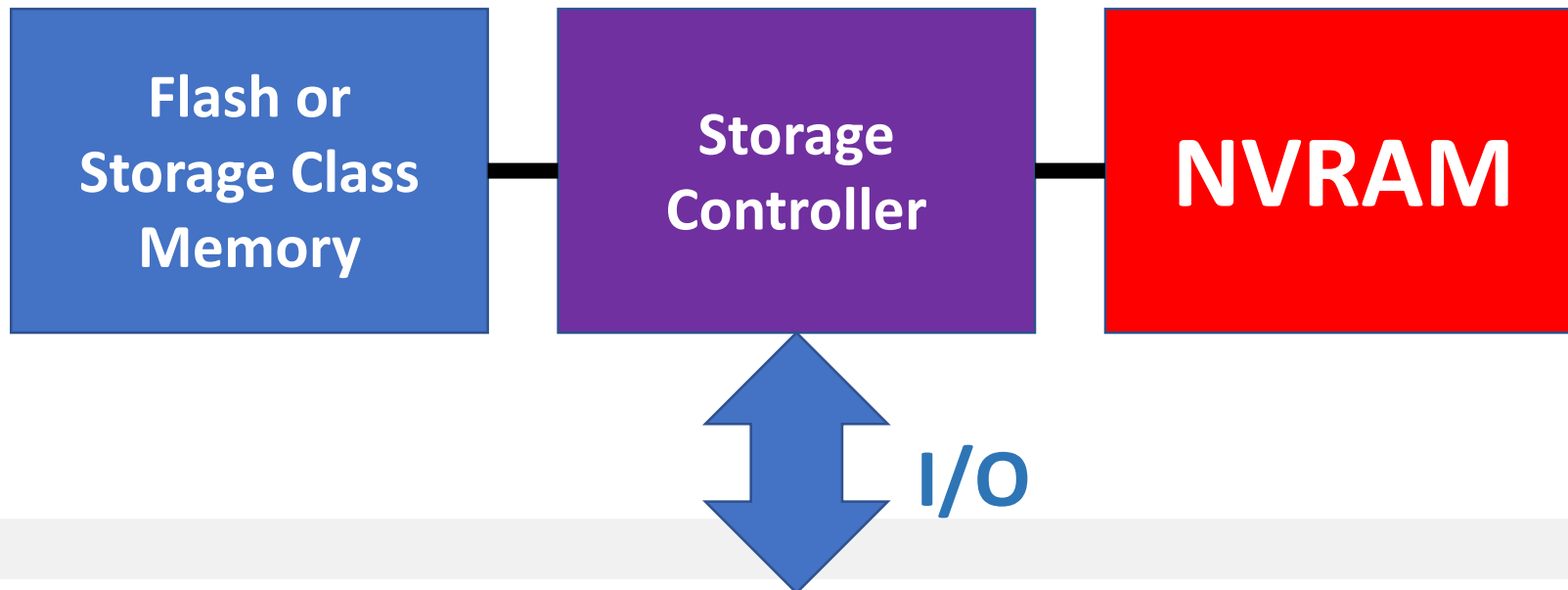


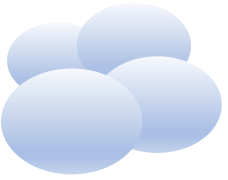


More room
for storage



Eliminate need
for backup
energy



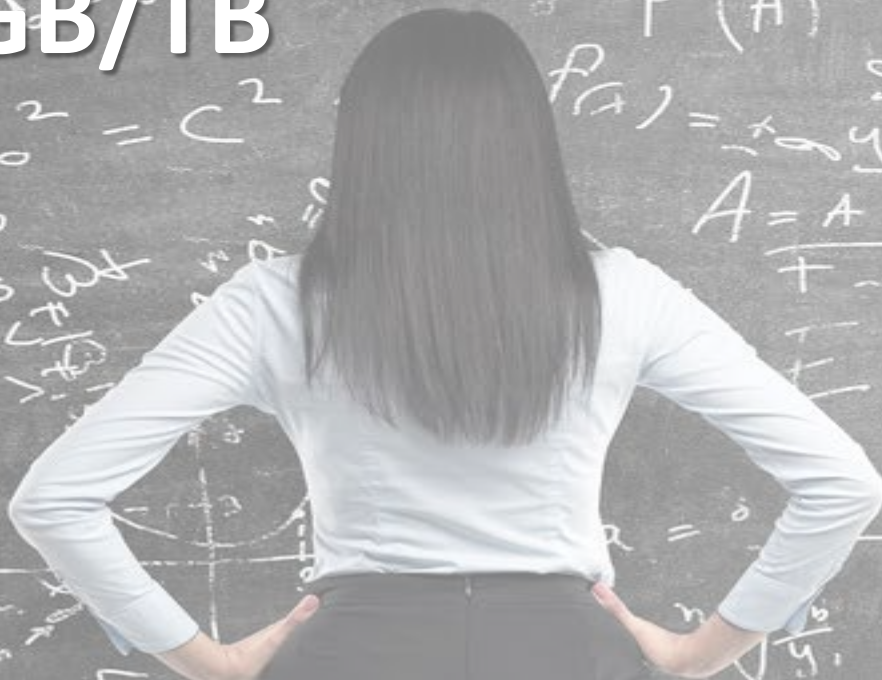


NVRAM Changes the Math

1GB/TB

DRAM cache limited by
energy available

No DRAM? Cache size
dictated by
cost/performance





Switching gears again...

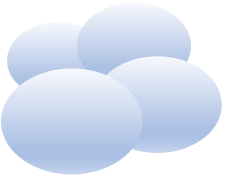
...to Systems Evolution



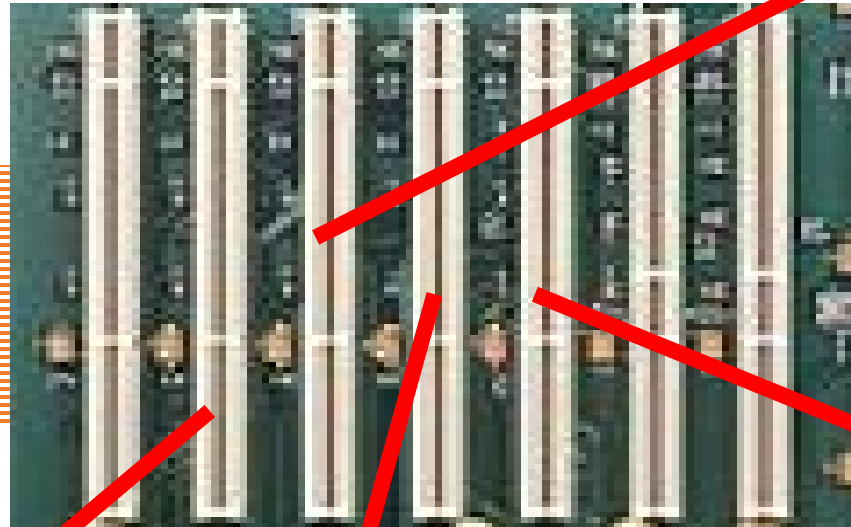
Pop quiz

How many
CPUs in a
1980s PC?

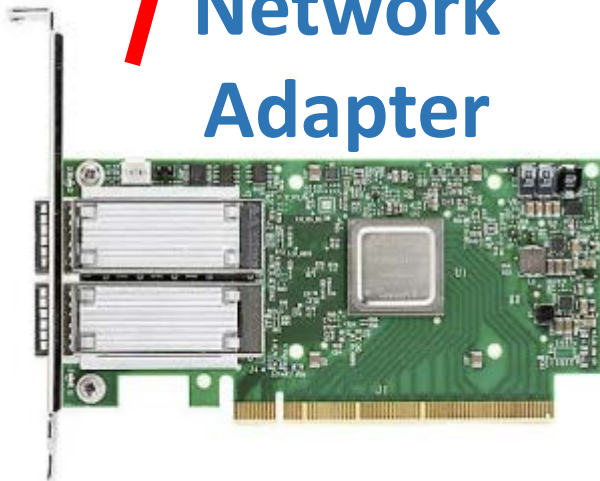




Graphics Adapter



Network Adapter

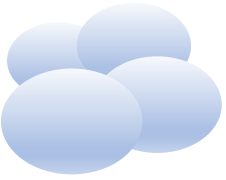


Modem



Sound Blaster



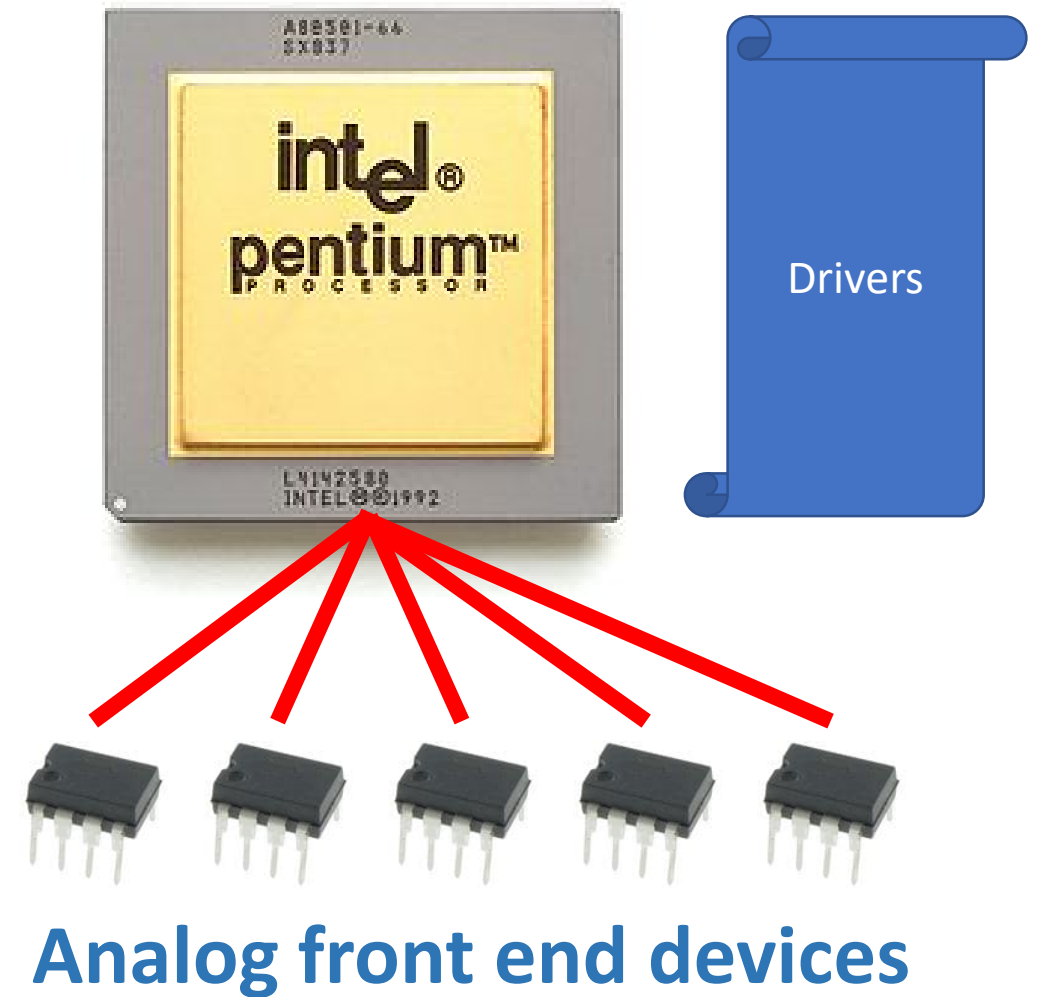


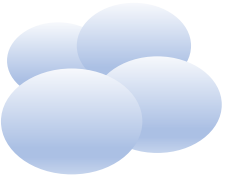
They were killed by “Native Signal Processing”

They were called “DSPs”
Digital Signal Processors

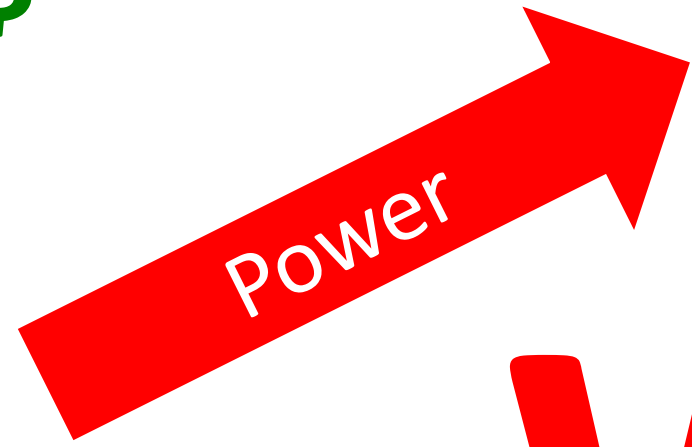


They put processing
next to the data

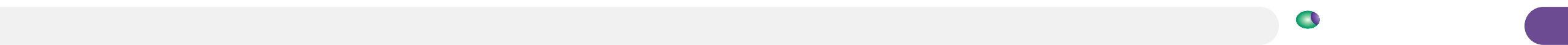


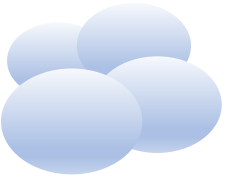


With NSP...

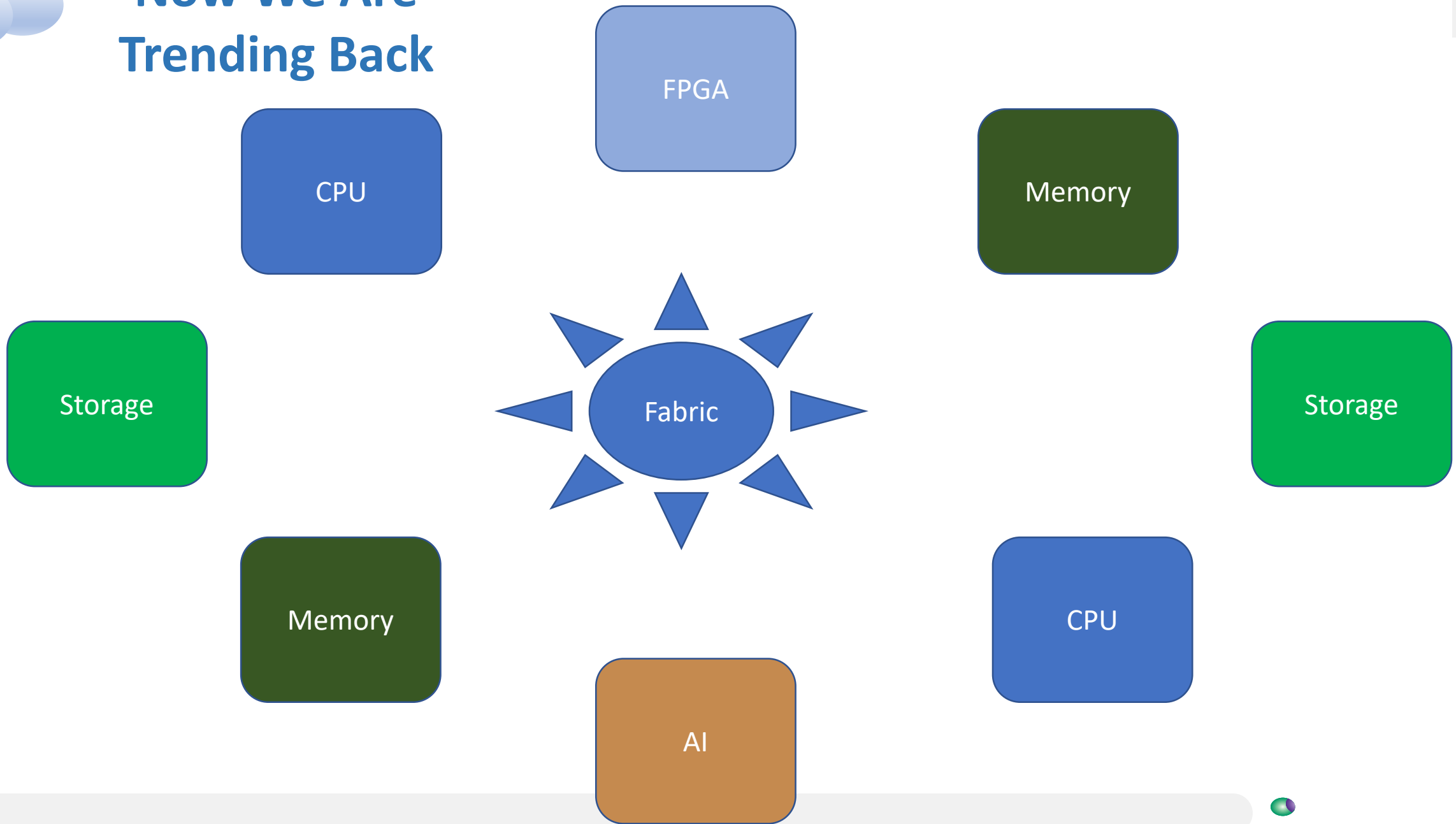


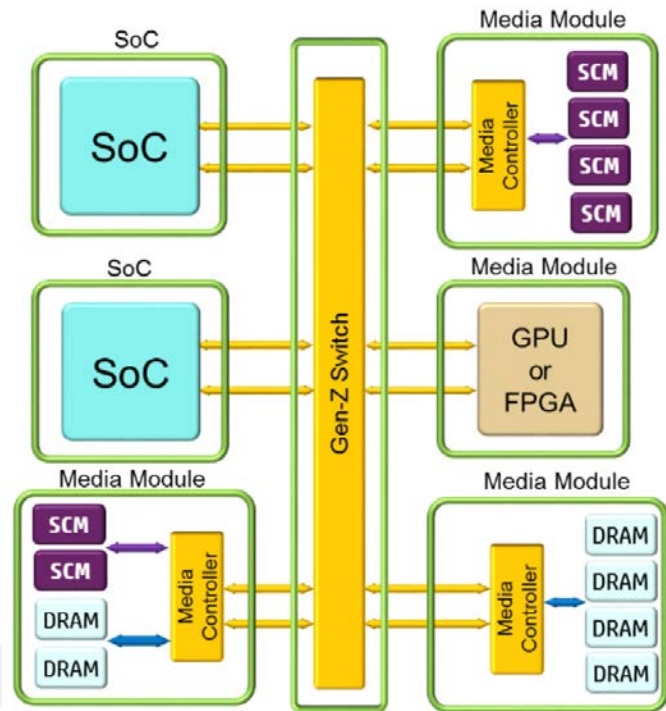
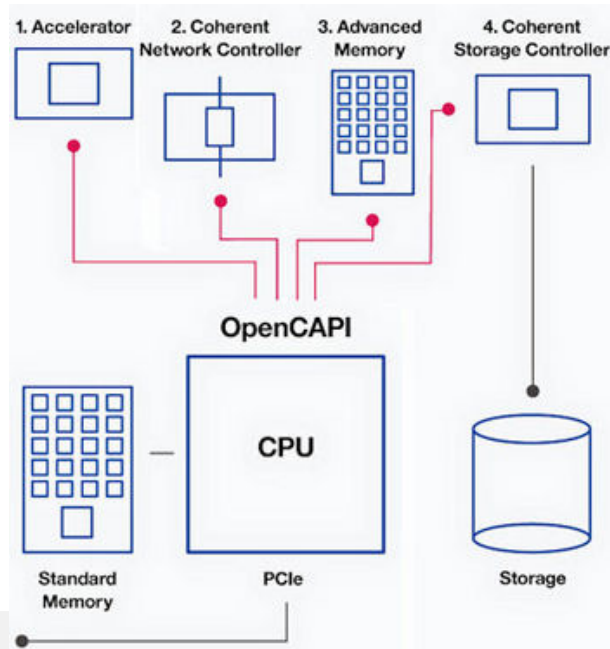
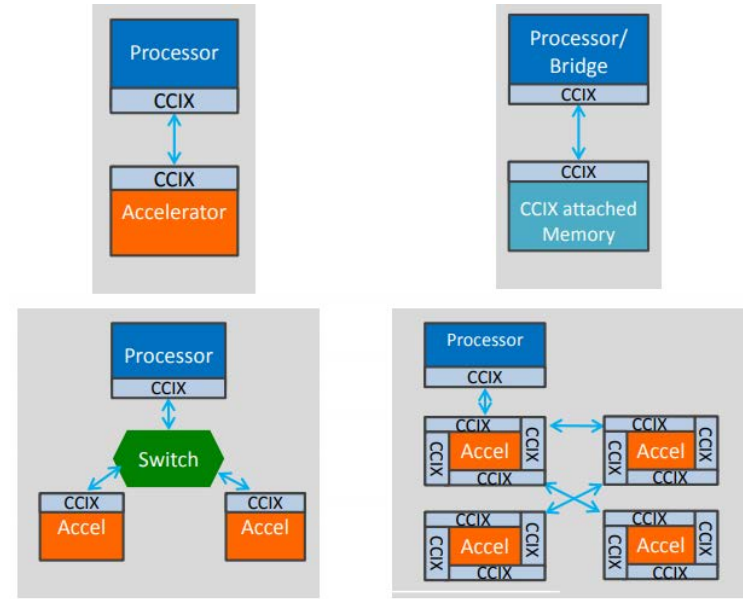
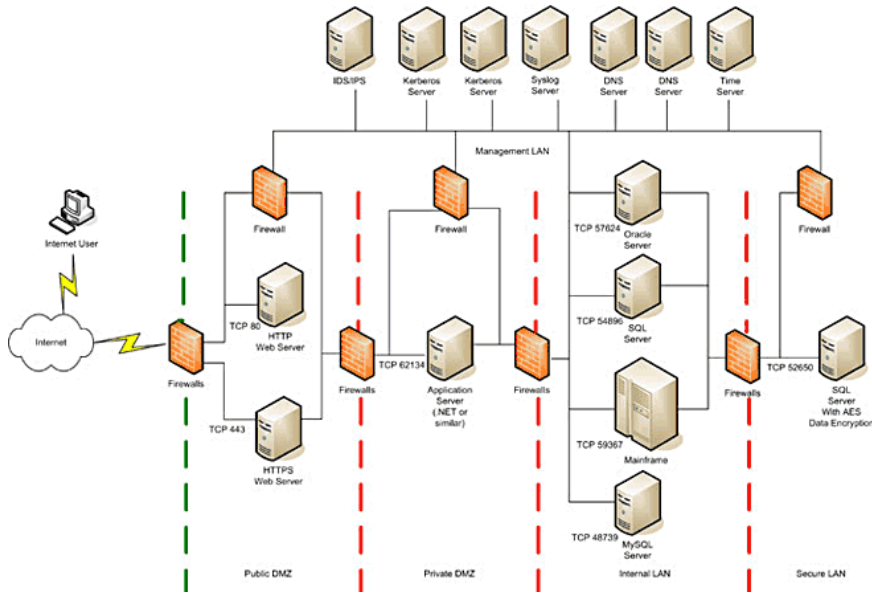
So why do it?

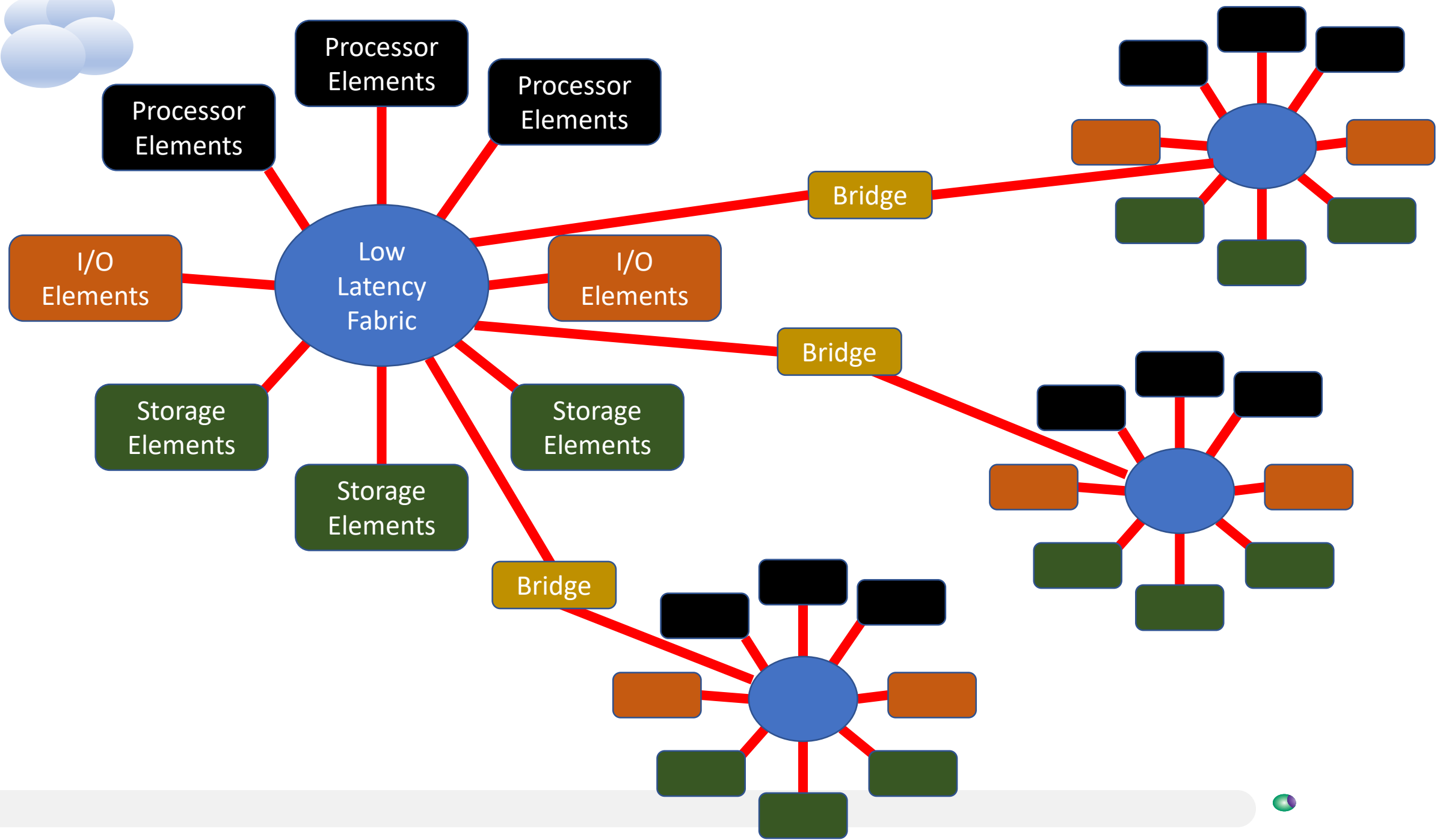
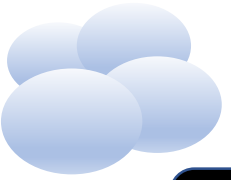


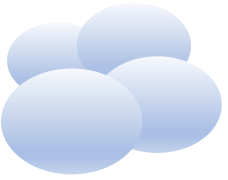


Now We Are Trending Back









Distributed resources

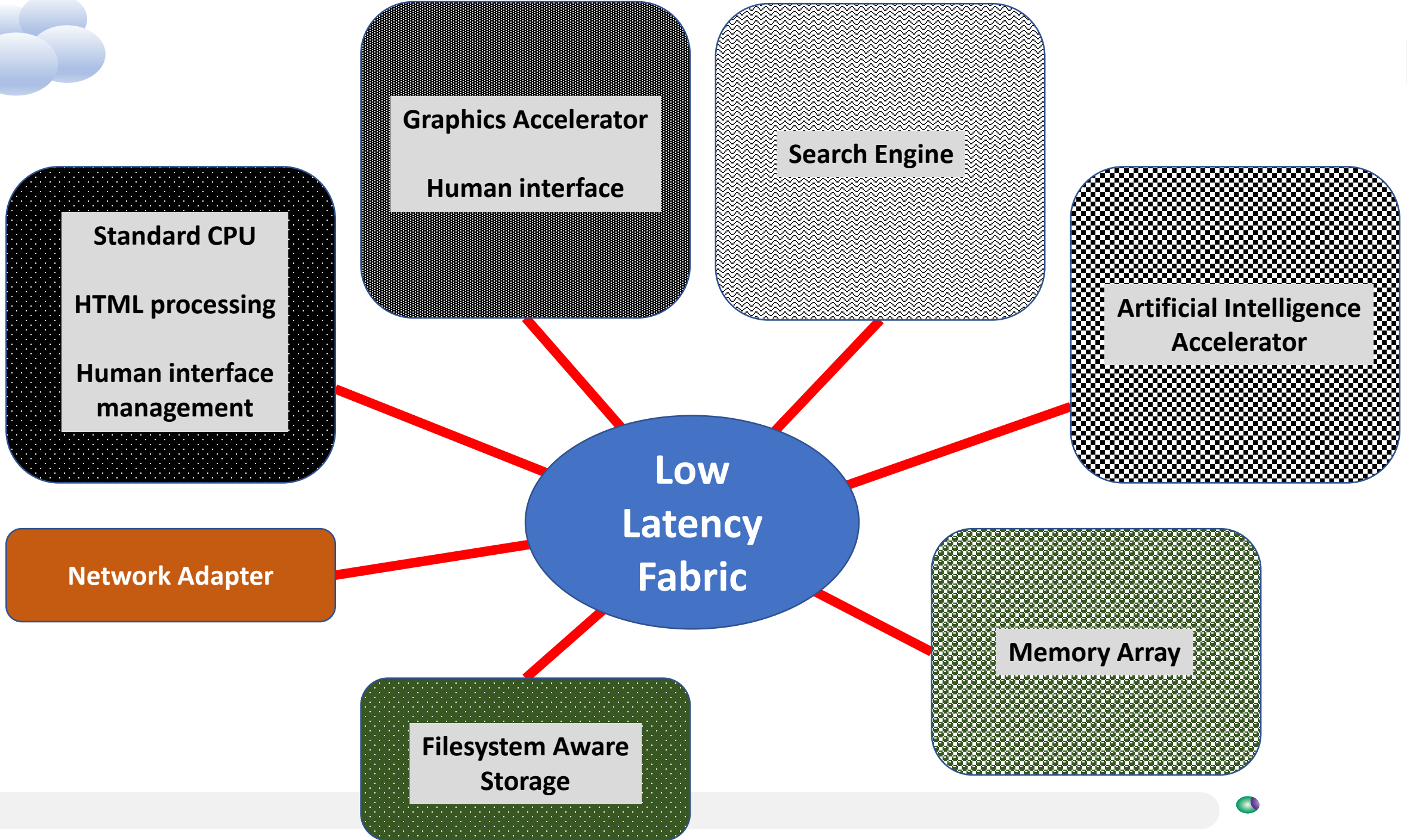
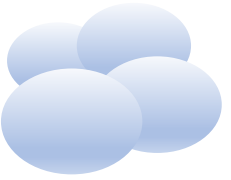
Application-specific computing

In-memory computing

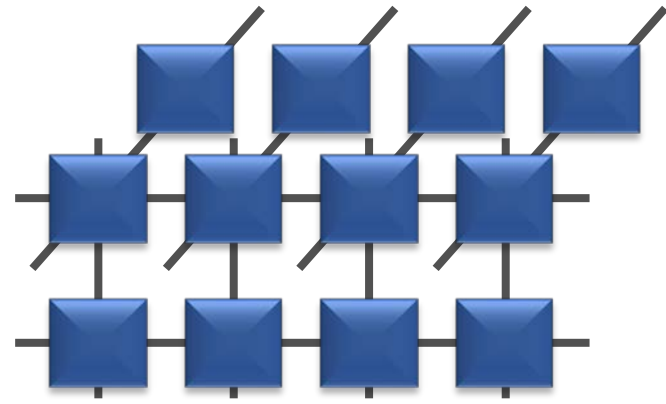
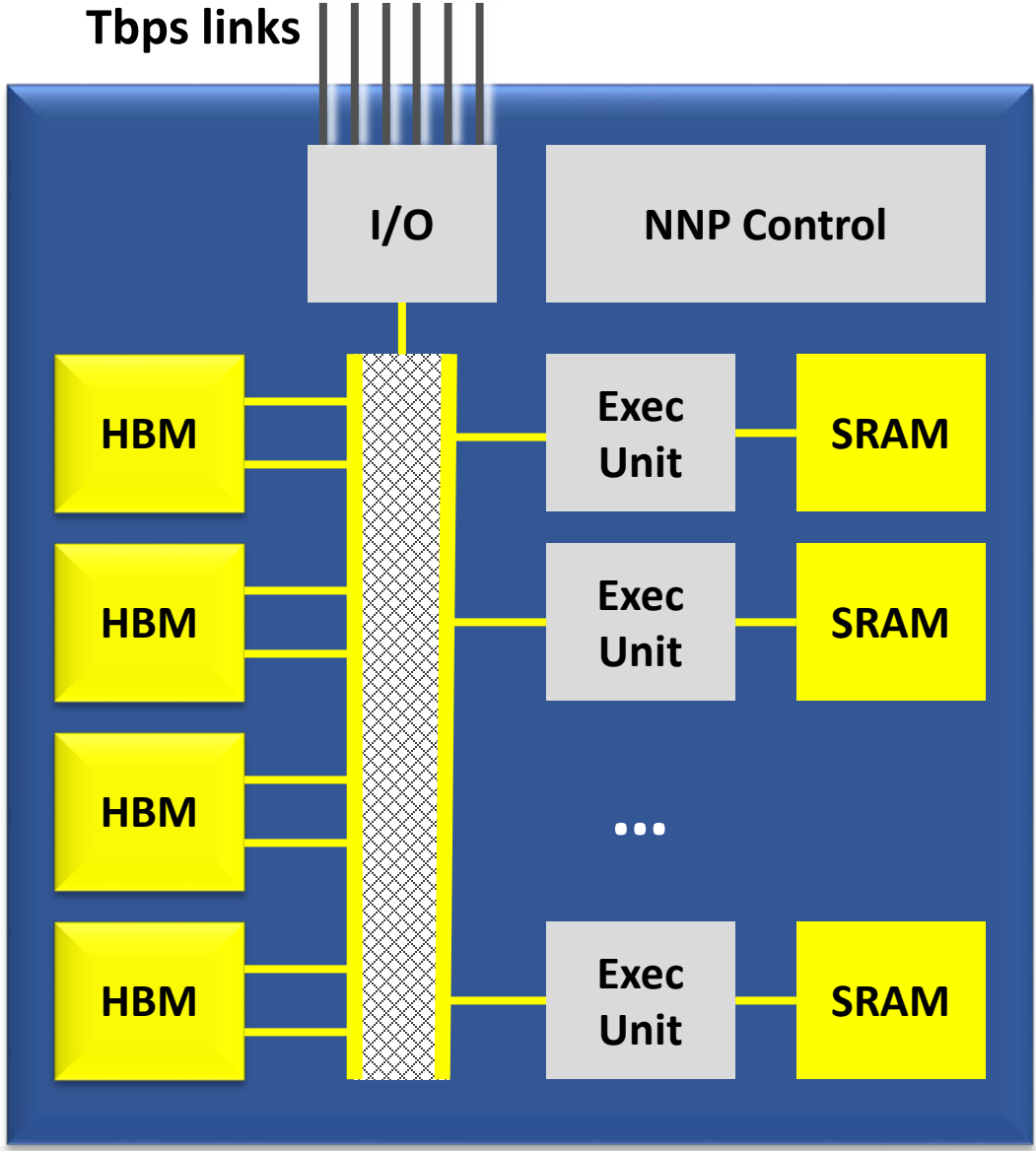
Artificial intelligence and deep learning

Security





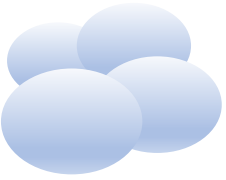
Example AI accelerator



SIMD architectures
Matrix interconnections
Fast pipes still limit load/save time

- Challenges:**
- Model checkpointing
 - Data loss on power fail
 - Temperature sensitivity

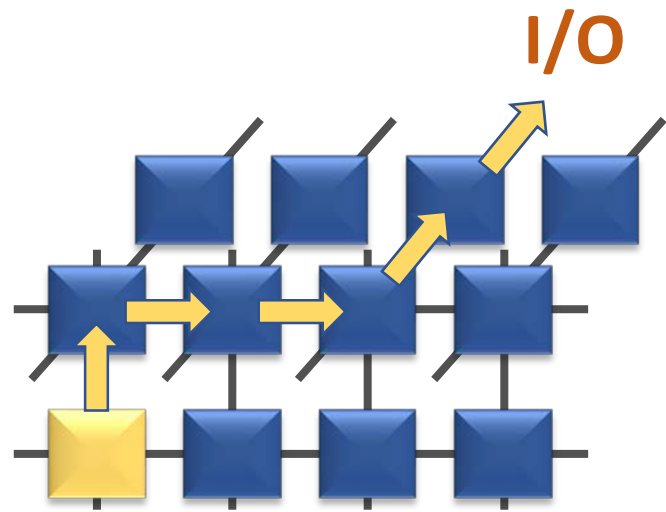
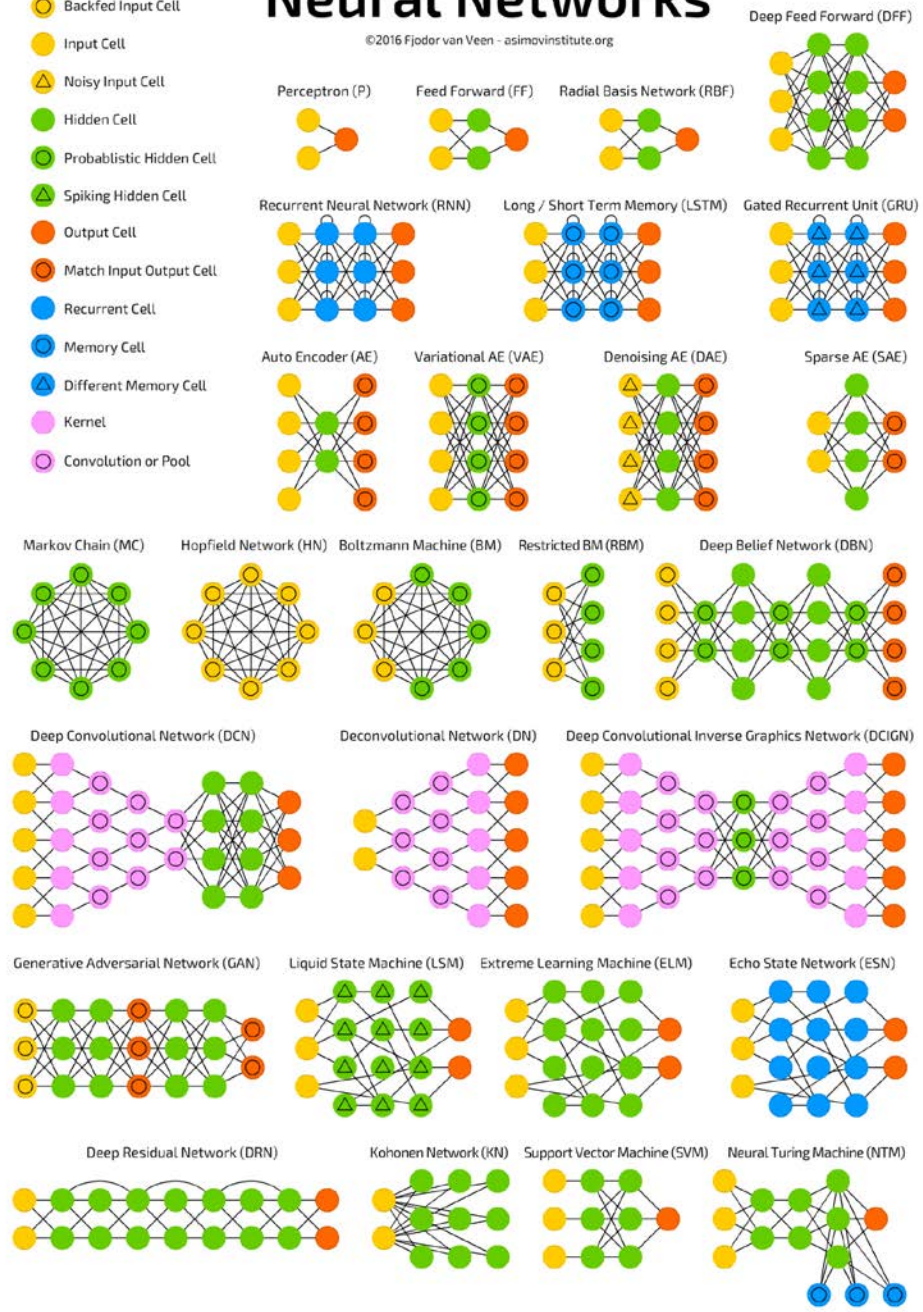




A mostly complete chart of Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org

- Backfed Input Cell
- Input Cell
- △ Noisy Input Cell
- Hidden Cell
- Probabilistic Hidden Cell
- △ Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- △ Different Memory Cell
- Kernel
- Convolution or Pool

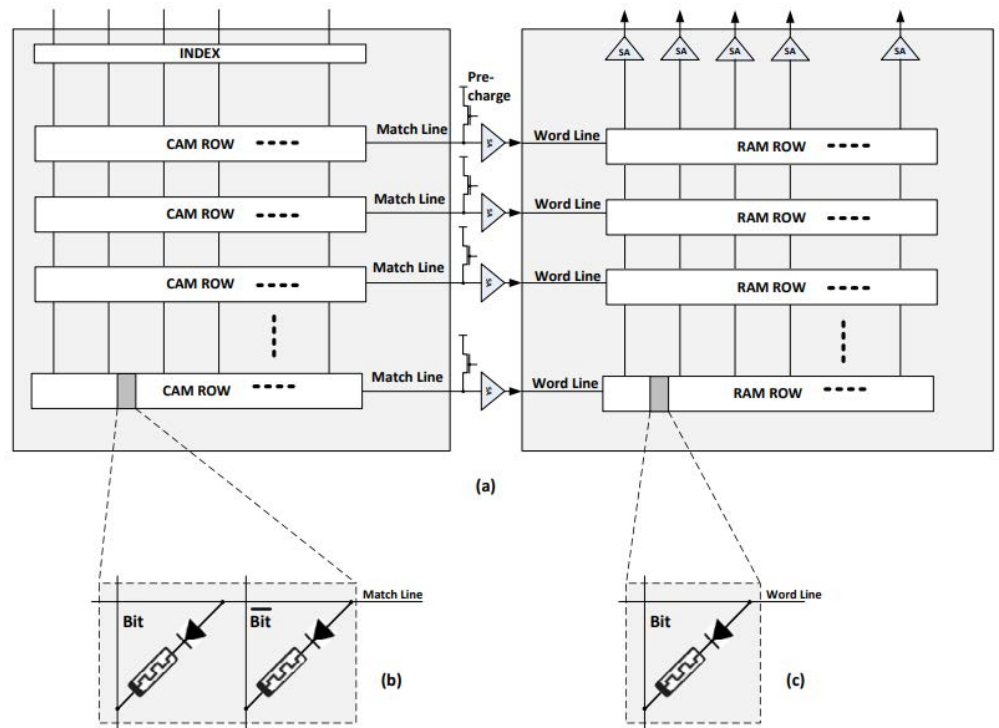
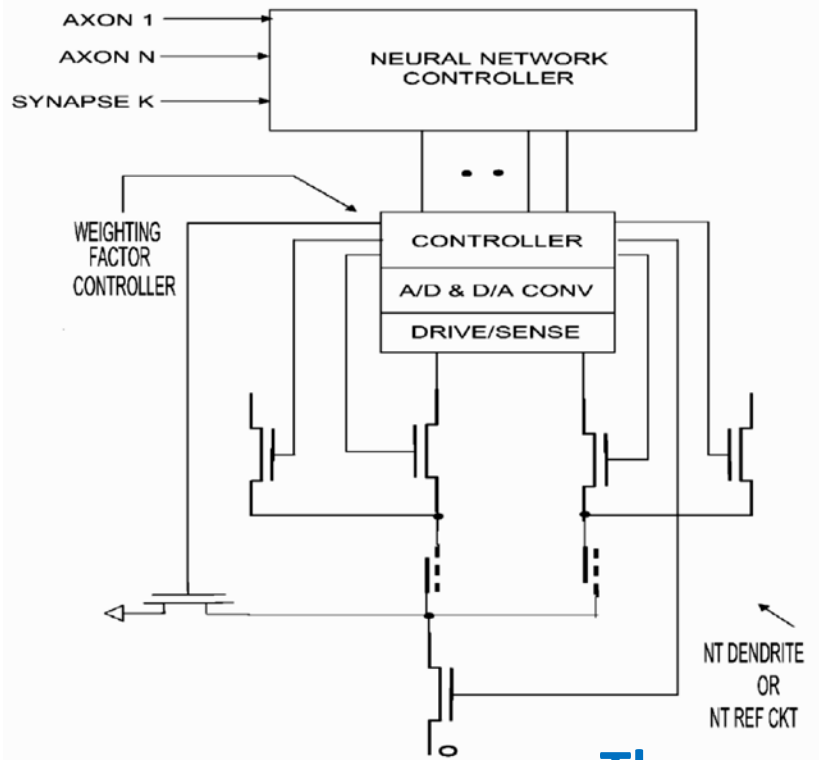
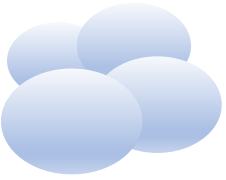


Back propagation algorithms
complicate things

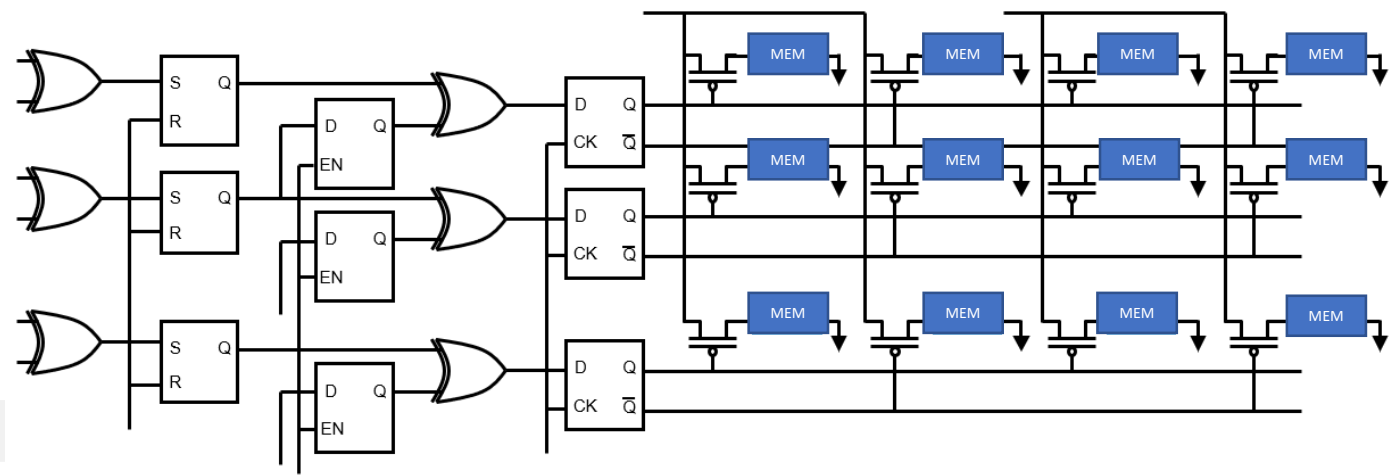
Data loss problems are amplified

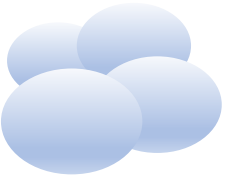
Checkpointing highly time and
bandwidth consuming





The more distributed memory gets,
the harder to load and unload

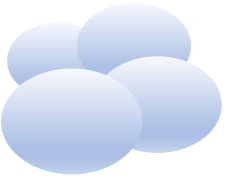




NVRAM TO THE RESCUE!

Replacing dynamic memory with persistent memory resolves the data loss issues

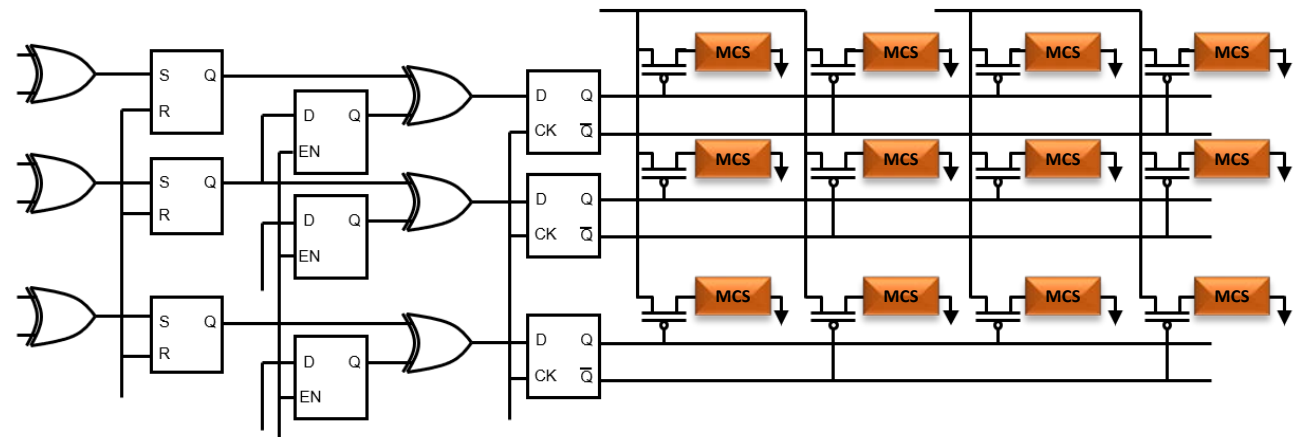
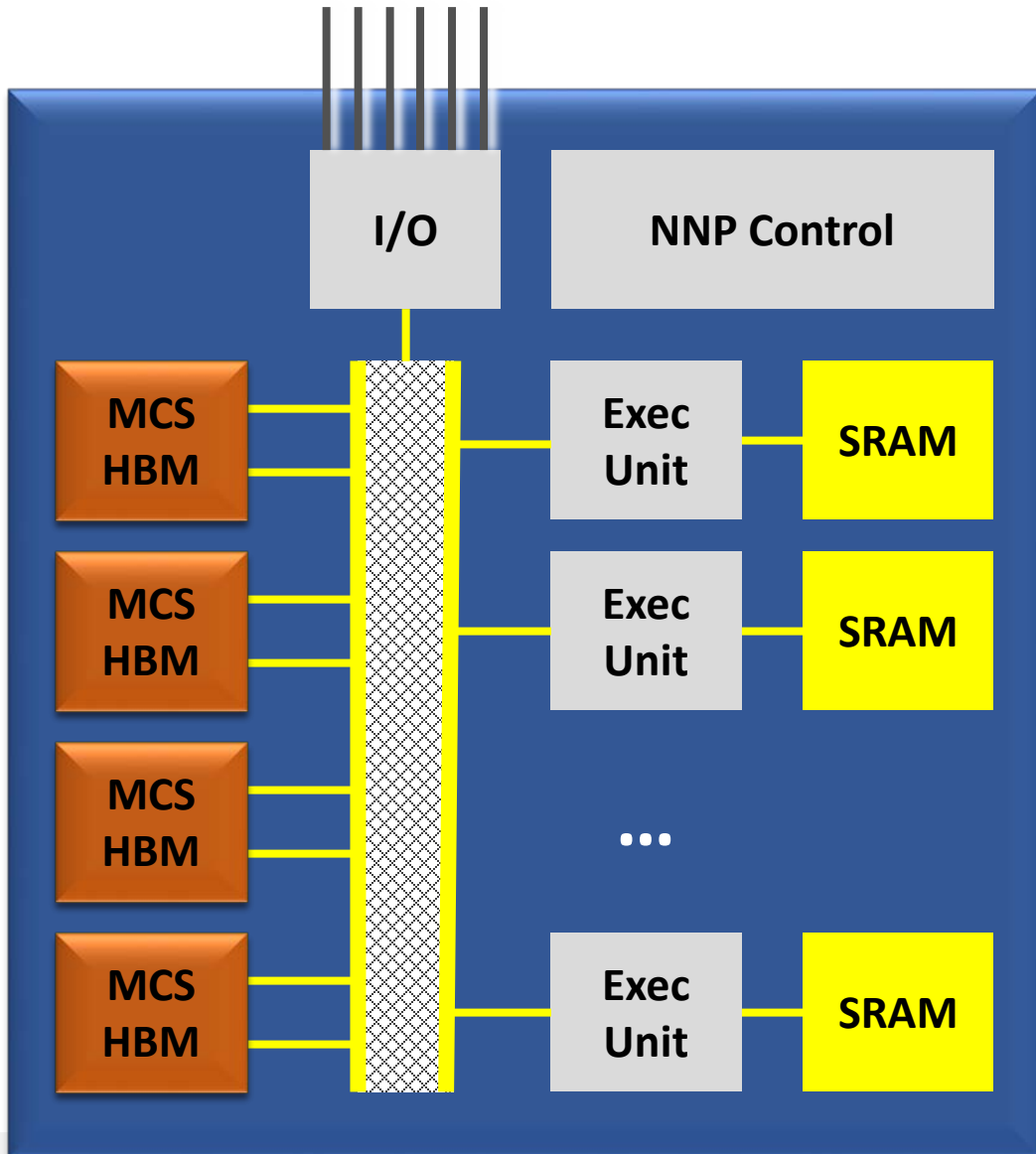




Just leave the data in place
as long as you want

Replace DRAM with NVRAM

Replace eRAM with NVRAM

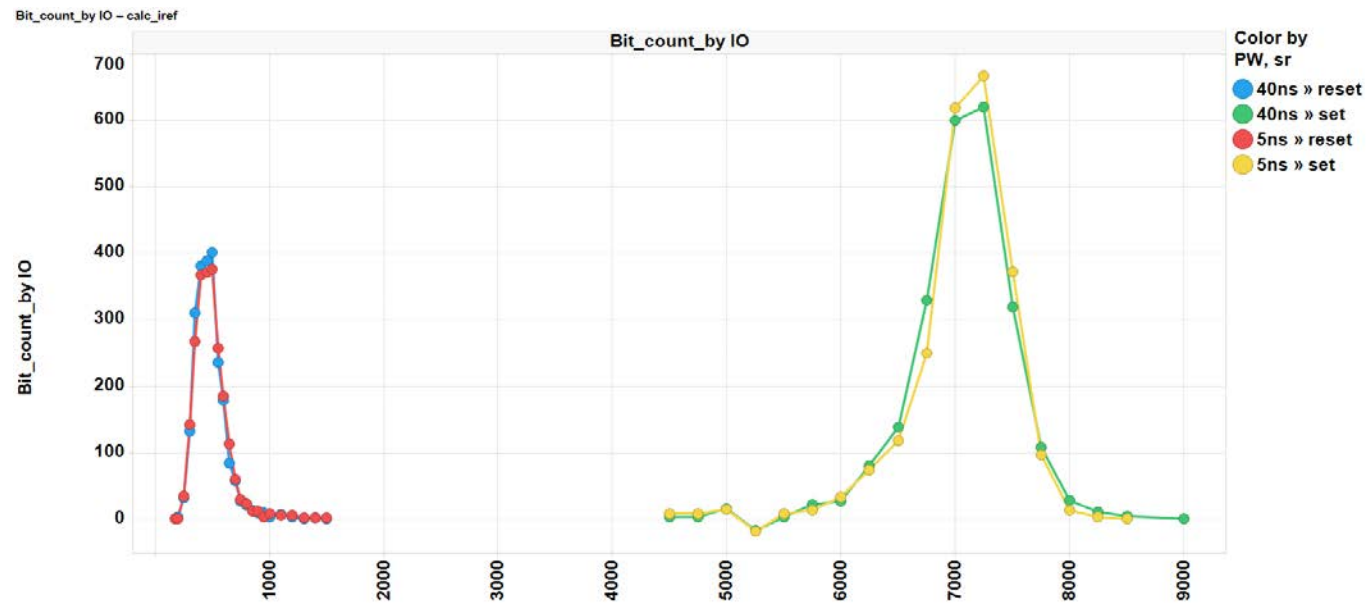




The final frontier...

SRAM & Registers

Continuing to look for ways to bring Memory Class Storage down under 1ns



Voltage adjustment

Faster edge rates

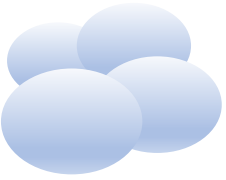
Better error check

Shadow registers

Getting smarter

It will happen





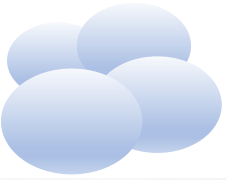
When we no longer fear
power failure...



DATA PERSISTENCE

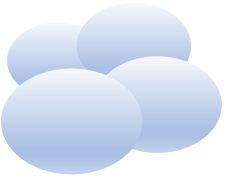
Full **END TO END** persistence



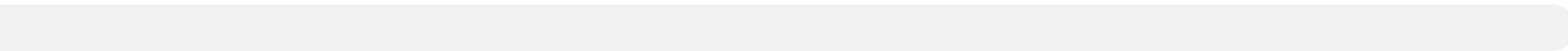


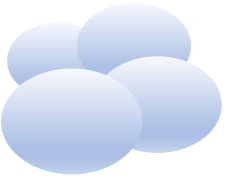
**Are we getting near the
day when we look back
at volatile memory...**

...and LAUGH?



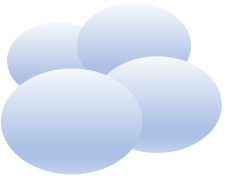
...but...





Persistent data introduces challenges, too



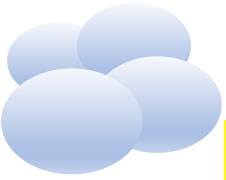


**Data is
ALWAYS
there!**



**Data security
is a growing
concern**





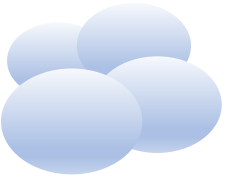
Application opens data from previous application

Memory moved from one system to another

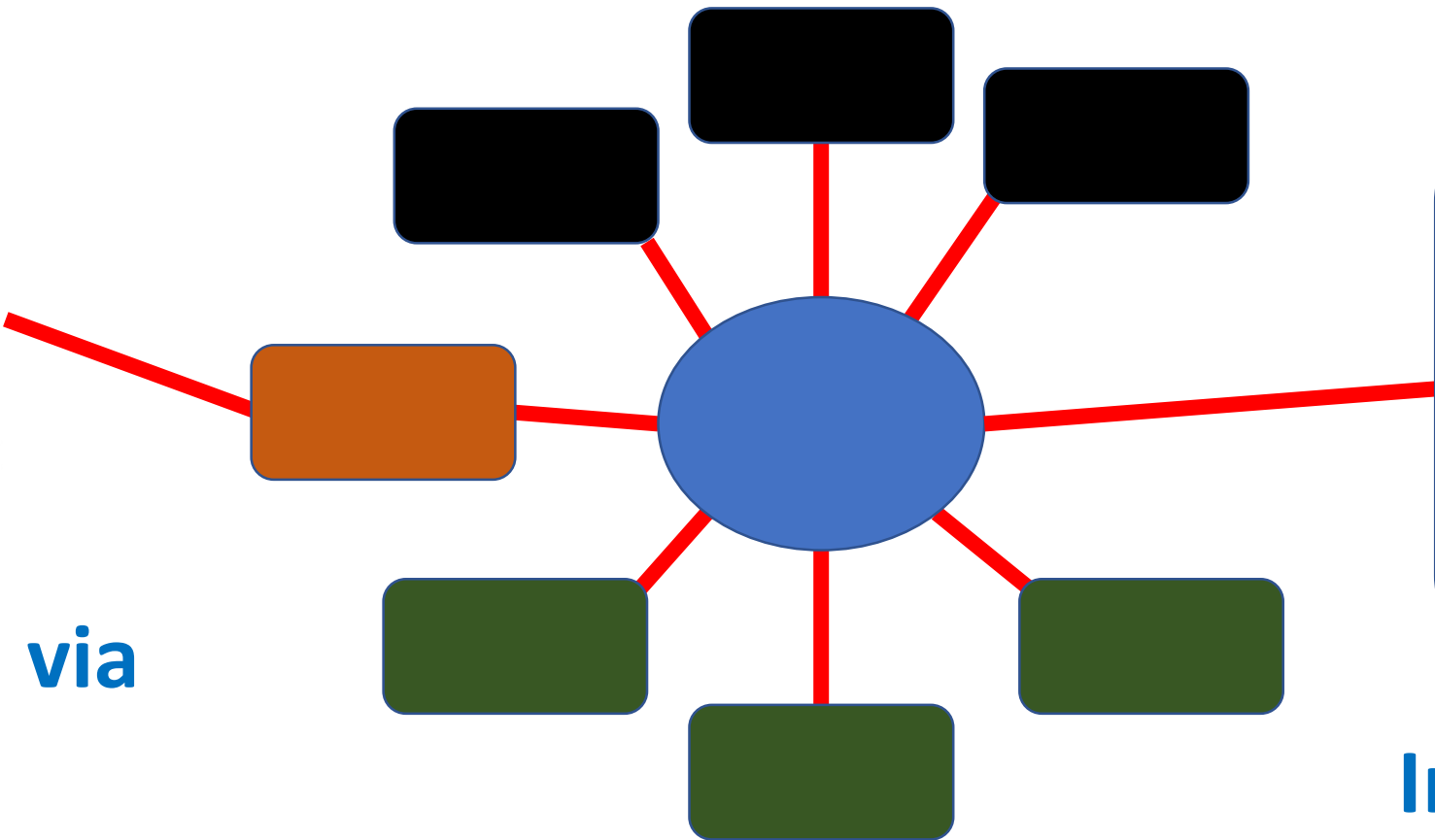
Spy devices on memory buses

So many potential breaches





**Infection via
hack**



**Infection via
spy devices**



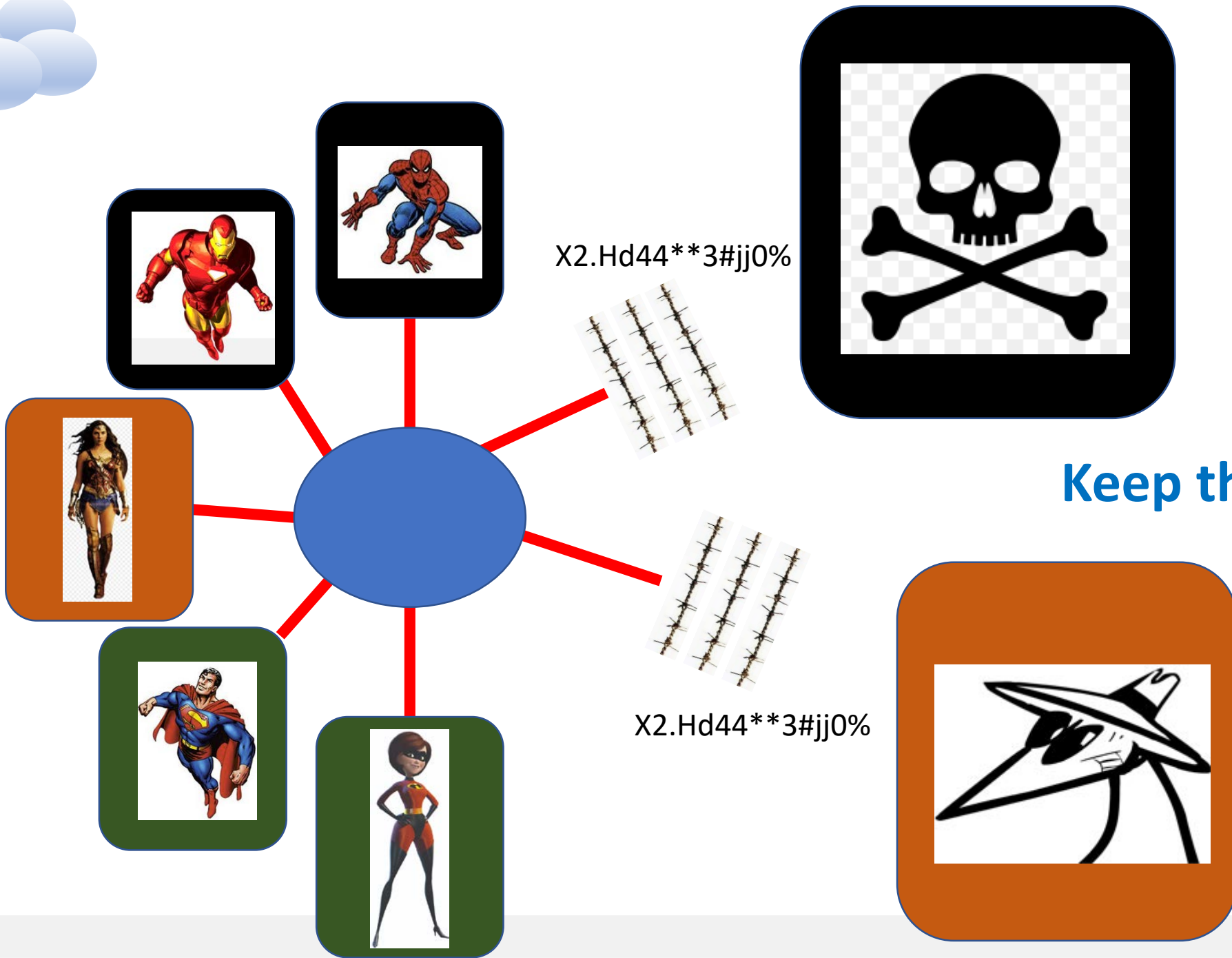
Password:



X2.Hd44**3#jj0%

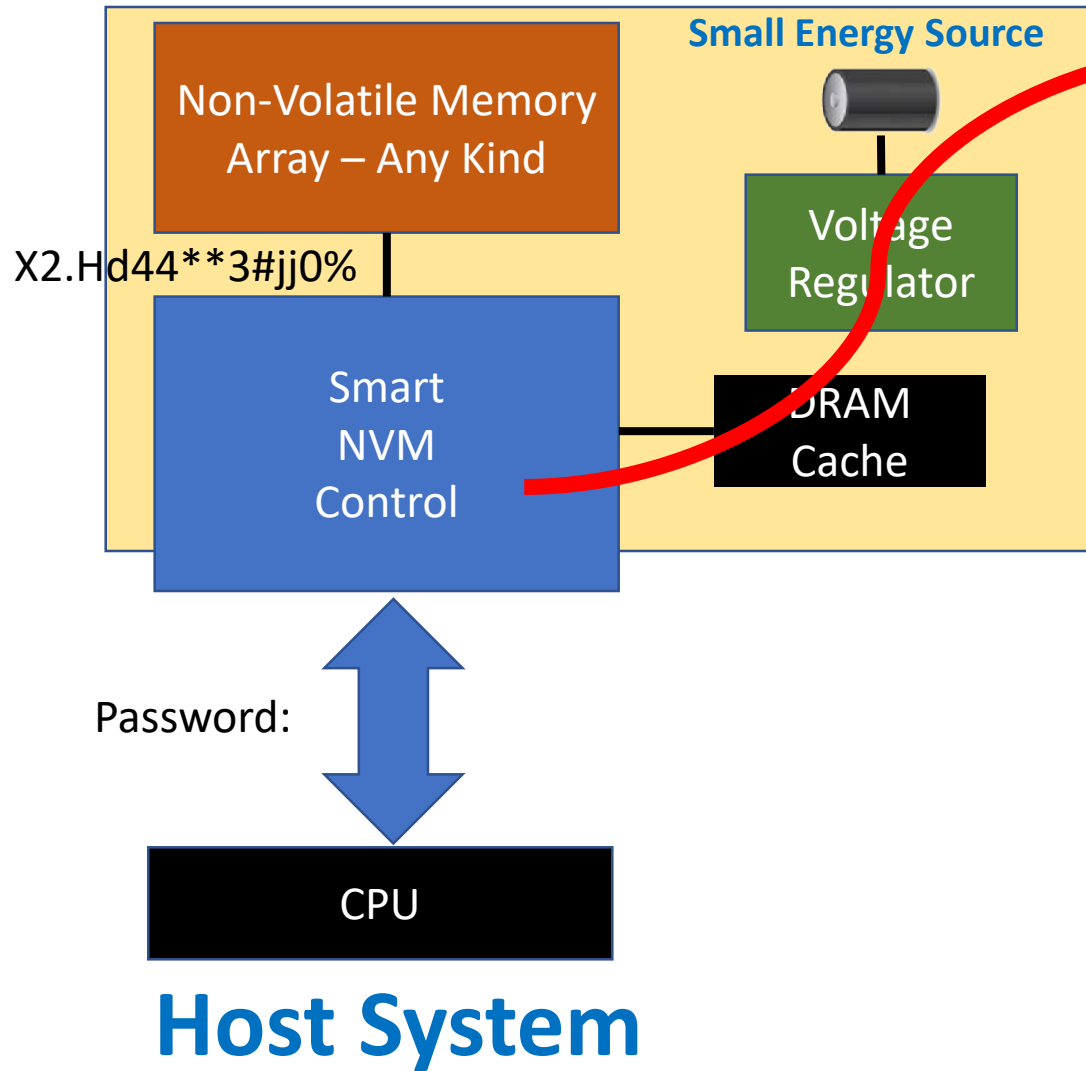
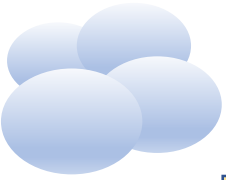
General trend is to encrypt data before transmission or storage





Keep the bad guys out





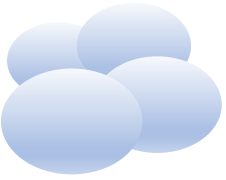
Some are adding in-memory compute functions including encryption

Works as long as the bus is secure

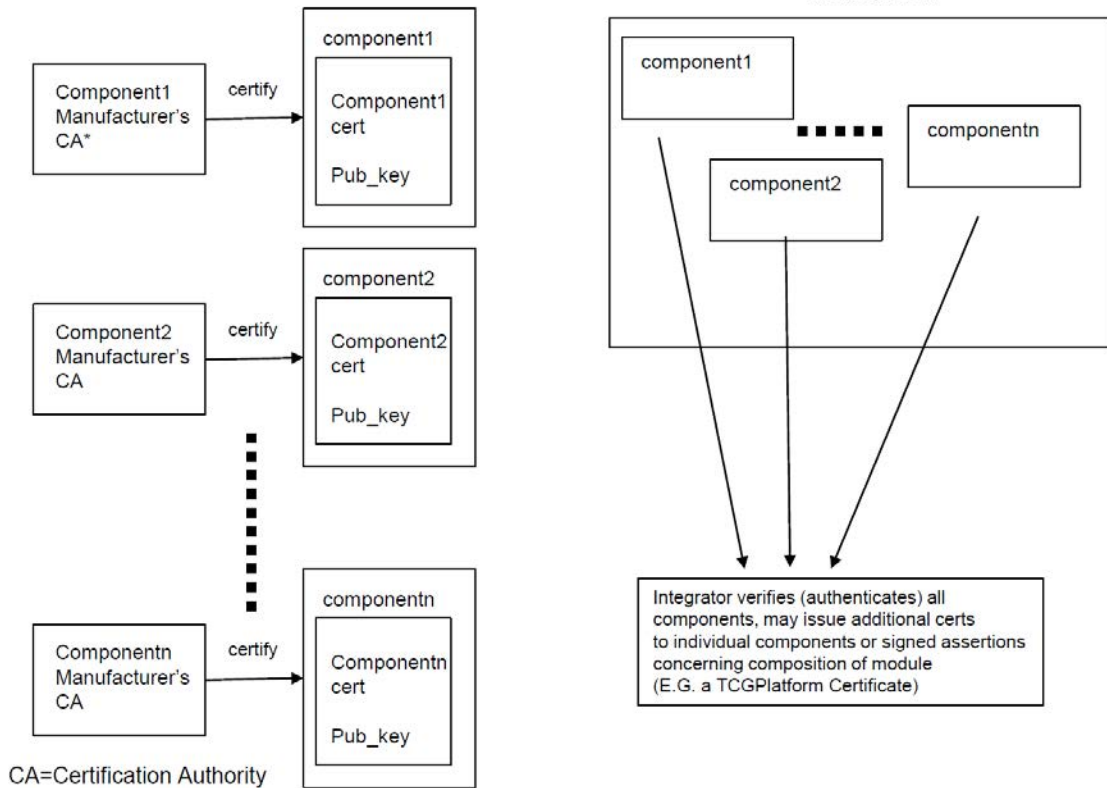
Encryption quality may be limited by block transfer size

Management of many keys can get complicated quickly





ISO/IEC 11889



ASN.1 - ISO-822-1-4;

- ITU-T X.680
- ITU-T X.681
- ITU-T X.682
- ITU-T X.683

DER - ISO-8825-1; ITU-T X.690

X509v3 - ISO-9594-8; ITU-T X.509

Common Criteria:

- Common Criteria for Information Technology Security Evaluation, Parts 1-3, Version 3.1, Revision 4, September 2010
- ISO/IEC 15408 Evaluation criteria for IT security Parts 1-3

ECDSA:

- ANSI X9.62; NIST-FIPS-186-4, Section 6
- ISO/IEC 14888-3 Digital signatures with appendix -- Part 3: Discrete logarithm based mechanisms (Clause 6.6)

NIST P256, secp256r1:

- Certicom-SEC-2, NIST-Recommended-EC
- ISO/IEC 15946 Cryptographic techniques based on elliptic curves (NIST P-256 is included as example)

SHA256:

- NIST-FIPS-180-4
- ISO/IEC 10118-3 Hash-functions -- Part 3: Dedicated hash-functions (Clause 10)

OID - ITU-T X.402

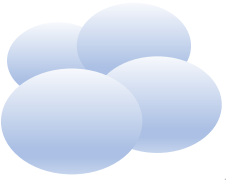
SP800-90A:

- NIST-SP-800-90A

SP800-90B:

- NIST-SP-800-90B





**Power Fail
Sucks**

**Saving Data is
a Pain**

**Need tiers
of memory
& storage**

**Persistence
is Essential**

**Sharing
Time**

Summary

**Today's
Solutions
Help**

**Persistence
Complications**

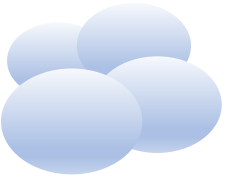
**But We
Can Do
Better**

**Data
Distribution
Challenges**

**Mix &
Match
Memories**

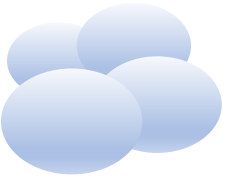
**DDR5
NVRAM Spec
in Progress**





Thank you for your time

Bill Gervasi
bilge@Nantero.com



I'm here to learn too



What do you deal with?

