NANTERO

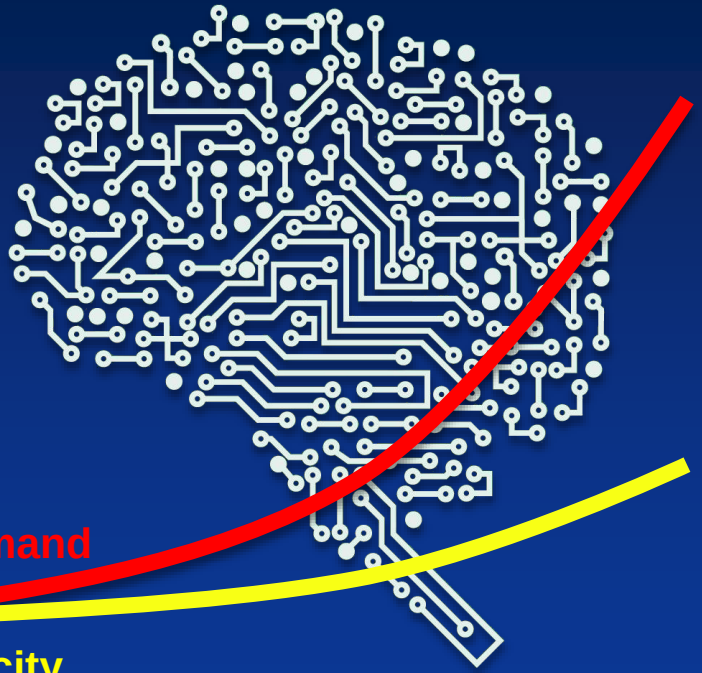
Non-Volatile Memory Modules (NVDIMMs)

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com

Demand Outpacing Capacity



Memory Demand

DRAM Capacity

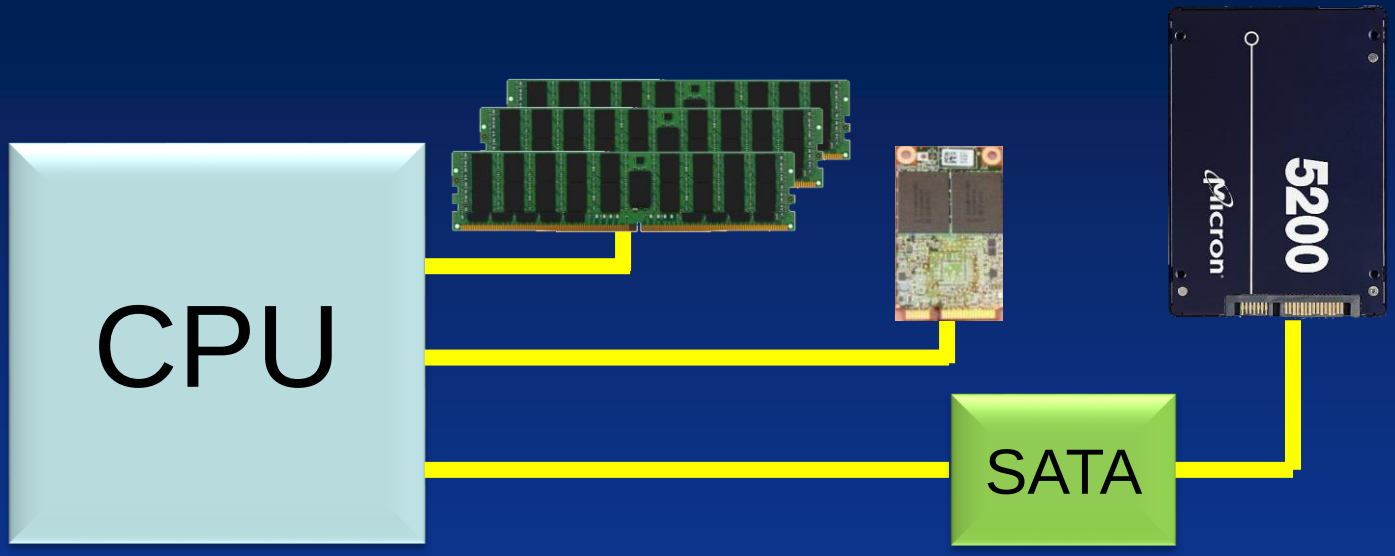
In-Memory Computing

Artificial Intelligence

Machine Learning

Deep Learning

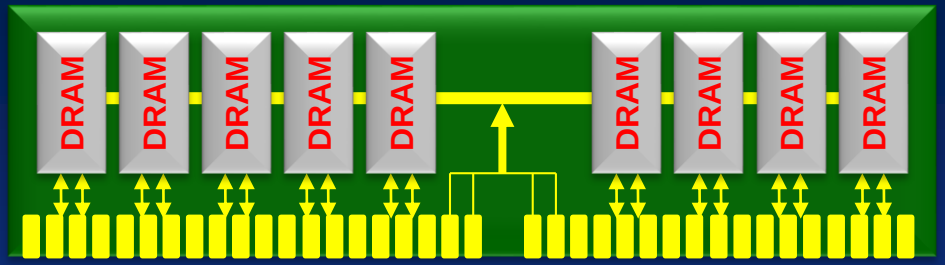
Chumming Up to the CPU



Lower latency, higher throughput

Mass storage moving closer to the CPU

DRAM Channel Protocol



Designed around direct connection to a DDR device
Fully deterministic operation required
Multiplexed address bus with:

- **Chip selects (ranks)**
- **Rows**
- **Columns**
- **Commands**

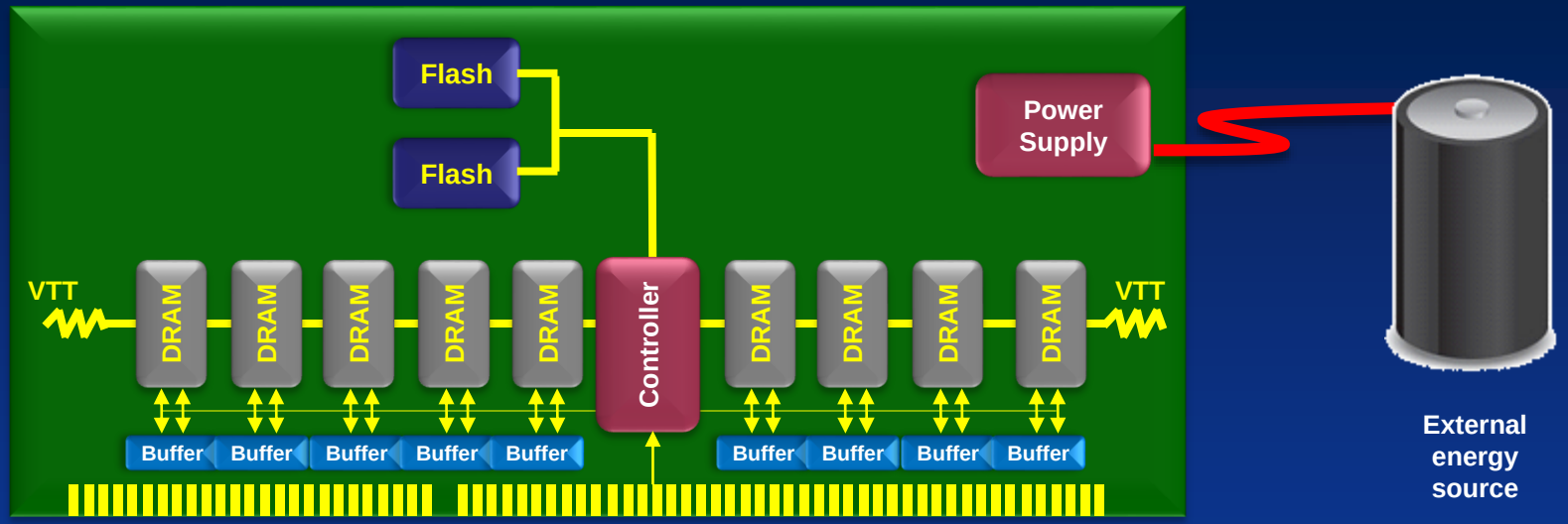
DDR4 limits:

- **16 Gb per chip**
- **144 chips per module**
- **256 GB per DIMM**

DDR5 limits

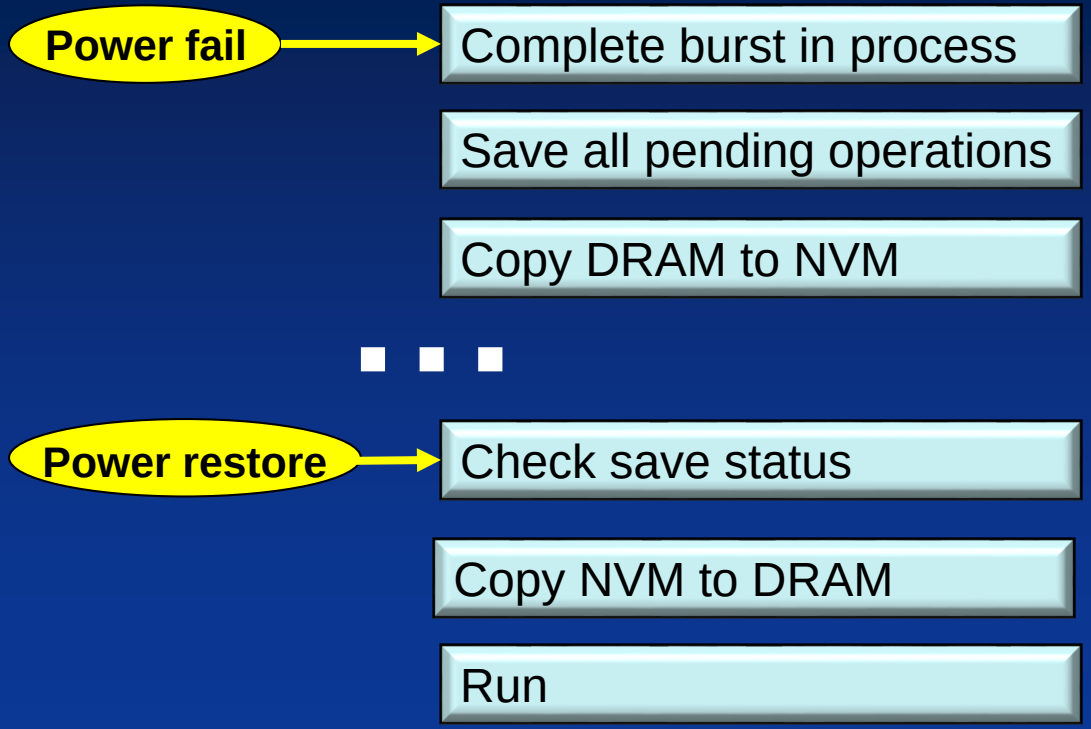
- **32 Gb per chip**
- **288 chips per module**
- **1 TB per DIMM**

NVDIMM-N, The Simplest Hybrid



CPU communicates with DRAM only
On power fail, Controller copies contents to Flash
External energy source powers NVDIMM until backed up

NVDIMM-N Backup Protocol

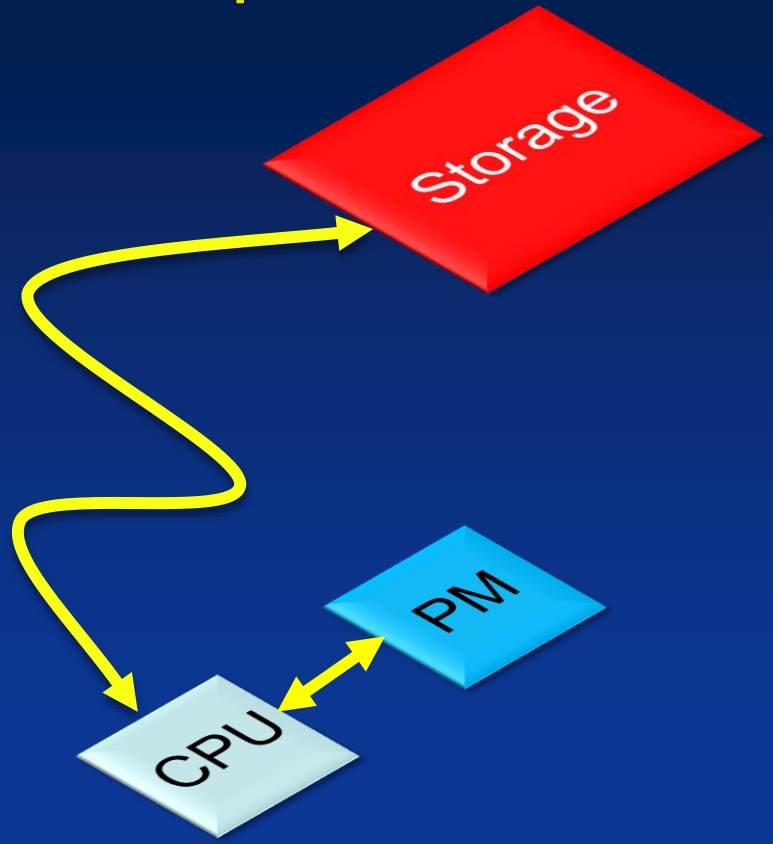


Why is Persistence Important?

Power failure is a key factor in server software design

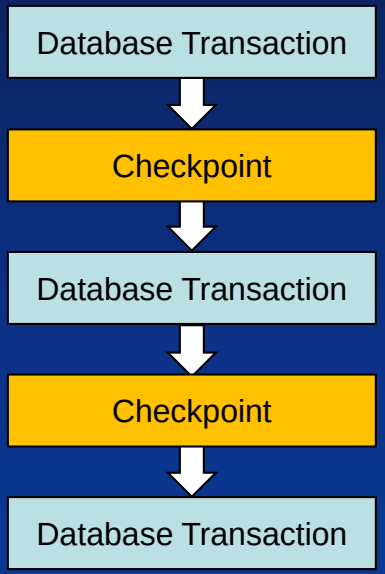
Checkpointing intermediate results to storage affects performance

Data persistence near the CPU is a huge improvement in systems architecture

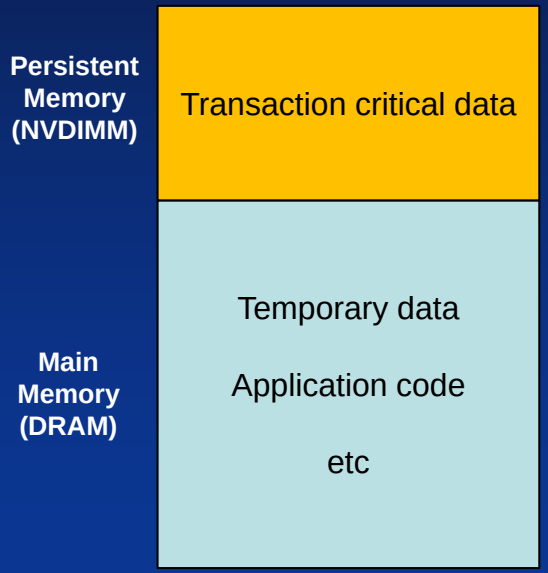


Persistence in Main Memory

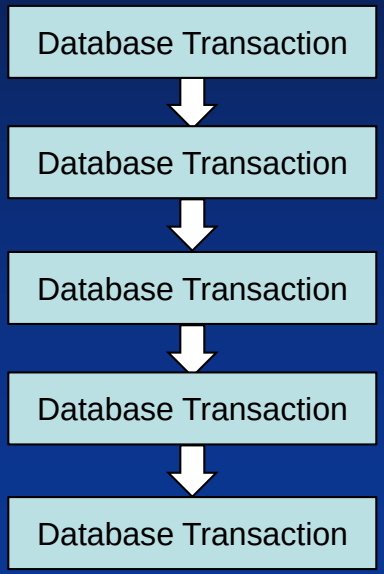
Old Process



Add Persistence Memory



New Process

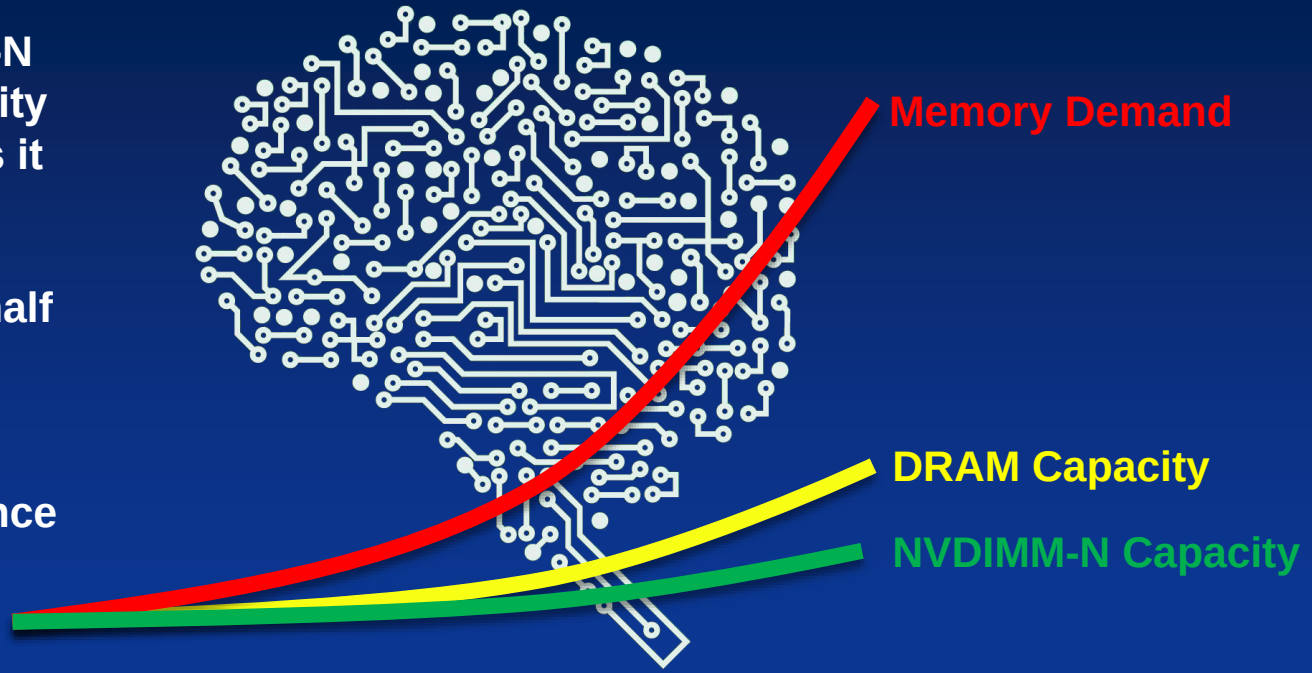


NVDIMM-N Capacity Limitations

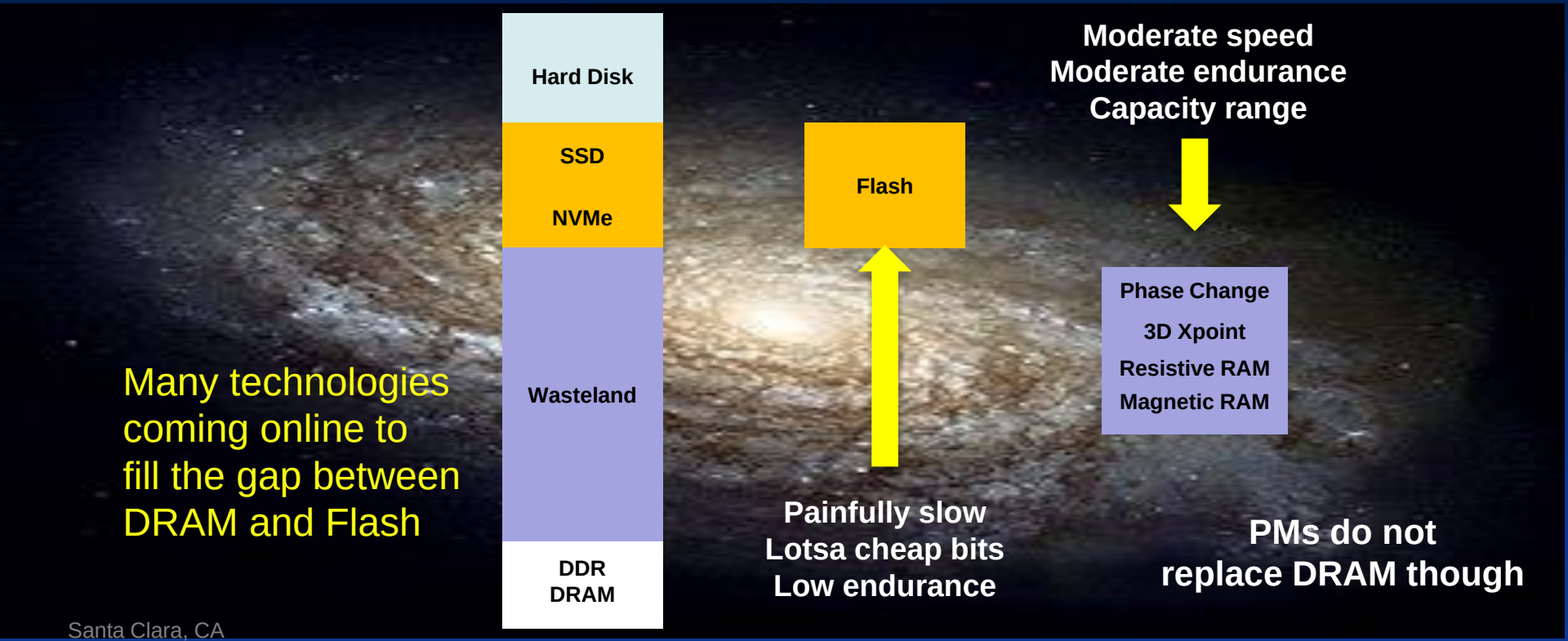
Unfortunately, NVDIMM-N doesn't solve the capacity demand… in fact makes it worse

NVDIMM-N capacity is half of the equivalent DRAM module capacity

Does add data persistence



Universe of Persistent Memories



Virtualizing the DRAM Channel

Incorporating PM into the DRAM channel requires:

- **Mapping devices into the DRAM address range**
- **Allowing for non-determinism for bookkeeping operations**

RAS & CAS Redefined

Mapped Blocks

Limited write endurance forces PMs to go offline for operations such as wear leveling

Media agnostic; any PM can be on the local bus

Virtualizing the DRAM Channel

DDR4

Function	Abbreviation	CKE Previous Cycle	CKE Current Cycle	CS_n	ACT_n	RAS_n/A16	CAS_n/A15	WE_n/A14	BG0-BG1	BA0-BA1	C2-C0	A12/BC_n	A17, A13, A11	A10/AP	A0-A9
Mode Register Set	MRS	H	H	L	H	L	L	L	BG	BA	V	OP Code			
Refresh	REF	H	H	L	H	L	L	H	V	V	V	V	V	V	V
Self Refresh Entry	SRE	H	L	L	H	L	L	H	V	V	V	V	V	V	V
Self Refresh Exit	SRX	L	H	H	X	X	X	X	X	X	X	X	X	X	X
				L	H	H	H	H	V	V	V	V	V	V	V
Single Bank Precharge	PRE	H	H	L	H	L	H	L	BG	BA	V	V	V	L	V
Precharge all Banks	PREA	H	H	L	H	L	H	L	V	V	V	V	V	H	V
RFU	RFU	H	H	L	H	L	H	H	RFU						
Bank Activate	ACT	H	H	L	L	Row Address(RA)			BG	BA	V	Row Address (RA)			
Write (Fixed BL8 or BC4)	WR	H	H	L	H	H	L	L	BG	BA	V	V	V	L	CA
Write (BC4, on the Fly)	WRS4	H	H	L	H	H	L	L	BG	BA	V	L	V	L	CA
Write (BL8, on the Fly)	WRS8	H	H	L	H	H	L	L	BG	BA	V	H	V	L	CA
Write with Auto Precharge (Fixed BL8 or BC4)	WRA	H	H	L	H	H	L	L	BG	BA	V	V	V	H	CA
Write with Auto Precharge (BC4, on the Fly)	WRAS4	H	H	L	H	H	L	L	BG	BA	V	L	V	H	CA
Write with Auto Precharge (BL8, on the Fly)	WRAS8	H	H	L	H	H	L	L	BG	BA	V	H	V	H	CA
Read (Fixed BL8 or BC4)	RD	H	H	L	H	H	L	H	BG	BA	V	V	V	L	CA
Read (BC4, on the Fly)	RDS4	H	H	L	H	H	L	H	BG	BA	V	L	V	L	CA
Read (BL8, on the Fly)	RDS8	H	H	L	H	H	L	H	BG	BA	V	H	V	L	CA
Read with Auto Precharge (Fixed BL8 or BC4)	RDA	H	H	L	H	H	L	H	BG	BA	V	V	V	H	CA
Read with Auto Precharge (BC4, on the Fly)	RDAS4	H	H	L	H	H	L	H	BG	BA	V	L	V	H	CA
Read with Auto Precharge (BL8, on the Fly)	RDAS8	H	H	L	H	H	L	H	BG	BA	V	H	V	H	CA
No Operation	NOP	H	H	L	H	H	H	H	V	V	V	V	V	V	V
Device Deselected	DES	H	H	H	X	X	X	X	X	X	X	X	X	X	X
Power Down Entry	PDE	H	L	H	X	X	X	X	X	X	X	X	X	X	X
Power Down Exit	PDX	L	H	H	X	X	X	X	X	X	X	X	X	X	X
ZQ calibration Long	ZQCL	H	H	L	H	H	H	L	V	V	V	V	V	H	V
ZQ calibration Short	ZQCS	H	H	L	H	H	H	L	V	V	V	V	V	L	V

NVDIMM-P

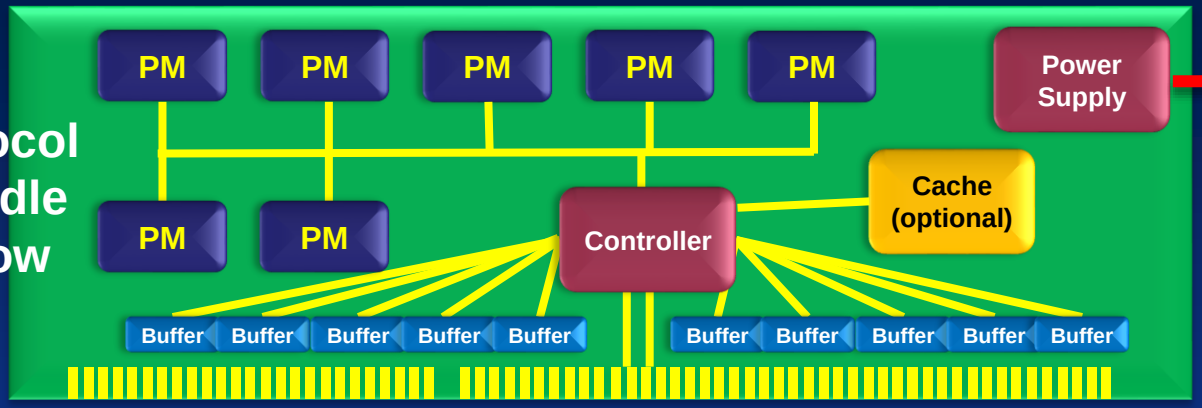
Function (Reference)	CKE_0 Previous	CKE_0 Current	CS_n	ACT_n	RAS_n / A16	CAS_n / A15	WE_n / A14	BG1-BG0	BA1-BA0	C2-C0	A12 / BC_n	A17	A13	A11	A10 / AP	A9-A0
MRS (Mode Register Set)	H	H	L	H	L	L	L	V	V	V	OP CODE					
XADR	H	H	L	ADDR[22:12]							XREAD & SREAD : RID [4:0] PWRITE : WGID [4:0] XWRITE: RFU UNMAP:LENGTH[4:0]					ADDR[11:2]
XWRITE	H	H	L	H	H	L	L	ADDR[39:33]			RFU			L	RFU	ADDR[32:23]
PWRITE	H	H	L	H	H	L	L	ADDR[39:33]			WGID[7:5]			H	Persist	ADDR[32:23]
SEND	H	H	L	H	H	L	H	RFU			RFU			L	L	RFU
SEND-W_PER	H	H	L	H	H	L	H	RFU			RFU			L	H	RFU
SREAD	H	H	L	H	H	L	H	ADDR[39:33]			RID[7:5]			H	RFU	ADDR[32:23]
XREAD	H	H	L	H	L	H	H	ADDR[39:33]			RID[7:5]			L	RFU	ADDR[32:23]
UNMAP	H	H	L	H	L	H	H	ADDR[39:33]			L	L	L	H	OPCODE[0]	ADDR[32:23]
FLUSH	H	H	L	H	L	H	H	RFU			H	H	L	H	~~RFU~~ Final	FL[1:0]+ WGID[7:0]
IOP	H	H	L	H	L	H	H	RFU			L	H	H	H	RFU	RFU[2:0]+ IOP TS[1:0]+ IOP TU[4:0]
NOP	H	H	L	H	H	H	H	V			V	V	V	V	V	V
DESELECT	H	H	H	X	X	X	X	X			X	X	X	X	X	X
POWER DOWN ENTRY	H	L	H	X	X	X	X	X			X	X	X	X	X	X IOP TS[1:0]+ IOP TU[4:0]
POWER DOWN EXIT	L	H	H	X	X	X	X	X			X	X	X	X	X	X
ZQ Calibration Long	H	H	L	H	H	H	L	V			V	V	V	V	H	V
ZQ Calibration Short	H	H	L	H	H	H	L	V			V	V	V	V	L	V

Different protocols but can share the same wires

NANTERO

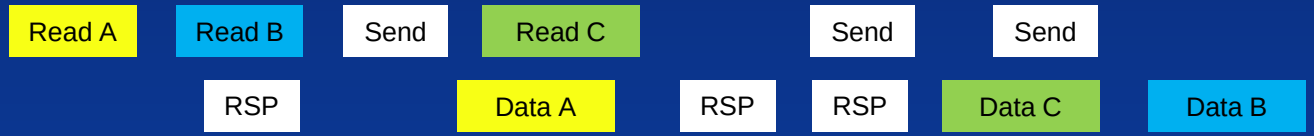
The NVDIMM-P Protocol

NVDIMM-P protocol invented to handle memory with low endurance



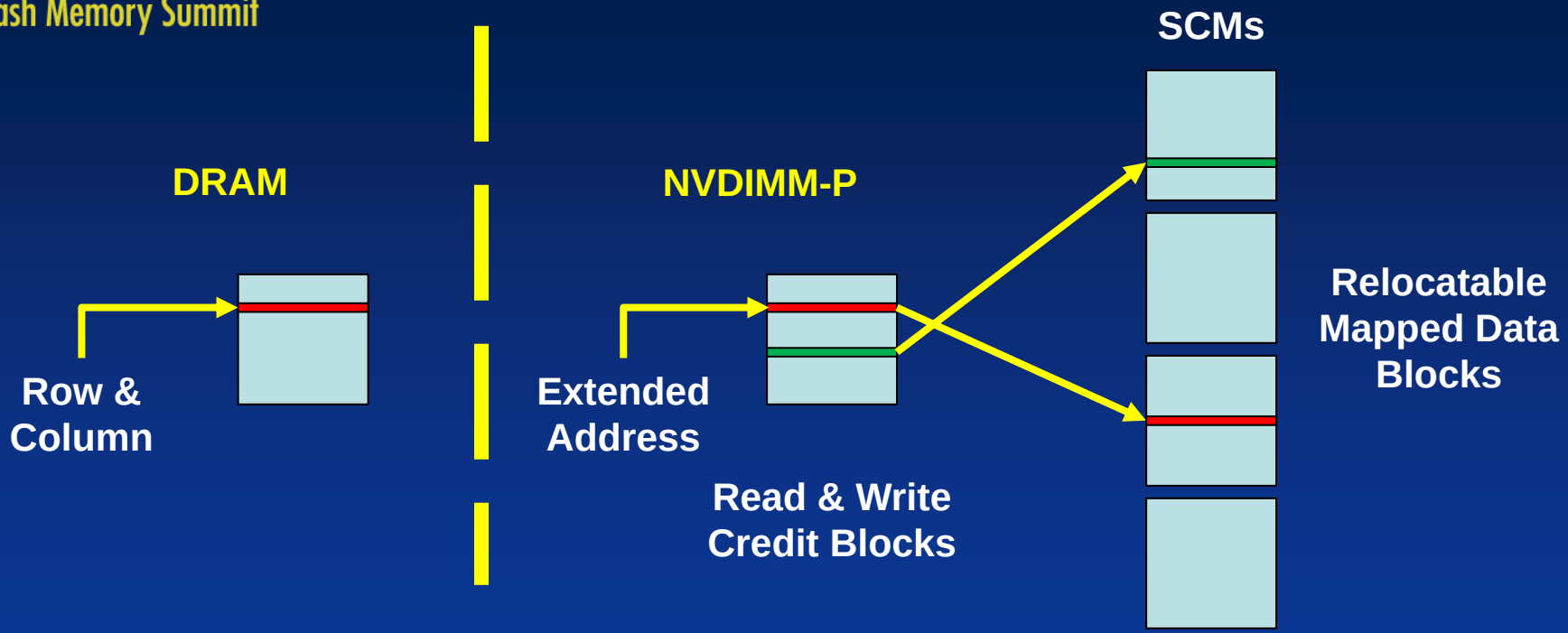
Optional external energy source

Non-deterministic credit based system allows time for bookkeeping



Out-of-order data returned with ID

Big Data Over -P Protocol



-P Protocol is NOT DRAM
Coexists by supporting DRAM timing and ODT decoding

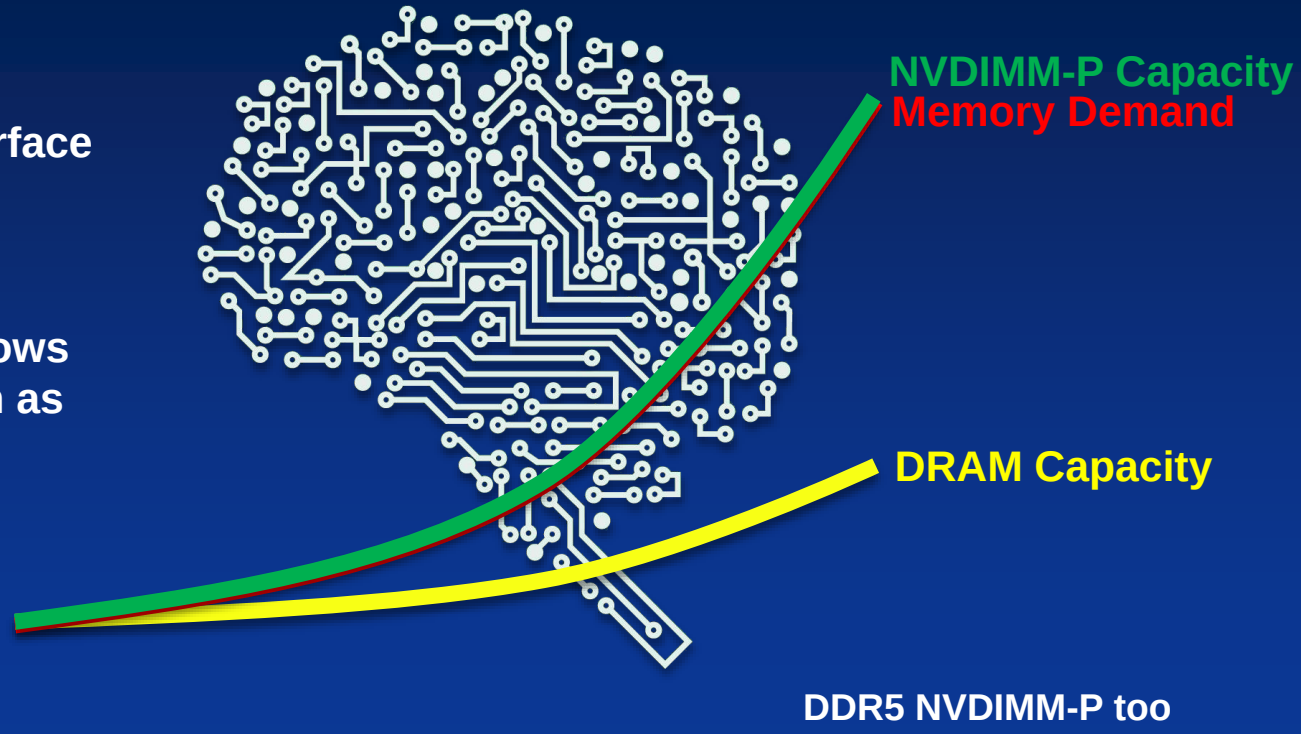
NANTERO

NVDIMM-P Capacity

NVDIMM-P Protocol extends the DDR interface to enable big data

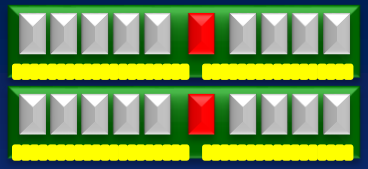
Out-of-order non-deterministic data allows for bookkeeping such as wear leveling

Requires new CPU



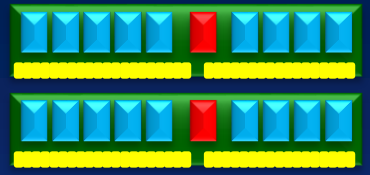
DDR5 NVDIMM-P too

Software Issues



All NVDIMM-N

No problem, all memory persistent, all memory has same performance

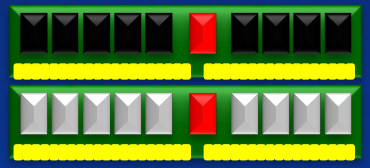


All NVDIMM-P

No problem, all memory persistent, all memory has same performance

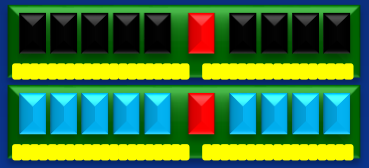
Symmetric solutions are simplest; no software changes, accept the performance you get

Asymmetric solutions are more complicated, software partitioning required, many solution punt by mounting NVDIMM as an SSD



Mix of NVDIMM-N & DRAM

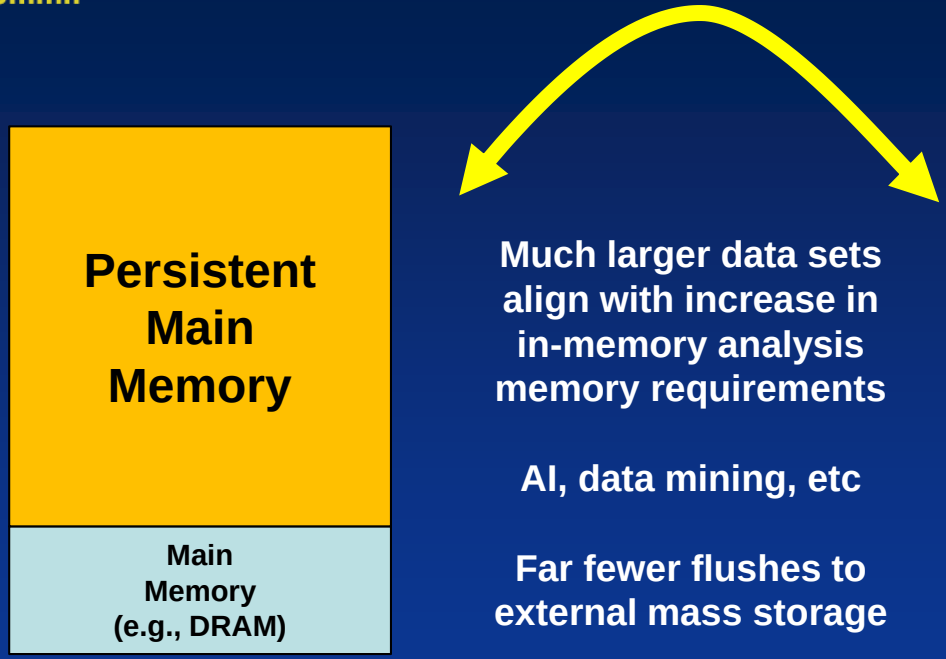
Complicates the solution
Software must separate persistent data from ephemeral data



Mix of NVDIMM-P & DRAM

NVDIMM-P can mount as extended memory with asymmetric performance or simply as SSD

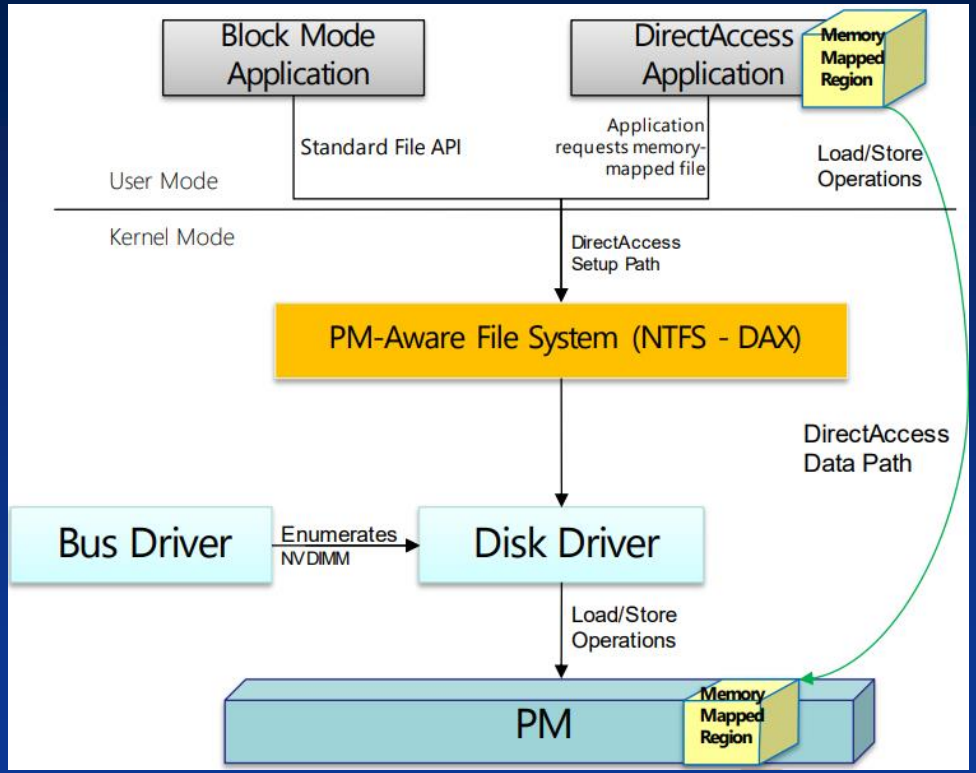
Advantage of Large Capacity PM



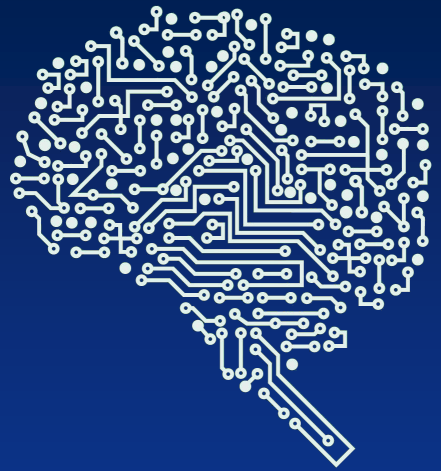
Tuning Software for Persistence

Operating systems have "new" hooks for persistent memory

Both disk mount and direct access enabled



Data Security

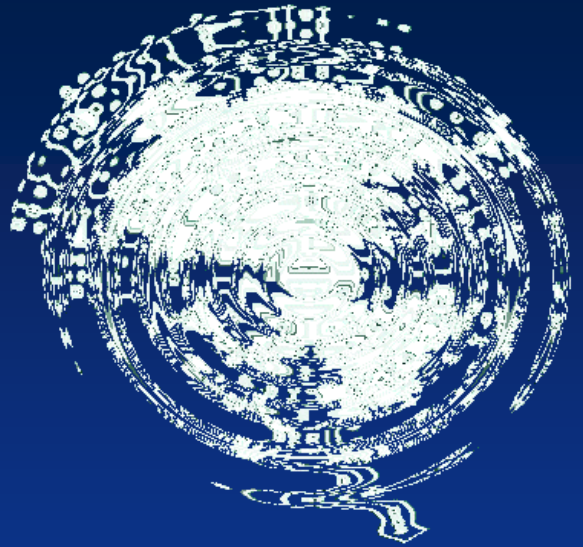


Persistent memory has generated concerns about data security

Some systems prefer to encrypt in the CPU

NVDIMMs specifications adding on-DIMM encryption option

May be required for systems with DMA to the DRAM channel



Summary

Memory capacity demands exceeding DRAM roadmap

DRAM protocol limited to 16 Gb for DDR4, 32 Gb for DDR5

NVDIMM-N adds data persistence

NVDIMM-P allows media independent expansion

Software must deal with performance/feature asymmetry

Data encryption coming

Questions?

Bill Gervasi
Principal Systems Architect
bilge@Nantero.com