Expanding the World of Heterogenous Memory Hierarchies

The Evolving Non-Volatile Memory Story

Bill Gervasi
Principal Systems Architect
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 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

- **CPU**: Faster, lower latency
- **I/O**: Network
- **Memory Control**: ...
The Search for

THE HOLY GRAIL
When we no longer fear power failure...
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.
From vacuum tubes

To core memory

To DRAM

To NVRAM
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.
Not all “persistence” is created equal
“Write endurance” determines HOW persistent

Wear leveling needed if writes are limited
Temperature sensitivity impacts long term retention

<table>
<thead>
<tr>
<th>Application Class</th>
<th>Workload</th>
<th>Active Use (power on)</th>
<th>Retention Use (power off)</th>
<th>Functional Failure Rqmt (FFR)</th>
<th>UBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client</td>
<td>Client</td>
<td>40°C 8 hrs/day</td>
<td>30°C 1 year</td>
<td>≤3%</td>
<td>≤10^{-15}</td>
</tr>
<tr>
<td>Enterprise</td>
<td>Enterprise</td>
<td>55°C 24hrs/day</td>
<td>40°C 3 months</td>
<td>≤3%</td>
<td>≤10^{-16}</td>
</tr>
</tbody>
</table>

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
Memory Class Storage = NVRAM

For now...

In the future?
Storage Class Memory Is NOT a Memory Class Storage
Flash Storage
Magnetic RAM
Resistive RAM
3DXpoint
Phase Change
3D NOR
Storage Class Memory

≥ DRAM performance
= DRAM endurance
≥ DRAM capacity

DDR NVRAM

Memory Class Storage

Wasteland
SSD
NVMe
Hard Disk

Storage

DDR DRAM

= DRAM performance
= DRAM endurance
≥ DRAM capacity
Deterministic

Non-Deterministic

Deterministic

Non-Deterministic

Deterministic

Non-Deterministic

Deterministic
DRAM

Optane

NVDIMM-N

NVRAM Memory Class Storage

NVDIMM-P
Itty bitty leaky capacitors lose charge
On power fail, you lose

Refresh time consumes up to 15% of bandwidth
Run

FAIL!

DRAM
**NVDIMM-N**

Use DRAM normally

On Power Fail, copy to Flash

Power restored, copy to DRAM
NVDIMM-N

Run

FAIL!

Switch to Battery Power

Copy DRAM to Flash

RESTORE

Copy Flash to DRAM

Run

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

One power fail cycle pays for a LOT of protection

NVDIMM-N

Copy DRAM to Flash

1-2 MINUTES

Copy Flash to DRAM

1-2 MINUTES
Optane

Reads are slow

Faster than Flash!!!

But vs DRAM? Meh

Decent capacity, though

 Writes are deathly slow

Could be used as a very slow DRAM but more common as expansion
3DXpoint Array

NVM Control

CPU

Host System

3DXpoint Array

NVM Control

DRAM as Cache

CPU

Host System

Optane

App Direct

Memory Mode

512GB = 512GB

512GB + 64GB = 512GB
New non-deterministic protocol

Not backward compatible with DDR

Requires NVDIMM-P aware CPU
NVDIMM-P Persistence Options

- Volatile Mode No Persistence
- Battery Backup ala NVDIMM-N
- 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
NRAM™

ReRAM *

PCM *

MRAM *

* Future generation devices
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
NRAM™ ≠ PCM
ReRAM ≠ MRAM

Timings
Precharge requirement
Persistence definition

DDR5 NVRAM Specification brings coherence
Refresh command is not needed
Decoded as NOP for compatibility
DRAM

- IDLE
- SELF REFRESH
- REFRESH
- FREQUENCY CHANGE

Power burned

NVRAM

- IDLE
- SELF REFRESH
- FREQUENCY CHANGE

“No” power burned
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
Data is persistent

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
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
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
System designers have a lot of options to balance Performance and Persistence.
Homogenous Main Memory

- DRAM
- MCS
- NVDIMM-N
- NVDIMM-P
- Optane
Heterogenous Main Memory

DRAM + Optane

MCS + NVDIMM-P

MCS + Optane
When capacity meets persistence

- NVDIMM-N
- DRAM
- NVRAM
- Memory Class Storage

- NVDIMM-P
- Optane
- 32GB
- 64GB
- 512GB
<table>
<thead>
<tr>
<th></th>
<th>Data Safe</th>
<th>Performance</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>No</td>
<td>Best</td>
<td>1.0 X</td>
</tr>
<tr>
<td>NVDIMM-N</td>
<td>Yes</td>
<td>Best</td>
<td>0.5 X</td>
</tr>
<tr>
<td>Optane</td>
<td>Yes</td>
<td>Worst</td>
<td>10 X</td>
</tr>
<tr>
<td>NVDIMM-P</td>
<td>Yes</td>
<td>Mid</td>
<td>10 X</td>
</tr>
<tr>
<td>MCS</td>
<td>Yes</td>
<td>Best+</td>
<td>1 X+</td>
</tr>
<tr>
<td>Heterogeneous Main Memory Combinations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Safe</strong></td>
<td><strong>Performance</strong></td>
<td><strong>Capacity</strong></td>
<td></td>
</tr>
<tr>
<td>DRAM + Optane</td>
<td>No</td>
<td>High</td>
<td>6 X</td>
</tr>
<tr>
<td>DRAM + NVDIMM-P</td>
<td>No</td>
<td>High</td>
<td>6 X</td>
</tr>
<tr>
<td>MCS + Optane</td>
<td>Yes</td>
<td>High</td>
<td>6 X</td>
</tr>
<tr>
<td>MCS + NVDIMM-P</td>
<td>Yes</td>
<td>High</td>
<td>6 X</td>
</tr>
</tbody>
</table>
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
System Power

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

Full presentation on Wednesday...
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
2,500 years ago

4,500 years ago

10,000 years ago

NRAM Data Retention = 12,000 Years
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
Carbon Nanotube Arrays

Row Decode
Column Decode
Bank Decode

SECDED ECC Engine

72 bits
64 bits

FIFO
FIFO

x4/x8

Data Strobe DataStrobe FIFO FIFO

Address

Chip ID

Die Selector

DDR4, DDR5 NRAM
 DDR4/DDR5

15-20%

DDR4/DDR5 NRAM

Base throughput
Elimination of refresh
Elimination of tFAW restrictions
Elimination of bank group restrictions
Elimination of power states
Elimination of inter-die delays

Bandwidth: larger is better

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

- Fully deterministic
- Non-deterministic

Persistence
Would you rather...

Step on broken glass?

A LEGO?

Or some jacks?
...about those energy stores...

Batteries

Supercapacitors

Tantalums (etc.)
Batteries
- High capacity
- High energy density
- Low reliability

Supercapacitors
- Medium capacity
- Low energy density
- Degrade over time

Tantalums (etc.)
- Low capacity
- Low energy density
- ...but stable
Flash or Storage Class Memory

Storage Controller

DRAM

Energy needed for backup of DRAM cache

Energy

I/O
Flash or Storage Class Memory

Storage Controller

NVRAM

More room for storage

Eliminate need for backup energy

I/O
NVRAM Changes the Math

DRAM cache limited by energy available

No DRAM? Cache size dictated by cost/performance

1GB/TB
Switching gears again...

...to Systems Evolution
Pop quiz

How many CPUs in a 1980s PC?
Graphics Adapter

Network Adapter

Modem

Sound Blaster
They were called “DSPs”
Digital Signal Processors

They put processing next to the data

They were killed by “Native Signal Processing”

Analog front end devices
With NSP...

Cost

Performance

Power

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
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
SRAM & Registers

The final frontier...
Continuing to look for ways to bring Memory Class Storage down under 1ns

It will happen

Voltage adjustment
Faster edge rates
Better error check
Shadow registers
Getting smarter
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.
So many potential breaches

Application opens data from previous application

Memory moved from one system to another

Spy devices on memory buses
Infection via hack

Infection via spy devices
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

Password: X2.Hd44**3#jj0%
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
Today’s Solutions Help
But We Can Do Better
Persistence Complications
Sharing Time
Data Distribution Challenges
Mix & Match Memories
DDR5 NVRAM Spec in Progress
Summary
Summary
Thank you for your time

Bill Gervasi
bilge@Nantero.com
I’m here to learn too

What do you deal with?