The NVRAM Standard, Bringing Coherence to the Crazy World of Persistent Memory

Bill Gervasi
Principal Systems Architect
Nantero, Inc.
Data Processing Challenges

Checkpointing

Memory Tiers

Persistence

Applications

A New Standard

Agenda
Data processing is great
Data processing is great

Until something goes wrong
...the volatile nature of DRAM
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

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
Run
DRAM

Checkpoint
Storage
Checkpoint

Run

FAIL!

RESTART

Run

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
Faster, lower latency
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?

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 24 hrs/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
Existing infrastructure

Memory Class Storage = NVRAM

For now...

In the future?

Memory Class Storage

NVRAM

New controllers & memories
Storage Class Memory

Is NOT a

Memory Class Storage
Flash
Storage
Phase Change
3DXpoint
Resistive RAM
Magnetic RAM
3D NOR

Memory Class

Storage
≥ DRAM performance
= DRAM endurance
≥ DRAM capacity

Storage Class

NVRAM
DDR
DRAM
Wasteland
SSD
NVMe
Hard Disk
DRAM speed

Non-volatility

Unlimited write endurance

Wide temperature range

Scalable beyond DRAM

Flexible fabrication

Low power

Low cost

NVRAM
Drop in replacement for DRAM

Fully Deterministic

Permanently persistent

Always available
NRAM™

DDR5

NVRAM

PCM *

ReRAM *

MRAM *

* Future generation devices
DDR5 NVRAM is “like a DRAM and...”
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
Common DDR5 NVRAM Feature Set

Profile 1 Features
Profile 2 Features
Profile 3 Features
Profile 4 Features

Differences captured in the SPD
Refresh command is not needed
Decoded as NOP for compatibility
Power burned

“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
Intrinsic Persistence

Data is persistent

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
### Bank Buffer Organization

<table>
<thead>
<tr>
<th>ROW</th>
<th>Bank Buffer 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW</td>
<td>Bank Buffer 31</td>
</tr>
</tbody>
</table>

- **DDR5 SDRAM**
- **NVRAM**

"ROW" includes bank group & bank...
Row Extension Example
Row Replacement Example
NRAM: Carbon Nanotube Cell Memory
Crosspoint tiles translated to DDRx
Full DRAM speed: DDR4, DDR5

Non-volatility: 12,000 year data retention

Unlimited write endurance

Wide temperature range: -55°C to +300°C tested

Beyond DRAM: 512Gb/die in DDR5 window

Flexible fabrication: Logic or memory

Low power: 15% lower power than DRAM

Low cost: cheaper to build than DRAM
Evolution?

Or Revolution?
It’s About the Software

Old paradigm

New paradigm
DAX Mode
Phase 1: √ to MCS

Phase 2: No √

Memory Class Storage

Run

Checkpoint

Storage

Memory Class Storage

Run

Checkpoint

Storage

Memory Class Storage

Run

Checkpoint

Storage

Memory Class Storage

Run

Checkpoint

Storage
I/O

Storage-free Systems Enabled

CPU

Memory Control

Network

Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage

... Memory Class Storage
(Psst...)  (I’m going with revolution...)
DATA PERSISTENCE

- Registers
- Cache
- Memory
- Storage Class Memory
- SSD
- NVMe
- Tape
- Hard Drive
Memory Class Storage

CPU

Flash

NVMe

Memory Class Storage

CXL

Memory Class Storage

AI

Deep Learning

Memory Class Storage
Power failure drives systems architectures

Checkpointing is a costly way to deal with system failure

Memory tiers balance safety and performance

Persistence is moving closer to the CPU

Applications evolving to exploit MCS

DDR5 NVRAM standard brings coherence

Summary
Thank you for your time

Bill Gervasi
bilge@nantero.com