MPS

You Can’t Fix

What You Can’t Measure

Bill Gervasi, Principal Memory Solutions Architect

Monolithic Power Systems

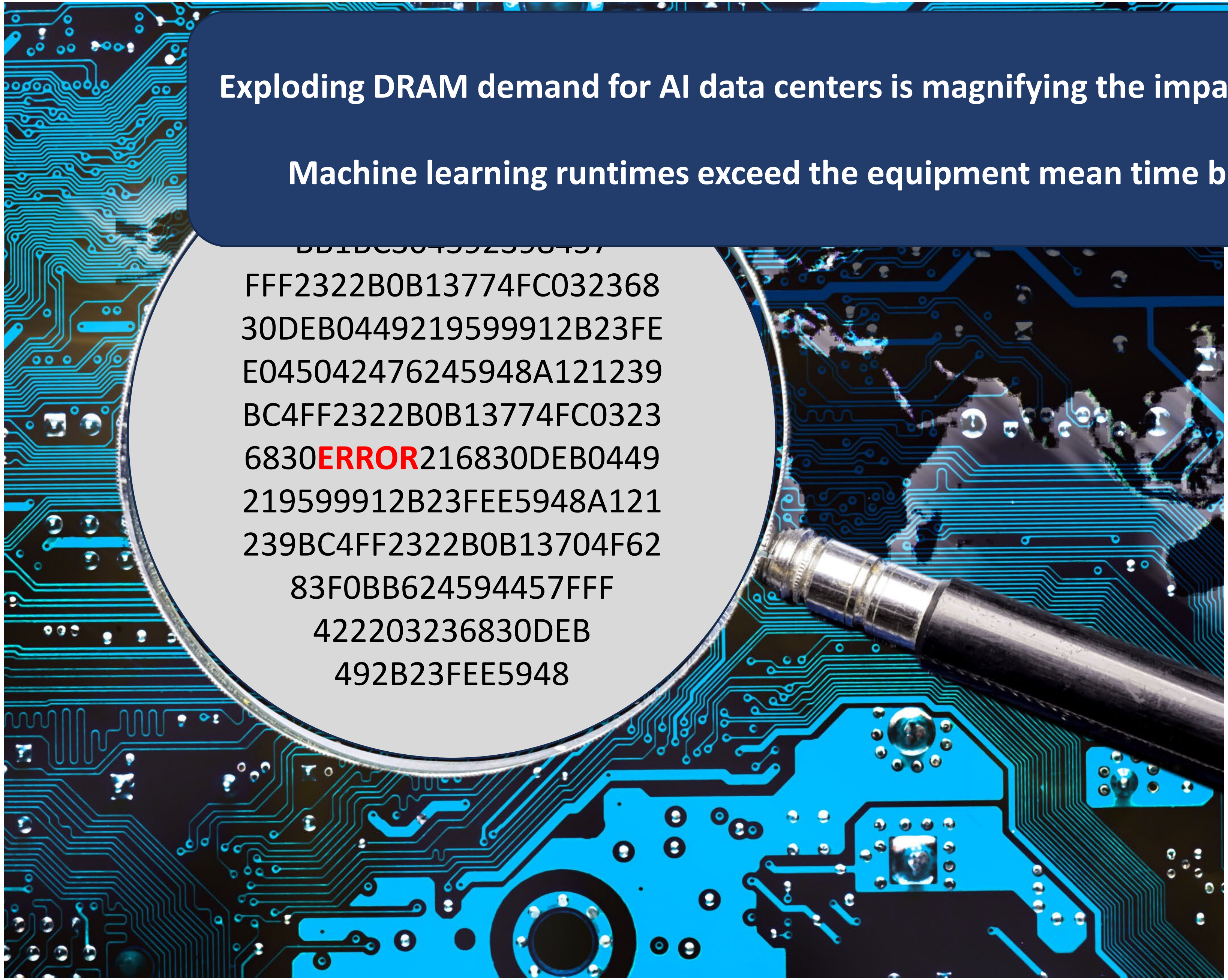
bill.gervasi@monolithicpower.com

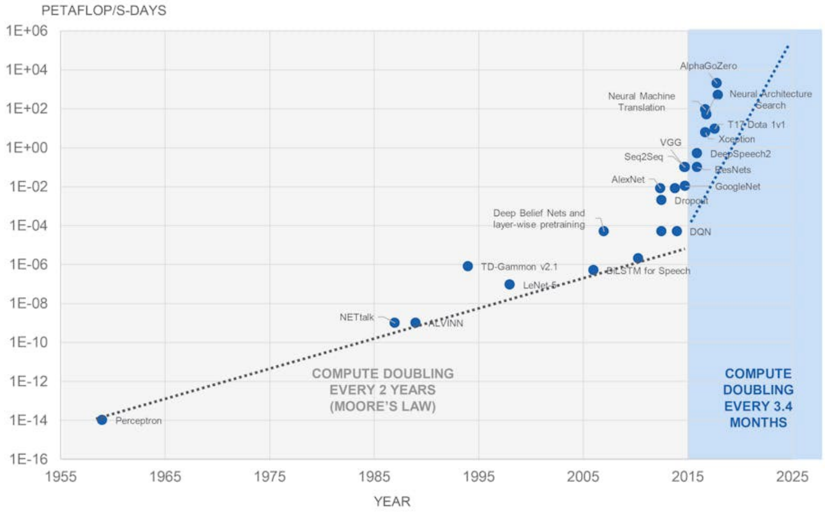
FMS the Future of Memory and Storage

(

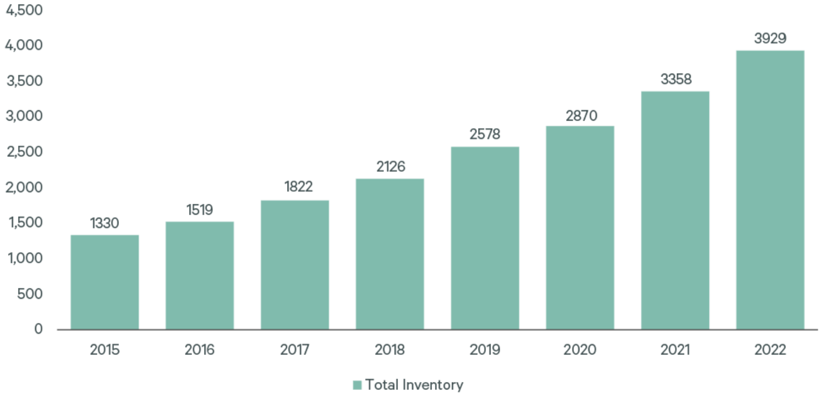
Exploding DRAM demand for AI data centers is magnifying the impact of memory errors

Machine learning runtimes exceed the equipment mean time between failures





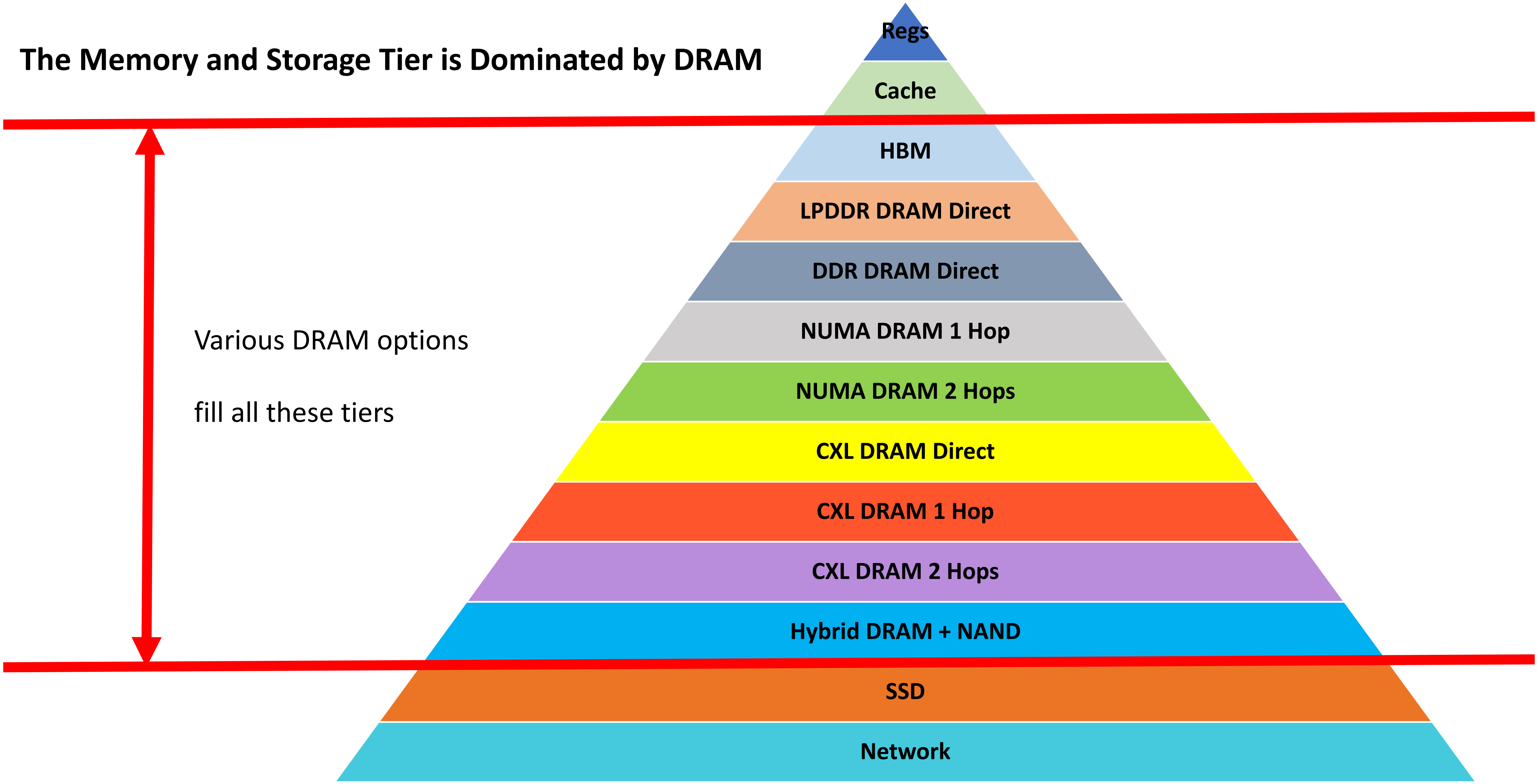
Source: OpenAI; publicly available on Openai.com: AI and Compute Research Report



The Memory and Storage Tier is Dominated by DRAM

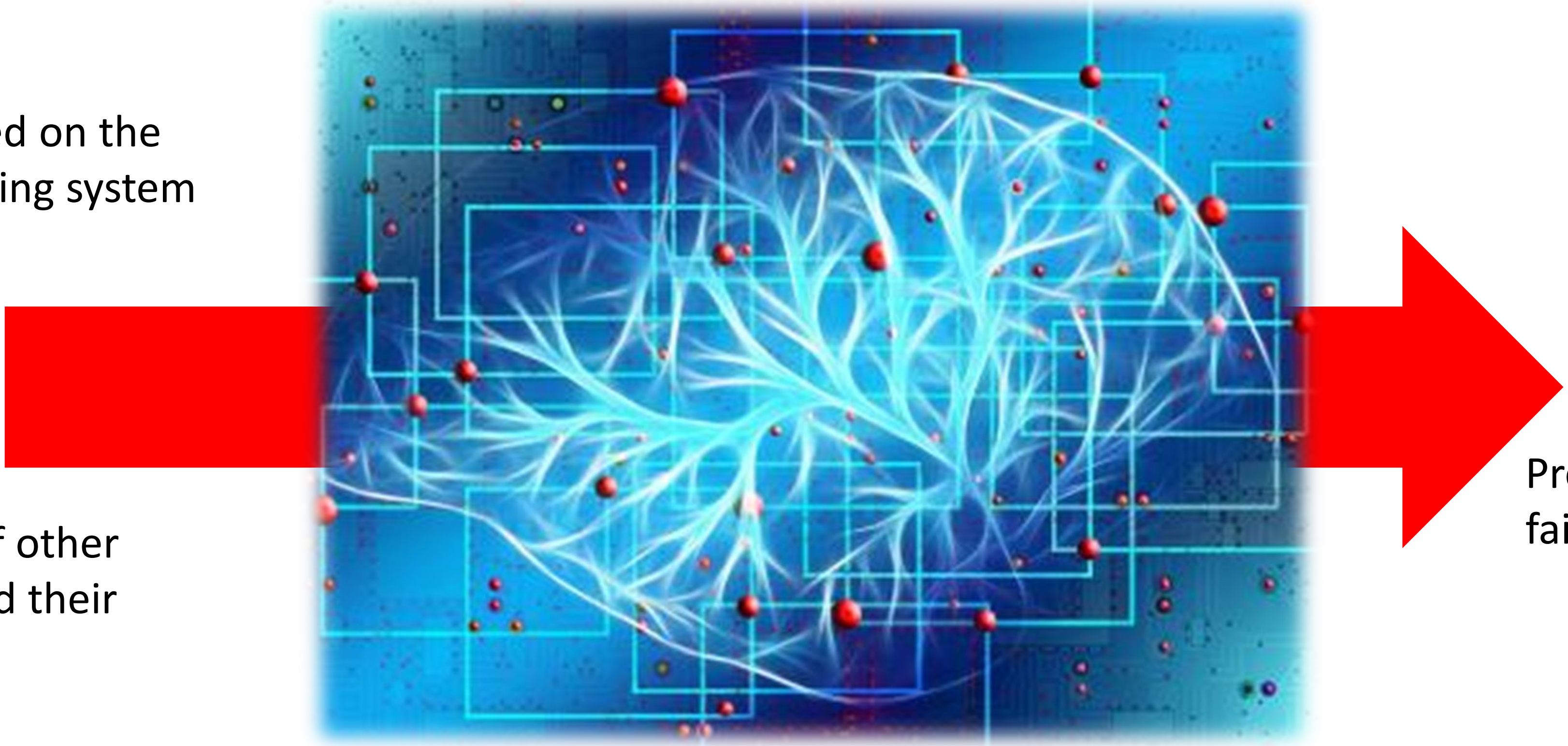
Various DRAM options

fill all these tiers



Truism That Preventing an Error is Cheaper Than Fixing It

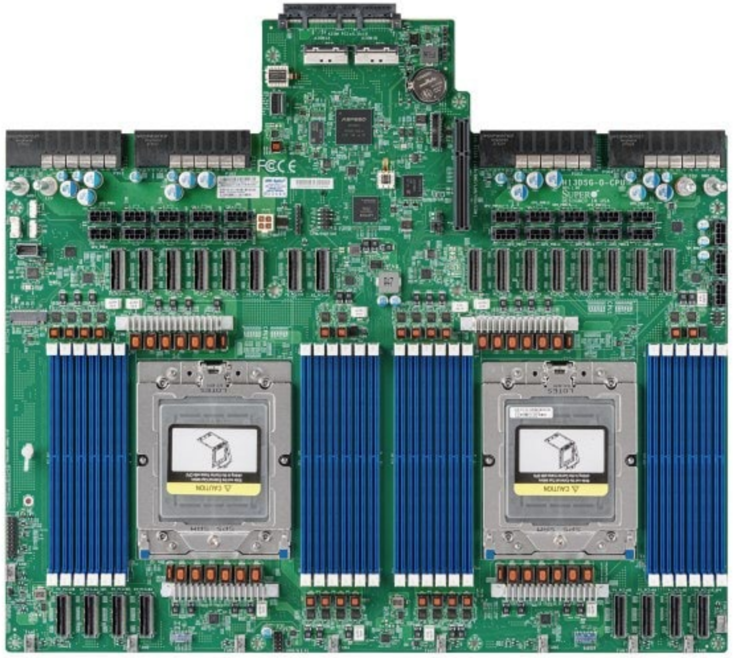
Data collected on the current running system

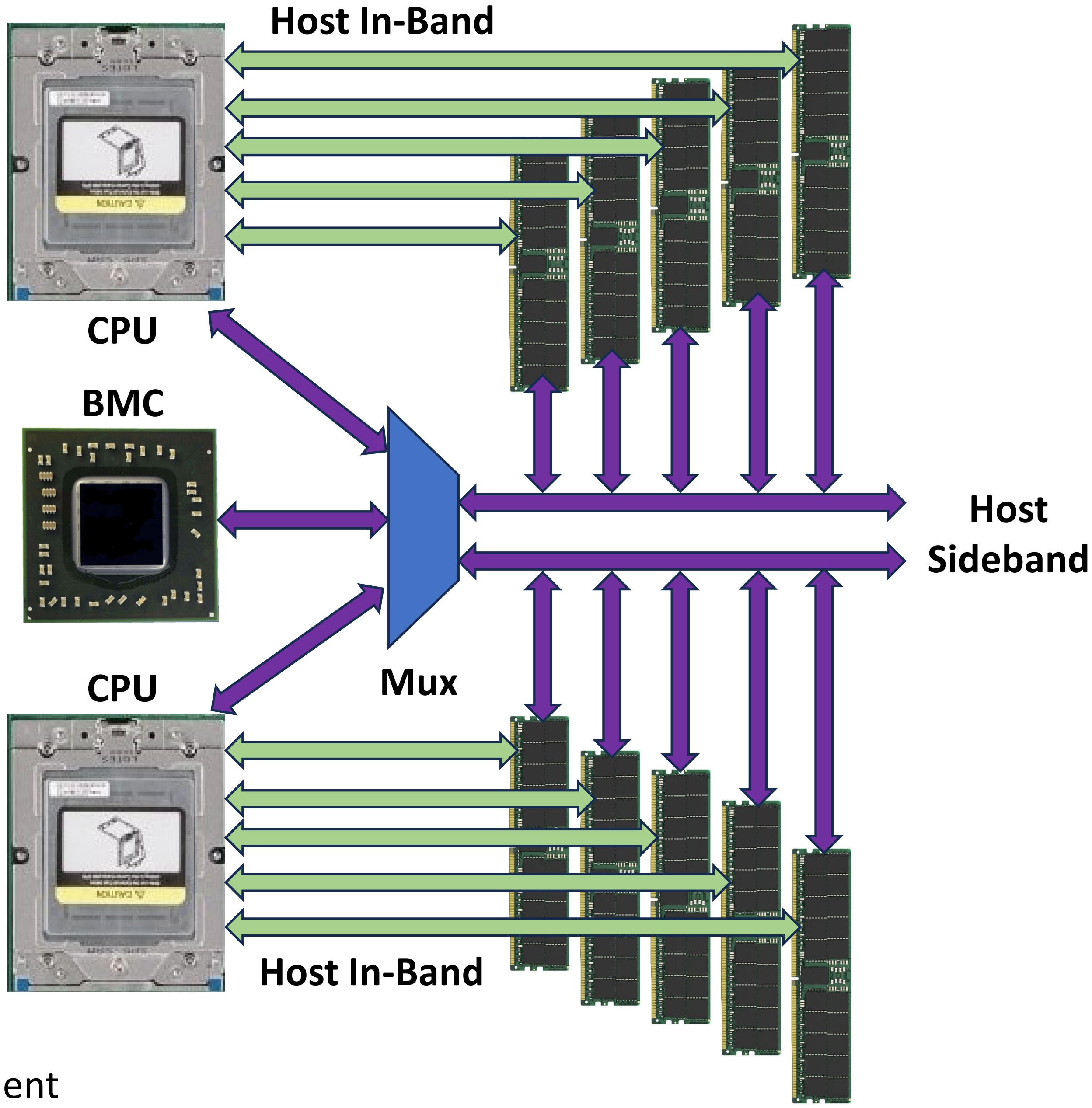


AI model of other systems and their failures

Predictive decisions about failure probability

This talk will focus on telemetry gathering for today's DRAM modules





Host In-Band allows interrogating DRAMs directly – DRAM access must be halted

Host Sideband allows interrogating module support logic – DRAM access is not interrupted

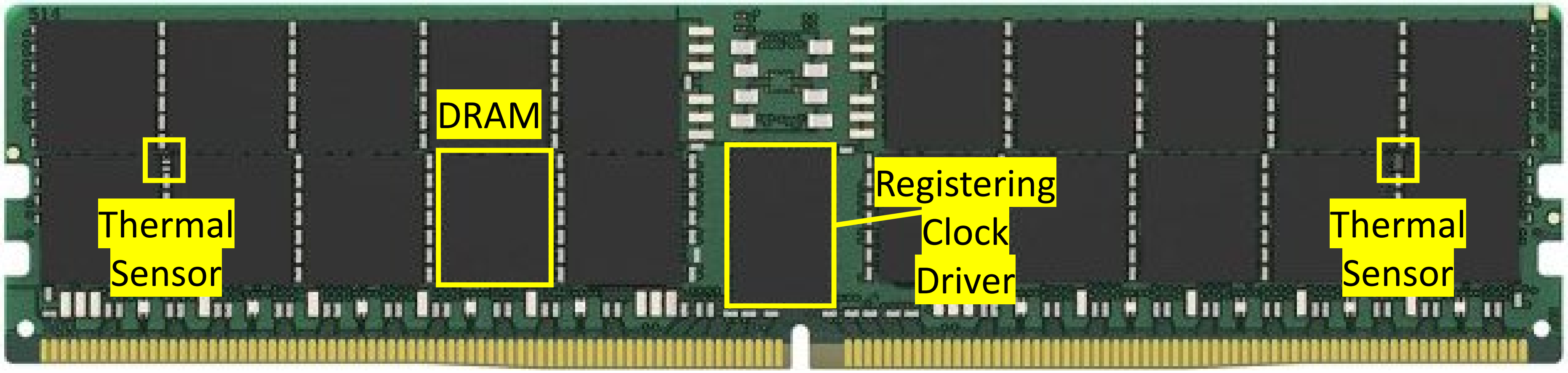
Mux allows

- CPUs and BMC to share sideband
- Supporting multiples of 8 modules per bus segment

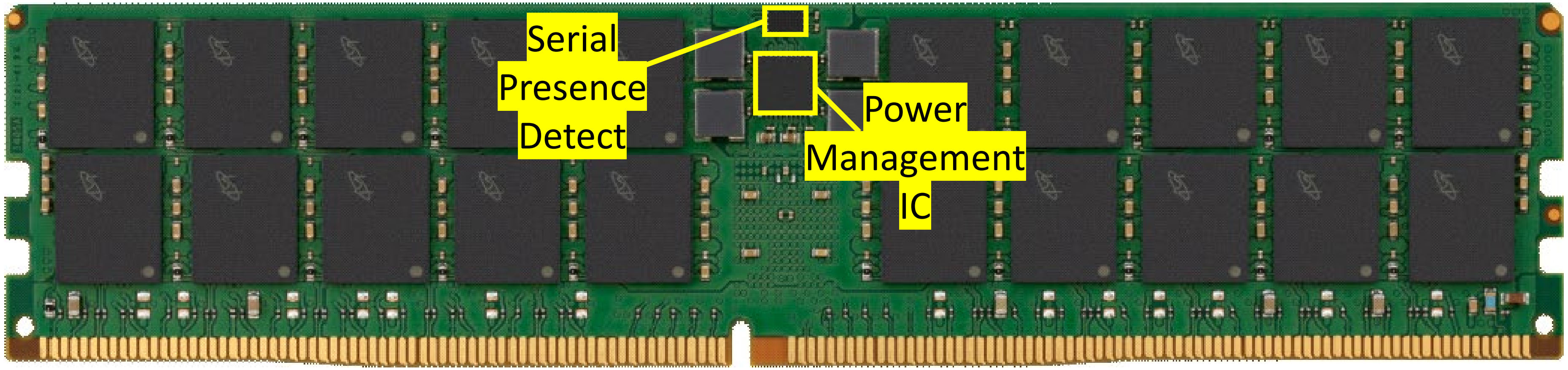
BMC: Baseboard Management Controller

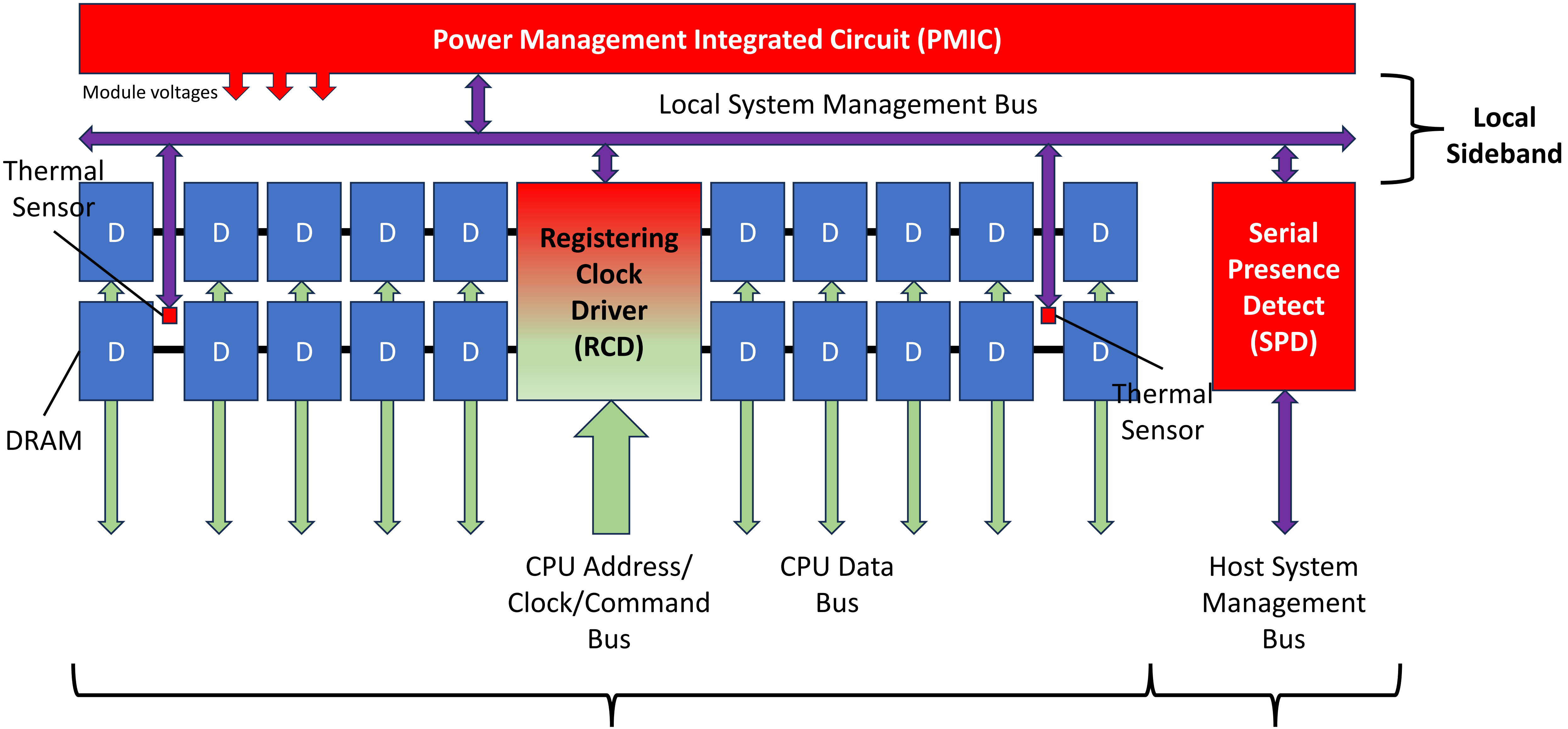
Anatomy of a Memory Module

Front



Back





Power Management Integrated Circuit (PMIC)

Module voltages

Local System Management Bus

Local Sideband

Thermal Sensor

D

Registering Clock Driver (RCD)

Serial Presence Detect (SPD)

DRAM

Thermal Sensor

CPU Address/ Clock/Command Bus

CPU Data Bus

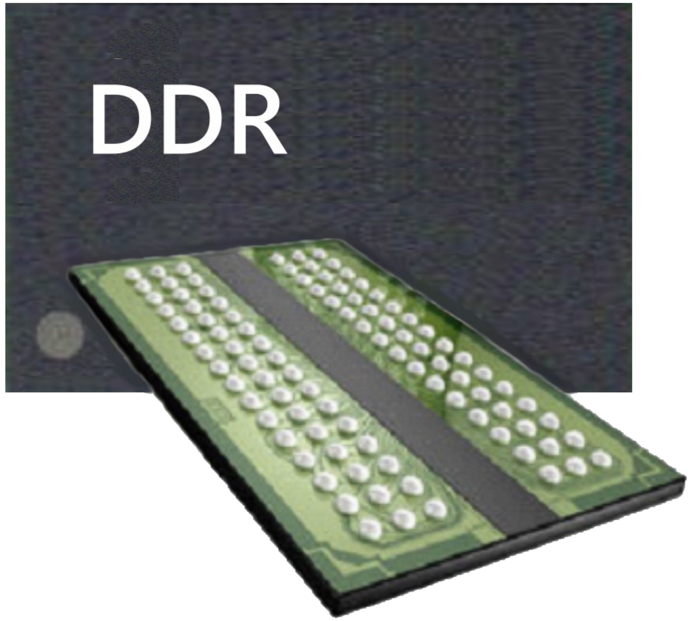
Host System Management Bus

🟦 **Host In-Band Devices**

Host Sideband Devices 🟥

the Future of Memory and Storage

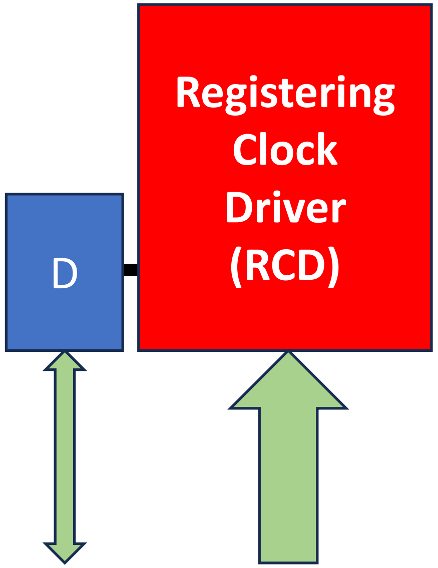
Reliable DRAM operation starts with effective calibration



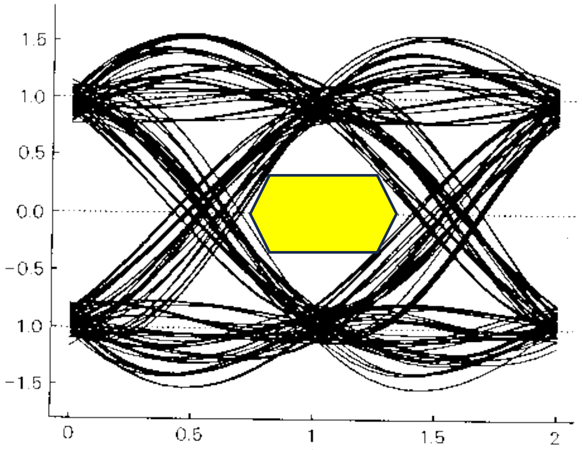
Each signal type is calibrated independently

- Data & strobes
- Address/commands
- Clocks
- Chip selects

Settings in dozens of mode registers

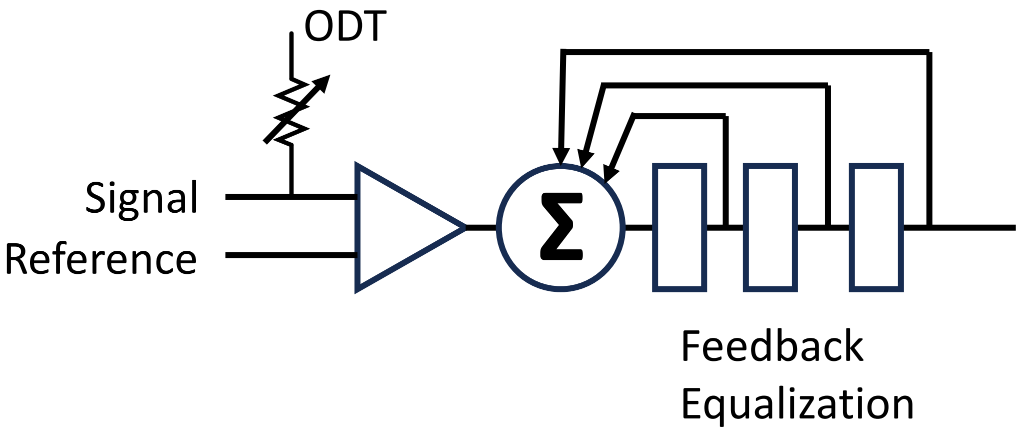


Settings done in-band

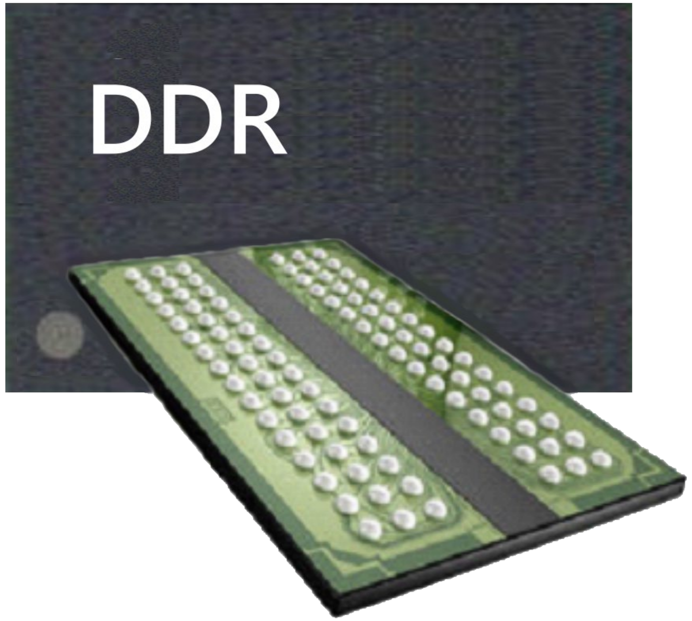


Input capture eye maximized by

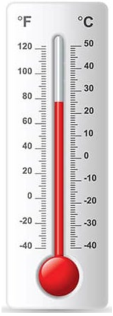
- Schmooing
- Test patterns
- On-die termination
- Voltage margining



When good memories go bad



Each DRAM provides hints about internal status



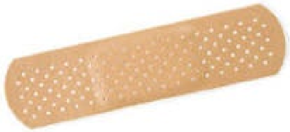
Internal temperature monitor

Suggested refresh rate

1X up to 85˚

2X from 85˚ to 95˚

On-Die Error Correction Code (ECC) is supported



Reads correct single bit errors before sending to host

Writes save ECC codes along with data

Runtime error transparency

Maximum number of errors

Worst 3 rows error count

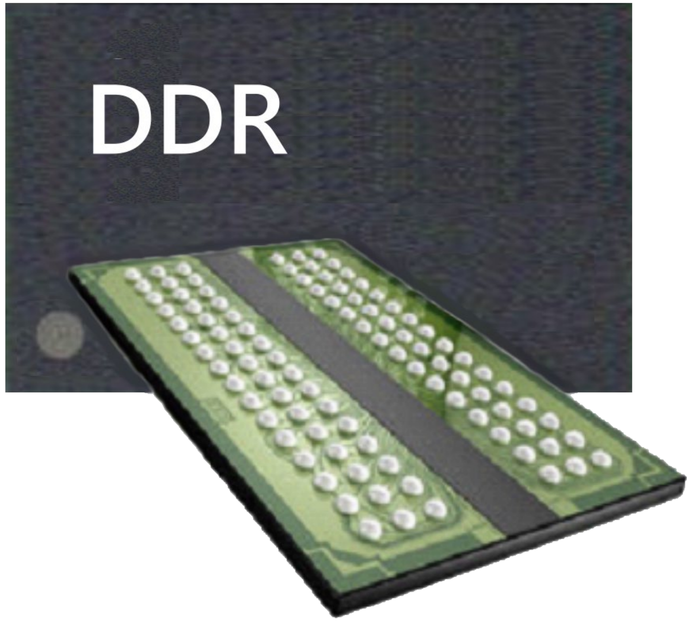
Error check scrub (ECS) error repair

Reads data, corrects errors, writes it back

Total error counter

Rows with errors counter

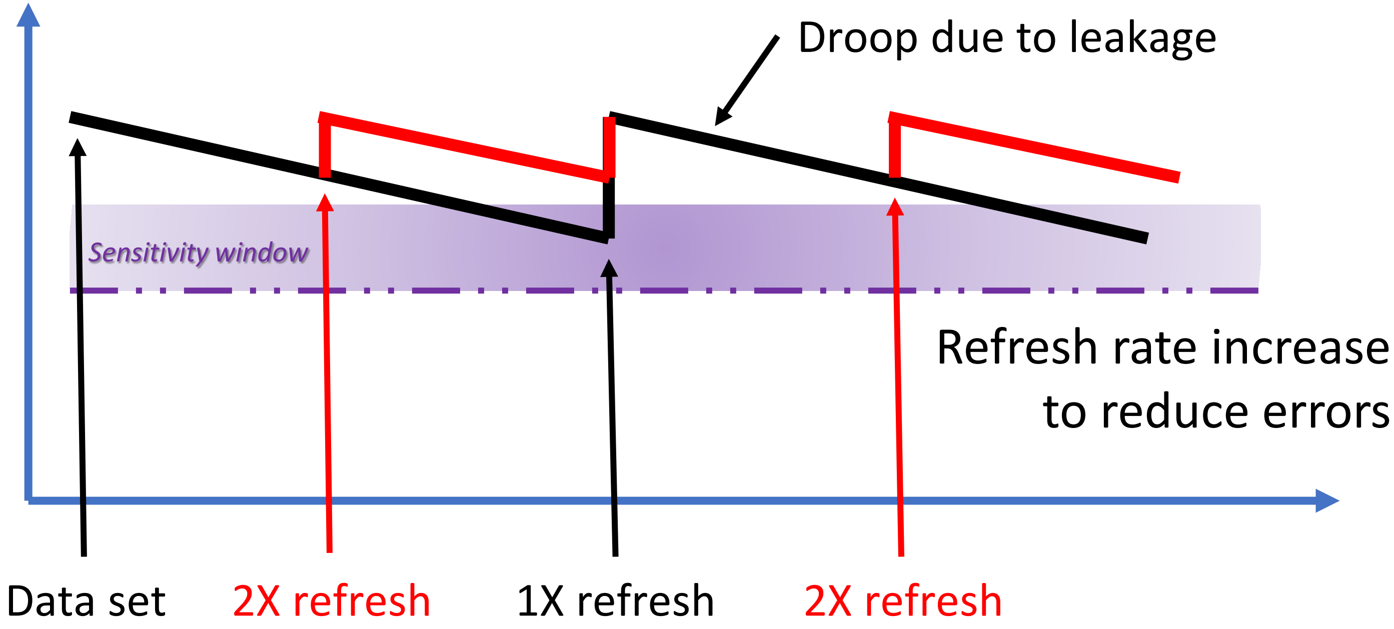
Dealing with errors



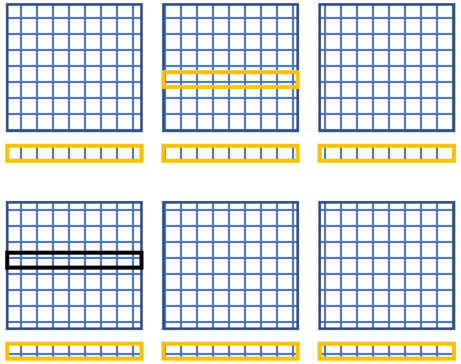
When an error occurs, the DRAM drives an ALERT signal to the host

- CRC error
- Excessive activation threshold exceeded

Losing data? Increase the refresh rate

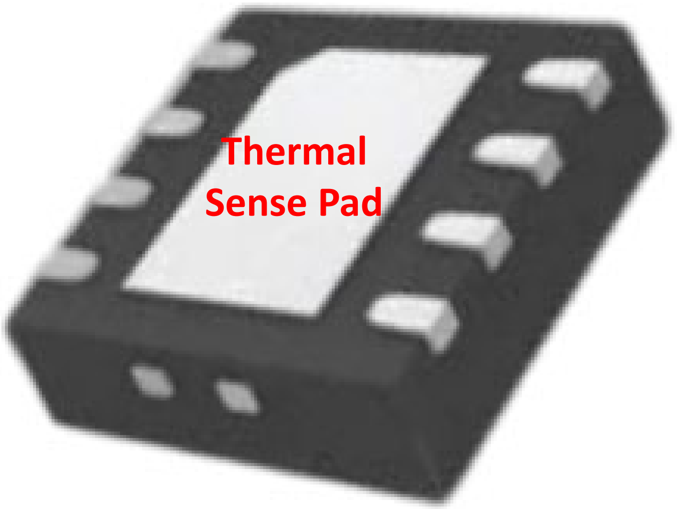


Refresh rate increase to reduce errors



Post-package repair of naughty rows

- If a row is misbehaving, it can be swapped out
- Combine mPPR with Memory Built-In-Self Test (MBIST)



Serial Presence Detect (SPD)

Powered via LDO independent of other module circuits

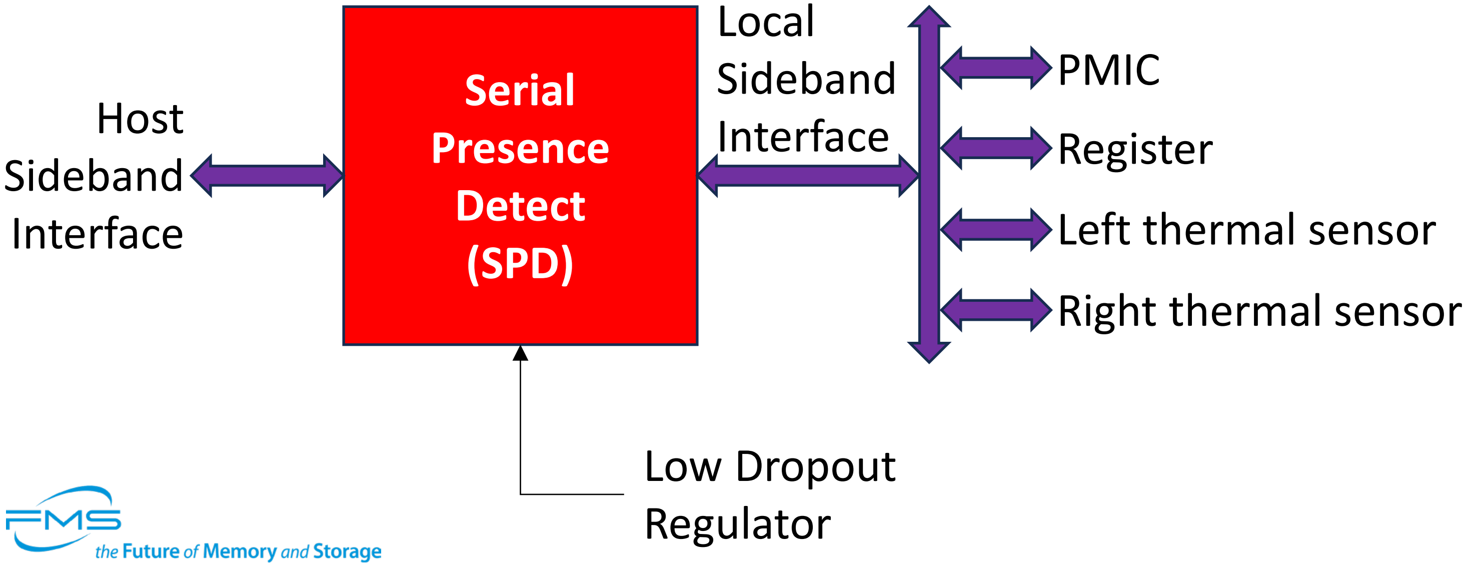
Host communication via sideband bus

1 KB of non-volatile memory contains module parameters

Operates as a Hub from the host to local bus devices

Integrated thermal sensor with low, high, and critical settings

Host interrupt capability for itself and all local bus devices



Host Sideband Interface

Serial Presence Detect (SPD)

Local Sideband Interface

PMIC

Register

Left thermal sensor

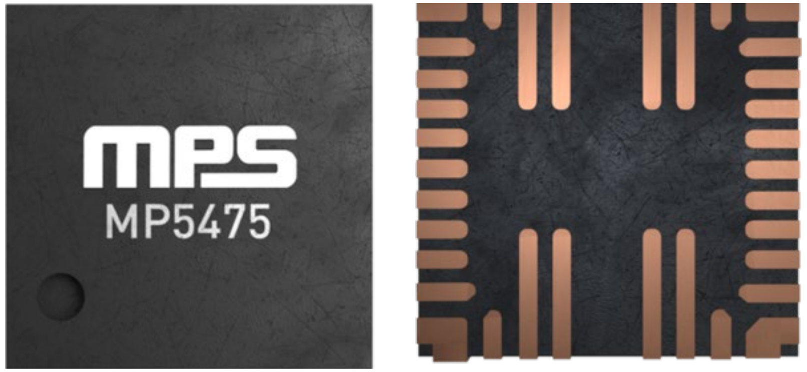
Right thermal sensor

Low Dropout Regulator

E.G., PMIC sees high temperature

1. Sends interrupt to the SPD
2. SPD interrupts the Host
3. Host interrogates PMIC

Power Management ICs (PMICs)



PMICs contribute to calibration and telemetry gathering
Communication over sideband interface

Calibration:

Each voltage rail can be adjusted based on device and signaling limits*
These may compensate for corner conditions as power planes cover large distances

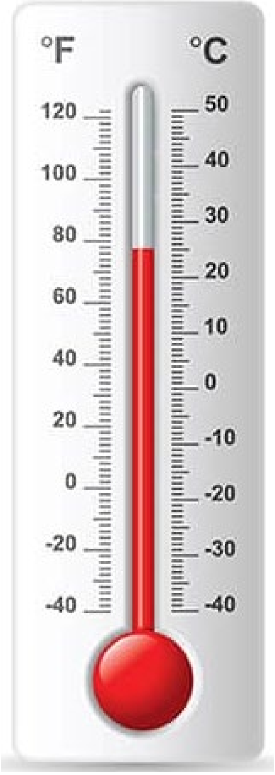
Telemetry:

Each voltage rail can be interrogated to measure voltage and current
Total module wattage may be calculated on the fly
Coupled with operational test patterns, power per operation type may be calculated

Warnings and Error reporting:

Each voltage rail can be configured with warnings at high and critical levels
Error counts are kept in non-volatile memory
Overvoltage and overcurrent treated separately

* Overclockers do this as standard procedure



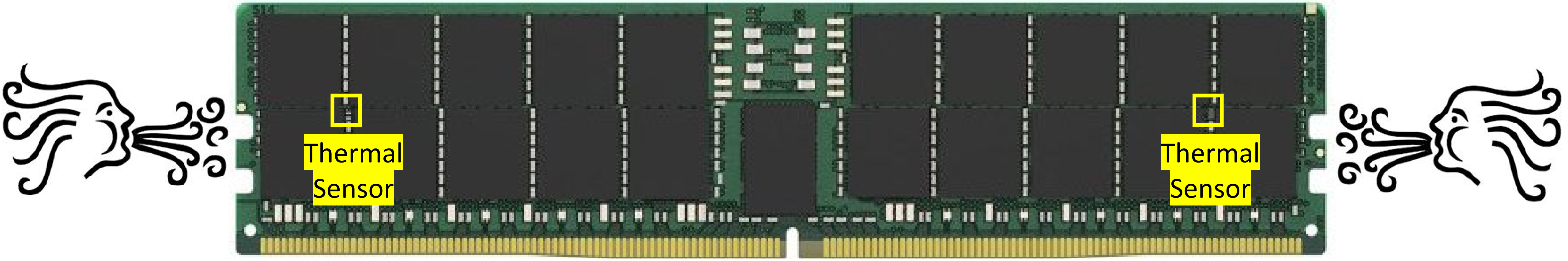
Thermal Sensors

Communication over sideband interface

Thermal sensors tie directly to the module ground plane

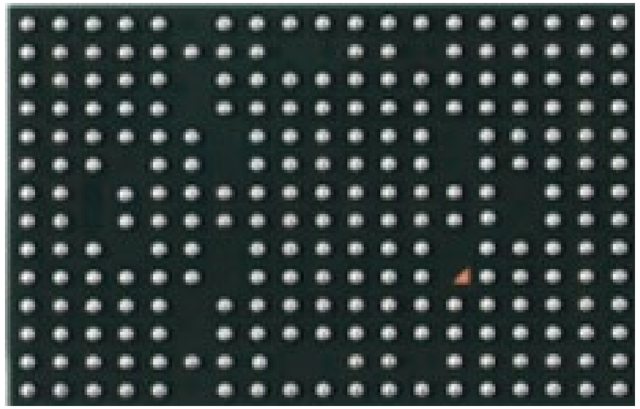
Direct thermal path from the DRAMs

Positioned at both ends of the module since airflow direction is 50/50



Registering Clock Driver (RCD)

RCDs contribute to calibration and telemetry gathering
Communication over in-band and sideband interfaces



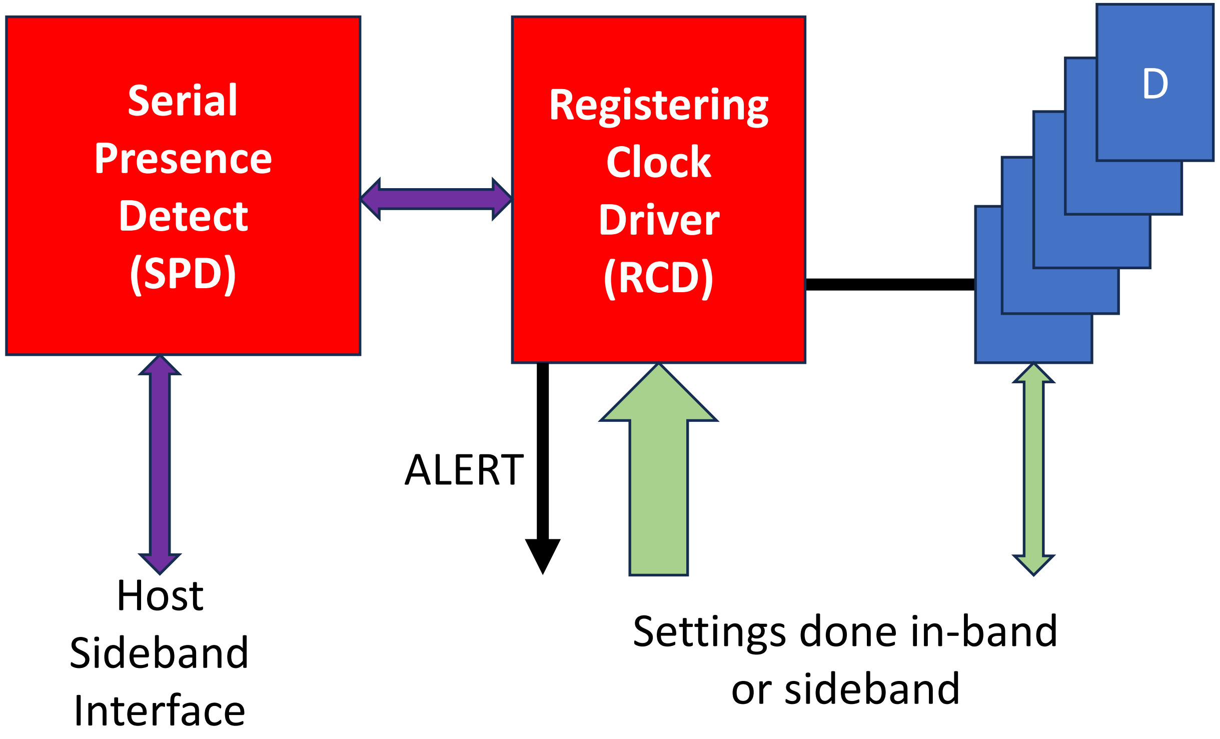
RCD Event Handling

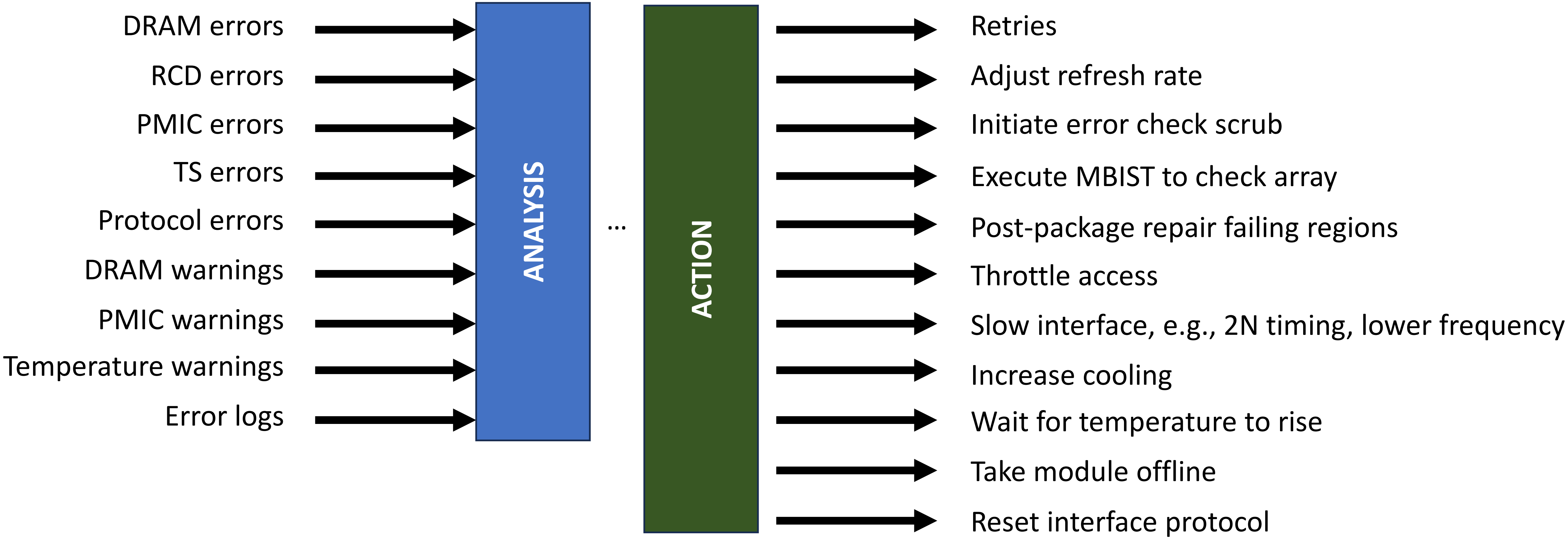
Event	DERROR_IN	ALERT	IBI	NACK
DRAM data CRC error	√	√		
DRAM PRAC alert back-off	√	√		
DRAM activation count violation	√	√		
DRAM ALERT verification	√	√		
RCD address parity violation		√		
RCD DCS training		√		
RCD DCA training		√		
RCD DFE training		√		
RCD DES training		√		
SidebandBus PEC error			√	√
SidebandBus Parity error			√	√
Handling	INBAND		SIDEBAND	

Interface between RCD and DRAMs calibrated

Detects and reports
DRAM errors
Register errors

On-chip error log reports what it saw on the address bus


Serial Presence Detect (SPD)
Registering Clock Driver (RCD)
D
ALERT
Host Sideband Interface
Settings done in-band or sideband



Each error or warning type has its own set of reactions and mitigations

At the system level, a point of failure can have major impact…
should the fan be faster for all modules in a rack to deal with one warning?

Collecting data at the rack, cage, hall, and building level can improve:
Dealing with errors and warnings
Predicting failures before they occur

This sounds like a good problem for AI

“Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field”

Non-DRAM memory failures from the memory controller and memory channel cause the majority of errors

Newer DRAM cell fabrication technologies have substantially higher failure rates, increasing by 1.8 over the previous generation

Using lower density DIMMs and fewer cores per chip can reduce failure rates of a baseline server by up to 57.7%

https://ieeexplore.ieee.org/document/7266869/

“Improving Memory Reliability at Data Centers “

AI to create a model of predictive patterns by comparing thousands and thousands of memory error logs from the field, then compares this model with scans from an operator’s data center to determine where problems may exist to support data center operation and workload continuity

Predictive memory resilience technology can reduce uncorrectable error rates by nearly 50%

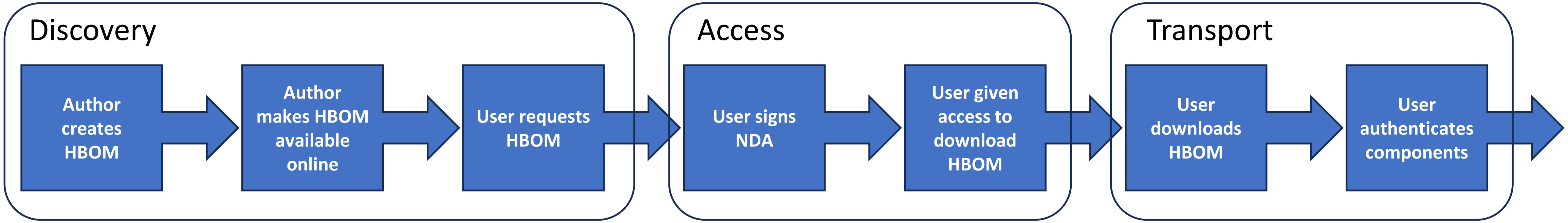
https://www.intel.cn/content/dam/www/public/us/en/documents/intel-and-samsung-mrt-improving-memory-reliability-at-data-centers.pdf

America's Cyber Defense Agency

NATIONAL COORDINATOR FOR CRITICAL INFRASTRUCTURE SECURITY AND RESILIENCE

CISA.gov

Electronically available hardware bill of materials is under deployment for security


Discovery
Author creates HBOM
Author makes HBOM available online
User requests HBOM
Access
User signs NDA
User given access to download HBOM
Transport
User downloads HBOM
User authenticates components

Once the HBOM is downloaded, error information can be added to the models and tracked to specific components

"Problem" parts can indicate potential for future failures

https://www.cisa.gov/resources-tools/resources/hardware-bill-materials-hbom-framework-supply-chain-risk-management
https://www.cisa.gov/sbom

Conclusions

Increasing demand for memory has exacerbated sensitivity to errors

Memory subsystems provide a collection of reports

- Errors detected
- Warnings

Using this data to mitigate an error is useful, but…

…Predicting the next error before it happens is essential

Thank you for your time

Any questions?

Bill Gervasi, Principal Memory Solutions Architect

Monolithic Power Systems

bill.gervasi@monolithicpower.com