OPEN POSSIBILITIES.

Optimizing Performance and Reliability of Datacenter Infrastructure with Alternative Memory Failure Protection Solutions

NOVEMBER 9-10, 2021

SI (Strategic Initiatives)

Optimizing Performance and Reliability of Datacenter Infrastructure with Alternative Memory Failure Protection Solutions

Zachary Bobroff, Sr Director of Product Management, AMI

OPEN PLATINUM™

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

Memory Failures

Memory failures are not new, why are we discussing them now?

- Memory failures continue to be one of the most costly causes of server downtime
- Memory failure rate has increased with every generation of DDR due to higher density and speed increases
- Current methodologies for memory error classification have reached their limits for effectively identifying a failing DIMM
- Can new methods improve system uptime and actually predict a DIMM failure before it occurs?

What are memory errors?

SERVER

DIMM Faults

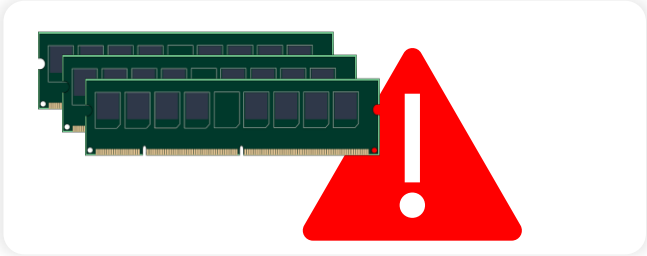
- The unobservable underlying causes of an “error”
  - Soft Faults: Particles, cosmic rays – restorable
  - Hard Faults: ware-out, manufacture defect – repeatable

DIMM Errors

- An observed symptom of a fault. Reported from FW or OS. (*e.g.*, MCELog, SEL log)
  - Correctable Errors (CE): Errors that can be corrected by ECC or chip-kill, etc.
  - Uncorrectable Errors (UE): Catastrophic failures, typically resulting in a crash

DIMM Failures

- The deviation from the expected behavior. Many errors can be caused by 1 failure
- Combined effects of DRAM wear level & implicit runtime context



Memory Failures Lead to Server Downtime

SERVER

- Data centers utilizes system hardware for running workloads for themselves and customers
- With digital infrastructure in such high demand, data centers must avoid downtime as much as possible
- Memory failures cause unexpected downtime that requires manual intervention for the installation of new DIMMs

OPEN POSSIBILITIES.

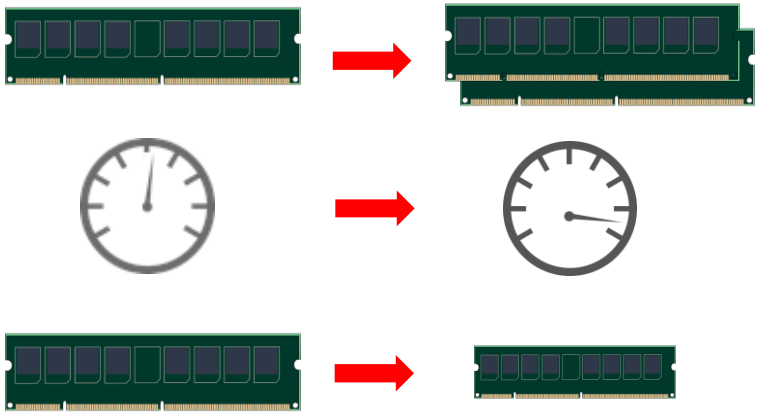
OCP GLOBAL SUMMIT 5

NOVEMBER 9-10, 2021

Memory Error Increase with New DDR Generations

With every generation of DDR:

- DRAM capacity increases
- DRAM clock speeds increase
- It is common for DRAM vendors to shrink the process technology

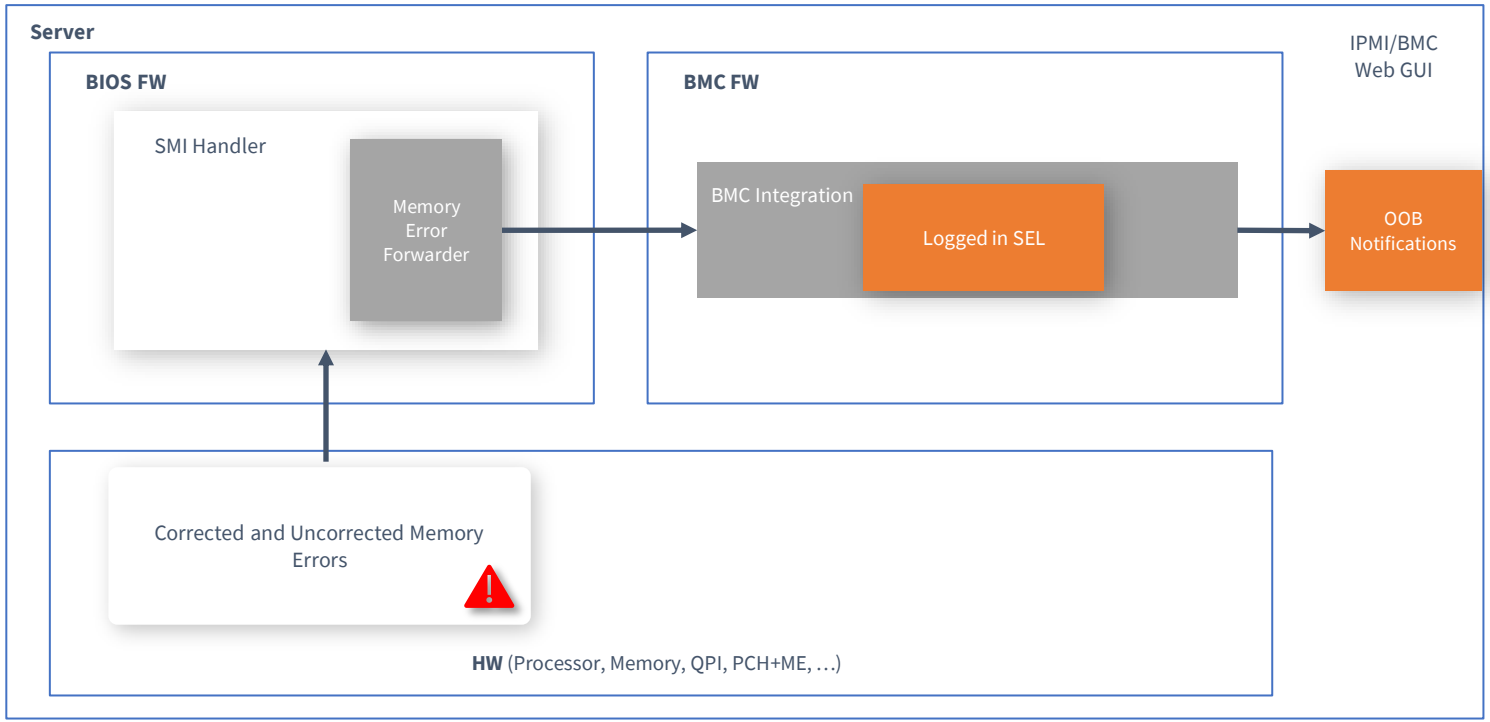


With higher speed, higher capacity and process shrink, single bit error likelihood increases

OPEN POSSIBILITIES.

Source: https://www.synopsys.com/designware-ip/technical-bulletin/error-correction-code-ddr.html

Traditional Memory Handling Process



Traditional Memory Error Handling Process

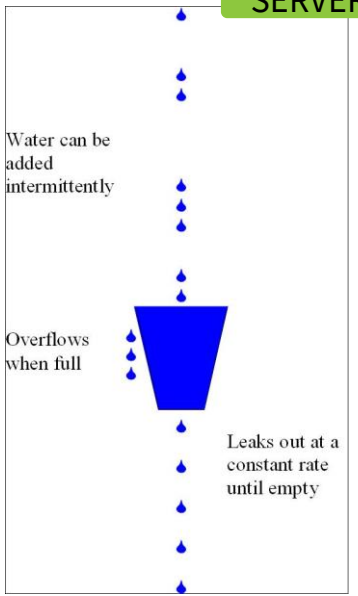
Expanded Memory Handling Techniques

SERVER

Commonly, leaky bucket methodologies are used to determine if a memory stick is failing

Leaky bucket mythology is rooted in a threshold being reached

Many complex algorithms are built on top of leaky buckets, but are still limited by reaching a threshold



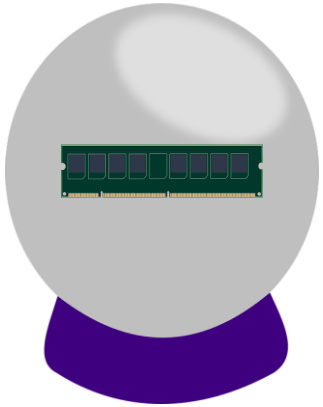
The Need for Memory Failure Prediction

SERVER

Using threshold methodology lets you know when a DIMM has gone bad, but is there a way to predict a memory failure before it has actually failed?

If a memory failure can be predicted:

- The system off-load work from that DIMM before failure
- The system can be scheduled for maintenance before it has system crashes



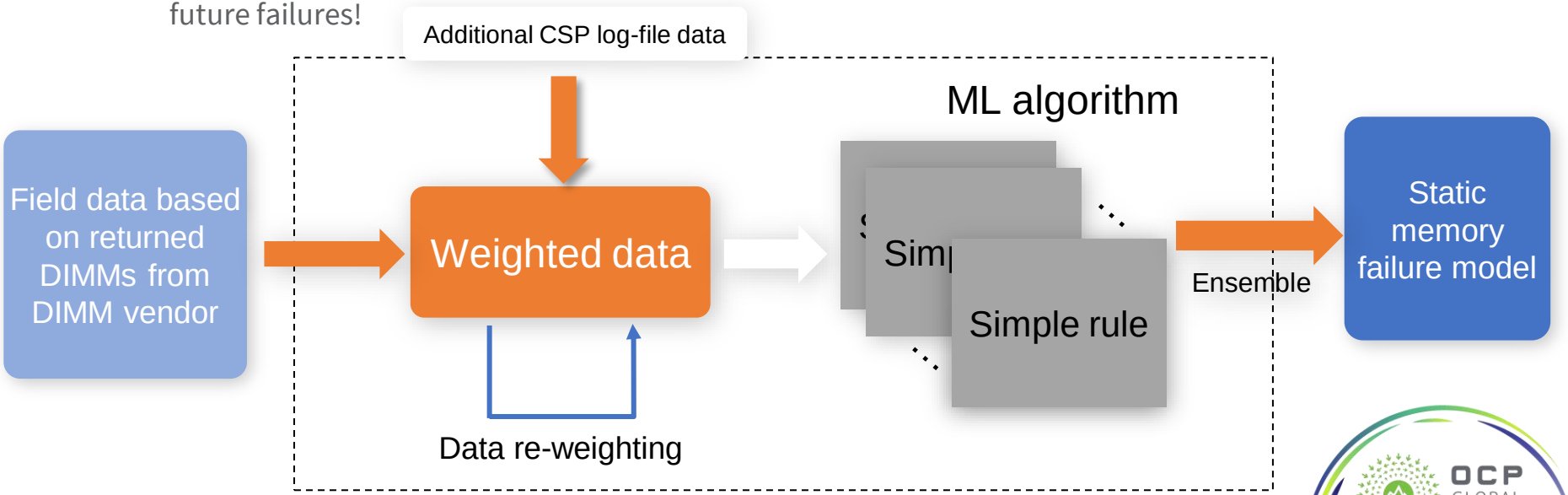
Using Modeled data, a score can be assigned each DIMM for overall health

Building a Memory Prediction Algorithm

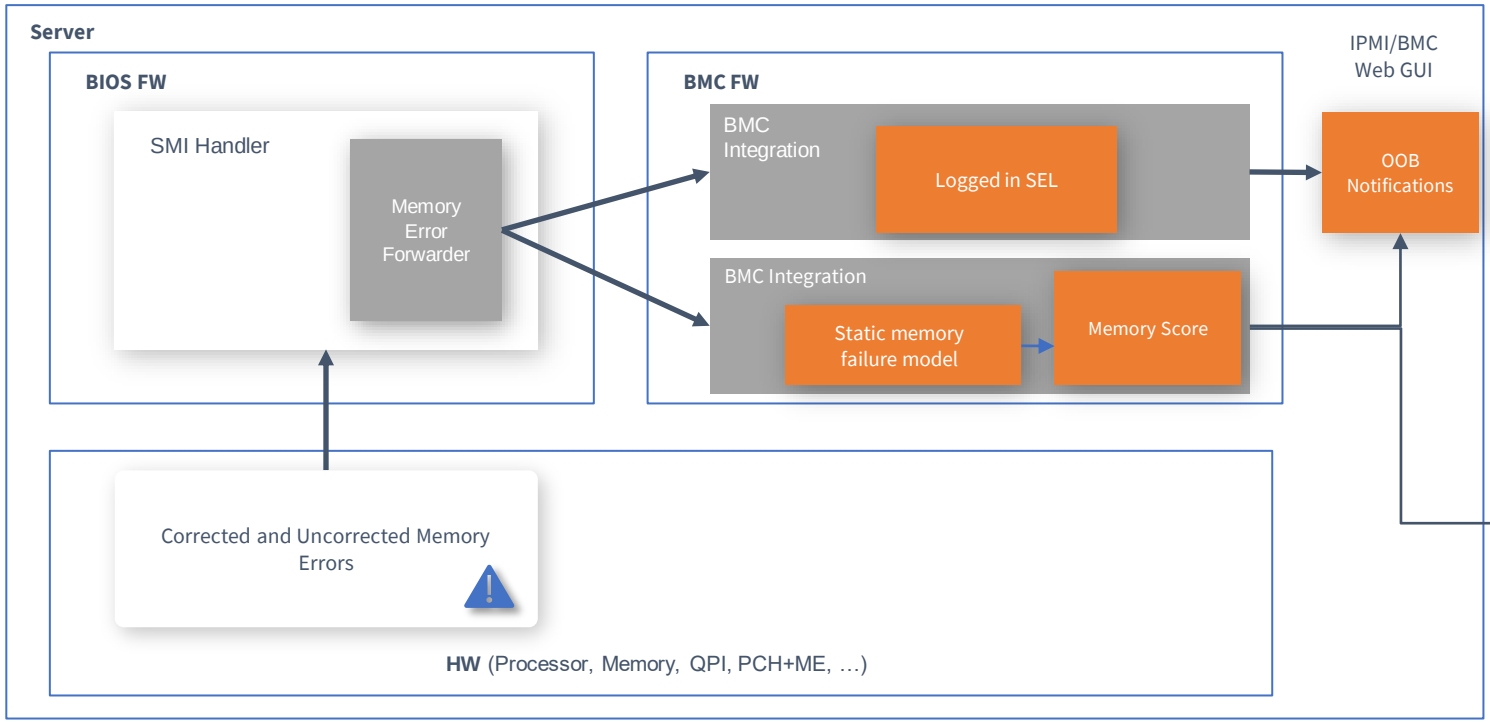
SERVER

Building a memory prediction algorithm is like building any ML model, you just need the data!

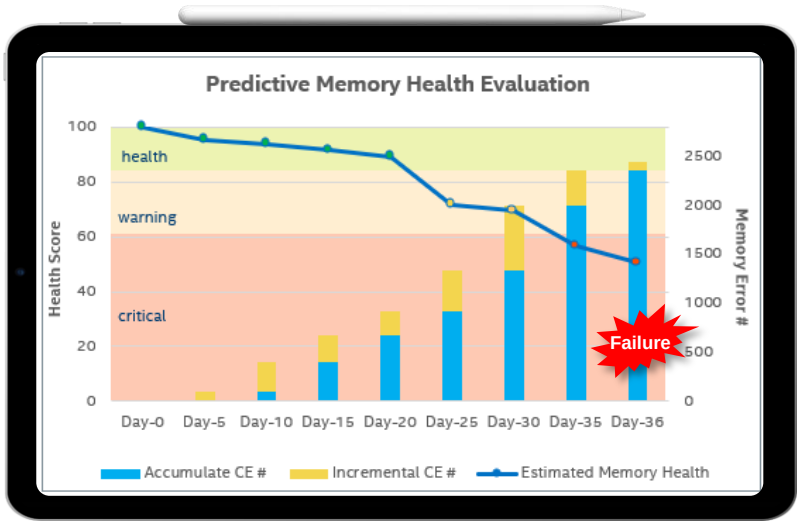
The model can simply look for known error patterns of failed DIMMs to discover known patterns of future failures!



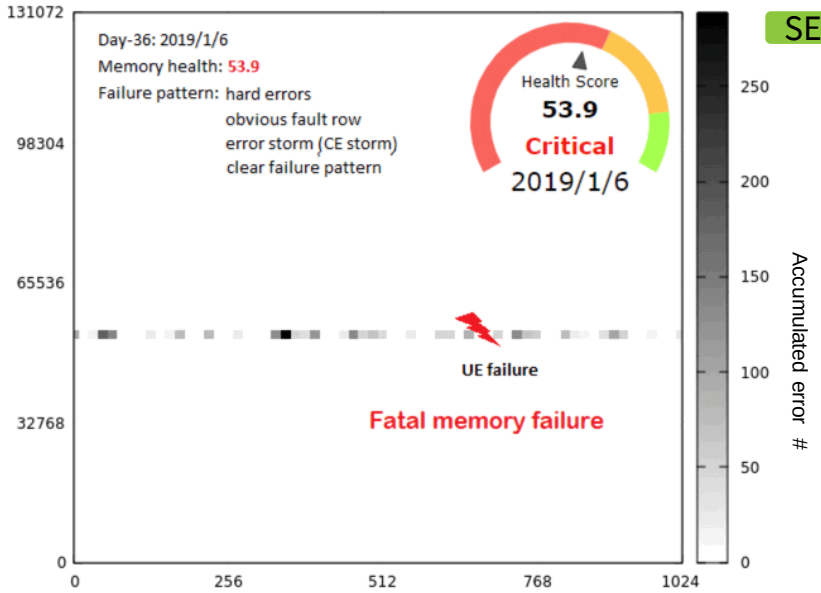
Using The Memory Model



Existing Memory Model in Action



SERVER



Source: Intel Corporation, November 2019.

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Existing Real-World Results

SERVER

Tencent

(Top APAC Cloud Service Provider)

Results

- 5X improvement in DIMM failure prediction and reduced downtime
- Simplified workload migration policies
- Optimized Page off-lining policies
- Reduced unnecessary expenses in DIMM replacement and upgrading

Reduced uncorrectable memory errors

Predicted memory failures based on historical data

Optimized page off-lining policies

Simplified workload migration decision making

Improved dim toss & purchase decision

Reduced downtime caused by server memory failures

Whitepaper: https://www.intel.com/content/www/us/en/software/intel-memory-failure-prediction-tencent.html

Meituan

(Top APAC eCommerce Vendor)

Results

- Intel® MFP helped Meituan analyze server memory health and predict failures before they happen
- Intel® MFP could help Meituan reduce server crashes caused by memory failures by 40%

Real-time visibility and predictive analysis into dram

Significantly reduced downtime caused by hardware failure to 40%

Reduced server failure through memory page offlining

Increased failure predictions & the resulting reduction of uncorrectable errors (UE)

Whitepaper: https://www.ami.com/download/intel-memory-failure-prediction-at-meituan-uses-machine-learning-to-send-potential-memory-failure-alerts-prior-to-hardware-failure-significantly-reducing-downtime/

OPEN POSSIBILITIES.

Further Thoughts for Improvement

Having all error handling of UE and CE handled by SMIs will still use valuable CPU cycles

- Handling of CE errors can potentially be gathered by BMC directly using platform interfaces
- Handling of UE errors should still be handled via SMIs to ensure proper capture in the case of a system crash

Integrating the prediction algorithm and health score aggregation with current infrastructure is key to finding real benefits

Call to Action

- Memory failures will continue to be a problem for the industry
- Predicting a memory DIMM failure has real world benefits
- Creating a model only takes time and data
- Let's work together to find new ways to utilize memory failure prediction and improve overall compute performance!

Open Discussion