OPEN POSSIBILITIES.

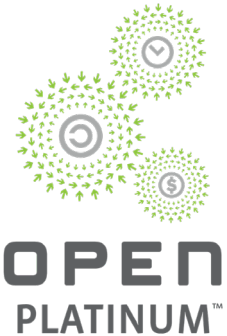
Memory Corrected Error profiling (via Linux EDAC Driver) within large-scale cloud infrastructure

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Memory Corrected Error profiling (via Linux EDAC Driver) within large-scale cloud infrastructure

Anil Agrawal, Hardware Systems Engineer, Meta
Stephanie Stickel, Hardware Analytics Engineer, Meta
Jonathan Zhang, Software Engineer, Meta

OPEN POSSIBILITIES.



Agenda

HW FAULT MGMT

- Data Center Pain-point to Address
- EDAC Enhancements – recap
- Memory Corrected Error Profiling
- EDAC extension proposal

Data Center Pain-points

- System Performance Degradation
- No-Trouble-Found in RMA
- Failure Mitigation Scaling

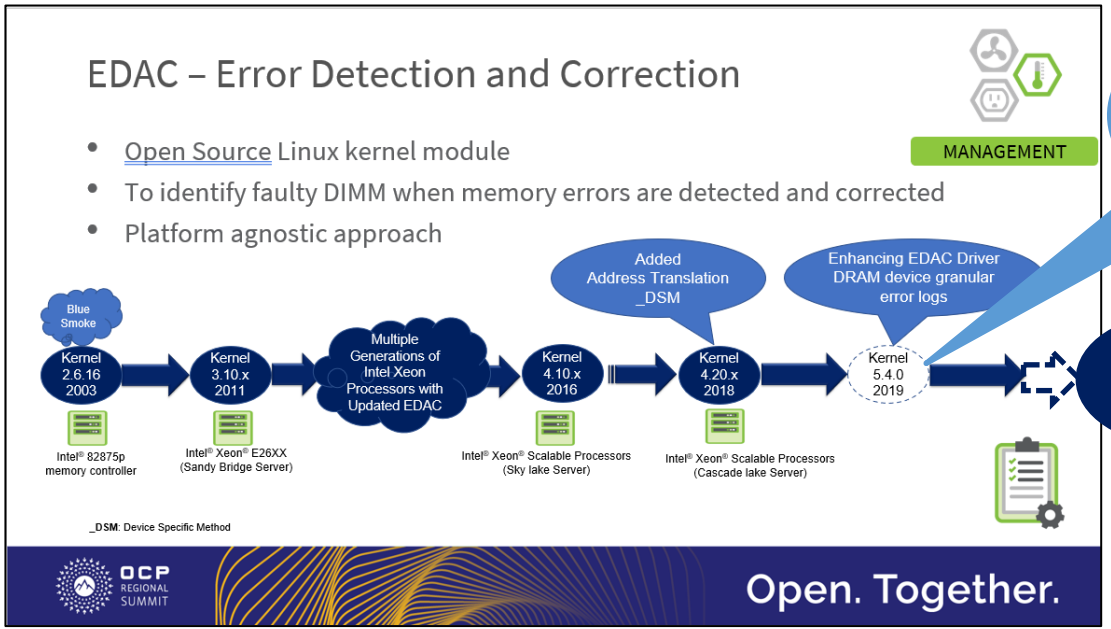
Shared Problem, Shared Solution

EDAC Drive based error reporting: Addressing the pain-point

Acronyms:
EDAC: Error Detection and Correction

EDAC Enhancements -Recap

Presented at 2019 OCP Regional Summit in Amsterdam



Eventually committed in kernel version 5.5

Kernel 5.15 2021

Added the capability to clear the logs after reading

OPEN POSSIBILITIES.

Acronyms:

EDAC: Error Detection and Correction

http://bluesmoke.sourceforge.net/

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

Agenda

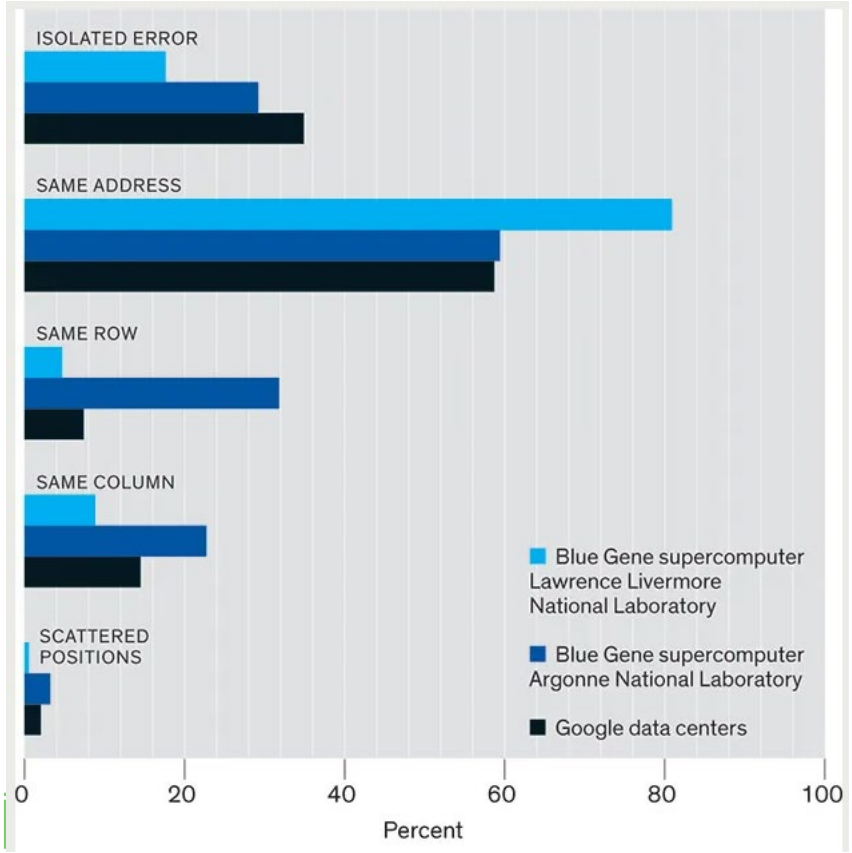
HW FAULT MGMT

- Data Center Pain-point to Address
- EDAC Enhancements – recap
- Memory Corrected Error Profiling
- EDAC extension proposal

Memory Corrected Error Profiling

Industry's known failure category/profile[1]

Limited data in public domain



Reference#1: https://spectrum.ieee.org/drams-damning-defects-and-how-they-cripple-computers

Memory Corrected Error Profiling

Methodology

Collect raw EDAC data from each host and dump it into a big database.

EDAC MC1: 1 CE memory read error on CPU_SrcID#0_MC#1_Chan#1_DIMM#0 (channel:1 slot:0 page:0xXXXXXX offset:0xXXX grain:XX syndrome:0xX - err_code:0xXXXX:0xXXXX socket:X imc:X rank:X bg:X ba:X row:0xXXXX col:0xXX retry_rd_err_log[XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX] correrrcnt[XXXX XXXX XXXX XXXX XXXX XXXX XXXX XXXX])

Convert into structured data with multiple columns

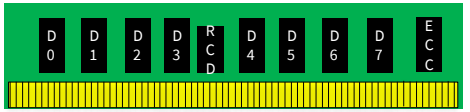
Define Distinct_error_Categories based on the syndrome, deviceID, page, row, column, total records…

Random_MBE, Random_SBE, Persistent_SBE, Persistent_Row, Single_DQ, Persistent_Bank, Block_error, Device_error, Single_burst, No matching pattern

Memory Corrected Error Profiling

Categorization Methodology Example

#	Category	Failure Cause	Observation Methodology (Same Dev)		
			Syndrome	Row Addr	Col Addr
1	Random_SBE	High energy Particle strike, marginal cells	Single-bit	Random	Random
2	Random_MBE		Multi-bits	Random	Random
3	Persistent_SBE	Retention time degradation	Single-bit	Same	Same
4	Persistent_Row	Word Line issue	Single/Multi-bits	Same Row	Random
5	Persistent_Bank	Peripheral issue	Single/Multi-bits	Same Bank	Random
6	Single_DQ	Peripheral issue	Column-bits	Random	Random
7	Miscellaneous	Block, device, burst, Col,..	Varies	Varies	Varies



One x4 Device Example

DQ3	DQ2	DQ1	DQ0
	b6	b5	b4
b11			
	b14	b13	
b19	b18		b16
b23			
b27		b25	b24
	b30		b28

Acronyms:

- MBE: Multi Bit Error
- SBE: Single Bit Error

OPEN POSSIBILITIES.

Memory Corrected Error Profiling

Results of past 9 months study

#	Category	Industry[1]	Meta Study
1	Random_SBE	32%	16%
2	Random_MBE	--	27%
3	Persistent_SBE	48%	12%
4	Persistent_Row_Error	8%	21%
5	Persistent_Bank_Error	--	4%
6	Single_DQ_Error	--	12%
7	Miscellaneous	12%	8%
	Total	100%	100%

computers

Key Learnings and future mitigation ideas:

- Random (transient) errors: ~50%
- Single bit persistent error can be mitigated by page–offline. Should not result in DIMM swap.
- Row persistent error can be mitigated by PPR
- Can debug vendor specific issues at scale

Acronyms:

- MBE: Multi Bit Error
- PPR: Post Package Repair
- SBE: Single Bit Error

Agenda

HW FAULT MGMT

- Data Center Pain-point to Address
- EDAC Enhancements – recap
- Memory CE Profiling
- EDAC extension proposal

EDAC Driver Extension Proposal

1. Extend the capability to capture additional errors using OS-first method
   a. Add Missing Memory corrected error type
   b. Add PCIe corrected error types
2. Tooling improvement: Add OS based SEL logging capability
   a. Problem to solve: FW-first method of error reporting allows logging events in platform's persistent storage. No existing mechanism available for such event logging in OS-first method (e.g., EDAC based).
3. Future efforts to implement OS-first based RAS action, e.g., triggering PPR.

Plan to develop requirements/specifications

Acronyms:

- PPR: Post Package Repair
- SEL: System Event Log

EDAC Extension Proposal#1

Existing EDAC logs Coverage

1. Demand read CE/UCR (error code: 0x101:0x009n)
2. Patrol Scrubber read CE (error code: 0x0008:0x00Cn)
3. Patrol Scrubber read UCR (error code: 0x0010:0x00Cn)
4. Demand Partial write CE/UCR (error code: 0x0104:0x00An)

Additional Coverage Proposed

1. Command/Address Parity error (error code: 0x0200:0x00Bn)

EDAC Extension Proposal #1:

Add capability to report Command/Address parity error

EDAC Extension proposal#1: Workflow

- Add OS-first handling for memory corrected, uncorrected-recoverable error.
- The workflow:
  - Host firmware configures CMCI to be triggered instead of SMI, when there are correctable errors.
  - Host Firmware configures interrupt policy, such as leaky bucket thresholds and time period length.
  - When CMCI is triggered, EDAC driver interrogates the error syndrome registers, and generates system trace messages as appropriate.

Acronyms:
- CMCI: Corrected Machine Check Interrupt
- SMI: System Management Interrupt

EDAC Extension Proposal#2

Add PCIe Error Coverage

Hardware Corrected Error	Uncorrected Non-fatal Error
1. Receiver Error	1. Poison TLP
2. Bad TLP	2. Unsupported Request
3. Bad DLLP	3. Completer Abort
4. Replay Timer Timeout	4. Unexpected Completion
5. Replay Number Rollover	5. Completion Timeout
6. Advisory Non-Fatal	6. ACS violation
	7. Completion with UR,CA

EDAC Extension Proposal #2:

Add capability to report PCIe errors

Call to Action

HW FAULT MGMT

- Collaborate on EDAC extension development
  - Design requiremets/specification, implementation, and upstreaming
- Where to find additional information
  - Wiki with latest information: https://www.opencompute.org/wiki/Hardware_Management/Hardware_Fault_Management
  - Mailing list: https://ocp-all.groups.io/g/OCP-HWFaultMgt/
  - Contact directly: Jonathan Zhang (jonzhang@fb.com) or Anil Agrawal (anilagrawal@fb.com)

Thank you!

NOVEMBER 9-10, 2021

Open Discussion

Abstract

HW FAULT MGMT

Title: A study of Linux EDAC drive based hardware corrected error reporting and profiling methods in a large-scale cloud infrastructure

Abstract: As a result of prior OCP initiative in developing 'Enhanced EDAC driver' to collect fine-granular memory corrected error logs, we can observe and create error profile that was not feasible in the past. It is helping us in making data-driven decision to manage such faults more effectively in our cloud infrastructure.
This presentation consists of two parts: first we plan to share the results of large-scale study of memory corrected errors over past nine-months, then we will share our proposal to extend similar approach to other types of hardware corrected errors, e.g., PCIe, cache, and inter-CPU link errors. This is in line with OCP's OS-first methodology for Hardware corrected error reporting to eliminate dependency on SMI based methods. We will also discuss various mitigation options we are considering managing the impact of such memory corrected errors.

Speakers: Anil Agrawal, Stephanie Stickel, Jonathan Zhang