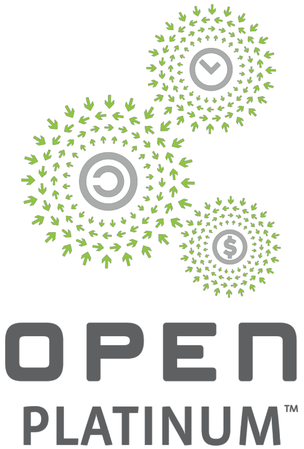
Open. Together.

OCP
SUMMIT

Hardware
Management

Performance Oriented Firmware Implementation of Error Handling on OCP Platforms for Large Scale Data center



Zhengyu Yang, Hardware System Engineer, Facebook
Chris Davis, Core Data Production Engineer, Facebook

MANAGEMENT

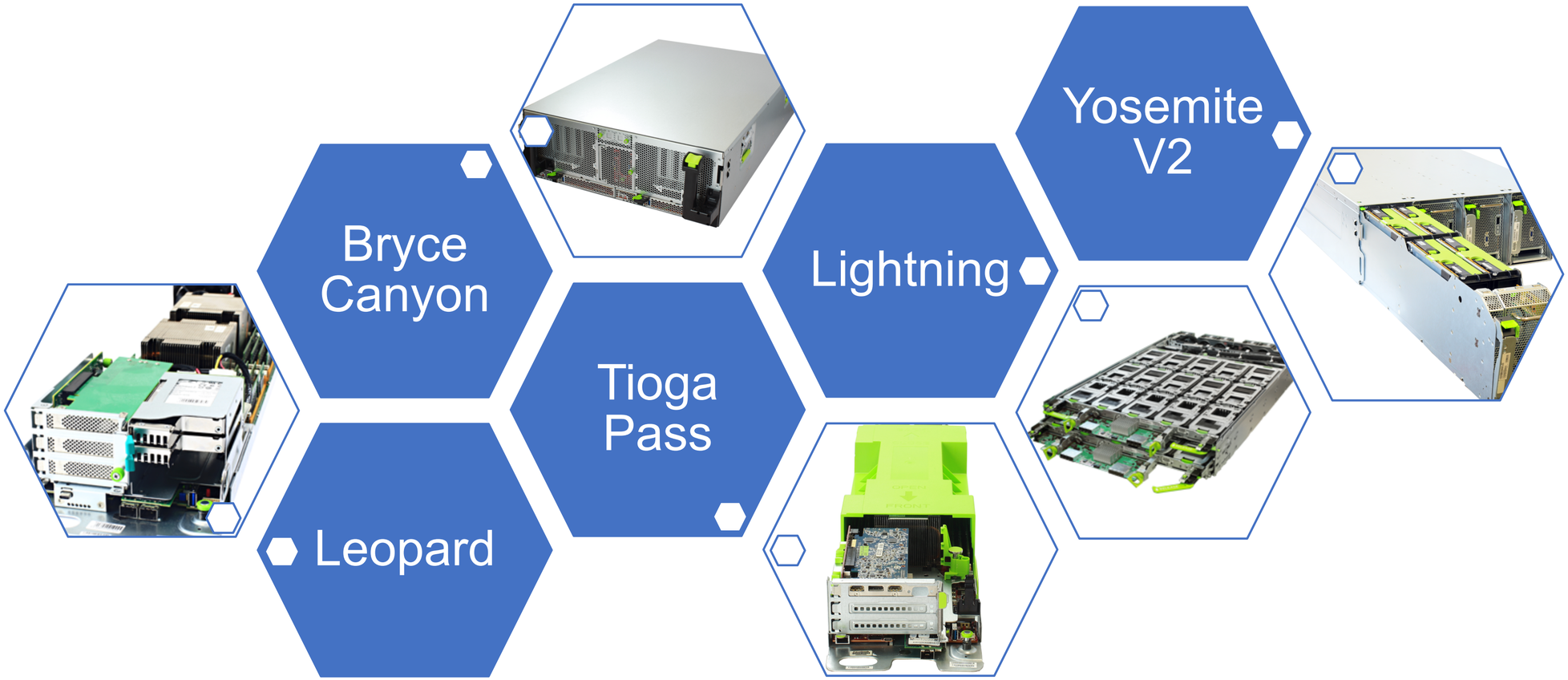
Agenda

- Generic performance problem on OCP platforms
- Problem diagnose and generic firmware solution
- Results and verification strategy
- Summary

White Papers

Performance Issues Affects Various OCP Platforms

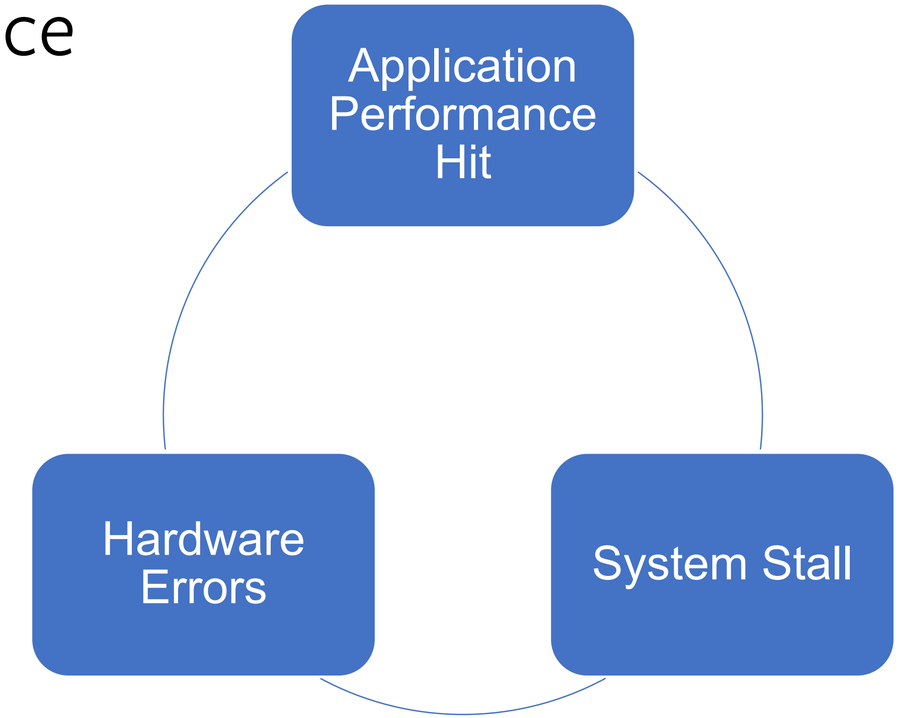
Platforms Affected in Facebook Fleet



Performance Issue Degrades Availability & Latency

Symptoms Observed in FB production

- Low latency systems like caches experience timeout spikes
- Failures in time critical processes like DB master promotion
- Packet loss, especially impacting new connection attempts
- Higher latencies & error rates in general



Diagnosing & Detecting System Stalls

System Unresponsiveness (a.k.a System Stall)

- Machine functions correctly but stalls for hundreds or even thousands of milliseconds at a time
- “stall” means the machine does 0 work for this time period

Task	Runtime ms	Switches	Average delay ms	Maximum delay ms	Maximum delay at
ksoftirqd/0:3	261.106 ms	166	avg: 25.445 ms	max: 570.087 ms	max at: 12725714.754445 s
rcu_sched:8	640.922 ms	8048	avg: 0.108 ms	max: 381.426 ms	max at: 12725671.219097 s
scribe_cat:(99)	2739.591 ms	1171	avg: 0.349 ms	max: 269.621 ms	max at: 12725714.454340 s
xargs:(18)	2541.290 ms	22096	avg: 0.018 ms	max: 254.727 ms	max at: 12725683.015549 s
kworker/3:0:244396	2.960 ms	121	avg: 2.193 ms	max: 254.498 ms	max at: 12725691.376932 s
ksoftirqd/3:29	3.871 ms	99	avg: 2.751 ms	max: 254.495 ms	max at: 12725691.376942 s
cat:(21992)	23824.856 ms	66206	avg: 0.020 ms	max: 254.491 ms	max at: 12725690.099208 s
carbon-global-s:(3)	446.988 ms	492	avg: 1.647 ms	max: 254.489 ms	max at: 12725691.376856 s
LadDxChannel:2861439	660.121 ms	5147	avg: 0.065 ms	max: 254.454 ms	max at: 12725691.376797 s
FutureTimekeepr:(146)	1963.384 ms	6555	avg: 0.283 ms	max: 254.442 ms	max at: 12725691.376753 s
[xarexec] /usr/:(21)	14159.697 ms	9236	avg: 0.036 ms	max: 254.427 ms	max at: 12725691.376745 s
kworker/1:0:1303233	2.927 ms	131	avg: 2.058 ms	max: 254.426 ms	max at: 12725691.376798 s
ksoftirqd/1:17	5.037 ms	109	avg: 2.447 ms	max: 254.412 ms	max at: 12725691.376817 s

Diagnosing & Detecting System Stalls

Detection

- Detection process simply sleeps and measures how long it actually slept.
- Stalls manifest as a significant delay in waking up.
- Doing this on all cores increases signal

Script for Detecting Stalls

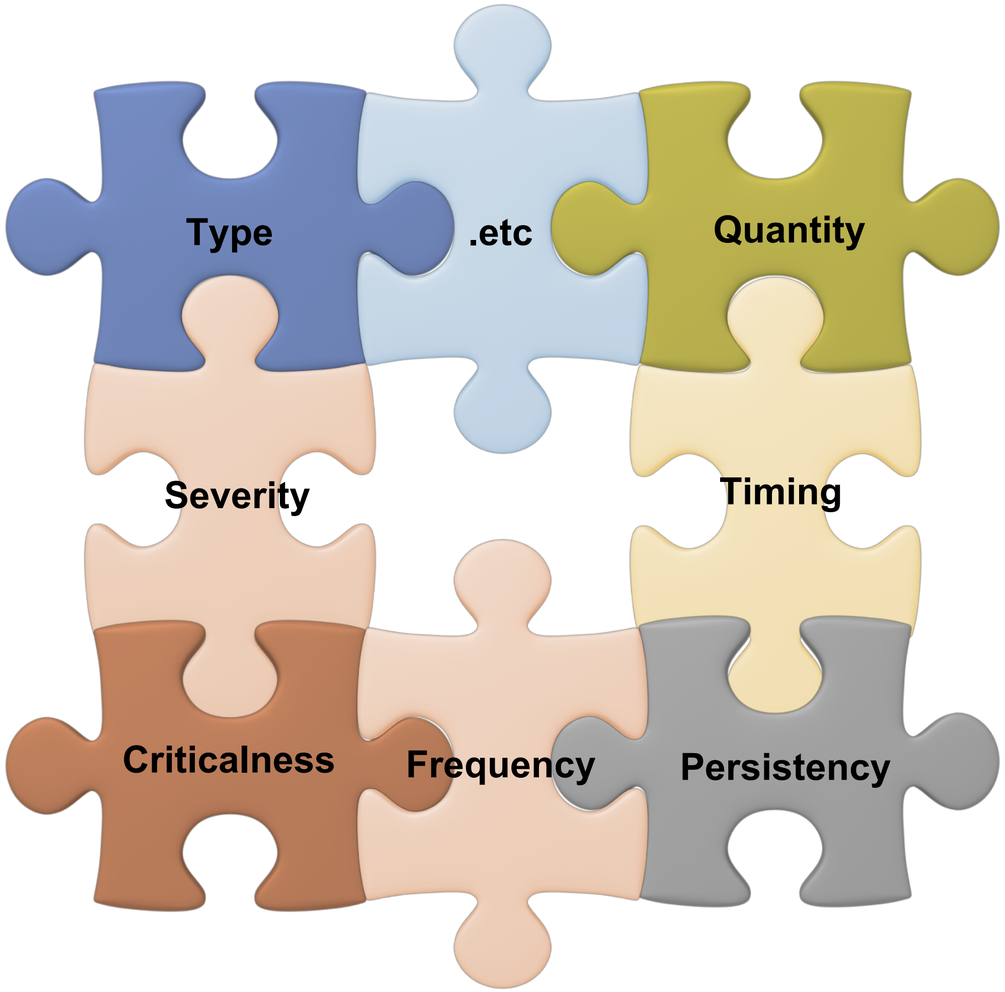
```
while True:
  start = time()
  sleep(100ms)
  elapsed = time() - start
  if elapsed > 150ms:
    print(“Stall Detected”)
```

System Stalls Correlate with Hardware Errors

Correctable Errors are common

```
[12725978.954807] {1861667}[Hardware Error]: Hardware error from APEI Generic Hardware Error Source: 0
[12725978.972915] {1861667}[Hardware Error]: It has been corrected by h/w and requires no further action
[12725978.991358] {1861667}[Hardware Error]: event severity: corrected
[12725979.003874] {1861667}[Hardware Error]:  Error 0, type: corrected
[12725979.016388] {1861667}[Hardware Error]:   section_type: PCIe error
[12725979.029095] {1861667}[Hardware Error]:   port_type: 4, root port
[12725979.041610] {1861667}[Hardware Error]:   version: 1.16
[12725979.052407] {1861667}[Hardware Error]:   command: 0x0010, status: 0x0547
[12725979.066313] {1861667}[Hardware Error]:   device_id: 0000:00:01.0
[12725979.078838] {1861667}[Hardware Error]:   slot: 0
[12725979.088590] {1861667}[Hardware Error]:   secondary_bus: 0x01
[12725979.100438] {1861667}[Hardware Error]:   vendor_id: 0x8086, device_id: 0x6f02
[12725979.115221] {1861667}[Hardware Error]:   class_code: 040600
[12725979.126871] {1861667}[Hardware Error]:   bridge: secondary_status: 0x2000, control: 0x0013
```

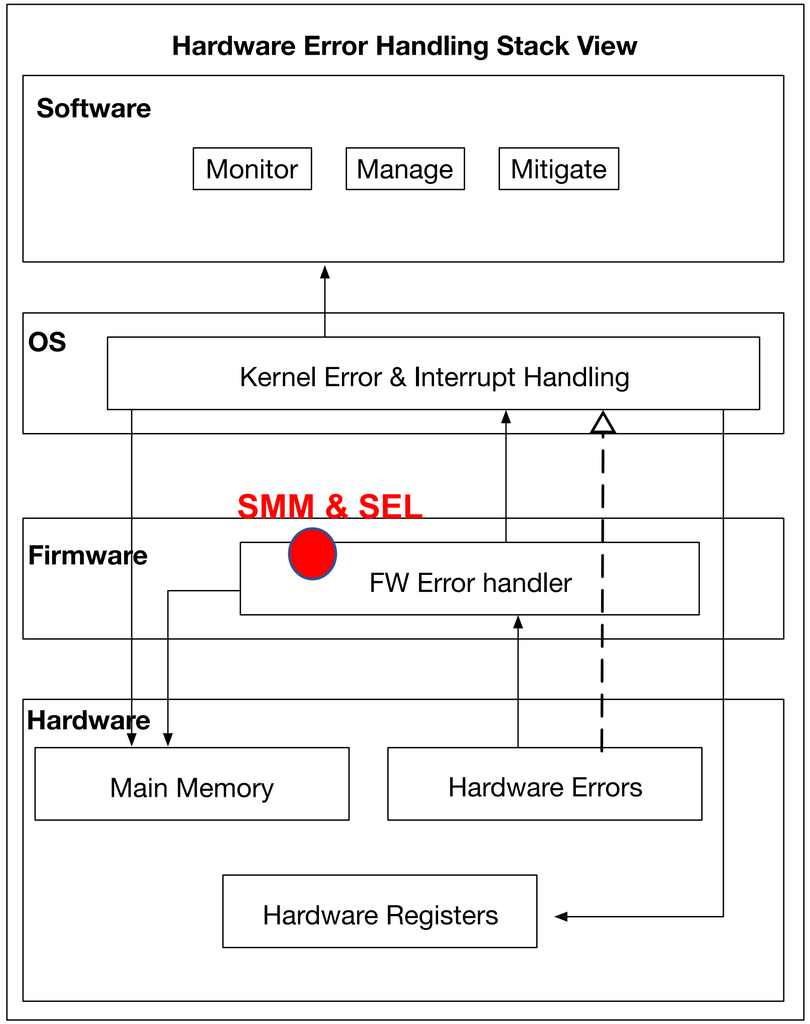
Breakdown of HW Error's Performance Impacting Aspects



Hardware Error Handling Overview

Stack view

- Components
  - Detect, Report and Remediate
- Contributors/Owners
  - **OS & Software** : Harvest failure information and apply remediation method if any
  - **Firmware** : Implement HW error handlers
  - **Hardware**: Provide HW capabilities and error information for error handling

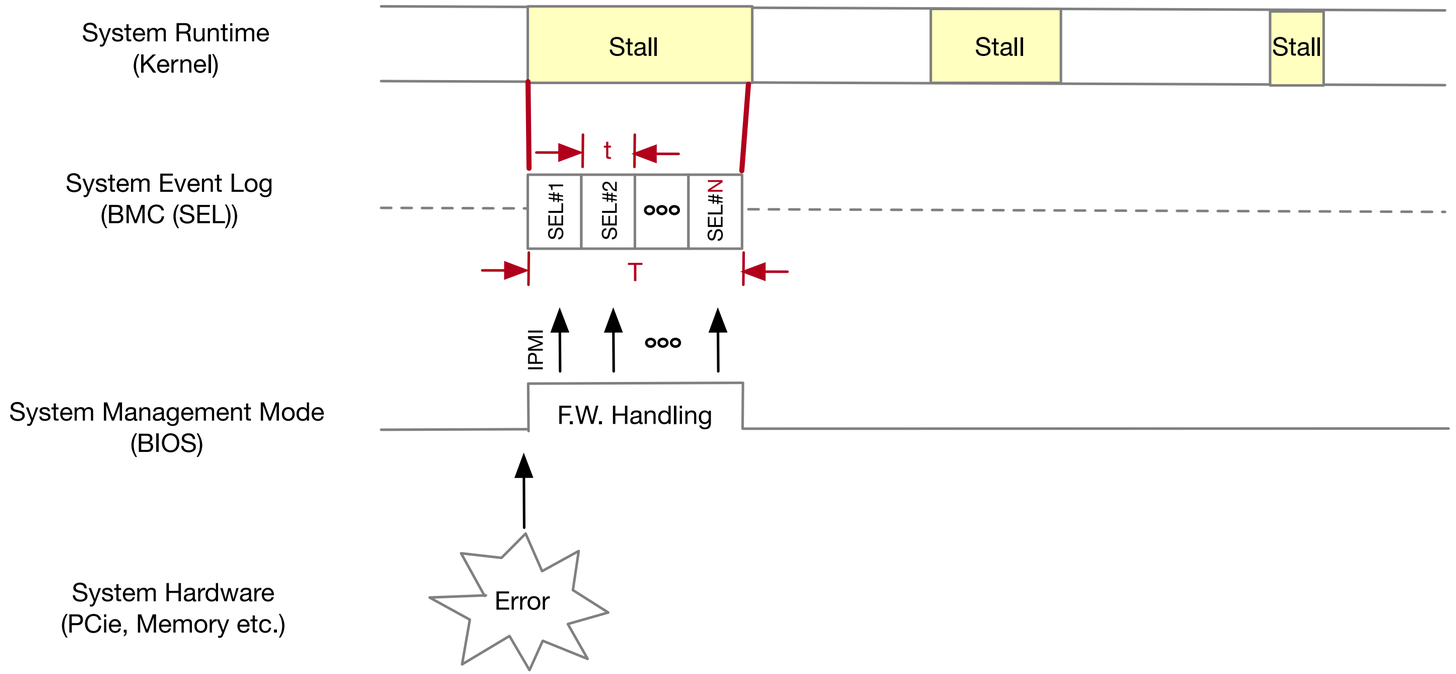


SMM and SEL

- System Management Mode (SMM)
  - Contains SMM HW error handlers
  - Context switch all cores when entered
  - Suppose to be fast
- System Event Log (SEL)
  - Industry standards, allow content customization
  - Created and sent out by BIOS using IPMI commands within SMM
  - IPMI commands can be expensive

Performance Issue Root Cause

System Stall due to SMM & SEL



T	=	N	*	t
Long Latency (a.k.a System stall)		Multiple SELs used for logging single PCIe Error		Long latency for single PCIe error logging

Generic problem due to FW, across OCP platforms!

Performance Focused Hardware Error Handling

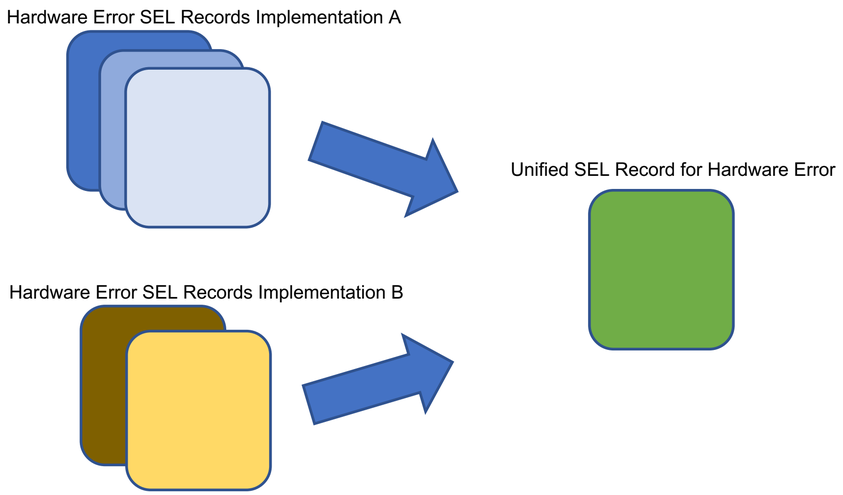
FW Solution Strategy

- Failure event based reporting vs. Error occurrence based reporting
- Consolidated, Standardized, Scalable error record
- Overhead minimized error information communication

Performance Focused Hardware Error Handling

FW Solution Example -- BIOS

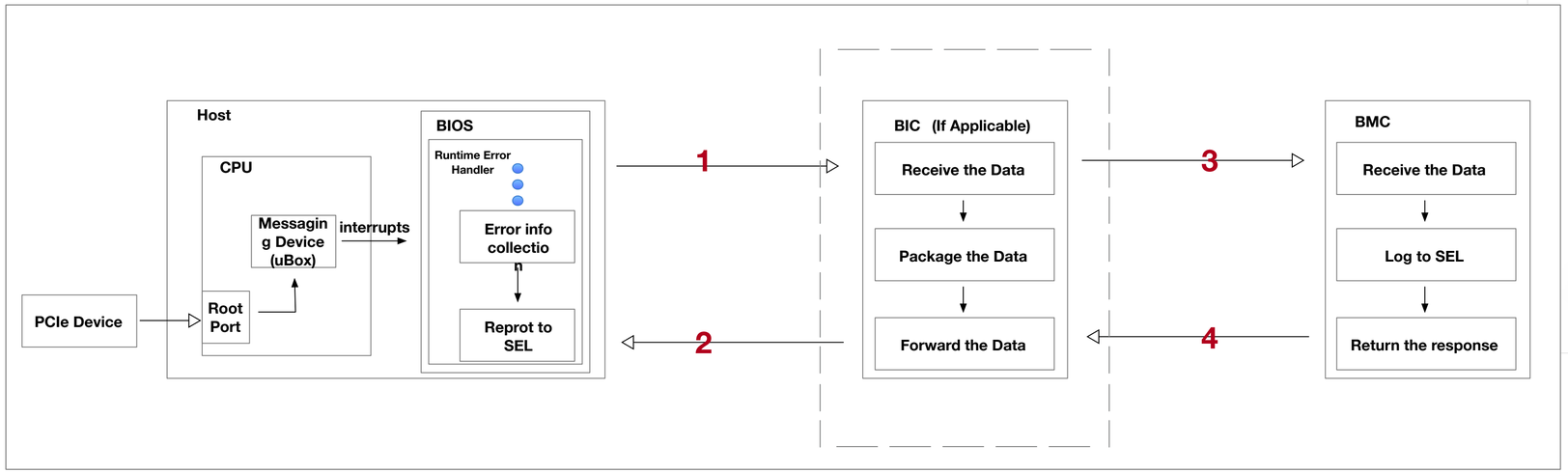
- Failure Event
  - High Frequency Error Burst
  - Low Frequency Error Tracking
- Unified failure record format



Performance Focused Hardware Error Handling

FW Solution Example – Bridge IC

- Overhead minimized error information communication
  - Ex. “Return first” Model in BIC FW

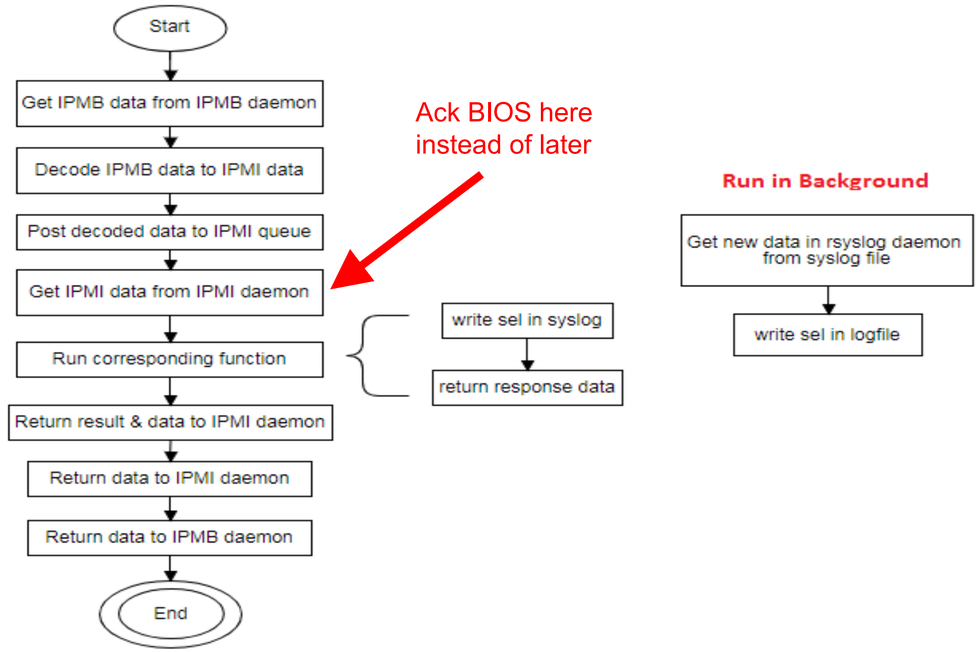


High-level Flow of HW Error SEL Logging with BIC

Performance Focused Hardware Error Handling

FW Solution Example – BMC

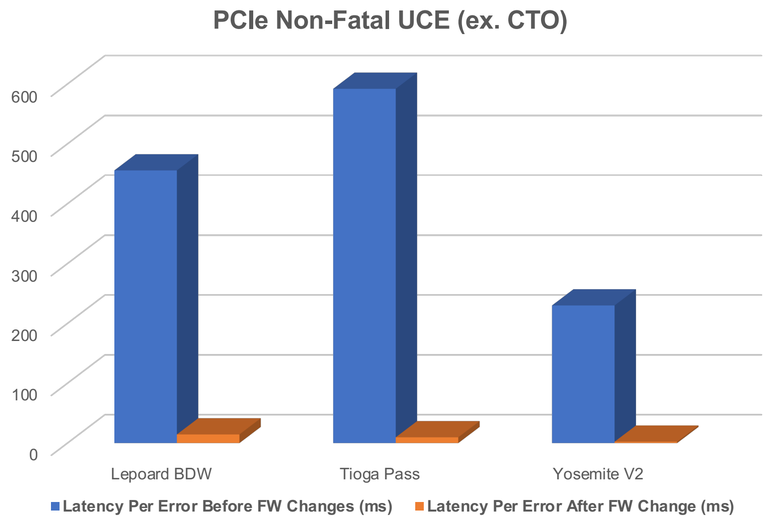
- Overhead minimized error information communication
  - Ex. Enhanced BMC SEL handling flow -- ”Ack Quick”*

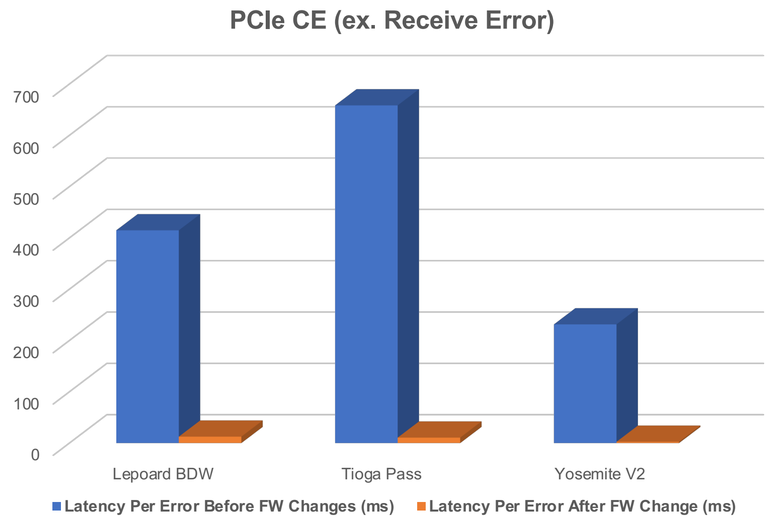


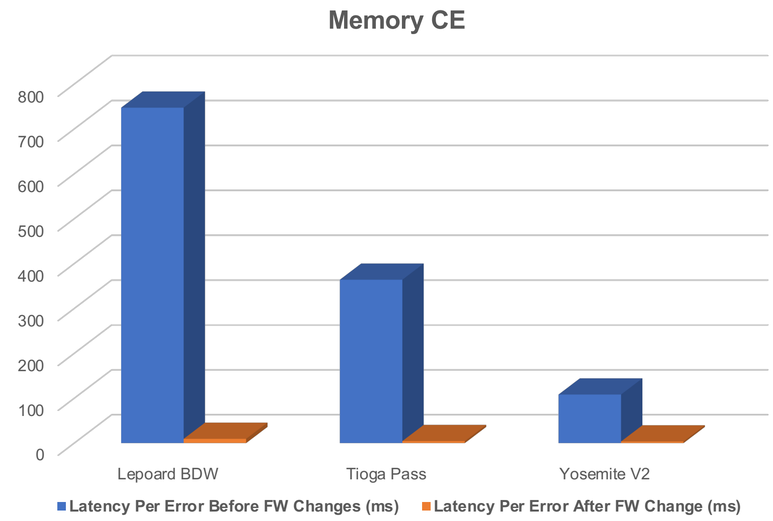
* Submitted to oBMC committee and pending upcoming spec adoption

Performance Focused Hardware Error Handling

FW Solution Example Results from 3 OCP platforms



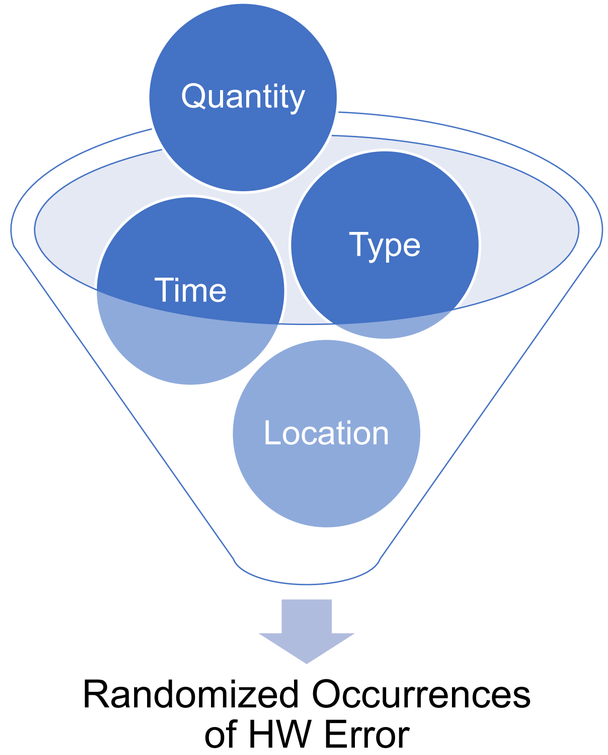




Hardware Error SEL logging latency down to as low as 2.34ms

FW Implementation Verification Strategy

- Error Injection
  - HW capability support
  - Error type coverage
- Performance focused
  - Use case driven metrics
  - Quantifiable criteria
- Verification model
  - Performance measured with hardware failures/errors
  - Error occurrences' randomness



Summary

- Hardware error handling caused system performance impact is a common problem across OCP platforms
- Improvements (ex. logging strategy, processing efficiency) can be made at the common firmware layer and applicable to various OCP platforms
- Development and validation of firmware solutions for hardware error handling needs to be system performance aware

Call to Action

MANAGEMENT

- Contribution to OCP (Planned Q3 2019)
  - IPMI SEL HW Error record format and content
  - BIC FW “Return First” Model
  - BMC FW “Ack Quick” SEL processing flow

- Check the fleet
  - OCP system stall in the fleet
  - Correlate system stall with hardware error occurrences
  - Adopt the presented fixing strategy to see improvements


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT