
OPEN
Compute Project
OCP U.S. SUMMIT 2017
Santa Clara, CA

Lightning (PCIe JBOF): Update, challenges, and solutions

Chris Petersen, Hardware Systems Technologist, Facebook
Wesley Yung, Applications Technical Lead, Microsemi
Bob Pebly, Platform Architect, Intel

OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

OPEN Compute Project

Lightning update

Design and validation status = Complete!

OCP contributions

- Hardware:
  - Wiwynn is contributing full Lightning design package
  - PCIe retimer specification available and design package coming
- Software:
  - OpenBMC (https://github.com/facebook/openbmc)
  - Switch management
  - Drivers

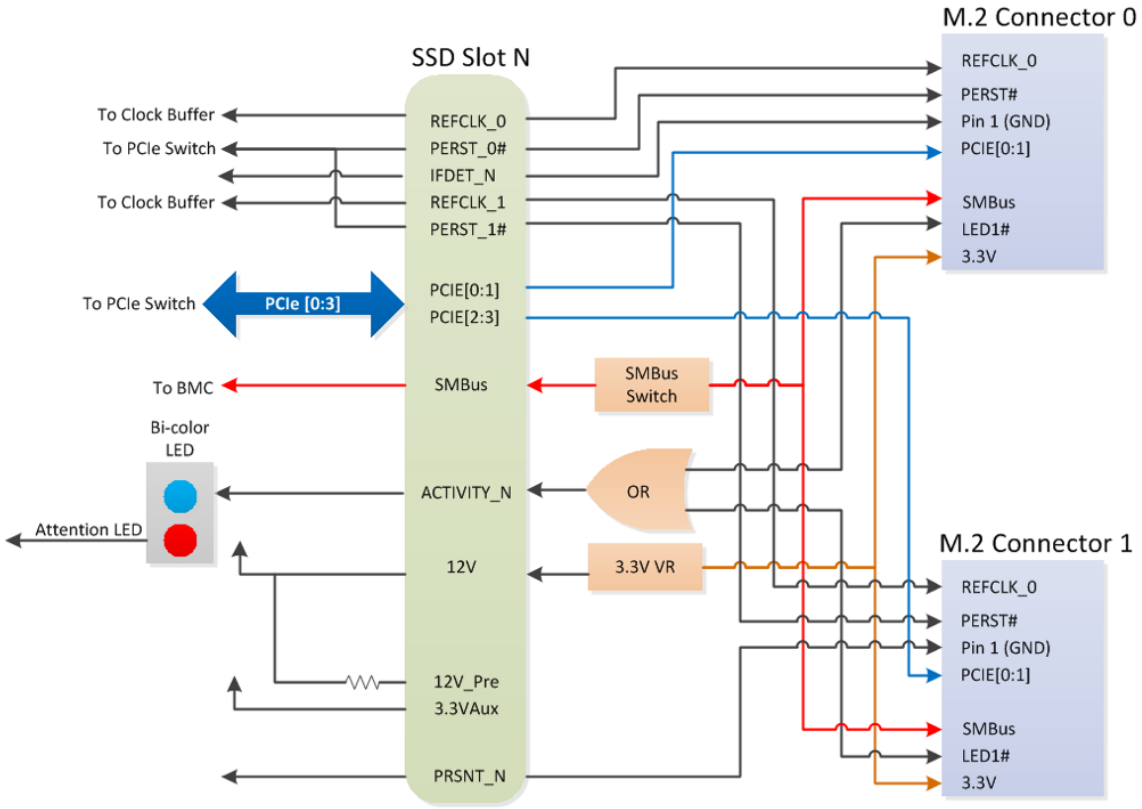
PCIe JBOF Challenges

- M.2 support
- Enclosure management
- PCIe Hot-plug

M.2 solution

- Thermal

- PCIe Hot-plug



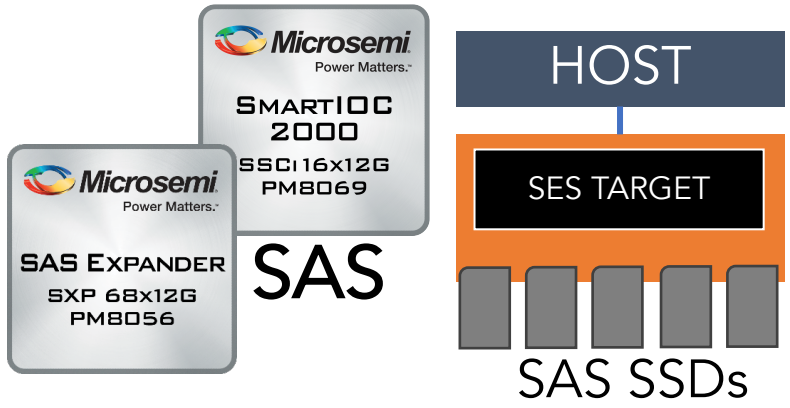
JBOF Enclosure Management

Wesley Yung, Applications Technical Lead, Microsemi

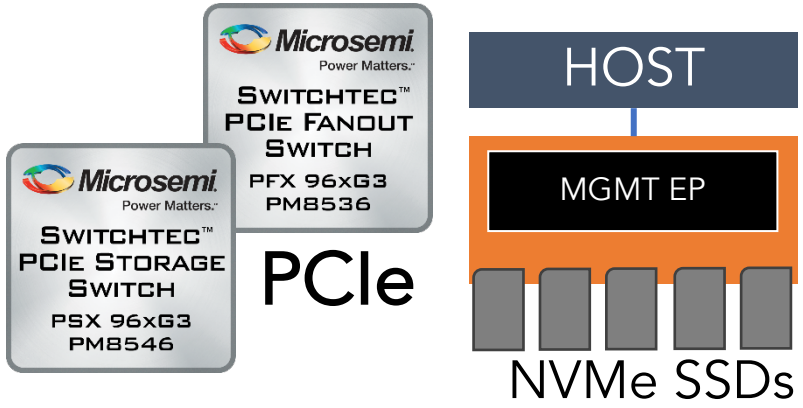
OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

OPEN Compute Project

Aligning PCIe Switch management to SAS capabilities



- Utilizes SCSI-Enclosure Services (SES) to manage the *JBOD* enclosure
- SAS expander supports in-band and out-of-band management of endpoints



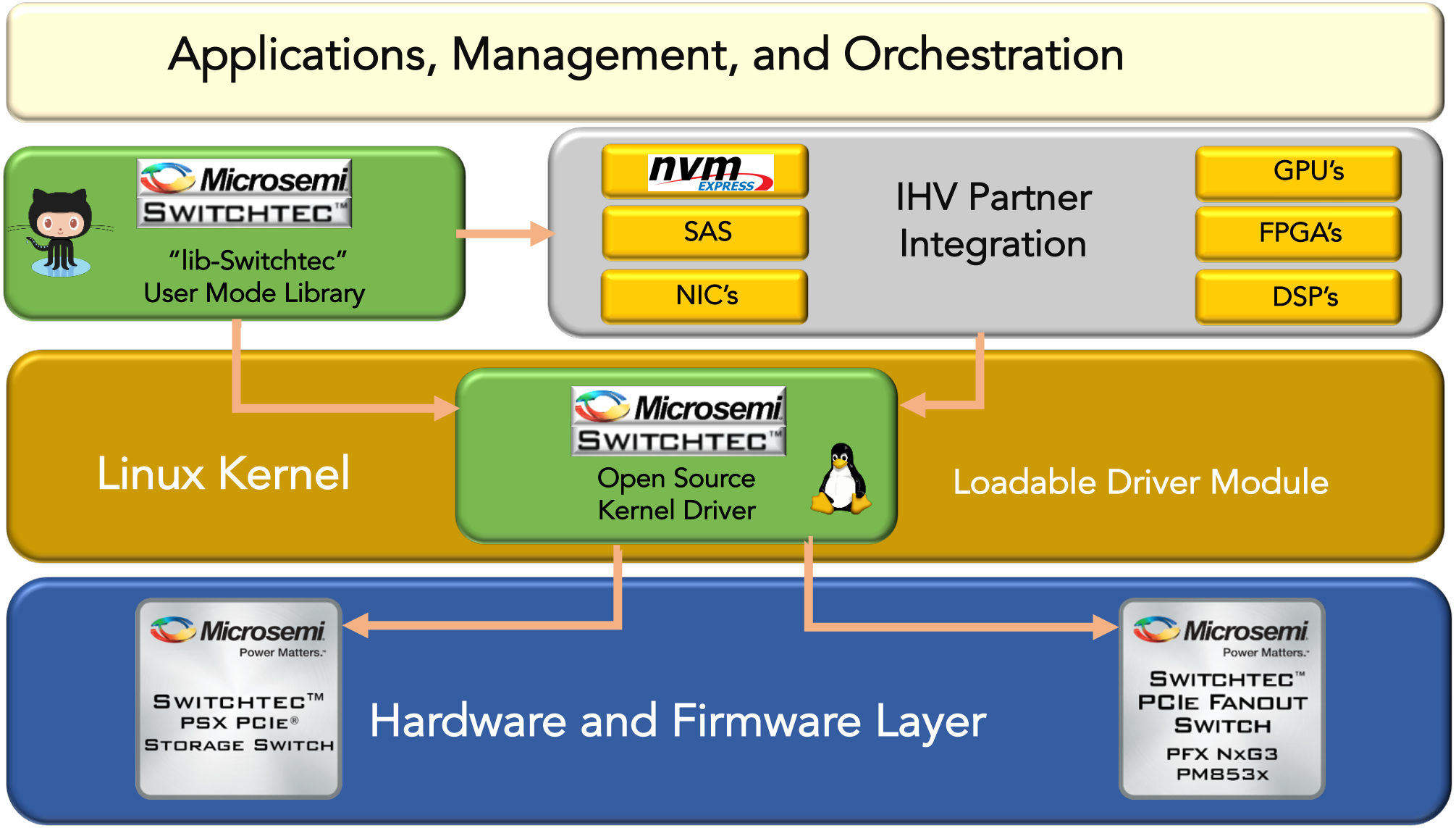
- Utilizes a PCIe management endpoint to manage the *JBOF* enclosure
- Supports in-band and out-of-band management of endpoints through Memory-mapped Remote Procedure Calls (MRPC)

Up-stream/open-source driver and utilities

- Kernel Driver
  - https://github.com/sbates13027 2/switchtec-kernel
  - Lightweight shim driver providing access to switch internal management endpoint
  - *Targeting Kernel 4.11*
  - https://lkml.org/lkml/2017/2/2/4 45
- User-space utility
  - https://github.com/sbates130 272/switchtec-user
  - FW Download / Upload
  - Error, Performance counters
  - Temperature monitor
  - Port Status / Health

```
gunthorp@cgy1-donard:~$ switchtec version
switchtec version 0.4
gunthorp@cgy1-donard:~$ switchtec list
/dev/switchtec0         PSX 48XG3       RevB    1.06 B03F       0000:03:00.1
gunthorp@cgy1-donard:~$ switchtec test /dev/switchtec0
/dev/switchtec0: success
gunthorp@cgy1-donard:~$ switchtec temp /dev/switchtec0
48.1 °C
```

Driver Architecture / Integration



PCIe & NVMe Hot Plug

Bob Pebly, Platform Architect, Intel

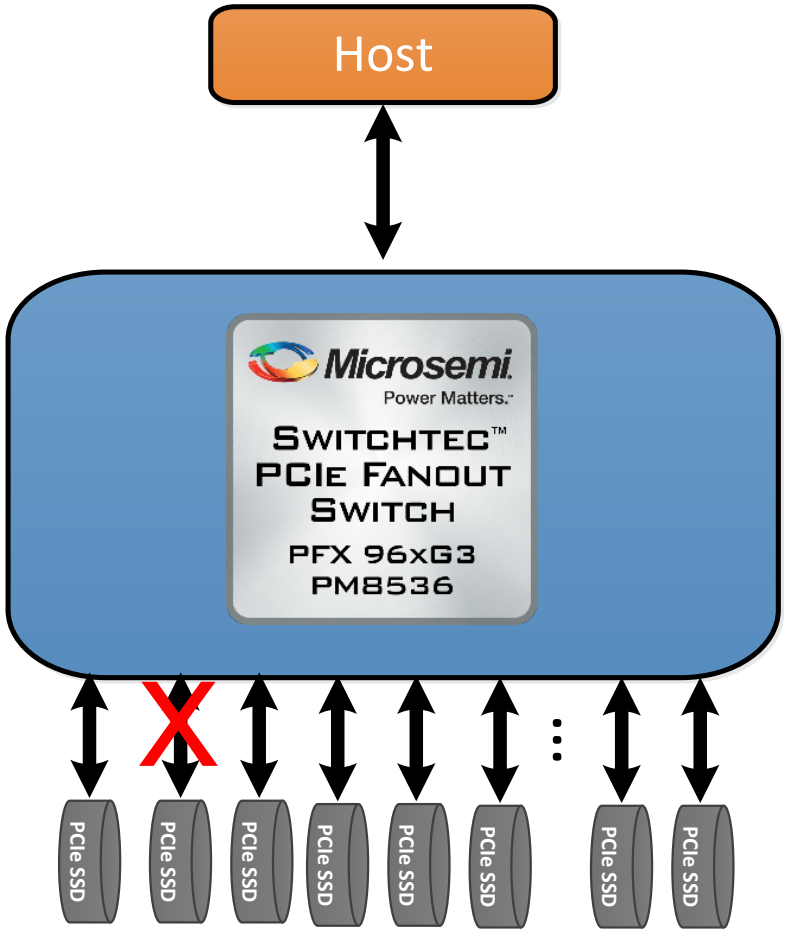
Wesley Yung, Applications Technical Lead, Microsemi

OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

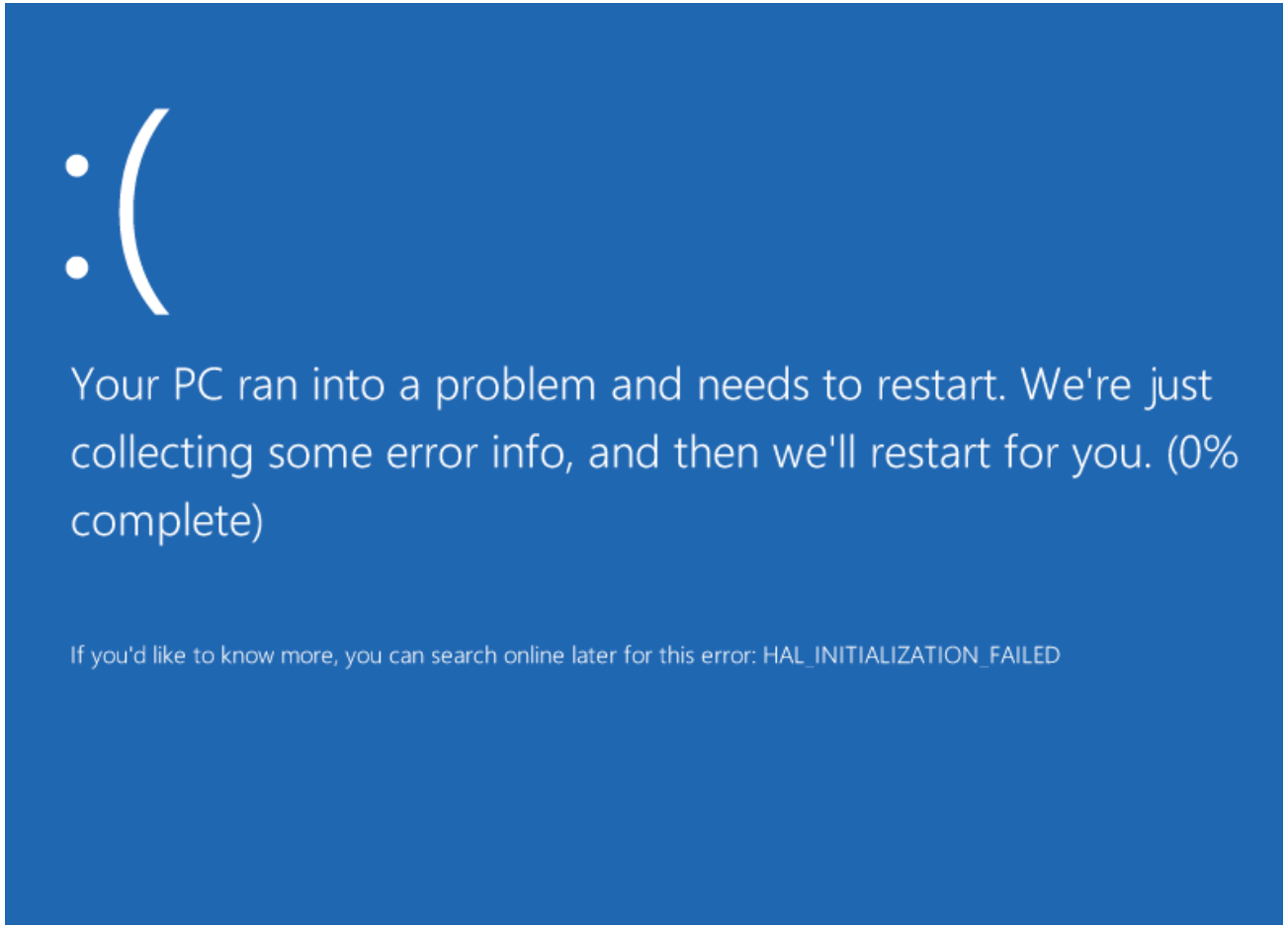
OPEN Compute Project

High level goal

Do this…

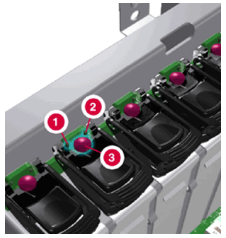


…and THIS doesn't happen



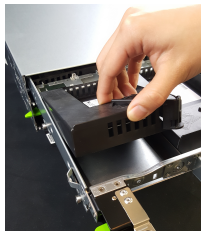
Key Terminology

Managed Hot Add



Managed Hot Remove

Surprise Hot Add



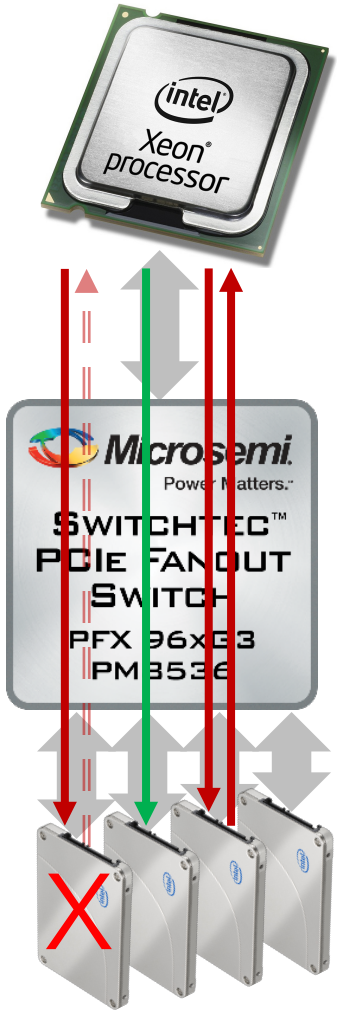
Surprise Hot Remove

Too many steps! Too many sidebands!

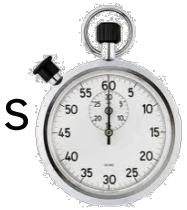
OPEN Compute Project

Facebook Lightning Hardware System

Challenges: Completion Timeout



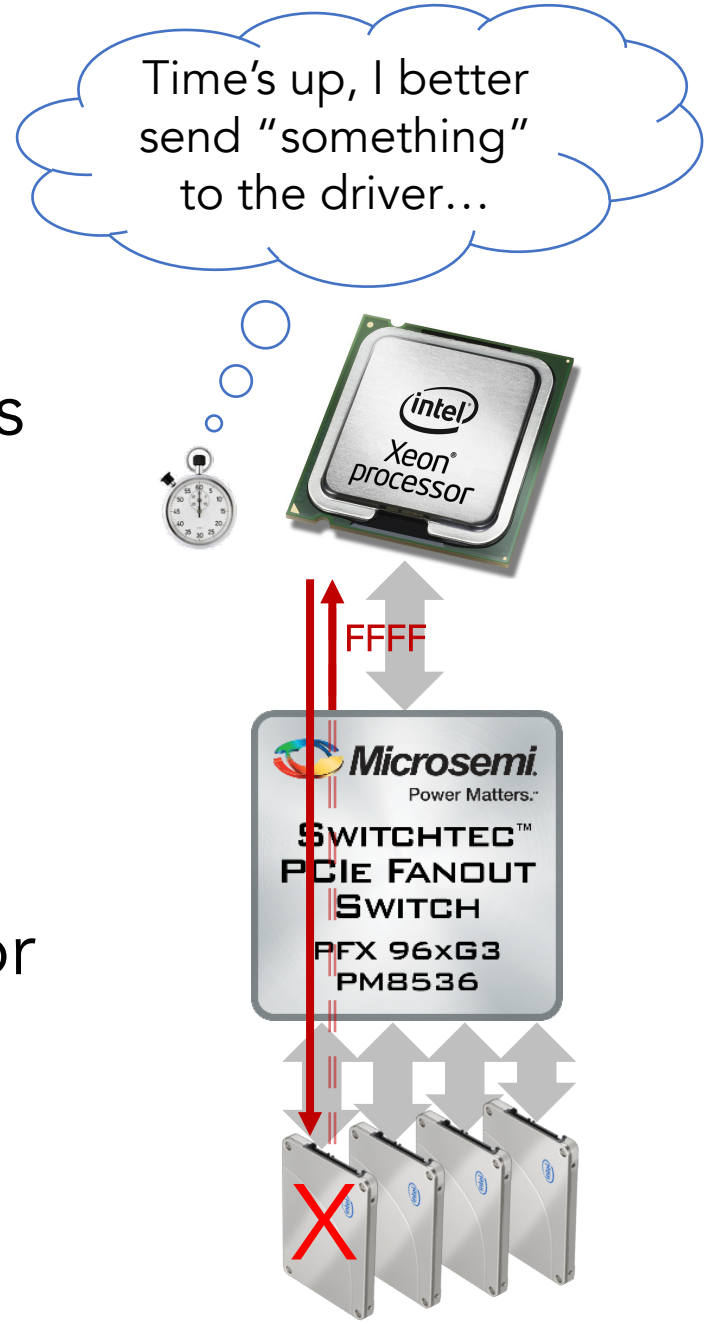
- Posted vs. Non-posted Transactions
  - Posted - Request only, no completion response
  - Non-posted – Split Transaction, Request & Completion
- Surprise hot plug can (& will) leave <u>many</u> transactions incomplete
  - Completion Timeout (CTO) is the PCIe mechanism to terminate incomplete transactions



Challenges:
All 1's (All F's) Completions

- A completion with data (CplD) where the data is All 1's
  - Memory Read, Config Read
- Happens when a completion never returns to protect requester from timing out
- Prior to Lightning - ***NO*** support for All 1's Completions in NVMe or PCIe service drivers (or most other Linux drivers)
- Otherwise…

```
[265862.256129] Uhhuh. NMI received for unknown reason 39 on CPU 0.
[265862.268121] Do you have a strange power saving mode enabled?
[265862.279578] Dazed and confused, but trying to continue
```

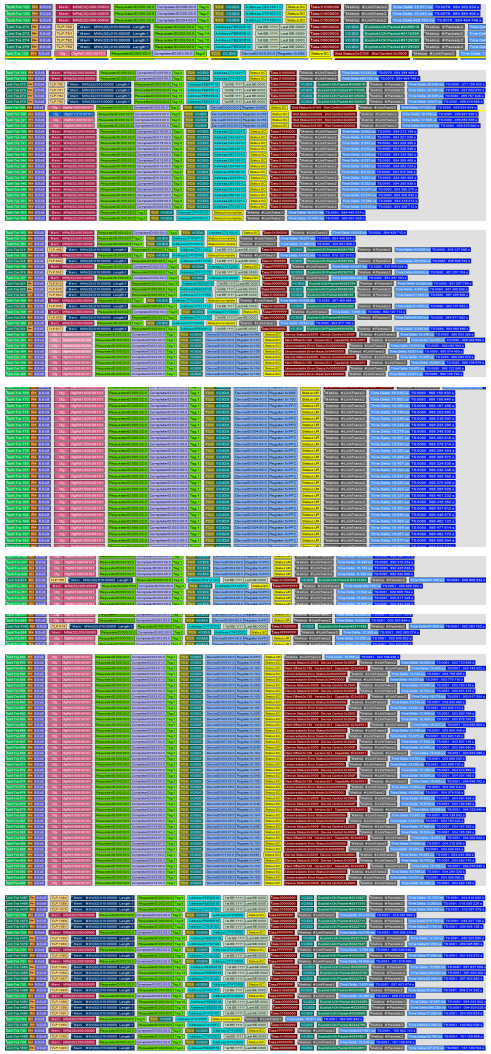


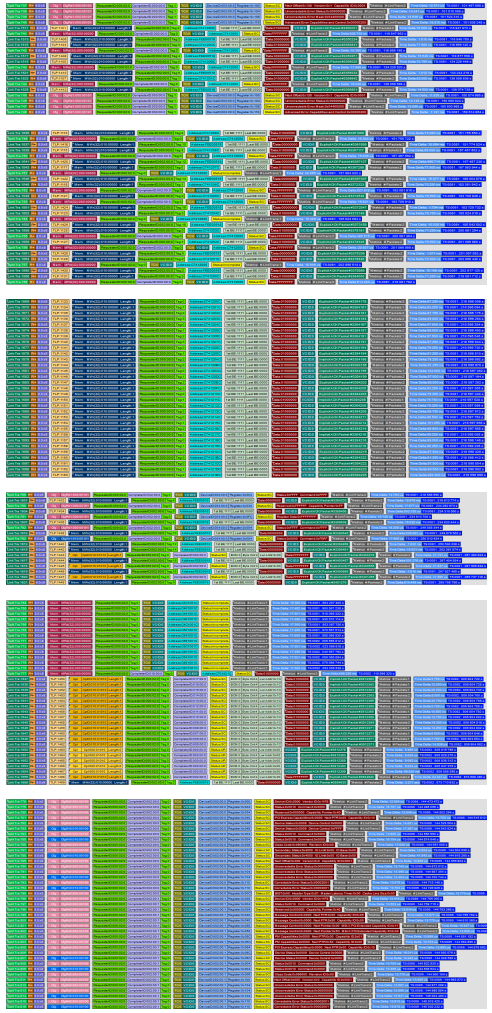
Solutions: Intel Surprise Hot Plug Linux Contributions

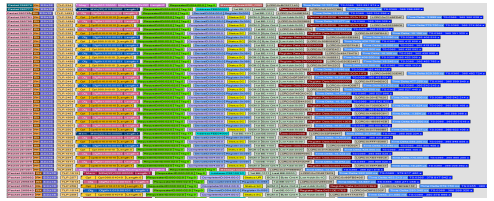
Contribution	Kernel Version
New PCIe Downstream Port Containment (DPC) driver	4.7
Enhancements & optimizations to PCI driver Recognizing All 1's as a missing device on key config registers	pending
Enhancements & optimizations to AER driver Caching of extended capability pointers	4.4
Enhancements to NVMe driver Recognizing All 1's as a missing device Cleaning up after hot remove without further IO	4.7
Enhancements & fixes in the block multi-queue driver Dealing with errors returned on IO following surprise removal	4.7

Kudos to Keith Busch for Linux NVMe & PCIe driver enhancements

Before & After:
NVMe Surprise Remove PCIe IO Trace







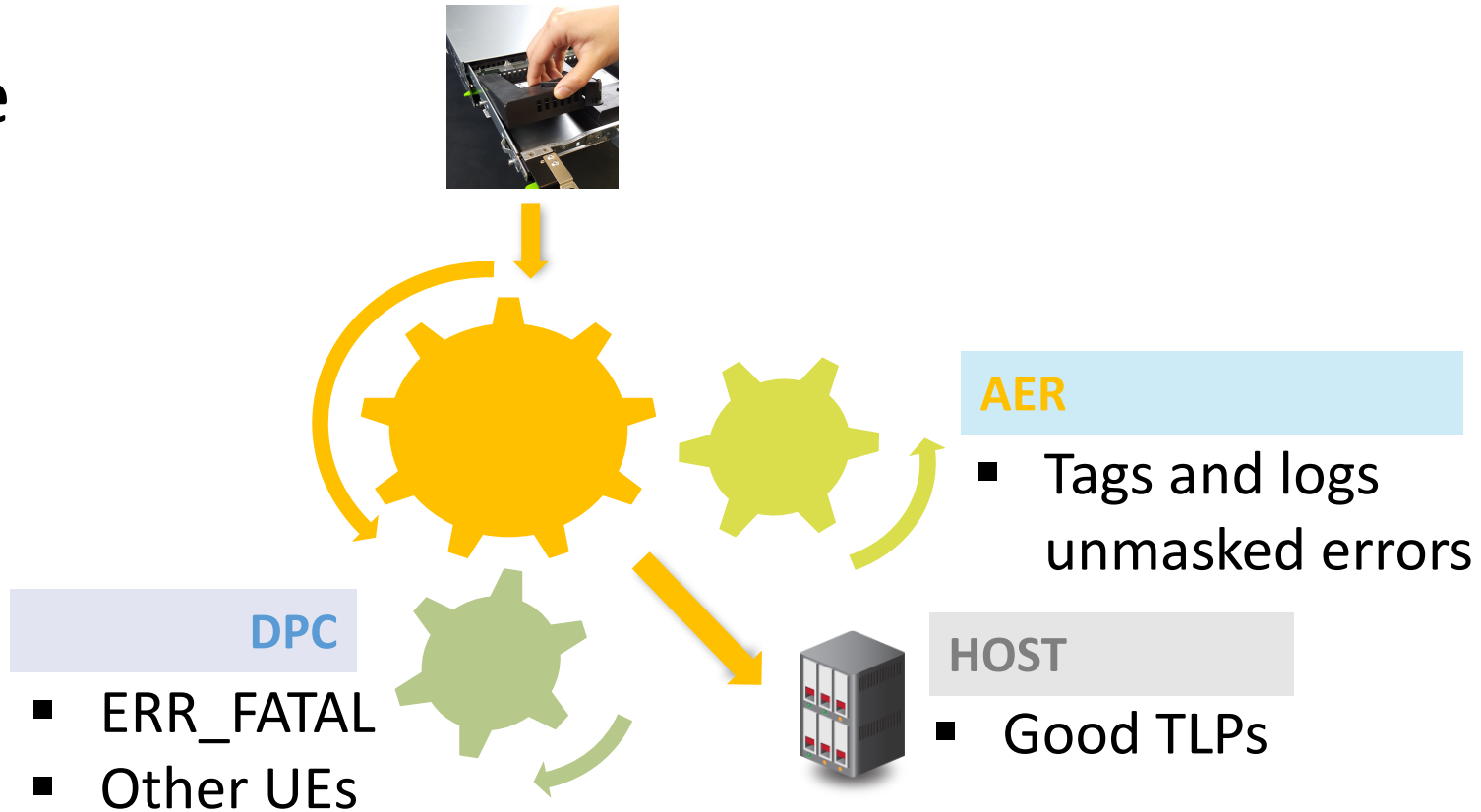
↑ After adding
All 1's support

←Before All 1's support

1000's of IO's reduced to ~20

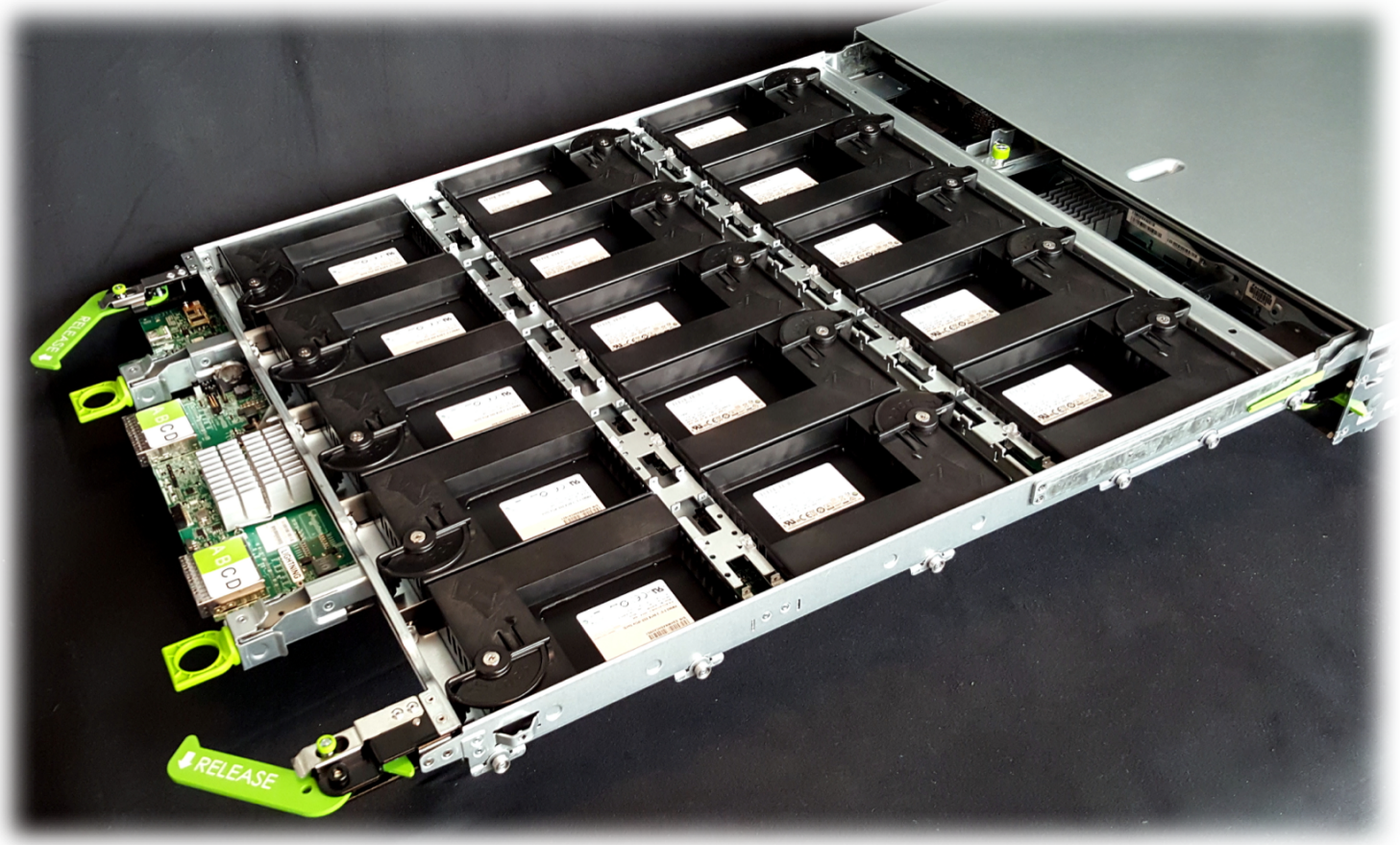
Challenges / Solutions: Downstream Port Containment (DPC)

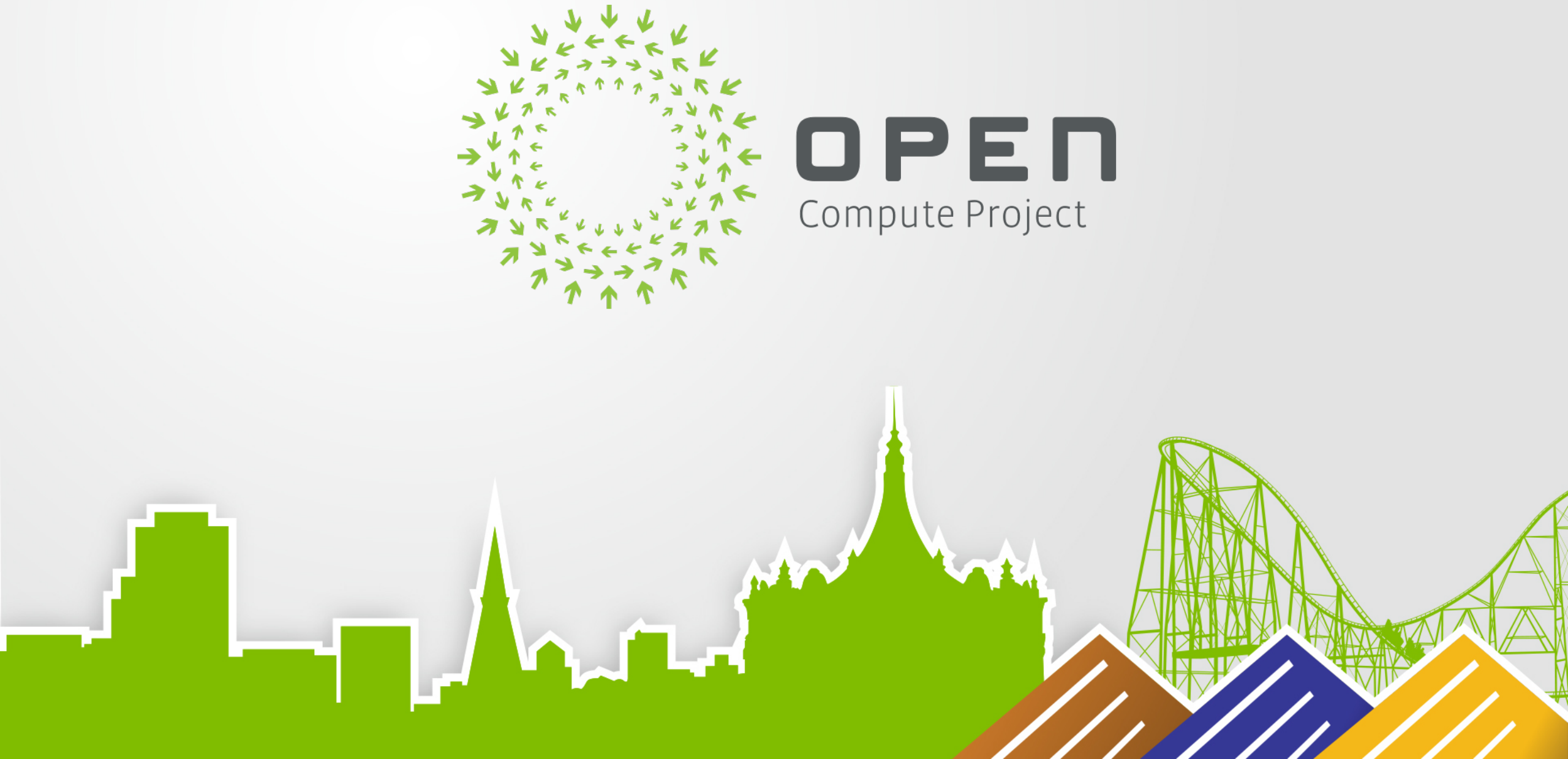
- Defined in PCI-SIG Base Specification 3.1
- Enabled by *new* DPC Driver (see slide 15)
- Allows for *ErrFatal+* to be supported in the host



PCIe JBOF Challenges Solved!

- ✓ M.2 support
- ✓ Enclosure management
- ✓ PCIe Hot-plug




OPEN
Compute Project