Enabling PIM Card Hot Swapping in SONiC for Edgecore Minipack

Deming Liu, Software Architect, Edgecore Networks

OPEN
PLATINUM™

OCP
REGIONAL
SUMMIT

Open. Together.

Agenda

- **Objectives**
- **Minipack introduction**
- **PIM hot swapping challenges in SONiC on Minipack**
- **Hot swapping implementation**
- **Experimental results**
- **Future works**

NETWORKING

Case Studies

Open. Together.

Objectives

- **Capture Minipack PIM card plug/unplug events in SONiC on Minipack**
- **Bring up/remove the ports accordingly in SONiC**

NETWORKING

Case Studies

Open. Together.

Minipack Introduction

- 20%+ switches in data centers are white box + open source software
- First OCP 400G port white box switch
- A new building block switch based on Broadcom Tomahawk 3 ASIC
- Support 2 types of port interface modules (PIM), 100G and 400G
- PIM card plug and play hardware enabled
- Smaller size
- Lower power

NETWORKING

Tomahawk 3

PIM Cards

Traditional Switch

Minipack Switch

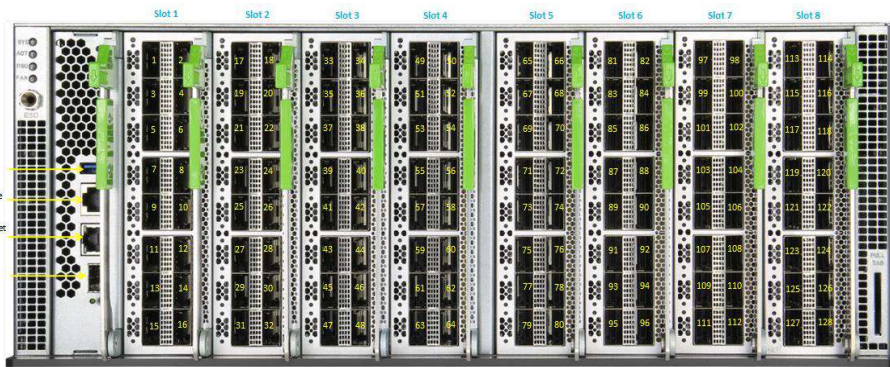Case Studies

Open. Together.

Minipack PIM Card Types

NETWORKING

- **2 types of PIM cards**
  - 128 X 100G, 12.8T switching bandwidth
  - 32 X 400G, 12.8T switching bandwidth
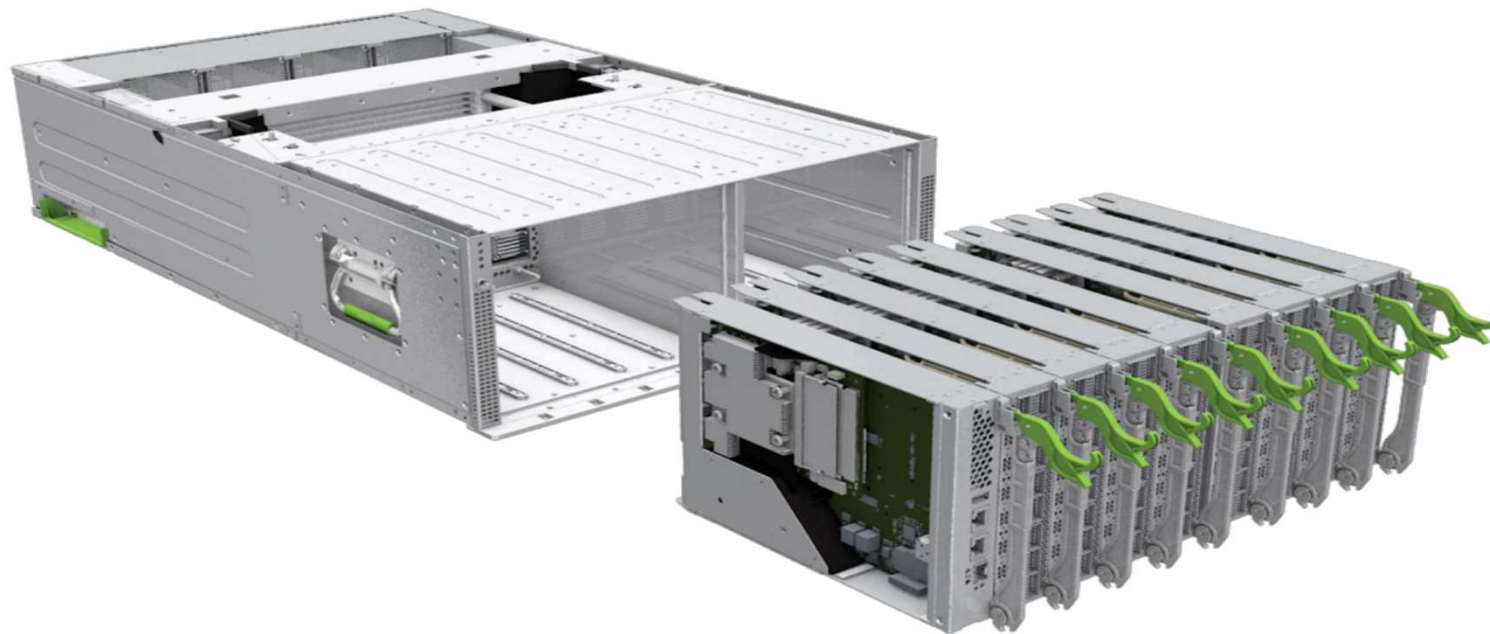

PIM-16Q 100Gb

PIM-4DD 400Gb

Case Studies
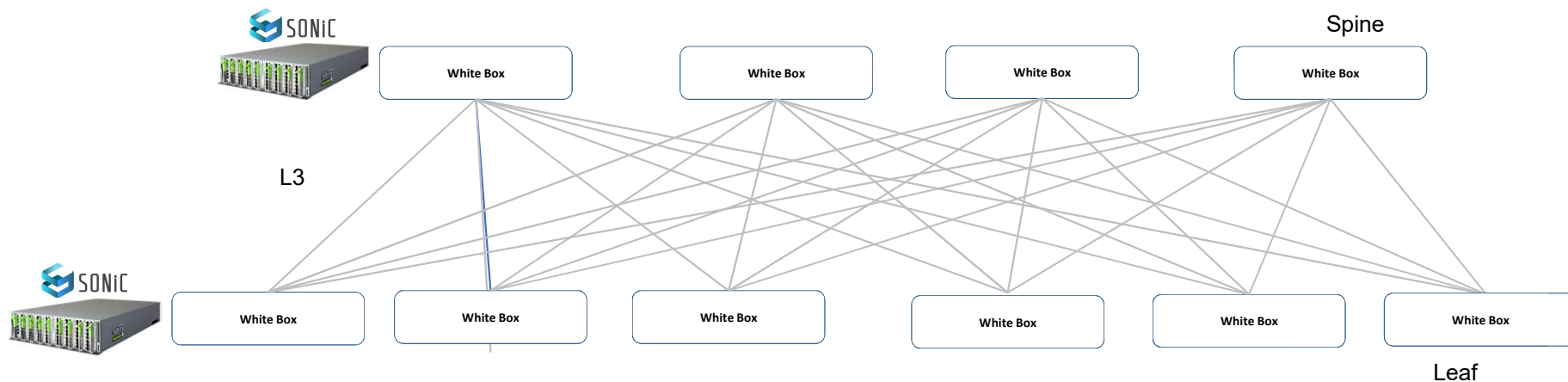
Minipack Plug and Play Support

NETWORKING

Case Studies

Minipack Applications

- **Can be used either in leaf layer or in spine layer**
  - Modular design
  - High port density

Spine

SONiC

| White Box | White Box | White Box | White Box |

L3

SONiC

| White Box | White Box | White Box | White Box | White Box | White Box |

Leaf

Case Studies
Case Studies

Open. Together.

PIM Card Hot Swapping Requirements

- **Even ported Minipack by Edgecore, SONiC doesn't support dynamically adding/removing ports**

- **SONiC reboot on Minipack takes quite a long time, several minutes**

- **Spine layer switches in data center need the flexibility of hot swapping**

Case Studies

OCP
REGIONAL
SUMMIT

Open. Together.

PIM Hot Swapping Challenges in SONiC on Minipack – Event Detection

- **Linux udev can't detect PIM probably because PIM cards are not directly attached on CPU bus**
- **Netlink doesn't capture PIM events (need kernel changes)**
- **Broadcom SAI/SDK has no PIM knowledge**
- **Polling IOB (Input/Output Block) FPGA for PIM card presence**
  - Other options, BMC (Baseband Management Controller)/EEPROM

Case Studies

Open. Together.

PIM Hot Swapping Challenges in SONiC on Minipack – Port Bring-up

- **Broadcom SAI doesn't yet support dynamic port adding/removing**
  - A static ASIC config file is loaded to configure all front panel ports when swss/syncd booting up
- **orchagent/portsorch in swss doesn't allow to dynamically add/remove ports for Broadcom ASIC**
- **Have to figure out a way to dynamically configure ports based on PIM card presence and types**

Case Studies

Open. Together.

Hot Swapping Implementation

- **Detect PIM card plug/unplug events**
- **Bring up/drop off ports of PIM cards**

NETWORKING

Case Studies

Detect PIM Card Plug/Unplug Events

- **Polling for presence of PIM cards**
- **A dedicated periodical daemon, pimd, polls IOB FPGA on PIM to get PIM presence status (plugged/unplugged) and types (100G or 400G)**
  - A kernel driver is implemented to access IOB FPGA
- **Download firmware to PIM card and call PIM initialization routine if cards are inserted**
- **Call Redis database sync routine to add the ports of PIM card config database**
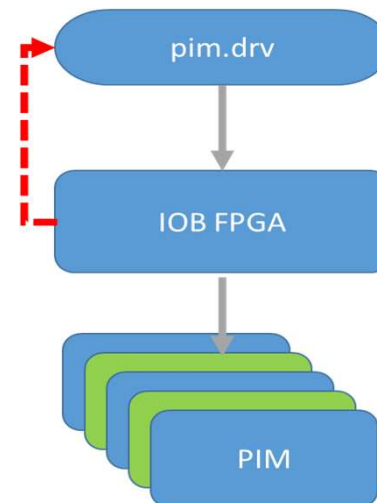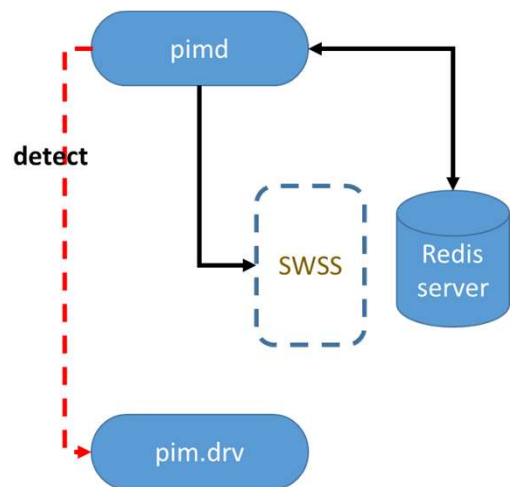 - Select a right lane set from lane mapping tables

NETWORKING

Case Studies

Detect PIM Card Plug/Unplug Events

- **Polling mode event capture**
  - Since platform dependency, pimd is put in platform monitor container (pmon)

Case Studies

Open. Together.

Dynamic Port Bring-up in SONiC

- **Modify Broadcom SAI 3.5.2.3 and make port add/remove APIs ready**
  - Change port breakout in SAI to support port add/remove
- **In swss, modify portsynd to make it handle port removing and orchagent/portsorch to allow calling SAI port add/remove APIs for Broadcom ASICs**
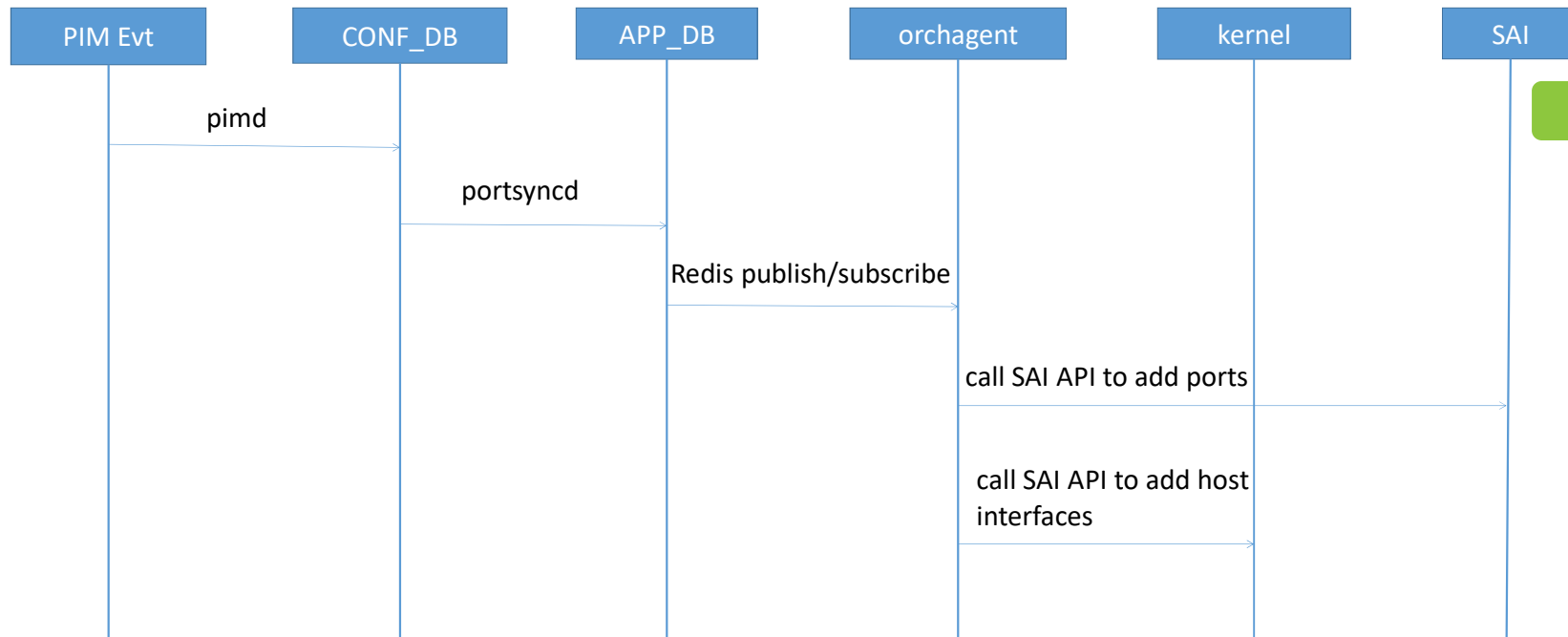
Case Studies

Open. Together.

Dynamic Port Bring-up in SONiC

PIM Evt	CONF_DB	APP_DB	orchagent	kernel	SAI

pimd

portsyncd

Redis publish/subscribe

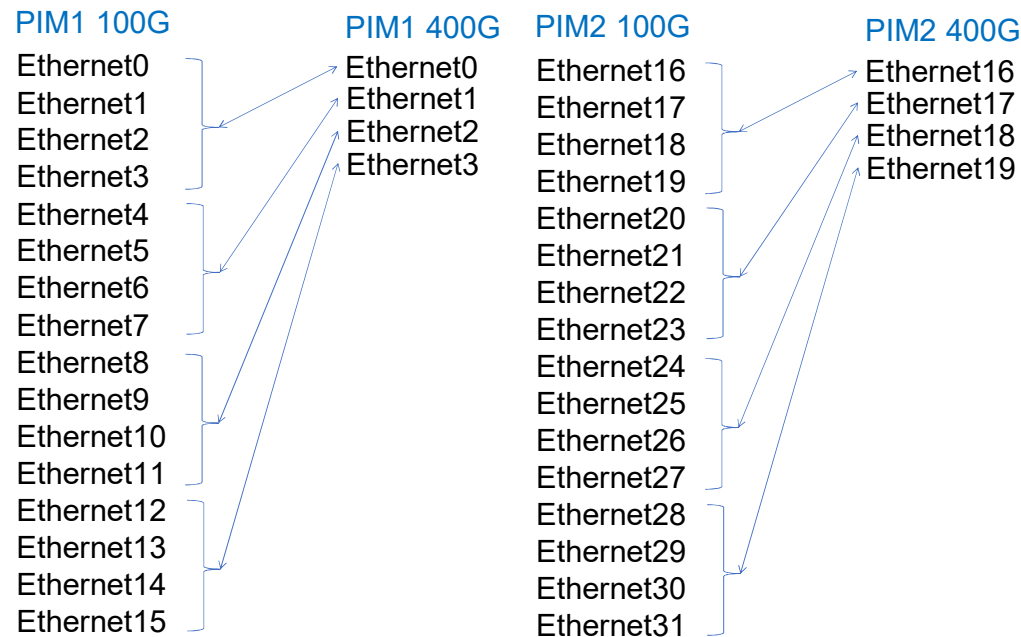call SAI API to add ports

call SAI API to add host interfaces

Case Studies

Lane and Port Name Mapping

- **Change port lane sets and names when PIM cards plugged/unplugged**
- **Lanes of every four 100G ports are mapped to one 400G port**

PIM1 100G
Ethernet0
Ethernet1
Ethernet2
Ethernet3
Ethernet4
Ethernet5
Ethernet6
Ethernet7
Ethernet8
Ethernet9
Ethernet10
Ethernet11
Ethernet12
Ethernet13
Ethernet14
Ethernet15

PIM1 400G
Ethernet0
Ethernet1
Ethernet2
Ethernet3

PIM2 100G
Ethernet16
Ethernet17
Ethernet18
Ethernet19
Ethernet20
Ethernet21
Ethernet22
Ethernet23
Ethernet24
Ethernet25
Ethernet26
Ethernet27
Ethernet28
Ethernet29
Ethernet30
Ethernet31

PIM2 400G
Ethernet16
Ethernet17
Ethernet18
Ethernet19

NETWORKING

Case Studies

Open. Together.

Experimental Results

- **PIM6 and PIM7 are present, Ethernet80-111**

```
admin@sonic:~$ show platform summary
Platform: x86_64-accton_minipack-r0
HwSKU: Accton-MINIPACK
ASIC: broadcom
admin@sonic:~$
admin@sonic:~$ ls /sys/class/net <====== Before unplugging card
bcm0        Ethernet102 Ethernet109 Ethernet84 Ethernet91 Ethernet98
Bridge      Ethernet103 Ethernet110 Ethernet85 Ethernet92 Ethernet99
docker0     Ethernet104 Ethernet111 Ethernet86 Ethernet93 lo
eth0        Ethernet105 Ethernet80  Ethernet87 Ethernet94 usb0
eth1        Ethernet106 Ethernet81  Ethernet88 Ethernet95
Ethernet100 Ethernet107 Ethernet82  Ethernet89 Ethernet96
Ethernet101 Ethernet108 Ethernet83  Ethernet90 Ethernet97
admin@sonic:~$ i2c i2c-130: delete_device: Deleting device tmp75 at 0x48
i2c i2c-131: delete_device: Deleting device tmp75 at 0x4b
i2c i2c-132: delete_device: Deleting device adm1278 at 0x10
i2c i2c-134: delete_device: Deleting device max34461 at 0x74

admin@sonic:~$
```

```
admin@sonic:~$ ls /sys/class/net <====== After unplugging PIM7
bcm0    eth1      Ethernet83 Ethernet87 Ethernet91 Ethernet95
Bridge  Ethernet80 Ethernet84 Ethernet88 Ethernet92 lo
docker0 Ethernet81 Ethernet85 Ethernet89 Ethernet93 usb0
eth0    Ethernet82 Ethernet86 Ethernet90 Ethernet94
admin@sonic:~$
admin@sonic:~$
admin@sonic:~$ lm75 130-0048: hwmon56: sensor 'tmp75'
i2c i2c-130: new_device: Instantiated device tmp75 at 0x48
lm75 131-004b: hwmon57: sensor 'tmp75'
i2c i2c-131: new_device: Instantiated device tmp75 at 0x4b
i2c i2c-132: new_device: Instantiated device adm1278 at 0x10
i2c i2c-134: new_device: Instantiated device max34461 at 0x74

admin@sonic:~$
```

NETWORKING

Case Studies

Experimental Results

```
admin@sonic:~$ ls /sys/class/net <====== After plugging PIM7 back
bcm0        Ethernet102  Ethernet109  Ethernet84  Ethernet91  Ethernet98
Bridge      Ethernet103  Ethernet110  Ethernet85  Ethernet92  Ethernet99
docker0     Ethernet104  Ethernet111  Ethernet86  Ethernet93  lo
eth0        Ethernet105  Ethernet80   Ethernet87  Ethernet94  usb0
eth1        Ethernet106  Ethernet81   Ethernet88  Ethernet95
Ethernet100 Ethernet107  Ethernet82   Ethernet89  Ethernet96
Ethernet101 Ethernet108  Ethernet83   Ethernet90  Ethernet97
admin@sonic:~$
```

Hot swapping demo with loading traffic is available at Edgecore booth B11

Welcome to visit us!

NETWORKING

Case Studies

OCP REGIONAL SUMMIT

Open. Together.

Future Works

- **Optimize PIM initialization process by using more parallel techniques**
- **Event driven solution detecting PIM card plug/unplug events such as udev**
- **Could employ dynamic port breakout to simplify the implementation**
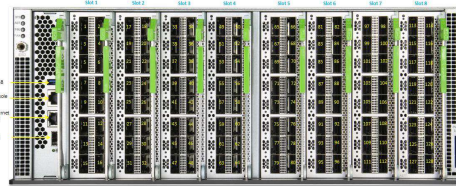
NETWORKING

Case Studies

Open. Together.

Product/Facility Info

- **Pictures of Product/Facility**

- **Links to SP's Site**
  https://www.edge-core.com
- **Marketplace Link**

  https://www.edge-core.com/productsInfo.php?cls=1&cls2=349&cls3=351&id=608

Call to Action

- **Port breakout support for Broadcom ASICs if not scheduled in SONiC community roadmap**
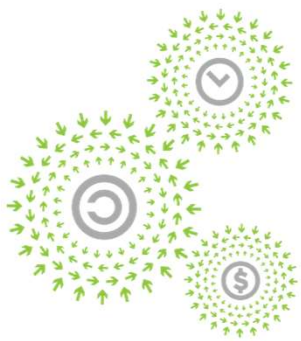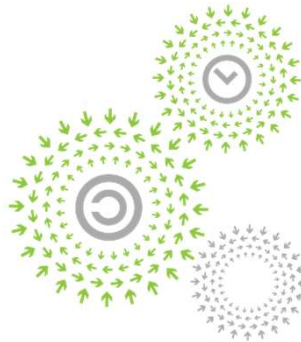- **Dynamic port add/remove support in Broadcom SAI**

  https://github.com/Azure/SONiC/wiki/Port-Breakout-High-Level-Design

  https://www.edge-core.com/productsInfo.php?cls=1&cls2=349&cls3=351&id=608
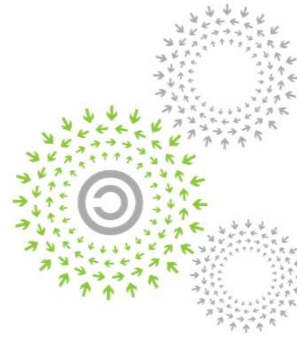
Please use one of these membership logo's to designate your company's membership level.

OPEN
PLATINUM™

OPEN
GOLD®

OPEN
SILVER™

OPEN
COMMUNITY®

Please use this logo if you or your supplier is an OCP Solution Provider.

Please use this logo if your Facility is an OCP Ready™ facility

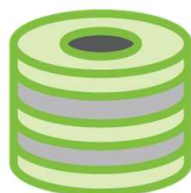Please use if your product has been recognized as an OCP validated product

OPEN
ACCEPTED™

OPEN
INSPIRED™

Please use the appropriate icon representing the Project Group

SECURITY	SERVER	STORAGE	NETWORKING	RACK & POWER
MANAGEMENT	HPC	TELCO	DATA CENTER FACILITIES	OPEN SYSTEMS FIRMWARE

The following project group logos are missing: OpenEdge, OpenRMC, ACS. If you need one of these, contact Archna@opencompute.org

Please use the appropriate icon representing the Regional Project Group

CHINA
Regional
Community

JAPAN
Regional
Community

TAIWAN
Regional
Community

KOREA
Regional
Community

Please use the appropriate icon representing your type of contribution

Specifications

Reference
Architecture

Embedded
Software

Tested
Configurations

Case Studies

White
Papers

Design Files

Product
Recognition

Facility
Recognition

Workshops
Summits

Testimonials
Seminars

Videos