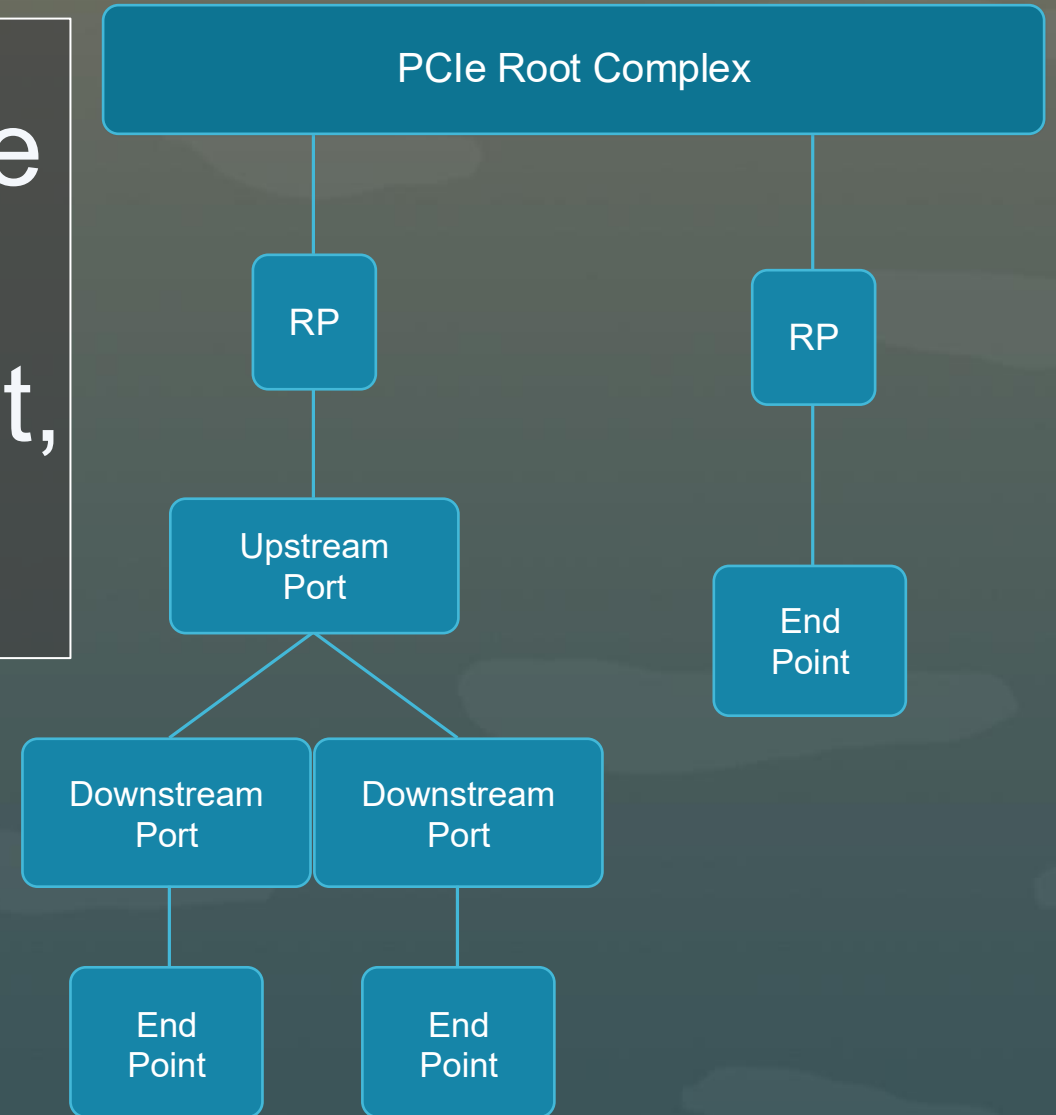


Two Years of Making PCIe “Hotplug” Reliable: What Landed, What Didn't, and Where Do We Go

** "Hotplug" => topology changes while the system is live*



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

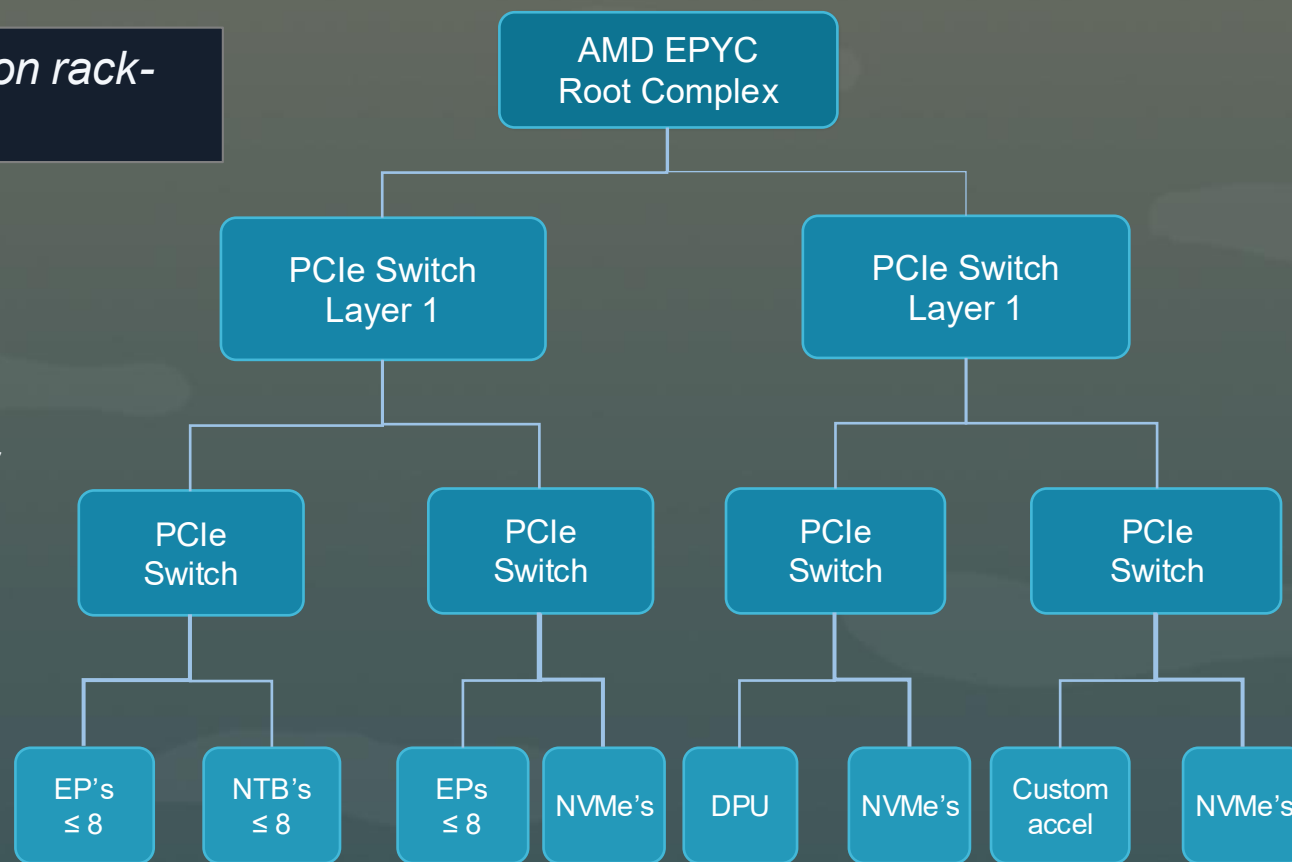
Maciej Grochowski · SONY SIE · LPC 2026

Current production generation

PCIe is already a rack-scale fabric in service

Today's rack already uses PCIe as a production rack-scale fabric, not only as local server I/O.

- 100+ PCIe endpoints per rack
- 2–3 switch layers
- Hot insertion and removal are supported
- Peer-to-Peer enabled between endpoints
- Switch firmware configuration is under our control
- Pain point: recovery and isolation at scale



100+ endpoints / rack

Main challenge
hotplug · error handling · recovery · FW issues (isolation)



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

Next-generation direction

Gen6 density turns rare PCIe edge cases into normal operations

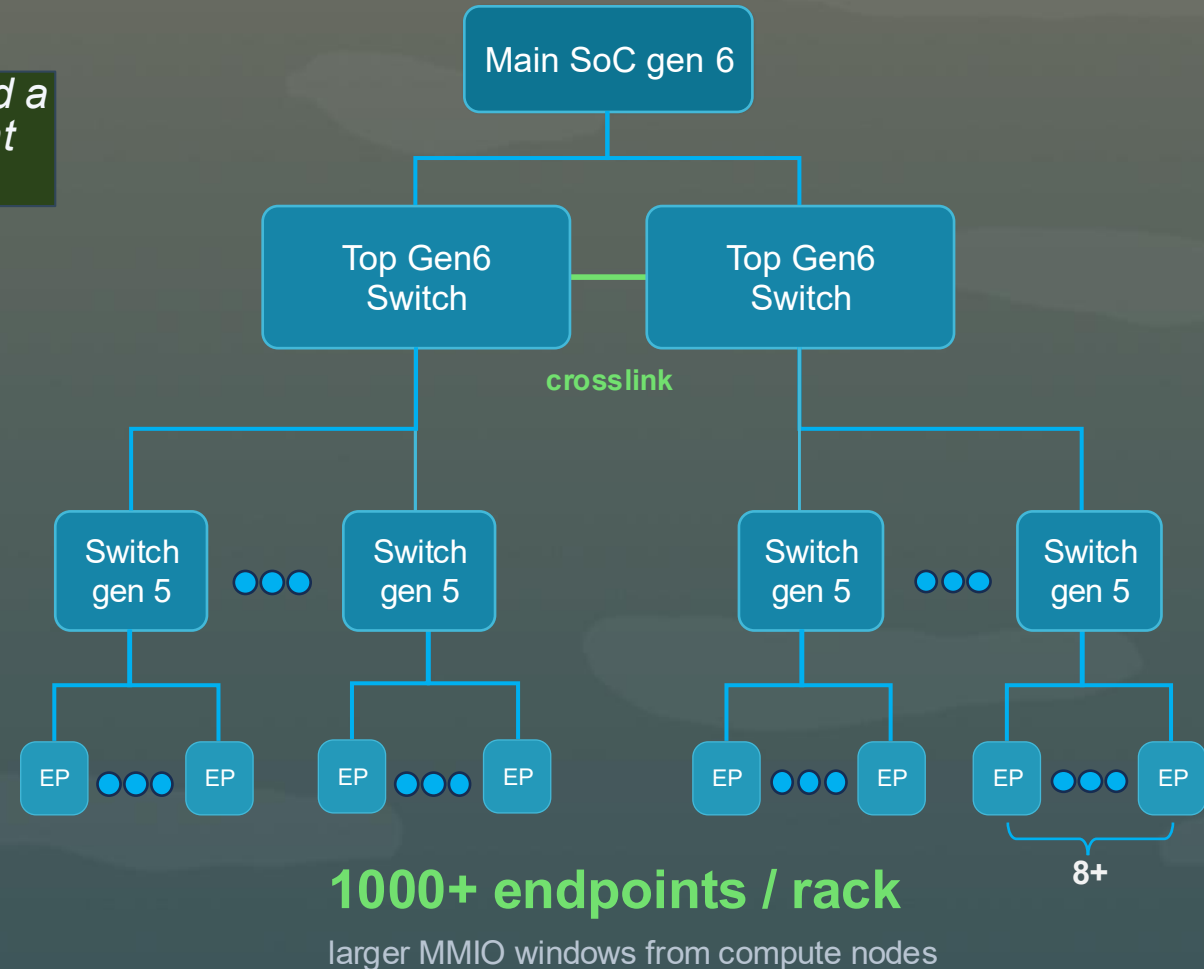
Cost and scaling constraints are pushing the rack toward a much larger PCIe fabric with significantly higher endpoint density.

- PCIe Gen6 switches
- Multiple types of EndPoints (EP): Storage (NVMe), Compute Modules, DPU's, NTB's
- Endpoint density grows by roughly an order of magnitude
- Larger PCIe MMIO exposure from compute nodes
- More peer-to-peer and device-to-device traffic
- New bottleneck: routing, isolation, and recovery semantics



LINUX
PLUMBERS
CONFERENCE

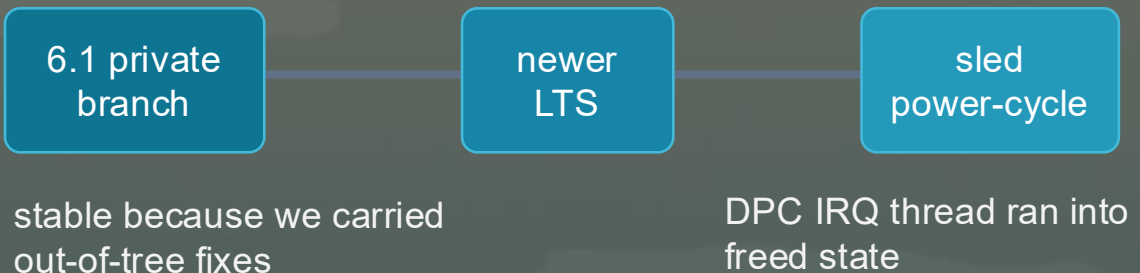
PRAGUE, CZECHIA
OCT. 5-7, 2026



The topology is becoming a fabric challenge,
not a server-I/O problem.

Where we were — and what the LTS update cost us

Enumeration was clean; DPC recovery exposed a lifetime bug



```
BUG: kernel NULL pointer dereference @ 0x10
RIP: pci_dev_wait+0x45
    pci_bridge_wait_for_secondary_bus
    dpc_reset_link → pcie_do_recovery →
dpc_handler
kernel tried to execute NX-protected page
"Fixing recursive fault but reboot is needed!"
```

2024 state

- private branch
- hotplug fixes carried locally
- PCIe-side stability good enough for service

After LTS move

- enumeration and hotplug looked clean
- DPC recovery hit UAF
- upstream already had parts of the fix

**The crash was easy to read.
The cause sat three layers down.**

- backport gap, not a new discovery
- correct commit looked like the culprit
- field mitigation re-armed a related bug



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

What upstream fixed since 2024

The good news is real: the point fixes landed

v6.11		11a1f4bc4736	Child UAF in pci_dev_wait	Lukas Wunner
v6.16		2af781a9edc4 + 2	Suppress reset/DPC link events in pciehp	Lukas Wunner
v6.18		45bc82563d55 + 3	Recovery state-machine rework	Lukas Wunner
v7.0		3958bf16e2fe	PCI: Stop over-estimating bridge window size	Ilpo Järvinen
v7.1		a1ed752bc7cb	DPC holds a port reference	Sizhe Liu
		8238cb69c01f	reset_subordinate becomes hotplug-safe	Keith Busch
		9036bd0efcb6	PCI: Align head space better	Ilpo Järvinen
v7.3		97ca178c899d	DPC decoupled from AER driver binding	Darshit Shah



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

What is still open?

Four tracks, each at a different maturity.

TRACK	WHERE IT IS NOW	NEXT DECISION
MMIO placement PATCH v2 - TESTING	General placement and composite-sizing series is under cross-machine testing.	Does it fix the Sony 8-endpoint case without moving a regression elsewhere?
Maximum Payload Size RFC - BUILD-TESTED	Scan-time hierarchy programming is reworked and the known gaps are listed.	What is the policy for hot-add below a switch that is already in use?
Delayed nested subtree RFC - HW-TESTED	Per-bridge reservations work on the tested two-level Switchtec topology.	How should three or more switch layers describe future resource demand?
Locking and lifetime RFC - DESIGN OPEN – NOT ALL CONCERNS ADDRESSED CURRENTLY	Referenced snapshot fixes the reproduced bus-reset race; topology ownership remains open.	Snapshot every reset path, or exclude topology mutation for the whole reset? Snapshots close races; we still need a subsystem-wide topology-ownership contract.

PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

MMIO: from local workaround to general series

The upstream exit path is already happening.

FIELD FAILURE

**Raw sum fits.
Placement does not.**

8 children x 20 MiB
16 MiB alignment

Budget: 160 MiB
Observed span: 244 MiB
Result: only 5 / 8 BARs



LOCAL STOPGAP

max () -> ALIGN ()

Restores 8 / 8 on the fixed Sony topology.

Useful evidence, but it can reserve too much for simpler cases.



GENERAL FIX

v2 - seven patches

Place resources at either edge.
Fix composite sizing.
Honor alignment overrides.

Cross-machine testing is underway.

SONY 8-ENDPOINT VALIDATION: PENDING / IN PROGRESS

Evidence -> public review -> a fix broader than either local workaround.



LINUX
PLUMBER
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

MPS: safe before bind, hard after bind

The RFC repairs scan-time policy; live hot-add below an active switch remains open.

**Root Port
MPS 512**



**Switch path
MPS 512**



**New endpoint
MPSS 128**

PATH MISMATCH -> MALFORMED TLP RISK

RFC FIXES - BEFORE DRIVERS BIND

Update saved MPS state

Configure the Root Port after the hierarchy is scanned

Lower an unused hierarchy as one unit

STILL OPEN - ACTIVE HIERARCHY

Hot-add below a live switch cannot safely lower MPS

Bound drivers may cache MPS or have DMA in flight

Policy: warn and bind - or leave the device unbound?

MPS is a path-wide contract. Live repair needs hierarchy-wide quiesce.



**LINUX
PLUMBER
CONFERENCE**

PRAGUE, CZECHIA
OCT. 5-7, 2026

Nested hotplug: the child arrives late

Linux allocates the branch before its future demand is visible.

T0 - HARD RESET

Both switches disappear

The top switch and the nested switch are removed from the PCI tree.



T1 - PARENT FIRST

Empty ports are sized now

Five hotplug ports are empty.
The selected branch receives:
3 buses + about 206 MiB.



T2 - CHILD LATE

The subtree no longer fits

Needs 11 buses and 1032 MiB aligned to 1 GiB.
The Switchtec function cannot bind.

RFC RESULT

pci=hypreserve= reserves bus numbers and aligned MMIO for one selected empty bridge.

Cold boot, two top-switch resets and one nested-switch reset passed on the two-level system.

OPEN DESIGN

For 3+ layers, who describes future demand?

Kernel command line, firmware, platform data - or PCI core policy?

The failure is timing: resources are committed before the subtree exists.

Locking: snapshot the reset set

The RFC fixes the reproduced race; topology ownership is still undecided.

TWO MUTABLE WALKS



Device removed between walks
locked set != unlocked set

ONE REFERENCED SNAPSHOT



Same objects, same order
lifetime held through reset

OPEN DESIGN

Extend snapshots to every reset path - or exclude topology mutation for the whole reset?

The targeted repair is backportable. The topology contract is still open.



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

Discussion / Questions

What would let us to not apply next private PCIe patch?

1

How can we get more PCIe fixes upstream instead of carrying private workarounds?

What needs to change in review, ownership, hardware access, or validation?

2

What evidence turns a local workaround into an upstreamable fix?

Reproducer, second topology, traces, regression coverage, or fleet testing?

3

Which PCIe hotplug or recovery issue should we work on next?

Where would maintainer direction and a second validation platform help most?

GOAL: one priority · one next action · one follow-up owner

Sony can provide: Switchtec hardware · reproducible reset/hotplug workloads · traces · KASAN/lockdep · fleet validation



**LINUX
PLUMBERS
CONFERENCE**

PRAGUE, CZECHIA
OCT. 5-7, 2026

Thank you!



**LINUX
PLUMBERS
CONFERENCE**

PRAGUE, CZECHIA
OCT. 5-7, 2026

Keith's 2024 bus-locking branch

The prep merged; the lifetime and locking guarantees did not

Pre-2024 design problems

8-patch branch, base v6.11-rc1

Upstream outcome

#	Problem	kbusch commit	Upstream?
1	pci_stop_dev() not safe under concurrent teardown	9c202911a2ae	merged v6.13
2	pci_destroy_dev() not safe under concurrent teardown	77926252efe6	merged v6.13
3	walk-bus lock held too widely	1e1c99339cb1	merged v6.13
4	__pci_walk_bus() iterative, fragile	ca06c468cf25	merged v6.13
5	pci_walk_bus_locked() needlessly exported	280e759c5ef1	merged v6.13
6	teardown open-coded in 4 places	7d9b713529f2	never merged
7	subordinate read under one lock, freed under another	3883c485d5e4	never merged
8	one global lock deadlocks on concurrent removals	8c90ad32ddbc	never merged

Merged 1–5 change no locking semantics.

The three that carry the real lifetime and locking guarantee never landed.

LINUX
PLUM
CONF

PRAGUE, CZECHIA
OCT. 5-7, 2026

The arc: bisect → v6.18 → the real fix

The bisect was right, the stopgap was evidence, and the real fix was decoupling

1 · Bisect

d8d2b65a940b
“Allow AER service only for
Root Ports & RCECs.”

Correct patch; bad side effect:
DPC was gated on per-device
AER-service bit, so switch
Downstream Ports silently lost
native DPC.

2 · v6.18 check

**Crashes patched,
root coupling still standing**

v6.18 is a practical baseline, not
the cure. The DPC gate was still
AER-service-coupled.

3 · Real fix

97ca178c899d
Gate DPC on host->native_aer

Decouple ownership of AER from
whether an AER driver binds the
port. Re-arms native DPC on
switch Downstream Ports.

```
- (pcie_ports_dpc_native || (services & PCIE_PORT_SERVICE_AER)))  
+ (pcie_ports_dpc_native || host->native_aer))
```

In one line: the fix was a decoupling.



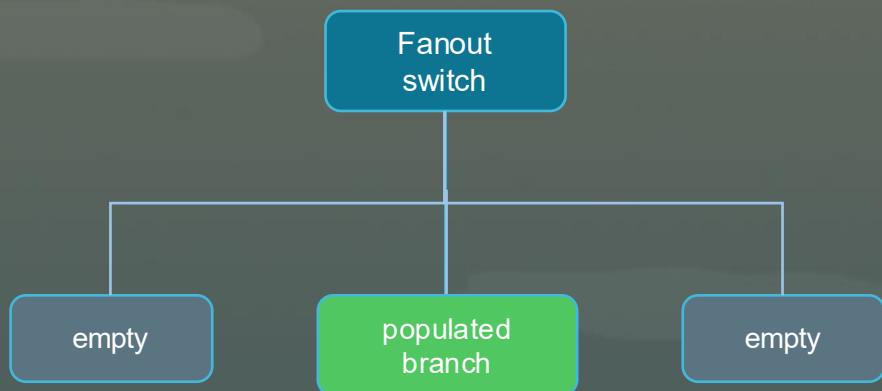
**LINUX
PLUMBERS
CONFERENCE**

PRAGUE, CZECHIA
OCT. 5-7, 2026

The other track: two MMIO allocation bugs

Topology is sound; Linux asks the parent for too small a window

Issue A · equal-split starvation



Linux split ~1032 MiB equally across 5 hotplug siblings: ~206 MiB each. Populated branch needs the full 1032 MiB for a 1 GiB-aligned BAR.



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

Root cause: equal distribution, not need-based sizing.

Issue B · alignment-gap undersizing



```
sum: 8×20 = 160 MiB  
real: 8×32 = 256 MiB  
3 of 8 endpoints get no BAR
```

```
Stopgap: size += ALIGN(r_size, align)
```

Both bugs scale against Gen6 endpoint density.