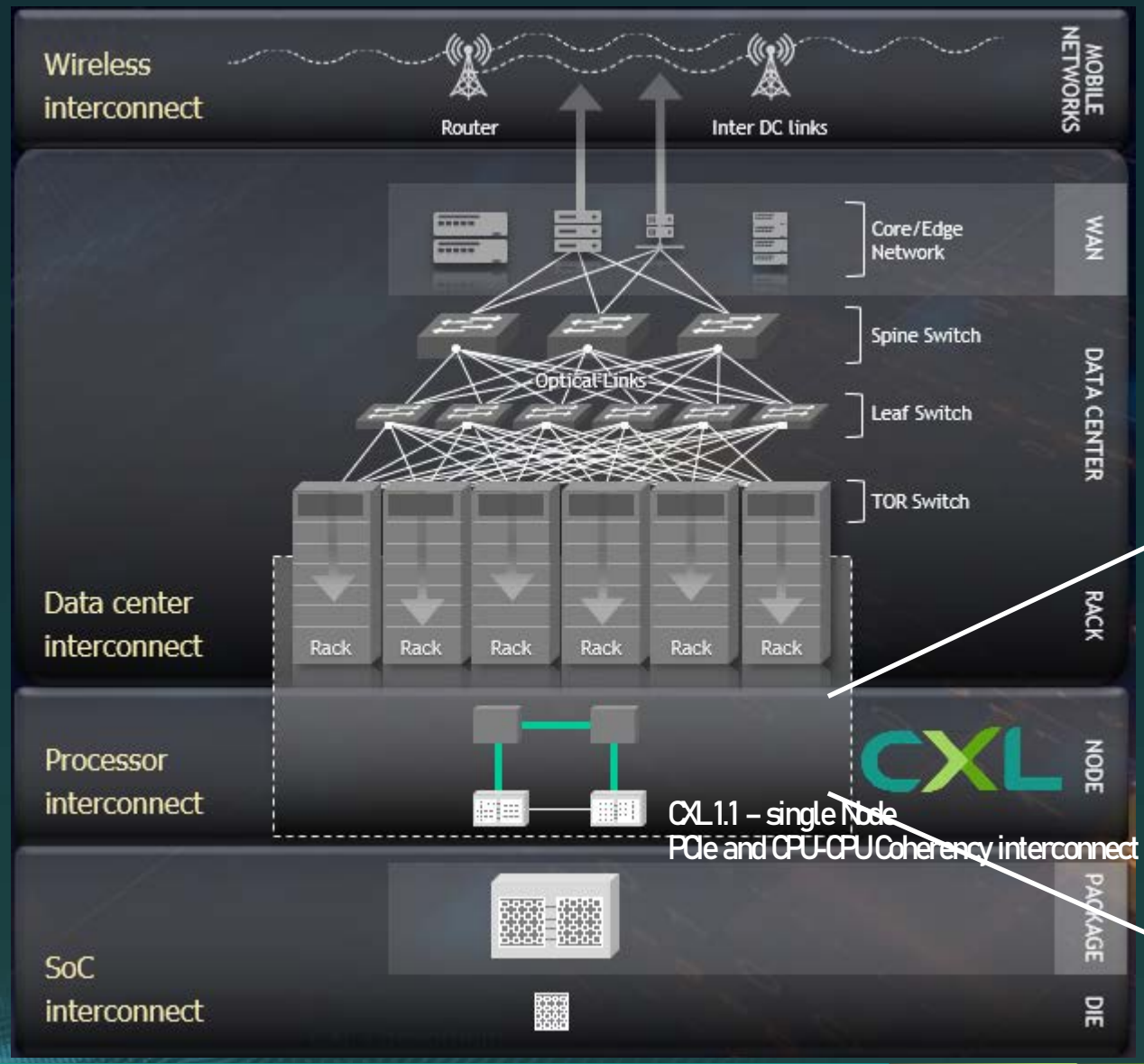
CXL 2.0 Switch for a Composable Memory System

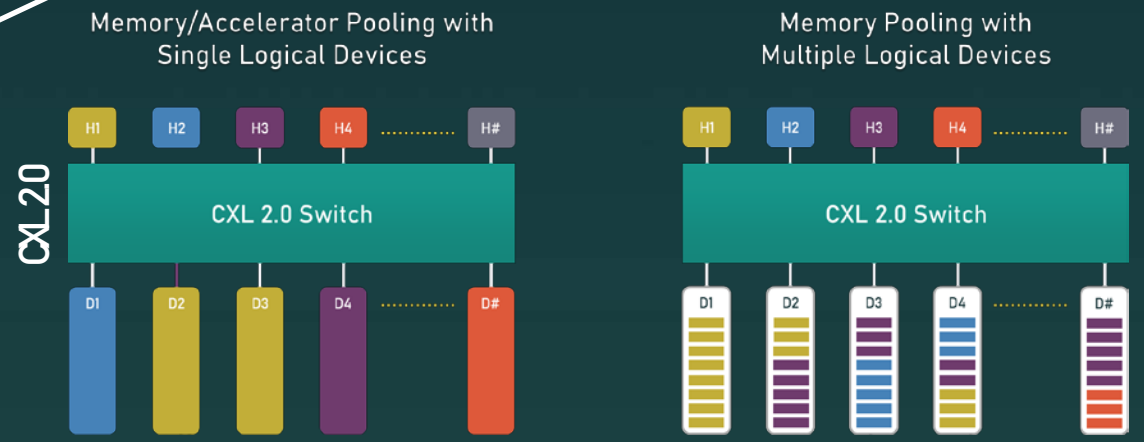
Presented by: Jim Kao, Xconn Technologies

Agenda

- Data Center Composable Memory System
- CXL 2.0 Switch for Memory Pooling/sharing
- Hot Plug for CXL 2.0 Switch
- Use cases for CXL 2.0 Switch
- Linux Ecosystem for CXL 2.0 Switch
- Wrap up

Composable Memory System Scope of CXL 2.0 over CXL 1.1

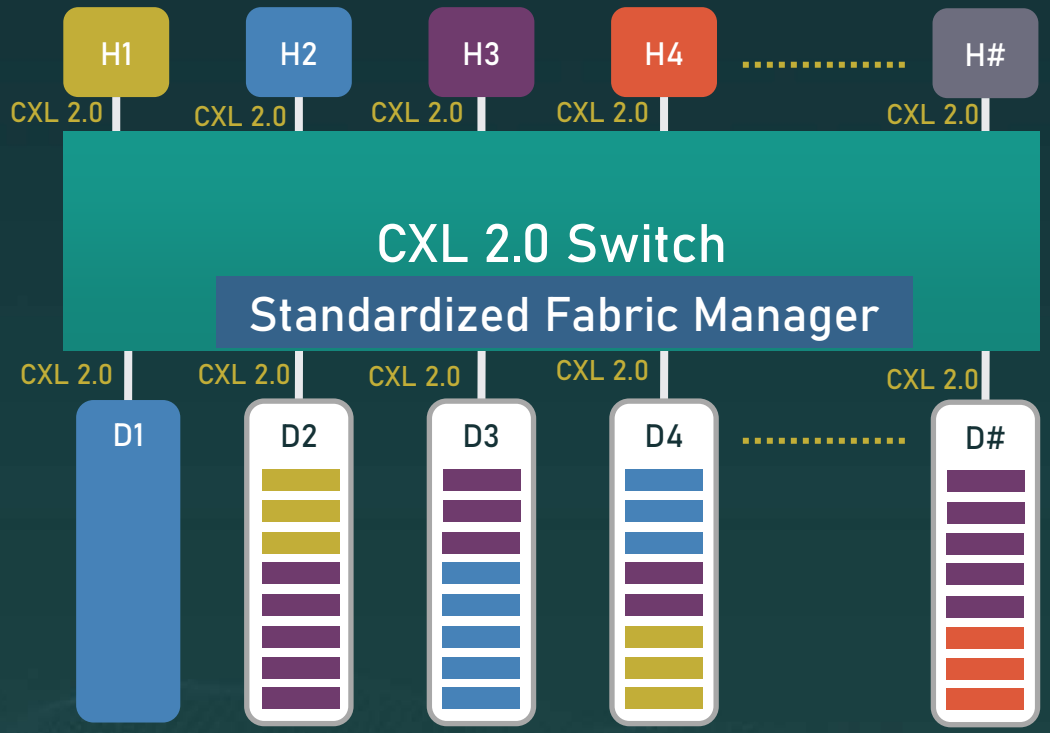




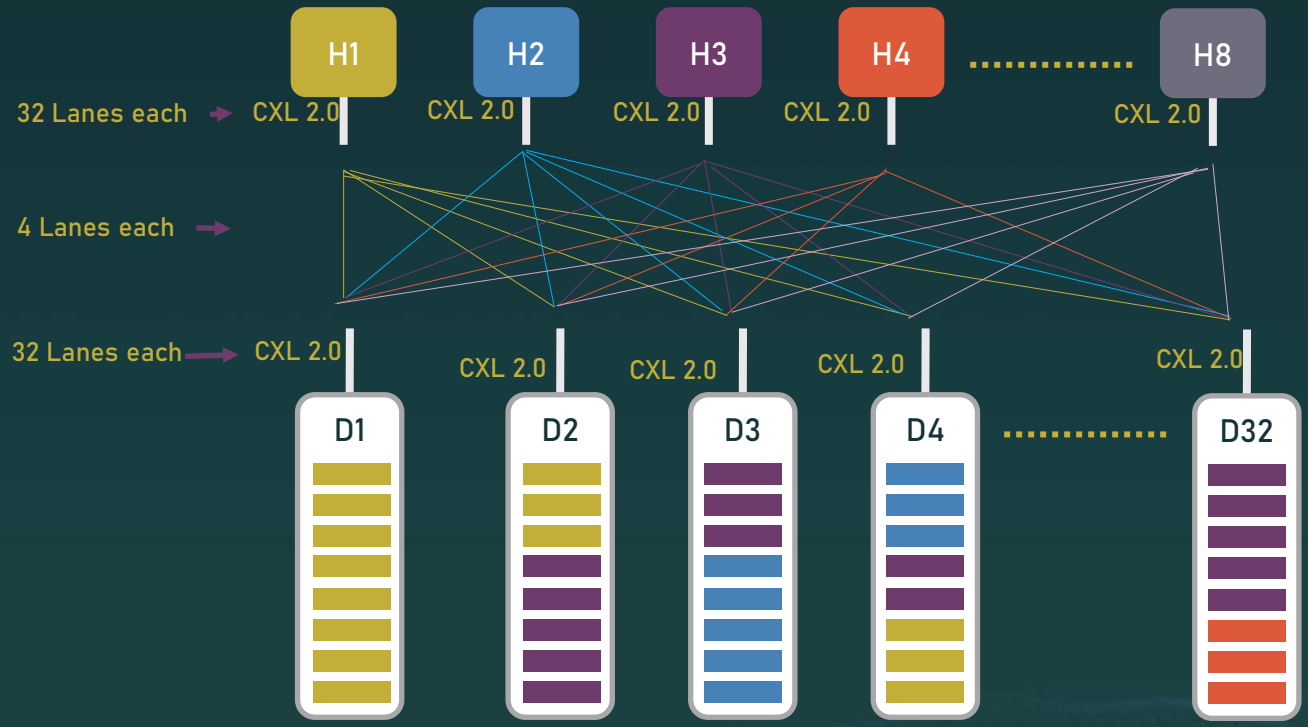
CXL 2.0 across Multiple Nodes inside a Rack/ Chassis supporting pooling of Composable resources

CXL 2.0 Pooling with vs without Switch

Memory Pooling with Single/ Multiple Logical Devices



Memory Pooling with Multiple Logical Device through direct connect



XConn's CXL 2.0 Switch

World's First CXL2.0 (XC50256) & PCIe 5.0 (XC51256) switch IC

2,048 GB/s total BW with 256 lanes

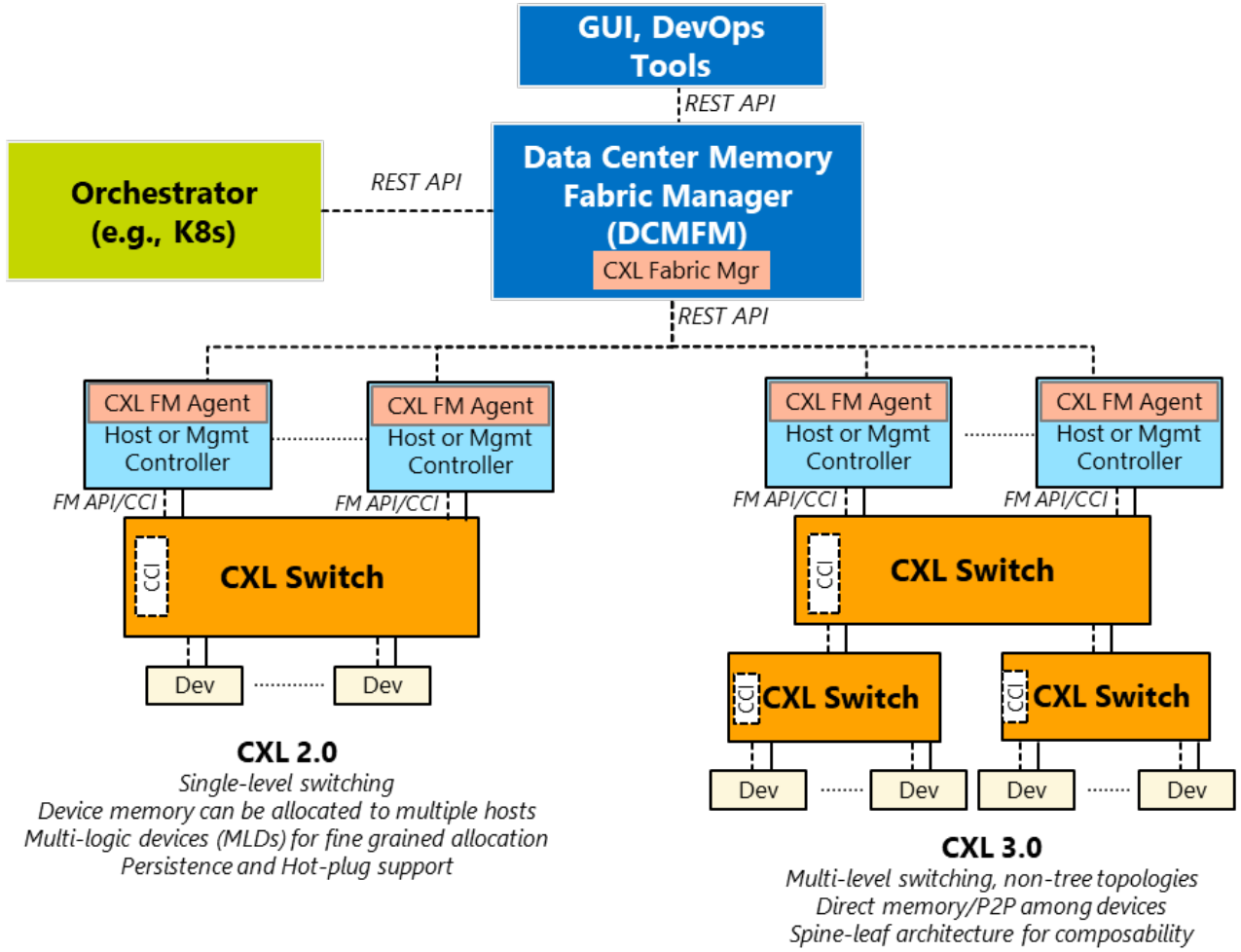
Lowest port-to-port latency

Lowest power consumption/port

Reduced PCB area
Lower TCO

- Works in CXL 1.1 and CXL 2.0 Configurations
- Supports CXL Memory Expansion/Pooling/Sharing
- Works in hybrid mode (CXL/PCIe mixed)
- MP version silicon CS (customer samples) now, MP Sept'24

Data Center Fabric Manager (DCMFM) Plans



- Agent based architecture
  - Fabric Manager(FM) agent software runs on host or management controller
  - Responsible for converting REST API to CXL protocol (MMIO/CCI)
  - High availability architecture with fail-over capability
- Supports CXL 2.0 and CXL 3.0 Fabric topologies
- DMTF Redfish profiles for device management
- CXL.IO/CCI, etc. for time sensitive mgmt. operations
- CMS focus – fabric manager reference implementation

Legend
——— ***Data plane*** *(cxl.mem/cxl.cache)*
------- ***Control plane*** *(cxl.io, FM API/CCI, MCTP/SMBus/I2C, REST etc.)*

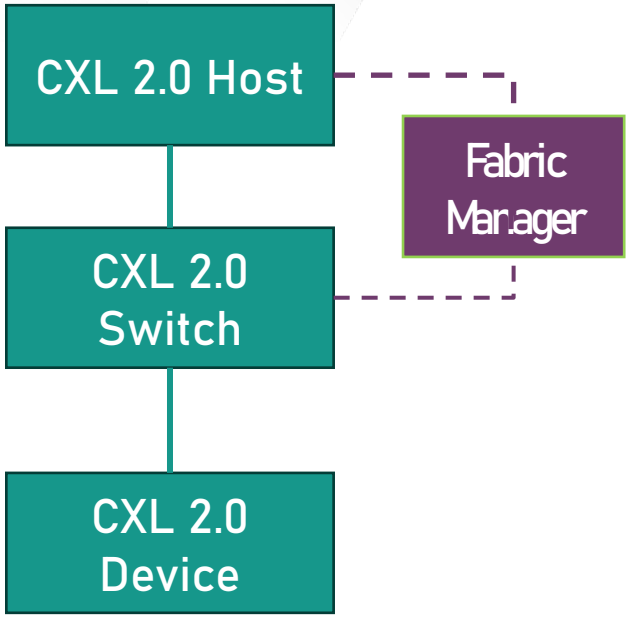
CXL 2.0 Hot Plug

- CXL 2.0 adds support for hot-add and managed hot-remove
  - Supported by all components - hosts, switches and devices
  - Managed hot-remove implies software is able to prepare the system for removal of the device
    - flush caches
    - offline pages mapped to HDM
  - Surprise removal is not supported and will result in unpredictable behavior
  - Capacity add and managed capacity removal are also supported
- Leverages Hot-plug model and Hot-plug elements as defined in PCI Express Specification and appropriate Form Factor Specification

Hot Added for Type 3 Device

When a Device is Hot-Added to an unbound port on a Switch

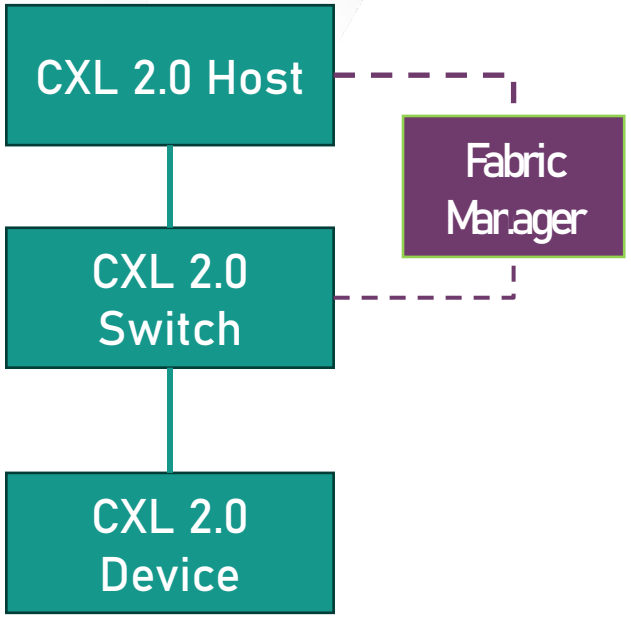
- The CXL 2.0 Switch notifies the FM by generating Physical Switch Event Records as the Presence Detect sideband signal is asserted and the port links up.
- MGMT FM issues the Get Physical Port State command for the physical port that has linked up to discover the connected device type.
- The MGMT FM can now bind the PPB to a vPPB.
- After completion of the binding process, the FM needs to notify Host using Managed Hot Add process for SLD or PCIe Device.
- Linux on Host
  - Create CXL region.
  - Add memory blocks to Linux kernel
  - Online memory block



Managed-Removed for type 3 device

When a Device is Managed-Removed from a bound port on a Switch

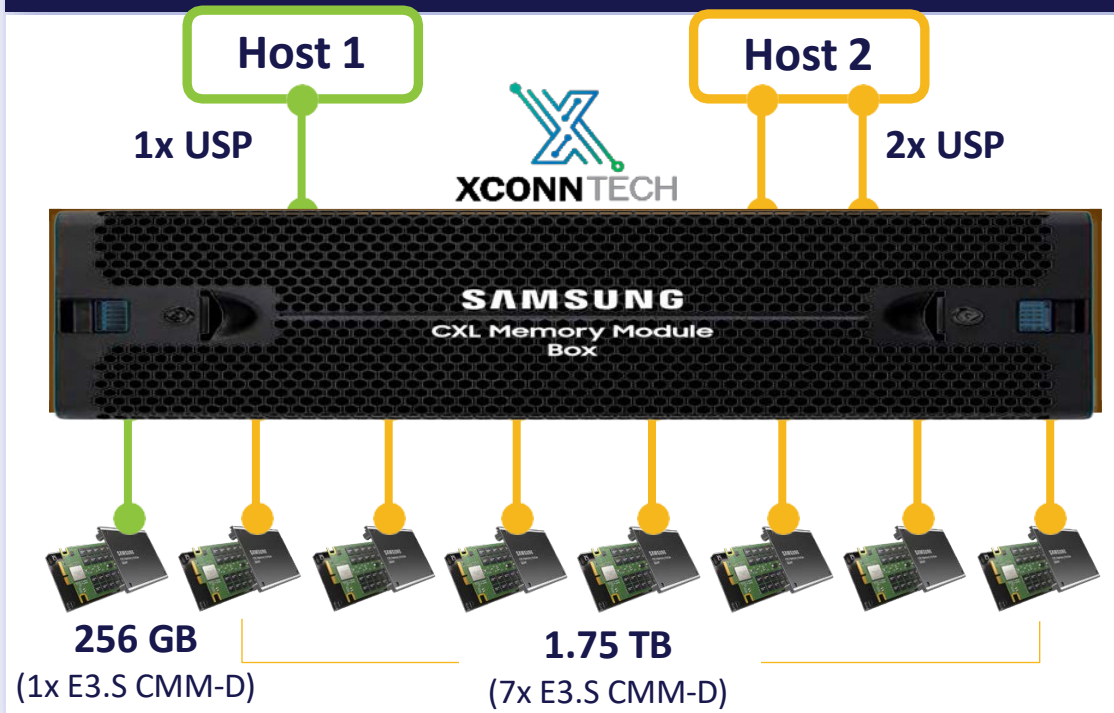
- When the Attention Button sideband is asserted.
- The switch notifies the FM and hosts with the MSI/MSI-X interrupts assigned to affected vPPB and a Virtual CXL Switch Event Record is generated.
- The host remove the memory blocks/region(s) on the device from the Linux kernel.
- When a host is ready for CXL device be removed, it will set the Attention LED bit in the associated vPPB's CSR to indicate CXL device can be physically removed.
- Removed the CXL device.
- The CXL 2.0 Switch notifies the FM by generating Physical Switch Event Records as the Presence Detect sideband is de-asserted and the associated port links down.
- After completion of the unbinding process, the FM notifies Host using Managed Hot Removal process.



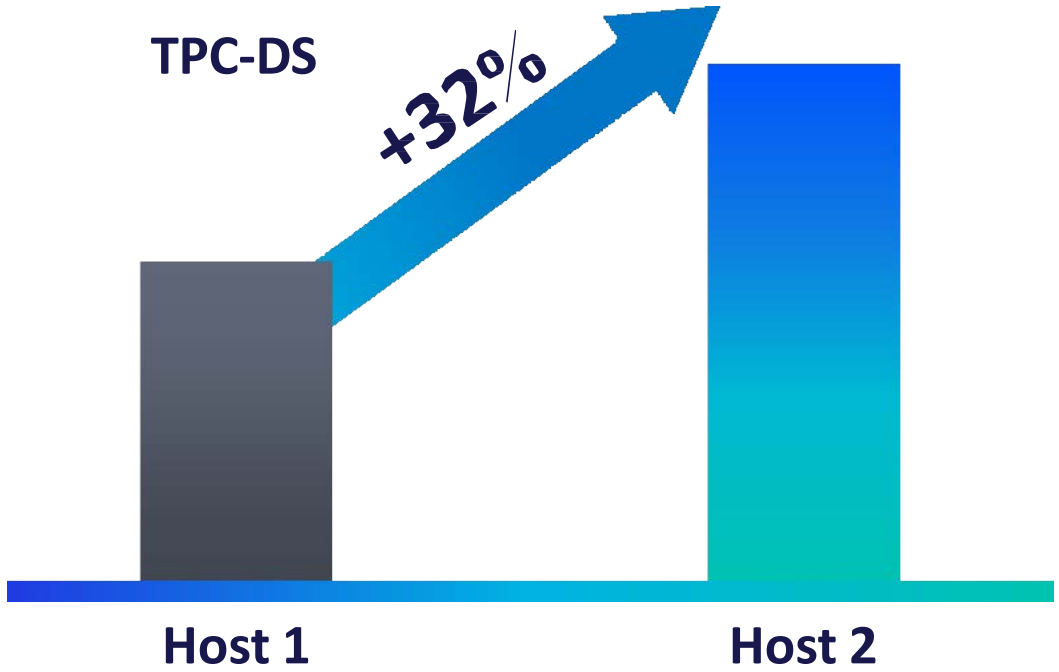
In-Memory Database in Action

- Dynamic CXL memory allocation & on-demand performance scaling
- 3 GEN5x8 USP and 8 GEN5x8 are used out of total 32 GEN5x8 port.

SAP HANA Test Scenario

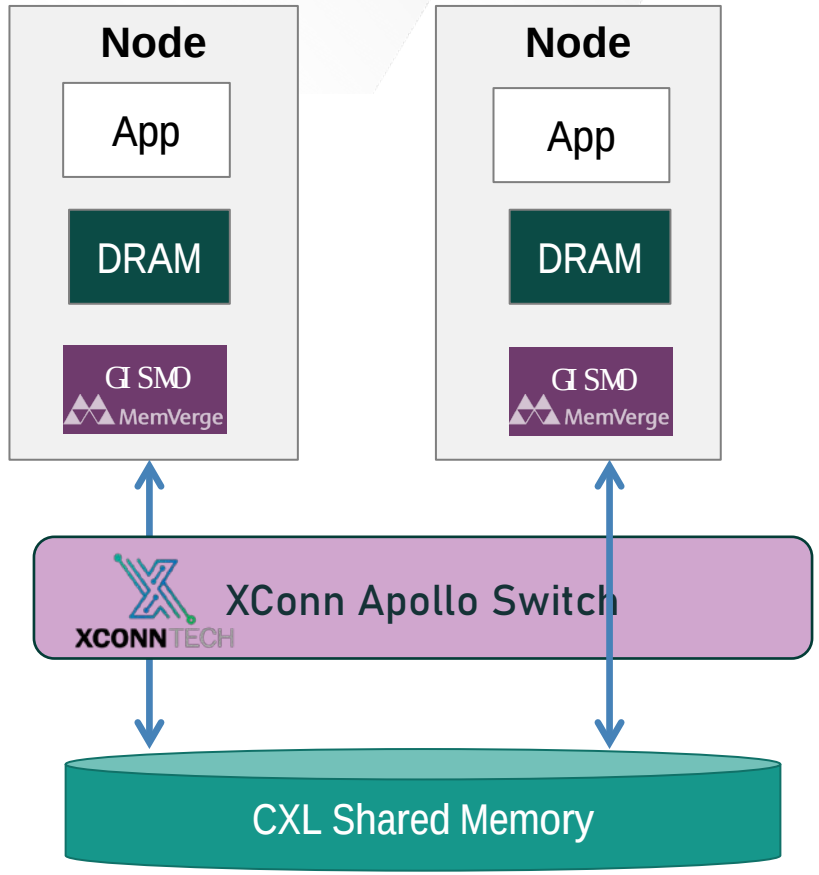


Benchmark Result (SQL/min)

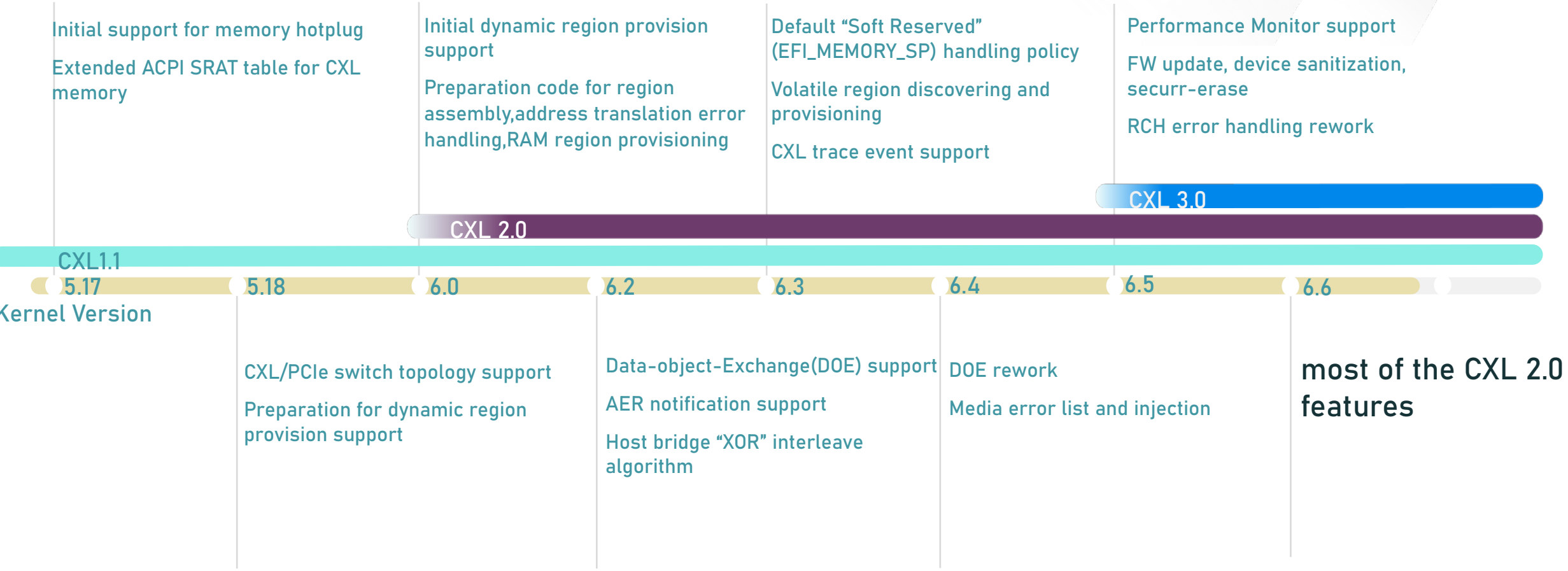


CXL Shared Memory Ecosystem

- Each node connects to shared memory over CXL 2.0 Switch
- Eliminates data duplication
- Memory speed direct byte-addressable memory access
- Application instances on any compute node can Read, Write, and Modify data in-place
- Enables pass-by-reference and zero-copy transfer of objects across multiple nodes
- Eliminates data skew



Linux Kernel CXL Features



[kernel/git/cxl/cxl.git - Compute Express Link Development]

Wrap up

- In-memory database is in action
- Linux ecosystem is ready for CXL 2.0
- Fabric Manager collaboration kicked off in OCP CMS

Thank You