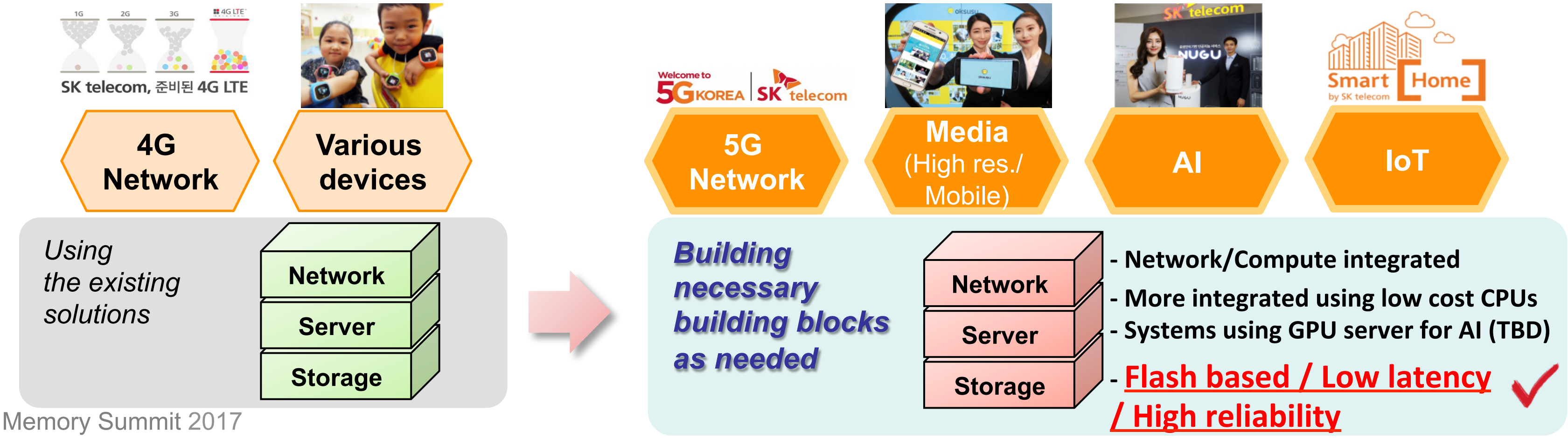
New NVMe DAS Pool with Reliability and Sharing capability

Eric H. Chang, Program Manager

New Computing Lab / SK Telecom

Why does SKT Build Systems?

- **Conventional Telcos have focused on providing differentiated services to end users.**
- **The requirements of the 5G era and 4th industrial revolution are forcing Telcos to change their delivery infrastructure.**
- **SK Telecom will lead the industry with new services that rely upon advanced infrastructure.**
    - We are building systems aligned with upcoming usage models:



4G Network

Various devices

Using the existing solutions

Network
Server
Storage

5G Network

Media (High res./ Mobile)

AI

IoT

Building necessary building blocks as needed

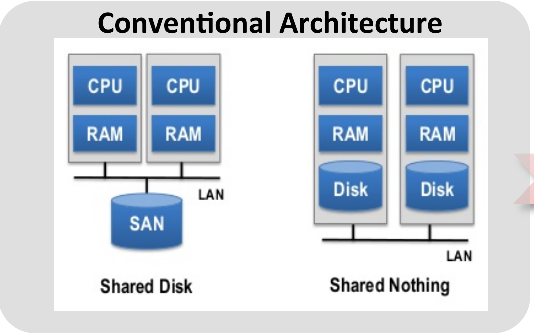
Network
Server
Storage

- **Network/Compute integrated**
- **More integrated using low cost CPUs**
- **Systems using GPU server for AI (TBD)**
- **Flash based / Low latency / High reliability** ✓

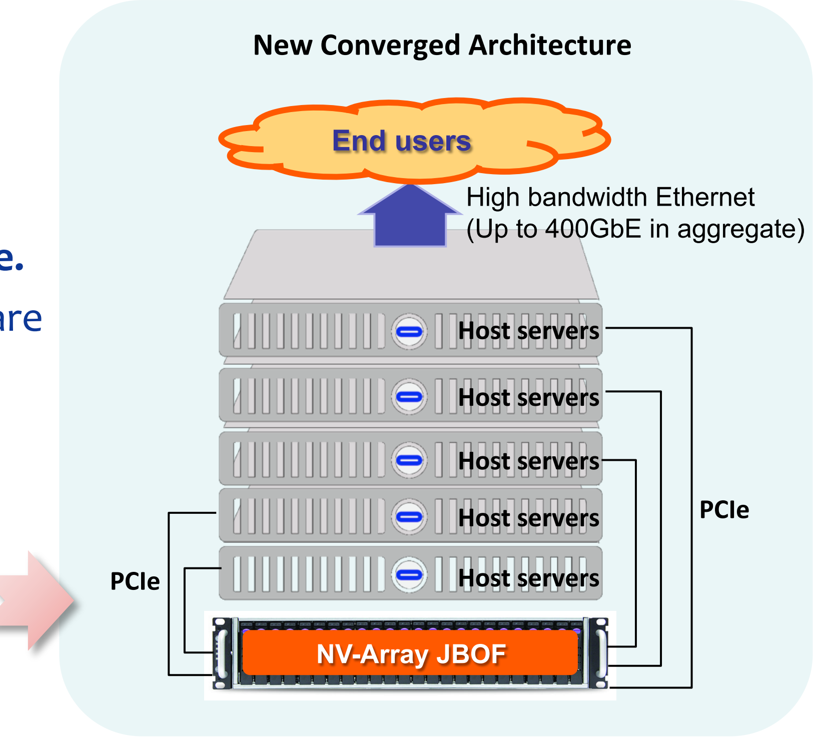
System Requirements - Storage

- **Drawbacks of the conventional architecture**
  - Shared-Disk: High complexity
  - Shared-Nothing: Large network overhead
- **New arch. maximizes advanced resource capability such as high bandwidth networks and Flash storage.**
  - Lower latency and minimized data movement are essential
- **Failover required.**

Conventional Architecture



New Converged Architecture

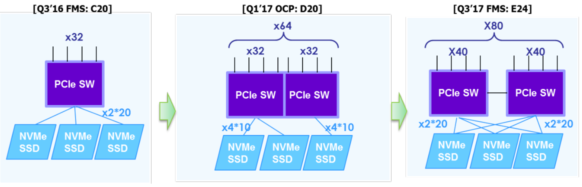


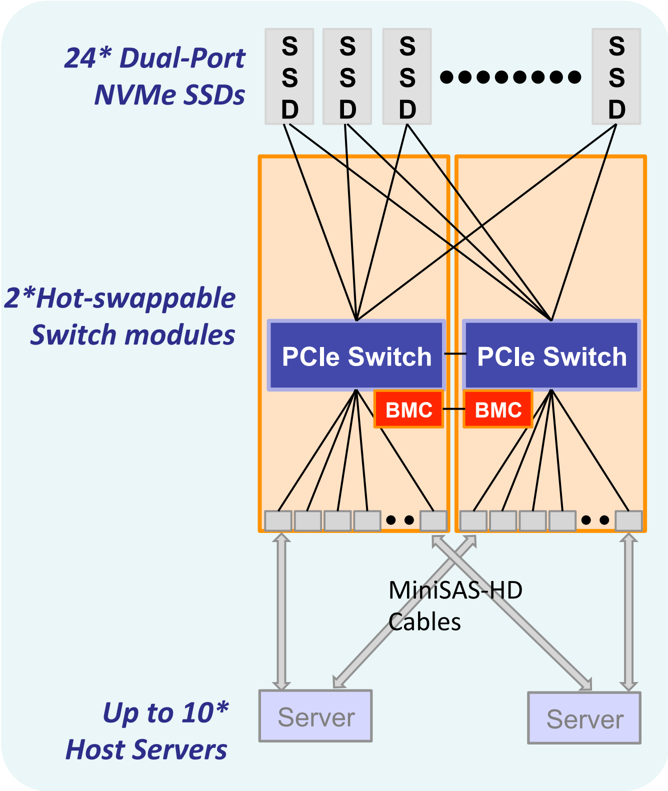
New NV-Array E24 Hardware

E24 - High Availability Architecture

- **NV-Array HA hardware support**
  - Hot pluggable NVMe SSDs
  - Hot swappable PCIe switch boards
  - Hot swappable fans and power supplies
- **Path failover with data re-routing (assisted by software)**
  - Implemented in each host as with multi-pathing software
- **Very high performance**
  - 80 lanes of PCIe Gen3, 10 hosts (8 lanes/host)
  - Up to 66GB/s bandwidth, 16M Random IOPS

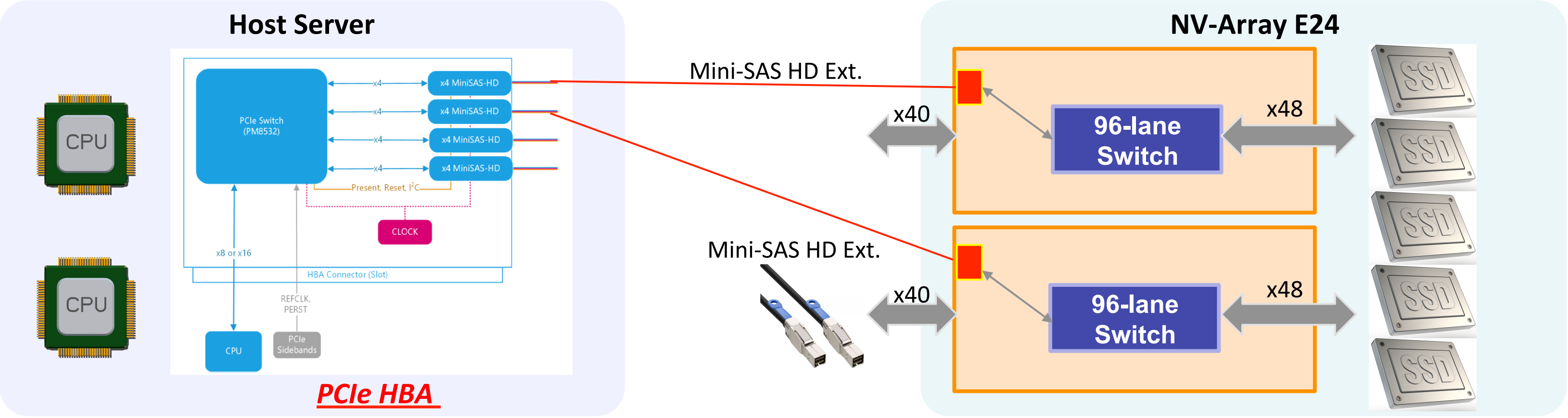
[Snapshot of OCP US summit 2017]





SKT HBA and Host Connectivity

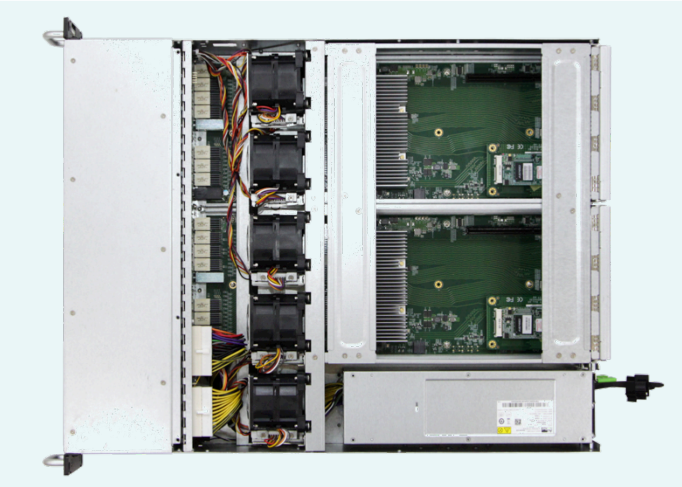
- **SKT Host Bus Adaptor provides cable connectivity to the NV-Array**
  - Supports all host servers regardless of BIOS level and Spread Spectrum support
  - PCIe x8 and x16 host slot options
  - A single HBA can provide two cables to the NV-Array for HA support
    - For servers that have a single free PCIe slot



Appearance

NV-Array Front and Top views

- Leverage the off-the-shelf enclosure with custom designed boards



HBA Top View

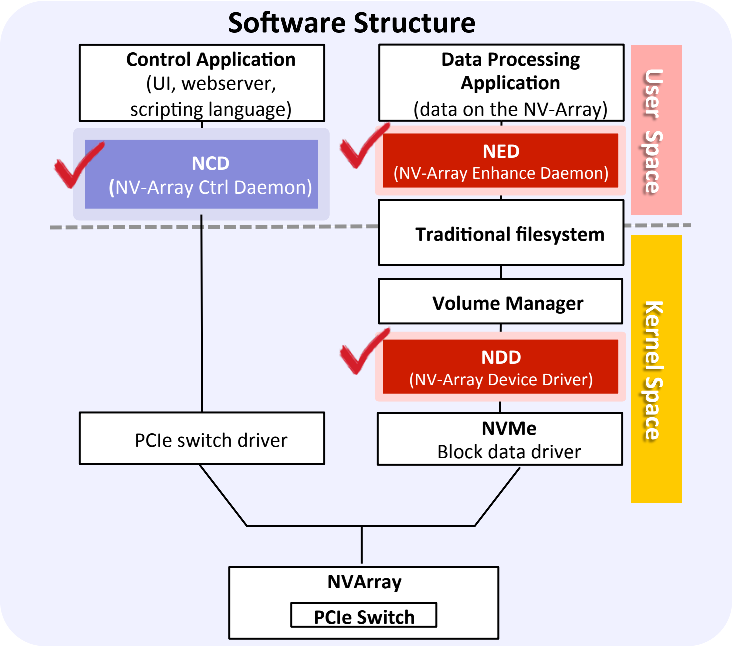
SKT adds values on JBOF boards and HBAs

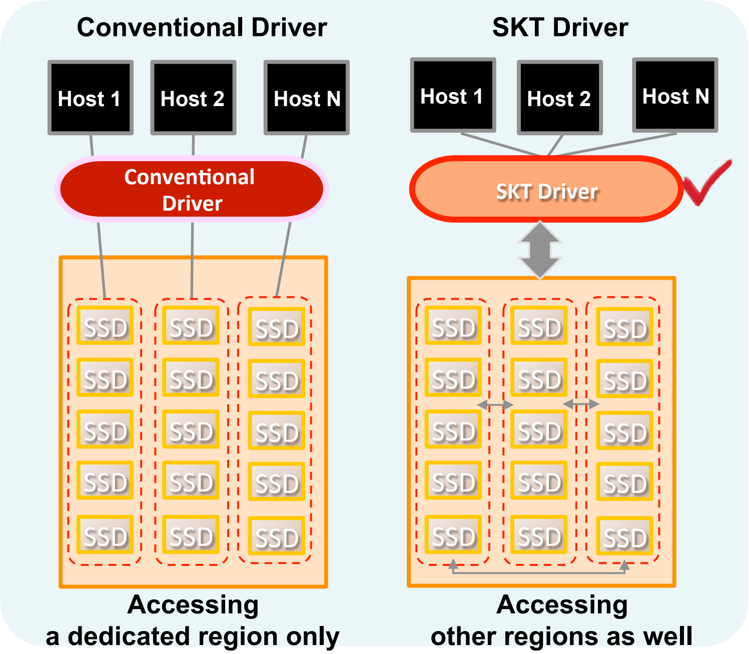
Check out the NV-Array JBOF at Booth 107

NV-Array Software to Enable New Features

NV-Array Software[NxD] Stack

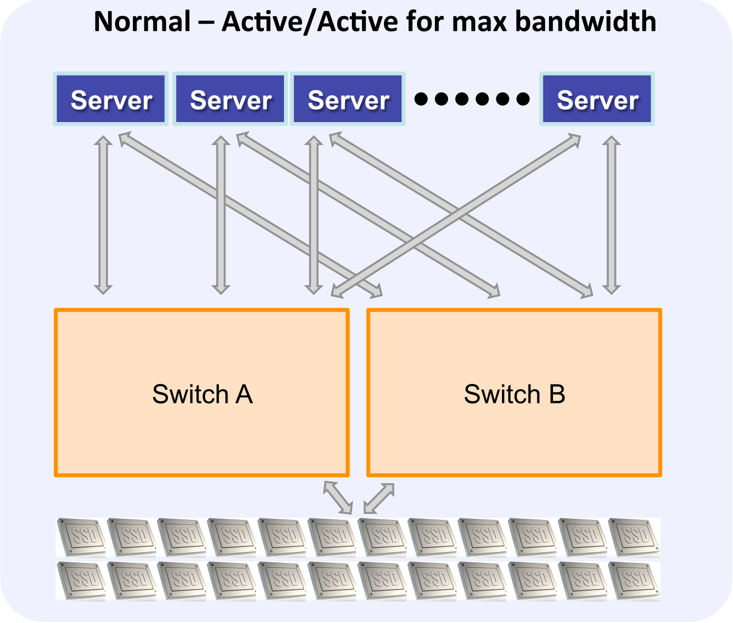
- **NCD – Provides user-mode control of all software/hardware operations**
- **NDD(data) / NED(control) – Allows block level data sharing/routing among hosts**
  - Can be overlaid with a distributed file system



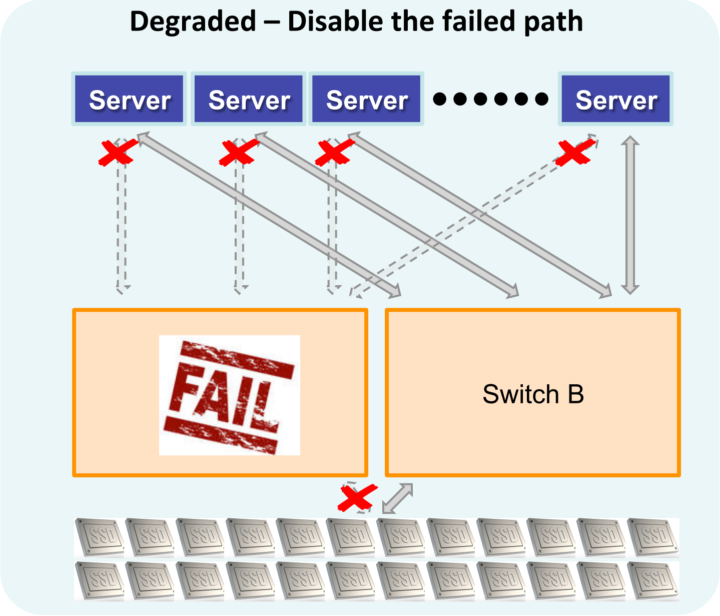


Data Path Re-routing at Switch failure

- On a cable or switch board failure, the device driver on the Host server re-routes data to t he operational path - bandwidth will be reduced by half.



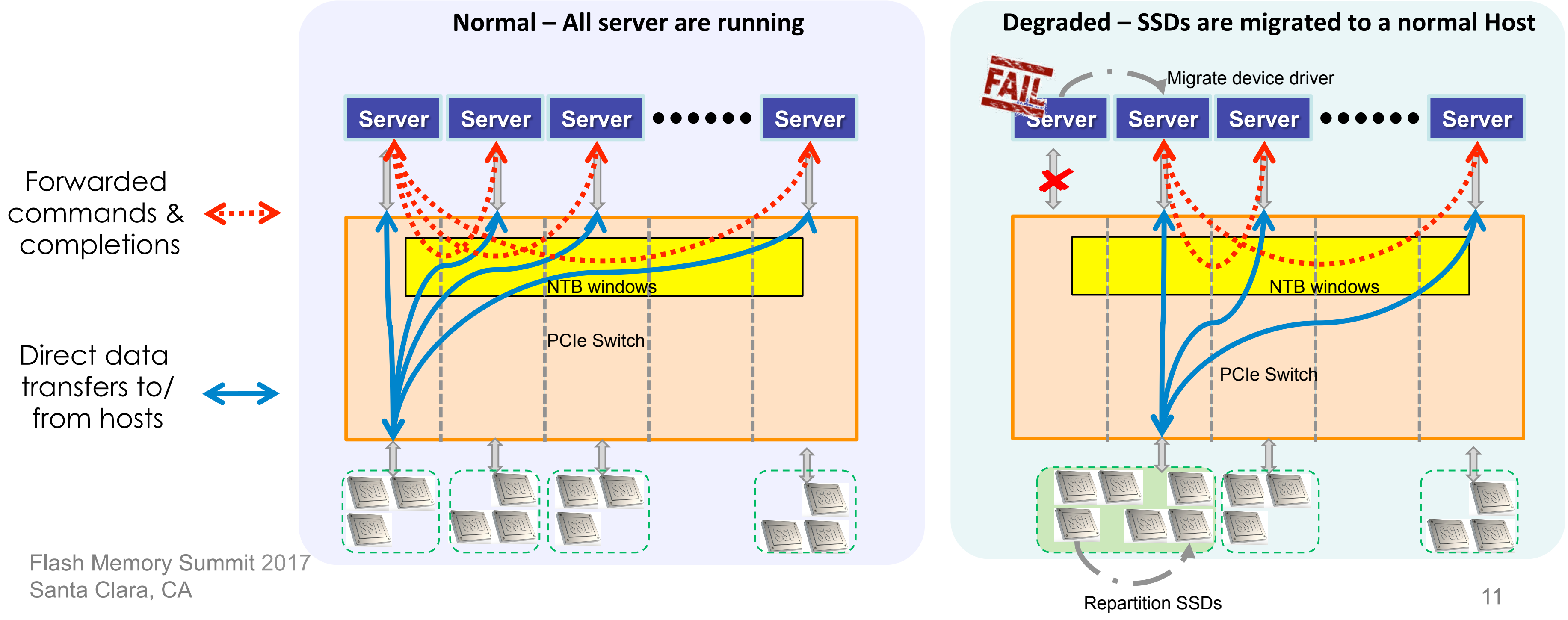
Normal – Active/Active for max bandwidth



Degraded – Disable the failed path

Data Path Re-routing at Server failure

- **When a host fails, the SSDs are dynamically repartitioned to another host - the bandwidth will be reduced by 1/(original number of hosts)**



Forwarded commands & completions

Direct data transfers to/ from hosts

Normal – All server are running

Server Server Server •••••• Server

NTB windows

PCIe Switch

Degraded – SSDs are migrated to a normal Host

Migrate device driver

Server Server Server •••••• Server

NTB windows

PCIe Switch

Repartition SSDs

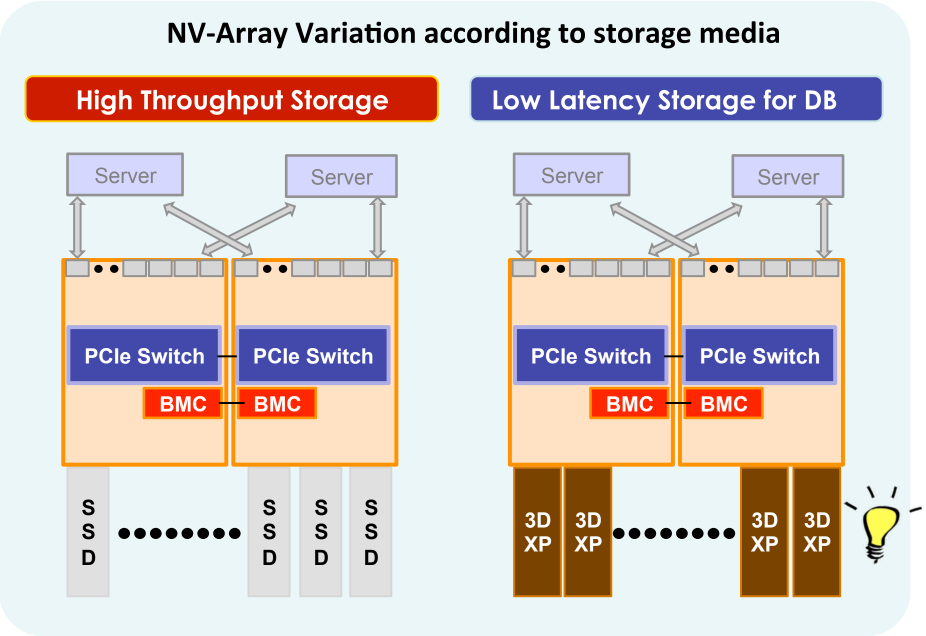
Findings: NVMe Hot-plug

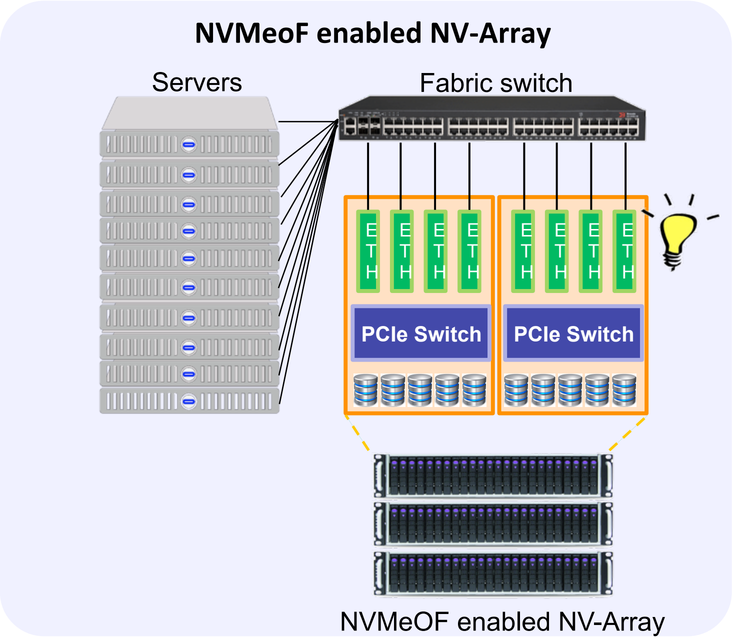
- **Reliable Hot Plug requires complex interaction between system hardware and software comp onents**
- **All system components must be properly configured and their operation validated**
  - Linux kernel version must be very recent (we used version 4.11.8)
    - Versions prior to 4.7 have no DPC support at all
    - There are changes after V4.7 - and after V4.11.8 …
  - Kernel build configuration must be set to include DPC drivers
  - Signals from the SSD slot (PRESENT, POWER CONTROL, RESET) must be properly defined a nd configured in the PCIe switch
  - NVMe SSDs must be verified for proper operation after a hot insertion
    - Not all initialize properly

Future works and Summary

Future Works (2018-19, TBD)

- 3D XP based NV-array for the highly consistent top tier storage for Database.
- NVMe over fabrics(NVMeoF) enabled for NV-Array for scalability





Summary

- The new NV-Array E24 offers the advanced I/O performance and the reliability that Telco/Ente rprise users require.
- SKT Drivers enable host servers to share data with others connected to the same NV-Array.
- SKT Drivers manage the data and control paths to support HA.
- SKT's two-year roadmap includes enhancing the NV-Array with top-tier, low latency storage a nd adding NVMeOF for improved scalability.

Thank you!

Please come visit Booth #107