KALRAY

Enabling Next-Generation Storage Fabrics with NVMe-oF I/O processors

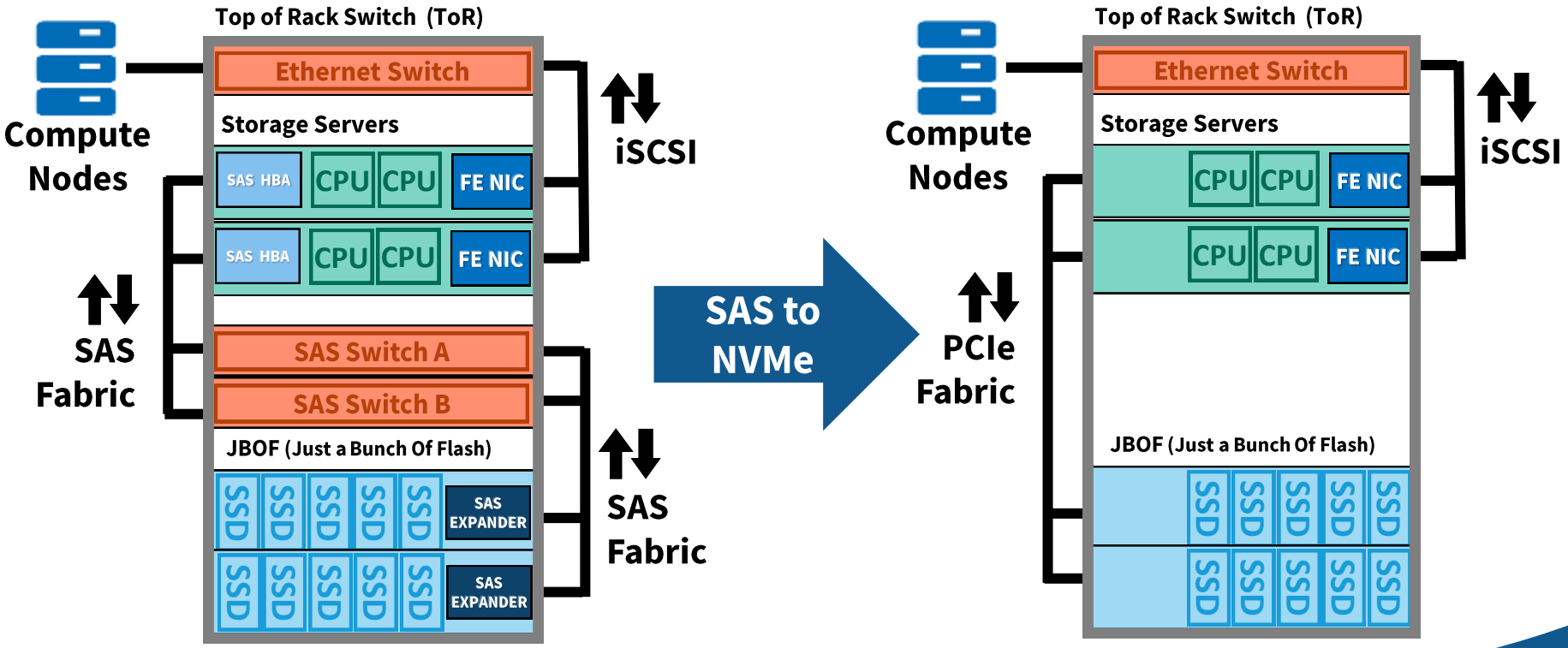
Flash Memory Summit

Thursday August 8, 2017

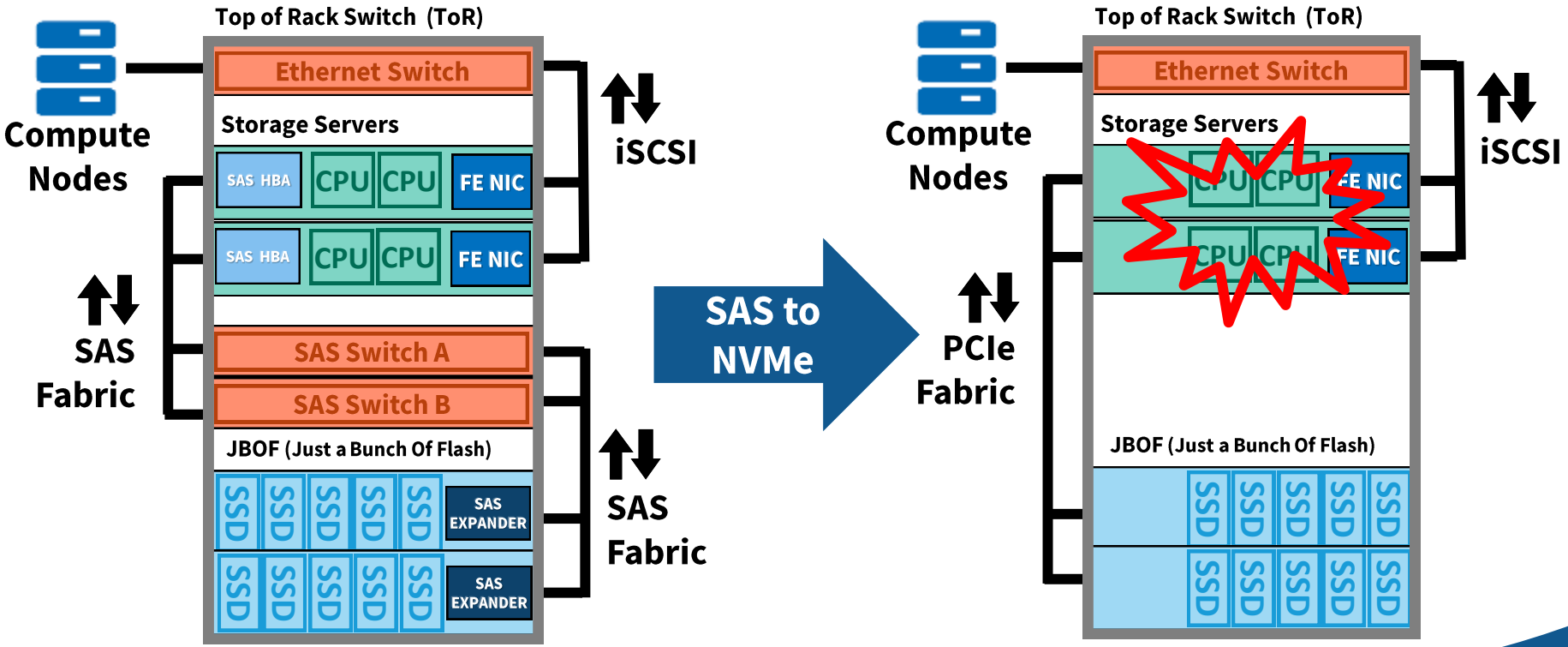
bganne@kalrayinc.com

www.kalrayinc.com

FROM SAS TO NVMe

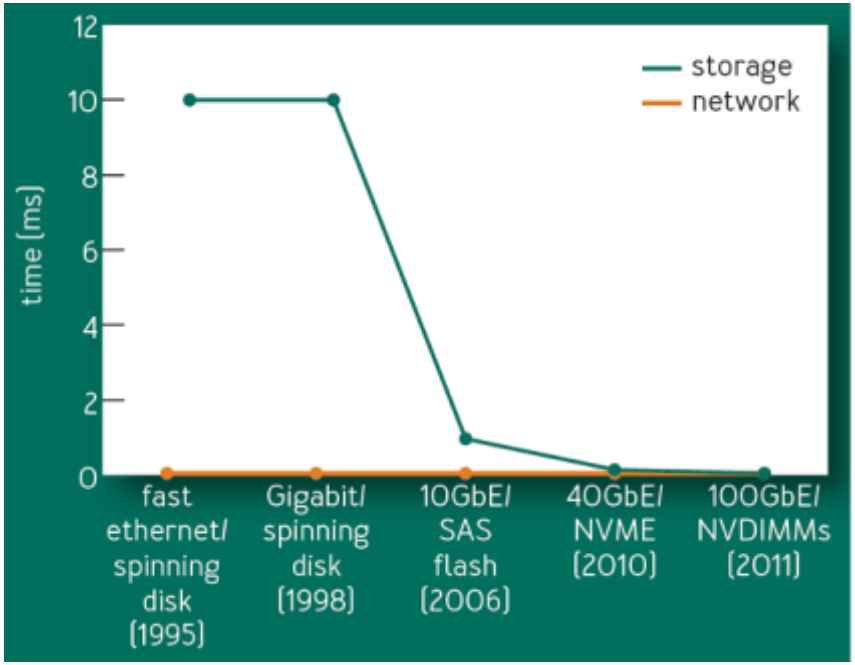


FROM SAS TO NVMe
CPU + DDR BOTTLENECK



CPU IS NO LONGER FAST ENOUGH
- or -
NVMe + NETWORK IS TOO FAST

Memory Tier	Latency
L1 cache reference	1 ns
L2 cache reference	10 ns
DDR reference	100 ns
Send 1kB over 10 Gbps network	1,000 ns
SSD read latency	100,000 ns
Read 1 MB sequentially from SSD	1,000,000 ns
HDD seek	10,000,000 ns
Read 1 MB sequentially from HDD	100,000,000 ns



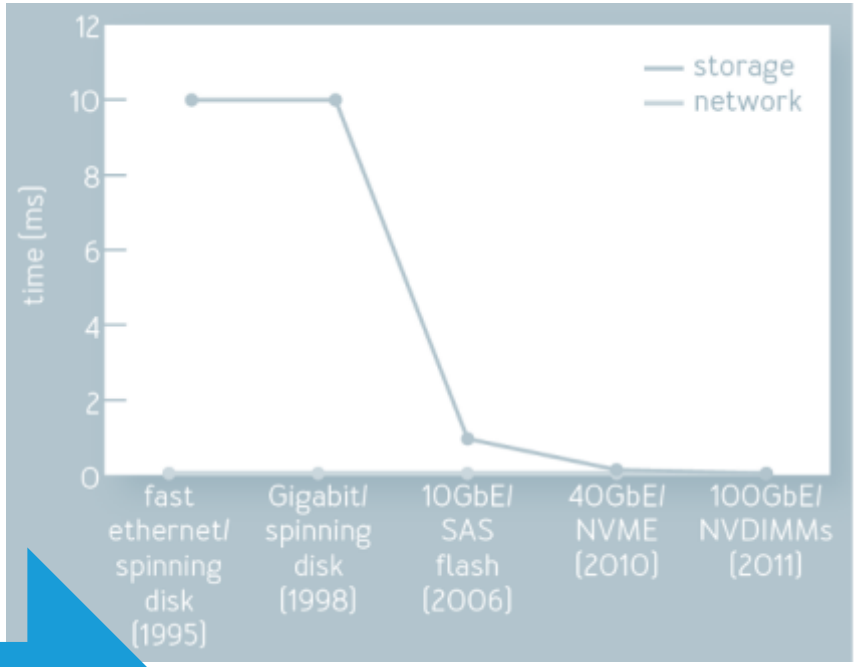
Time budget to service an I/O

Mihir Nanavati et al., Non-volatile Storage - Implications of the Datacenter's Shifting Center, acmqueue, File Systems and Storage, volume 13, issue 9, January 5, 2016

KALRAY

CPU IS NO LONGER FAST ENOUGH
- or -
NVMe + NETWORK GROW TOO FAST

Memory Tier	Latency
L1 cache reference	1 ns
L2 cache reference	10 ns
DDR reference	100 ns
Send 1kB over 10 Gbps network	1,000 ns
SSD read latency	100,000 ns
Read 1 MB sequentially from SSD	1,000,000 ns
HDD seek	10,000,000 ns



Observation #1: it only takes a few cache misses to ruin your I/O cycles budget...

dget to service an I/O

ati et al., Non-volatile Storage - Implications of the er's Shifting Center, acmqueue, File Systems and Storage, e 13, issue 9, January 5, 2016

KALRAY

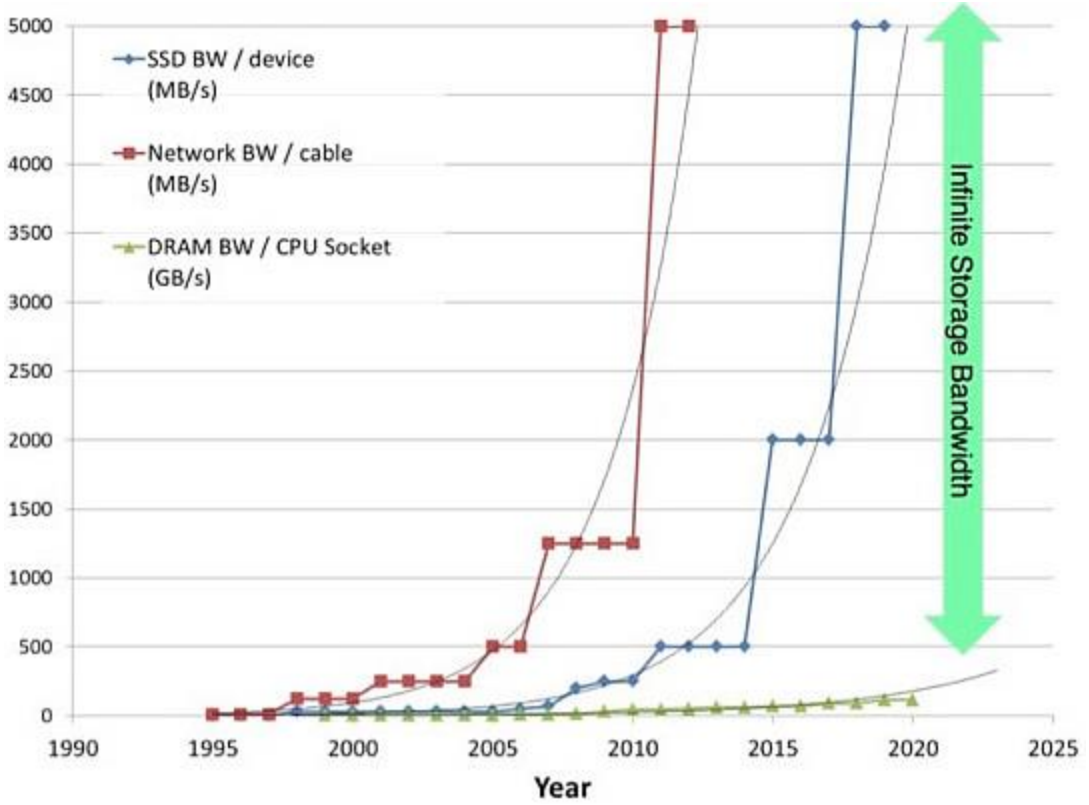
DDR BANDWIDTH DOES NOT SCALE FAST ENOUGH
- or –
NVMe + NETWORK GROW TOO FAST

NVMe, next-gen SCM and network scales at the same exponential speed

- Doubling every 18 month

This is not the case for DRAM

- Doubling every 26 month



Network, storage and DRAM trends

SanDisk® Fellow, Fritz Kruger, CPU Bandwidth – The Worrisome 2020 Trend

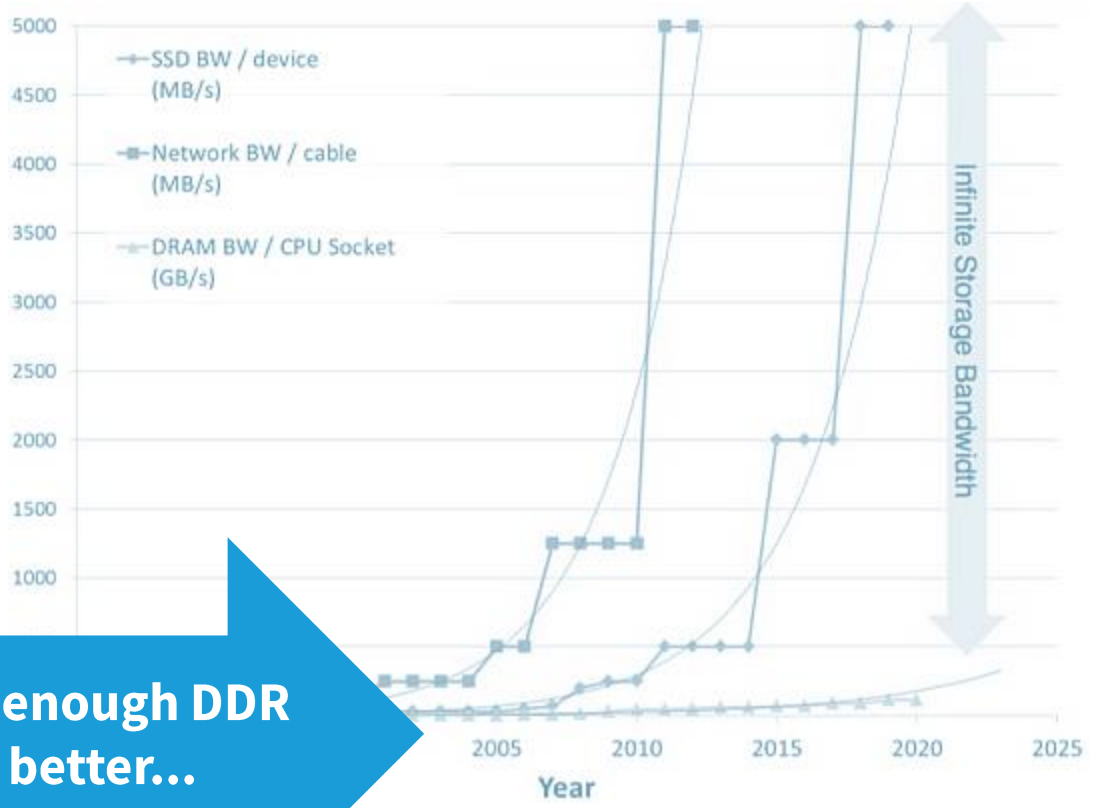
DDR BANDWIDTH DOES NOT SCALE ANYMORE
- or –
NVMe + NETWORK GROW TOO FAST

NVMe, next-gen SCM and network scales at the same exponential speed

- Doubling every 18 month

This is not the case for DRAM

- Doubling every 26 month



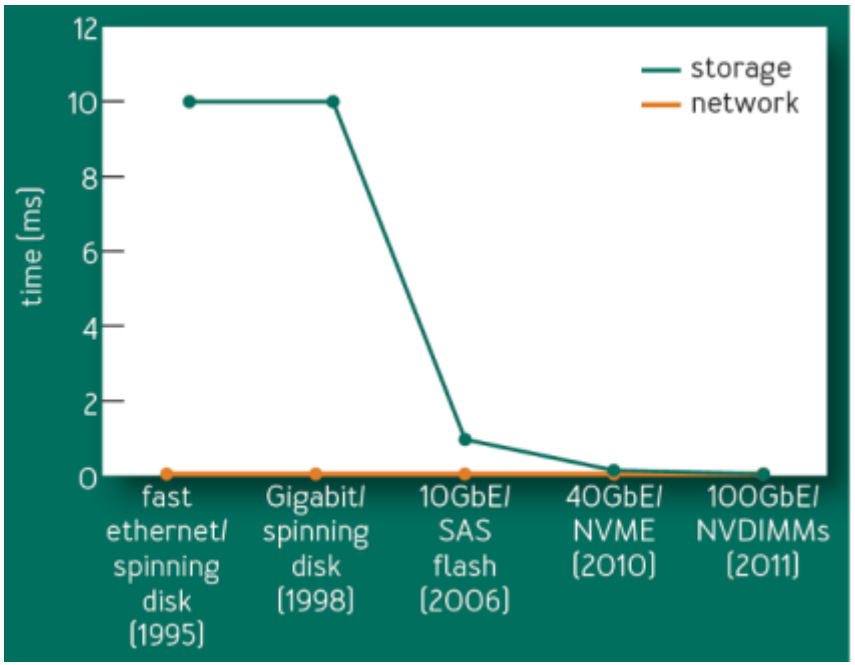
Network, sto and DRAM trends

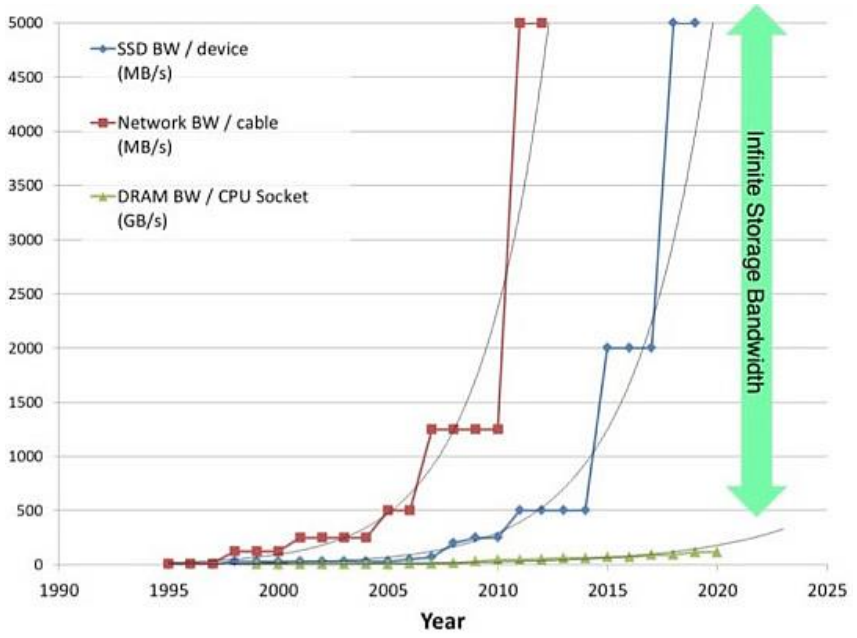
SanDisk® Fellow, Fritz Kruger, CPU Bandwidth – The Worrisome 2020 Trend

<u>Observation #2</u>: you never have enough DDR channels, and it is not going better...

KALRAY

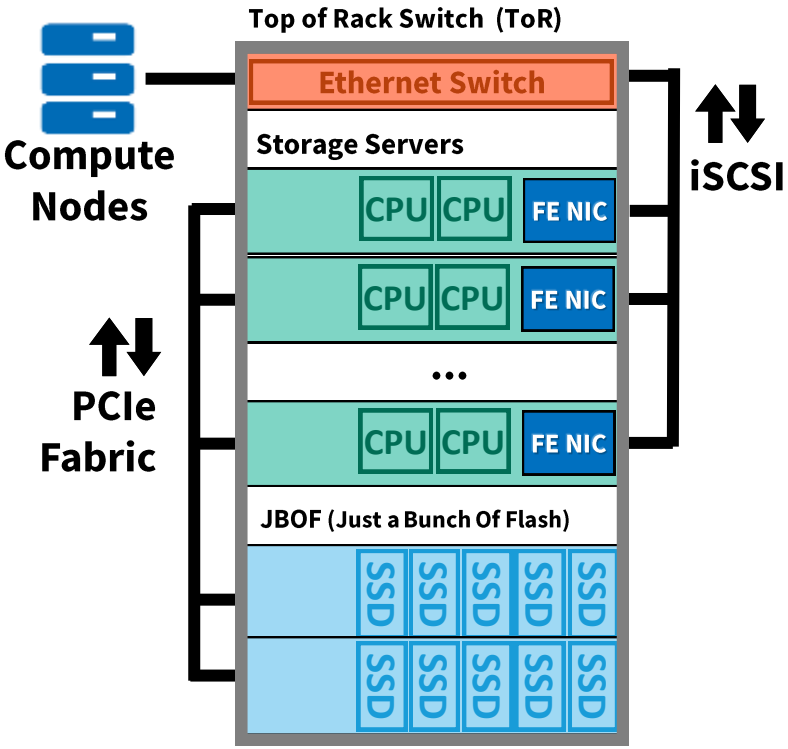
CPU + DDR = BOTTLENECK





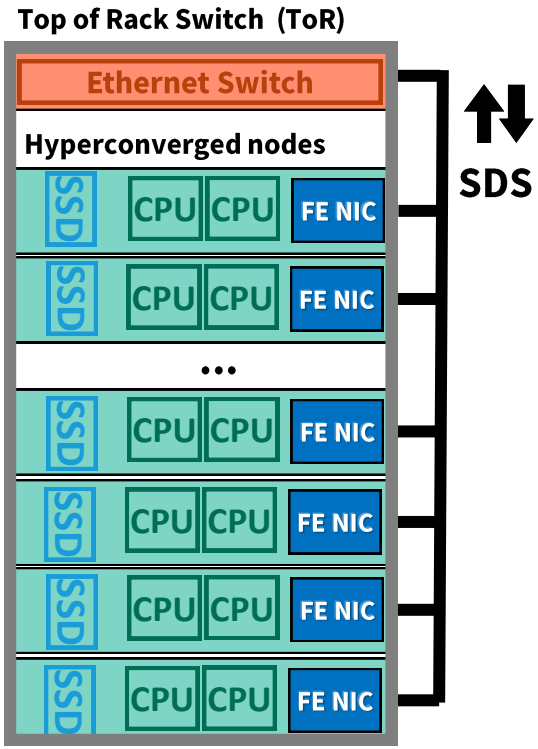
KALRAY

SO, HOW CAN WE SOLVE THESE ISSUES?



Option #1: you scale up

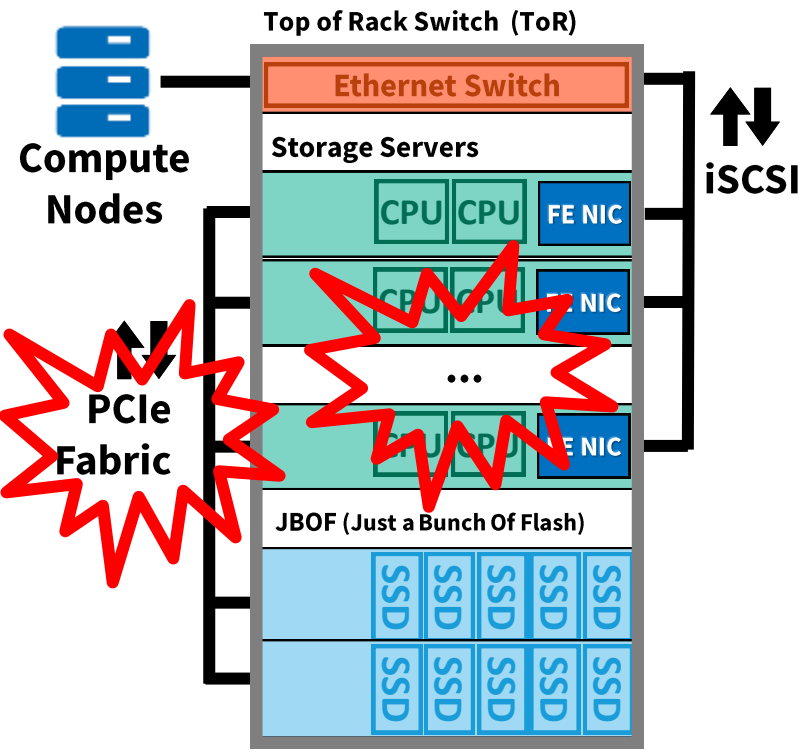
☺ More head nodes = more CPU + DDR



Option #2: you scale out

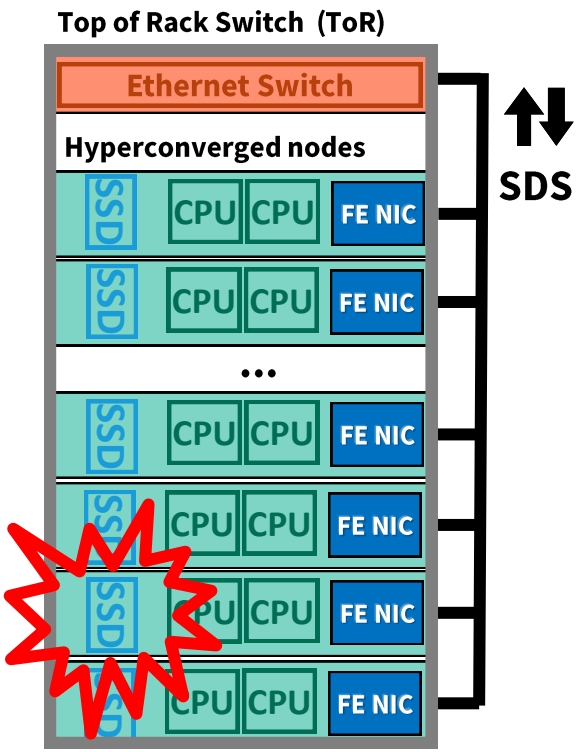
☺ Hyperconverged/SDS scales naturally

SO, HOW CAN WE SOLVE THESE ISSUES?



Option #1: you scale up

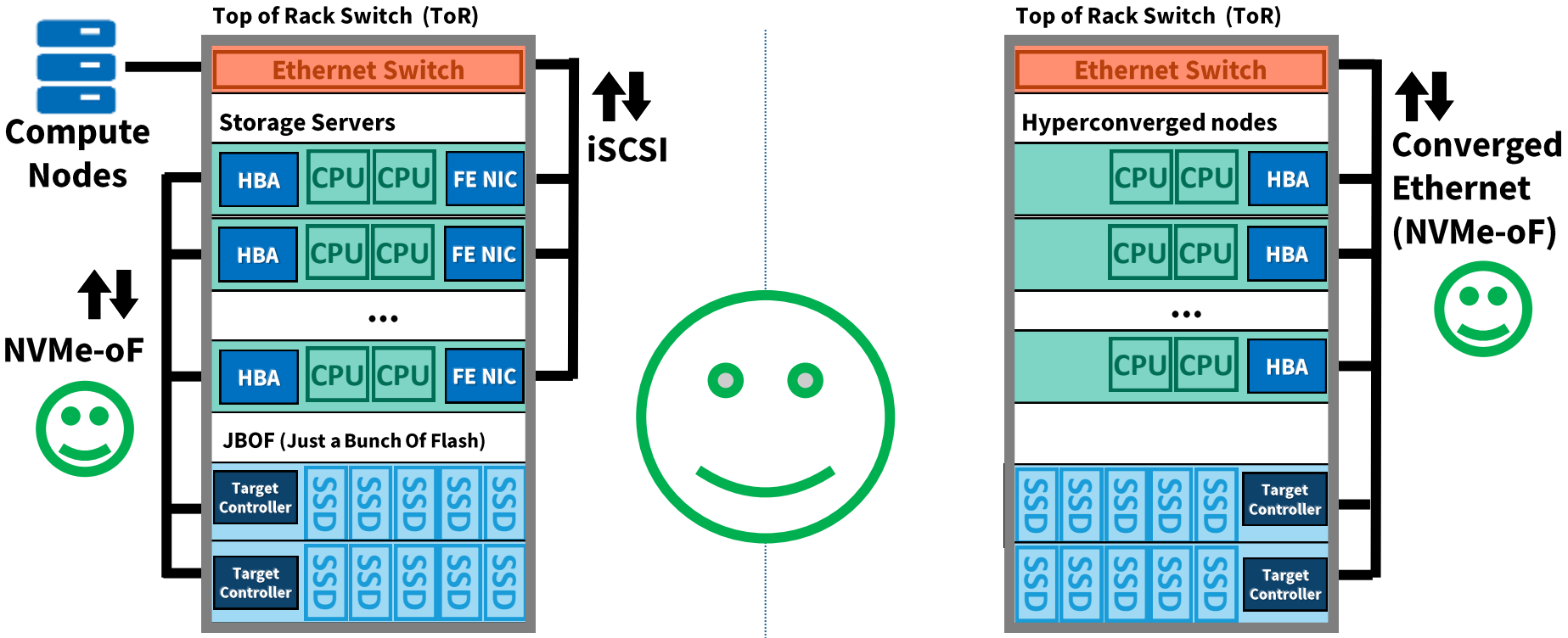
- ☺ More head nodes = more CPU + DDR
- ☹ PCIe has limited scaling



Option #2: you scale out

- ☺ Hyperconverged/SDS scales naturally
- ☹ Compute/storage ratio is fixed
- ☹ DAS is expansive

SO, HOW CAN WE SOLVE THESE ISSUES?
NVMe-oF WAS DESIGNED FOR THAT!



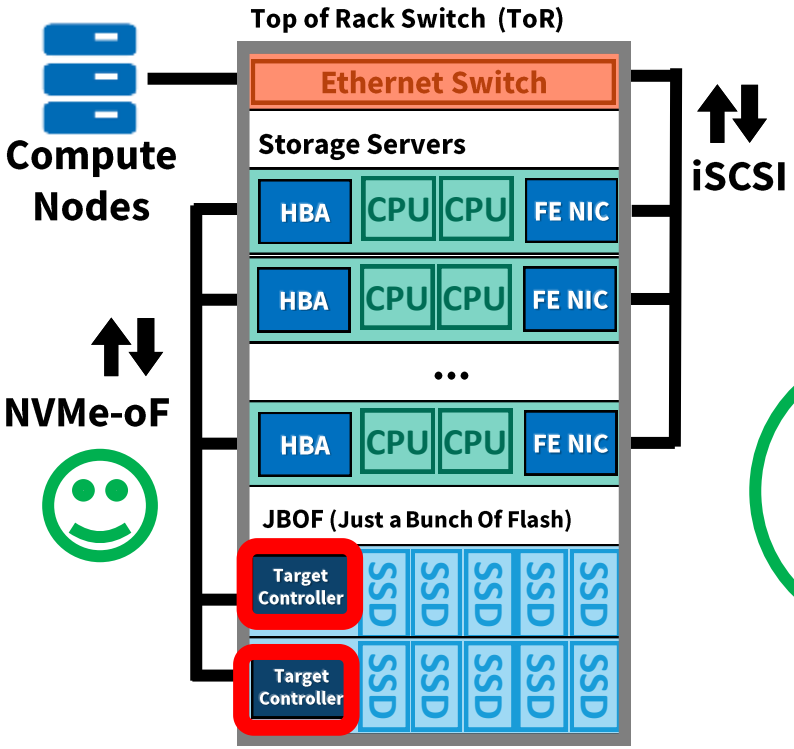
Option #1: Enterprise SAN/NAS

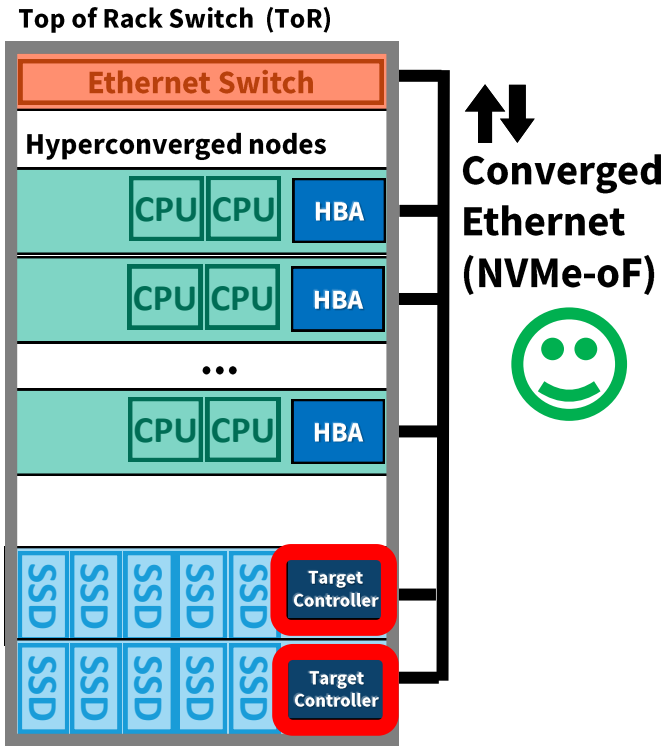
- ☺ Scale head nodes based on services
- ☺ Scale storage as needed

Option #2: Disaggregated SDS

- ☺ Scale compute & storage independently

SO, HOW CAN WE SOLVE THESE ISSUES?
NVMe-oF WAS DESIGNED FOR THAT!



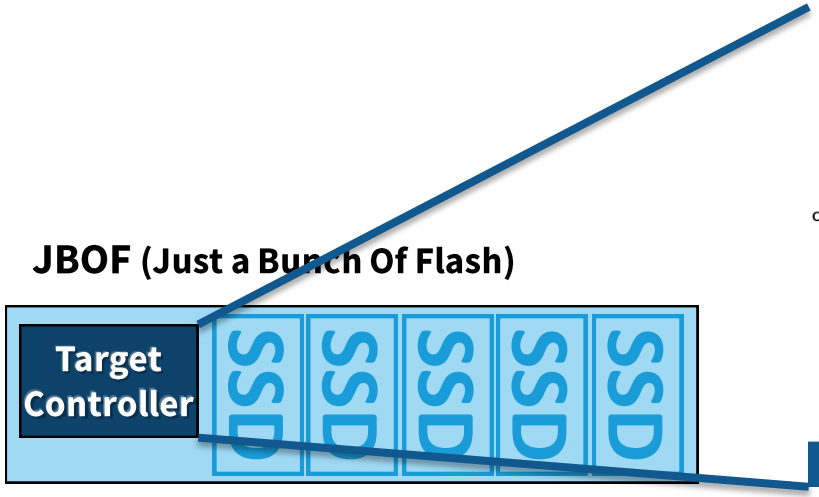


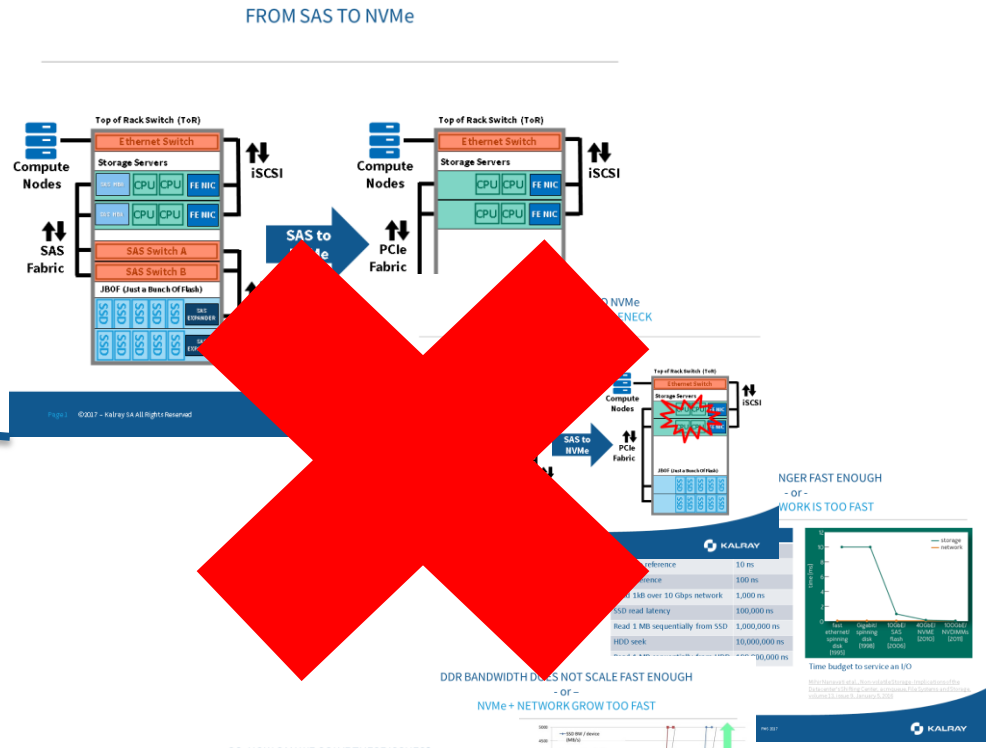
Option #1: Enterprise SAN/NAS
- ☺ Scale head nodes based on services
- ☺ Scale storage as needed

Option #2: Disaggregated SDS
- ☺ Scale compute & storage independently

KALRAY

SO, HOW CAN WE SOLVE THESE ISSUES?
REMOVE CPU + DDR FROM NVMe-oF DATAPATH





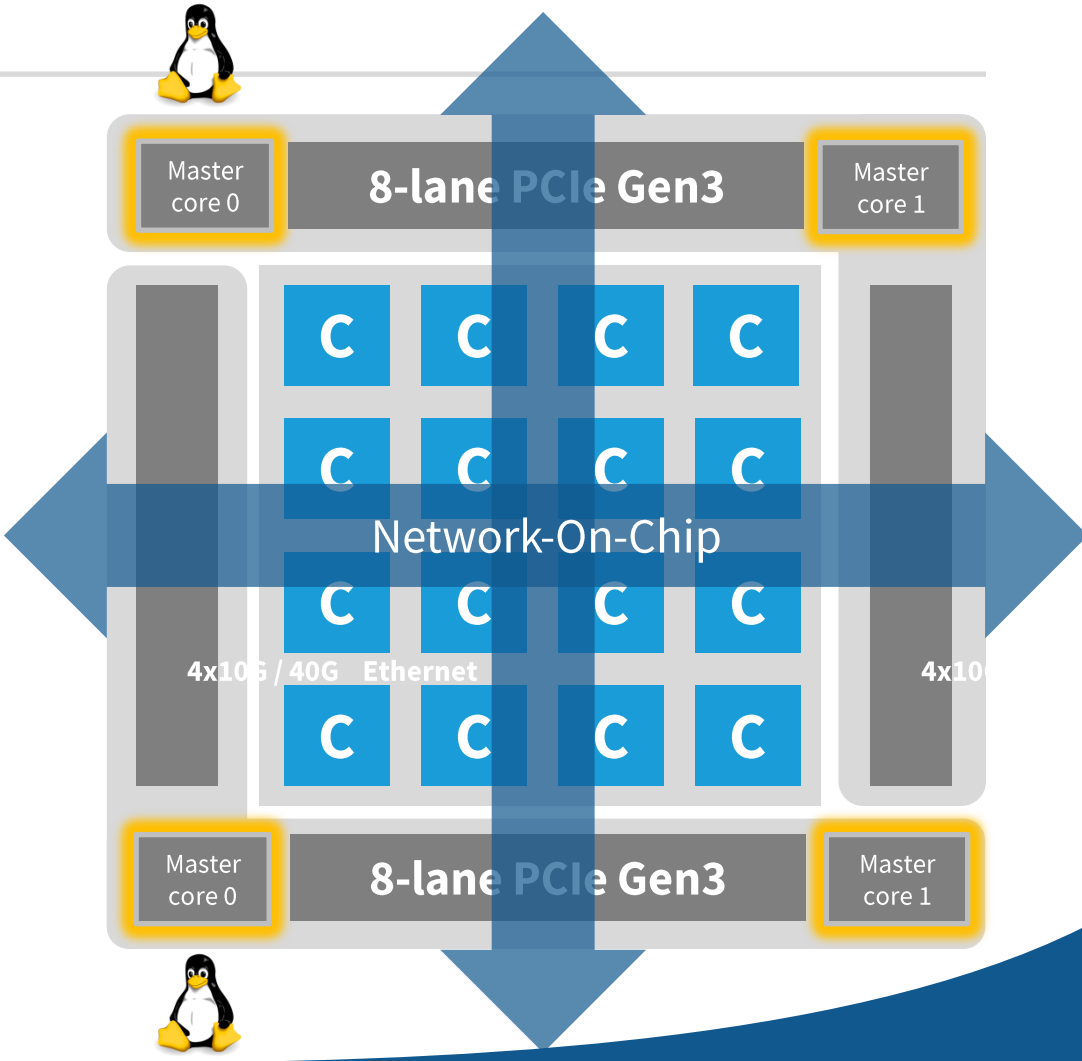
KALRAY I/O PROCESSOR
MERGING IOs, MEMORY AND COMPUTE TOGETHER

HIGH-SPEED INTERFACES:
- 2x 40GbE
- 2x PCIe Gen3 8-lanes (EP/RC)

CONNECTED TO A LARGE ARRAY OF PROCESSING
- Full C/C++ Programmable
- Dataplane execution

VIA A HIGH BANDWIDTH LOW LATENCY NETWORK ON CHIP
- Direct packet-to-core delivery
- Direct core-to-core transfers
- Direct connect between multiple MPPAs

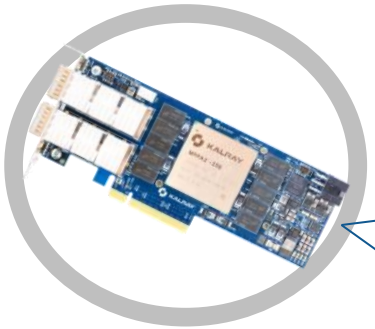
AND I/O MASTER CORES
- Runs Linux
- Runs control plane

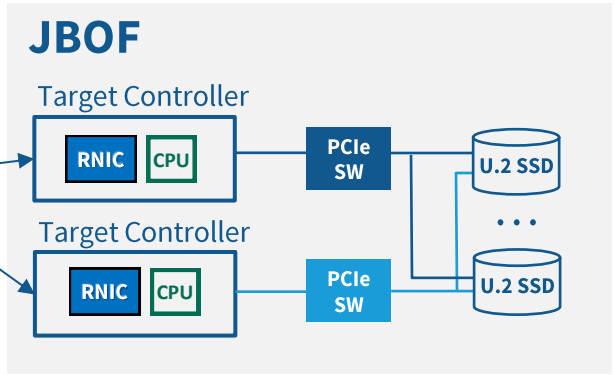


KALRAY

KALRAY I/O PROCESSOR
ENABLING AN NVMe-oF JBOF (AND MORE)

KALRAY TARGET CONTROLLER





Manages all the storage functions of the new generation storage JBOF

TARGET CONTROLLER FEATURE

PCIe RC MODE FOR DIRECT SSD CONTROL
- Standard Linux with NVMe Driver
- **Any NVMe SSD supported – no need for CMB**
- Control up to 255 PCIe endpoints
- SSD Hot Plug Support

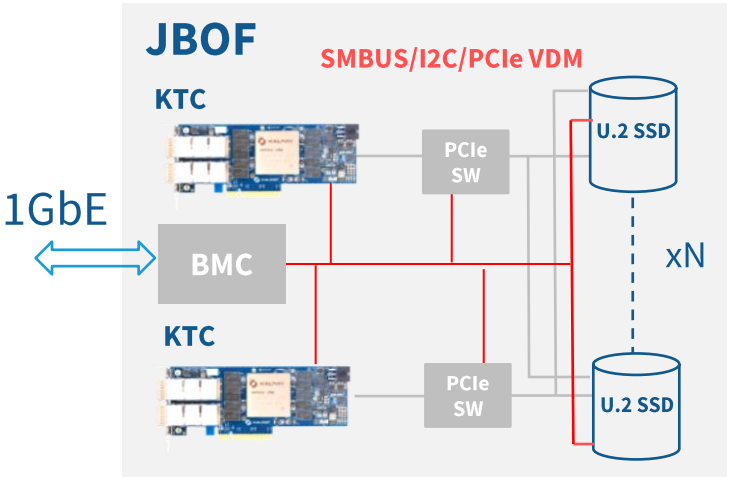
NVMe-oF PROTOCOL OVER RoCEv1/v2
- 4x + performant than SAS (IOPs &throughput)
- Scalability: Connect up to 2048 initiator cores
- standard ethernet connectivity

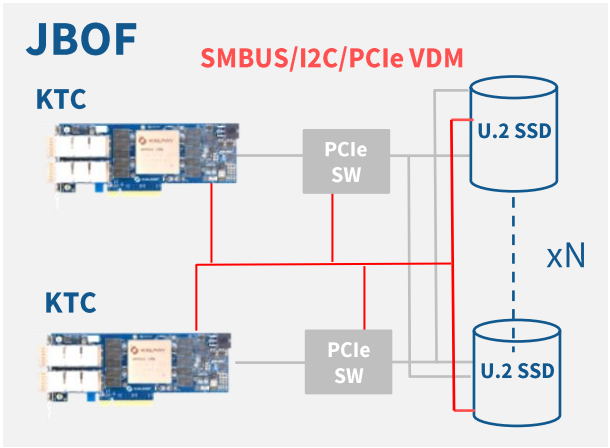
BOARD MANAGEMENT CONTROL (BMC)
- Supervise enclosure

HIGH AVAILABILITY
- Multipath architecture

END USER INLINE PROCESSING
- Encryption, erasure coding...

BOARD MANAGEMENT CONTROL (BMC)





OUT-OF-BAND WITH EXTERNAL BMC

- KTC seen as a component attached to BMC
- BMC exposes out-of-band management interface over 1GbE

IN-BAND WITH INTEGRATED BMC

- KTC implements in-band management on Linux

ENCLOSURE MANAGEMENT

- Sensor monitoring (Temperature, voltage, …)
- Fan Control
- NVMe-MI

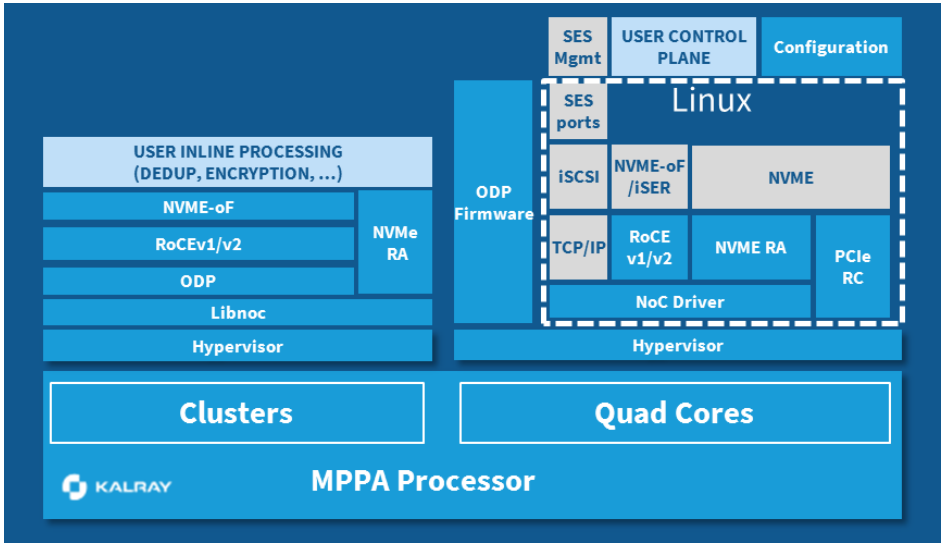
FABRIC MANAGEMENT

- Fabric configuration (network address...)
- NVMe-oF discovery
- NVMe-oF namespaces, zoning

MANAGEMENT API

- Redfish Fabric Extension
- Swordfish
- IPMI
- OpenBMC
- Ansible + nvmetcli

END USER CUSTOMIZABLE SOLUTION



CUSTOMIZABLE FUNCTIONS

INLINE PROCESSING

- Compression
- Encryption
- Deduplication
- Erasure Coding

BOARD MANAGEMENT CONTROL (BMC)

- Redfish/Swordfish
- SES
- OpenBMC
- ...

FLEXIBLE IO SCHEDULING POLICIES

- Implement optimized Read/write scheduling to improve performances and determinism

KALRAY

KTC40 & KTC80 HARDWARE SPECIFICATION

KALRAY

KTC80 *

* UNDER CONSOLIDATION

- MPPA®2.5-256 (Bostan2 processor)
- 80 GbE sustained throughput
- 2 x QSFP+ ports
- Integrated 16-lane PCIe Gen3 Switch
- 2 x DDR3-1866 with ECC (4GB)
- FHHL (Full-Height, Half-Length)

KALRAY

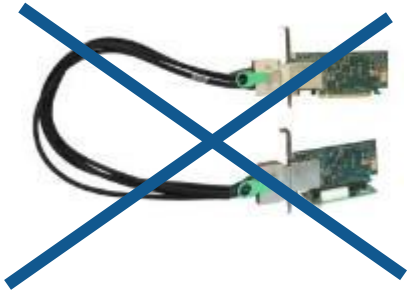
KTC40

- MPPA®2.5-256 (Bostan2 processor)
- 40GbE sustained throughput
- 2 x QSFP+ ports
- 8-lane PCIe Gen3
- 2 x DDR3-1866 with ECC (2GB)
- LP (Low-profile)

KALRAY

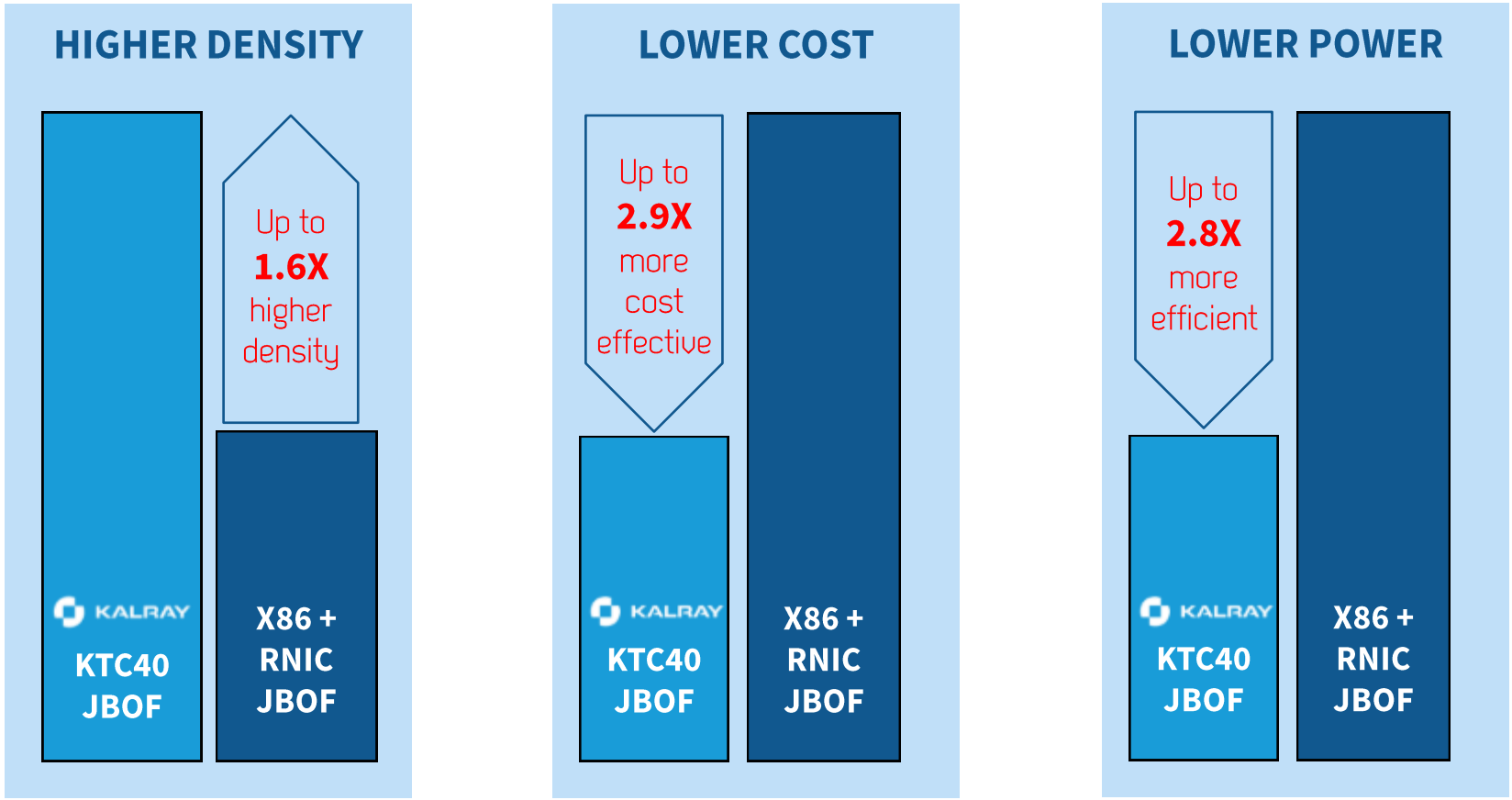
YOUR PCIe JBOF EASILY BECOMES AN ETHERNET JBOF WITH KALRAY TARGET CONTROLLER

No modifications



KTC ENABLES FAST TIME-TO-MARKET TO BUILD NVMe-oF JBOF

KALRAY
HIGHER DENSITY, LOWER COST, LOWER POWER



KALRAY

KALRAY S.A. - GRENOBLE - FRANCE
445 rue Lavoisier,
38 330 Montbonnot - France
Tel: +33 (0)4 76 18 09 18
email: info@kalray.eu

KALRAY INC. - LOS ALTOS - USA
4962 El Camino Real
Los Altos, CA - USA
Tel: +1 (650) 469 3729
email: info@kalrayinc.com

KALRAY