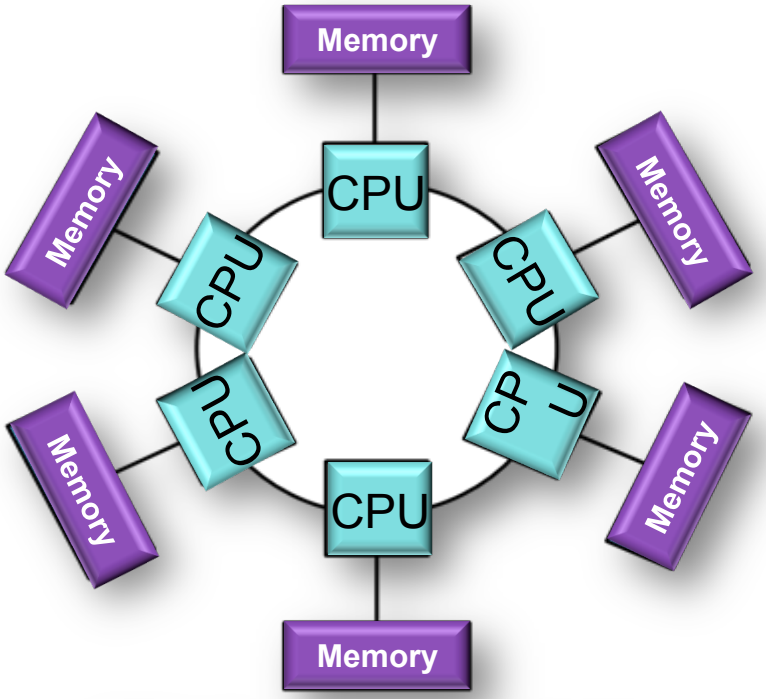
Flashmatrix™ Technology

All-flash Super-Converged Platform

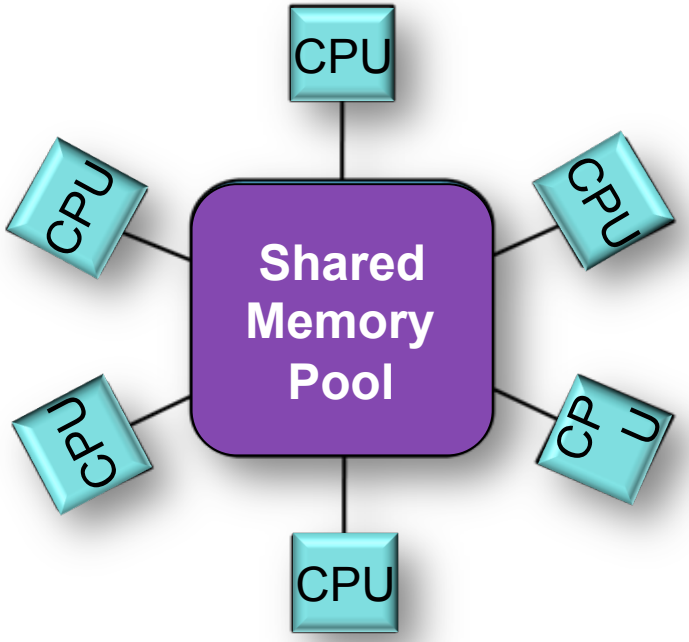
By

Ram Johri

Traditional Von Neumann vs. Data Centric Architecture



CPU-Centric Architecture

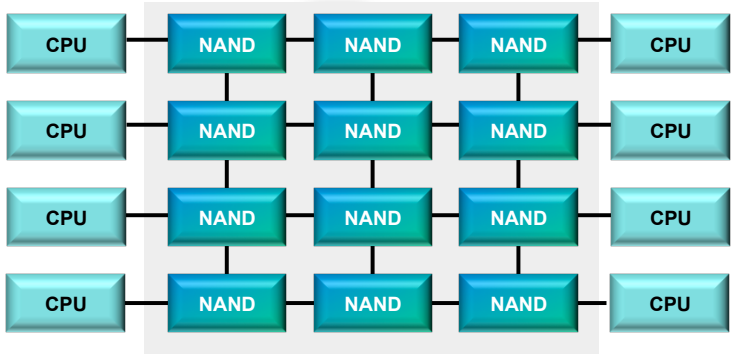


Data-Centric Architecture

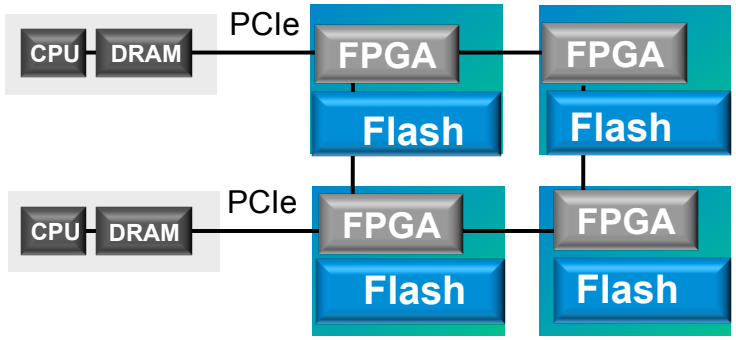
Flashmatrix: Data Centric Architecture

Flash Memory Summit

NAND flash memories in a matrix



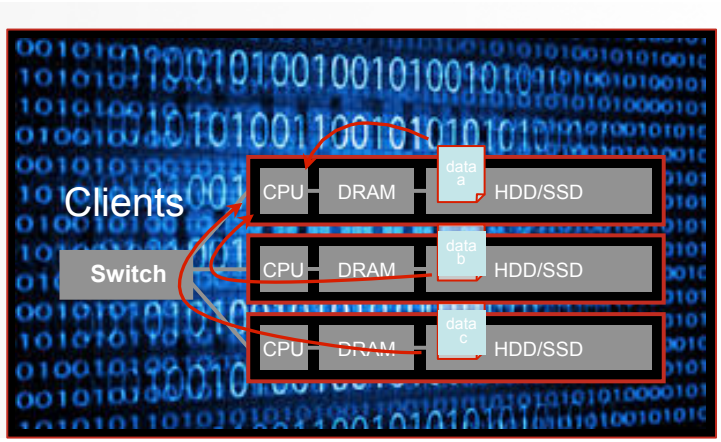
Parallel data transfer & processing



Shared Everything

All Flash

The Data Challenge of Existing Architectures

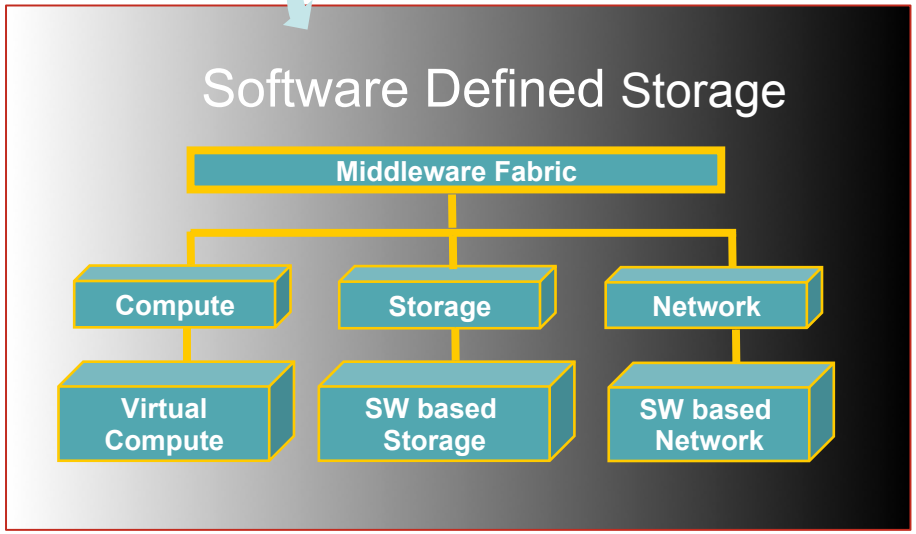


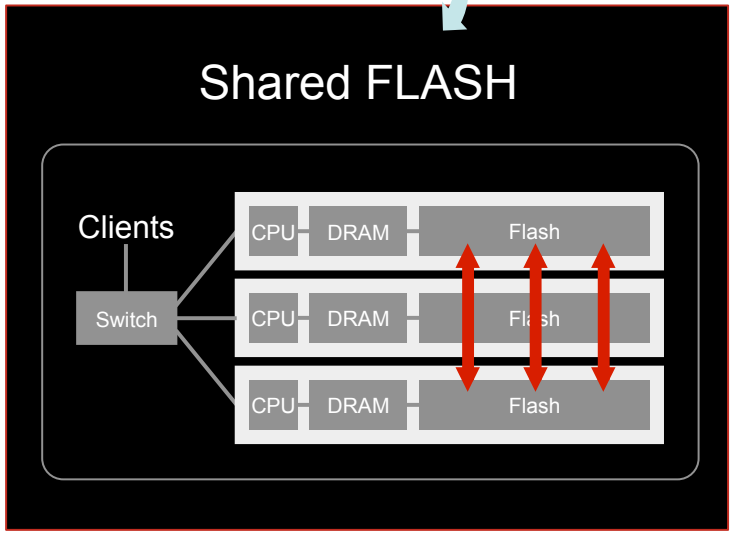
- Switching CPU cores carries a latency penalty due to data transfers in storage
- Performance strongly depends on the physical location of the stored data
- Performance varies with application work loads
- There is an “All or Nothing” scalability

Flash Memory Summit

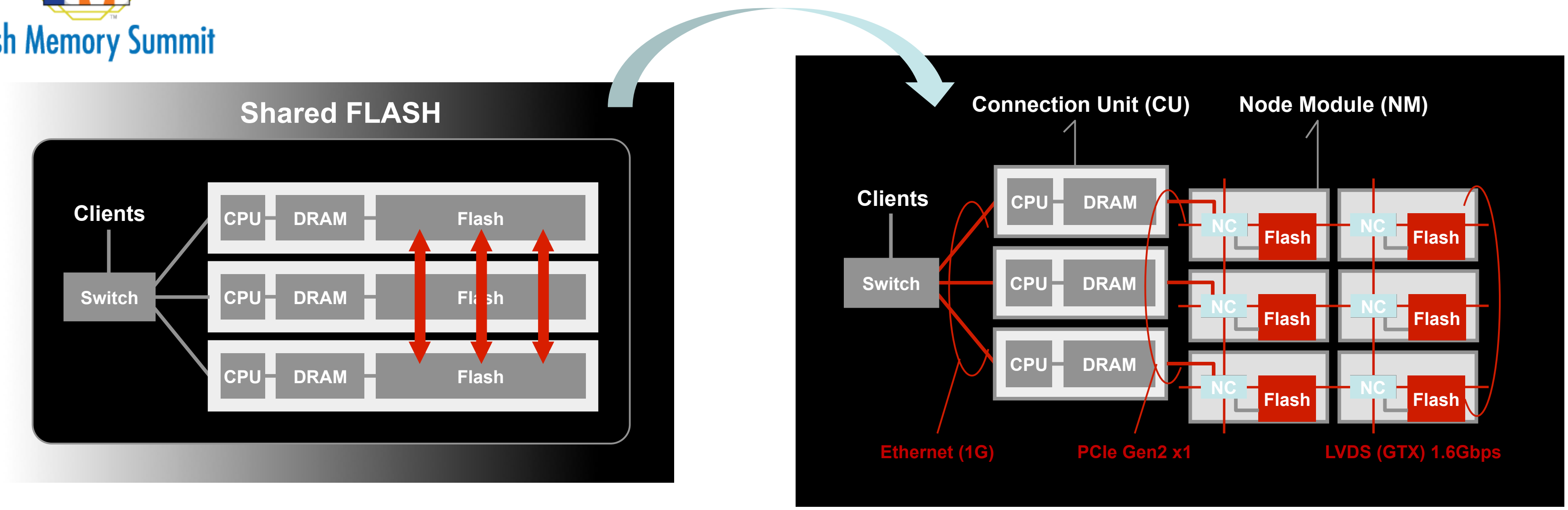
Possible Solutions

The best way to address the "Big Data" problem is to enable any CPU core access the entire Storage via





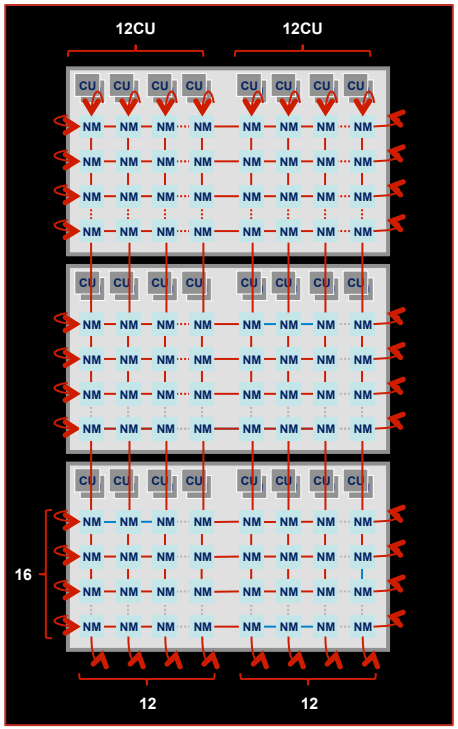
Shared Everything Architecture



Eliminates Shortcomings of "Software Defined Storage" Namely:

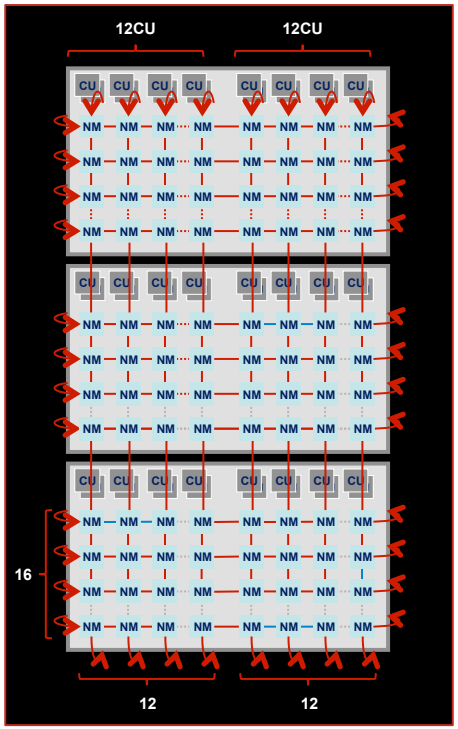
- Switching CPU cores latency
- Performance degradation due to data location and application work loads

2D Torus Network Implementation



- Topology allows for high traffic
- Data can be transmitted simultaneously from different nodes resulting in high speed and low latency.
- Scalability achieved without system disturbance & it is theoretically “unlimited” practically up to 14 enclosures.

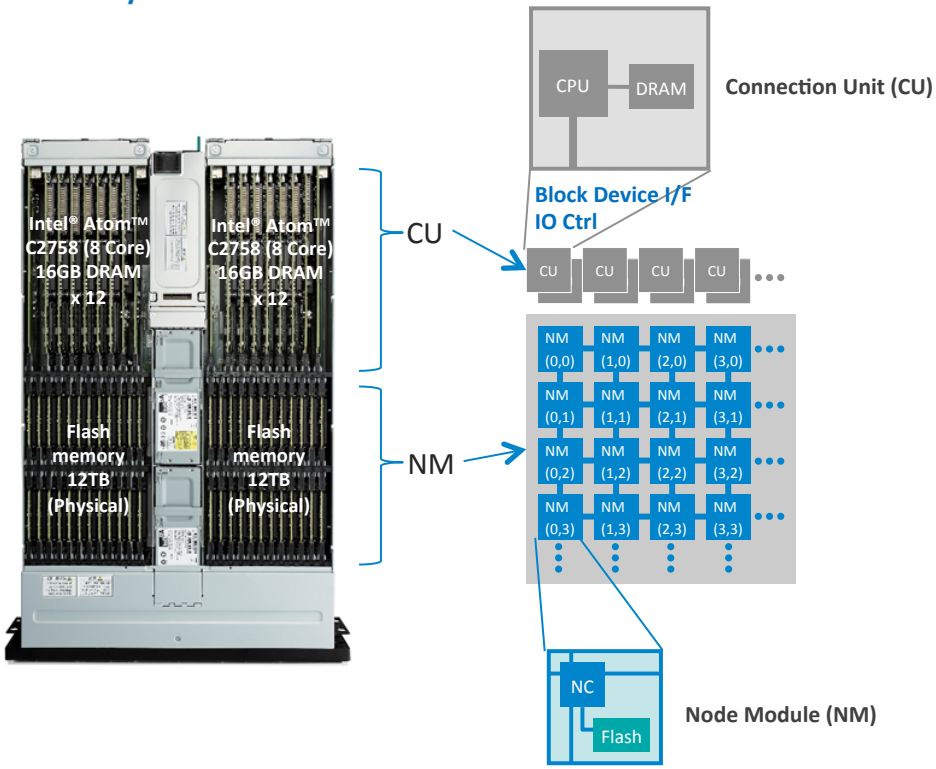
2D Torus Network Implementation



- Single node failure does not affect data transfers.
- Presents a more efficient routing & provides better power efficiency due to less data transfers.
- Removes some of the switches used in a conventional clusters, which are points of failure, thus increases reliability.

Flash Memory Summit

Flashmatrix: Super-Converged Platform



> Orchestrated Containers

- The Super Orchestration Layer (SOL) enables stateless Docker® container apps for the edge.
- SOL seamlessly provisions, de-provisions and resurrects your apps based on system demand.
- Flashmatrix and SOL are co-optimized for edge computing use cases and workloads.

> Shared Everything All Flash

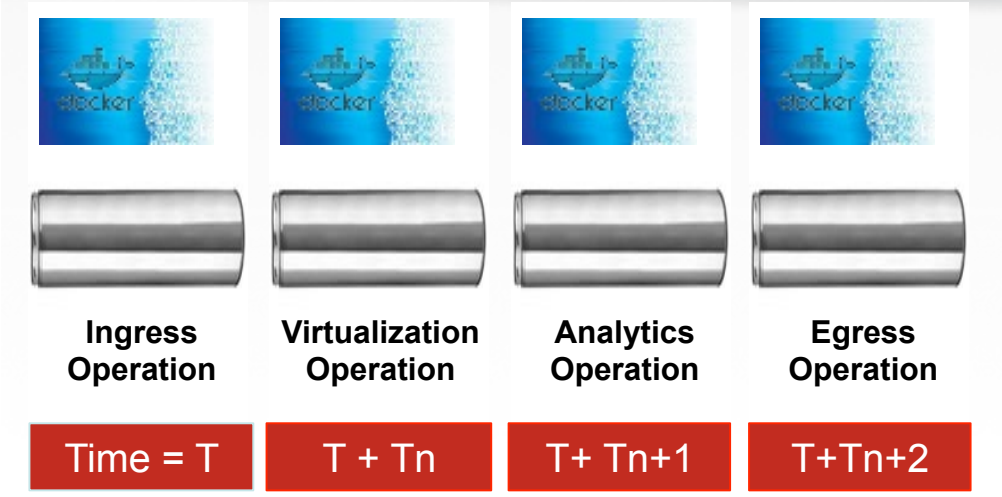
- 24 terabytes of flash, 24 microservers, 192 CPU cores, all converged in a 2U enclosure.
- Toshiba's "shared everything" architecture enables all data to be accessible to all compute.
- Storage is fully virtualized and integrated into a converged mesh network switch.

> Highly Scalable

- Expand compute and storage to meet your growing demands (14 enclosures in one rack).
- Grow shared memory and compute without degradation in performance.

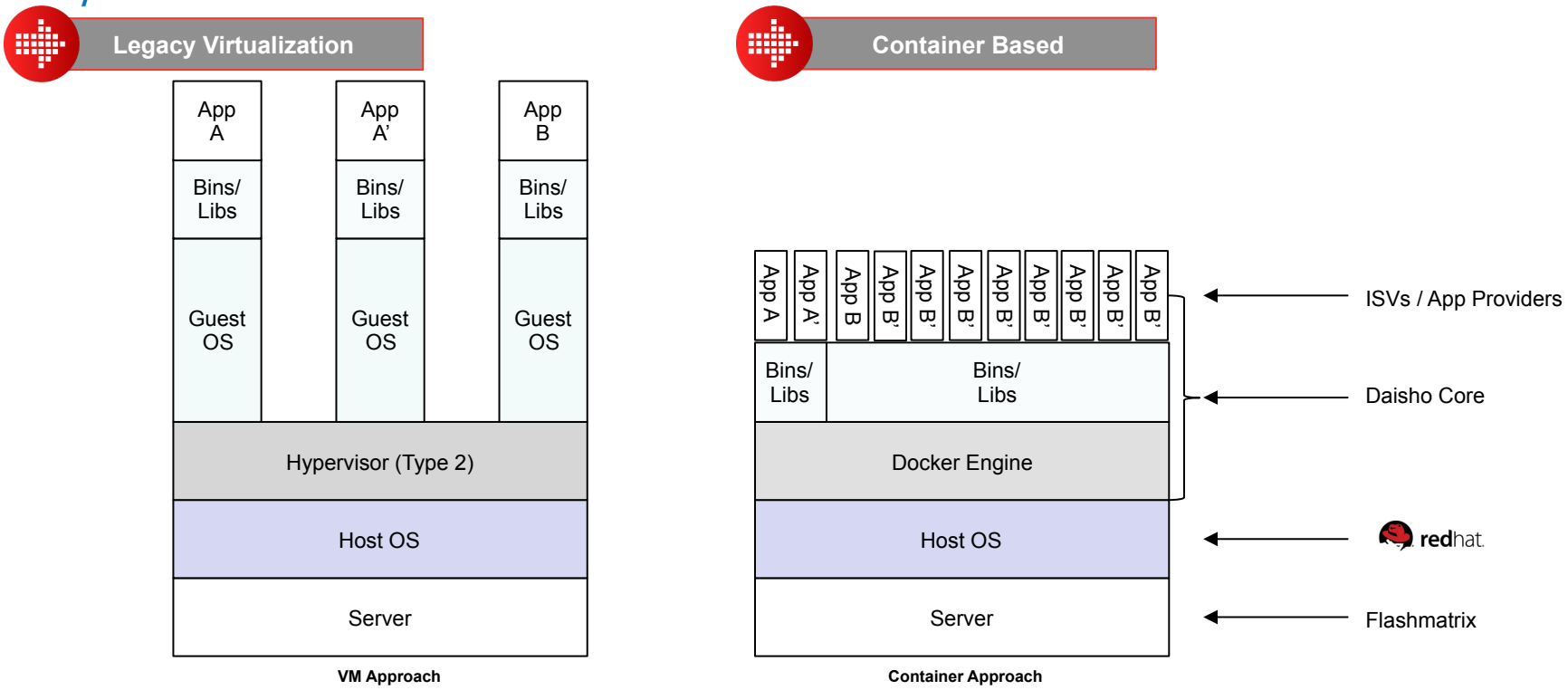
The Hidden Distributed Application Latency

Applications requiring multi-node operations carry a hidden latency in todays “Non Shared Storage” architectures



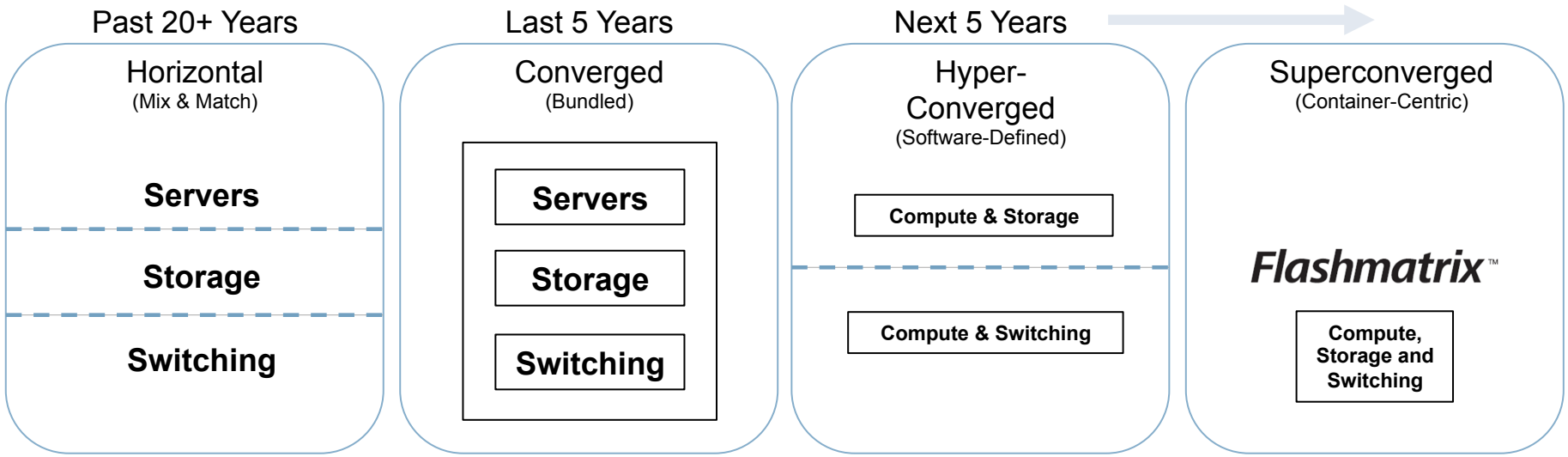
Latency Eliminated By Shared Memory Systems

Simplification of the Stack



Optimized for Multiple Microservers in Flashmatrix

Flashmatrix Driving Change in the IT Consumption Model



TOSHIBA
Leading Innovation

Flashmatrix Overview

Super-Converged for the Edge that's…

> Shared Everything All Flash

- 24 terabytes of flash, 24 microservers, 192 CPU cores, all converged in a 2U enclosure.
- Toshiba's "shared everything" architecture enables all data to be accessible to all compute.
- Storage is fully virtualized and integrated into a converged mesh network switch.

> Orchestrated Containers

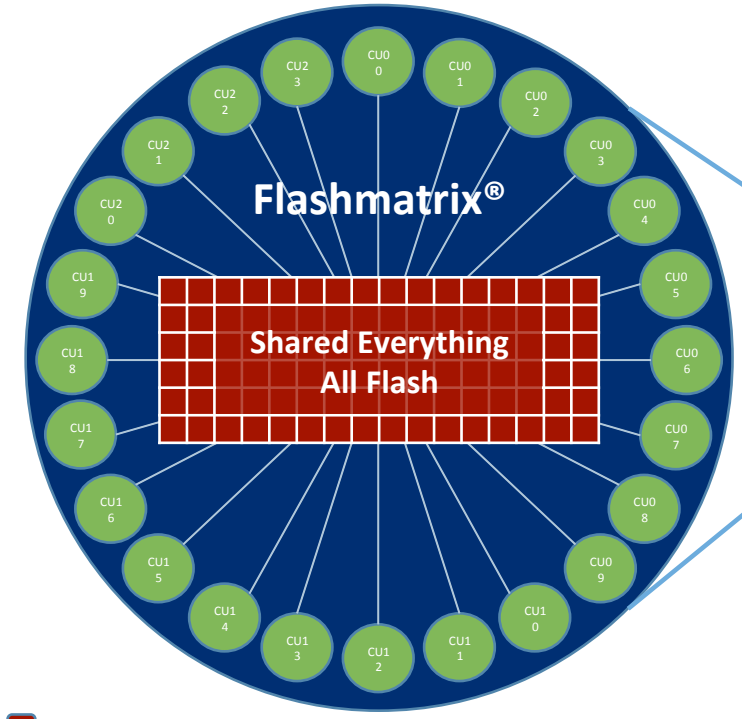
- Daisho Core enables stateless Docker® container apps for the edge.
- Daisho Core seamlessly provisions, de-provisions and resurrects your apps based on system demand.
- Flashmatrix and Daisho Core are optimized for edge computing use cases and workloads.

> Highly Scalable

- Seamlessly scale edge mesh networks with 14 enclosures in a single rack
- Massive capability at the edge with 2,688 cores and 336TB per rack
- Large data throughput for scalable applications and microservices

Super-Converged for the Edge: Shared Everything All Flash

Flash Memory Summit



24 Terabytes Shared Flash
24 µServers

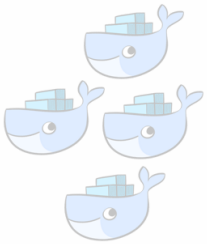
Super-Converged for the Edge: Orchestrated Containers

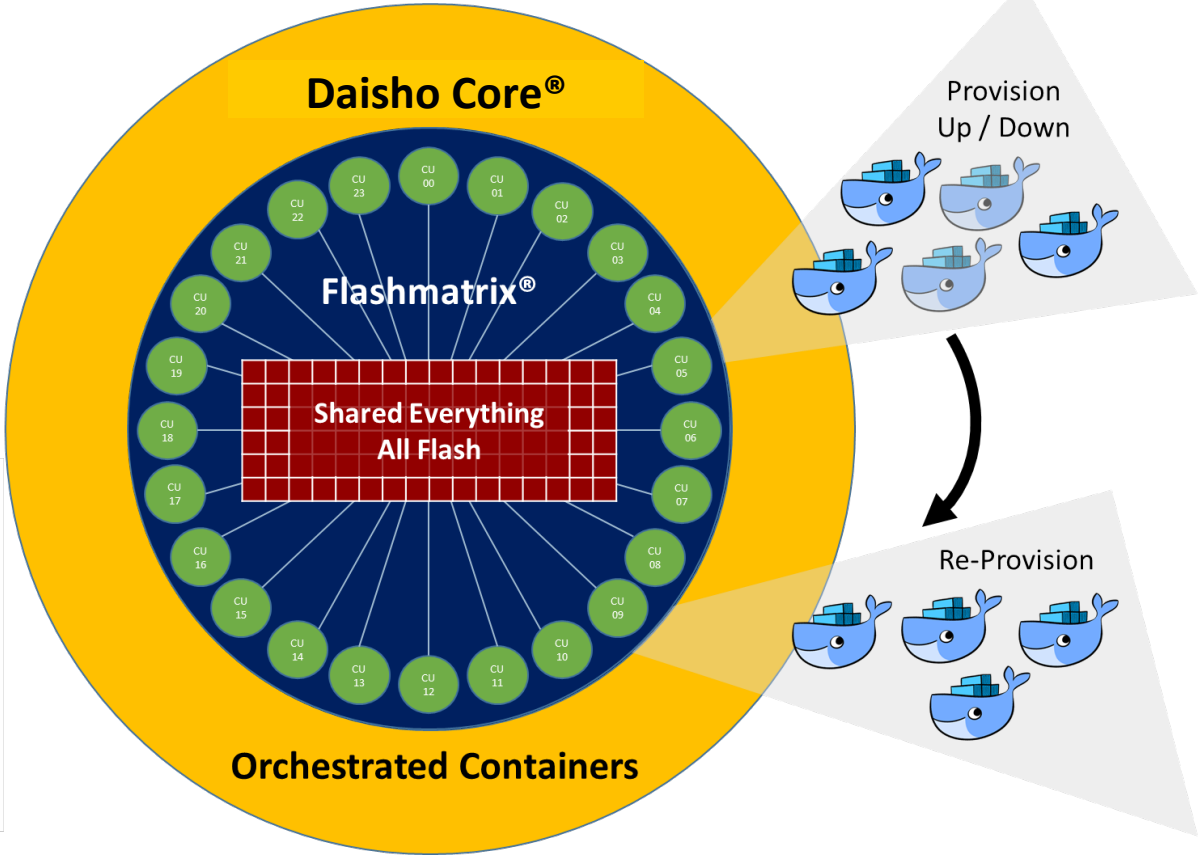
Flash Memory Summit

100s of Containers

1,000s of Apps

De-Provision

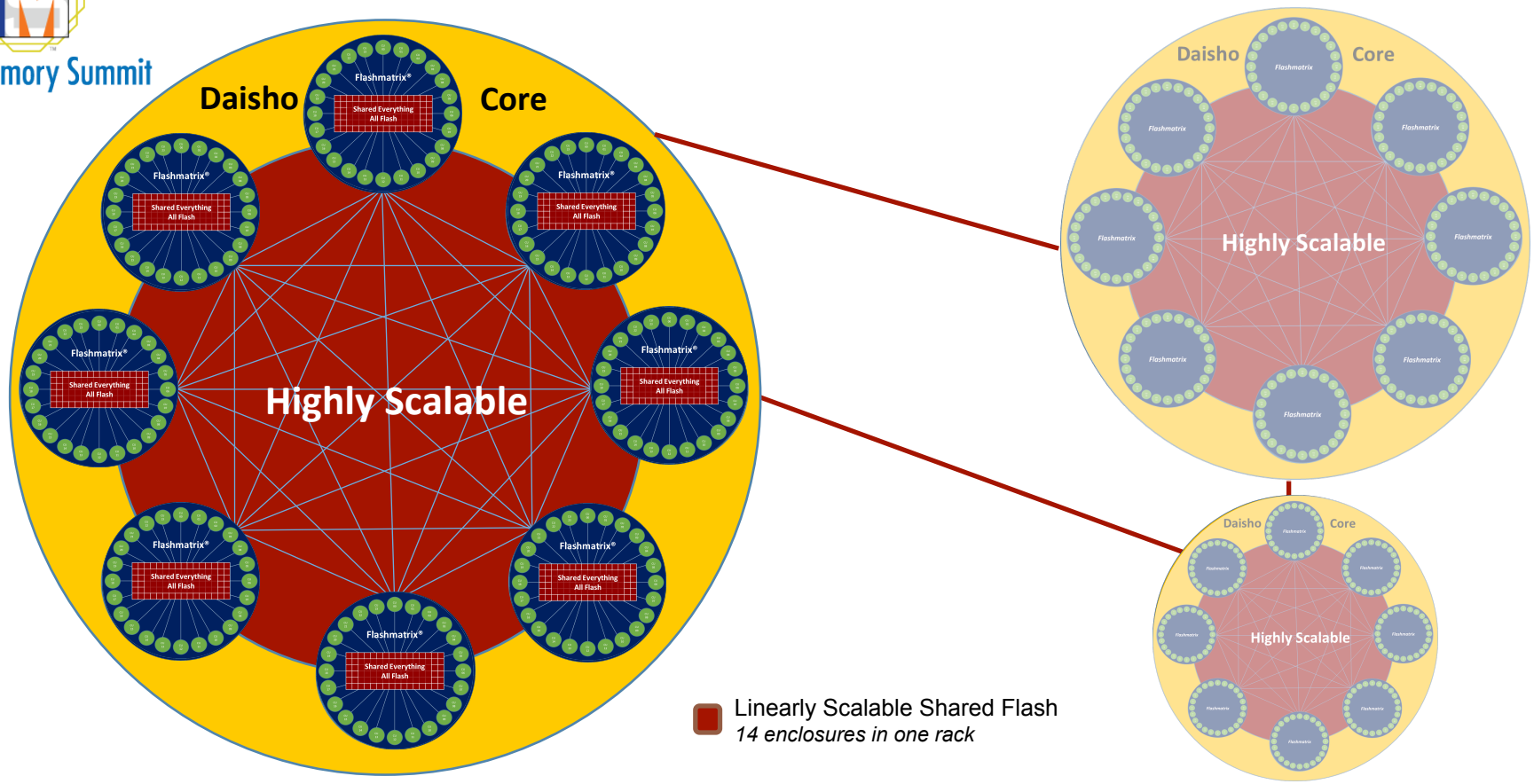




Provision
Up / Down

Re-Provision

Super-Converged for the Edge: Highly Scalable



TOSHIBA

Leading Innovation