Realizing the next step in storage/converged architectures

“Imagine having the same data access and processing power of an entire Facebook like datacenter in a single rack of servers”

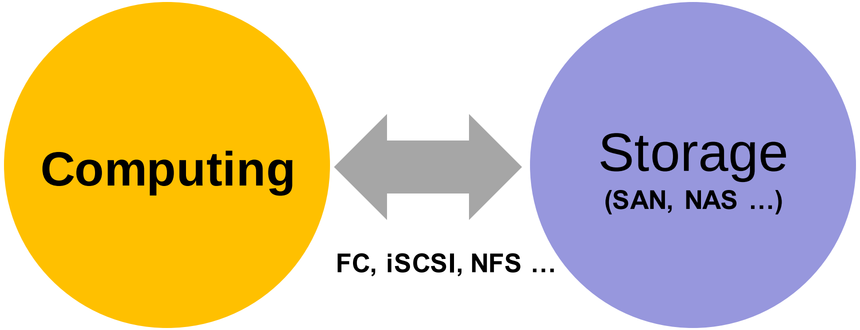
E.Billi
A3CUBE CTO

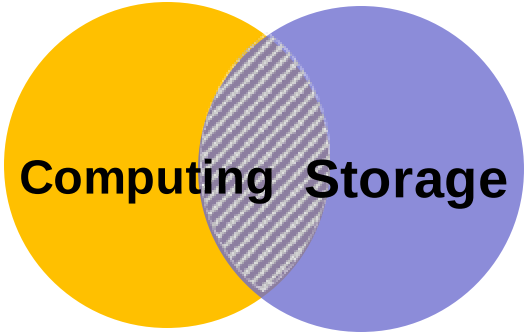
The Storage Paradigm Shift

The role of storage has changed dramatically from being a simple data repository to now being an active part of the computation itself

Architecturally speaking, this forces us to think differently about how to build new generations of active storage systems that fit with modern software requirements:

Convergence, Parallelism, Low Latency, High IOPS





The Storage Paradigm Shift

Unfortunately, all existing storage approaches, including software defined only, are falling significantly short of their promise:

- Low parallelism
- No Convergence
- High latency network based scale out approaches
- High latency software defined architecture

Fortunately, the storage industry is introducing new disruptive devices based on Flash technologies:

- Fast
- Low Latency
- High IOPS

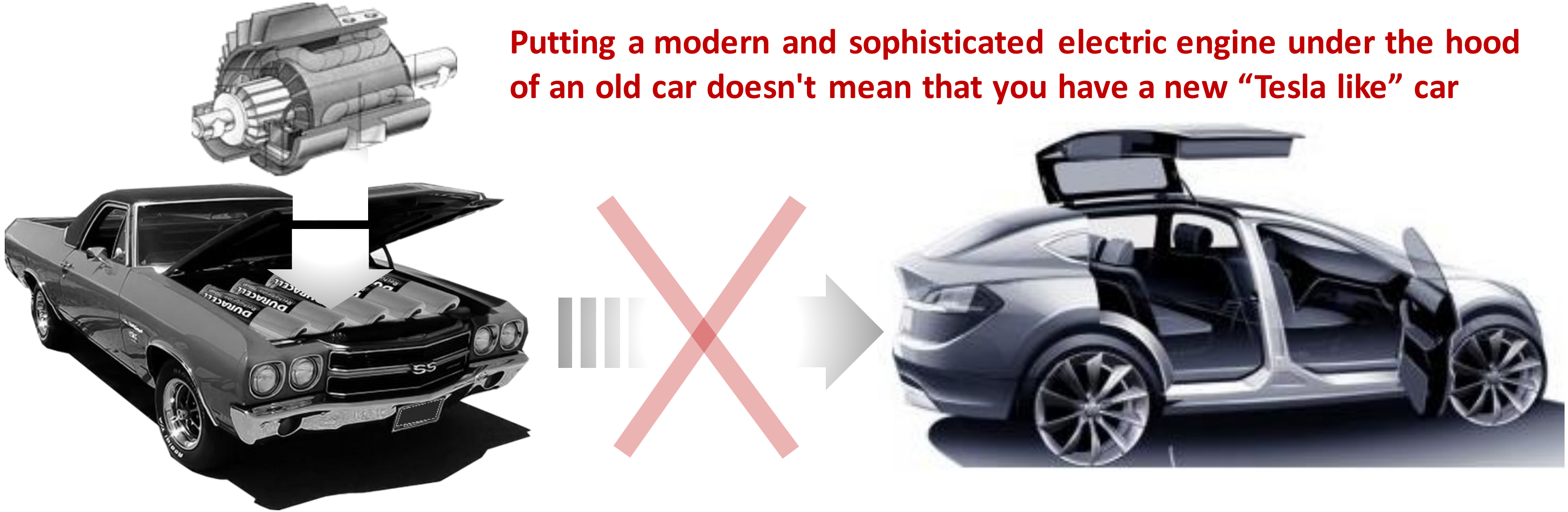
Unfortunately, current architectures are not able to utilize Flash technology to its full potential:

- Current market approaches rely on legacy architecture in combination with new flash technologies while claiming to be the next future storage architecture

The Storage Paradox

Why we need a new architectural approach instead of software only solutions

Putting a modern and sophisticated electric engine under the hood of an old car doesn't mean that you have a new “Tesla like” car



By putting new technologies (SSDs, NVMe, and adding only complex software) under the hood of an old storage scale OUT architecture doesn’t create a new system and doesn't extoll the potential of emerging technologies but …

Understanding the future

Modern software needs a converged approach
Latency in data access is now a critical factor

Emerging software architectures are massively parallel and require parallel access to data (Hadoop, Storm, Greenplum)

Existing storage architectures are not designed to meet these requirements!

Understanding the future

Today, capacity and throughput are “commodities”, while latency has become the new performance metric

Emerging technologies (Flash Drive, NVMe Flash, In-Memory Data Architectures) demonstrate how high the latency impact is on the performance of an application

Existing (Software only) architectures are not able to maintain good performance and low latency when they scale, as the software requires (Think about iSCSI, FC, ETH, …)

What is latency and Why it is so important

Latency is the most critical performance factor because it directly affects system data exchange time. In fact, latency means **losing time**; time that could have been spent more productively producing computational results, but it is instead spent waiting for I/O resources to become available.
Latency is the "**application stealth tax**", silently extending the elapsed times of individual computational tasks and processes, which then take longer to be execute. (*)

In a world in which data growth is disruptive (more than 5 Exabyte of content are created each day) response time is critically important.
Higher levels of application performance can be easily achieved with low latency platforms, on which new applications like machine learning and artificial intelligence can perform on much grander scales than ever before.

(*) Remember adding software stacks = Adding Latency to the system

Flash gives the opportunity to solve the latency problem

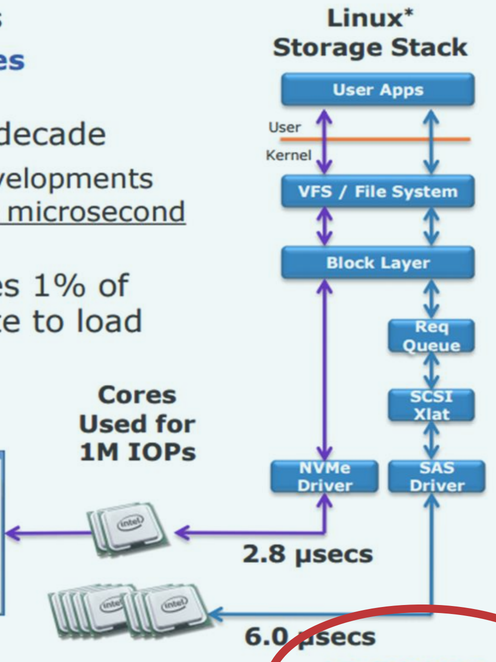
NVMe Flash Latency from Intel

2.8 us (hardware level)

- NVMe reduces latency overhead by **more than 50%**
  - SCSI/SAS: 6.0 µs 19,500 cycles
  - **NVMe: 2.8 µs 9,100 cycles**
- NVMe is designed to scale over the next decade
  - NVMe supports future NVM technology developments that will drive latency overhead below one microsecond
- Example of latency impact: Amazon* loses 1% of sales for every 100 ms it takes for the site to load

Chatham NVMe Prototype

Prototype Measured IOPS



IDF2012 INTEL DEVELOPER FORUM

NVMe = NVM Express
Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat* EL6.0 2.6.32-71 Kernel.

Why current scale out and converged approaches fail

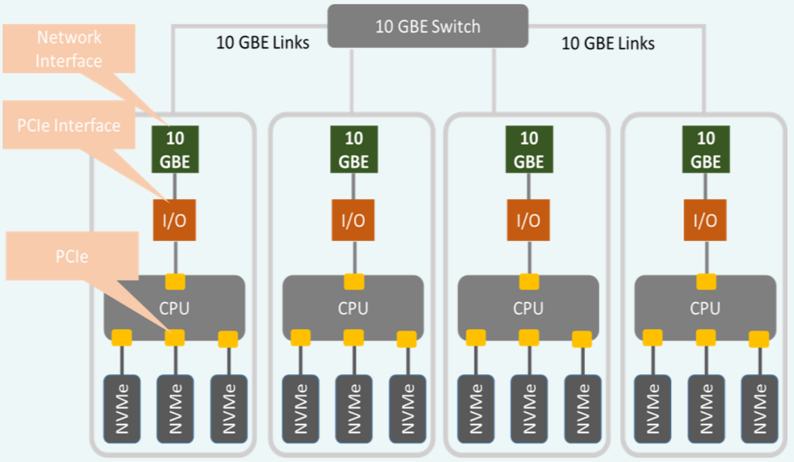
Storage is a complex architecture that includes multiple dimensions such as capacity, bandwidth, IOPS and massive scalability.
A *Scale-Out* storage architecture is certainly the right approach for addressing data problems within (present and future) datacenters, but need to be executed in a correct way!

Inter Storage Latency!

Intel 82598EB 10 Gigabit AT CX4

Message Size	One Way Latency 10GigE
1	20.0
4	20.0
8	19.9
12	20.0
16	20.0
24	20.0
32	20.0
48	20.0
64	20.1
96	51.3
128	51.3
256	51.2
512	51.2
1024	49.9
2048	62.6
4096	125.0
8192	124.9
16384	125.0

Standard Network Interconnection



Typical Scale out modern Approach

Software Defined Storage most common mistake

Why current scale out and converged approaches fail

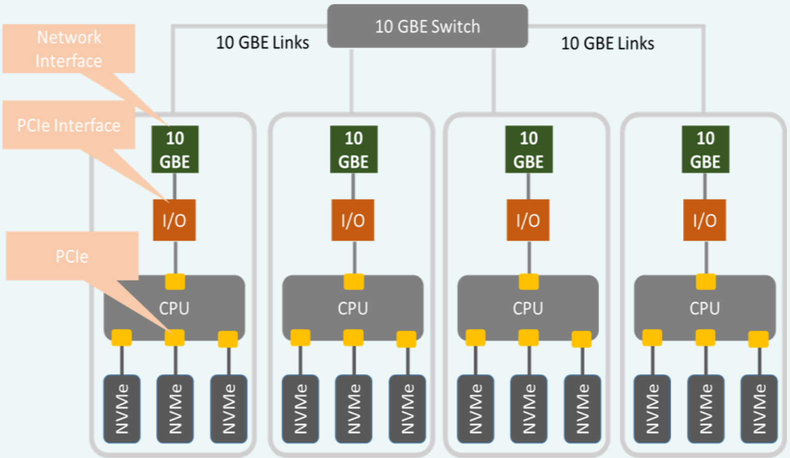
Storage is a complex architecture that includes multiple dimensions such as capacity, bandwidth, IOPS and massive scalability.
A *Scale-Out* storage architecture is certainly the right approach for addressing data problems within (present and future) datacenters, but need to be executed in a correct way!

Inter Storage Latency!

Intel 82598EB 10 Gigabit AT CX4

Message Size	One Way Latency 10GigE
1	20.0
4	20.0
8	19.9
12	20.0
16	20.0
24	20.0
32	20.0
48	20.0
64	20.1
96	51.3
128	51.3
256	51.2
512	51.2
1024	49.9
2048	62.6
4096	125.0
8192	124.9
16384	125.0

Standard Network Interconnection



Typical Scale out modern Approach

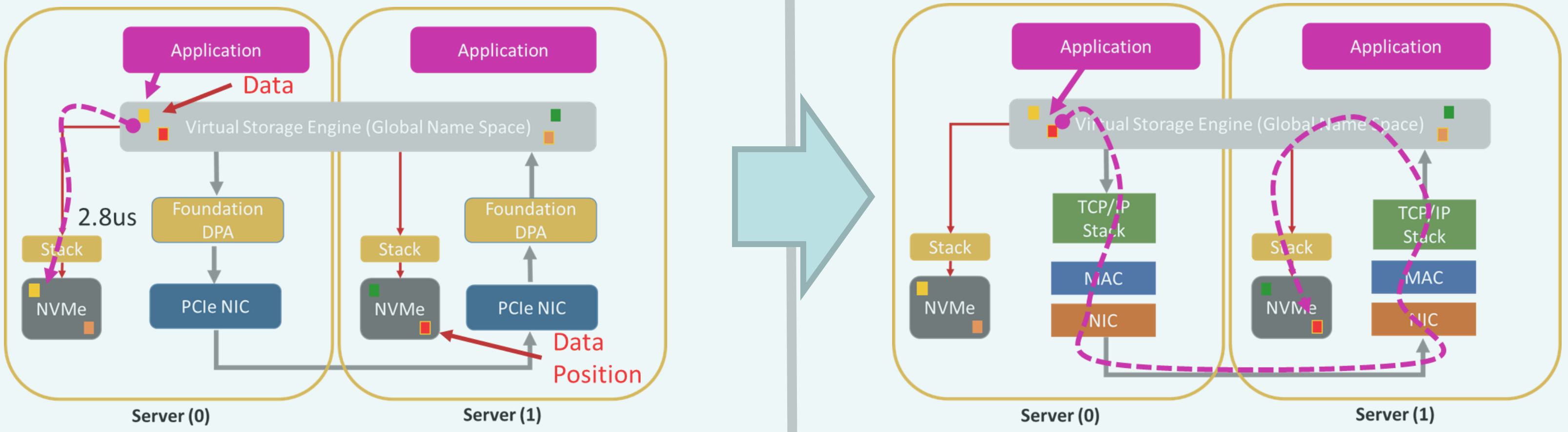
Software Defined Storage most common mistake

The problem with existing approaches in a simple picture!

Local App to Local NVMe

Local & Remote Access Time Comparison

Local App to Remote NVMe



Latency NVMe Stack **(A) 2.8 us @ 0 Byte**

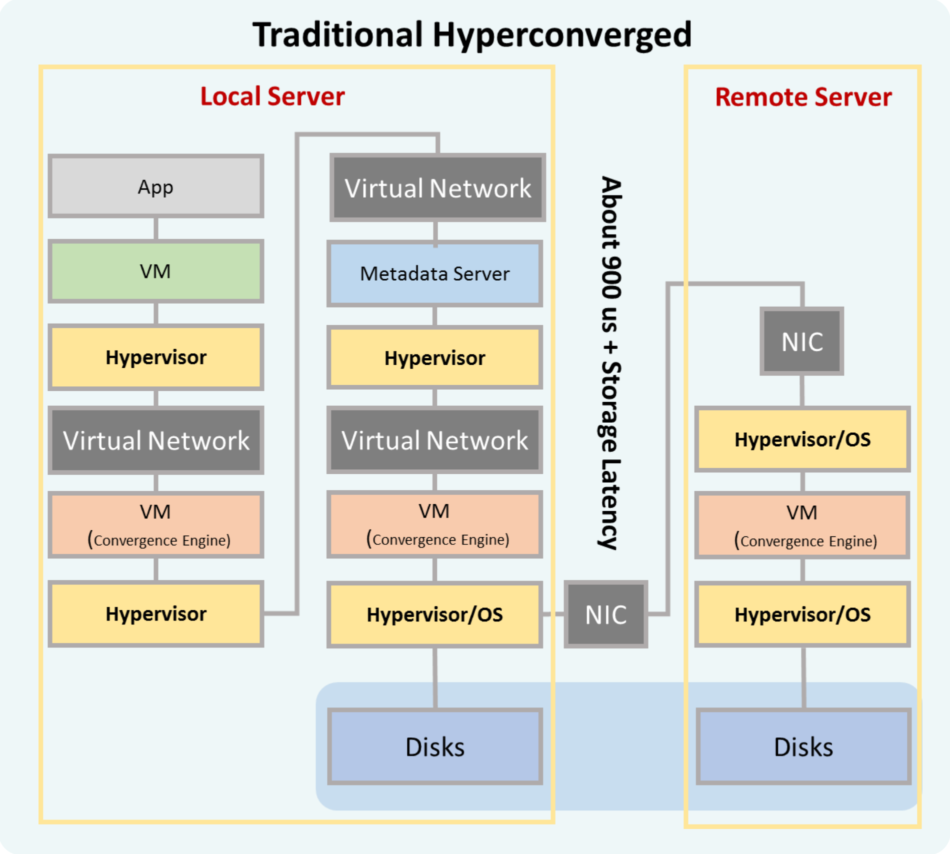
(A) 38.8 remote access @ 0 Byte (Over sockets 10GBE)

(A1) 34.30us remote access @ 0 Byte (Socket Over PCIe (PLX))

14x performance degradation

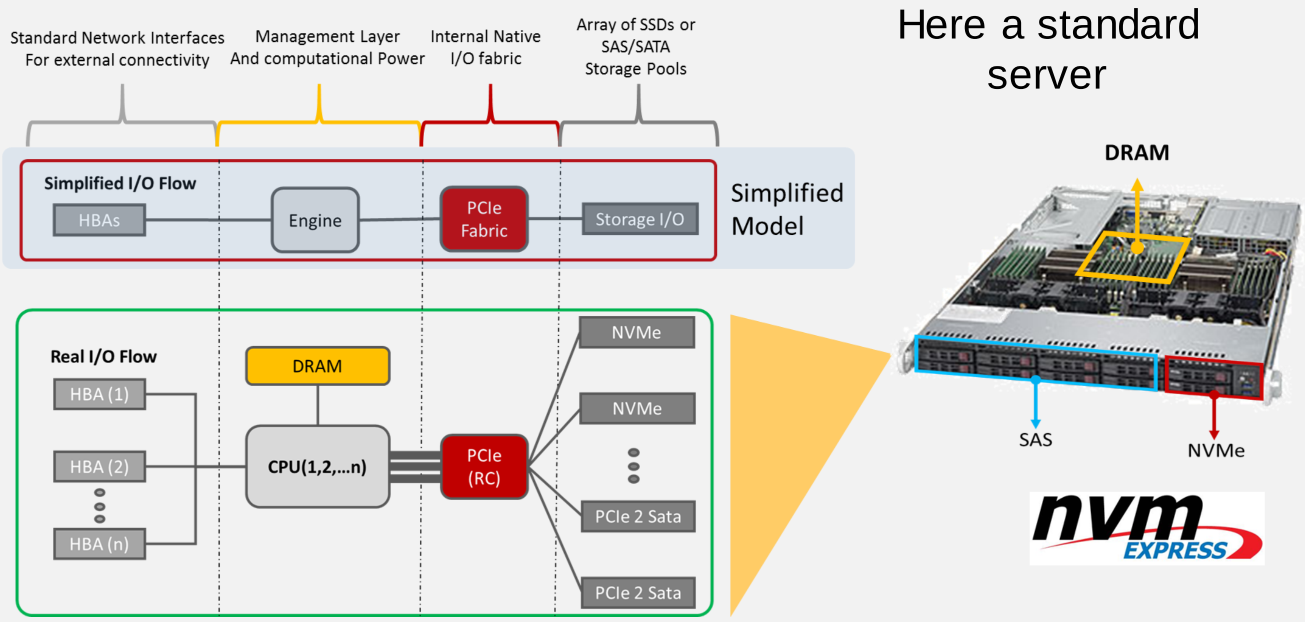
Network Latency serious impact on overall performance

Hyper-converged is an even worse situation



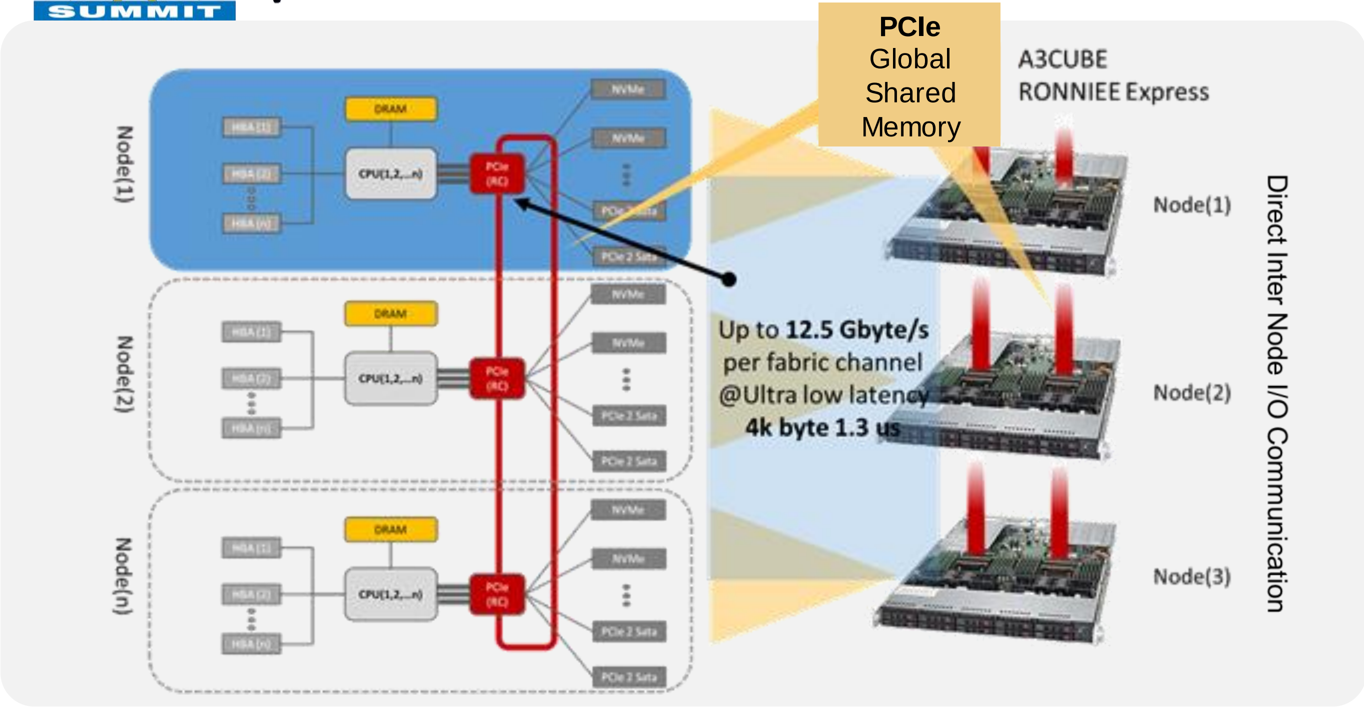
- In Hyper-converged solutions the latency is so high that these kind of approaches will never be viable
- Extremely bad performance, ultra low efficiency, NO real benefits
- Up to 100x slower than non Hyper-converged solutions

The solution: How it Works



Modern Server Architecture

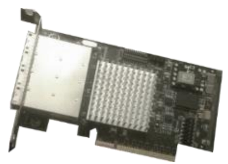
The solution: How it Works



nvm EXPRESS

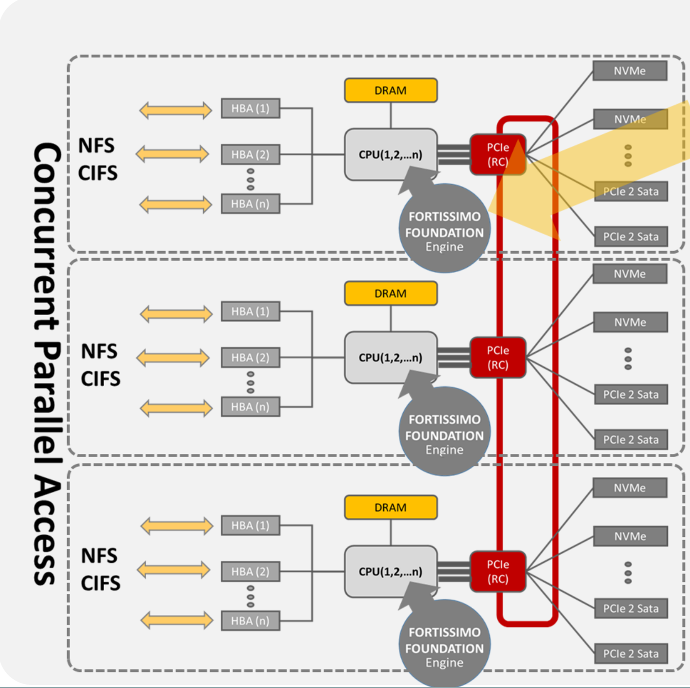
+

PCI Express
Direct I/O accelerator and shared memory generator



We aggregate the storage nodes with RONNIEE Express a PCI Express based shared memory fabric and IO accelerator used as a virtualized, flexible, rugged, unified, single network for all types of storage IO communications

The solution: How it Works



Fortissimo Foundation

No Metadata and global unified name space

Multivolume/ Multiuser with multilevel of HA

Multilevel Tiering (RAM, NVMe, SSDs ...)

Unique Direct "RamStor" with HA on SSDs and auto-recovery (Global RAM Storage)

Linear Scalability in Capacity and Performance in the same system.

Scale UP and Scale OUT without compromise.

nvm EXPRESS

+

A3C's RONNIEE Express Direct I/O

+

Hardware Accelerated IO Virtualization

Fortissimo Foundation, is a Hardware accelerated Software Defined System (HSDS™) that consolidates the computer-tier and the storage-tier into a single integrated extremely fast parallel storage and converged platform

Real Product Example

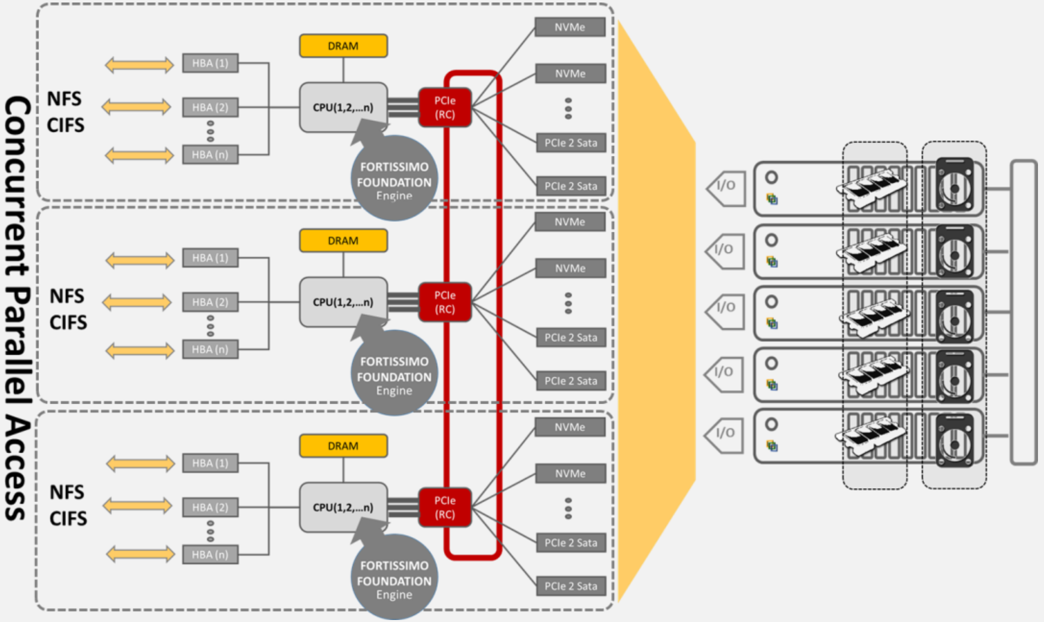
nvm EXPRESS

+

A3C's RONNIEE Express Direct I/O

+

Fortissimo Foundation

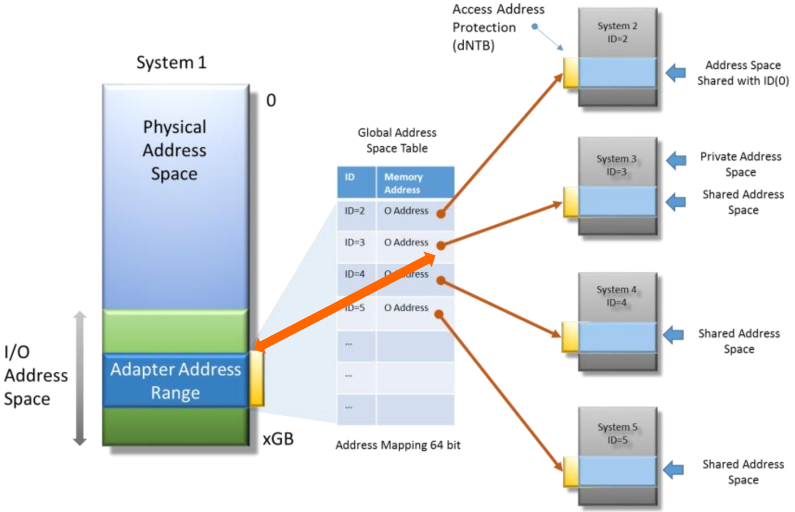


Data access and processing power of a Facebook-like datacenter in a single rack of servers

Direct Storage Nodes Communication

a³cube

Direct System to System / Memory to memory latency & bandwidth

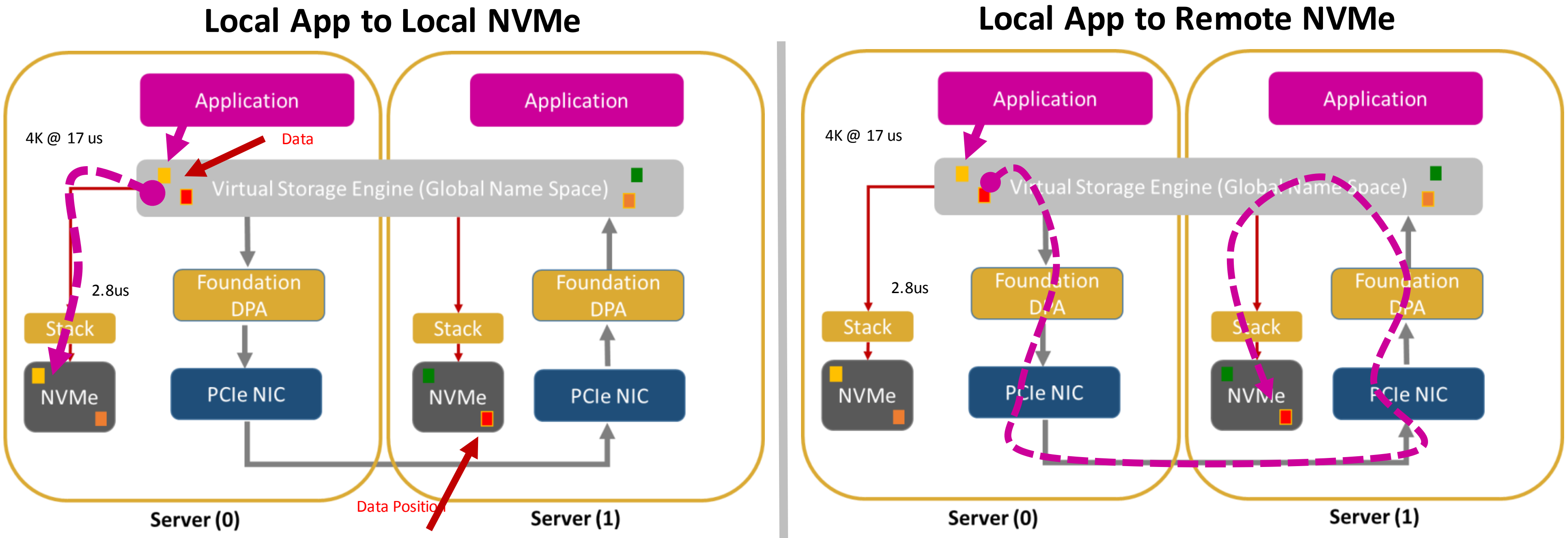


Message	Latency/2	Bandwidth
64	0.75 us	305.54 MBytes/s
128	0.80 us	619.46 MBytes/s
256	0.82 us	1198.27 MBytes/s
512	0.86 us	2308.34 MBytes/s
1024	0.88 us	4443.57 MBytes/s
2048	1.12 us	5924.76 MBytes/s
4096	1.30 us	6061.34 MBytes/s
8192	1.88 us	6137.05 MBytes/s
16384	2.65 us	6180.02 MBytes/s
32768	5.28 us	6210.10 MBytes/s
65536	10.51 us	6233.58 MBytes/s
131072	20.98 us	6245.99 MBytes/s
262144	42.11 us	6225.88 MBytes/s
524288	83.79 us	6257.32 MBytes/s

Remove the latency problem

Direct Storage Nodes Communication

Local & Remote Access Time Comparison



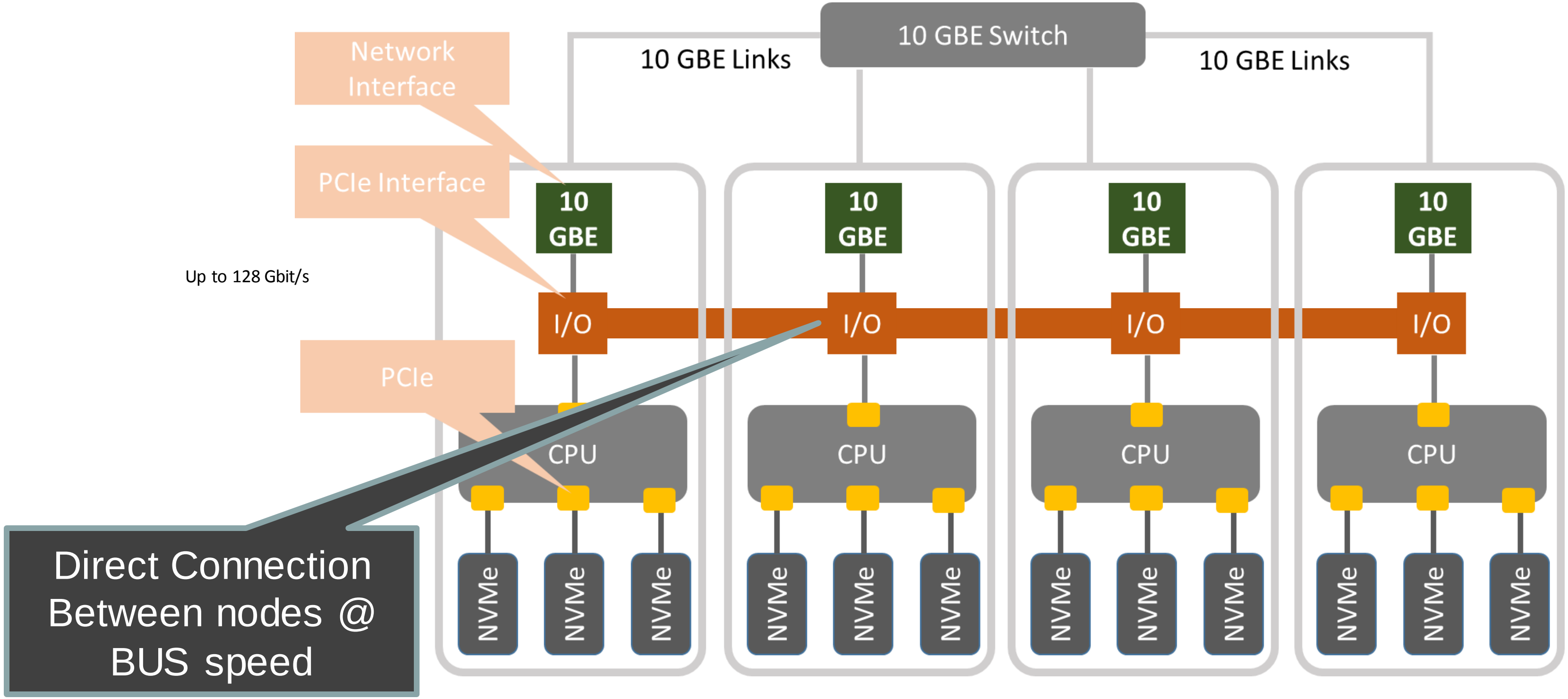
Latency NVMe Stack **(A)** **2.8 us**

3.5 us remote access (Direct PCIe Memory Access)

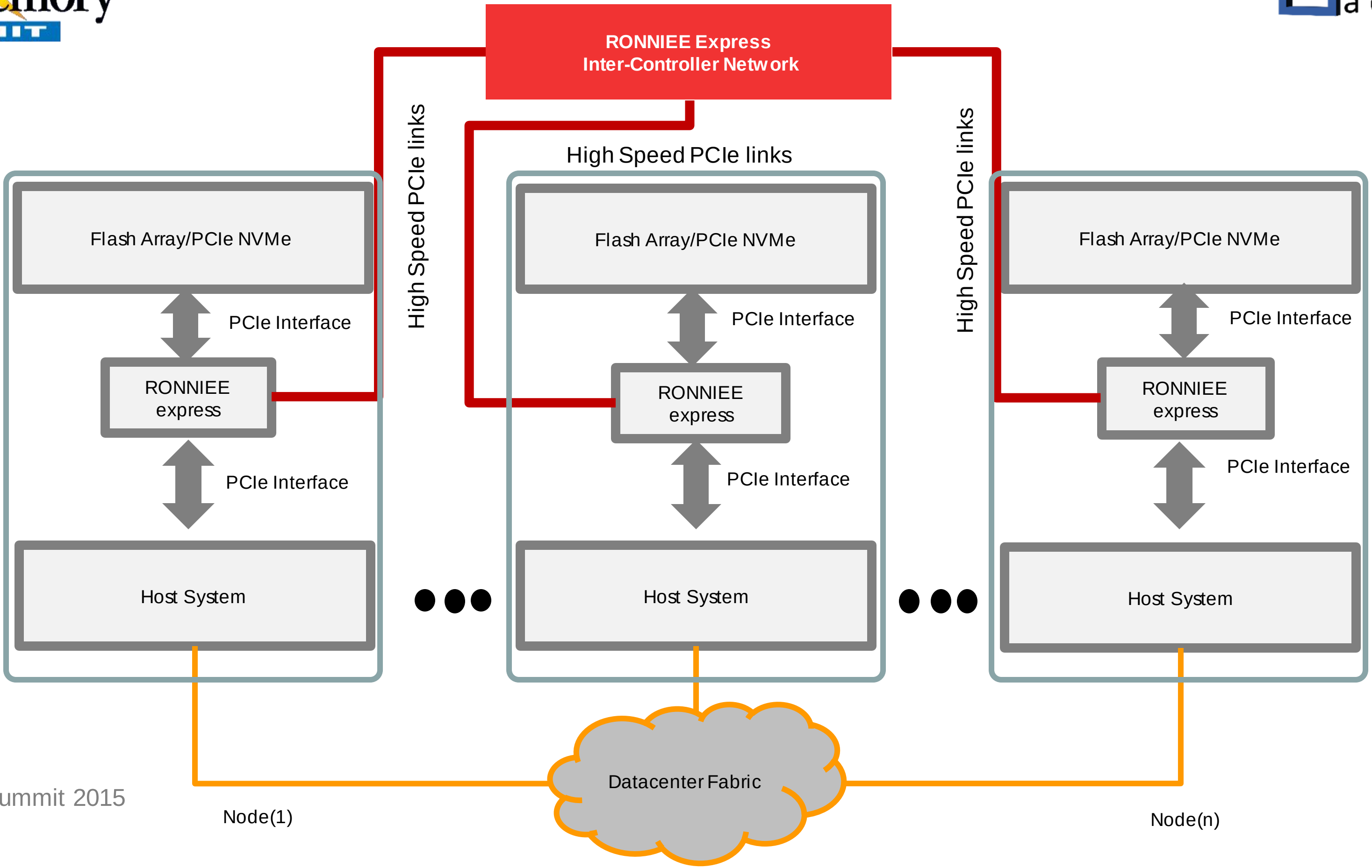
No performance degradation (Flash Drives side)

Direct Storage Nodes Communication

Network used for datacenter operation and not for storage



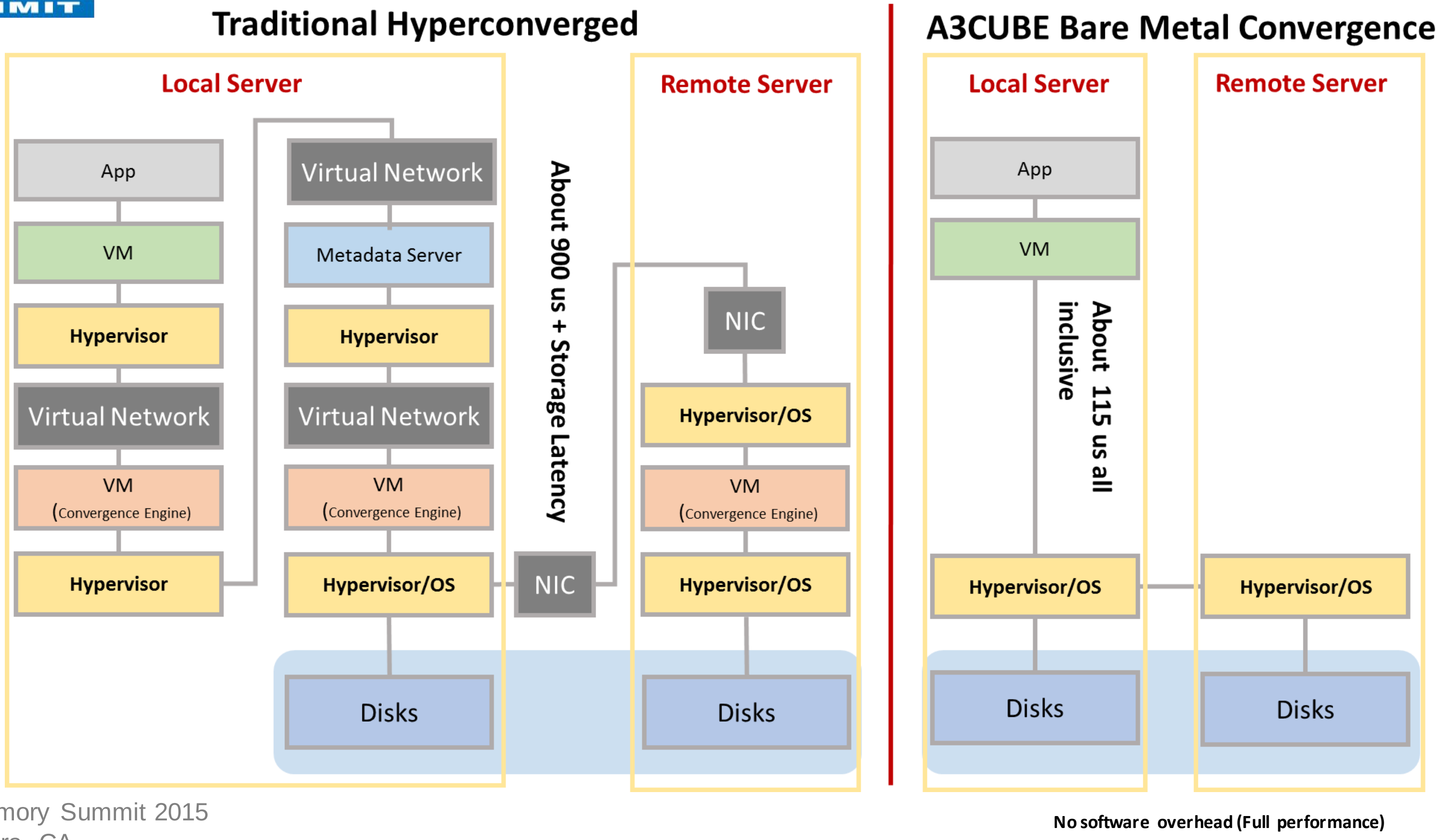
Or more in detail Works



Direct Storage Nodes Communication

- **NVMe and emerging storage technologies are PCIe driver**
- **PCIe and NVME support SR-IOV (Hardware virtualization sharing support)**
- **PCI Express memory mapped fabric is used as a virtualized, flexible, rugged, unified, single network for all types of communication**
  - Networking (Hardware Accelerated SDS)
  - Host to Host
  - IO Expansion
  - Host to IO
  - Peer to Peer - IO to IO
- **Network attached resources can be attached, migrated and removed**
- **Data flows direct between active components**
- **Take advantage of DMA, PIO, and NTB functions within PCI Express (No device modification (e.g. NVME use its driver unmodified)**

Realizing Bare Metal Convergence

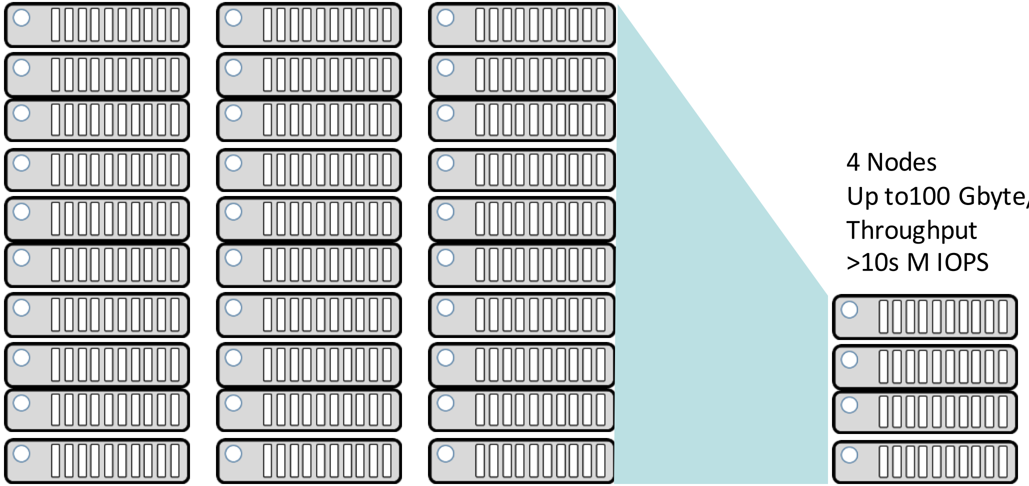


New Level of Efficiency

Realize the processing power of an entire large datacenter in just a few servers

Up to 100x performance or up to 10sx less hardware

100s nodes, < 100 Gbyte/s Throughput, < 10 M IOPS



4 Nodes
Up to100 Gbyte/s Throughput
>10s M IOPS

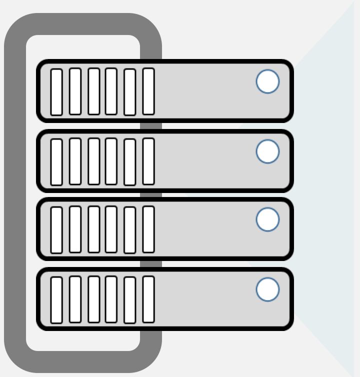
Less hardware, Less CAPEX, Less OPEX, Less complexity
Higher performance

New Level of Efficiency

Fortissimo Foundation

All NVMe or Hybrid

Converged System



Per Single Rack

Up to > 800 Gbyte/s of data access bandwidth

Up to 100s of M IOPS

1.3 us @ 4 Byte packets of internode latency

3 level of ultra fast automatic caching

We Provide a turn key solution for any of these applications

Converged Computation

Run your application with no bottleneck in IO access, Run Hadoop with 100x time faster IO (Run any application at the speed of memory)

Converged Virtualization

Run your VMs at a speed that you never imagined

Run in your VMs application at the bare metal speed!

Legacy Parallel NAS

Access to your data with millions of IOPS and more faster that you can imagine

Ultra fast System

Supercharge, far beyond your expectations, existing infrastructures maximizing the investment and remain competitive well into the future!

Some picture to visualize …

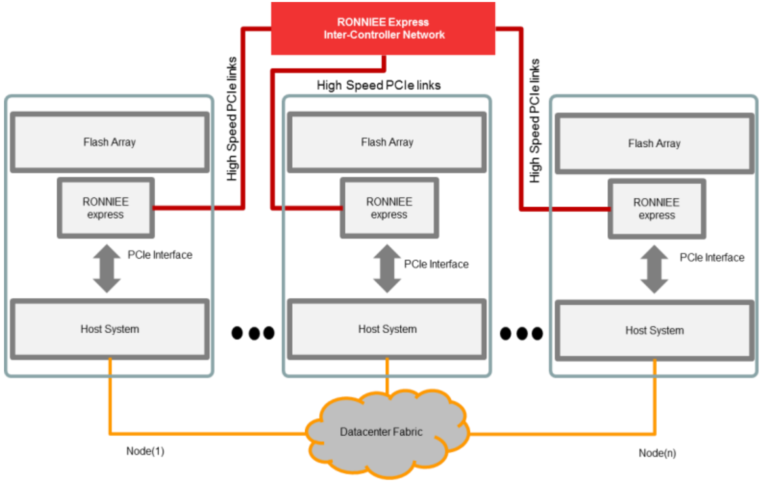
Two type of NVME drivers

Standard NVME controller

2.5” Drive

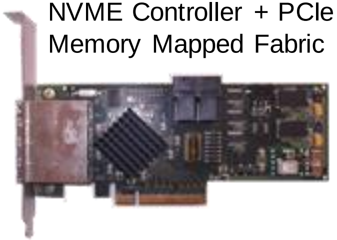
PCIe Form Factor

A Real Implementation



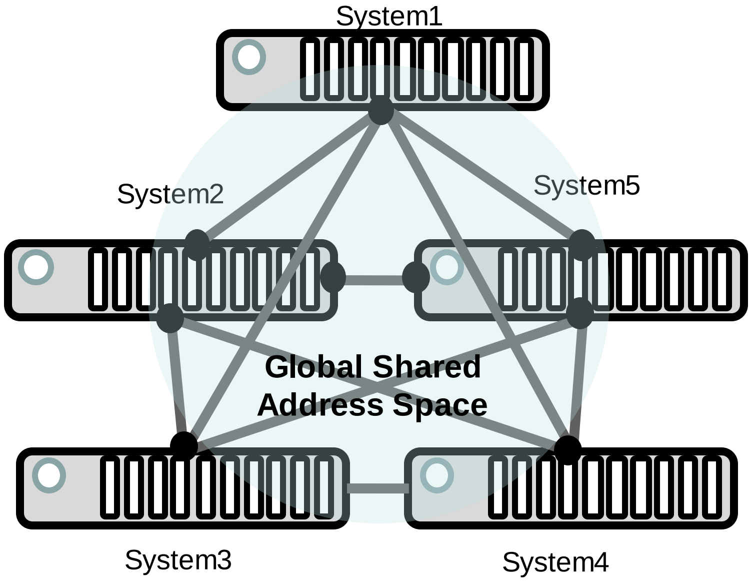
RONNIEE Express Family

NVME Controller + PCIe Memory Mapped Fabric

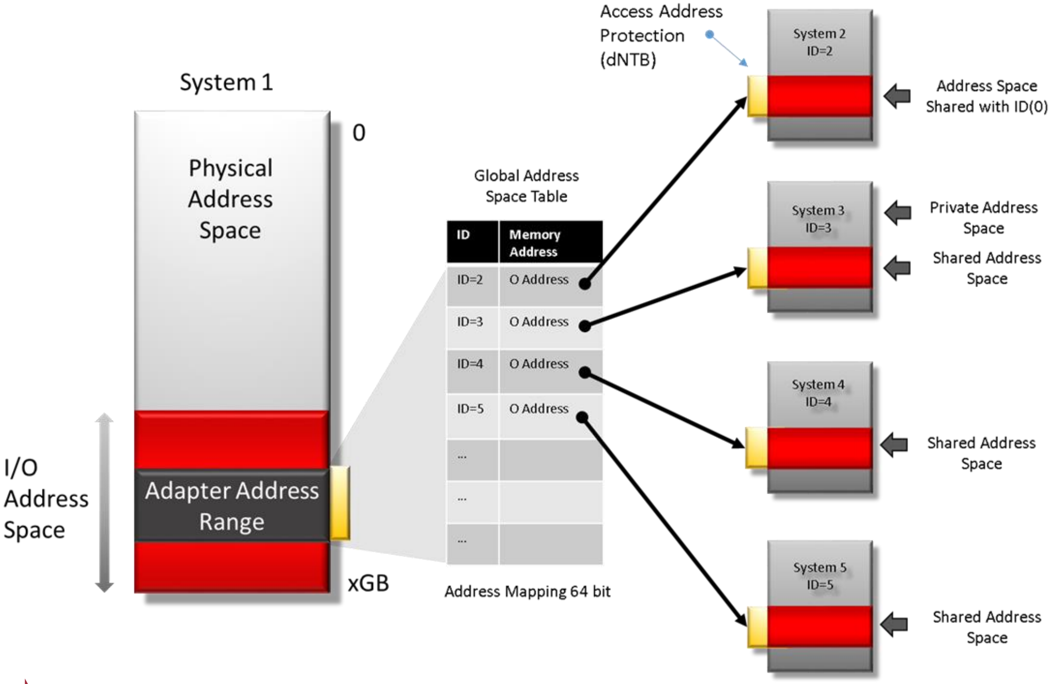


Universal PCIe Memory Mapped Fabric (only)

Some Topologies Supported

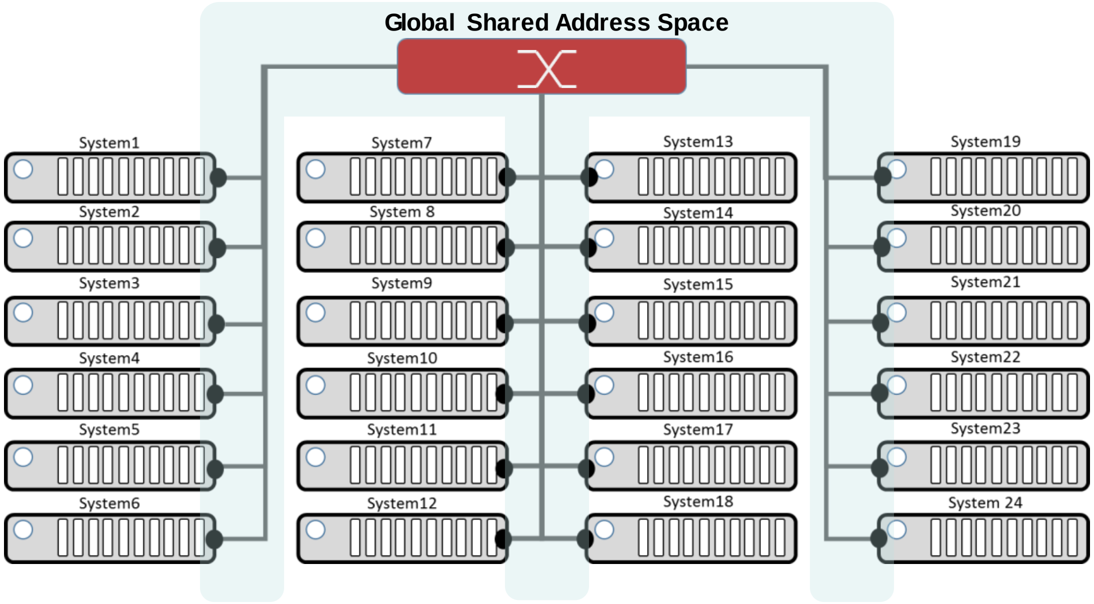


4 Bytes Latency 740ns (0.74 us) ultra low jittering



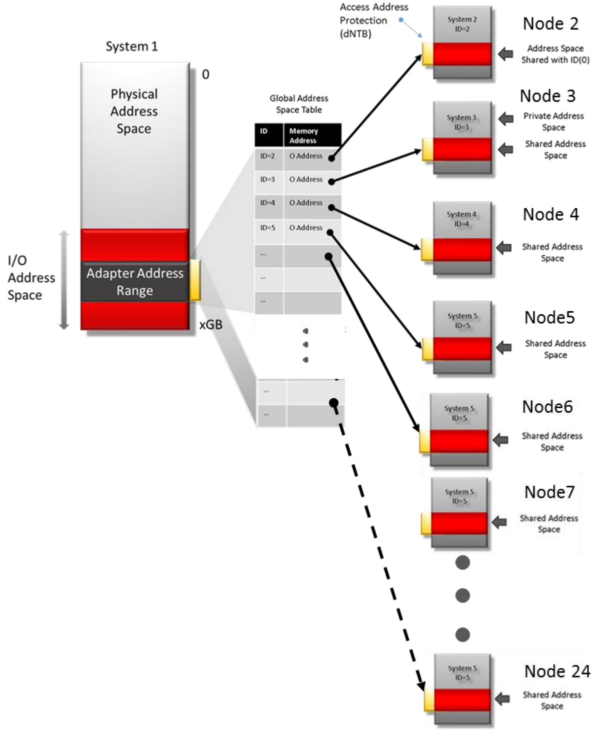
From 3.5 to 6.5 Gbyte/s (bi directional)

Some Topologies Supported



4 Bytes Latency 690 ns (0.60 us) ultra low jittering

From 6.5 to 12.5 Gbyte/s (bi directional)



Thanks for the Time

Questions?