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”
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
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 ...
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!
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 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
Cores Used for 1M IOPs

© 2015 A3CUBE Inc  www.a3cube-inc.com
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!
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!
The problem with existing approaches in a simple picture!

Local App to Remote NVMe

(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

Here a standard server

Simplified I/O Flow

- Standard Network Interfaces
  For external connectivity
- Management Layer
  And computational Power
- Internal Native I/O fabric
- Array of SSDs or SAS/SATA Storage Pools

Simplified Model

Real I/O Flow

- HBA (1)
- HBA (2)
- HBA (n)
- DRAM
- PCIe (n)
- NVMe
- PCIe 2 Sata
- SAS
- NVMe

Modern Server Architecture

© 2015 A3CUBE Inc  www.a3cube-inc.com
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.
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

Data access and processing power of a Facebook-like datacenter in a single rack of servers
Direct Storage Nodes Communication

Direct System to System / Memory to memory latency & bandwidth

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

Remove the latency problem
**Direct Storage Nodes Communication**

**Local & Remote Access Time Comparison**

**Local App to Local NVMe**

- Application
- Virtual Storage Engine (Global Name Space)
- Foundation DPA
- PCIe NIC
- NVMe

**Server (0)**

- 4K @ 17 us
- 2.8us

**Local App to Remote NVMe**

- Application
- Virtual Storage Engine (Global Name Space)
- Foundation DPA
- PCIe NIC
- NVMe

**Server (0)**

- 4K @ 17 us
- 2.8us

**Remote Access**

- Remote access (Direct PCIe Memory Access)

**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

Direct Connection Between nodes @ BUS speed

Up to 128 Gbit/s
Or more in detail Works
NVMe and emerging storage technologies are PCIe driver.
PCle 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

Traditional Hyperconverged

Local Server
- App
- VM
- Hypervisor
- Virtual Network
  - VM (Convergence Engine)
  - VM (Convergence Engine)
  - Hypervisor

Remote Server
- Virtual Network
  - Metadata Server
  - Hypervisor
  - Hypervisor

About 900 us + Storage Latency

A3CUBE Bare Metal Convergence

Local Server
- App
- VM
- Hypervisor
- Hypervisor

Remote Server
- NIC
- Hypervisor/OS
- VM (Convergence Engine)
- Hypervisor/OS

About 115 us all inclusive

No software overhead (Full performance)
Realize the processing power of an entire large datacenter in just a few servers.

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

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!

© 2015 A3CUBE Inc www.a3cube-inc.com
Two type of NVME drivers

2.5” Drive

PCle Form Factor

Standard NVME controller

Some picture to visualize ...
A Real Implementation

RONNIEE Express Inter-Controller Network

Flash Memory Summit 2015 Santa Clara, CA

© 2015 A3CUBE Inc  www.a3cube-inc.com
Some Topologies Supported

- System 1
- System 2
- System 3
- System 4
- System 5

Global Shared Address Space

Physical Address Space

Adapter Address Range

Address Mapping 64 bit

Access Address Protection (dINTB)

From 3.5 to 6.5 Gbyte/s (bi directional)

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

© 2015 A3CUBE Inc  www.a3cube-inc.com

Flash Memory Summit 2015
Santa Clara, CA
Global Shared Address Space

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?