SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Why Composable Infrastructure?

Live Webcast
February 13, 2019
10:00 am PT

Today's Presenters

Philip Kufeldt
University of California Santa Cruz

Mike Jochimsen
Kaminario

Alex McDonald
NetApp

© 2019 Storage Networking Industry Association. All Rights Reserved.

SNIA Legal Notice

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

© 2019 Storage Networking Industry Association. All Rights Reserved.

Forward Looking Statements

- This presentation contains forward looking statements
  - An intelligent & educated exposition of a new approach to compute, networking & storage technology
  - For any third party technologies mentioned here, these are our interpretations of those technologies
  - We assume you understand a little about how compute, networking & storage are currently bolted together

© 2019 Storage Networking Industry Association. All Rights Reserved.

SNIA-At-A-Glance

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

185
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

© 2019 Storage Networking Industry Association. All Rights Reserved.

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements Incorporate them into standards and programs

Collaborate with other industry associations

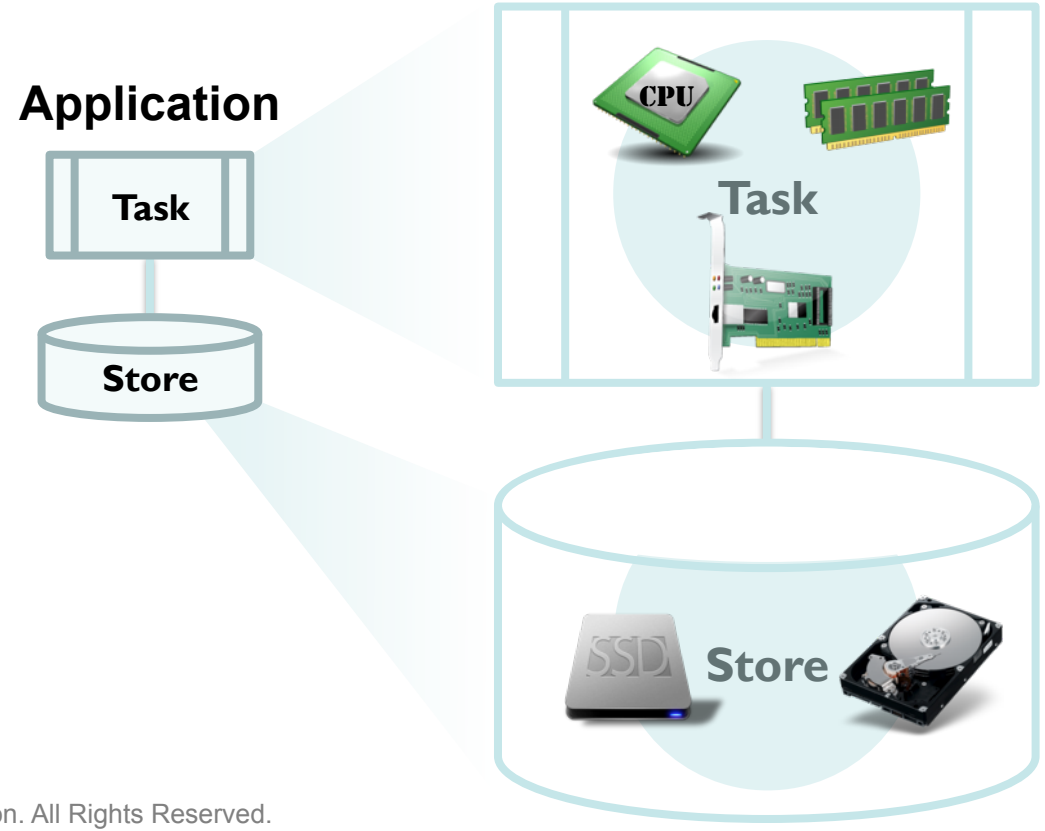
Agenda

- **What's driving current compute, network & storage developments?**
- What is Composable Infrastructure?
- What steps are we taking to realize it?

© 2019 Storage Networking Industry Association. All Rights Reserved.

What is an Application

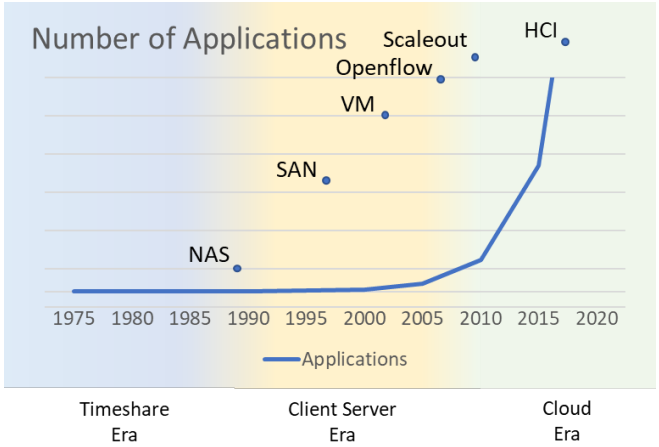
- Task
  - Apps need a system
    - Has requirements
      - CPU cores
      - Memory size
      - Network BW
      - Network Location
      - Availability
- Store
  - Most apps need a persistent store
    - Has requirements
      - BW
      - Latency
      - Capacity
      - Availability
- Examples
  - RDBMS
  - Web Servers
  - ML application



© 2019 Storage Networking Industry Association. All Rights Reserved.

Keeping up with the Apps

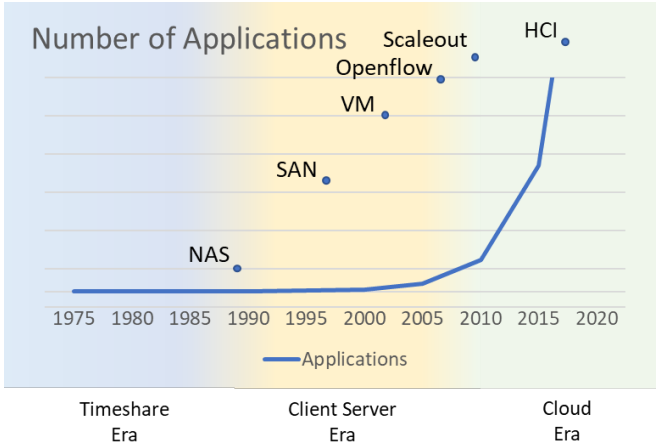
- Decades of history - Timeshare to Client Server to Cloud
  - Applications growing geometrically
    - Faster than technology
  - Causes an ever growing need for more HW
  - Growing configuration requirements
  - Growing management problem
- Flexibility and cost drive the direction
  - Overprovisioning reductions
    - Reduce costs
  - Disaggregation focuses management problem



© 2019 Storage Networking Industry Association. All Rights Reserved.

Keeping up with the Apps

- Timeshare to client server
  - How many cheap servers = one mainframe
    - Unit of allocation decreases
  - Flexibility
    - Management
    - Configuration
    - Availability
- Client Server to Cloud
  - Hitting limits of scale up
  - Elasticity without infrastructure investment
  - Location abstraction
  - Externally: Unit of allocation drops to App
    - But within the cloud the same problems

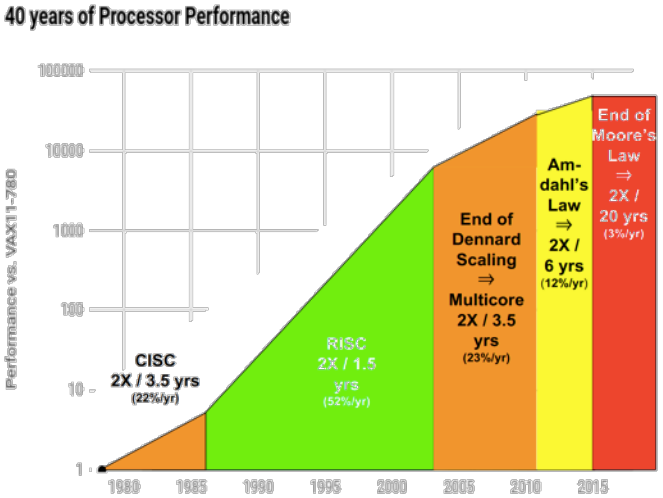


© 2019 Storage Networking Industry Association. All Rights Reserved.

While Applications are GROWING

- Dennard's scaling ended
  - Power leakage and heat prevent cycle scaling
- Multicore hit Amdahl's law
  - Applications can only be parallelized so far
- Moore's law is ending
  - Physical size limitations
- What's left then?
  - Domain Specific Architectures
    - Graphics Processing Unit (GPU)
    - Offloading Network Interface Controllers (NIC)
    - Tensor Processing Unit (TPU)
    - FPGA Based Accelerators
  - This increases configuration complexity

David Patterson's presentation at ISSCC2018
https://youtu.be/NZS2TtWcutc



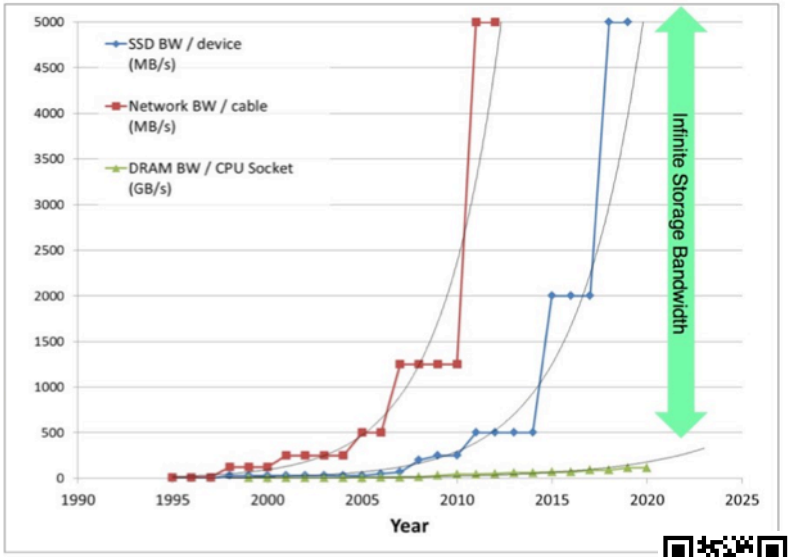
Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e 2018

© 2019 Storage Networking Industry Association. All Rights Reserved.

Multicore Problems

- Thanks to discussions started by Fritz Kruger and Allen Samuels
- Storage and Network throughput growth outstripping CPU
  - By 2020 only a couple SSDs per socket will be needed to outstrip ability to shuttle data in and out
    - SCM is going to make this worse
- This suggests that larger multicore systems will compete for DMA bandwidth
  - Fewer apps per server
  - Age of Hyper Converged Infrastructure may be closing

- **Network, Storage and DRAM Trends**
  - DRAM throughput is a proxy to CPU capability
  - Storage Bandwidth is not literally infinite
  - But the ratio of Network and Storage to CPU throughput is widening very quickly

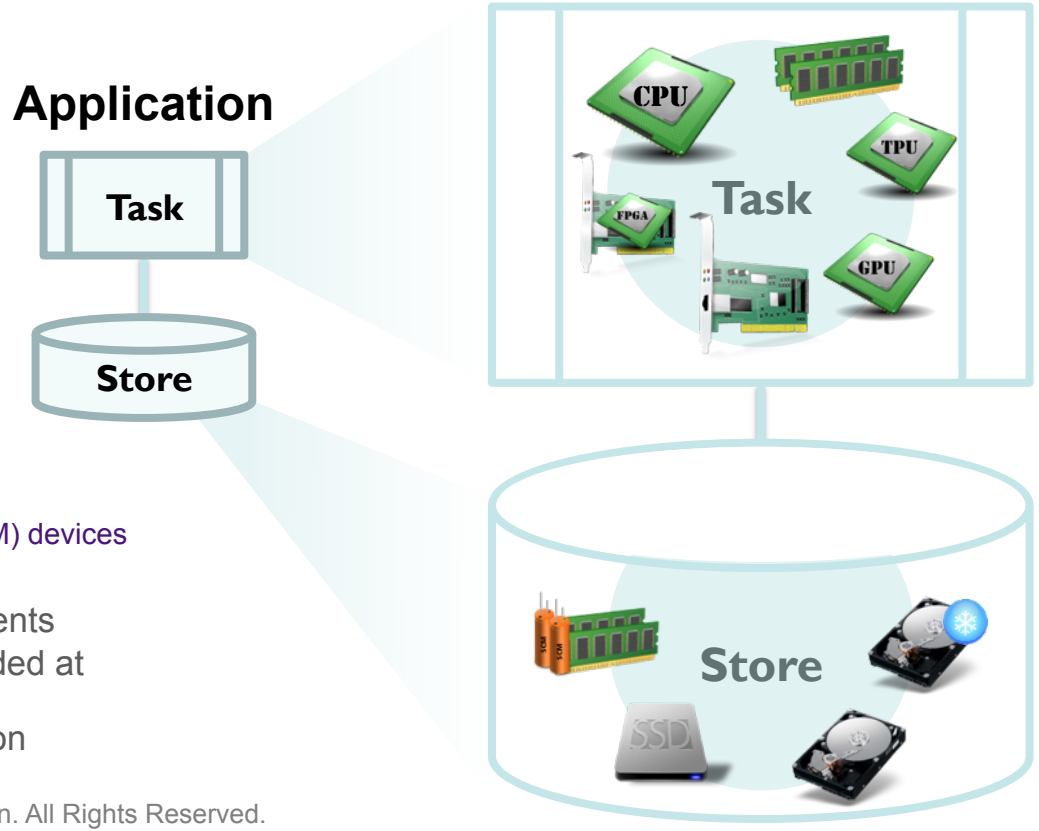


CPU Bandwidth – The Worrisome 2020 Trend, Fritz Kruger, Mar 2016

© 2019 Storage Networking Industry Association. All Rights Reserved.

Today's Applications

- Task
  - Many System options
    - CPUs/SoCs
    - Core counts
    - DDR Capacity
    - NICs
    - Accelerators
    - GPUs/TPUs
- Store
  - Many Store options
    - Cold stores
    - HDDs
    - SSDs
    - Persistent memory (SCM) devices
- All must go into a box
  - Dictated by App requirements
  - What and how much decided at purchase time
  - No going back, no evolution



© 2019 Storage Networking Industry Association. All Rights Reserved.

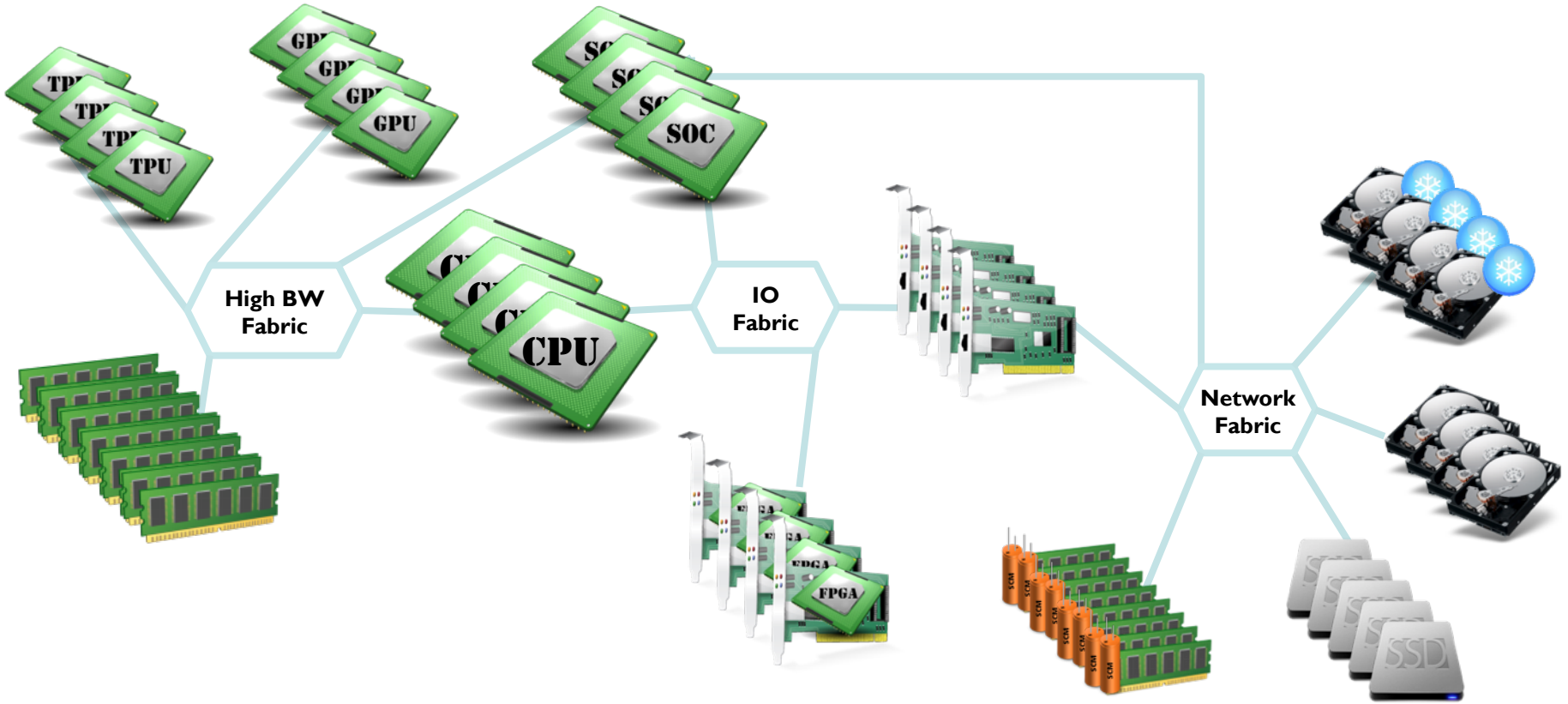
What is the Problem

- Application requirements wide and varied
  - Complicated set of hardware requirements
- Must map these requirements onto physical hardware
  - Due to core counts, multiple apps must be mapped to single system
  - Forces IT managers to be system designers
  - Forces overprovisioning inside the system
  - Availability and Competition issues
- Application requirements quickly and constantly evolve
  - Mapping occurs at purchase time and cannot evolve
  - Invalidates system design requirements
- Growth rate of apps
  - Forces overprovisioning system counts for elasticity
- Ever growing classes of hardware systems
  - Lifecycle management (scaling, EOL, etc) becomes a multi-vectored problem

The multicore server as the unit of app allocation is now too big and complicated

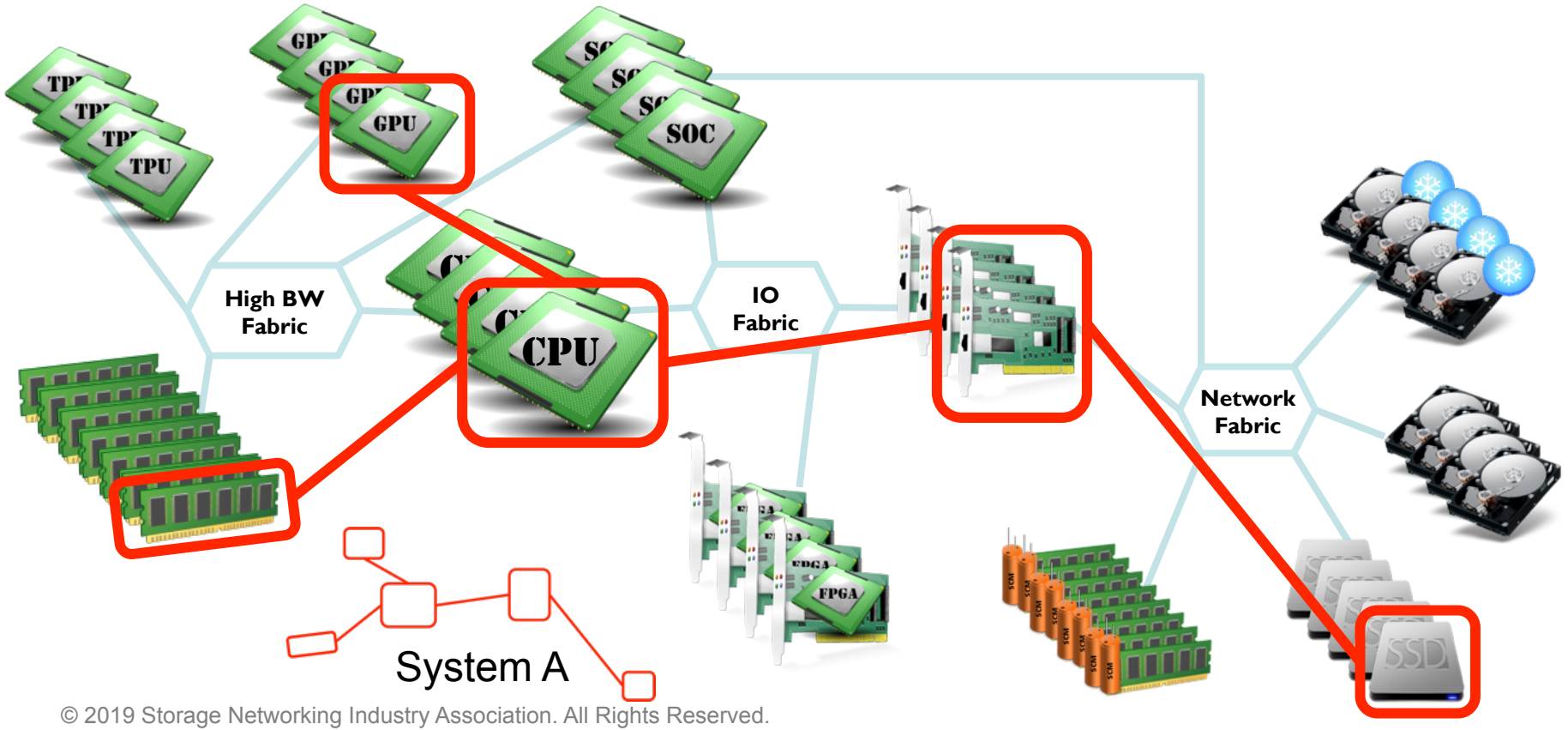
© 2019 Storage Networking Industry Association. All Rights Reserved.

What could a solution look like



© 2019 Storage Networking Industry Association. All Rights Reserved.

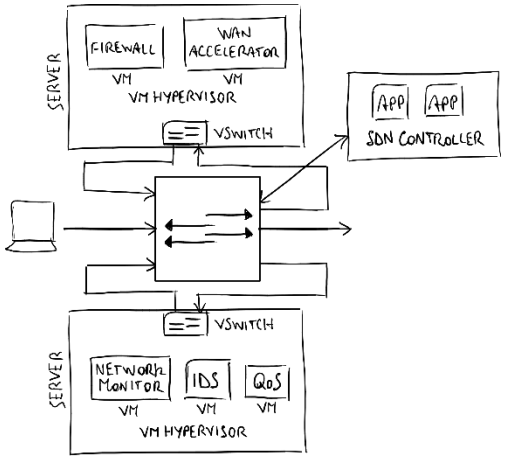
What could a solution look like



© 2019 Storage Networking Industry Association. All Rights Reserved.

We are already moving this way

- Software Defined Networking
  - Virtual networks created dynamically
- Software Defined Storage
  - Disaggregation exists and broadening
    - NAS, SAN
    - Scale out
    - NVMeoF
- Fabric Attached Persistent Memories
- Storage Accelerators
  - SoC and FPGA based



https://commons.wikimedia.org/wiki/File:SDN_NFV.svg

© 2019 Storage Networking Industry Association. All Rights Reserved.

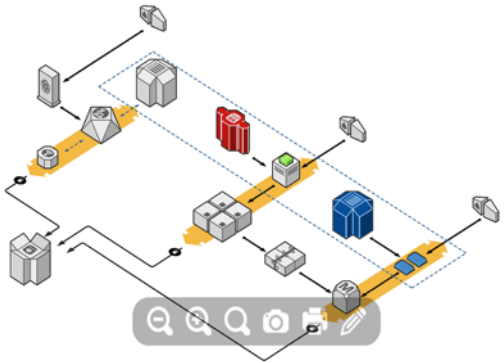
Agenda

- What's driving current compute, network & storage developments?
- **What is Composable Infrastructure?**
- What steps are we taking to realize it?

© 2019 Storage Networking Industry Association. All Rights Reserved.

What is Composable Infrastructure?

- Compose – to form by putting together
- Infrastructure – the underlying foundation or basic framework (as of a system or organization)



Source: Merriam Webster dictionary

© 2019 Storage Networking Industry Association. All Rights Reserved.

What is Composable Infrastructure?

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

- "Composable infrastructure treats compute, storage, and network devices as pools of resources that can be provisioned as needed, depending on what different workloads require for optimum performance."[1]
- "A composable infrastructure is a framework whose physical compute, storage and network fabric resources are treated as services. In a composable infrastructure, resources are logically pooled so that administrators don't have to physically configure hardware to support a specific software application. Instead, the software's developer defines the application's requirements for physical infrastructure using policies and service profiles and then the software uses application programming interface (API) calls to create (compose) the infrastructure it needs to run on bare metal, as a virtual machine (VM) or as a container."[2]
- "Composable infrastructure brings together compute, storage and network fabric into one platform, similar to a converged or hyperconverged infrastructure. It also integrates a software-defined intelligence and a unified API to "compose" these fluid resource pools."[3]

1. https://www.networkworld.com/article/3266106/data-center/what-is-composable-infrastructure.html
2. https://www.zdnet.com/article/composable-infrastructure/
3. https://www.itprotoday.com/business-resources/just-what-heck-composable-infrastructure-anyway

© 2019 Storage Networking Industry Association. All Rights Reserved.

What is Composable Infrastructure?

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

- "Composable infrastructure treats **compute, storage, and network devices as pools of resources** that can be **provisioned as needed**, depending on what different **workloads require** for optimum performance."[1]
- "A composable infrastructure is a framework whose **physical compute, storage and network fabric resources** are treated as services. In a composable infrastructure, resources are **logically pooled** so that administrators don't have to physically configure hardware to support a specific software application. Instead, the software's developer defines the application's requirements for physical infrastructure using policies and service profiles and then the **software uses application programming interface (API) calls to create (compose) the infrastructure it needs** to run on bare metal, as a virtual machine (VM) or as a container."[2]
- "Composable infrastructure brings together **compute, storage and network fabric** into one platform, similar to a converged or hyperconverged infrastructure. It also integrates a software-defined intelligence and a **unified API to "compose" these fluid resource pools**."[3]

1. https://www.networkworld.com/article/3266106/data-center/what-is-composable-infrastructure.html
2. https://www.zdnet.com/article/composable-infrastructure/
3. https://www.itprotoday.com/business-resources/just-what-heck-composable-infrastructure-anyway

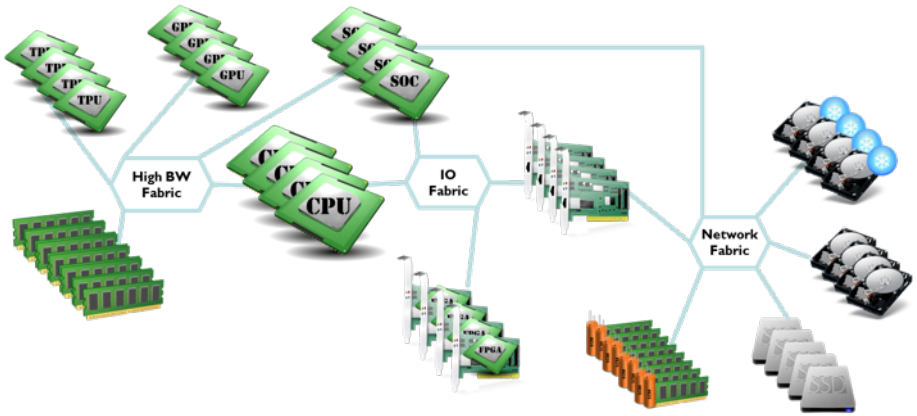
© 2019 Storage Networking Industry Association. All Rights Reserved.

What is Composable Infrastructure?

- ✓ Separate compute, storage, networking components
- ✓ Pools of resources
- ✓ Don’t need to be physically proximate
- ✓ Compose as needed via orchestration
- ✓ Scalability is not one way 
- ✓ API driven (autonomous operation)
- ✓ Driven by application needs

© 2019 Storage Networking Industry Association. All Rights Reserved.

Disaggregation



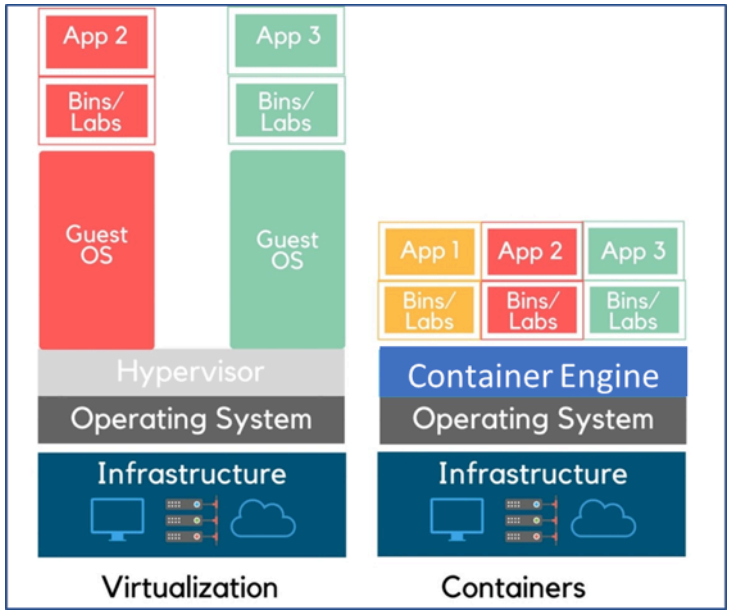
- Disaggregate – to separate into component parts (Merriam Webster)
- Systems – server, storage device, network switch
- Components - CPU, memory, discrete disks, PCI devices
- Anything that can be accessed on a network fabric

© 2019 Storage Networking Industry Association. All Rights Reserved.

But what about virtualization/ containerization?

Fundamentally the opposite!

- Virtualization
  - Take one instance of a resource (e.g., a server) and slice it up
  - Optimal use of a single resource by sharing it among multiple apps
  - Each app isolated above the kernel level through the use of VMs
- Containerization
  - OS level virtualization method
  - Multiple apps share the same kernel without having to launch VMs

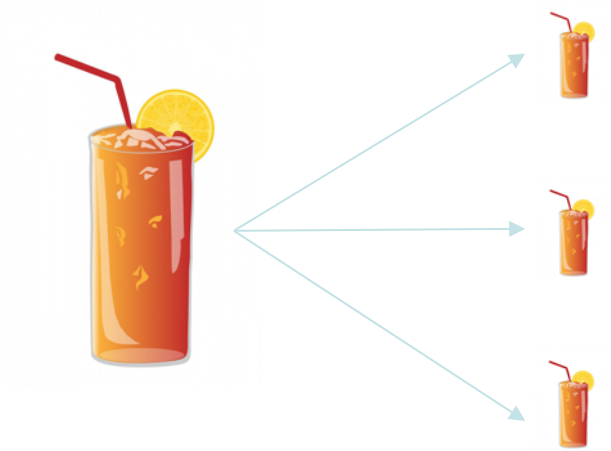


© 2019 Storage Networking Industry Association. All Rights Reserved.

Virtualization & Containerization vs. Composability

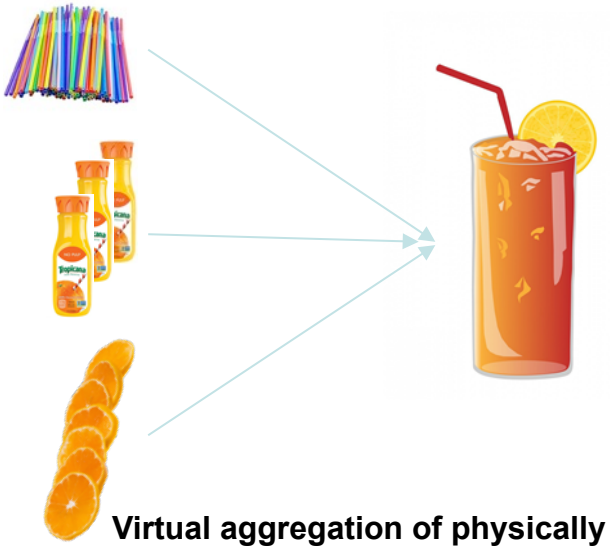
SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Virtualization/Containerization



Virtual disaggregation of physically aggregated components

Composability



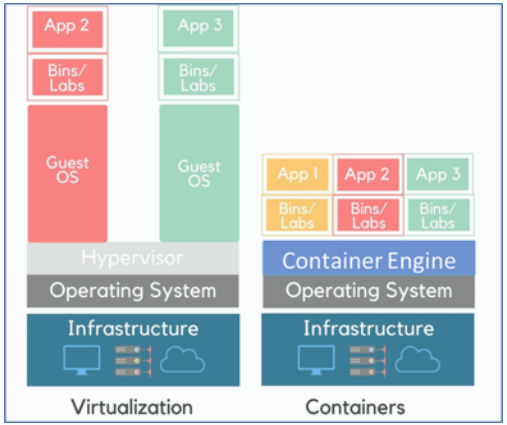
Virtual aggregation of physically disaggregated components

© 2019 Storage Networking Industry Association. All Rights Reserved.

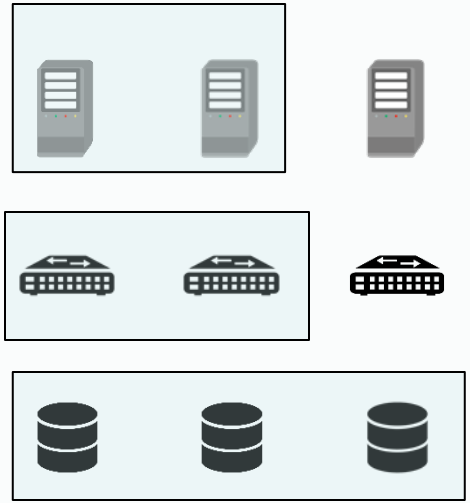
Where are we today?

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

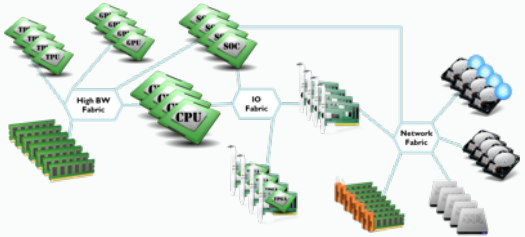
Where We've Been



Where We Are



Where We Are Going



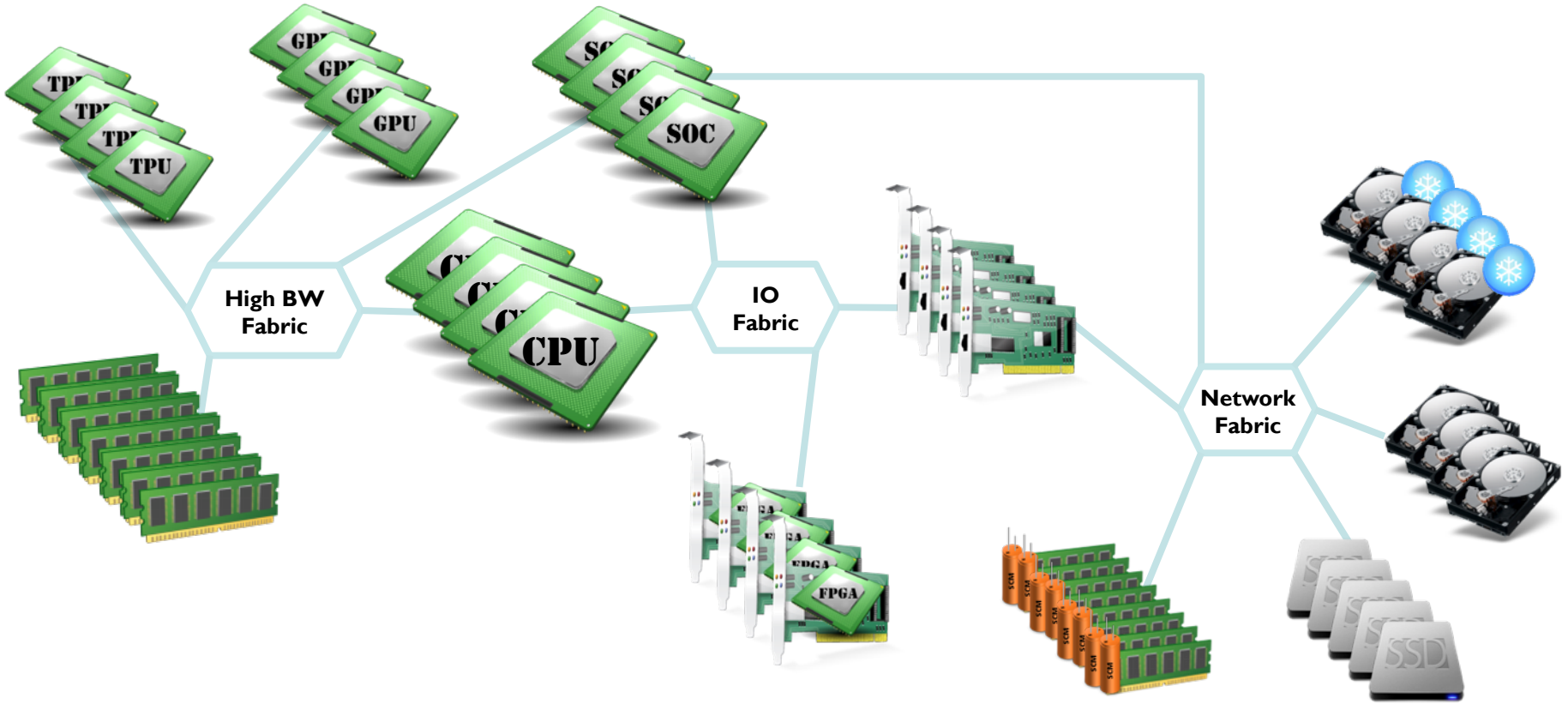
© 2019 Storage Networking Industry Association. All Rights Reserved.

Agenda

- What's driving current compute, network & storage developments?
- What is Composable Infrastructure?
- **What steps are we taking to realize it?**

© 2019 Storage Networking Industry Association. All Rights Reserved.

What could a solution look like



© 2019 Storage Networking Industry Association. All Rights Reserved.

- Peripheral Component Interconnect Express (PCIe)
- Standard Processor Device Interconnect
  - Extremely Broad support
- Once only a local interconnect
- Switches enabled creating a fabric
- Still Target/Initiator architecture
- Version 4 emerging, 5 on its heels

© 2019 Storage Networking Industry Association. All Rights Reserved.

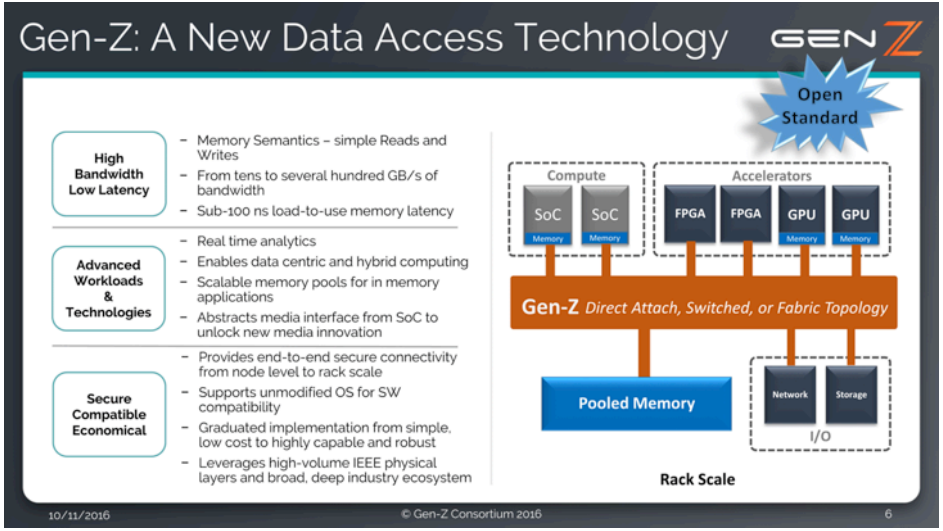
- Non-Volatile Memory express (NVMe)
  - Storage access protocol
    - Redesign due to high speeds flash and memory devices
  - Default interconnect is PCIe
- Non-Volatile Memory express over Fabrics (NVMe-oF)
  - Disaggregates from local fabrics
  - Permits the use of general switched and routed fabrics
    - RDMA
      - Infini-Band (IB)
      - RoCE v2
      - iWarp
    - Fibre Channel (FC and FCoE)
    - TCP

© 2019 Storage Networking Industry Association. All Rights Reserved.

- **Cache Coherent Interconnect for Accelerators** (**CCIX**), pronounced "see-six“
- Key Members: AMD, ARM, Huawei, Mellanox, Qualcomm, Xilinx
- CCIX is an open cache coherent interconnect architecture developed by the CCIX Consortium.
  - CCIX is designed to simplify the communication between the central processor and the various accelerators in the system through a cache-coherent extension to standard PCIe.

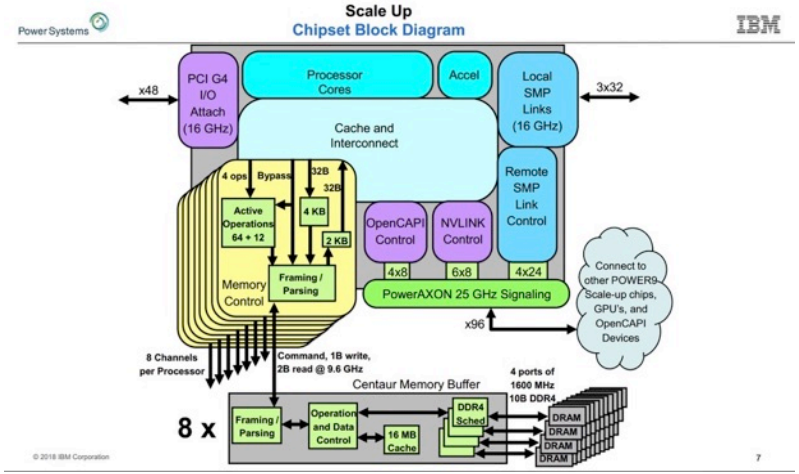
© 2019 Storage Networking Industry Association. All Rights Reserved.

- Founder Members
  - AMD, ARM, Broadcom, Cray, Dell EMC, Hewlett Packard Enterprise, Huawei, IDT, Mellanox, Micron, Microsemi, Samsung, SK Hynix, and Xilinx
- Gen Z is
  - An open systems Interconnect designed to provide memory-semantic access to data and devices via direct-attached, switched or fabric topologies



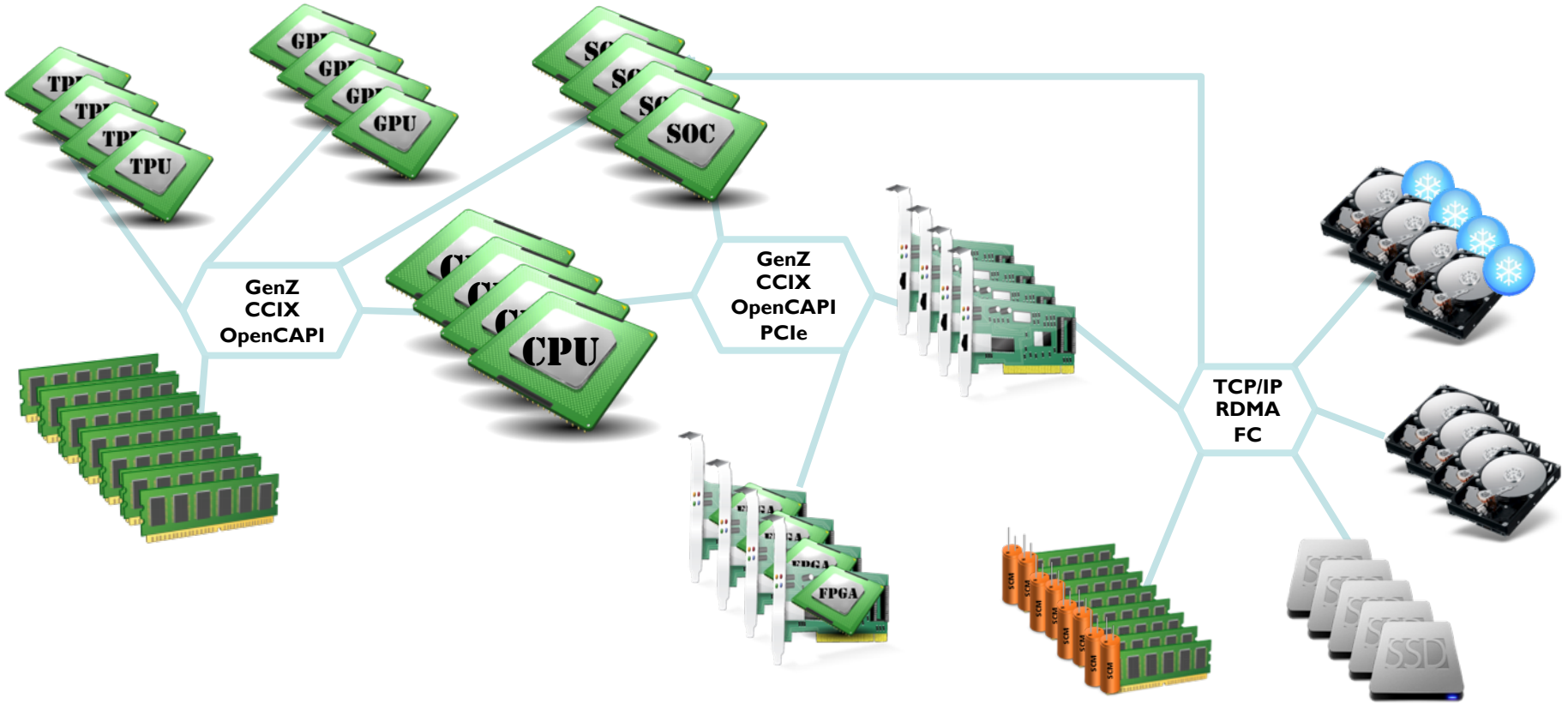
© 2019 Storage Networking Industry Association. All Rights Reserved.

- Open Coherent Accelerator Processor Interface
- Strategic Members:
  - AMD, Google, IBM, Mellanox, Micron, NVIDIA, Western Digital, Xilinx
- OpenCAPI is
  - An Open Interface Architecture that allows any microprocessor to attach to it
  - Provides
    - Coherent user-level accelerators and I/O devices
    - Advanced memories accessible via read/write or user-level DMA semantics
  - Intended to be agnostic to processor architecture



© 2019 Storage Networking Industry Association. All Rights Reserved.

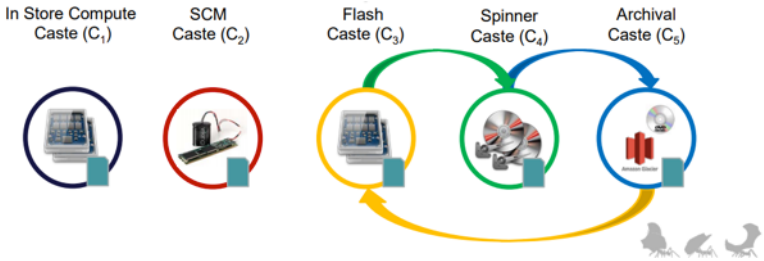
What could a solution look like



© 2019 Storage Networking Industry Association. All Rights Reserved.

Storage Disaggregation

- New Ideas
  - Eusocial storage devices
    - Fully autonomous, disaggregated and composable storage devices



Eusocial Storage Devices
Offloading Data Management to Storage Devices that Can Act Collectively

PHILIP KUFELDT, CARLOS MALTZAHN, TIM FELDMAN, CHRISTINE GREEN, GRANT MACKEY, AND SHINGO TANAKA

As storage devices get faster, data management tasks rob the host of CPU cycles and DDR bandwidth. In this article, we examine a new interface to storage devices that can leverage existing and new CPU and DRAM resources to take over data management tasks like availability, recovery, and migrations. This new interface provides a roadmap for device-to-device interactions and more powerful storage devices capable of providing in-store compute services that can dramatically improve performance. We call such storage devices "eusocial" because we are inspired by eusocial insects like ants, termites, and bees, which as individuals are primitive but collectively accomplish amazing things.

The Evolution of the Problem

Why Try Smart Storage Again, and Why Now?

Offloading storage processing has been around since the earliest days of computing. The idea of having a dedicated and cheaper I/O processor offloading the main processor complex made sense at a time when processor cycles were incredibly scarce and costly. However, over the years, processor cycle availability has geometrically increased and costs have plummeted making the utilitarian I/O processor costlier in terms of complexity in both hardware architecture and software. These fast and large CPU complexes permit general-purpose execution and I/O management, including data management. Data management tasks are beyond the basic tasks of storing and retrieving data, including services such as translation, mapping, deduplication, compaction, sorting, scrubbing, data movement, data redundancy, and recovery.

Including I/O management created a tight coupling of storage with the server system architecture. With such compute resources available, storage devices need only do the media management and map logical placement information to physical placement information, leaving essentially all data management relegated to the general-purpose processor. Furthermore, the simplistic API required to accomplish these goals treats every device as completely independent even though data management necessarily creates device relationships and dependencies, all of which have to be managed by the general-purpose processor.

This has driven the evolution of the storage component towards a highly cost-efficient model that has resisted most attempts to offload tasks to the component. Attempts to push some of data management back into the device, such as SCSI OSD or Kinetic, have all failed due to the need for additional compute and memory in the device pushing up per-GiB costs.

NAS Succeeds in Offloading

The one place where data management offloading was successful was Network Attached Storage (NAS). NAS environments offload all the data management to centralized servers on the network (Figure 1).

Client servers then use a network-based access protocol (NFS, CIFS) to store and retrieve data. The reason for the offloading was two-fold:

16 ;login: SUMMER 2018 VOL. 43, NO. 2 www.usenix.org

STORAGE

Eusocial Storage Devices: Offloading Data Management to Storage Devices that Can Act Collectively

Client Servers
TCP/IP
NAS/Scaleout Heads
SATA/SAS
Media

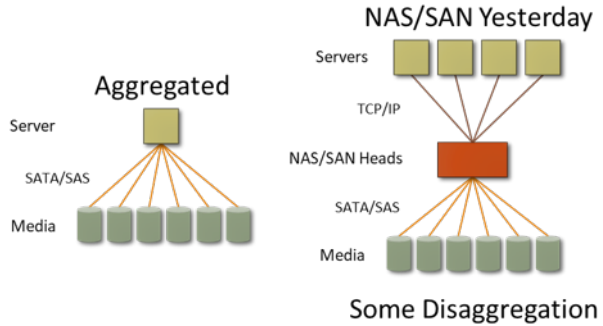
1. Centralizing storage management
2. Disaggregating the storage from the server

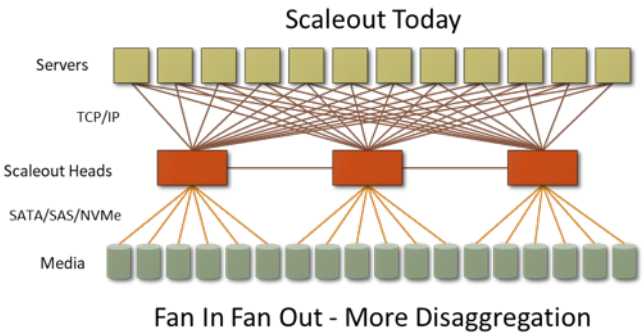
The former is a straightforward centralization gain; by consolidating all of the server management touch points, management man hours were reduced. Disaggregation was driven by the need for sharing and availability. By separating data resources from the local server and its processor complex, these resources could be placed on a general-purpose network. This provided two big wins: the data could be shared by many servers, and data resources could

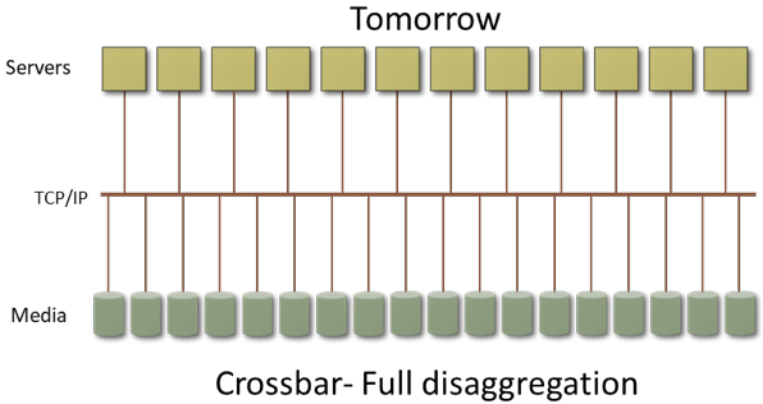
https://www.usenix.org/system/files/login/articles/login_summer18_05_kufeldt.pdf

© 2019 Storage Networking Industry Association. All Rights Reserved.

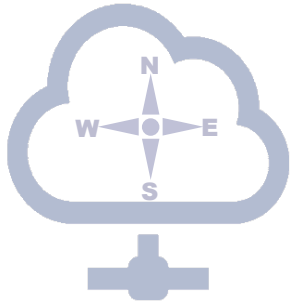
Where this might go







All elements can freely talk



North and South as well as East and West

© 2019 Storage Networking Industry Association. All Rights Reserved.

Our Next Webcast

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

What's New in Container Storage?

February 26, 2019

10:00 am PT

Register at:

https://www.brighttalk.com/webcast/663/345389

© 2019 Storage Networking Industry Association. All Rights Reserved.

After This Webcast

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Cloud Storage Technologies Initiative website and available on-demand at https://www.snia.org/forum/csti/knowledge/webcasts
- A full Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud

© 2019 Storage Networking Industry Association. All Rights Reserved.

Thank You

© 2019 Storage Networking Industry Association. All Rights Reserved.