\Orchestrating a brighter world

NEC

GIGAIO

Cloud Agility.
Half the TCO.

# GigaIO and NEC HPC/AI Solutions: Composable Architectures with Next-Gen Interconnects and Vector Engines

Marc Lehrer– VP Global Sales
mlehrer@gigaio.com

ISC
High Performanc

# Agenda

Who is GigaIO?

Changes in the data center

Challenges with heterogeneous computing

GigaIO FabreX™ – Universal Dynamic Fabric

NEC SX-Aurora TSUBASA performance results

Summary

# Company Overview

Established in 2016 by networking industry veterans

Headquartered in Carlsbad, CA

Highly skilled team with broad and deep network architecture, S/W, H/W and silicon development capability

Strong patent position; 6 issued patents; ~12 in process

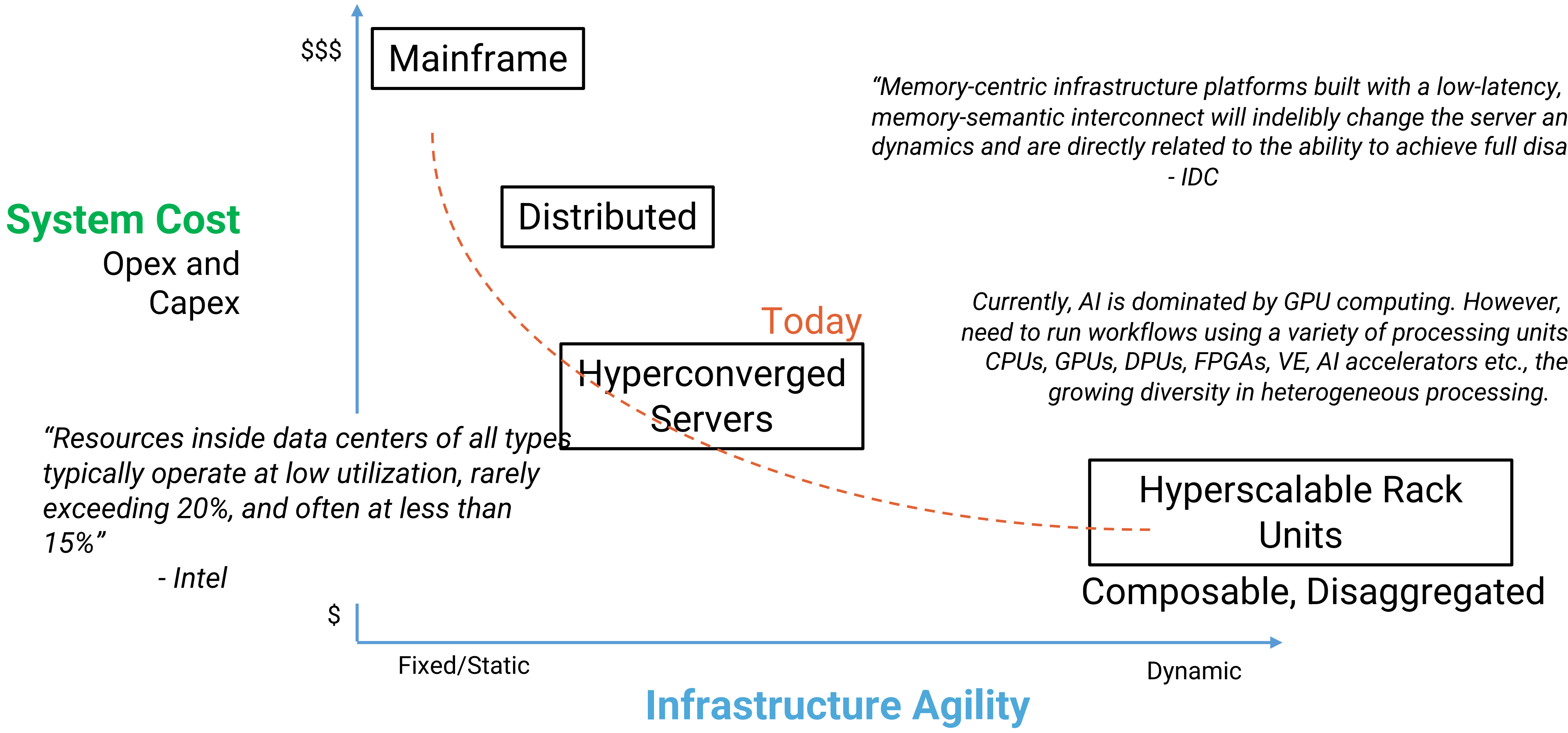
Well funded startup with knowledgeable investors and advisors

2nd Generation product developed and shipping!

# Fundamental Change is Happening in the Data Center

- End of Moore’s law
  - CPUs alone are no longer adequate for many computational tasks – rise of accelerators
- Explosive growth in data
  - Surge in amount of data to be processed and stored
- Dramatic increase in workload variation driven by machine learning pipelines
- Compute capability moving closer to where data is being generated (and used) - to the Edge

# Driving the Next Generation IT Architecture



System Cost

Opex and Capex

$$$

$

Mainframe

Distributed

Today

Hyperconverged Servers

Hyperscalable Rack Units

Composable, Disaggregated

Fixed/Static

Dynamic

Infrastructure Agility

*"Memory-centric infrastructure platforms built with a low-latency, high-bandwidth, memory-semantic interconnect will indelibly change the server and storage market dynamics and are directly related to the ability to achieve full disaggregation."*
*- IDC*

*Currently, AI is dominated by GPU computing. However, users need to run workflows using a variety of processing units. From CPUs, GPUs, DPUs, FPGAs, VE, AI accelerators etc., there is growing diversity in heterogeneous processing.*

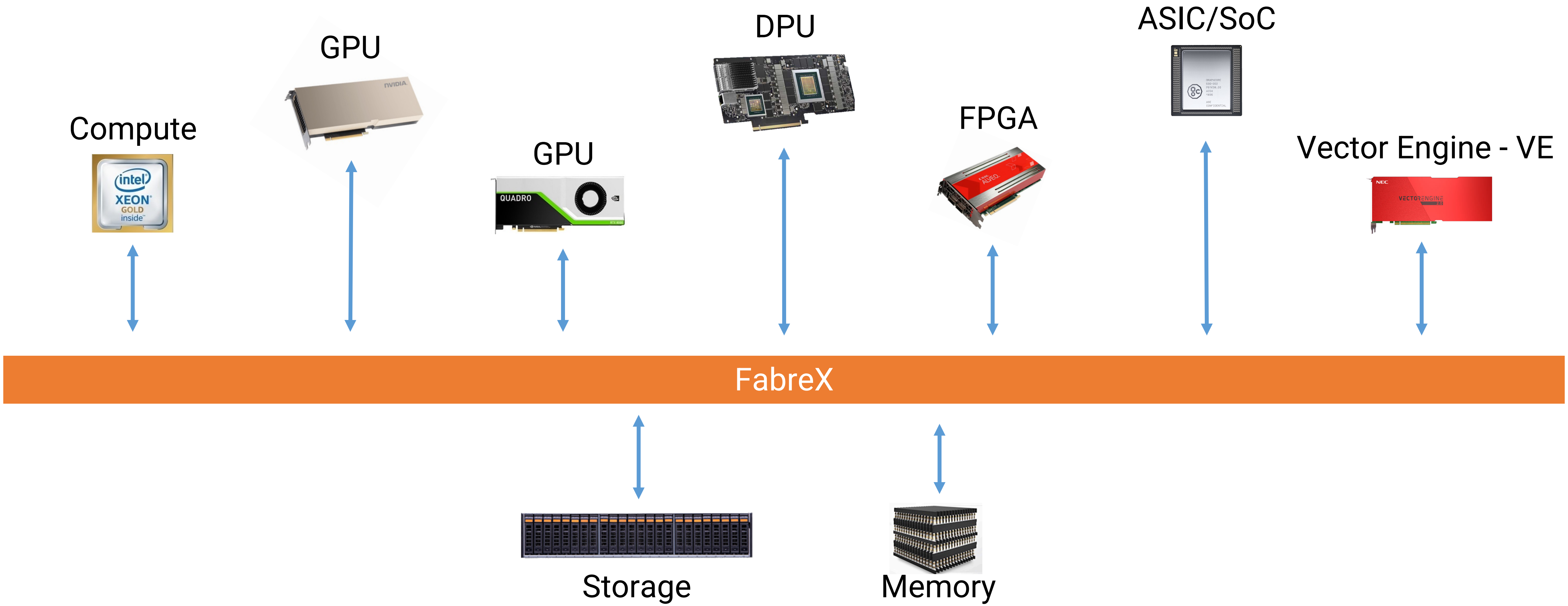
*"Resources inside data centers of all types typically operate at low utilization, rarely exceeding 20%, and often at less than 15%"*
*- Intel*

# New IT Challenges with Heterogeneous Compute

- The server refresh cycle does not match the accelerator innovation cycle:
- Workloads are diversifying and expanding:
  - Existing workloads in simulation and analysis are expanding
  - **And** - new workloads such as AI/ML and data analytics are being added
- **And** - the hardware technologies required for this expansion are diversifying and becoming more specialized (VE, GPUs, FPGAs, DPUs, etc.)
- **But** - budgets are NOT increasing accordingly
- **SO** – How to achieve more out of the existing infrastructure?

# Today’s Heterogeneous System

## Vast Choice of Accelerator Devices



Compute

GPU

GPU

DPU

FPGA

ASIC/SoC

Vector Engine - VE

FabreX

Storage

Memory

# I&O Decision Makers Face Difficult Choices

- Build multiple infrastructure silos – one for each application, OR
- Build a single configuration optimized for one application – and sub-optimal for everything else

➢ What if it were possible to dynamically change the infrastructure down to the component level based on workload?

# Fabric Computing Key to Heterogeneous Computing

What if a solution offered these characteristics:

- Delivers full performance – both latency and bandwidth
- Works with any workload and with any component (VE, GPU, FPGA, etc.)
- Is based on open standards to avoid vendor lock-in
- Is scalable, ideally down to the component level
- Enables new capabilities using existing infrastructure as well as easily accommodating new infrastructure
- Is easy to deploy and manage, ideally using existing tools

# FabreX™ is the New Universal Dynamic Fabric

## The Only <u>Routable</u> PCIe (and CXL) Fabric Throughout the Rack to Connect Both Resources and Servers

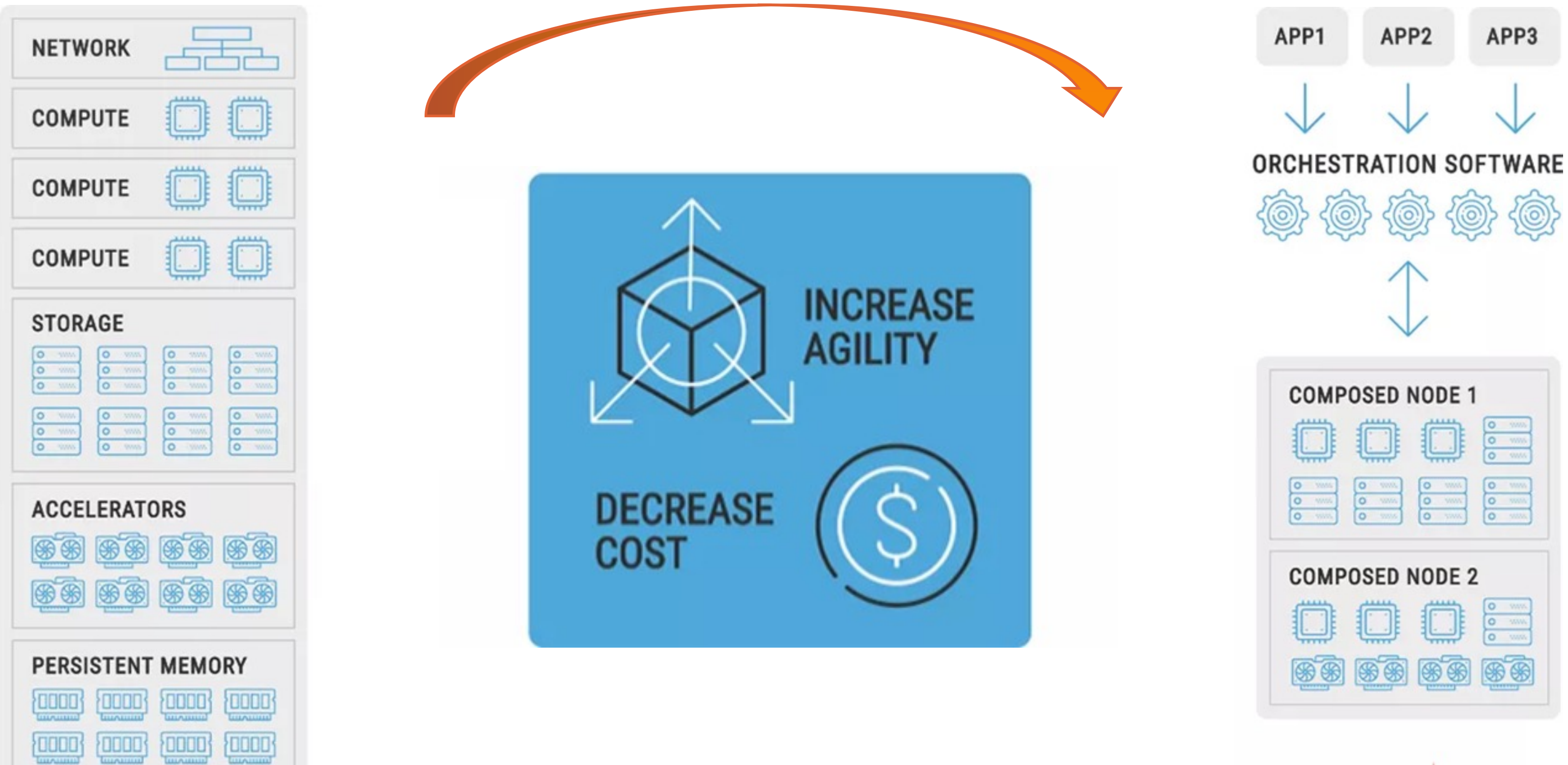
- Any workload
- Any component
- Based on open standards
- At full performance without overhead

- Composable
- Scalable
- Data centric

# FabreX Delivers Composable Infrastructure

## Create Disaggregated Pools

Buy only what you need, when you need it

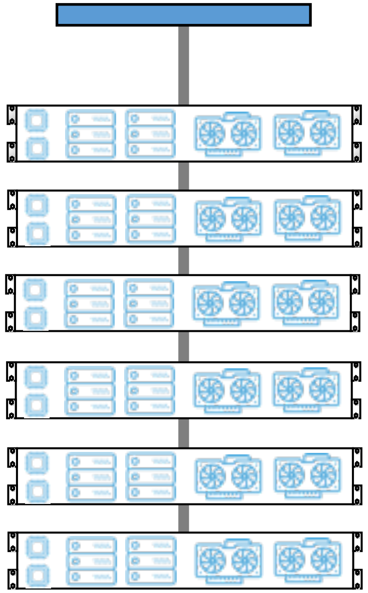


## Compose On-The-Fly

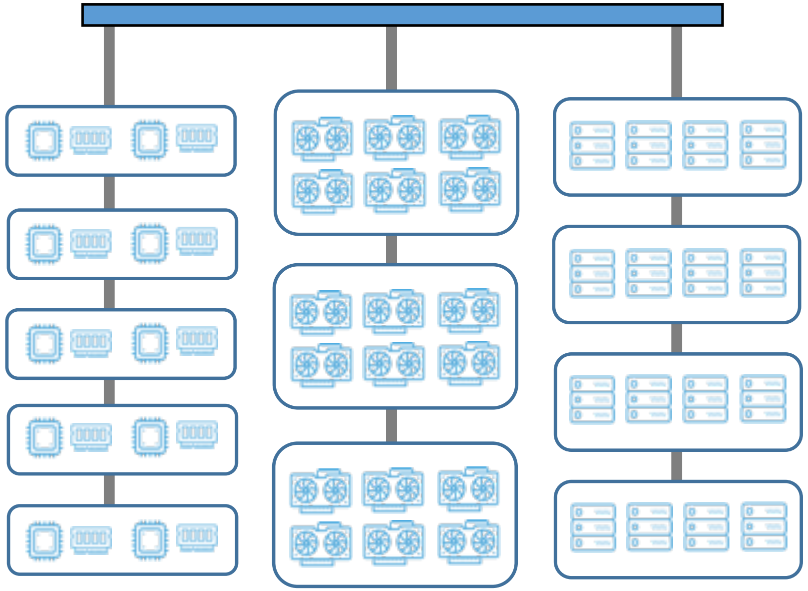
Create virtual machines
for
specific workloads

Boost utilization rates - 3 to 5 times
Reduce Total Cost of Ownership by half

# The Key to Eliminating Stranded Resources is Disaggregation…

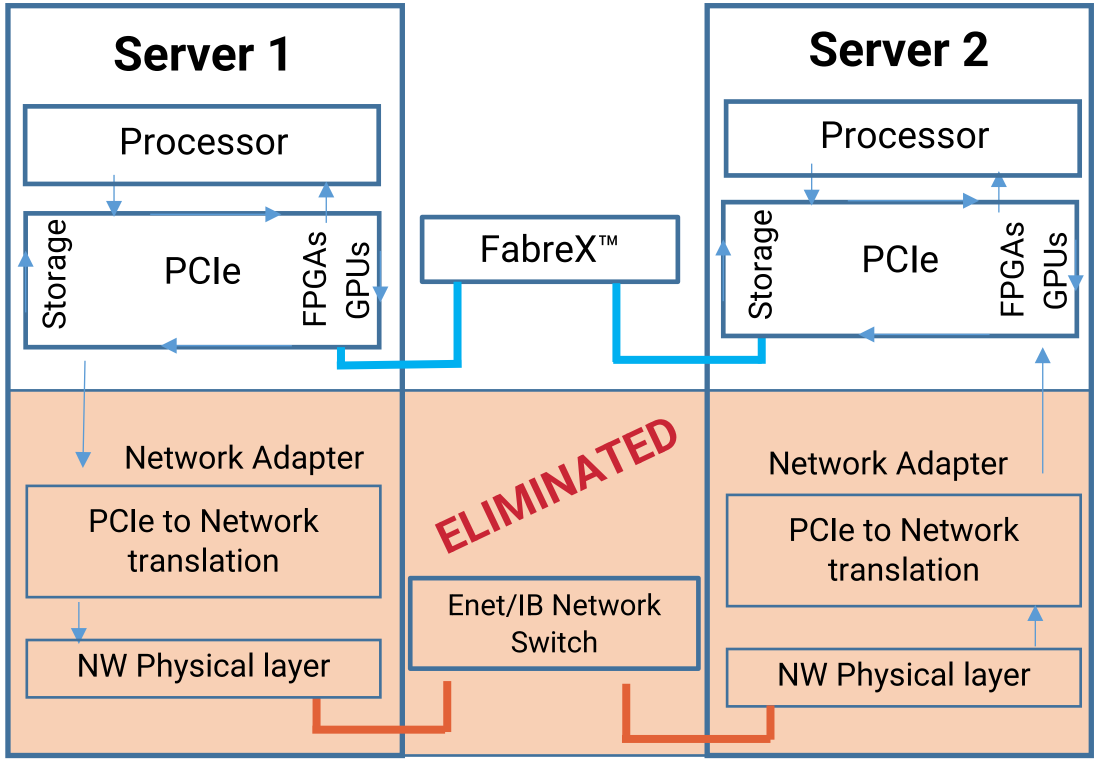


Enabling servers to use any device at any time



- Flexible resource pools, easily updated/upgraded
- Configure resources to precisely match each workload
- Precisely tune compute/storage/accelerator ratios for each job
- Higher utilization of expensive resource
- Different obsolesce paths for each resource

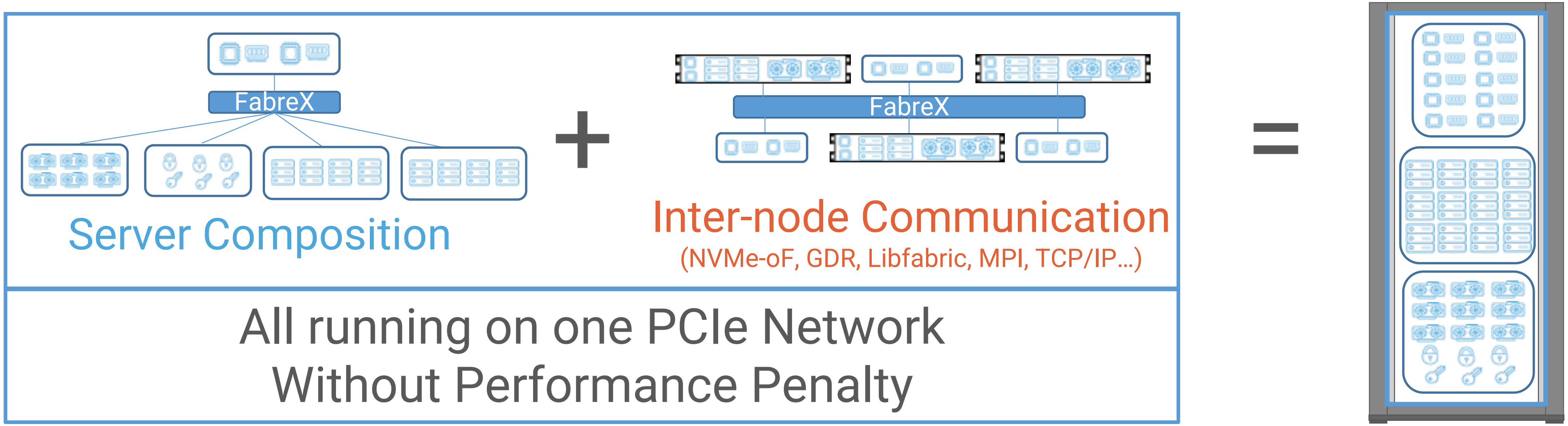
# FabreX Drives out Latency to Deliver Disaggregation



Sub-Microsecond Latency

The Only Routable PCIe (and CXL) Network Throughout the Rack to Connect Both Resources and Servers

# Only FabreX PCIe Delivers Cloud-Class™ Composition



Server Composition

Inter-node Communication

(NVMe-oF, GDR, Libfabric, MPI, TCP/IP…)

All running on one PCIe Network
Without Performance Penalty

Minimize TCO
Improve Serviceability

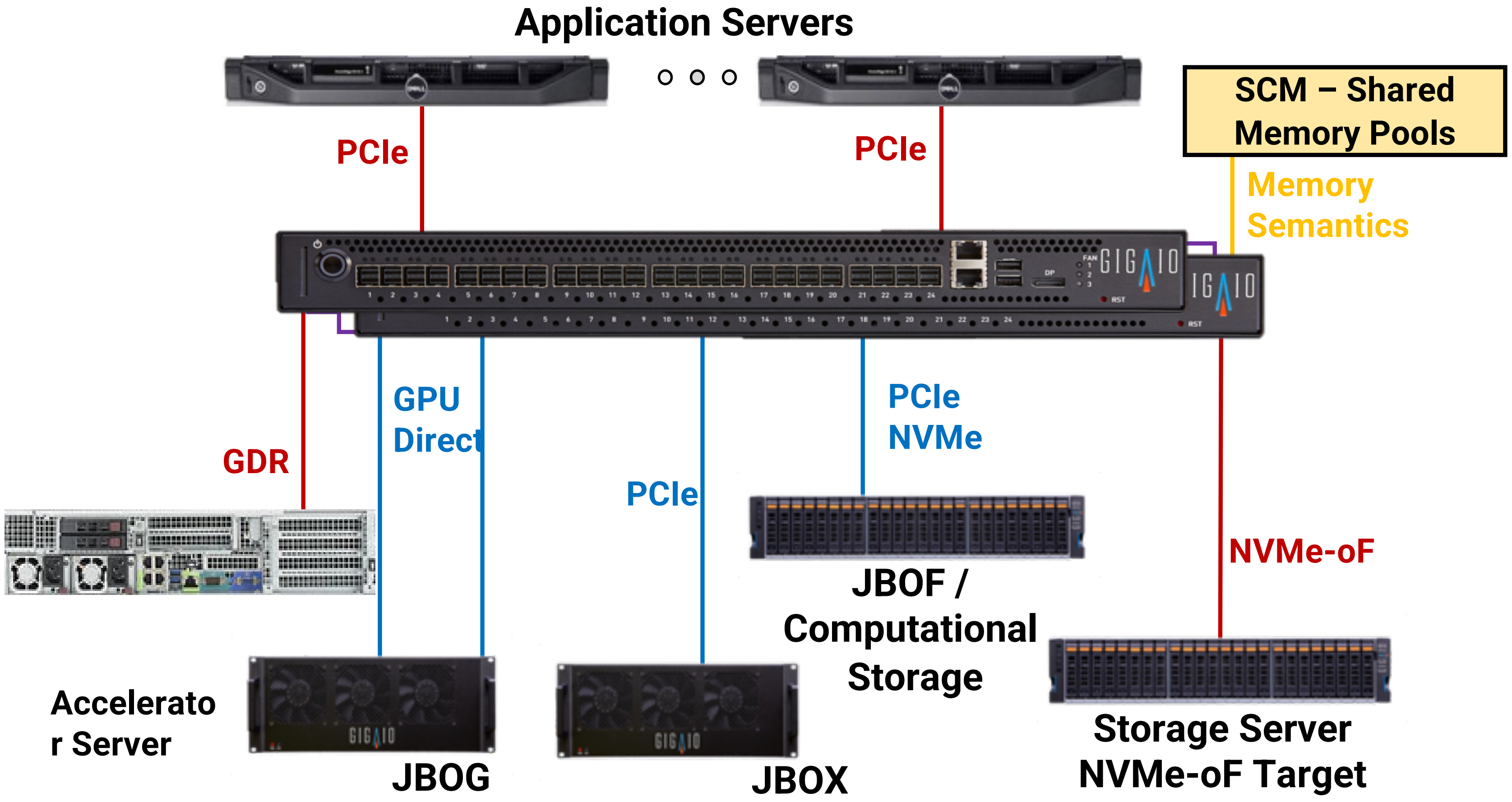
+

Deliver Scale
Ensure Easy Integration

=

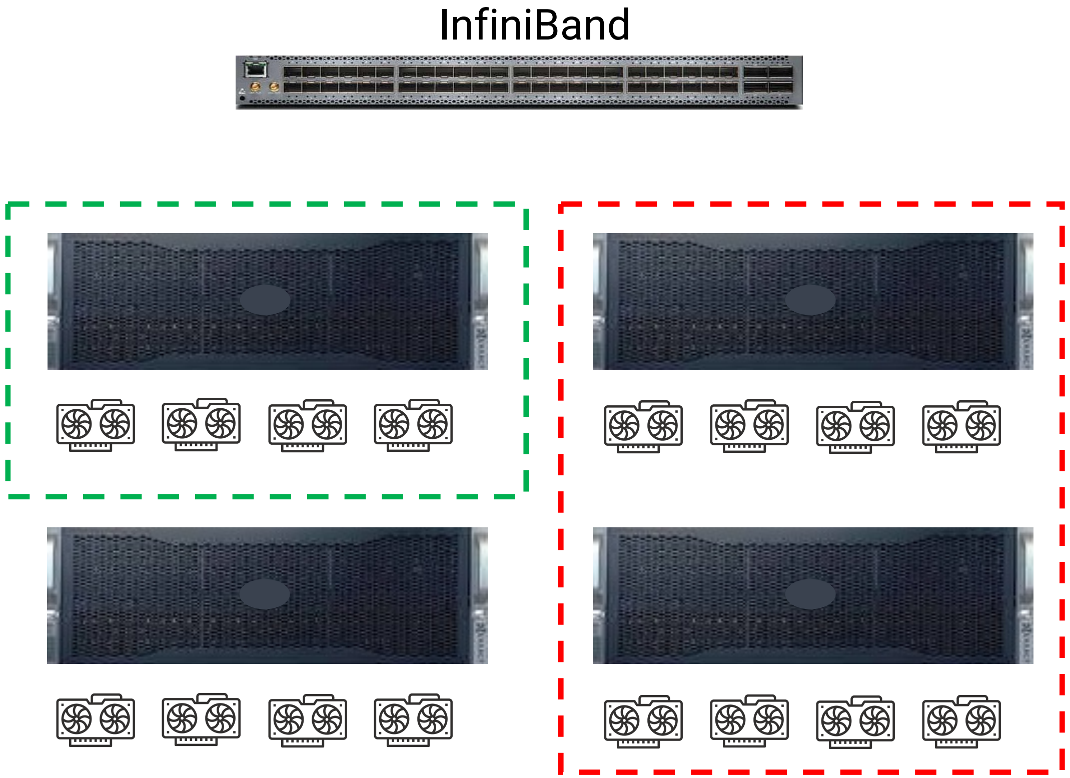
Rack Scale Composition
Any Server. Any Device. Any Time.

# FabreX: Scale Up + Scale Out

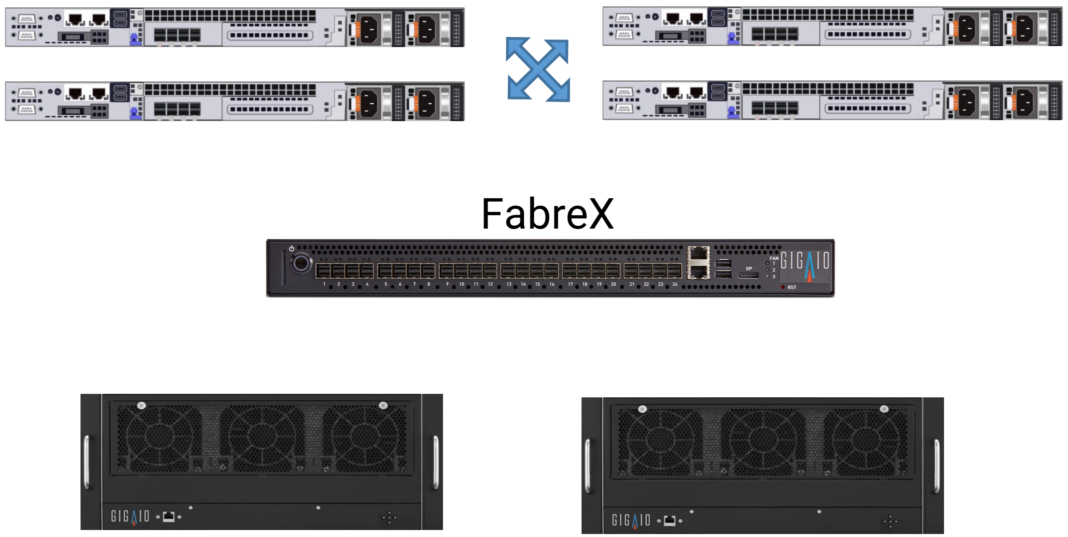


# Resource Composition in a Multi-tenant Environment

## Today’s Static Architecture

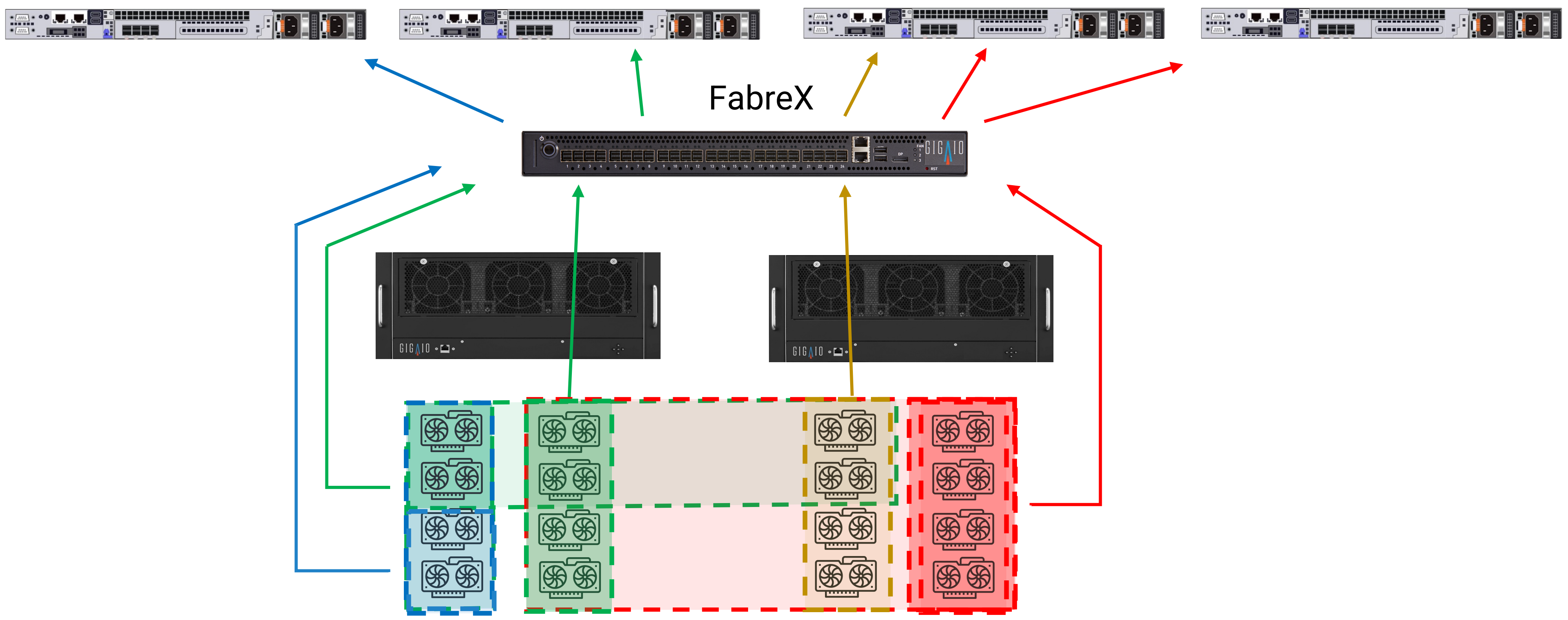


## Tomorrow’s Composable Architecture



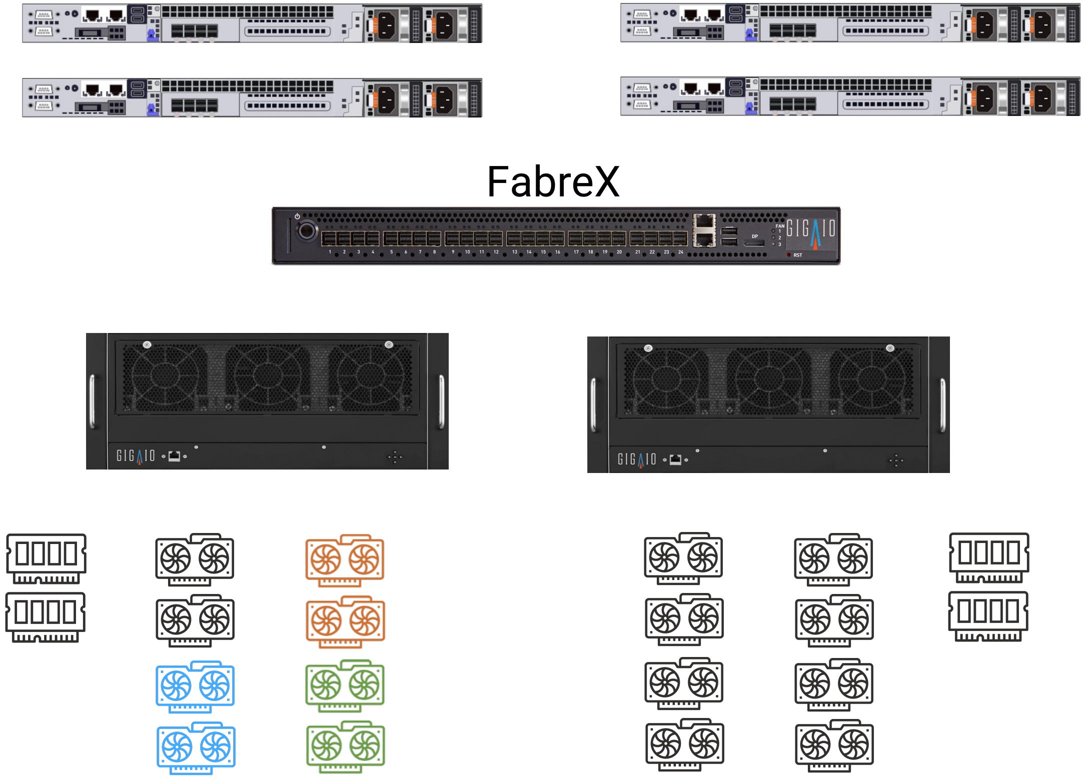
# FabreX Resource Composition

## Simple Composition and Resource Sharing

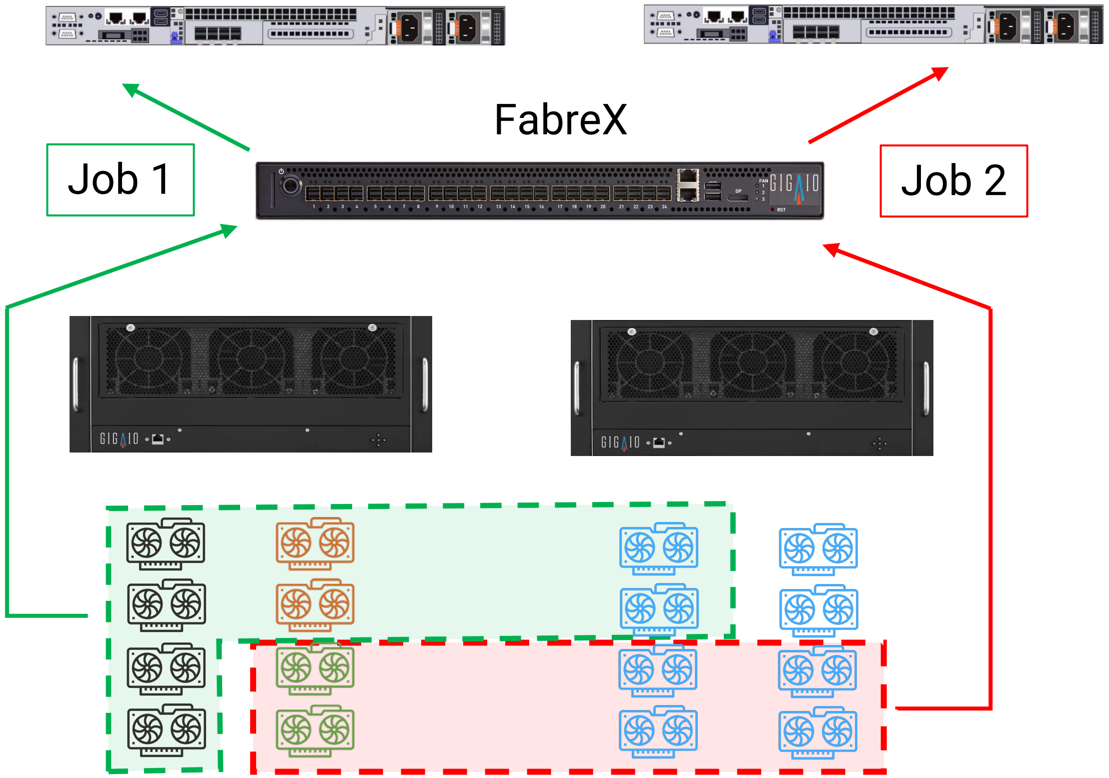


# FabreX Resource Composition

## Tremendous Flexibility in Selecting Accelerators

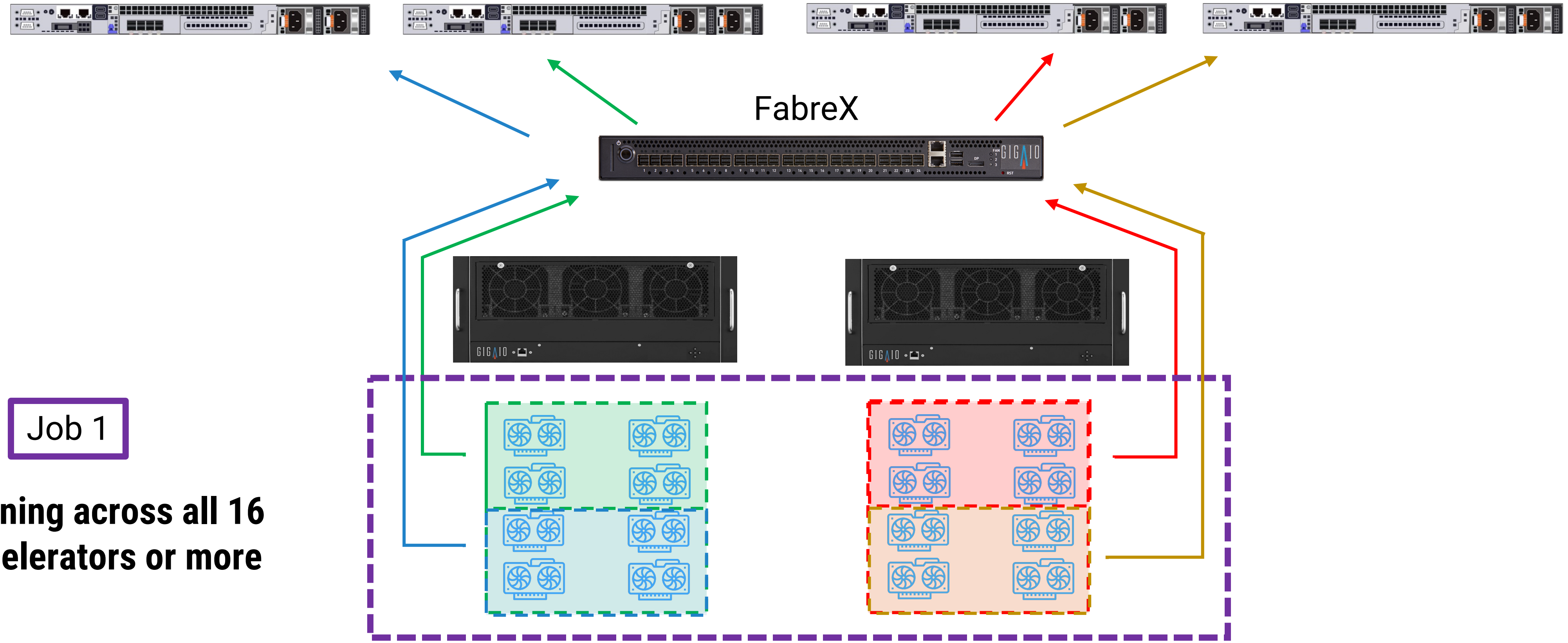


# Let the Workflow Drive the Optimal Composition



# FabreX Server to Server Capabilities

## Extended Composition – Achieving Larger Scale



Job 1

**Running across all 16 Accelerators or more**

# The Complete Solution

## Certified, Ready-to-Run Orchestration Software

Composition with FabreX is built right in

- Slurm – OpenStack – Containers – Virtual Machines – Private Cloud - Bare metal

## Software: Switch and Host

Composability
+ GPUDirect P2P

**Composer**

Composability
+ Multi Host
+ MPI
+ NVMe-oF
+ TCP/IP
+ GPUDirect RDMA

**Leader**

Composability
+ Beyond 2 Leader Switches

**Maestro**

## Hardware – Gen3 and Gen4

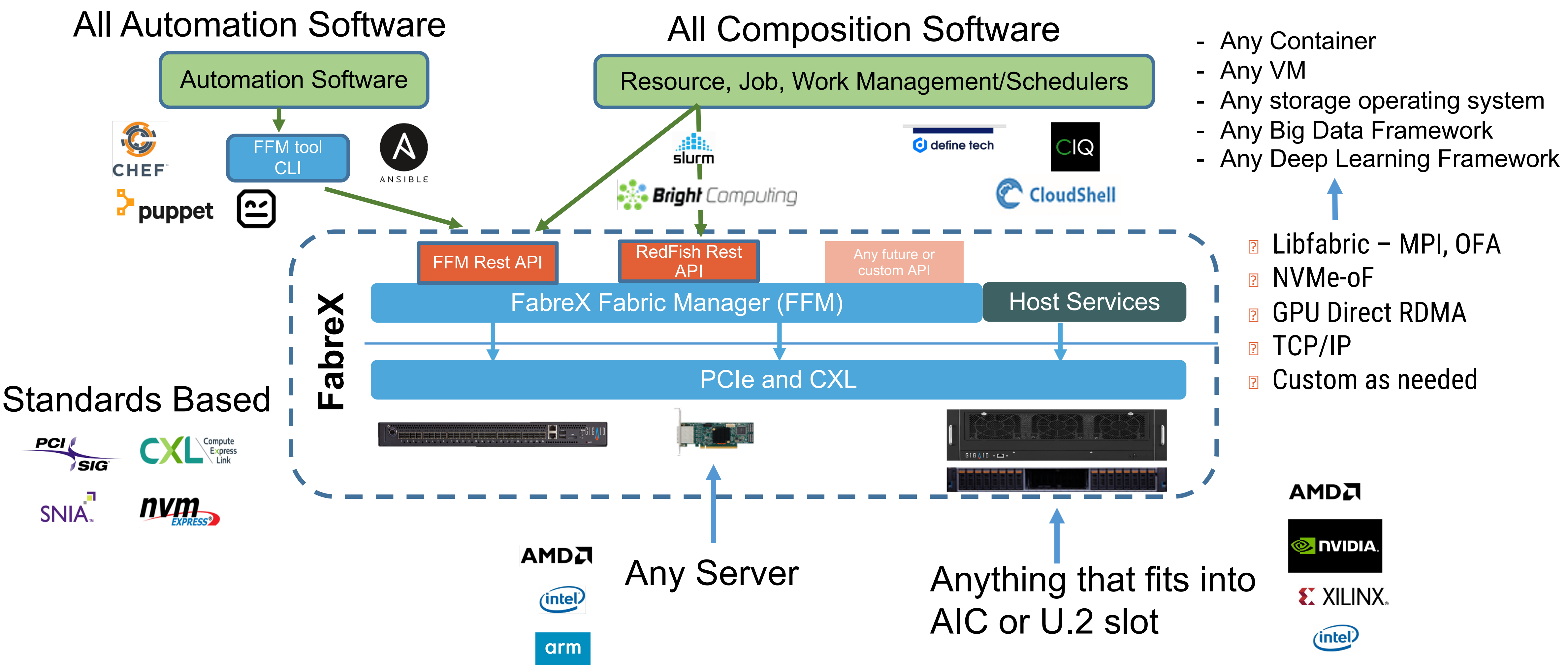
FabreX™ Switches

Link Cards

Cables

Resource Boxes

# FabreX Was Architected from the Beginning to be Open



All Automation Software

Automation Software

FFM tool CLI

All Composition Software

Resource, Job, Work Management/Schedulers

- Any Container
- Any VM
- Any storage operating system
- Any Big Data Framework
- Any Deep Learning Framework

FFM Rest API

RedFish Rest API

Any future or custom API

FabreX Fabric Manager (FFM)

Host Services

PCIe and CXL

FabreX

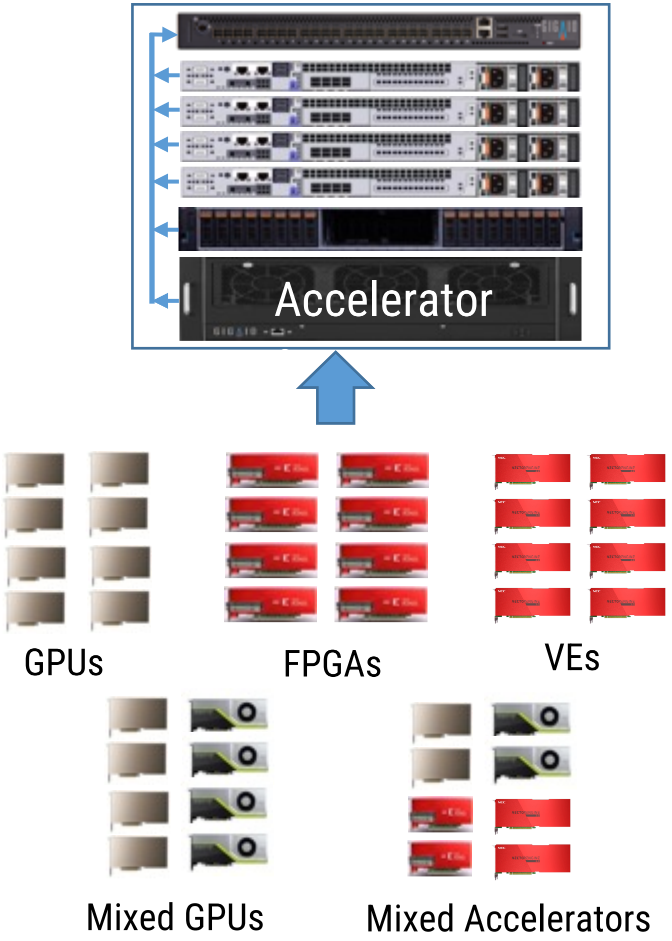
- Libfabric – MPI, OFA
- NVMe-oF
- GPU Direct RDMA
- TCP/IP
- Custom as needed

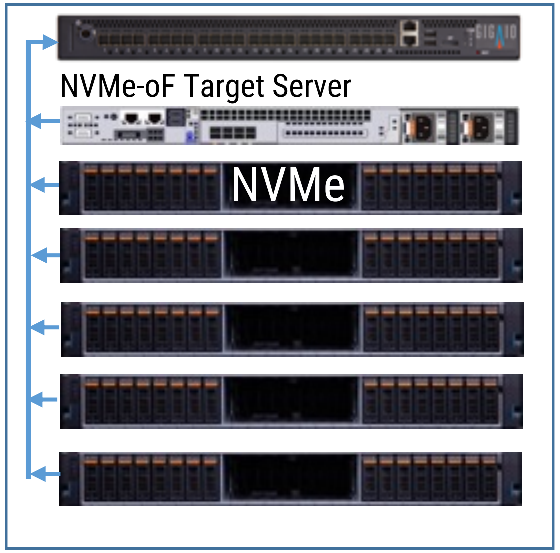
Standards Based

Any Server

Anything that fits into AIC or U.2 slot

# Limitless Variation to a GigaCell

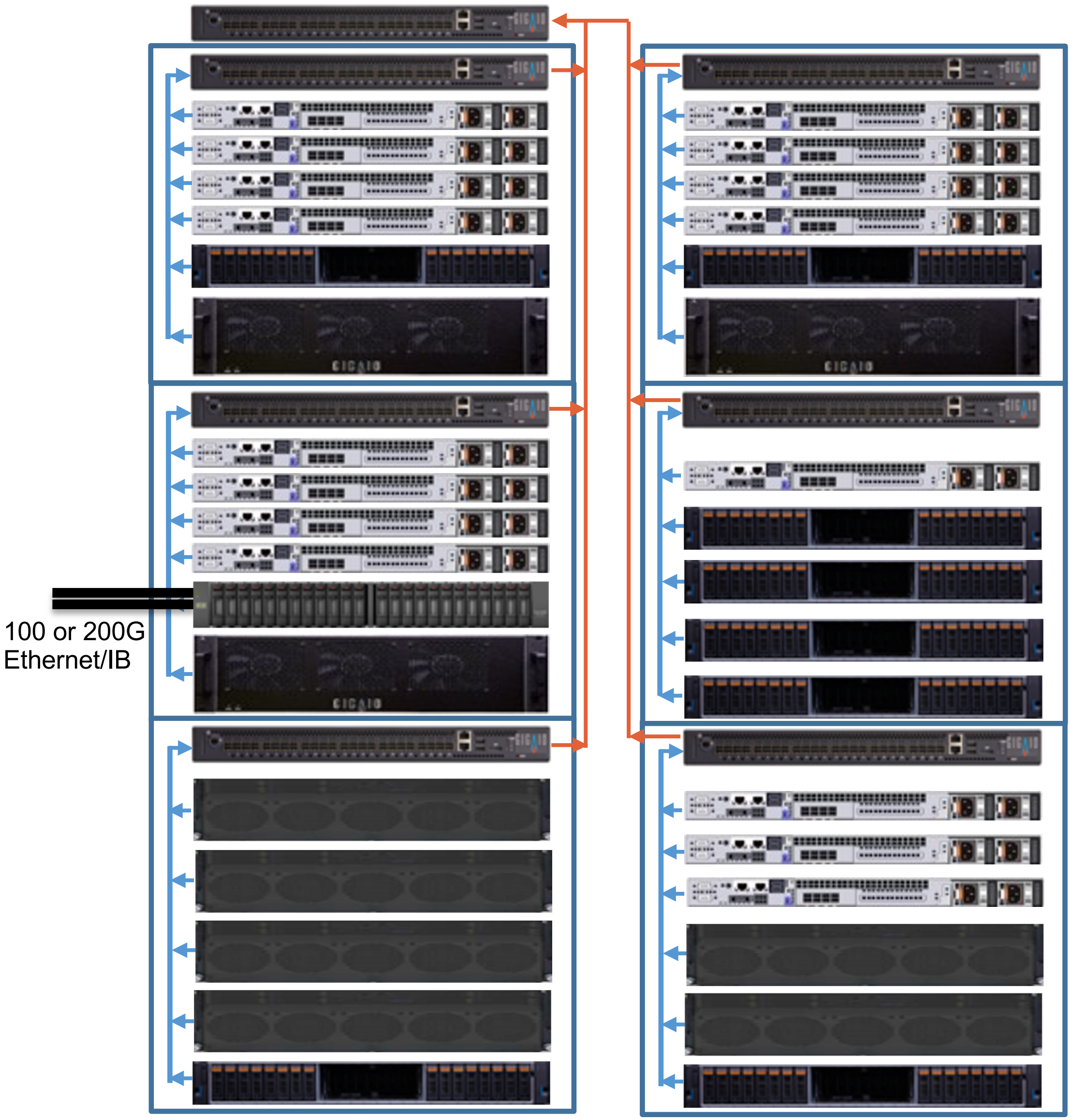




Storage Cell
With up to 3PB

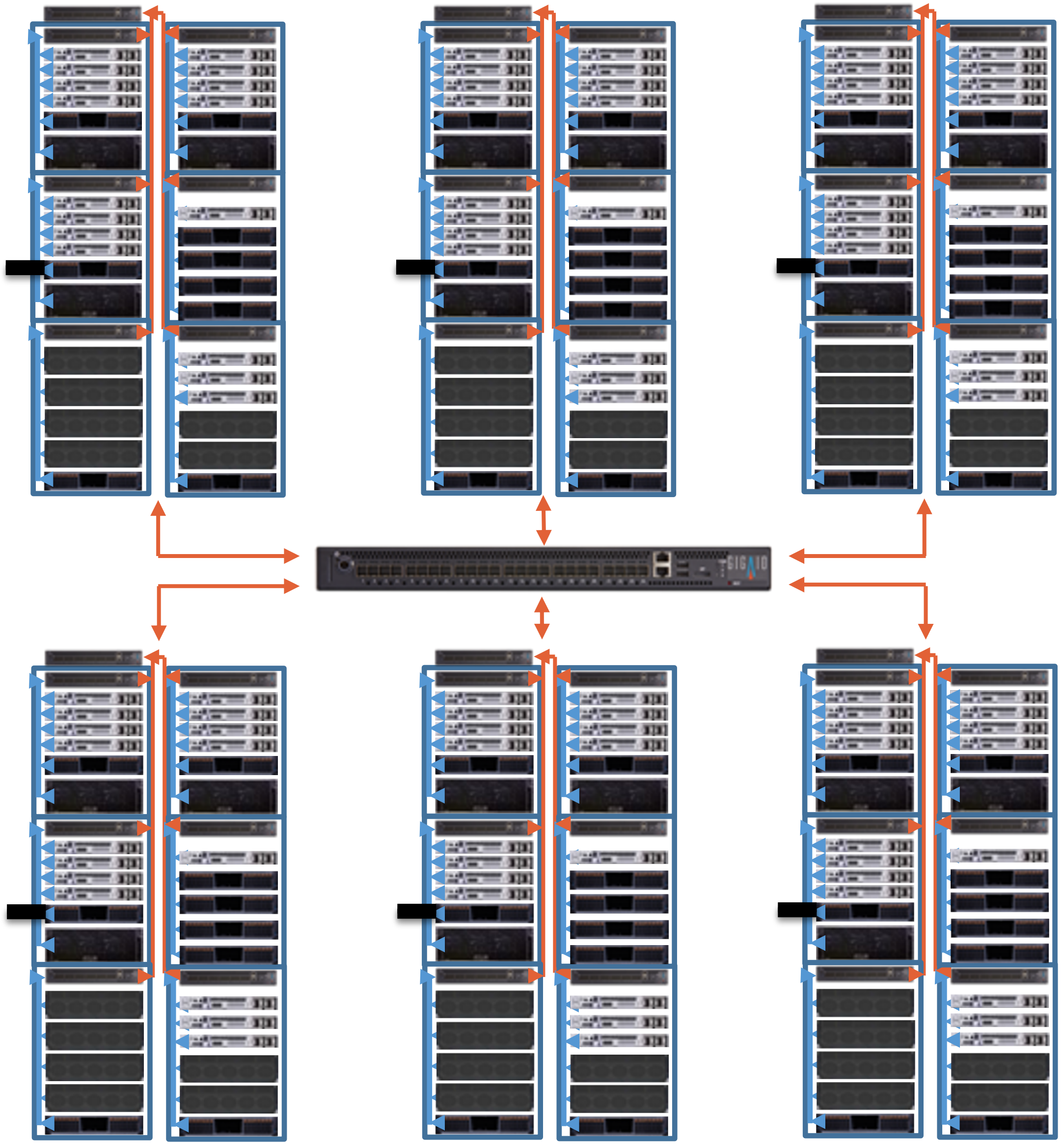
# Simply combine up to 6 cells via FabreX (PCIe / CXL) To create a GigaPod™

# Rack Scale Computing Made Simple



# To Scale Out

Combine up to 6 GigaPods to create a GigaCluster™



\Orchestrating a brighter world

NEC

GIGAIO

Cloud Agility.
Half the TCO.

# NEC Vector Engine Performance

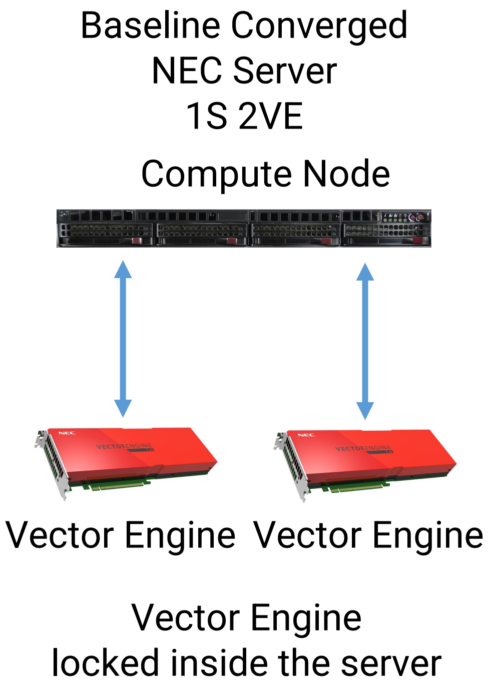
ISC
High Performanc

# Results - Summary

- Objectives:
  - Execute industry standard benchmarks in Converged and Composed configurations
    - Converged – all resource inside the server
    - Composed – all resources inside Accelerator Pooling Appliance and share across servers using GigaIO FabreX
  - Compare results

- Summary
  - Vector Engine is 100% PCIe compliant
  - Simply plugged, recompiled applications and it just worked
  - System software all worked
  - Vector Engines can be shared between multiple servers
  - Vector Engines can be dynamically reconfigured across servers
  - Performance identical in all configurations
    - No performance overhead with FabreX

# Test Configurations

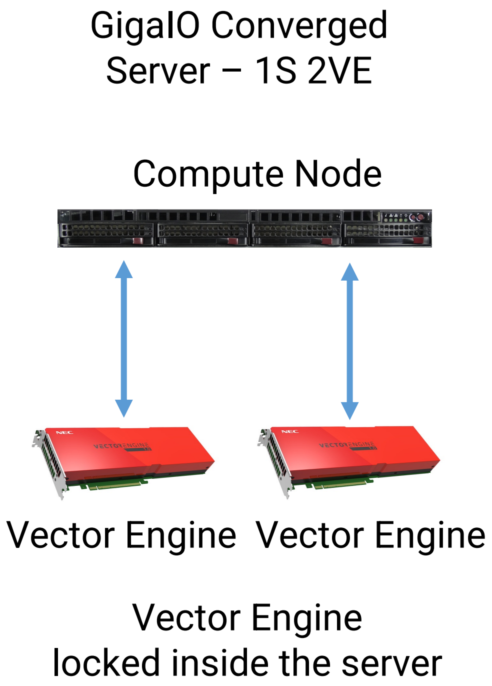


Baseline Converged NEC Server
1S 2VE

Compute Node

Vector Engine Vector Engine

Vector Engine locked inside the server

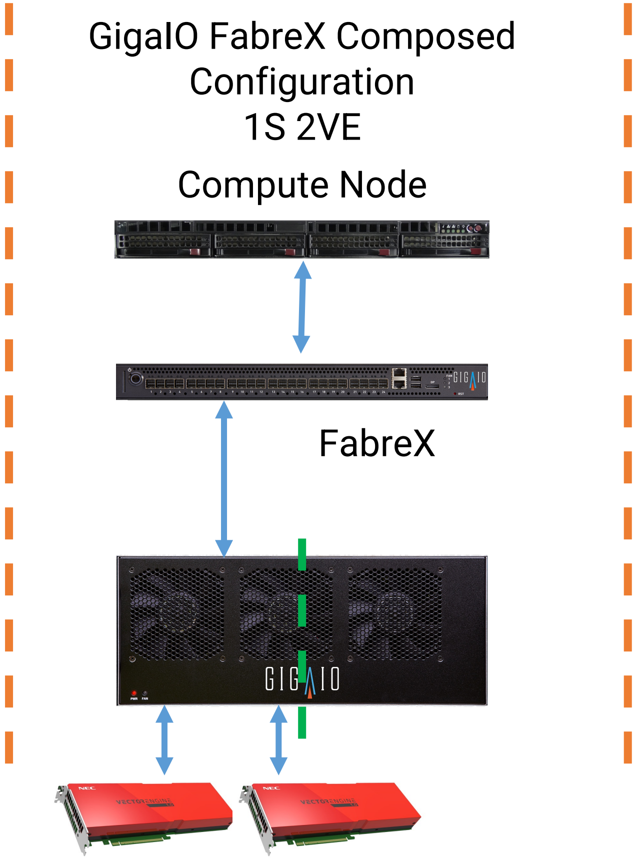


GigaIO Converged Server – 1S 2VE

Compute Node

Vector Engine Vector Engine

Vector Engine locked inside the server

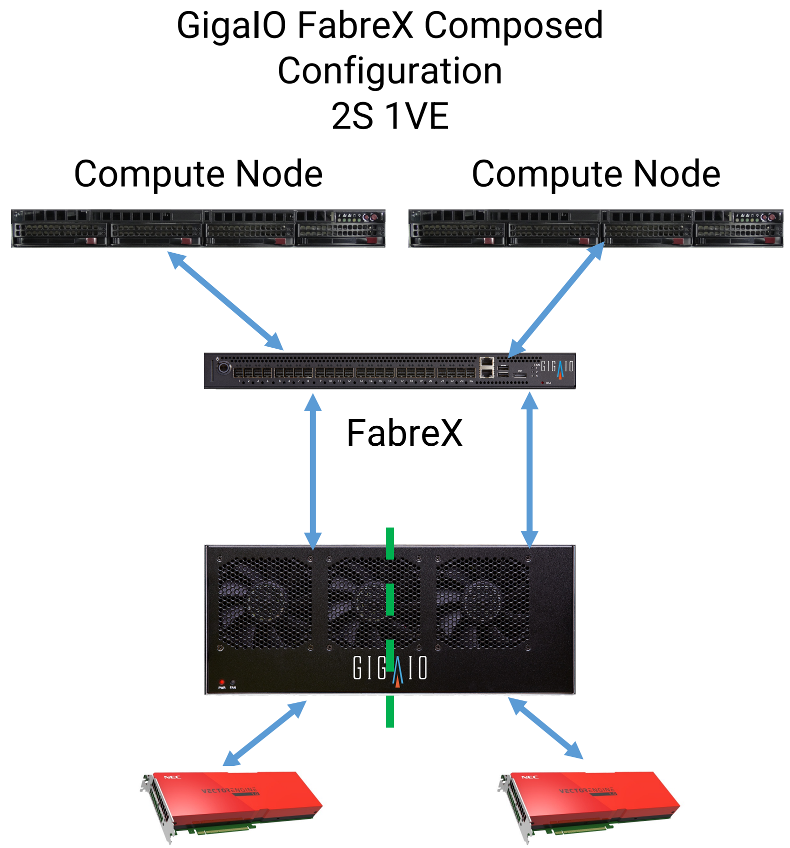


GigaIO FabreX Composed Configuration
1S 2VE

Compute Node

FabreX

Vector Engines inside the Accelerator Pooling Appliance and shared between all servers on FabreX



GigaIO FabreX Composed Configuration
2S 1VE

Compute Node Compute Node

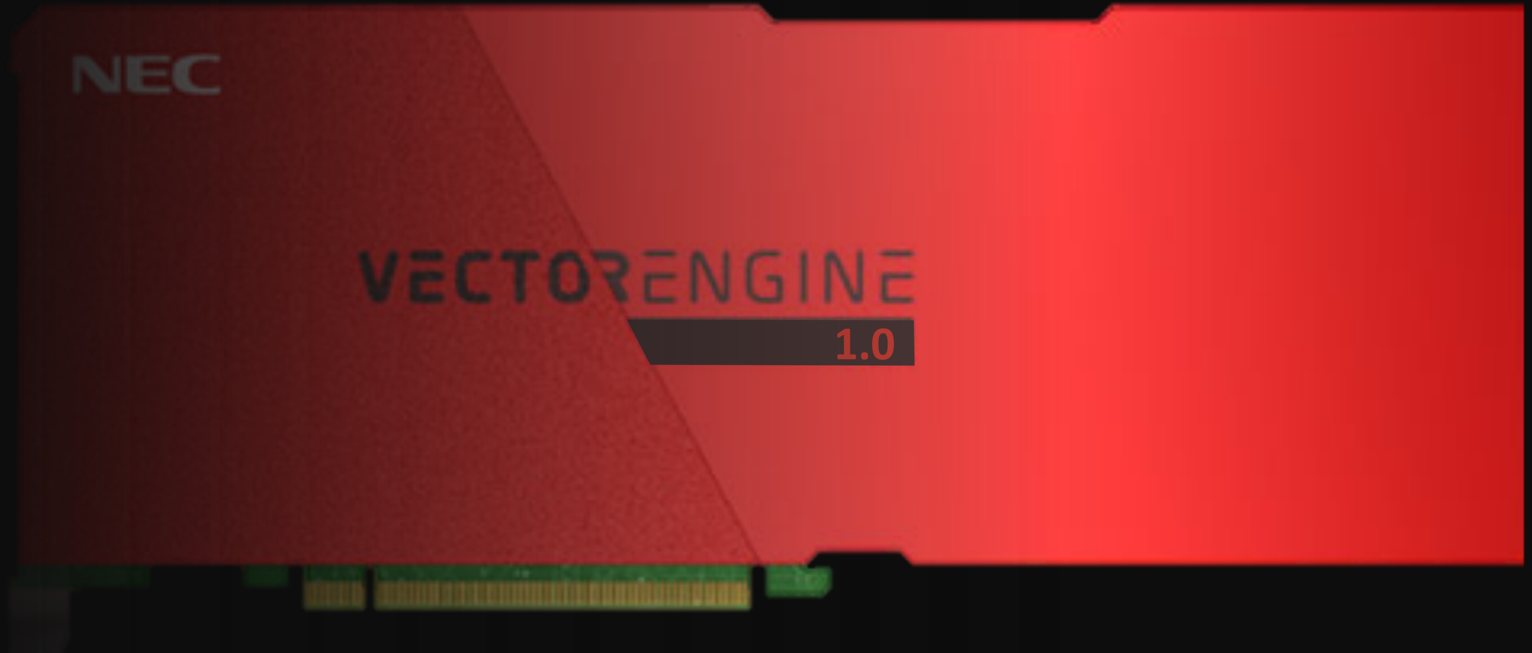
FabreX

Vector Engines inside the Accelerator Pooling Appliance and shared between all servers on FabreX

NOTE: Future test in Multi VE configuration supporting RDMA operation

# Benchmark Results

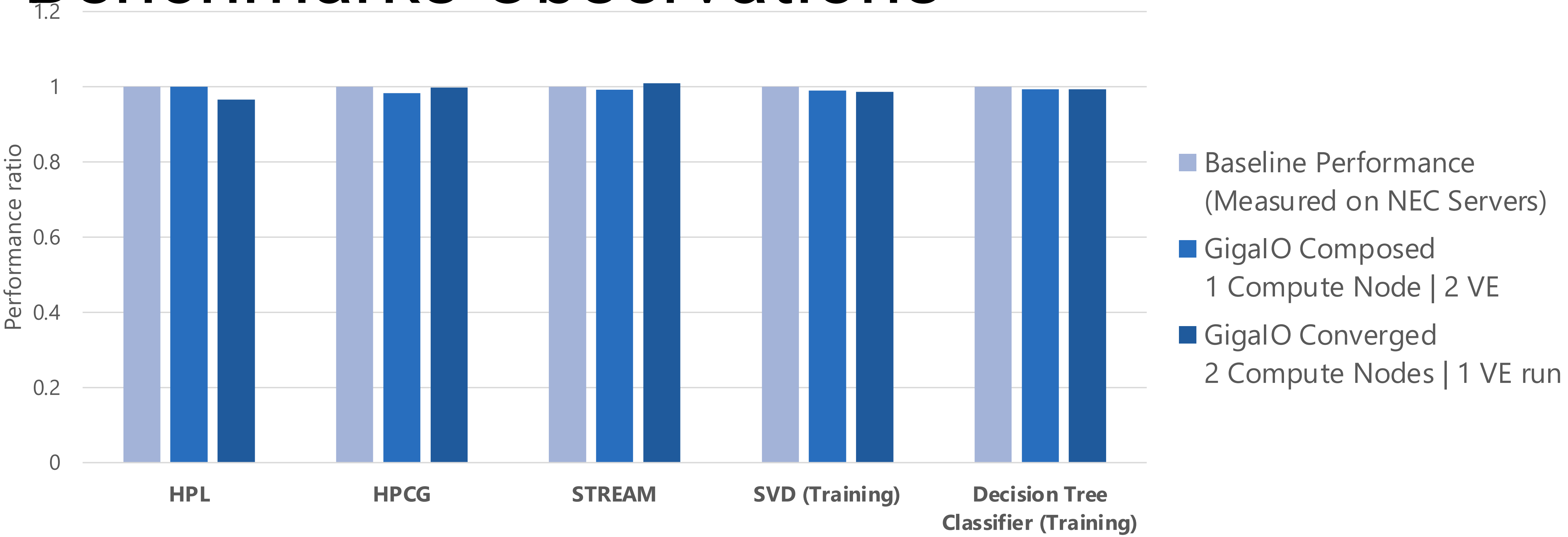
NEC Vector Engine



# Benchmarks Test Description

- **HPL** -- the High-Performance Computing LINPACK Benchmark solves a (random) dense linear arithmetic on distributed-memory computers.
- **HPCG** -- The High-Performance Conjugate Gradients (HPCG) complements the High Performance LINPACK (HPL) benchmark, currently used to rank the TOP500 computing systems.
- **STREAM** -- a simple synthetic benchmark program that measures sustainable memory bandwidth (in MB/s)
- **SVD** – Singular Value Decomposition (SVD), widely used matrix decomposition method.
- **Decision Tree Classifiers** – used successfully in many diverse areas including machine learning.

# Benchmarks Observations

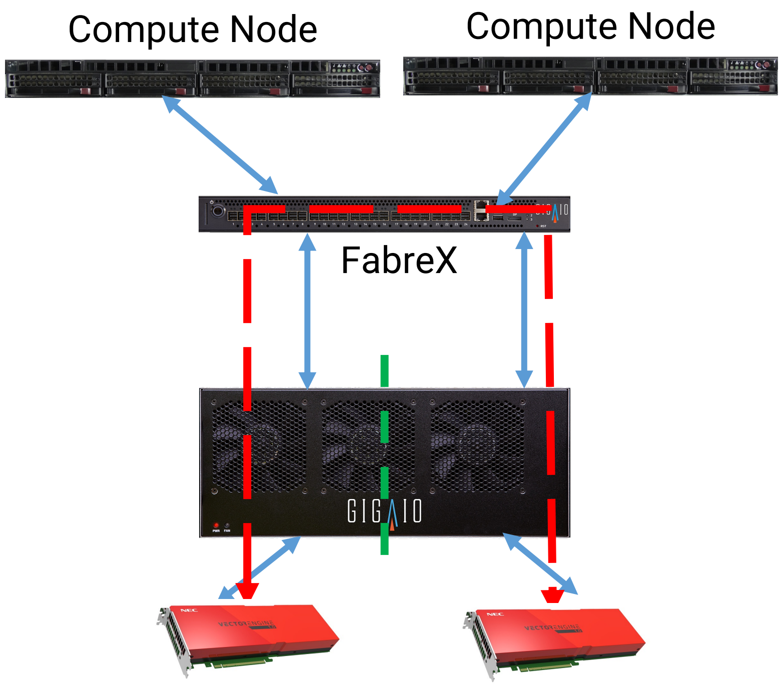


- Current performance on GigaIO composed and GigaIO converged configurations are almost identical, as well as the performance measured on NEC servers.
- More converged configurations need to be supported and evaluated.

# Next steps…….

- Test additional composed configurations
  - Multiple servers with Multiple VEs
    - RDMA mode with MPI traffic flowing between Vector Engines without going through the server
    - Higher performance due to lower latency

GigaIO FabreX Composed Configuration
2S 1VE (use 2 VE)



Vector Engines inside the Accelerator Pooling Appliance and shared between all servers on FabreX

# Summary

- IT is being asked to support ever expanding workloads and diversifying accelerated computing technology – on the same budget.
- Each workload is “lumpy” in its own way – and different architectures maximize performance for different applications.
- FabreX – the next-gen Universal Dynamic Fabric enables IT’s to improve system performance, incorporate the latest technology, revitalize existing infrastructure, and meet budget and sustainability goals.
- FabreX composable architecture with NEC Vector Engines delivers performance
  - Expect to improve performance running multiple VEs across FabreX……
- Available today in production

\Orchestrating a brighter world

NEC

GIGAIO

# *Compose. Compute. Achieve.*

## *Thank You*

For more information visit:

- www.gigaio.com
- Email: mlehrer@gigaio.com

Questions?

ISC
High Performanc