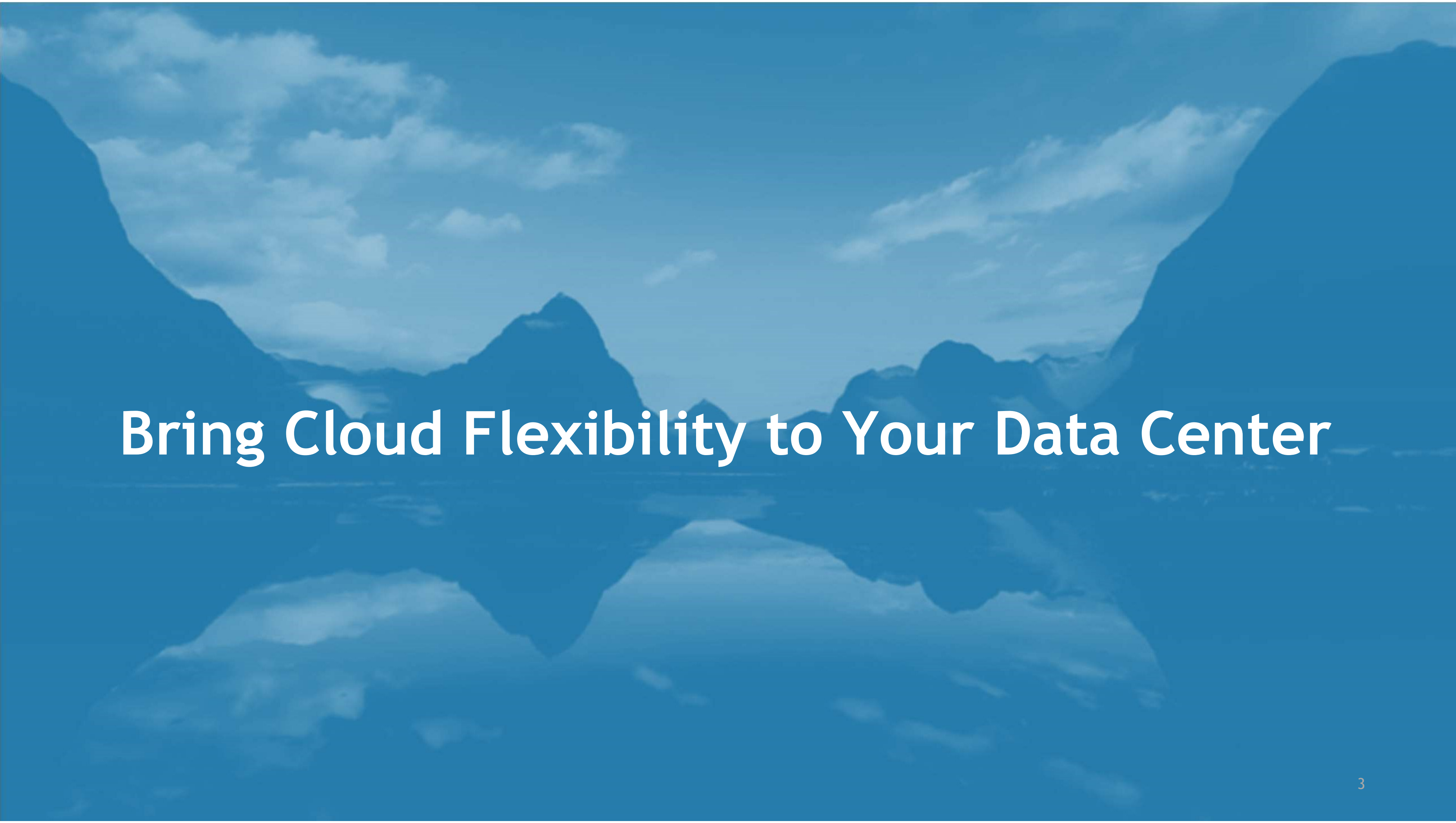
NEC Vector Engines on GigaIO’s Rack -Scale Computing Platform

AGENDA

- What is GigaIO
- Memory Fabric Introduction
- Composing Across a Memory Fabric
- Scaling out your Rack-Scale Computing Platform
- GigaIO + NEC Vector Engines = Performance



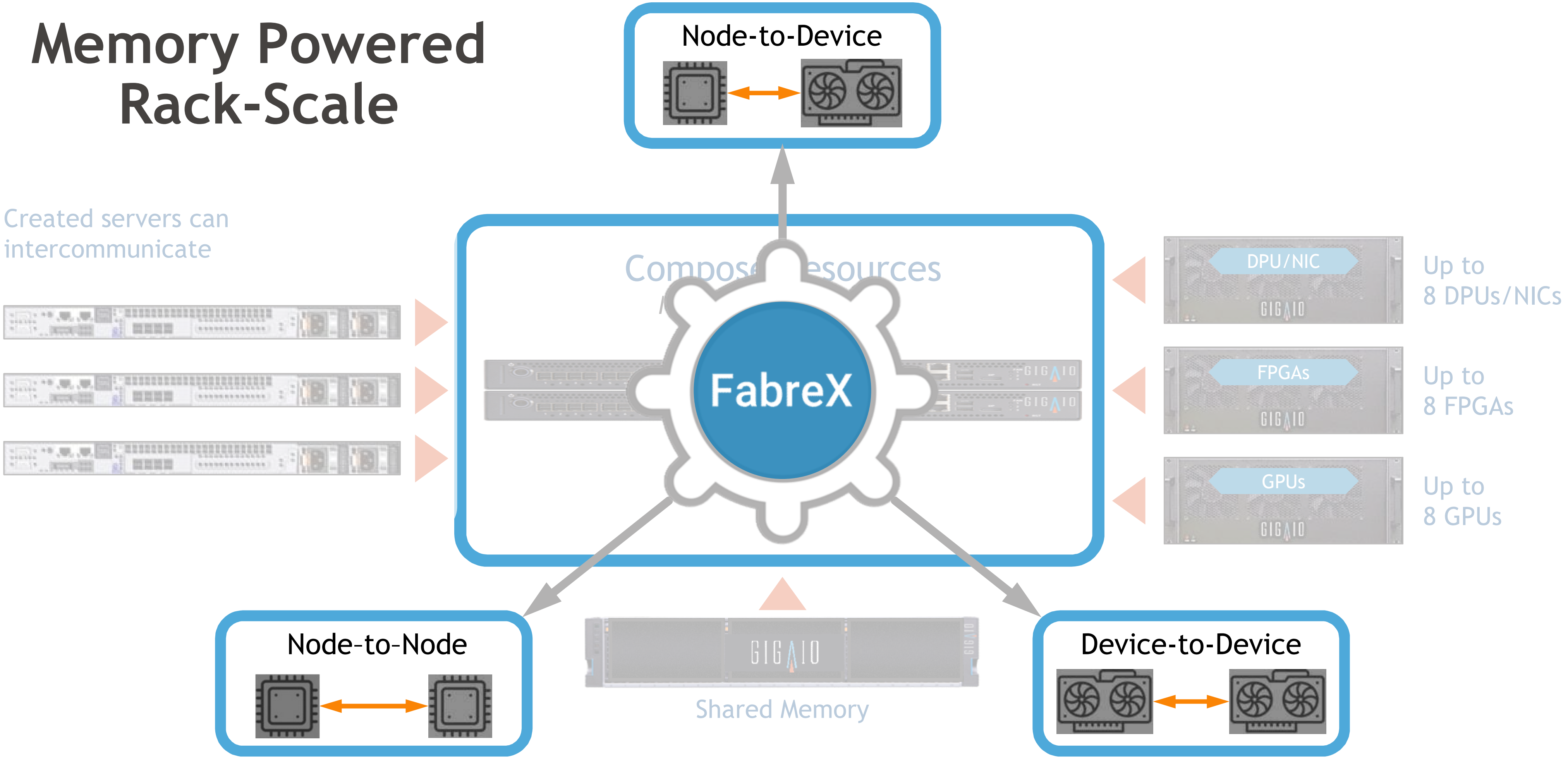
Bring Cloud Flexibility to Your Data Center

THE BEST OF BOTH WORLDS

GigaIO’s FabreX Rack-Scale Computing Platform

- Works with **all** workloads
- Works with **hybrid** and **multi-cloud** environments
- Brings “software-defined hardware” flexibility and agility to on-premise
- Creates “impossible servers” and combinations of accelerators not available in the cloud
- Optimizes resource utilization on-premise
- Maximizes device compatibility across infrastructure
- Decreases time-to-insight by democratizing access to specialized compute

Memory Powered Rack-Scale



Node-to-Device

Created servers can intercommunicate

Compos esources

FabreX

DPU/NIC

Up to 8 DPUs/NICs

FPGAs

Up to 8 FPGAs

GPUs

Up to 8 GPUs

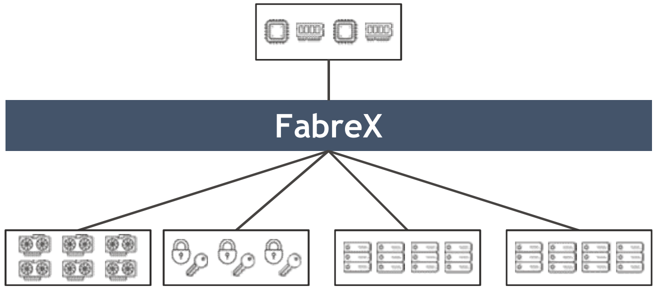
Node-to-Node

Shared Memory

Device-to-Device

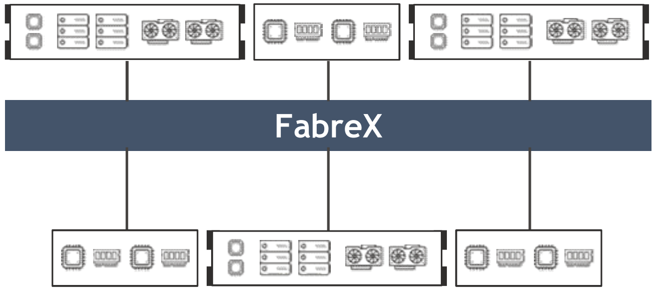
FabreX™

- Memory Simplifies all Communications



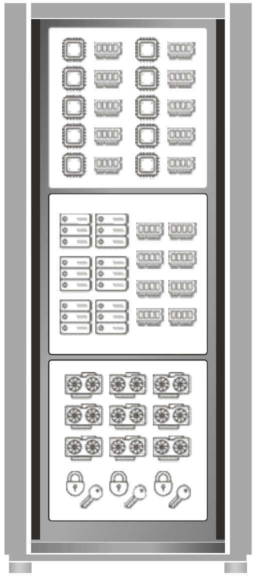
Server Composition

+



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

=



All running on one Memory Fabric without performance penalty

Minimize TCO
Improve Serviceability

+

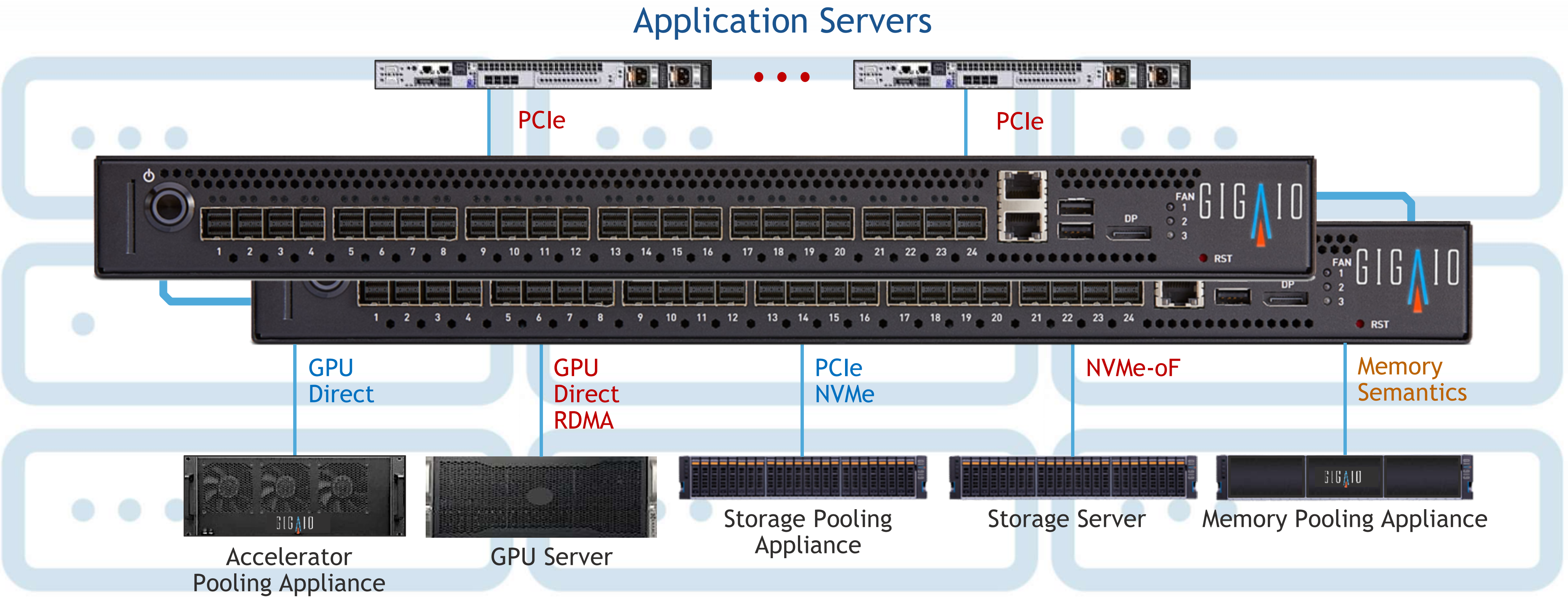
Deliver Scale
Ensure Easy Integration

=

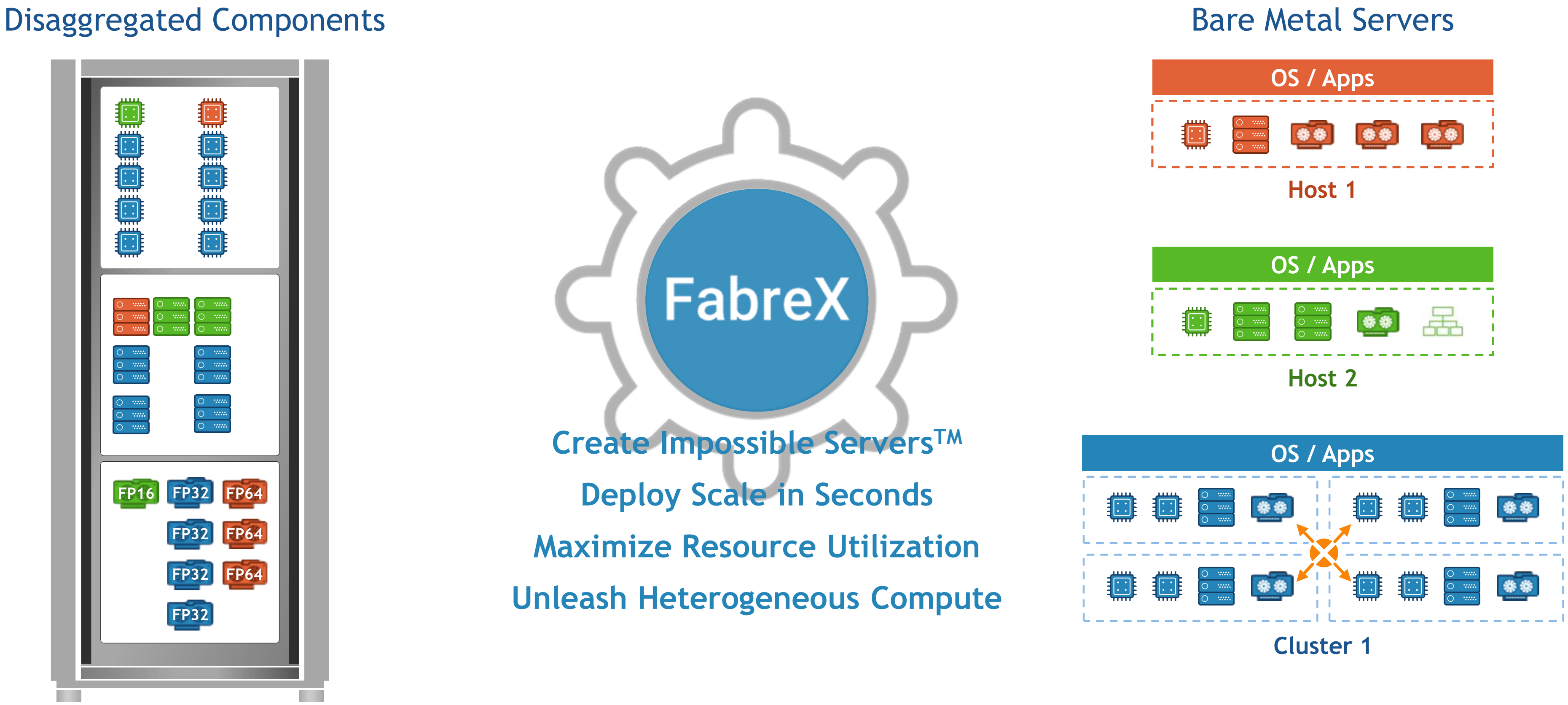
Rack Scale Composition

Any Server.
Any Device.
Any Time.

HARDWARE & CONNECTIVITY



Composing Resources in GigaIO's Rack-Scale Computing Platform



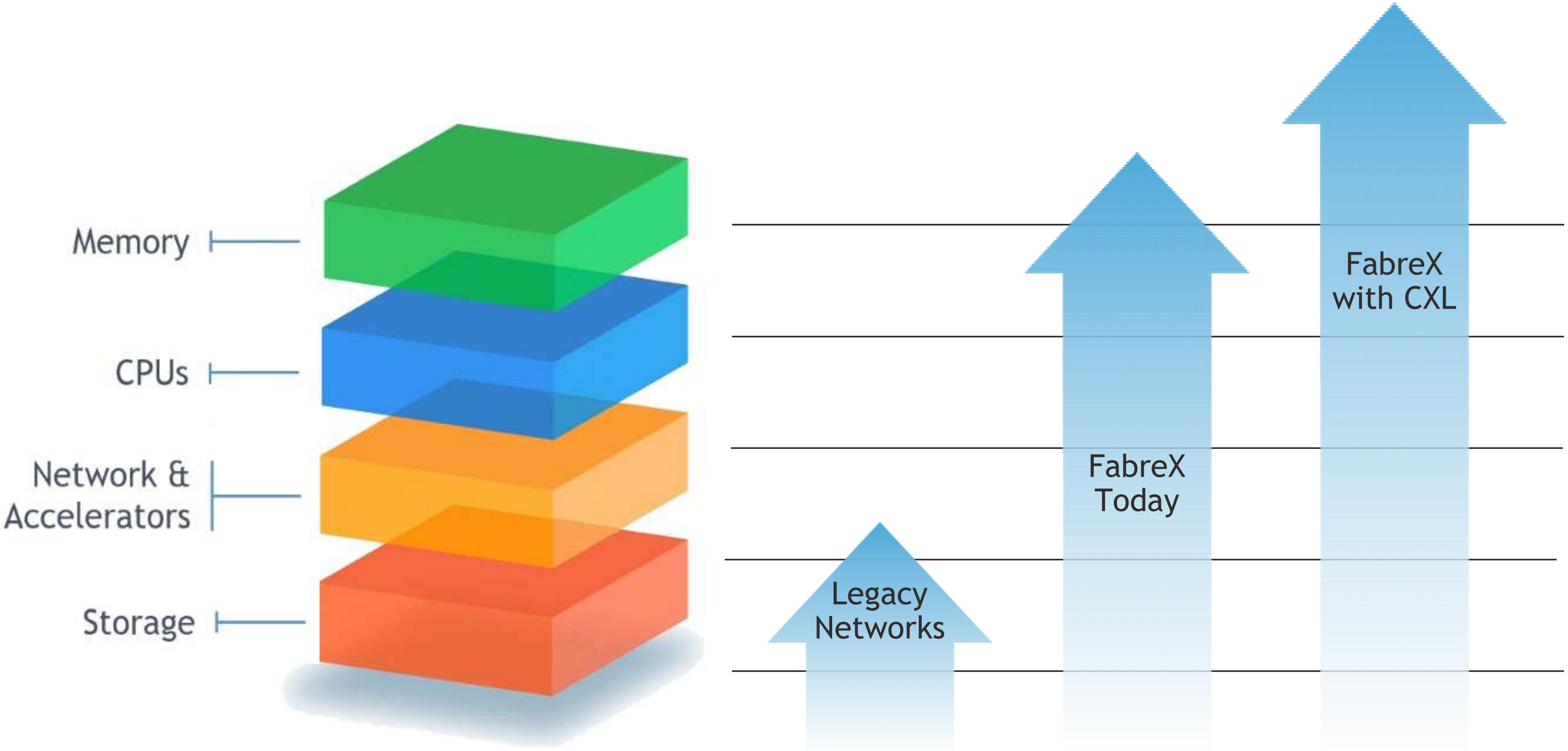
Create Impossible Servers™

Deploy Scale in Seconds

Maximize Resource Utilization

Unleash Heterogeneous Compute

THE MARCH OF COMPOSABILITY



BENEFITS OF RACK-SCALE COMPUTING PLATFORM

Optimize Capex and Opex

Pay as you Grow

Scale Resources Independently

Faster Time to Insight

Better Business Agility

Improved Sustainability

Impossible Servers

Reduced Power & Cooling (Sustainability)

Better Managed Life Cycles

ORCHESTRATION & MANAGEMENT INTEGRATIONS

Open Architecture

Open Ecosystem

Redfish APIs

Bring Your Own or Use a Pre-Integrated Solution

NORTH-BOUND INTEGRATIONS

Fully Vetted Jointly Engineered Solutions

Bright Computing

slurm workload manager

Quali CLOUDSHELL

SuperCloud Composer

ALTAIR Altair® PBS Professional®

kubernetes

CIQ

define tech

Any infrastructure automation tool

Red Hat Ansible

puppet

CHEF Progress

DMTF Redfish

Jenkins

ROBOT FRAMEWORK

GIGAIO

RACK SCALE OPTIONS

GigaCell

GigaPod

GigaCluster

GIGACELL

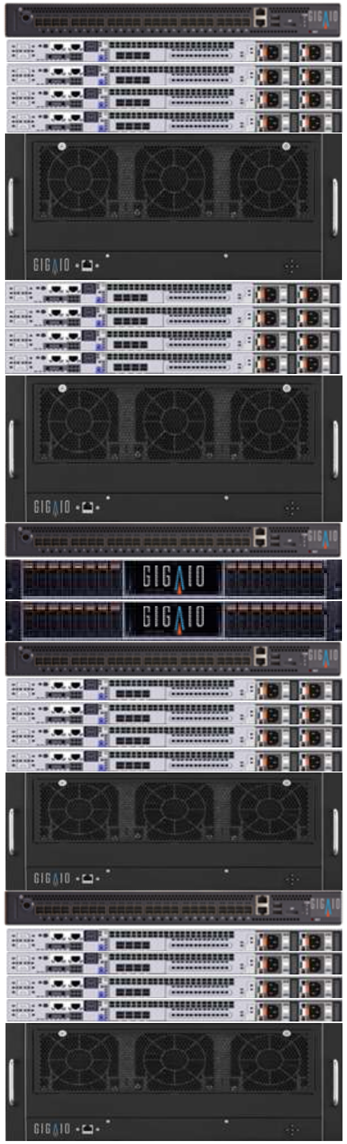
The power user or small workgroup supercomputer

- ¼ Rack – 12u
- 1 x FabreX Switch + FFM Software
- 1 x FabreX Storage Appliance
- (Up to 720 TB)
- 4 x 1U Servers
- 1 Accelerator Pooling Appliance
- (Up to 8 DW GPUs)

GIGAPOD

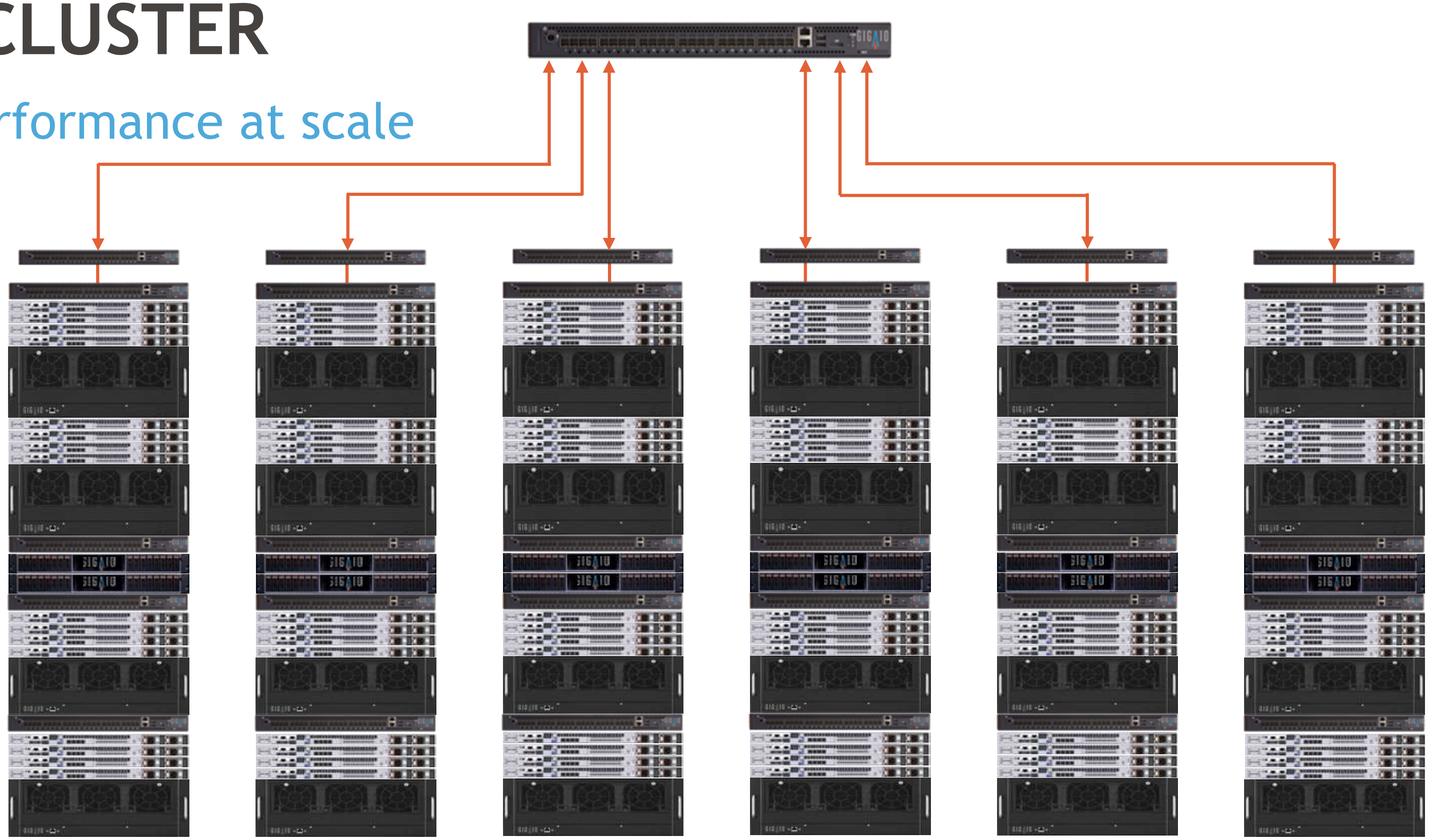
The data center rack-scale building block

- One Rack – 46u
- 5 x FabreX Switches + FFM Software
- 2 x FabreX Storage Appliances
- (Up to 1.44 PB)
- 16 x 1U Servers
- 4 Accelerator Pooling Appliances
- (Up to 32 DW GPUs)



GIGACLUSTER

High performance at scale



©2022 GIGAIO

GIGACELL XTREME

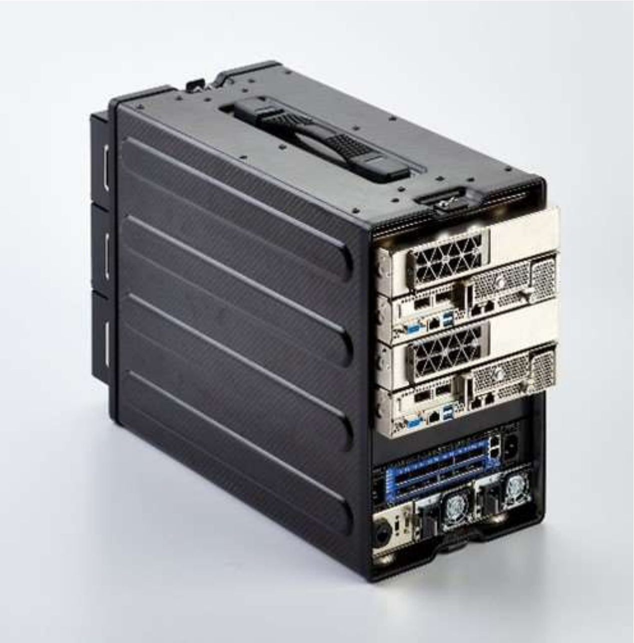
Modular AI-Ready Composable Edge System

- Up to 6 GPUs and and 30TB NVMe storage with Memory switch uplink
- Composable compute, storage, and network with reduced system SWaP
- Fits in overhead bin — TSA carry-on portable size
- All storage easily removable for portability

ISR data analysis

Cognitive visualization

Identity management



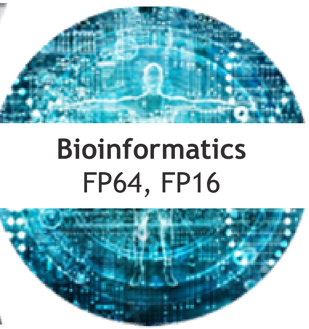
USE CASES

- AI/ML — Data analytics
- High Performance Computing (HPC)
- Visualization / Interactive graphics
- Simulation / Modeling — Digital twin
- Edge / 5G / 6G Networks
- Colo / Hybrid cloud
- Virtual Desktop Infrastructure (VDI)
- Brownfield Installations

"CONDO" CLUSTER
UNIVERSITY HPC

Grant Assistance Provided

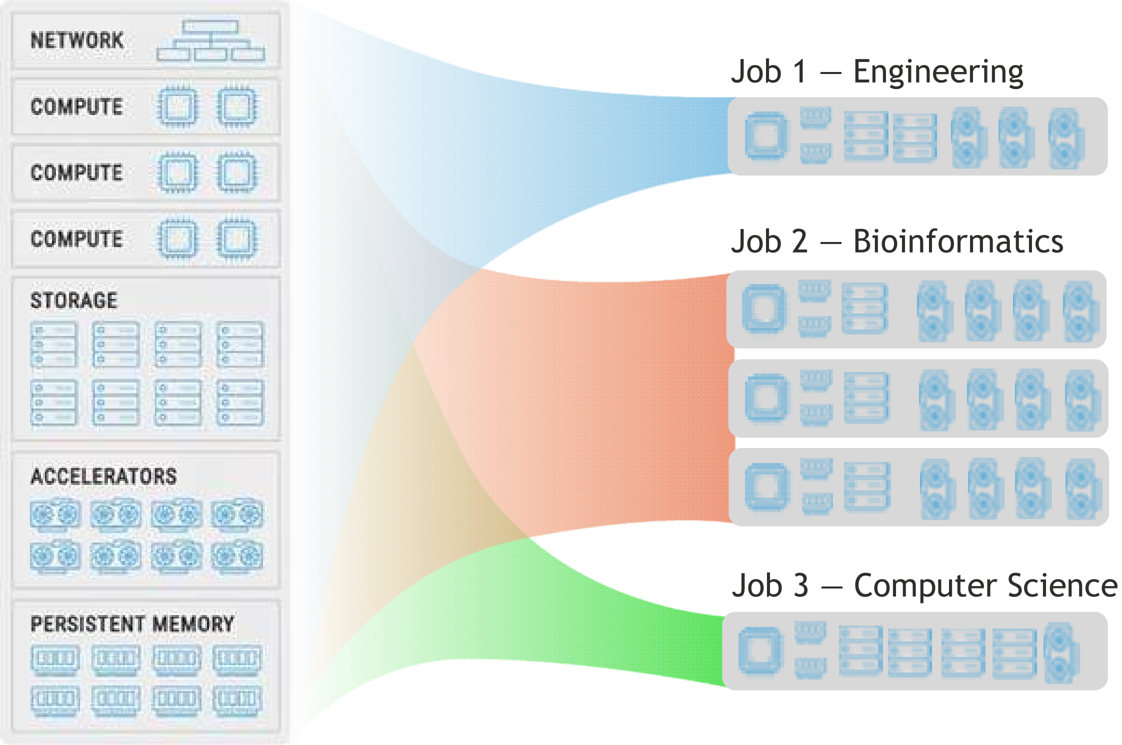
Mechanical Engineering
FP32, Vector Engines



Bioinformatics
FP64, FP16

Computer Science
FPGAS, FP16

Software defines hardware uniquely for each workload



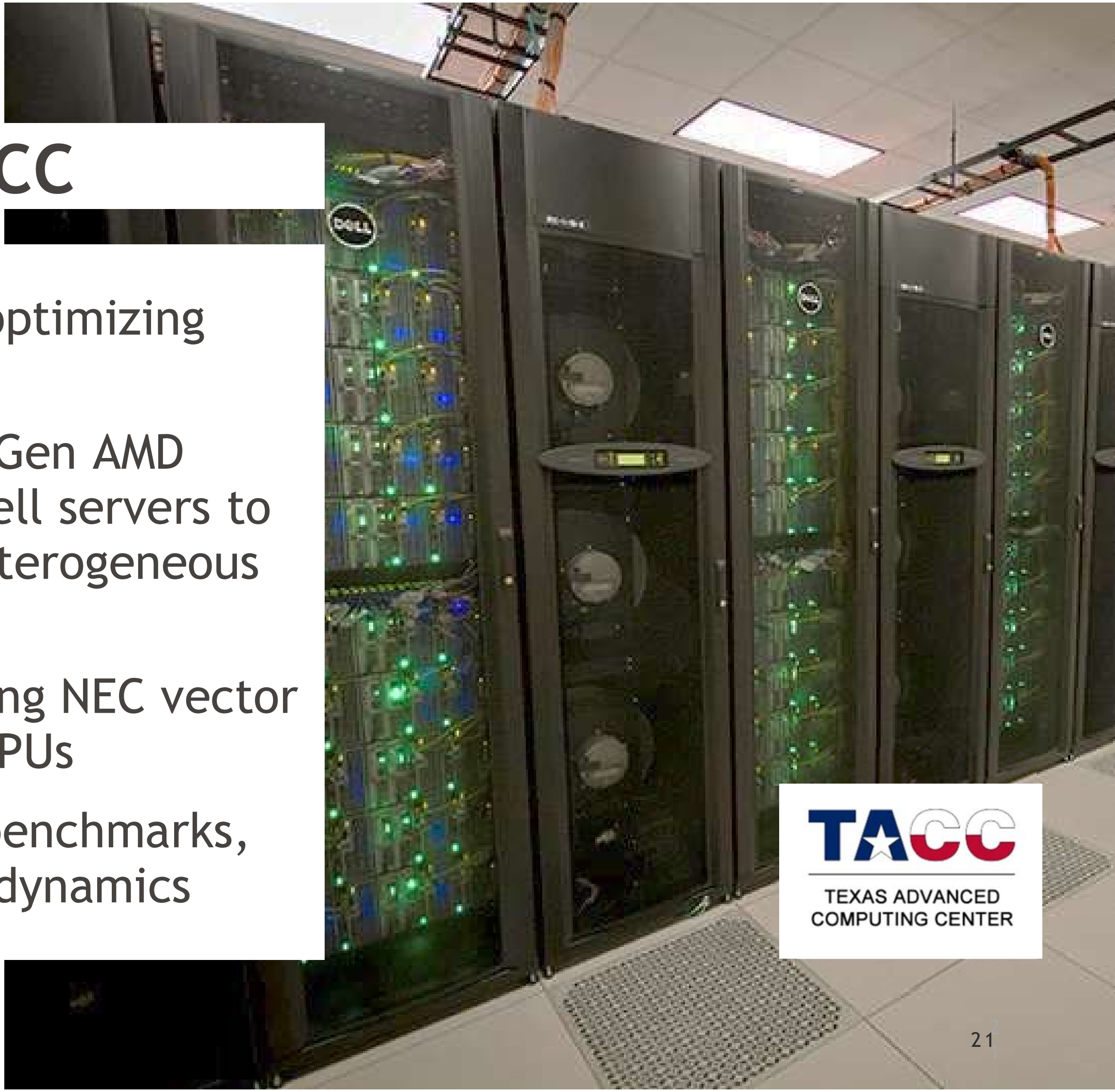
Cluster grows grant by grant

CUSTOMER STORY: INDEX TRADING

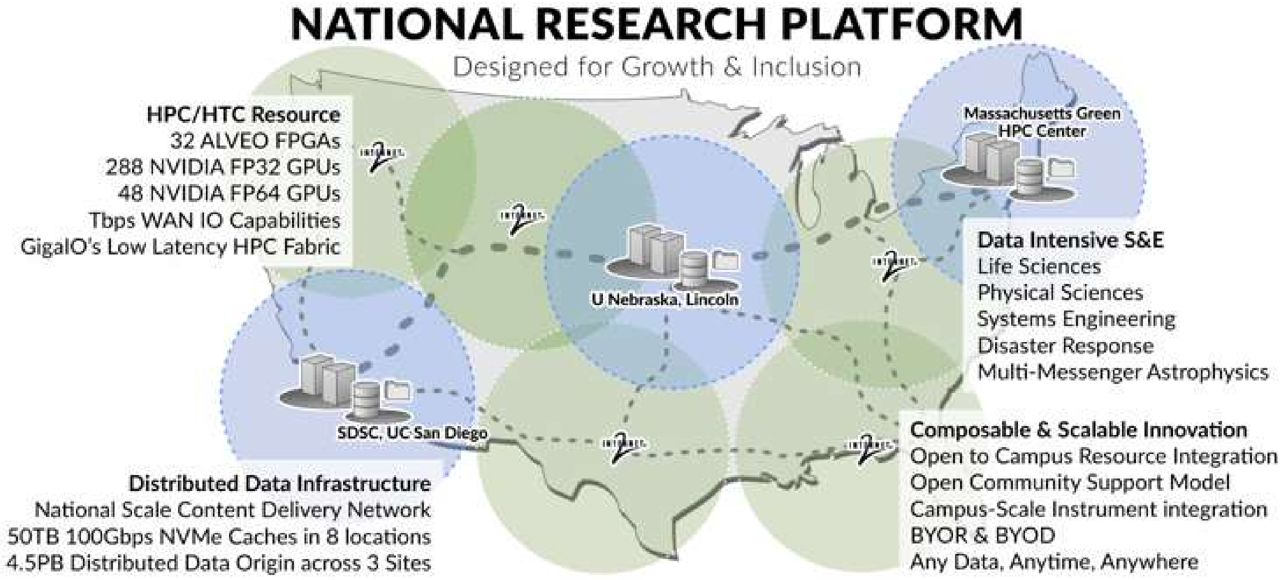
- Low latency is critical
- Separating FPGA’s allows servers to be closer to the patch panel
- Memory Fabric solution provides much lower latency than software-bases solutions
- Planned migration to CXL will reduce latency even more

CUSTOMER STORY: TACC

- Focus is green data center – optimizing resource utilization
- Lonestar 6 add-on, using 3rd Gen AMD EPYC™ processors-powered Dell servers to provide reconfiguration of heterogeneous compute
- Multiple accelerators, including NEC vector engines, FPGAs and various GPUs
- Suite of 20 applications and benchmarks, including CFD and molecular dynamics



CUSTOMER STORY: SDSC AND NSF



NATIONAL RESEARCH PLATFORM

Designed for Growth & Inclusion

HPC/HTC Resource
32 ALVEO FPGAs
288 NVIDIA FP32 GPUs
48 NVIDIA FP64 GPUs
Tbps WAN IO Capabilities
GigaIO's Low Latency HPC Fabric

Data Intensive S&E
Life Sciences
Physical Sciences
Systems Engineering
Disaster Response
Multi-Messenger Astrophysics

Distributed Data Infrastructure
National Scale Content Delivery Network
50TB 100Gbps NVMe Caches in 8 locations
4.5PB Distributed Data Origin across 3 Sites

Composable & Scalable Innovation
Open to Campus Resource Integration
Open Community Support Model
Campus-Scale Instrument integration
BYOR & BYOD
Any Data, Anytime, Anywhere

SDSC SAN DIEGO SUPERCOMPUTER CENTER

- Prototype platform centered on GigaIO's composable capabilities for HPC
- Multiple accelerators, including Xilinx FPGAs and various GPUs
- Being deployed across NSF data centers

NEC Vector Engine Performance with GigaIO FabreX

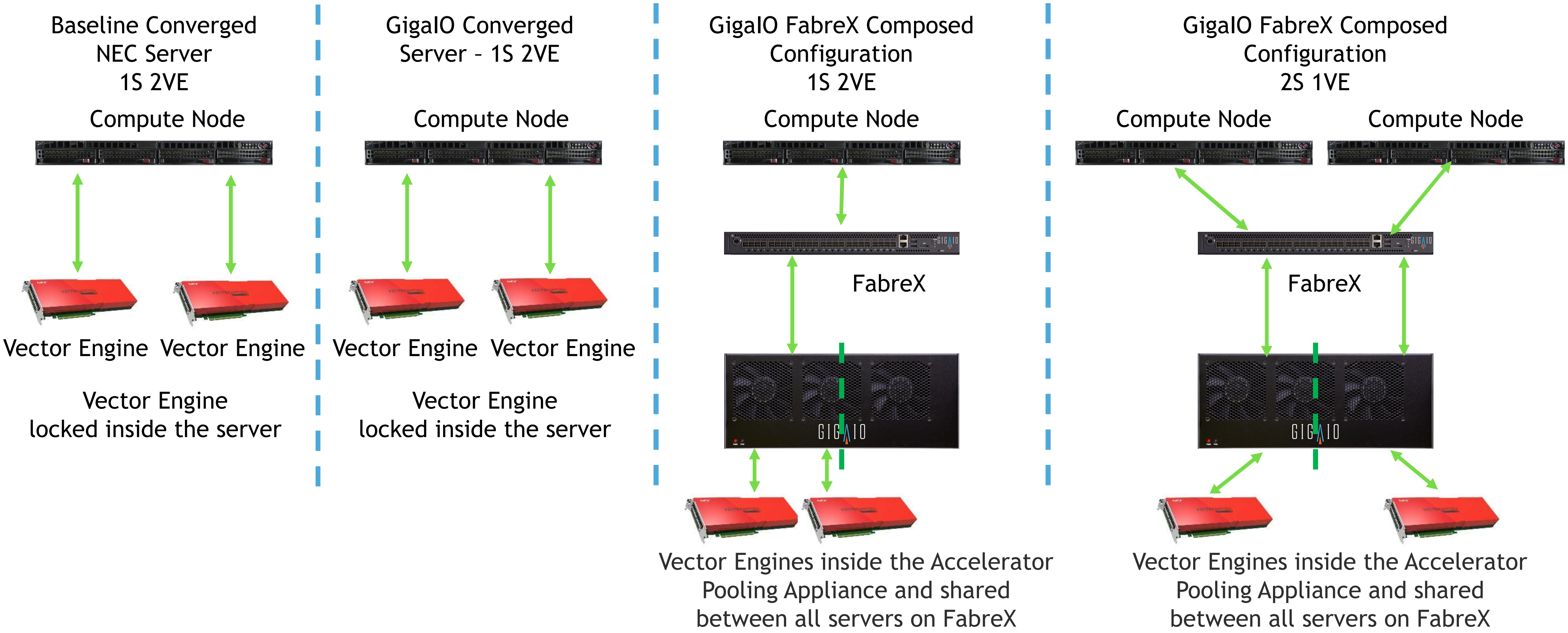
Objectives:

- Execute benchmarks in Converged and Composed configurations
  - Converged – all resource inside the server
  - Composed – all resources connected across memory Fabric
- 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

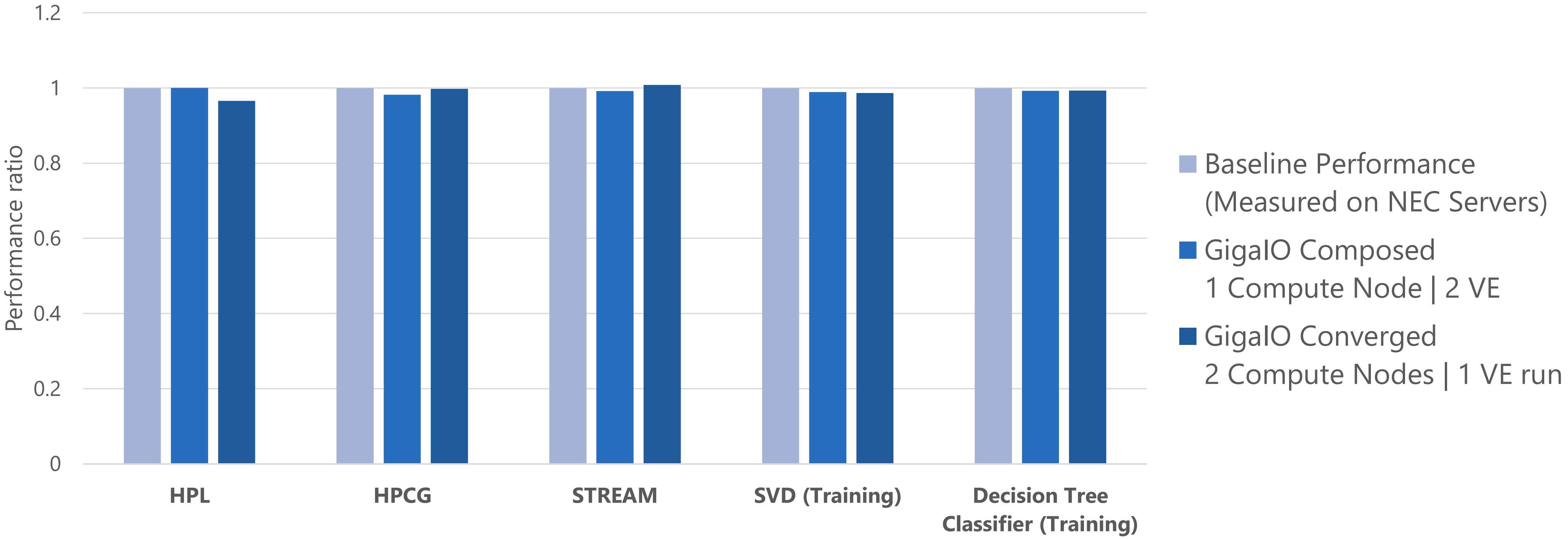
FabreX

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

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.

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 Rack-Scale Computing Platform enables IT’s to improve system performance, incorporate the latest technology, revitalize existing infrastructure, and meet budget and sustainability goals.
- FabreX Memory Fabric architecture with NEC Vector Engines delivers performance
  - Expect to improve performance running multiple VEs across FabreX
- Available today in production

THANK YOU

Questions?

Matt Demas, Field CTO

mdemas@gigaio.com

GIGAIO