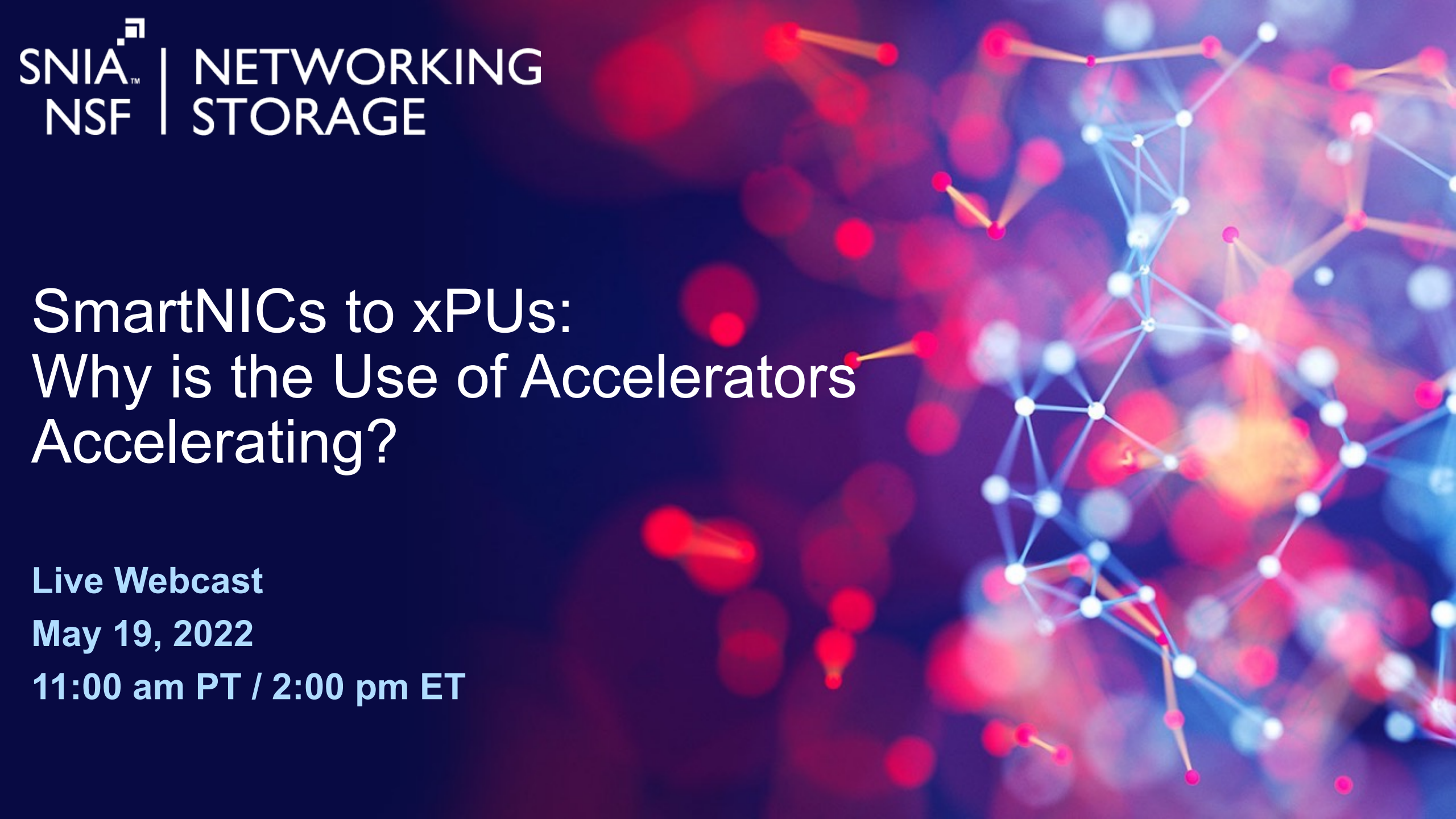
SNIA NSF | NETWORKING STORAGE

SmartNICs to xPUs: Why is the Use of Accelerators Accelerating?



Live Webcast

May 19, 2022

11:00 am PT / 2:00 pm ET

Today's Presenters

Alex McDonald
Independent Consultant

Dr. Joseph L. White
Fellow at Dell Technologies

Jai Menon
Chief Scientist
Fungible

Lior Khermosh
CTO
NeuReality

John Kim
SNIA NSF Chair
NVIDIA

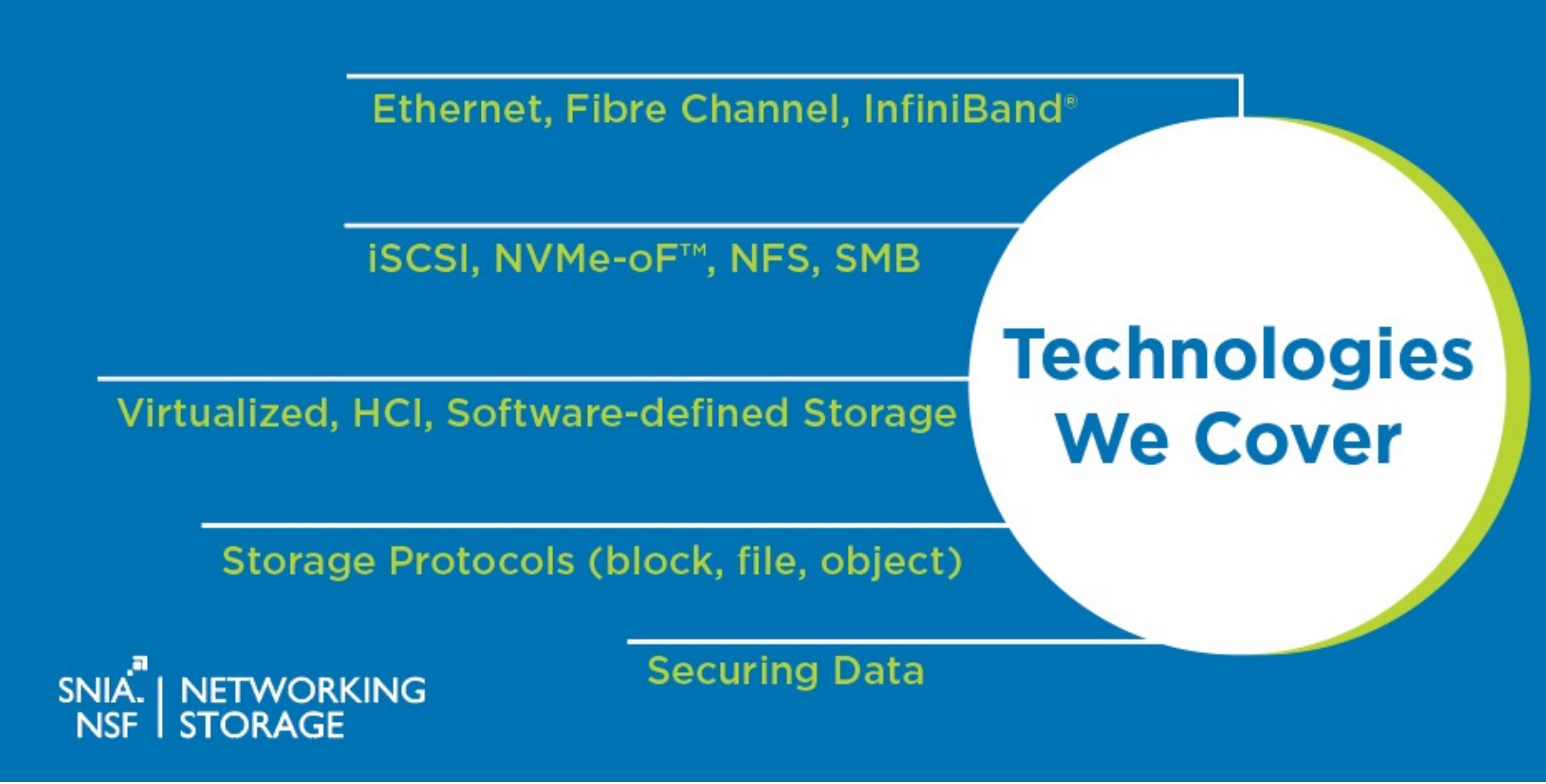
SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**



SNIA Legal Notice

- 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.

Agenda

- What is an xPU?
- Trends and workloads
- Deployment and Solutions
- Market landscape

What is an xPU?

Dr Joseph L White, Fellow, Dell

xPU Terminology: Take xPU where ‘x’ stands for…

SmartNIC	SmartNIC (Intelligent NIC has also been used)
DPU	Data Processing Unit
IPU	Infrastructure Processing Unit
FAC	Function Accelerator Card
NAPU	Network Attached Processing Unit
NPU	Neural Processing Unit
TPU	Tensor Processing Unit
GPU	Graphics Processing Unit
CPU	Central Processing Unit
NPU	Network Processing Unit
APU	Application Processing Unit (GPU + CPU)

These are the xPUs

We have new things to define!

☺

vector, matrix, and tensor processing accelerators

General purpose compute

Networking

These already have good definitions and semantics

Devices with a tightly coupled combinations of CPU, xPU, GPU, etc. will exist. The taxonomy gets interesting at that point…

xPU Definition

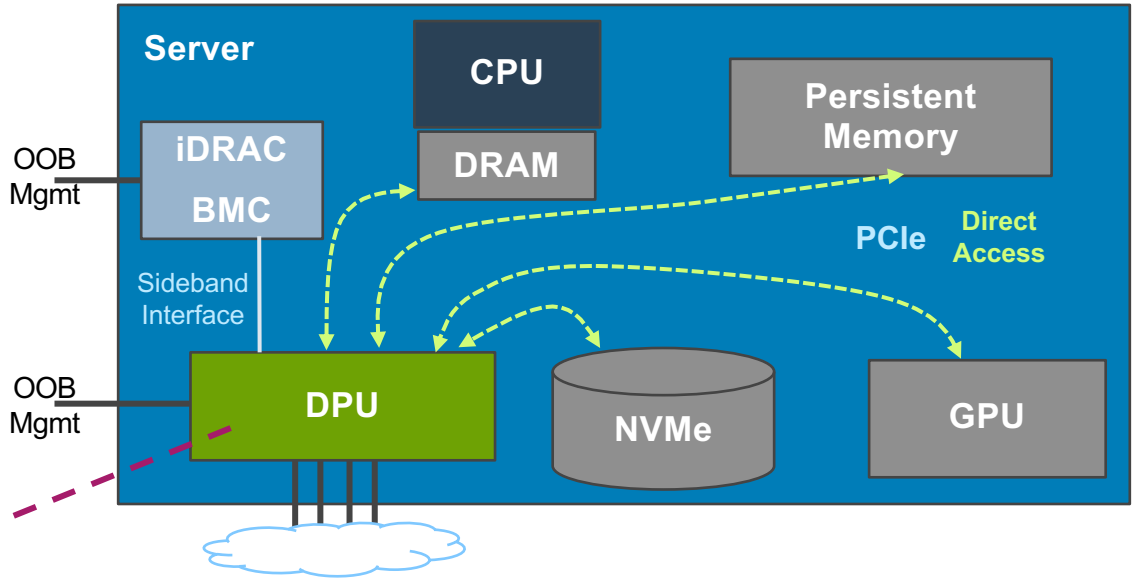
x Processing Unit

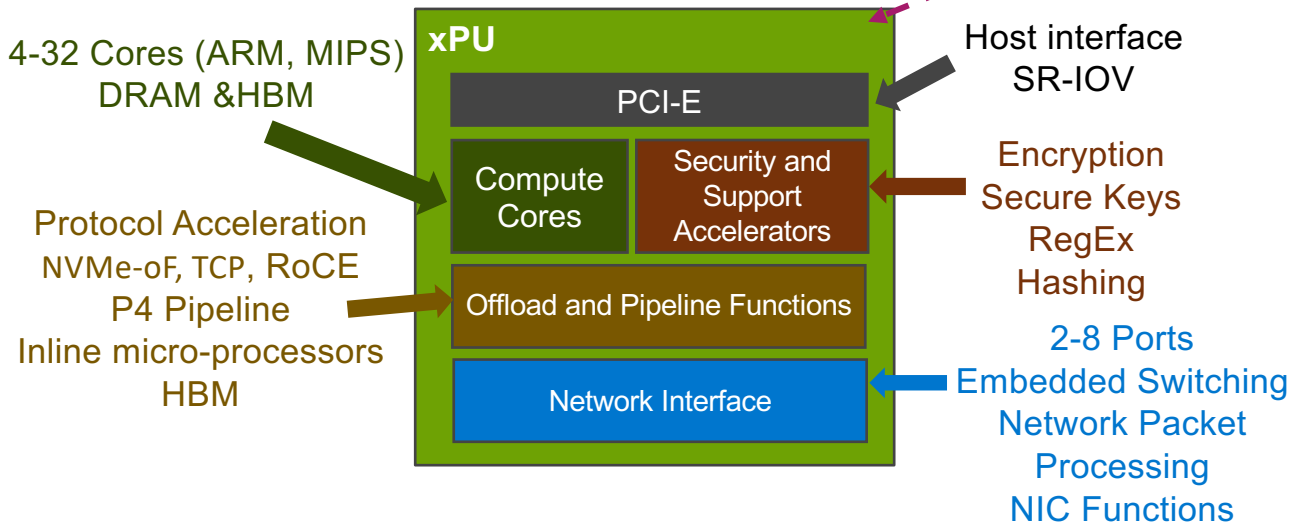
Effectively a micro-server optimized for dataflow and packet processing providing accelerators, offload engines, & local services

xPUs are (typically) a component in a host system but can be completely independent (e.g Edge, JBOF, Gateway)

Presents virtual functions to a host (looks like a NIC, GPU, etc)

Presents network interfaces





Representative Functions and Use Cases

NVMe-oF & Storage Offload

Optimize Solution Data-Flows (eg: access to GPUs & GPU pooling)

Control Plane Offload (saves x86 Cores for business relevant work)

Enhanced Security (Trust, Key Mgt, Encryption, Firewall, DPI)

Improved Telemetry Generation and Analytics Capabilities

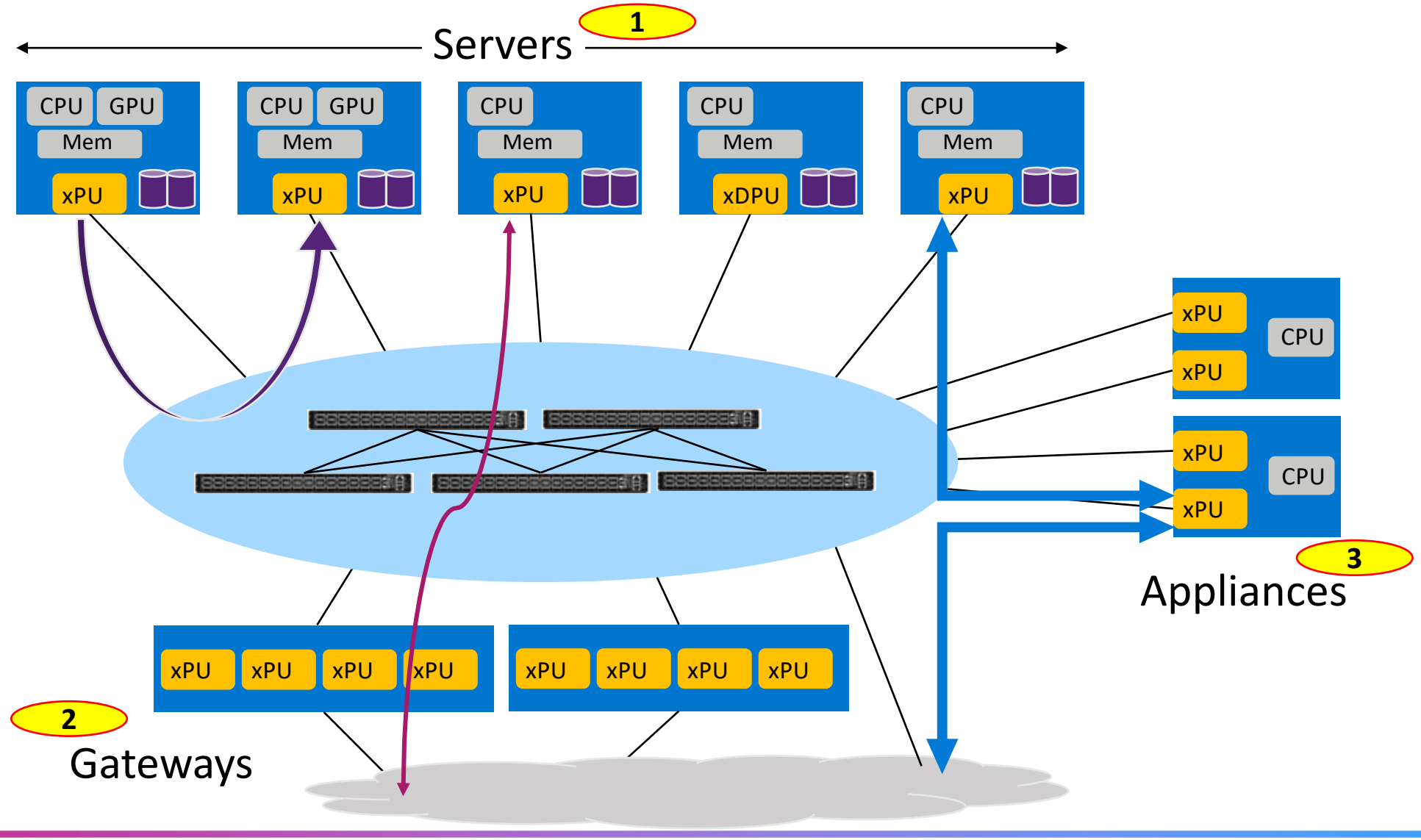
Application Specific Computation Offload (RegEx, EBPF, containers)

Physical and Virtual Networking and packet forwarding

Network Gateway & SD-WAN

Computational storage offloads

xPU Deployment



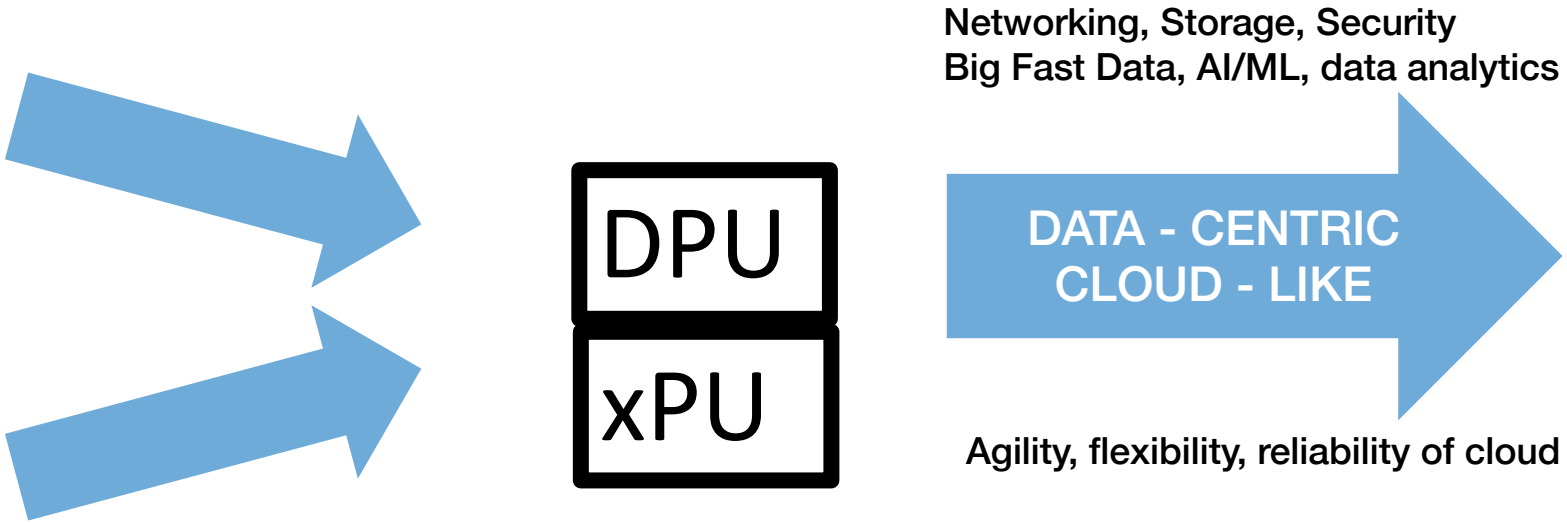
Trends & Workloads

Jai Menon

Chief Scientist, Fungible

xPUs/DPUs Address Two Important Data Center Trends

#1) Growth of Data-Centric Tasks Stateful processing of multiple high b/w data streams for networking, storage, security, AI/ML
(DPUs are optimized for data-centric tasks)



#2) Cloudification of On-prem and Public Data Centers
(xPUs/DPUs enable efficient disaggregation to improve TCO, power, footprint, agility, security of data centers
(DPUs optimize data flow processing)

Examples of xPU-Based Products

xPU enhances servers (2 examples)

1A

PCIe Host Cards
(offload storage, network, security)

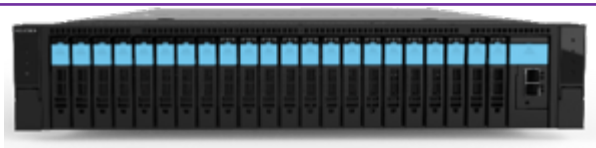
Save cores & improve perf. (see next slide)

1B

Disaggregated PCIe/CXL device pools (e.g. GPUs)

Enhanced disaggregation for cloudification

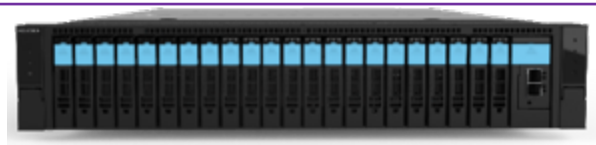
Standalone xPU products (3 examples)



2A

High performance storage target

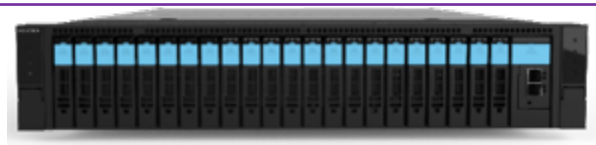
10X speedup for storage targets



2B

Computational storage

20X speedup for some TPC-H queries



2C

SDN appliance

Overlay network processing @3X scale & 5X CPS

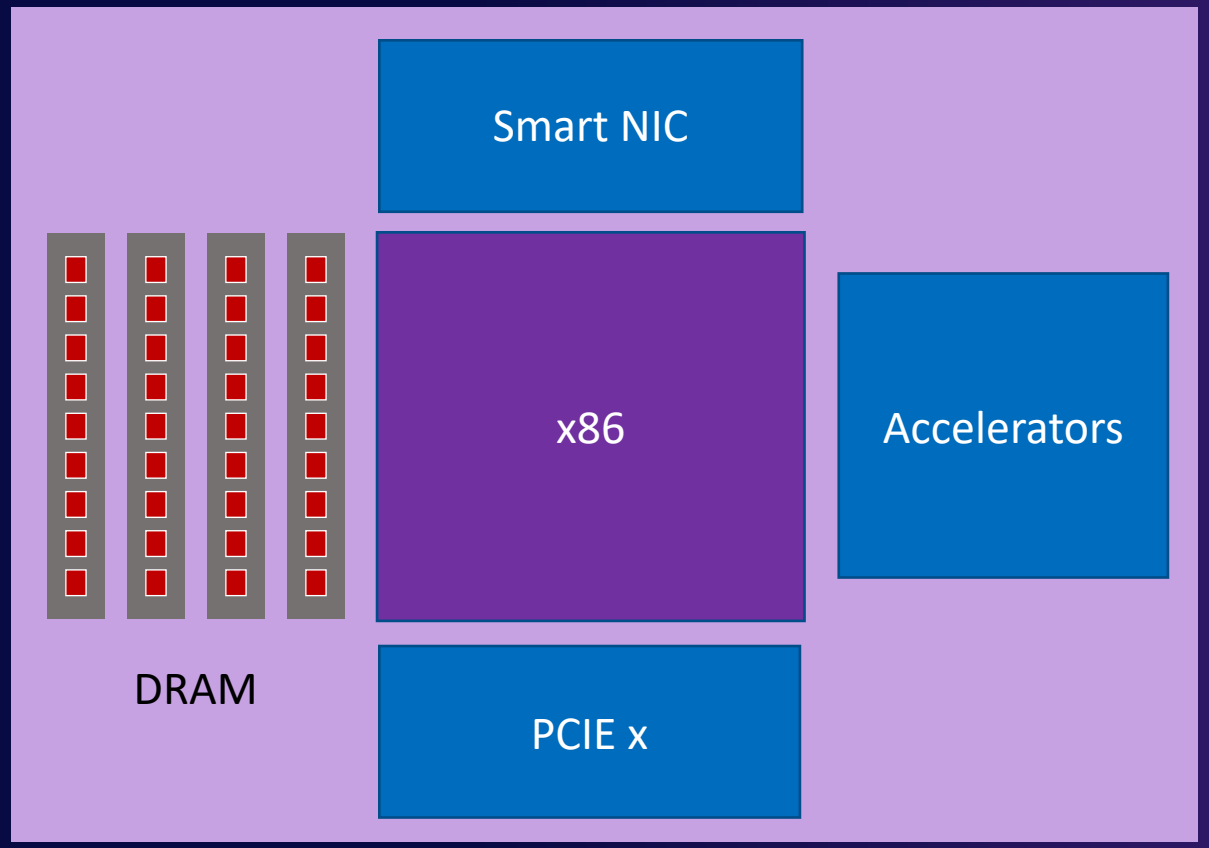
Example Capabilities of xPU-Based Host cards (1A)

Accelerate performance and offload data-centric work from servers

Domain	Sample Features	Value
Network	TCP, Network overlays, switching/routing (vRouter – e.g. save 7 cores, improve latency by 10X, throughput by 2X)	Offload and accelerate
Security	IPsec, SSL/TLS, kTLS, stateful firewall (NGFW – e.g. save 12 cores, improve perf by 6X)	Offload and accelerate
Storage	High speed NVMe/TCP offload DPUs are optimized for storage (Storage Initiator - e.g - save 96 cores, improve perf by 2X)	Offload and accelerate
Virtual PCIe Switch	GPU disaggregation	Efficient disaggregation

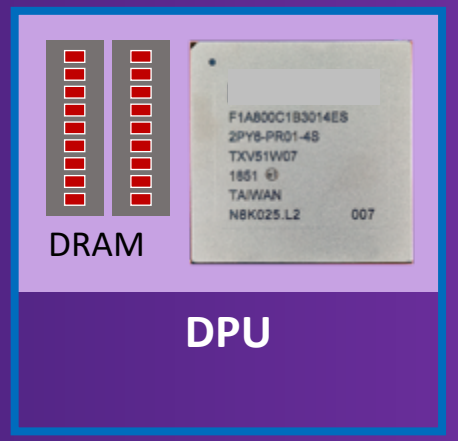
DPU-ENABLED HIGH PERFORMACE STORAGE TARGET (2A)

SINGLE CHIP INTEGRATION IS THE FUTURE OF STORAGE PROCESSING



Typical CPU-Based Discrete Implementation

Evolution



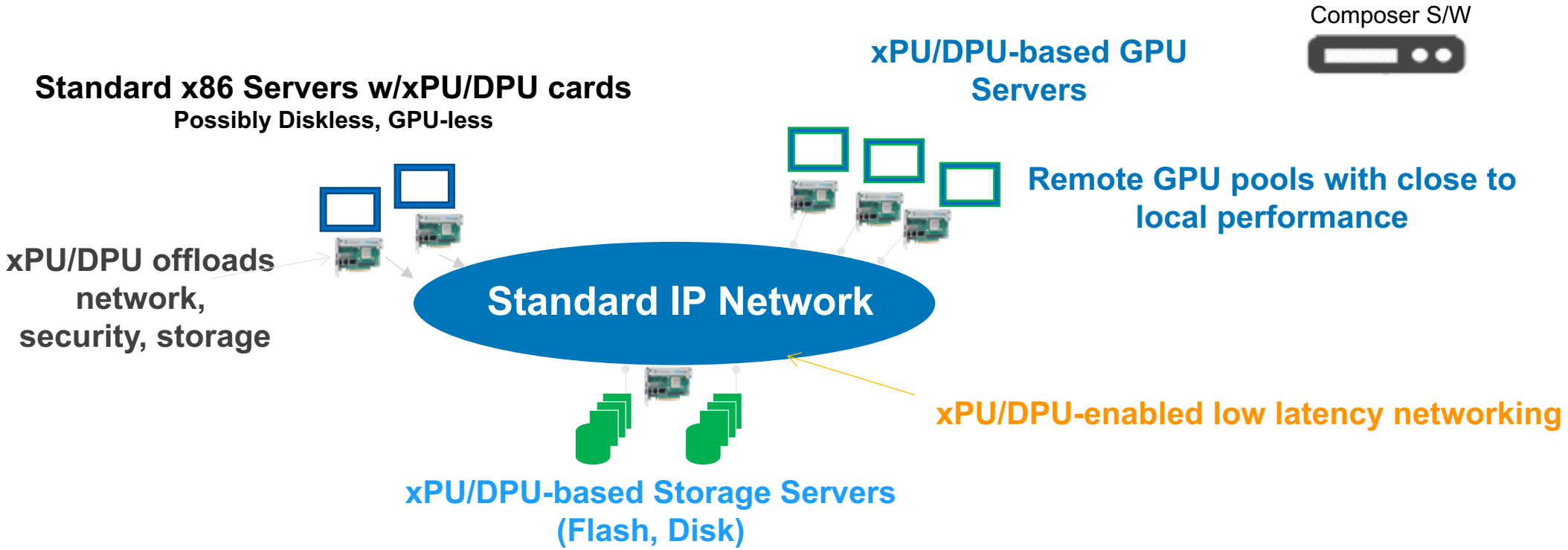
Example Implementation

Attribute	x86 Based Discrete System	DPU Based Integrated System	Improvement Factor
Performance/W	9.2K IOPS/W	90.2K IOPS/W	**10x**
Performance	5M IOPS	13M IOPS	**2.6x**
Power	540 W	144 W	**3.8x**

High Performance, Low Power, Full Featured

xPU/DPU-Based Next Gen Data Center

Emerging approach to cloudification using composable disaggregated infrastructure (CDI)



Remote storage pools with close to local performance

3 key elements of the solution – xPU/DPU, low-overhead disaggregation, and composability

Summary – Data centers will increasingly use CPUs, GPUs **and** xPUs

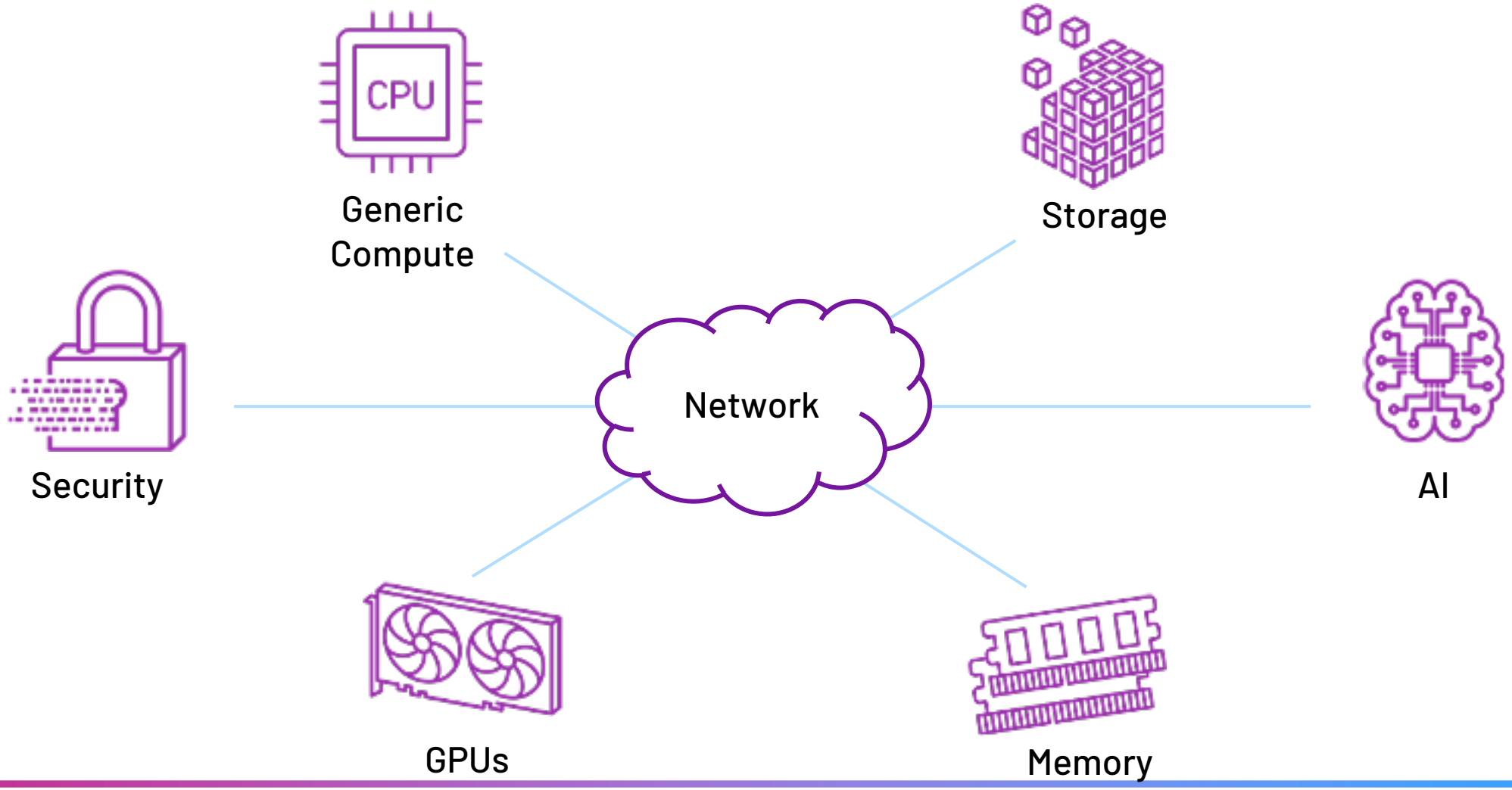
- Data Centers will use
  - xPUs/DPUs
    - To efficiently handle data-centric tasks (reduce power, TCO, footprint)
  - Efficient disaggregation of expensive resources (reduce TCO, power, footprint)
    - xPU/DPU-based storage
    - xPU/DPU-based GPU/TPU pools
    - xPU/DPU-based host cards
  - Composability software (agility, reduce # of SKUs)
    - To eliminate or reduce infrastructure silos

Deployment and Solutions

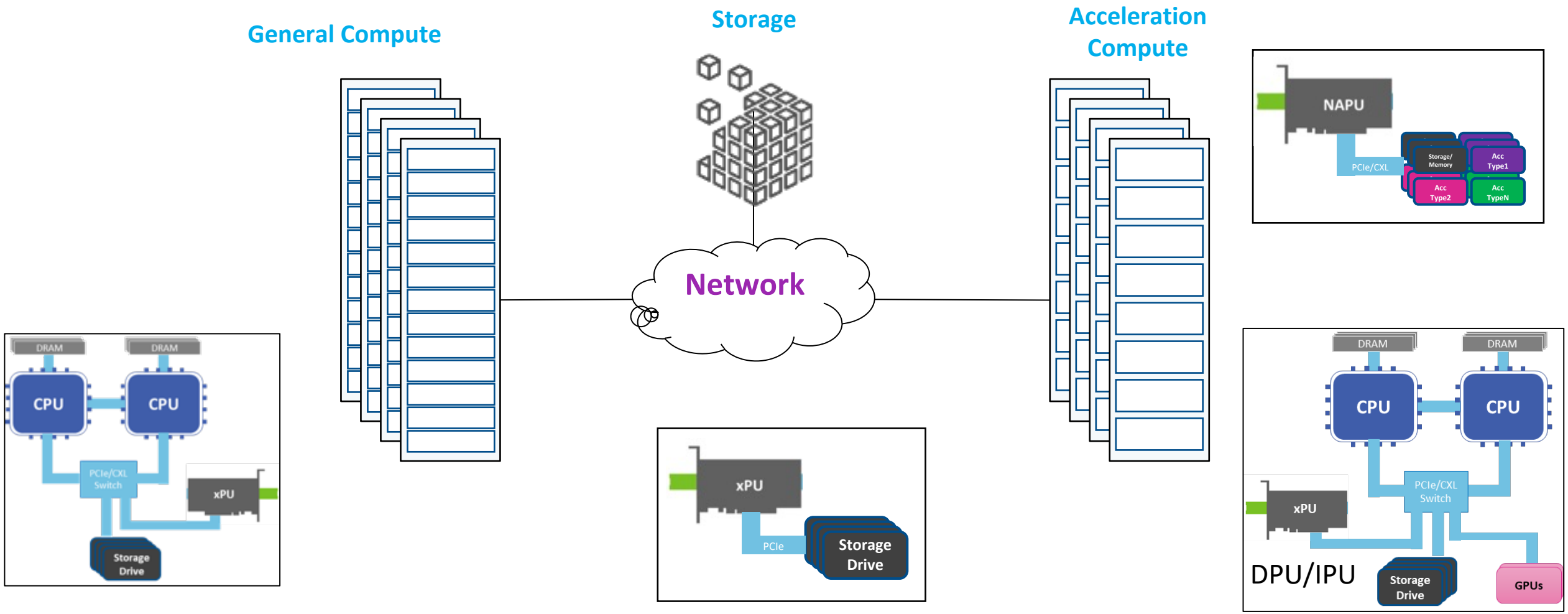
Lior Khermosh

NeuReality

Composable Infrastructure Deployment



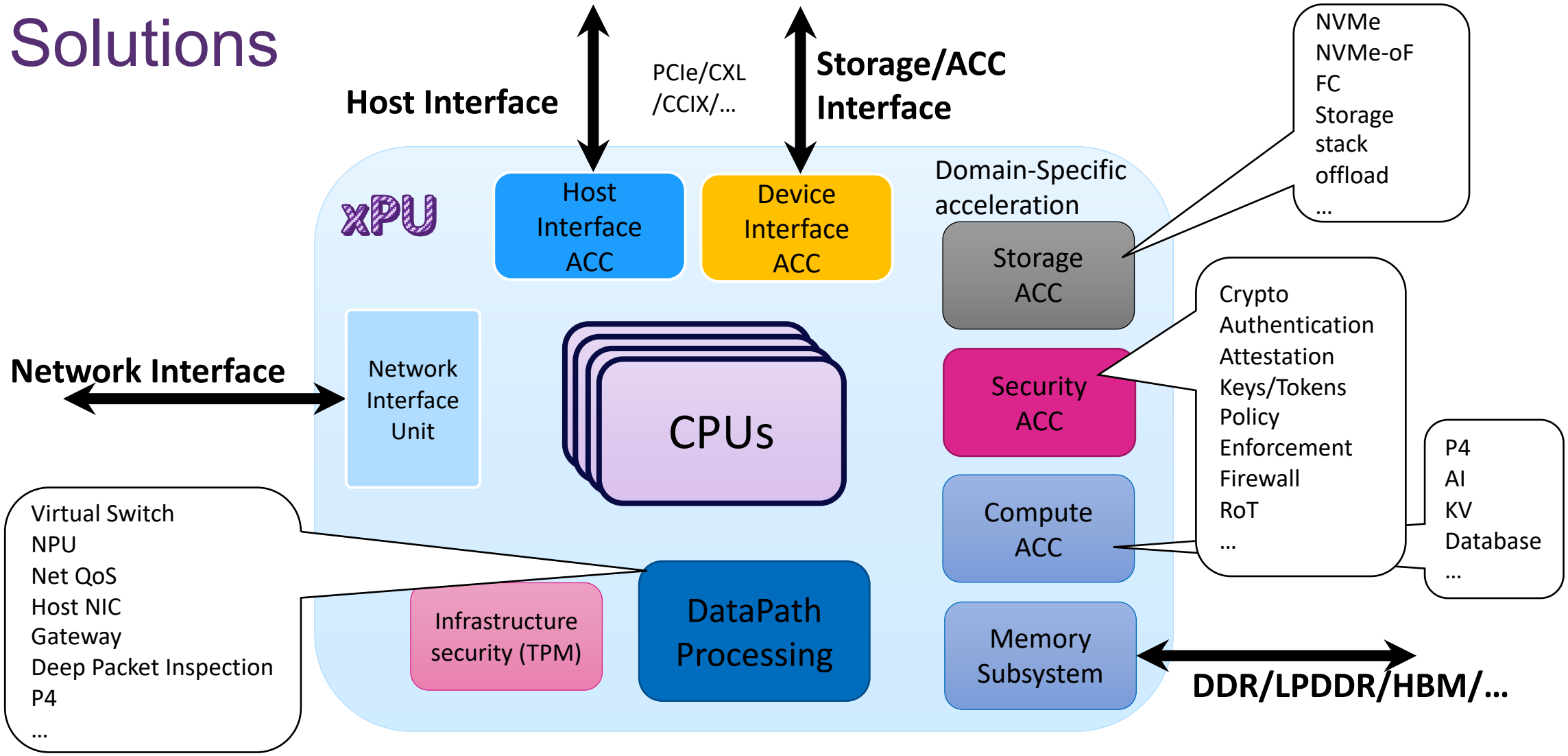
CI Deployment <-> Disaggregation Options



What is an NAPU?

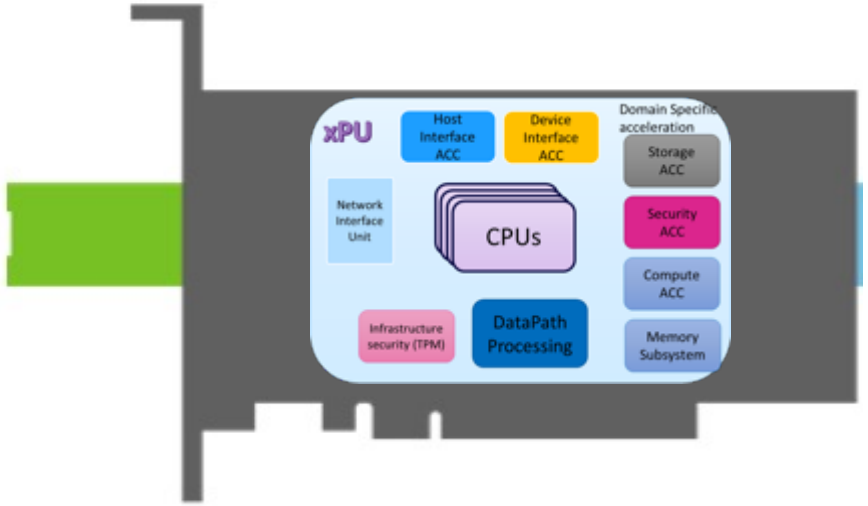
- An NAPU is a workflow-optimized hardware device that has the following attributes:
  - One or more specialized processing units
  - Native network capabilities
  - Streamlined hardware-based datapath
  - Virtualization capabilities
  - Managed resource abstraction layer
  - Self-sufficiency

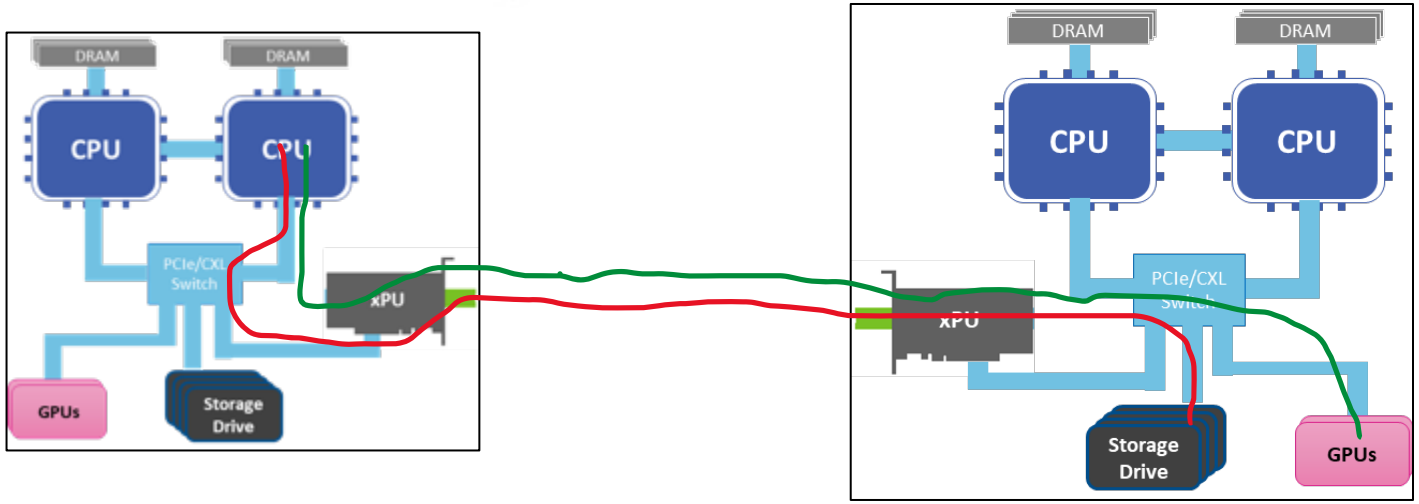
xPU Solutions



- Different flavors of xPUs vary in their specific block diagram

xPU Functions





- Service virtualization
- Performance
- Host Offloading
- Storage Frontend
- BM/VM/Container management
- Server Root of Trust
- P4 Programming
- Compute Offload
- Security Offload

Datapath Processor
- Virtual Switch
- NPU
- Network QoS
- Host NIC
- Gateway

xPUs are an extremely effective way to **Scale** and handle **Big Data**

xPU Advantages

- Host Offloading
  - Compute cores dedicated to Application, not for infrastructure and datapath movement
  - In many cases the CPU is stalled, waiting for data from memory or IO
- High Performance
  - Removing bottlenecks
  - Accelerating key complex components
  - QoS, Network processing, Firewall
  - Programmable @ Performance, ex: P4
- xPUs are the Joints of Composable Infrastructure
  - Using the right accelerator at the right place, in exactly the right portion, for the right part of the workload
  - Multiple types – Storage, Accelerators
  - Native virtualization and datacenter management
- Isolated embedded controlling entity
  - Excellent for Server Management and Security
  - Bare Metal/VMs management, RoT

Market Landscape

John Kim

NVIDIA

xPUs Sales to Grow Rapidly but not Dominate NICs

- By 2023, one in three NICs sold will be a FAC[1] – Gartner®
  - By 2026, SmartNIC revenue will reach $1.6 Billion – Dell'Oro[2]
  - By 2026, more than half the servers shipped to hyperscaler CSPs will be equipped with SmartNICs – Dell'Oro[2]
- Not all accelerator cards included in FAC/SmartNIC forecasts
  - i.e. CPU, TPU, GPU, APU or crypto card not counted as SmartNICs
- Managed appliances will drive additional xPU sales – not yet considered by most analysts

[1]Source: Gartner "Your Server is Eating Your Network –Time to Rethink Data Center Network Architectures," Gartner, G00742770 -- April 2021
[2]Source: Dell'Oro Group "Ethernet Adapter & Smart NIC 5-Year Forecast Report" – February 2022

xPU Vendors and Routes to Market

Who Makes xPUs

- Startups
  - Will they be absorbed?
- Big Vendors
  - Silicon vendors (AMD/Xilinx, Intel, Marvel, NVIDIA, etc.)
  - Networking vendors
- CSPs – design/codesign

Who Sells xPUs?

- xPU vendors
- Infrastructure OEMs and ODMs
- VARs, SIs
- Cloud SPs – indirectly

How Do We Program xPUs?

- Standard open interfaces to specific functions
  - E.g. DPDK, SPDK, OVS/OVN, NVMe-oF, P4
- Vendor-specific development platforms and SDKs
  - Some xPU functions work across most xPUs; others are vendor/card-specific
  - High-level and low-level programming interfaces
- Development of open platforms/SDKs
  - OPI Project and IPDK
  - DASH
- Programmers: vendor, end-user, ISV, infrastructure OEM, SI, or CSP

Tradeoffs and Considerations for Deploying xPUs

Yes, xPU in Every Server!

- Offloads infrastructure tasks
- Frees server CPU for apps
- Higher efficiency, lower TCO, flexible disaggregation
- Can improve security

No, Only in Some Servers

- Increases server CapEx
- Offloads specific workloads
- Higher power consumption
- Needs to be programmed; no one standard interface

Summary

- We've covered…
  - What is an xPU?
  - Trends and workloads
  - Deployment and Solutions
  - Market landscape
- This is a Series!
  - Join us for our next webcast:

xPU Accelerator Offload Functions

June 29, 2022

11:00 am PT/ 2:00 pm ET

https://bit.ly/xPUOffloads

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast, including answers to questions we couldn't get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on Twitter @SNIANSF

Thank You

