SNIA NSF | NETWORKING STORAGE

xPU Deployment and Solutions Deep Dive

Live Webcast

August 24, 2022

10:00 am PT / 1:00 pm ET

Today's Presenters

John Kim
SNIA NSF Chair
NVIDIA

Tim Michels
Distinguished Engineer
F5

Mario Baldi
Fellow
AMD Pensando Systems

Amit Radzi
Software Architect
NeuReality

SNIA - By the Numbers

Industry Leading Organizations

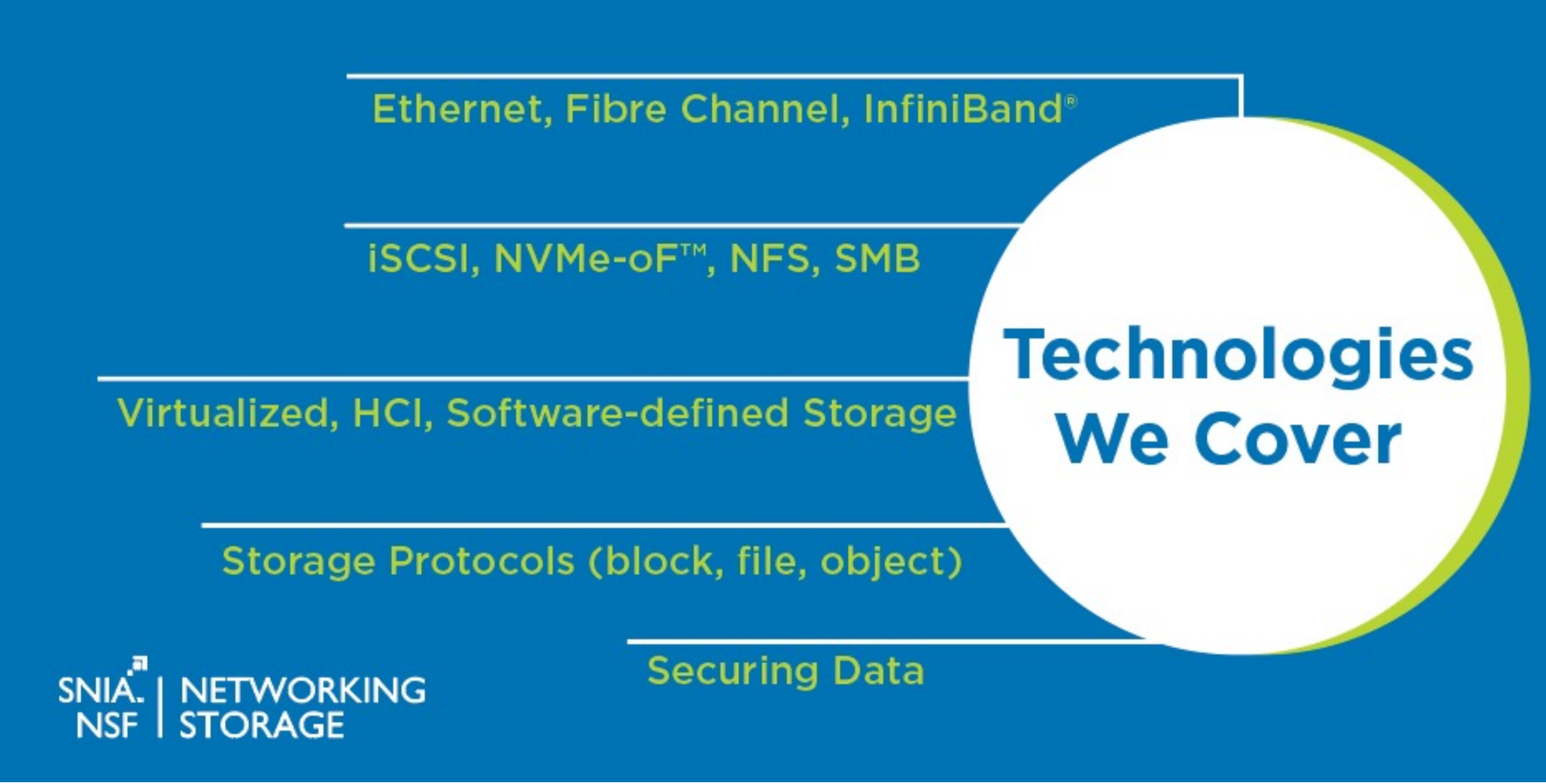
180

Active Contributing Members

2,500

IT End Users & Storage Pros Worldwide

50,000



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.

This is a 3-Part Series!

- 1st Webcast: “SmartNICs to xPUs: Why is the Use of Accelerators Accelerating?”
  - Watch on demand at: https://bit.ly/SNIAxPU1
- 2nd Webcast: “xPU Accelerator Offload Functions
  - Watch on demand at: https://bit.ly/SNIAxPU2

SNIA Networking Storage Forum presents

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

SNIA Networking Storage Forum presents

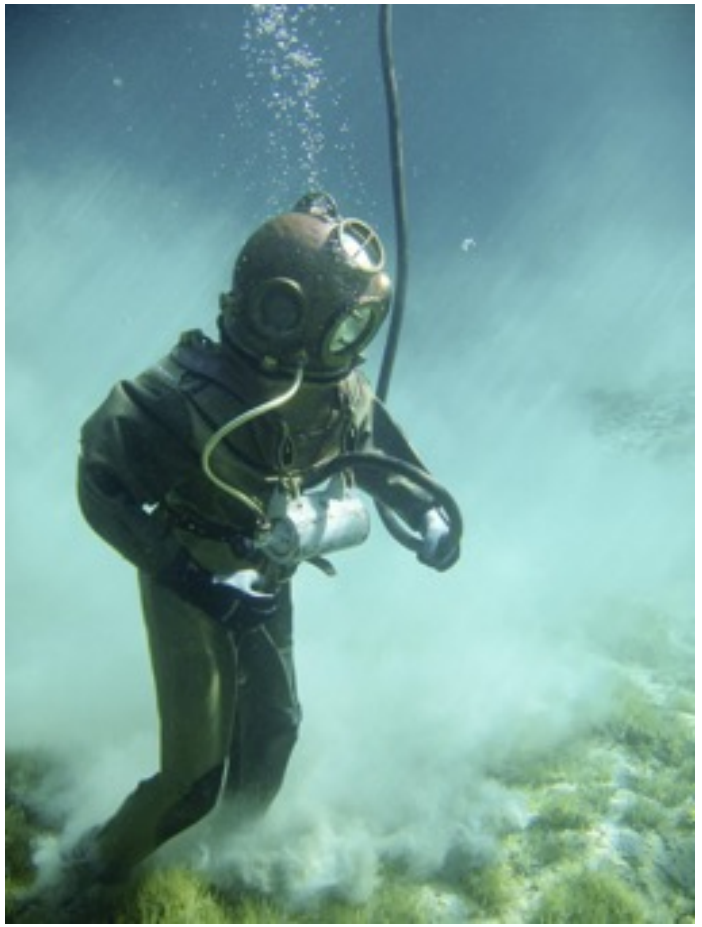
xPU Accelerator Offload Functions

June 29, 2022
11 AM PT

Register Here

Agenda

- xPU Deployment and Solutions Deep Dive
  - When to Deploy
  - Where to Deploy
  - How to Deploy



When to Deploy

Tim Michels

F5

xPU Value Depends on How it is Used

Incremental Use Cases

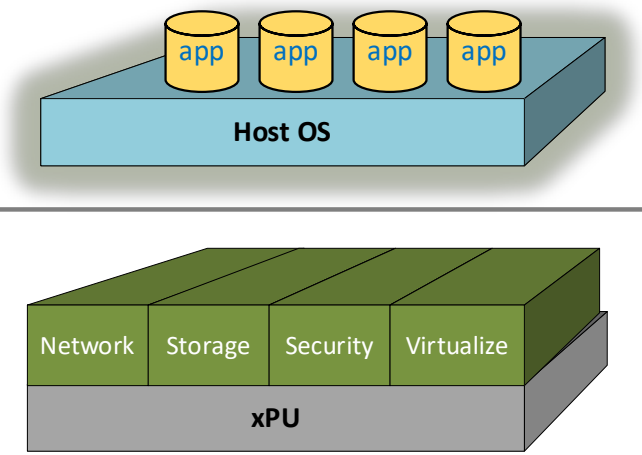
- In-band, ***tightly coupled***, HW offloads
- xPU compute used only for ***exception processing*** or control plane
- Provides only a ***closed*** SW stack

Revolutionary Use Cases

- Hosting *platform* for ***stand alone*** infrastructure services
- ***Abstraction layer*** for infrastructure services
- Hosting of ***multi-application*** services
- Hosting ***3rd party*** services

Separate Business Apps from Infrastructure Services

- **Business Apps** run on the Node
- **Infrastructure Apps** are Services running on the xPU
  - ❑ Network
  - ❑ Storage
  - ❑ Security
  - ❑ Virtualization



Business Applications

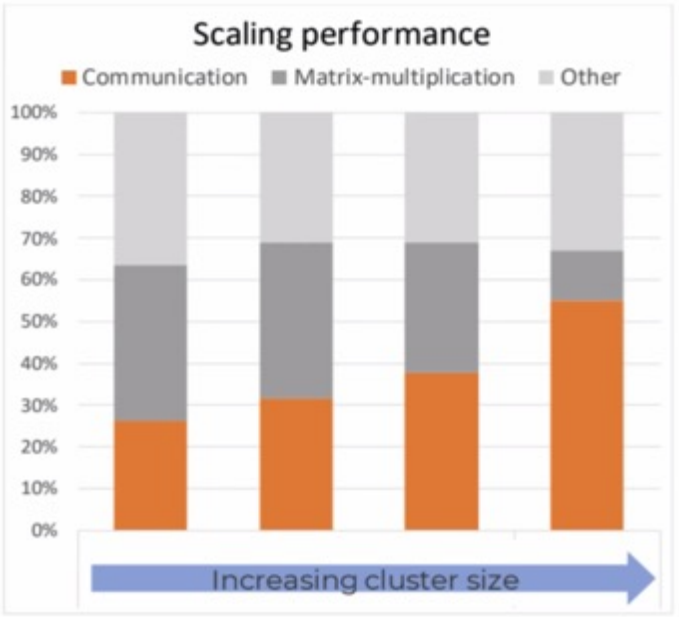
Infrastructure Services

- Why move Infrastructure off the node?

  - "30% of CPU cores are being used for datacenter infrastructure needs."

  -"It would take 125 cores to run all the Security, Network, and Storage offloads at 125Gbps"

  Jensen Huang, NVIDIA CEO, @ 2020 GTC Keynote



Meta

Deploy xPU Technology to Get …..

Application Simplification

- Standard APIs for xPU Services
- Abstract away from Infrastructure Complexity
- Confine HW Complexity to the xPU services

Performance Isolation

- Solves for the "noisy neighbor" problem
- Dedicate Server Cores to Application Monetization
- Infrastructure services enabled for SLA delivery

Security Isolation

- Confined Blast Radius
- Air-Gaped the Hypervisor
- Defensible Barrier for Application Security Metadata

Organizational Isolation

- Separation of Concerns
- Management Aligned with Organization
- Easier Path to Compliance

Service Isolation – "The Infrastructure Layer"

Trade Offs and Cautions

Be Careful

- X Cap Ex costs higher
- X Per server power higher
- X Server selection impacts
- X Deployment complexity
- X Application re-factoring

The Payoff

- Normalized Application APIs
- Enhanced Portability
- Composable Infrastructure
- Dramatically Increased Scalability
- Enhanced Manageability
- Lower Operational Complexity
- It all adds up to Lower TCO

Where to Deploy

Mario Baldi

AMD Pensando

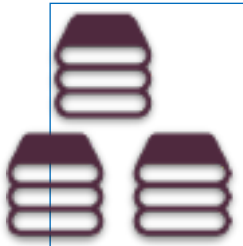
Deployment Opportunities

Enterprise data center

- Security (microsegmentation)
- Storage virtualization
- (Partial) hypervisor offload

Cloud Data Center

- Network virtualization (overlay)
- Security
- Storage virtualization



Colocation/Bare Metal as a Service

- Network virtualization (overlay)
- Storage virtualization

Service Provider Edge (5G) and Far Edge

- Virtual Network Function (VNF) offload

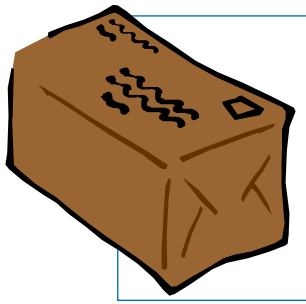
Storage provider

- Storage virtualization target offload

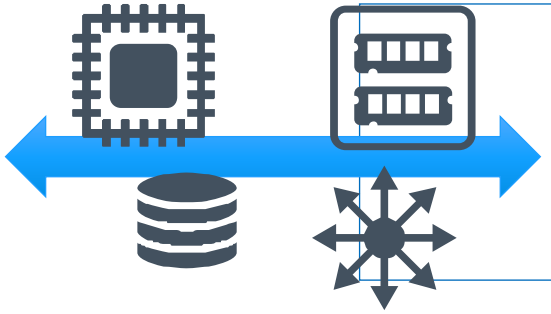
Where Exactly?

At the network edge

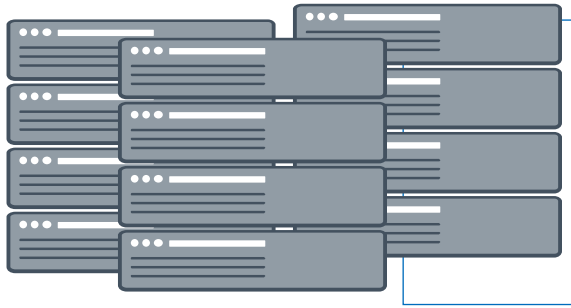
Point of contact between host and network



Access to traffic



Possibly access PCIe



Leverage scale out model

Form Factors

NIC

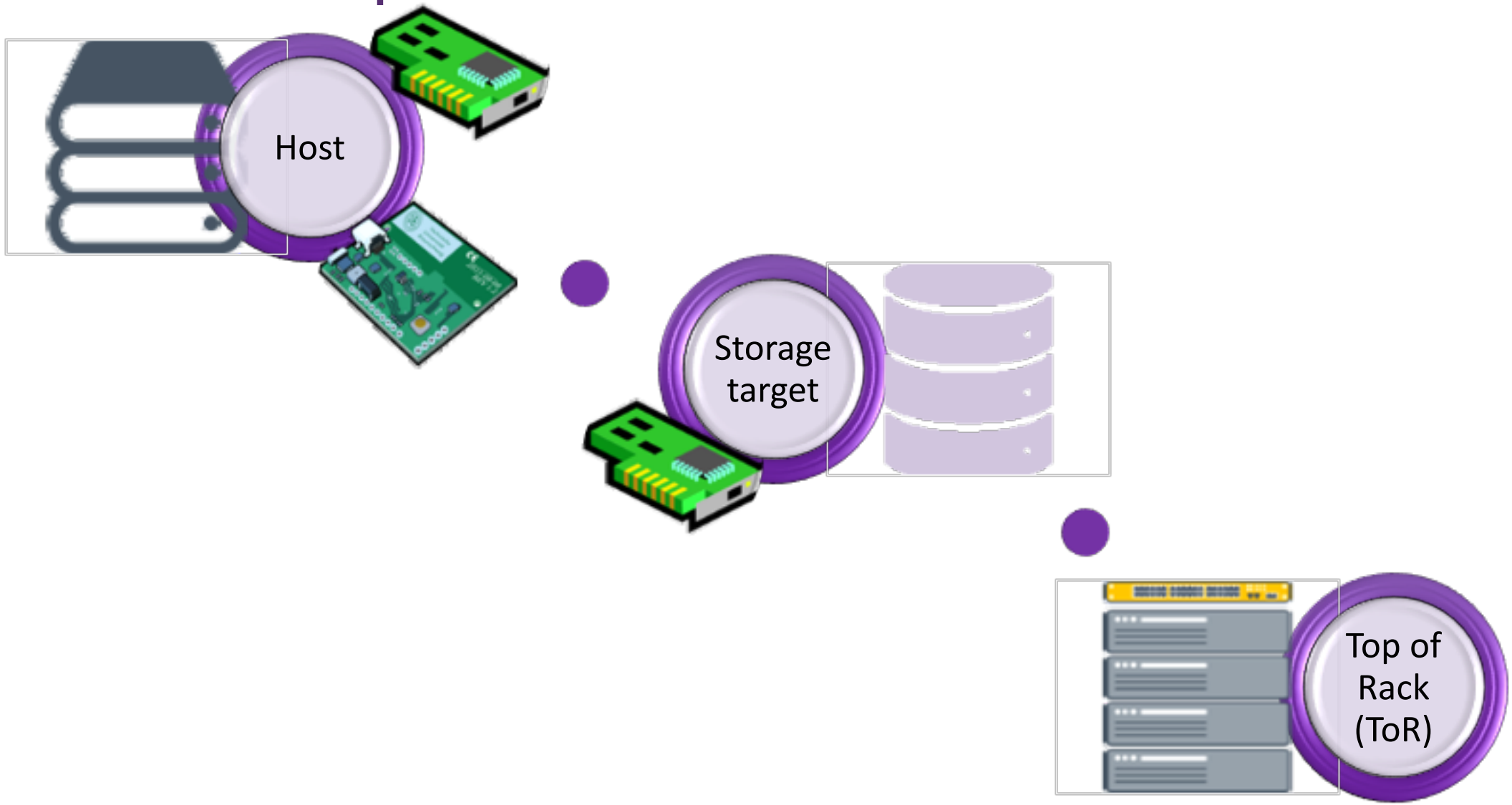
SmartSwitch

ExpEther for PCIe-based services

Motherboard

Chipset

Placement Options

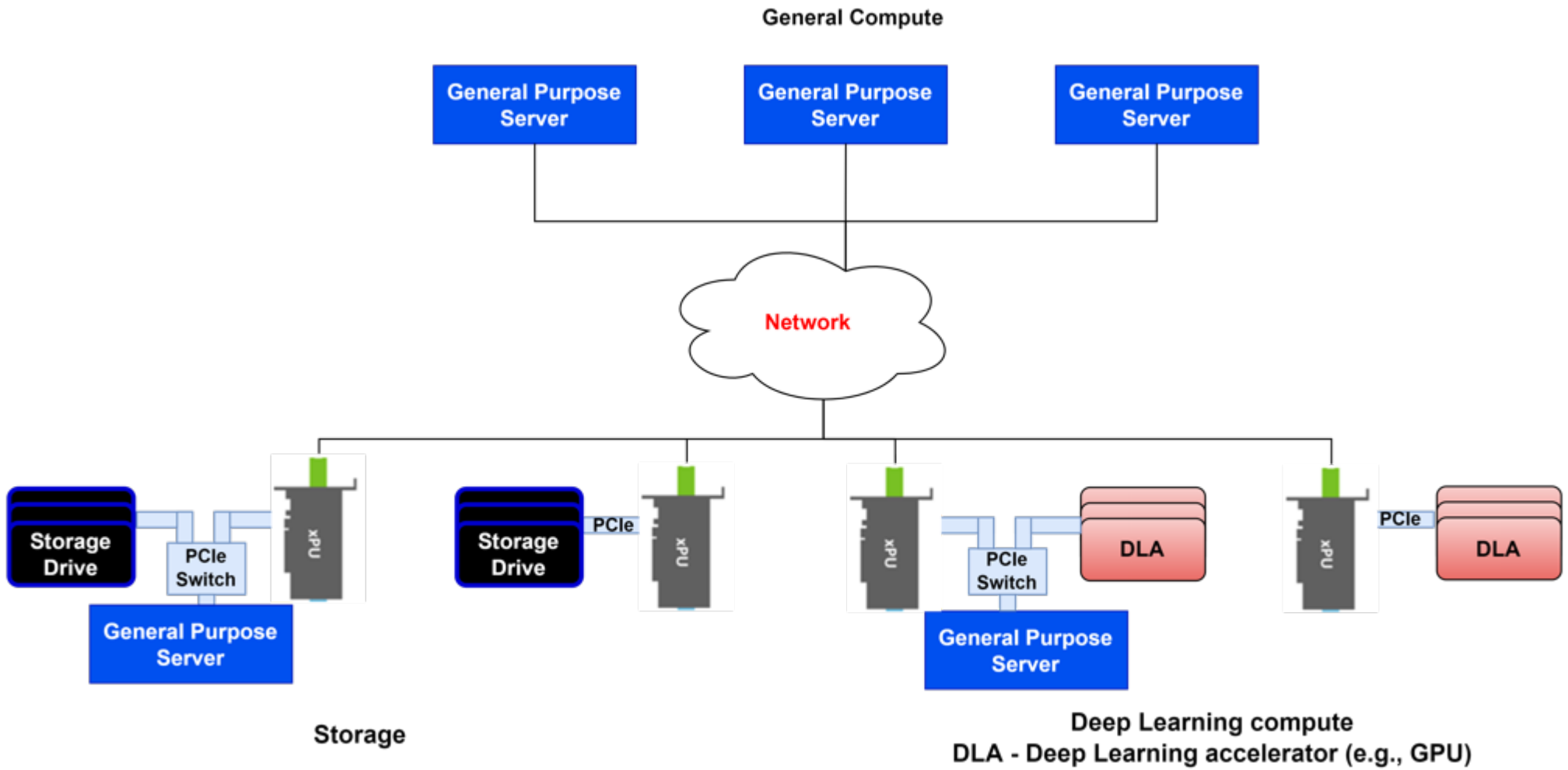


How to Deploy

Amit Radzi

NeuReality

Disaggregated Infrastructure



Workflow Mapping – Local to Disaggregated

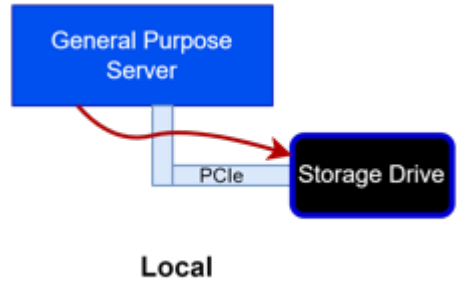
- Map workflow to disaggregated infrastructure
  - Choose workflow that can utilize from a composable data center by consuming disaggregated resources
  - Create a well-defined separation line between an initiator consuming the resources and target exposing the resources
    - Clear easy to use API (preferably an existing one)
    - Allow for virtualization and multi tenancy
    - Elastic – Can be added and removed dynamically

Workflow Mapping – Accelerating to xPU

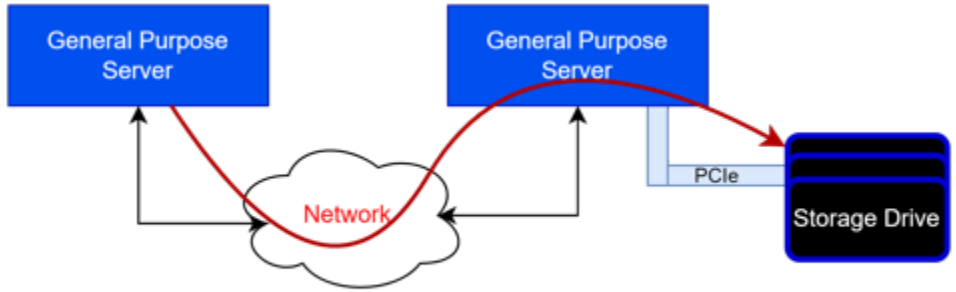
- Utilize xPU in workflow
  - xPUs inherently offloads data preparation and movement for the workflow in the disaggregated data infrastructure
- Accelerate workflow operations by offloading functions to xPU/s
  - Identify data centric tasks that natively can run efficiently inside xPUs
  - Identify bottlenecks in flow that can run in the xPU
  - Identify operations that consume host resources and can be offloaded by the xPU

Use Case Example: Storage Drives – Local to Disaggregated

- Use case
  - Clients on GP servers consume local NVMe storage drives
- Disaggregated flow
  - Client on GP server consumes storage using NVMe-oF driver
  - Remote storage drives are exposed as a virtual drive
    - Can include multiple and partial SSD drives
  - NVMe-oF target implemented on remote side
    - Implements NVMe-oF protocol
    - Implements mapping virtual drive to physical drives
    - Supports multiple tenants



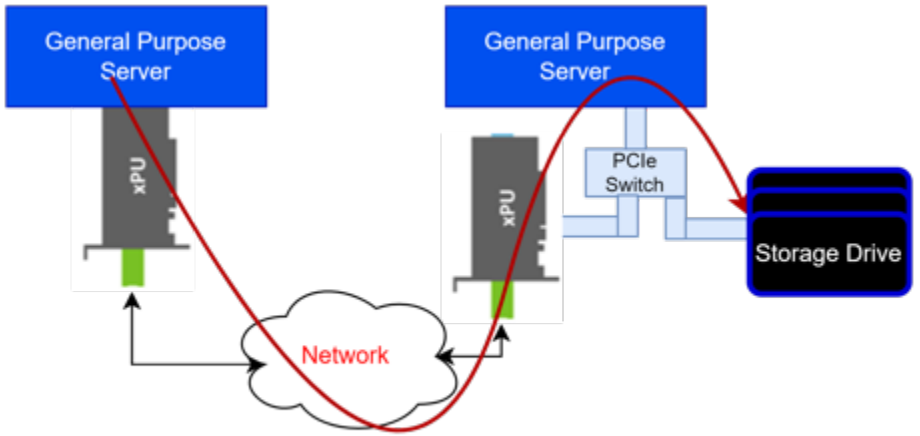
Local



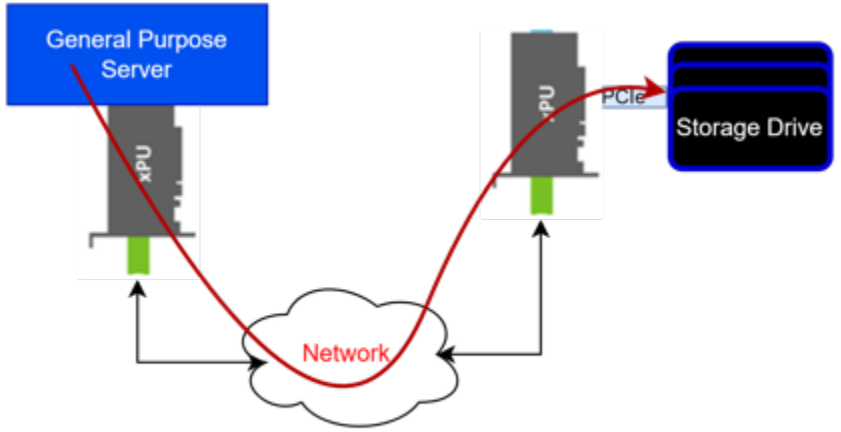
Disaggregated

Use Case Example: Storage Drives – Accelerating to xPU

- Integrate xPU into workflow
  - Use xPUs in both client and server network interface
  - Access from xPU in server side directly to storage drives (e.g. PCIe peer to peer)
- Offload capabilities to xPU
  - Offload on client side
    - NVMe-oF protocol offload
    - Virtualization and multi tenancy
  - Offload on target side
    - NVMe-oF protocol offload and termination
    - Storage offloads
      - Compression, encryption, etc.
      - Elastic and composable
    - Exposed memory as any storage drive topology (unrelated to actual HW)
      - Expose multiple NVMe-oF targets
- Full offload to xPU (target side)
  - Connecting xPU directly to SSDs without any general-purpose server in target side



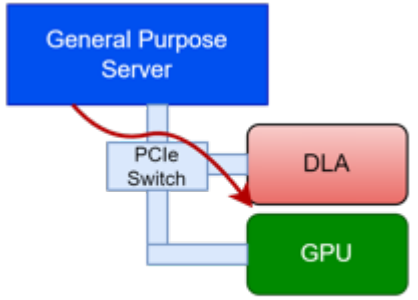
xPU accelerated



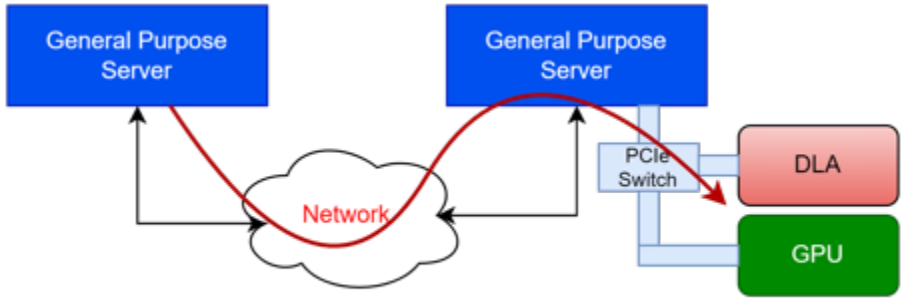
xPU full offload

Use Case Example: AI Inference – Local to Disaggregated

- Use case
  - Clients on GP servers consume deep learning compute resources for AI inference (e.g., GPU)
- Disaggregated flow
  - Client on GP server consumes AI inference resources using library (e.g., runtime serving)
  - Inference serving on server side
    - Parsing AI inference request and schedules it on deep learning compute resources
    - Executes AI inference on allocated hardware resources
    - AI inference result retrieved and returned to client
    - Supports multiple clients/tenants



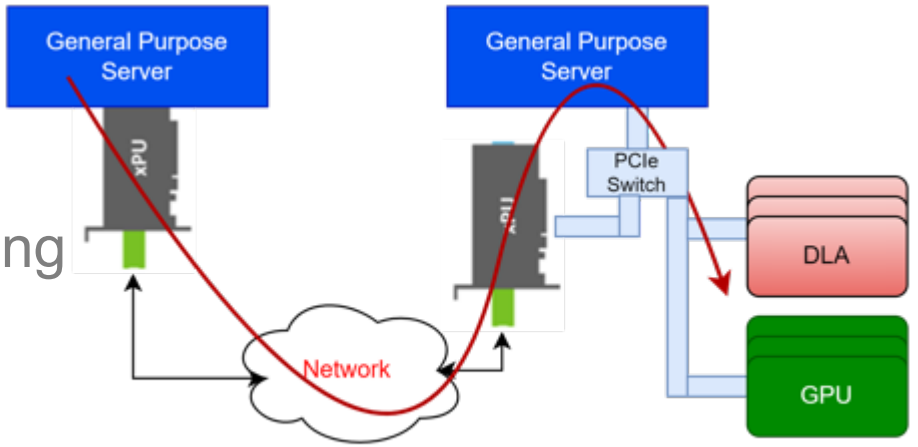
Local



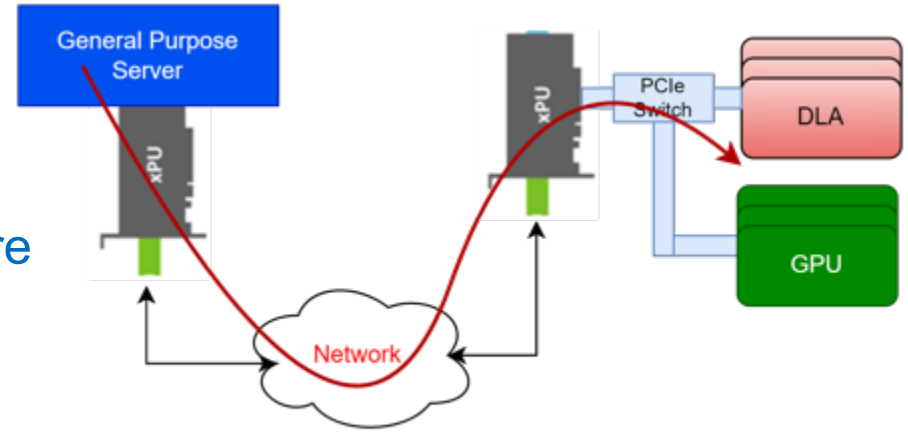
Disaggregated

Use Case Example: AI Inference – Accelerating to xPU

- Integrate xPU into workflow
  - Use xPUs in both client and server network interface
  - Access from xPU in server side directly to deep learning compute hardware (e.g., GPU)
- Offload capabilities to xPU
  - Network protocol offload (including serving protocol)
  - Driving deep learning compute hardware from xPU
  - Resource allocation, virtualization and multi-tenancy
- Full offload to xPU
  - Connecting xPU directly to deep learning compute hardware without any general-purpose server in target side



xPU accelerated



xPU full offload

Summary

- Mapping flow
  - Local -> Disaggregated
  - Disaggregated -> xPU integration and offload
  - Partial offload
    - Partial system flows (e.g., networking, security)
    - Complex functions on target side (e.g., storage failover and recovery)
  - Full offload
    - Target side operations can be fully contained in xPU and additional attached HW
    - xPU capabilities fully support system flow requirements
- Use cases
  - Storage target
  - AI inference
  - Many other flows (security, networking virtualization, general purpose compute, etc.)

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides are 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

