SNIA™ NSF | NETWORKING STORAGE

Storage Life on the Edge: Accelerated Performance Strategies

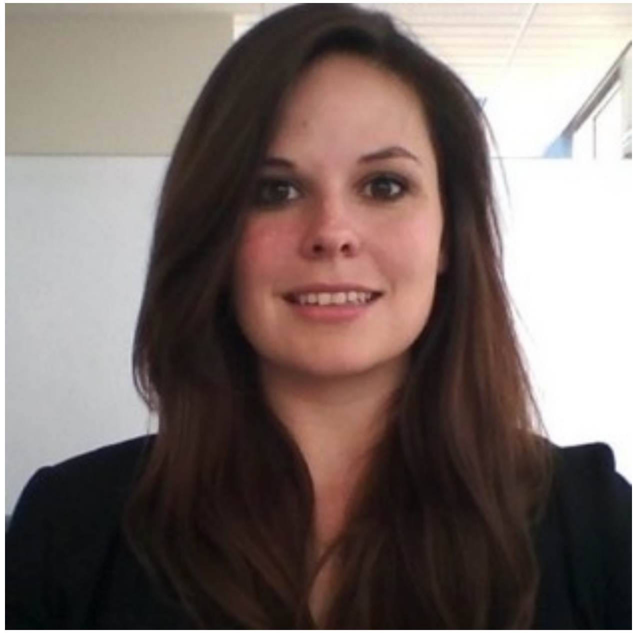
Live Webcast

July 12, 2022

10:00 am PT / 1:00 pm ET

Today’s Presenters

Alex McDonald
Independent Consultant



Amanda Saunders
Senior Manager
Edge AI Product Marketing
NVIDIA

Tushar Gohad
Principal Engineer
Storage Software Architecture
Intel

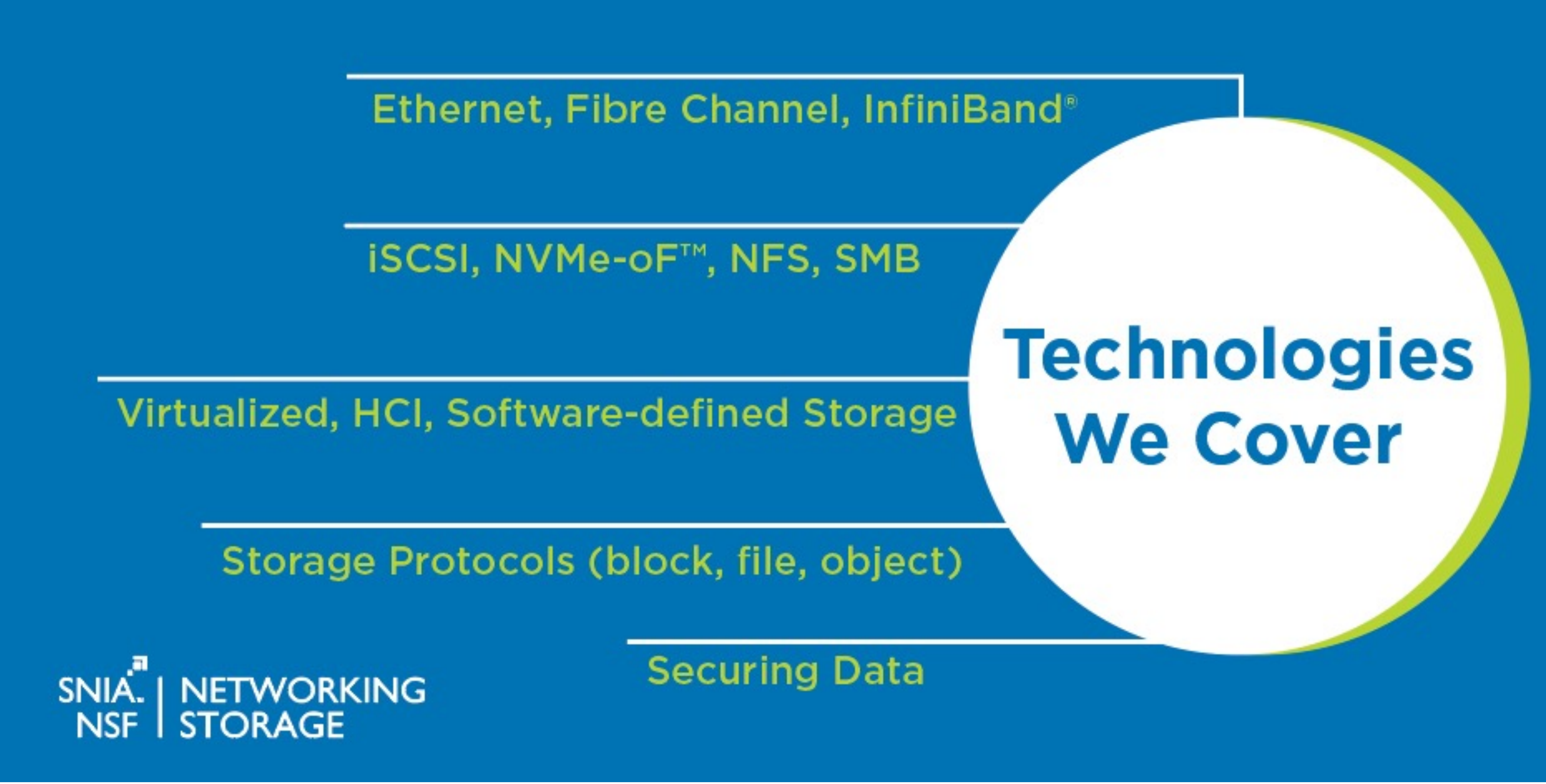
SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

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



Technologies We Cover

- Ethernet, Fibre Channel, InfiniBand®
- iSCSI, NVMe-oF™, NFS, SMB
- Virtualized, HCI, Software-defined Storage
- Storage Protocols (block, file, object)
- Securing Data

SNIA. NSF | NETWORKING STORAGE

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

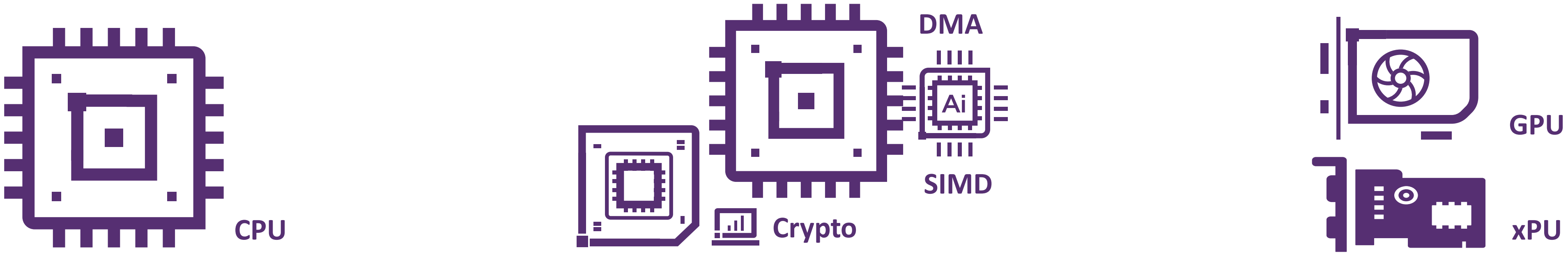
- Continuum of Acceleration Strategies
- Accelerating Edge AI
  - Moving to Intelligence at the Edge
  - Sizing AI Accelerators for the Edge - xPUs
- Accelerating Edge Cloud and CDNs
  - Storage Models at the Edge
  - Edge Locations, Requirements and Constraints
  - Choices Edge Storage and Caching Acceleration
  - Software Innovation for better Perf/Watt

Accelerating Edge AI

Amanda Saunders

NVIDIA

Continuum of Acceleration Strategies at the Edge



Optimized Software

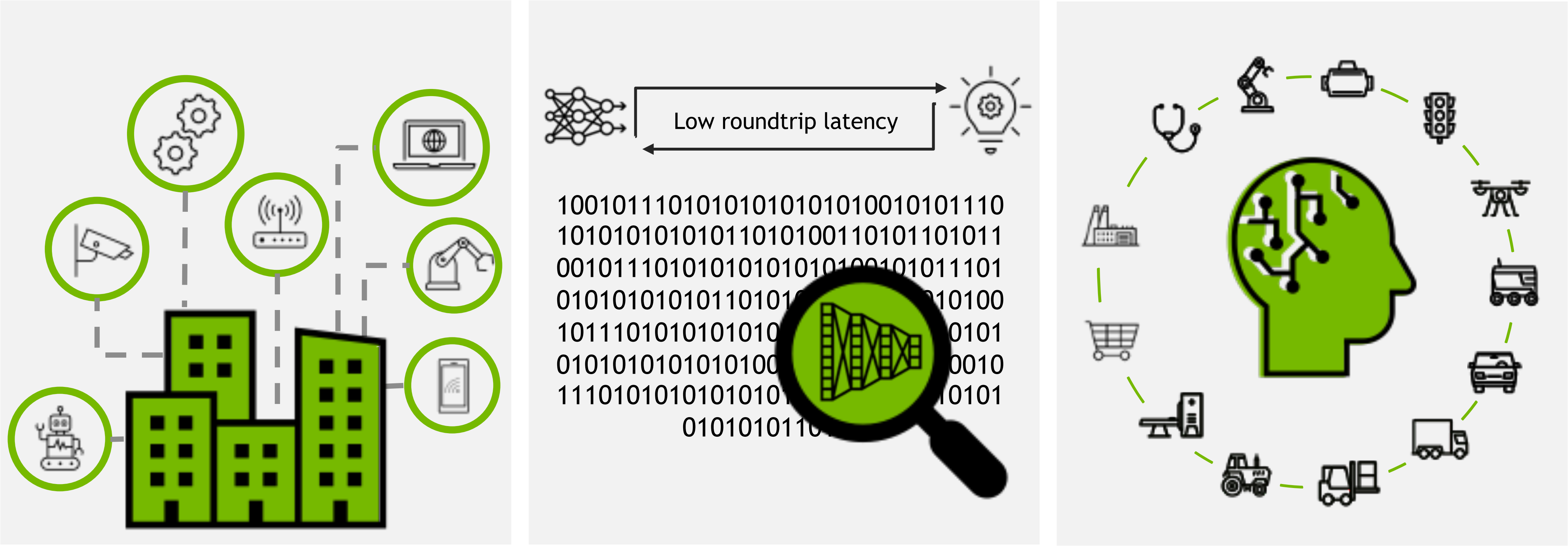
Integrated Accelerators

Discrete Accelerators

Workload Flexibility ⟷ *Performance*

Moving to Intelligence at the Edge

DATA COLLECTION → AI INFERENCE AT THE EDGE → INTELLIGENT SPACES



What are Intelligent Spaces?

1

Transportation and Logistics

Traffic flow management
Road Digital Signage
Suspicious Activity Monitoring
Warehouse Autonomous Mobile Robot

2

Smart Factory

Industrial Inspection
Perceptive Robotics
Materials Handling
Factory Floor Video Analytics
Digital Twin and Sensor Fusion
Preventive Maintenance
Additive Manufacturing

3

Smart Store

Automated Checkout
Store Traffic Analytics
Inventory Management
Shopper Analytics
Intelligent Digital Signage
Social Distancing Detection

4

Smart City

Traffic Analytics
Vehicle Counting
Number Plate Detection
Surveillance and Public Safety
Smart Parking System

5

Smart Hospital

Surgical Robot
Medical Image Assistant
Telepathology
Patient Health Monitoring
Digital Health System

6

Agriculture

Intelligent Robot Assistant for Harvesting
AI Pollinator
BeeHome with Robots and AI
Livestock Health Management
Selective Spraying system
Smart Farm Machines

Sizing Accelerators for the Edge - xPUs

GPU

Up to 100X increased AI performance

- Size, power, temperature
- GPU memory
- Inferences/second
- Encoders and decoders

CPU

Memory and System Function

- Single-socket for edge use cases
- Low power
- 6 physical CPU cores per GPU

DPU

Improved Security and CPU Offload

- Ultra-low latency
- Zero-trust security

Questions to Ask When Sizing Your Environment

APPLICATION QUESTIONS

What application(s) will I need to run at the edge?

How many streams of data will they process?

What is the latency needed for my application to be successful?

Are there software optimizations that can be done to reduce the hardware requirements of my application?

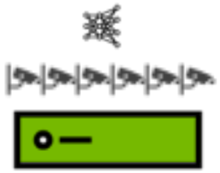
EDGE ENVIRONMENT QUESTIONS

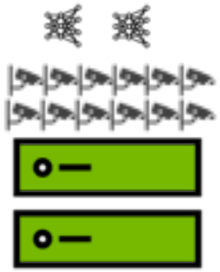
What are the power and space requirements at my location?

Do I have the same requirements at all my locations?

Is the edge environment connected, semi-connected, or disconnected?

Computer Vision: Small, Medium, and Large

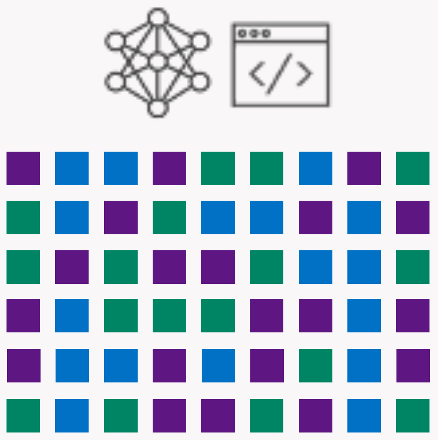




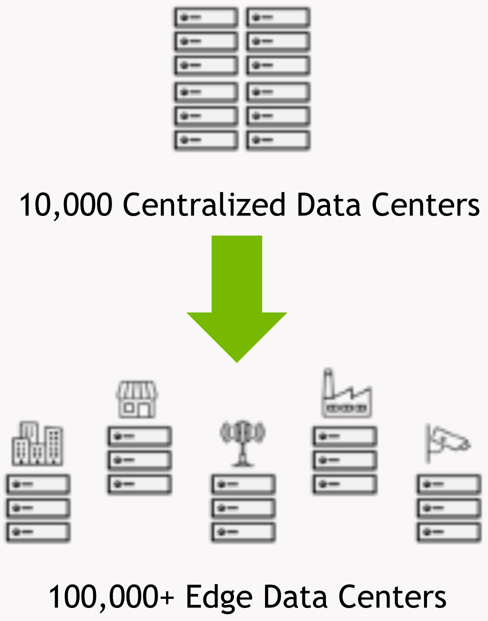
	Small configuration	Medium configuration	Large configuration
Scenario	6 to 7 video processing streams for people detection	10 to 12 video processing streams for inspection and people counting	20 video processing streams for inspection, people counting, and vehicle identification
Server model	1U	2U	2U
Processor	6 x CPU cores	12 x CPU cores	24 x CPU cores
Memory	8 x 8 GB	8 x 16 GB	16 x 16 GB
GPUs	1 Low Power GPU	2 Low Power GPU	1 Medium Power GPU
Storage	6 x 480 GB SAS SSDs in RAID 6	8 x 480 GB SAS SSDs in RAID 6	12 x 480 GB SAS SSDs in RAID 6
Trusted Platform Module	Trusted Platform Module 2.0	Trusted Platform Module 2.0	Trusted Platform Module 2.0

AI is Reshaping the Data Center

MASSIVE AMOUNTS OF DATA REQUIRED



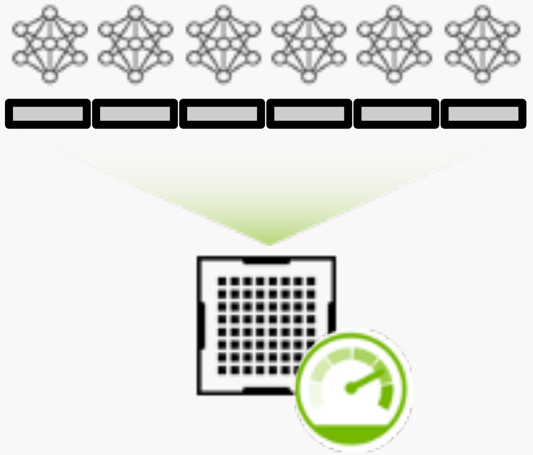
DATA CENTERS WILL BE DISTRIBUTED



10,000 Centralized Data Centers

100,000+ Edge Data Centers

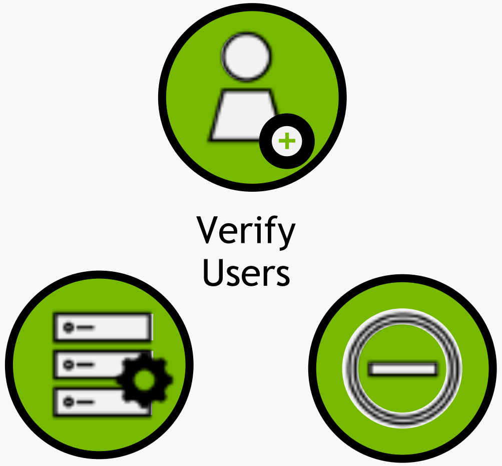
PERFORMANCE IS PARAMOUNT



NEW MODELS OF SECURITY

Foundation of Secure Edge

LEAST TRUST REMOTE ACCESS

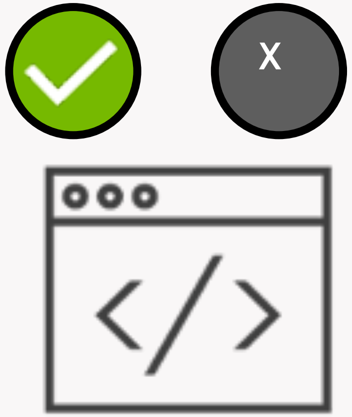


Verify Users

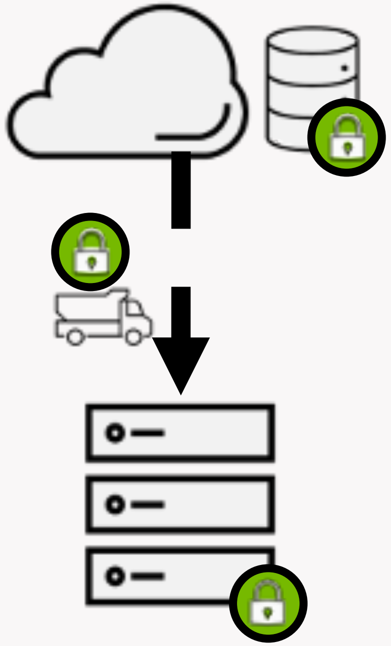
Validate Devices

Limit Privileged Access

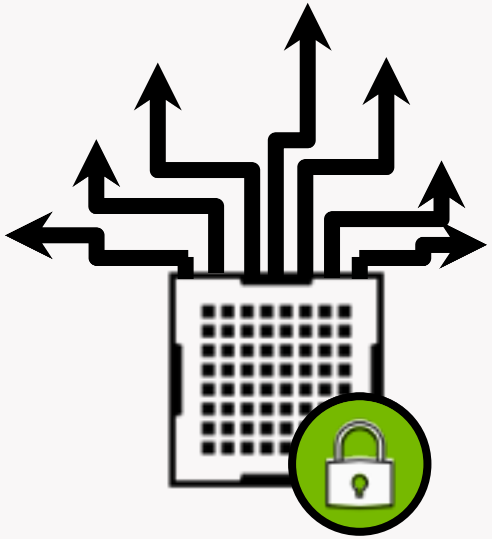
SIGNED/MEASURED SOFTWARE



ENCRYPTION IN TRANSIT AND AT REST



PROTECTING AI MODELS IN USE



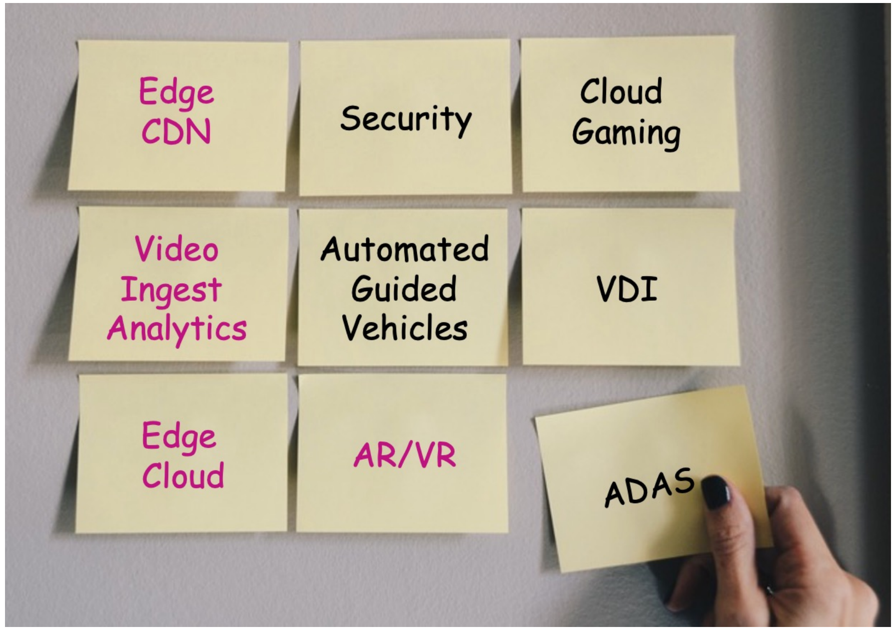
Accelerating Storage and Caching at the Edge

xPU and Software-based Acceleration for better Perf/Watt

Tushar Gohad

Intel

Edge Use Cases and Storage Models – Recap



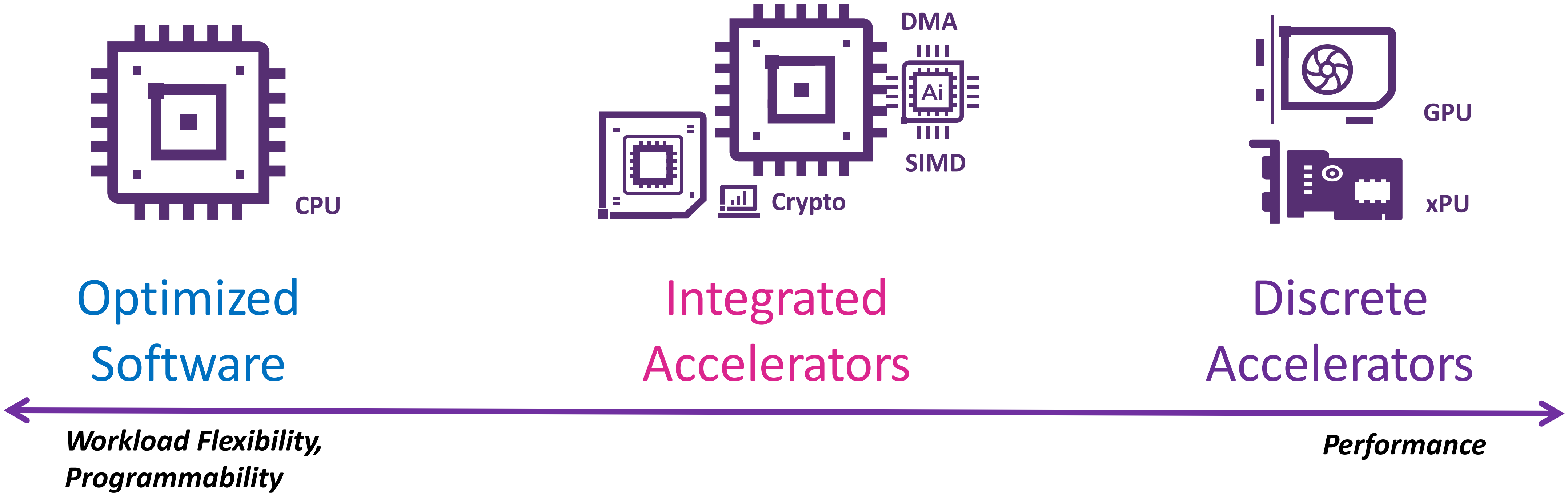
- Edge Storage
  - Edge Compute, Edge Cloud, Video Ingest and Analytics
  - S3 Object, KV (Cloudflare R2, WorkersKV, etc)
  - Local and disaggregated Block/File (NVMe-oF, SDS)
  - Computational Storage
- Edge Caching
  - CDNs (Video, AR/VR), Edge Storage Gateways
  - Cache upstream and downstream data
  - Improves performance by proximity
  - Reduces egress costs from Cloud
  - CDN to remain one of the top Edge use cases

https://medium.com/@mfcaulfield/edge-computing-9-killer-use-cases-for-now-the-future-ed2083ff4e6c
https://blog.min.io/storage-at-the-edge/

Continuum of Acceleration Strategies at the Edge

- Accelerators deliver enhanced benefit to a workload or use case through purpose-built hardware or software

- Workload acceleration strategies, not mutually exclusive:



Optimized Software

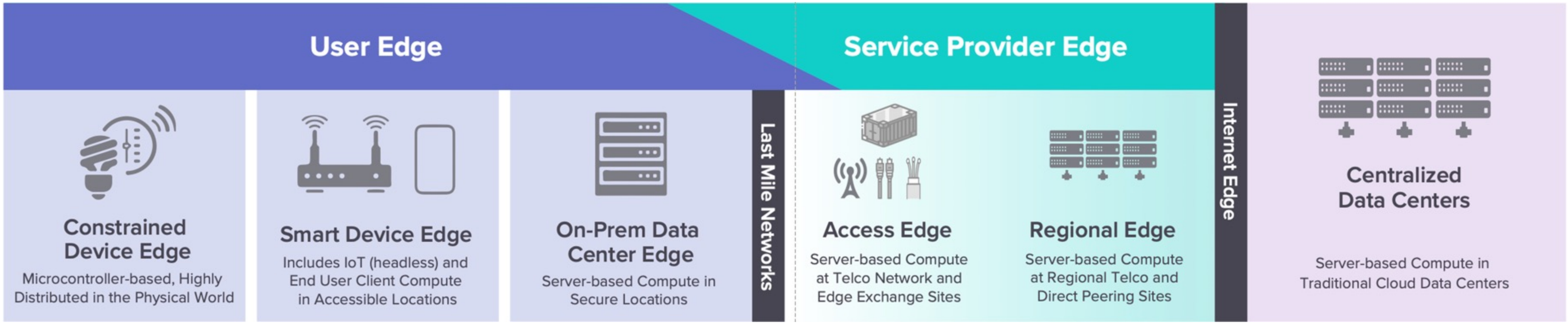
Integrated Accelerators

Discrete Accelerators

Workload Flexibility, Programmability

Performance

Edge Locations – Varying Requirements/Constraints



User Edge

Service Provider Edge

Constrained Device Edge
Microcontroller-based, Highly Distributed in the Physical World

Smart Device Edge
Includes IoT (headless) and End User Client Compute in Accessible Locations

On-Prem Data Center Edge
Server-based Compute in Secure Locations

Last Mile Networks

Access Edge
Server-based Compute at Telco Network and Edge Exchange Sites

Regional Edge
Server-based Compute at Regional Telco and Direct Peering Sites

Internet Edge

Centralized Data Centers
Server-based Compute in Traditional Cloud Data Centers

Physical Attributes	Custom Form-factors	Mix of specialized Form-factors with unique I/O	GP Server Infrastructure Some ruggedization regulatory considerations	Purpose-built radio + GP infra. Power, thermal, ruggedization and regulatory considerations	GP infra. Power, thermal, ruggedization and regulatory considerations	General purpose server infrastructure

Latency critical applications

Latency-sensitive applications

Increasing Resource Constraints, HW + SW Customization and deployment scale

https://www.lfedge.org/wp-content/uploads/2020/07/LFedge_Whitepaper.pdf

Edge Location and Choice of Acceleration Strategy

- Requirements
  - Workload Heterogeneity
  - Security and Privacy
  - Sustainability
  - Low Latency
  - Manageability
  - Elasticity

- Constraints
  - Lights-out Environments, Floor Space
  - Limited Power, Cooling
  - Total Cost of Ownership

- Optimization Targets
  - Performance per Watt
  - Performance per $$
  - Performance per RU
  - $$ per TB

Enviro Constraints and Optimization Targets drive the Choice of Acceleration Strategy

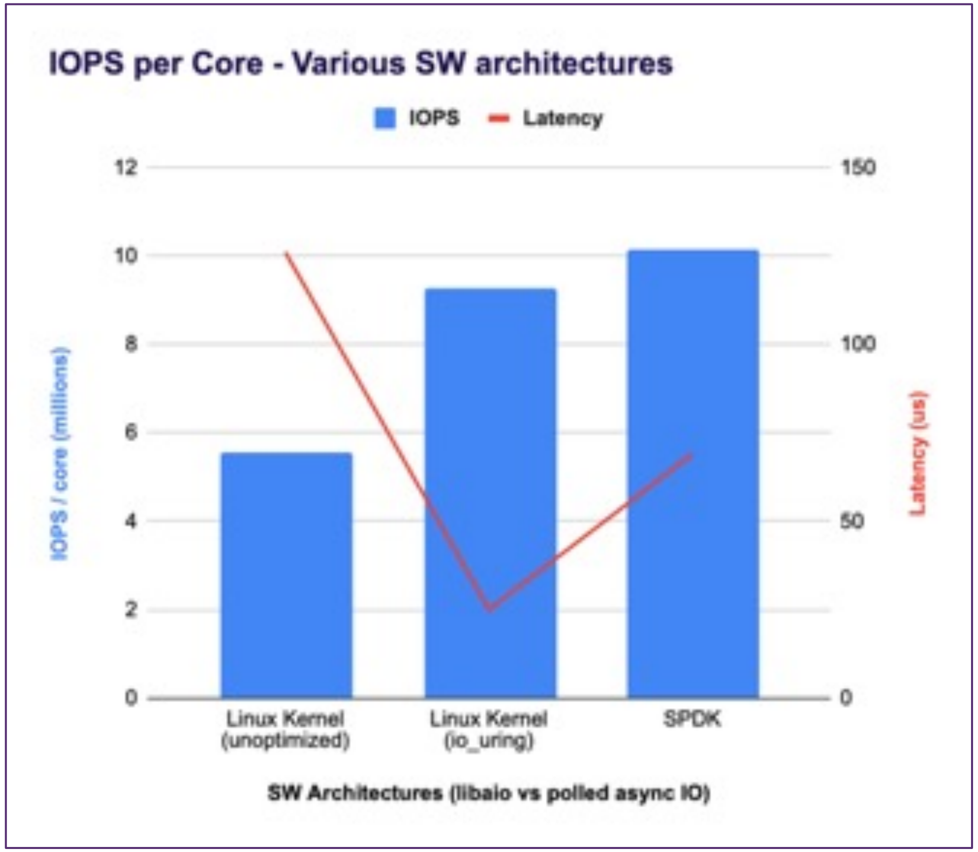
https://stlpartners.com/articles/edge-computing/what-is-edge-computing/
https://jelvix.com/blog/what-is-edge-computing
https://www.lfedge.org/wp-content/uploads/2020/07/LFedge_Whitepaper.pdf

Optimized Software

Software Innovation for Better Perf/Watt

- SPDK or io_uring programming model for Storage (Local or Disaggregated Storage)
  - Userspace, polling, asynchronous IO
  - Runtime detection of CPU/platform features
  - Dynamic, hugepages based memory allocation (SPDK)
  - Zero-copy TX (SPDK), io_uring for rcv (SPDK)

- NVMe-over-TCP Target Example
  - > 65% better Perf/Watt with io_uring
  - > 81% better Perf/Watt with SPDK over legacy libaio-based implementations



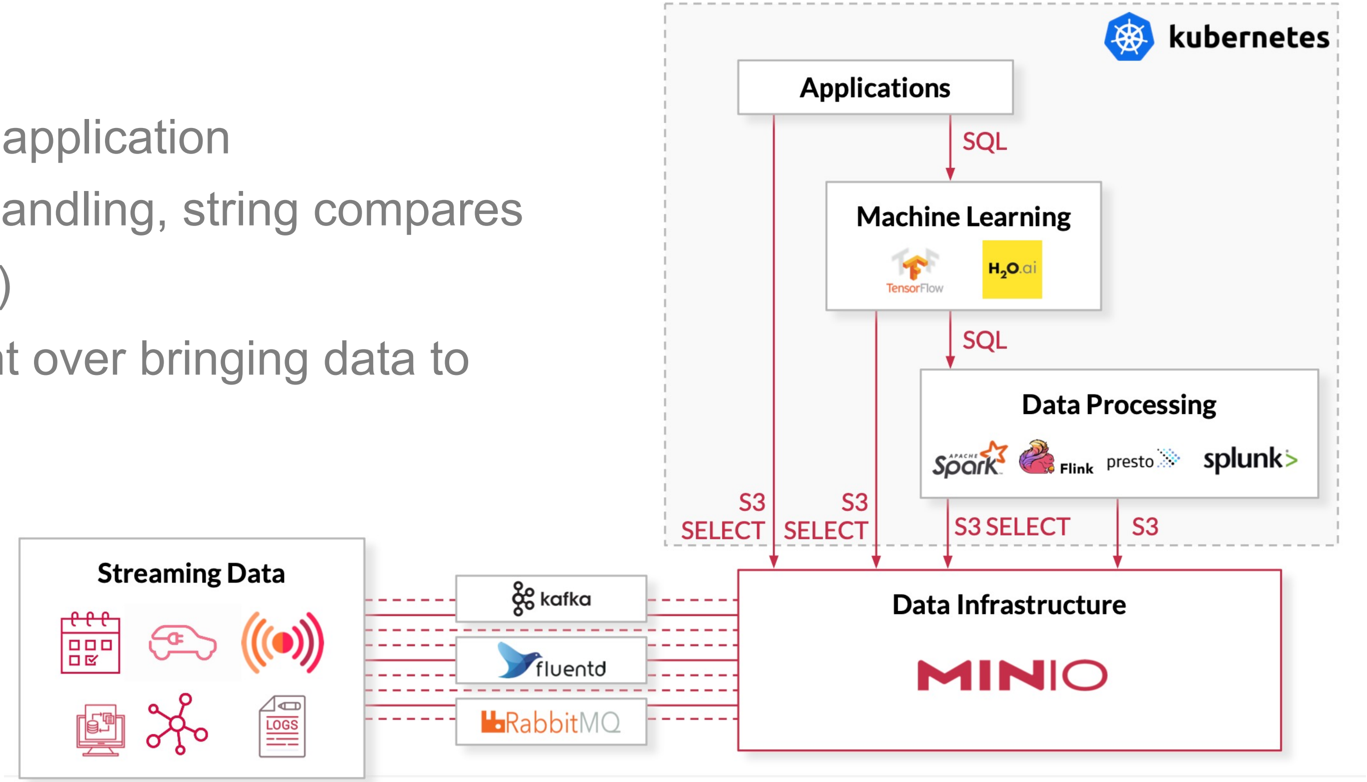
https://ci.spdk.io/download/performance-reports/SPDK_nvme_bdev_gen4_perf_report_2201.pdf

Optimized Software

Integrated Accelerators

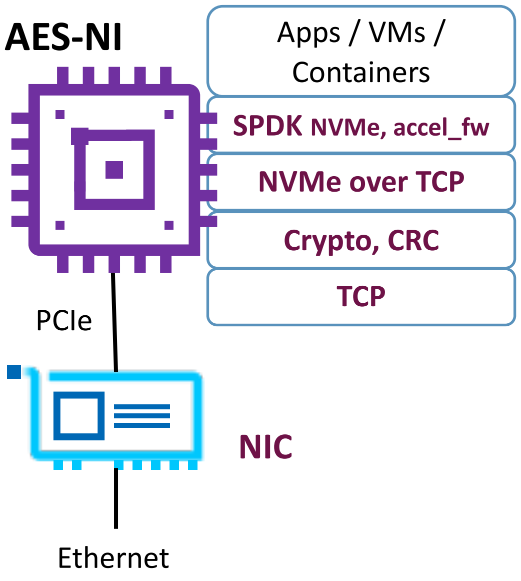
Software Innovation for Better Perf/Watt

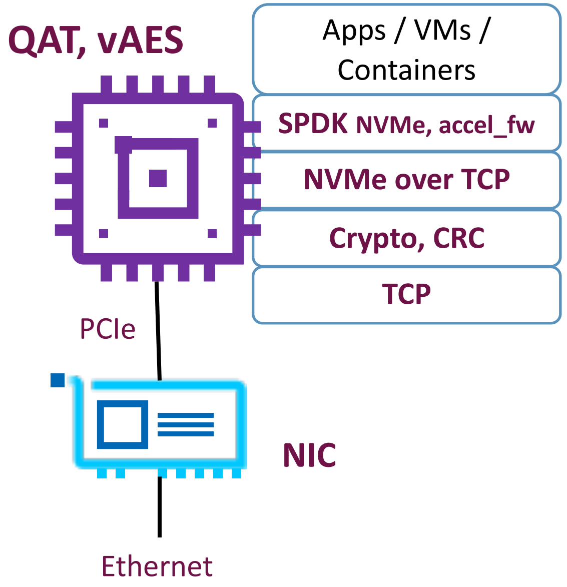
- Modern S3-Select architecture for Edge Analytics (Computational Storage)
  - Offload filtering to Storage
  - Retrieve only the data needed by your application
  - Vectorized (AVX512) parsing, bitmap handling, string compares
  - Erasure coding optimizations (AVX512)
  - Large (>30x) query speed improvement over bringing data to the local Spark node for processing

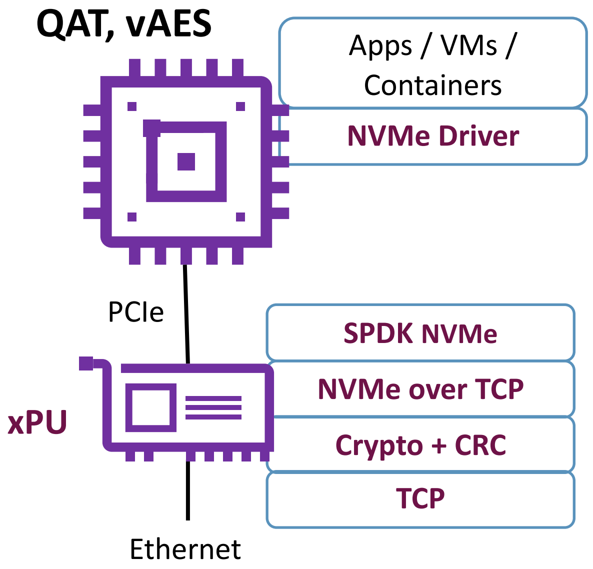


https://blog.min.io/running-peta-scale-spark-jobs-on-object-storage-using-s3-select

Edge Storage: NVMe-oF Initiator implementations







- Optimized SPDK + TCP SW stack on Host
- **Crypto and CRC on Host**

Optimized Software

- Optimized SPDK + TCP SW stack on Host
- **Crypto and CRC accelerated via AVX512 based vectorization + AES ISA**

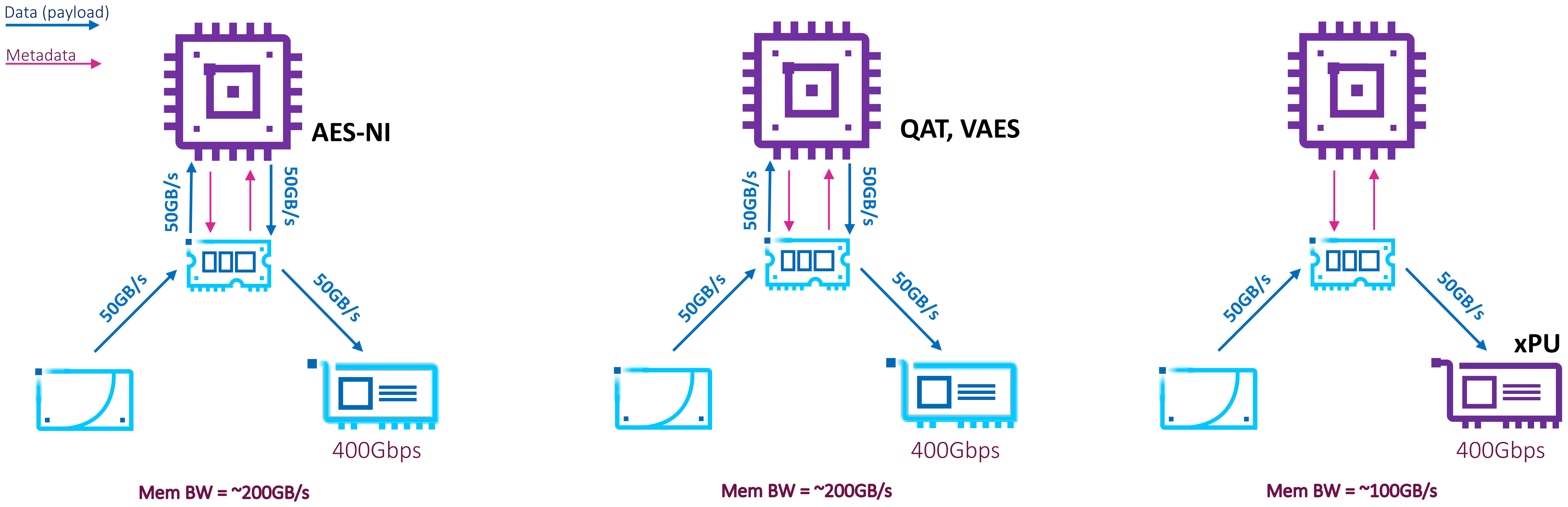
Integrated Accelerators

xPU based NVMe/TCP Initiator

- TCP stack on xPU cores, or in a HW path.
- **Inline Crypto and CRC offloads**

Discrete Accelerators

Edge Caching: Video Delivery (CDN VOD)



- **NUMA-aware** CDN stack
- **TLS** (RSA, AES) processing with OpenSSL SW-only
- **35% better Perf/W over baseline**

Optimized Software

- NUMA-aware CDN stack
- **TLS** processing with Vectorized Crypto (AVX512 + AES)
- **10-15% better Perf/W**

Integrated Accelerators

- NUMA-aware CDN stack
- **TLS** processing with xPU
- **50% less mem BW, best Perf/W***

Discrete Accelerators

* Based on recent measurements reported by Netflix

Summary

- Continuum of Acceleration Strategies
- Accelerating Edge AI
  - Moving to Intelligence at the Edge
  - Sizing AI Accelerators for the Edge – xPUs
- Accelerating Edge Cloud and CDNs
  - Storage Models at the Edge
    - Many; multiple choices of strategy
  - Edge Locations, Requirements and Constraints
    - Amplifies issues, makes constraints more apparent
  - Choices: Edge Storage and Caching Acceleration
  - Software Innovation for better Perf/Watt
    - Sustainability goals a bigger focus now than ever before

Storage Life on the Edge is a Series!

- Watch previous presentations at the SNIA Educational Library
  - Storage Life on the Edge: Security Challenges
  - Storage Life on the Edge: Managing Data from the Edge to the Cloud and Back
  - Storage Life on the Edge: Edge Use Cases

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!