



Bringing Intelligence to Storage

SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

Deployment of In-Storage Compute

Scott Shadley, Storage Technologist

NGD Systems, Inc

www.NGDSystems.com



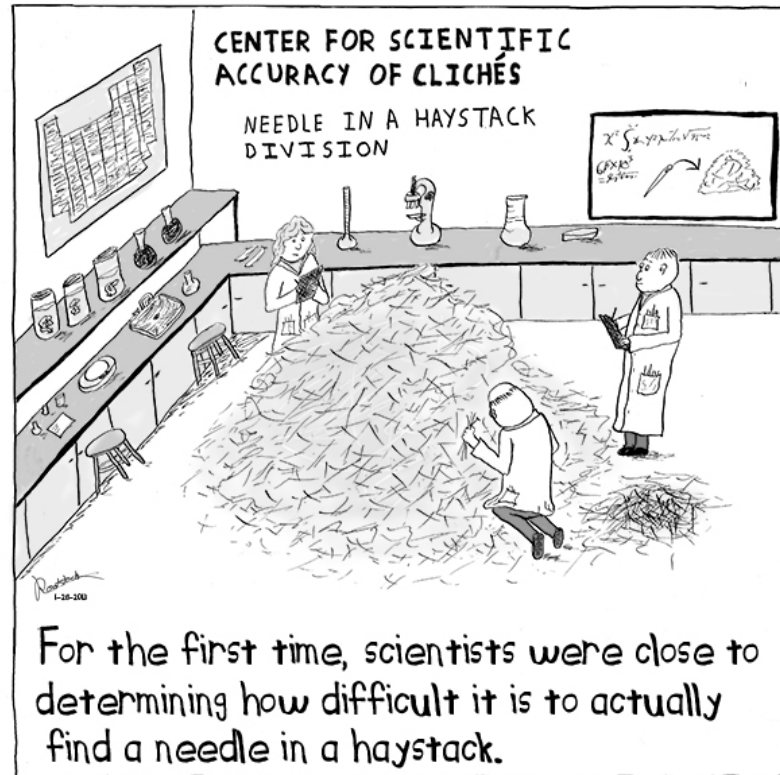
SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

A Market Driven Need Explained

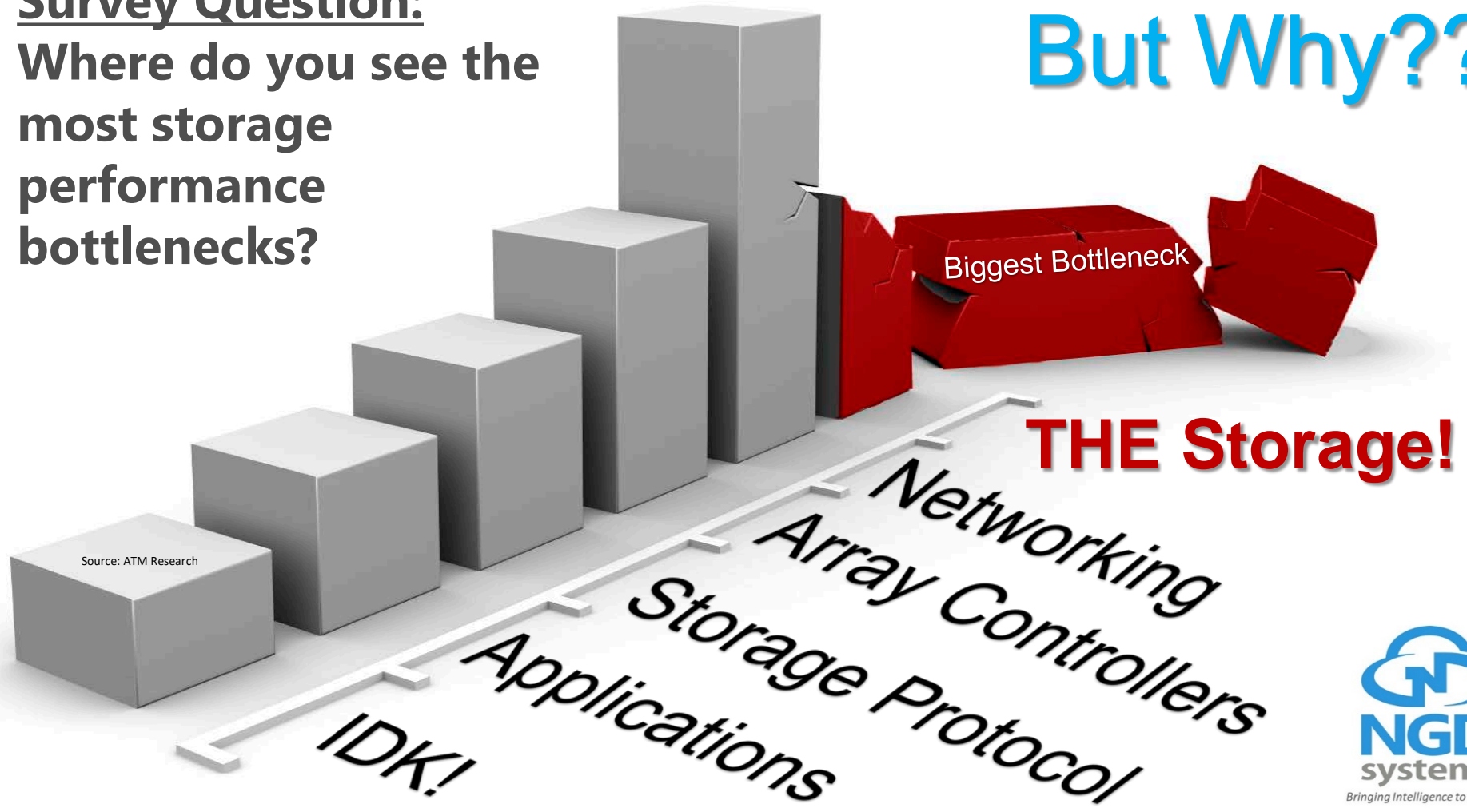
Finding the Needle in a Haystack



Survey Question:

Where do you see the most storage performance bottlenecks?

But Why??

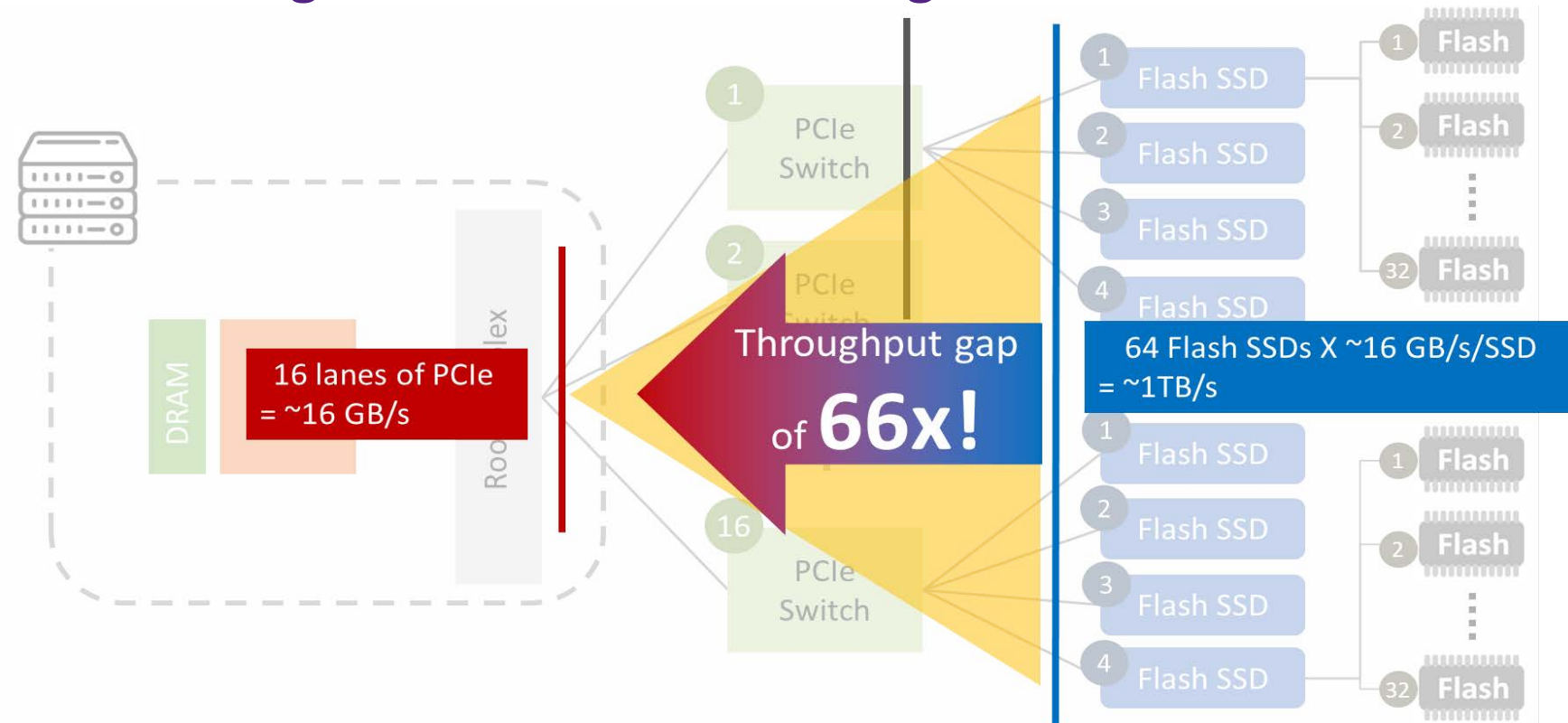


Architectures are not Addressing The Issue

- ❑ Software-defined control and management is an inevitable trend that is already touching other parts of the data center infrastructure
 - ❑ Software-Defined Networking (SDN) making network switches and server NIC cards increasingly programmable for enhanced network-wide functionality
 - ❑ Programmable GPGPUs and FGPAs leveraged by new generations of applications like deep learning
- ❑ Rapidly-changing requirements can be supported on-the-fly once DC infrastructures become dynamically programmable

Don't Leave Storage Behind!

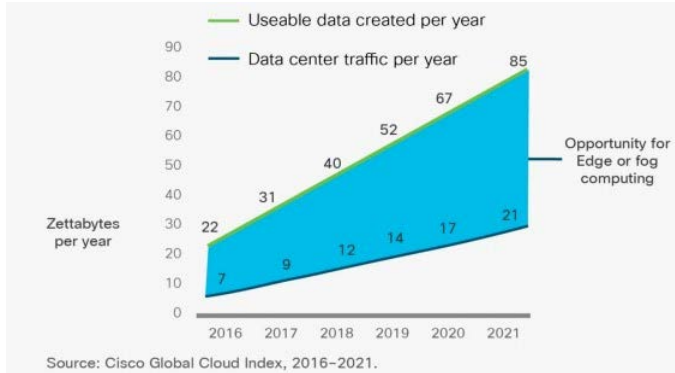
Challenges within Modern Storage Architecture



The Storage Problem... Lack of Near-Data Compute

PUSHED TO THE EDGE

February 19, 2018 Timothy Prickett Morgan



AI Weekly: Computing power is shaping the future of AI

KHARI JOHNSON @KHARIJOHNSON MAY 18, 2018 7:14 PM

NEAR-DATA PROCESSING: INSIGHTS

Near-Data Computation: Looking Beyond Bandwidth

Published in: [IEEE Micro](#) (Volume: 34, [Issue: 4](#), July-Aug. 2014)

Three motivating factors for using Edge Computing

IBM

Internet of Things blog

1. Preserve privacy
2. Reduce latency
3. Be robust to connectivity issues

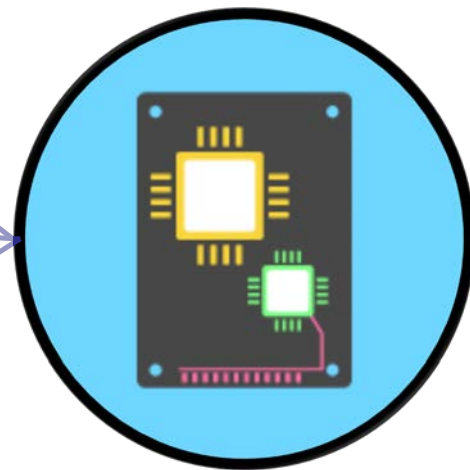
Value propositions

Moving compute close data

Agile, flexible Storage

Secure Computation

Computational Storage



Moving Computation to Data is Cheaper than Moving Data

- A computation requested by an application is much more efficient if it is executed near the data it operates on
 - minimizes network traffic
 - increases effective throughput & performance of the system
 - Example: the Hadoop Distributed File System
 - enables distributed processing
- Especially true for Big Data (analytics): large sets & unstructured data
- Traditional approach: high-performance servers coupled with SAN/NAS storage
- Eventually limited by networking bottlenecks

Dimensions of Computational Storage

operating system

bare metal

RTOS

64-bit OS

user application

firmware

application software

container virtualization

AI applications

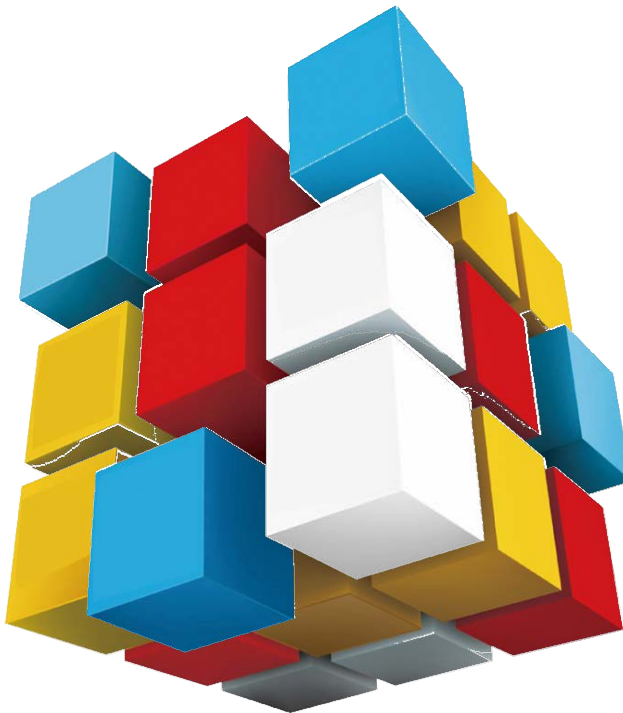
hardware

32-bit real-time processors

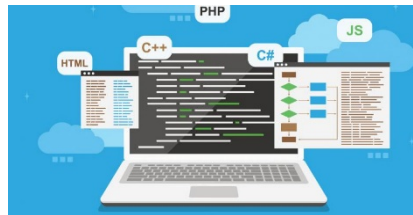
HW acceleration

64-bit application processors

AI acceleration



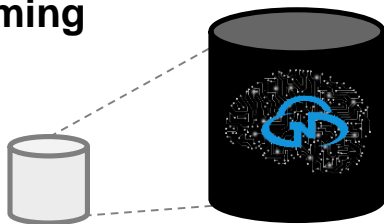
Using **IN-SITU PROCESSING** to Tackle the Mismatch



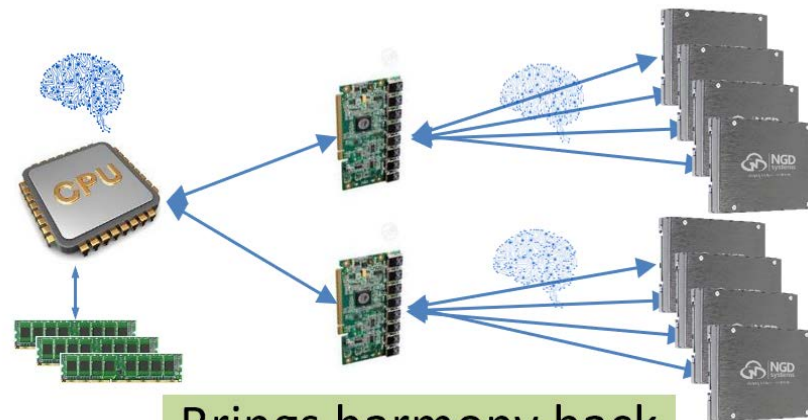
Seamless Programming Model



Scalability

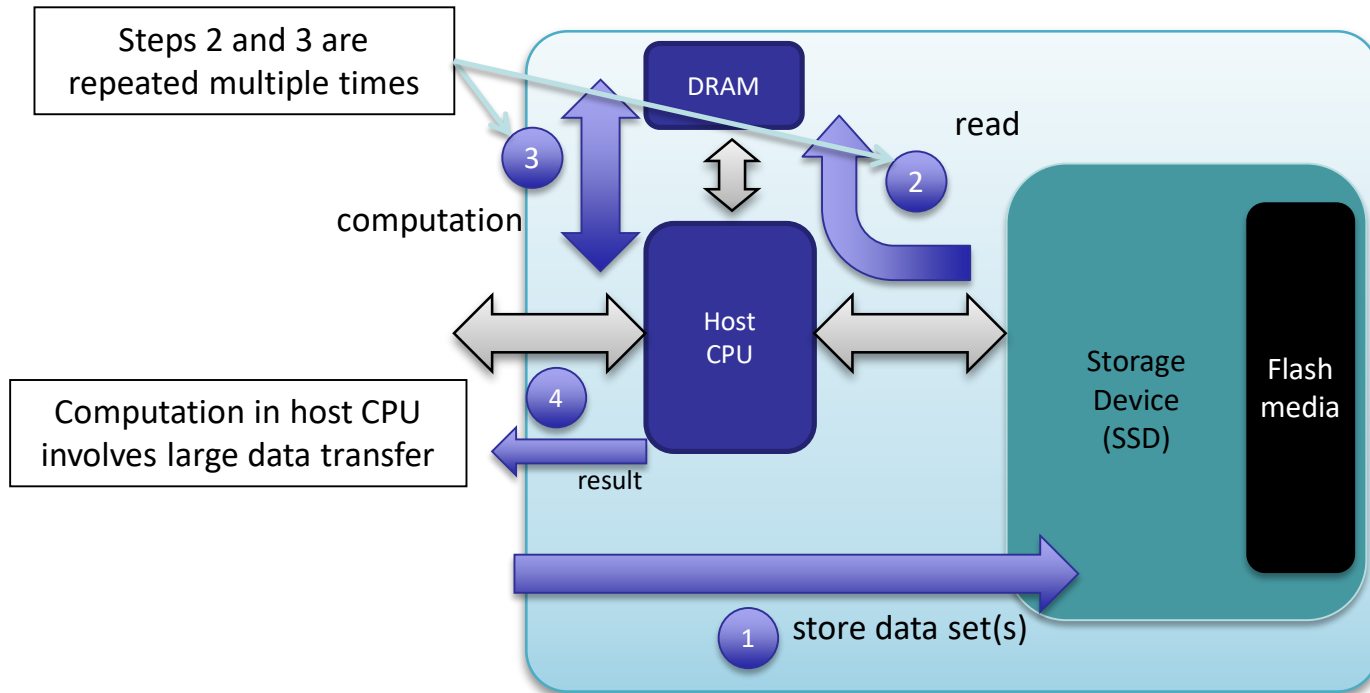


Manage Capacity Growth

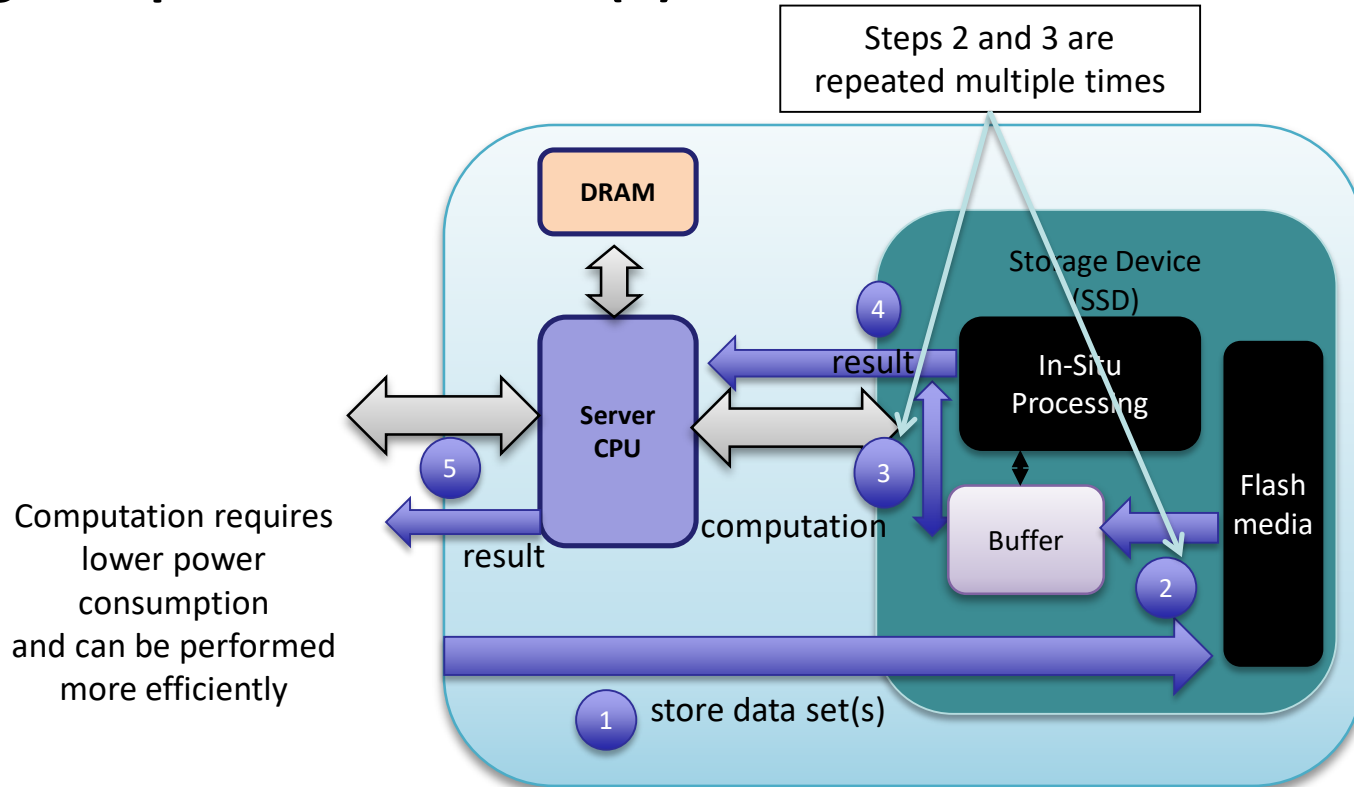


Brings harmony back to bandwidth needs

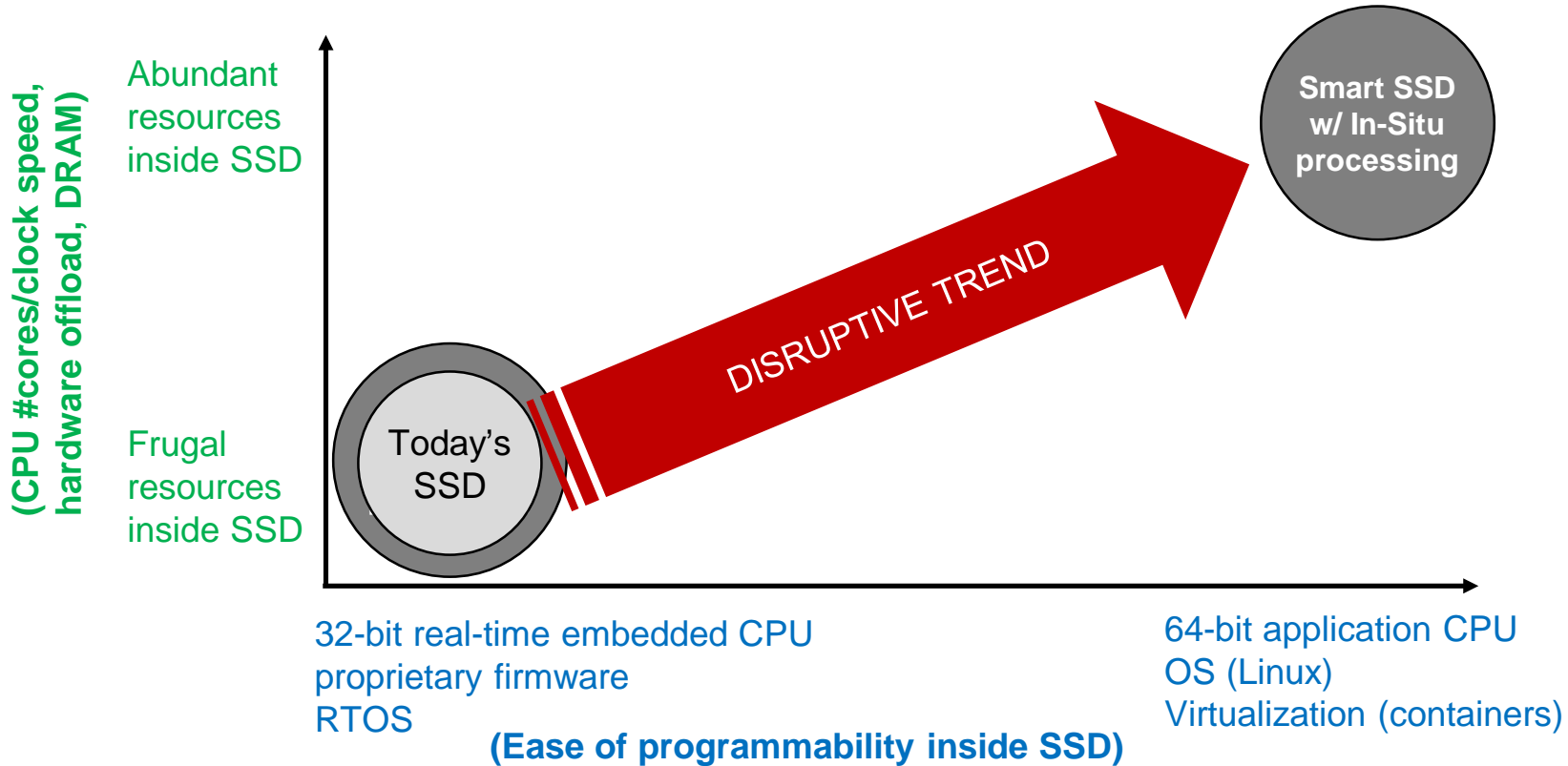
Moving Computation to Data (1)



Moving Computation to Data (2)



Disruptive trends that enable computational storage





SDC¹⁸

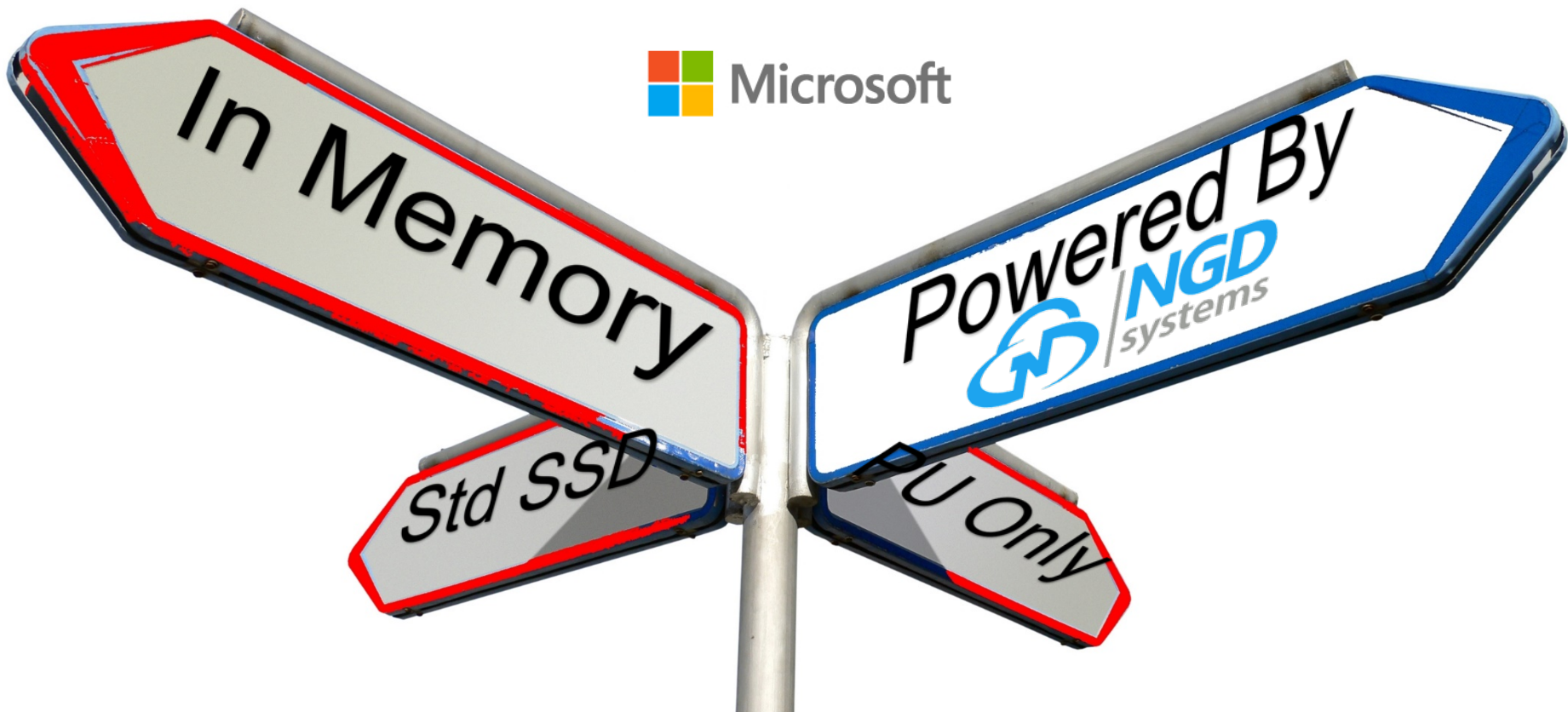
September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

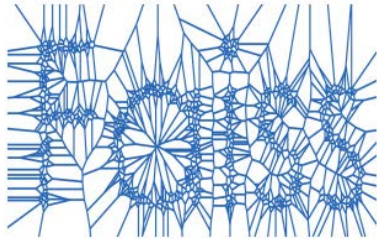
Solutions and Scale

Joint Application Research Programs

Powerful Partners Embrace the Technology



AI Use Case: Image Similarity Search at Data Center



Facebook **AI** Similarity Search



Computational Storage

10 M images

UCSB 2007

1 Billion images

Facebook 2017

1 Trillion images

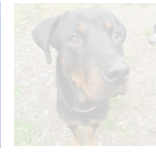
2019



Query



Data Set 1B --> 1T images

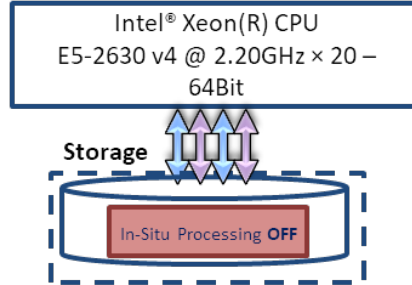


...

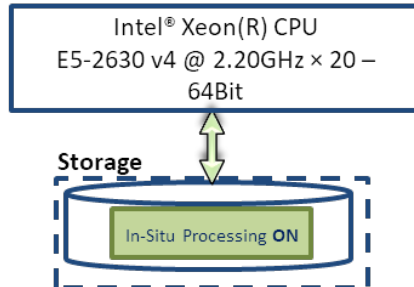


The Base Platform Addressed with Microsoft Research

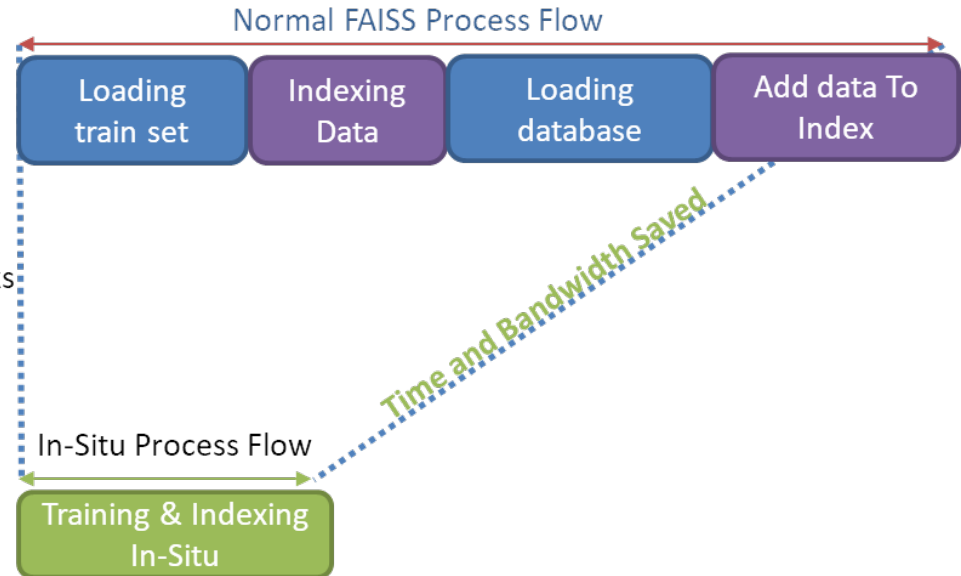
Top- HOST and In-Situ Disable NGD Drive



Bottom- HOST and In-Situ Enabled NGD Drive

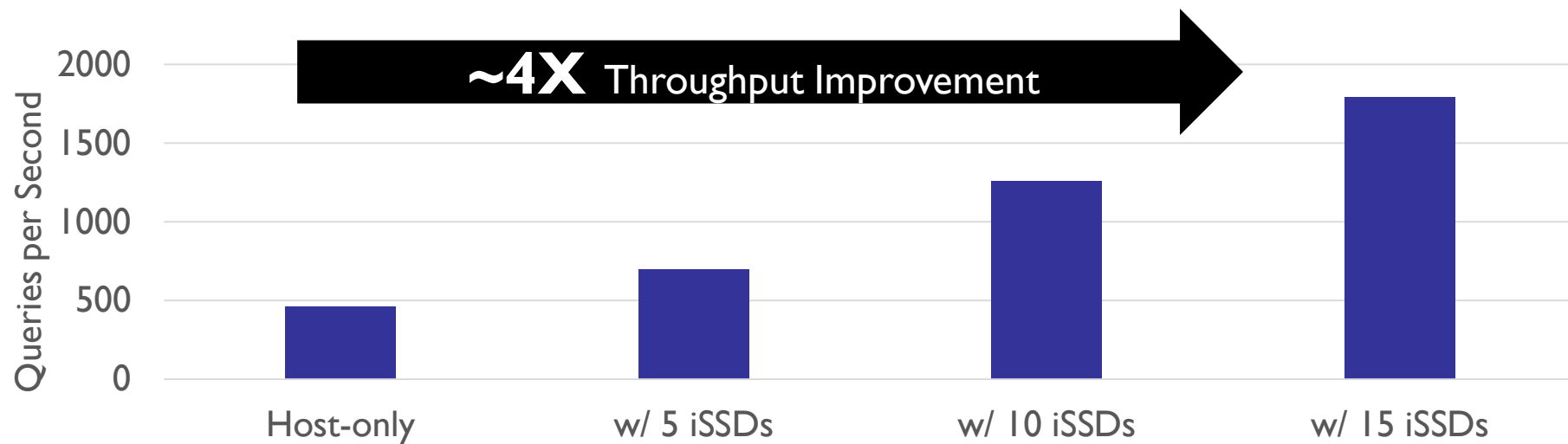


How FAISS works
→ → →

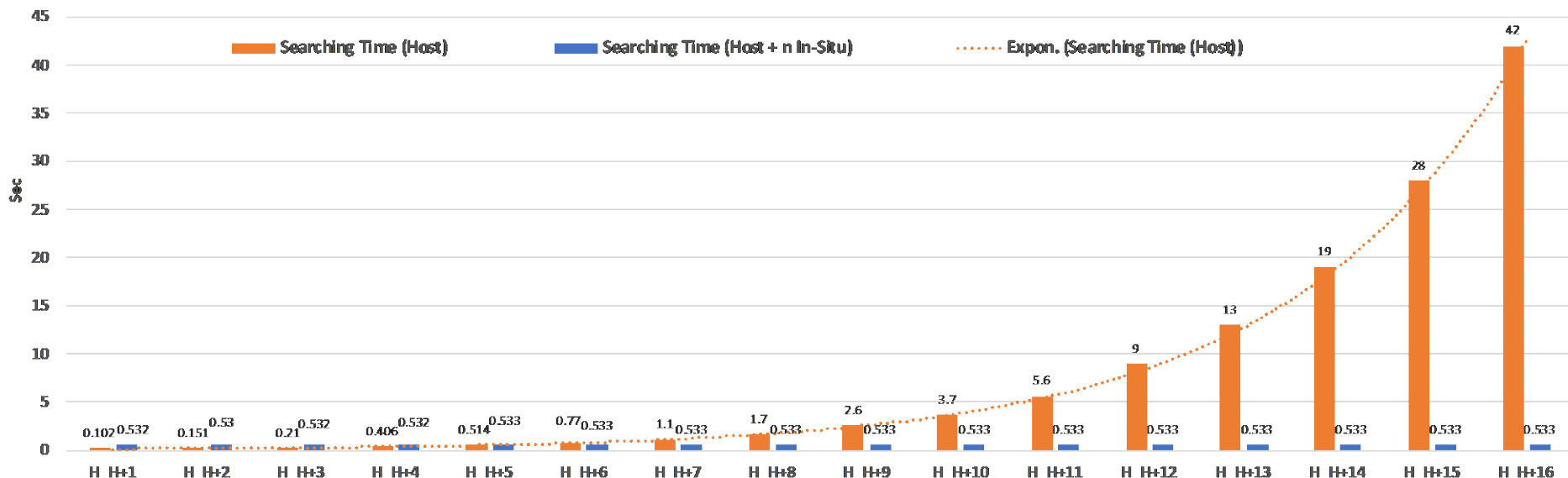


FAISS = Facebook Artificial Intelligence Similarity Search 

Results: Image Query Throughput

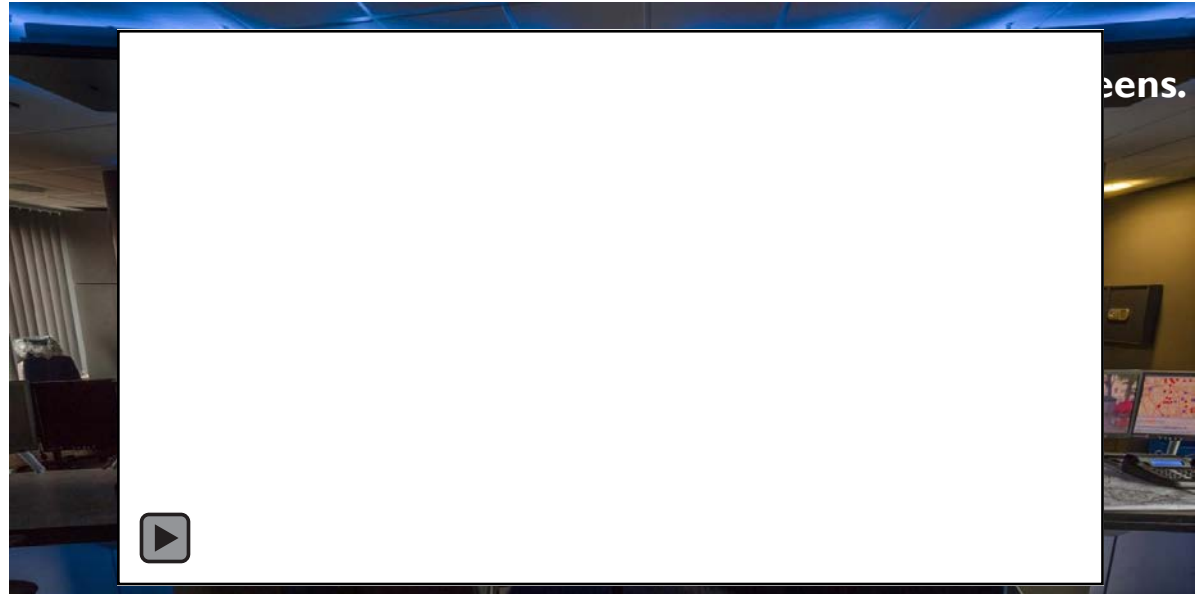
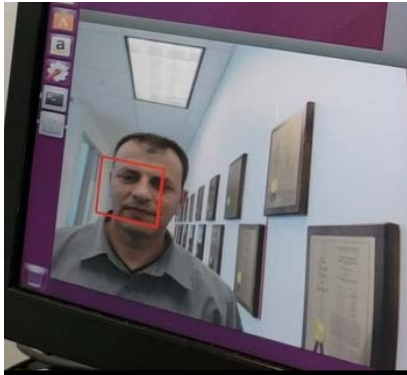


Results: Image Search Time



Object Tracking in Near real Time

At the Edge Object Tracking



Running **OpenALPR** on Drive



NGD's In-Situ Openalpr demo

10.1.1.2/cgi-bin/openalpr.sh

NGD systems
Bringing Intelligence to Storage

In-Situ Openalpr demo

This is a demonstration of the **In-Situ** Processing Capabilities of the NGD's SSD.

The **Openalpr** (Open Automated License Plate Recognition) runs on a **Docker** image inside the embedded system running on the **In-Situ**.

The **In-Situ** Embedded System and the Host System, share the same Filesystem using **OCFS2** Clustered filesystem.

Search License Plate Bank:

Add a new image:

plate0: 10 results

- VODKAA	confidence: 92.0216
- VODKAA	confidence: 90.6883
- VODKAA	confidence: 85.0866
- VODKAA	confidence: 84.0379
- VODKAA	confidence: 83.7532

No file selected.

Run the app on a image:

Choose an image by clicking it:



plate0: 10 results

- 2W1J305	confidence: 93.9312
- 2W1J305	confidence: 86.8626
- 2W1J305	confidence: 86.7148
- 2W1J305	confidence: 86.2435
- 2W1J305	confidence: 85.7515

Protein Sequencing – BLAST® Accelerated

- ❑ DNA and Protein alignment Database Management



The Basic Local Alignment Search Tool (BLAST) finds regions of similarity between sequences.

- ❑ Not balanced dataset
 - ❑ Computation time varies a lot across files
 - ❑ Different number of sequences per file
- ❑ By combining with Computational Storage SSDs and using the 4 cores per drive, you gain **up to 100% more** performance at **no cost** in CPU or Memory



Additional CPU Cores



SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

The Platform Defined and Delivered

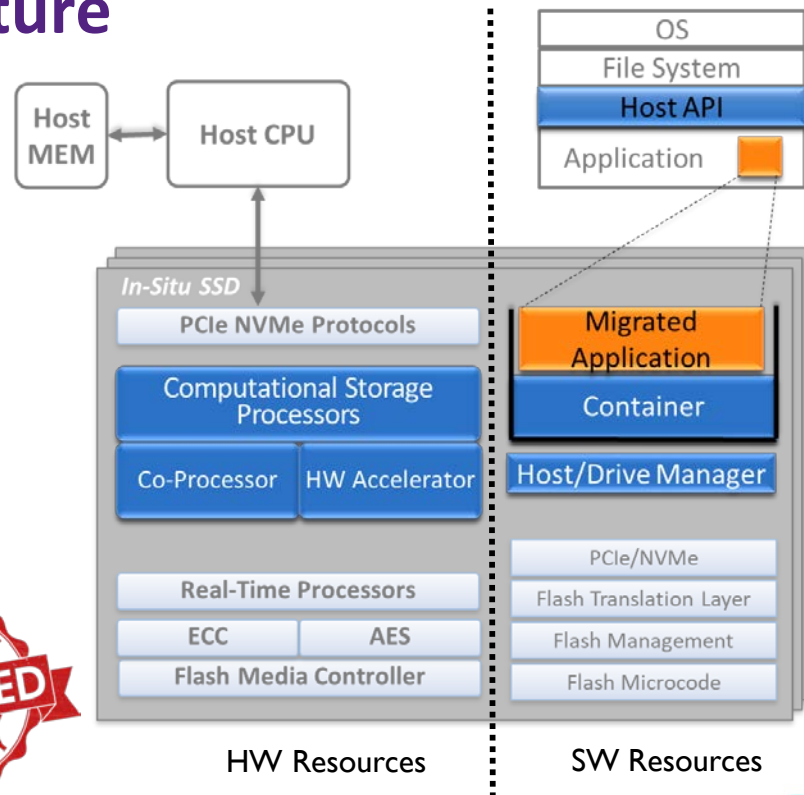
One Final Look at the Architecture

It's an NVMe SSD at the core

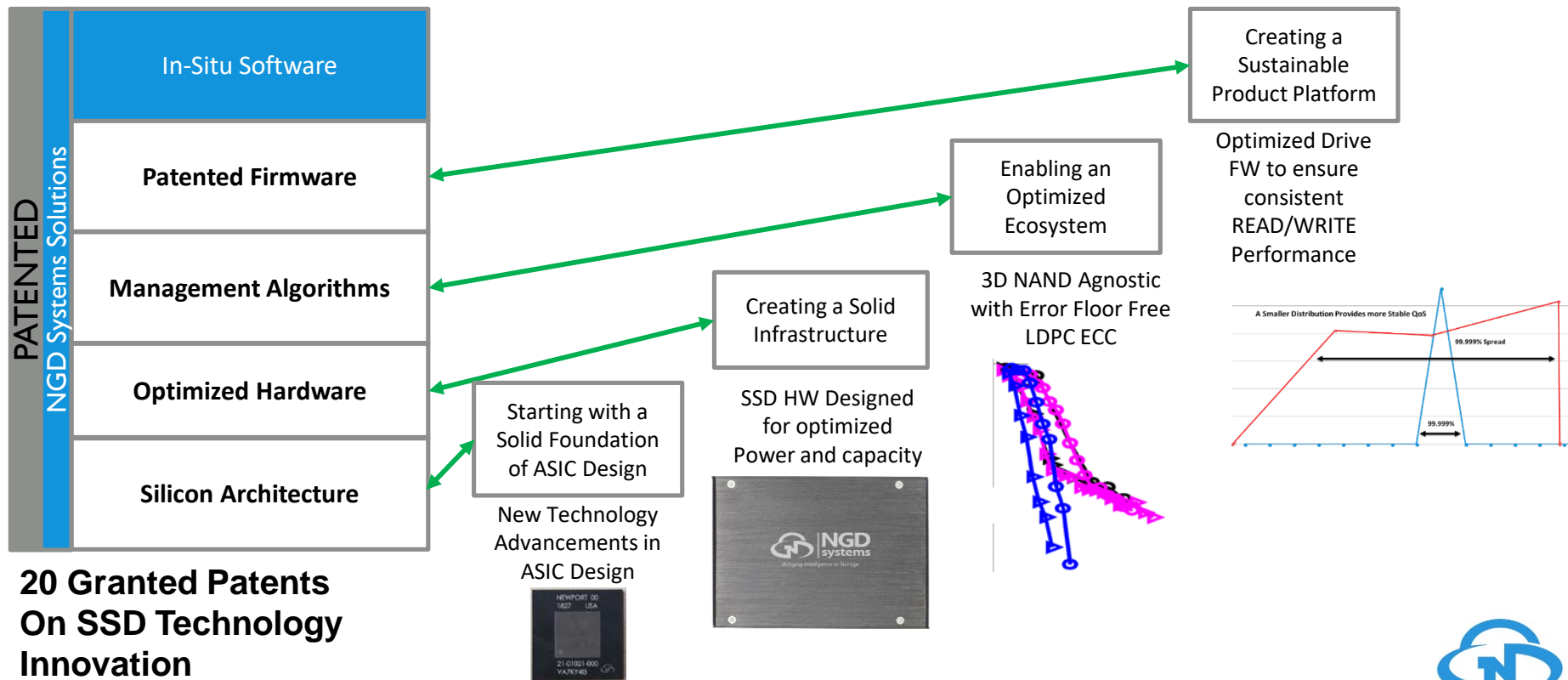
- No impact on host read/write
- No impact on NVMe driver
- Standard protocols

But then there is MORE (Patented IP)

- Dedicated compute resources
- HW acceleration for data analytics
- Seamless programming model
- **Scalable**



Core SSD and Computational Storage Solutions




Computational Storage Platform


PATENTED

NGD Systems Solutions


Patented Firmware
Flexible Management Algorithms
Optimized Hardware
Silicon Architecture




**U.2 15mm Gen3 (x4)
Up to 32TB**



**EDSFF/M.2
Up to 16TB**



Proprietary NVMe Controller






**AIC Gen3 (x4)
Up to 64TB**

Hardware acceleration

Quad-core 64-bit application processor

Full Fledged on drive OS

Light Virtualization



Newport Platform Provides

- ❑ Finding the Needle Faster

IN-SITU PROCESSING

- ❑ Bigger Pipes Feed Smaller Ones



- ❑ Smarter Storage Does Work



- ❑ Requires Intelligent Controllers



- ❑ Power is Factor - Always

Watts/Terabyte

On Drive Linux OS, Container Support

Dedicated Compute Cores

Mitigating Data Movement

Optimizing Application Execution

Partnerships for Success

Real World Implementation

Flash Agnostic – ONFI/Toggle, TLC/QLC

16 Channels - Capacities greater than 64TB

.35 W/TB @ 16TB



SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

THANK YOU!

**Follow us on LinkedIn, Twitter, Facebook or
on the Web @ NGDSystems.com**

SNIA.ORG/COMPUTATIONAL