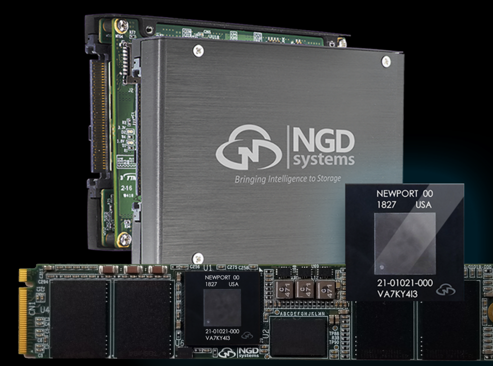
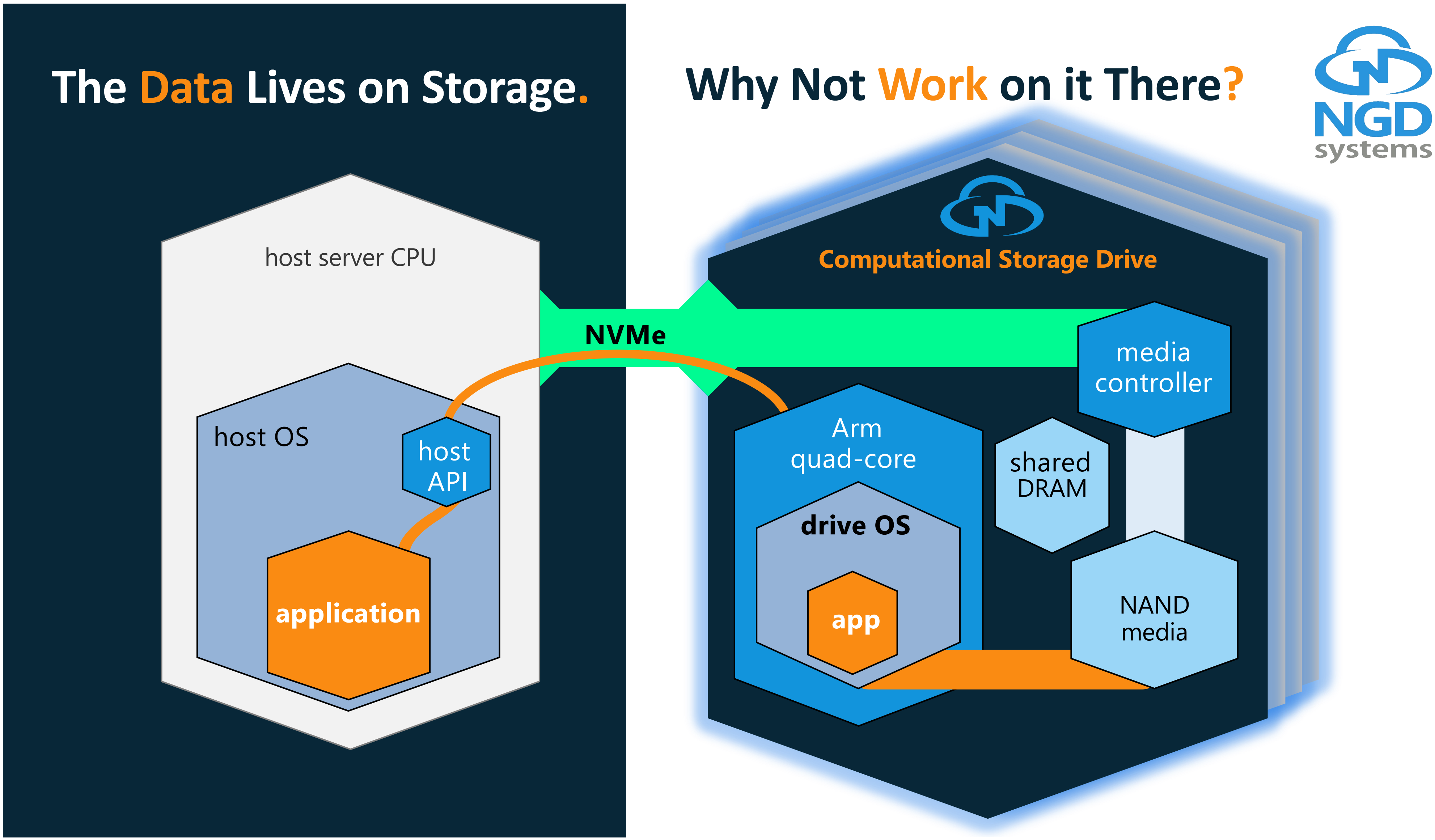
# Computational Storage Distributed AI with ML

A New way to Look at Storage



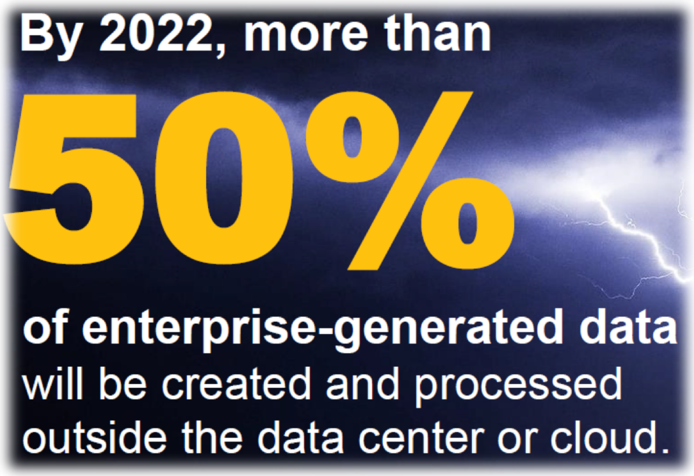
Scott Shadley, VP Marketing
Dr. Vladimir Alves, CTO

# Data, Data, Data. But Don't Take Our Word For it.

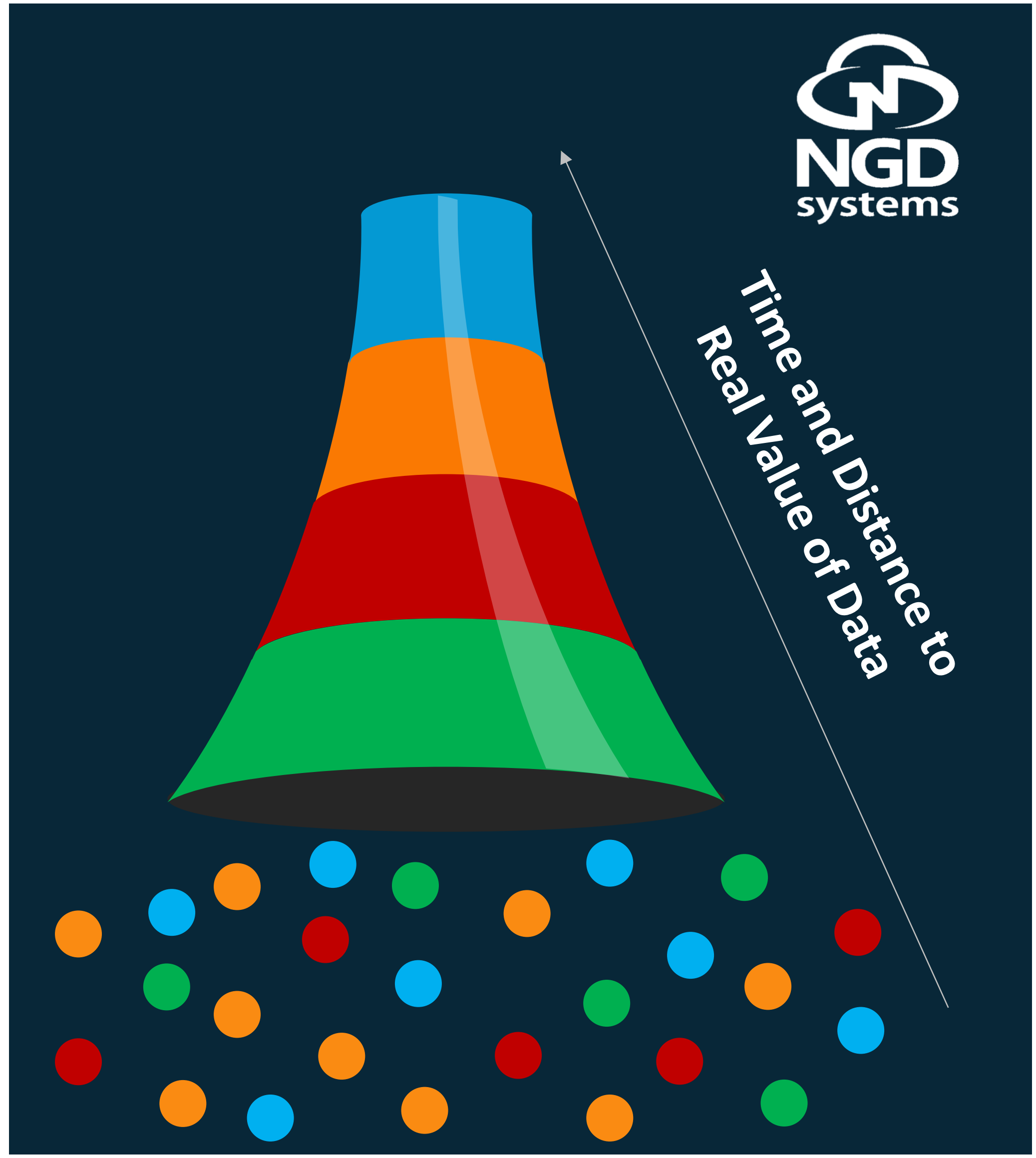

The Data Lives on Storage.
Why Not Work on it There?
NGD systems
host server CPU
host OS
host API
application
NVMe
Computational Storage Drive
media controller
Arm quad-core
drive OS
app
shared DRAM
NAND media

# What is Driving Our Data Analytics Issues.

## Weeding through the Noise at the Edge




Source: Gartner - Bittman



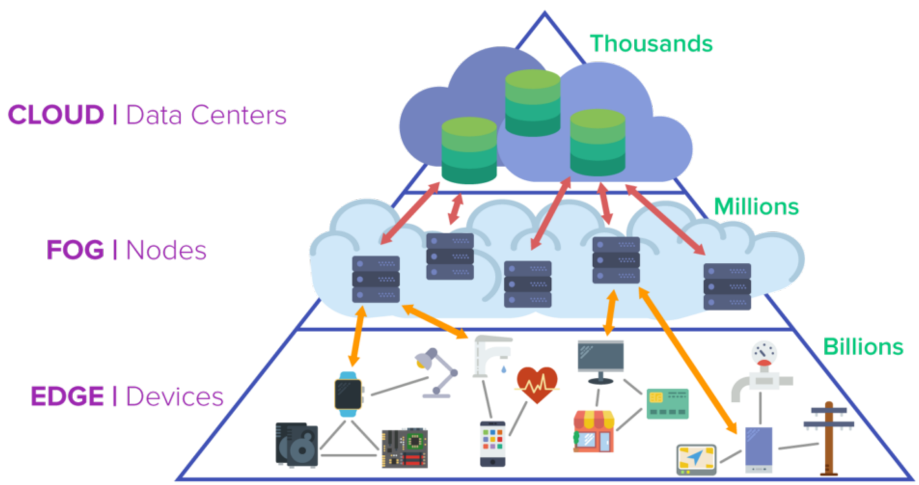
# The Sharp Edge

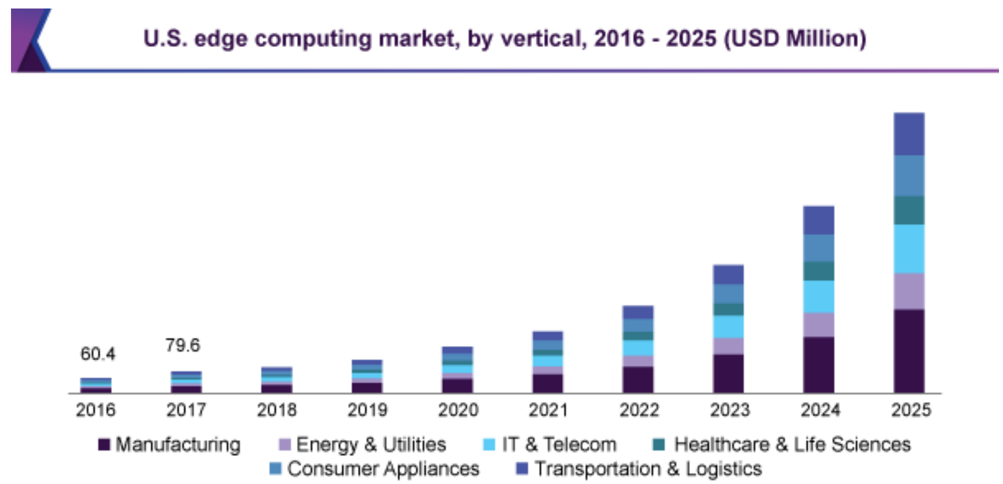
**Jonathan Hinkle**

Executive Director and Distinguished Researcher

Systems Architecture, Lenovo Research

# Why Do More at the Edge?



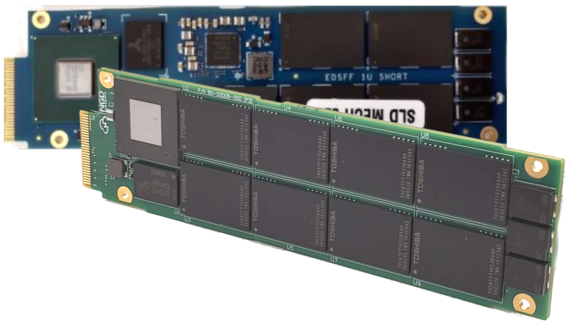

**Expensive (in cost, performance, power) to move all that data to the cloud**

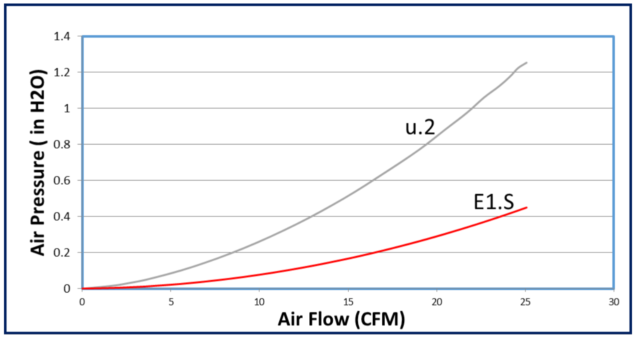
## Some needs are very different from data center

- **Space** – Need **Small** Form factors, Dense, Modular and easy to fit
- **Power** – Low power and energy requirements that are more like mobile, may have battery/energy limitations, especially needs **low power** for thermal reasons

# EDSFF 1U Short (E1.S) drives – well suited for the Edge

Industry Standard datacenter-optimized NVMe drive that provides significant new system benefits



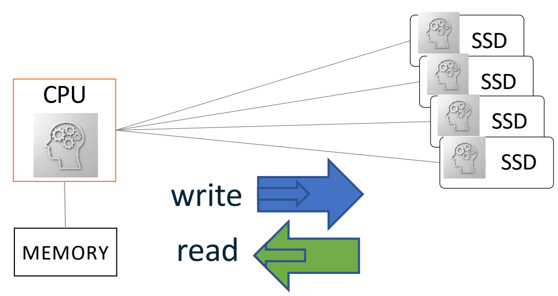

Significantly lower air pressure drop resulting in much better system cooling

- Key benefits:
  - Much **smaller** enabling high **density** storage
  - Significantly improved **system airflow** and thermal solution
  - Most **efficient** modular **scaling** of NVMe capacity and performance Enhanced feature set in **space-constrained** edge systems
  - **Low** base system infrastructure and drive **costs** (high volume, common building block)

# Computational Storage at the Edge

- **Computational Storage aligns well to many key paradigms for Edge computing**
  - Complements system processor to **avoid** requiring **highest power** CPU with additional cooling overhead
  - **Lowers power** required by moving data around less inside the system
  - Allows for insights from data to be developed in **parallel for faster responses**
  - Easier **scaling** of resources as system requirements vary widely between implementations
  - Provides additional compute to process data to **balance** performance as drives being added for storage capacity


EDSFF E1.S Drive with Computational Storage




Data movement to and from storage: reduced

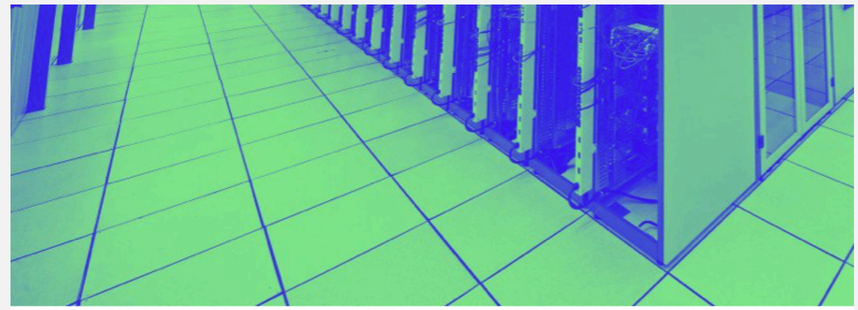
# Machine Learning where Data Resides

Dr. Vladimir Alves

**Computational Storage Takes over for Scale**

# The Case for Computational Storage Machine Learning.

- AI is an an indispensable dimension
- AI performance vs. energy efficiency
- Machine Learning needs DATA
- Where is data kept?
- Could ML apps run in-storage?



DEAN MOUHTAROPOULOS | GETTY; EDITED BY MIT TECHNOLOGY REVIEW

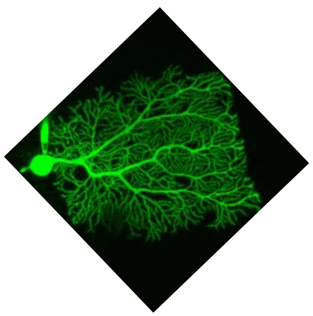
Artificial Intelligence / Machine Learning

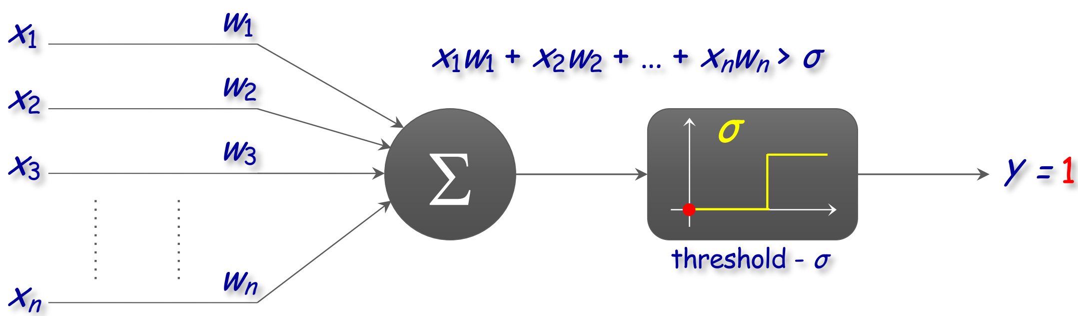
## Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by **Karen Hao** Jun 6, 2019

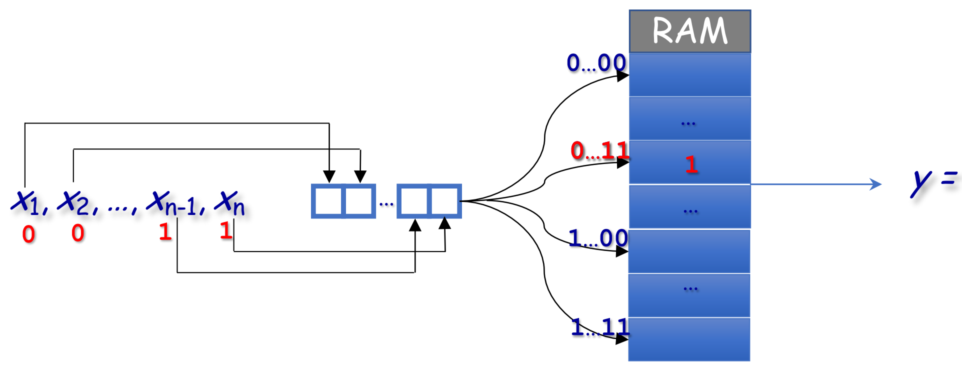
# Neuron Models: Trade-off in Compute vs. Memory.






***McCulloch & Pitts Neuron Model***

A logical calculus of the ideas immanent in nervous activity
*McCulloch and Pitts, 1943*




***WiSARD*** **Weightless Artificial Neural Network**

*N*-tuple sampling machine
*Bledsoe and Browning, 1959*

Universal logic circuit
*Aleksander, 1966*

Analog veto operation
*Boycott and Wässle, 1974*

# The Drive for Artificial Weightless Neural Networks.

NGD systems

**Speed**

Both training and inference are extremely simple. Less client time, processing and energy are used.

*Financial credit analysis via a clustering weightless neural classifier - Cardoso et al., Journal Neurocomputing, Jun 2015*

**Parallelism**

The model has multiple independent components which can be easily parallelized.

*WiSARD-based multi-term memory framework for online tracking objects - Nascimento et al., European Symposium on Artificial Neural Networks, Apr 2015*

**Accuracy**

Parallel training is similar to a parallel sum, which means that no training is lost when using the federated approach.

*Multilingual part-of-speech tagging with weightless neural networks - Carneiro et al., Journal Neurocomputing, Feb 2015*

**Small footprint**

Lightweight hardware thanks to an advantageous memory vs. compute trade-off.

*Design of Robust, High-Entropy Strong PUFs via Weightless Neural Network - Araújo et al., Journal of Hardware and Systems Security, Aug 2019*

# Weightless Neural Networks Used for Object Tracking.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

## A WiSARD-based multi-term memory framework for online tracking of objects

Daniel N. do Nascimento[1], Rafael L. de Carvalho[1,3], Félix Mora-Camino[4], Priscila V. M. Lima[2], Felipe M. G. França[1] *

1 – COPPE, 2 – NCE, Universidade Federal do Rio de Janeiro, BRAZIL
3 - Universidade Federal do Tocantins, UFT, BRAZIL
4 - Ecole Nationale de l'Aviation Civile - Laboratoire d'Automatique, FRANCE

**Abstract**. In this paper it is proposed a generic object tracker with real-time performance. The proposed tracker is inspired on the hierarchical short-term and medium-term memories for which patterns are stored as discriminators of a WiSARD weightless neural network. This approach is evaluated through benchmark video sequences published by Babenko et al. Experiments show that the WiSARD-based approach outperforms most of the previous results in the literature, with respect to the same dataset.

NGD systems

## Weightless Neural Networks Do Care.

MMI facial expression database

# Facial Emotion Classification

- CNN: 99.6% accuracy (Bucker et al.)
- WiSARD: 99.4% accuracy (Lusquino et al.)

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

### Near-optimal facial emotion classification using a WiSARD-based weightless system

Leopoldo A.D. Lusquino Filho[1], Felipe M.G. França[1] and Priscila M.V. Lima[2] *

1- PESC/COPPE 2- NCE
Universidade Federal do Rio de Janeiro – Brazil

**Abstract**. The recognition of facial expressions through the use of a WiSARD-based $n$-tuple classifier is explored in this work. The competitiveness of this weightless neural network is tested in the specific challenge of identifying emotions from photos of faces, limited to the six basic emotions described in the seminal work of Ekman and Friesen (1977) on identification of facial expressions. Current state-of-the-art for this problem uses a convolutional neural network (CNN), with accuracy of 100% and 99.6% in the Cohn-Kanade and MMI datasets, respectively, with the proposed WiSARD-based architecture reaching accuracy of 100% and 99.4% in the same datasets.

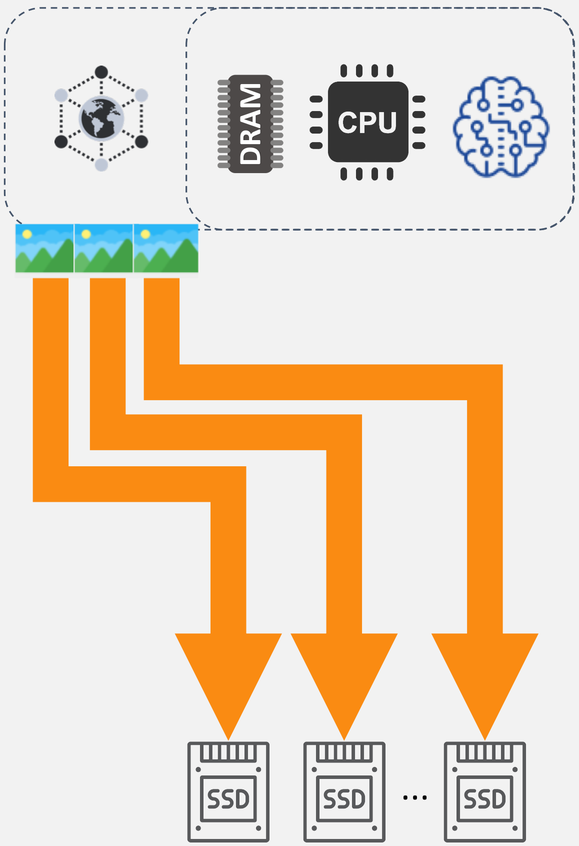
NGD systems

## Moving Beyond Traditional Models.

- Parallel & distributed Training in Computational Storage
- Federated/Transfer Learning
- Reduce data transfers by sending sparse model updates

What Next?

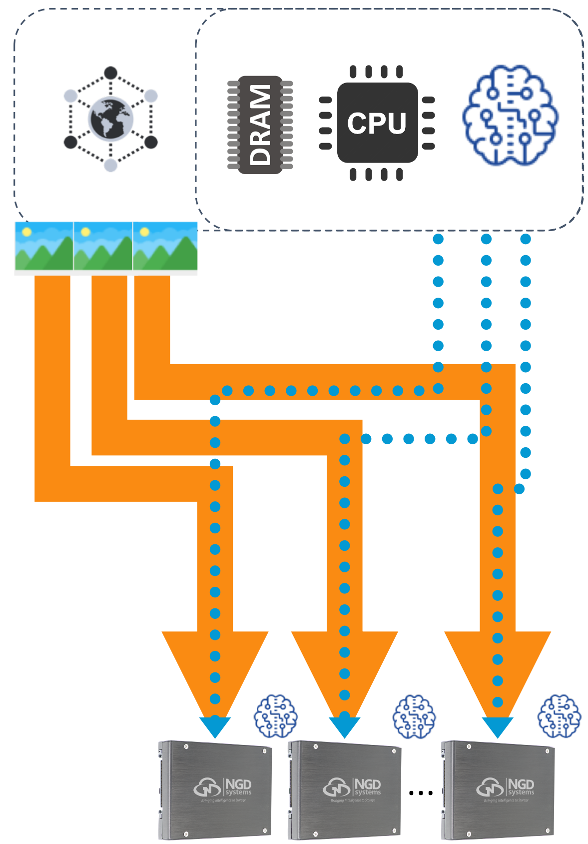
# ML Training with Traditional Approach.



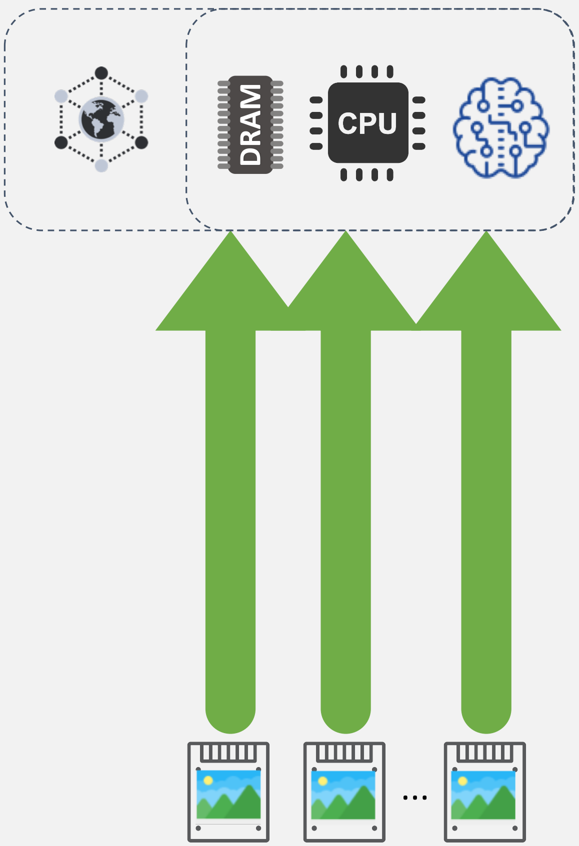

Load Data

# ML Training with Computational Storage.

NGD systems



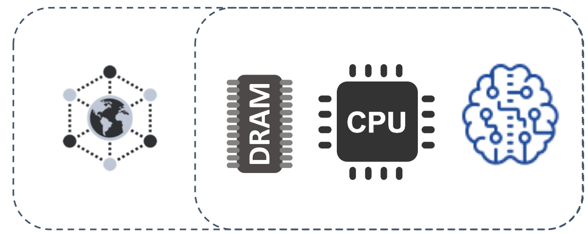
# ML Training with Traditional Approach.




Load Data

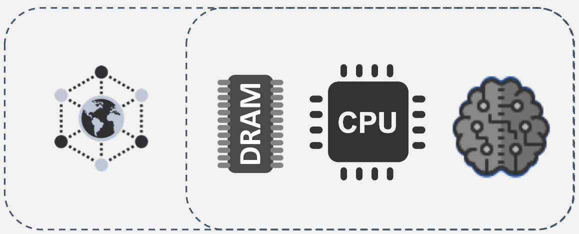
Train

# ML Training with Computational Storage.




- No data movement
- No host CPU needed
- Distributed training

## ML Training with Traditional Approach.




- Host CPU still needed
- No Parallelism


Load Data

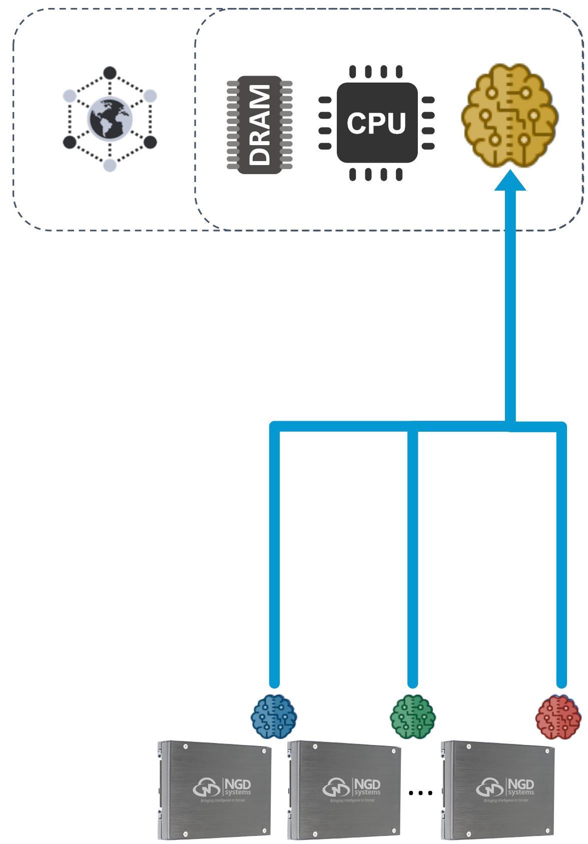
Train

Evaluate

Update

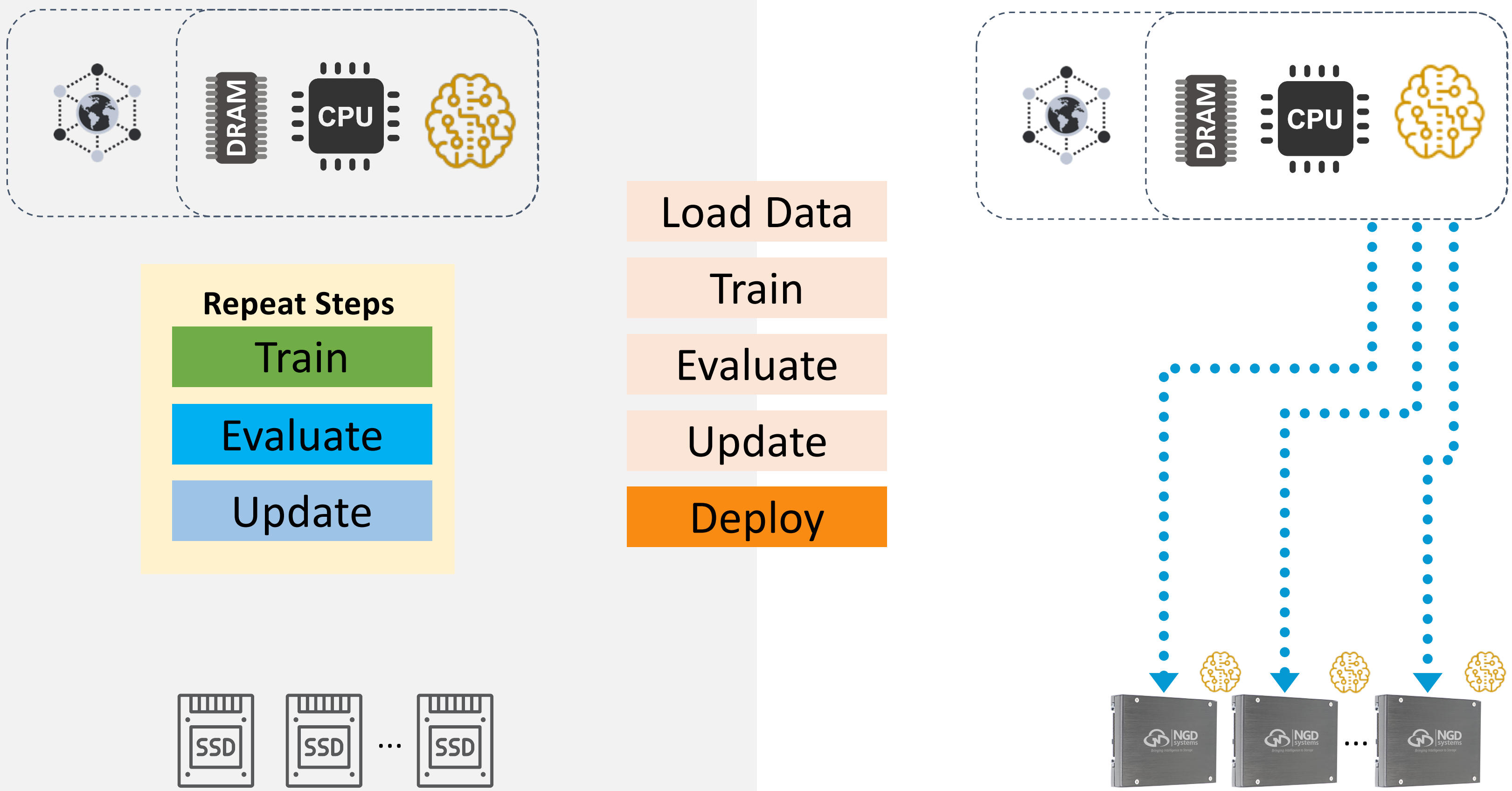
## ML Training with Computational Storage.

NGD systems



# ML Training with Traditional Storage.

# ML Training with Computational Storage.



# Federated/Transfer Learning.

## MNIST DATASET

60,000 samples

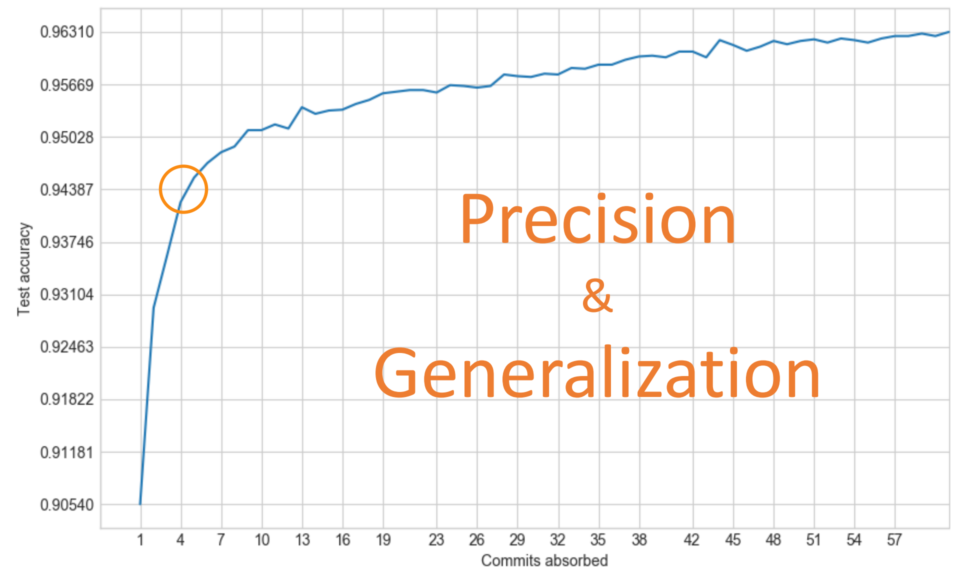
From the training set

61 updates

Model updates transferred

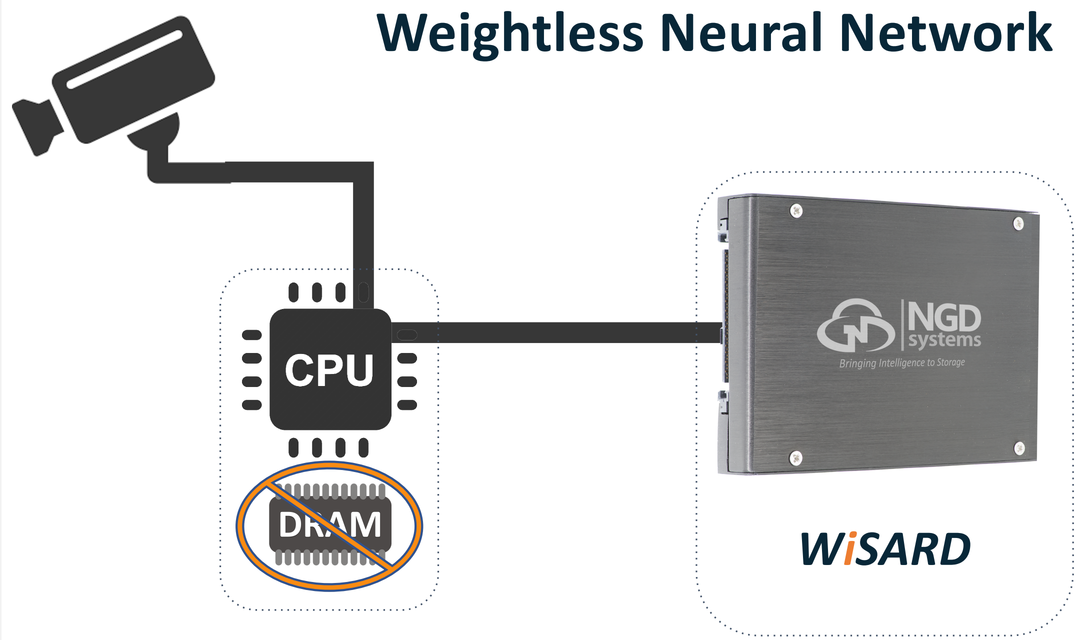
94% accuracy

With only 4 partial model updates

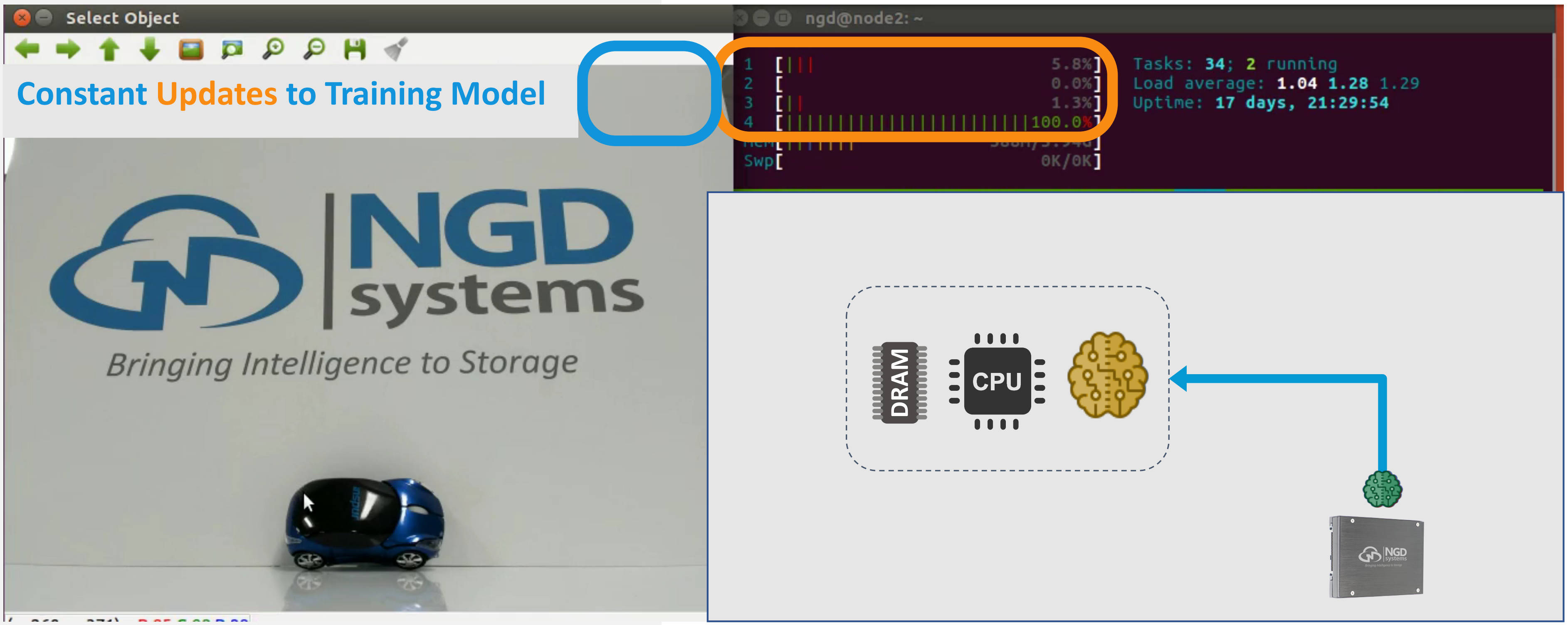


# Using Computational Storage Drives for ML.

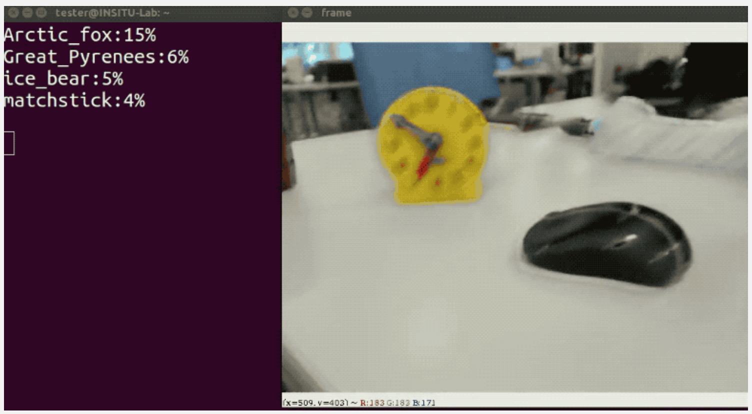
## Object Tracker.



# Using Computational Storage Drives for ML.

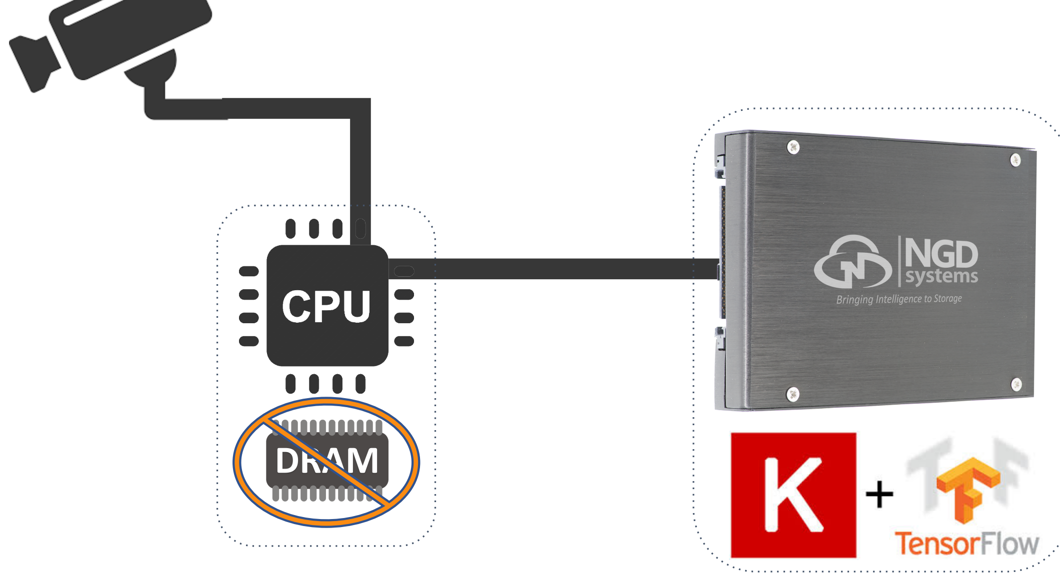


# Using Computational Storage Drives for ML.

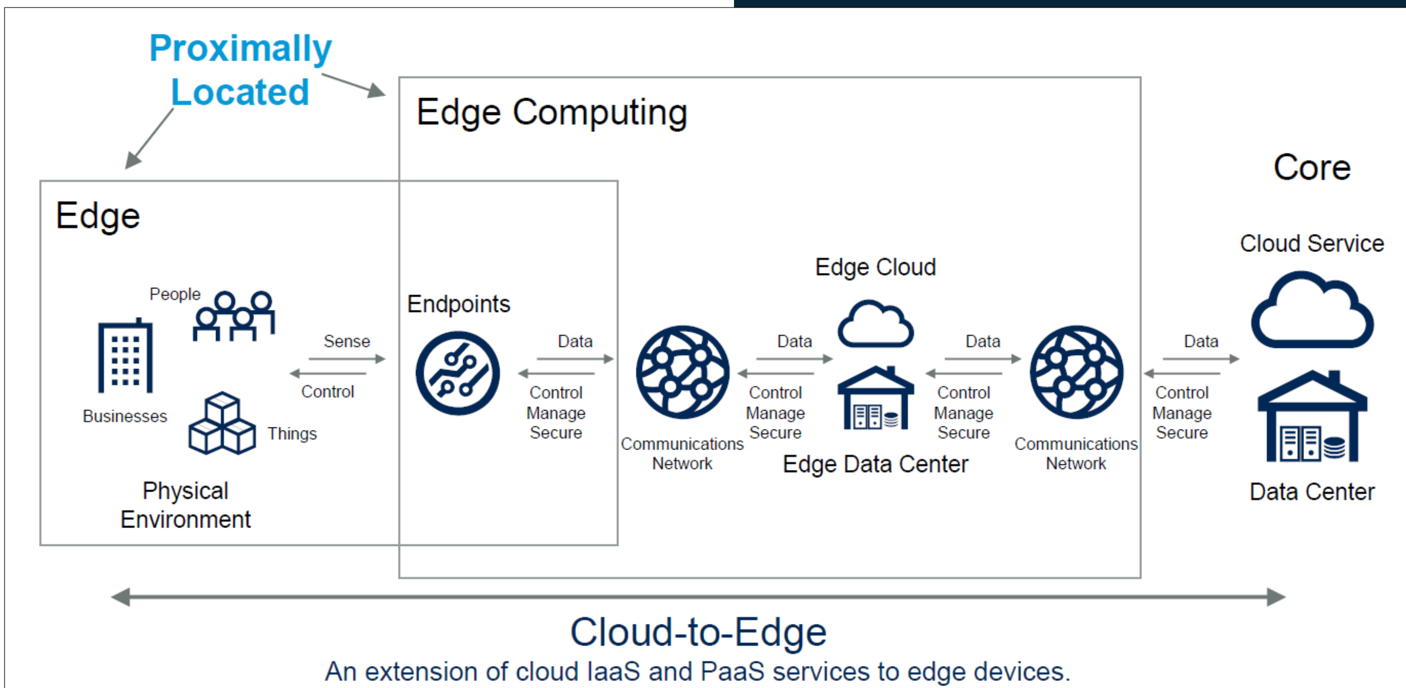



## MobileNetV2.

Keras + TensorFlow



# What Are You Doing with Your Data Today?




Source: Gartner - Bittman

It's No Longer Black and White.

# Scalable Computational Storage.

## A New Storage Paradigm is Here


- **The "New Cloud" needs the Distributed Edge**
  - There is no longer just a 'central' storage location
- **Edge data growth challenges HW platforms**
  - Innovative form factors and high capacity for the Edge
- **In-Situ Processing brings ML closer to data**
  - Exploit data locality and enable distributed processing

IDC Innovator

NGD systems

Gartner Cool Vendor 2018

Located in Booth 618

Live Demos of Computational Storage

Eli Tiomkin on Thursday in COMP-301B

FLASH MEMORY SUMMIT
BEST OF SHOW

# NGD systems

Gartner
Cool Vendor 2018

# World Leader in NVMe Computational Storage

More than Just Your Average SSD