HPC-AI ADVISORY COUNCIL

NETWORK OF EXPERTISE

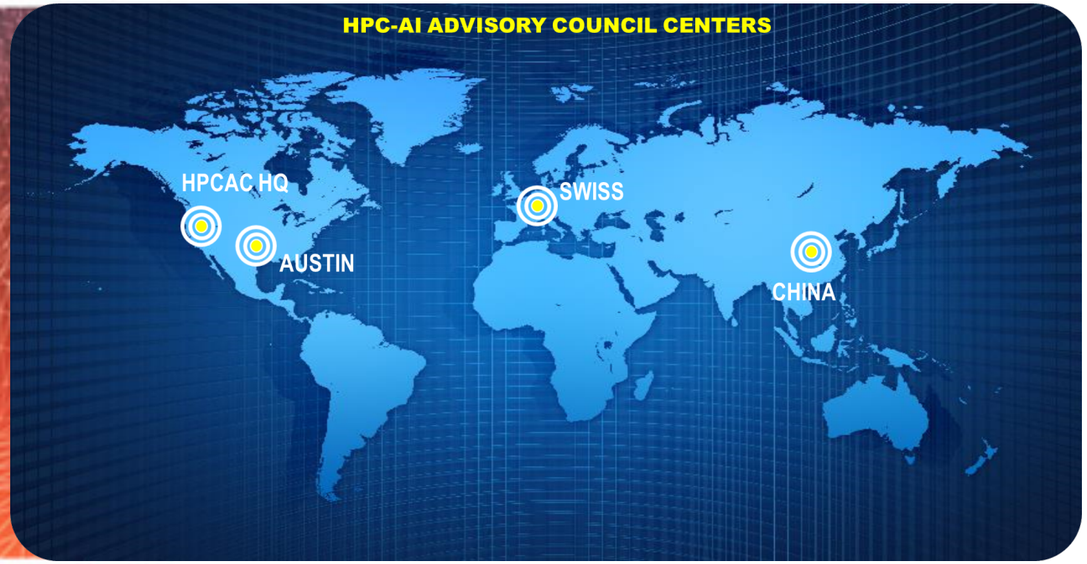
# Intelligent Data Center Architecture to Enable Next Generation HPC/AI Platforms

Gilad Shainer, HPC-AI Advisory Council

SCAsia 2018 Supercomputing

# The HPC-AI Advisory Council

- World-wide HPC-AI non-profit organization
- More than 400 member companies / universities / research centers
- Bridges the gap between HPC-AI usage and its potential
- Provides best practices and a support/development center
- Explores future technologies and future developments
- Leading edge solutions and technology demonstrations

# HPC-AI Advisory Council Members

HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

# HPC-AI Advisory Council Activities

- **Applications Best Practices**
  - Hundreds of cases published
- **Cluster Center and Advanced Technology center**
- **2019 Conferences**
  - USA (Stanford University) – February
  - Switzerland (Swiss Supercomputing Center) – April
  - Australia (Pawsey Supercomputing Center) – August
  - Spain (Barcelona Supercomputing Center) – Sep
  - UK (University of Leicester, DiRAC) – Sep
  - China (HPC China)
- **2019 Competitions**
  - APAC HPC-AI Competition – March
  - China - 7th Annual RDMA Competition – May
  - ISC Germany - 8th Annual Student Cluster Competition – June
- **For more information**
  - www.hpcadvisorycouncil.com
  - info@hpcadvisorycouncil.com

# Data as a Resource

20th Century


21st Century

# World of Data – World of Opportunities

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

| | | | | |
|---|---|---|---|---|
| >70% of data is generated by consumers | 80% of data stored | 3% structured | 0.5% being analyzed | <0.5% generates information |

# The Power of Data – for Everyone



NOKIA
Connecting People

NAVTEQ

- ~5M traffic sensors for $8.1B

waze

- 50M traffic sensors for free

# Data Processing Plant – Convert Data to Information



facebook | Google

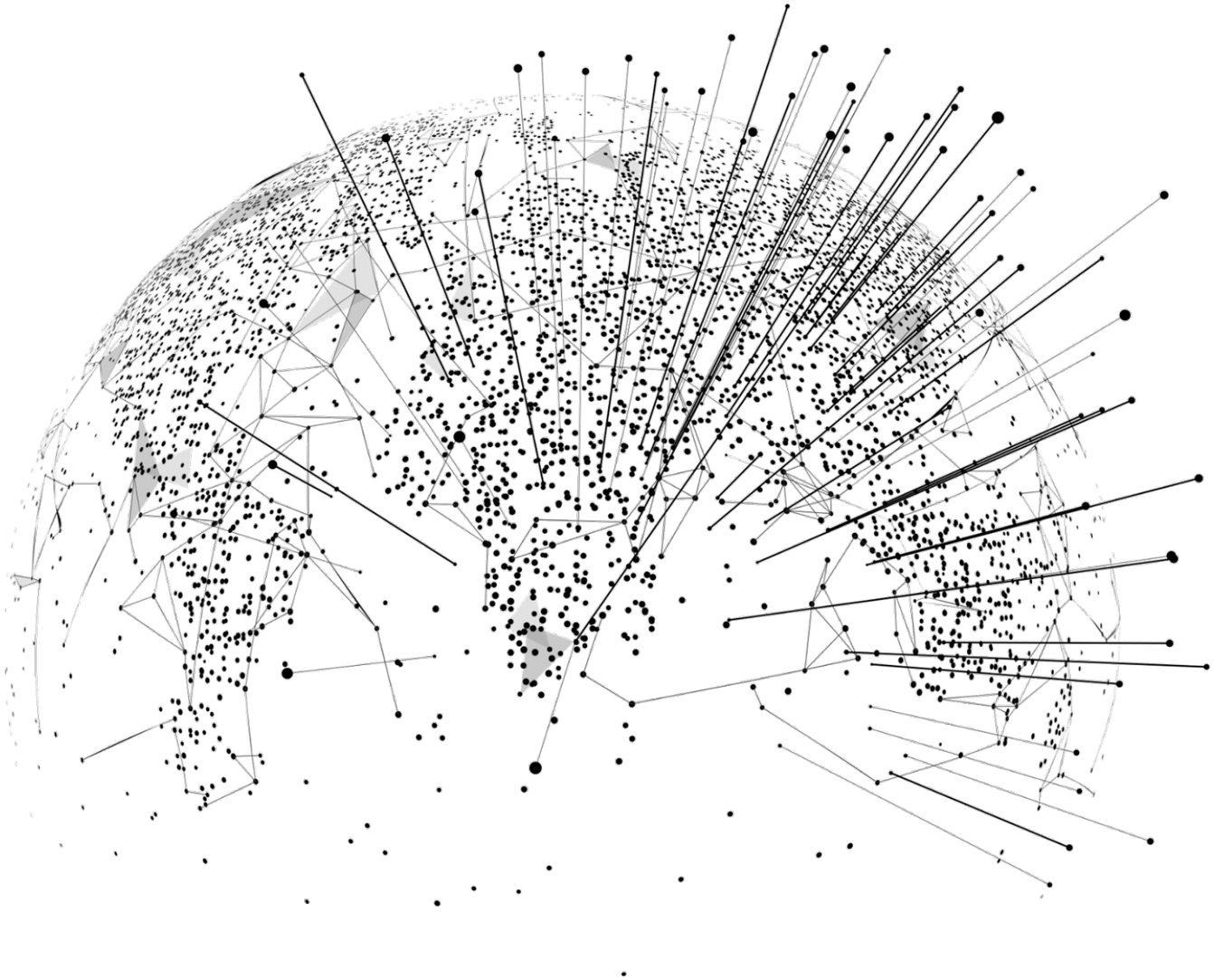
twitter | Windows Azure

Alibaba.com | Baidu 百度

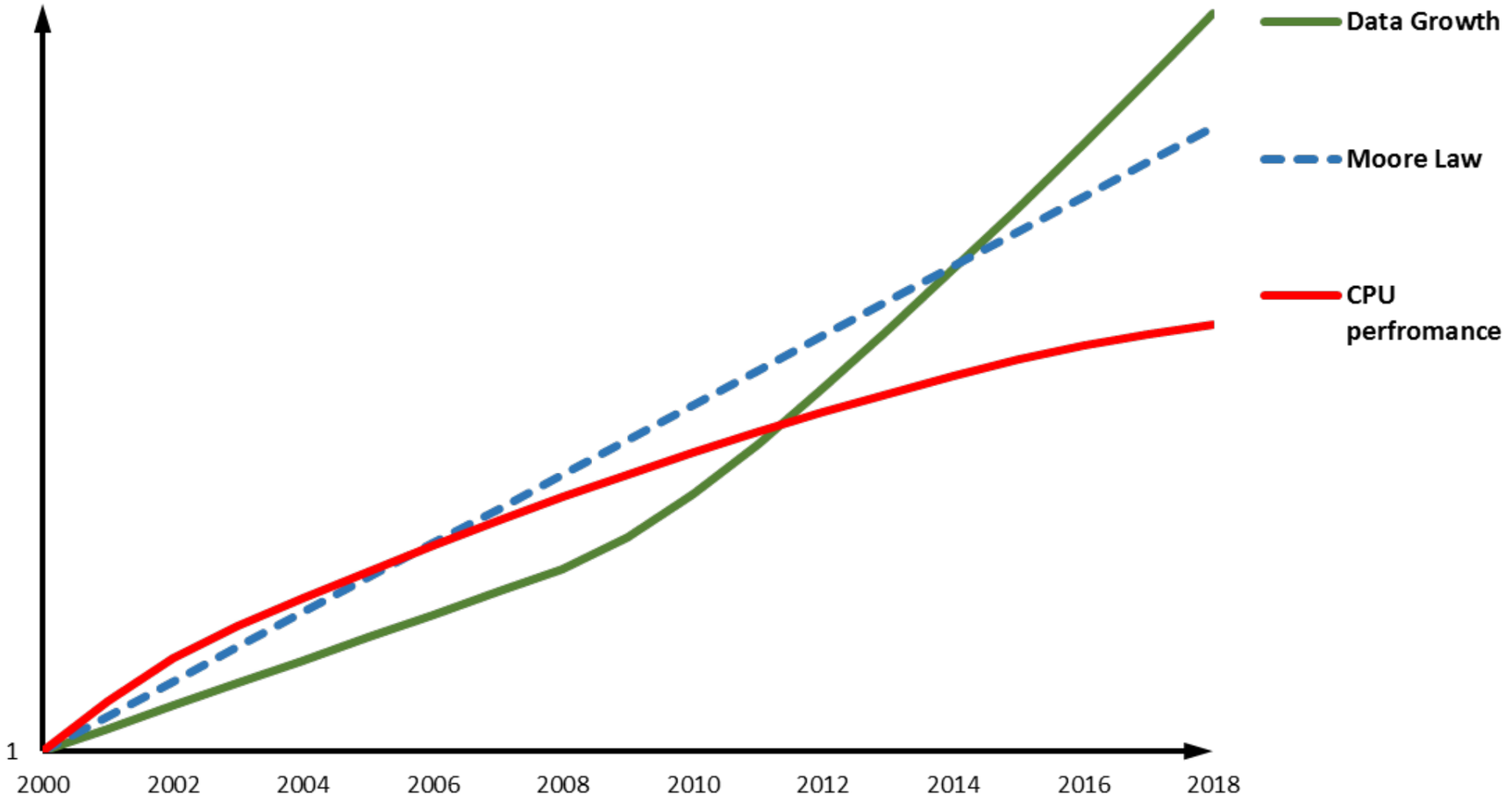
amazon | ebay

iCloud | Yandex

PayPal | waze

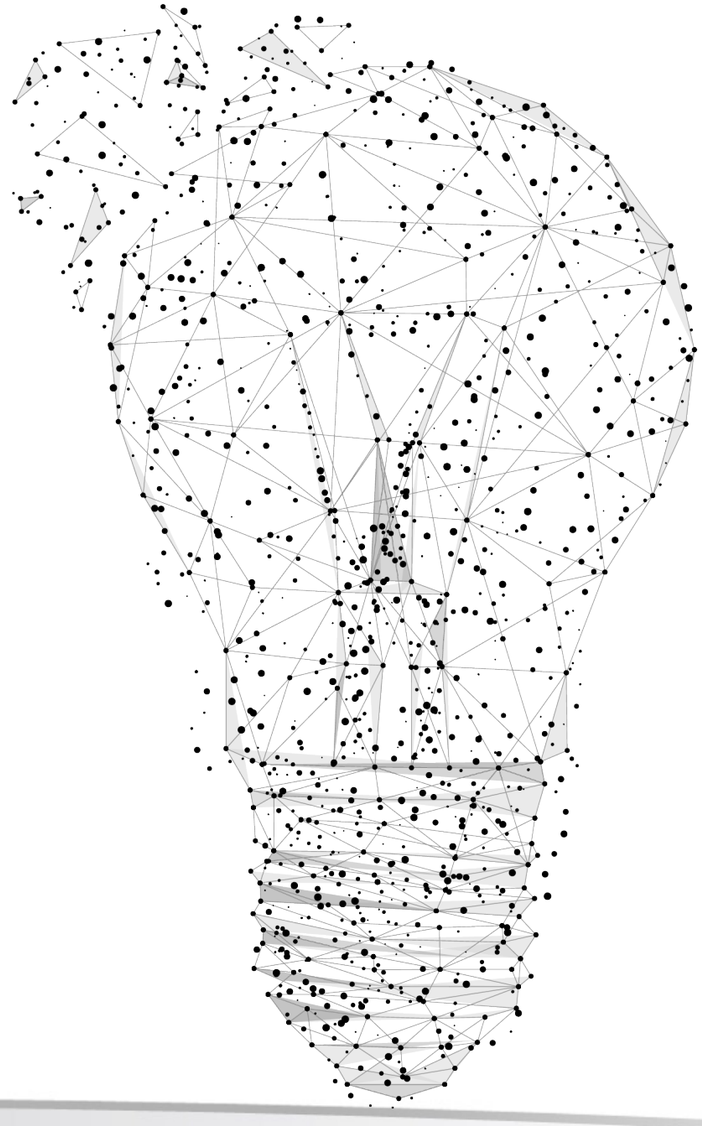


# Computing Challenge



# Breakthrough Needed

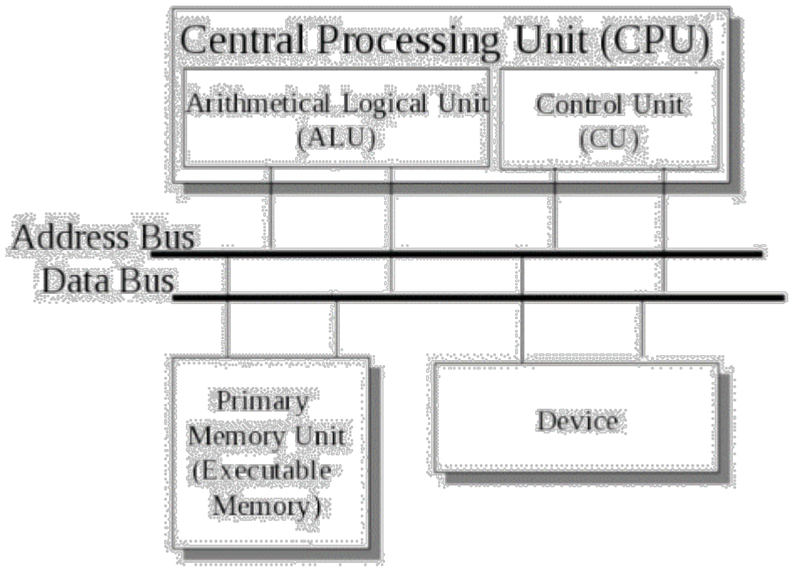
"The Electric Light
Did Not Come From
Continuous Improvement
Of Candles"



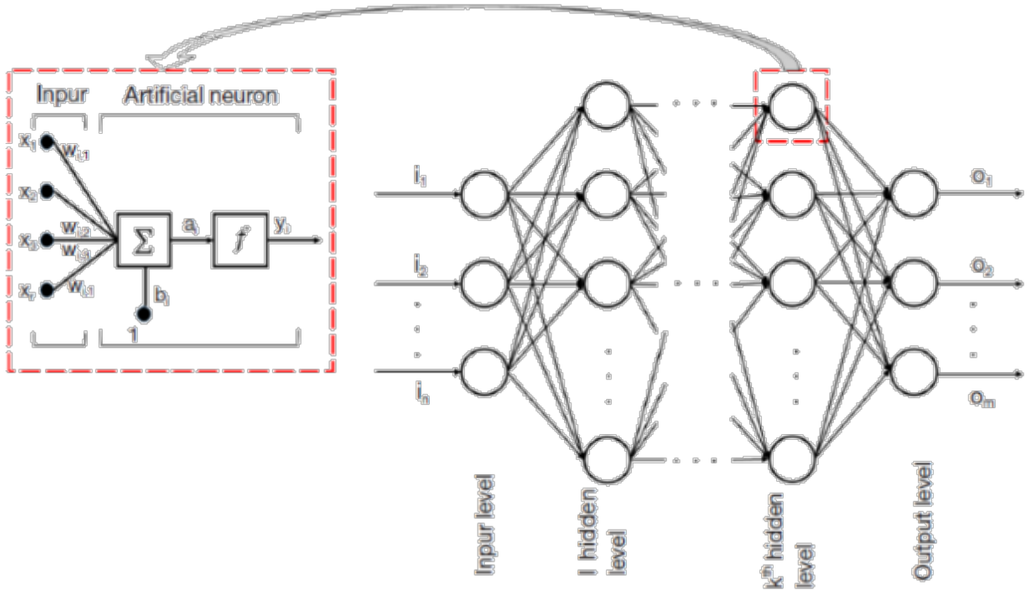
# Data Processing Revolution



## Compute-Centric




## Data-Centric



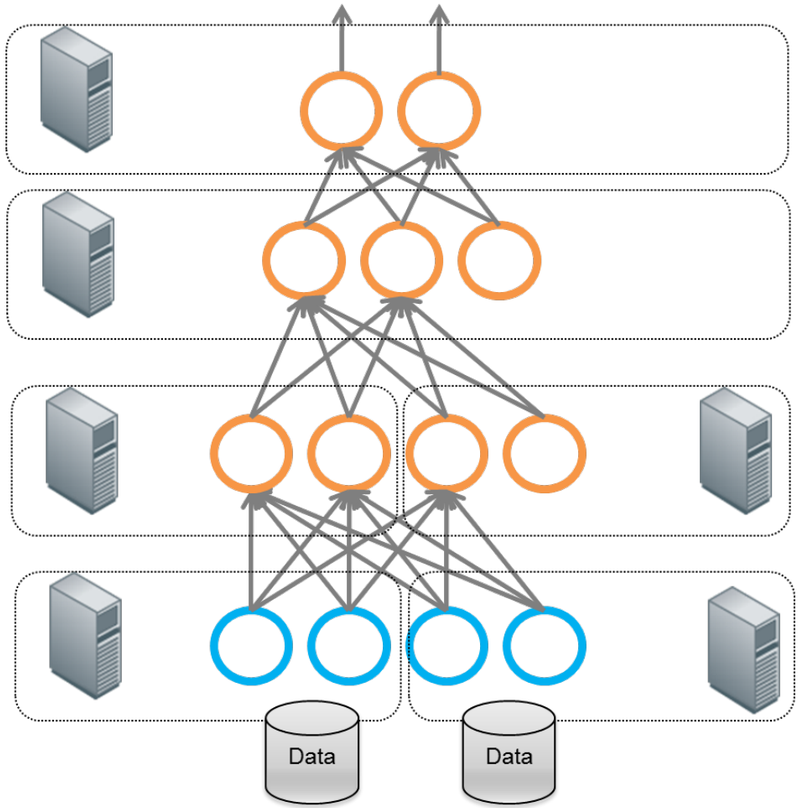

Von Neumann Machine

DataFlow Machine

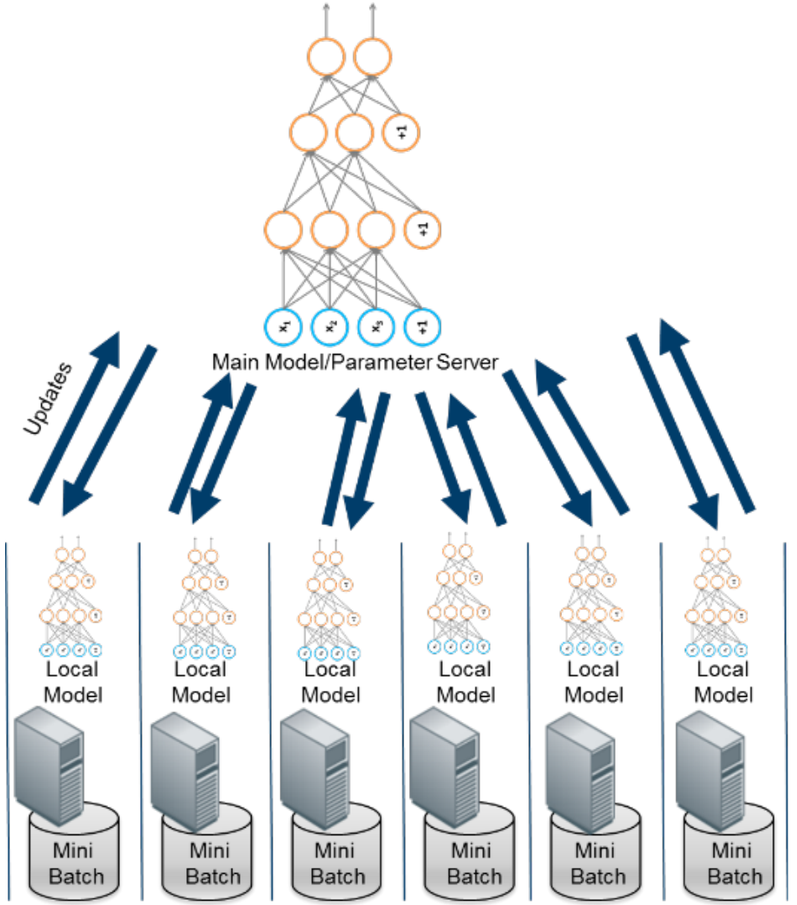
# Deep Learning Model and Data Parallelism



## Model Parallelism




## Distributed Training



# High Performance Computing and Artificial Intelligence



OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

JÜLICH FORSCHUNGSZENTRUM

**High Performance Computing**

**Artificial Intelligence**

**Same Technology**

AIST
NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

NASA

UNIVERSITY OF TORONTO

# The Ever Growing Demand for Higher Performance



Terascale

Petascale

1st

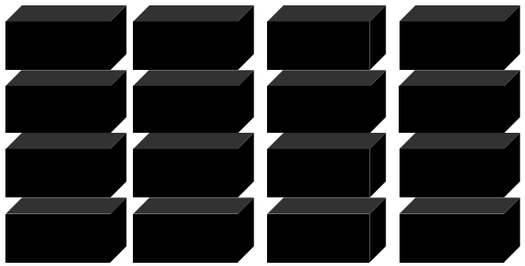
"Roadrunner"

Exascale

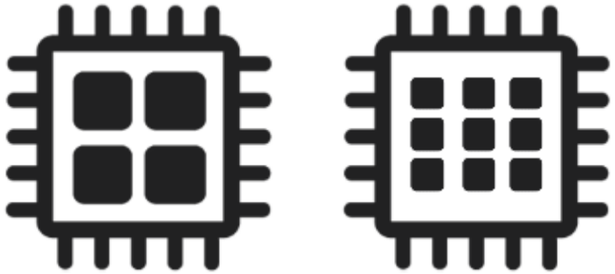
OAK RIDGE National Laboratory
"Summit" System

Lawrence Livermore National Laboratory
"Sierra" System

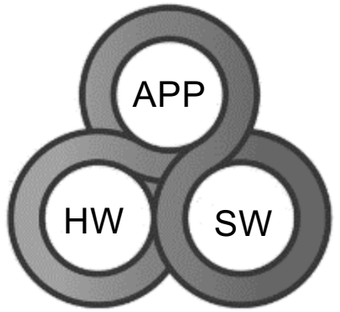
2000 2005 2010 2015 2020



SMP to Clusters



Single-Core to Many-Core



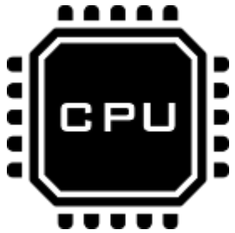

Application
Software
Hardware

Co-Design

# From CPU-Centric to Data-Centric Data Centers



Everything




**CPU**

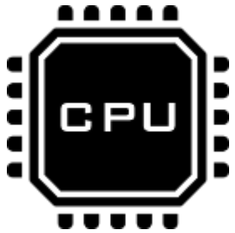
**Network**

**Storage**

# From CPU-Centric to Data-Centric Data Centers



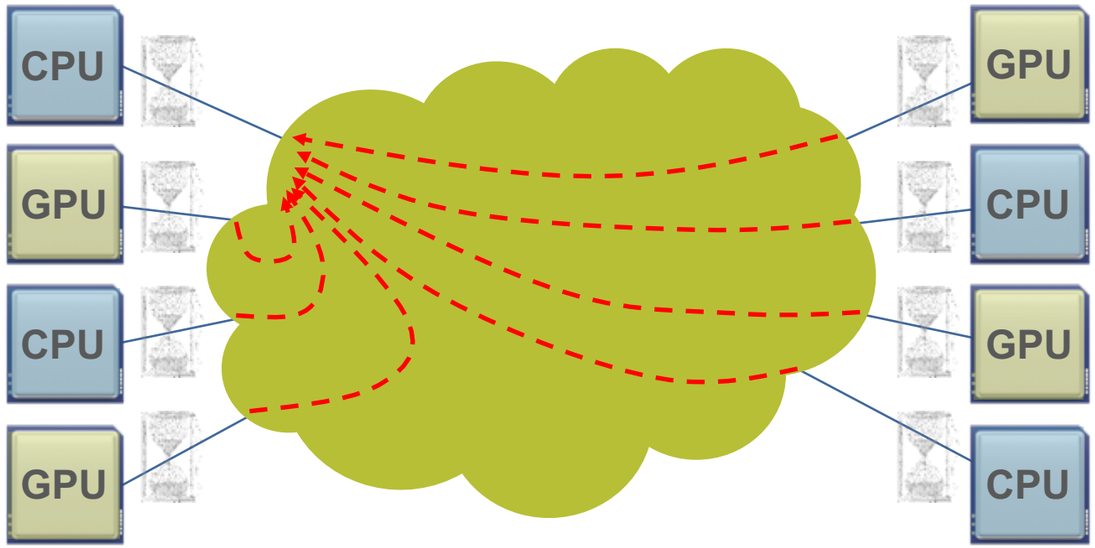
| In-CPU Computing | In-Network Computing | In-Storage Computing |
|---|---|---|
| Workload | Workload | Workload |
| | Communication Framework (MPI) | |
| CPU Functions | Network Functions | Storage Functions |




# In-Network Computing

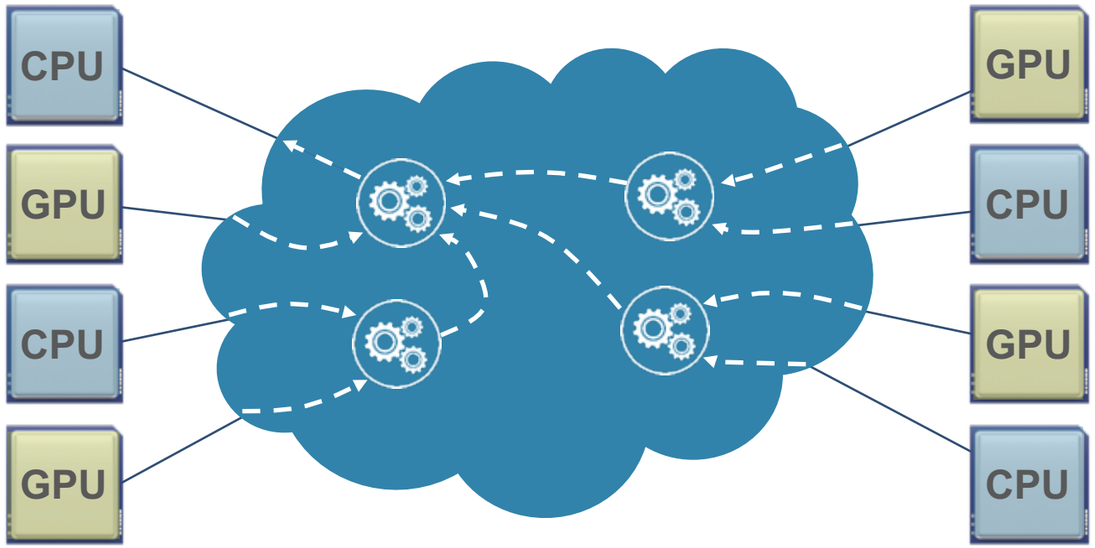


## CPU-Centric (Onload)




Communications Latencies of 30-40us

## Data-Centric (Offload)




Communications Latencies of 3-4us

# The Roadmap of In-Network Computing



**Application Framework**

- Data Analysis

**Communication Framework**

- Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)
- MPI Engines
- Software Defined Virtual Devices

MPI

UCF Unified Communication Framework

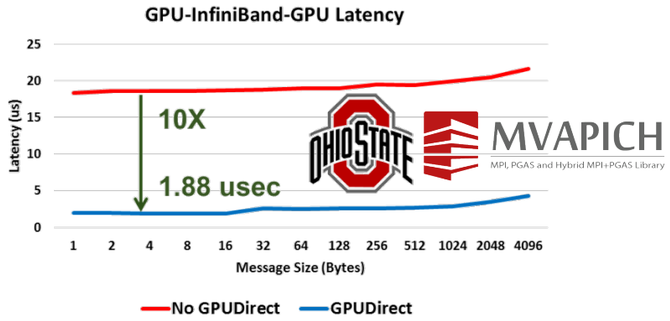
**Network Framework**

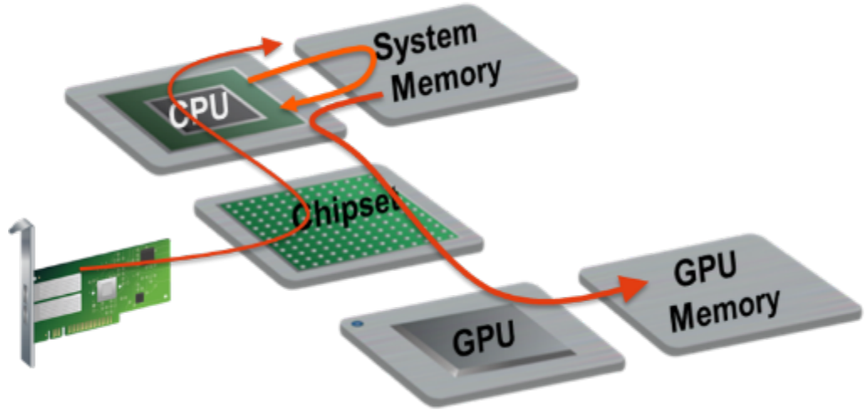
- Network Transport Offload
- RDMA and GPU-Direct
- Self-Healing Network

GPUDirect

RDMA

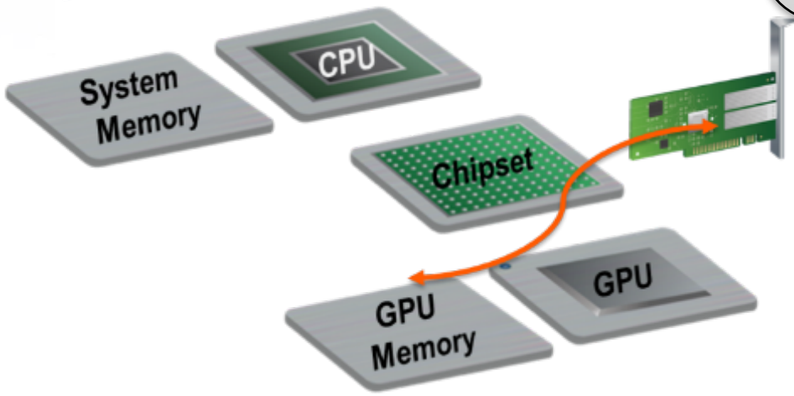
# GPUDirect RDMA Technology and Advantages



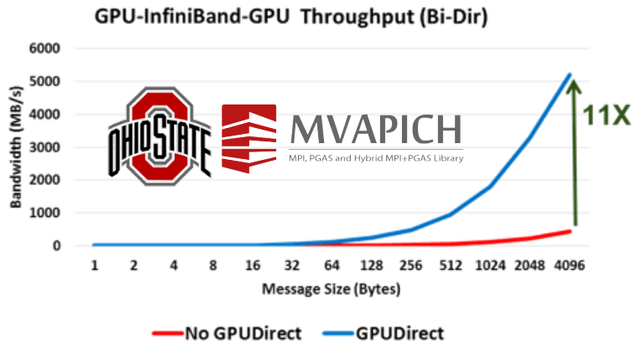




RDMA Enabled Network

**With GPUDirect**

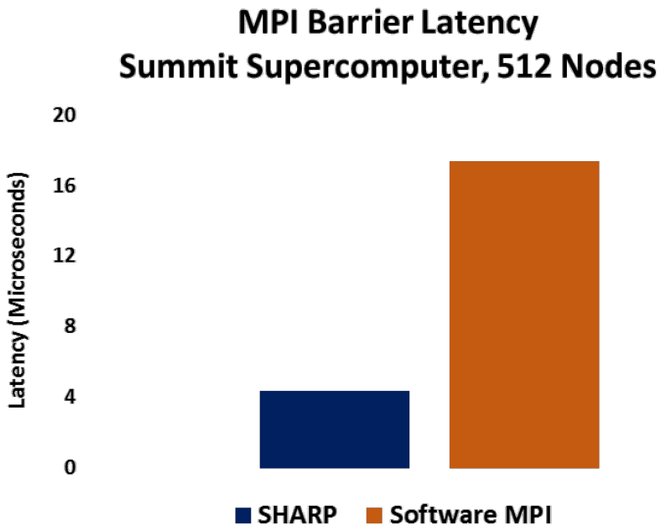



**Without GPU Direct**
**- Same Data Copied 3 Times**



# Scalable Hierarchical Aggregation and Reduction Protocol (SHARP™) Performance Advantages



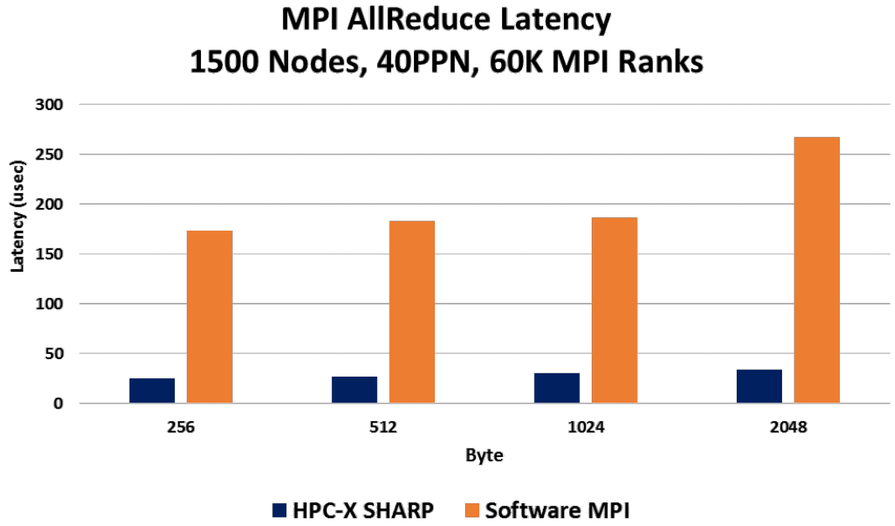



OAK RIDGE National Laboratory

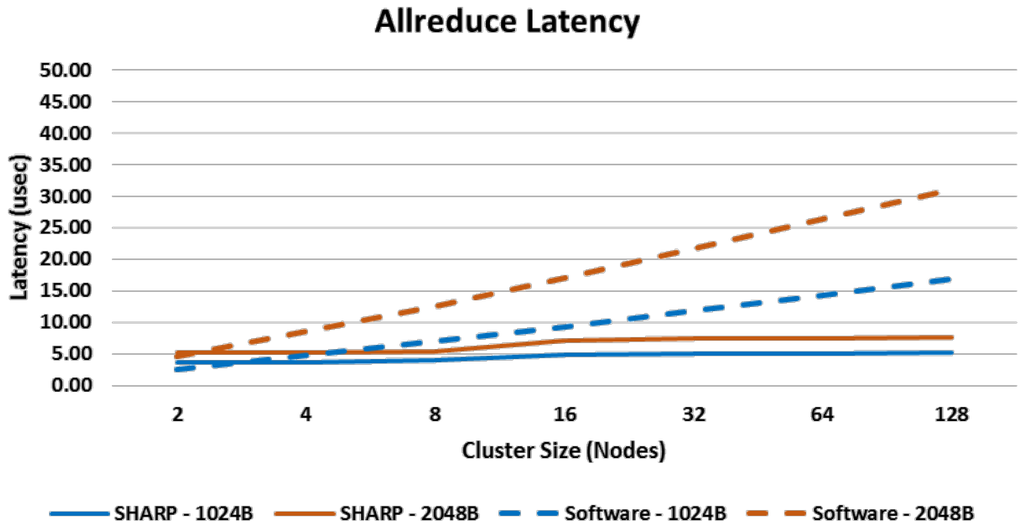
Mellanox TECHNOLOGIES

R&D 100

SHARP
Scalable Hierarchical Aggregation and Reduction Protocol



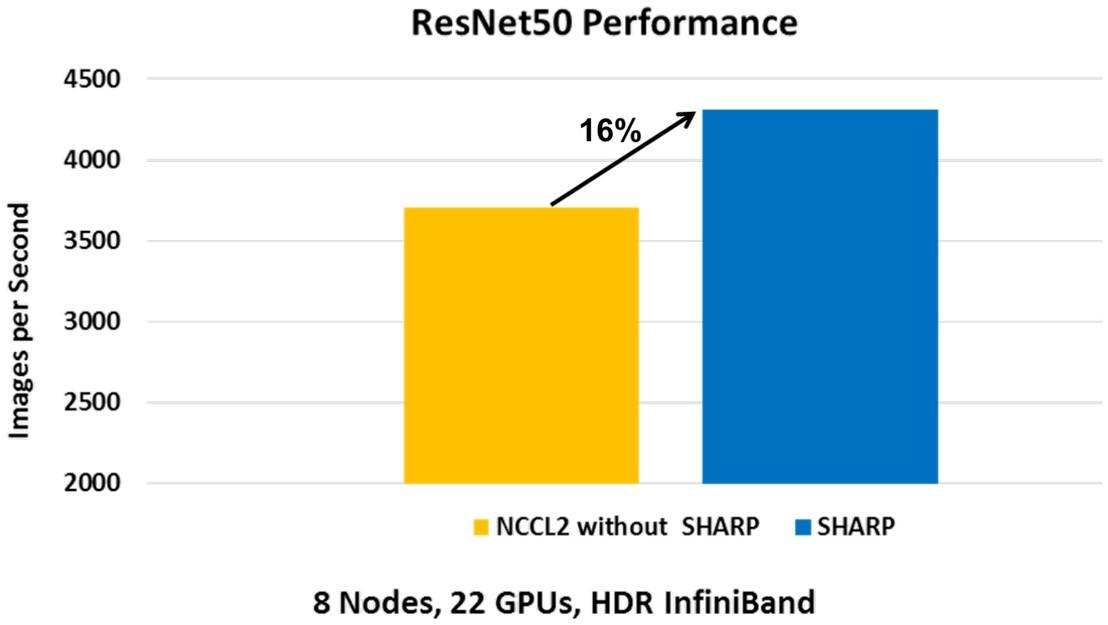

OAK RIDGE National Laboratory

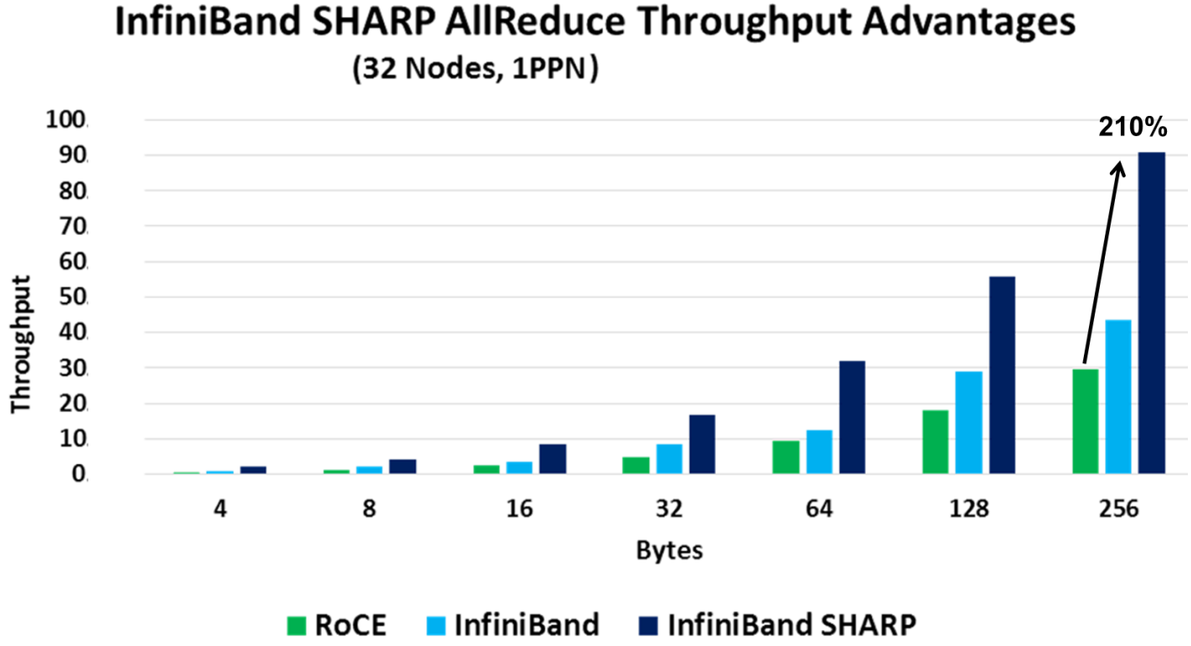




UNIVERSITY OF TORONTO

# SHARP Performance Advantage for Deep Learning

- SHARP provides 16% Performance Increase for deep learning, initial results
- TensorFlow with Horovod running ResNet50 benchmark, HDR InfiniBand

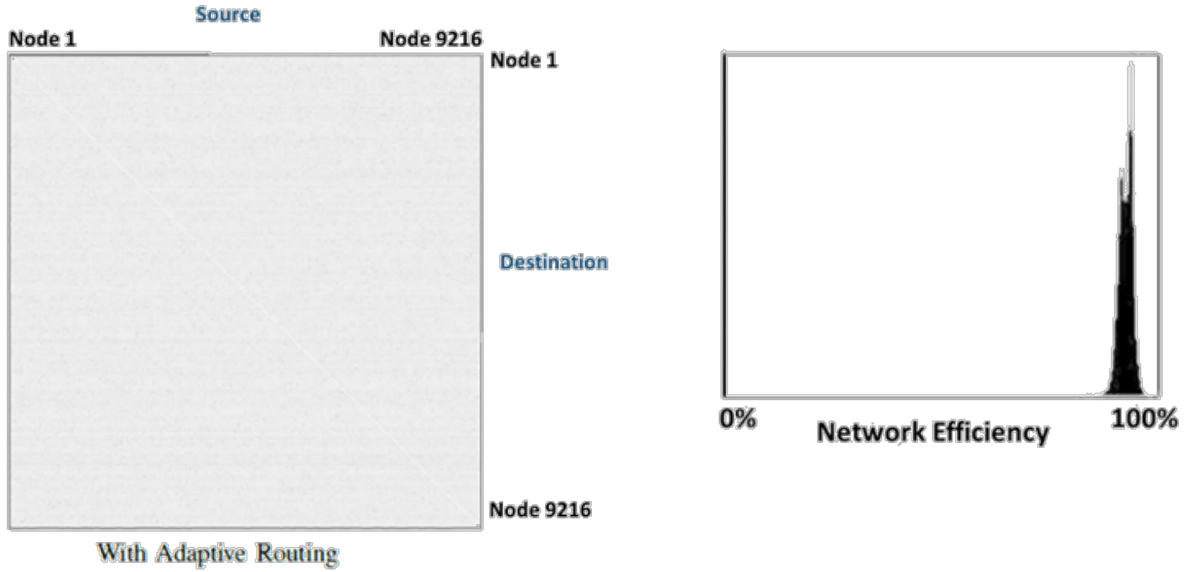


# Smart Routing
## Adaptive Routing Performance – ORNL Summit

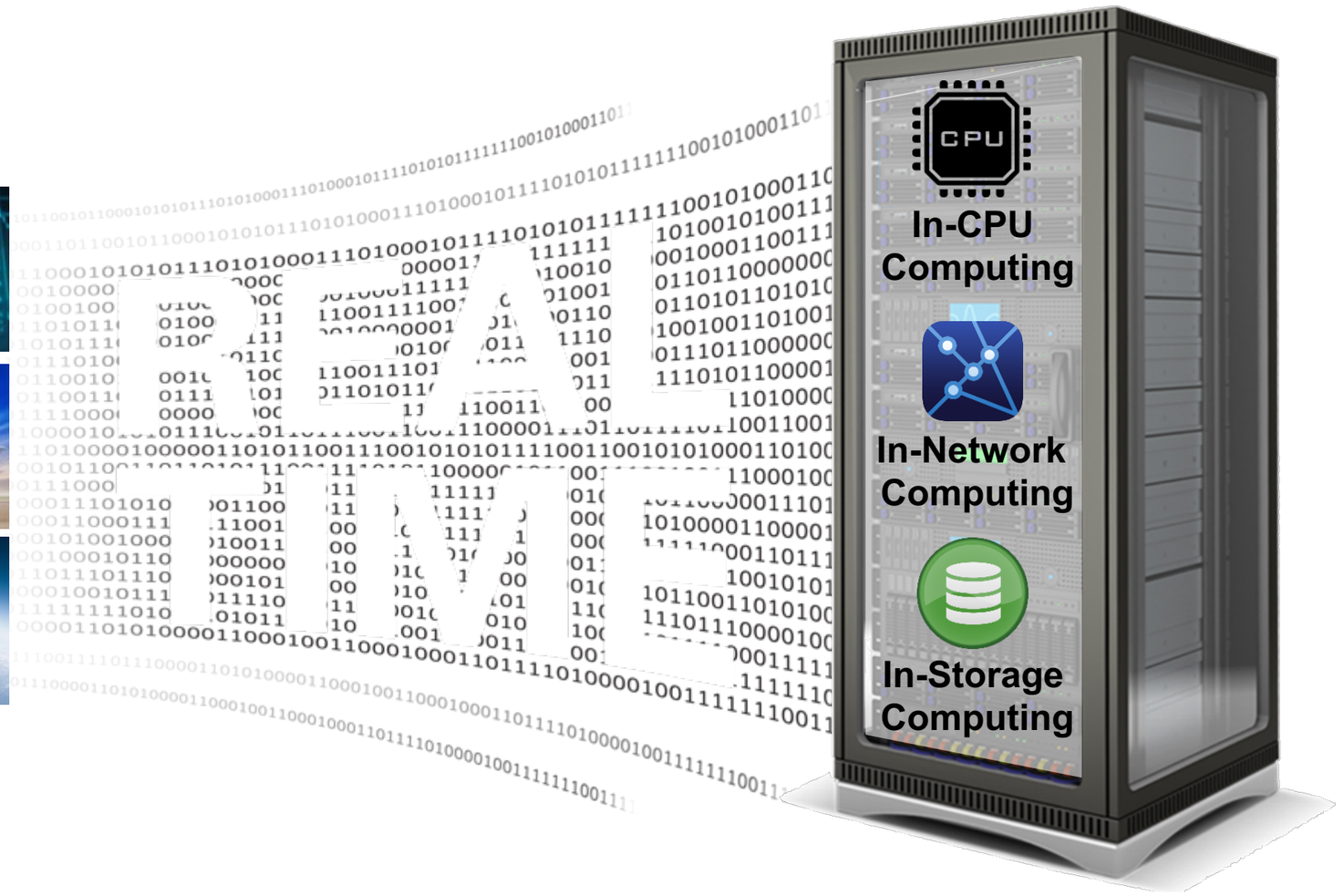
- Oak Ridge National Laboratory – Coral Summit supercomputer
- Bisection bandwidth benchmark, based on mpiGraph
- InfiniBand demonstrates an average performance of 96% network utilization

mpiGraph explores the bandwidth between possible MPI process pairs. In the histograms, the single cluster with adaptive routing indicates that all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion limits bandwidth, negatively impacting overall application performance.




Source: "The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems" paper

# Unleashing the Power of Data

HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

# Thank You