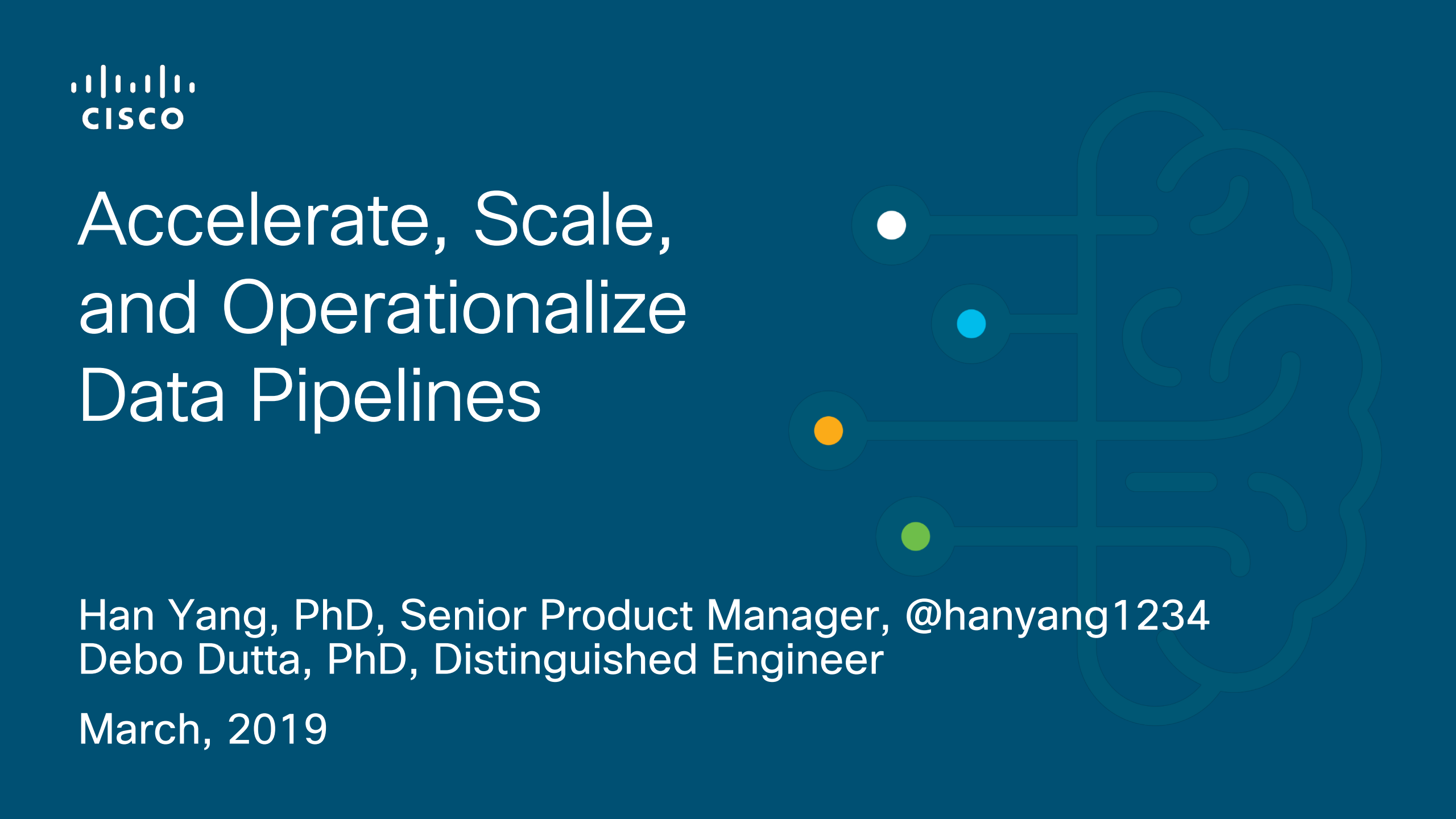


Accelerate, Scale, and Operationalize Data Pipelines

Han Yang, PhD, Senior Product Manager, @hanyang1234
Debo Dutta, PhD, Distinguished Engineer

March, 2019

A Top CXO Priority

2.5 quintillion[1]

bytes of data generated per day in 2017

$3.9 trillion[2]

AI derived business value by 2022

85%[3]

of CIOs will be piloting AI programs by 2020

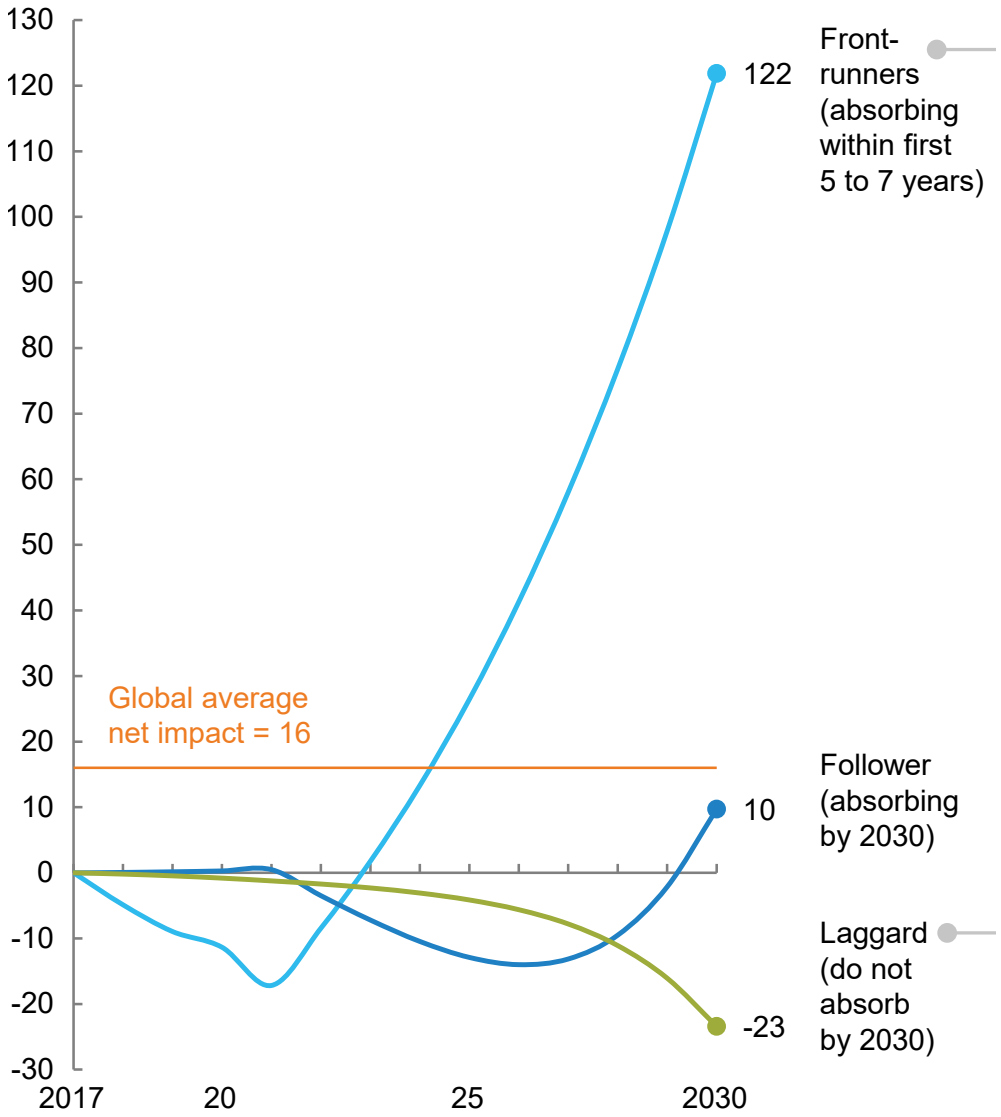
Sources: 1. IBM

Sources: 2. Smarter with Gartner. "2018 Will Mark the Beginning of AI Democratization", ~~Laurence Goasduff~~. Dec 2017, https://www.gartner.com/smarterwithgartner/2018-will-mark-the-beginning-of-ai-democratization/

Sources: 3. Gartner. "Forecast: The Business Value of Artificial Intelligence, Worldwide, 2017-2025."Published 12 March 2018 - ID G00348137

Relative Change in Cash Flow by AI Adoption Cohort

Relative changes in cash flow by AI adoption cohort
% change per cohort, cumulative



NOTE: Numbers are simulated figures to provide directional perspectives rather than forecasts.

SOURCE: McKinsey Global Institute analysis

https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-frontier-modeling-the-impact-of-ai-on-the-world-economy?cid=other-eml-alt-mgi-mck-oth-1809&hlkid=5ebe957bb3594f96bedda5695e4664fd&hctky=10366723&hdpid=677435cb-04b0-445e-afba-4588aa47d2fe

AI Projects and Inquiries Across All Industries

Finance

Fraud Detection

Cryptocurrencies

Algorithmic Trading

Healthcare

Cancer Cell Detection

Drug Discovery

Medical Research

Media and Entertainment

Video Captioning

Content Based Search

NLP, VR and AR

Security and Defense

Face Recognition

Crowd Analytics

Cyber Security

Retail

Theft Detection

Auto Checkout

Targeted Marketing

Manufacturing

Reduce Product Defects

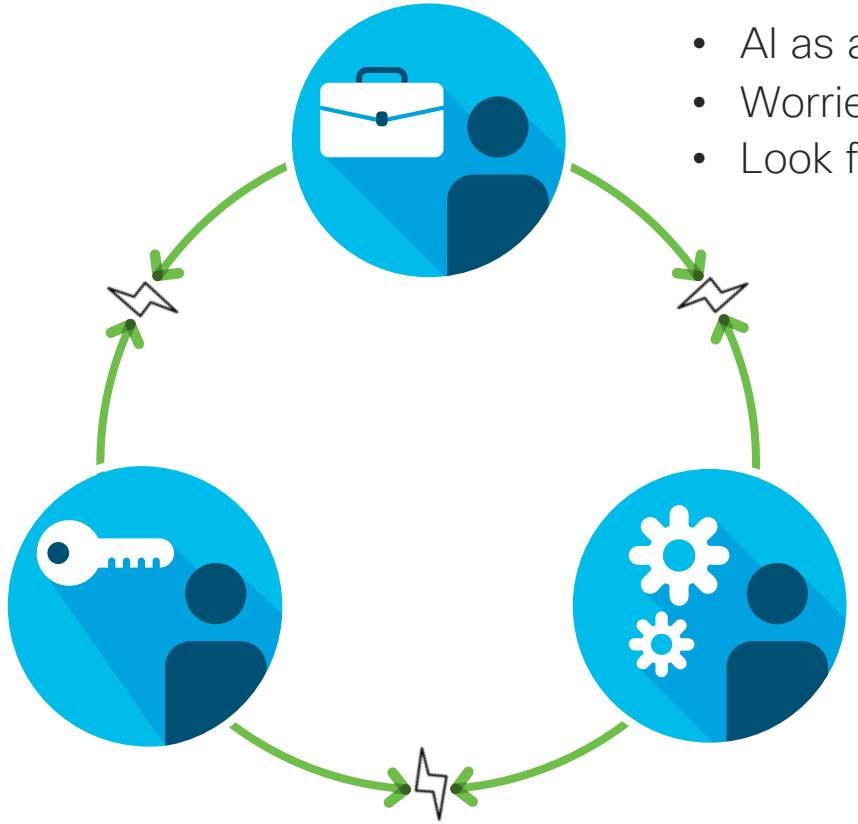
Increase Production Speed

Shorten Downtime

Major gaps exist between key stakeholders in realizing ML impact

CXOs Business Leaders

- AI as a top agenda, wants a strategy
- Worried about falling behind
- Look for use cases, resource and impact



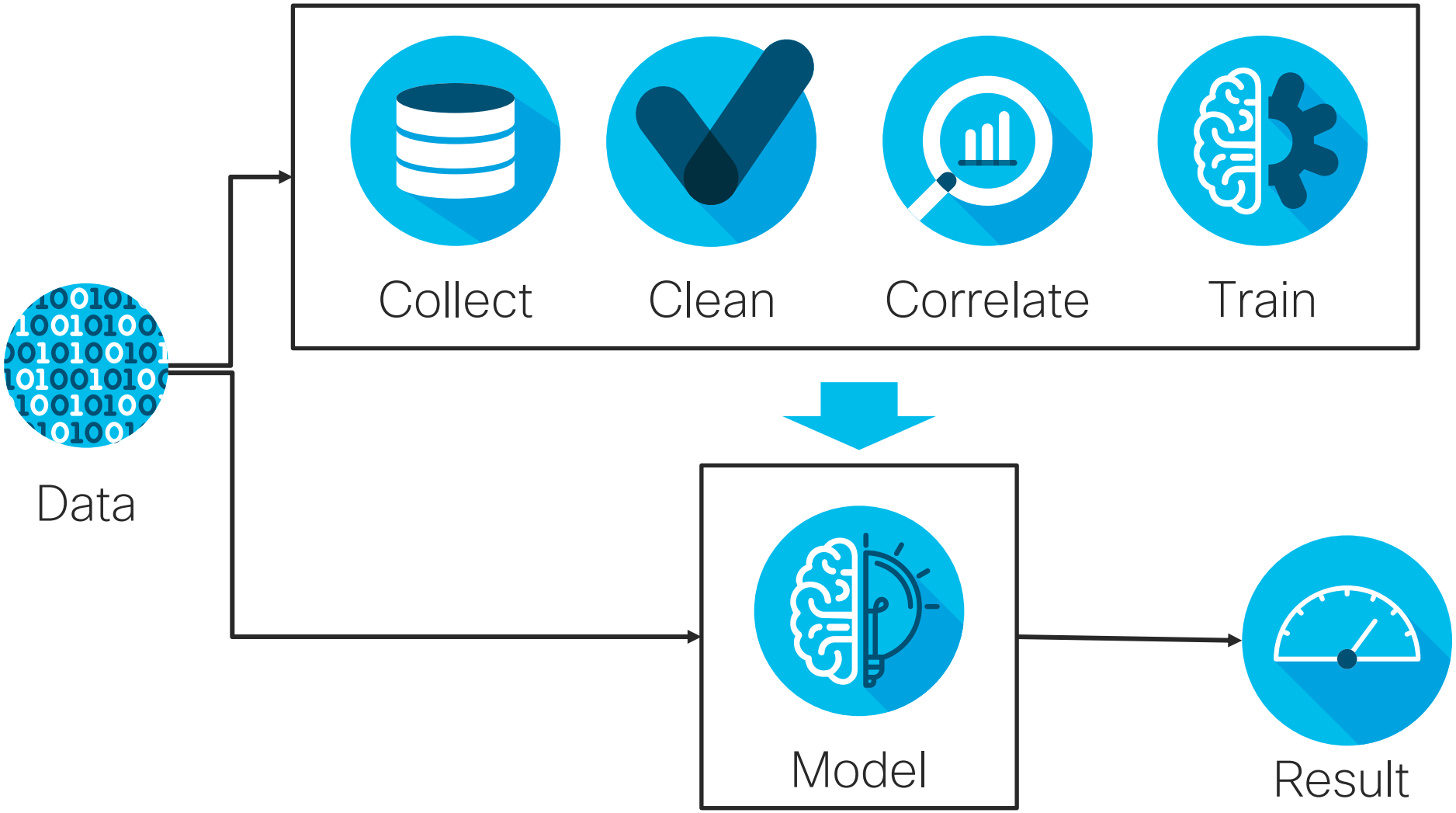
Data Scientists
Data Engineers

- Rapidly evolving open source ML frameworks
- Start with workstation or cloud
- Hard to scale for production

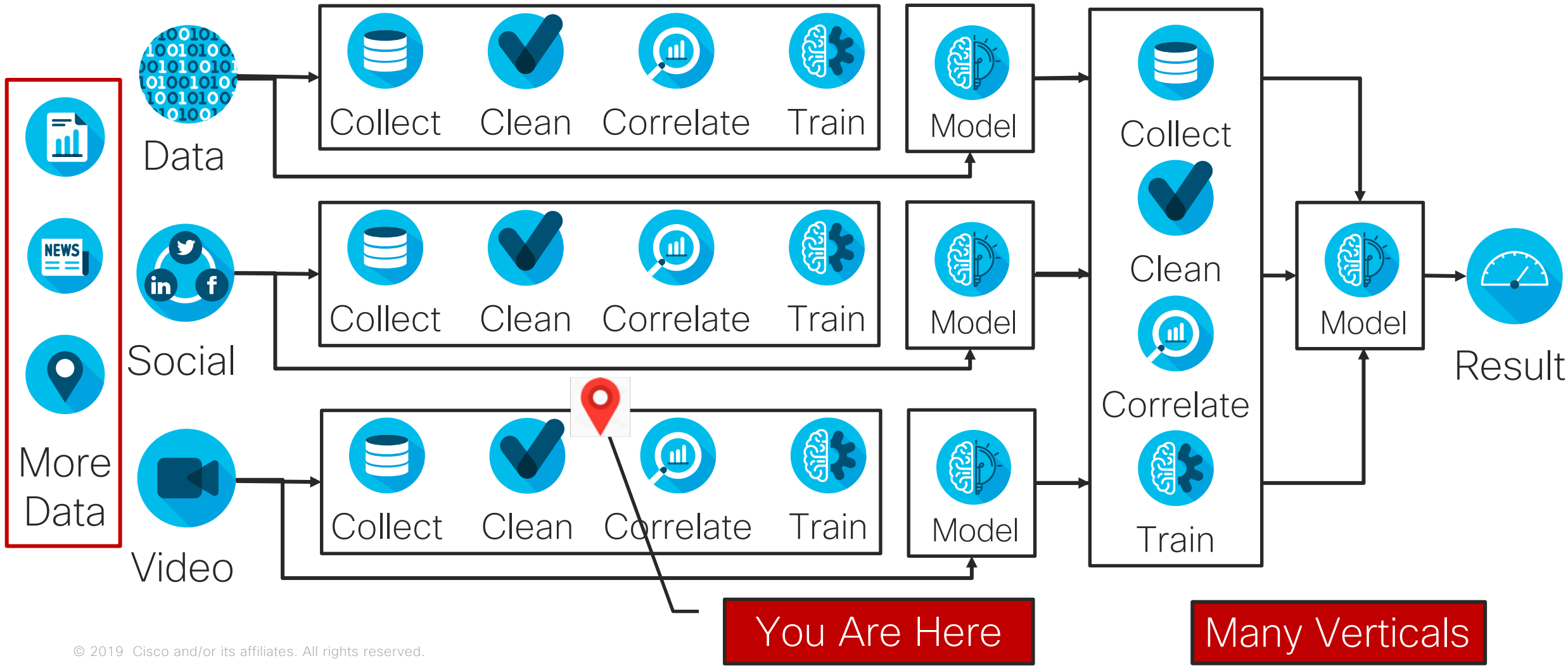
IT Team

- Lack of AI/ML expertise
- Need new infra architecture
- Need to solve silos and manageability
- Need enterprise readiness at scale

Data Pipeline for Single Data Source

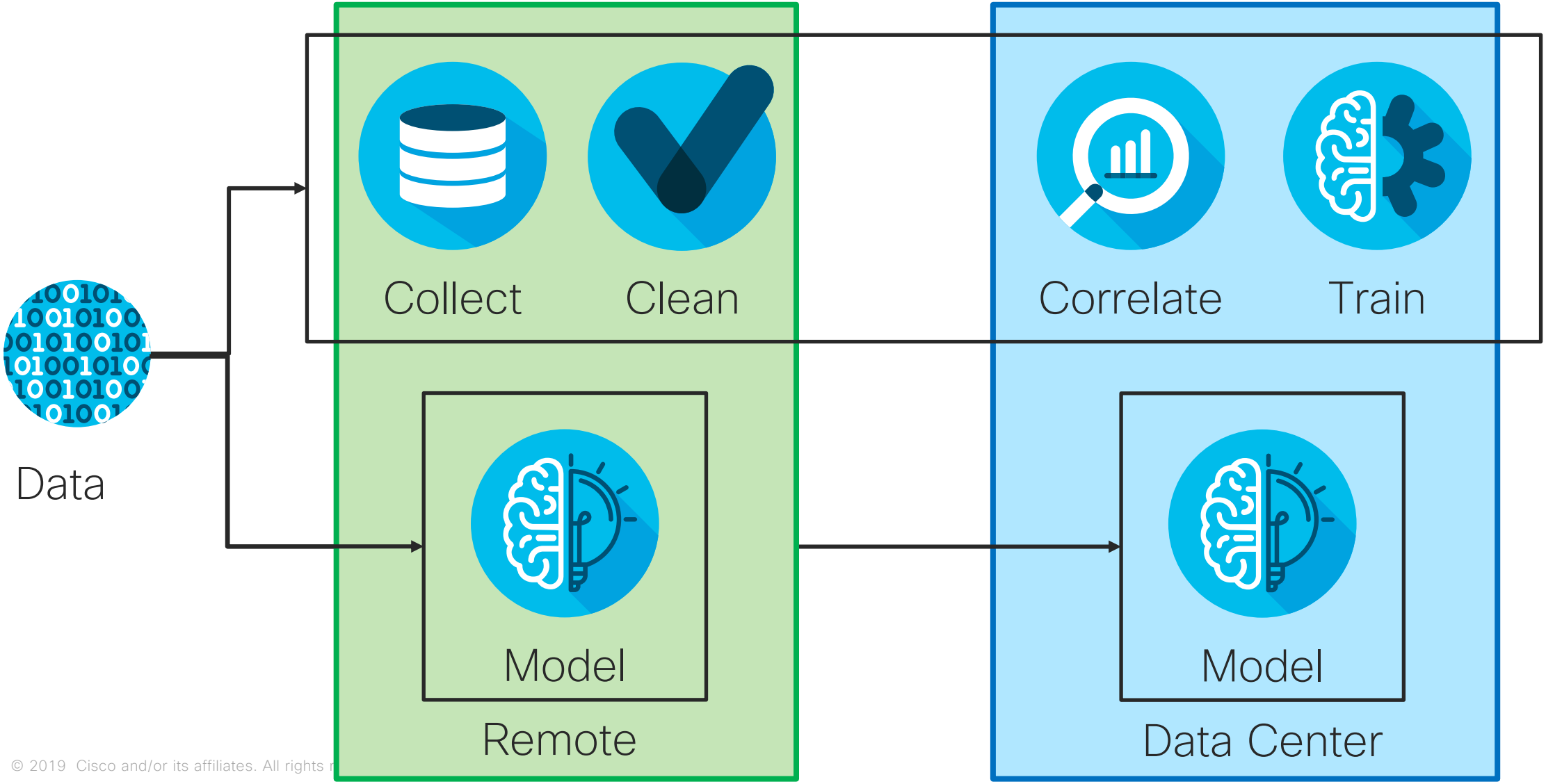


Data Pipeline for Multiple Data Sources



Where are Your Data Sources?

Remote and Data Center Inferencing



Cloud-Powered Systems Management

Centralized Management

Global Policies

Intuitive Experience

Enhanced Support

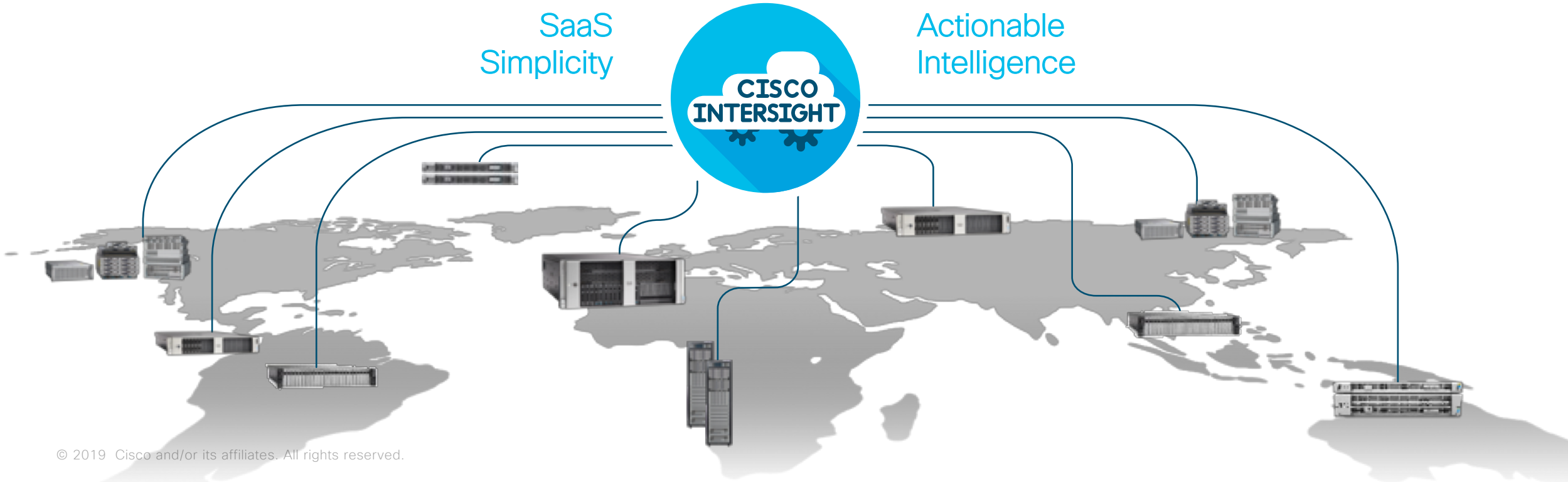
Proactive Guidance

Secure and Extensible

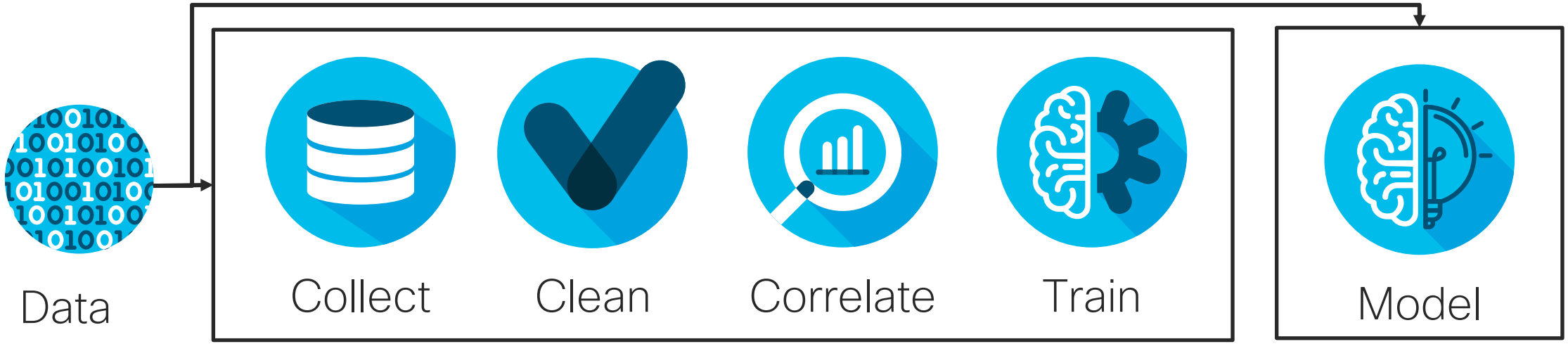
SaaS Delivered

Comprehensive Automation

Single Pane of Glass



Data Pipeline Software Tools



Infrastructure Solutions for the Data Pipeline

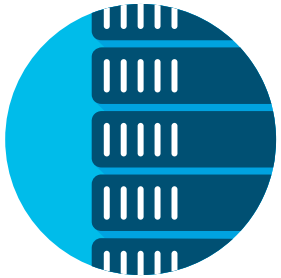
Ingestion

Apache Spark

Apache kafka

jupyter

hadoop



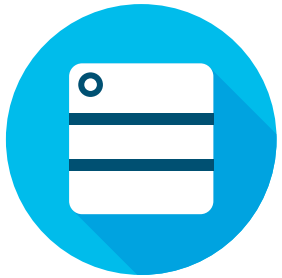
Compute Intensive

TensorFlow

PyTorch

Kubeflow

kubernetes



Storage

SCALITY

SwiftStack

VersaStack™ Solution by Cisco and IBM

FlexPod A Cisco and NetApp Solution

FLASHSTACK CI

UCS and HyperFlex AI / ML Solutions

NGC TensorFlow on Hortonworks 3

TensorFlow

HORTONWORKS
POWERING THE FUTURE OF DATA™

UCS

NGC TensorFlow on OpenShift

TensorFlow

OPENSHIFT

UCS

Kubeflow on UCS and GCP

Kubeflow

Kubeflow

UCS

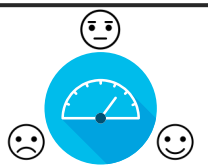
Google Cloud

FlexPod
A Cisco and NetApp Solution

FLASHSTACK CI

NVIDIA

UCS C480 ML



Customer Sentiment Inference on HyperFlex 4.0

Performance

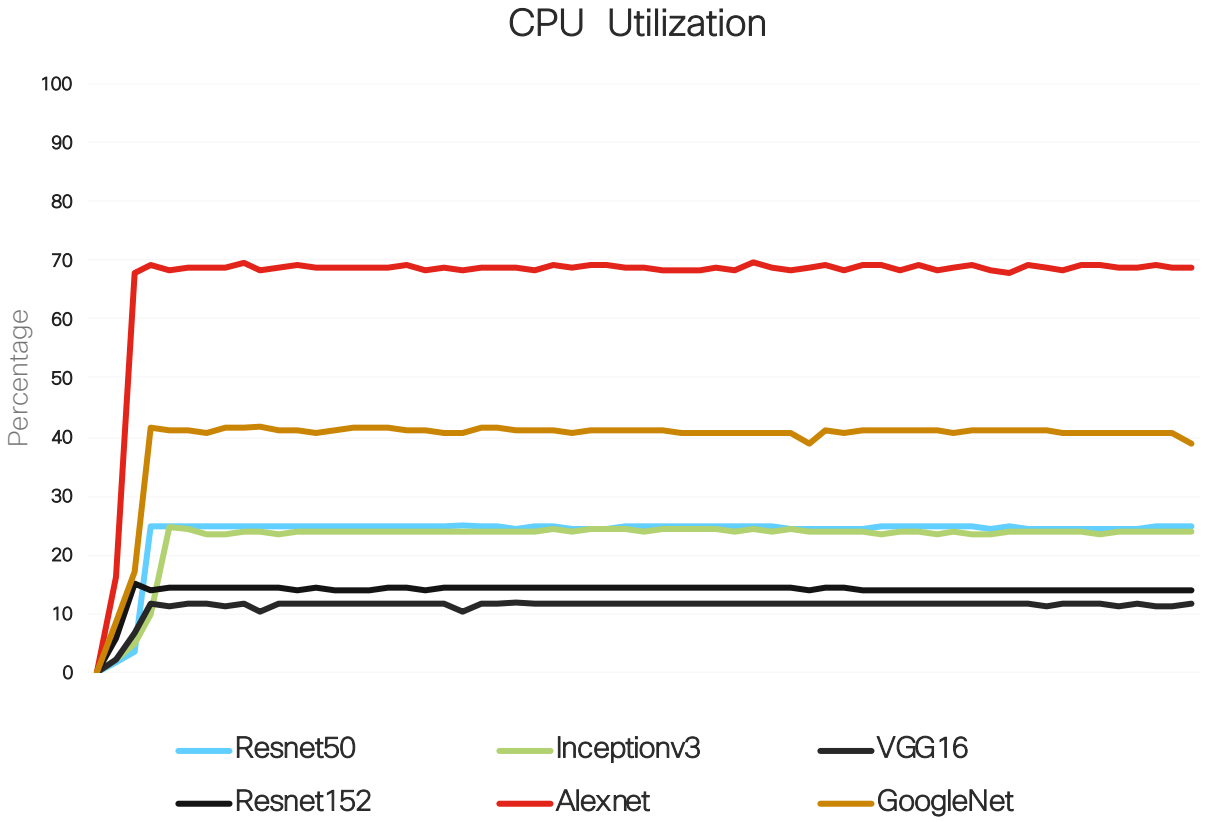
Cisco UCS C480 ML Rack Server

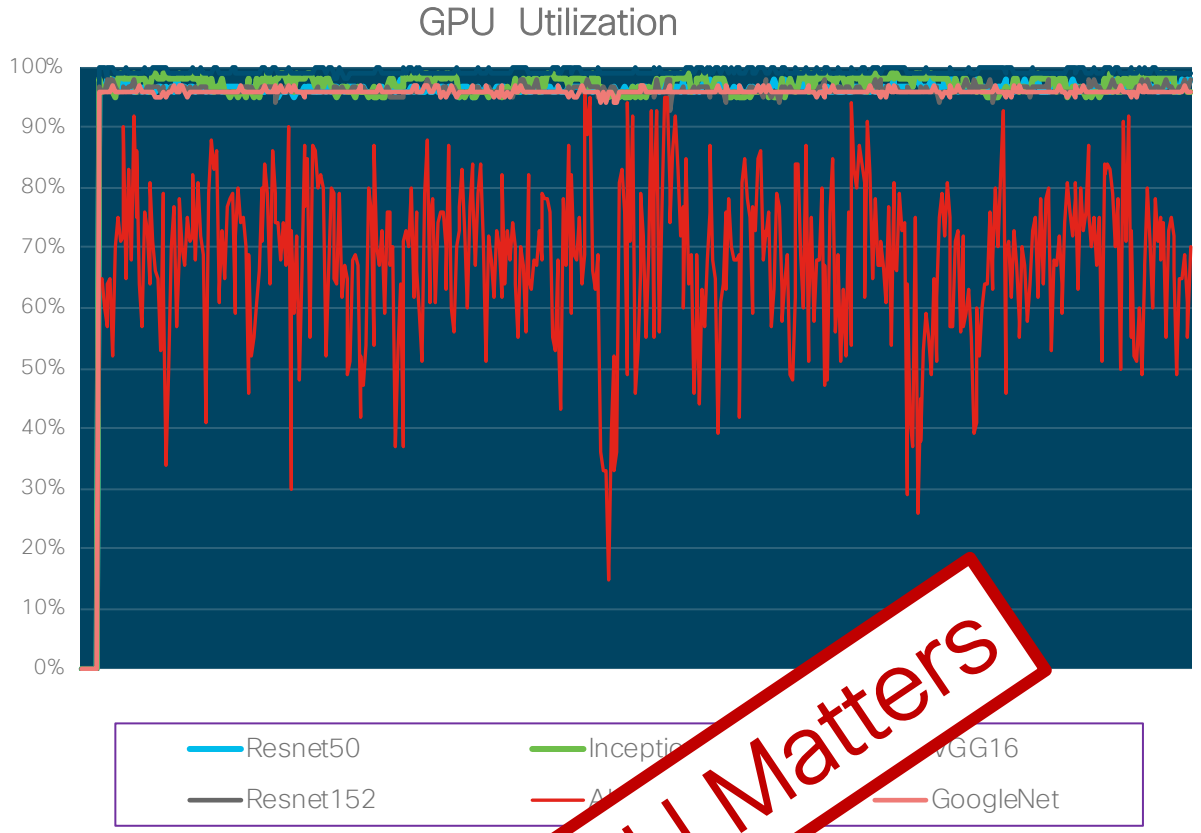
No-Compromise Purpose Built Server for Deep Learning

- Eight SXM2 Nvidia V100 GPUs
- NVlink GPU Interconnect
- Intel Skylake Processor Choices
- Up to 24 drives with 6 NVMe
- UCSM & Intersight managed
- Cisco Validated Designs

- 40,000+ CUDA Cores

C480 M5 ML Training: CPU and GPU Utilization

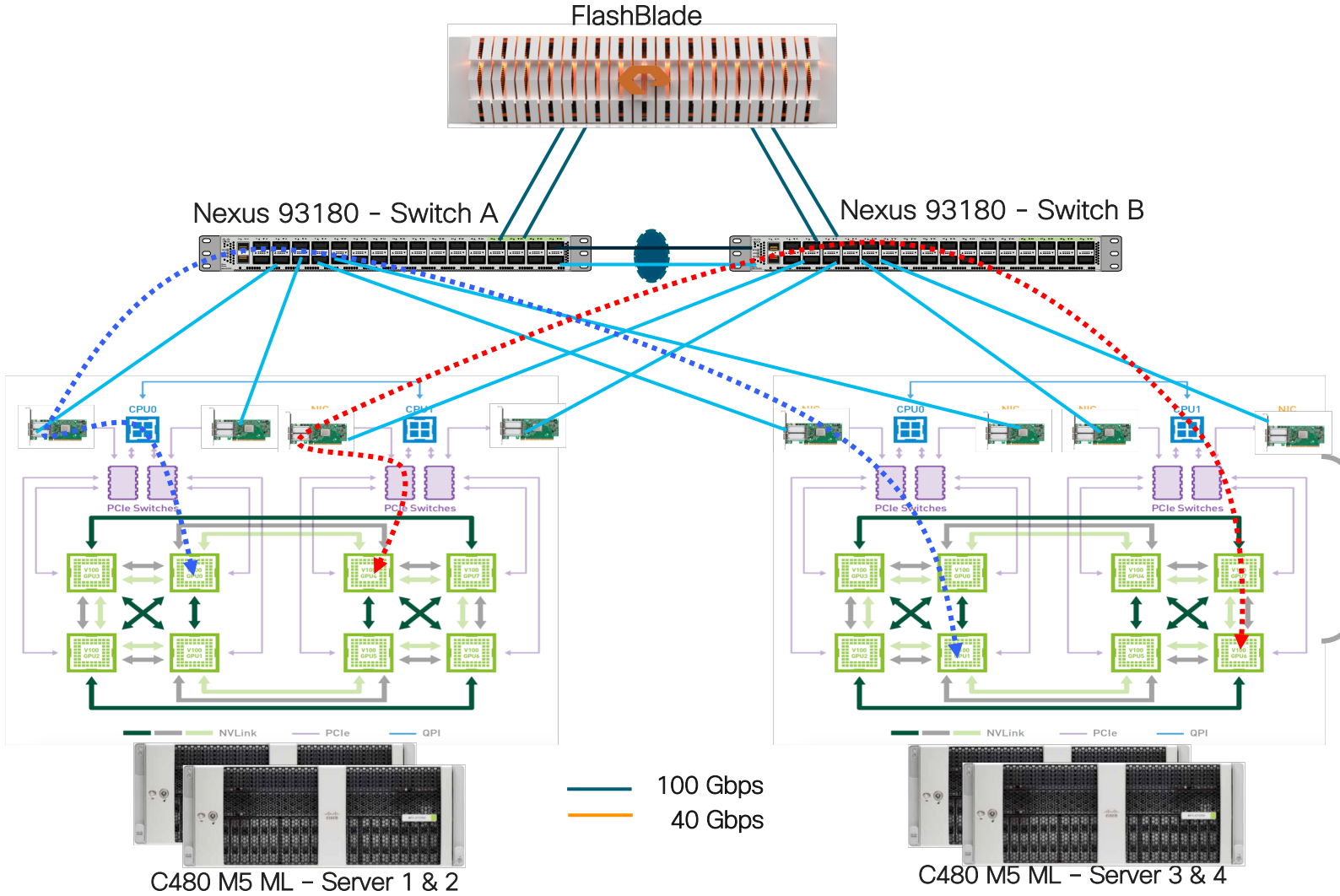




CPU Matters

- Highest CPU utilization for Alexnet, VGG16 uses least.
- Observed even core utilization (across all CPU cores)
  - Higher # of cores might give slight edge compared to frequency optimized CPU.

Distributed TensorFlow Performance

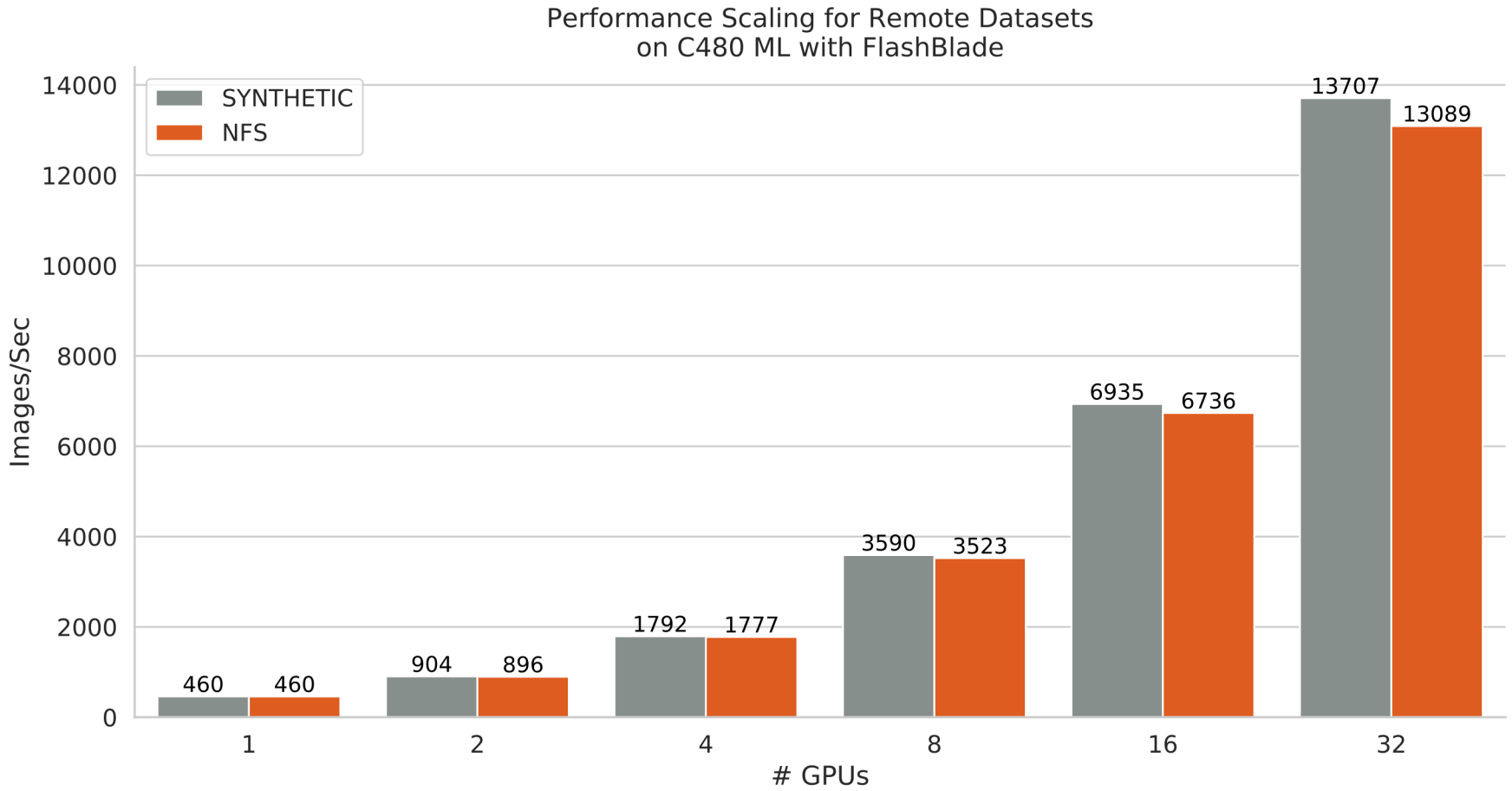


C480 <-> FlashBlade: TCP
Simplicity from unified Ethernet fabric

GPU to GPU: RDMA
via NVIDIA NVlink Interconnect traffic (

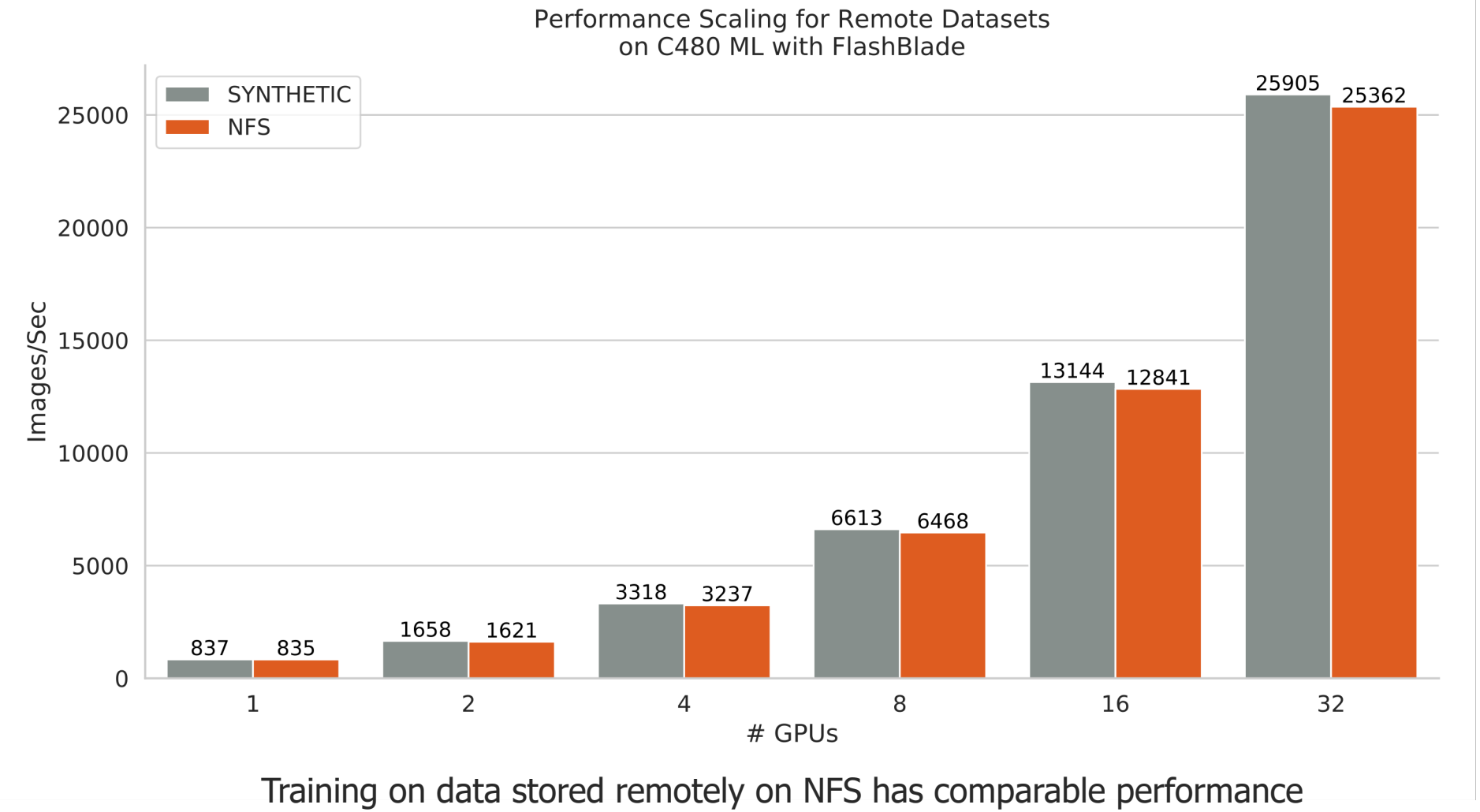
C480 <-> C480: RoCE v2
RoCE V2 traffic between nodes via Nexus (Blue & Red Dotted lines)

Distributed TensorFlow Performance (Horovod) VGG



Training on data stored remotely on NFS has comparable performance to training with synthetic data (no input pipeline).

Distributed TensorFlow Performance (Horovod) Resnet 50



Training on data stored remotely on NFS has comparable performance to training with synthetic data (no input pipeline).

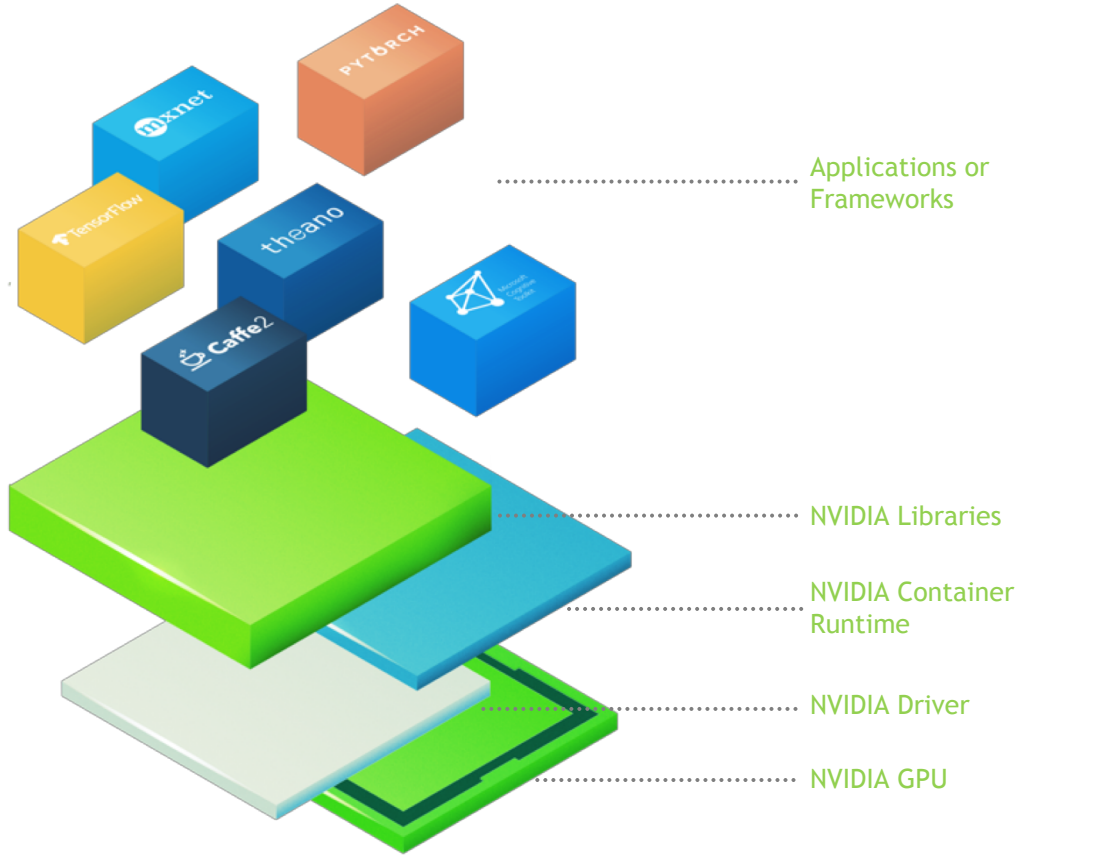

NVIDIA
NGC-READY

CHALLENGES WITH COMPLEX SOFTWARE

Current DIY GPU-accelerated AI and HPC deployments are **complex** and **time consuming** to build, test and maintain

Development of software frameworks by the community is moving **very fast**

Requires high level of **expertise** to manage driver, library, framework dependencies



Cisco UCS C480 ML Rack Server

No-Compromise Purpose Built Server for Deep Learning

- Eight SXM2 Nvidia V100 GPUs
- NVlink GPU Interconnect
- Intel Skylake Processor Choices
- Up to 24 drives with 6 NVMe
- UCSM & Intersight managed
- Cisco Validated Designs

- 40,000+ CUDA Cores

NGC-Ready

THE DESTINATION FOR GPU-ACCELERATED SOFTWARE

HPC	Deep Learning	Machine Learning	Inference	Visualization	Infrastructure
BigDFT	Caffe2	H2O Driverless AI	DeepStream	Index	Kubernetes on NVIDIA GPUs
CANDLE	Chainer	Kinetica	DeepStream 360d	ParaView	
CHROMA	CUDA	MATLAB	TensorRT	ParaView Holodeck	
GAMESS	Deep Cognition Studio	OmniSci (MapD)	TensorRT Inference Server	ParaView Index	
GROMACS	DIGITS	RAPIDS		ParaView Optix	
LAMMPS	Microsoft Cognitive Toolkit				
Lattice Microbes	MXNet				
MILC	NVCaffe				
NAMD	PaddlePaddle				
PGI Compilers	PyTorch				
PicOnGPU	TensorFlow				
QMCPACK	Theano				
RELION	Torch				
vmd					

10 containers

SOFTWARE ON THE NGC CONTAINER REGISTRY

42 containers

October 2017

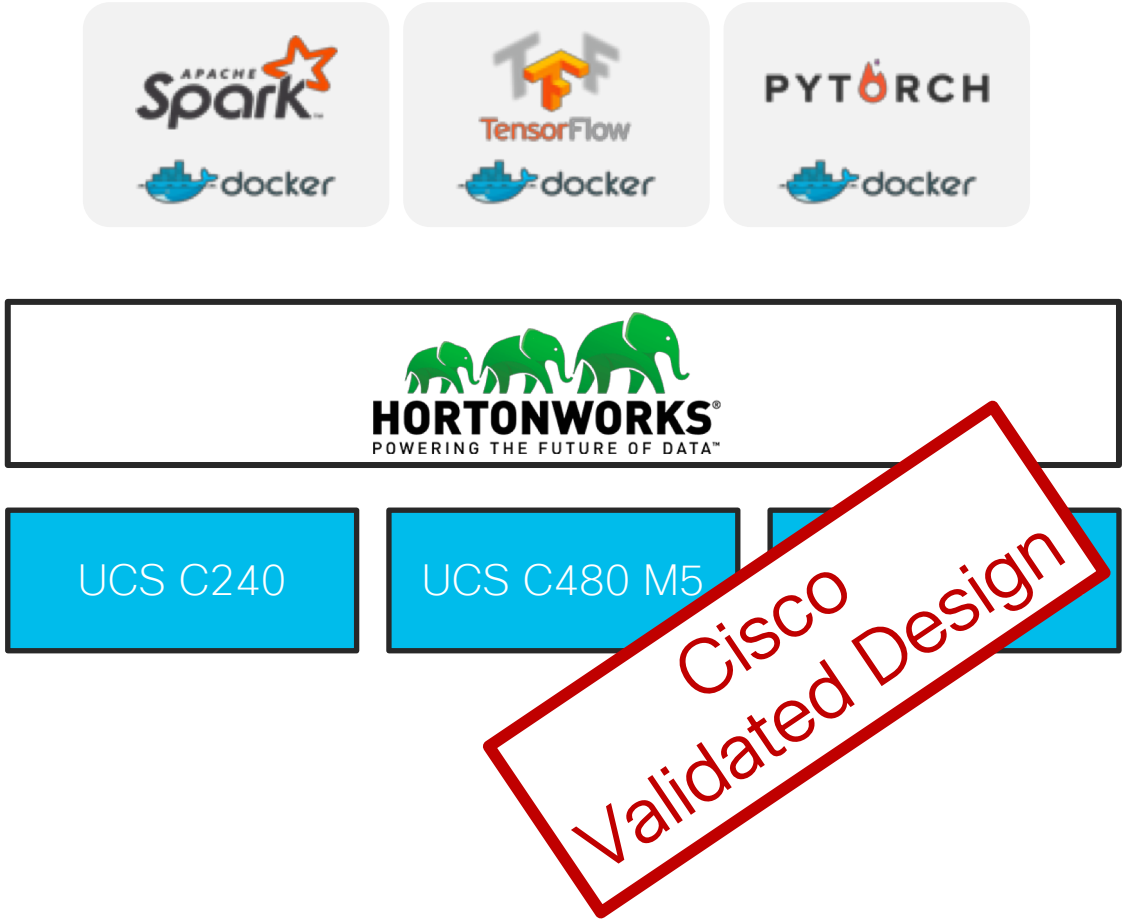
November 2018

UCS AI / ML Solutions:

Hortonworks
Red Hat OpenShift
FlexPod
FlashStack
Inferencing on HyperFlex
Kubeflow

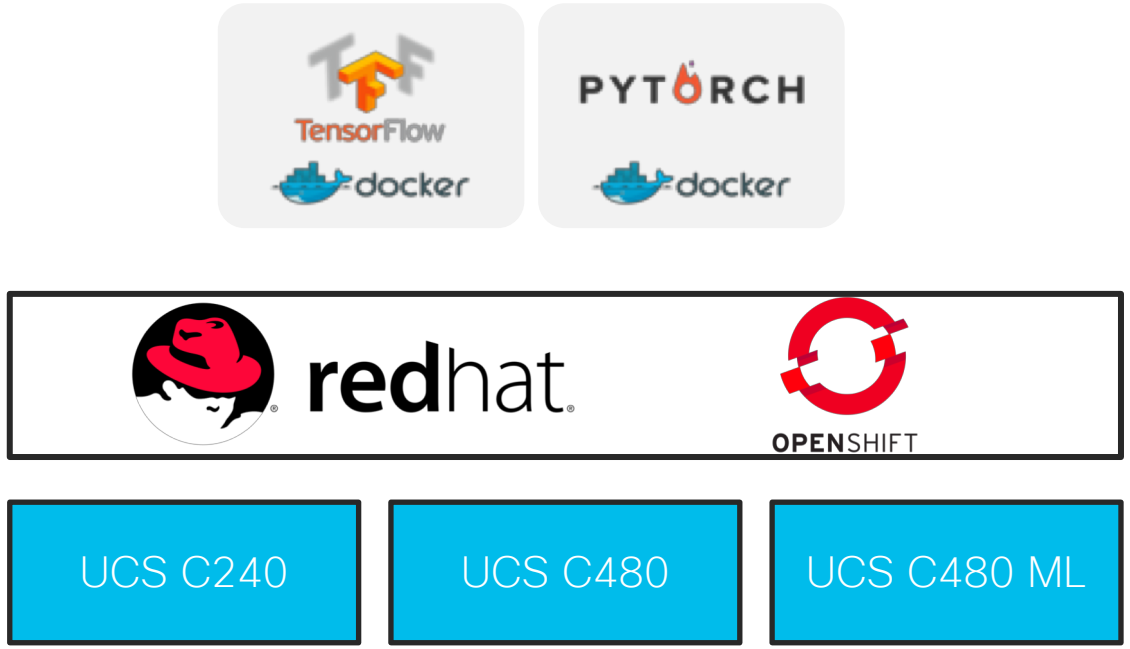
NGC on Hortonworks 3

- Hortonworks 3 schedules
  - Docker container workloads on
  - Servers with CPU and GPU
- Run NGC
  - TensorFlow
- Scale CPU, GPU, and Storage
- Mix and Match Different UCS
  - UCS C240: Up to 2 PCIe GPUs
  - UCS C480 M5: Up to 6 PCIe GPUs
  - UCS C480 ML: 8 NVLink GPUs



NGC on Red Hat OpenShift

- Scale CPU and GPU on Kubernetes with Enterprise support
- Mix and Match Different UCS
  - UCS C240: Up to 2 PCIe GPUs
  - UCS C480: Up to 6 PCIe GPUs
  - UCS C480 ML: 8 NVLink GPUs
- Run NGC
  - TensorFlow



FlashStack for AI

NEW!

Next-Gen Stack
Platform for AI/ML Workloads

Built by Industry Innovators
AI/ML in a Box

Next Gen Solutions for Next Gen Opportunities

FlashStack Built for AI/ML

For Modern Workloads
Ingest, Analyze, Access at Massive Scale

FlexPod AI: Platform for Innovation

AI/ML/DL Workloads with UCS C480 ML + NetApp A800

Cisco Nexus 9K

High speed fabric

Cisco UCS FI

Unified compute fabric and management

NetApp A800

The world's fastest, cloud-connected flash for AI/DL

Cisco UCS 480 ML M5

Optimized for AI/ML with 8 NVIDIA SXM2 V100 32G modules and NVLink interconnect

Simple

- Extend your existing FlexPod to support AI/ML/DL
- Consistent operational model with single vendor support

Flexible

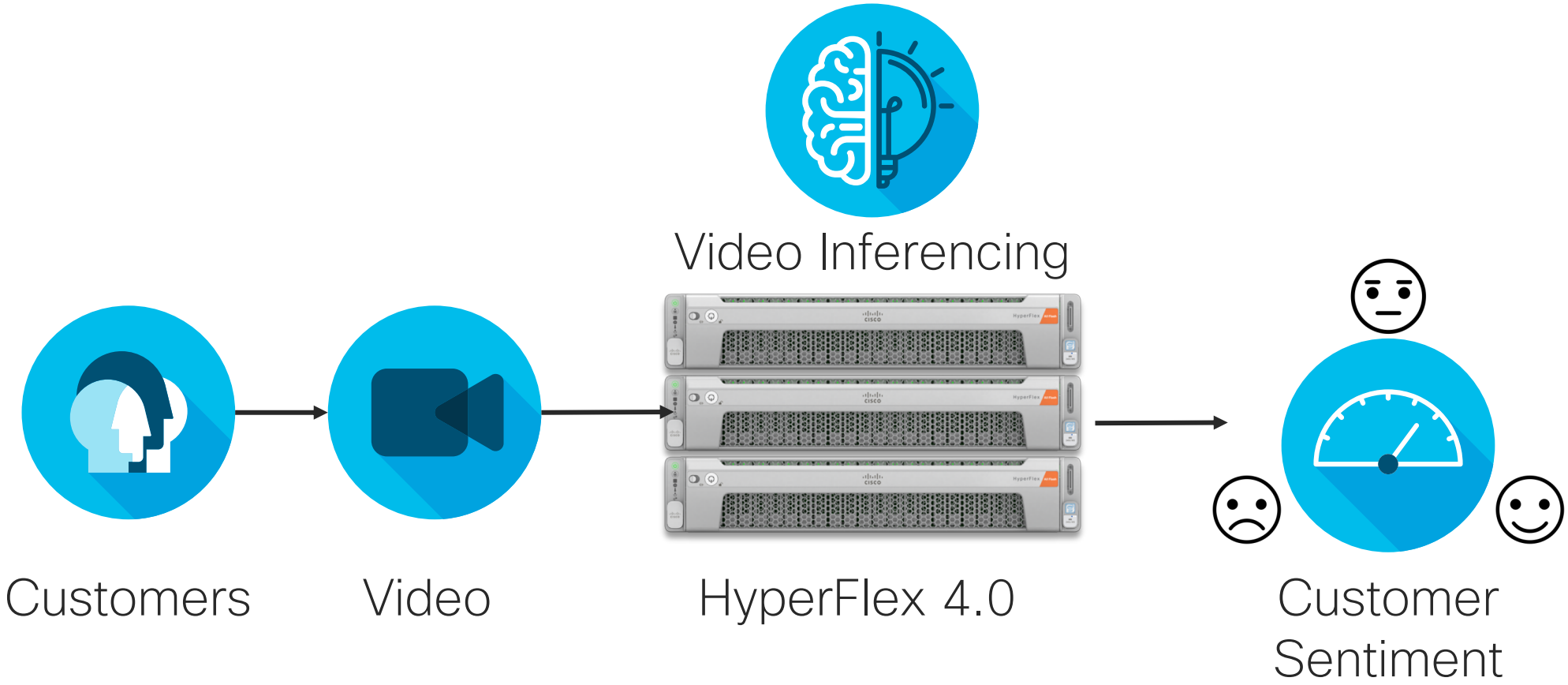
- Intelligently manage data and compute across edge, core & cloud
- Deploy AI Frameworks with confidence

Chainer · mxnet · TensorFlow · Caffe2 · theano · PYTORCH

Powerful

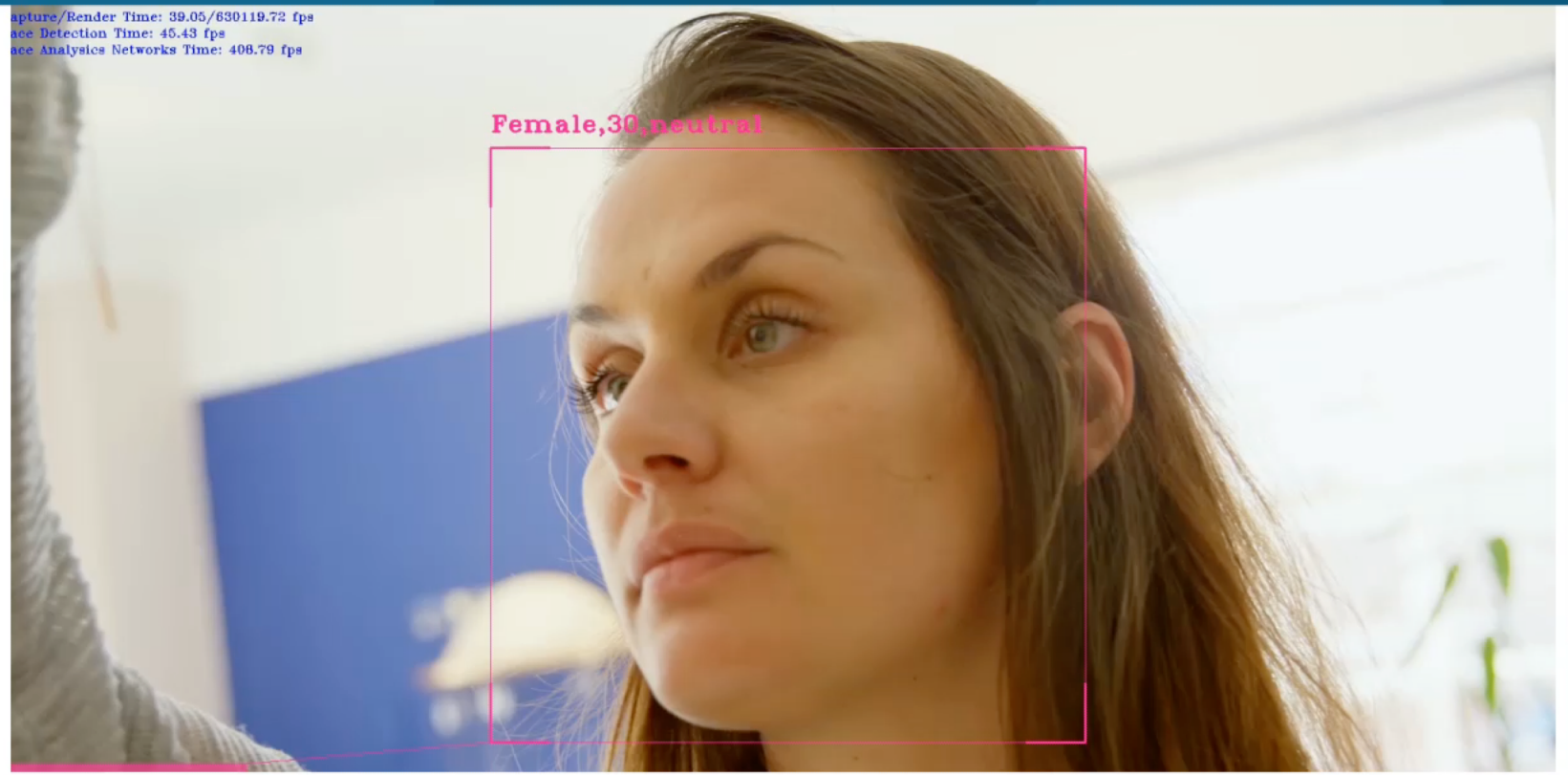
- GPU optimized compute with massively scalable flash
- Start small and grow non-disruptively - Scale without limits

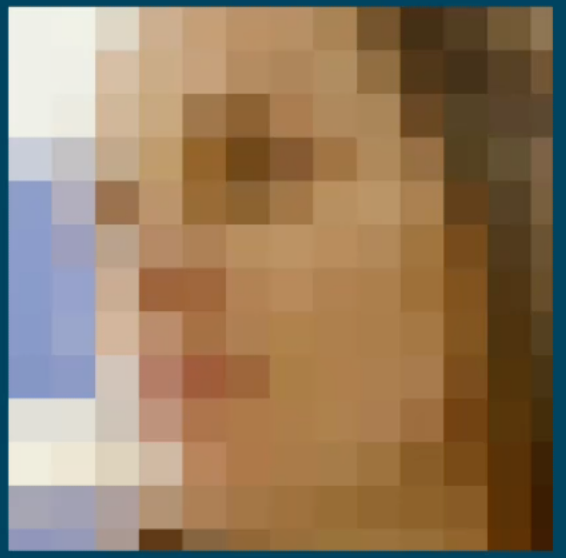
HyperFlex 4.0 for Inferencing on the Edge



Retail Customer-Experience Intelligence

Powered by Cisco® HyperFlex™, Intel® Xeon® processors, and Intel® Optane™ DC SSDs





Gender: Female Age: 27 Expression: Surprise

00 male

00 female

00 total

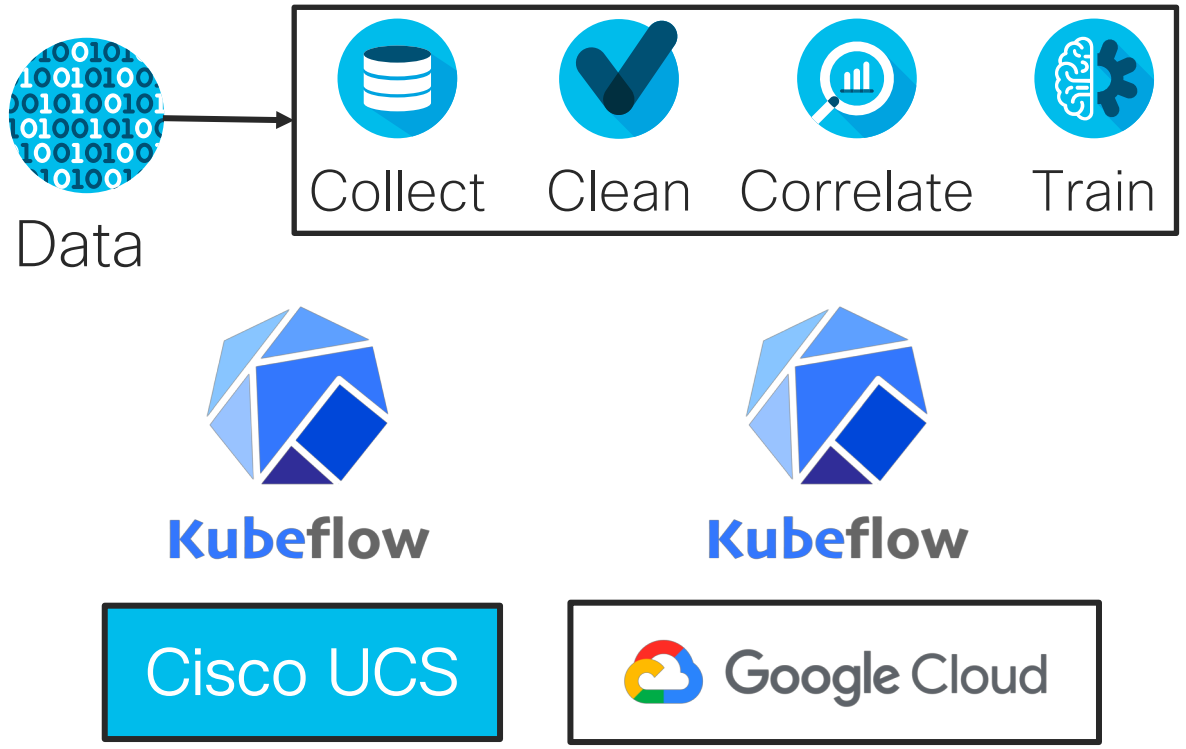
Male

Female

Gender	Age	Expression	Date	Time
Female	27	Surprise	011519	10:00AM

Kubeflow Pipelines on UCS and Google Cloud

- Kubeflow
  - Integrating TensorFlow and Kubernetes
- Kubeflow Pipelines:
  - Reusable software components to build complete data pipeline
- Kubeflow Pipelines on UCS and Google Cloud
  - Hybrid cloud architecture for data pipeline and machine learning



Sina Chavoshi

Google Technical Program Manager

cisco

Kuberflow

Scalable ML Services on Kubernetes

Easy to get started

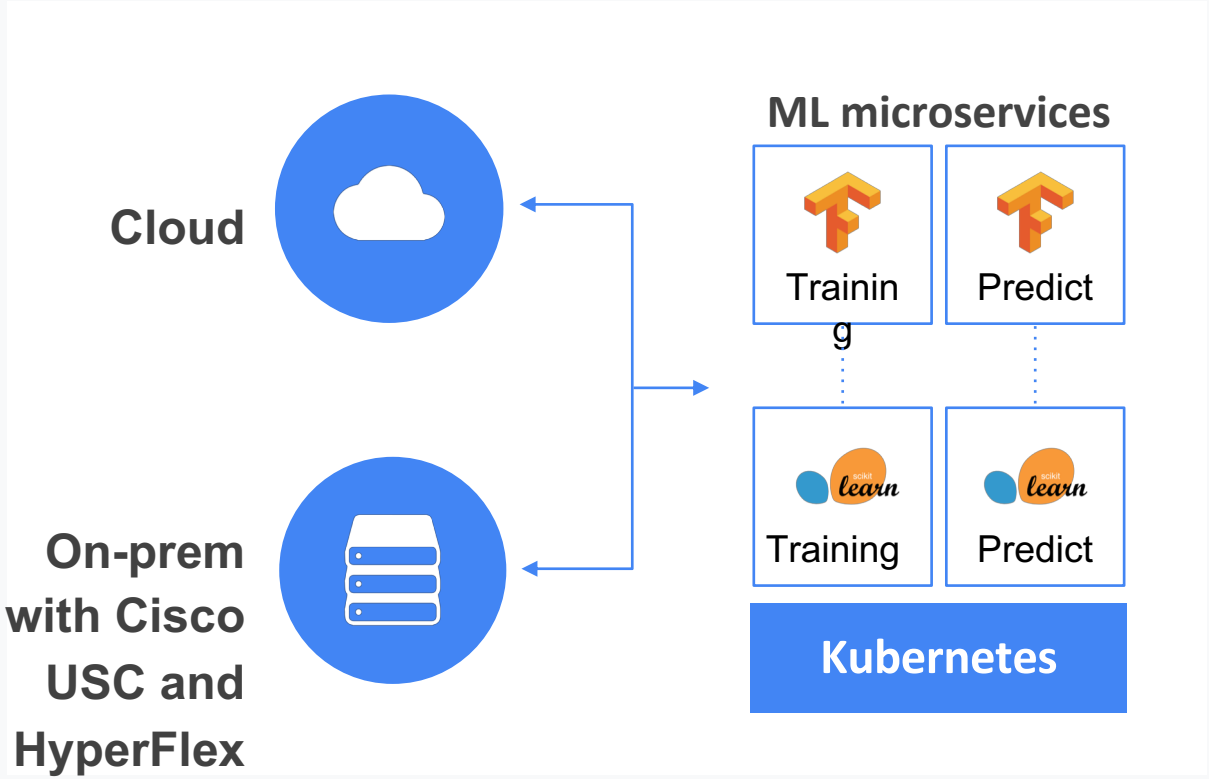
- Out-of-box support for top frameworks
  - pytorch, caffe, tf and xgboost
- Kubernetes manages dependencies, resources

Swappable & Scalable

- Library of ML Services
- GPU support
- Massive Scale

Meet customer where they are

- GCP
- On-prem



Google Cloud

CISCO

Debo Dutta

PhD, Cisco Distinguished Engineer

cisco

Cisco: One of the Leading Contributors to Kubeflow

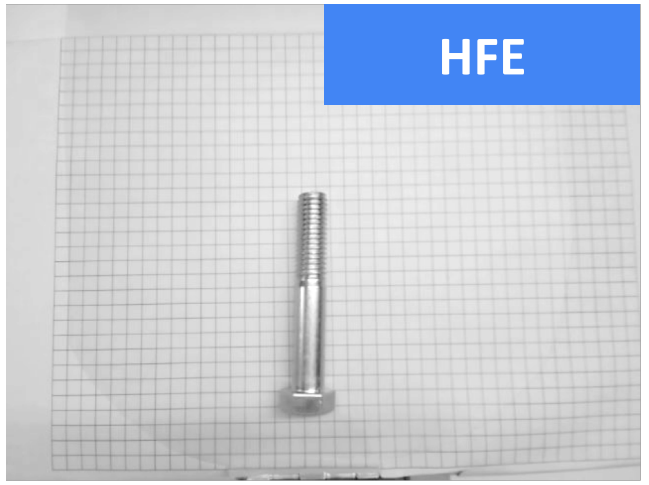
Over 2.8M Lines of Code with 3 Major Proposals

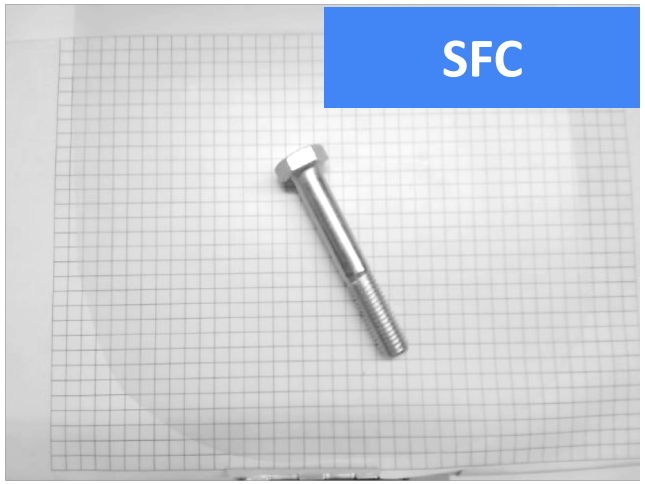
- **#ConsistentAI**: Thought leadership to expand Kubeflow charter to include hybrid cloud
- **Kubebench**: Originated and implemented benchmark for Kubeflow implementation
- **PyTorch** Operator: Continuous improvement and maintenance

- **Katib**:
  - Hyperparameter search
  - AutoML with Neural Architectural Search
- Improve on-premise user experience
- Two of top 5 contributors were from Cisco for version 0.3

Recognizing Bolts Based on Inches vs. Centimeters

- Bolts based on inches vs. centimeters are hard to distinguish: Wrong bolt can ruin equipment
- Use machine learning image classification to identify different types of bolts
- Kubeflow workflow for training, model evaluation, and inferencing
- Run on Cisco UCS and Google Cloud





Kubeflow Demo

cisco

Summary

- Speed of AI/ML deployment is critical to enterprise success
- Cisco solutions can help data scientists and IT teams to accelerate AI/ML Deployment
- Regardless of location of data pipeline from edge to data center to cloud: Cisco has the solution to help You

Call to Action

- Google / Cisco joint webinar on March 21, 10:00 AM PDT: Deploying the AI/ML Data Pipeline Anywhere
- Cisco UCS AI/ML Solutions
- Cisco C480 ML Performance Whitepaper
- 4 x Cisco C480 ML Performance Whitepaper
- Cisco Validated Design with Hortonworks 3
- Other Cisco Validated Designs