Using Kubernetes to Host Machine Learning Workloads in a Cloud Native Way

And what you need to know to get there…..

Michael Doherty
Technical Architect, WW DC Team
DGTL-BRKCLD-2148

cisco Live!

#CiscoLive

cisco

Topics Covered

- Opportunity knocks!
- So how do WE view Machine Learning?
- How and where is Machine Learning done?
- So where does kubernetes & kubeflow fit in?
- Cisco ML Solutions
- Opportunities to look out for
- The new frontier….The Edge!

Opportunity Knocks!

So let’s set a baseline

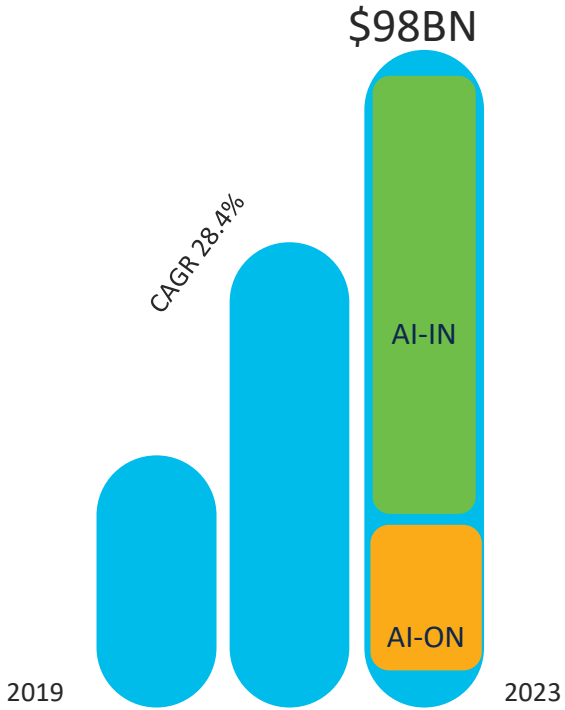
Humans learn from experience

Computers follow instructions

AI/ML learns from data

- AI set to be the biggest technology transition in history.
- We do have a role to play here.
- AI sales enablement will yield greater returns in AI.

The AI Opportunity - Cisco

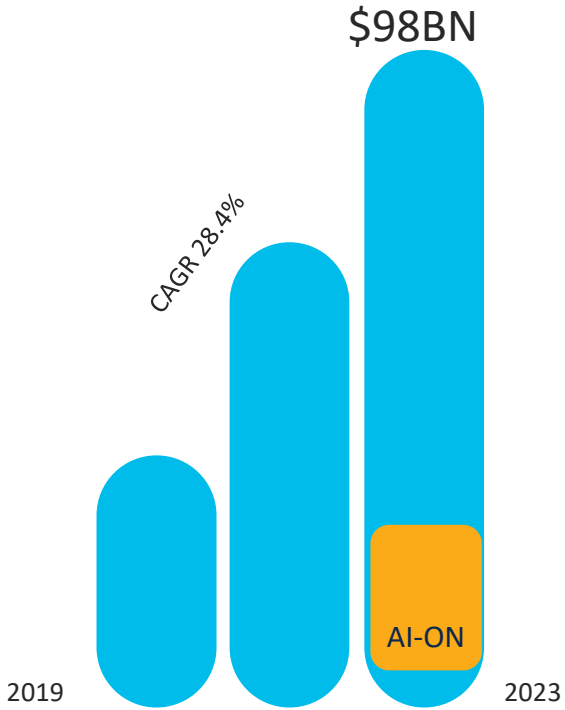


Source: IDC AI systems spending guide 2019

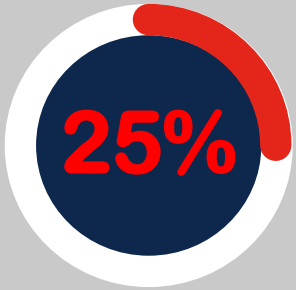
Infrastructure for AI Platforms
DC | IoT | Meraki

AI Powered Infrastructure Features
DNAC | Collab | Security

The AI Opportunity - Customers



Source: IDC AI systems spending guide 2019



of a **Data Scientist** time is spent on infrastructure tasks.

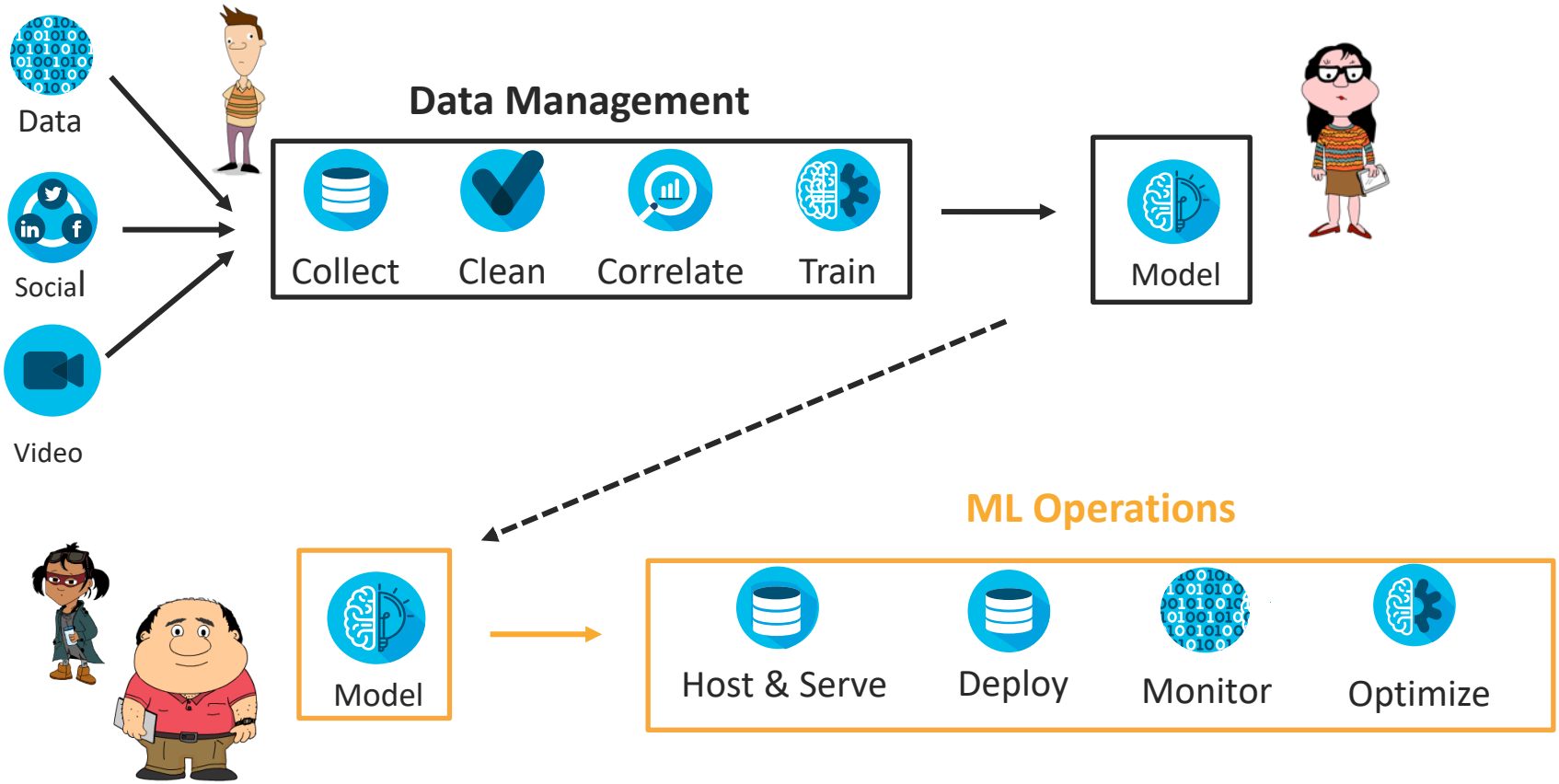
MIT Technology Review

"..majority of teams developing ML capabilities are doing so using **open-source** tooling because of the **dearth of viable commercial options**".

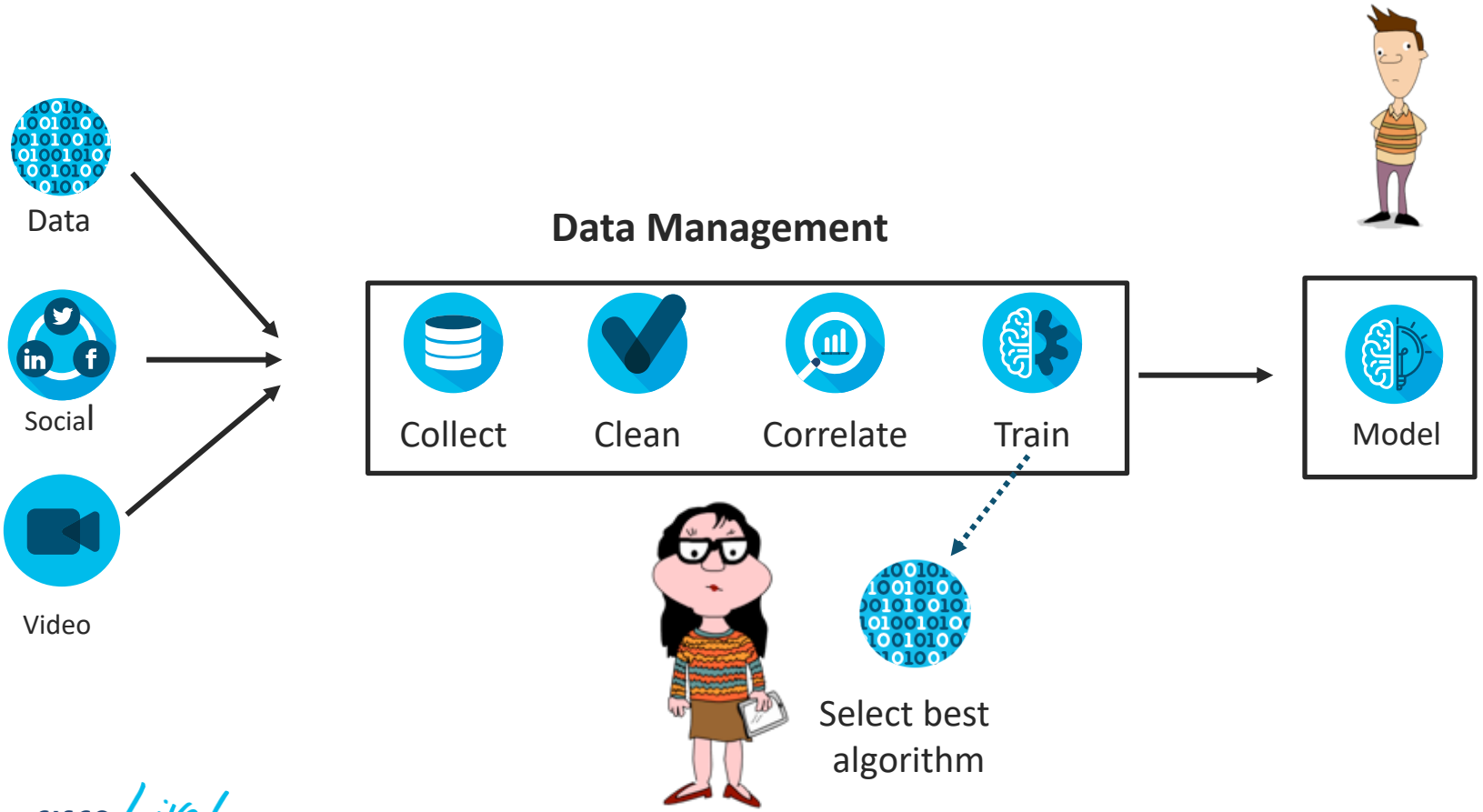
Gartner


So how do WE see
Machine Learning?

There 2 main parts to consider

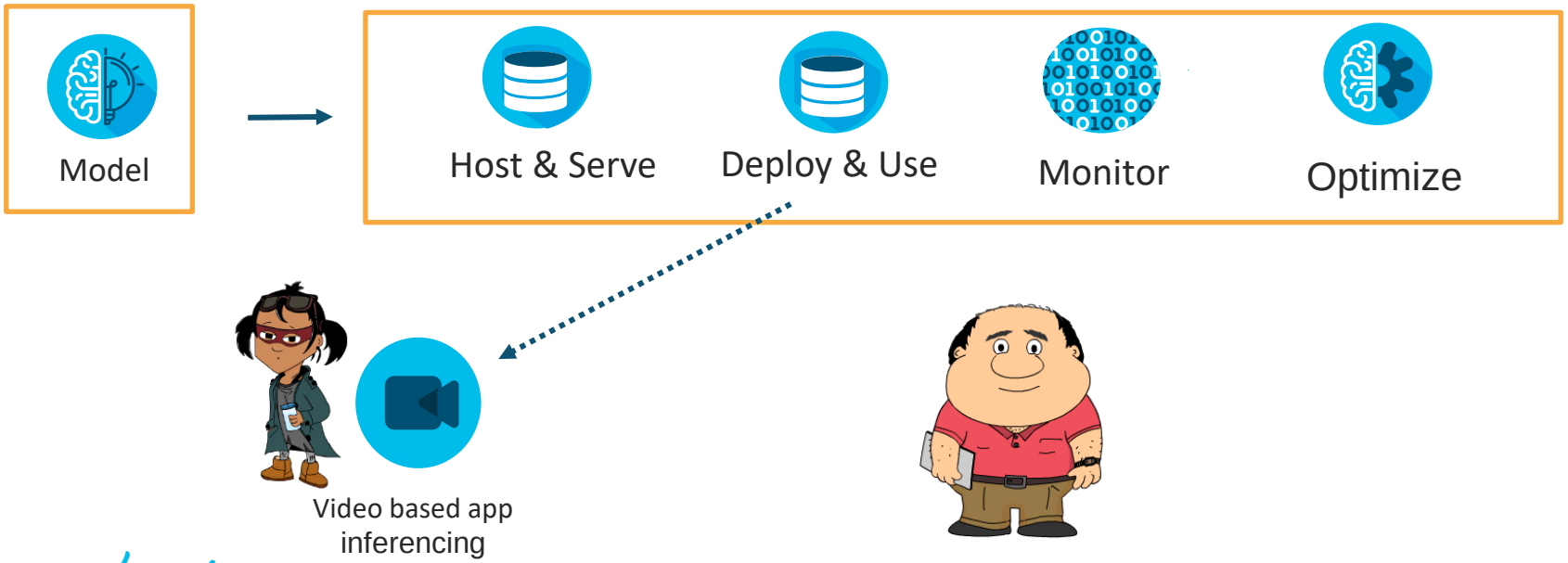


The Data Scientists/Engs



The Operations Teams

ML Operations

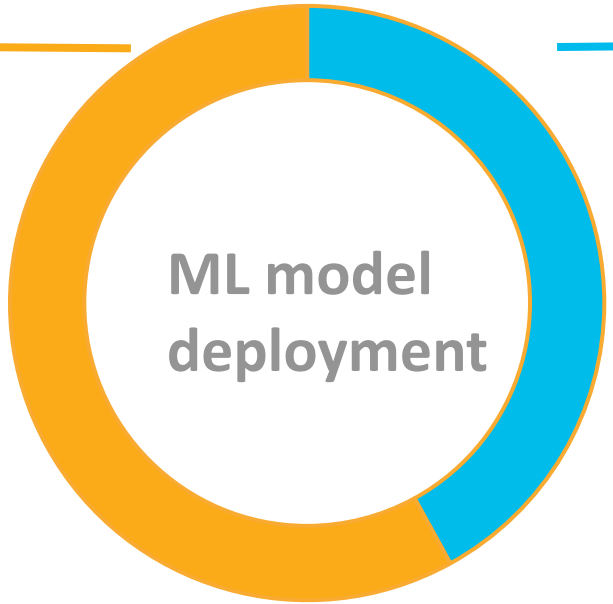



How and where is
Machine Learning done?

On-Premise vs. Cloud

~60%
On Prem

Data gravity and integration
Performance
TCO
Governance
Remodel, retraining
At scale production



~40%
Cloud

Fast deployment
Test-dev
Simplicity
APIs
One time

Source: Gartner, Market Guide for Machine Learning Compute Infrastructures, Sep 2108, ID: G00362287; Figure 3

Public Cloud Appeal

IBM WATSON

Amazon SageMaker

Azure Machine Learning

TensorFlow Enterprise

ML Frameworks

GPUs/TPUs

jupyter

ML Collaboration & Notebooks

amazon web services

IaaS

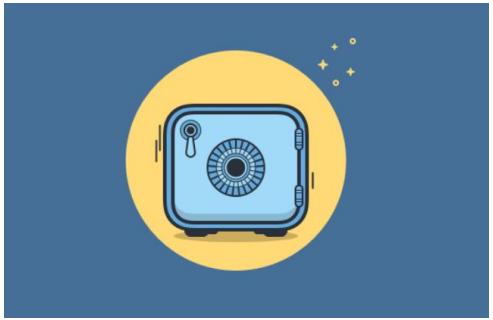
ML Models

Data Services

CI/CD

Private Cloud Drivers

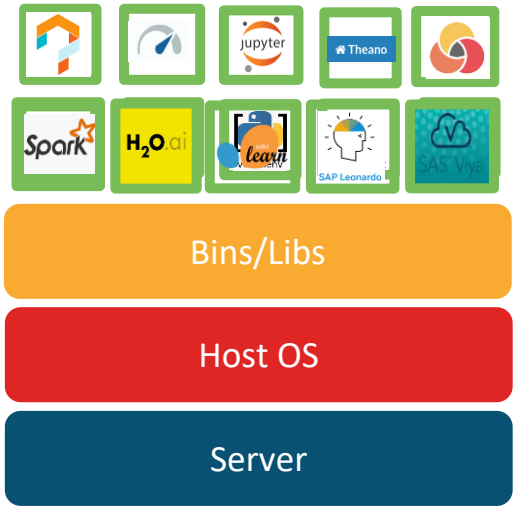
Data gravity and integration



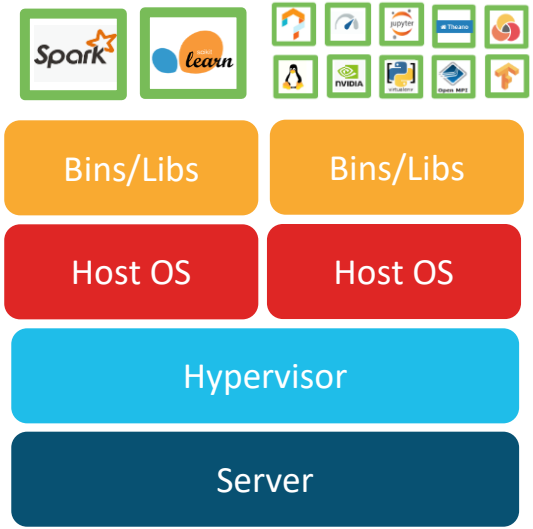
Data Privacy

Costs @ Scale

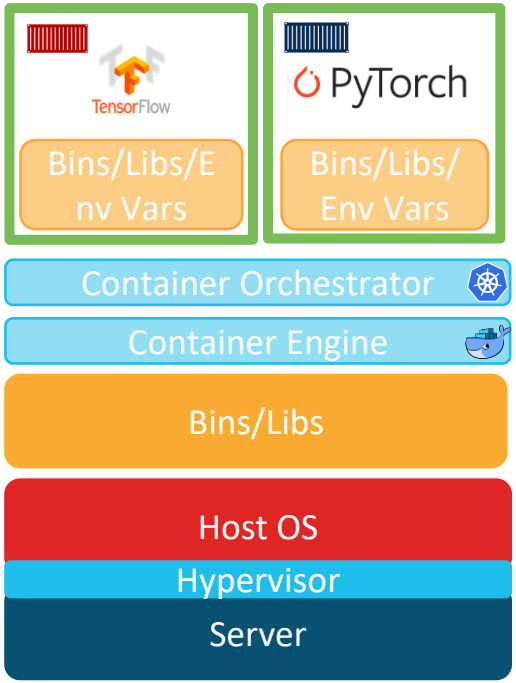
Traditional vs Cloud Native



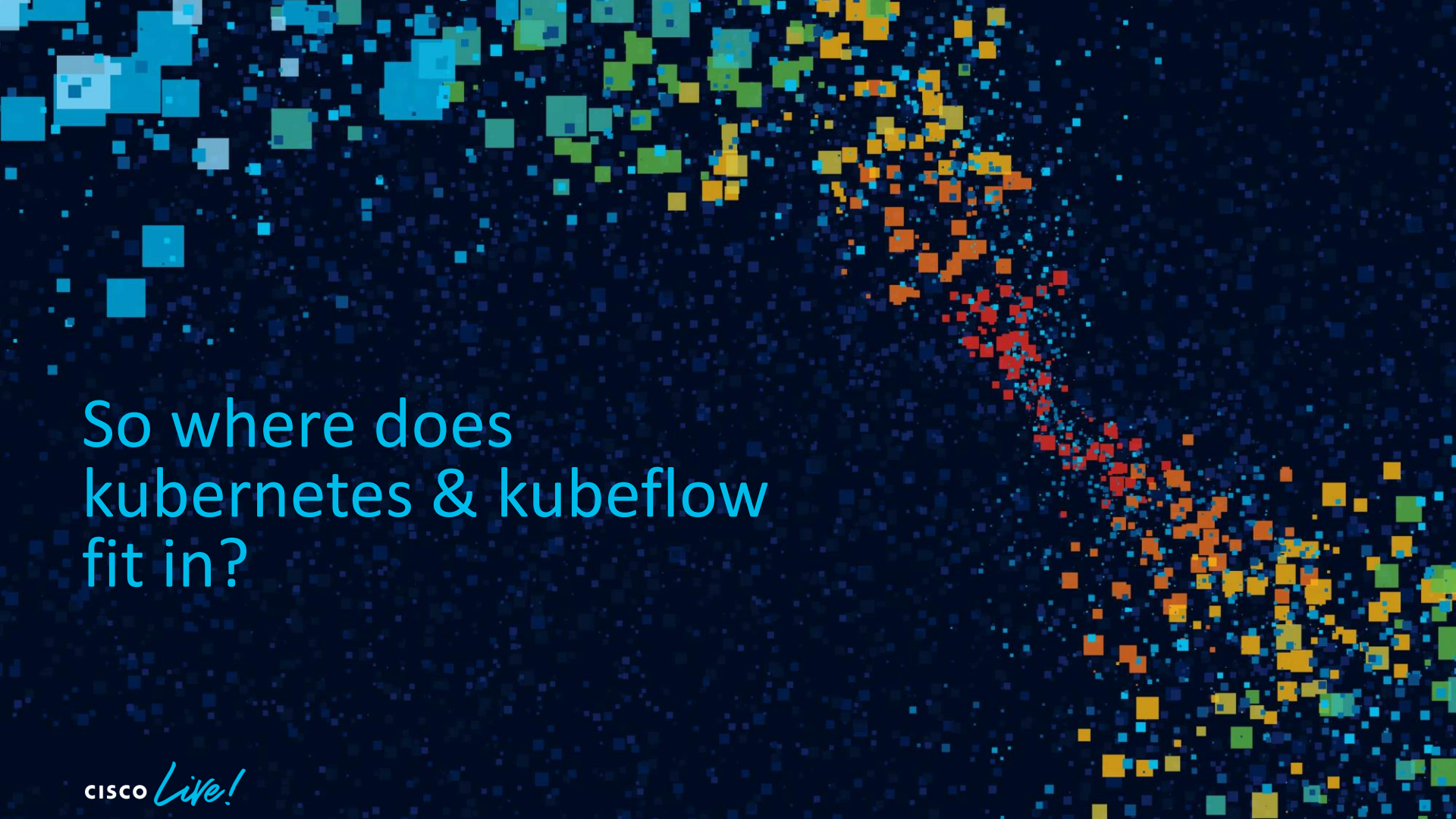
Bare Metal



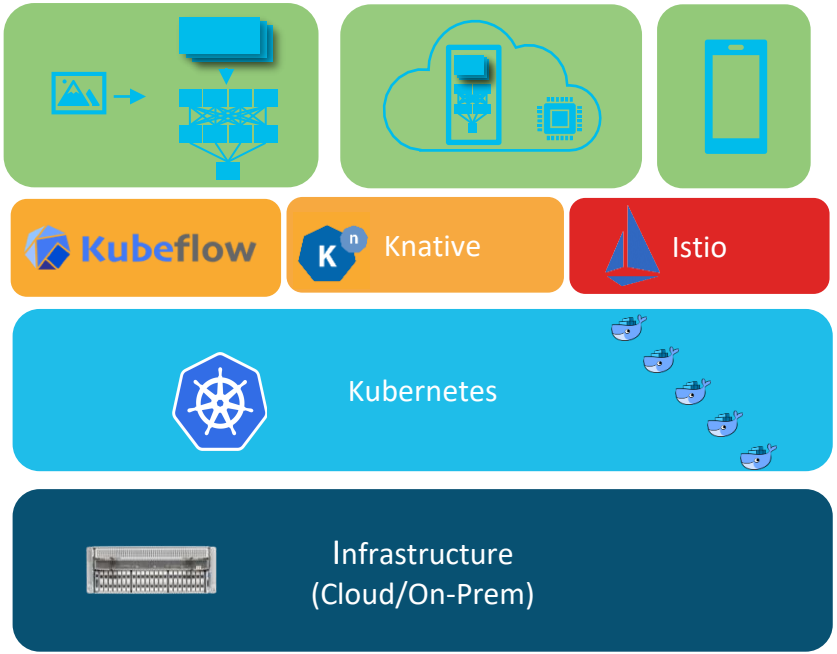
Virtual Machines



Containers

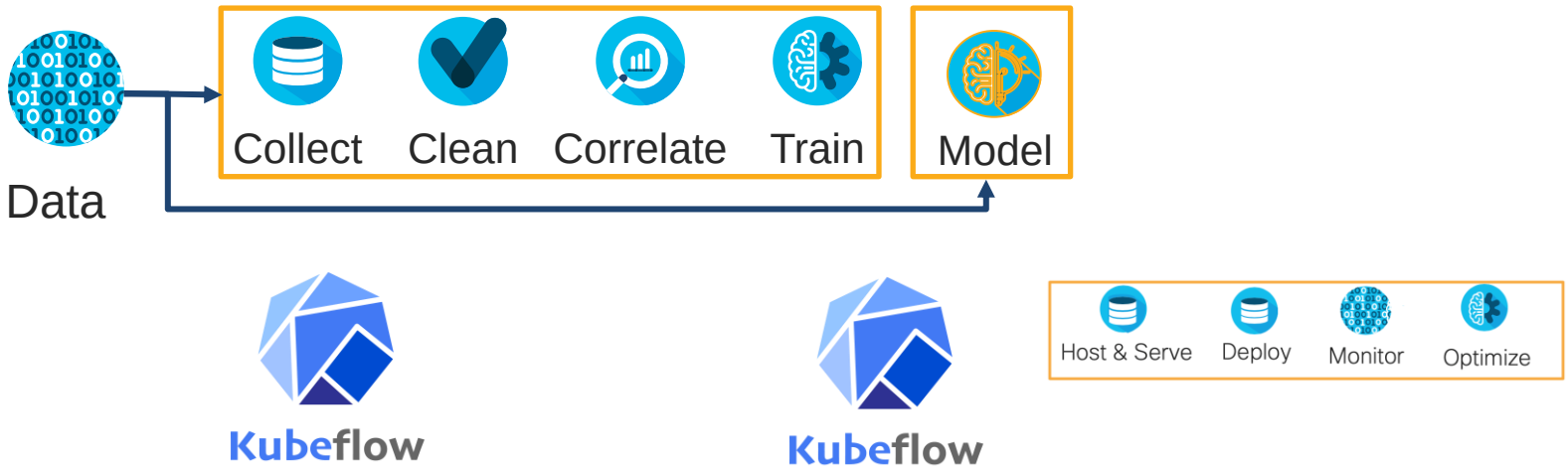

So where does
kubernetes & kubeflow
fit in?

K is the magic letter…well almost!



Kubeflow

Machine learning lifecycle manager that makes it easy to develop, deploy and manage portable, scalable end-to-end ML workflows everywhere



Kubeflow

Scalable ML Services on Kubernetes

Easy to get started

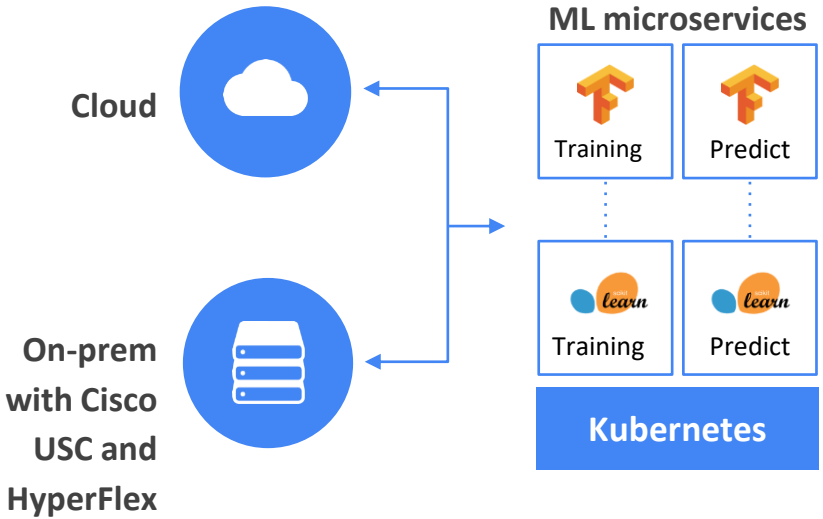
- Out-of-box support for top frameworks
  - pytorch, caffe, tf and xgboost
- Kubernetes manages dependencies, resources

Swappable & Scalable

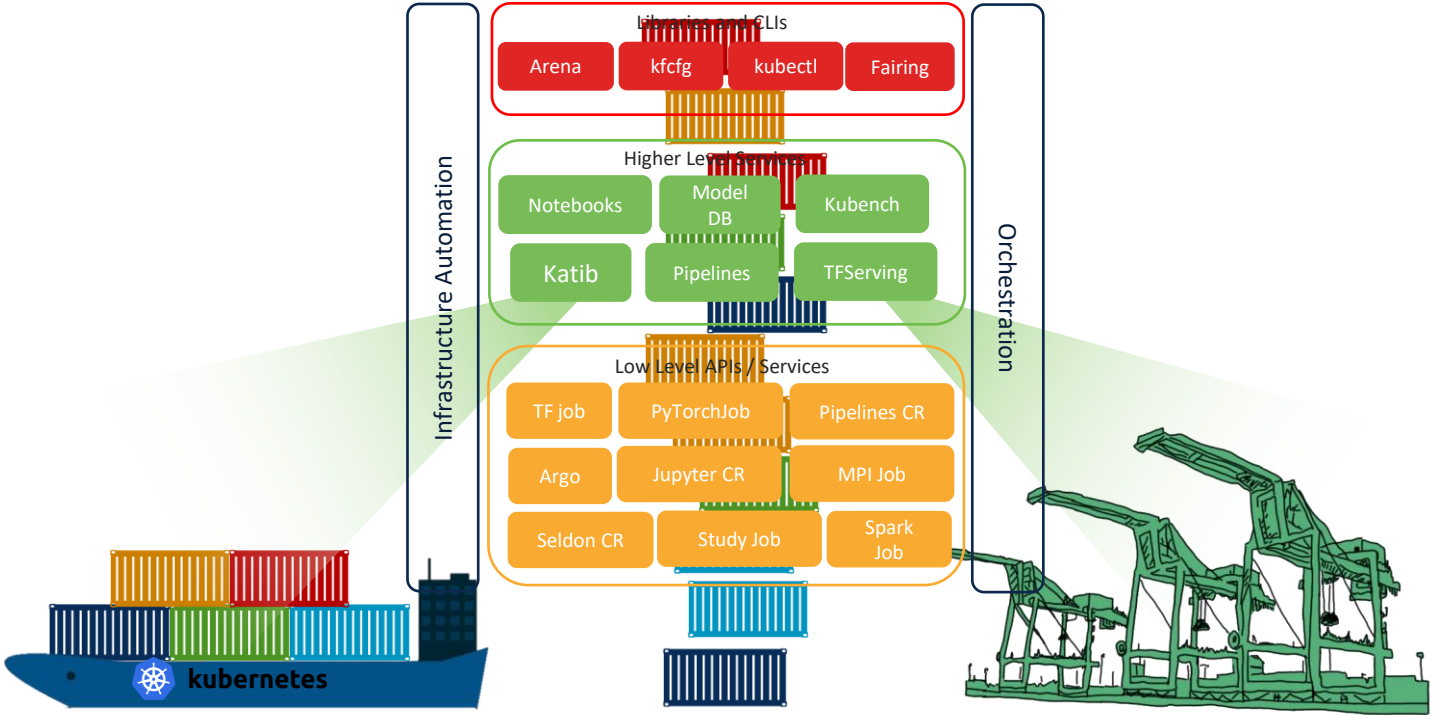
- Library of ML Services
- GPU support
- Massive Scale

Meet customer where they are

- Public Cloud
- On-prem



Kubeflow is made up of multiple components

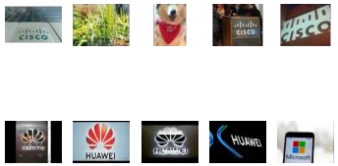


High Level Design PoV for Kubeflow Pipeline

Training a Neural Net to recognize logos

Inputs (collect)

~2k images of logos scraped from the web, labelled ”Cisco” or ”Not-Cisco”



Pre-processing (clean)

- add noise & other pertubations
- rescale to 400x300px
- hold out some data for validation.

Model Specification

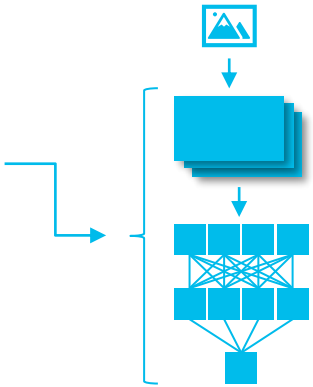
- Convolutional ”funnel” architecture feeding into densely connected binary classifer.
- Specification of hyperparameters

Model Training

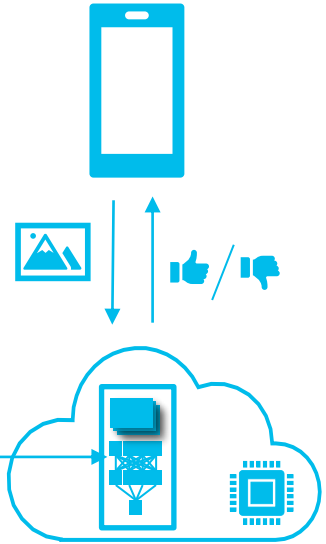
Model Evaluation

Repeat

Export trained model



Deploying the N.N. in an app

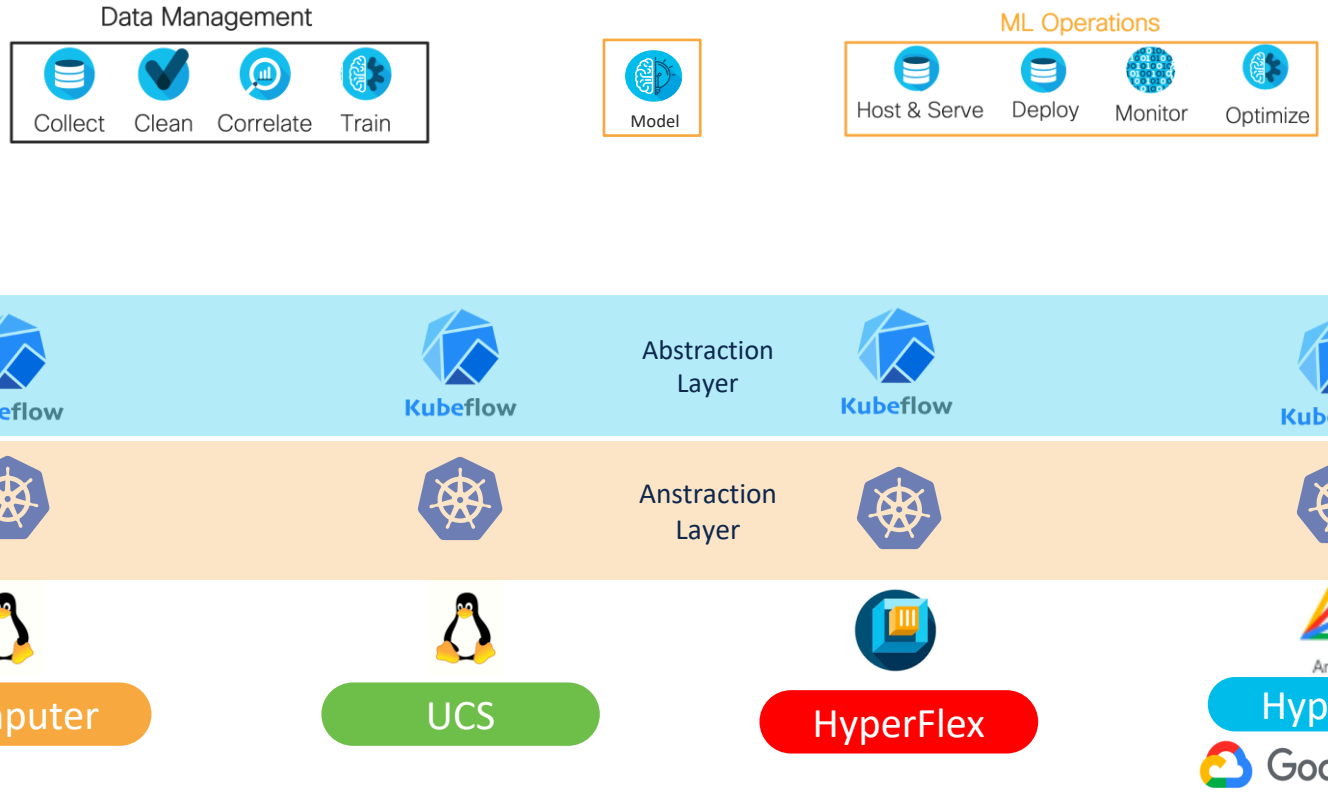


Frontend

API

Backend

Consistent Cloud and On-Premise Machine Learning Experience

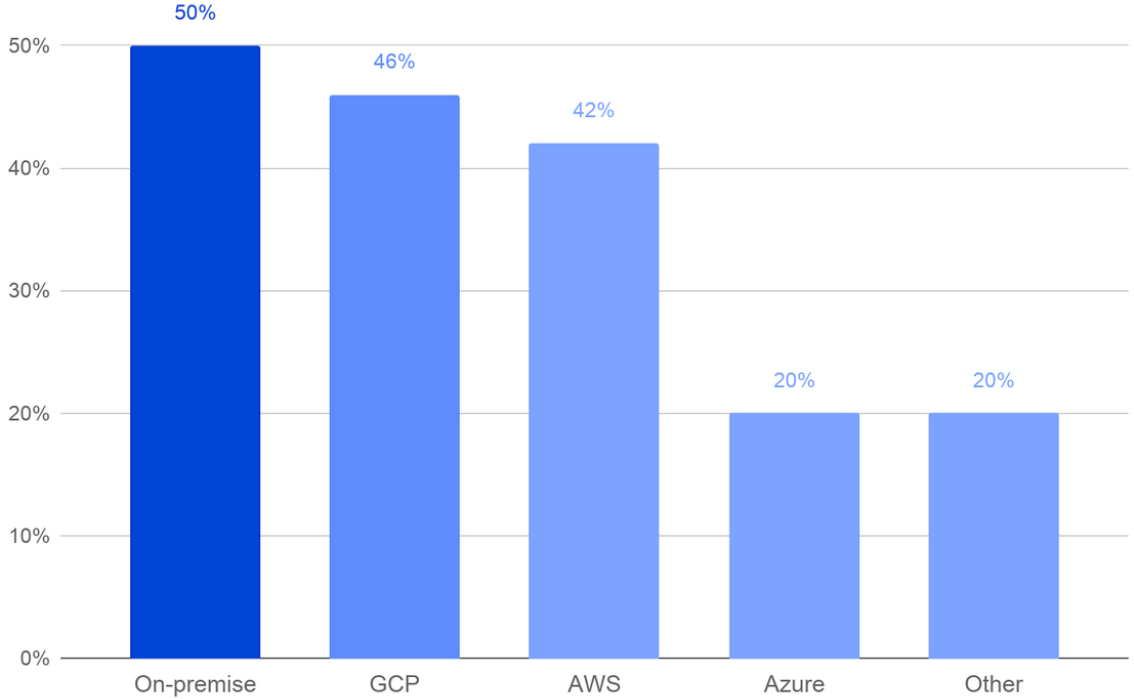


Demo Intro



“Data Scientists have to wait on average 3 months to get an ML Platform delivered to them by IT”

Combined with……

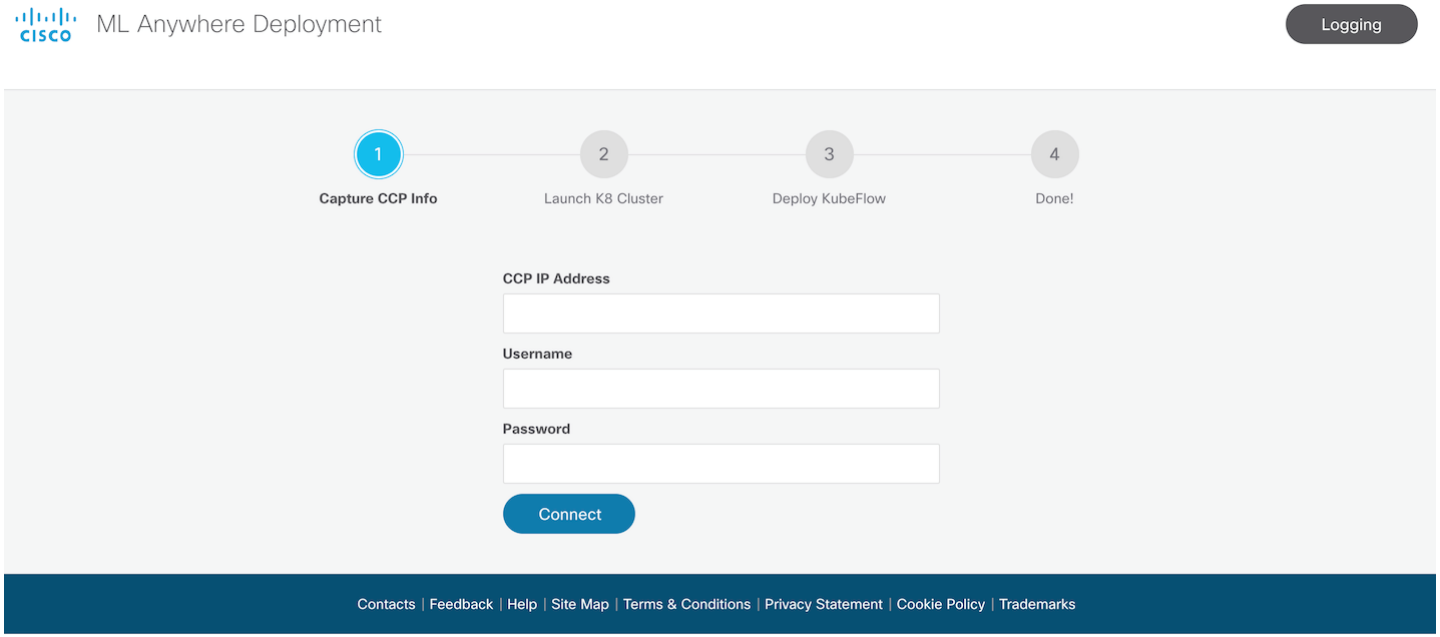


Pic. 3. Where do you run your AI/ML workloads? (Multiple select)? N = 50

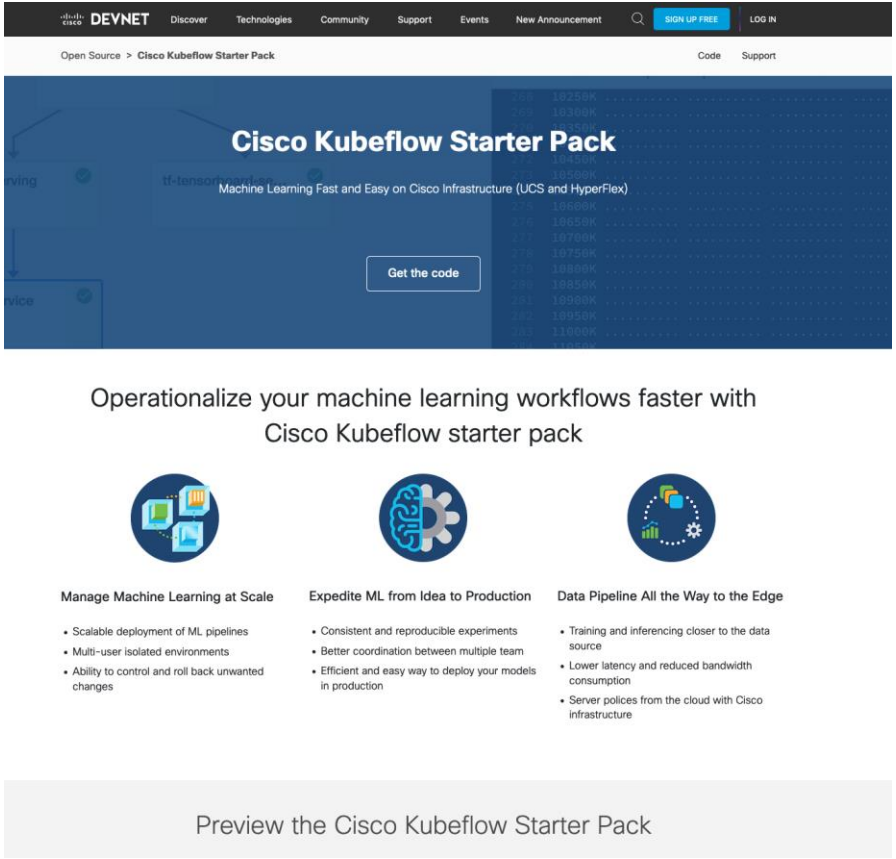
Kubeflow community 2019 fall survey
https://medium.com/kubeflow/kubeflow-community-user-survey-fall-2019-a84776c71743

Introducing ML Anywhere

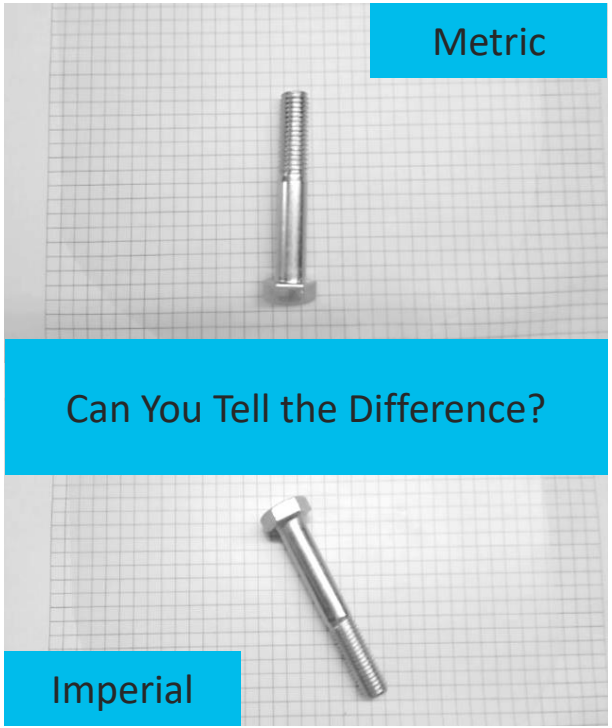
A ML platform deployment ‘easy button’ built on Kubernetes and Kubeflow

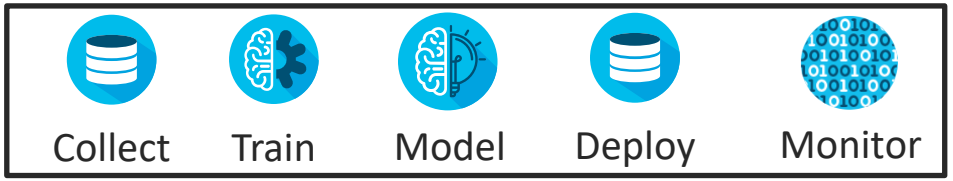


The Cisco kubeflow Starter Pack



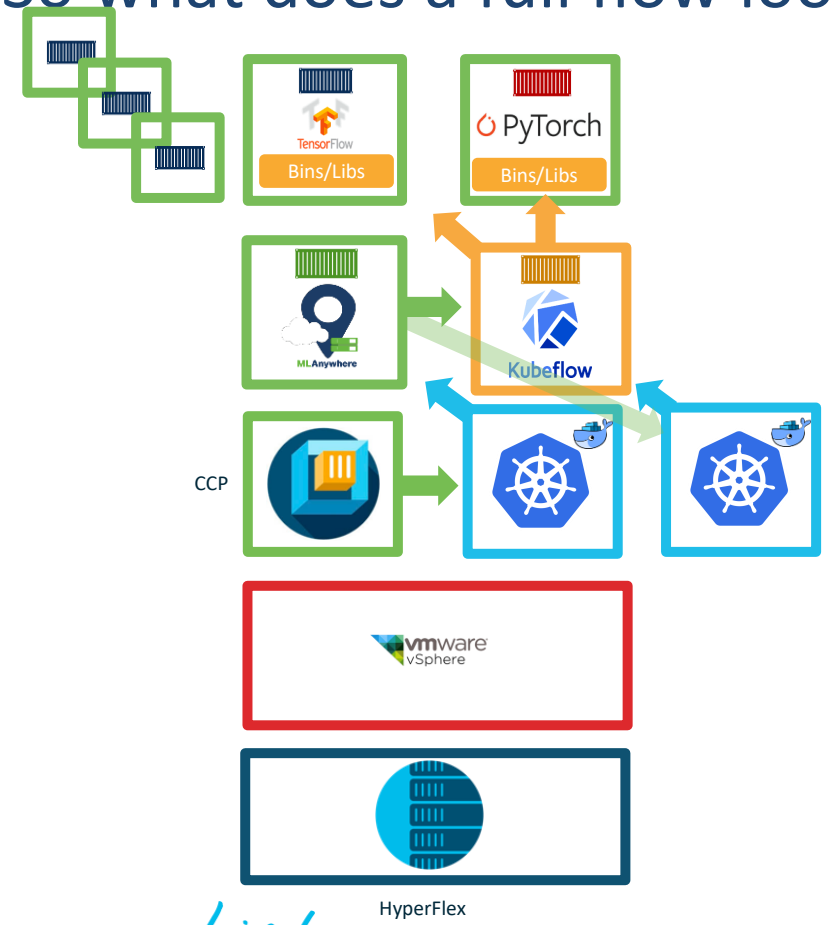
Recognizing Bolts Based on Imperial vs. Metric

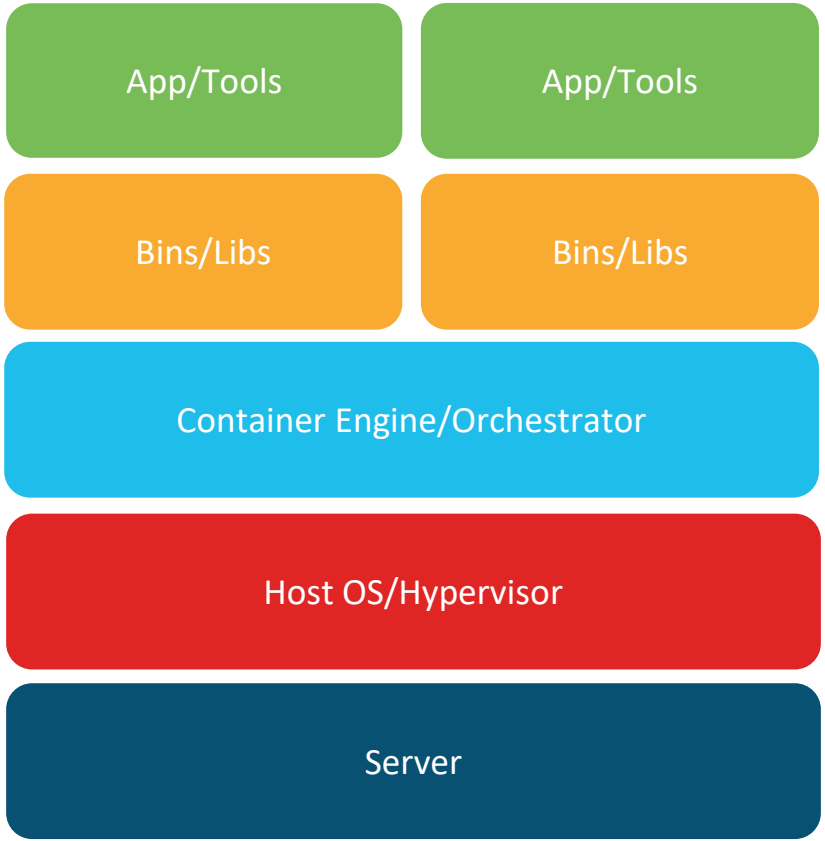




- Bolts based on (Imperial) inches vs. (Metric) centimeters are hard to distinguish: Wrong bolt can ruin equipment
- Use machine learning image classification to identify different types of bolts
- Kubeflow workflow for training, model evaluation, and inferencing
- Run on Cisco HX & CCP

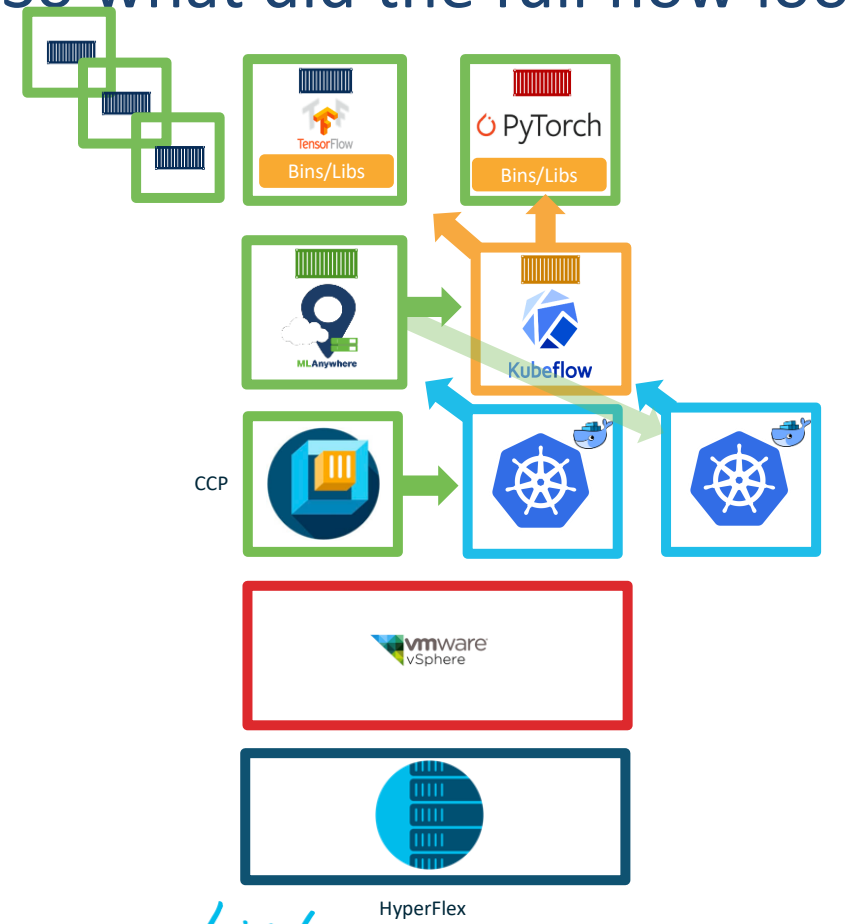
So what does a full flow look like?

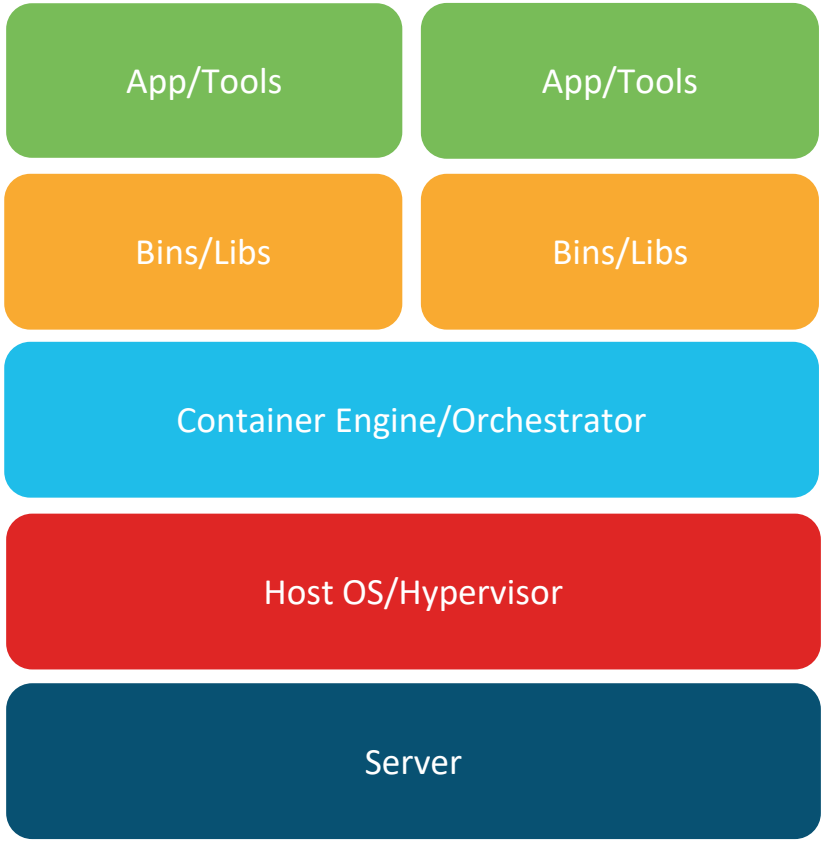




Demo Time

So what did the full flow look like?





Cisco ML Solutions

Cisco Container Platform

Turnkey Solution
for Production-Grade Container
Environments

Native Kubernetes (100% Upstream)
Direct updates and best practices from open source community

Multicloud Optimized
Deploy on-premise and native Amazon EKS, Azure AKS, Google GKE clusters

Integrated
Networking | Storage | Management | Registry | Security | AI/ML

Flexible Deployment Model
VMware (air gapped) | OpenStack (air gapped) | Public cloud

Easy to acquire, deploy and manage | Open and consistent | Extensible platform | World-class advisory and support

Cisco Container Platform Feature Set

Kubernetes-as-a-Service

Setup

- Deploy Kubernetes clusters on HyperFlex, vSphere, OpenStack (CVIM), EKS, AKS
- CNI and Istio service mesh
- Persistent storage
- L4 / L7 Load Balancing
- Container Registry

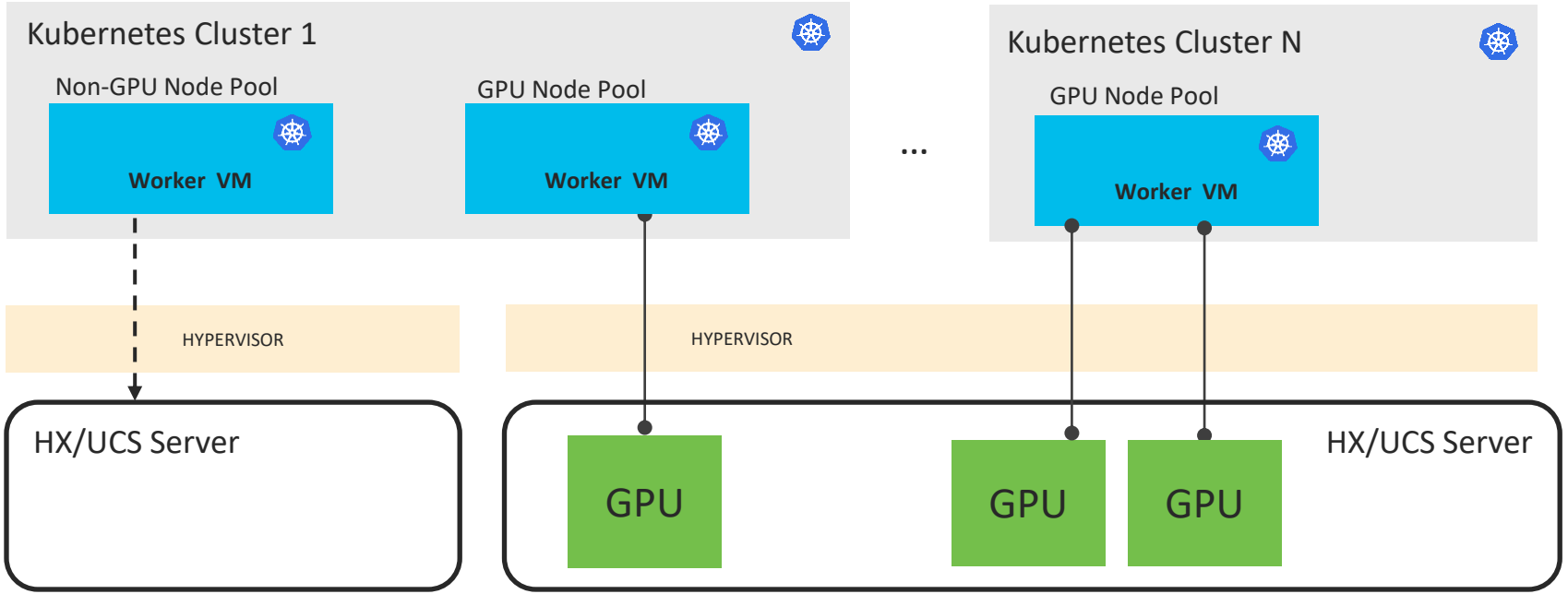
Consume

- AD Authentication / RBAC
- Resource based node pools
- Multi-GPU –as-a-Service
- Kubeflow (tech preview)
- UI – Kubernetes, API
- Security (policies, encryption)

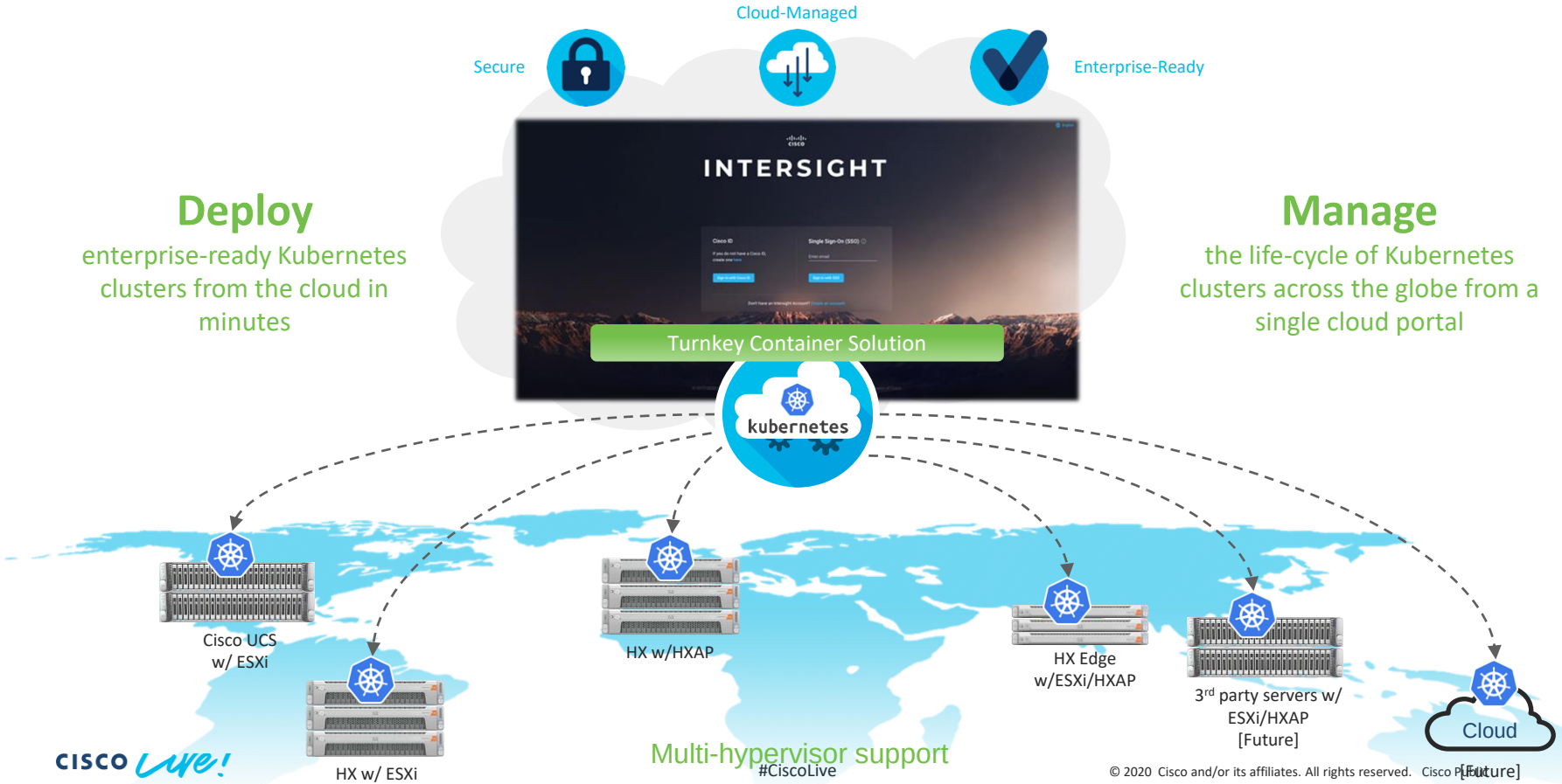
Manage

- Add / remove Kubernetes nodes
- Lifecycle management (OS updates, Kubernetes upgrades)
- Prometheus/Grafana Monitoring
- EFK Logging
- Self-healing Kubernetes clusters
- Multi-master nodes

Multi-GPU as a Service



Cisco Intersight Kubernetes Service (formerly CCP)



Deploy

enterprise-ready Kubernetes clusters from the cloud in minutes

Manage

the life-cycle of Kubernetes clusters across the globe from a single cloud portal

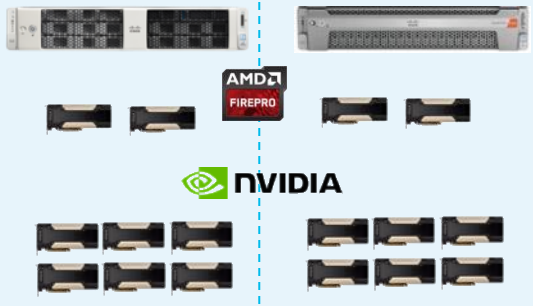
Multi-hypervisor support

UCS AI/ML Portfolio Compute:

UCSM and Intersight Managed | Cisco IMC | XML API | python SDK

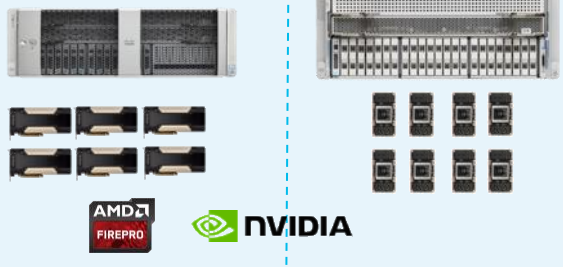
Test & Dev and Model Training

C240 | HyperFlex 240



Deep Learning/ Training

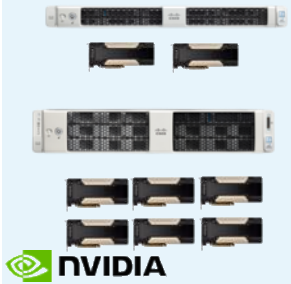
C480 | C480 ML



Inferencing

C/HX 220
C/HX 240

AMD FIREPRO



Validated AI/ML SW For Turnkey (Working with Partners)

NVIDIA | ANACONDA | HORTONWORKS POWERING THE FUTURE OF DATA | cloudera

Better Together, Customer Choice, Cisco Validated Design with Eco-system

Opportunities to look out for

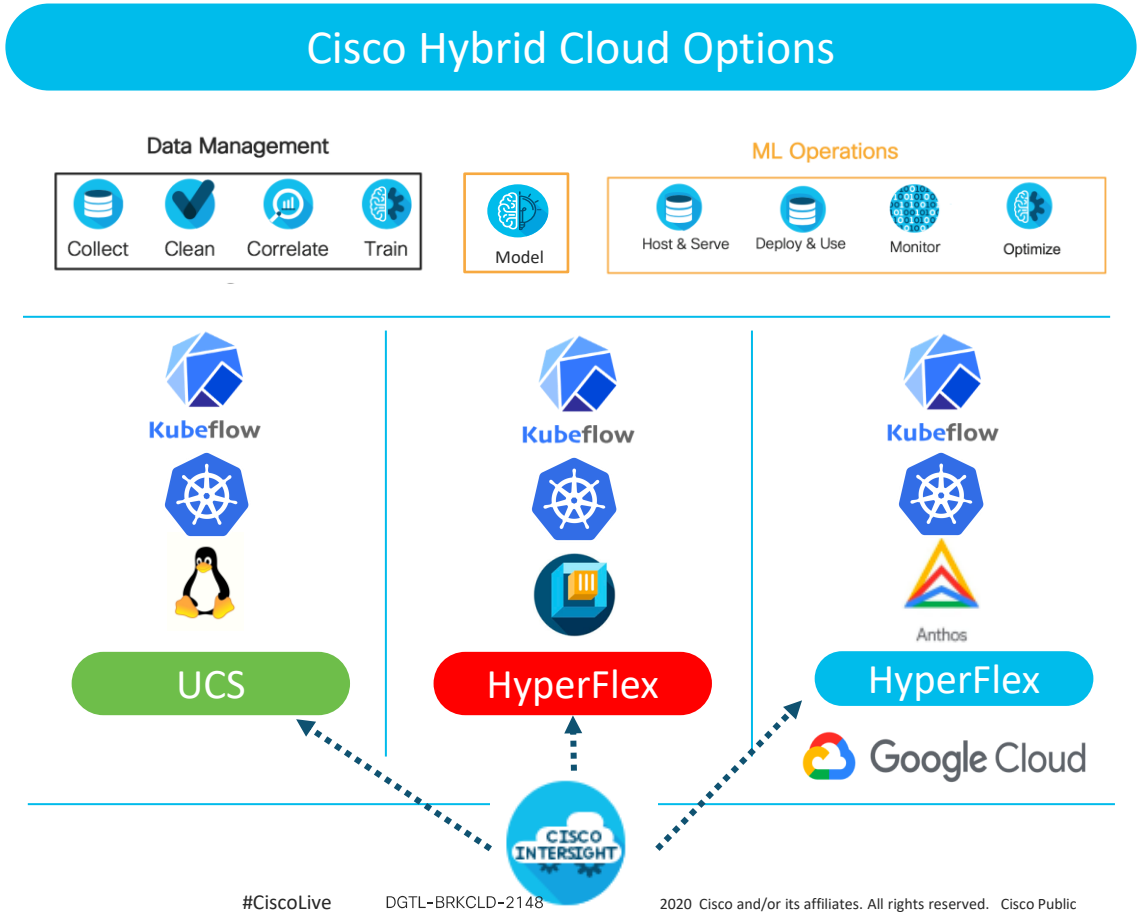
Consistent Cloud and On-Premise Machine Learning Experience

Cloud

- Fast deployment
- Test-dev
- Simplicity
- APIs
- One time

On Prem

- Data gravity and integration
- TCO and Performance
- Governance
- Remodel, retraining
- At scale production



Dedicated vs. Shared Infrastructure

Data Scientists Experience

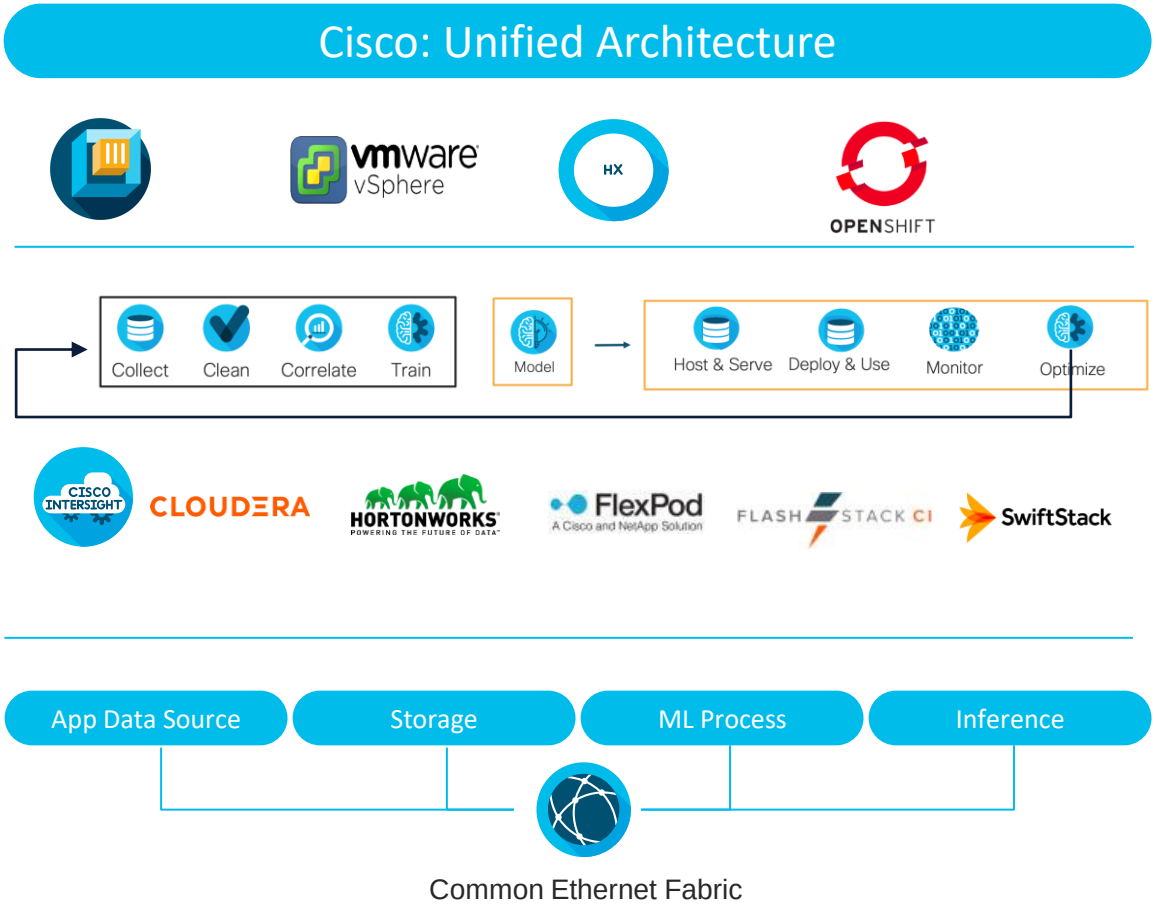
- Dedicated workstation vs. Share ML infrastructure

Data Center Architecture

- Silo ML infra vs. Integrated Data Center Architecture

Networking Architecture

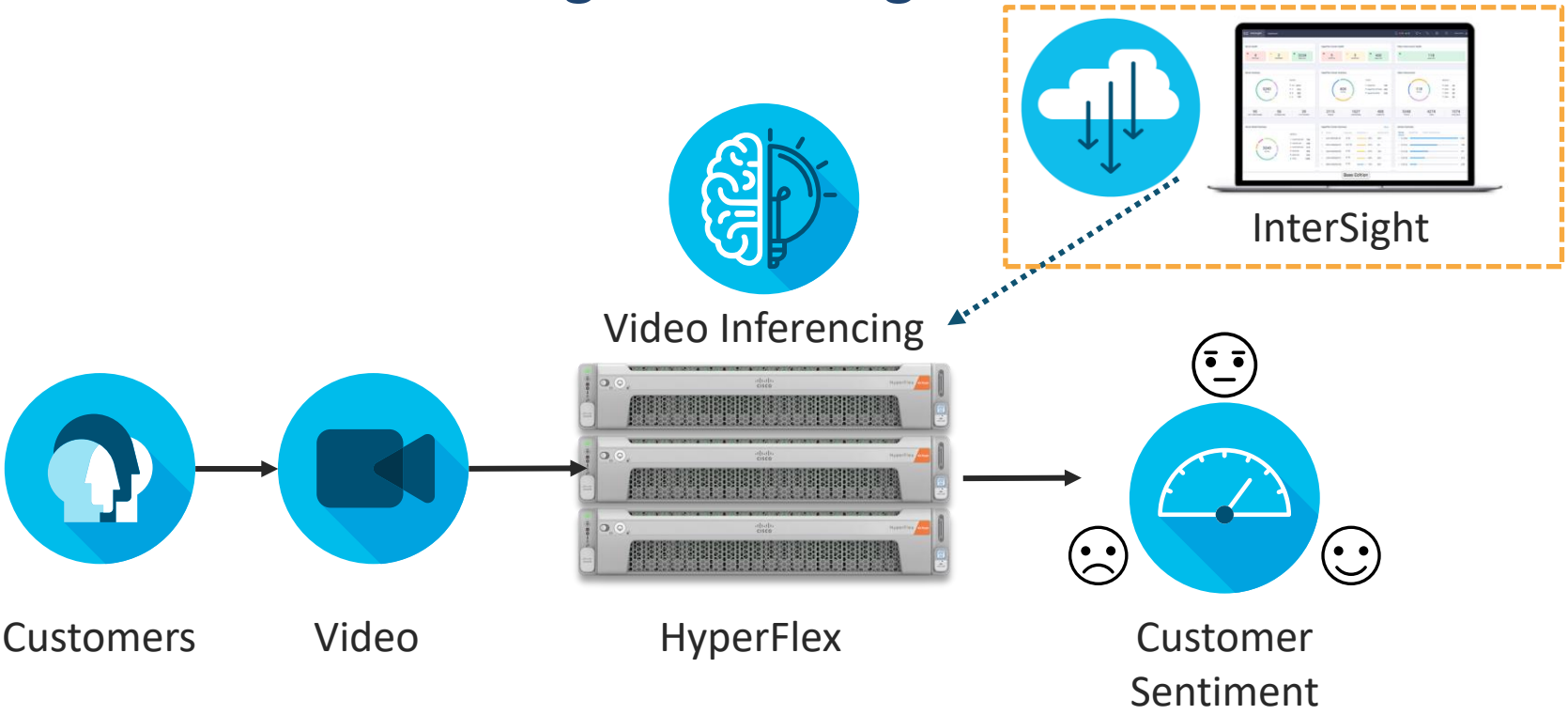
- InfiniBand vs. Scalable Ethernet



https://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/big-data/solution-overview-c22-742735.pdf


The new frontier….The Edge!
cisco Live!

HyperFlex for Inferencing on the Edge



Resources

- Cisco AI/ML Page
  https://www.cisco.com/c/en/us/solutions/data-center/artificial-intelligence-machine-learning/index.html#~resources

- Recent NVIDIA blog re: a100
  https://blogs.cisco.com/datacenter/nvidia-a100-gpu-maximizes-ai-infrastructure-for-cisco-customers

- Enterprise Study: Accelerating theAI Journeywith Cisco
  https://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/unified-computing/esg-wp-cisco-ai-ml-ucs.pdf

Thank you

Possibilities

#CiscoLive