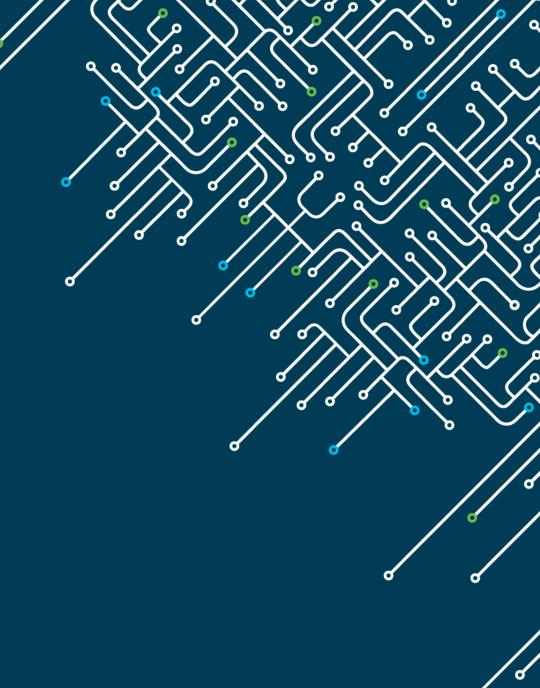


UCS AI/ML Update

Ravi Mishra Technical Marketing Engineer May 2019



Artificial Intelligence, Machine Learning, and Deep Learning (AI/ML/DL)

Artificial Intelligence

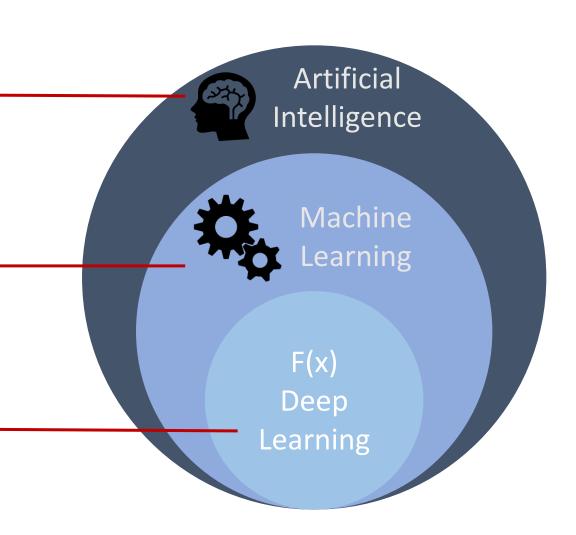
Technique where computer can mimic human behavior

Machine Learning

Subset of AI techniques which use algorithms to enable machines to learn from data

Deep Learning

Subset of ML techniques which uses multi-layer neural network to learn



Examples of Artificial Intelligence, Machine Learning, and Deep Learning

Artificial Intelligence

Play Chess: Find all moves and prune off

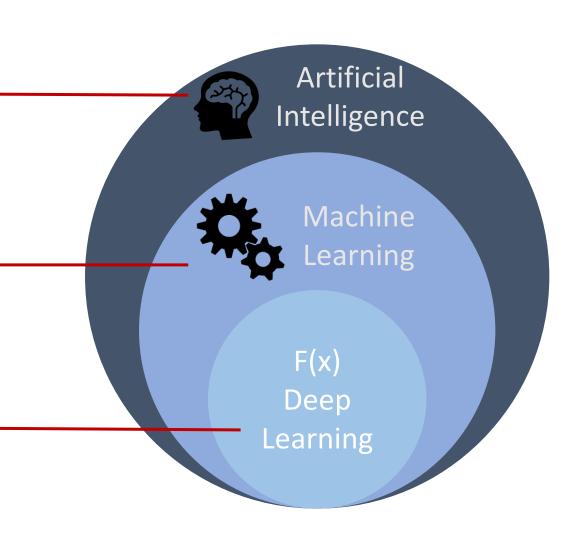
bad moves

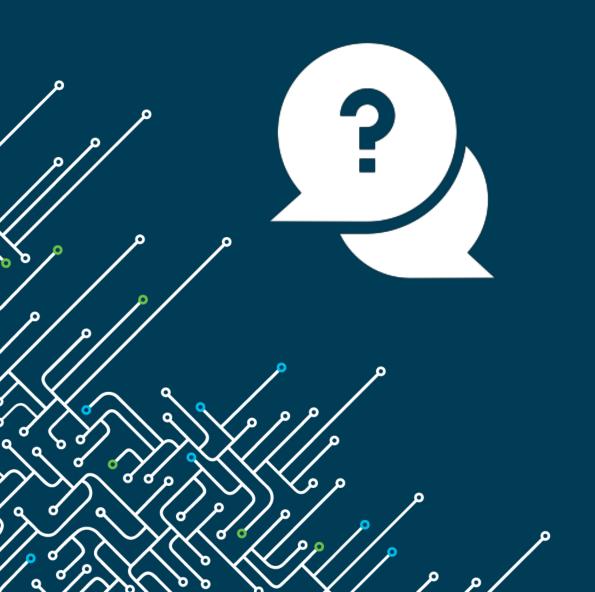
Machine Learning

Based on size and price of neighboring houses, predict price of new house on market

Deep Learning

Given lots of images of cats and dogs, is there a cat or dog in a new photo?





What are the use cases of AI/ML for our customers?

Business Drives Al













Duainaga	
Business	
Question	

Expert Al Task Finance

Healthcare

Media and Entertainment Security and Defense

Retail

Manufacturing

Is "it" present or not?

Detection

Identify Access Indications of Anomalies Anomalous Care

Content Based Search Identify Security
Breaches

Identify Events in Store Surveillance

Detect Manufacturing Flaws

What type of thing is "it"?

ng Classification Fraud detection

Medical Imagery
Diagnostics

Content Labeling

Facial Recognition

Identify Returning vs New Shoppers

Enable Robots to Track Objects

is "it"?

To what extent is

nt is ? Segmentation

Sentiment Analysis Condition Analysis

Improved Product
Placement

Crowd Analytics

Segment by Customers Actions Sort Components by Quality

"it" present?

What is the interpretation?

Natural Language Processing

Chatbot Advisors

Expert Diagnosis from Notes

Video Captioning

Real Time Language Translation In Store Personal Assistants Assembly Build Instruction Translation

What is the likely outcome?

the objective?

outcome? Prediction

What will satisfy Recommen-

ecommendations Algorithmic Trading

Credit Profiling

Length of Stay
Forecasting

Treatment
Recommendations

Targeted Content
Generation

Effective Content
Recommen-

dations

Equipment Health
Assessment

Risk Management

Customer Churn and Retention

"Magic Mirror"

Maintenance
Assembly Process
Improvements

Proactive Machine

Finance Use Cases



Detect Fraud

Financial transactions are now performed 24 hours a day from a variety of devices on a global scale. Being able to stop fraud without negatively impacting a customer's ability or experience is an essential capability for all financial institutions that requires AI.



Create Competitive Advantages

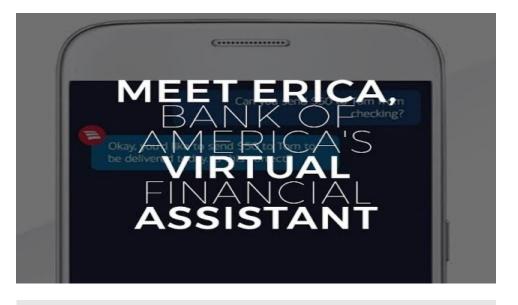
Al allows financial institutions to create intelligent applications that provide customer experiences that are powerful yet intuitive. Chatbots, for example allow customers to ask questions, like "How much was my vacation to Disneyland?"



Grow Without Increased Costs

Through the super-human capabilities of AI, financial institutions are finding insights in more data, engage more customers, all without adding more headcount. Virtual assistants help employees and customers find instant, accurate answers to their questions.





- Separate Fraudulent Claims from Valid Claims
- Financial Data Mining
- Credit Card Churn Analysis
- Stock Selection
- Assess Stock Prices
- Identify Fraudulent Claims
- Stock Performance Predictions
- Loan Default Prediction
- Smart Indexing
- Automated Document Processing/OCR
- Signature Recognition
- Chart Analysis
- Client Facing Chatbots
- High Frequency Trading

Healthcare Use Cases

Remove Distractions

Healthcare environments deal with complicated data, many devices, alerts and most importantly people and the uniqueness of how individuals communicate. All separates what is important from the noise so that people can get required information quickly.

Simplify Diagnoses

Al scans through/processes massive amounts of medical images, medical literature and/or medical records/past cases to help doctors diagnose conditions accurately and plan treatments. Al can detect even the smallest abnormalities to save lives.

Predict Challenges

Al can predict conditions before they become apparent. Al spots abnormalities in sensor data and the way people move, speak, or look among other characteristics, before people can. According to the CDC, 40% of all deaths by the 5 leading causes could be prevented.





- Discovering Disease Subtypes
- Disease Progression Modeling
- Condition Diagnoses
- Clinical Decision Analysis
- Treatment Selection
- Healthcare Cost Modeling
- Patient Readmission Reduction
- Predict Length of Stay
- Medical Imagery Segmentation
- Computer Aided Diagnostics
- Medical Imagery Diagnostics
- DNA Sequence Binding Analysis
- Early Identification of High Risk Conditions
- Model Usage of Beds
- Recommend Billing or Diagnostic Codes

Security and Defense Use Cases



Protect from Bots

Malicious bots can wreak havoc. From DDoS attacks to stealing of sensitive data, bots are a serious security threat that are getting more sophisticated. Al enabled bots are able to easily break safeguards like CAPTCHA techniques, so it takes Al to stop Al.



Video Analysis

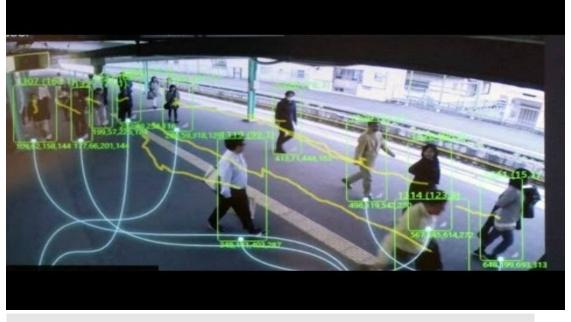
It can take a large number of cameras to safely cover a moderate area. People are unable to watch every video stream and track every object within video streams. All is used to "see" everything across large numbers of streams to bring attention to threats.



Intelligent Authentication

Protecting digital assets and identities can be challenging, especially when balancing security with convenience. Al can see even the most obscure patterns for authentication. Al is used to identify people by voice, typing rhythm and human motion.





- DDoS Defense
- Detect Bots
- Identify Abnormal Network and System Usage
- Identify SPAM vs HAM Email
- Malware Detection
- Predict Crime and Optimize Patrolling
- Profile Cyber Crime
- Defense Spending Analysis
- Network Intrusion Detection
- Adversarial Defense
- Computer Aided Radar
- Camera to Camera Identity Tracking
- Identify Unsafe Areas from Satellite Images
- Electricity Theft Detection
- Network Risk Scoring

Manufacturing Use Cases

Improved Product Design

Product designers use AI to scan massive amounts of images, product data and social media to create products that customers desire, are easier to use and are more durable.

Optimize Warranties

Manufacturers use AI to identify and model product usage to provide warranties that allow for the longest duration and coverage without impacting profitability.

Predictive Maintenance

Al allows retailers to use dynamic pricing that is adjusted to make sales based on the location, characteristics and habits of customers. Even after a purchase is made, Al can make offers for accessories, warranties or other related items that are logical.



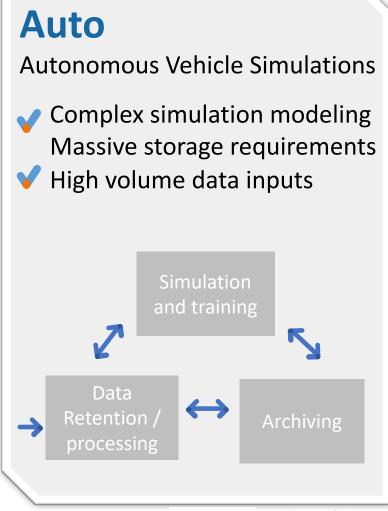


- Control Rules Analysis
- Supply Chain Analysis
- Sort Components by Specification
- Proactive Equipment Health Monitoring
- Manufacturing Process Selection
- Production Cost Estimation
- Quality Predictions
- Defect Detection
- Sort Components by Quality
- Enable Robots to Track Objects
- Identify Similar Parts for Query
- Catalog Components
- Predictive Maintenance
- Process Optimization
- Manufacturing Instructions Translation

Real use case example

Banking Customer-Centric Marketing Product recommendation Experience personalization Attrition prediction **Existing UCS Big** Al Farm: UCS Data Cluster GPU cluster

Operation Improve customer experience Predicting Failures Automatically Position Spares at Depots Optimizing Supply Chain and Customer Experience Component Failure Prediction Al Platform (for Develop / Train) Communicat Sources achine Learning Libraries



















Why should we care about AI/ML?

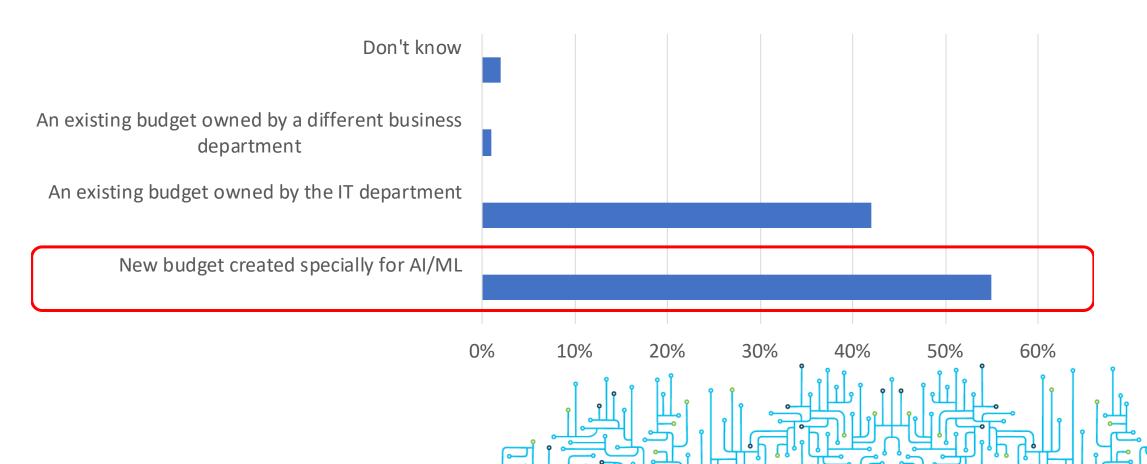
Because our customers care ...

Gartner 2019 CIO survey

Typical rs Performers (n = 2,329)	Trailing Performers (n = 276)	Top Performers (n = 165)	Typical Performers	Trailing
0500			(n = 2,476)	(n = 193)
25%	24%	7%	5%	1%
25%	21%	28%	26%	20%
10%	14%	15%	9%	8%
9%	7%	17%	14%	12%
6%	5%	7%	6%	5%
2%	1%	Not listed in survey for 2018 report		
10%	11%	6%	6%	6%
5%	5%	2%	4%	3%
1%	0%	4%	1%	1%
	10% 9% 6% 2% 10% 5%	10% 14% 9% 7% 6% 5% 2% 1% 10% 11% 5% 5%	10% 14% 15% 9% 7% 17% 6% 5% 7% 2% 1% Not liste 10% 11% 6% 5% 5% 2%	10% 14% 15% 9% 9% 7% 17% 14% 6% 5% 7% 6% 2% 1% Not listed in survey for 2018 report 10% 11% 6% 6% 5% 5% 2% 4%

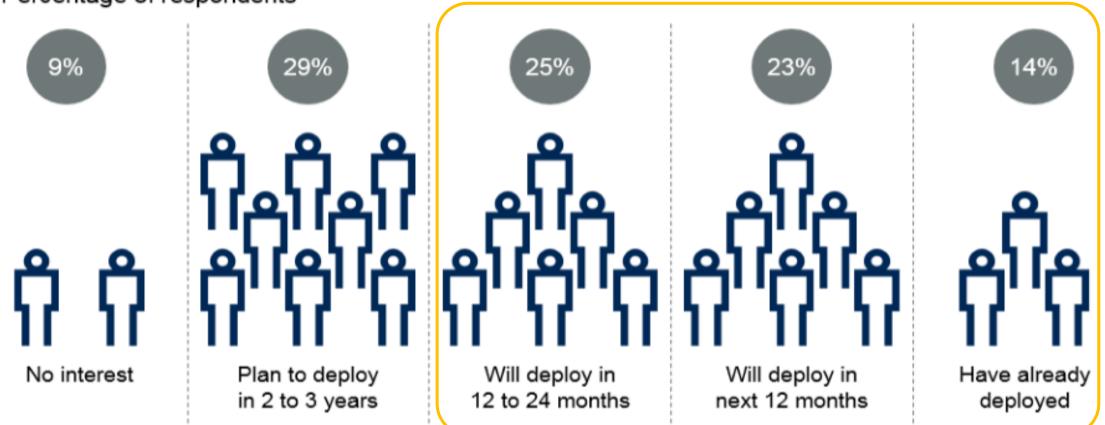
Because it represents new wallet...

ESG 2018 AI/ML Survey: Where does your organization's budget for investments in AI/ML initiatives come from?



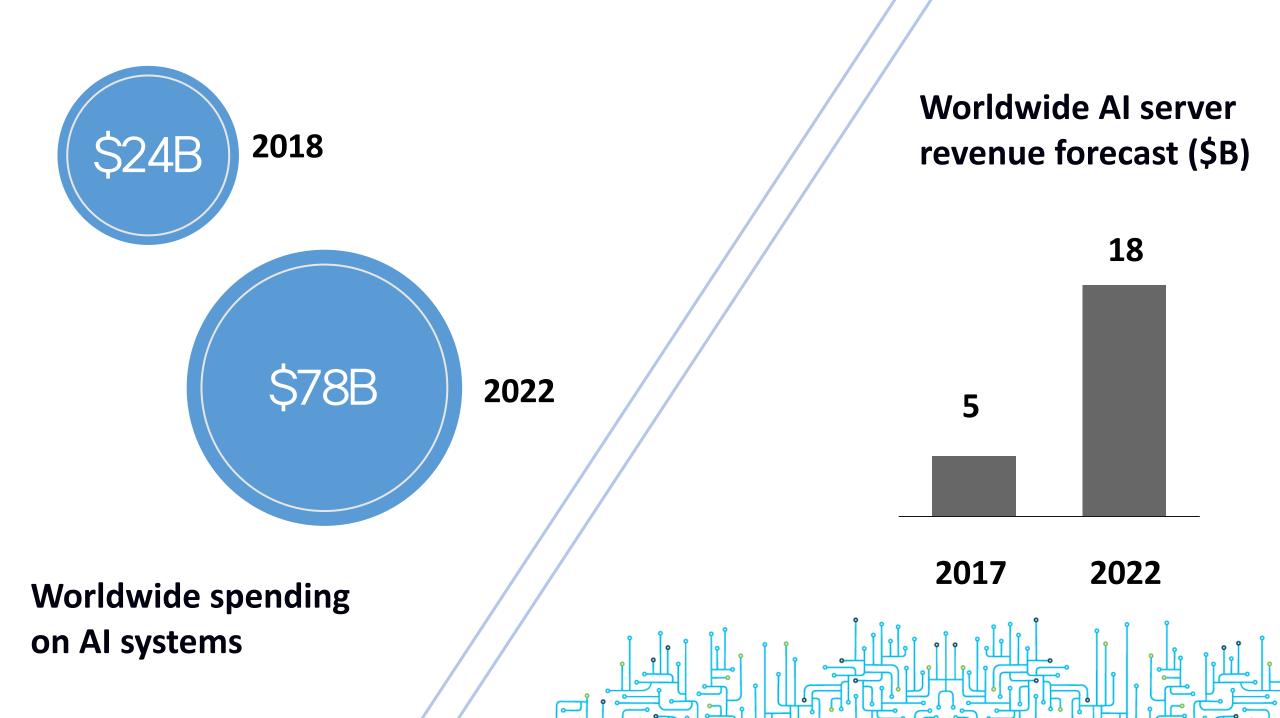
Enterprise Plans to Deploy Al

Percentage of respondents

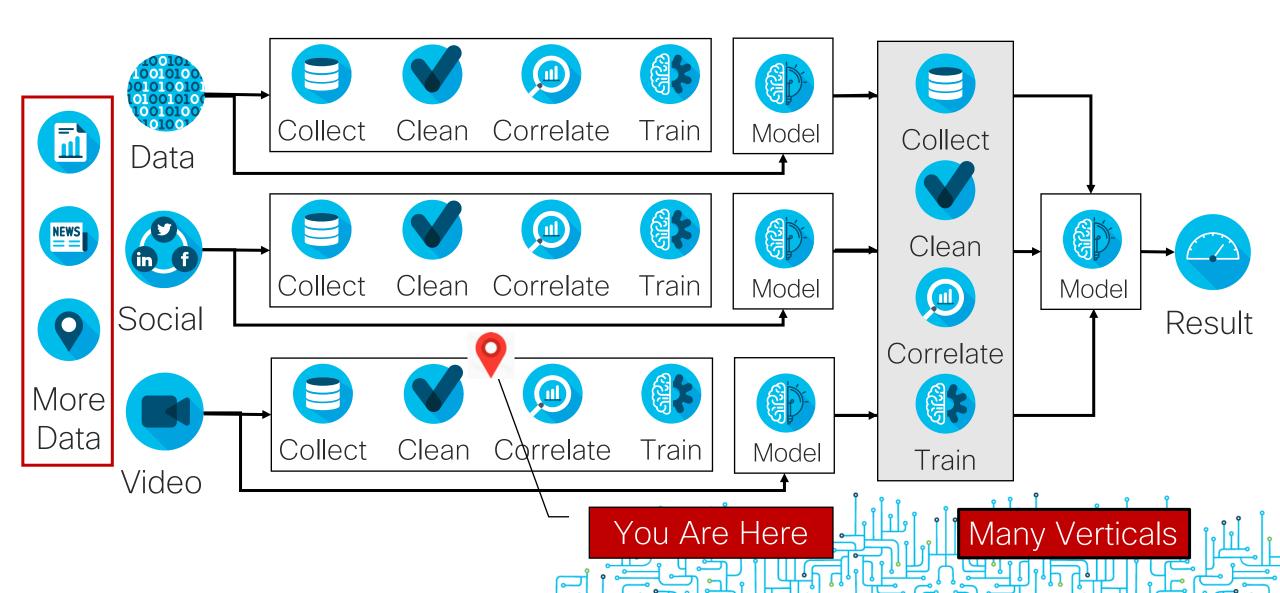


Base: All answering, n = 2,882

What are your organization's plans in terms of artificial intelligence?



Data Pipeline for Multiple Data Sources

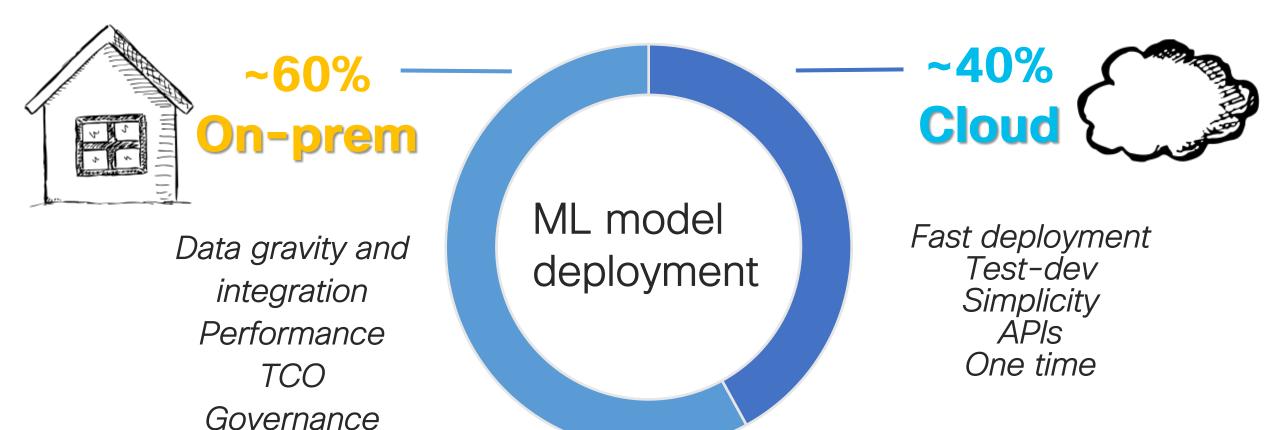




The benefits of both

Remodel, retraining

At scale production





What are the major challenges in AI/ML deployment?

Shift in AI/ML questions

- How to solve the issue?
- How shall I analyze the data?
- What's the best algorithm?
- How to build the best model / most accurate model?

How can we operationalize AI/ML?



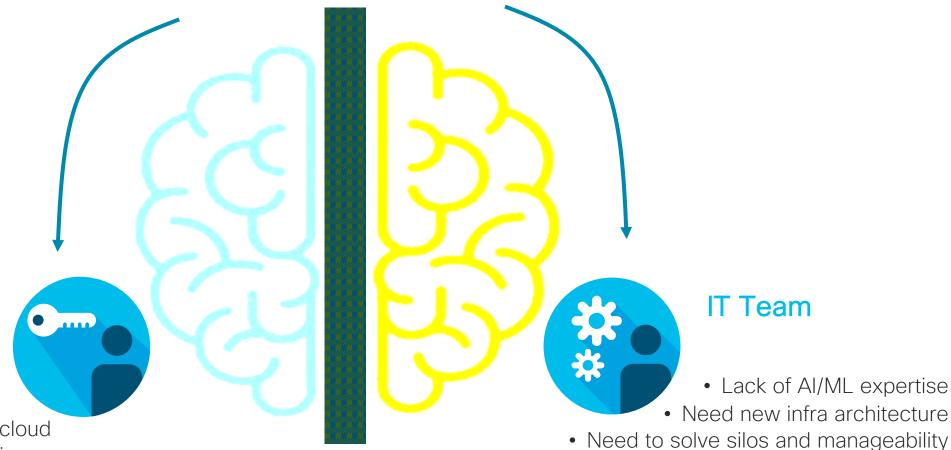
Data Scientists
Data Engineers





Major challenge gaps exist between key stakeholders in realizing ML impact

CIOs Business Leaders



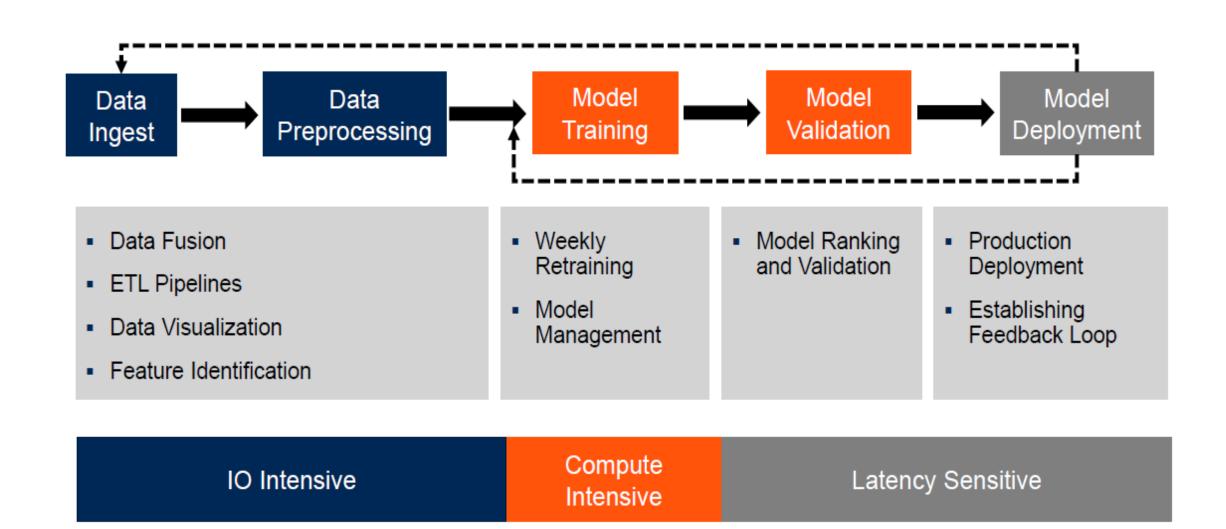
Need enterprise readiness and

security

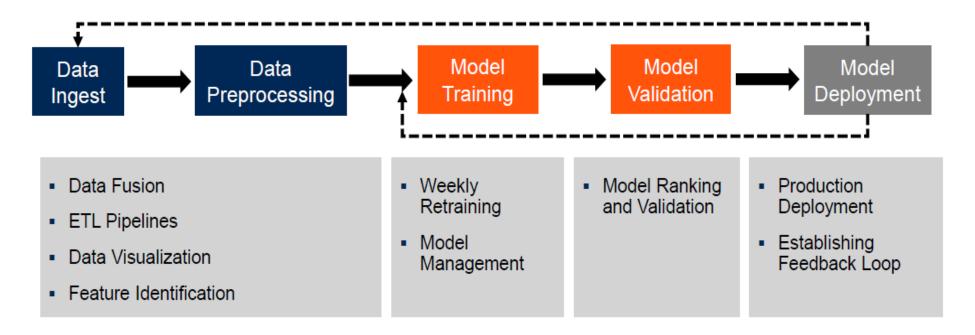
Data Scientists Data Engineers

- Rapidly evolving open source ML frameworks
- · Start with workstation or cloud
- Hard to scale for production
- Wanticloud like experience on premised. Cisco Confidential

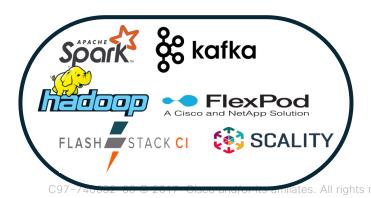
Infrastructure Implications of Al Pipelines



Framework and SW instruments need to be supported



Data Infrastructure



ML/DL Framework / Infrastructure



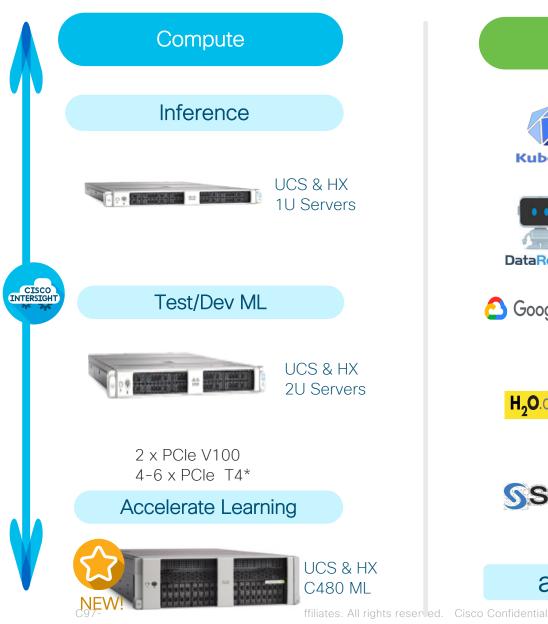
Inferencing & Ingestion End Point

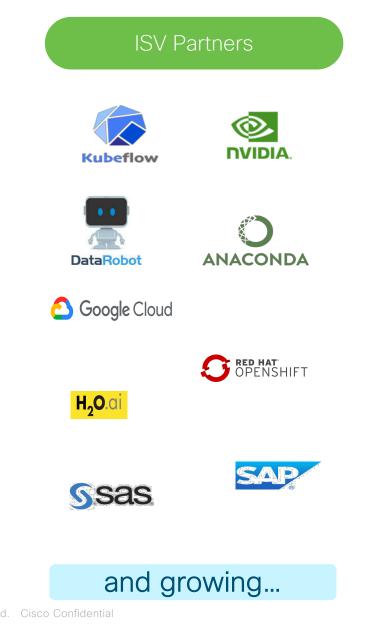




What's Cisco solution?

Cisco Al/ML Computing Platforms & Eco-System Partners







Cisco UCS AI/ML Compute HW Portfolio



Test/Dev.



UCS C240 Servers (with GPUs options)



HyperFlex Nodes (with GPUs options)

Deep Learning



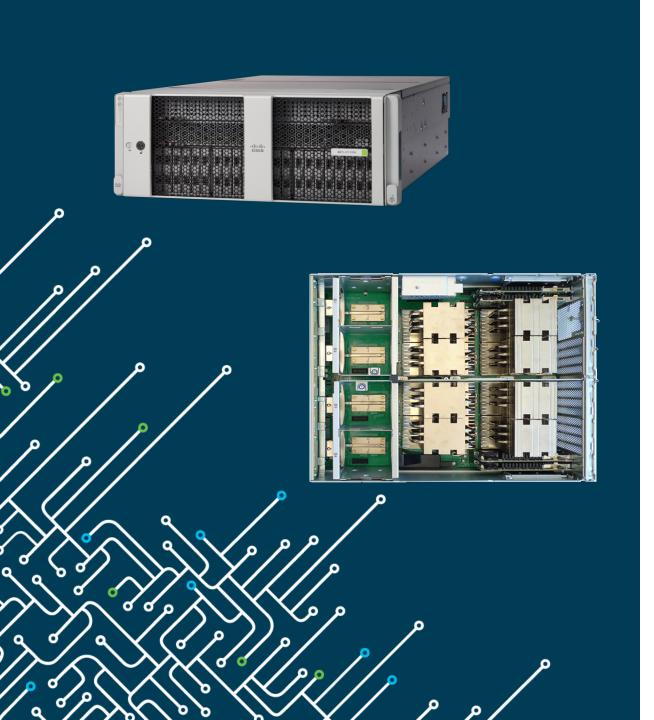
(8 Nvidia GPUs; NVLink)

Inferencing



UCS and HyperFlex Nodes (with GPUs options)

UCS platforms (Big Data, Enterprise Mission Critical, Edge)



UCS C480 ML Rack Server A No-Compromise Purpose Built Server for Deep Learning

Eight SXM2 Nvidia V100 GPUs

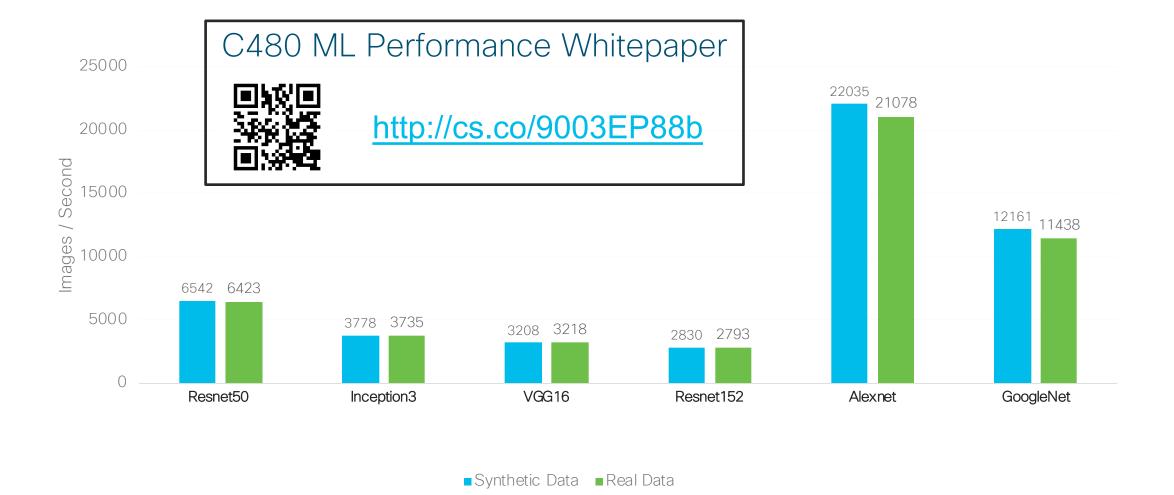
NVIink GPU Interconnect

Intel Skylake Processors

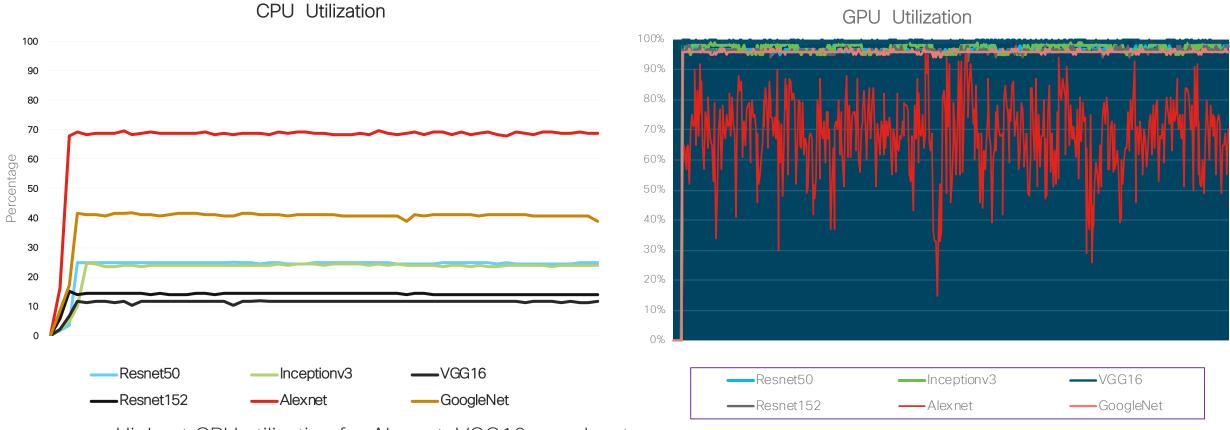
Up to 24 drives with 6 NVMe

UCSM & Intersight managed

UCS C480 ML TensorFlow Training Performance



C480 M5 ML Training: CPU and GPU Utilization



- 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.

CPU Choice matters: C480 M5 ML gives Skylake CPU choices.

Cisco Al/ML Solution Architecture - Bridge the gap

AI/ML SW platforms





RED HAT OPENSHIFT





bitfusion



Big Data Clusters









Virtualization layer





Converged infrastructure Al









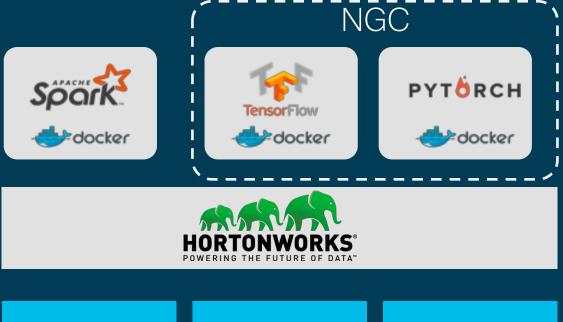






Channel partners & System Integrators

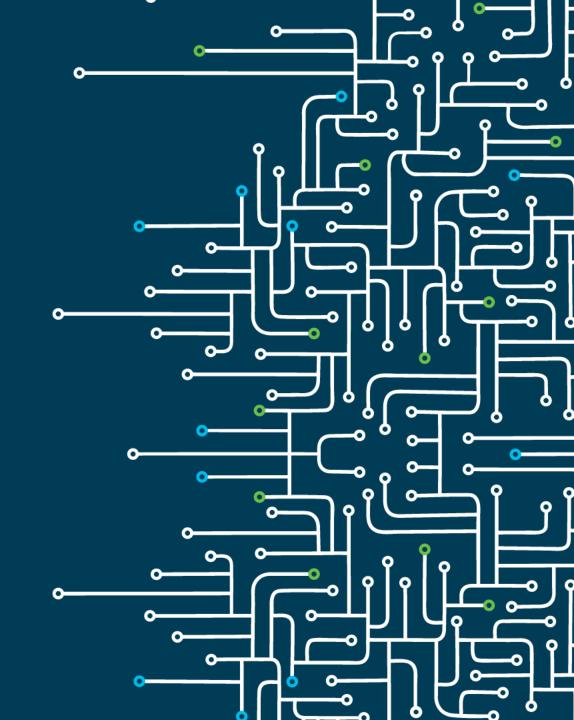
NGC + Hortonworks 3 + UCS



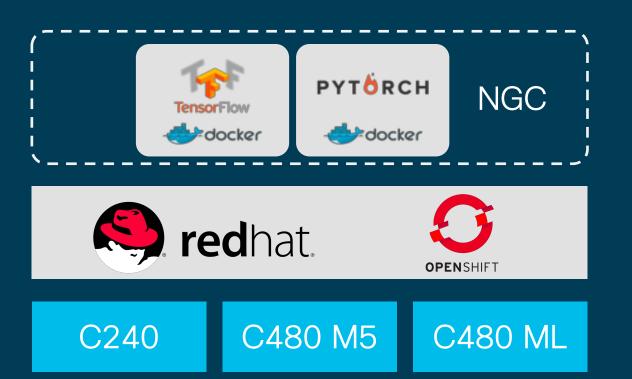
C240

C480 M5

C480 ML

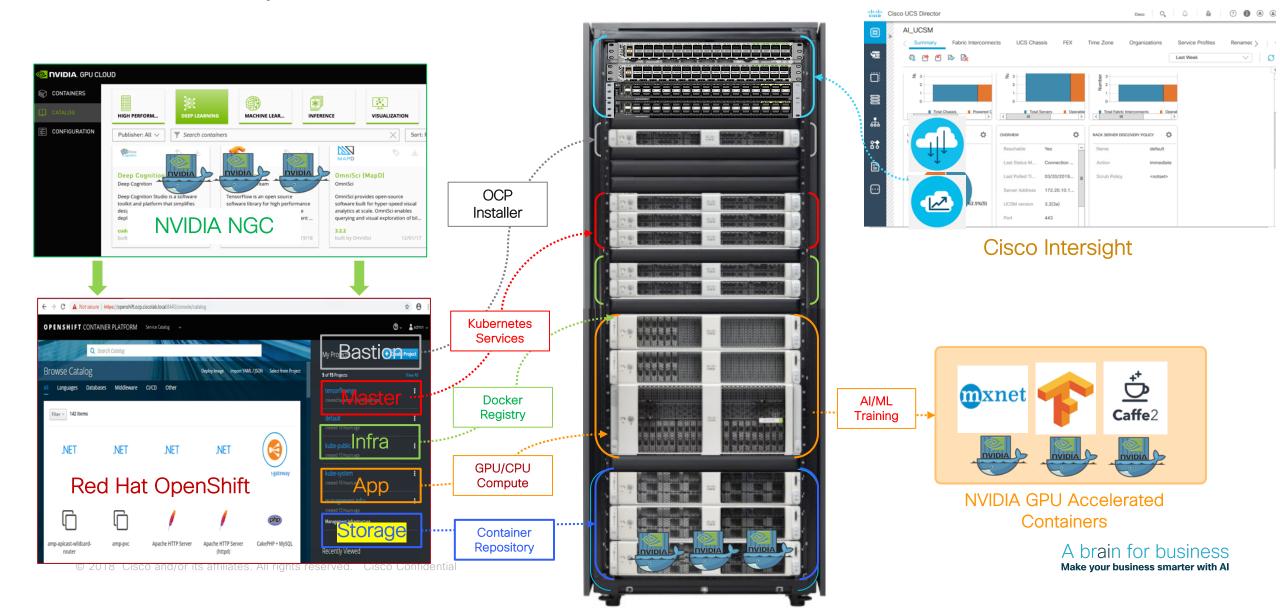






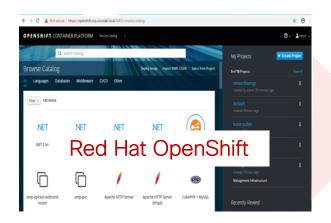


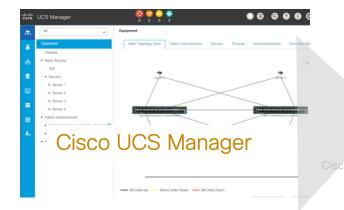
Coming Soon.... Deploying At Scale Red Hat OpenShift Software Architecture With AIML Containers



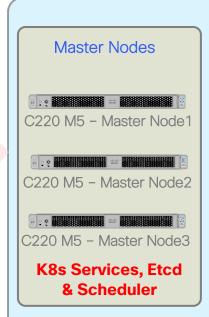
Coming Soon..

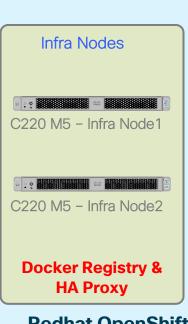


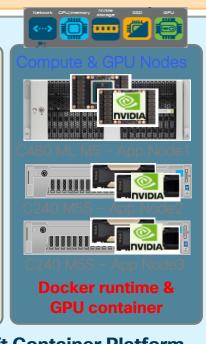


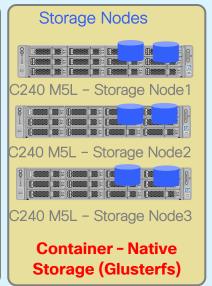








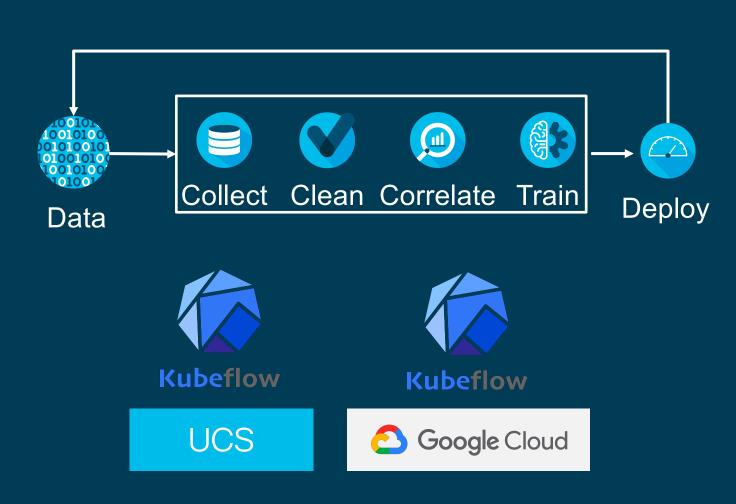


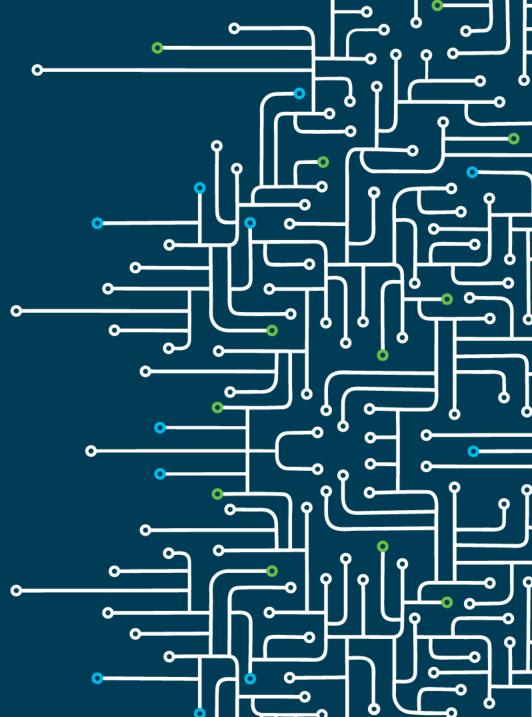


Redhat OpenShift Container Platform



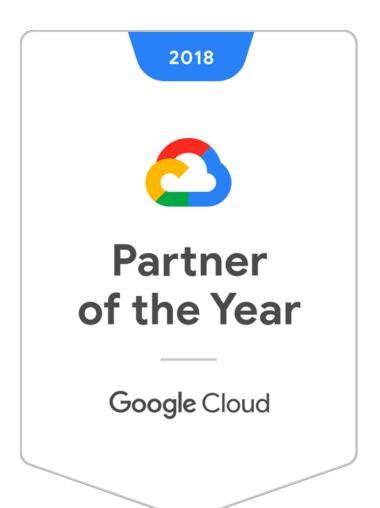
KubeFlow on UCS and GCP



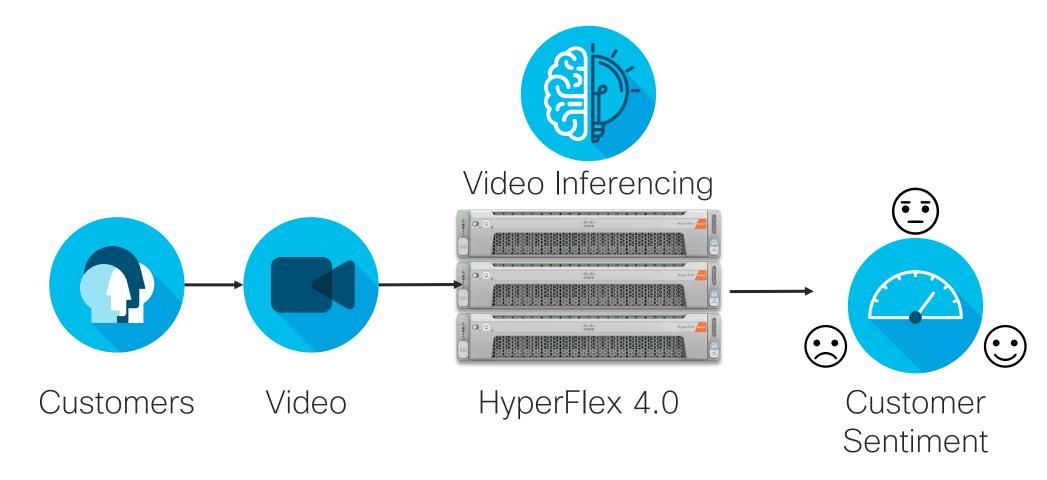


Cisco: One of the Leading Contributors to Kubeflow Over 2.8M Lines of Code with 3 Major Proposals

- #ConsistentAl: 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
- #2 contributor in terms of code



HyperFlex 4.0 for Inferencing on the Edge



cisco



Retail Customer-Experience Intelligence

OpenVINO

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







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

Announcing FlashStack For Al



Cisco UCS C480ML

- Fully Integrated Platform Designed to Accelerate Deep Learning
- Highly dense, High Core Count (1000s of GPU cores)
- Eight NVIDIA Tesla V100s with NVIDIA NVLink Interconnect

Cisco Fabric: State-of-the-art data center fabric switching

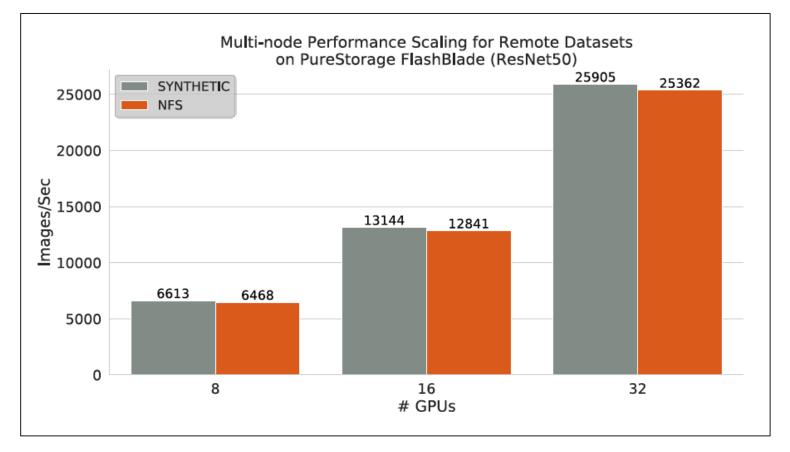
- Fibre Channel or Ethernet/iSCSI storage interconnect
- Massive Network bandwidth with Nexus switches- 4X100GB.

Pure FlashBlade Storage: Fast, reliable and shared storage

- Designed for AI and ML with all-flash storage in 4U form factor
- Scale out storage from 190 to 1607 TB
- Delivers elastic performance as data grows

Flexible deployments for the ML and DL data pipeline

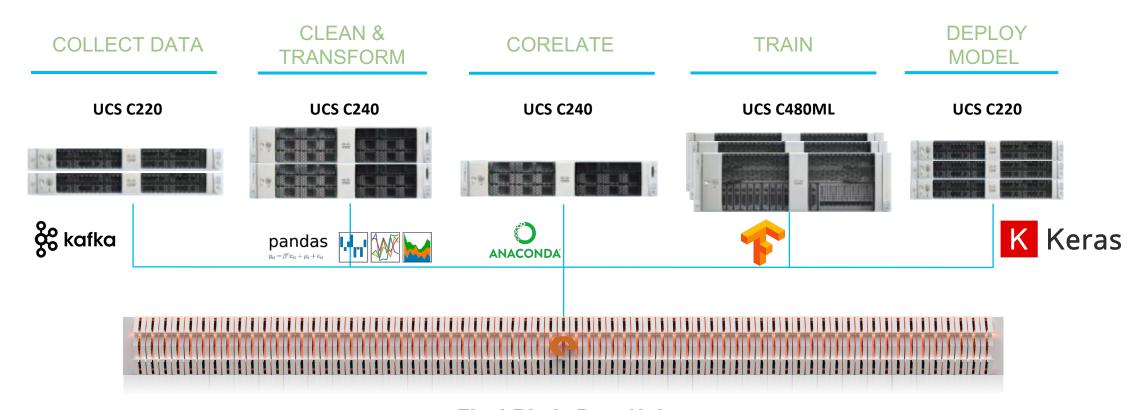
Multi Node Performance White Papers...



- C480 ML + FlashBlade system Scaleout AI, throughput is nearly identical between training on remote datasets saved to FlashBlade and training on synthetic data (generated right on the GPUs).
- Scaling C480 ML nodes achieved near-linear training throughput for multi-node jobs

NOT STOPPING THERE.. NEXT STEPS: SIMPLIFY ENTIRE DATA PIPELINE*

Powered by FlashStack for Al



FlashBlade Data Hub
Share Data Not Just Store them

FlexPod AI: Platform for Innovation

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



Cisco Nexus 9K

High speed fabric

Cisco UCS FI

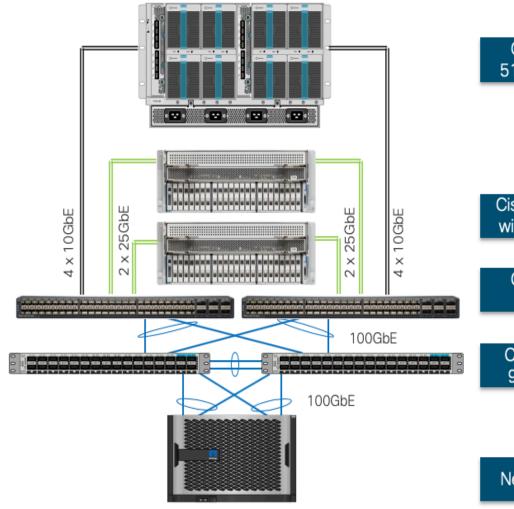
Unified compute fabric and management

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



Cisco UCS 5108 Chassis

Cisco C480ML with VIC 1455

Cisco UCS 6454 FI

Cisco Nexus 9336c-FX2

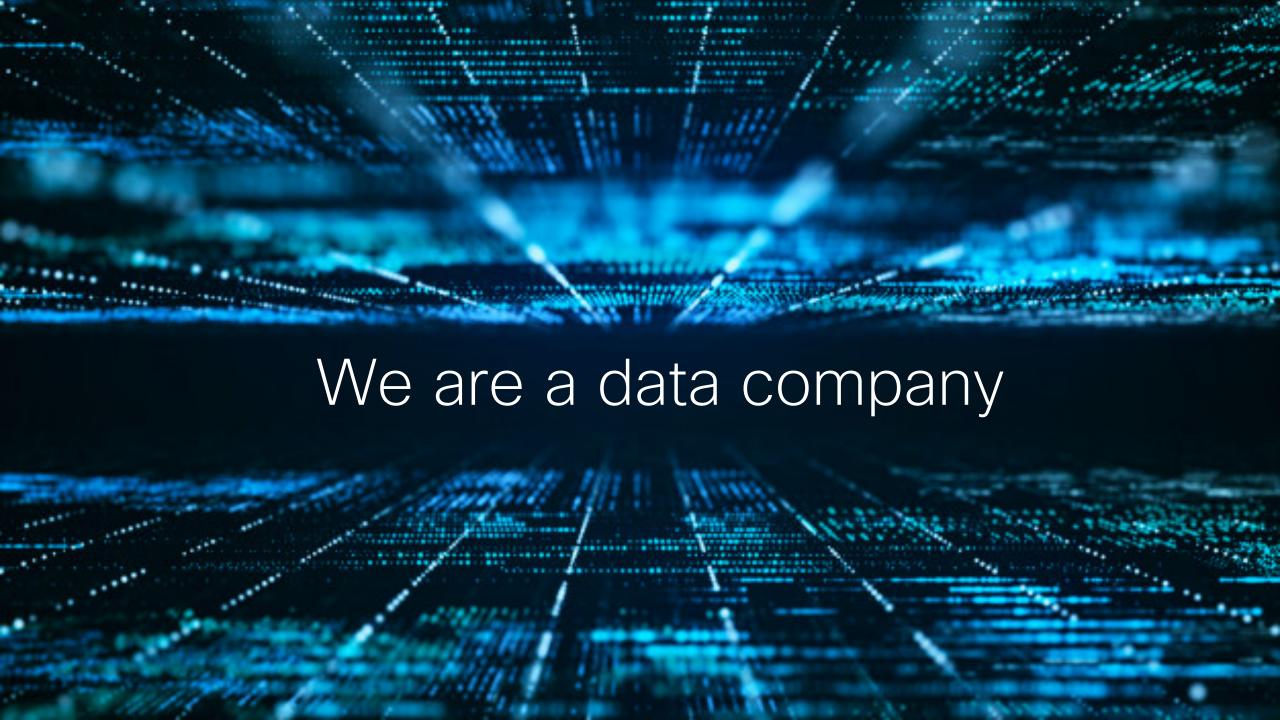
NetApp A800

A brain for business
Make your business smarter with AI



Why Cisco?

How do we do comparing to other players?









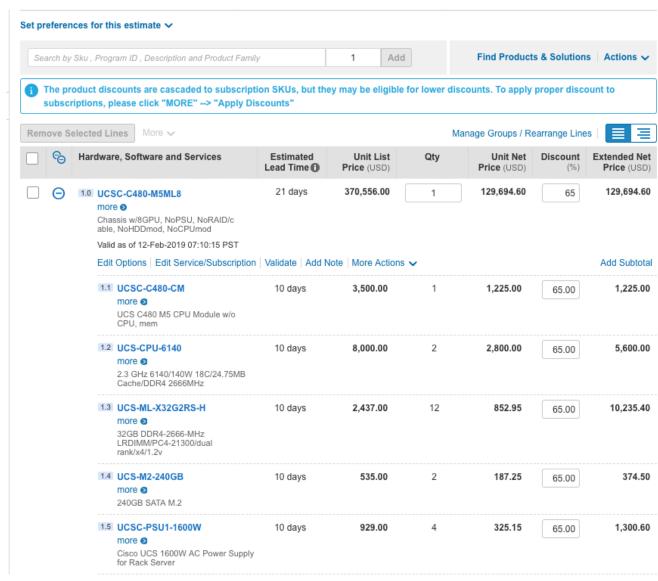
k k 1	best thermal design	HPE Oneview SW, Apollo 6500 and	1U 2CPU 4GPU box lower footprint Customers complains about thermal and melt down
Hign density GPU servers	UCSM Intersight mgmt	OEM SuperMicro box	Not ideal CPU:GPU ratio
	End to end portflio, HX CPU and GPU offerings PCIe and SXM2	End to end portflio CPU and GPU offerings PCIe and SXM2	End to end portflio CPU and GPU offerings PCIe and SXM2
Enterprise experience	Strong	Strong	Strong
H	NGC, Google, Anaconda, H2O, Big data softwares, Openshift	Acquired BlueData	Open
F	Flashstack for AI Flexpod for AI Swiftstack (WIP) Weka.io (WIP)	Weka.io, own storage	EMC



What resources do we have to support our partners?

How to order C480ML

- Order Rack level PID
 - UCS-C480-M5ML8
- Pick CPU, Memory (12 or 24 DIMMs recommended)
- Select HDD Drive Module and storage drives
- 8x V100 GPUs are added by default
- Select VIC 10/25 or 40/100G



New C480 ML BOOST Offering

ONLY FOR FIRST

50 Units

(MAX 2 SERVERS PER CUSTOMER)

SPECIAL PROMOTION
ONLY FOR Q3 FY19

UCS BOOST Promotional offer is to accelerate UCS C480 ML customer acquisition by showcasing the value of our AI ML architecture offering.

Reduce sales cycle time with the best, pre-adjusted, up-front pricing available from Cisco

Promotional offer and pricing align with all Cisco channel partner incentive pricing strategies and policies, creating a minimal learning curve for partners.

NET STARTING PRICE LOWER THAN \$100K (per server)

Assuming qualified for OIP

Pricing gets more aggressive for AB/HPE competitive programs

Terms and Conditions

	C480 ML BOOST
Quoting End Date	July 27, 2019
Ordering End Date	July 27, 2019 or First 50 Booked Units
Qty Limit	2 Servers
AB/HPE Competitive	Up to 68%
OIP	Up to 65%
TIP	Up to 65%
MIP	Up to 68%
SeedIT	NA
VIP	Yes
DSA	No
Try and Buy	No
NFR (Not for Resale)	No

SKU	CONFIGURATION	GPL in \$US
C480 M5L BOOST UCS-SP- C480MLBOOST	 2x Intel® Xeon® Processor 6140 2.3 GHz 6140/140W 18C/24.75MB Cache 32GB DDR4 RDIMM (12x 32GB 2666 MHz) 4 x 1.9TB 2.5 inch Enterprise Value 6G SATA SSD 8xV100 SXM2 GPUs 	\$264,600

This product is subject to Compliance Hold (Max two servers for qualified new C480 ML customers only. Order must be booked the sooner of (a) the end of Q3 FY19 or (b) within the first 50 server units bookings. Limited number of units). For questions about the hold and customer qualification, please email c480ml-boost-approvers@cisco.com

Smartplay Bundles are Discount Restrict Product (DRP). Contact your Cisco Sales Rep or Cisco Finance (Internal) for T&C

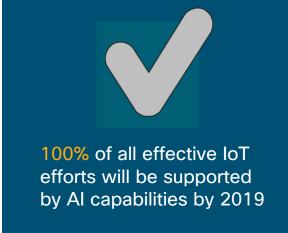
Partner for success - resources

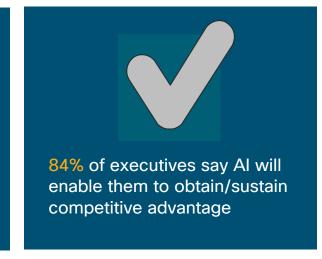


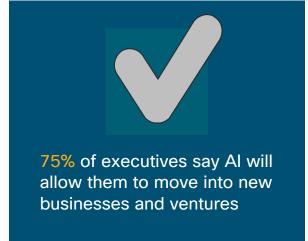
ask-ucs-aiml-tme@cisco.com; Ask-ucs-aiml-pm@cisco.com

Executive Summary Why Al/ML and Why Now









Why Cisco

- Flexible most complete solution from H/W to S/W
- AI/ML starts with the data; storage, connectivity and pipelining (lifecycle)
- Unified Management simplified mgmt., customer choice, validated design

The Opportunity

Creating New Revenue Streams

- Leverage existing buying centers to get to new ones
- Hardware & Services will represent 80-90% of customer spend
- Land & Expand Create new H/W, S/W, Services revenue streams
- Sell with motion with Cisco, Cisco DSI partners and NVIDIA

How Cisco Is Different

Competitive Level Set

- Flexibility in H/W, software and OS configs (Not a fixed appliance)
- Enterprise scale / IT deploying at scale
- Cloud like on-prem capability
- Cisco software partner ecosystem
 - Big Data CVD's with Cloudera and Hortonworks
 - o Partner ecosystem with NGC, Kubeflow, Datarobot etc.

· I | I · I | I · I CISCO