From Packets to Predictions:

Transforming Network from OSI-to-AI

Aavrinder Singh
Rosario Medrano

BRKOPS-1079

About us

Rosario Medrano

Aavrinder Singh

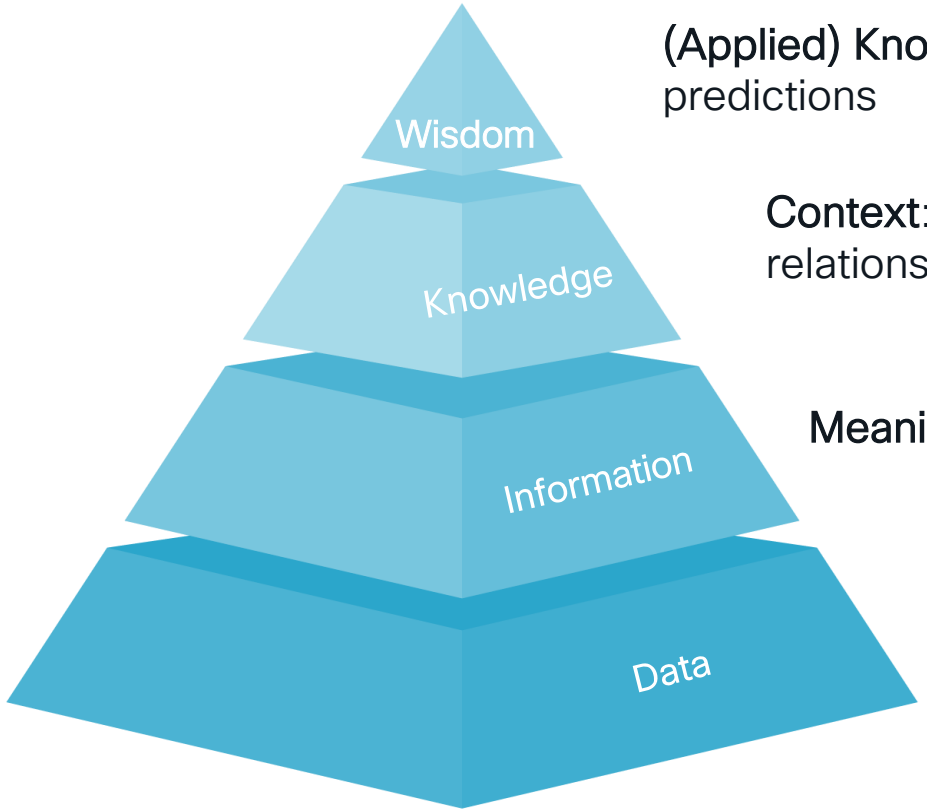
Agenda

- Anthropomorphize
- Understanding AI with OSI Layers
- Connecting the Layers
- Call to Action

Anthropomorphize

How we got here?

The Role of Data



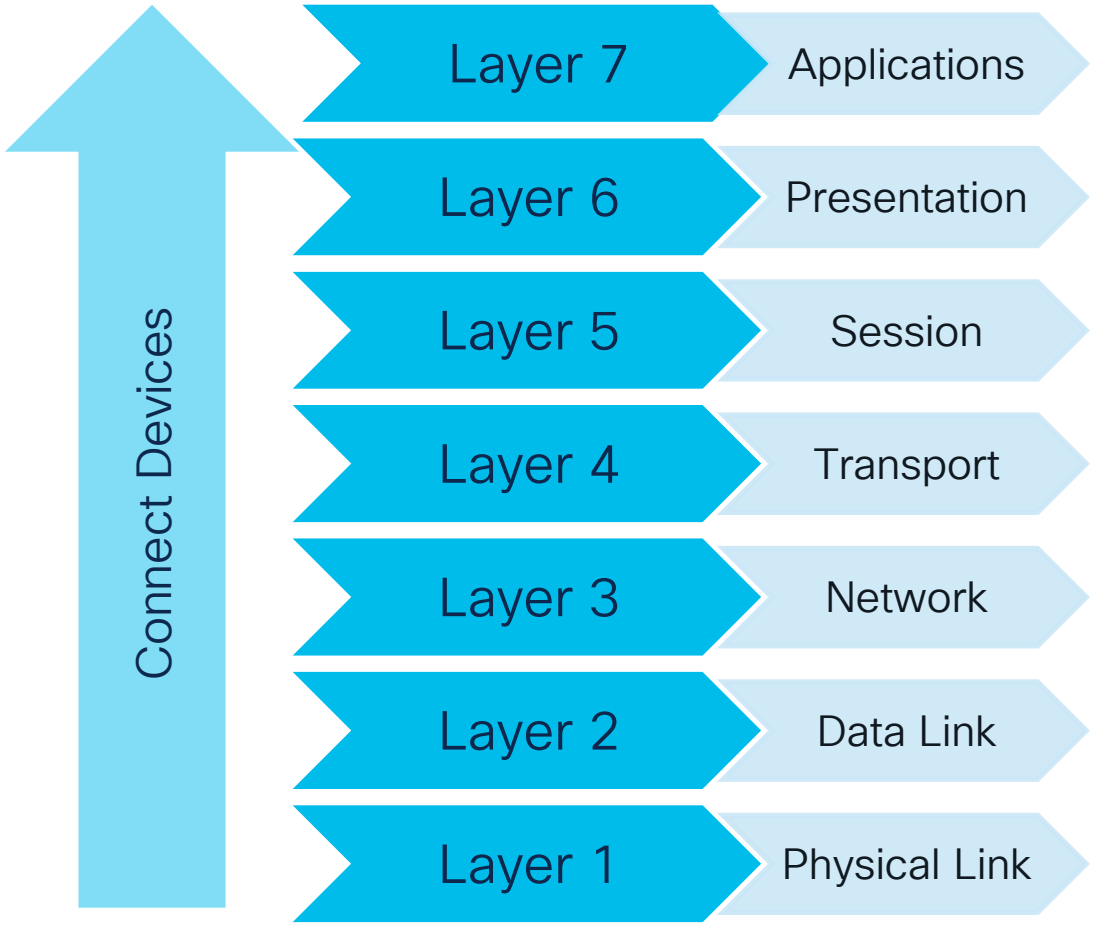
(Applied) Knowledge: Informed decisions or predictions

Context: Understanding patterns and relationships by analyzing information

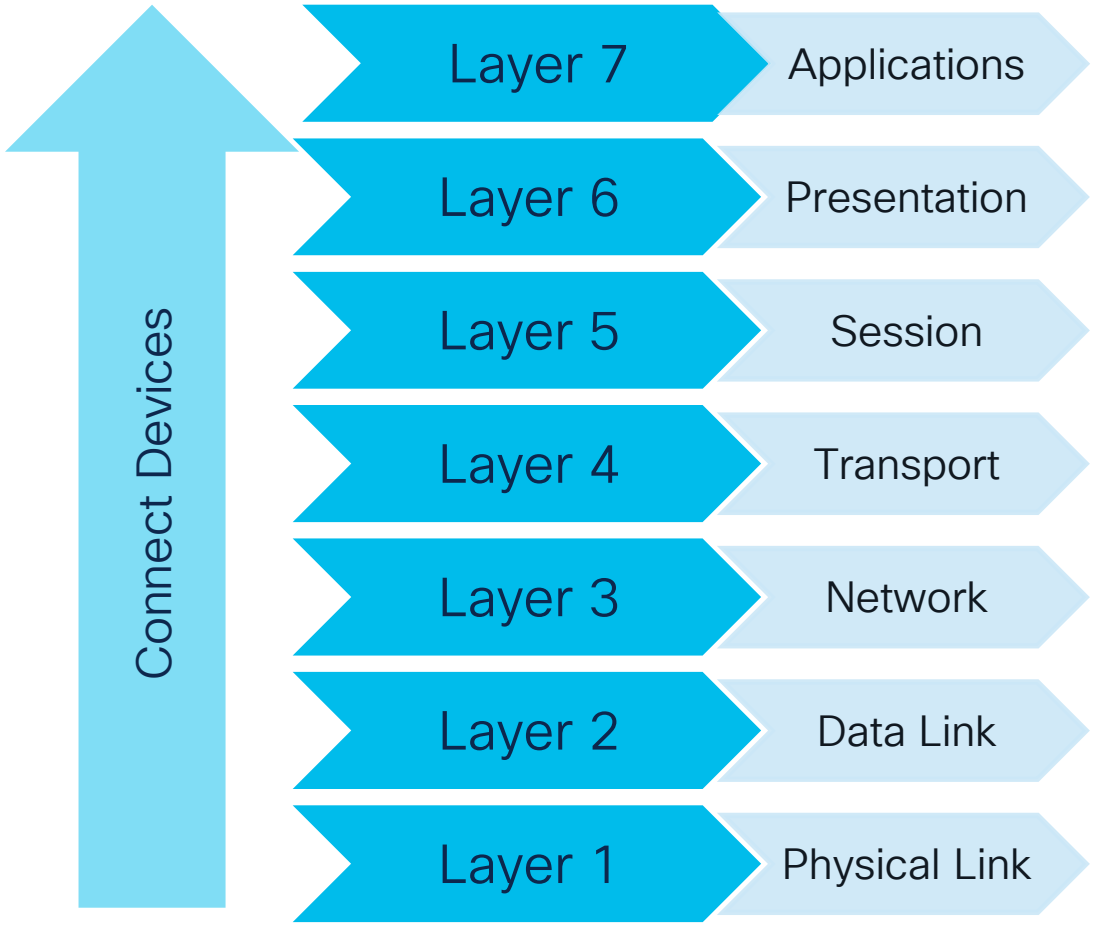
Meaning: Data organized and given context

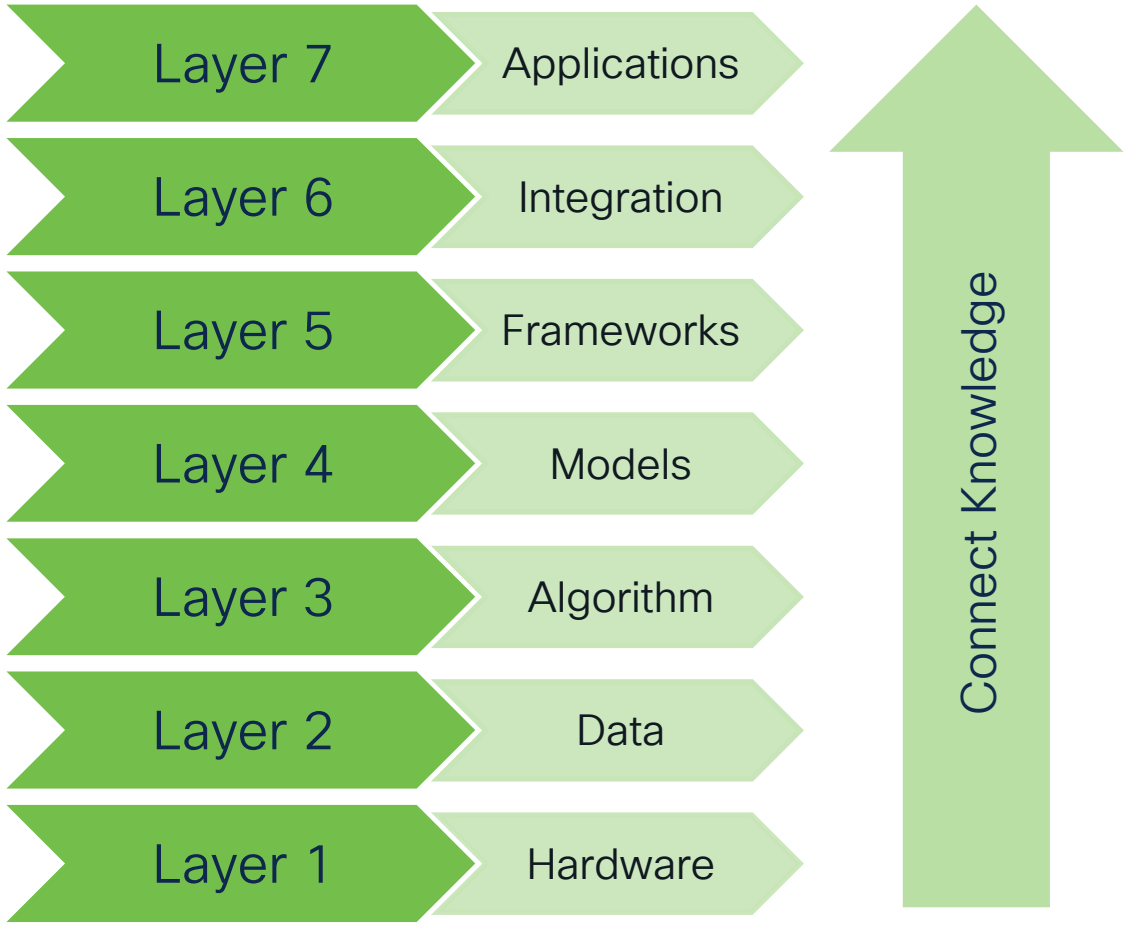
Raw Data: unprocessed facts

From OSI to AI

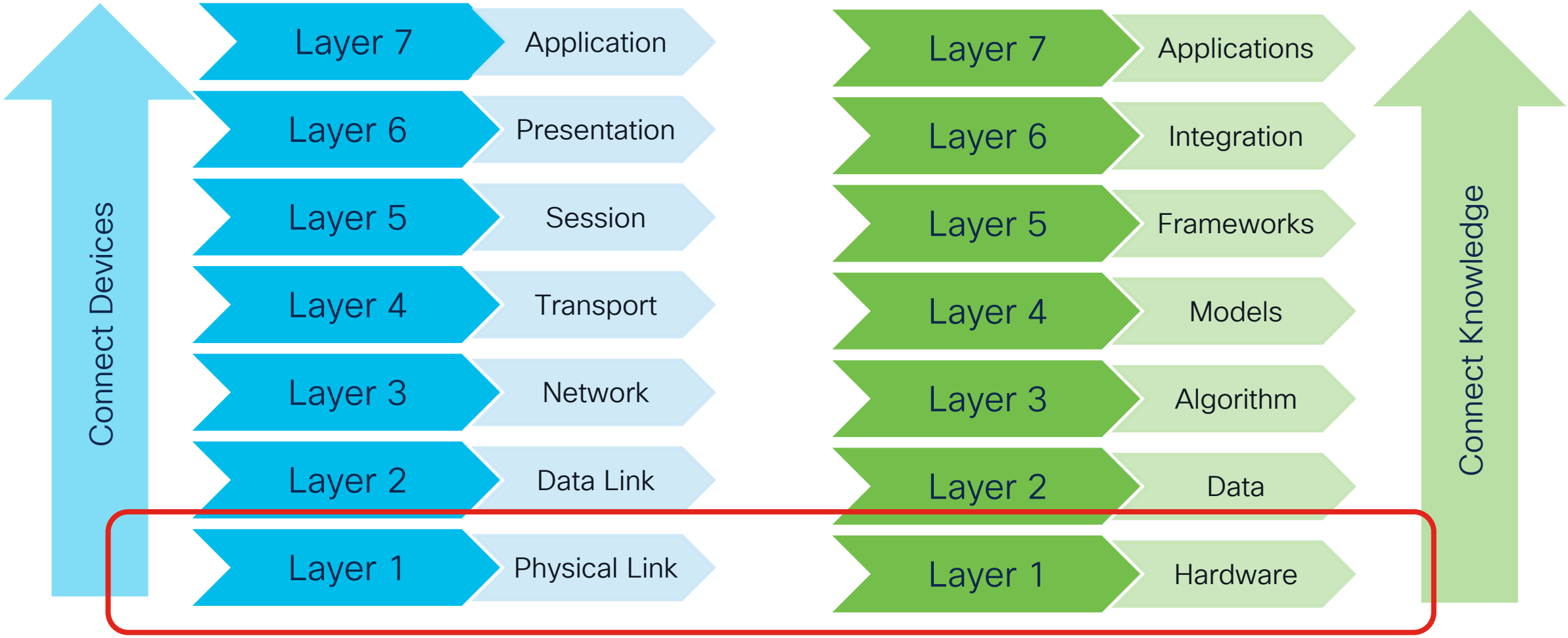


From OSI to AI





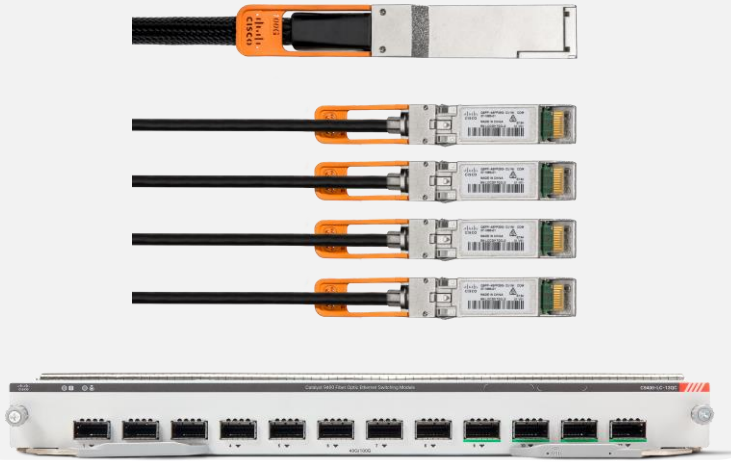
From OSI to AI



Layer 1

OSI

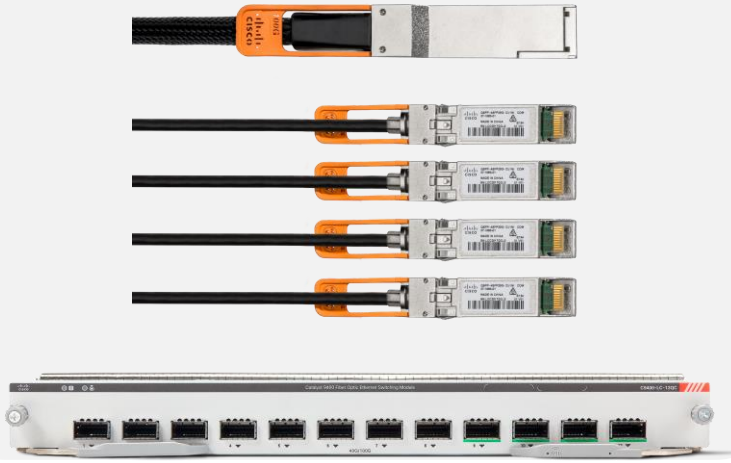
- Transmission of raw data (bits)
- Cables, connectors, radio signals



Layer 1 - Hardware

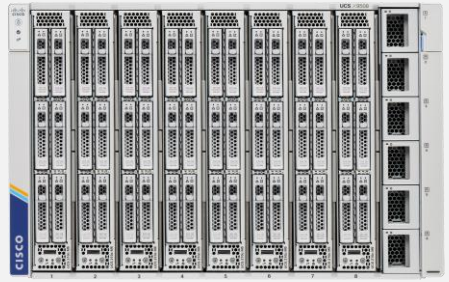
OSI

- Transmission of raw data (bits)
- Cables, connectors, radio signals



Artificial Intelligence

- Computational Resources to process data and perform tasks.
- Lossless networks



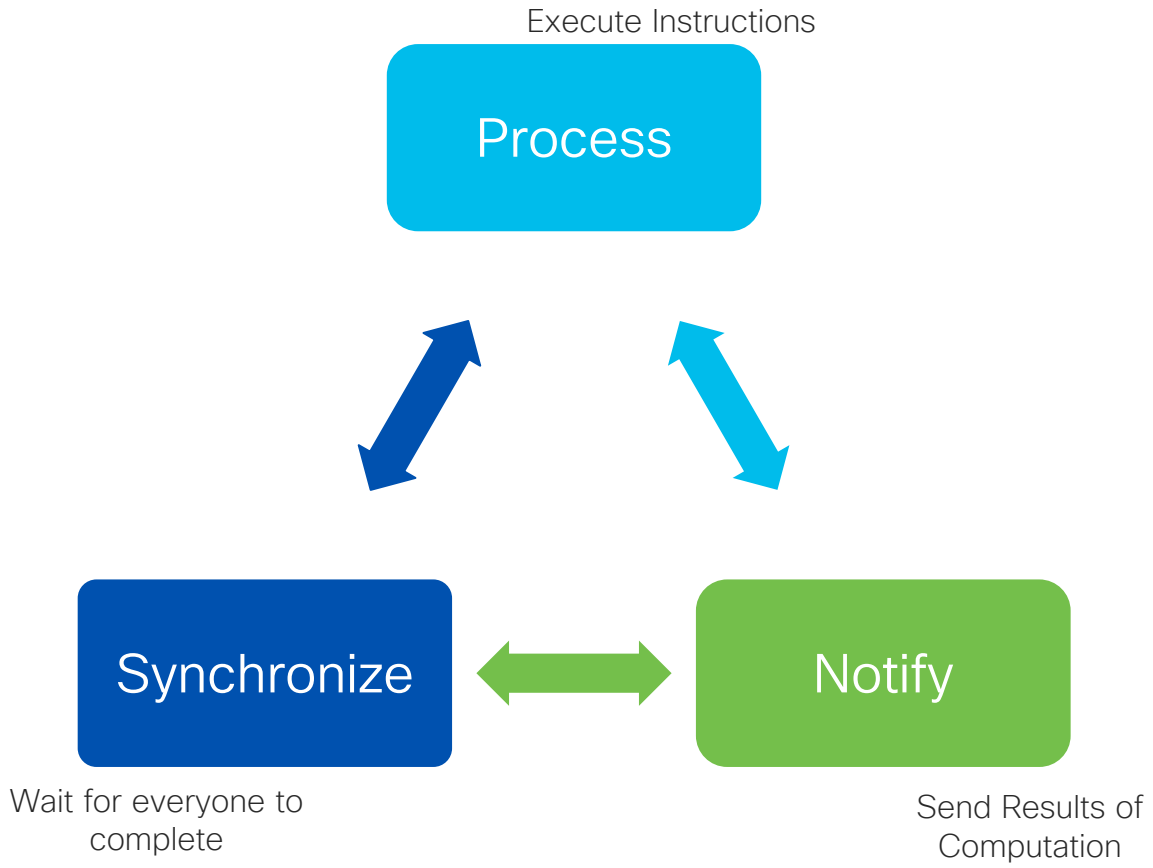
AI Hardware – The Need for Speed

CPU

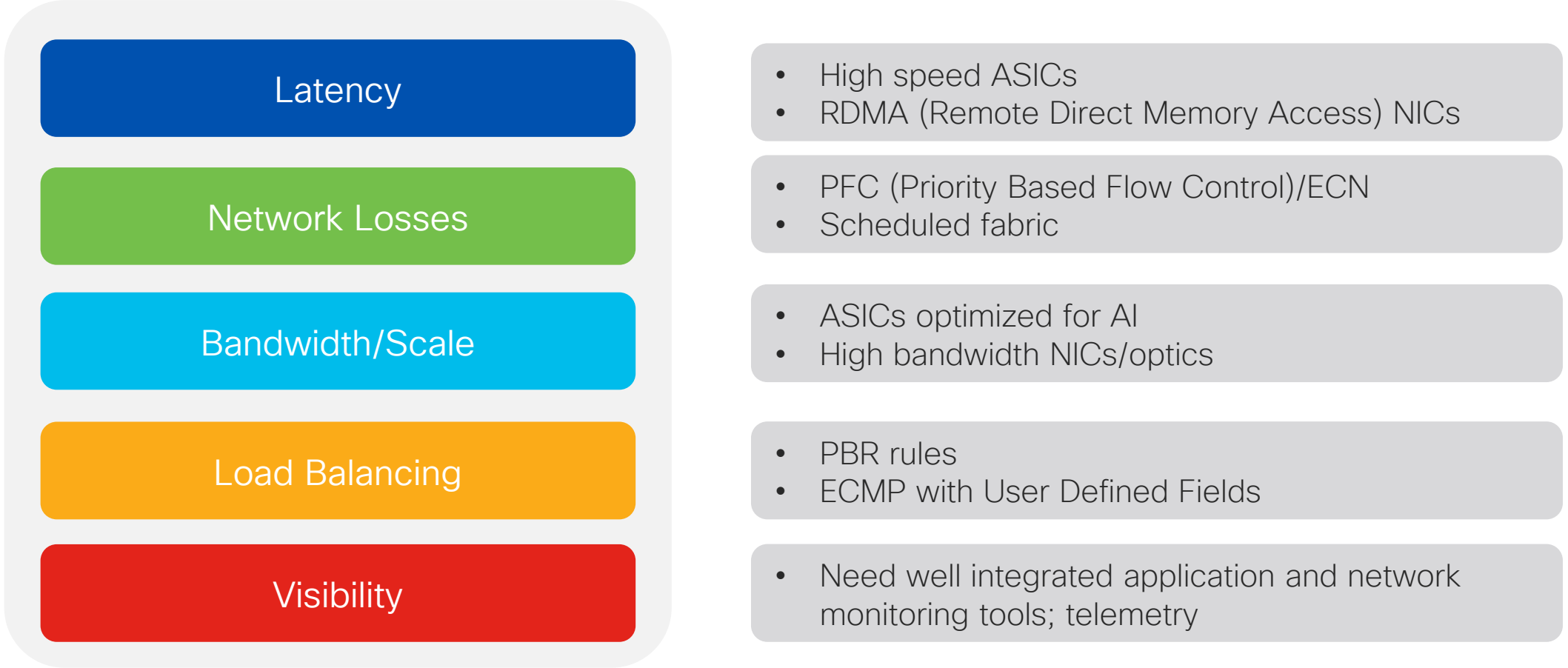
General-purpose computing
- Running applications
- Virtualization
- OS System Operations

GPU

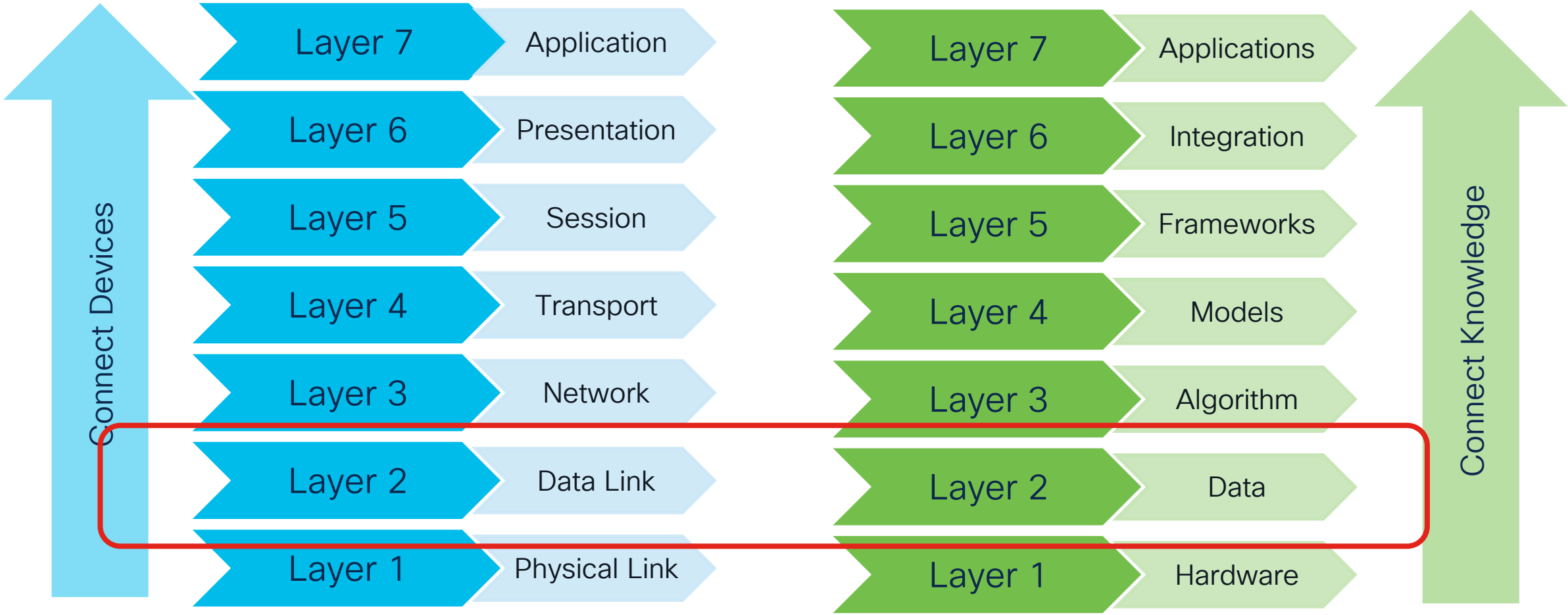
Parallel Processing
- Perform many calculations simultaneously



AI Network Requirements



From OSI to AI



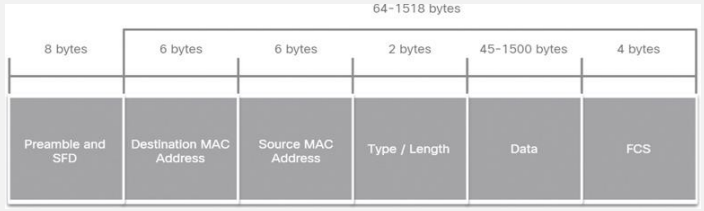
Layer 2

OSI

- Node to Node Data Transfer

```
SW1#show mac address-table
          Mac Address Table
-------------------------------------------

Vlan    Mac Address       Type        Ports
----    -----------       --------    -----
  20    aabb.cc00.0100    BLOCKED     Et0/2
  20    aabb.cc00.0200    DYNAMIC pv  Et0/0
  20    aabb.cc00.0300    BLOCKED     Et0/1
  10    aabb.cc00.0100    DYNAMIC pv  Et0/2
  10    aabb.cc00.0200    DYNAMIC     Et0/0
  10    aabb.cc00.0300    DYNAMIC pv  Et0/1
  10    aabb.cc00.0400    DYNAMIC pv  Et1/0
  10    aabb.cc00.0600    DYNAMIC pv  Et0/3
  30    aabb.cc00.0200    DYNAMIC pv  Et0/0
  30    aabb.cc00.0400    DYNAMIC     Et1/0
  30    aabb.cc00.0600    DYNAMIC     Et0/3
Total Mac Addresses for this criterion: 11
```



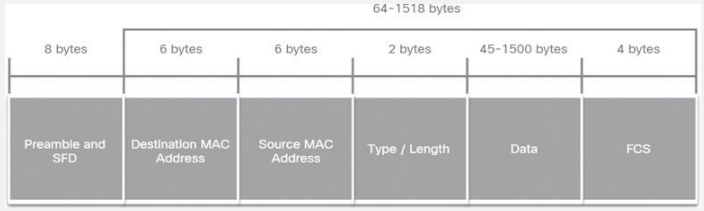
Layer 2 – Data

OSI

- Node to Node Data Transfer

```
SW1#show mac address-table
          Mac Address Table
-------------------------------------------

Vlan    Mac Address       Type        Ports
----    -----------       --------    -----
  20    aabb.cc00.0100    BLOCKED     Et0/2
  20    aabb.cc00.0200    DYNAMIC pv  Et0/0
  20    aabb.cc00.0300    BLOCKED     Et0/1
  10    aabb.cc00.0100    DYNAMIC pv  Et0/2
  10    aabb.cc00.0200    DYNAMIC     Et0/0
  10    aabb.cc00.0300    DYNAMIC pv  Et0/1
  10    aabb.cc00.0400    DYNAMIC pv  Et1/0
  10    aabb.cc00.0600    DYNAMIC pv  Et0/3
  30    aabb.cc00.0200    DYNAMIC pv  Et0/0
  30    aabb.cc00.0400    DYNAMIC     Et1/0
  30    aabb.cc00.0600    DYNAMIC     Et0/3
Total Mac Addresses for this criterion: 11
```



Artificial Intelligence

- Data Ingestion and processing

Layer 2- Unstructured Data Streams

Data Before	Data Now
Structured data at rest	Unstructured Data in movement
Inventory DB, Financial Records	Video, logs, sensor feeds, machine generated data
Batch-oriented, processed at scheduled intervals	Real-time processing

Layer 2- Unstructured Data Streams

Data Before	Data Now	Data Wrangling
Structured data at rest	Unstructured Data in movement	Cleaning, organizing, and transforming raw data into a structured and usable format for analysis
Inventory DB, Financial Records	Video, logs, sensor feeds, machine generated data	
Batch-oriented, processed at scheduled intervals	Real-time processing	

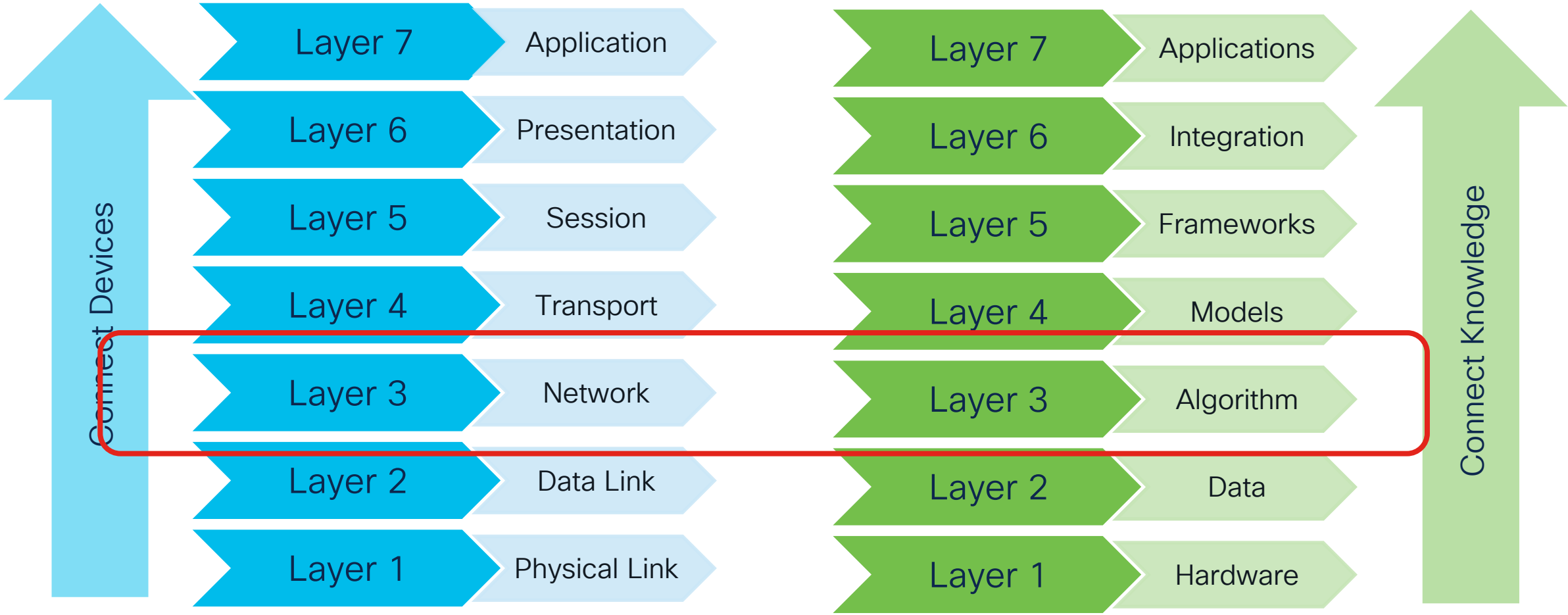
Layer 2- Unstructured Data Streams

Data Before	Data Now
Structured data at rest	Unstructured Data in movement
Inventory DB, Financial Records	Video, logs, sensor feeds, machine generated data
Batch-oriented, processed at scheduled intervals	Real-time processing
Connection Oriented	Flow-Centric

Data Wrangling

Cleaning, organizing, and transforming raw data into a structured and usable format for analysis

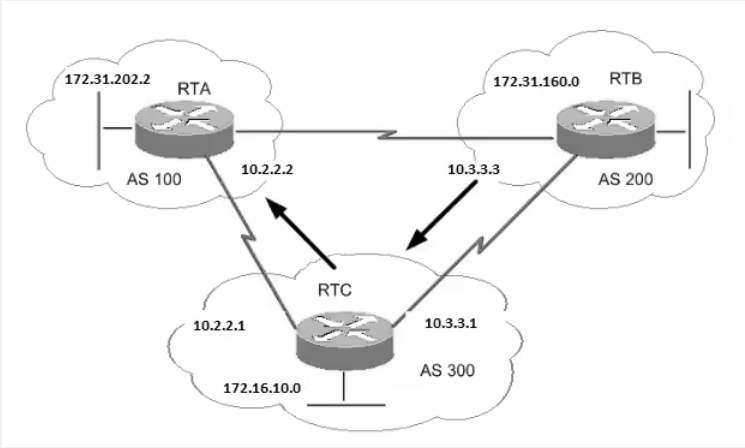
From OSI to AI



Layer 3

OSI

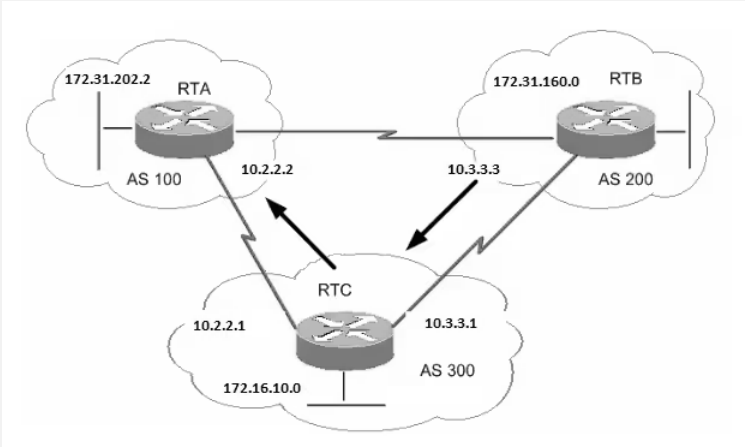
- Routing Addressing, Delivering Decisions



Layer 3 – Algorithm

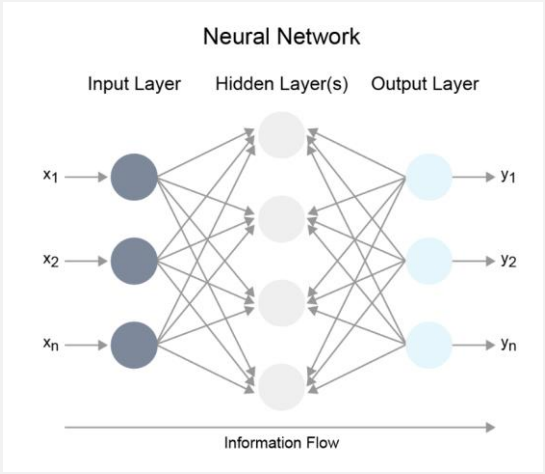
OSI

- Routing Addressing, Delivering Decisions

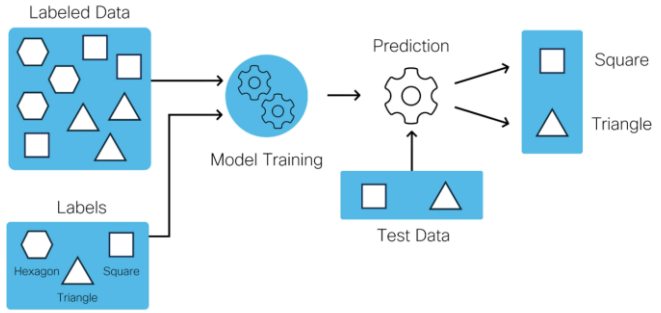


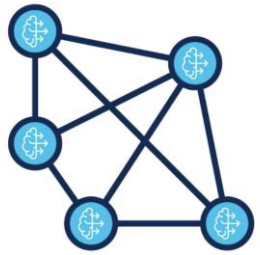
Artificial Intelligence

- Learning mechanisms and techniques



Layer 3 – Machine Learning or Deep Learning

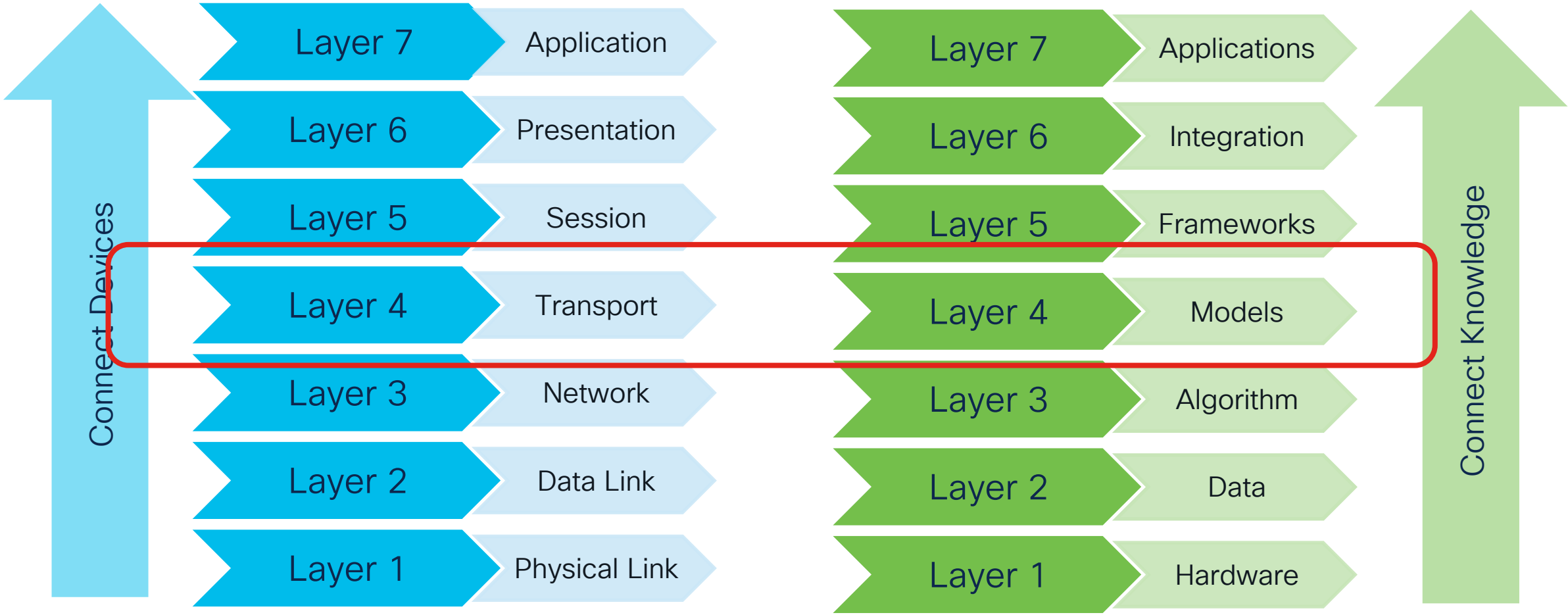




Neural Network

Machine Learning	Deep Learning
Learn independently, need human adjustments	Advanced computing neural networks
Statistical techniques	Subset of ML, analyze raw data
Need less data, structured data	Large amounts of data

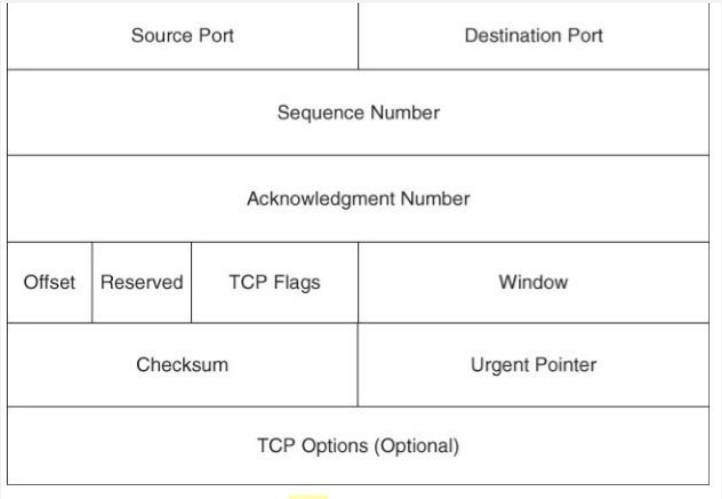
From OSI to AI



Layer 4

OSI

- Transport, Reliable data transfer



Layer 4 – Models

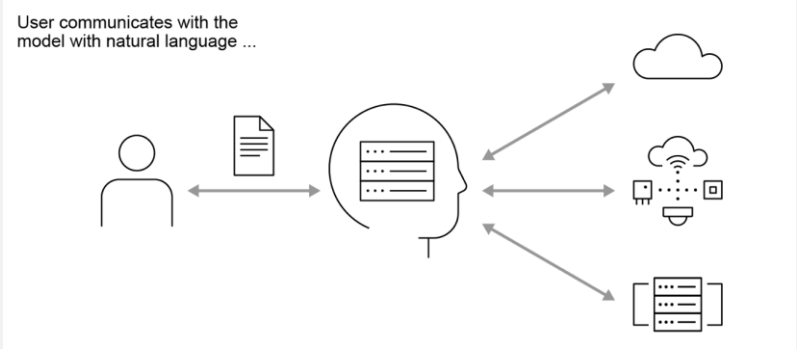
OSI

- Transport, Reliable data transfer

Source Port			Destination Port
Sequence Number			
Acknowledgment Number			
Offset	Reserved	TCP Flags	Window
Checksum			Urgent Pointer
TCP Options (Optional)			

Artificial Intelligence

- Model for the right purpose



Layer 4 – Models

Machine Learning Models

Learn from data to make predictions

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning

Deep Learning Models

Complex neural networks for advanced tasks

- Convolutional Neural Networks
- Recurrent Neural Networks
- Generative Adversarial Networks

Generative Models

Create new data based on existing data

- Convolutional Neural Networks
- Recurrent Neural Networks
- Generative Adversarial Networks

LLMs

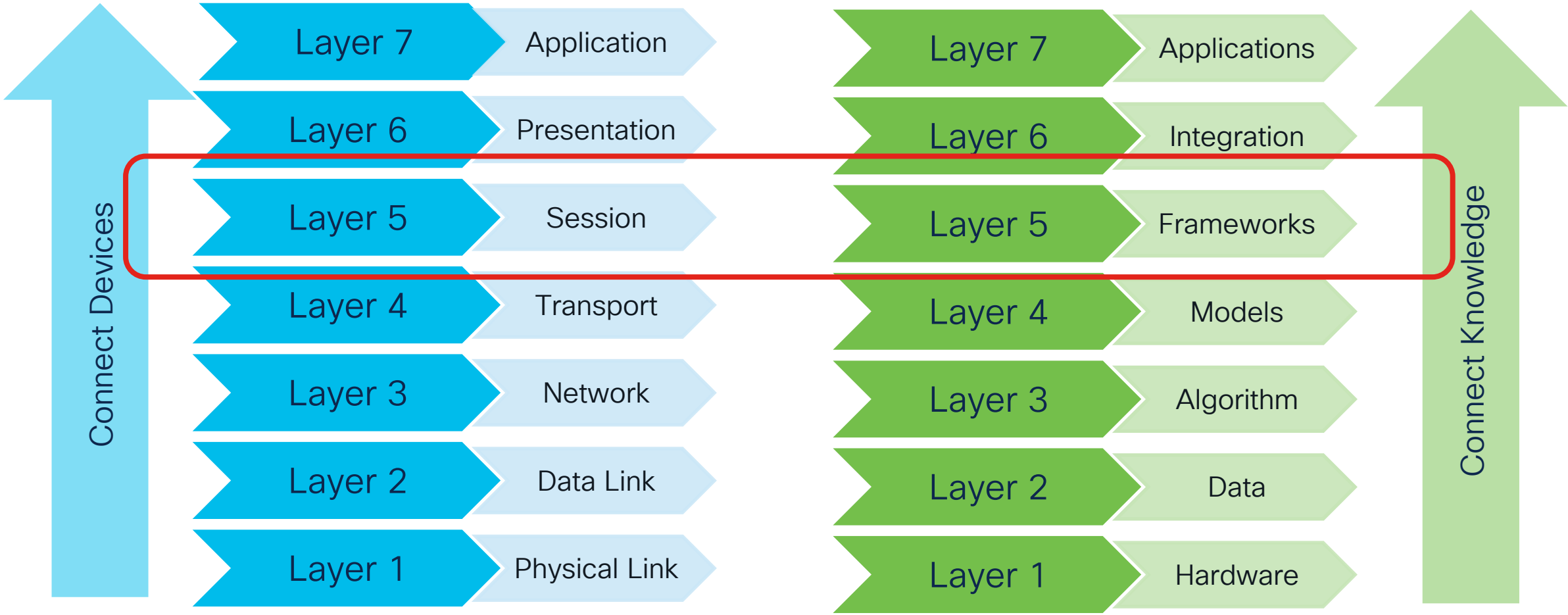
RAG

Reinforcement Learning

Learn through trial and error

- Q-Learning
- Deep Q-Learning
- Reward, Penalty Methods

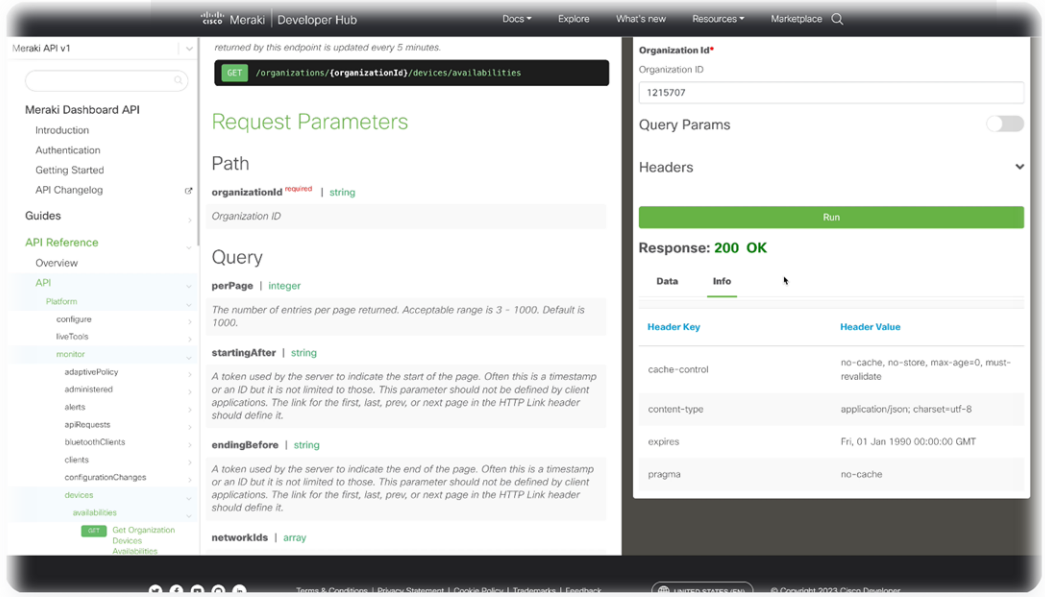
From OSI to AI



Layer 5

OSI

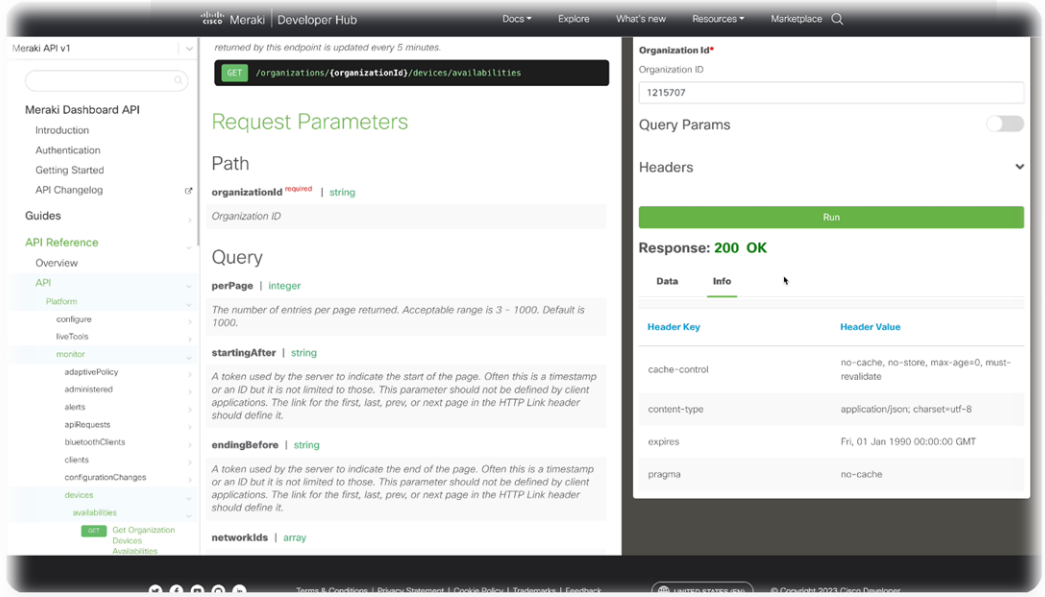
- Manage sessions between applications. APIs and RPCs



Layer 5 – Software Framework

OSI

- Manage sessions between applications. APIs and RPCs



Artificial Intelligence

- Software Framework

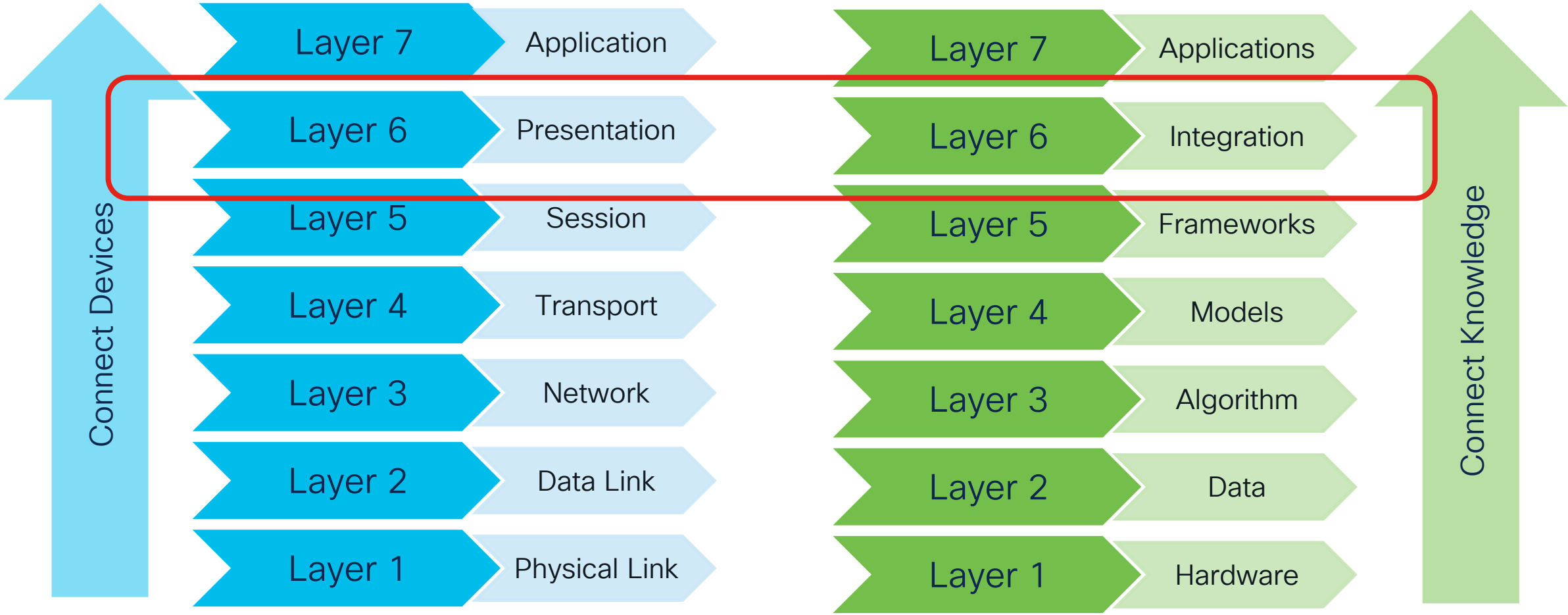
PyTorch

TensorFlow

Layer 5 – Software Framework

TensorFlow	PyTorch	Scikit-learn	LangChain
Google open-source machine learning library	Open-source ML library developed by FAIR	Data mining and analysis	Open-source framework for apps based on LLMs
• Image and Speech Recognition • Natural Language Processing, reinforcement learning	• Computer vision • Natural Language Processing • Generative Modeling	• Classification • Regression • Clustering • Dimensionality Reduction	• To create powerful apps with AI LLMs

From OSI to AI



Layer 6

OSI

- Presentation, Translates data formats between application and network

JPEG/ASCII/EBDIC/TIFF/GIF/PICT

Layer 6 – Integration

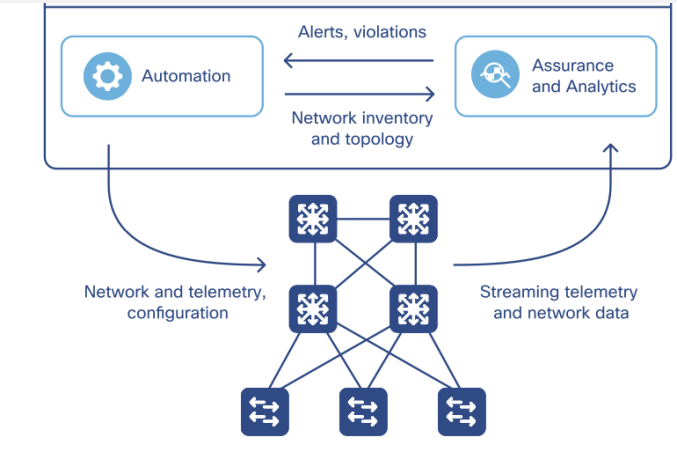
OSI

- Presentation, Translates data formats between application and network

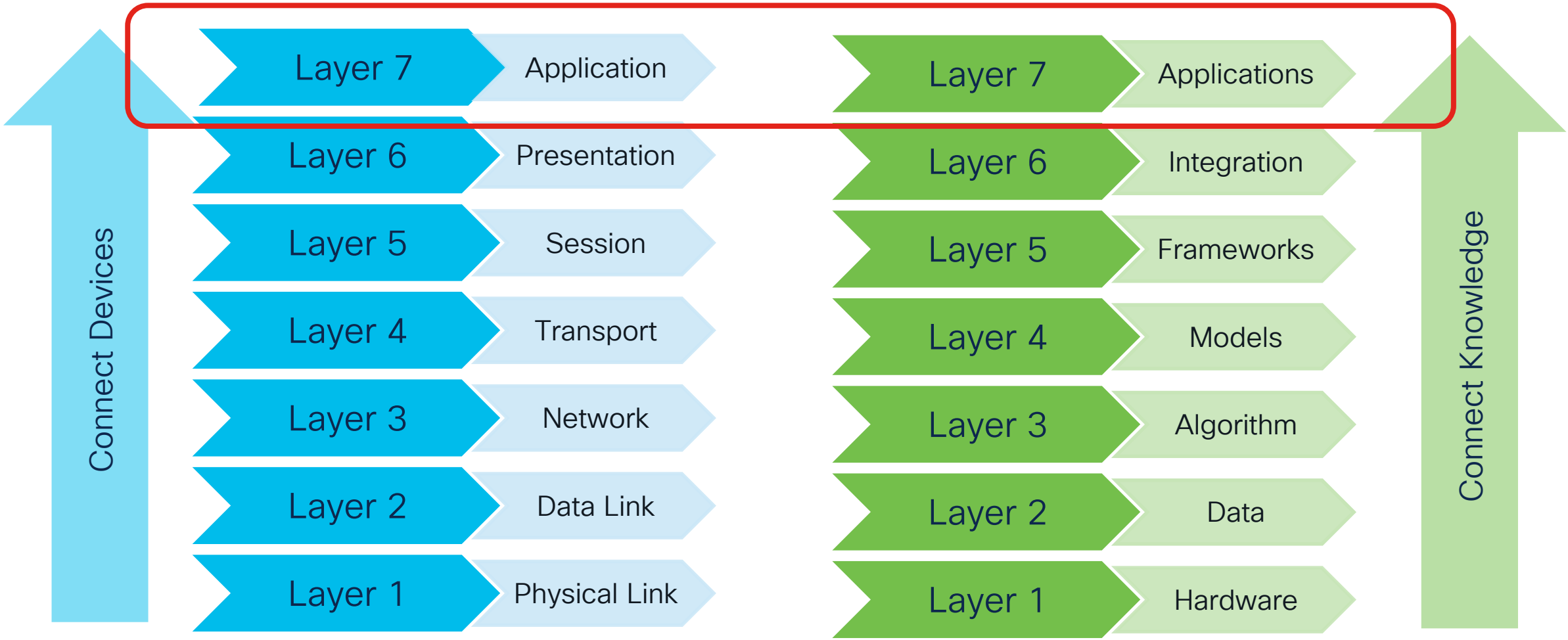
JPEG/ASCII/EBDIC/TIFF/GIF/PICT

Artificial Intelligence

- Integration, Ensures AI integrates into systems and applications



From OSI to AI



Layer 7

OSI

- Application, Interfaces directly with user applications for network services

HTTP, FTP, SMTP, DNS, SNMP

Layer 7 – Application

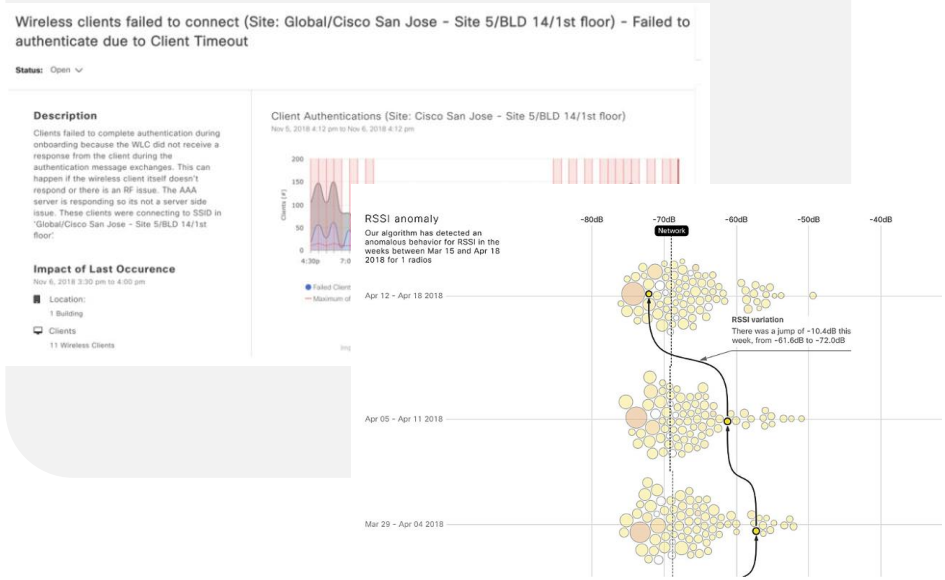
OSI

- Application, Interfaces directly with user applications for network services

HTTP, FTP, SMTP, DNS, SNMP

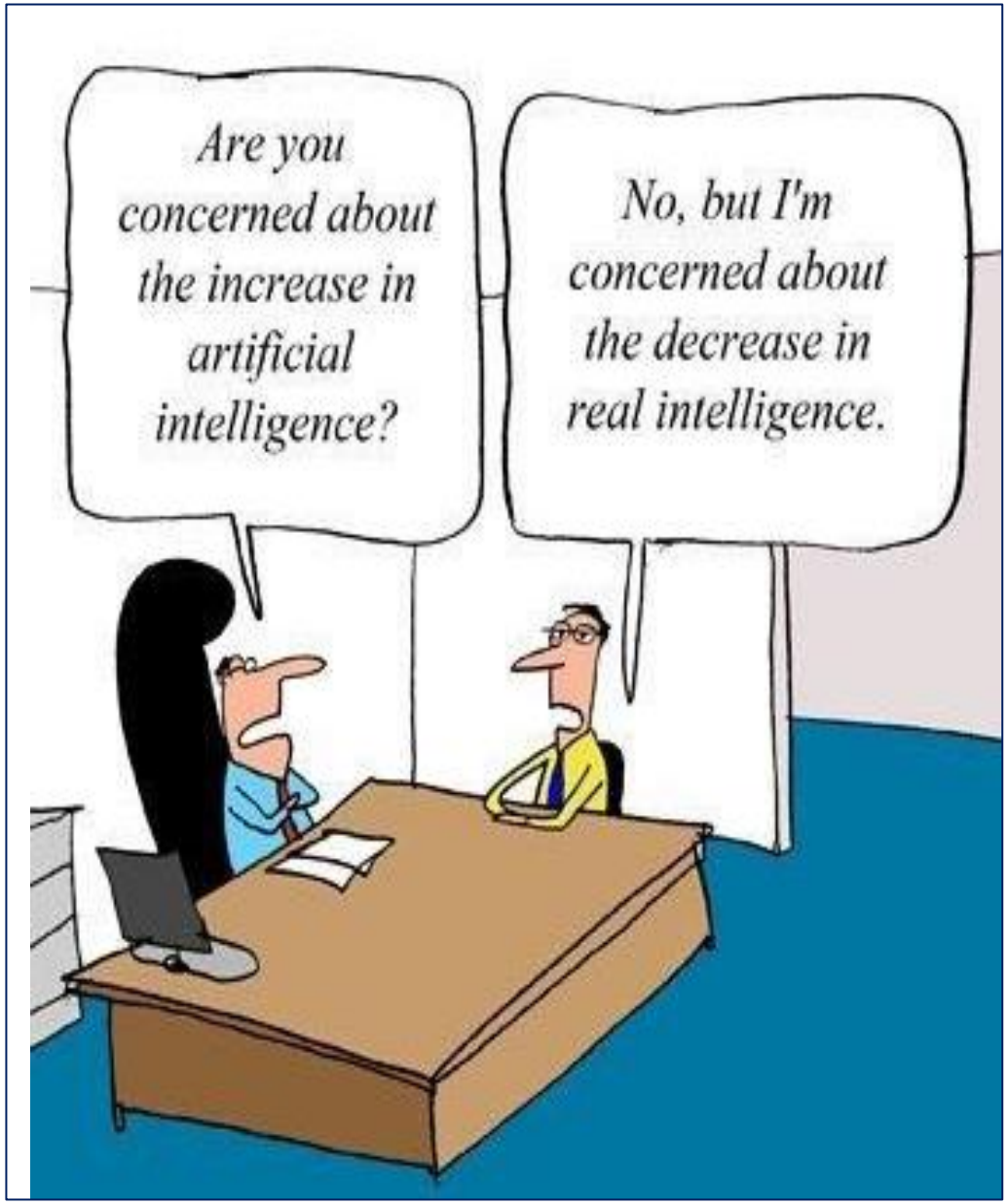
Artificial Intelligence

- Application, Provides end user functionality and value derived from AI

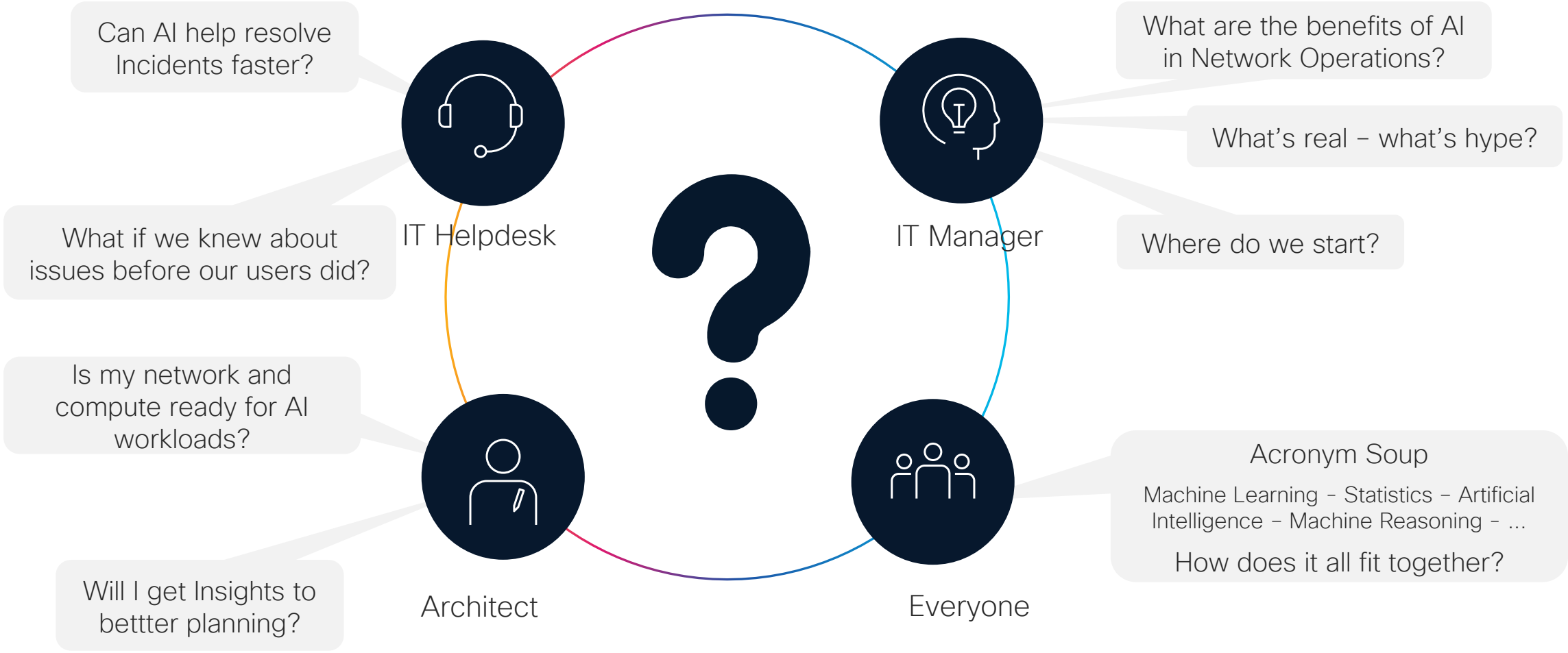


What does it mean it to me?

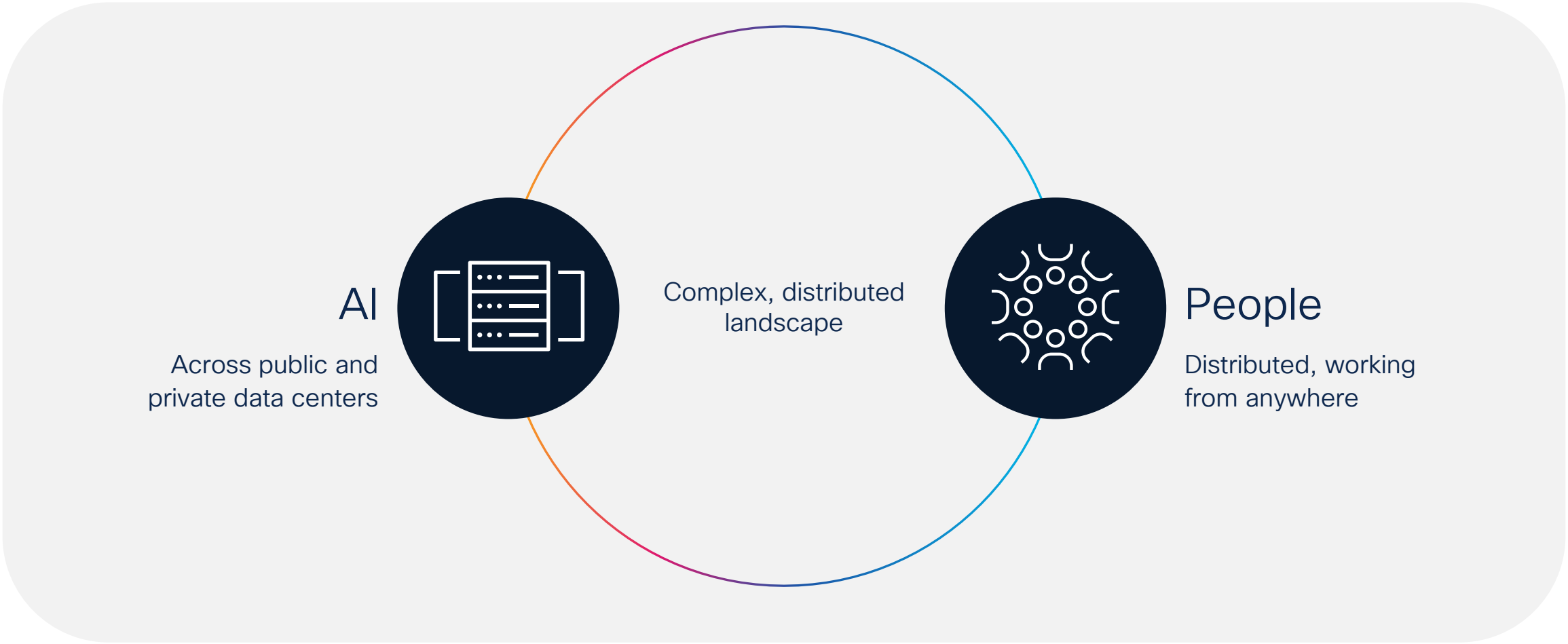
Why should I Care?



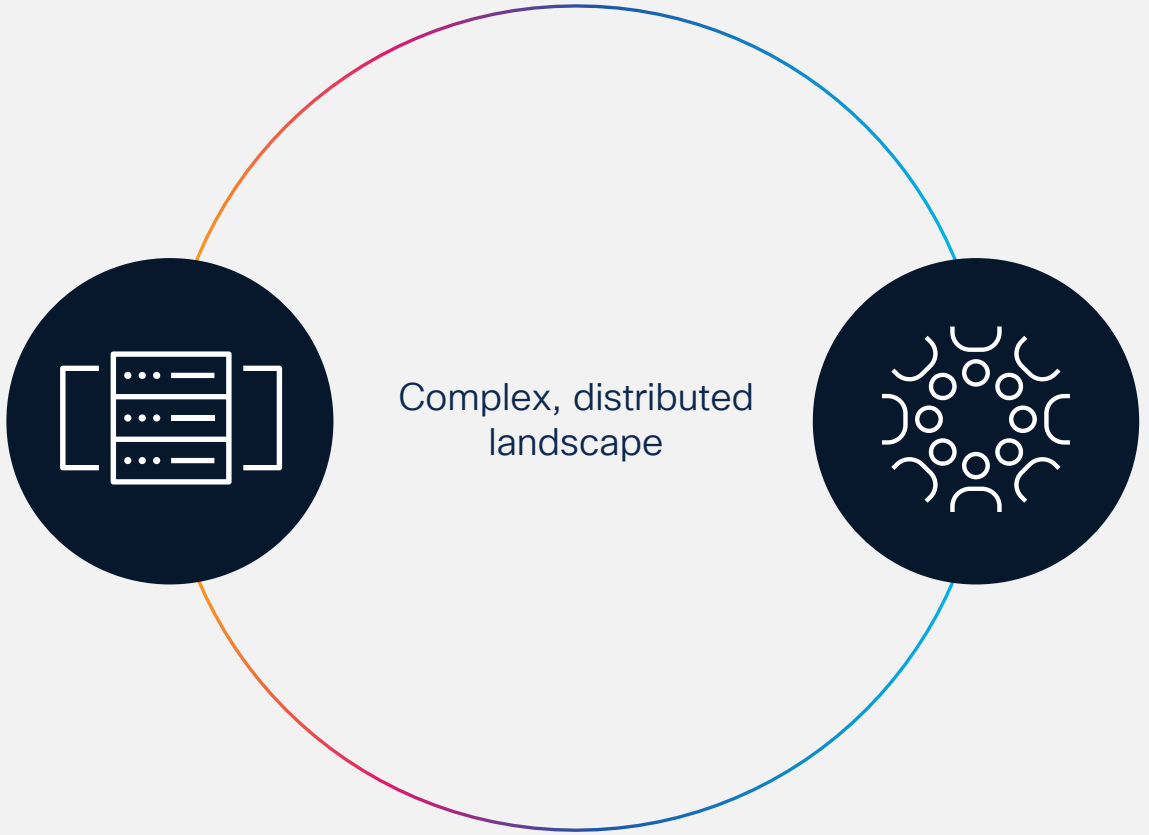
Why Should I Care About Artificial Intelligence (AI)?



AI and People



Yesterday’s approaches don’t support today’s reality



AI

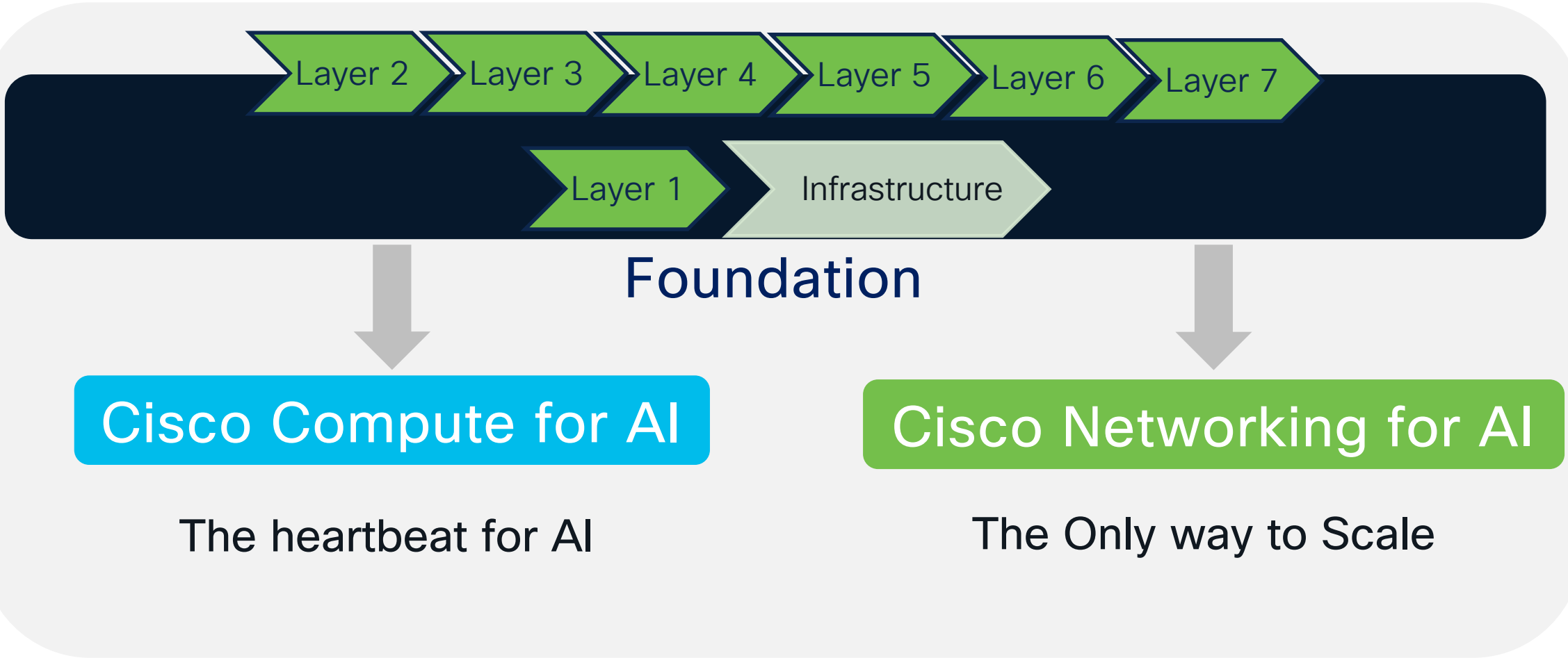
High latency and low efficiency were nuisances – now, they can undermine massive AI investments

Complex, distributed landscape

People

Secure, connected experiences have become a mandate – expectations have changed faster than many have adapted

Connecting the Layers



Cisco Compute for AI

The heartbeat for AI

AI Compute Challenges

Feeding neural networks

- Inference vs Fine-Tuning vs Training requirements
- Sizing
- Power consumption

From Training to Inferencing: A Songwriting Analogy

Training:

Composition: Crafting the melody and lyrics. This is the most resource-intensive stage, similar to building an AI model from scratch.

Fine Tuning:

Mixing: Refining the song with adjustments and effects to suit a particular style, akin to enhancing an AI model for specific tasks.

Inferencing:

Performance: Playing the completed song for an audience, much like using a trained AI model to interact with a user.

Sizing AI Workloads

Considerations

- Number of model parameters
- Concurrent requests
- Context size
- Knowledge-base size
- Precision compression level

A single 70B parameter model is 240GB alone

Single model memory footprint

Model component	Bytes @ full precision
Model Parameters	4 bytes per parameter
Activations	4 bytes per token * context size
Optimizer states	8 bytes per parameter
Gradients	4 bytes per parameter

Training Overhead

Power Considerations

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

	8U HGX H100 Server	(4) 2U UCS C245 M8 server	2U UCS C245 M8 server
GPU	(8) H100 80GB SXM5 700W	(8) H100 80GB PCIe 350W	(2) H100 80GB PCIe 350W
LAN	CX-7 dual 400G	(4) VIC quad 10/25/50G	VIC quad 10/25/50G
Storage Network	CX-7 dual 400G IB	(4) CX-6 dual 100G	CX-6 dual 100G
GPU Network	(8) CX-7 dual 400G IB	N/A	N/A
Power Draw	9,900 watts max	5,600 watts max	1,400 watts max
System Watts per GPU	1,237	700	700

What are your per-rack power requirements?

Portfolio built for choice

Validated solutions for AI with compute, network, storage, and software

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

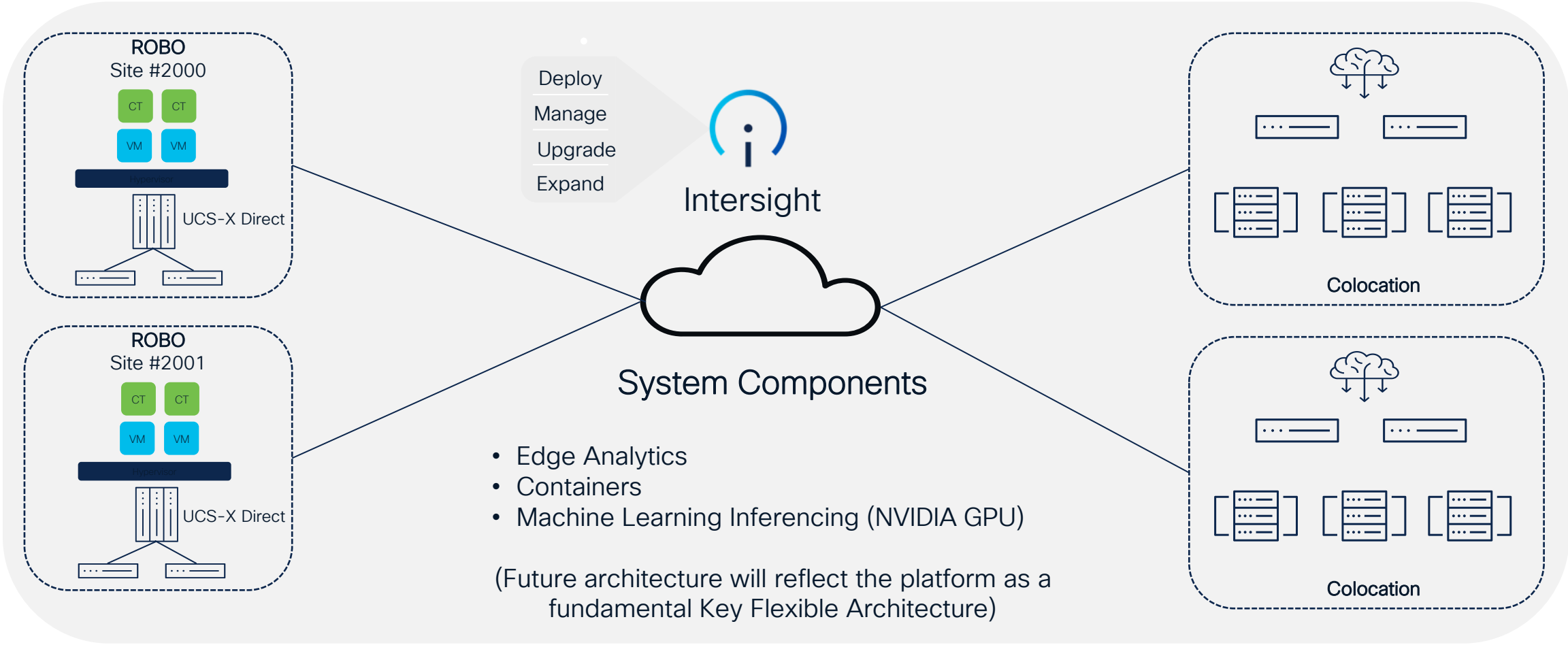
UCS C885A Dense GPU Servers

UCS Blade (w/GPU Expansion) and Rack

Dense compute for demanding AI

Full stack AI with compute and networking

Global Infrastructure Management



Cisco Networking for AI

The only way to scale

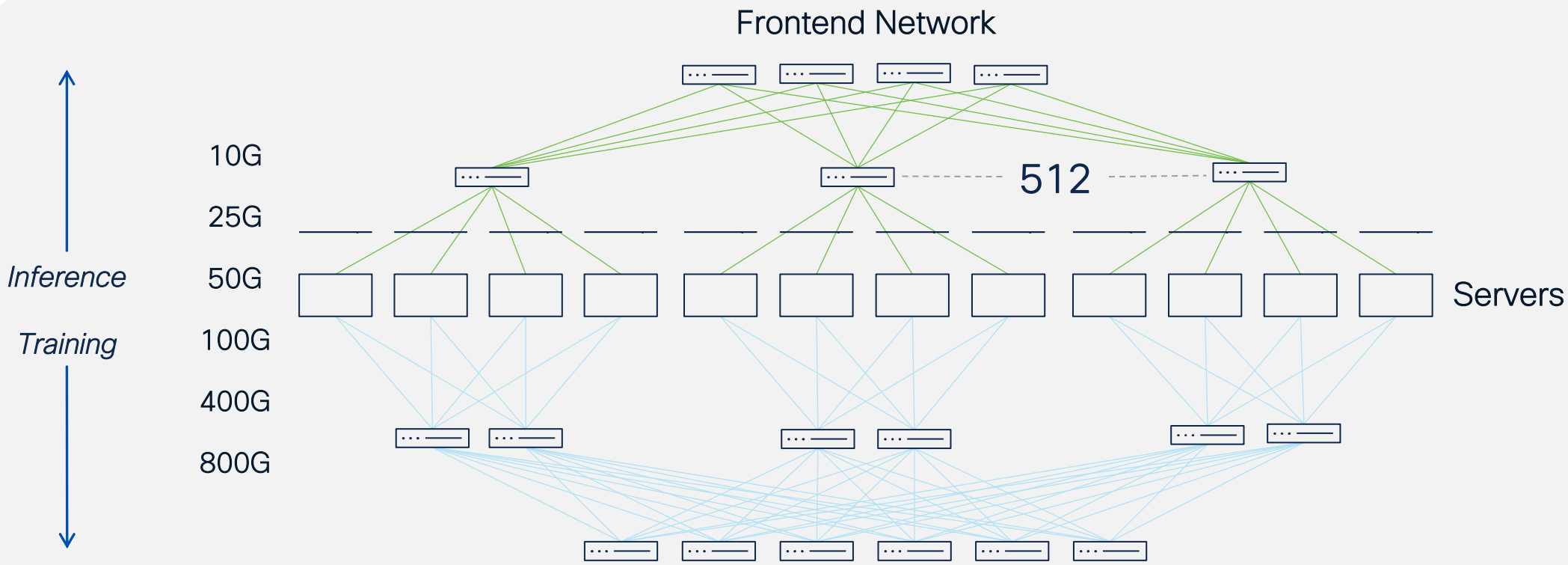
AI Networking Challenges

Connecting the hivemind

- Different sub-networking
- Training flow characteristics
- Cabling

AI Datacenter

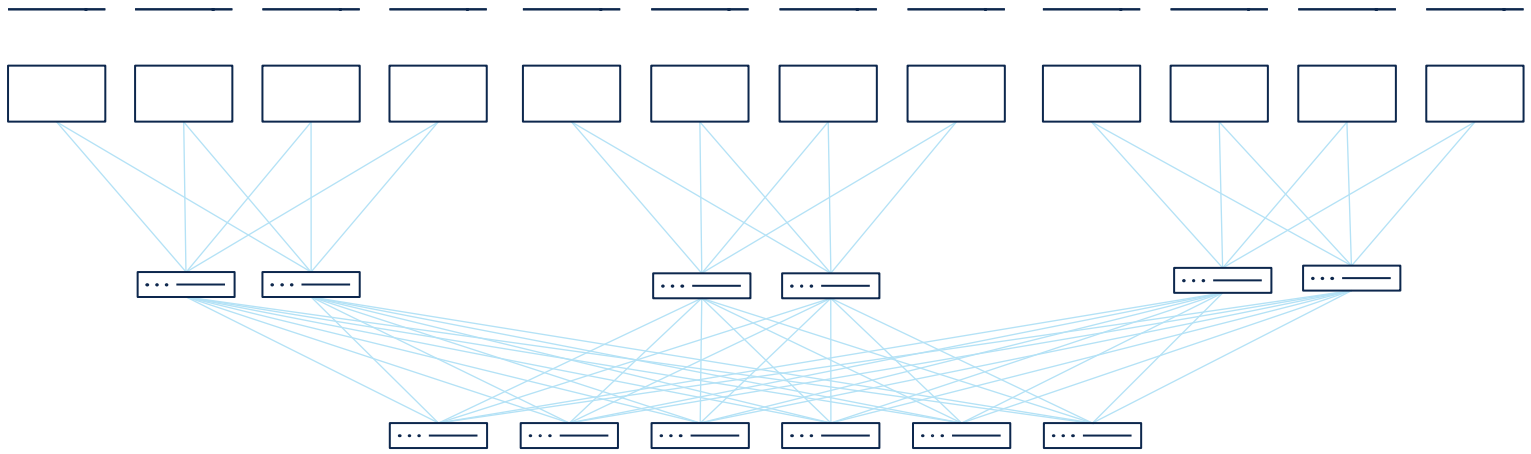
Different networks for different purposes



Backend Ethernet AI Fabric

"Just move data and get out of the way"

Lossless | High-Throughput | Low Jitter | Low-Latency



Backend (GPU Direct) & Storage Networks

400/800G Ethernet Transition (25.6T & 51.2T)

High-bandwidth fabrics with reduced footprint and energy savings

Network card for every GPU

RDMA over Ethernet (RoCEv2)
Lossless network (PFC + ECN)

Non-blocking network

Low latency, congestion management

Load balancing enhancements

AI fabric templates, AI analytics, telemetry, congestion scores

---

Validated designs for networks and ecosystem partners

Spine/Leafe, 3-tier, & Rail-Optimized architectures

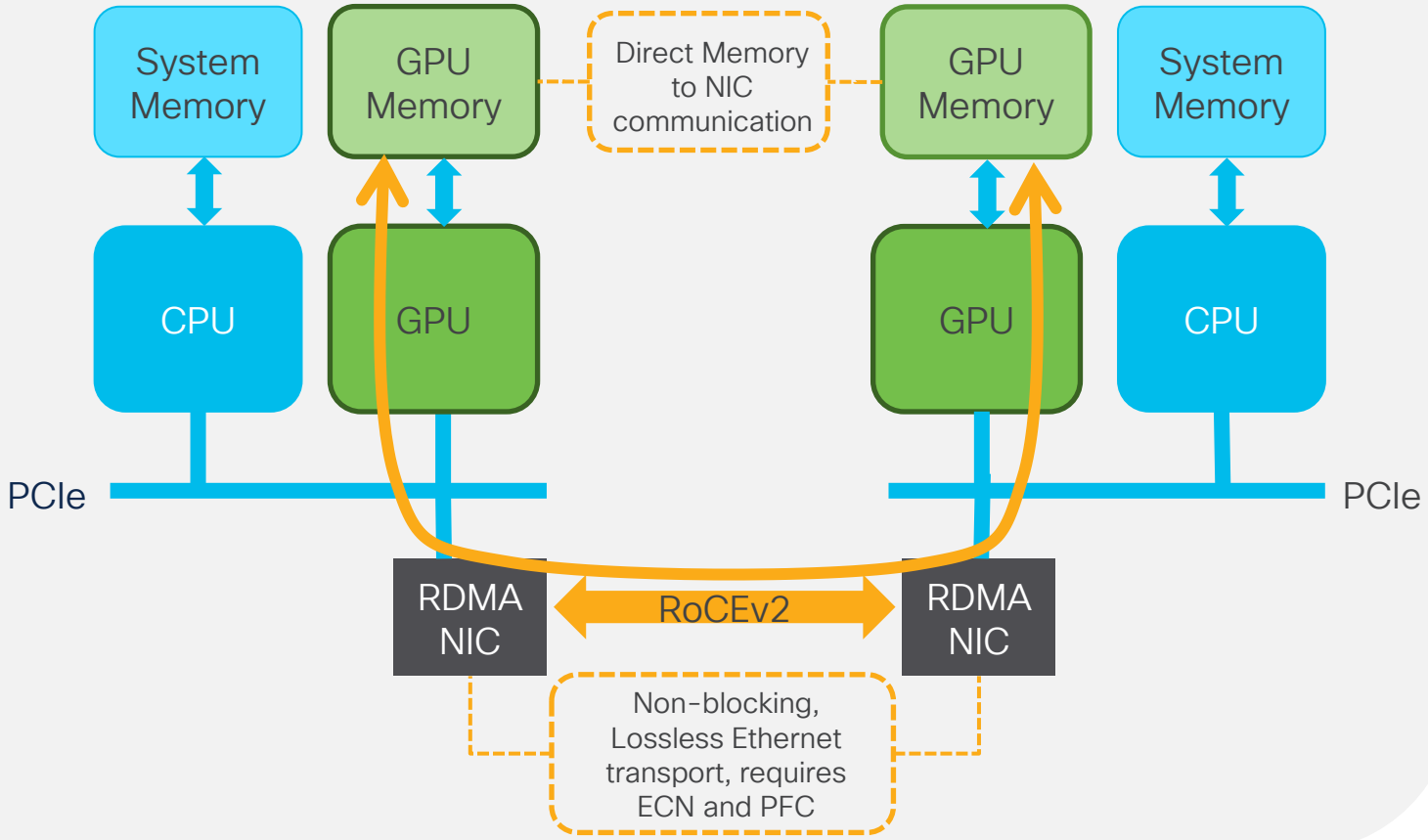
PFC: Priority Flow Control

ECN: Explicit Congestion Notification

Remote Direct Memory Access (RDMA)

over Converged Ethernet

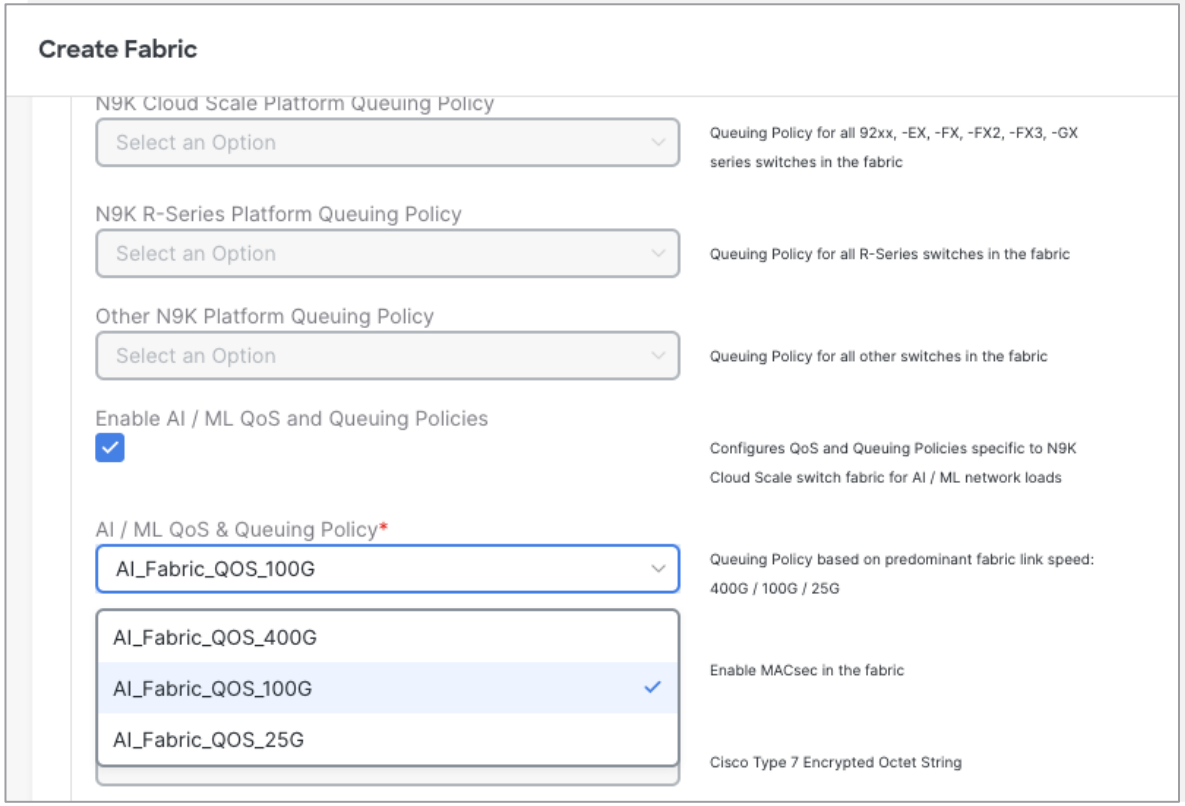
- RDMA allows GPUs to access the bytes directly in each others VRAM
- Latency is very low as CPU and kernel can be bypassed
- Does not handle loss well

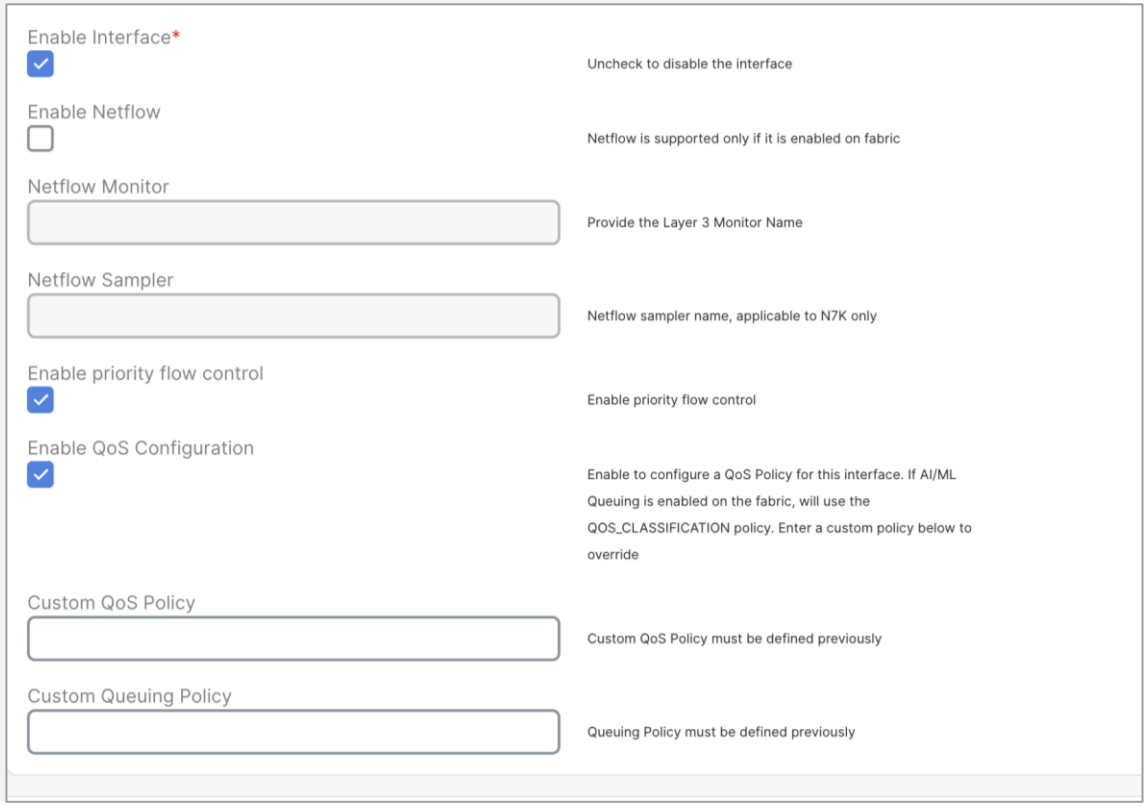


Nexus Dashboard

Automation

"Easy button" templates for lossless ethernet fabrics

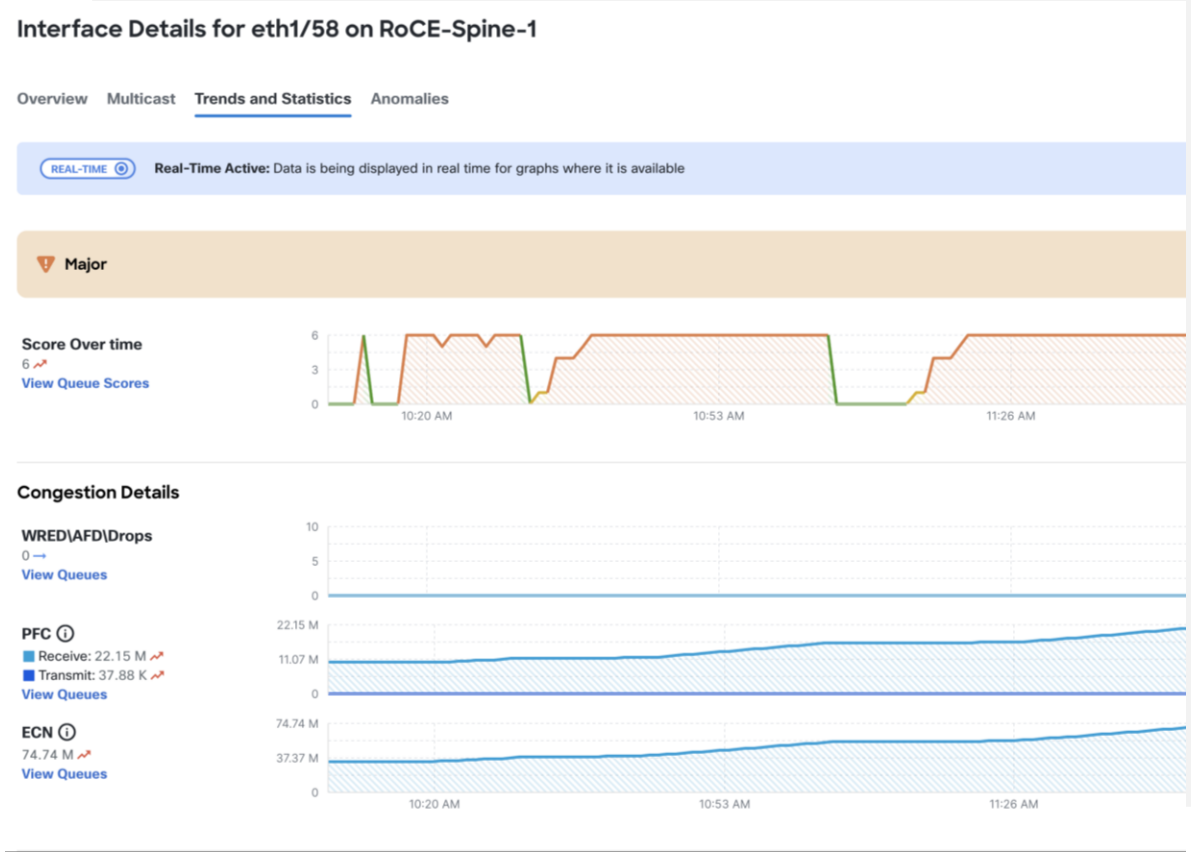


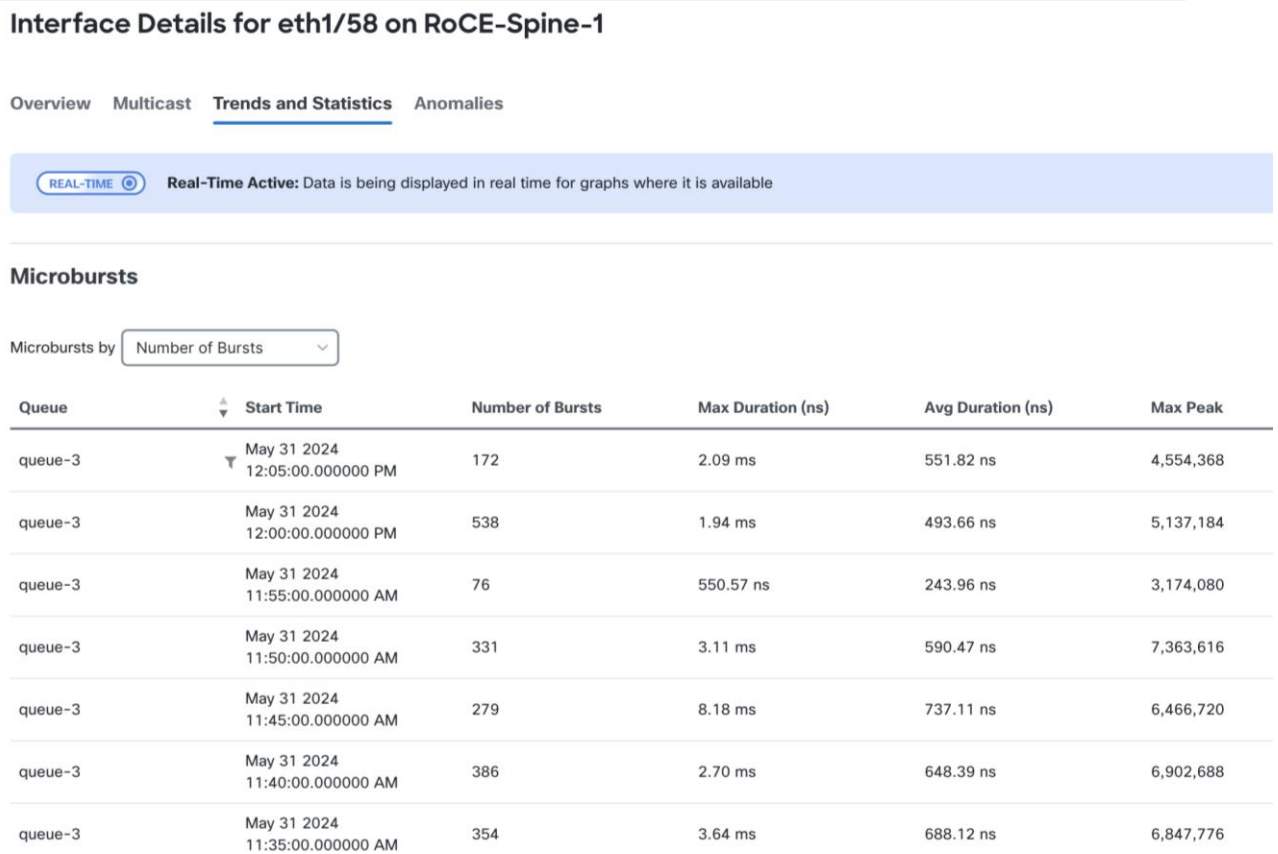


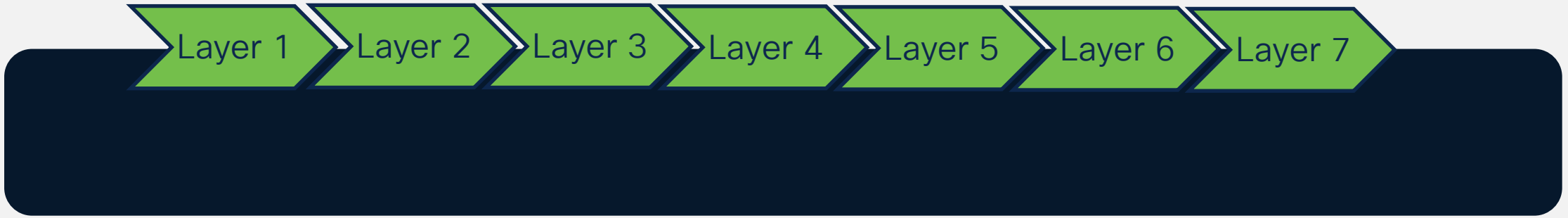
Nexus Dashboard

Visibility

Real-time performance metrics







Bringing it Together

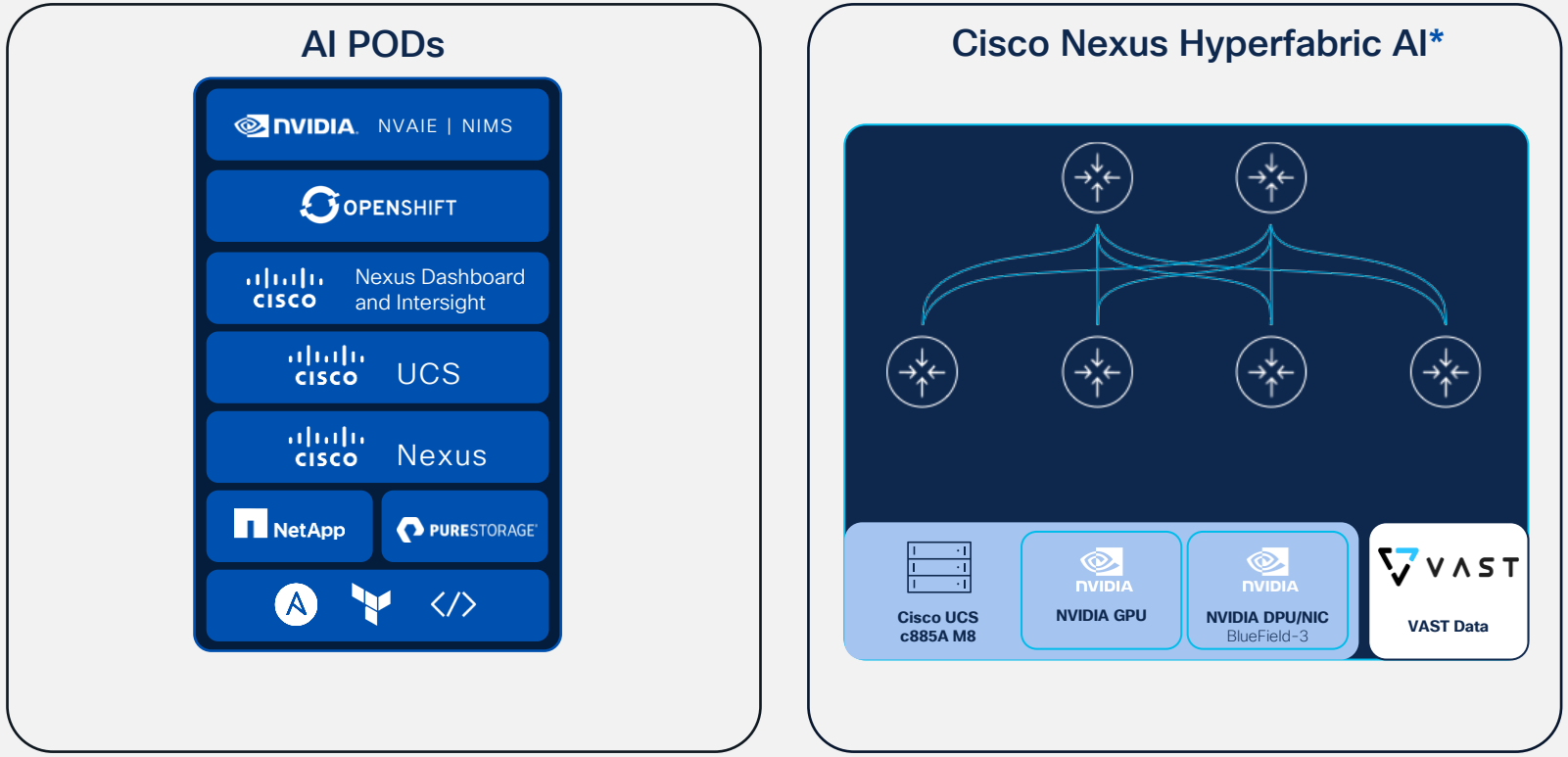
Full Stack Systems

Full Stack Systems to accelerate time to value

Use fully integrated, modular, and scalable stacks for AI workloads

Realize fast, reliable deployments based on validated AI infrastructure

Gain unmatched performance wherever AI lives



*Roadmap

Full Stack Systems

AI PODs

Inference & Fine-tuning with confidence

- Confidently deploy AI-ready infrastructure with pre-designed full stack architecture bundles for targeted AI use cases.
- Leverage automation frameworks for rapid deployment and adoption of infrastructure.
- Operate with best-in-class single-support model for your AI deployment architecture, include enterprise support for select Operations Support System (OSS) tools and libraries

AI Model

AI Tooling

Containers

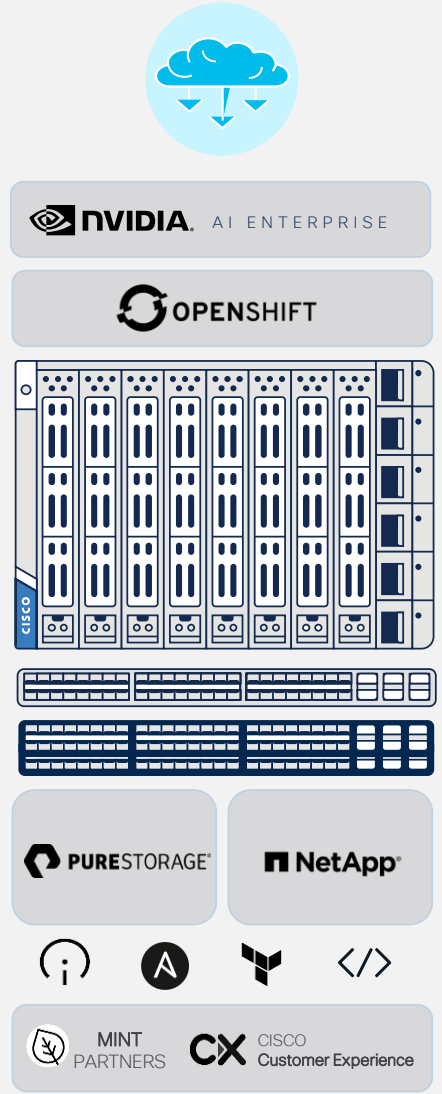
Accelerated Compute

Networking

Converged Infrastructure

Management & Automation

Adoption & Support Services



Cisco Nexus Hyperfabric AI Cluster

in partnership with NVIDIA

Democratize AI Infrastructure

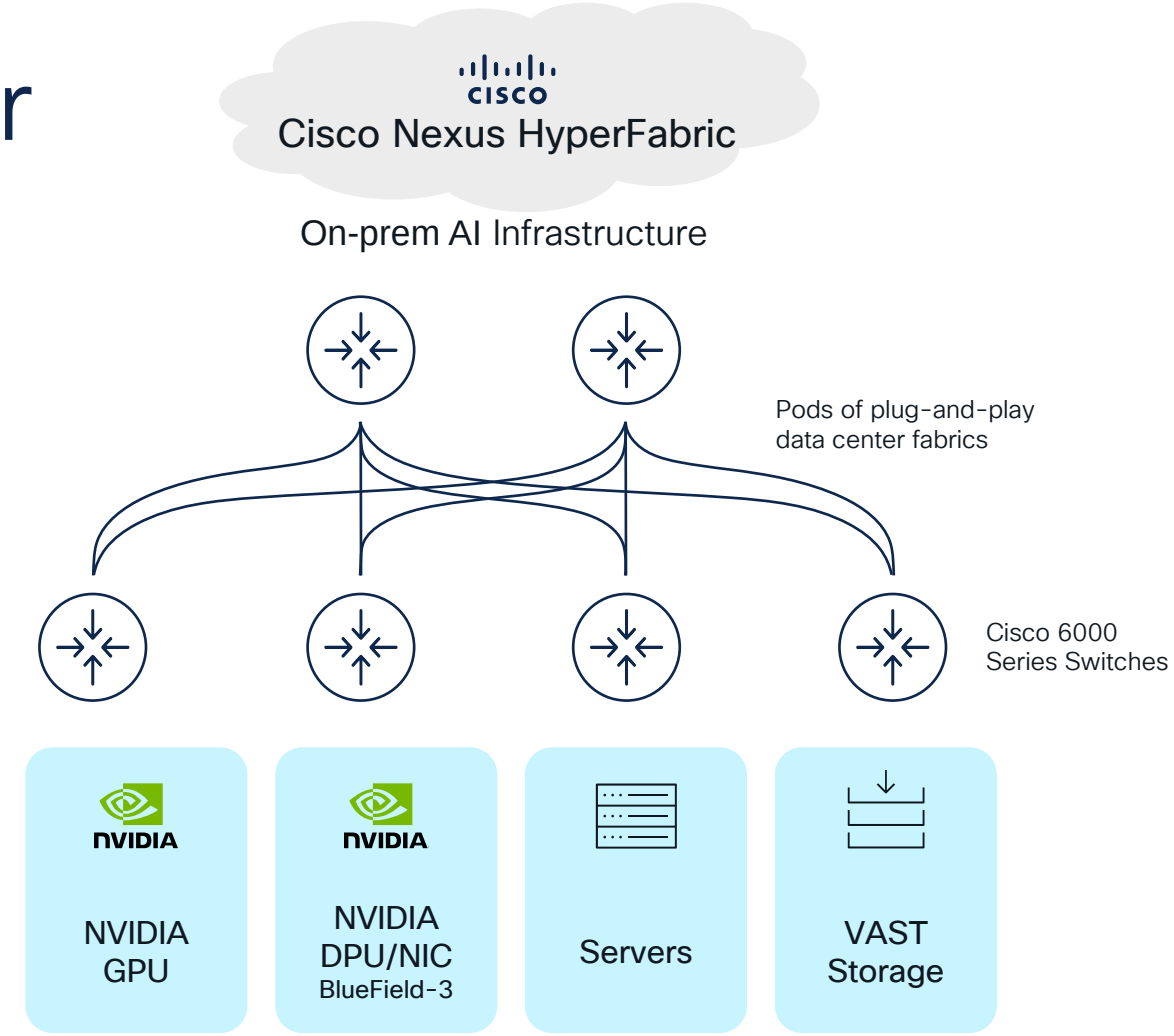
Visibility into full stack AI

Unified stack Including NVAIE

AI-native operational model

High-performance Ethernet

Cloud managed operations

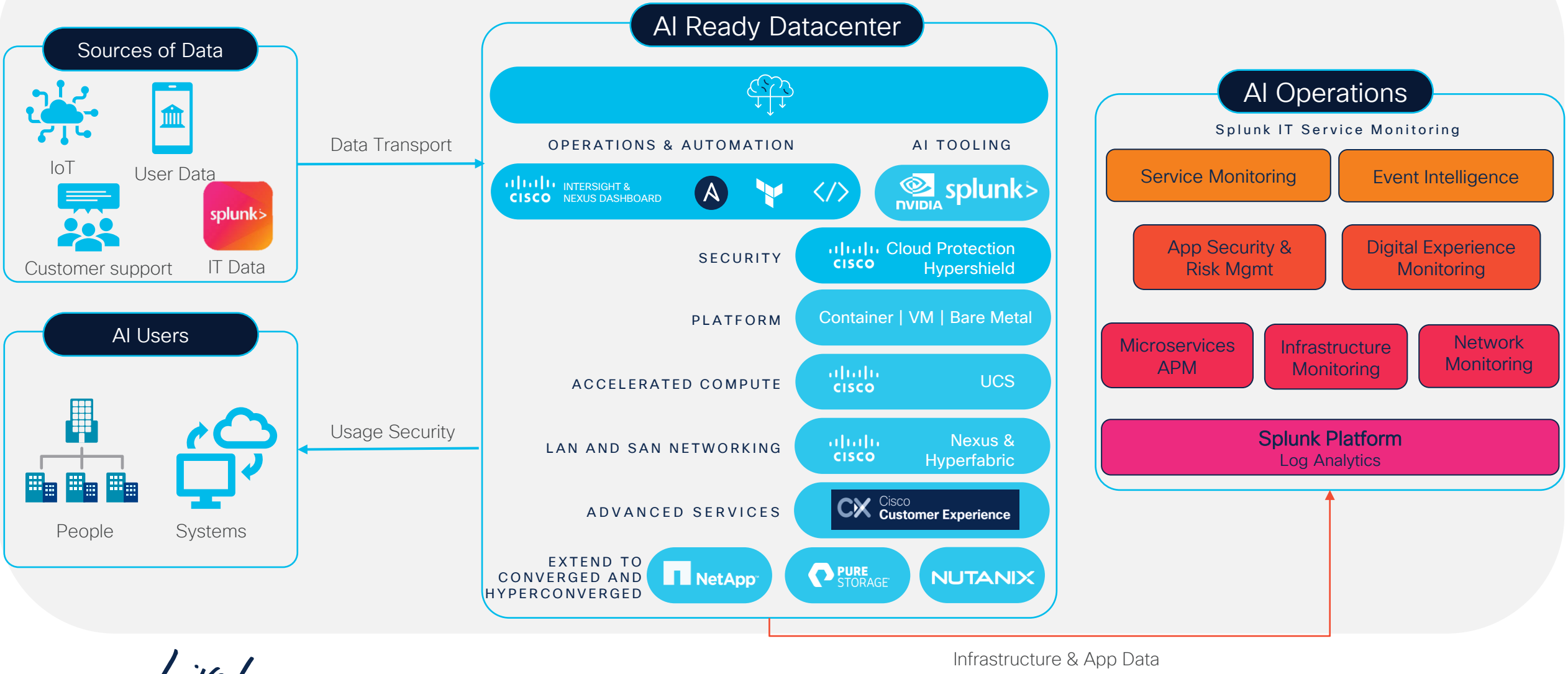


A solution that will enable you to spend time on AI innovation—not on IT.

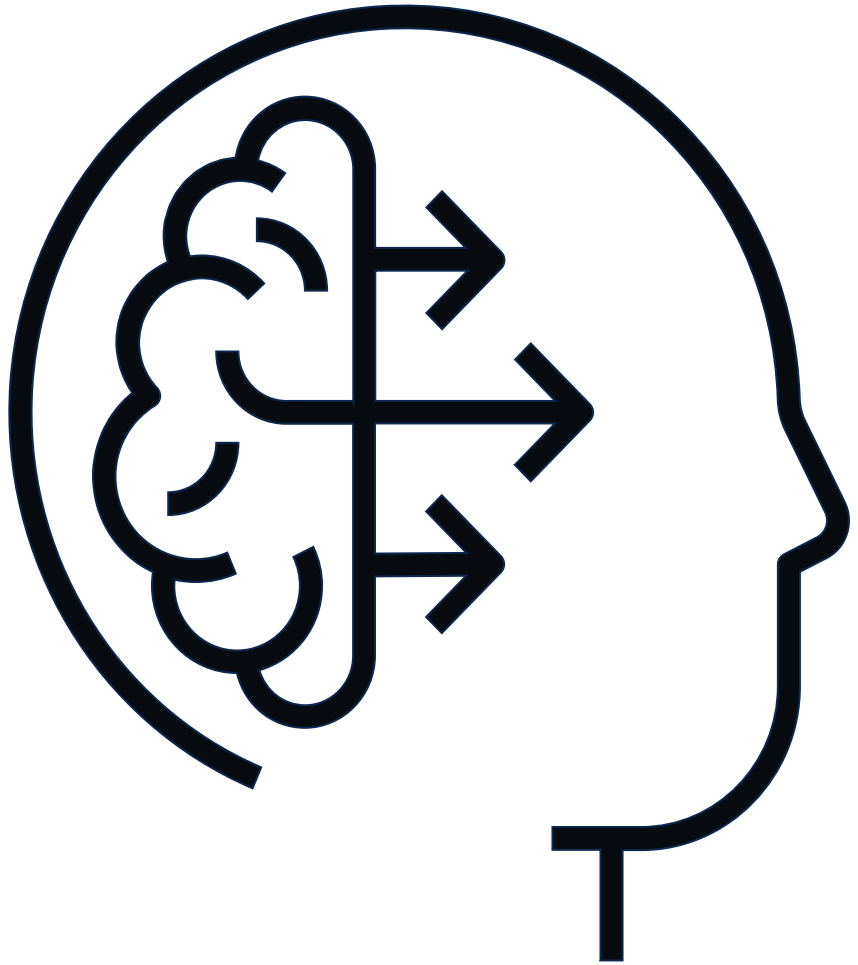
Design › Order › Deploy › Validate › Monitor › Upgrade › Collaborate

Connecting the Layers

Follow the data

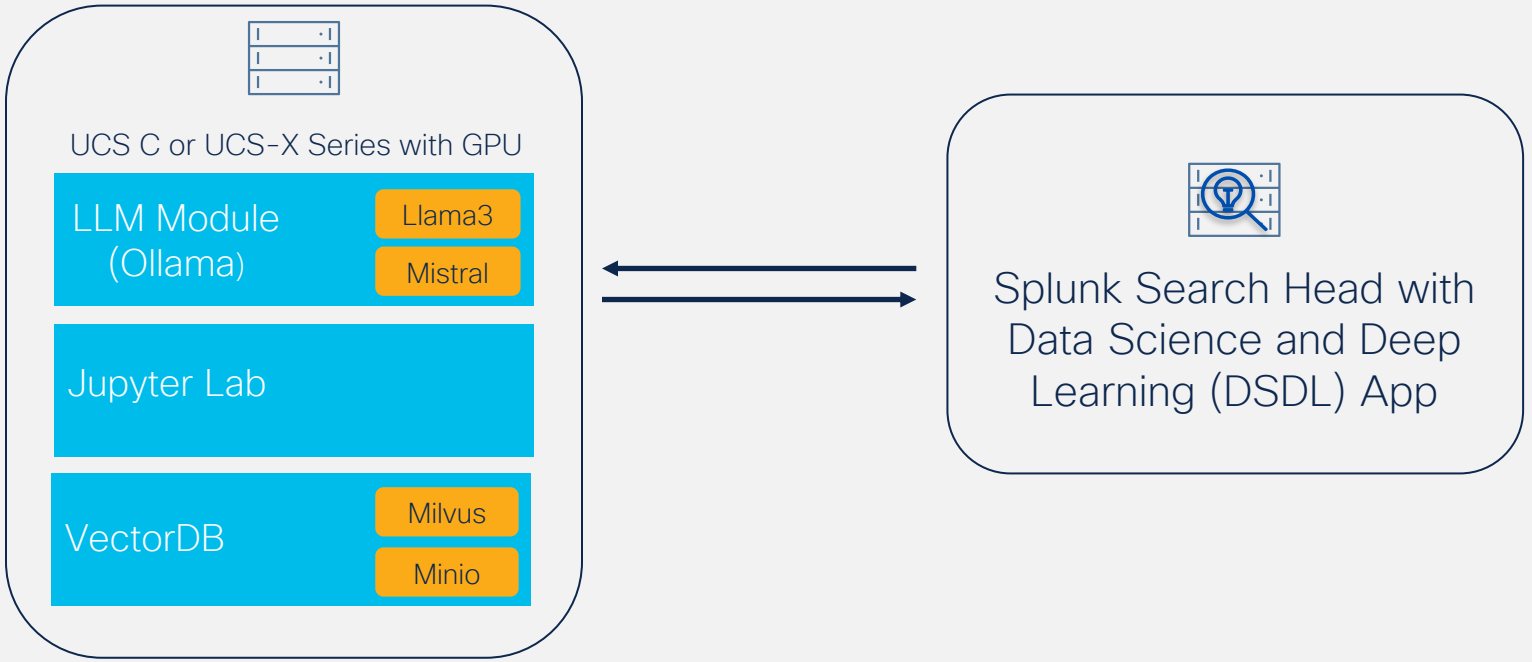


Building Your OWN IT Operations AI Assistant



Workload Architecture for Splunk + UCS

Leveraging the Splunk Data Science and Deep Learning App

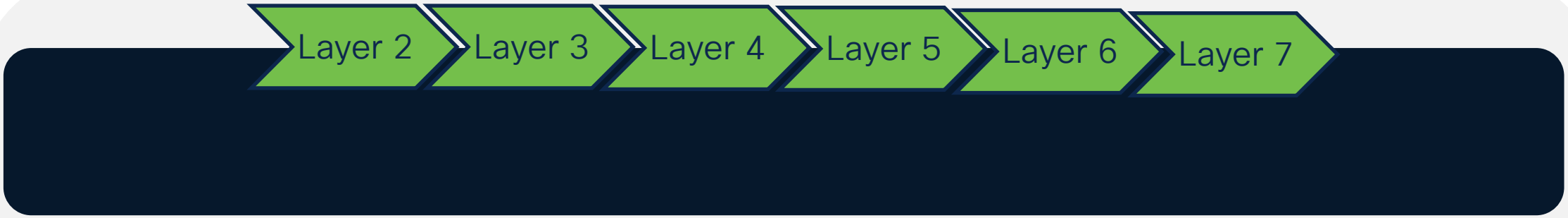


Example Query and Response

LLM-RAG Result

_time	query	RAG_Response	RAG_References
2024-11-20 17:22:07	what's happening to my app and why was it down? How was it fixed?	Based on the provided log messages, here is a detailed explanation of what happened to your application, why it was down, and how it was fixed: The Order Processing App was experiencing issues, leading to an alert being triggered from the Monitoring Agent. The app entered maintenance mode (switched to warning state) at 14:01:45 on ServerA. Upon further investigation, we can see that the app encountered critical errors, halting all processing (critical error log at 14:01:25). It attempted to reconnect to the database but was unsuccessful due to a timeout or connection failure (error logs at 14:01:10 and 14:01:05). Specifically: 1. The app could not connect to the database at IP address `192.168.2.10` on port `5432`, resulting in a "Connection timed out" error. 2. After three retries, the connection failed again, returning a "Service Unavailable" status. The primary issue here appears to be a communication problem between the application and the database server. The app was unable to establish a stable connection, likely due to network issues or configuration problems on ServerA. However, after some time (approximately 15 minutes), the application recovered from its critical errors (INFO log at 14:02:10). This indicates that the underlying infrastructure issues were resolved or mitigated. To further understand what happened, let's examine the events leading up to the downtime: * At 14:01:05, Interface FastEthernet0/24 on Switch1 went down, which might have introduced a temporary network issue affecting ServerA. However, this is just a separate event and not directly related to the database connection or application issues. * The Monitoring Agent's alert at 14:01:30 indicates that the Order Processing App was down, prompting further investigation. The app's recovery appears to be primarily due to the fact that the underlying network issue (Interface FastEthernet0/24 on Switch1) was addressed, and the database connection problem was resolved. It is unclear if any other issues were present during the downtime, but it is clear that the critical errors caused a halt in processing. To fix this issue, you should address both the potential network configuration problems and ensure reliable communication with your database server: * Verify the database server's availability and stability. * Investigate the cause of the intermittent connectivity issues between ServerA and the database (IP address `192.168.2.10`). * Ensure that any necessary patches or updates are applied to prevent similar incidents in the future. **Please note**: Without more detailed information about your specific environment, it is challenging to provide a definitive diagnosis. However, this analysis should give you a good starting point for understanding what happened and potential steps to take for prevention and resolution.	2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA 2024-10-10 14:01:30,MonitoringAgent,ALERT,OrderProcessingApp is down on ServerA

Connecting the Layers

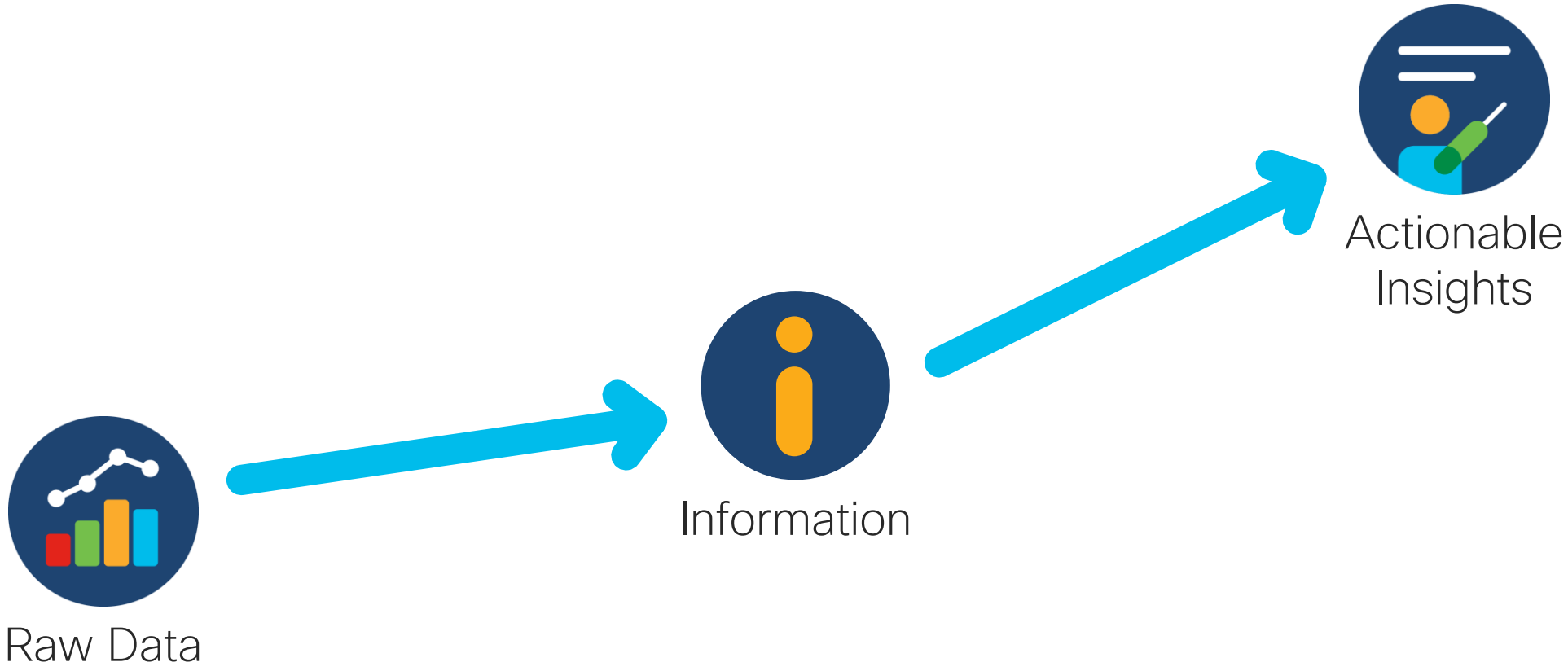


The People

What can AI do for Network Analytics?

The Path to Predictive Networks

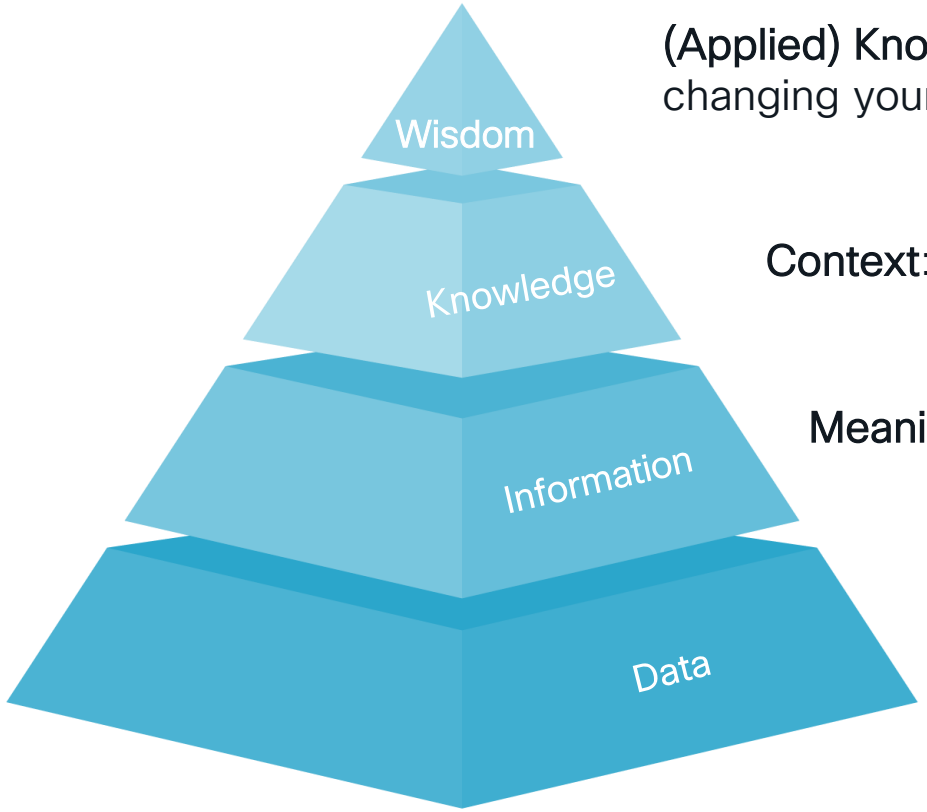
Accelerating Data Driven Operations



Expose Actionable Insights to Operations Team

The Path to Predictive Networks

The Role of Data



(Applied) Knowledge: You can avoid an outage by changing your config.

Context: Microsoft is experiencing an outage.

Meaning: An Office365 test in London UK has failed.

Raw Data: unprocessed facts.

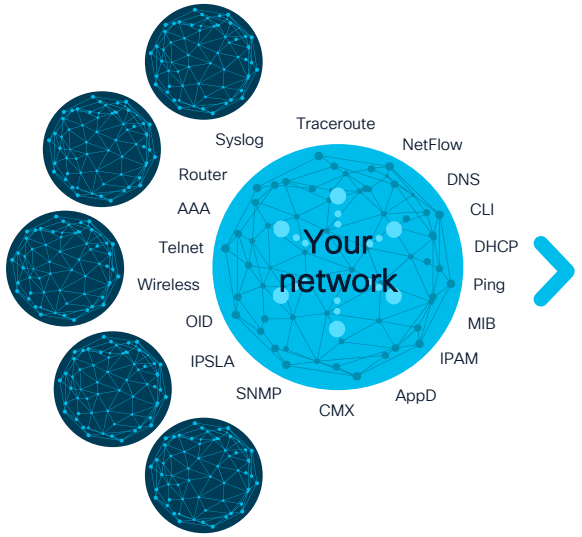
Cisco Advantage

Best Data, Best Knowledge Base

Diverse data:
Network, application, security

Diverse networks:
Local and global

Streaming telemetry :
Cloud connected



Worldwide data platform
Anonymized data

AI/ML

supervised vs unsupervised, reinforcement/active learning, adversarial or ensemble methods, Convolutional Neural Networks

Accurate insights

Improved performance

IT Experience

40 years of top engineering knowledge

Communities

Cisco Fellows

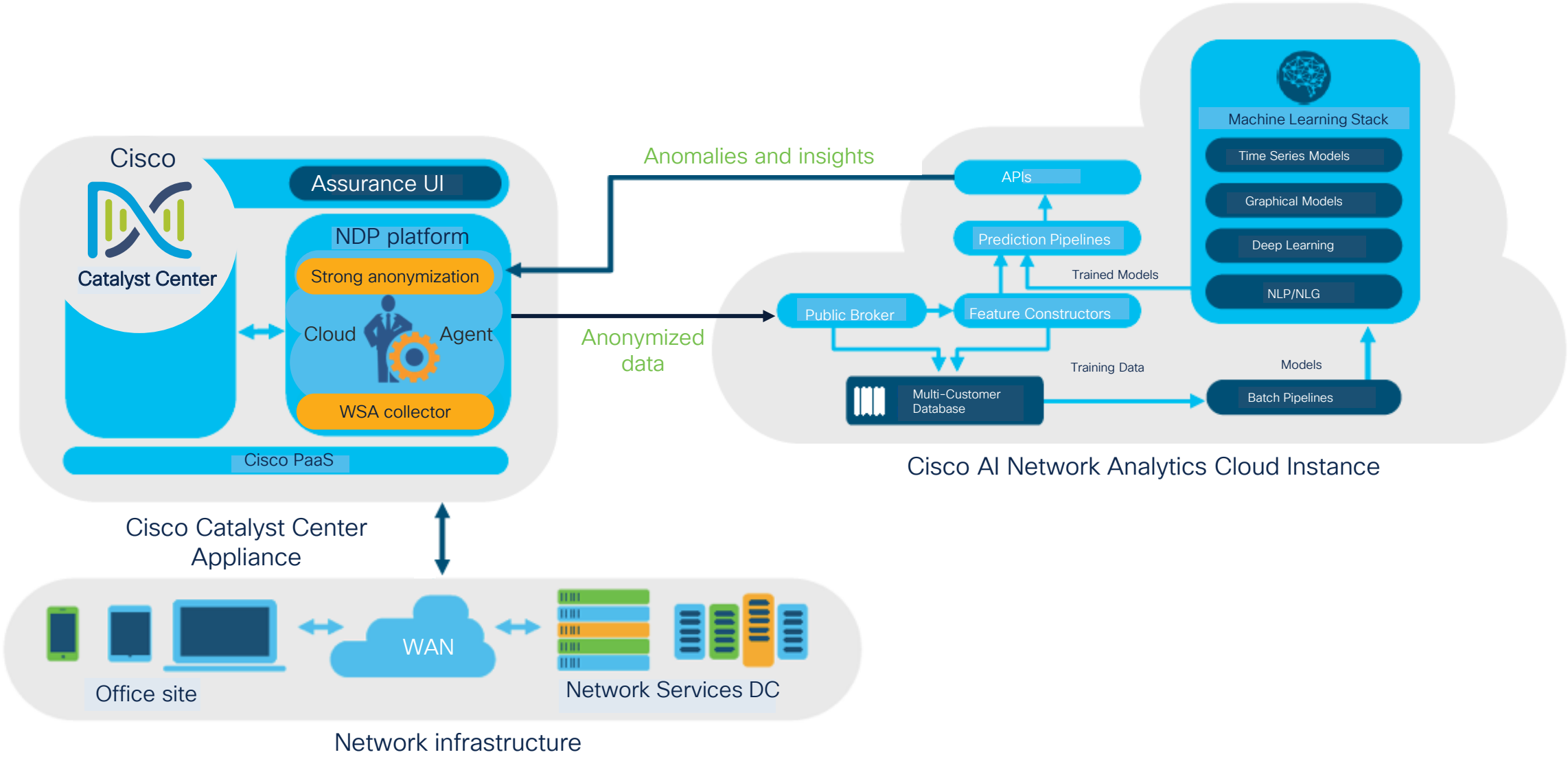
CX

TAC

Cisco Validated Designs

Knowledge base

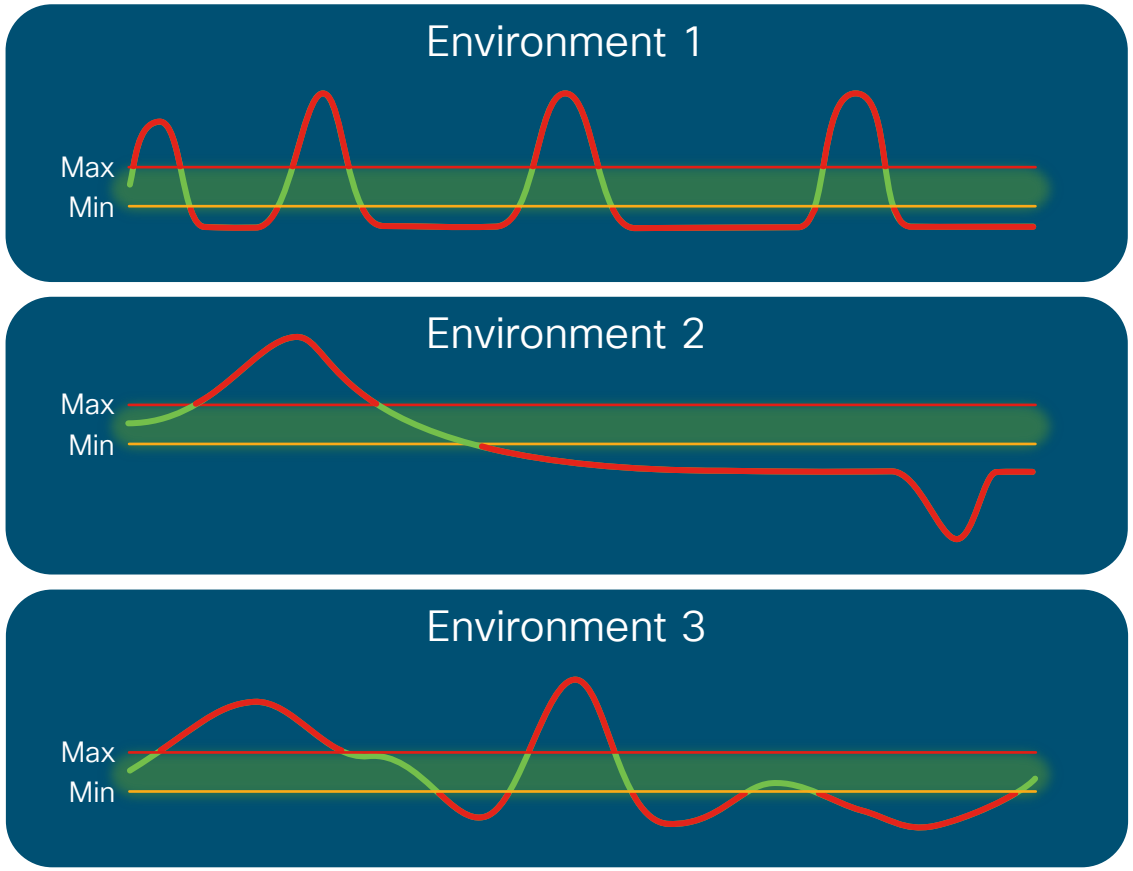
Cisco AI Network Analytics architecture



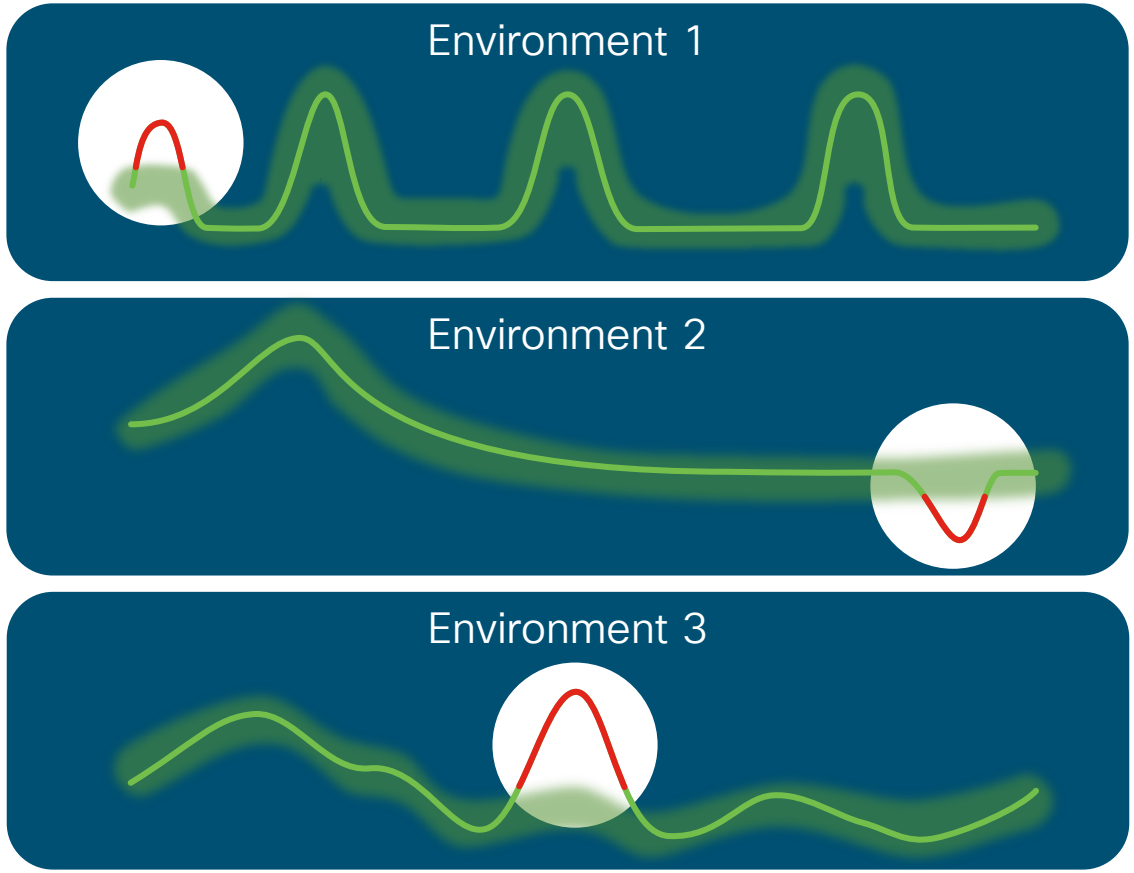
Improve incident alert fidelity

Personalized baselining

Before: Custom thresholds = Alert overload



AI-driven: Dynamic baselines = relevant anomalies



AI driven wireless problem identification

Wireless onboarding

Wireless user failed to connect
Wireless user took too long to connect

- Excessive Onboarding Time
- Excessive Failures
- Excessive DHCP Time
- Excessive DHCP Failures
- Excessive AAA Time
- Excessive AAA Failures
- Excessive Assoc. Time
- Excessive Assoc. Failures

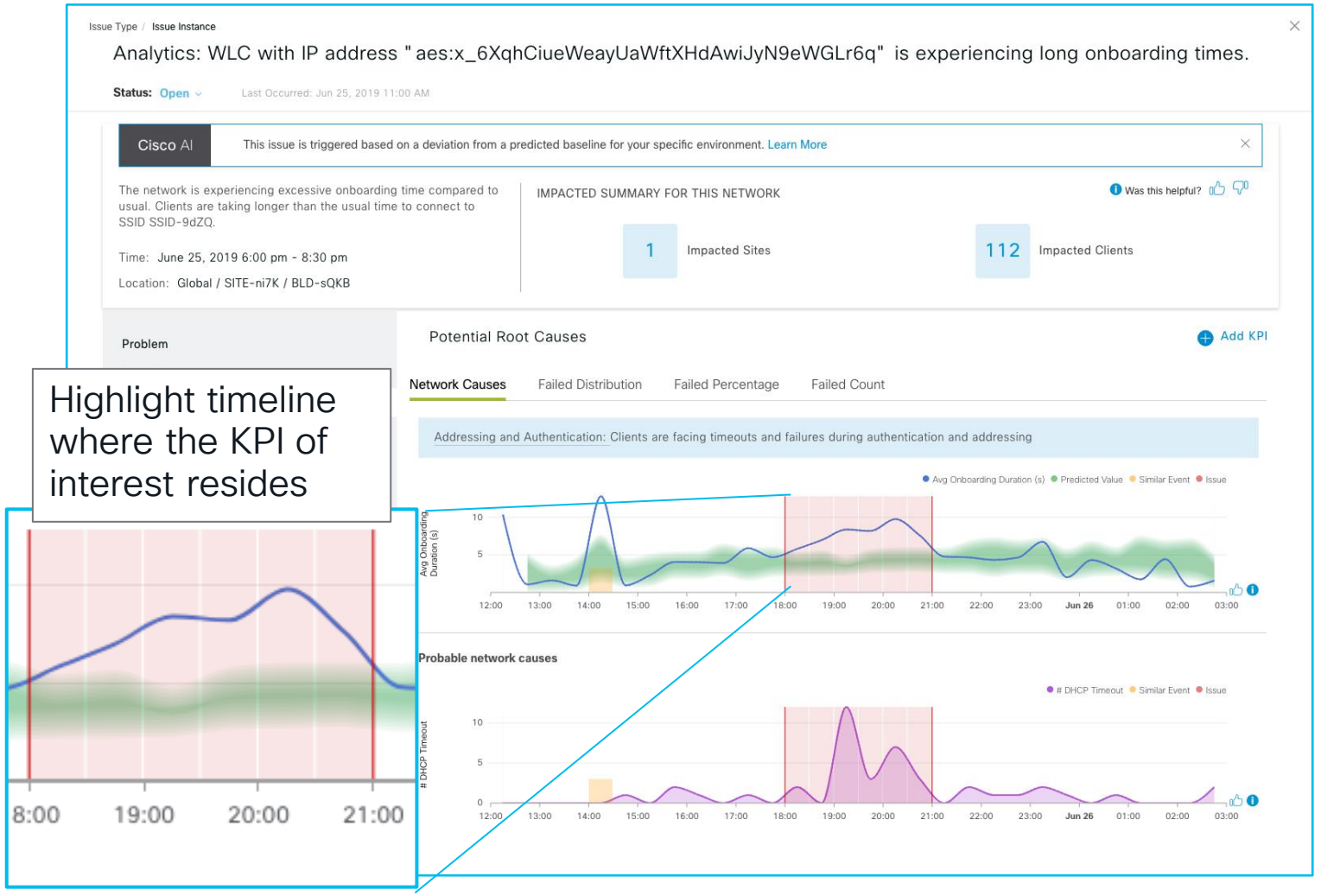
Application experience

Wireless user's application throughput is declining

- Total Radio Throughput
- Media Application Throughput
- Cloud Application Throughput
- Social Application Throughput

Use Case: Accelerated Remediation

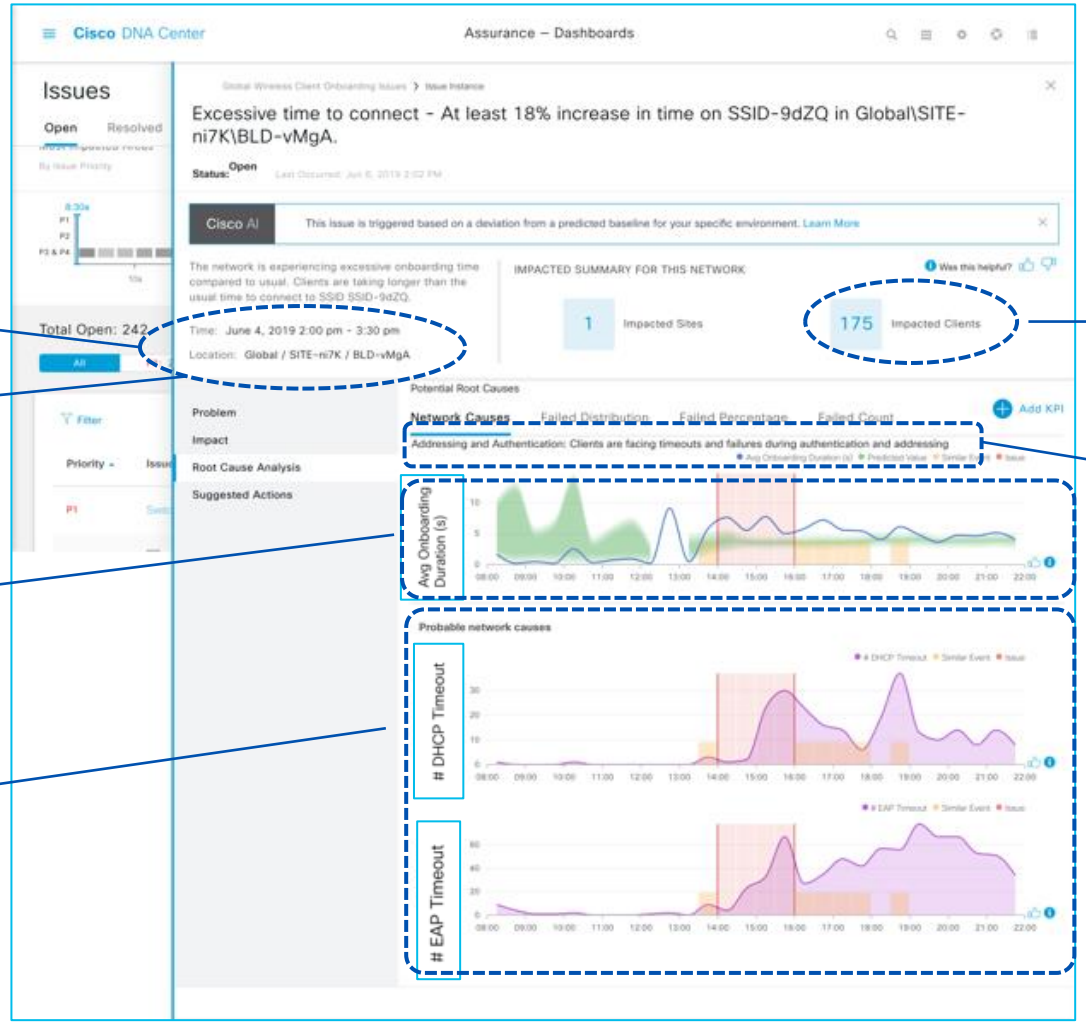
AI driven issue identification



- Dynamically generated “Green Band”, expected normal range of client experience
- Predictive model that derived from number of variables and KPIs
- Identify 5W1H – When, Where, Who, What, Why and How

Use Case: Accelerated Remediation

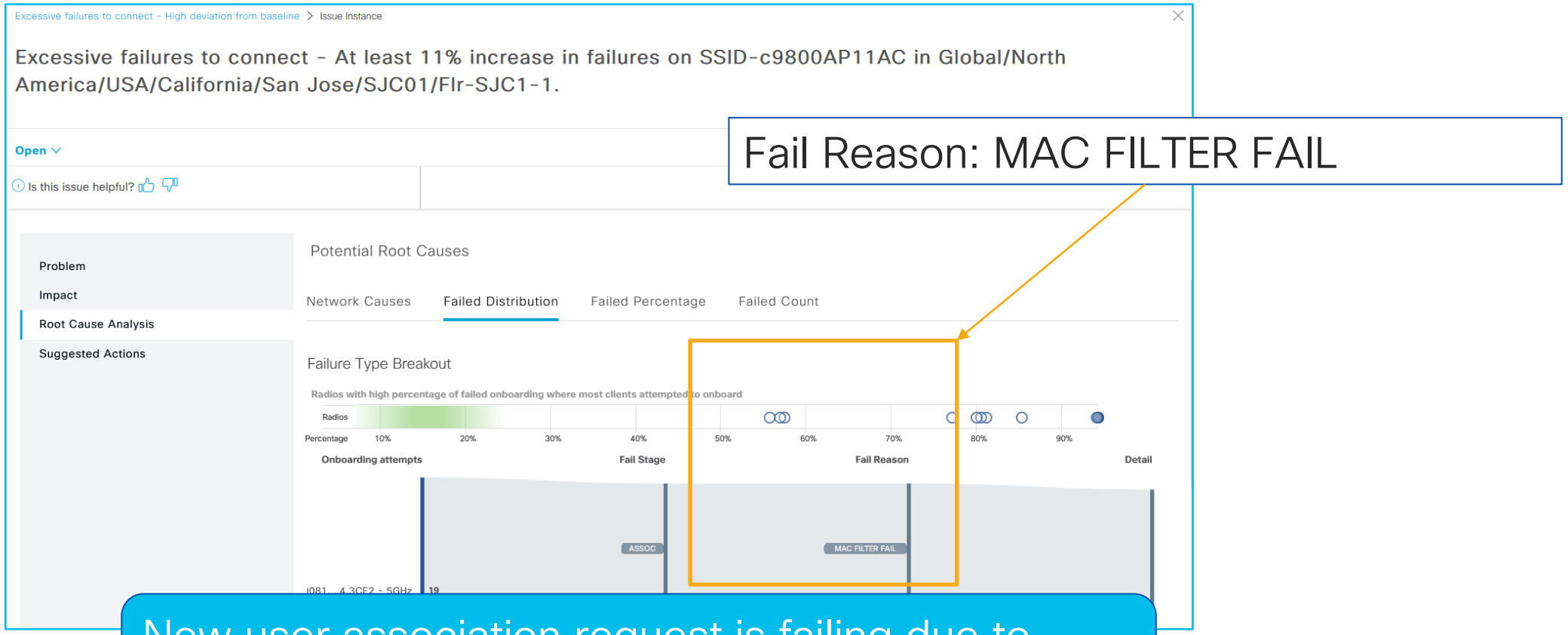
Root cause analysis



"Clients are facing timeouts and failures during authentication and addressing"

Use Case: Accelerated Remediation

Go deep into why

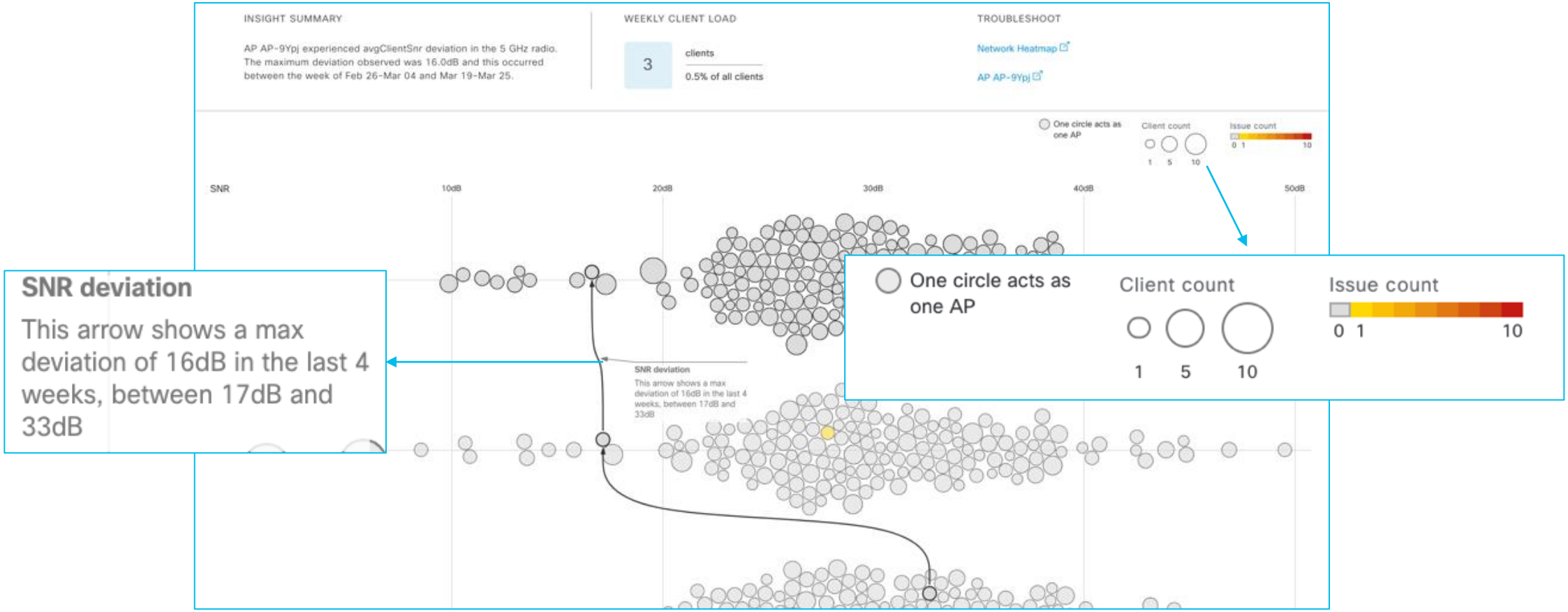


Fail Reason: MAC FILTER FAIL

New user association request is failing due to MAC Filtering

Use Case: Proactive & Predictive Insights

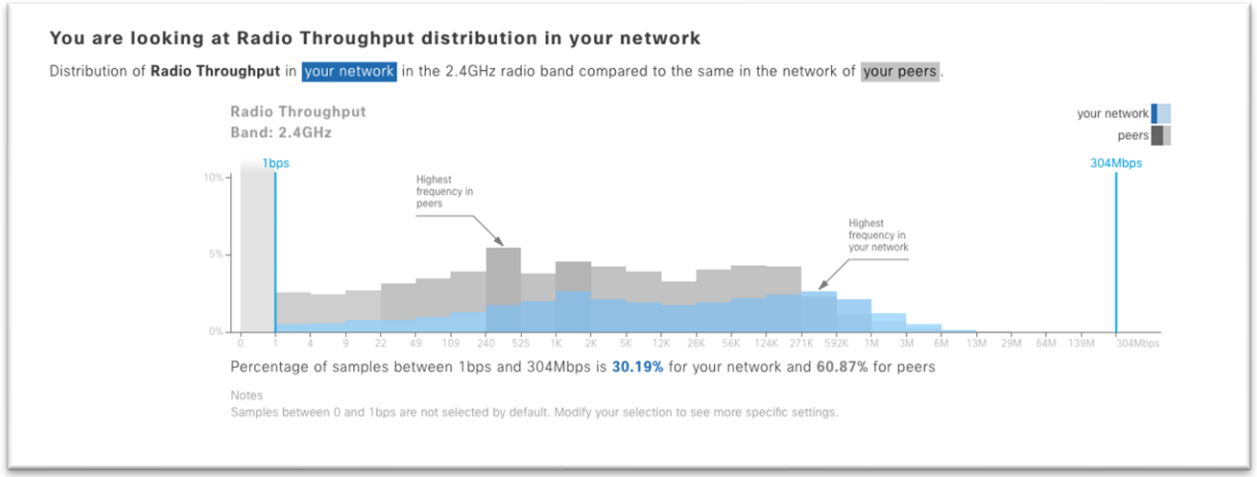
Example: Long-term Trending and Behavior Change



Insight – Found AP exposed to HUGE noise for last 2 weeks

Use Case: Proactive & Predictive Insights

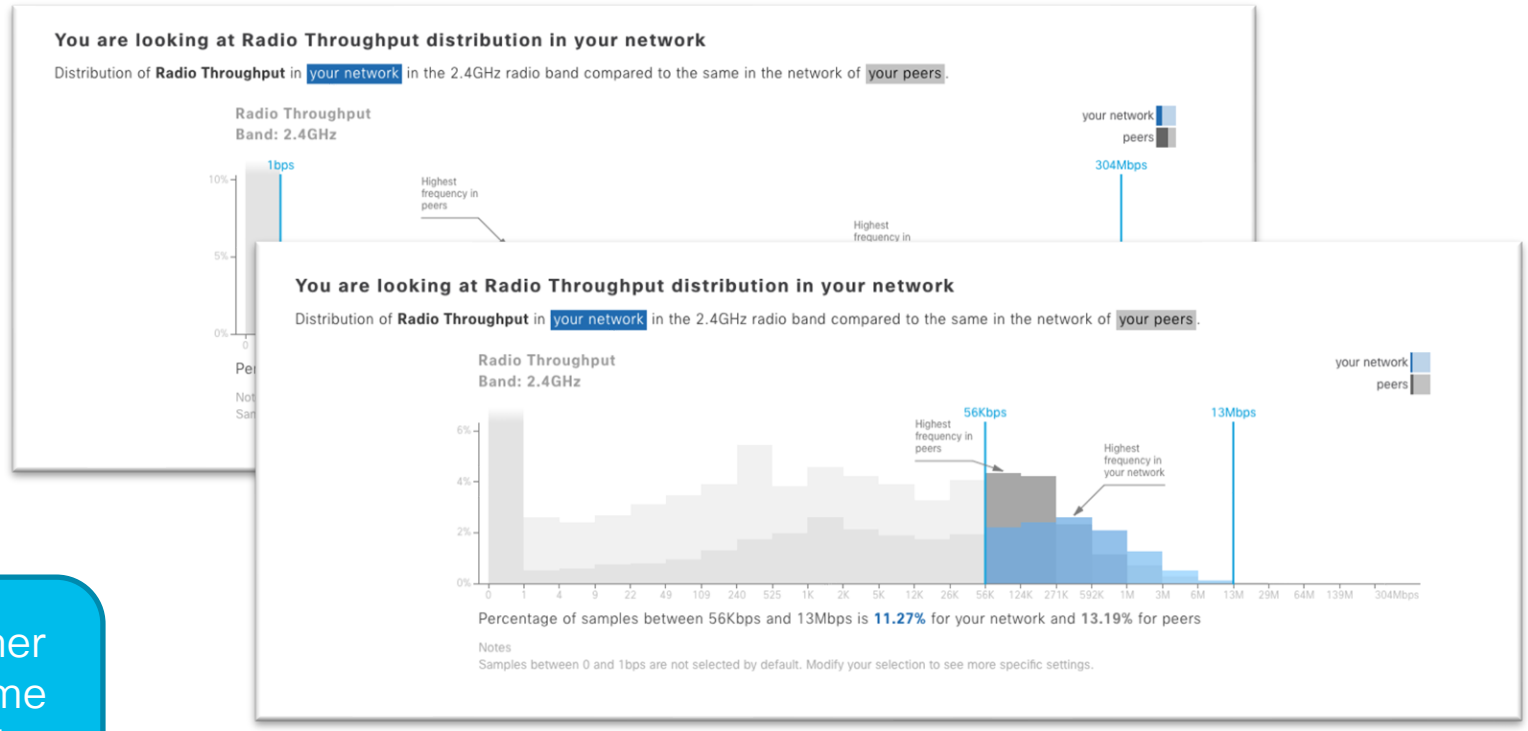
Compare KPIs with peers



Compare with other devices in the same network, time and ... comparable networks

Use Case: Proactive & Predictive Insights

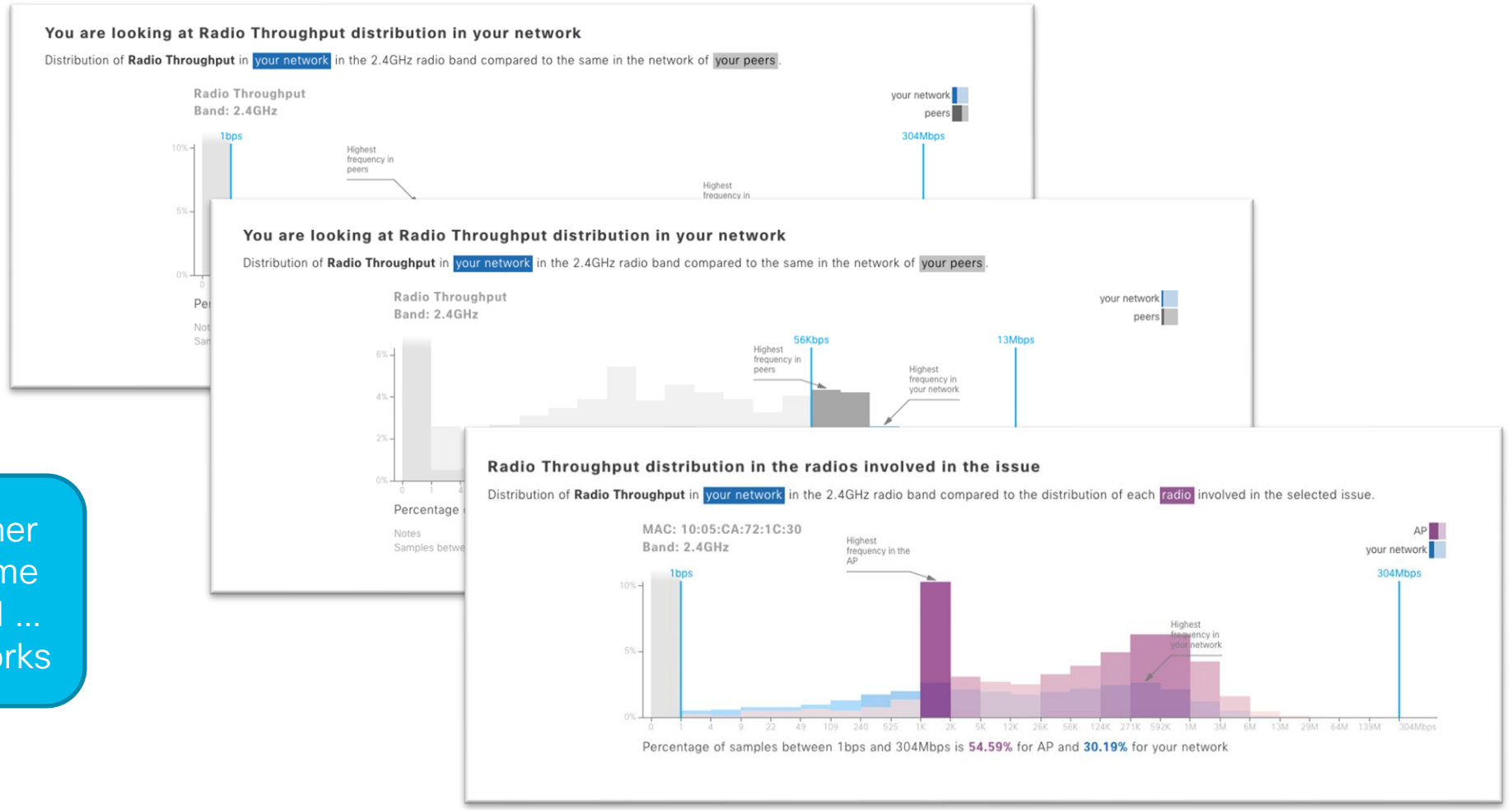
Compare KPIs with peers



Compare with other devices in the same network, time and ... comparable networks

Use Case: Proactive & Predictive Insights

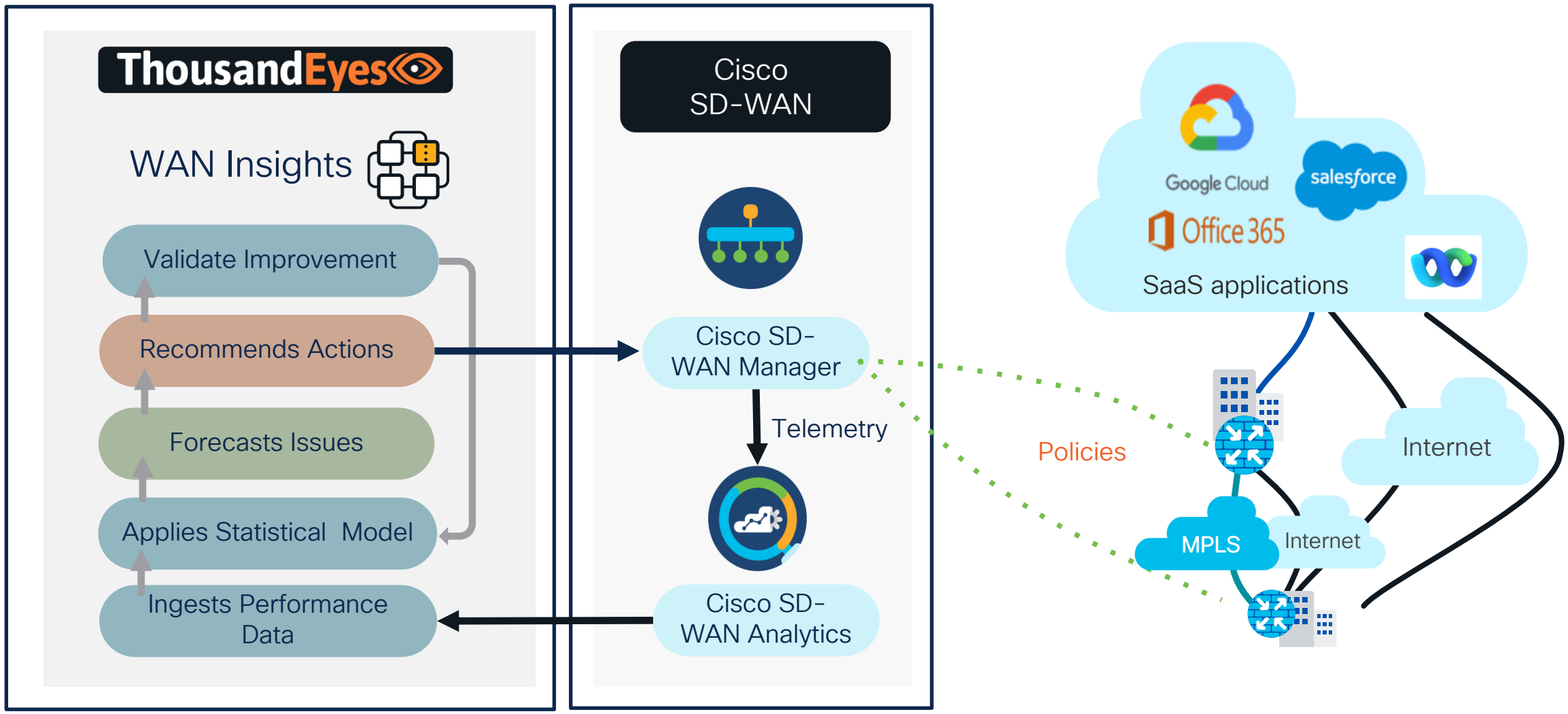
Compare KPIs with peers



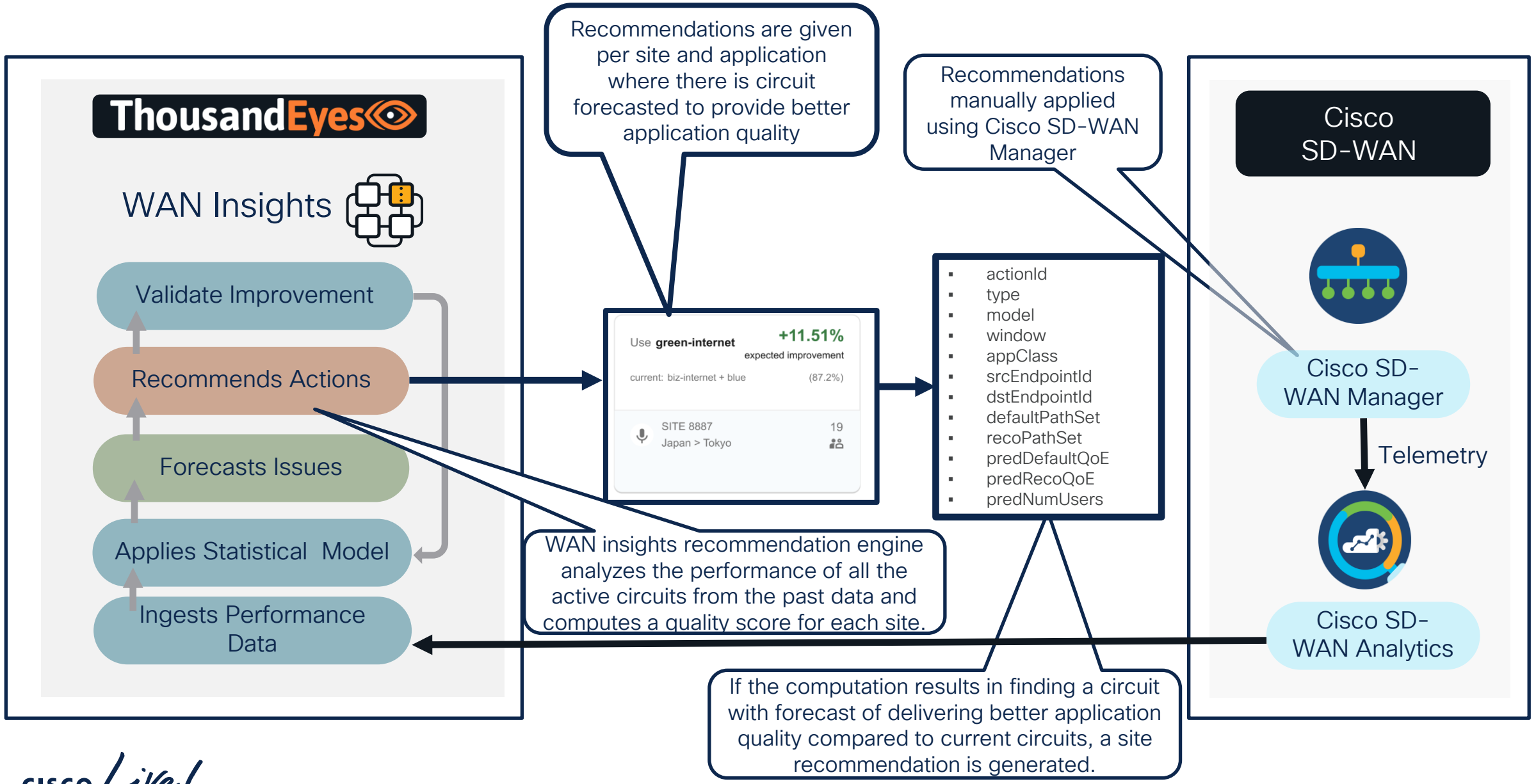
Compare with other devices in the same network, time and ... comparable networks

Predictive optimization with Cisco Catalyst SD-WAN

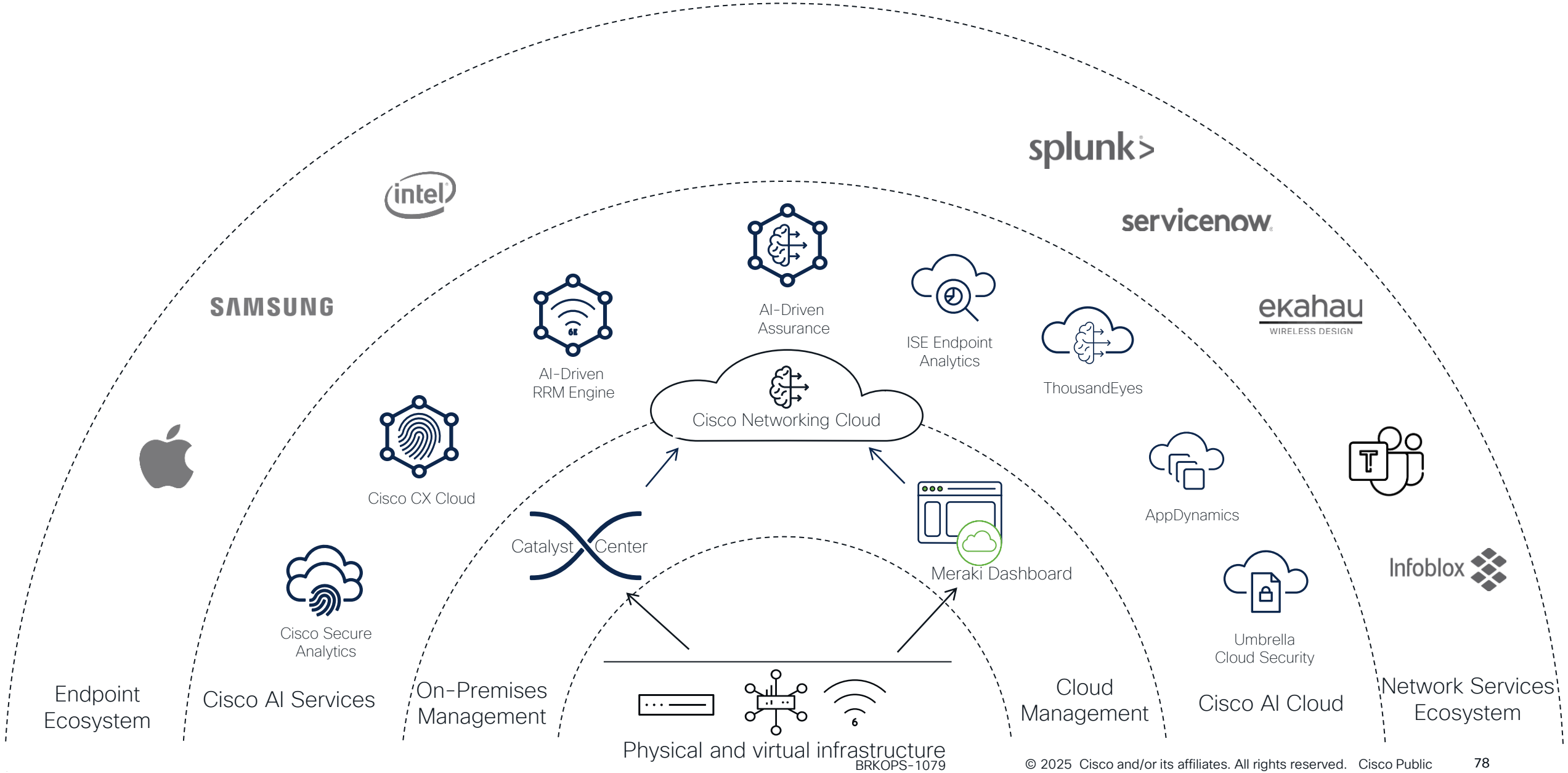
Improved user experience



Understanding and applying recommendations



Full-Stack Network Intelligence



Call to Action:

- Engage with business stakeholders
- Ask the questions about the AI outcomes your organization wants to implement
- Leverage AI in Cisco solutions and lead the technology conversation for the AI transformation

Webex App

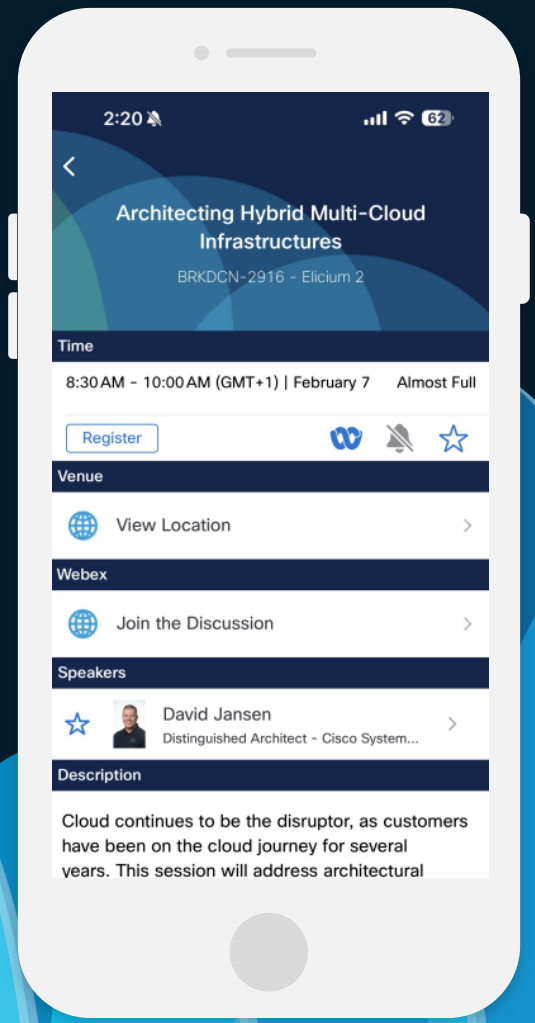
Questions?

Use the Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events mobile app
2. Click “Join the Discussion”
3. Install the Webex app or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until February 28, 2025.



Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

All surveys can be taken in the Cisco Events mobile app or by logging in to the Session Catalog and clicking the ‘Participant Dashboard’

Content Catalog

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at ciscolive.com/on-demand. Sessions from this event will be available from March 3.

Thank you

cisco Live!