Cisco Silicon for AI -
Capabilities, Designs, and Results

CISCO Live!

Peter Jones
Distinguished Engineer

Dave Zacks
Distinguished Engineer

BRKARC-2095

Cisco ~~Silicon~~ *Hardware* for AI - Capabilities, Designs, and Results

CISCO Live!

Peter Jones
Distinguished Engineer

Dave Zacks
Distinguished Engineer

#HighBitRate

BRKARC-2095

Cisco Webex App

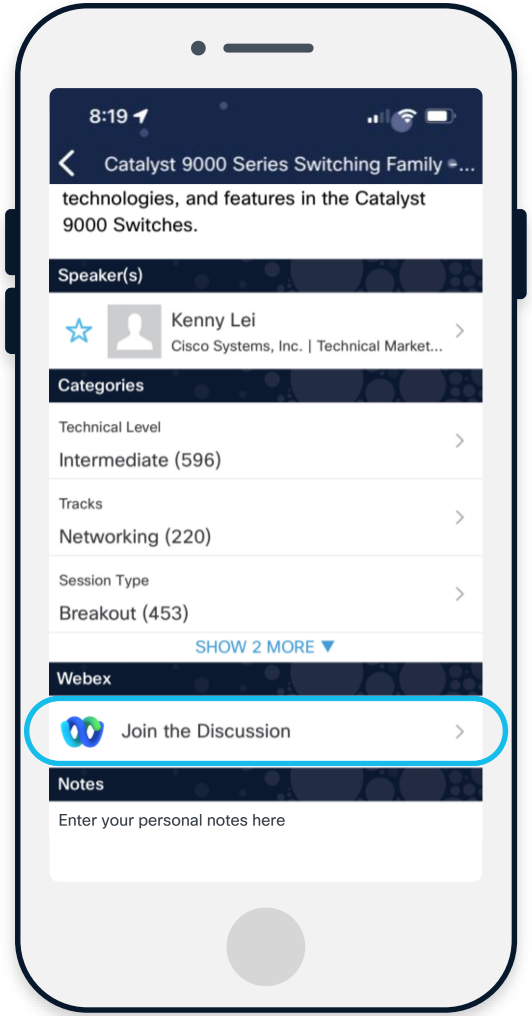
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

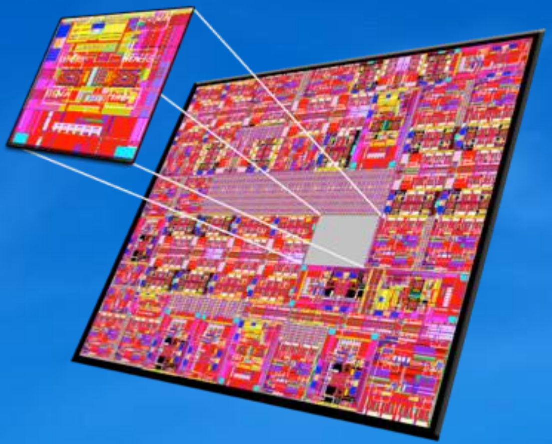
1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 13, 2025.



Agenda

By Way of Introduction ...



I am a **Distinguished Engineer** in the Cisco Security Innovations CTO team, and have been with Cisco for 25 years.

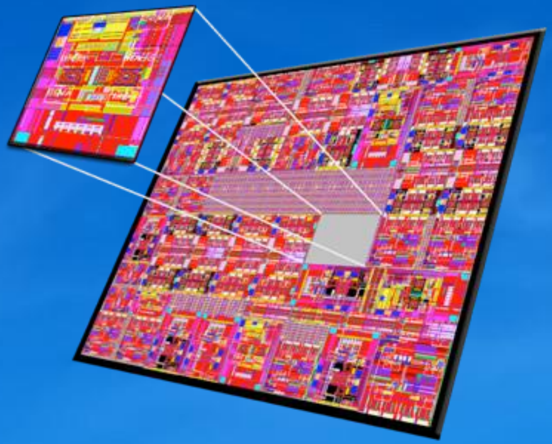
I work primarily with large, high-performance Enterprise network architectures, designs, and systems. I have over 30 years of experience with designing, implementing, and supporting solutions with many diverse network technologies.

I have a strong background in, and focus on, customer requirements, and integrating these into the products and solutions Cisco builds.
I have a special interest in **Flexible Hardware, Fabrics, Assurance** and **ML/AI**.

Dave Zacks
Distinguished Engineer

Email: dzacks@cisco.com
Bluesky: davezacks.bsky.social
LinkedIn: In/dave-zacks-43677474/

By Way of Introduction ...



I am a **Distinguished Engineer** in the Cisco Networking Hardware team and have been with Cisco since 2005.

I work on system architecture and standards strategy across the portfolio.
I was a key figure in the development of the UADP switching ASIC architecture and the Catalyst switches that use it.

I work in defining and promoting new Ethernet standards in IEEE 802.3 and as Ethernet Alliance Chairman.

I am passionate about **Network Evolution, Adoptable Technology** and **Ethernet.**

Peter Jones
Distinguished Engineer

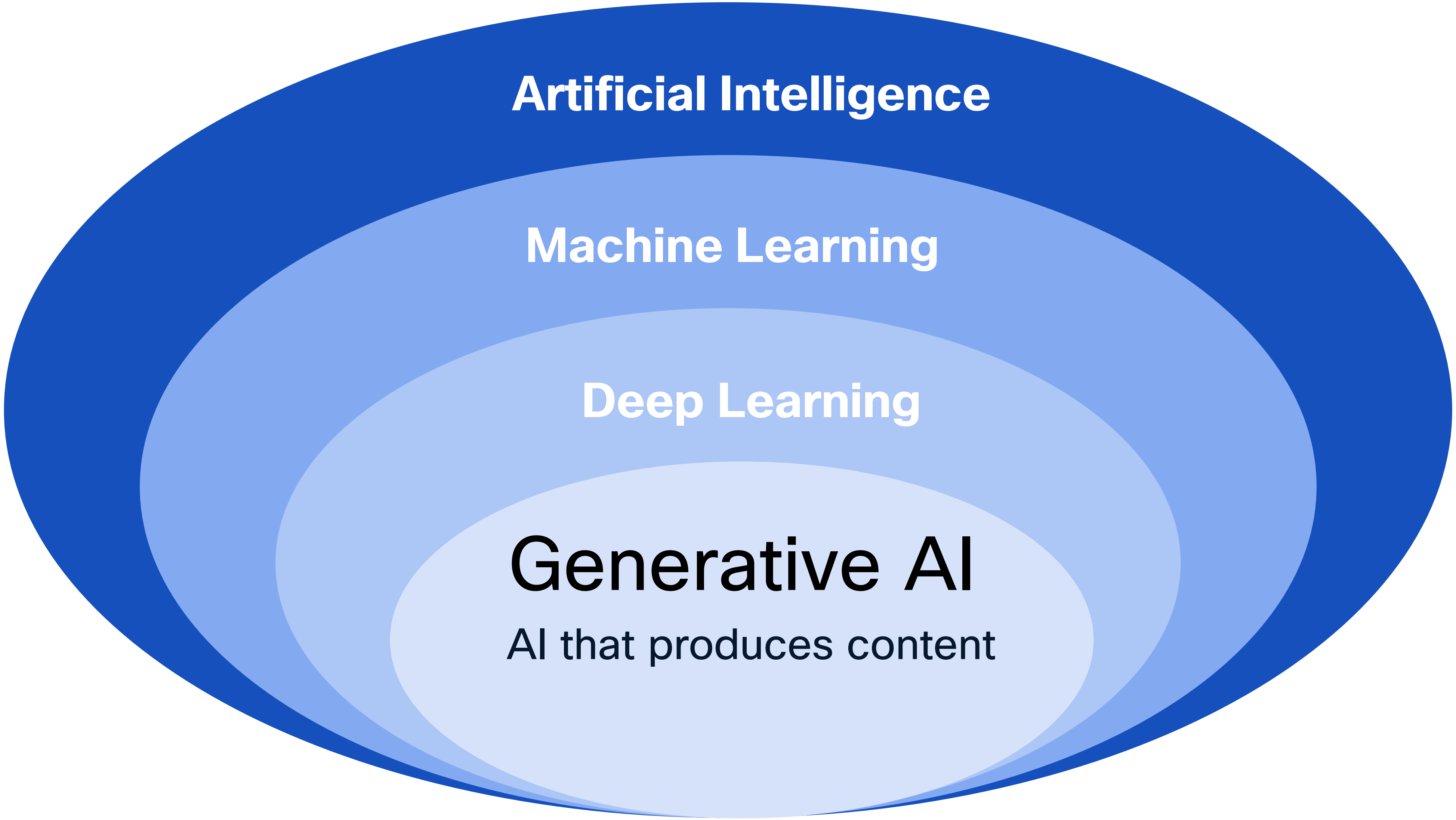
Email: petejone@cisco.com
Bluesky: petergjones.bsky.social
LinkedIn: in/petergjones/

What's this AI thing?

Wikipedia: Blind men and an elephant

AI – Overview

The Breakdown of Artificial Intelligence



How Is AI Different From Regular Algorithms?

Regular Algorithms

Input

Rule-Based

Deterministic

Output

Artificial Intelligence (AI)

Input

Learning-Based

Pattern-Based

Self-Improve

Adaptive to Input

Output

Supervised Learning

Supervised Learning

Using past "labeled" data to predict future trends

- Spam email identifier
- Stock price prediction
- Sales forecast

Note: Labeled data is data that has been tagged with the correct answer or output

Scenario: Predicting if an Email is Spam

Email 1
- To/From
- Subject
- Content

Email 2
- To/From
- Subject
- Content

Email 3
- To/From
- Subject
- Content

Email 4
- To/From
- Subject
- Content

Labels → Spam Not Spam

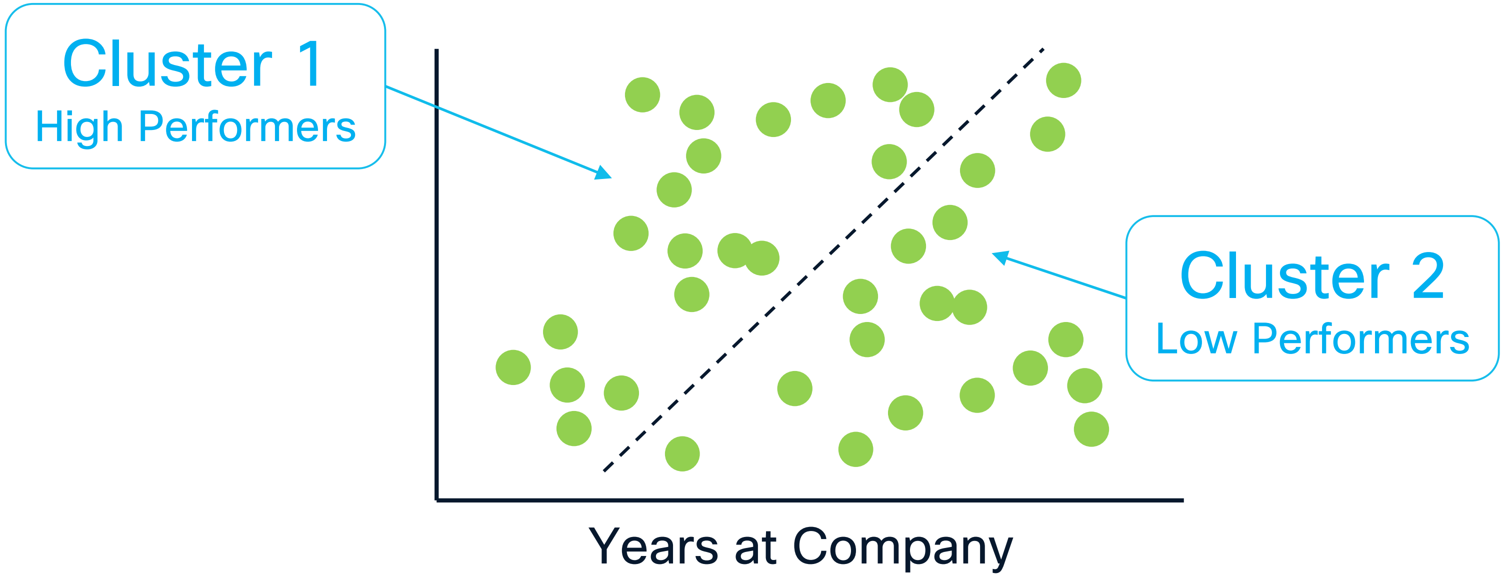
Unsupervised Learning

Scenario: Predicting if an employee is going to be a top performer

Unsupervised Learning

Using “Unlabeled” data to learn patterns

- User segmentation
- Anomaly detection
- Image/Video analysis



Note: Unlabeled data refers to data that does not have predefined categories or outputs

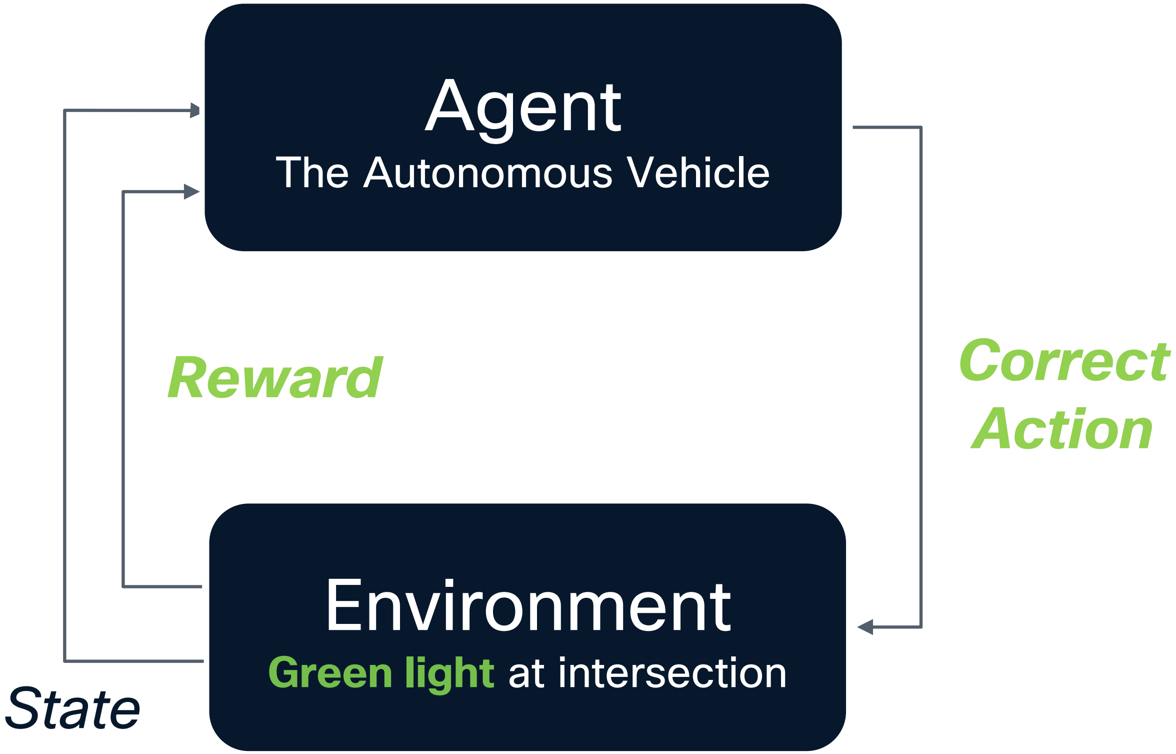
Unlabeled Data → Employee Data

Reinforcement Learning

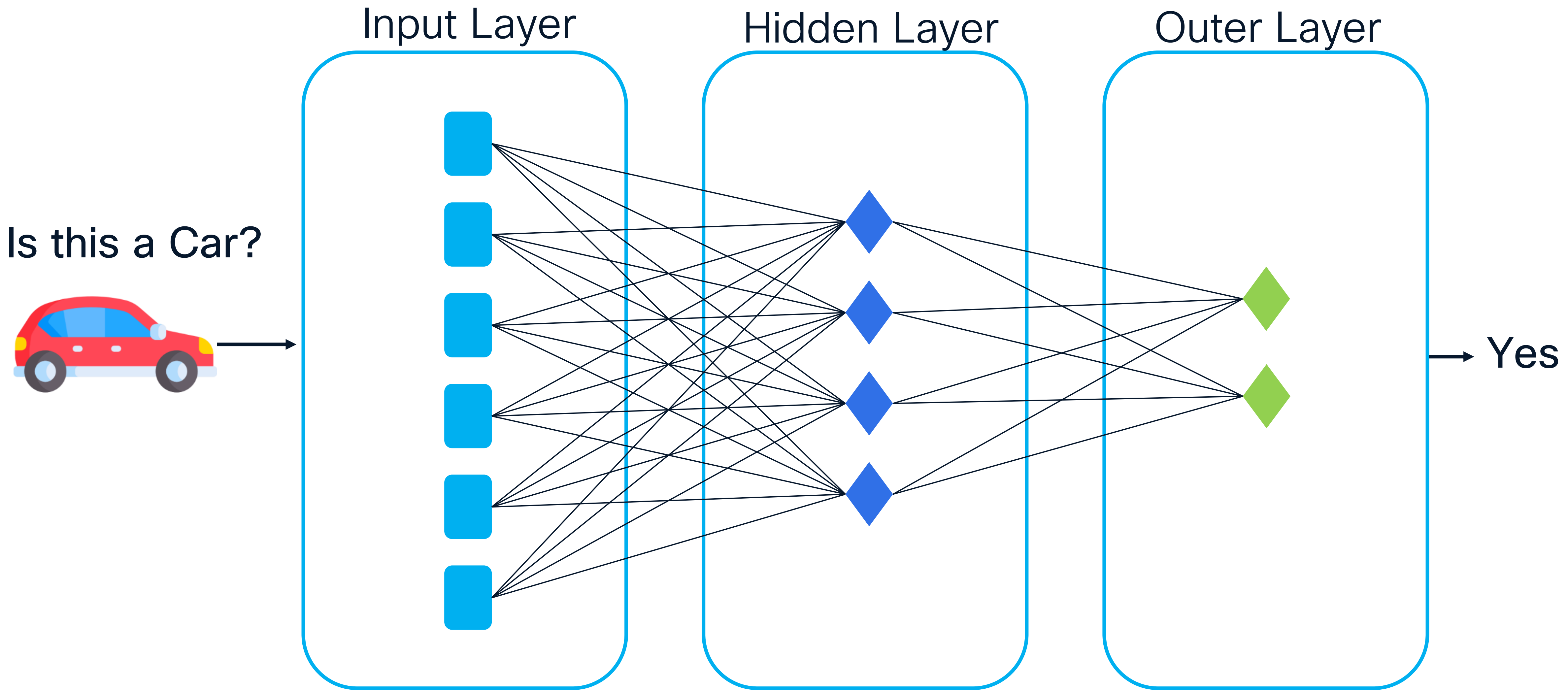
Reinforcement Learning

Trained on reward or penalty feedback loop based on its actions during simulations.

- Autonomous vehicles
- Robotics
- Resource management



Neural Networks – Identify Patterns with Deep Learning
Divide and conquer large amounts of complex data



Large Language Models and Diffusion Models

Large Language Models

Trained to create text content.

Ex: ChatGPT 4o

Diffusion Models

Trained to create image and video content.

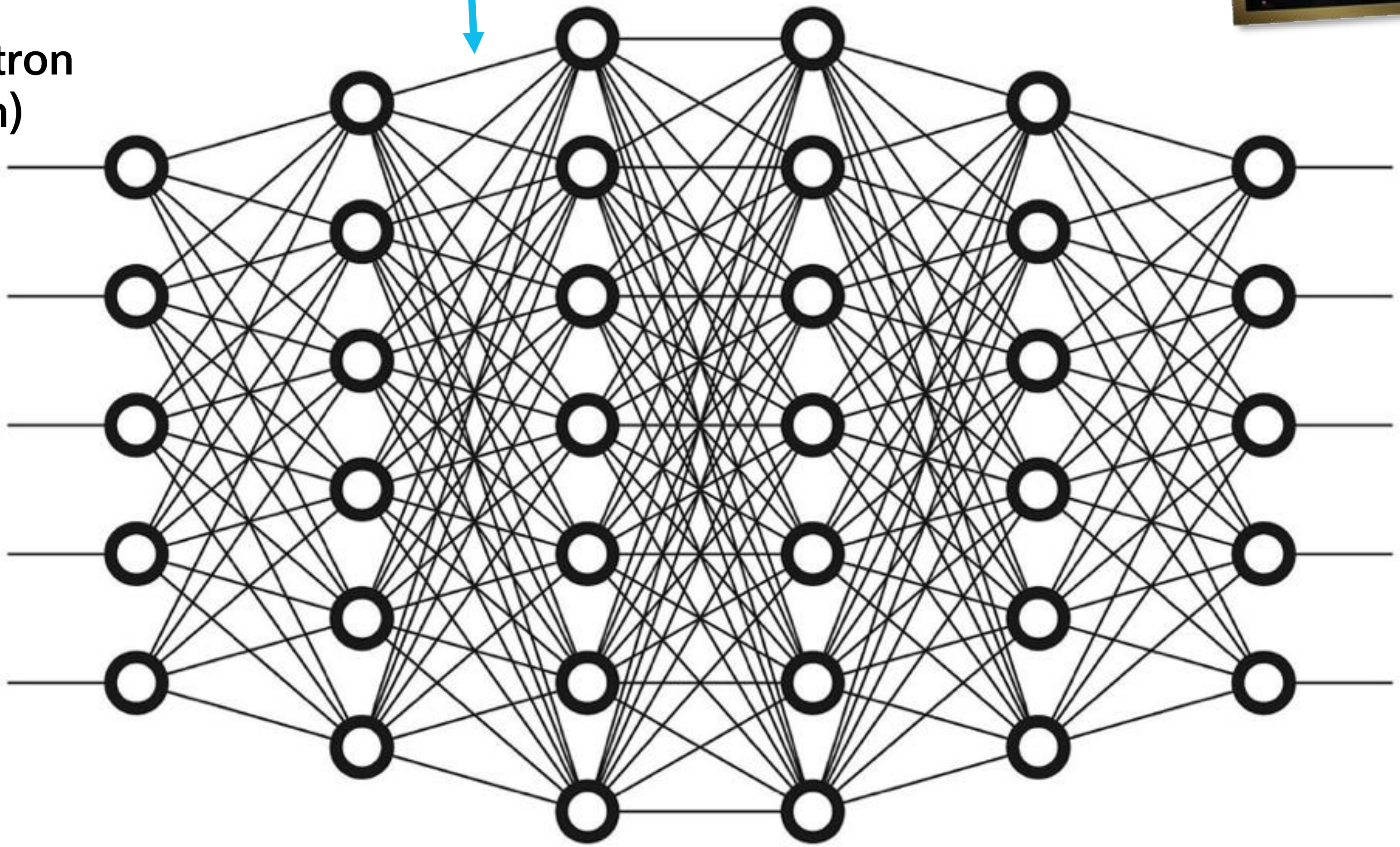
Ex: DALL·E 3

Why is this happening now?

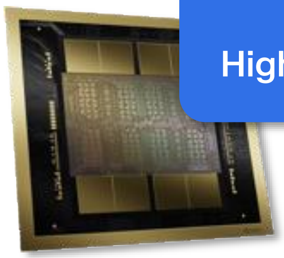
Scale

Perceptron (Neuron)

Parameter (Synapses)



NVIDIA Blackwell



Advances in Silicon –
High-density, High-performance GPUs

NVIDIA Flagship Accelerator Specification Comparison

	B200	H100	A100 (80GB)
FP32 CUDA Cores	A Whole Lot	16896	6912
Tensor Cores	As Many As Possible	528	432
Boost Clock	To The Moon	1.98GHz	1.41GHz
Memory Clock	8Gbps HBM3E	5.23Gbps HBM3	3.2Gbps HBM2e
Memory Bus Width	2x 4096-bit	5120-bit	5120-bit
Memory Bandwidth	8TB/sec	3.35TB/sec	2TB/sec
VRAM	192GB (2x 96GB)	80GB	80GB

Geoffrey Hinton – the "Godfather" of Deep Learning

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

arXiv:1706.03762 [cs.CL]

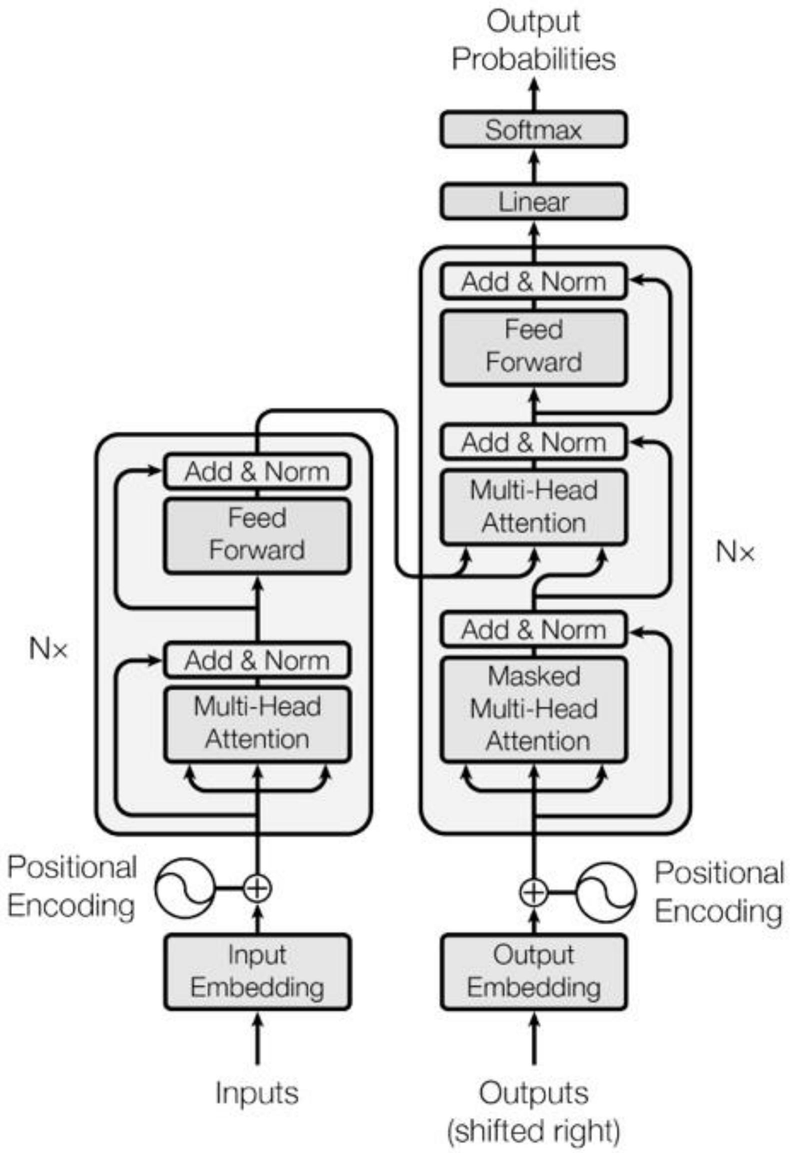


Figure 1: The Transformer - model architecture.

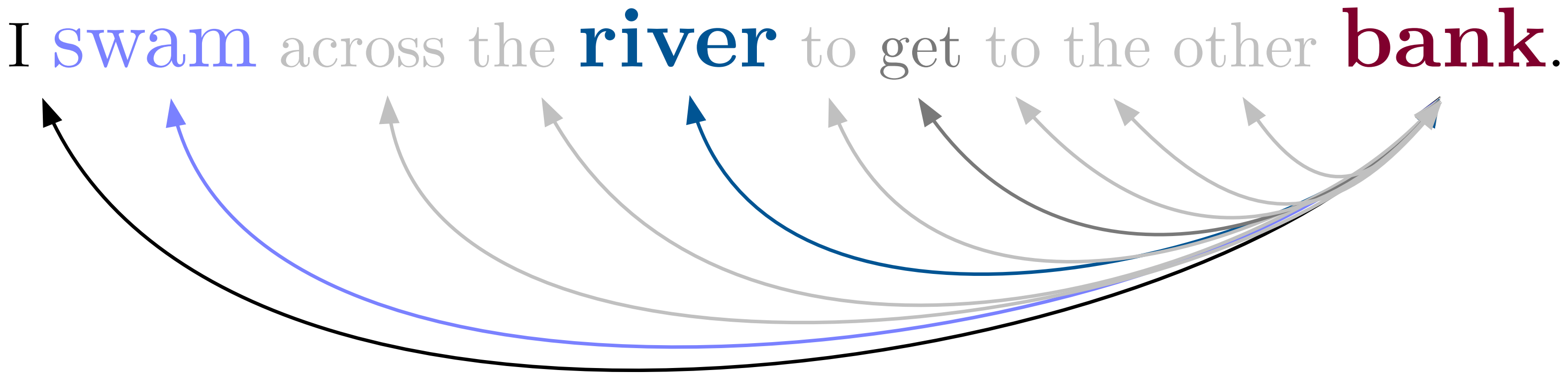
Attention Mechanism - Overview

Foundation for Transformer Models

You have no problem interpreting "bank" in the following sentence:

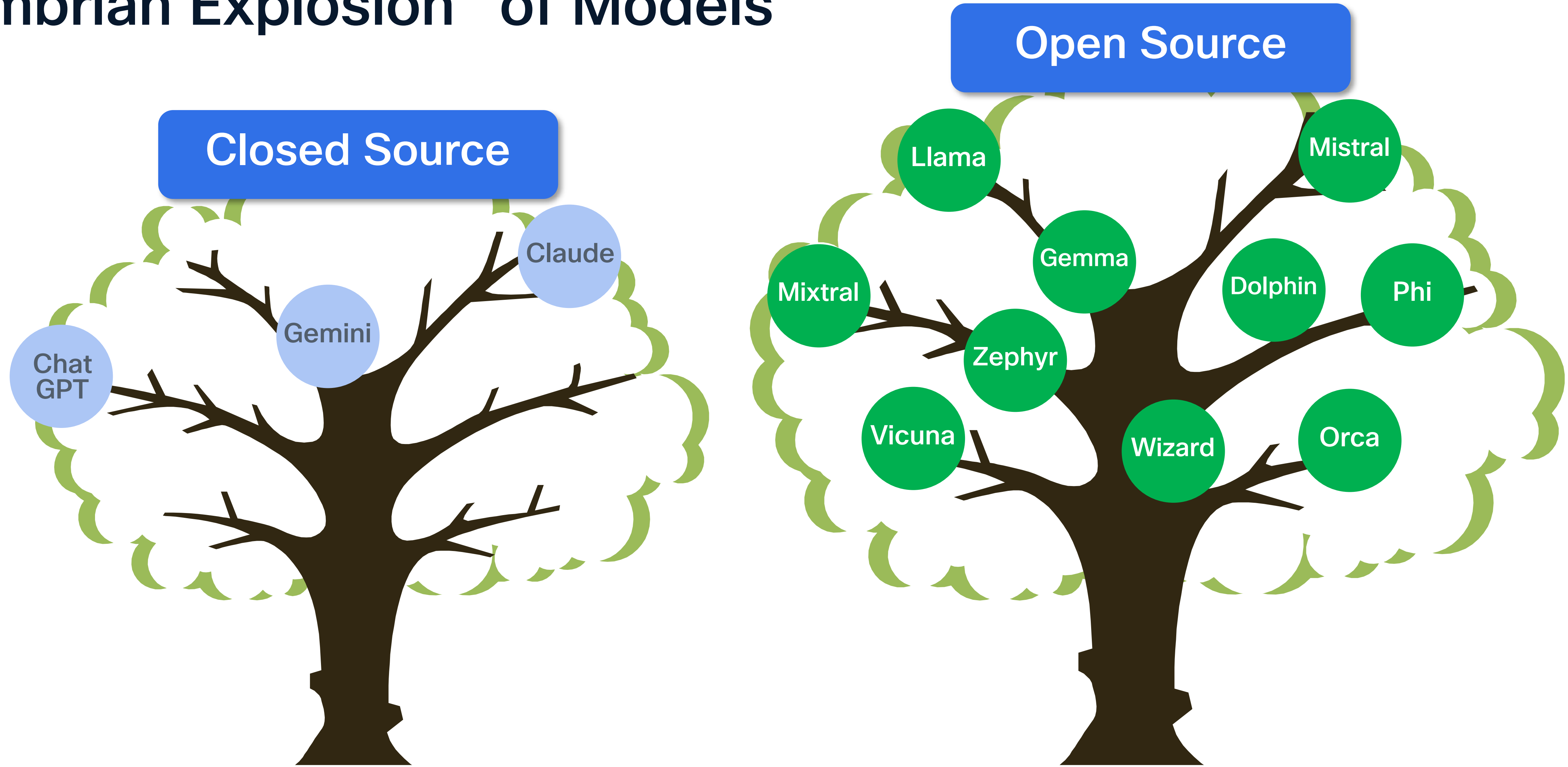
"I swam across the river to get to the other bank."

A machine needs some help...



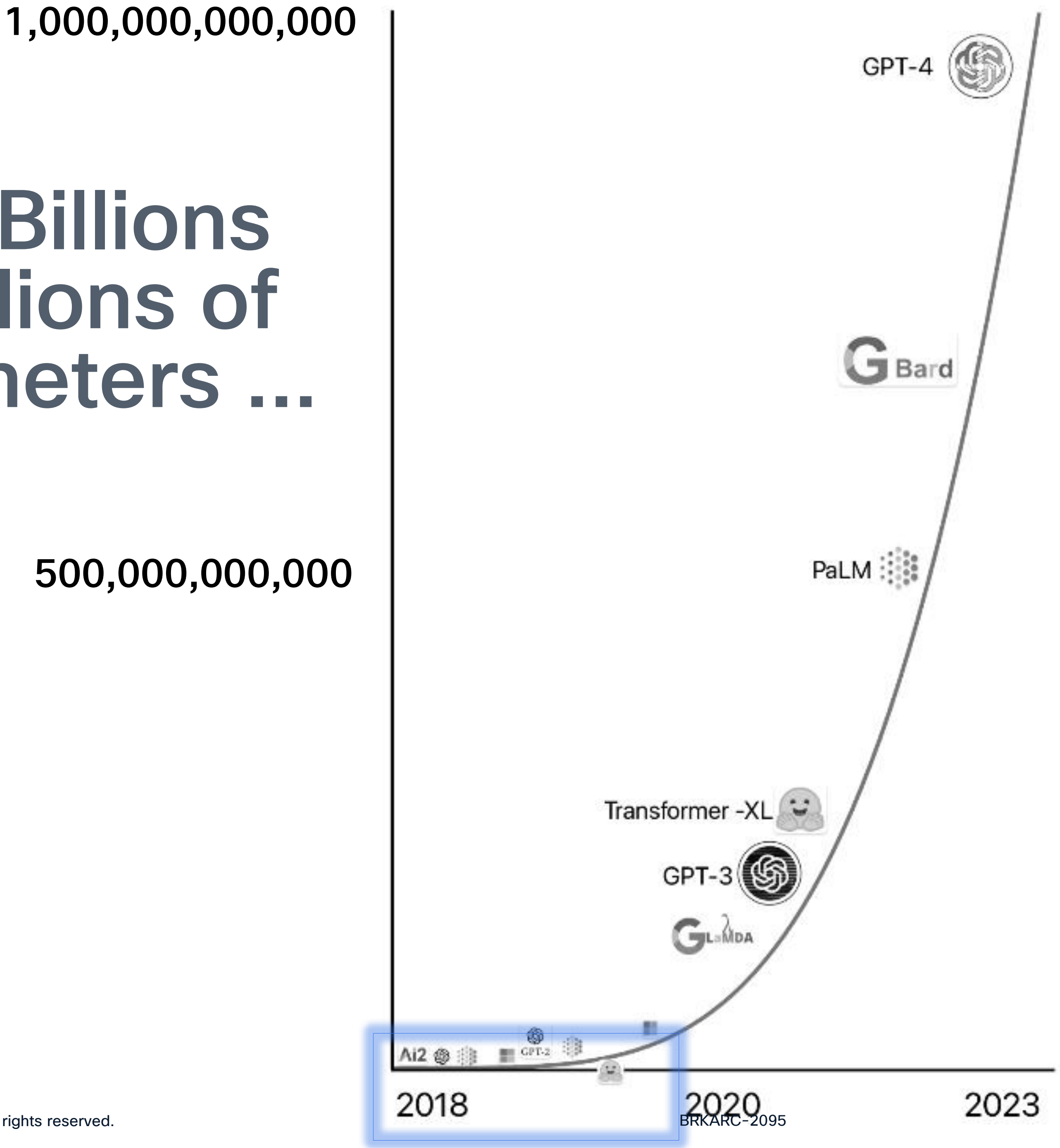
The goal of the attention mechanism is to add **contextual information** to words in a sentence.

“Cambrian Explosion” of Models



Models – Various types, sizes, focus, ...

From Billions to Trillions of Parameters ...



FUN FACT!

The human brain contains **86 billion neurons**, and over **100 trillion synaptic connections**

How are LLMs Trained for Text and Code?

Step 1: Data Collection
(Feeding Knowledge)

Step 2: Tokenization
(Breaking It Down)

Step 3: Parameter Learning
(Storing Knowledge)

Step 4: Fine-Tuning
(Specialized Learning)

Step 1: Data Collection (Feeding Knowledge)

What Happens?

- LLMs are trained on massive amounts of text data – books, articles, websites, and more.

Analogy:

- Giving a child access to a library of books, the more they read, the more they learn.

***Fun Fact**: GPT-4 was trained on terabytes of text, equivalent to hundreds of millions of books.*

Step 2: Tokenization and Vectorization
Breaking it Down

How It Works:

- The text is split into **tokens** (words, subwords, or characters) so the model can process it.
- Tokens are further split into vectors (numerical values)

Analogy:

- Teaching a child to break down sentences into words & letters.

Raw Text

“My name is Dave"

Tokenized Text

[”My", ”name", ”is”, “Dave”]

Vectorized Tokens

```
"My" -> [0.12, -0.43, 0.33, 0.85, -0.17]
"name"-> [0.52, 0.10, -0.21, 0.44, -0.09]
"is"  -> [0.09, -0.15, 0.47, 0.13, 0.56]
"Richard” -> [0.67, -0.25, -0.33, 0.78, 0.45]
```

Step 3: Parameters Learning (Storing Knowledge)

What Happens?

- Vectors flow through neural networks; parameters learn token relationships.

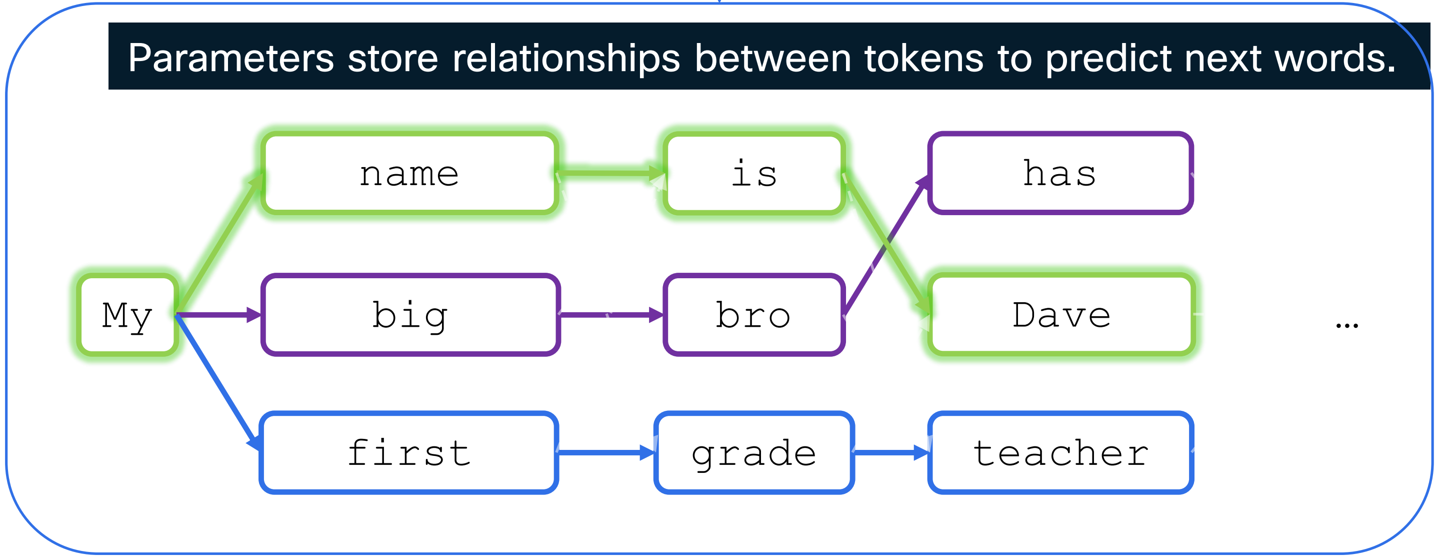
Analogy:

- A child learns how words fit together to form sentences.

Vectorized Text

```
"My" -> [0.12, -0.43, 0.33, 0.85, -0.17]
"name"-> [0.52, 0.10, -0.21, 0.44, -0.09]
"is"  -> [0.09, -0.15, 0.47, 0.13, 0.56]
”Dave” -> [0.67, -0.25, -0.33, 0.78, 0.45]
```

Neural Network



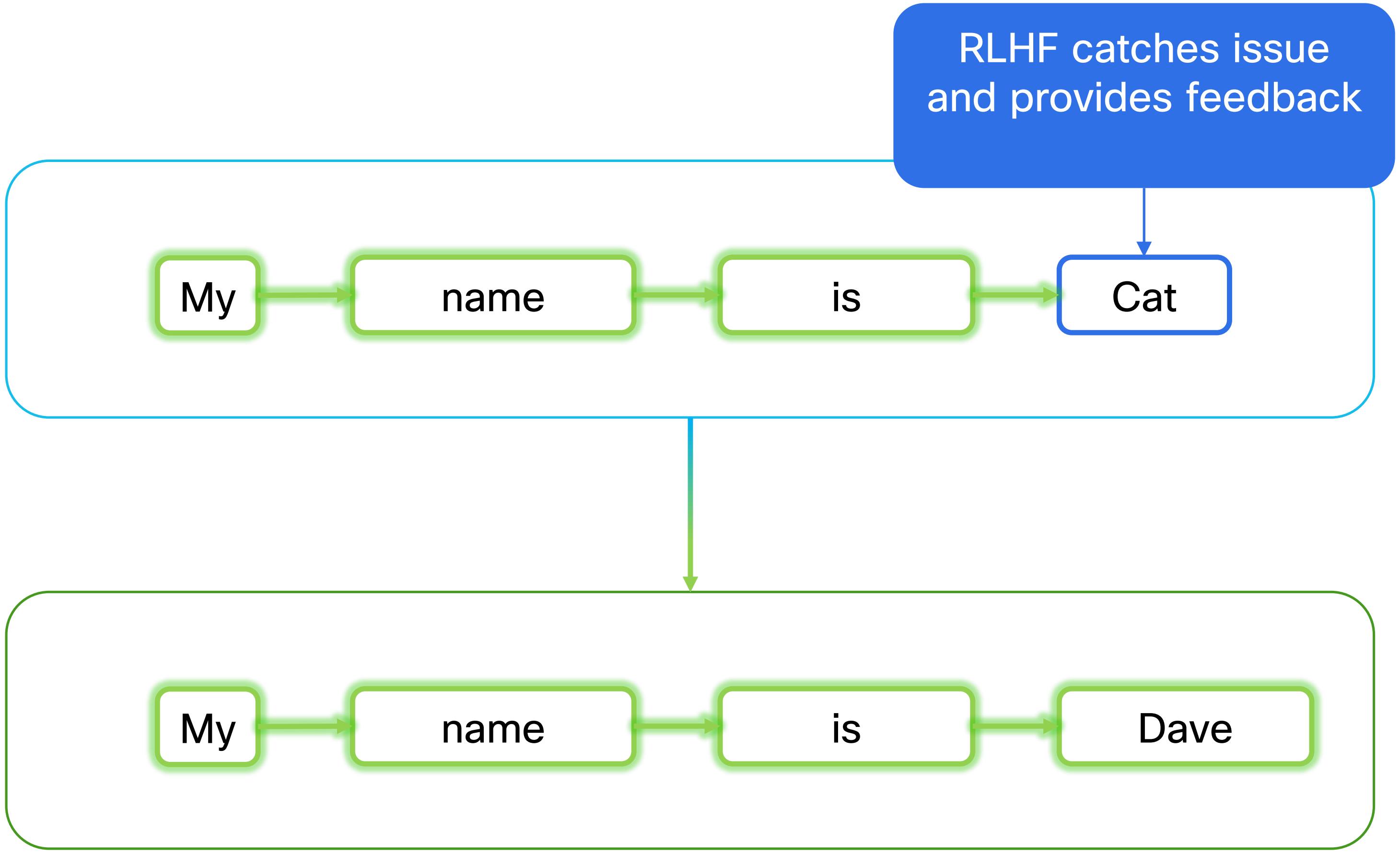
Step 4: Fine-Tuning the Model (Optimizing Predictions)

What happens?

- Parameters are adjusted to minimize prediction errors.
- The model improves by learning from its mistakes

Analogy:

- A child practices speaking by receiving feedback & adjusting.



A Foundational Generative AI Model!

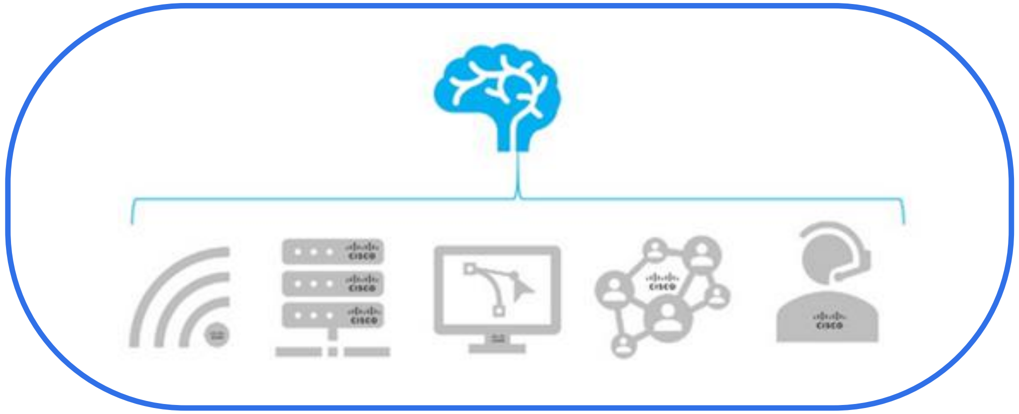
Jack of All Trades Model:

- Pre-trained on vast datasets including text, images, code, etc.
- Can handle a broad array of questions across domains.

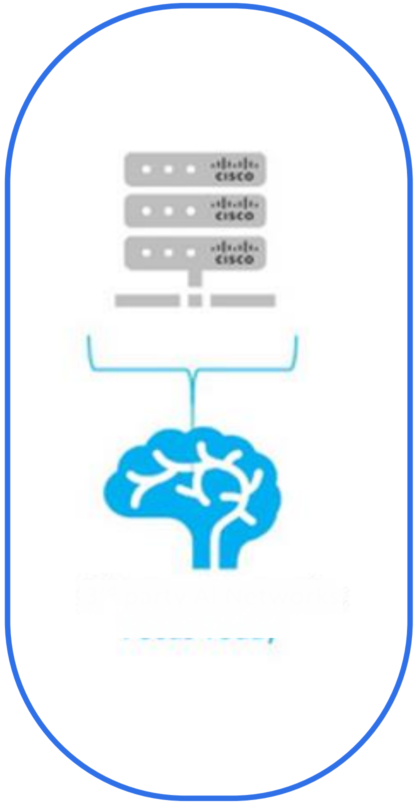
AI in Cisco – Products and Solutions

Artificial Intelligence and Cisco

AI in Cisco –
AI to improve products



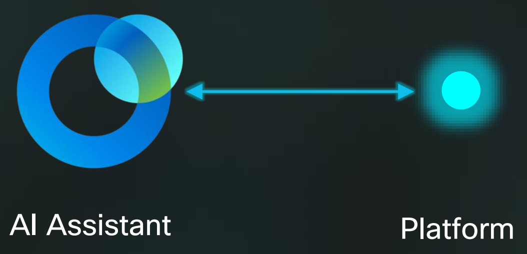
AI on Cisco –
Products to improve AI



AI Assistants Have "Skills", Not Features

AI Skills

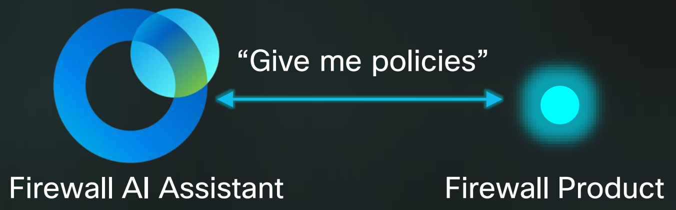
- **Definition: Any action that a Cisco AI Assistant can performance.**
- **Skills:** Troubleshooting, configuration, recommendations, etc.



AI Assistants Native Skills
Enhance Intra-Product Experience

Native Skills

- **Definition: Capabilities of an AI Assistant for the local product it's integrated with**



Documentation Summarization

Answers to questions about a product sourced from its documentation.

Troubleshooting

Insights into issues and guided resolution for accelerated remediation.

Optimization

Recommendations into how a user could better fully utilize their product.

Configuration

Guided workflows helping users to configure what they need to optimally.

Native Skills Across Products Examples

Secure Firewall

1. Connection & Security logs
2. Policy inquiry
3. Policy creation

Splunk Platform

1. SPL generation
2. SPL querying
3. Data summarization

Duo

1. User activity timeline
2. Device info & compliance
3. Authentication logs

Meraki

1. Client troubleshooting
2. Device troubleshooting
3. App troubleshooting

ThousandEyes

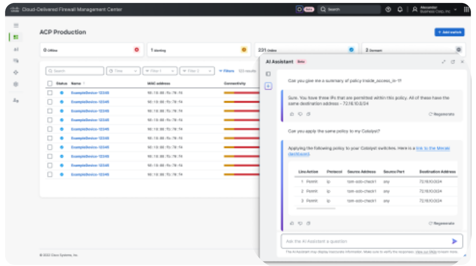
1. Internet outages
2. Network events
3. User to app troubleshooting

CX

1. TAC case management
2. Field notices
3. Vulnerability & PSIRTs

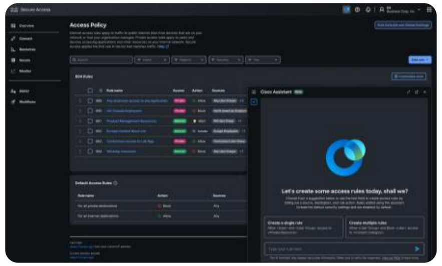
Cisco Security's Suite of AI Assistants

Firewall



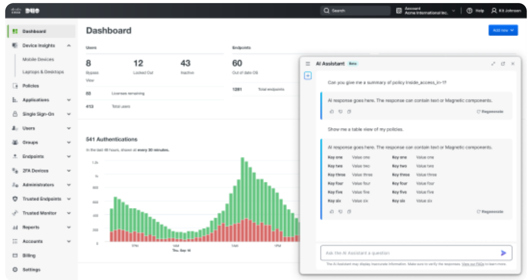
Block any outbound exfiltration to the IP address identified from the C&C

Secure Access



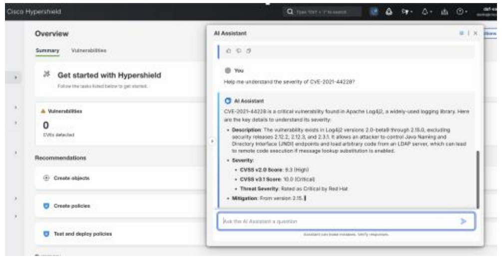
Ensure users access only resources they need securely

Duo



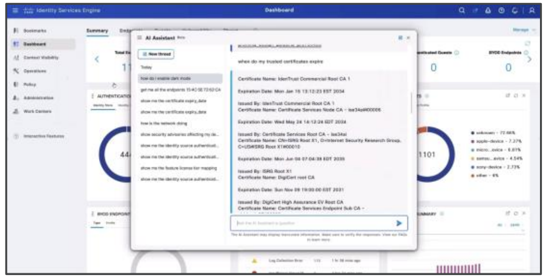
Lock affected user out of critical applications

Hypershield



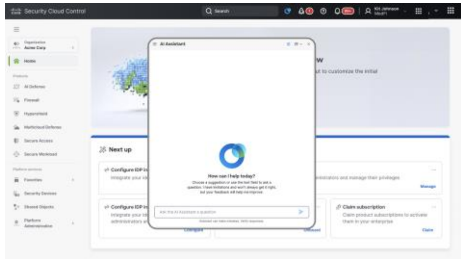
Autonomous segmentation and exploit protection

Identity Service Engine



Enforces identity-based access policies, ensuring secure network access and compliance

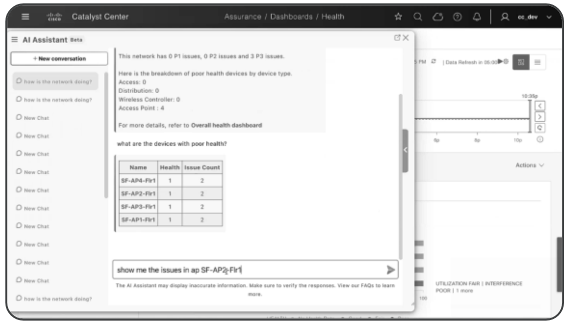
Security Cloud Control



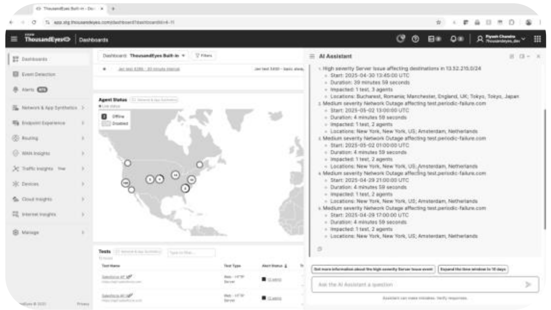
Manage all security products in a single place

Cisco Networking’s Suite of AI Assistants

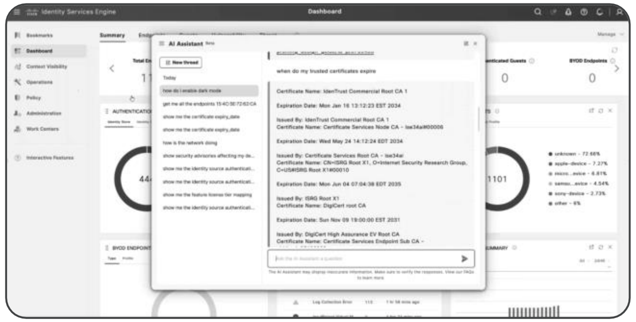
Cloud-managed networking with security, visibility, and device control



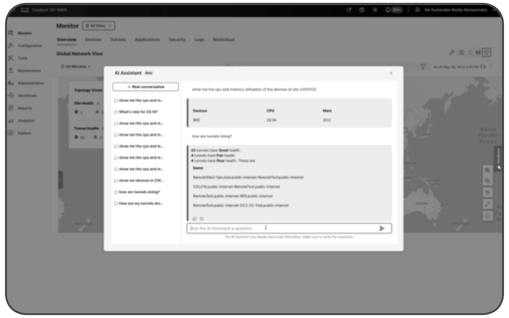
On-prem network management for automation, policy, security & assurance



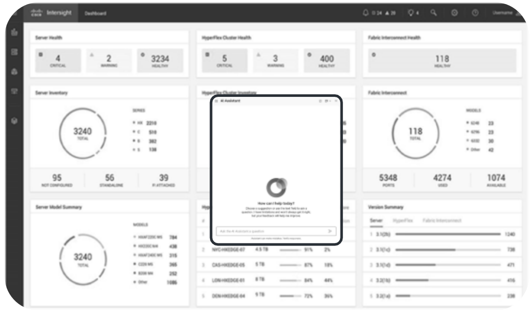
Monitors network and application performance across the internet



identity-based access policies, ensuring secure network access and compliance



Optimizes WAN traffic and security across remote sites



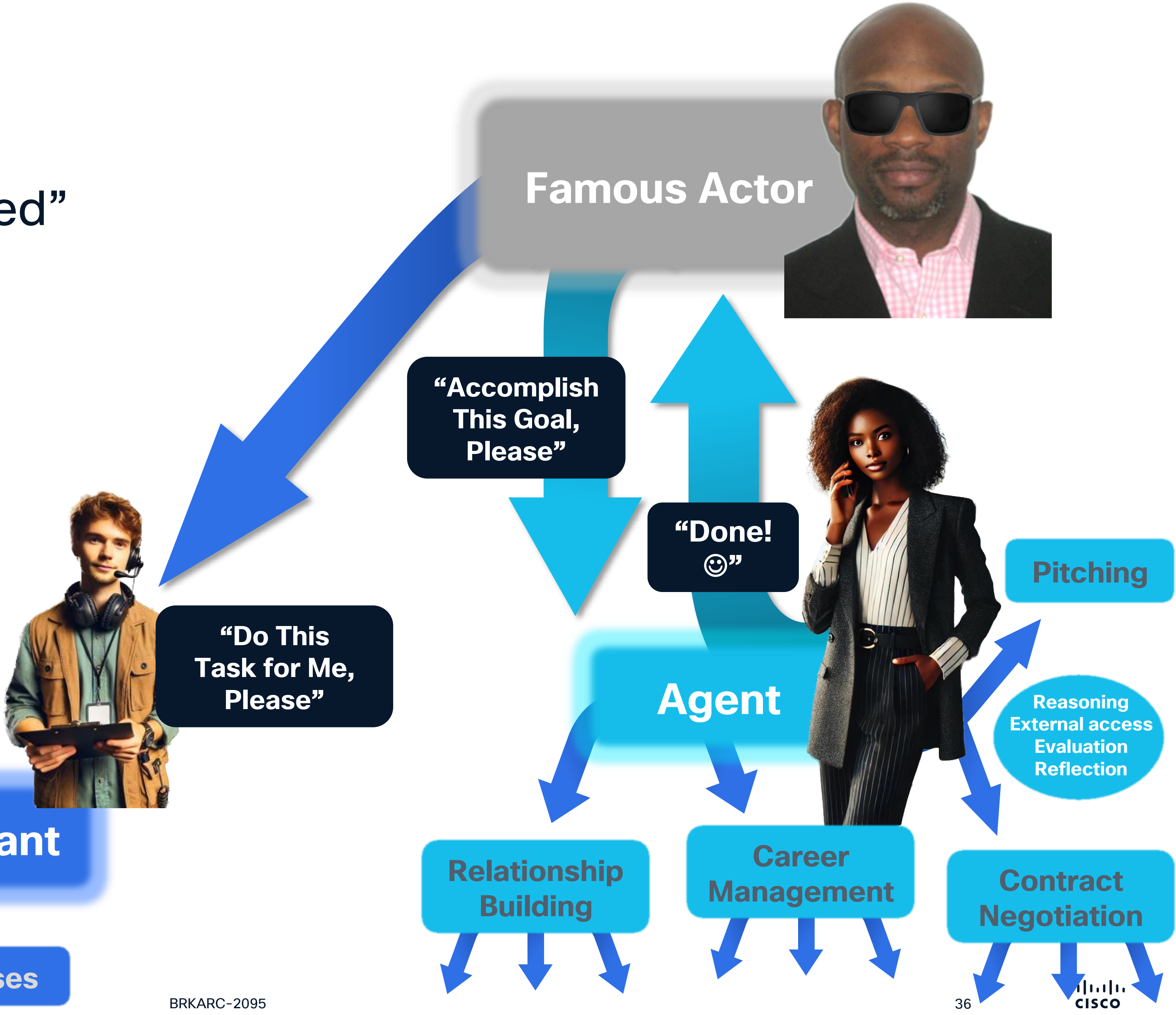
Optimizes WAN traffic and security across remote sites

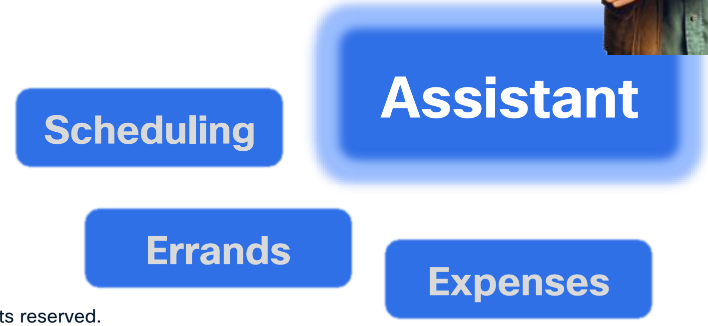
Individual AI Assistants Are Integrated Across Cisco

Security	Firewall, Secure Access, Hypershield, Duo, Identity Intelligence, Splunk Enterprise Security, ISE
Networking	Meraki, Catalyst Center, Catalyst SD-WAN, ThousandEyes, Intersight, Mobility Services
Observability	Splunk Observability (Cloud, ITSI, AppDynamics)
Data	Splunk Platform
Collaboration	Webex Control Hub
Service Ops	Customer Experience

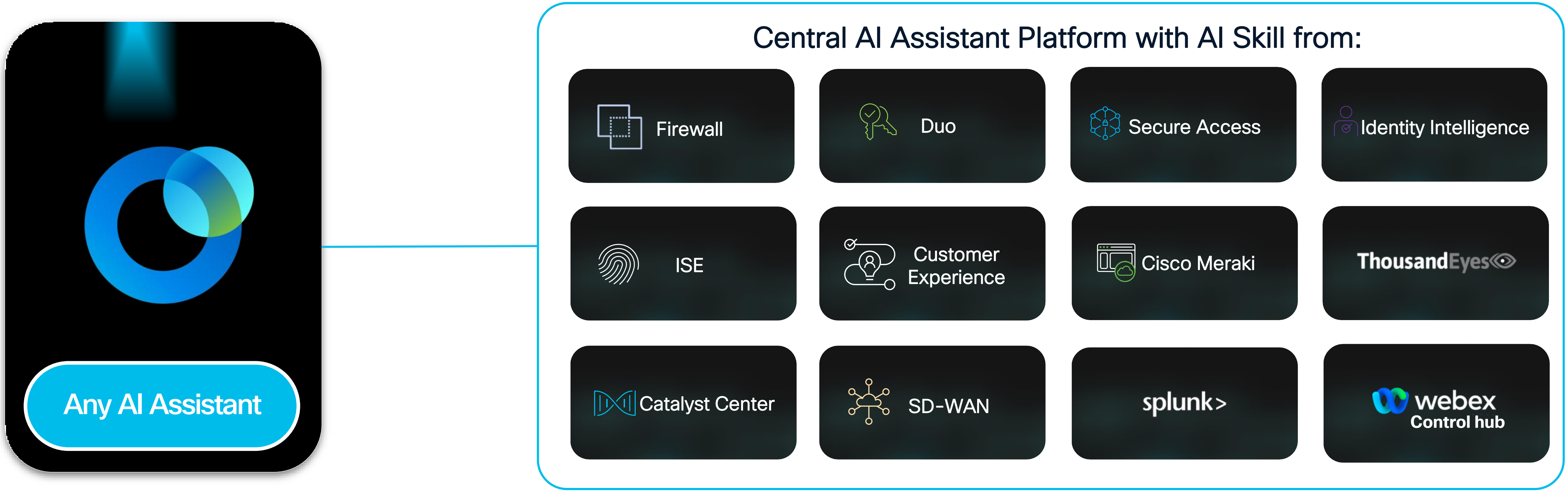
What is an AI Agent?

- An Autonomous system “skilled” to accomplish specific task(s)
- LLM accompanied with:
  - Tools / Functions
  - Memory
- Core capabilities:
  - Planning and Reasoning

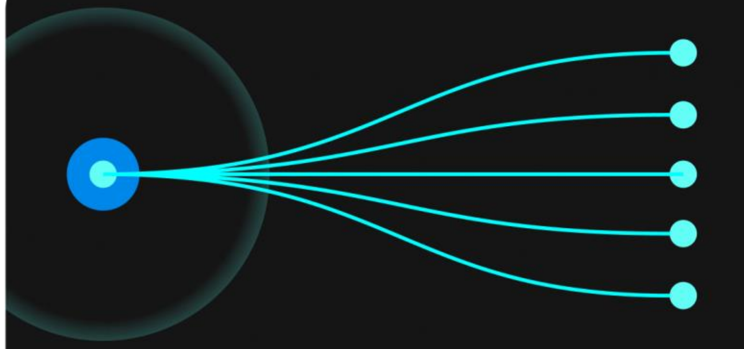




Unify Cisco AI Assistants to enable a network of AI Agents that can use cross-product AI Skills

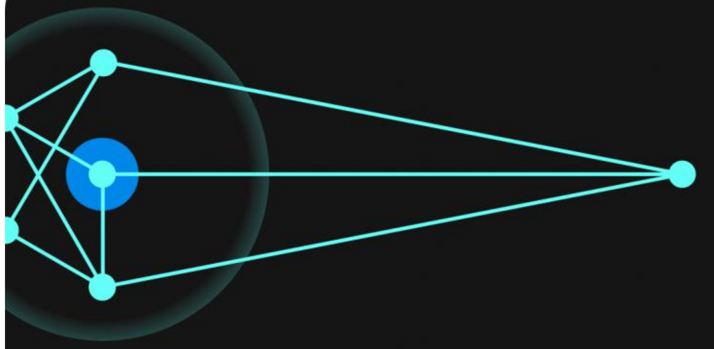


Benefits of the Unifying AI Assistants into a Network of AI Agents



One assistant, many skills

Each Cisco product enhances the Unified AI Assistant with additional “simple” skills to troubleshoot issues.



Compounding value

Combines cross-platform ‘simple’ skills into ‘Composite’ skills–more Cisco products mean exponentially richer context and smarter recommendations.

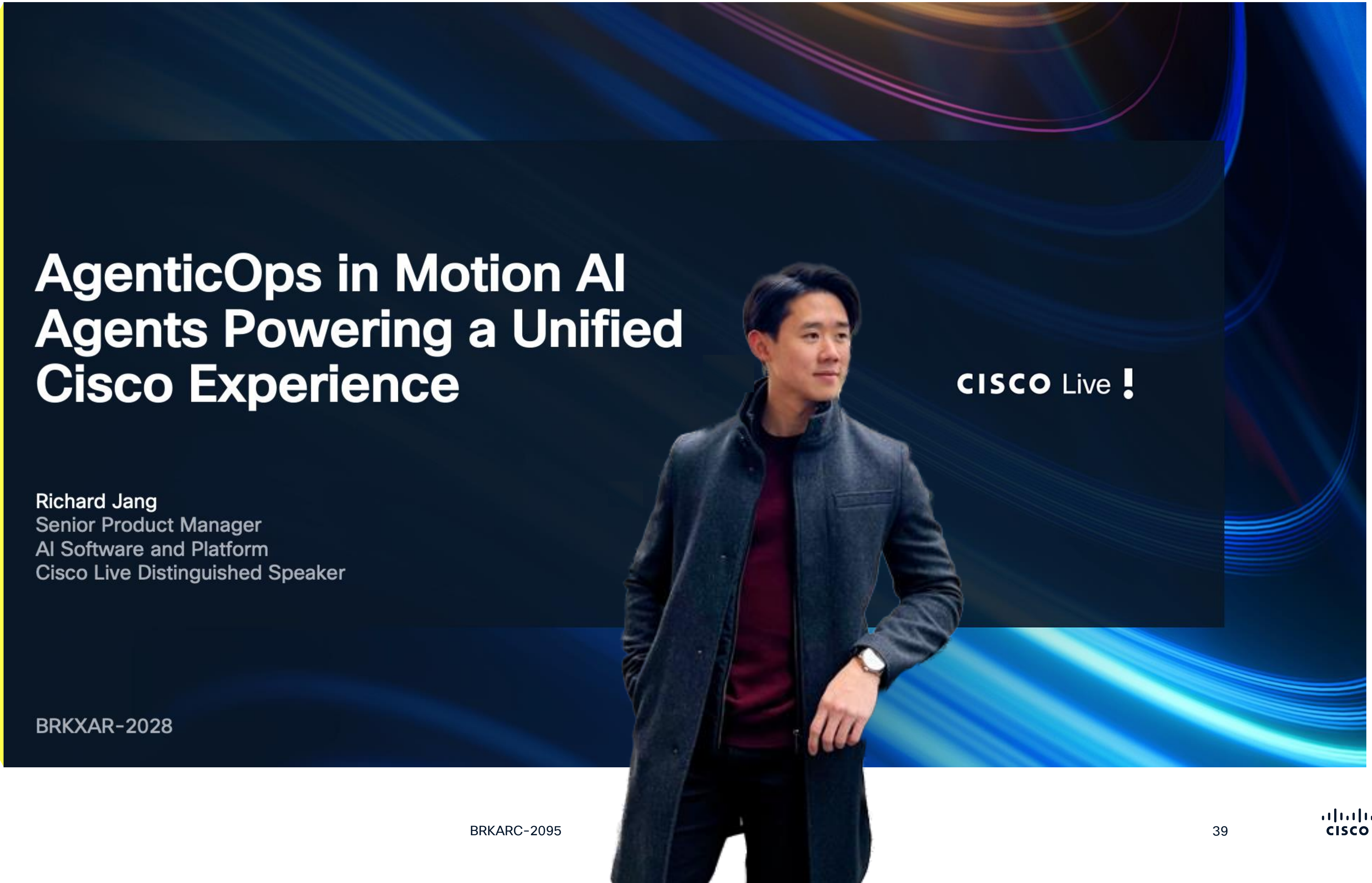
Accelerated resolution

Troubleshooting is consuming, but the AI Assistant enables RCA in minutes by correlating cross-domain insights!

More about Cisco Assistants, GenAI, ...

Tuesday,
June 10th

2:00 –
3:30pm



Why Networking is Relevant to AI Deployments

LLMs are orders of magnitude more intensive than DLRM

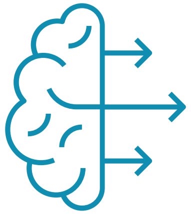
Deep Learning Recommendation Models

Search, Feed ranking. Ads & content recommendation

Inference needs a few Gigaflops for 100ms TTFT

Narrower scope, domain specific

Training: ~100 Gigaflop/ sentence



Large Language Models

Intricacies of human language

Inference needs 10s of Petaflops for 1 sec TTFT

Generate intelligent, creative responses

Training : ~1 Petaflop/ sentence

An Improved user experience means *a* ***faster time to first token,*** making ***distributed inference an imperative***

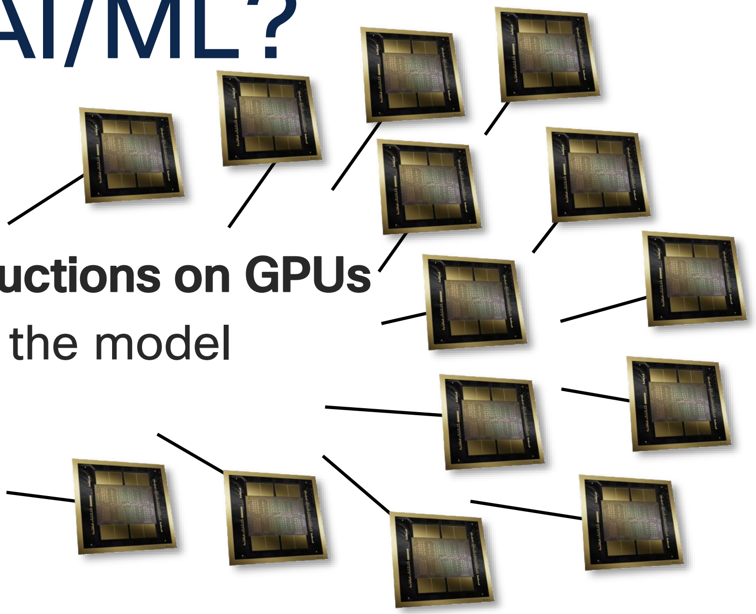
AI on Cisco - Building Networks for AI/ML

GenAI is upending the global IT spend.

- The Hyperscalers spent ***~$180B*** in infrastructure alone in 2024[1].
- AI accelerator silicon revenue grew ***130%*** in 3Q 2024[2].
- DC switching and NIC markets will double to ***>$50B*** in 5 years[3].
- A ChatGPT query takes ***~10x*** the power of a Google search[4].
- Nuclear power is becoming a ***critical*** DC energy source[5].
- Goldman Sachs forecasts global DC power demand may increase ~***165%*** by 2030[6].

1) CIO Dive: Big tech on track to pour more than $180B into data centers this year
2) Dell'Oro: US Hyperscalers Set to Deploy Over 5 Million AI Training-Capable Accelerators in 2024
3) Crehan Research: Ethernet switch and NIC market to reach $50 Billion in the next five years
4) Kanoppi: Search Engines vs AI: energy consumption compared
5) Power: The SMR Gamble: Betting on Nuclear to Fuel the Data Center Boom
6) Goldman Sachs: AI to drive 165% increase in data center power demand by 2030

Why does the network matter for AI/ML?



Execute instructions on GPUs

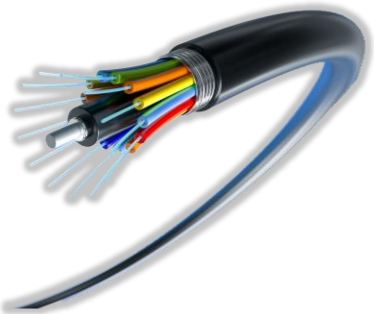
Training the model

Process

Notify

Share results

Everyone sends to everyone



Synchronise

Wait for everyone to complete

Job Completion Time (JCT) is based on the ***worst-case tail latency***

Barrier operation

How do I get the most out of $Bs of GPUs and Faculties

Giorgio Trovato Unsplash

GDJ Open Clipart

The network exists to enable the GPUs do ***their work***

A ***minute*** occupied by the network is a ***minute*** the GPUs are idle

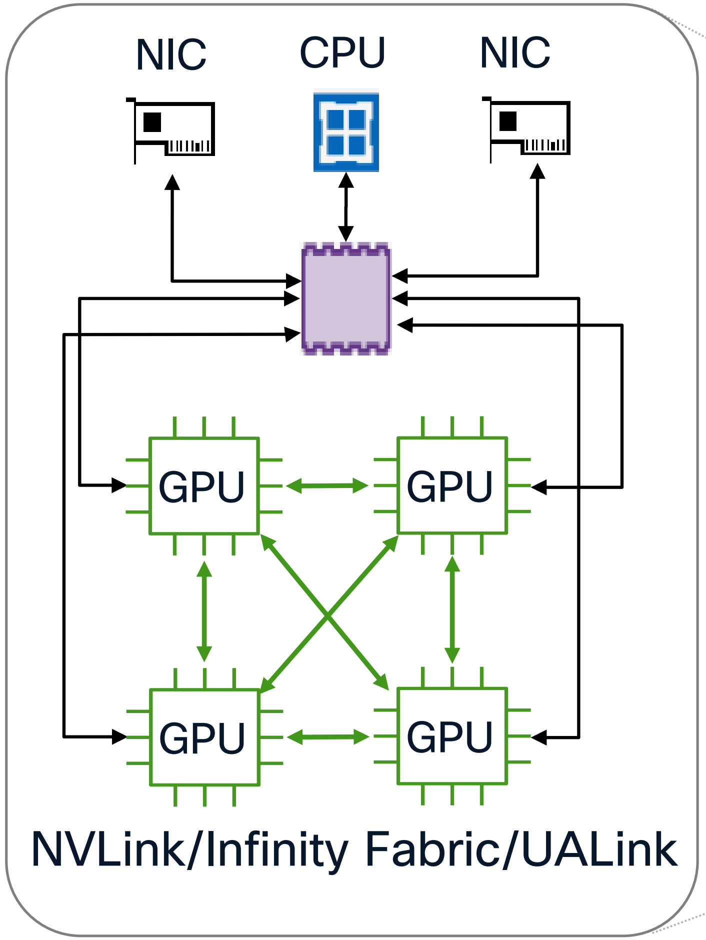
A ***watt*** spent on the network is a ***watt*** not spent on the GPUs

What matters?

- Throughput under full load
- Reliability/Resilience
- Power

Networking for GenAI

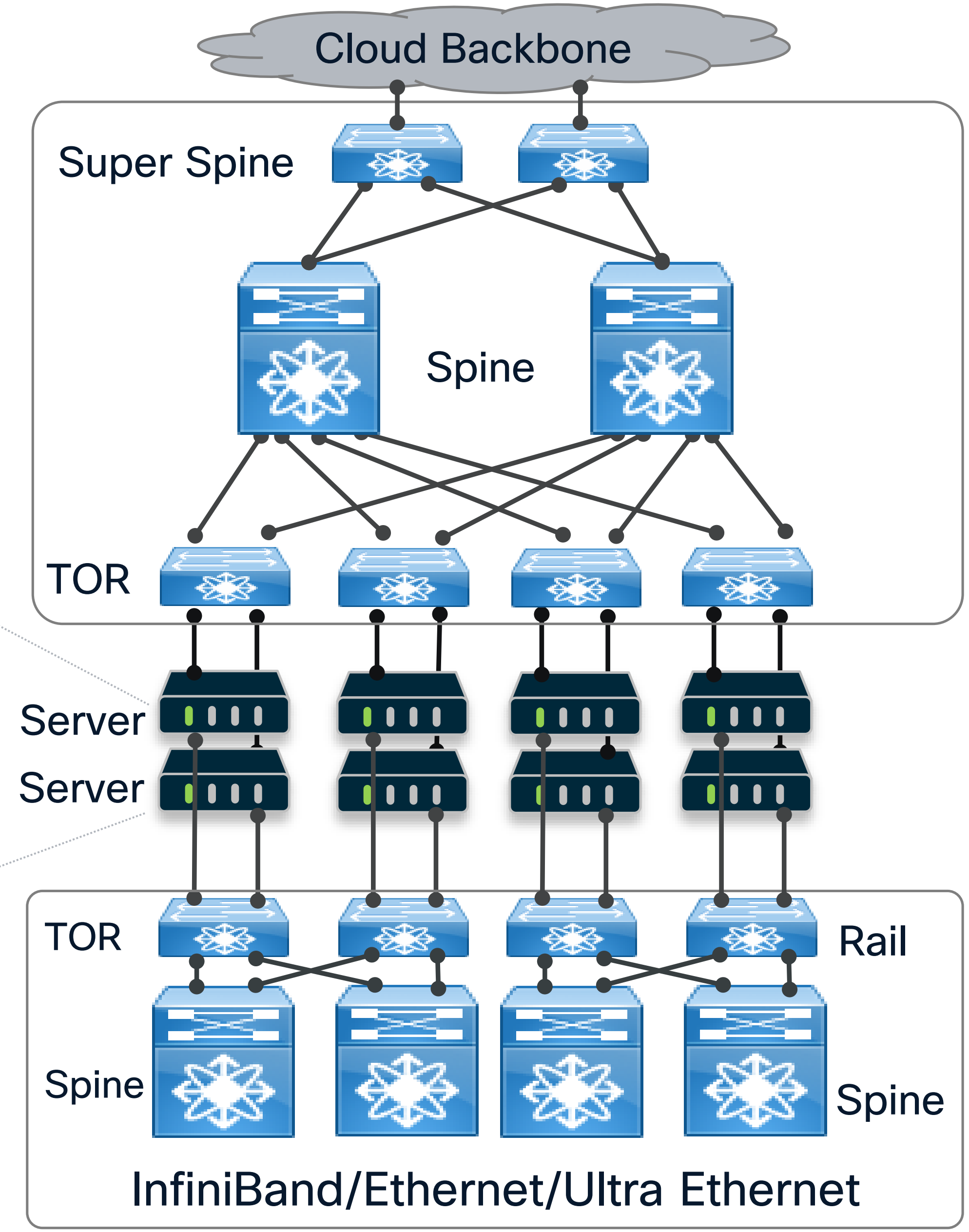
AI Network Fundamentals



Back-end
Scale-up
Network

Front-end
Network

Back-end
Scale-out
Network



Ethernet vs InfiniBand

- Google search "Ethernet InfiniBand benchmark" - AI Overview:

  "In benchmarks, InfiniBand generally outperforms Ethernet in terms of latency and bandwidth, especially in HPC and AI environments. However, Ethernet is rapidly closing the gap, with newer standards like UltraEthernet offering substantial performance improvements. In some cases, especially with optimized Ethernet and larger, more complex workloads, ***Ethernet can even outperform InfiniBand****."*

- WWT: The Battle of AI Networking: Ethernet vs InfiniBand[1]

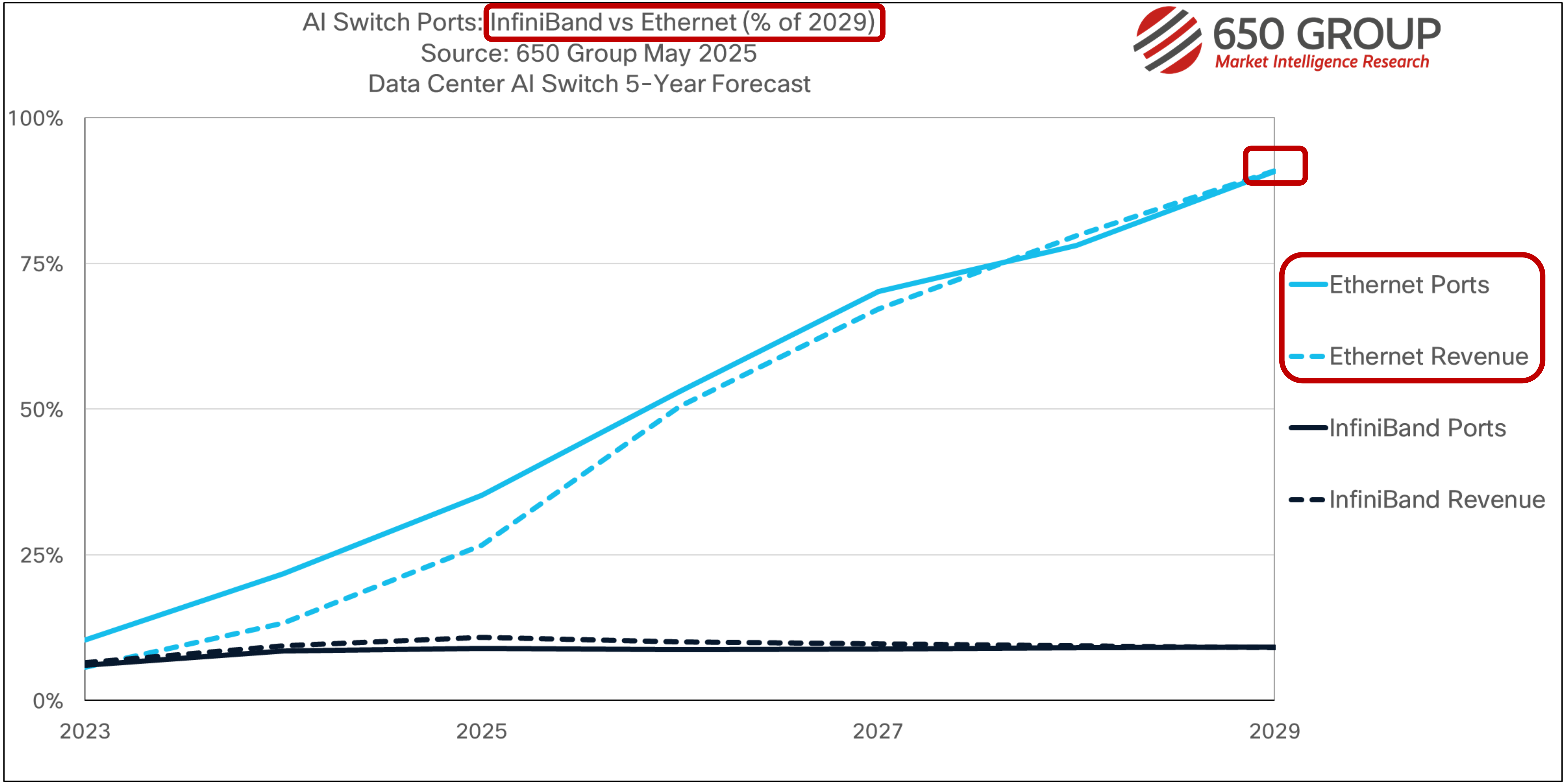
  Q: "is Ethernet ***good enough****?"*

  A: "Across generative tests and OEMs, the performance delta between InfiniBand and Ethernet was ***statistically insignificant*** *(< 0.03%)"*

  "WWT views Ethernet as a ***wholly viable alternative*** *to InfiniBand for most generative and inference use cases"*

1: https://www.wwt.com/blog/the-battle-of-ai-networking-ethernet-vs-infiniband

Ethernet vs InfiniBand – AI Backend Network Switch Ports



What's driving Ethernet?

Scale

- Hyperscalers are looking to build very large training clusters (300,000+) [1], have clusters span multiple DCs[1], and InfiniBand has scaling limitations.

Supplier Diversity

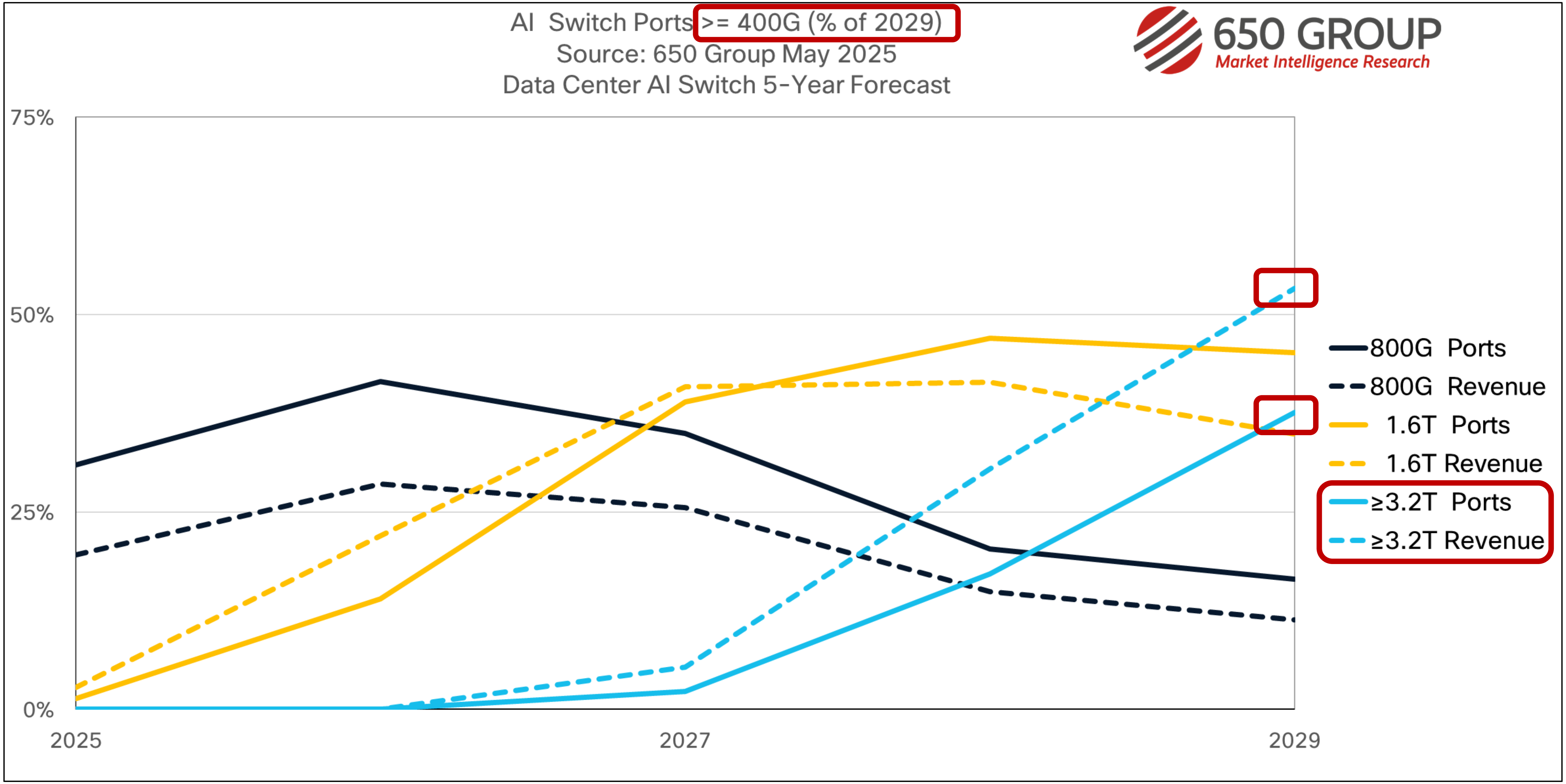
- Nvidia(Mellanox) dominates the InfiniBand market[2].

Cost of Operations

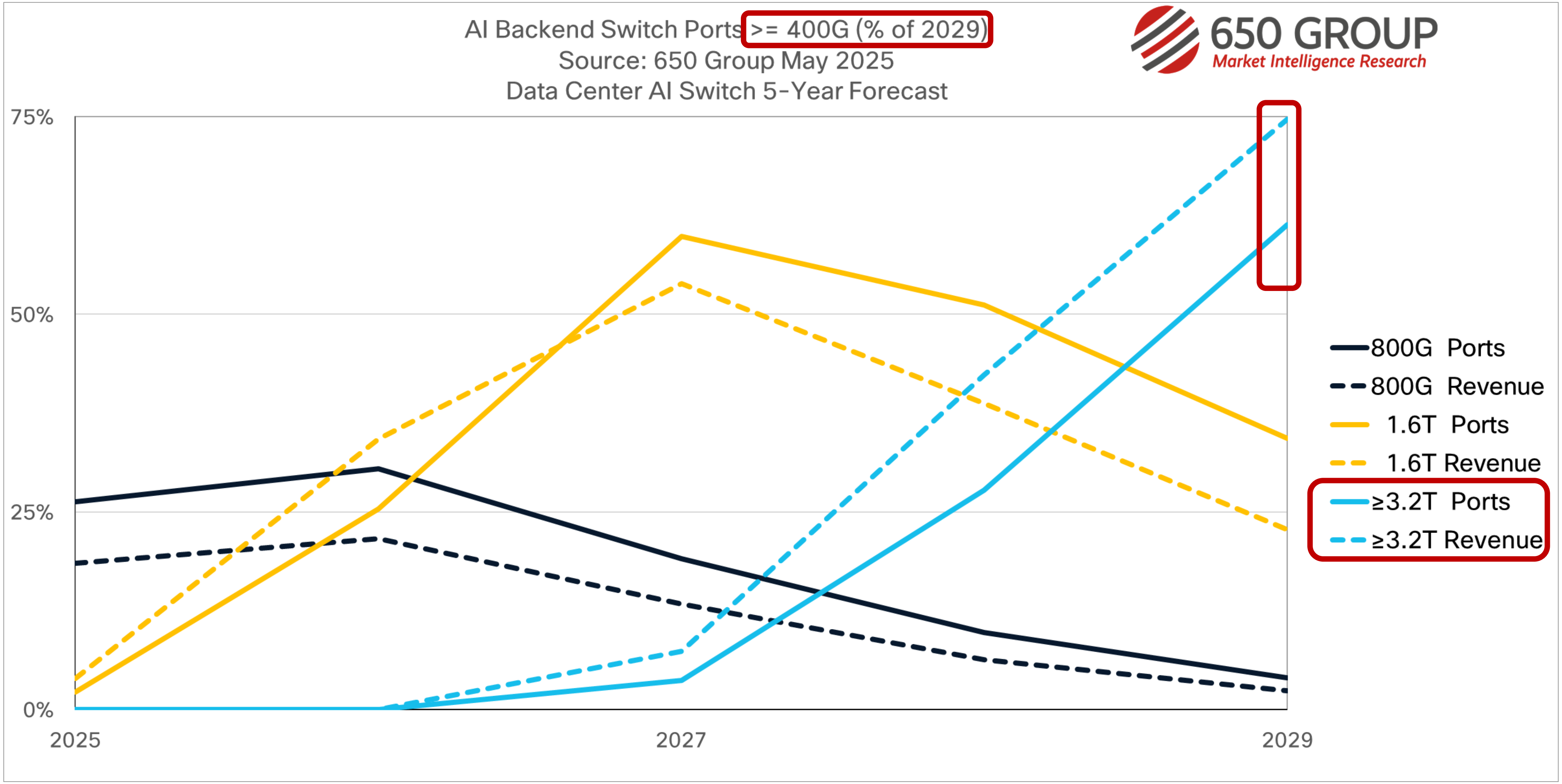
- History shows that Ethernet becomes less expensive to own and operate than the technologies it replaces.
- Everyone has Ethernet, using one technology reduces operational cost.

1. SemiAnalysis: Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure
2. NADDOD: Where to Buy Infiniband products

Ethernet Speed Trends – AI Network Switch Ports



Ethernet Speed Trends – AI Backend Network Switch Ports



RFC 1925: The Twelve Networking Truths

```
Network Working Group                                    R. Callon, Editor
Request for Comments: 1925                                            IOOF
Category: Informational                                       1 April 1996
```

Abstract

This memo documents the fundamental truths of networking for the Internet community. This memo does not specify a standard, except in the sense that all standards must implicitly follow the fundamental truths.

Acknowledgements

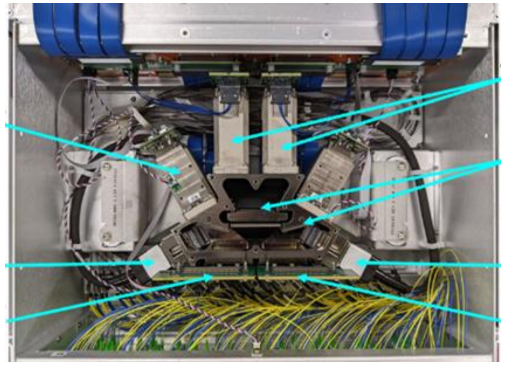
The truths described in this memo result from extensive study over an extended period of time by many people, some of whom did not intend to contribute to this work. The editor merely has collected these truths, and would like to thank the networking community for originally illuminating these truths.

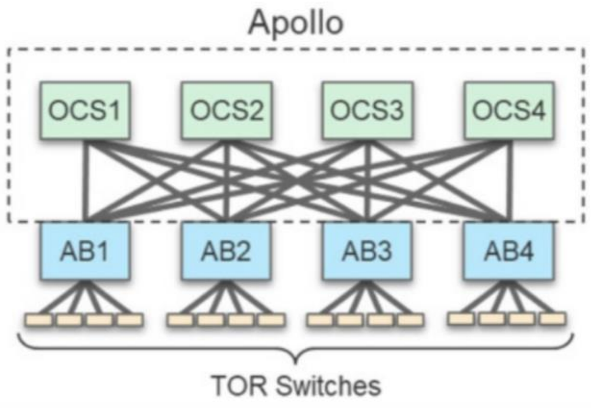
1. Introduction

This Request for Comments (RFC) provides information about the fundamental truths underlying all networking. These truths apply to networking in general, and are not limited to TCP/IP, the Internet, or any other subset of the networking community.

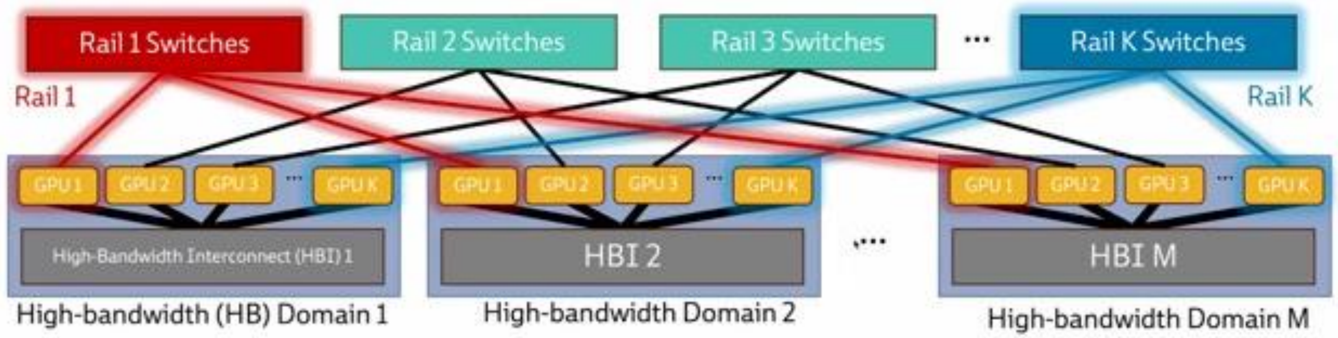
RFC 1925 rule 10 - “One size never fits all”.

Google uses Custom Optical Switches[1] in its Jupiter network architecture[2].





Meta has a “Rail-only” design. [2]



1. SemiAnalysis: Google OCS Apollo: The >$3 Billion Game-Changer in Datacenter Networking
2. Google: Speed, scale and reliability: 25 years of Google data-center networking evolution
3. NextPlatform: This AI Network Has No Spine – And That’s A Good Thing

Ethernet for AI Networks: Who's doing What

Ethernet Alliance

- Building cross industry consensus, e.g., TEF 2024: Ethernet in the Age of AI

IEEE 802.3

- IEEE P802.3dj is writing the 200G/lane standard
- NEA investigating 400G/lane and AI bandwidth needs

Optical Internetworking Forum(OIF)

- Exploring technology problems/solutions, e.g., 448Gbps Signaling for AI Workshop

Storage Networking Industry Alliance(SNIA)/Small Form Factor Committee (SFF)

- Exploring technology problems/solutions, e.g., 400G AI Workshop

Ultra Ethernet Consortium(UEC)

- Open standard for scale-out Ethernet networks
- UE 1.0 specification expected soon

Adjacent: Ultra Accelerator Link™ (UAL)

- Open standard for scale-up Accelerator-to-Accelerator communication
- UALink 1.0 defines 200G/lane for 1,024 accelerators within an AI pod

Ethernet for AI Networks: Who's doing What

Ethernet Alliance

- Building cross industry consensus, e.g., TEF 2024: Ethernet in the Age of AI

IEEE 802.3

- IEEE P802.3dj is writing the 200G/lane standard
- NEA investig

Optical Inter

- Exploring te

Storage Net

- Exploring te

Ultra Ethern

Lots of Activity!

- Open standa
- UE 1.0 specification expected soon

Adjacent: Ultra Accelerator Link™ (UAL)

- Open standard for scale-up Accelerator-to-Accelerator communication
- UALink 1.0 defines 200G/lane for 1,024 accelerators within an AI pod

Ethernet for AI Networks: Who’s doing What

Ethernet Alliance

Building cross industry consensus, e.g., TEF 2024: Ethernet in the Age of AI

IEEE 802.3

IEEE P802.3dj is writing the 200G/lane standard

NEA investig

Optical Inter

Exploring te

Storage Net

Exploring te

Ultra Etherne

Open standa

UE 1.0 specification expected soon

Adjacent: Ultra Accelerator Link™ (UAL)

Open standard for scale-up Accelerator-to-Accelerator communication

UALink 1.0 defines 200G/lane for 1,024 accelerators within an AI pod

RFC 1925 rule 12:
“In ~~protocol~~ network design, perfection has been reached not when there is nothing left to add, but when there is nothing left to take away”.

Ultra Ethernet Consortia (UEC)

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

THE NEW ERA NEEDS A NEW NETWORK

Ultra Ethernet

As ***performant*** as a supercomputing interconnect

As ***ubiquitous*** and ***cost-effective*** as Ethernet

As ***scalable*** as a cloud data center

AI Ethernet Fabric Options

	Ethernet	Enhanced Ethernet		Ultra Ethernet	Scheduled Ethernet
Load Balance	Stateless ECMP	Stateful Flow/ Flowlet	Spray & Re-order in SmartNIC	Endpoint Controlled adaptive packet spraying	Spray & Re-order in leaf
Congestion Management	Congestion Reaction with ECN/PFC	Adjust distribution based on congestion		Congestion Management	Congestion Avoidance
Link Failure	Software	Hardware		Hardware	Hardware
JCT	Good	Better		Even Better	*Best*
NIC and Fabric Coupled	No	No	Yes	Yes	No
Place in Network	Frontend, Backend	Frontend, Backend		Backend	Frontend, Backend

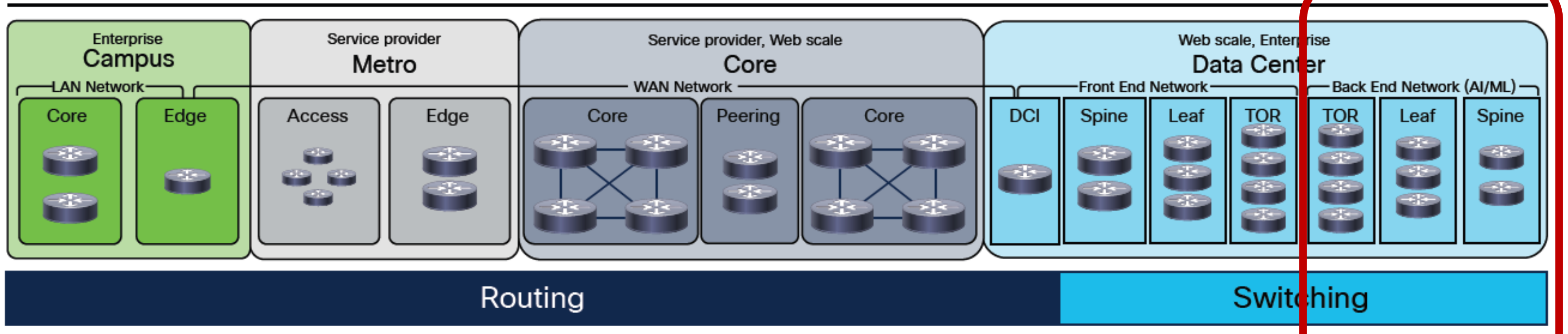
Performance *DEPENDENT* on Traffic Characteristics

Performance *NOT DEPENDENT* on Traffic Characteristics

AI on Cisco –
Silicon One in AI Networks

Cisco Silicon One - Convergence without compromise

Cisco Silicon One



	Campus Core	Campus Edge	Metro Access	Metro Edge	WAN Core	Peering	WAN Core	DCI	Front End Spine	Front End Leaf	Front End TOR	Back End TOR	Back End Leaf	Back End Spine
High efficiency switching – Silicon One G-Series									☑	☑	☑	☑	☑	☑
Complex feature switching – Silicon One E-Series											☑			
High bandwidth Core and lean edge routing – Silicon One P-Series				☑	☑	☑	☑					☑	☑	☑
Super-set Routing and switching – Silicon One Q-Series	☑	☑			☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Complex feature Edge routing – Silicon One K-Series				☑										
Complex feature Access routing and switching – Silicon One A-Series			☑											

One Architecture, One SDK, One Experience

Silicon One in AI

Q200L - 12.8T
32x400GE

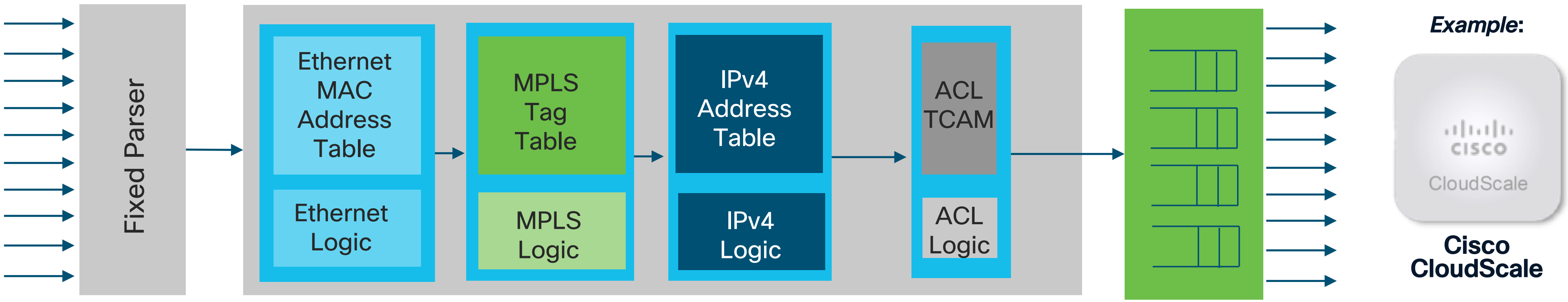
P100 – 19.2T
24x800GE

G100 – 25.6T
24x800GE

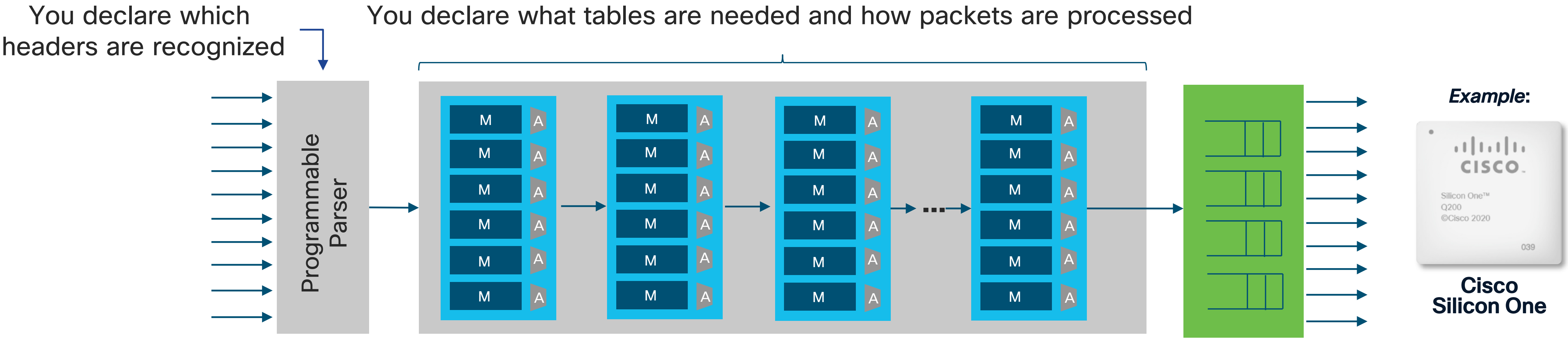
G202 – 25.6T
64x400GE

G200 - 51.2T
64x800GE

Fixed vs. Programmable Packet Processing



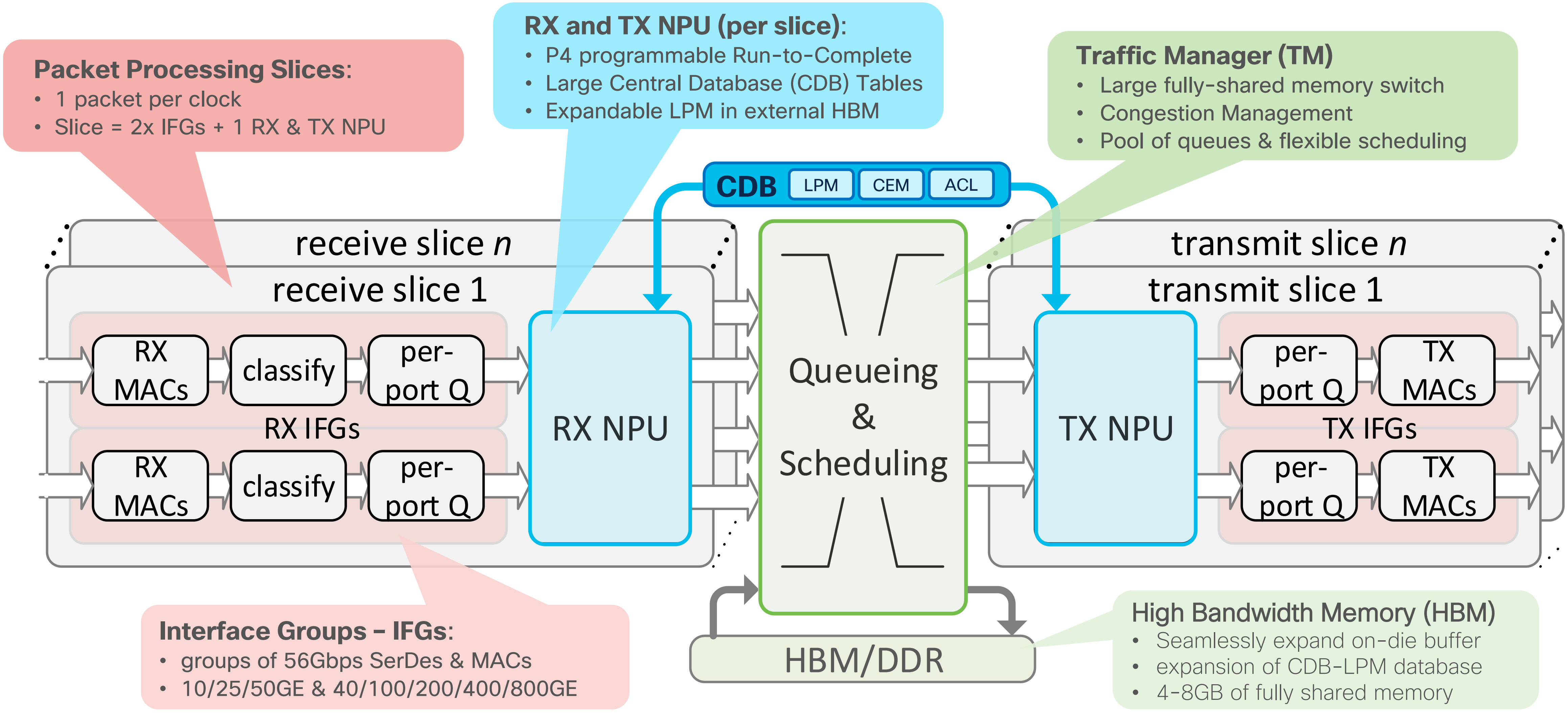
Fixed Pipeline: features and functionality are baked-in at design time



Programmable Pipeline: all stages identical, customer-defined match-action logic

Silicon One

Top Level



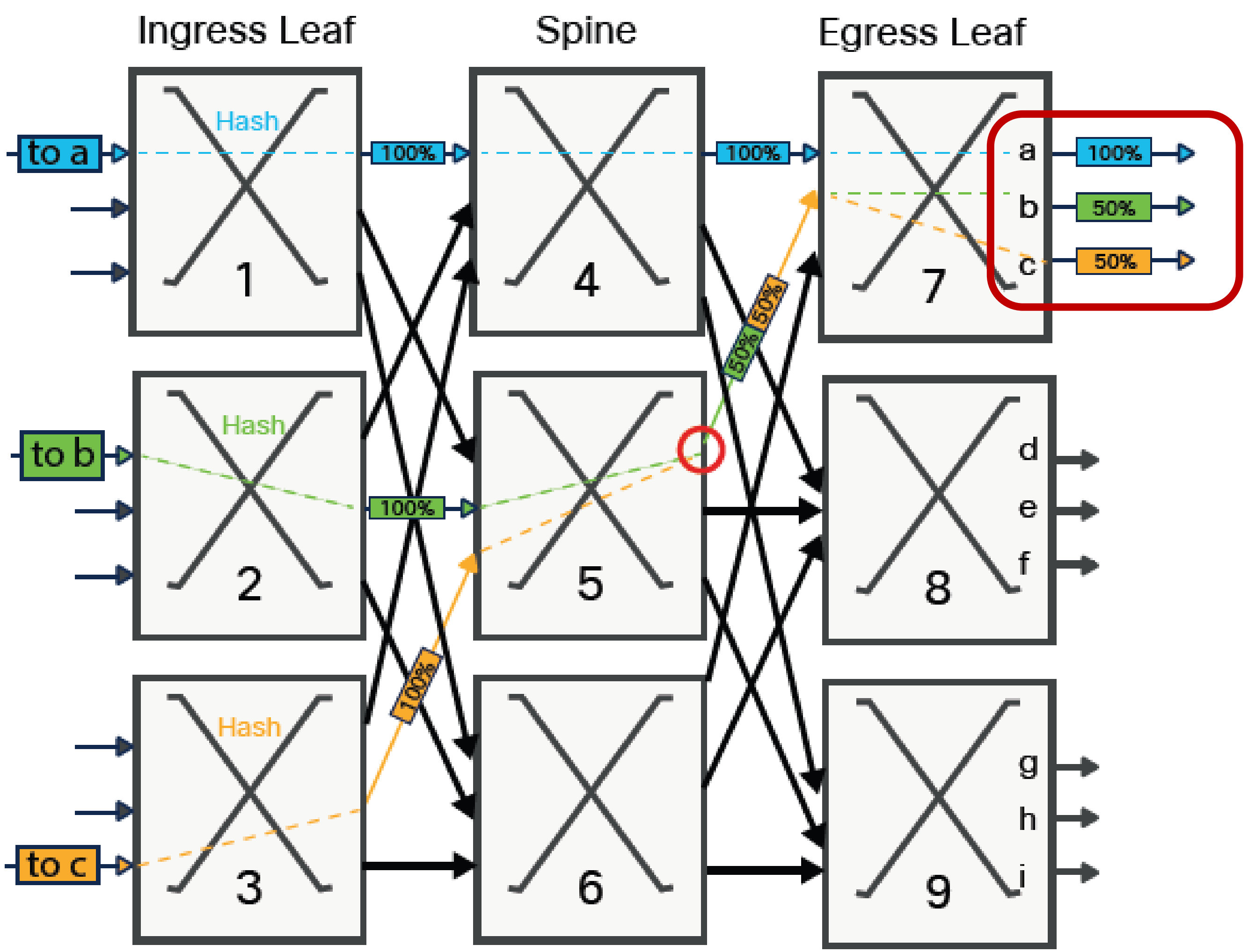
ECMP and Congestion

All-to-All flows
- smaller number of bigger flows
- low header entropy

ECMP
- unaware of network load/congestion
- needs entropy in packer headers
- assumes most flows are short lived

Result
- traffic/network inefficiency as flows “collide” in the network

Congested Links



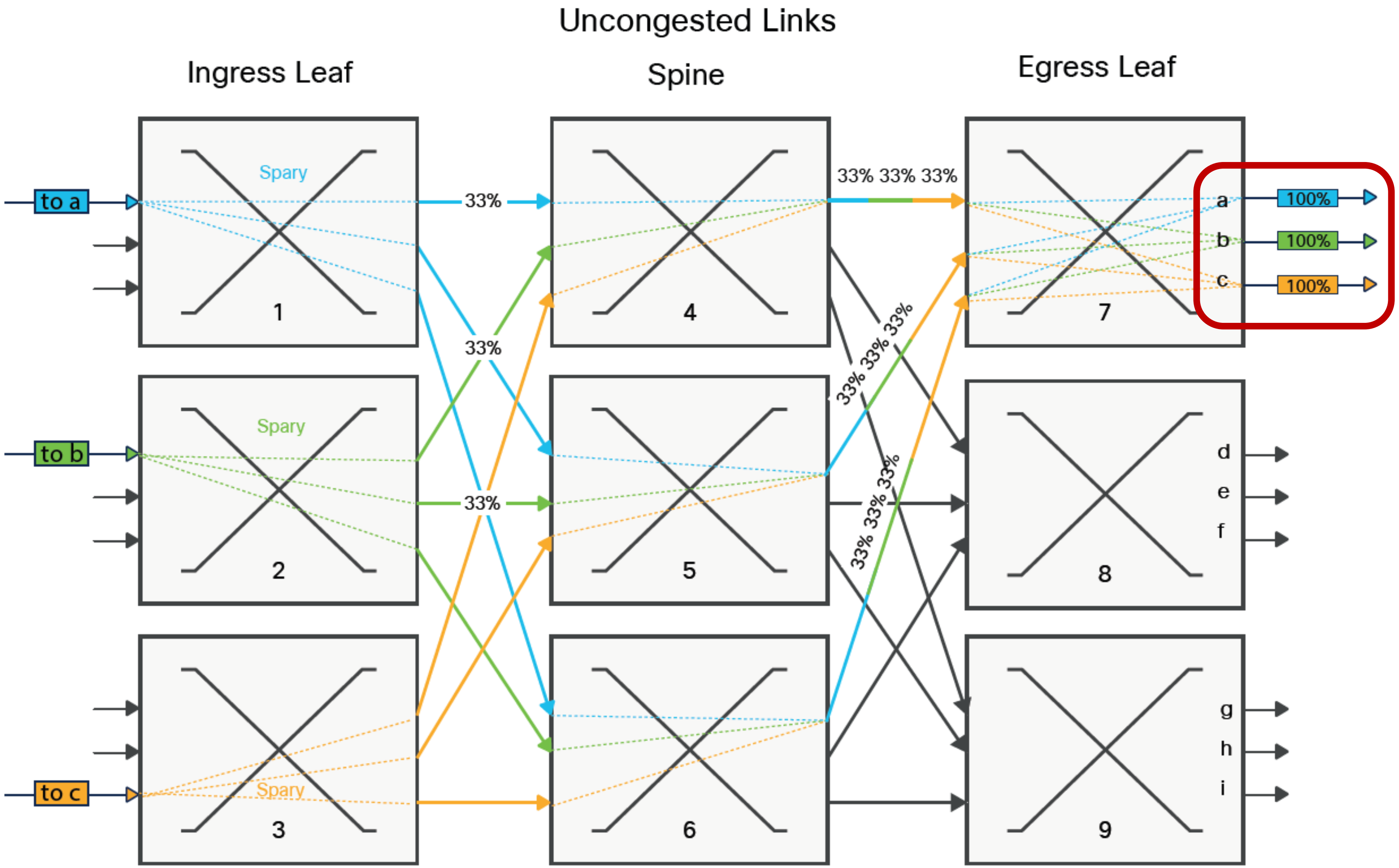
https://www.cisco.com/c/en/us/solutions/collateral/silicon-one/evolve-ai-ml-network-silicon-one.html

Fully Scheduled Network

All-to-All flows
- smaller number of bigger flows
- low header entropy

Distributed Switch Model
- VoQs in ingress leaf
- active congestion control
- re-ordering at egress

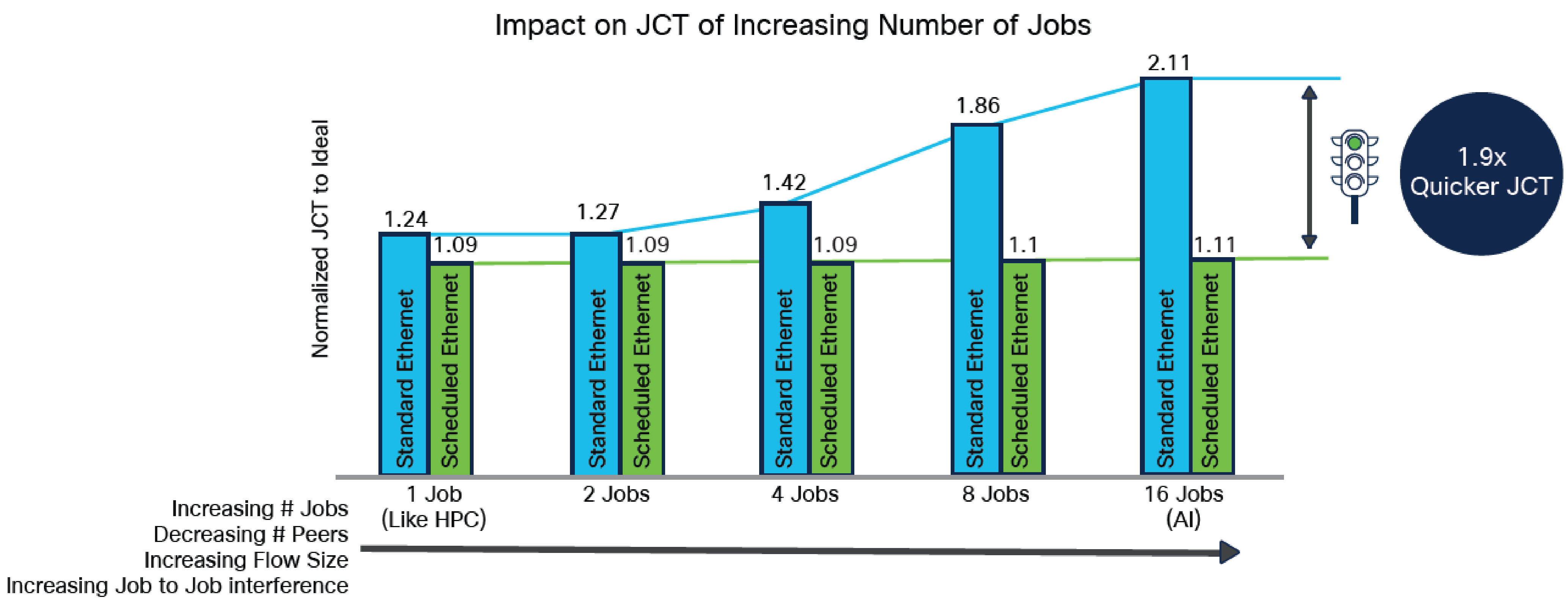
Result
- optimal network performance



https://www.cisco.com/c/en/us/solutions/collateral/silicon-one/evolve-ai-ml-network-silicon-one.html

Building Networks for ML/AI Workloads

Optimized Job Completion Time(JCT) with Fully Scheduled Fabric



Scheduled Ethernet provides exceptional performance, providing lower job completion time

https://www.cisco.com/c/en/us/solutions/collateral/silicon-one/evolve-ai-ml-network-silicon-one.html

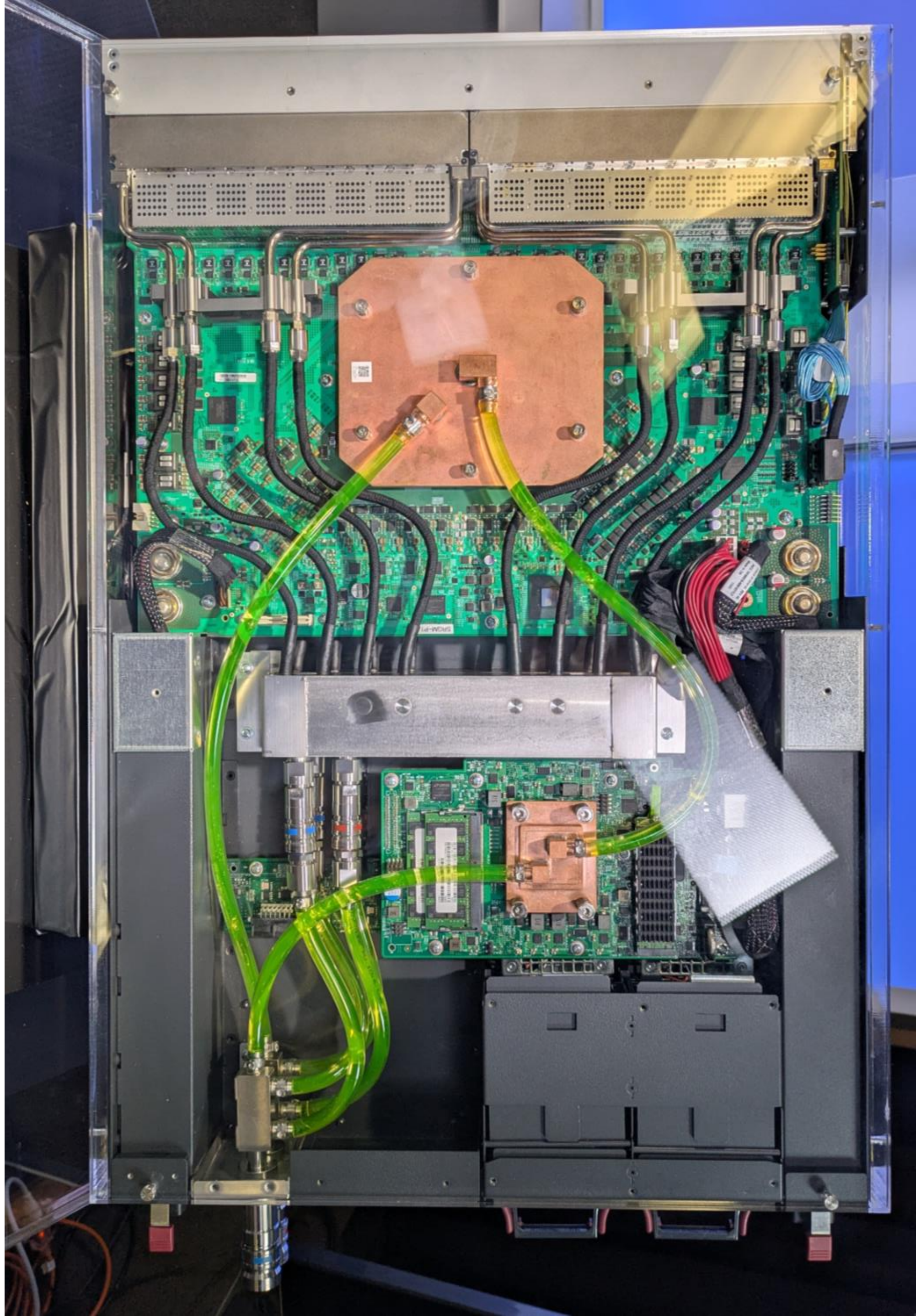
AI on Cisco –

Cisco Hardware in AI Networks

Cold Plate Liquid Cooling

- Power density/cooling is becoming the limiting constraint
- NVIDIA GB200 NVL72 is ***~1.2kW per GPU*** and ~***120kW per rack***[1]
- Microsoft and Meta Mount Diablo design uses ***400Vdc***[2] into the rack
- Google is planning for racks up to ***1MW***[3]
- Power savings
  - ~10-15% from system fans
  - ~60% facility power (chillers etc)
- Improves Power Usage Effectiveness(PUE)[4] ~20%

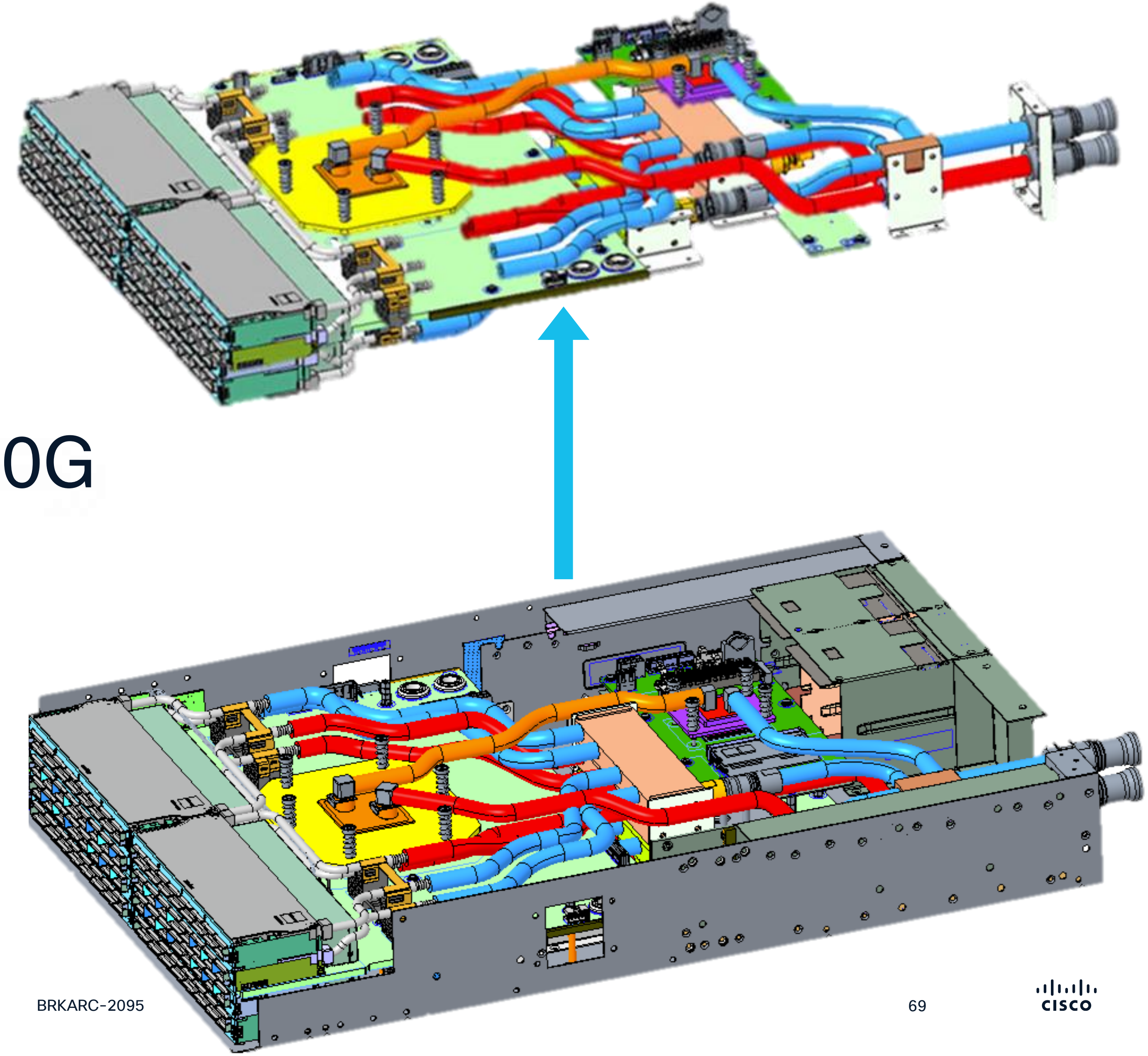
1. NVIDIA GB200 NVL72: https://training.continuumlabs.ai/infrastructure/servers-and-chips/nvidia-gb200-nvl72
2. Mount Diablo: https://www.datacenterdynamics.com/en/news/microsoft-and-meta-reveal-open-ai-rack-design-with-separate-power-and-compute-cabinets/
3. Google 1MW rack plans: https://cloud.google.com/blog/topics/systems/enabling-1-mw-it-racks-and-liquid-cooling-at-ocp-emea-summit
4. Power usage effectiveness: https://en.wikipedia.org/wiki/Power_usage_effectiveness



Liquid Cooling 51.2T Switch Technology Demonstration

Liquid cooled components:

ASIC, CPU, 64 x OSFP 800G



Liquid Cooling removes up to ***80%*** of system heat

25.6T Co-Packaged Optics(CPO) at OFC 2023

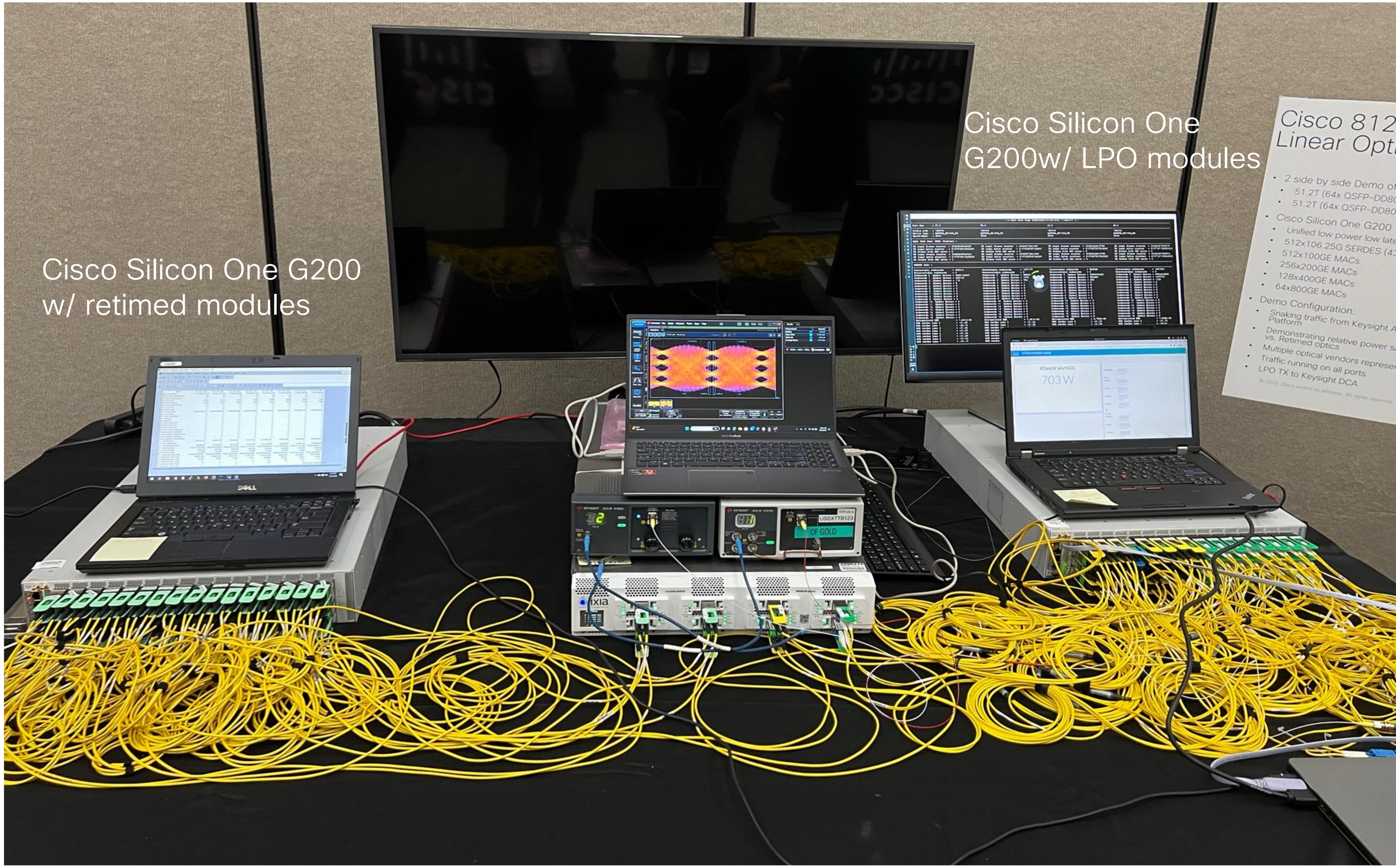
Retimed optics

CPO

CPO power reduction: ~270W

51.2T Linear Pluggable Optics(LPO) at OFC 2024

LPO power reduction: ~700W



Fault Managed Power: Touch Safe High Voltage DC

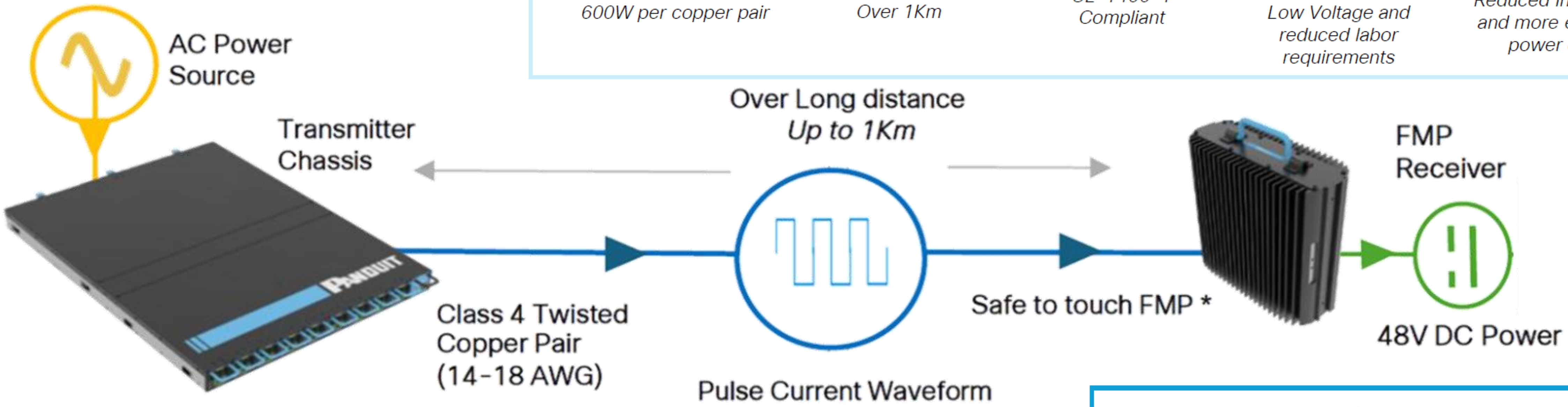
Significant Power
600W per copper pair

Long Distance
Over 1Km

Safety
UL-1400-1 Compliant

Speed to Deploy
Low Voltage and reduced labor requirements

Sustainable
Reduced Infrastructure and more efficient DC power delivery



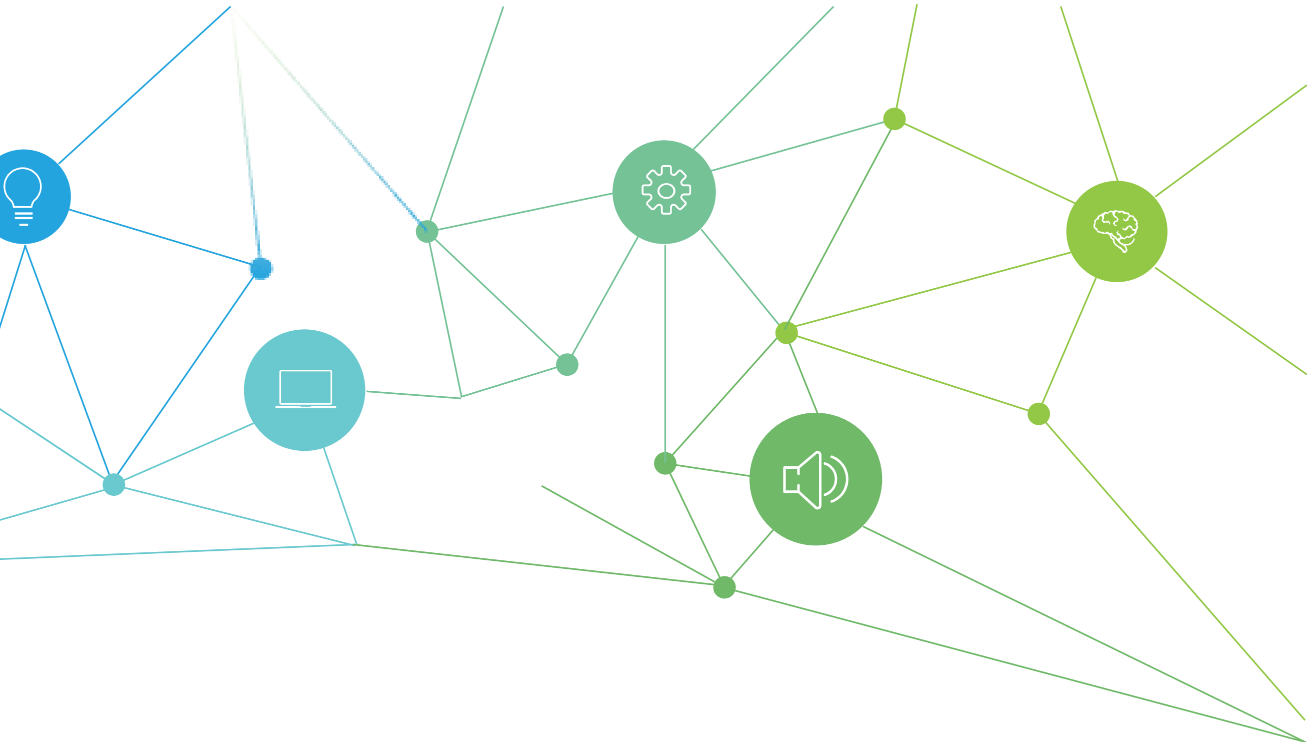
5-30% increase in energy savings in buildings with widespread adoption of DC power
US Dept Energy

10-20% increase in energy efficiency by eliminating AC to DC conversion
www.energy.gov

https://www.panduit.com/en/products/featured-products/panduit-fault-managed-power-system.html
https://www.cisco.com/c/en/us/td/docs/engineering_alliances/panduit_fmps_and_cisco_implementation_guide.html

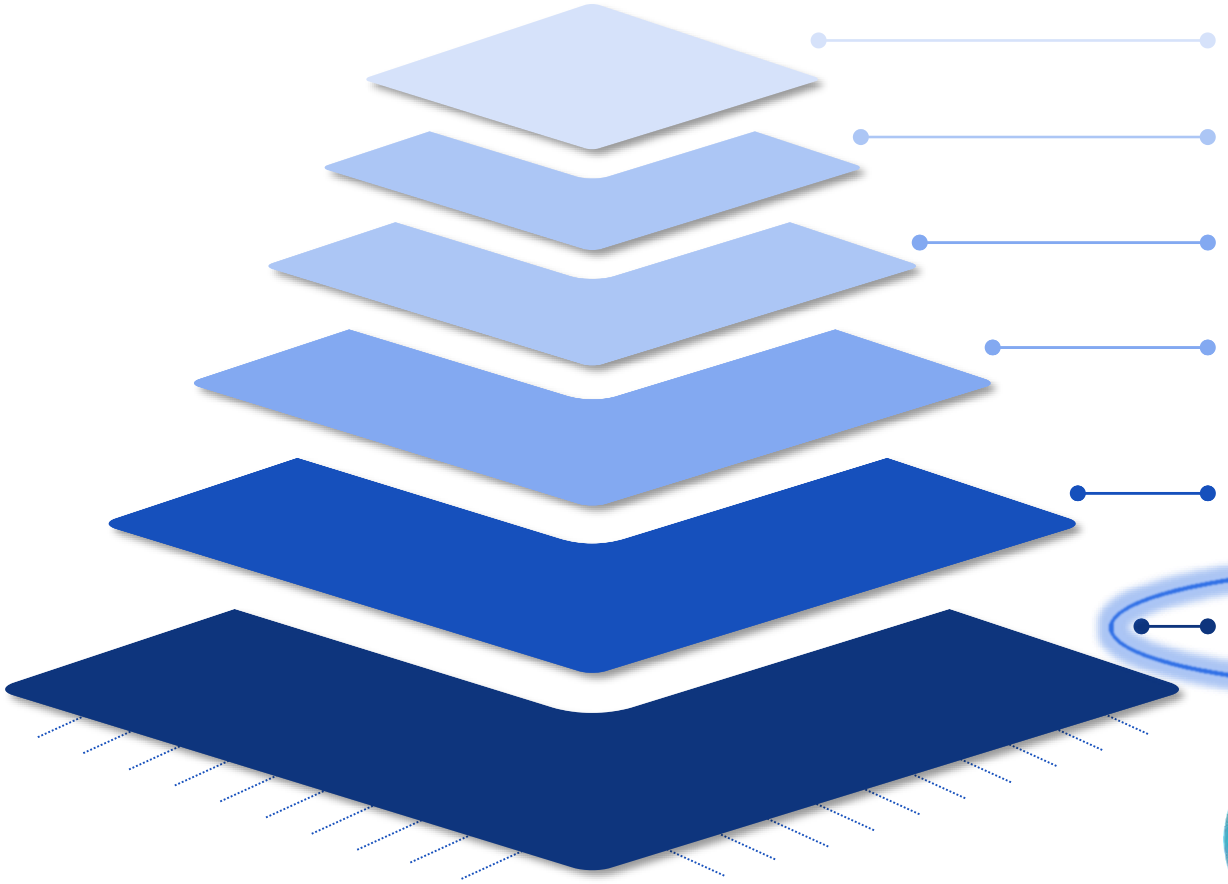
Summary – Cisco ~~Silicon~~ *Hardware* for AI

Cisco and AI



- Cisco is **investing** in AI capabilities
- We have a focus on **creating AI solutions for use by customers**
- We have a focus on **creating solutions that support AI workloads**

Cisco ~~Silicon~~ Hardware for AI
Foundational Elements to Support AI Growth



Applications and Zero Trust

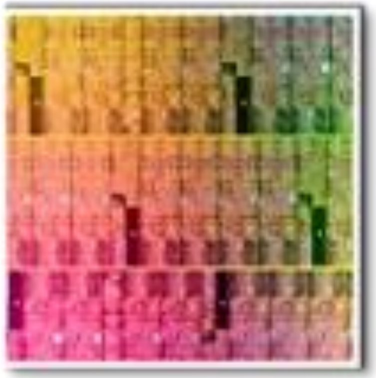
Highly Programable APIs

Cloud Management
On-Prem Management

Modern OS stacks

Physical and virtual infrastructure

Cisco Application-Specific Integrated Circuit (ASIC)



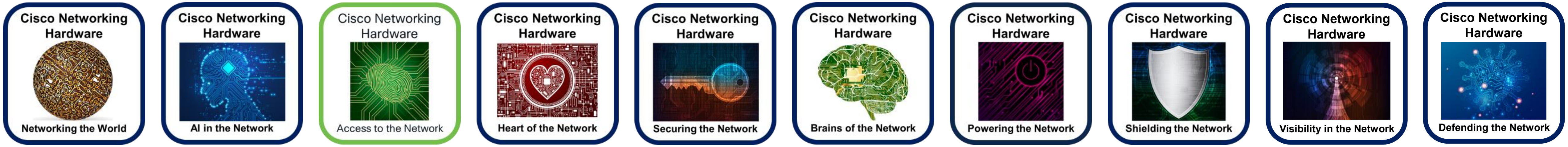
Best-In-Class Hardware

Secure Networks

Sustainable

Highly Programmable

Cloud Ready


Cisco Networking Hardware
Networking the World
Cisco Networking Hardware
AI in the Network
Cisco Networking Hardware
Access to the Network
Cisco Networking Hardware
Heart of the Network
Cisco Networking Hardware
Securing the Network
Cisco Networking Hardware
Brains of the Network
Cisco Networking Hardware
Powering the Network
Cisco Networking Hardware
Shielding the Network
Cisco Networking Hardware
Visibility in the Network
Cisco Networking Hardware
Defending the Network


YOUR NETWORK
IS OUR LIFE'S WORK
CISCO
Silicon One™
G200
©Cisco 2023
CISCO
Silicon One ™
P100
©Cisco 2021
CISCO
Silicon One™
Q201
©Cisco 2020

How Did We Do?

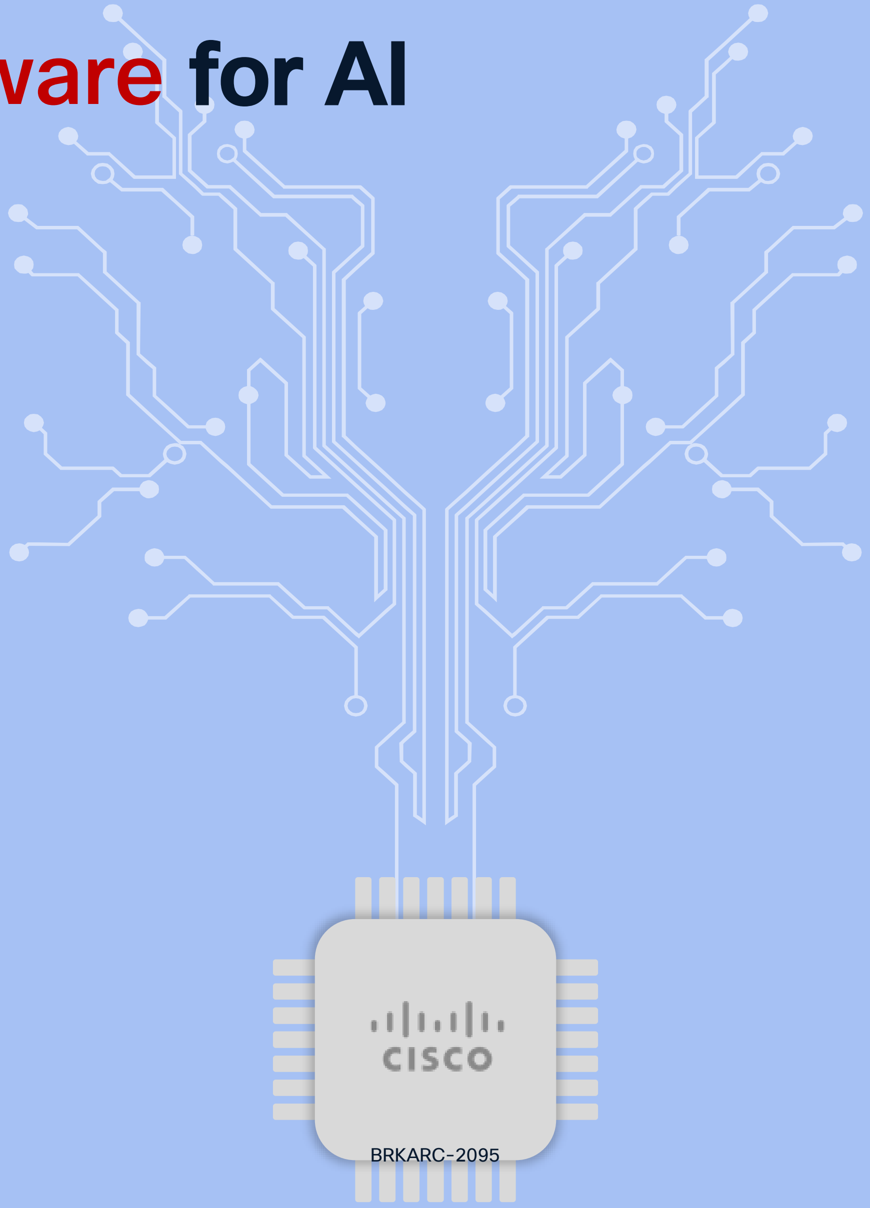
Cisco ~~Silicon~~ Hardware for AI

Do You Have a Better Understanding …

… what **Cisco is doing in AI** and why it matters?

… of why **Hardware Functionality and Flexibility** are Key for AI Solutions …

… and how You can **Leverage Cisco's Latest Flexible Hardware and Advanced Capabilities** in Your Own Network Designs?



Complete your session evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

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 www.CiscoLive.com/on-demand

Contact us at: email: dzacks@cisco.com
bluesky: petergjones.bsky.social

What else to see

- Silicon One
  - Networking for AI - DEMCPA-09
  - Networking for AI | Silicon One - DEMAIDC-04
  - Redefine your AI/ML networks with Silicon One - PSODCN-1005
  - Redefine your AI/ML networks with Silicon One - AIHUB-1004
  - SILICON ONE & ULTRA ETHERNET FOR AI INFRASTRUCTURE - BRKNWT-2508
  - Preparing for AI-Ready Infrastructure with Silicon One - ITLGEN-2065
  - Silicon One - DEMCPA-10
  - Ethernet Fabrics for AI clusters - Silicon One and Nexus - ultra high performance, scalable & non-blocking ethernet fabric. - BRKCOC-3005
- Liquid Cooling
  - WoS demonstration – Sustainability Booth
  - Integrated Rack Design | Liquid Cooling for Networking, Linear Pluggable Optics, and Rack System Cooling - DEMAIDC-02
  - The AI-Revolution – Cooling Technologies for the Data Center & Edge - WOSGEN-2100
  - Improving Power Usage Effectiveness | Immersion Cooling and Energy Management - DEMAIDC-06
  - Next generation power and cooling technologies in the datacenter - IBOCOM-2101
- Optics
  - Optics for AI Infrastructure - WOSGEN-2102
  - Optics for AI Connectivity - DEMSGC-03
  - Integrated Rack Design | Liquid Cooling for Networking, Linear Pluggable Optics, and Rack System Cooling - DEMAIDC-02
  - 400G, 800G, and Terabit Pluggable Optics: What You Need to Know - BRKOPT-2699

Thank you

cisco Live!