



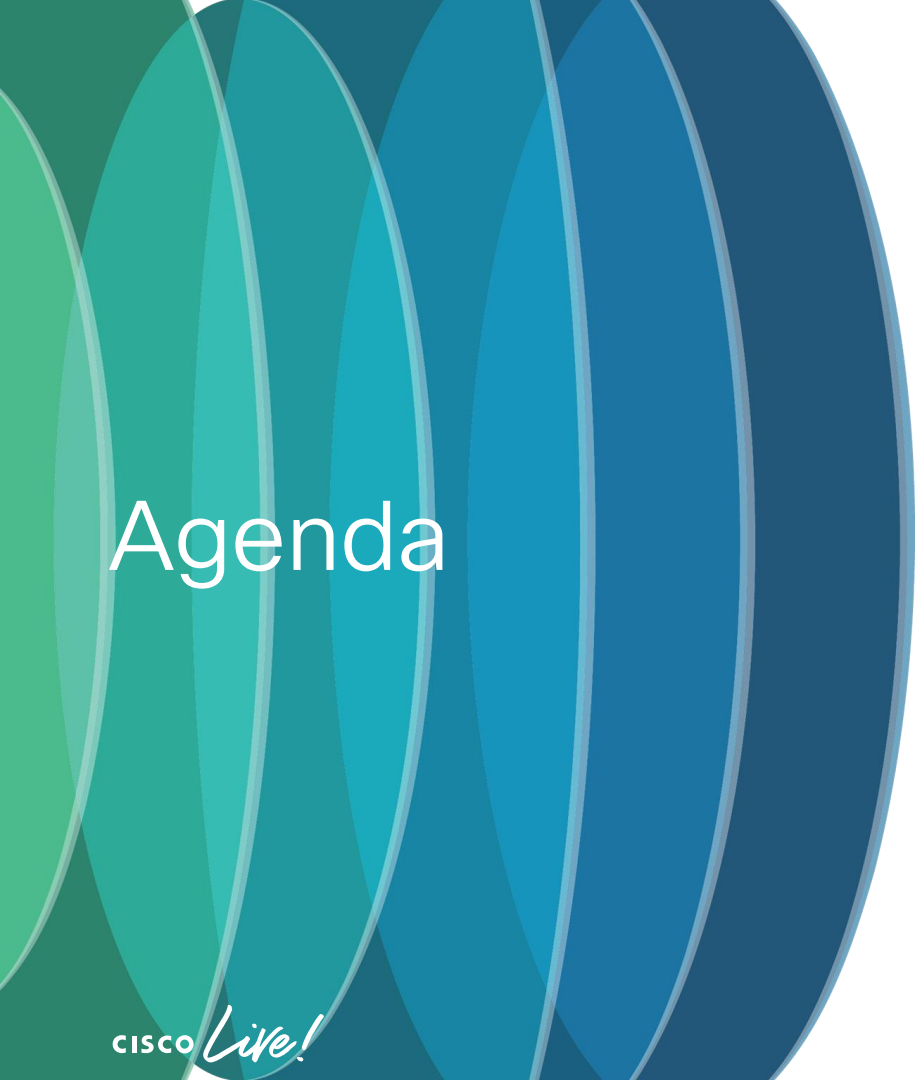
Cisco Silicon One

Redefine your AI/ML Network

Yaron Agami - Senior Manager, Product Marketing

AIHUB-1004

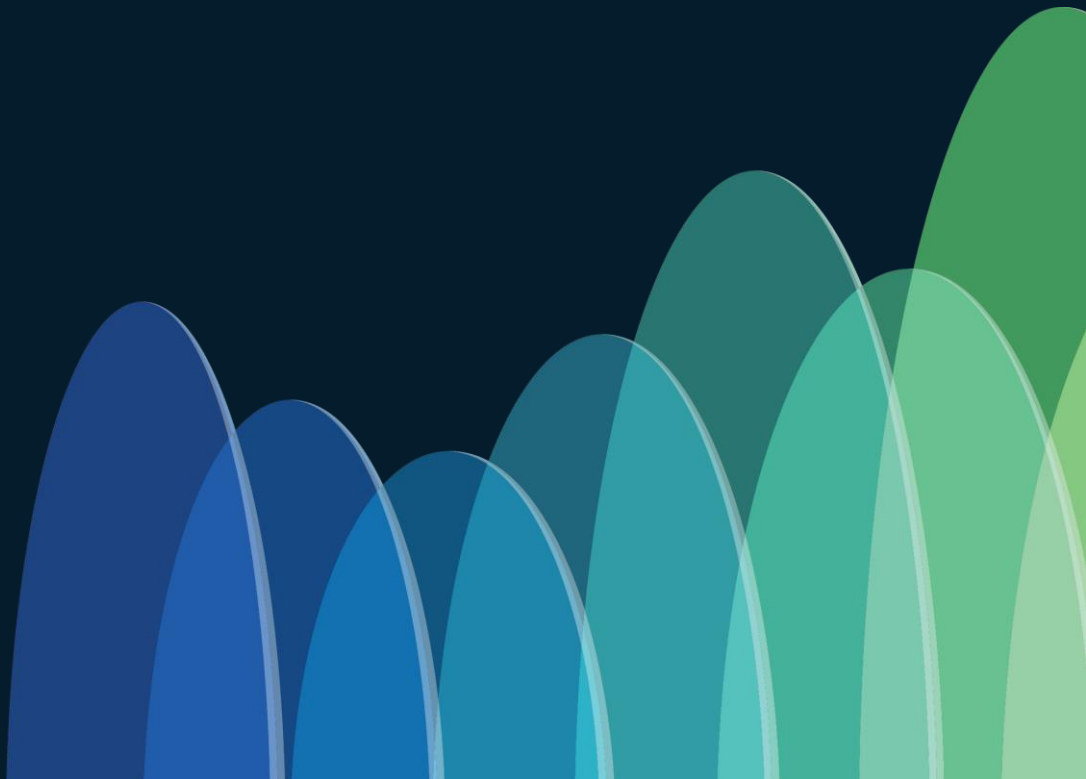
CISCO *Live!*



Agenda

- Introduction to Cisco Silicon One
- Networking in AI / ML clusters
- Cisco in AI / ML clusters
- Cisco Silicon One AI / ML differentiators

Introduction to Cisco Silicon One



Experience the power of a unified silicon architecture

Enable convergence without compromise
and redefine AI/ML networking

One Architecture. One Experience. Unmatched Capabilities

- 17 silicon chips were already developed
- Used in 60 Cisco platforms across all customer segments
- Powers
 - All of Cisco 8000 systems, every new Nexus and Catalyst systems
 - 5 out of 6 Tier1 Web Scales are using Cisco Silicon One, with more than 20 use cases
 - Running Variety of operating systems, including SONiC OS



Cisco 8000 portfolio

Cisco Silicon One

One Architecture. Multiple Devices. No Compromise

Switching

Routing



Use Cases

- Spine
- Leaf
- TOR

- DCI
- Core
- Peering
- Spine
- Leaf
- TOR
- Edge
- Campus
- Core

- Pre Agg
- Agg On
- Access
- Access

- Data Center

CISCO *Live!*

Service Provider

AI-ML Web Scale

Enterprise

Breakthrough innovation in performance, flexibility, and sustainability

Cisco Silicon One is the only unifying architecture that delivers the highest bandwidth routing and web scale switching silicon in the industry



Get performance at scale

Meet demands for scale with a choice of a high-performance, feature-rich, and low-power networking silicon—that provides a consistent experience across the entire Cisco Silicon One portfolio.



Leverage programmability

Evolve your AI/ML infrastructure with programmability while still providing low latency, cost, and power.



Boost sustainability

Significantly reduce costs, power, and latency of network deployments with fewer network layers. That means a more sustainable network with fewer devices that last longer and are more efficient.

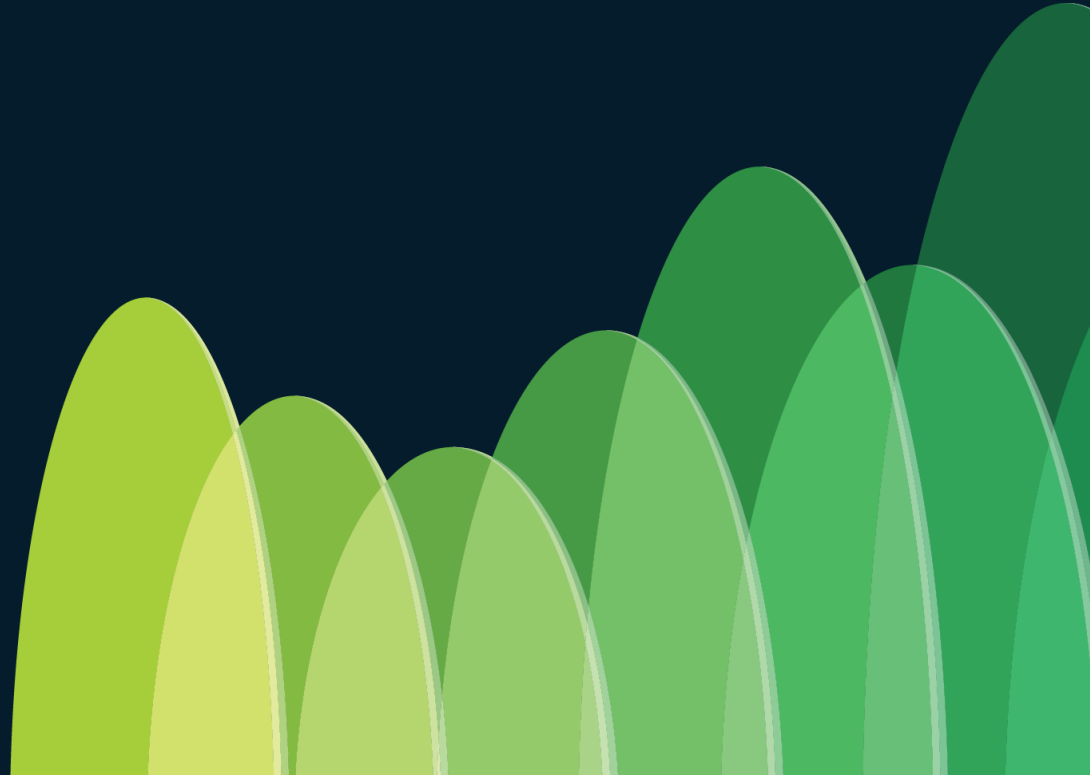


Manage and secure

Get end-to-end visibility, so you can see everything and respond when needed—with advanced trustworthy telemetry capabilities.

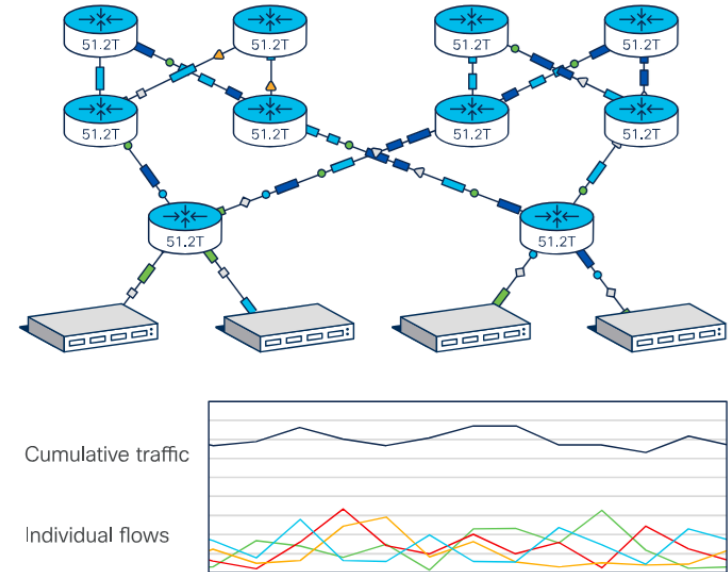
From top-of-rack (ToR) switches, web scale data centers and across service providers, to enterprise networks and data centers with a fully unified routing and switching portfolio, our devices power Cisco and third-party systems

Networking in AI / ML clusters



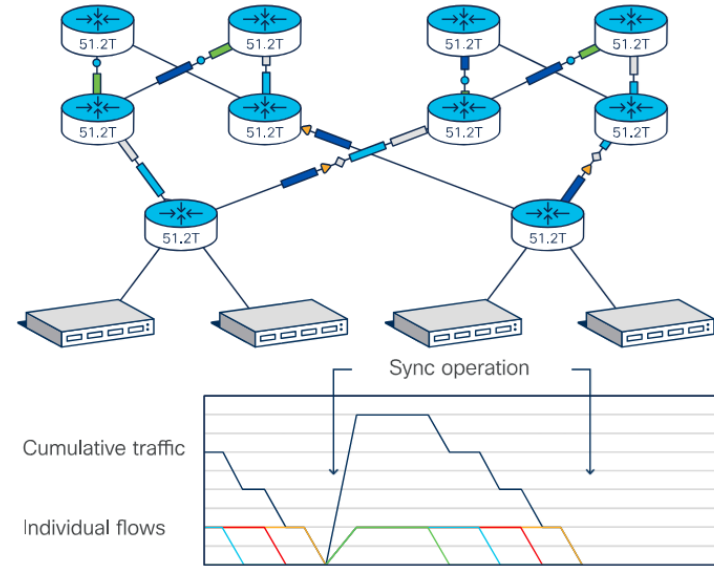
Traditional Data Center traffic patterns

- Server applications send data with unique traffic patterns and timing
- This leads to random, asynchronous, small bandwidth flows
- On average, this creates a consistent network load



AI Data Center traffic patterns

- AI/ML uses fewer, high-bandwidth, synchronized flows, leading to sharp network load fluctuations
- Latency and congestion cause some GPUs to receive data sooner, execute, and then stall, waiting for others to finish
- Optimal network performance is critical to maximize GPU efficiency and reduce job completion times



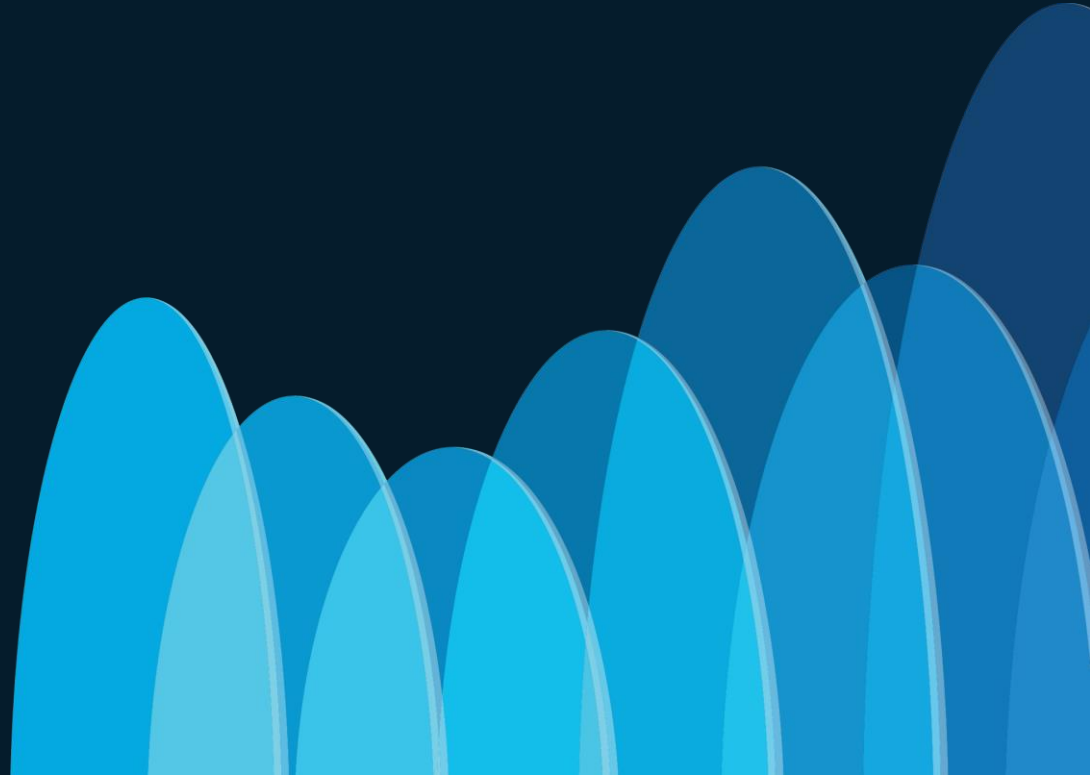
AI Networking Ethernet Fabric Choices

Lower latency, high bandwidth and superior multi-job efficiency, ensuring optimal AI workload performance

Standard Ethernet	Enhanced Ethernet	Ultra Ethernet consortium	Scheduled Ethernet
Foundational networking, for general-purpose data communication	Upgraded Ethernet for better reliability and efficiency for AI networking	AI-focused Ethernet initiative optimizing Ethernet for large-scale AI workloads	Ethernet with deterministic scheduling, providing short job completion time
Scalability – support high speed connections, cost-effectiveness, broad industry adoption	Lossless transport, improved Quality of Service (QoS), better congestion control.	AI-aware congestion control, low latency, scalability, industry compatibility, improved RDMA performance	Predictable latency, time-synchronized packet delivery
High bandwidth but lacks low-latency and congestion management for AI workloads	Reduces packet loss, improves stability for AI training clusters	Designed for multi-job AI training, dynamic resource allocation, scalability, large AI training clusters	Useful for AI real-time and time-sensitive applications

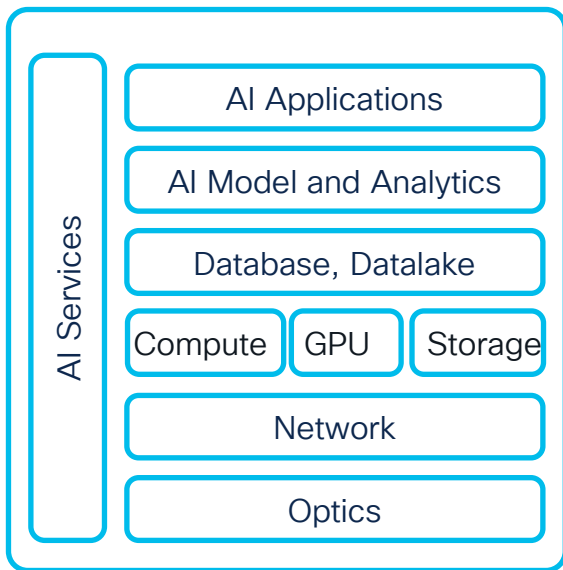
Silicon One programmable architecture gives customers Flexibility to program networks into various modes

Cisco in AI / ML clusters

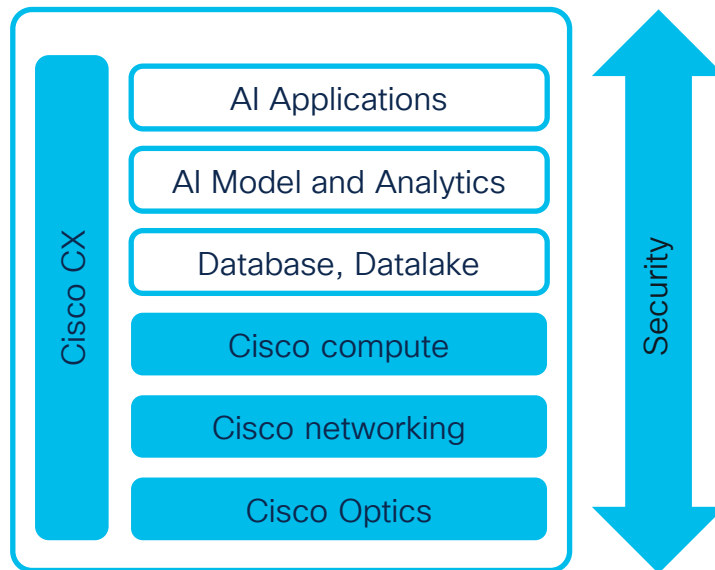


Breaking Down the AI Stack

AI stack model

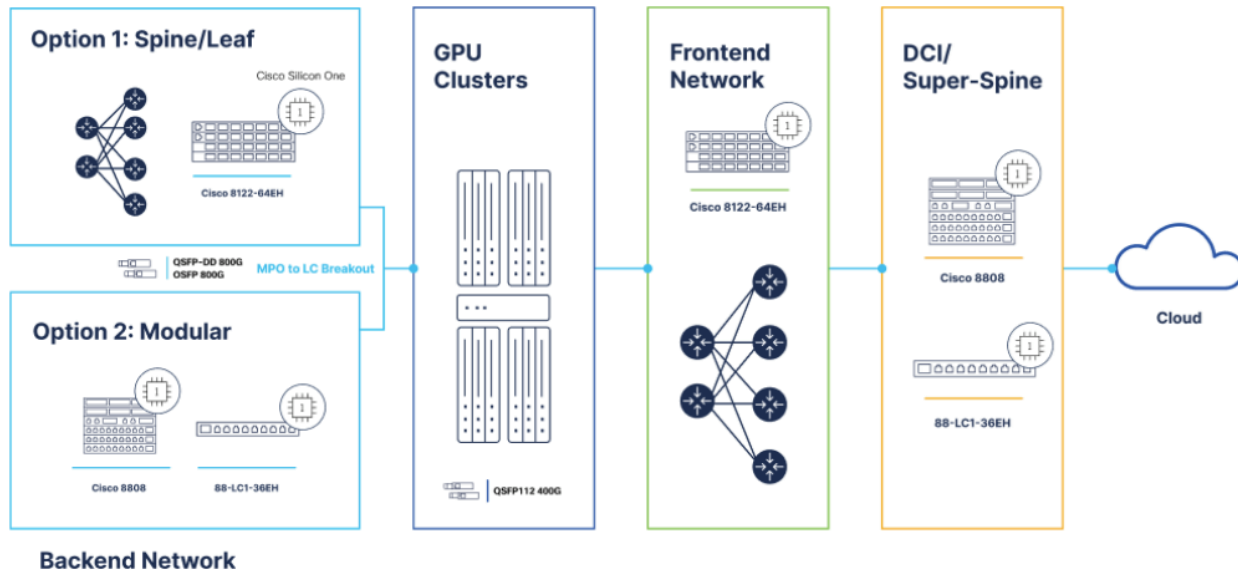


Cisco in the AI stack



AI Native Infrastructure

The demands of AI/ML workloads are surging, and the solution is clear: modern Ethernet-based AI-native infrastructure. You can now deliver open, high-performance, scalable infrastructure that supports AI/ML workloads, no matter what.



Cisco Nexus Hyperfabric AI

- ✓ Design, deploy and operate on-premises fabrics located anywhere
- ✓ Easy enough for IT generalists, application and DevOps teams
- ✓ Outcome driven by a purpose-built vertical stack

High-performance Ethernet

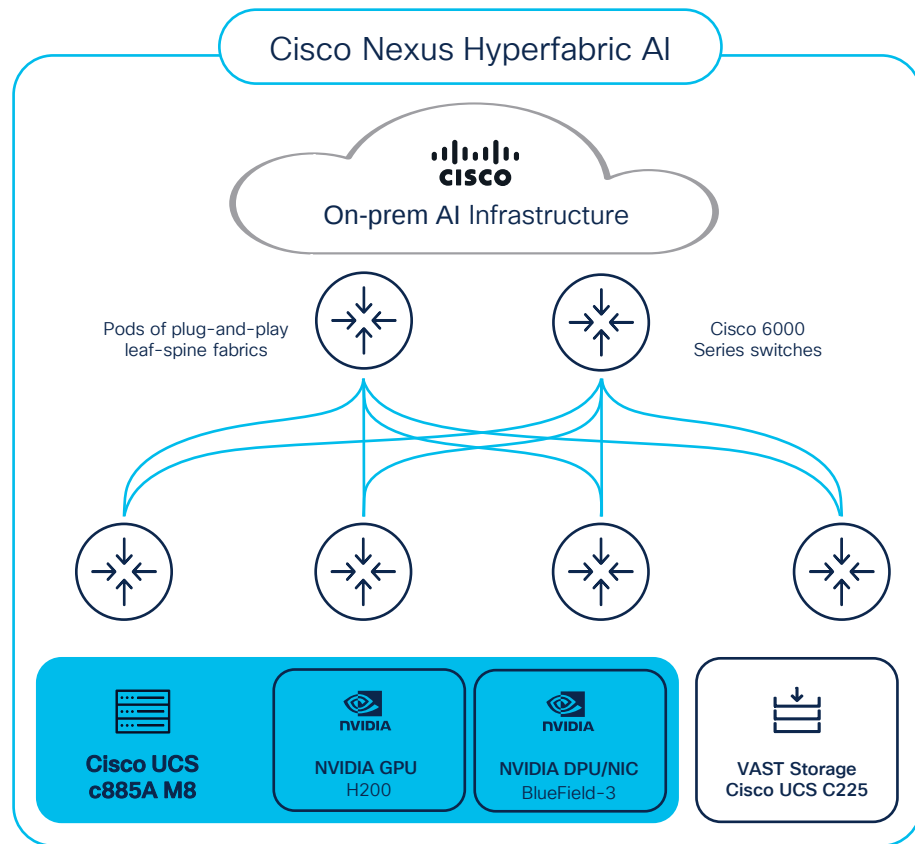
Cloud-managed operations

Democratize AI infrastructure

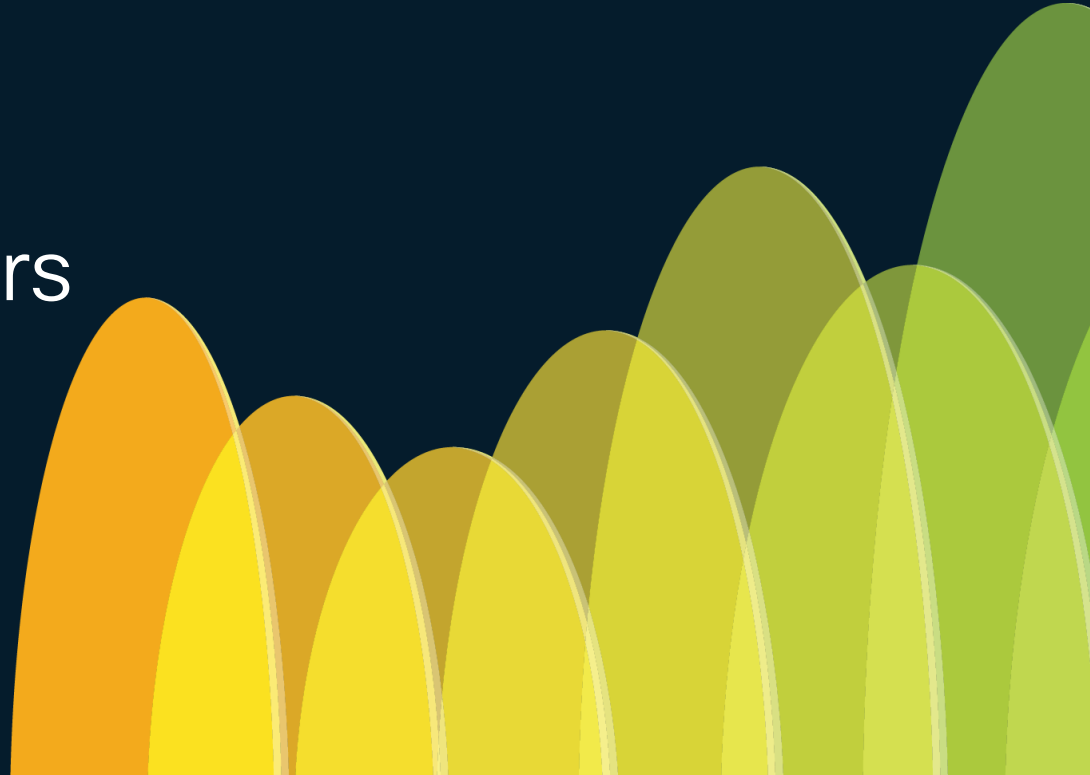
Unified stack including NVAIE

AI-native operational model

Visibility into full stack AI



Cisco Silicon One AI/ML differentiators

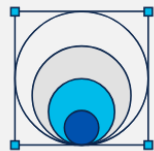


Cisco Silicon One Differentiators For AI networking



Efficiently process
AI / ML workloads

Achieve ultra-low latency
and optimal GPU cluster
utilization for fastest job
completion



Build scalable
AI/ML clusters

Flexibly program networks
with top Ethernet solutions



Establish sustainable
data centers

Optimize network efficiency
with high Radix and
advanced SerDes for power
savings



Manage and control

Enhance network utilization
with telemetry,
troubleshooting
and serviceability



Security

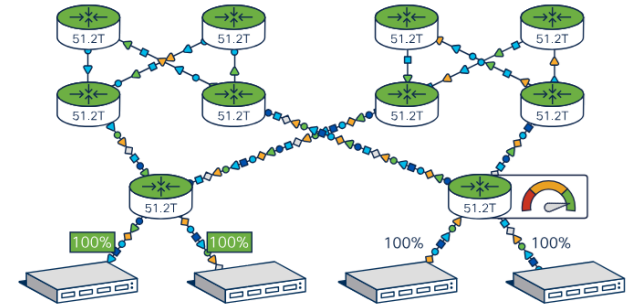
Ensure secure hardware
and effective encryption

Efficiently process AI / ML workloads

Load balancing alternatives



- Traditional – ECMP
 - Distributes data evenly across multiple network paths for smooth flow
 - Uses stateless Equal-Cost Multi-Path (ECMP) routing
 - Relies on a unique code from data source and destination
 - Ensures data travels the best path, maintaining order
- Dynamic stateful – Flowlet
 - Unlike stateless ECMP, maintains state per flow and re-balances paths during idle times
 - Adapts to conditions to optimize link utilization and performance
- Dynamic stateless – packet spray
 - The switch sprays data packets randomly in a round-robin fashion across all available network paths based upon port utilization
 - Optimizes resources and maintains packet order for reliable transmission
 - Enhances performance by adapting to network conditions and maximizing link use



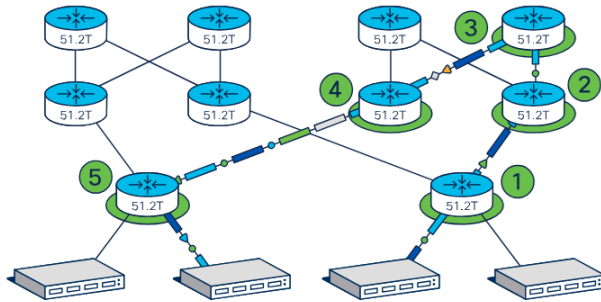
Dynamic stateless - packet spray



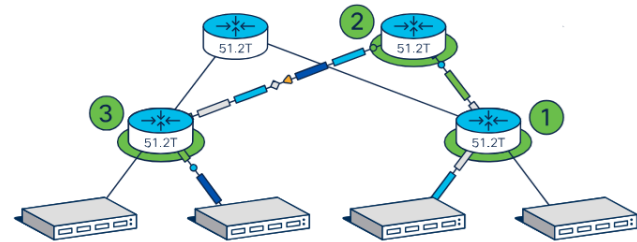
Efficiently process AI / ML workloads

Radix in Silicon Chips Influences Latency

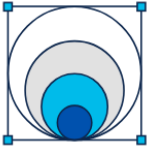
- Radix is the number of ports on a network chip, affecting job completion time and latency
- Higher radix allows more connections, essential for scalability
- Cisco Silicon One has the highest radix with 512 x 100GE ports
- In our example it reduces latency from five hops to three, enhancing agility in large-scale networks



256 Radix enforce more layers



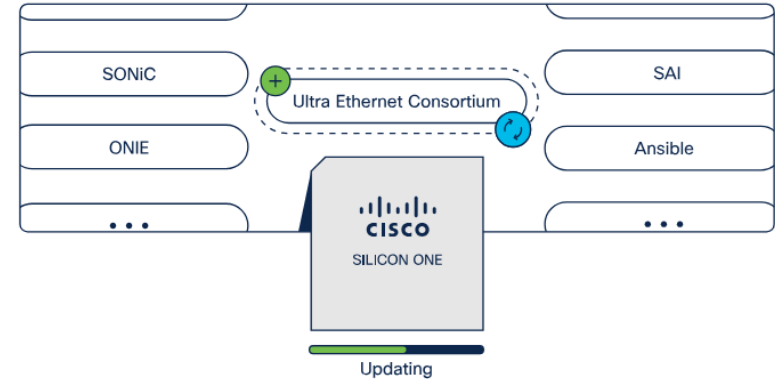
512 Radix flattening the network and reduces latency



Build scalable AI/ML clusters

Leading Programmable Networking

- Cisco supports industry standards, open-source models, and open networking
- Cisco Silicon One offers full programmability, adapting to current and future standards
- The SDK simplifies ASIC updates, impacting the Network Operating System
- Seamlessly implements new methods, like UEC congestion control, to enhance performance

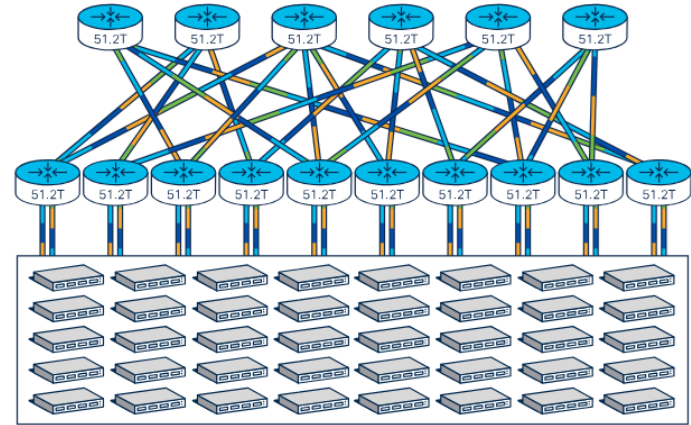


Establish Sustainable Data Center

Advanced SerDes innovations



- AI-optimized networks offer 100 to 1000 times more bandwidth than traditional ones
- Increased bandwidth leads to substantial power demands
- Focus on reducing system power, especially optics
- Cisco's Silicon One uses SerDes for efficient data conversion



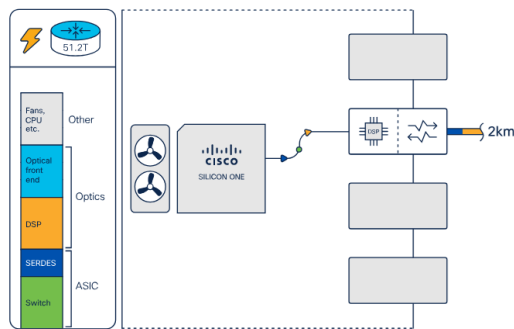
Optics account for approximately 50% of the total power used in AI networking.



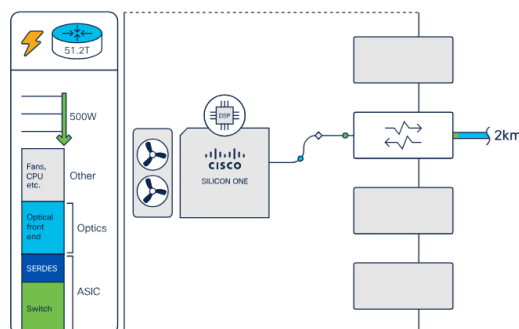
Establish Sustainable Data Center

Advanced SerDes innovations

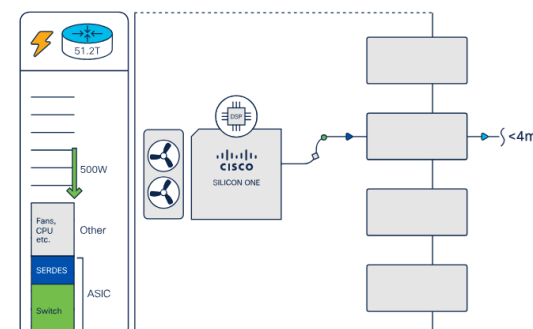
Unique capabilities of Silicon One enable the use of lower power linear pluggable optics (LPO) or passive direct attach copper cables (DAC)



Traditional optics



Linear pluggable optics



Direct attached cable

Establish Sustainable Data Center

Sustainable high-radix networking

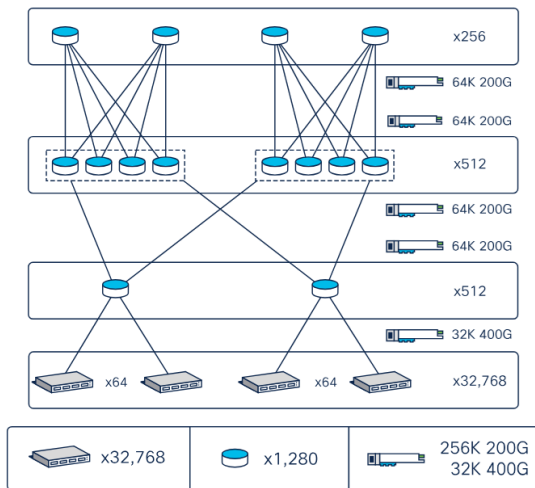


- Radix: Number of ports a network chip supports; crucial for scalable AI networking
- Cisco Silicon One has the highest radix with 512 x 100GE Ethernet ports
- High radix reduces a large AI cluster's network layers from three to two
- Results: 33% fewer layers, 50% less optics bandwidth, 40% fewer switches
- Saves up to 9 million kWh/year, equivalent to over 6,000 metric tons of CO2 or burning 7.3 million pounds of coal

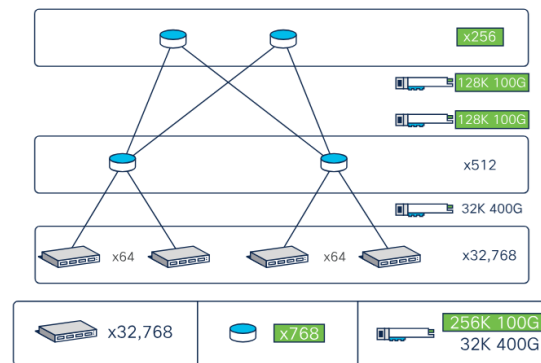


Establish Sustainable Data Center

Sustainable high-radix networking



Three layers network



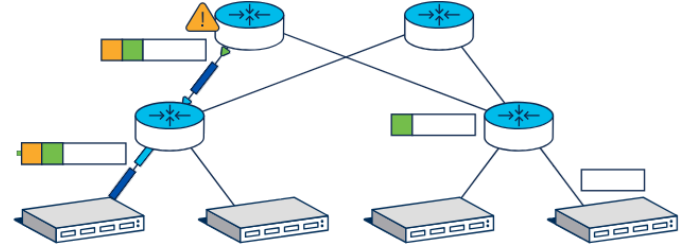
Two layers network



Manage and Control

In band telemetry - optimizing networks with smart packet stamping

- Cisco Silicon One stamps packet headers with congestion level, delay, and output port info.
- Network Interface Controllers (NIC) use this data for end-to-end congestion control.
- Benefits include reduced Job Completion Time (JCT) and lower latency.

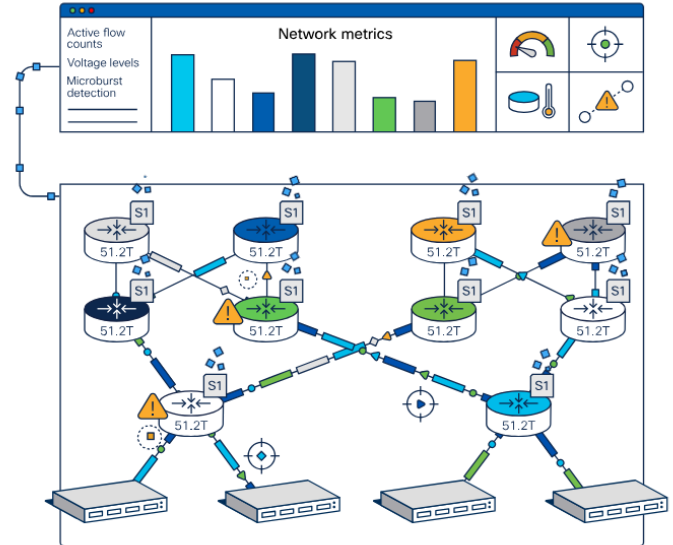




Manage and Control

Streaming telemetry – proactive network optimization

- Proactive network maintenance optimizes performance by addressing weak points and risks
- Telemetry data collection from the control system, particularly from the Cisco Silicon One chip, is vital for this maintenance approach
- Cisco Silicon One collects data on temperature, voltage, link status, congestion, traffic type, active flows, and microburst detection
- The control system updates routing paths using this data to enhance job completion time and reduce latency.





Manage and Control

Improving network troubleshooting

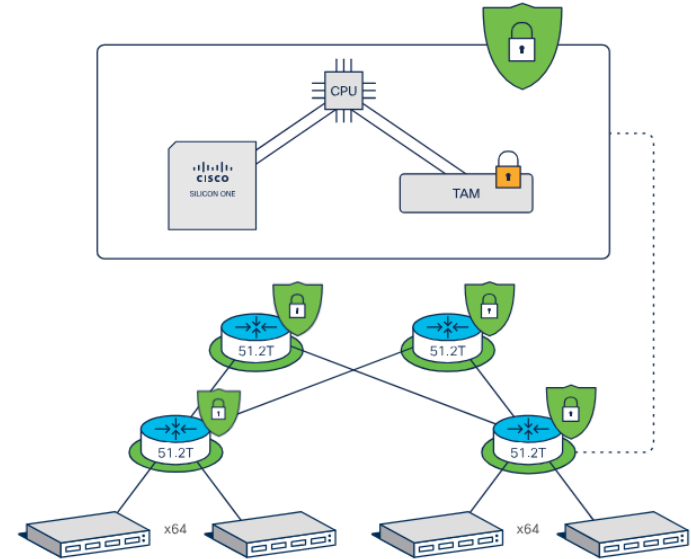
- Cisco Silicon One allows access to trace records of packet flows across switches
- Provides insights into processing stages, table interactions, and queue usage
- Helps identify causes of packet drops, delays, unbalanced loads, and congestion
- Enables real-time identification of elephant flows for optimized load balancing
- Improves troubleshooting and serviceability significantly



Security

Establishing Network Trust

- Cisco Silicon One ensures device security with built-in mechanisms to verify chip authenticity.
- Trust Anchor Module checks CPU software, and the Silicon One device identifies itself to the software.
- Confirms correct installation in intended systems, ensuring trusted systems and networks.

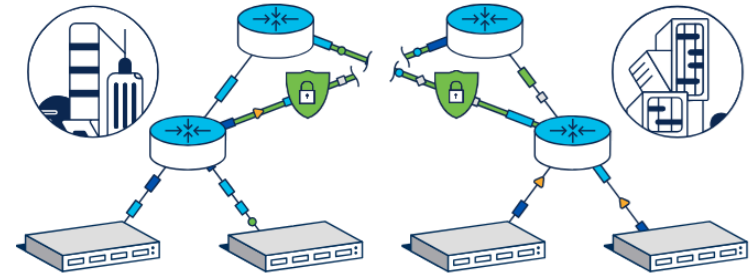


Security

Optimized Encryption and Performance



- Cisco Silicon One supports IPsec and MACsec for secure AI workloads in scale-out networks
- Features a unique format combining MACsec and IPsec benefits, tailored for RDMA workloads
- Provides native RDMA support and encapsulates all packet formats beyond IPsec limitations
- Ensures low latency with cut-through functionality and improved load balancing with visible entropy fields
- Enhances job completion time (JCT) compared to IPsec and MACsec



Webex App

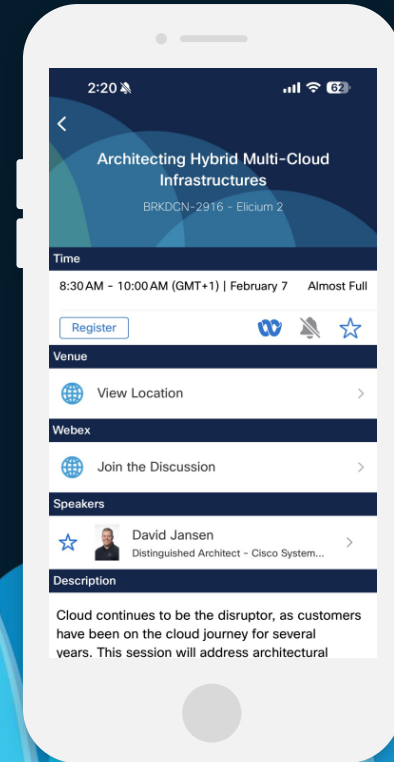
Questions?

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.



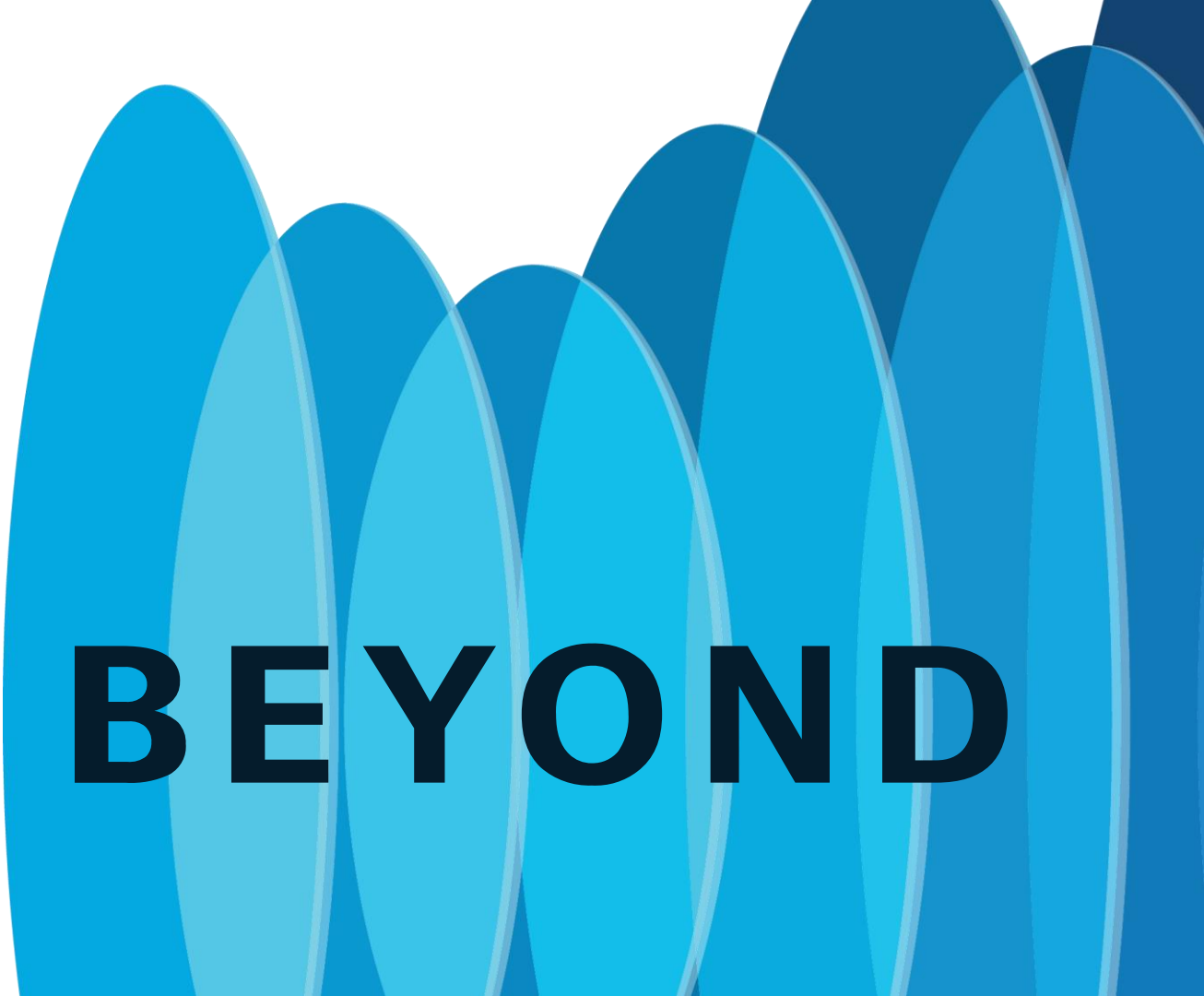
Thank you

CISCO *Live!*



CISCO *Live!*

GO BEYOND

The background of the slide features a series of overlapping, teardrop-shaped elements in various shades of blue, ranging from light sky blue to deep navy blue. These shapes are arranged in a way that creates a sense of depth and movement, resembling a stylized horizon or a series of waves. The overall aesthetic is clean and modern.