Impact of AI traffic in transport networks

Virginia Teixeira – Principal Solutions Engineer
Global Sales Specialists

BRKSPG-1180

Agenda

- AI Traffic -What's Different
- AI Connectivity Scenarios
- Foundational Capabilities
- Building Differentiation
- Conclusion

Is AI traffic moving the needle?

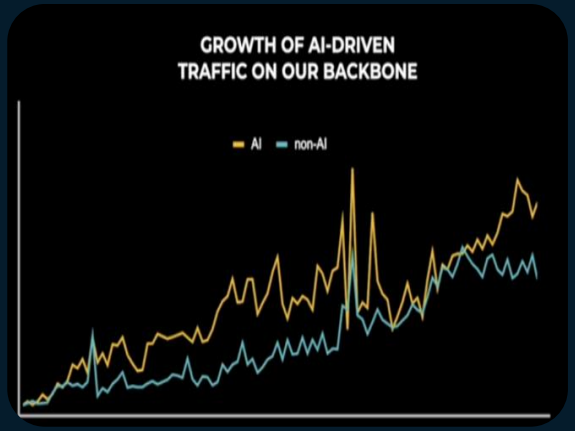
46%

of AI processing by 2027 will be inferencing,

66%

of enterprises list GenAI workloads as one of their top use cases for using multi cloud networking

IDC report



Meta backbone traffic grew >30% YoY for the last 4 years
AI-driven traffic grew faster and overtook non-AI traffic

Meta, @Scale conference, September 2024

36x

23/24 YoY AI traffic growth

22x

23/24 YoY user request growth

60%

23/24 YoY AI transaction size growth

openrouter.ai/rankings

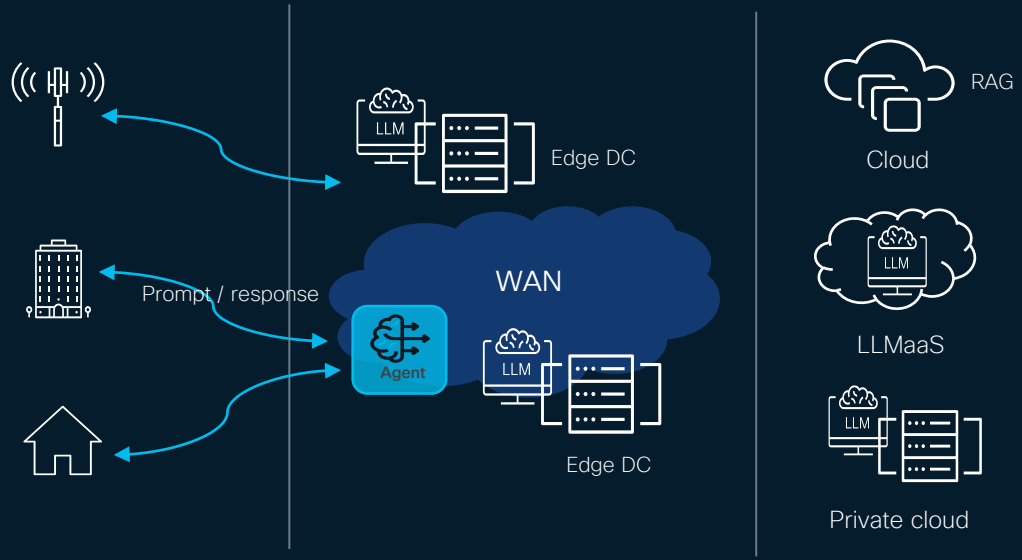
New AI-Assistants

will drive an increase in uplink traffic that is unprecedented, beyond the capacity of current 5G networks as soon as 2028

Mobile Experts, September 2024

AI Inference

@edge, RAG, AI agents, AI assistants, Split Inference



Bandwidth Demand

- AI traffic is unique, dynamic, non-cacheable
- Upstream 10x downstream – impact on the access
- AI traffic is growing fast
- Peering links can be expensive and hard to scale

Latency

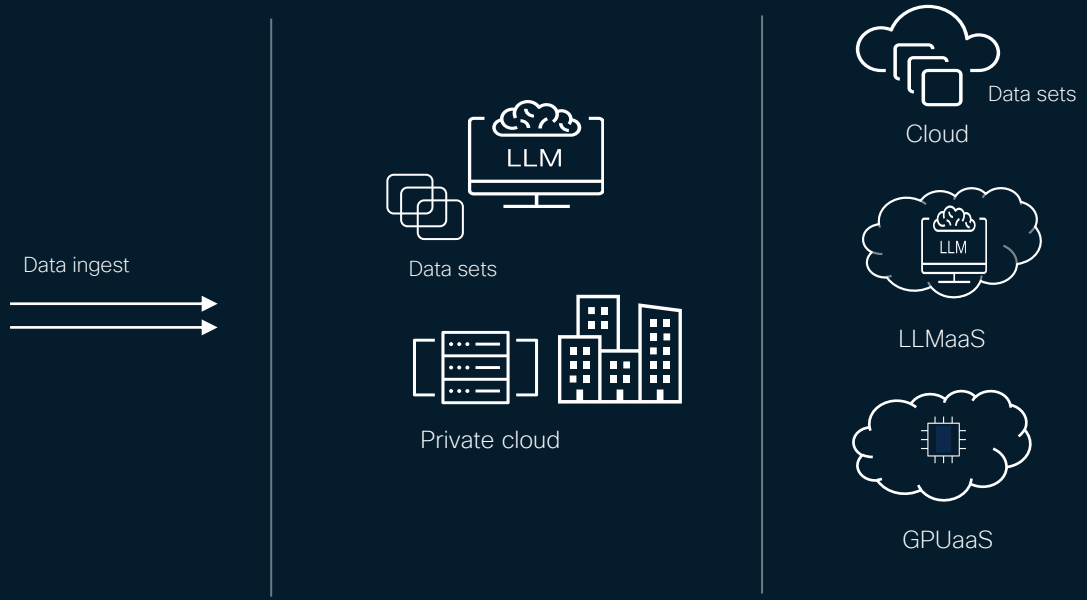
- AI assistants are chatty, interactive – every 300msec increase leads to 30% drop engagements
- AI agents require multiple processing steps before reply– latency x10s or higher with multi-agents
- Less predictable latency due to multi-modal varying request/response sizes
- For example, Meta has the goal to have sustained <30msec rtt for real-time AI/immersive experiences

Resiliency & security

- Reduce blast radius – service continuity
- Process data locally, regulatory compliance, protect IP
- Failures/error rate can impact significantly a real-time, interactive experience`

AI Training

Training, Fine-tunning, Federation, Swarm



Dataset movement & replication

- Very High BW needs with high peak-to-average ratio
- AI Federation require iterations of training with transfer of model, e.g 70B parameters model = 150GB
- Fresh data ingest – collect from where it's generated to where it will be used
- AI uses a lot of data and instigates the generation of more data

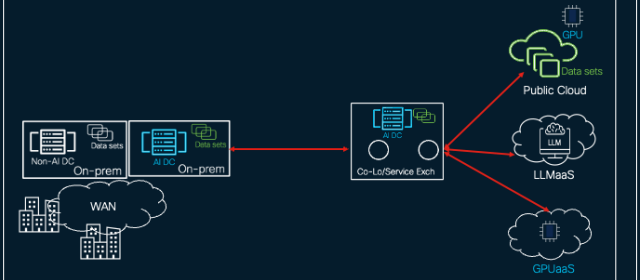
Secure high throughput transfers

Reliable & Resilient

Good quality data and network resiliency is critical for training

AI Connectivity scenarios

DCI Expansion



What's driving it

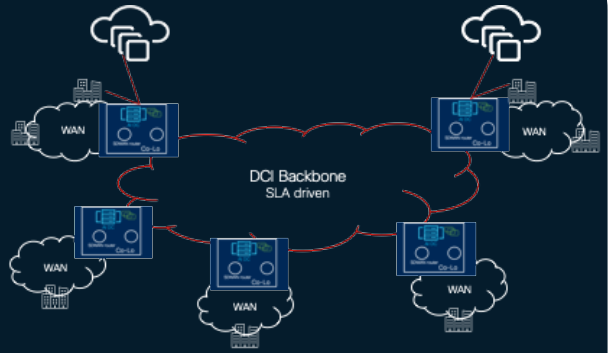
Dominated by training- Centralised public & private cloud, public/internet data

What's required

BW expansion, High Resiliency, Quantum-safe encryption, Visibility, Automation

- DCI aaS model (MOFN, RON)
- IP and optical solutions

Any-to-Any DCI



What's driving it

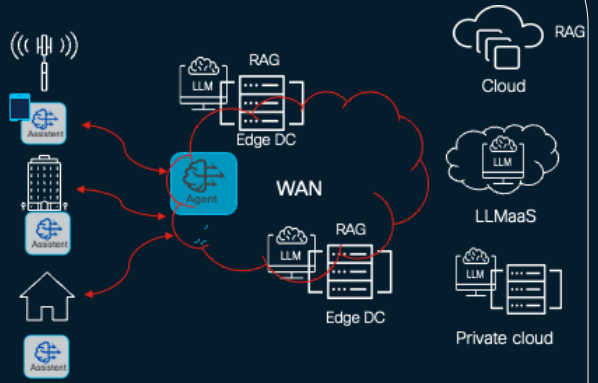
Dominated by inference- Distributed DCs, more sites, private data, move to private cloud

What's required

DCI backbone providing any-to-any, on-demand connectivity

- SR-based converged DCI backbone over shared infrastructure

Intelligent AI connectivity



What's driving it

Inference@edge – AI in production, customised models, more data over WAN

What's required

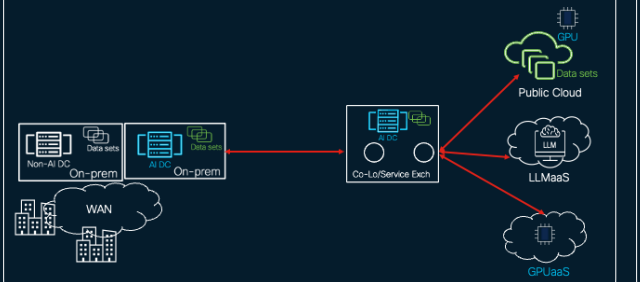
Guaranteed WAN SLAs
Leverage Agile Services Networking
AI vs non-AI traffic visibility

AI Connectivity Scenarios

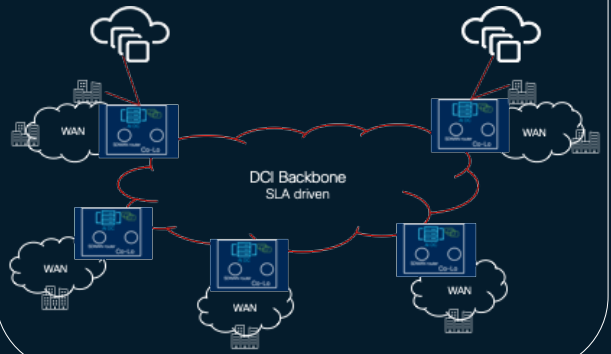
cisco Live!

AI Connectivity Scenarios - Technical Solutions

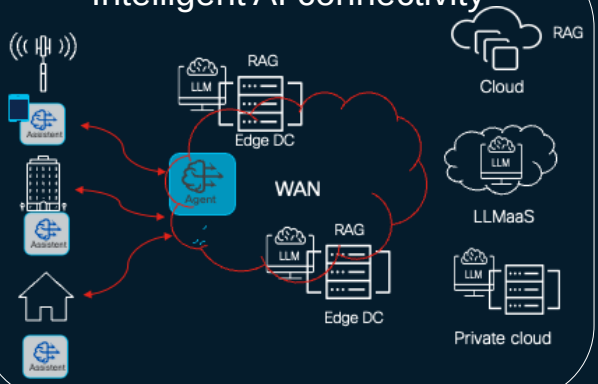
DCI Expansion



DCI Any-to-Any



Intelligent AI connectivity



High-capacity links 100G+ private or shared (p2p, hub&spoke) → **Optical DCI**

- MOFN: WaaS, spectrum aaS – open APIs enabling aaS, richness of metrics, assurance
- Built-in resiliency and security

Tiered BW IP services, RON, shared (multi-service, multi-tenant), up to 100G → **IP DCI**

- NaaS for DCI – SR based network, open APIs, closed-loop assurance
- Built-in resiliency and security

From Access to DC

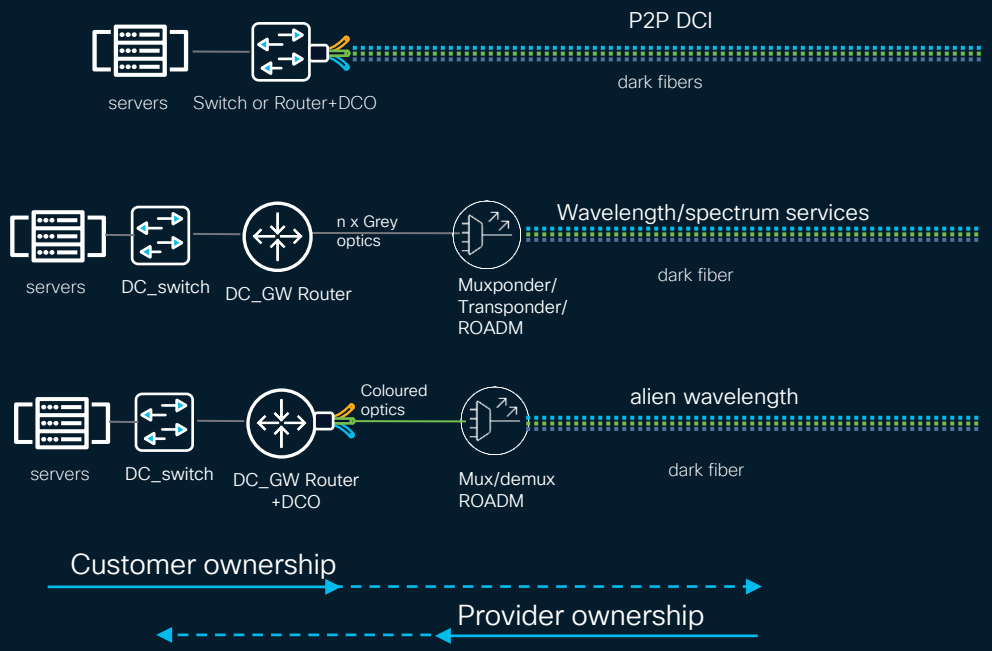
SR/SRv6 underlay integrated with SDWAN overlay

SRv6 to the CPE

AI traffic detection, AI KPIs

Agile Services Networking

Optical DCI solution for AI
High-capacity, High-resiliency



Need

Scalable and high capacity (>100G) DC interconnection and cloud connectivity

Simple network topology

Emerging use case is RoCEv2 over DCI

aaS models

Solution

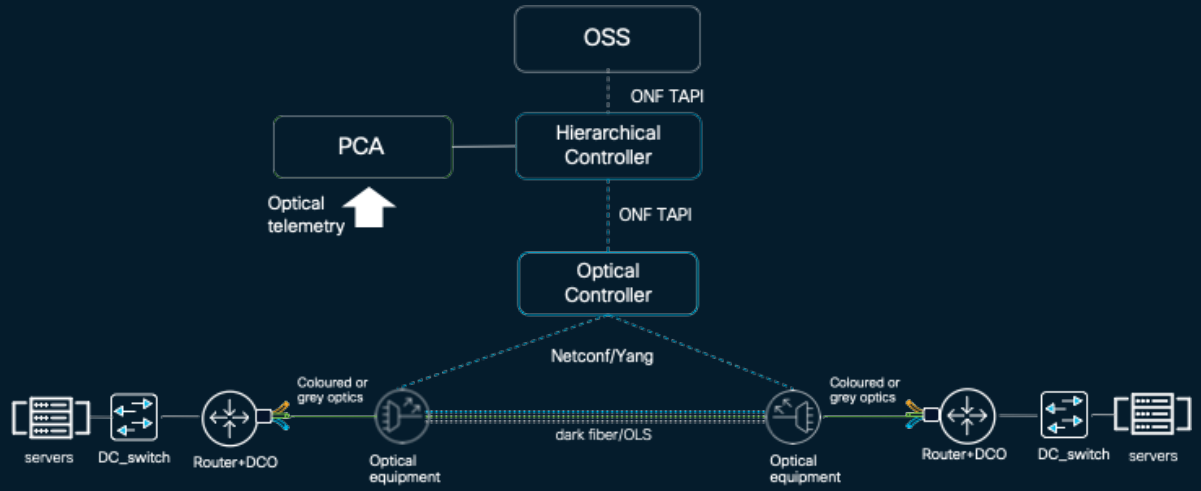
Complete optical portfolio for diverse capacity, distance and topology requirements

Resiliency/availability with highest quality optics in the industry and carrier-grade equipment

Security at optical and network node level

Optics and optical is a big portion of DCI infrastructure for AI

Optical DCI solution for AI
Automation and Visibility for Assurance and aaS



Automation and visibility are foundational to
- automated assurance
- WaaS, spectrum aaS

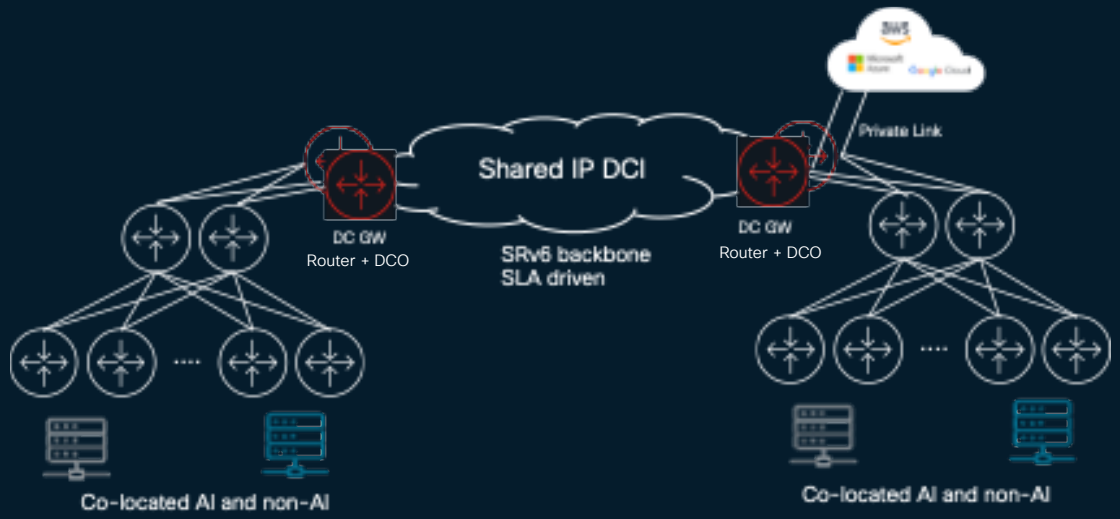
Cisco optical automation is compliant with IETF ACTN and standard ONF TAPI interfaces

Automated service delivery with integrated provisioning and service validation

Rich and granular visibility of optical KPIs and optical parameters aggregated in PCA – assurance, resiliency and security

PCA – Provider Connectivity Assurance

IP DCI solution for AI
Shared IP DCI



Need

DC GW w/ advanced routing for DCI, WAN, cloud connectivity – any-to-any connectivity

Shared IP DCI for AI and non-AI traffic

Tiered IP services <100G

Solution

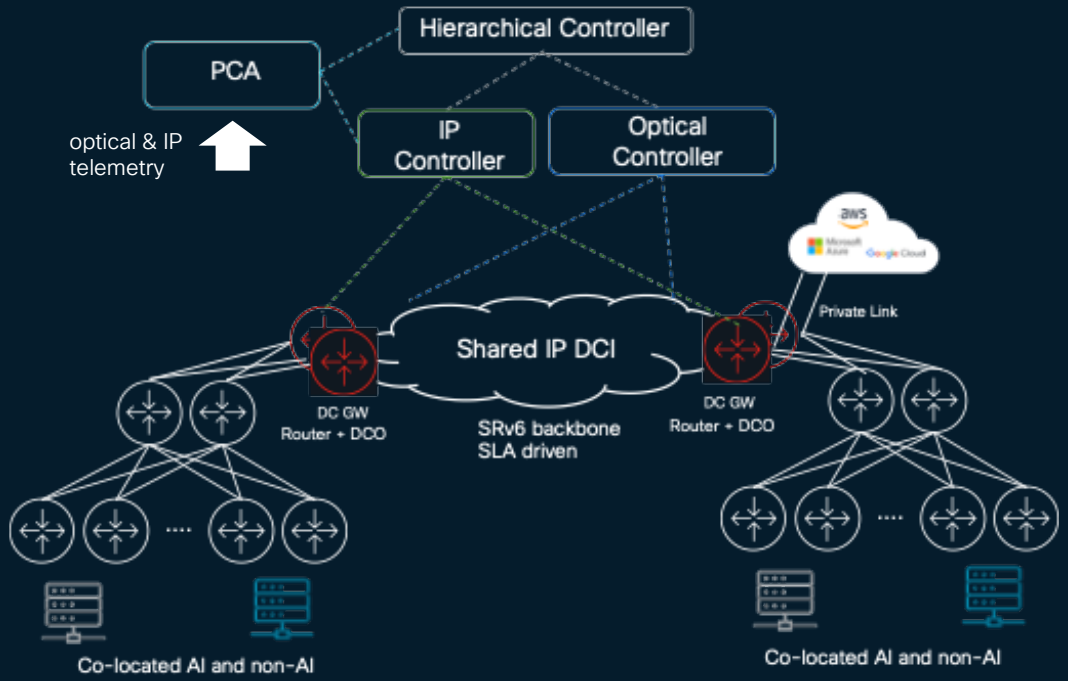
Traffic differentiation /network slicing with flexible any-to-any connectivity
- SR policies, SR-TE, FlexAlgo
- SR programmability

High resiliency/availability
- SR HA capabilities, disjoint paths
- QoS best practices
- Quality of optics matter
- Visibility to improve resiliency

Security/sovereignty of data in-transit
- Quantum-safe MacSec w/ SKIP
- SR-TE/FlexAlgo secure path enforcement

Sustainability & space optimization - RON

IP DCI solution for AI
Automation and Visibility for Assurance and aaS

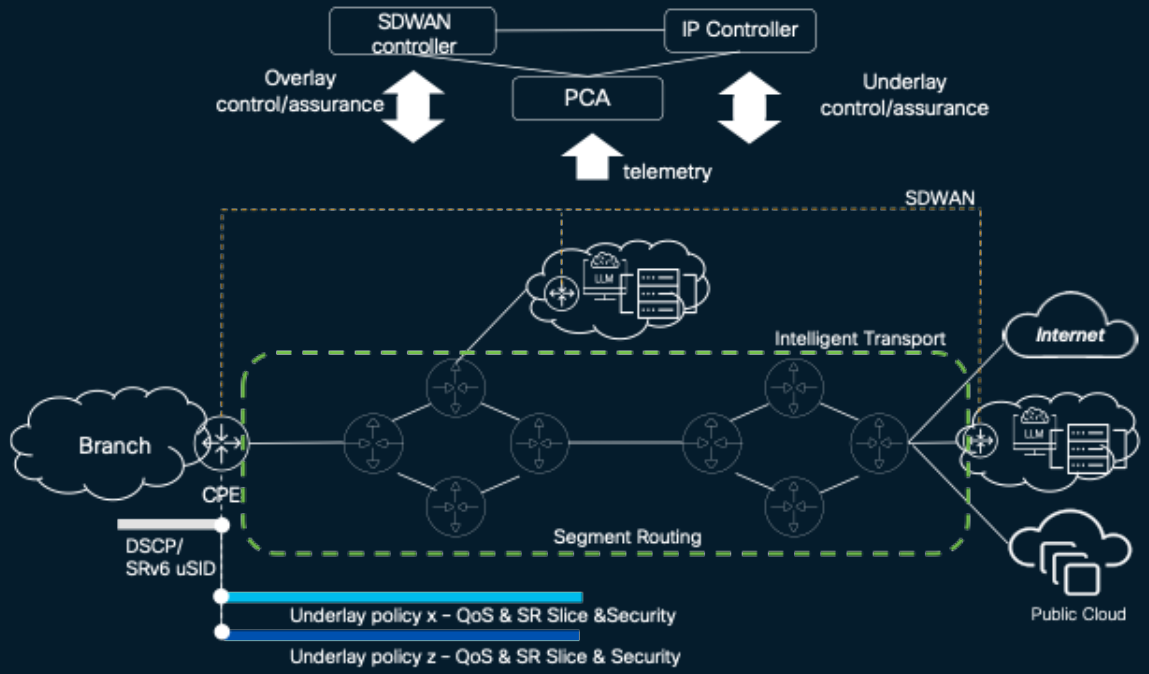


Visibility of high-speed links with PCA
- SR/SRv6-PM, SRv6-IPM, HW/SW telemetry, HW/SW sensors
- Correlate metrics/KPIs with topology

Integrated optical and IP insights for RON systems

Automated service health and network health
Open APIs for a complete NaaS model

Intelligent AI Connectivity
SLA based WAN



Need

AI in production w/ new AI techniques that are further distributed and need stricter performance

Best effort WAN no longer good-enough - shift to SLA based WAN

Solution

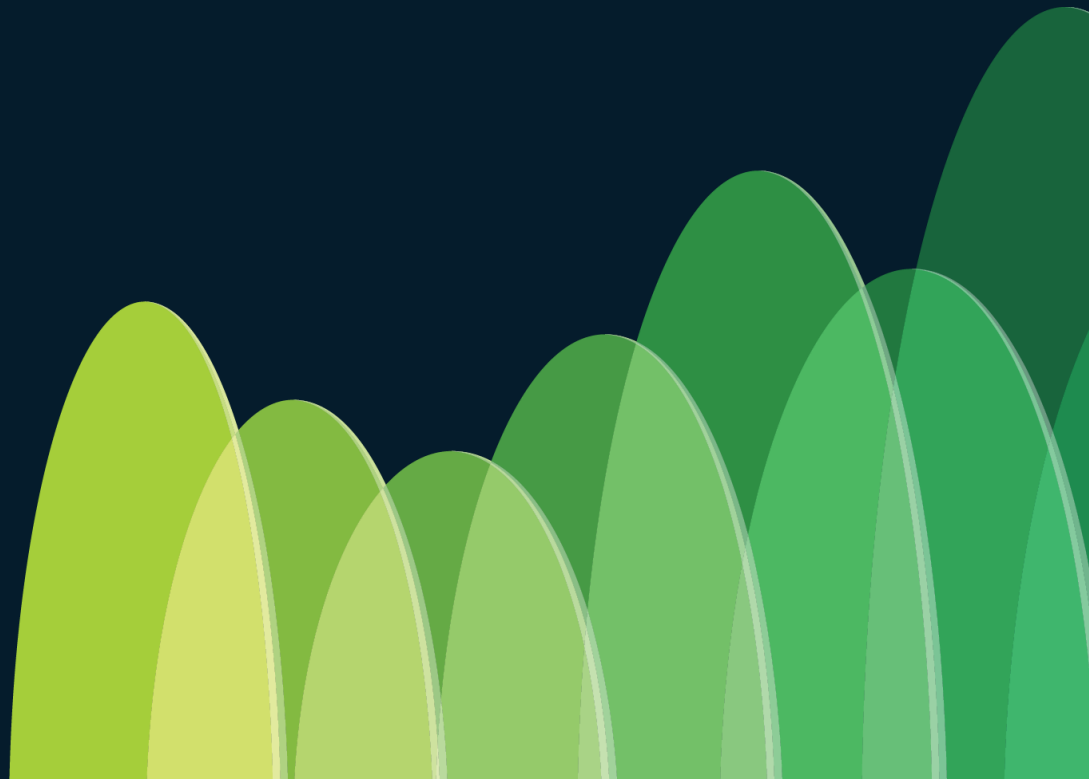
SDWAN overlay integrated with SR underlay

- Only way to guarantee SDWAN SLA intent is enforced end2end
- SDWAN service SLA is mapped to SR slice, QoS and security policy
- Uses DSCP outer packet marking or SRv6 uSIDs from the CPE to determine transport policy

PCA collects network and SDWAN controller data for an integrated overlay/underlay service view – generates insights for closed-loop assurance

Monitor AI transport KPIs and LLM KPIs

Foundational Capabilities



AI Connectivity Needs

Resiliency

Uninterrupted access to the AI App is critical for user adoption and business relevance

Mitigate the impact of failures, errors, congestion, security attacks

Security

Tampered data is useless data – Protect data in transit

Secure the network components, the links, the paths

Control and account for data paths

Visibility

Baseline for automation and assurance

Understand the network and the traffic for effective QoS, policies, traffic engineering choices

Improves resiliency, availability and security

Performance

High BW with a twist – higher peak to average ratio, symmetric BW

Low Latency considering agentic AI multiplying factor

Scalable any-to-any connectivity

Resiliency/Availability

Components

Optics – high quality optics is a must - highest single component failure in an AI cluster severely impacting cluster efficiency

Nodes – balance built-in HA with blast radius

Redundant links, redundant nodes

Architecture

SR HA features - TI-LFA, microloop avoidance, ECMP/UCMP, AnyCast SID

Multiple forwarding planes - SR-TE, SR FlexAlgo

SR disjoint paths, fast failure detection SRv6 IPM

QoS on every link to enforce priorities and quotas

Observability & Automation

Continuous and sub-sec granular visibility of network performance as input to closed loop assurance to predict failures, detect degradation and act on it

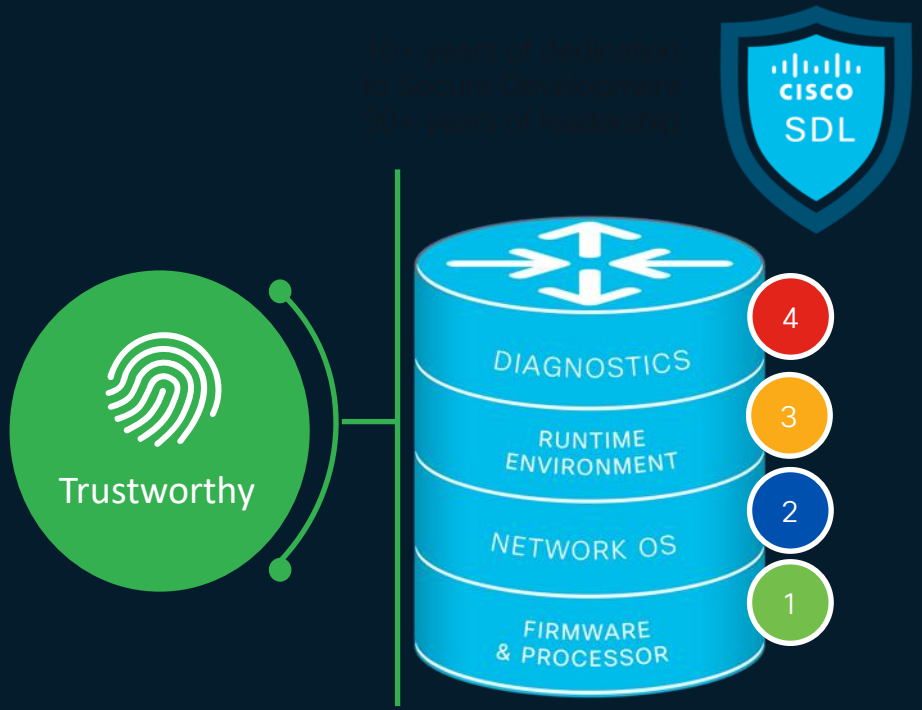
HW/SW sensors, SR-PM, SRv6 IPM, HW/SW telemetry,

Network Simplification

Less technology layers and protocols – Agile Service Networking

- SR unified forwarding plane w/ BGP unified service plane
- Routed Optical Networking converged layers
- SR-TE/FlexAlgo

Security Built-in
Cisco's Trustworthy Stack – Network Element



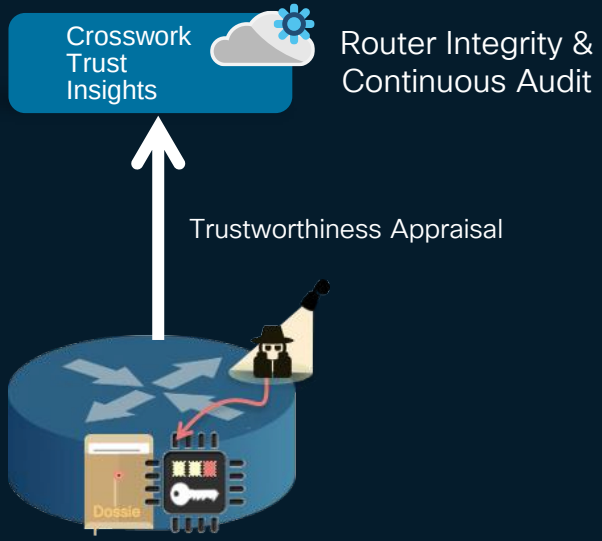
1 Trust Begins in Hardware
Tamper-proof Trust Anchor as Root of Trust

2 Verifying Trust in the Network OS
Image signing and Secure Boot infrastructure

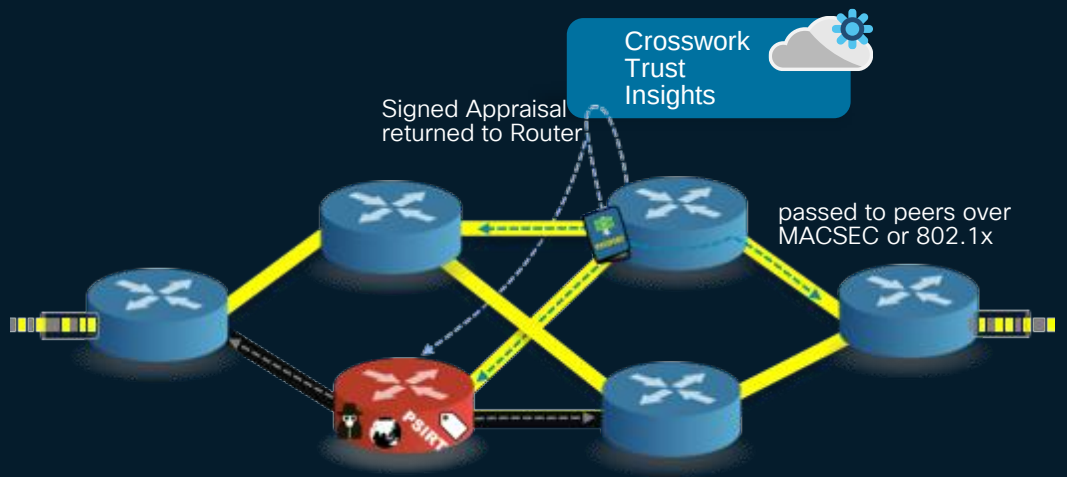
3 Maintaining Trust at Run time
Run-time defense & integrity measurements

4 Visualize and Report on Trust
Audit production network with cryptographically secured data collection

Secure Networking
Trusted Path – Building on Secure Nodes and Secure Links



Remote Attestation



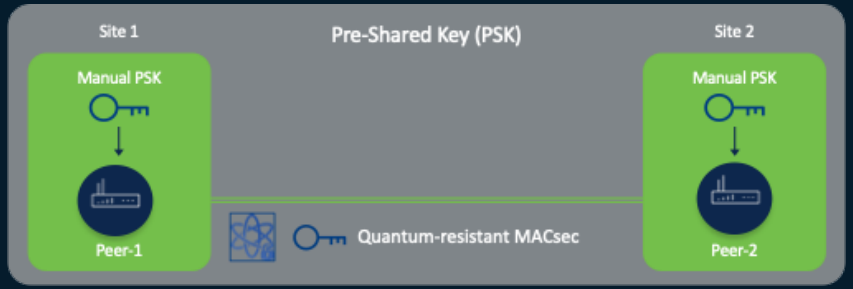
Trusted Path Routing

draft-voit-rats-trustworthy-path-routing

Secure Networking
Quantum-safe MACsec

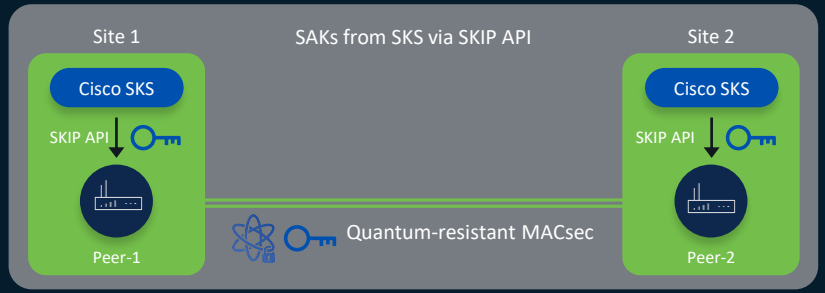
Symmetric crypto with pre-shared-key based session keys is Quantum-resistant
MacSec is quantum-safe but requires Quantum safe key distribution

Manual PPK



- MACsec with PSK option is already used
- No need for hardware or software upgrade
- Quantum-safe as this is based on symmetric cryptography

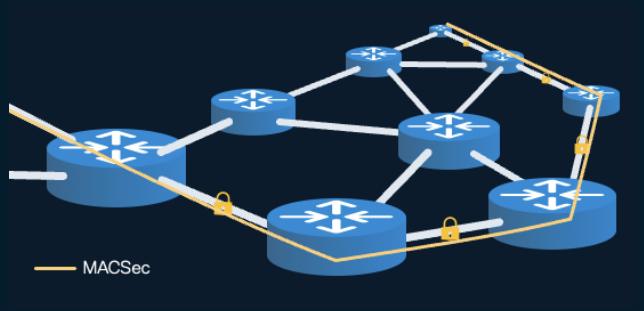
Dynamic PPK



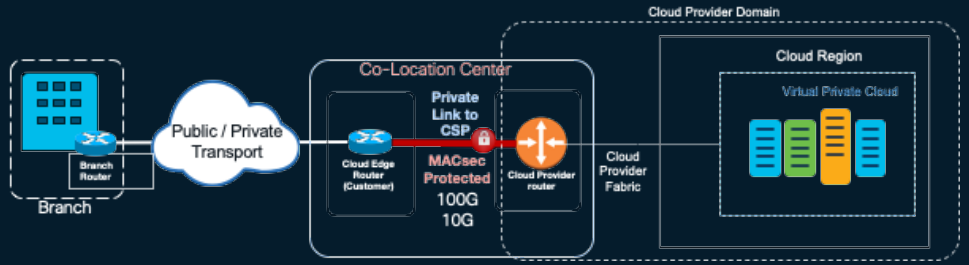
- Software-based key source
- No dedicated circuit or distance limitations
- No additional hardware requirement
- No additional cost

Secure Networking
Encrypted Path with MacSec

Quantum-safe Path



Secure High-Speed Private Links to the Cloud



Path control over secure links and within defined boundaries

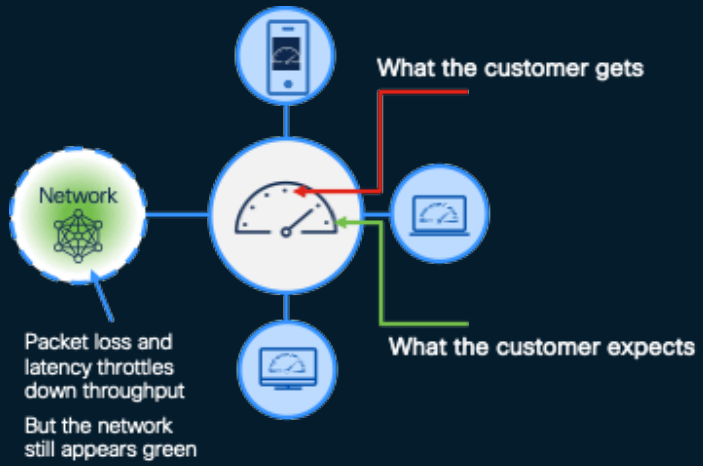
- Link affinity indicating MacSec
- Routing metric MacSec
- SR-TE/FlexAlgo
- Observability of the path for assurance

MACsec Protection

- dedicated BW and encryption over private link
- Leverage enhancements provided by WAN MacSec - Extended packet numbers for Secure Association Key (SAK)

Visibility
Understand the Network

Accurate network visibility is foundational to effectively detect and understand the source of network problems



0.53% packet loss → 50% decrease in data throughput*

5 msec delay → 10% decrease in data throughput*

10 msec jitter → 10% decrease in data throughput*

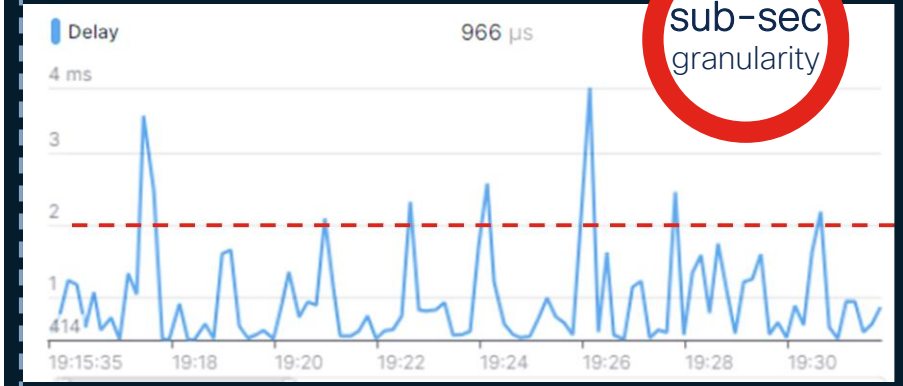
Data ingest, data replication, data transfer for AI training and production AI inference, drive high peaks of BW usage

Packet loss, error rate, deep buffers have implications in the effective network latency

*Source: Tier-1 provider

Visibility
Understand the Network

Perception



Reality

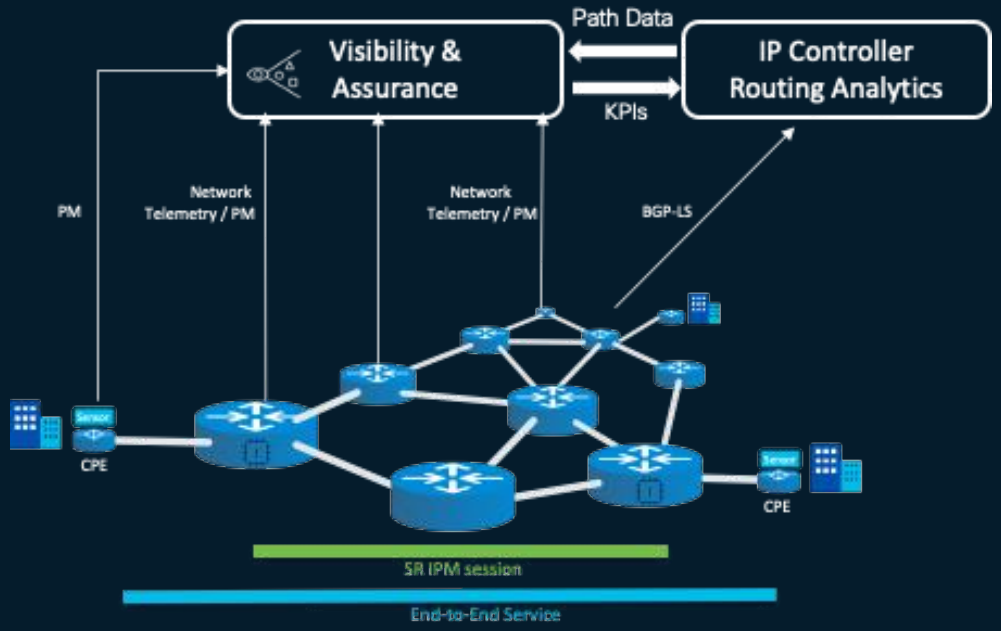
Visibility at Scale
Foundational to Automation and Assurance

Correlate customer service experience with the actual traffic path and its characteristics in real-time

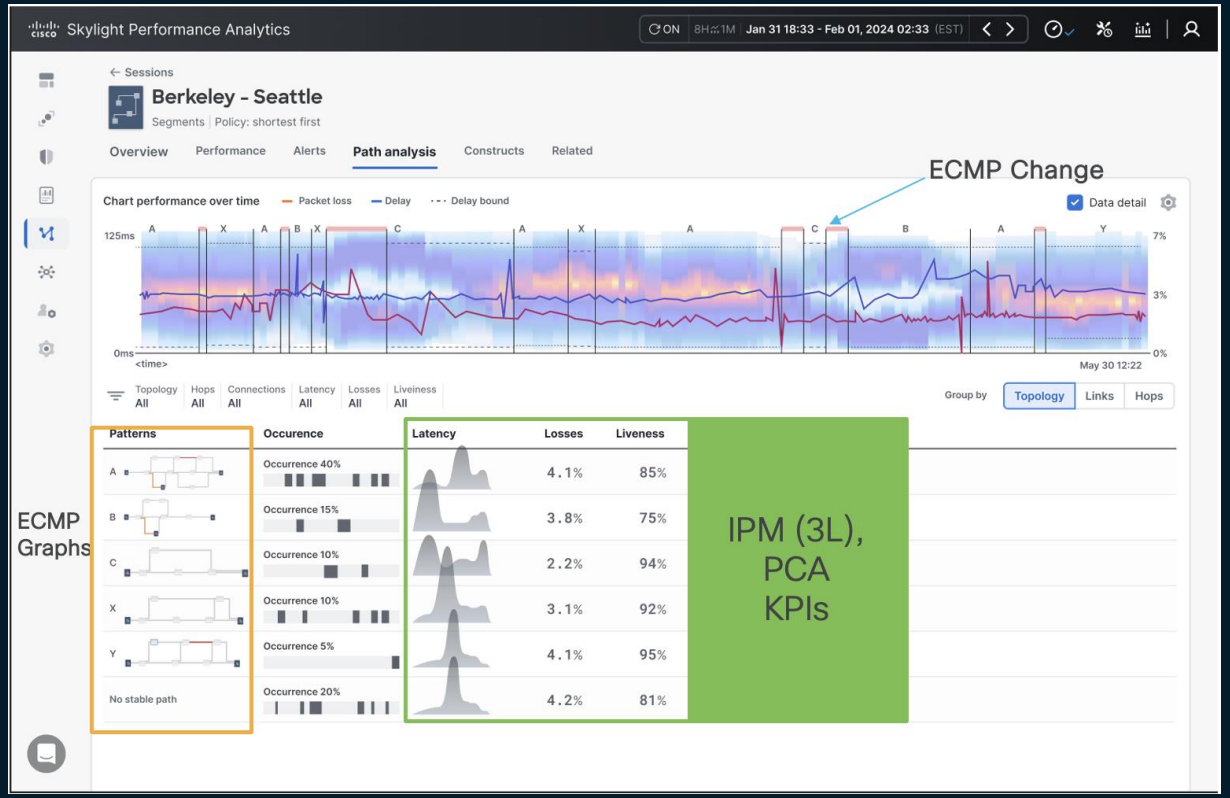
Insights for closed-loop assurance – network health and service health

Understand the network behaviour to effectively determine QoS, TE and policies best practices

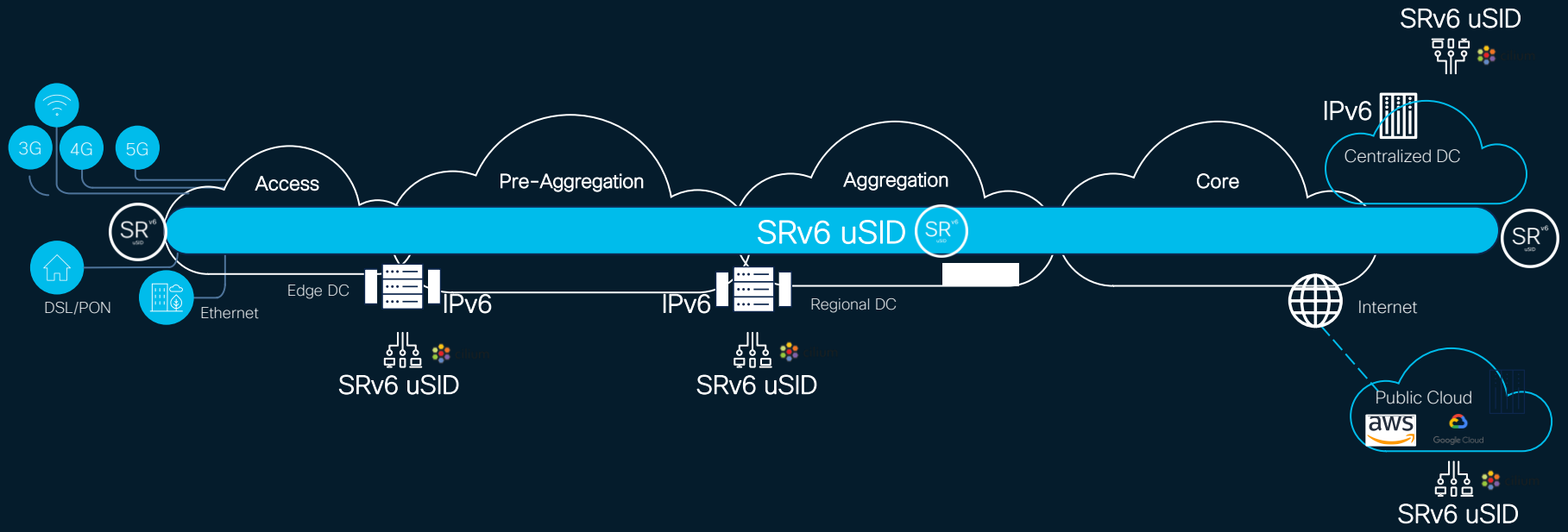
Granular visibility requires millions of probes at sub-msec speed - scale required only possible with silicon support
- Measure all ECMP paths



Visibility at Scale
Drilling Down to the Details



SRv6 uSID
Universal IP Solution: Any Service Anywhere



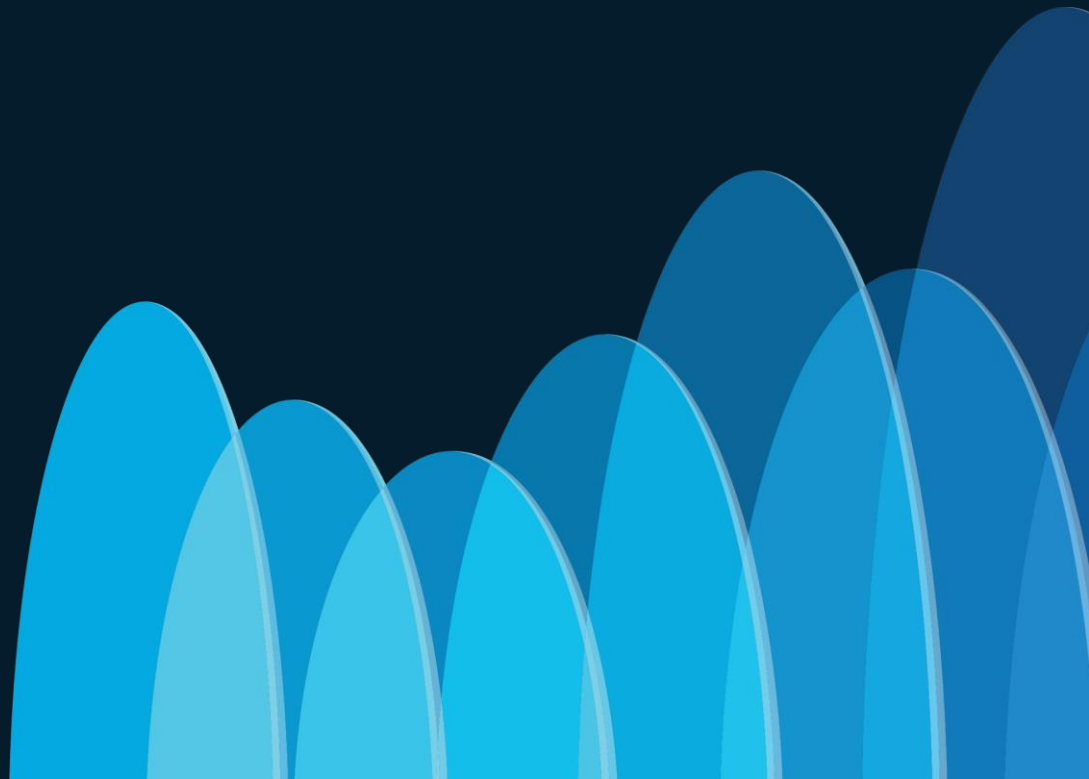
Any Service over IP without any shims

Unified Solution with Better Reliability

Seamless Brownfield Deployment

Native Host and Cloud

Building Differentiation

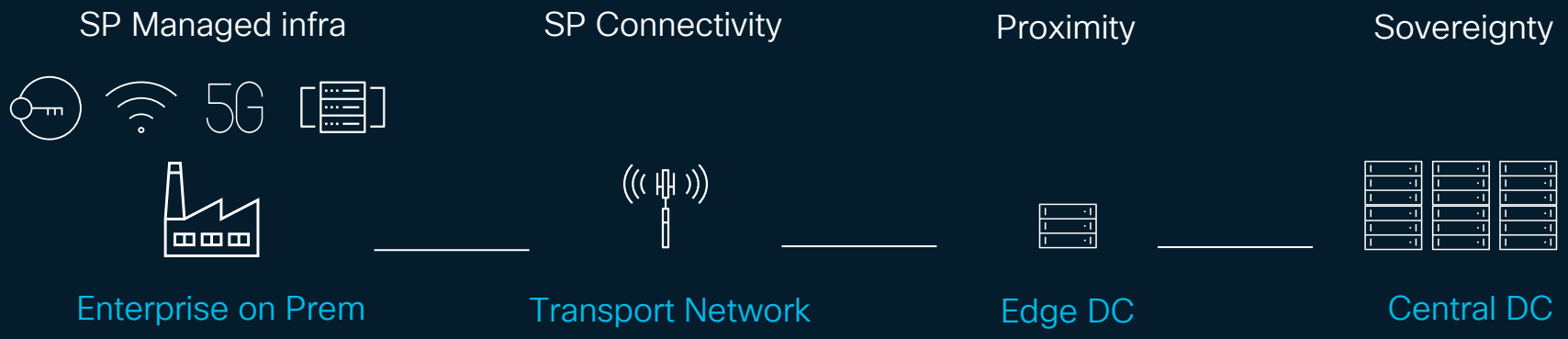


cisco Live!

Cisco Innovations - Incubation

- AI visibility – understand AI traffic trends vs non-AI traffic even for encrypted traffic
- LLM Performance and AI Network Performance:
  - Collect LLM performance KPIs – Time to first token (TTFT), Tokens per second, TPOT (time per output token), total latency, requests/sec, input tokens/sec, output tokens/sec
  - Correlate LLM and network performance
- Intelligent LLM routing at the network edge in a multi-LLM world - Ability for an application to select the best LLM based on diverse criteria, such as, cost, network and LLM performance, security/sovereignty, etc

SP unique value for AI infrastructure delivery



Trust

Conclusion

- AI is a massive consumer, producer and promoter of data – the new era of AI is driving a flood of new traffic to transport networks
- Good news is the foundational network capabilities to tackle the AI wave are already available, positioning you to be fully prepared for what lies ahead
- Service Providers (SPs) are at the center of it all, with numerous opportunities to differentiate, create value, and play a pivotal role in the AI ecosystem

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’

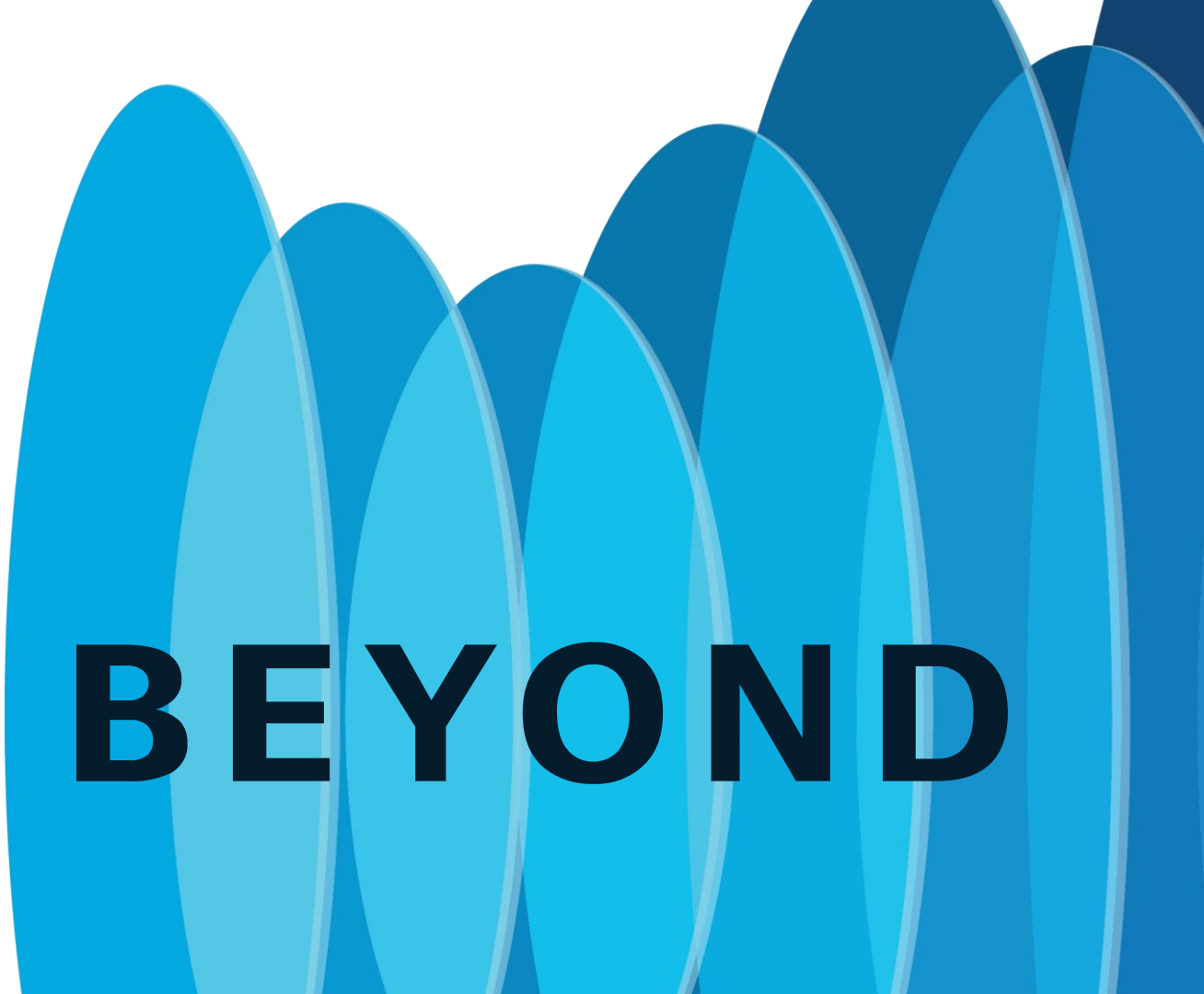
Content Catalog

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.


cisco
Thank you
cisco Live!

cisco Live!



GO BEYOND