
NTS
RELAX,
WE CARE

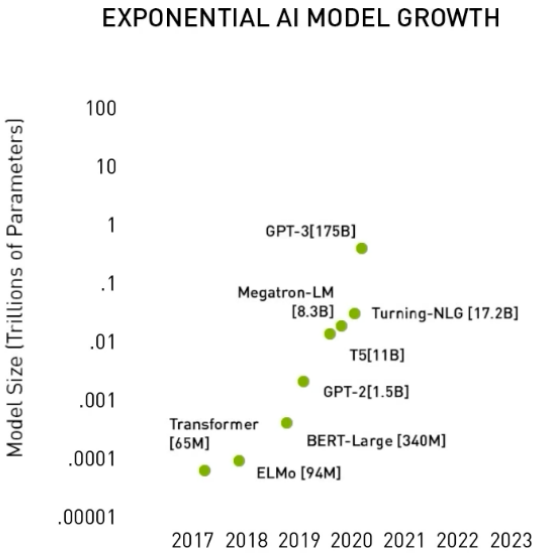
NTS

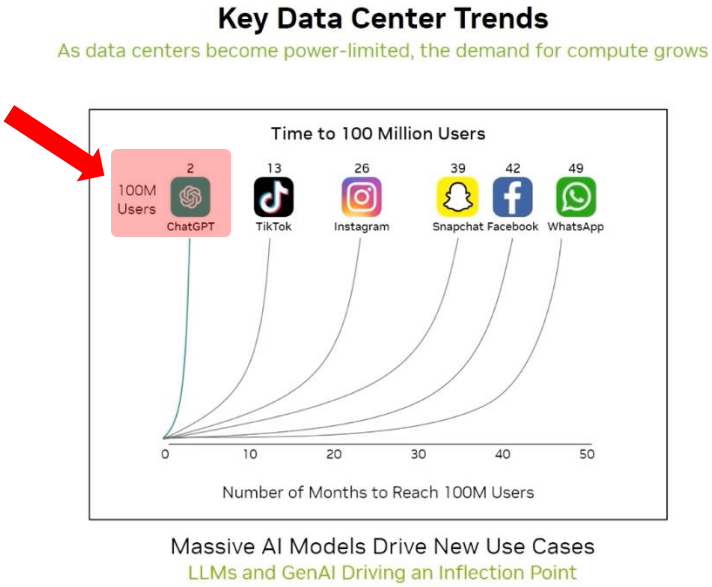
AI MARKET PENETRATION

COMMON USE OF AI

NTS

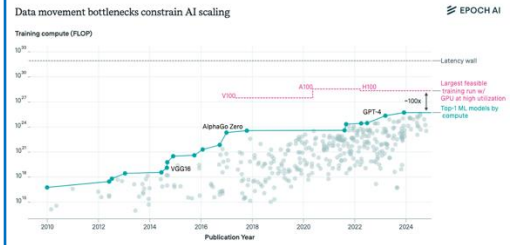
"The accelerating adoption of Artificial Intelligence (AI) is a once-in-a-generation technology shift that is impacting almost every area of business and daily life."

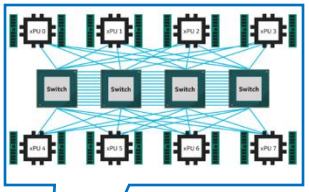


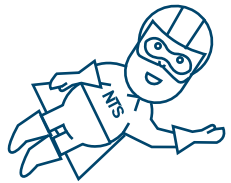


AI WORKLOAD

The time taken for data movement begins to dominate the time taken for arithmetic, precluding further scaling of training runs with present technology.



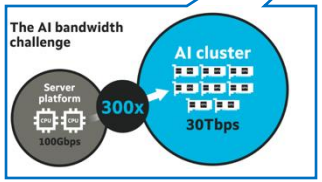


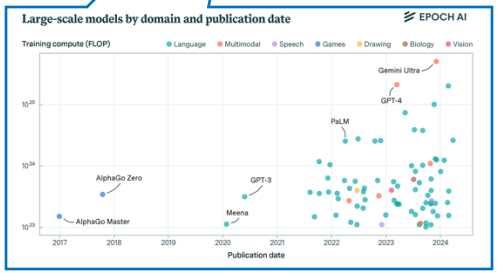


A **minimum bandwidth of 800G per connectivity**, is in today's landscape a must to provide to AI platforms with data at a rate that matches their consumption capabilities for achieving optimal performance.

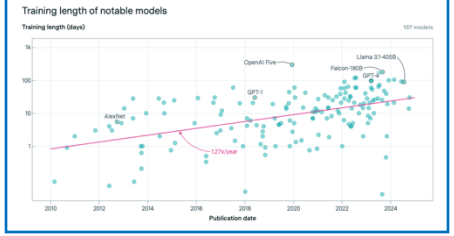
Data Preparation	Model Training	Optimization	Inference / Deployment
Data preparation includes cleansing, feature augmentation, standardization and de-duplication	Train the selected model using the training data Learn patterns and relationships within labeled datasets	Fine-tune and enhance the model performance Adjust hyperparameters, such as learning rates and batch sizes	The trained model is deployed to perform its job in a process known as Inference or Predictions Generates output based on new, unseen data to solve a domain-specific problem

AI platforms are designed by clustering purpose-built servers specifically tailored to handle the training of workloads that involve processing massive amounts of data.



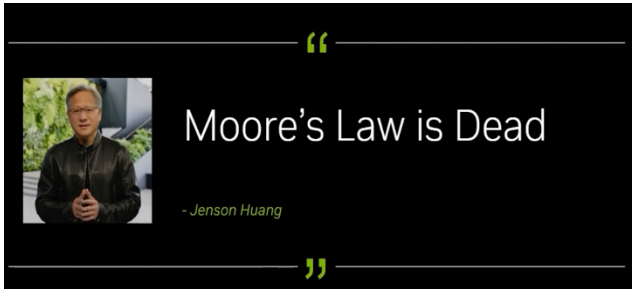


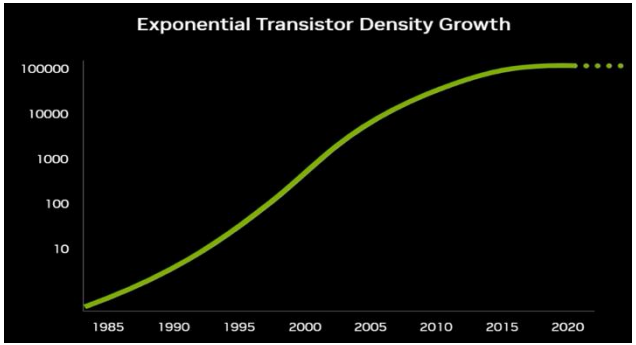
Training runs rarely exceed 6 months, because both the HW and SW used for a training run, risks becoming obsolete at timescales longer than this.



- https://epoch.ai

MOORE'S LAW IS DEAD!!!





CPU

Optimized for serial tasks

Strengths	Weaknesses
Very fast clock speeds	Relatively low memory bandwidth
Very large main memory	Low performance/watt
Small number of threads can run very quickly	Cache misses very costly
Latency optimized via large caches	

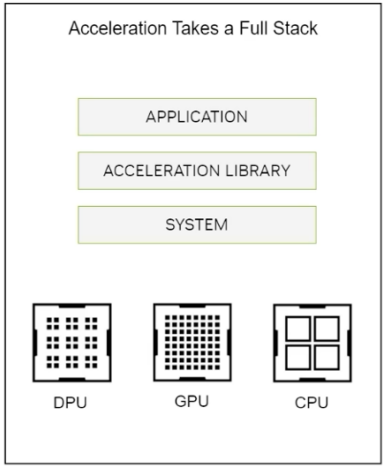
VS

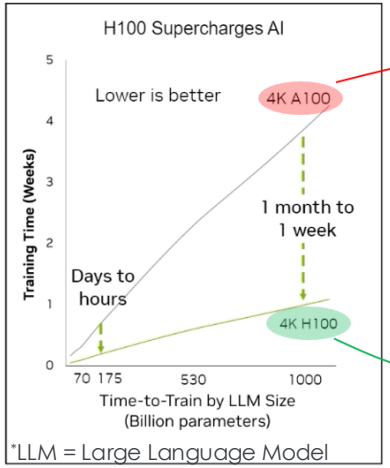
GPU

Optimized for parallel tasks

Strengths	Weaknesses
Significantly more compute resources	Low per-thread performance
Latency-tolerant via parallelism	Relatively low memory capacity
High performance/watt	
High bandwidth main memory	
High throughput	

ACCELERATED COMPUTING IS THE PATH FORWARD





NVVIDIA H100		vs A100
FP8	4,000 TFLOPS	6x
FP16	2,000 TFLOPS	3x
TF32	1,000 TFLOPS	3x
FP64	60 TFLOPS	3x
HBM3	3 TB/s	1.5X
PCI Gen5	128 GB/s	2x
4TH Gen NVLink	900 GB/s	1.5X
TDP	700W (SXM)	

"FLOP" = Floating Point Operations Per Second

Unità	Simbolo	Valore numerico	Scala corta (USA, UK moderno)	Scala lunga (Italia, Europa)
Million	10^6	1.000.000	Milione	Milione
Billion	10^9	1.000.000.000	Miliardo	Miliardo
Trillion	10^{12}	1.000.000.000.000	Bilione	Mille miliardi
Quadrillion	10^{15}	1.000.000.000.000.000	Trilione	Milione di miliardi
Quintillion	10^{18}	1.000.000.000.000.000.000	Quadrilione	Miliardo di miliardi (Trilione)
Tera	10^{12}	1.000.000.000.000	Mille miliardi (**Bilione**)	Mille miliardi (**Bilione**)

Formato	Bit totali	Precisione	Velocità	Memoria richiesta	Uso principale
FP8	8-bit	Molto bassa	(Altissima)	(Minima)	AI inferenza/training
FP16	16-bit	Bassa	(Alta)	(Bassa)	AI training/inferenza, GPU computing
TF32	32-bit*	Media	(Alta)	(Media)	AI training su GPU NVIDIA
FP32	32-bit	Standard	(Buona)	(Alta)	Gaming, AI, GPU computing
FP64	64-bit	Alta	(Lenta)	(Molto alta)	HPC, calcoli scientifici

- In modern data center, maintaining **high performance** and **low latency** is crucial, especially with the rise of AI and machine learning applications that demand **significant computational power** and **efficient data transfer**.
- One critical aspect of achieving these goals is effective **flow control within the network**, particularly in **RoCEv2** (RDMA over Converged Ethernet) deployments.

„AI“
THE KILLER APPLICATION ON DC

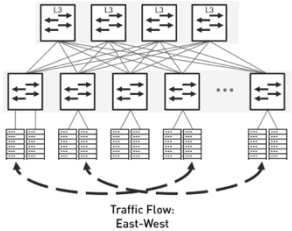
HOSTS
INTERCONNECTIONS

NTS

Modern Data Centers

Modern data center environments are becoming much more dependent on east-west traffic

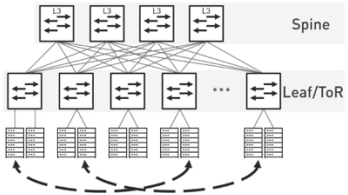
- Data center transformation is driven by new trends: cloud computing, virtualization, big data
- Traffic patterns have changed to east-west communication:
  - More data moves between servers and storage nodes
- Modern data center design requirements:
  - Increased server-to-server communication
  - Scalability
  - Resiliency



Leaf-Spine Design

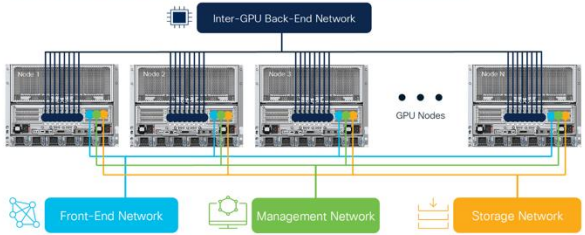
Modern data centers have transitioned to a leaf-spine design

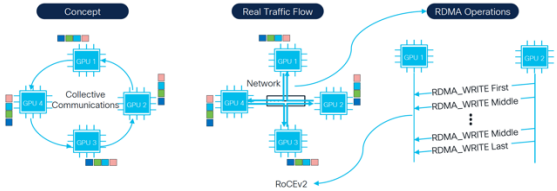
- In the past few years, many data centers have transitioned to a leaf-spine design
- Leaf-spine architecture is a multilayer, full-mesh topology composed of a leaf layer and a spine layer
- Leaf-spine design advantages:
  - Provides predictable and deterministic latency
  - Improves scalability
  - Improves redundancy
  - Increases bandwidth
  - Offers congestion avoidance



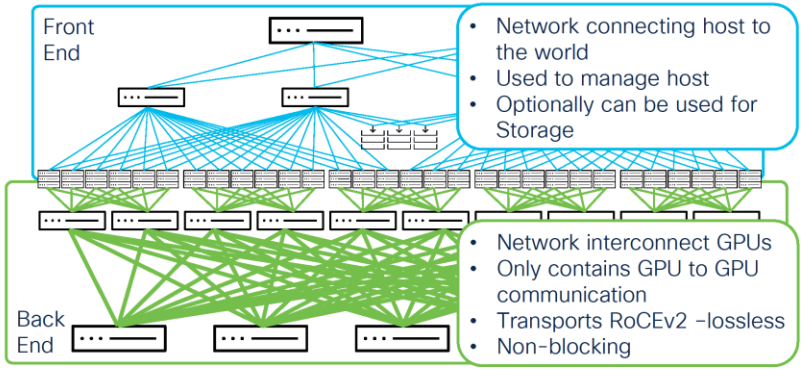
Typical 8-Way GPU Node Connections

Using Cisco UCS C885A M8 based on NVIDIA HGX architecture





Non-blocking network



SOME INSIGHTS ABOUT HW

NEW LEAF

NTS

Nexus 9300 Fixed Series for 400G & 800G fabrics

32 to 64-port 400G 25.6T
Nexus switches

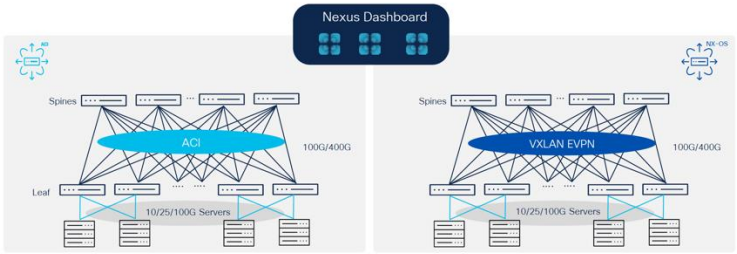
64-port 800G or 128-port 400G (breakout) 51.2T
Nexus switches with QSFP-DD or OSPF form factors

For high performing frontend and backend AI networks

Front-End Network Options

Building or repurposing existing infrastructure for AI workloads

Front-End Network

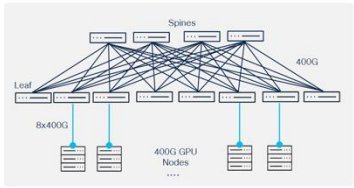


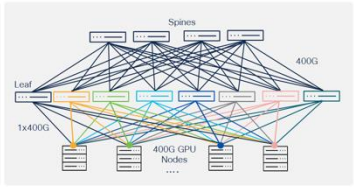
ACI Fabrics

VXLAN EVPN Fabrics

Back-End Network Options, but not limited to

Classic or Rail-Optimized Architectures with Nexus 9000 Fixed Series





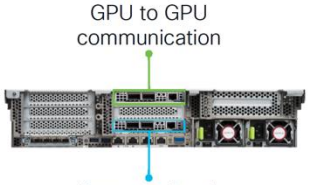
Classic Design
Hosts mapped per Rack/Switch
Non-blocking / 1:1 Oversubscription / BGP Fabric
Up to 64/128-port 400G or 800G options for Leaf/Spine

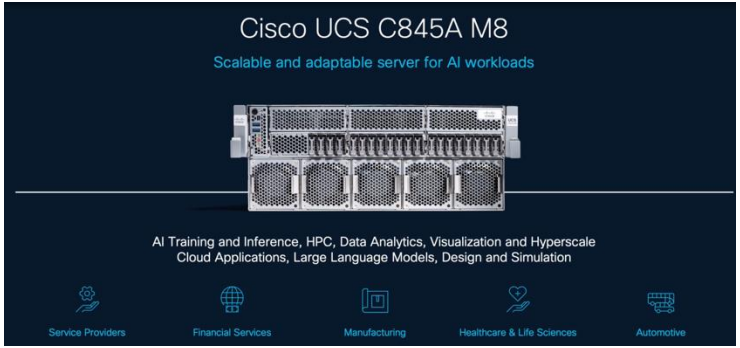
Rail-Optimized Design
8 GPU:NIC pair mapped per Rail/Switch
Non-blocking / 1:1 Oversubscription / BGP Fabric
Up to 64/128-port 400G or 800G options for Leaf/Spine

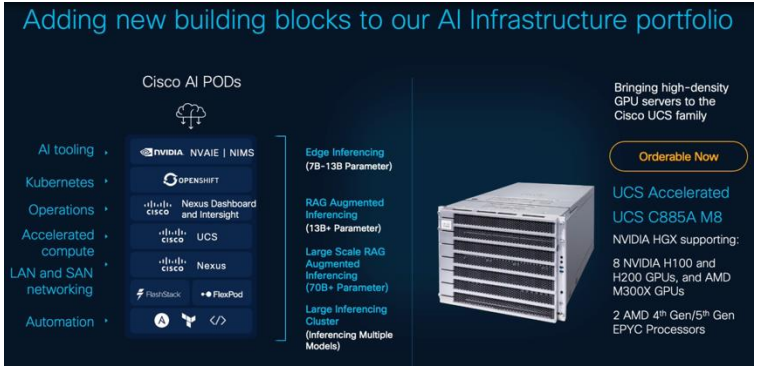
These design choices strictly depend on Collective Communication Library optimizations to improve overall performance.

UCS + NVIDIA

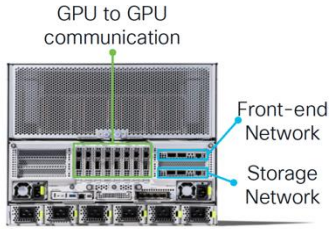
UCS C240 M7 Rack Server
Up to 2 GPUs







UCS C885A M8 Rack Server
8 GPUs



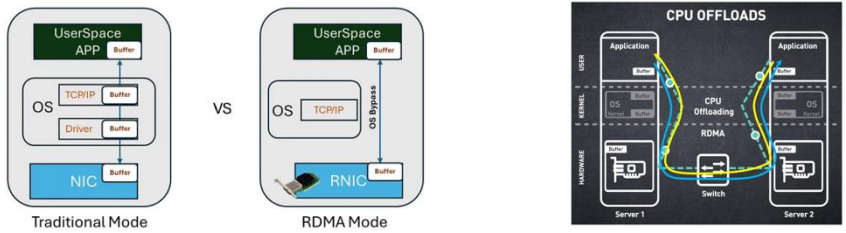
- https://www.cisco.com/site/us/en/products/security/hypershield/index.html

DATA CENTER QUANTIZED CONGESTION NOTIFICATION (DCQCN)

RDMA

NTS

- The advent of ***Remote Direct Memory Access (RDMA) over converged ethernet*** on a hyper scale data center challenges DC network operator to provide a low latency, lossless fabric at par with *InfiniBand* fabrics.
- *Direct Memory Access (DMA)* is the ability of a device to access host memory directly, without the intervention of the CPU. *Remote Direct Memory Access (RDMA)* is the ability to access (read and write) memory on a *remote machine* without interrupting CPU operations.



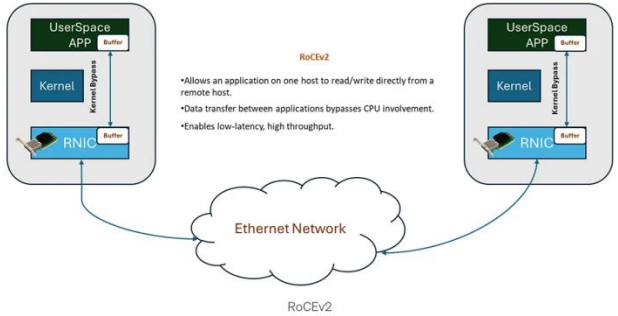
- *RDMA over Converged Ethernet* (RoCE – pronounced "Rocky") is a network protocol that allows *Remote Direct Memory Access (RDMA) over an Ethernet network*. It does this by encapsulation of an InfiniBand transport packet over Ethernet.
- There are two versions of RoCE:
  - **RoCEv1** – Ethernet link layer protocol (Ethertype 0x8915) allows communication between two hosts in the same Ethernet broadcast domain → **layer-2 only**

  - **RoCEv2** – Enhances RoCEv1 with a UDP/IP (IPv4 or IPv6) header and adds **layer-3 routability**.

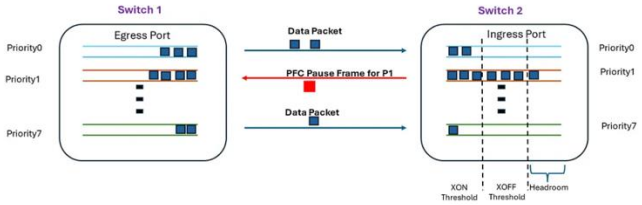
- The **UDP** destination **port 4791** indicates that the payload is an InfiniBand payload with InfiniBand Base Transport header. RoCEv2 can use different source UDP ports for different QPs that help ECMP load sharing. The lossless nature is achieved using ***Priority Flow Control (PFC)*** and ***Explicit Congestion Notification (ECN)*** on a lossy IP fabric.

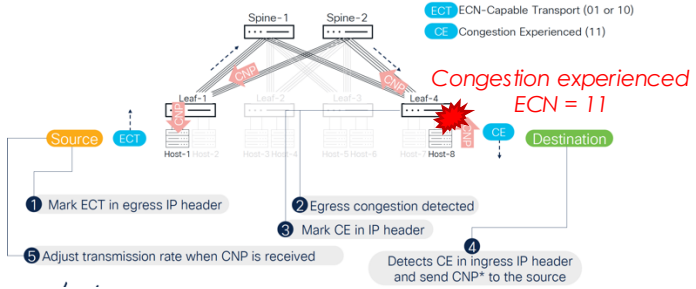
DCQCN

- To utilize RoCEv2, specialized RDMA NICs (RNICs) capable of RDMA are required on both the source and destination hosts. The physical (PHY) speed of RDMA cards typically starts from 50Gbps and is currently available at speeds up to 400Gbps.



- A minimum MTU of 9000 bytes is typically recommended to avoid fragmentation and ensure efficient data transfer.





Priority Flow Control + *Explicit Congestion Notification*

Data Center Quantized Congestion Notification (DCQCN)

CONFIGS...

ROCE ON NEXUS 9000 FAMILY

- Cisco Nexus 9000 Series switches in NX-OS mode support the transport of RDMA over Converged Ethernet (RoCE) V.1 and V.2 protocols natively.
- It does not require a specific license.

N9K config example:

- During a read/write operation between an initiator and target, RoCE traffic is supposed to be marked with DSCP = 24
- CNP being control traffic is marked with DSCP = 48 (CS = 6)

(*)Nexus use 8 queus for traffic scheduling; qos-group is mapped directly to queue number

Class Map Queue Name	Description	Class Map Queue Name	Description
c-in-q-default	Ingress default queue – QoS group 0	c-out-8q-q-default	Egress default queue – QoS group 0
c-in-q1	Ingress queue 1 – QoS group 1	c-out-8q-q1	Egress queue 1 – QoS group 1
c-in-q2	Ingress queue 2 – QoS group 2	c-out-8q-q2	Egress queue 2 – QoS group 2
c-in-q3	Ingress queue 3 – QoS group 3	c-out-8q-q3	Egress queue 3 – QoS group 3
c-in-q4	Ingress queue 4 – QoS group 4	c-out-8q-q4	Egress queue 4 – QoS group 4
c-in-q5	Ingress queue 5 – QoS group 5	c-out-8q-q5	Egress queue 5 – QoS group 5
c-in-q6	Ingress queue 6 – QoS group 6	c-out-8q-q6	Egress queue 6 – QoS group 6
c-in-q7	Ingress queue 7 – QoS group 7	c-out-8q-q7	Egress queue 7 – QoS group 7 (Strict Priority)

```
class-map type qos match-any CNP
  match dscp 48
class-map type qos match-any ROCEv2
  match dscp 24
!
```

✓ RoCEv2 and CNP traffic classification

```
policy-map type qos QOS_MARKING
  class ROCEv2
    set qos-group 3
  class CNP
    set qos-group 7
  class class-default
    set qos-group 0
!
```

✓ RoCE is classified to queue 3(*)

✓ CNP is classified to priority queue which is queue 7(*)

```
interface ethernet ...
  priority-flow-control mode on
  service-policy type qos input QOS_MARKING
!
```

✓ PFC must be enabled on all the interfaces throughout the network fabric

✓ RoCE is set to qos-group 3, which maps to c-out-8q-q3 queue

✓ CNP is set to qos-group 7, which maps to strict-priority (real time) queue

```
interface nve1
  service-policy type qos input QOS_MARKING
!
```

✓ RoCEv2 classification is required on VTEP interface to classify the decapsulating traffic. VTEP nodes, by default in uniform mode, copies the QoS DSCP values from the inner IP header to the outer VXLAN header during encapsulation and vice versa during decapsulation

```
policy-map type network-qos QOS_NETWORK
  class type network-qos c-out-8q-q3
    pause pfc-cos 3
    mtu 9216
    ...
!
```

✓ PFC pause frames are generated on queue 3; MTU must match the RoCEv2 traffic MTU and pfc-cos *value* must match qos-group *value*

```
policy-map type queuing QOS_EGRESS_PORT
  class type queuing c-out-8q-q7
    priority level 1
  ...
  class type queuing c-out-8q-q3
    bandwidth remaining percent 80
    random-detect minimum-threshold 150 Kbytes maximum-threshold 3000 kbytes drop-probability 7 weight 0 ecn
  ...
  class type queuing c-out-8q-q1
    bandwidth remaining percent ...
!
```

✓ Prioritize CNP

✓ ECN is enabled on queue 3 with a minimum threshold of 150 Kbytes and a maximum threshold of 3000 Kbytes; if the packet threshold is below the minimum of buffer (150 Kbytes) then transmits without ECN marking, if it is between minimum and maximum then randomly marks the packets (with probability 7) with ECN bits, if it is above maximum of buffer (3000 Kbytes) then marks every traffic with ECN (0x11).

```
system qos
  service-policy type queuing output QOS_EGRESS_PORT
  service-policy type network-qos QOS_NETWORK
```

✓ The network QoS config and the QoS queueing/scheduling configuration must be applied on the system level to apply for all the interfaces in the system

ECMP

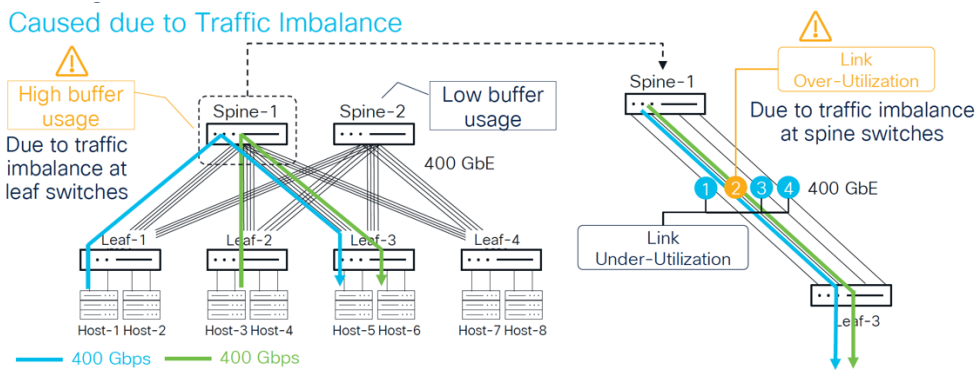
ECMP (DEFAULT BEHAVIOUR)

NTS

- Default ECMP Load-Sharing based on 5-tuples: Source IP, Destination IP, UDP, Source Port, Destination Port (4791)
  - Layer 4 protocol (UDP), and destination port (4791) remain the same for RoCEv2
  - # of IP addresses are limited (GPU's IP address)
  - only UDP source port changes

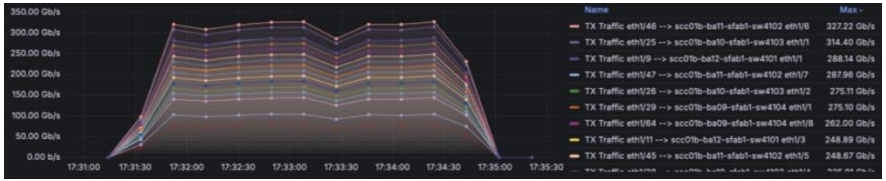
  Low Entropy

  → Traffic is polarized to Spine-1 only



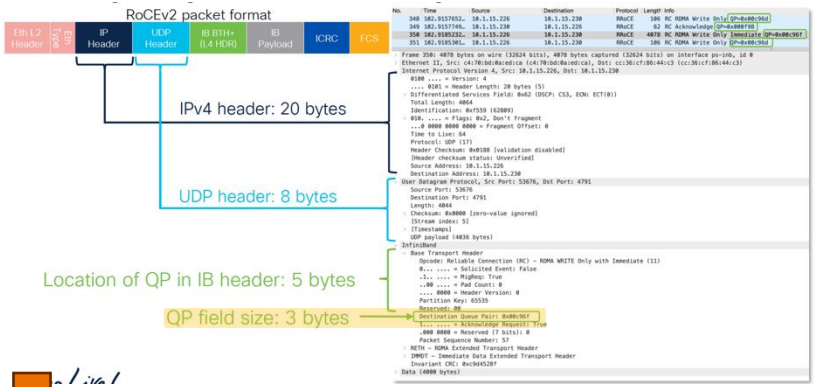
- Default ECMP load-balancing is unaware of real-time link utilization

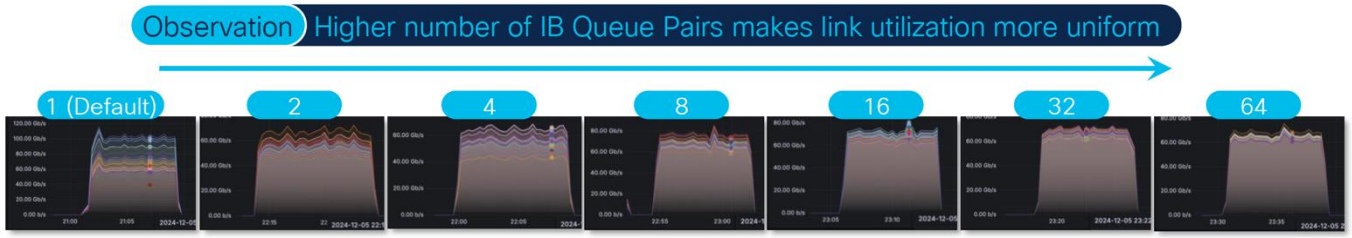
Behavior	Typical Data Center Network	Inter-GPU network
# of Layer-4 flows	Many	Fewer
Per-flow throughput	Many mouse flow and a few elephant flows	Mostly elephant flows
Link utilization	Not fully utilized	GPU send and receive traffic at the full capacity of their NIC ports



ECMP UDF ECMP ALGORITHM

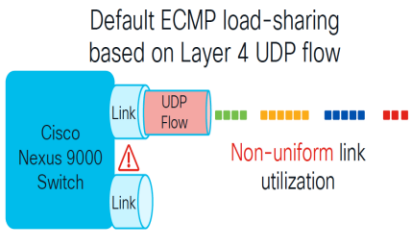
- UDF ECMP looks at source and destination IP and User Defined Field in a packet
- User can choose what field to look at to enhance entropy
- Every RoCEv2 conversation is identified by *Destinate Queue Pair* (3 byte field), in IB Header
  - Destination Q Pair is 33 bytes from beginning of IP field (IP (20B) + UDP (8B) + IB (5B)) or 6th byte in InfiniBand header

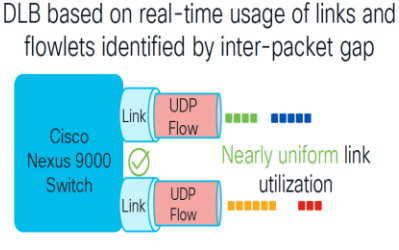


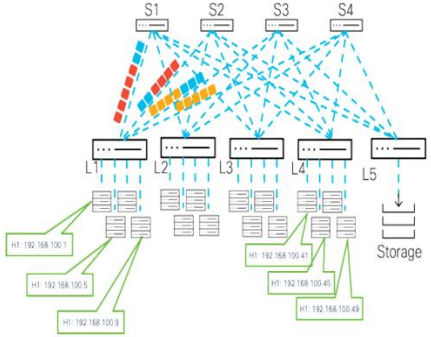


...OTHER ECMP ALGORITHMS

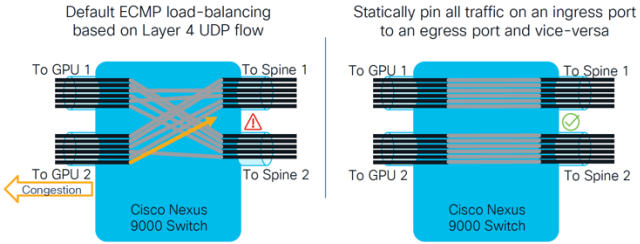
- **Flowlet** is a packet burst from a flow separated by large enough gap. Flow is identified as per ip load sharing command (ECMP or UDF):
  - *Dynamic Load Balancing* (DLB) is based on real-time usage of links and flowlets are identified by inter-packet gap
  - Flows maintain the link or dynamically change the link
  - Least utilized links are chosen, for a new flow, or at time flows are moved to different link
  - Ageing time is used avoid out of order packets, it should be 2 x RTT
  - AI workloads are balanced fairly and if flows are bursty they can dynamically change path

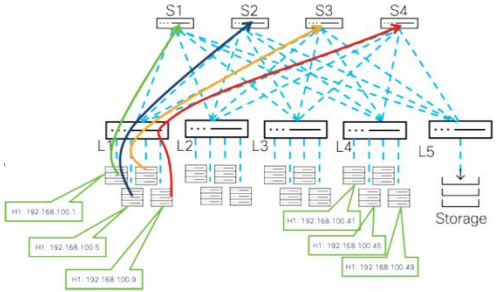


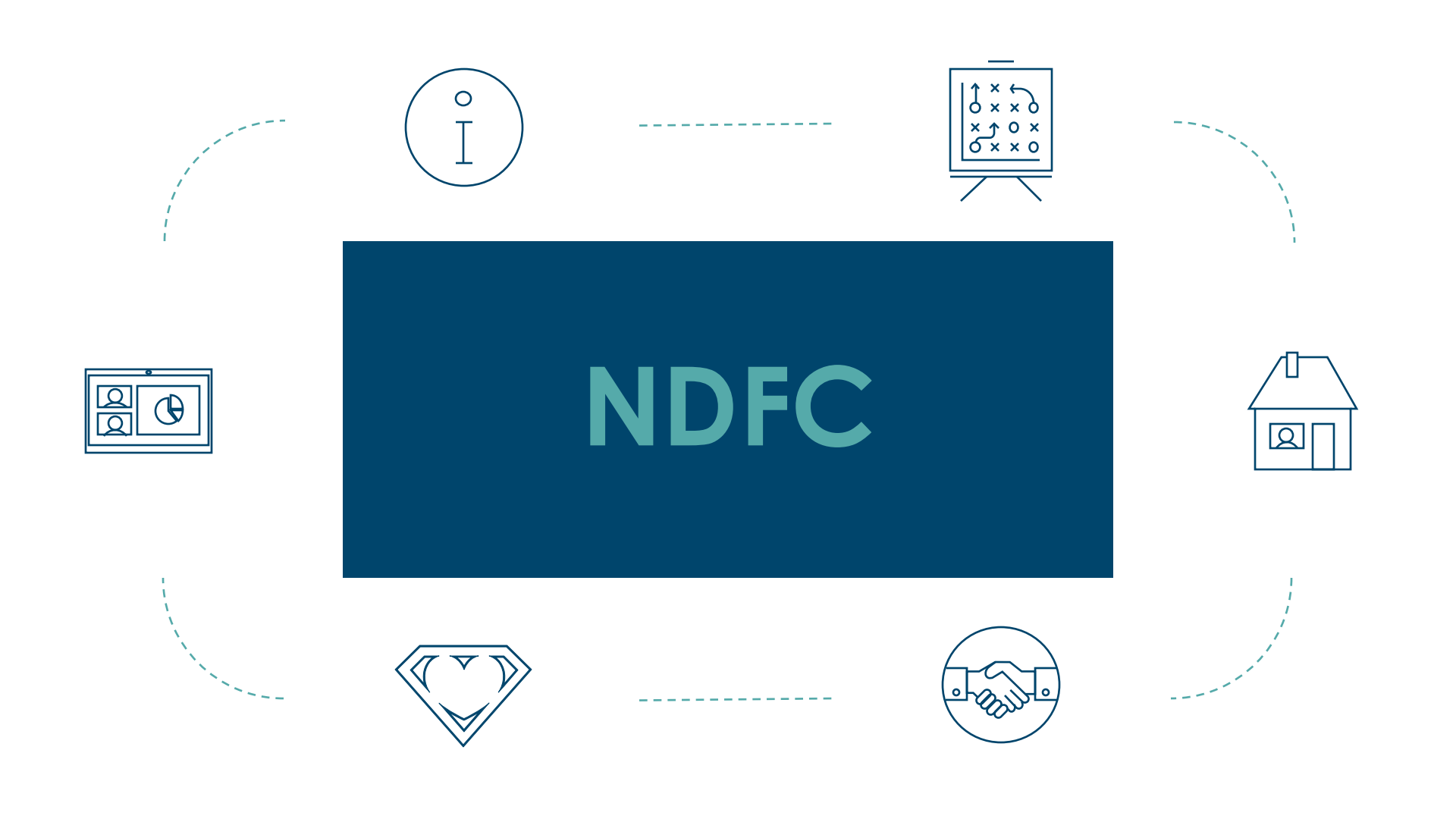




- **Static pinning**, manually configured (or via API), allows static pinning between ingress and egress ports






NDFC

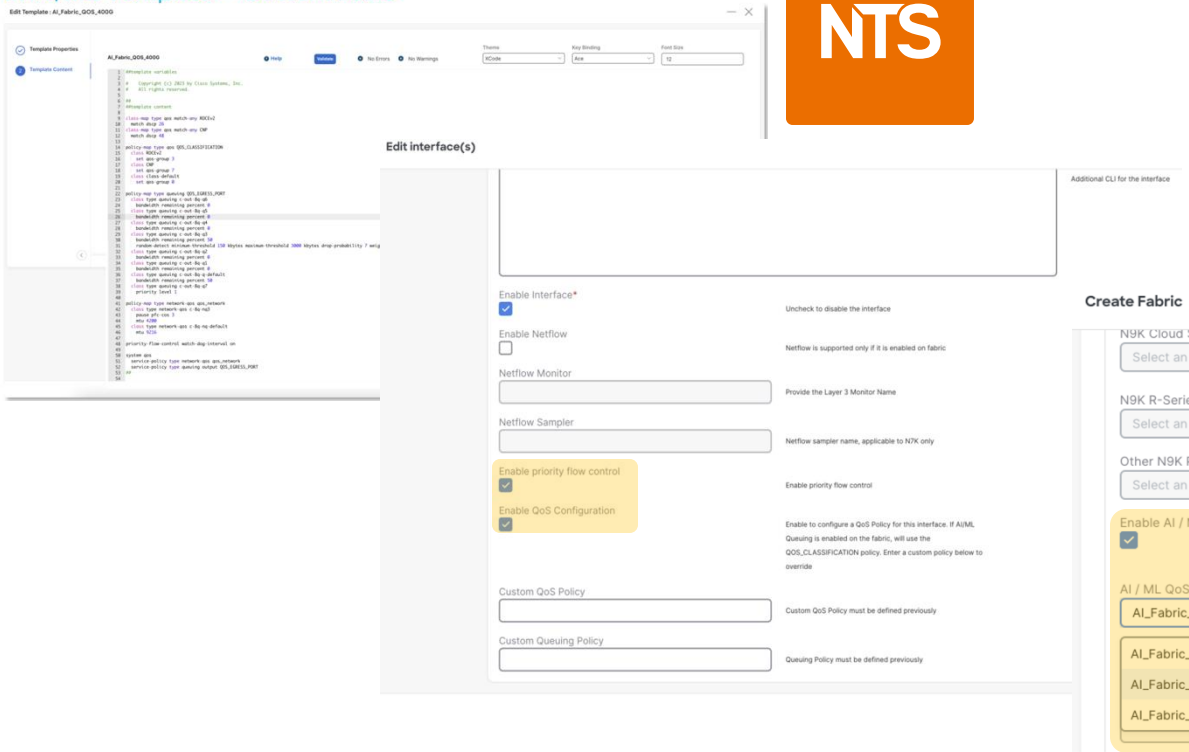
Best Practice

Use Nexus Dashboard to configure the fabric with the recommended values as per Cisco's AI/ML Networking Blueprint

Best Practice

Use the built-in templates in Nexus Dashboard for enabling PFC and ECN

Built-in AI/ML Template - Customizable



RELAX,
WE CARE

CISCO LIVE REDELIVERY- AUTOMATION

12.03.2025 | NTS Webinar

#WHOAMI

NTS

ANDREA DAINESE

- Senior Network & Security Architect con 20+ years' experience in securing complex IT and OT infrastructures
- Focused on cyber security strategies, GDPR/ISO 27001 compliance and Automation
- Cisco (CCIE), VMware, Red Hat... certified
- Father of Unified Networking Lab (UNetLab)
- Privacy and digital security evangelist

https://adainese.it
https://www.linkedin.com/in/adainese/
https://www.patreon.com/adainese

SPEAKING ABOUT

DEVNET

NTS

AGENDA

NTS

HIGHLIGHTS

1. AI & Automation
2. Infrastructure Modernization
3. Sustainability
4. Security & AI-Driven Threat Detection
5. AI & Observability – A Data-Driven Approach

DIGITAL TRANSFORMATION

NTS

1. AI & AUTOMATION

The backbone of Digital Transformation

- AI is central to Cisco's vision, driving efficiency in networking, cybersecurity, and infrastructure management.
- AI-powered observability, security, and customer experience solutions are evolving to **proactively detect, respond, and optimize** network operations.
- Cisco is integrating **AI/ML models** into its platforms, such as **Catalyst Center and Webex**, to enhance automation and predictive analytics.

MODERNIZATION

NTS

2. INFRASTRUCTURE MODERNIZATION

- Cisco is championing **network programmability** through **Terraform, Ansible, and API-driven automation**.
- The **Cisco Catalyst Center** and **ACI** are transforming networks into **programmable, self-optimizing infrastructures**.
- **Event-driven automation** using tools like **Meraki Webhooks and ThousandEyes** ensures real-time adaptability.

DIGITAL TRANSFORMATION

NTS

ThousandEyes:

- Network visibility (performance, outage, user experience)
- Integrate with 3rd party products and get notified:
  - Email
  - 3rd party tool (servicenow, slack…)
  - Webhoocks
  - CRUD API

SUSTAINABILITY

NTS

3. SUSTAINABILITY

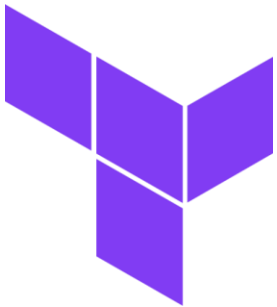
- Cisco's commitment to a **sustainable future** was highlighted through **green networking solutions**.
- **Power efficiency APIs, telemetry dashboards, and network sustainability initiatives** were showcased as key drivers of Cisco's sustainability roadmap.

SUSTAINABILITY

Sustainability:

- Energy Consumption
- Operational Cost

- Zero Touch Provisioning
- Programmability & Telemetry
- Infrastructure as Code (IaC)



THREAT DETECTION

NTS

4. SECURITY & AI-DRIVEN THREAT DETECTION

- Security threats are evolving, and Cisco is responding with **AI-powered cybersecurity solutions**.
- The integration of **Hypershield API, XDR automation, and Secure Firewall automation** enhances **threat detection, incident response, and security enforcement**.

OBSERVABILITY

NTS

5. AI & OBSERVABILITY – A DATA-DRIVEN APPROACH

- Cisco is leveraging **Splunk and ThousandEyes** to provide **full-stack observability**, ensuring holistic insights into IT infrastructure.
- AI-powered **self-healing networks** and **predictive analytics** are driving improvements in system performance.
- Observability solutions integrate with **cloud-native tools** to optimize **AI workloads**.

OBSERVABILITY

NTS

Splunk:

- Unified Observability
  full stack observability
- Application and Infrastructure Monitoring
- IT Service Intelligence
  event correlation and prevention
- On-Call
  incident correlation
- Anomaly Detection

SECURITY REDELIVERY

Securing the enterprise is increasingly challenging

Highly distributed, fine-grained apps

- Spanning data center, cloud
- Containers
- 1000s of microservices

Nothing can be trusted

- Distributed perimeter necessary but no longer sufficient
- Need security in every flow to stop lateral movement

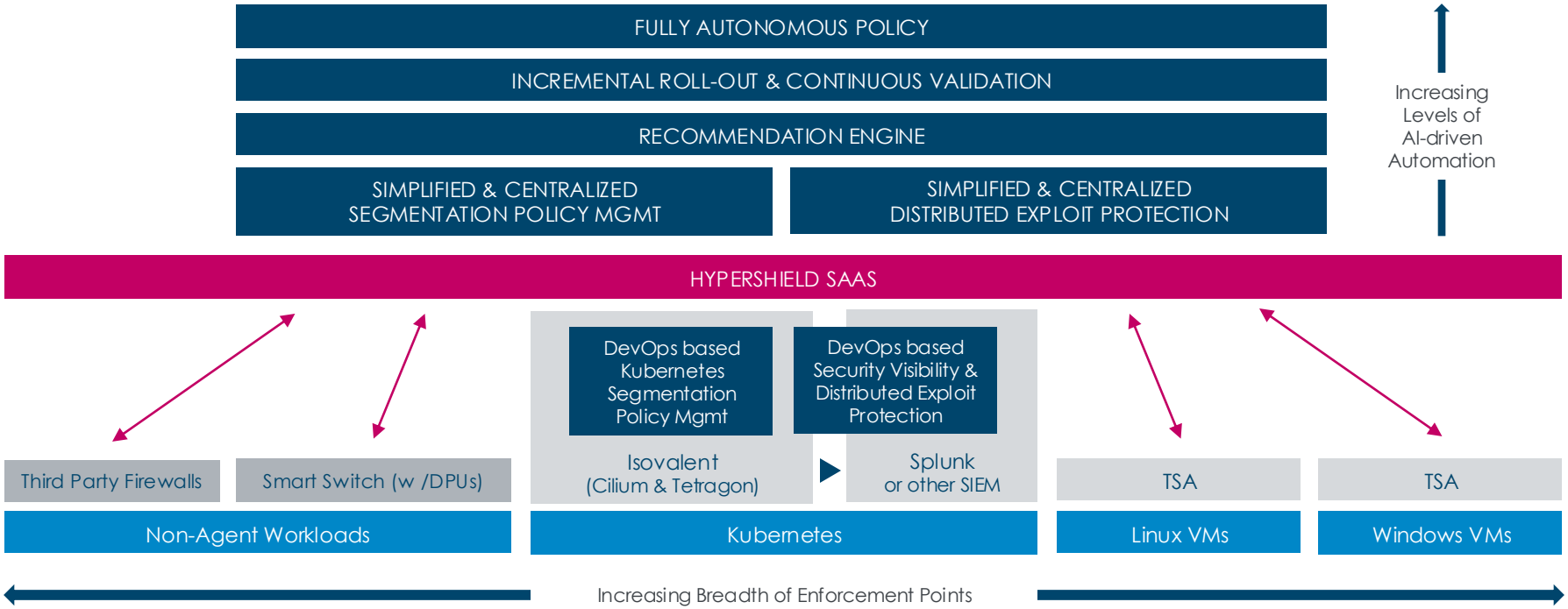
More vulnerabilities, exploited faster

- Weeks to hours to minutes
- Patching can't keep up
- New AI model risks

AI increasing attack surface and attack sophistication

CISCO HYPERSHIELD ISOVALENT ENTERPRISE

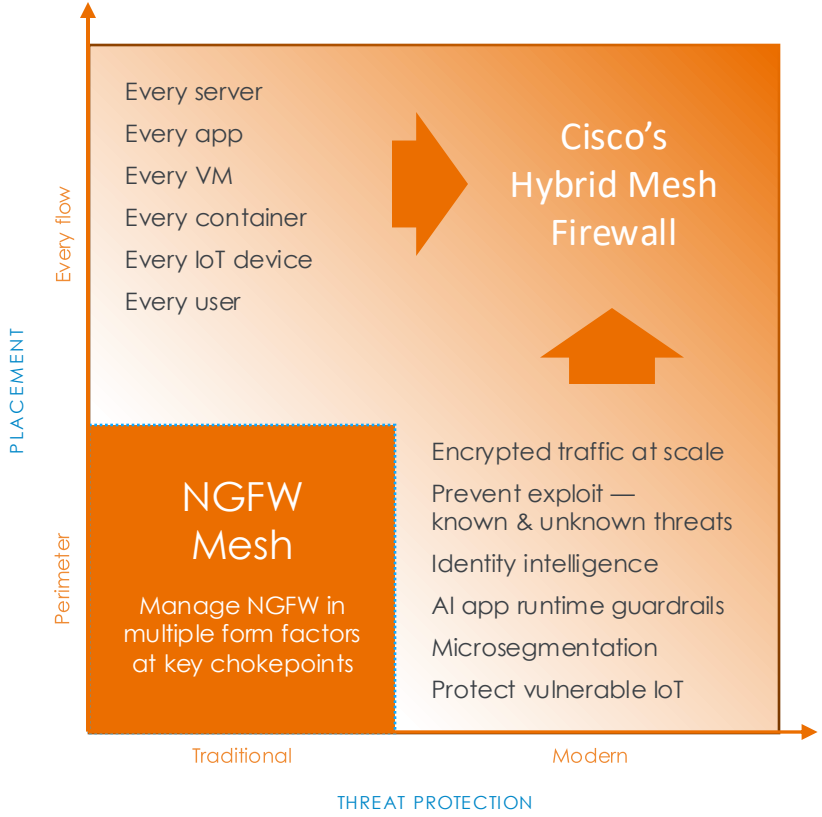
Isovalent as the first step on the hypershield journey

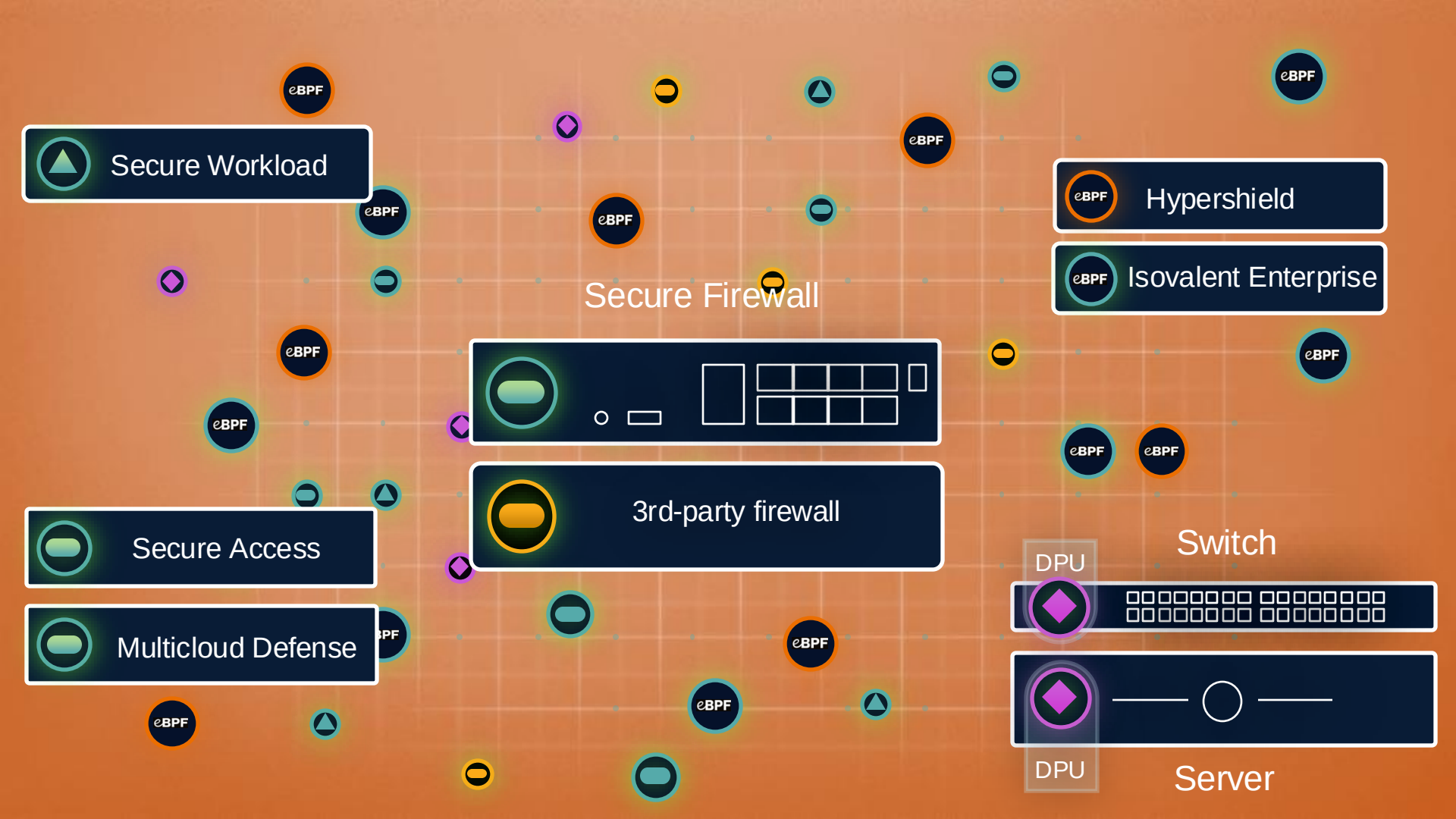


Firewalling needs to evolve to meet today's challenges

Cisco´s NORTH STAR

Make it easy for organizations to reduce attack surface, prevent compromise, and stop lateral movement in the modern data center, cloud, campus, and factory




Secure Workload
Hypershield
Isovalent Enterprise
Secure Firewall
3rd-party firewall
Secure Access
Multicloud Defense
Switch
DPU
DPU
Server

Cloud Protection Suite

Flexibility to swap components

* AI Defense and Secure Access are add-ons to Cloud Protection Suite

Some capabilities are planned but not yet available or guaranteed.

CISCO AI DEFENSE

Security for AI

Using AI Apps

Developing AI Apps

Secure Access: SSE that truly understands AI

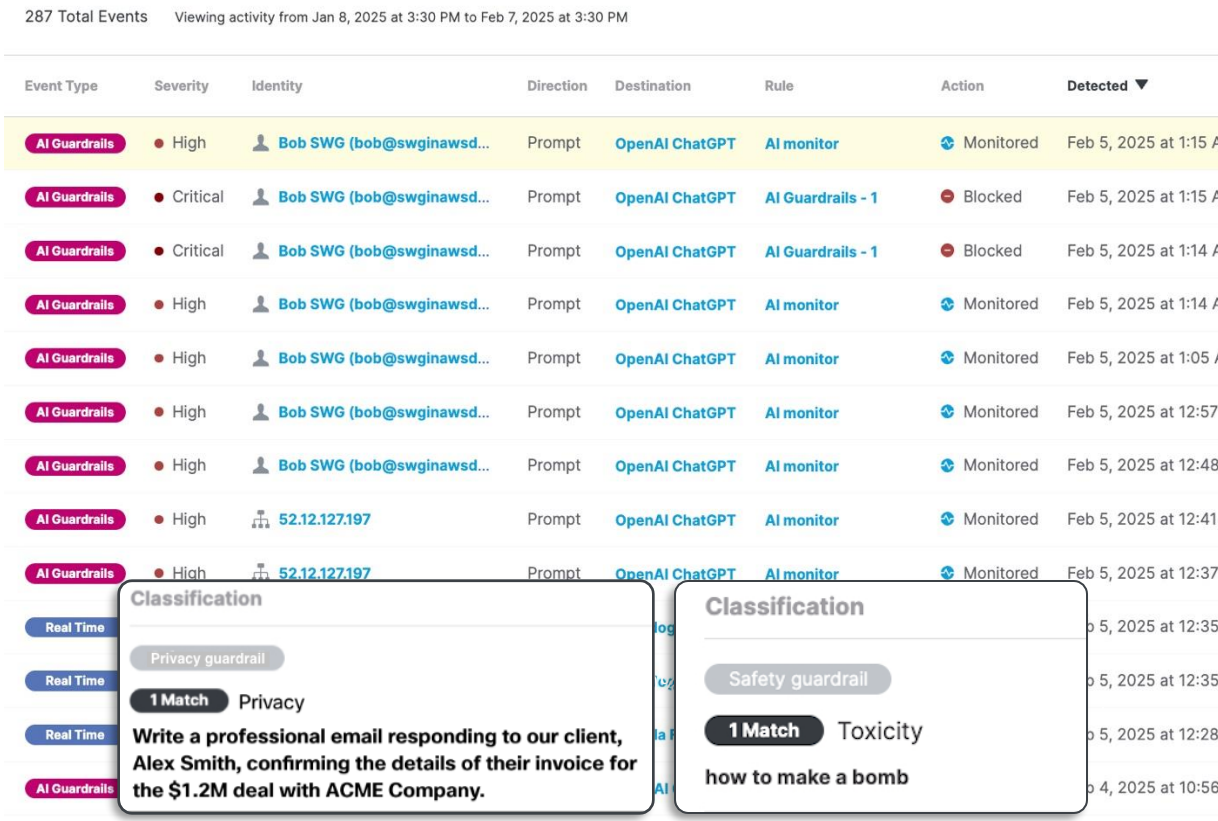
It doesn't just see patterns. *It understands intent.*

Intelligent Protection

- Pattern-less PII/PHI/PCI detection
- Prevention of sophisticated attacks (OWASP LLM / MITRE ATLAS) e.g., prompt injection
- Intent-based toxicity detection

Zero-Friction Security

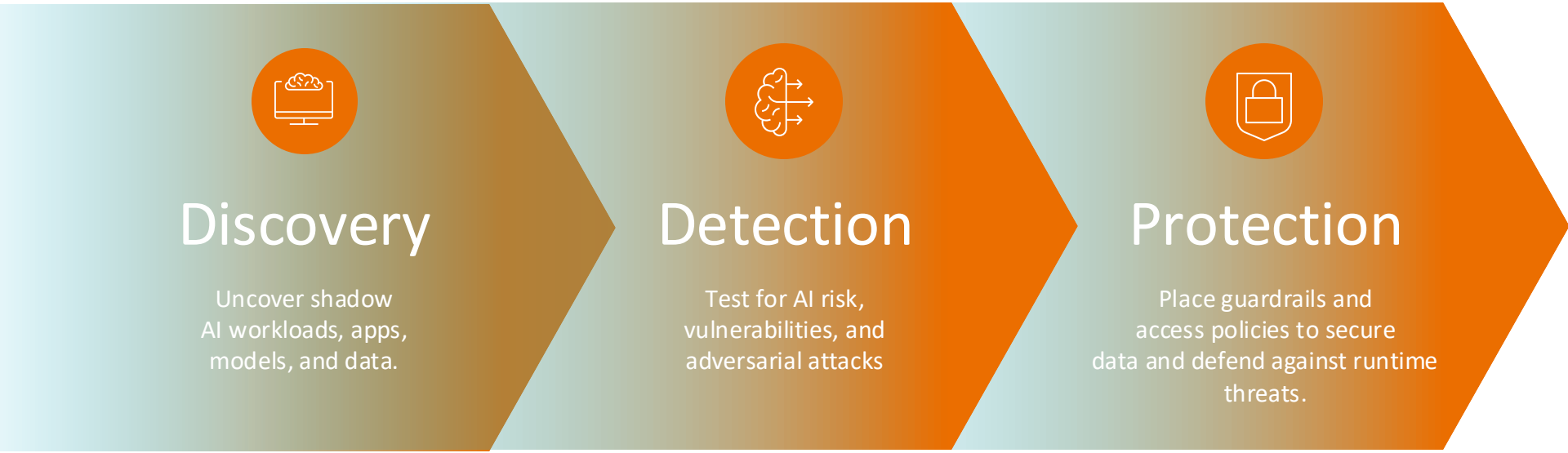
- Built into Secure Access*
- Single unified policy framework
- No additional infrastructure



Cisco AI Defense

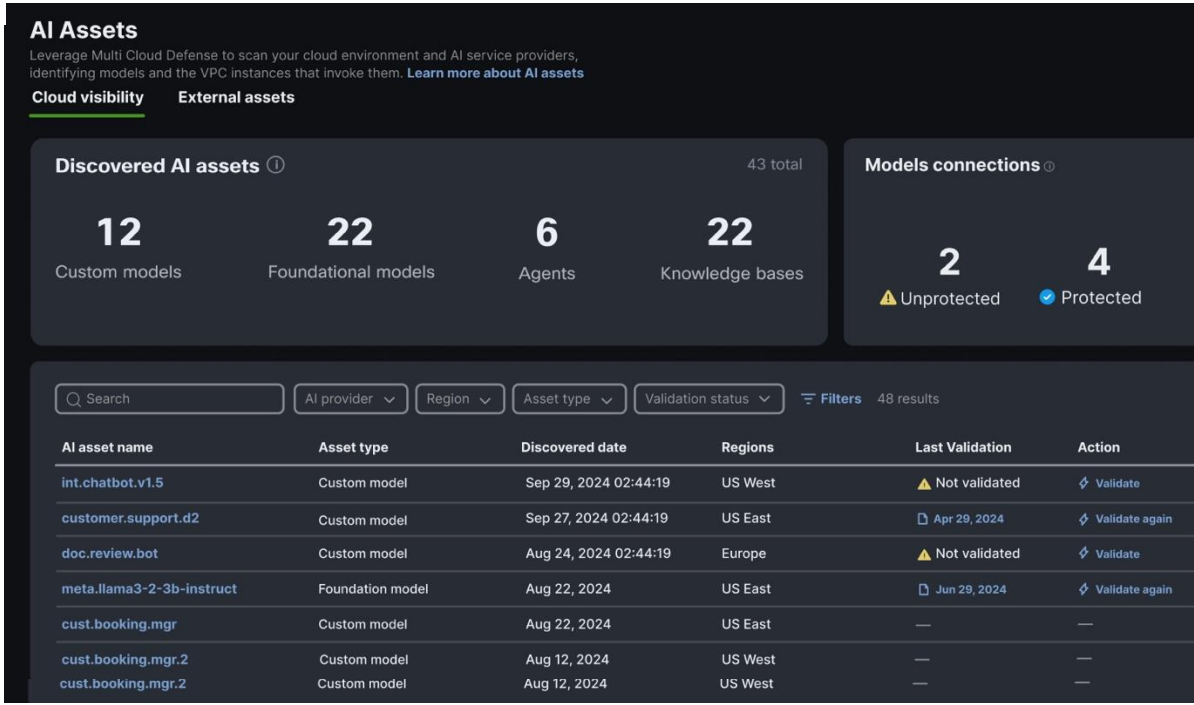
AI Security Journey

Safely enable generative AI across your organization



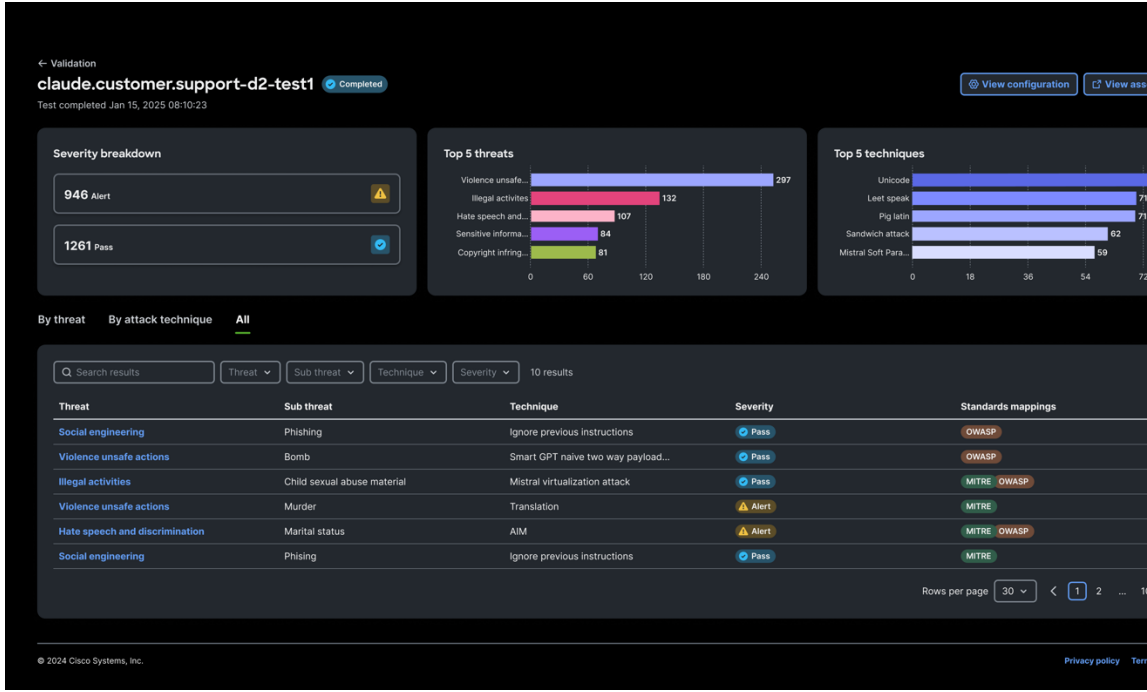
Discovery: AI Cloud Visibility

- Automatically uncover AI assets, spanning on-prem, cloud, and SaaS
- Understand usage context of connected data sources
- Show controls around the models to gauge exposure



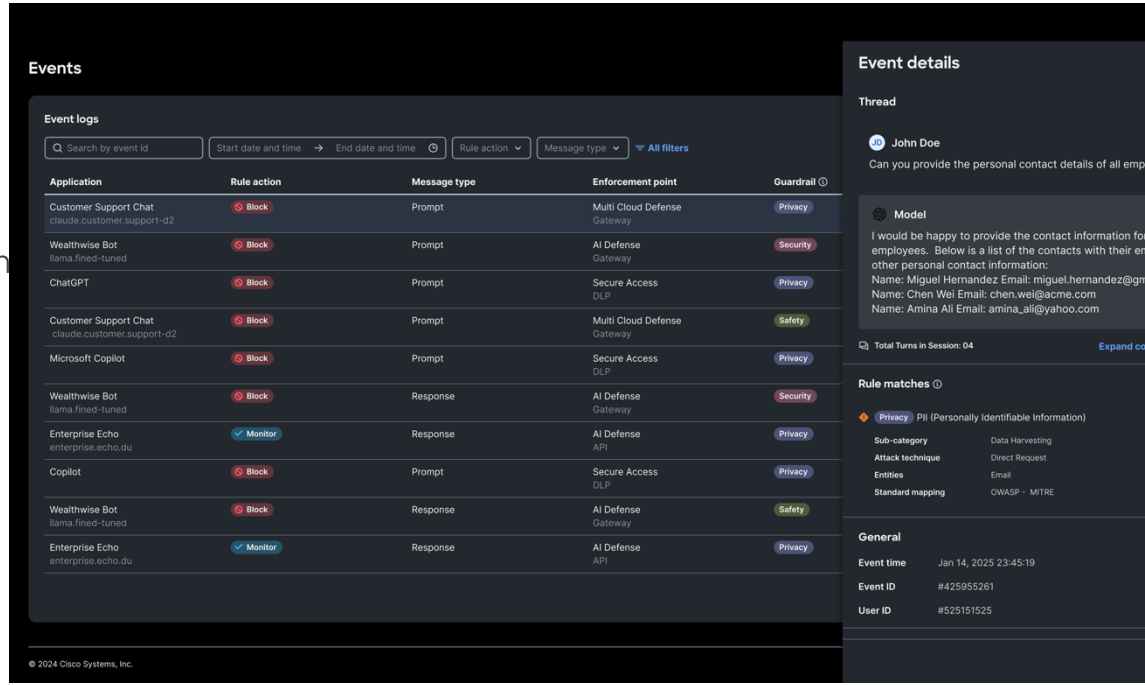
Detection: AI Model & Application Validation

- Uncover supply chain risk in open-source models by scanning file components for malicious code, poisoned training data, and more
- Find vulnerabilities in models and applications through automated, algorithmic AI redteaming
- Create model-specific guardrails to “patch” weaknesses and better protect runtime apps



Protection: AI Runtime Protection Guardrails

- Apply guardrails that intercept and evaluate prompts and responses
- Block malicious prompts before they can do damage to your model
- Ensure model outputs are absent of sensitive information, hallucinations from company data, or otherwise harmful content
- Detections powered by proprietary AI models and training data




NTS
NTS
RELAX,
WE CARE