Building secure AI data centers

Luka Markota, Pero Gvozdenica

22.10.2025.


The **next era** of AI is here

Evolution of AI

Chatbots

Humans talk to AI

Agentic

Workflows get automated


AI will make our world of **8B** people
feel like one with the capacity of **80B**

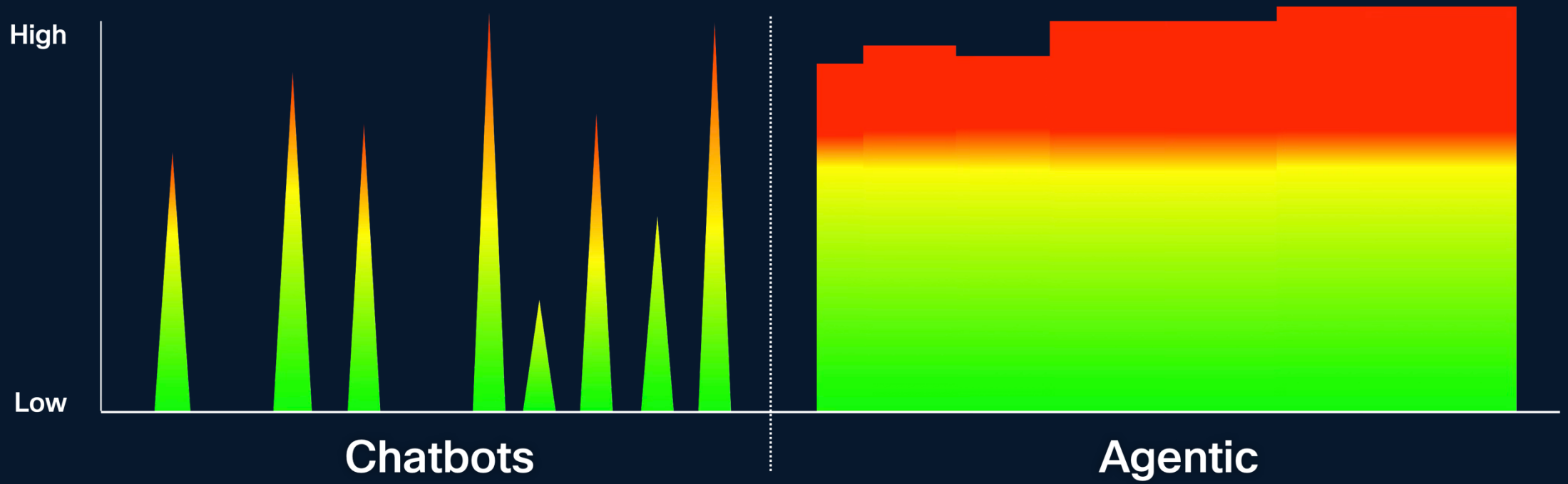
All of this has **massive implications** for our customers' **technology architectures**


Infrastructure
constraint
Trust
deficit
Data
gap

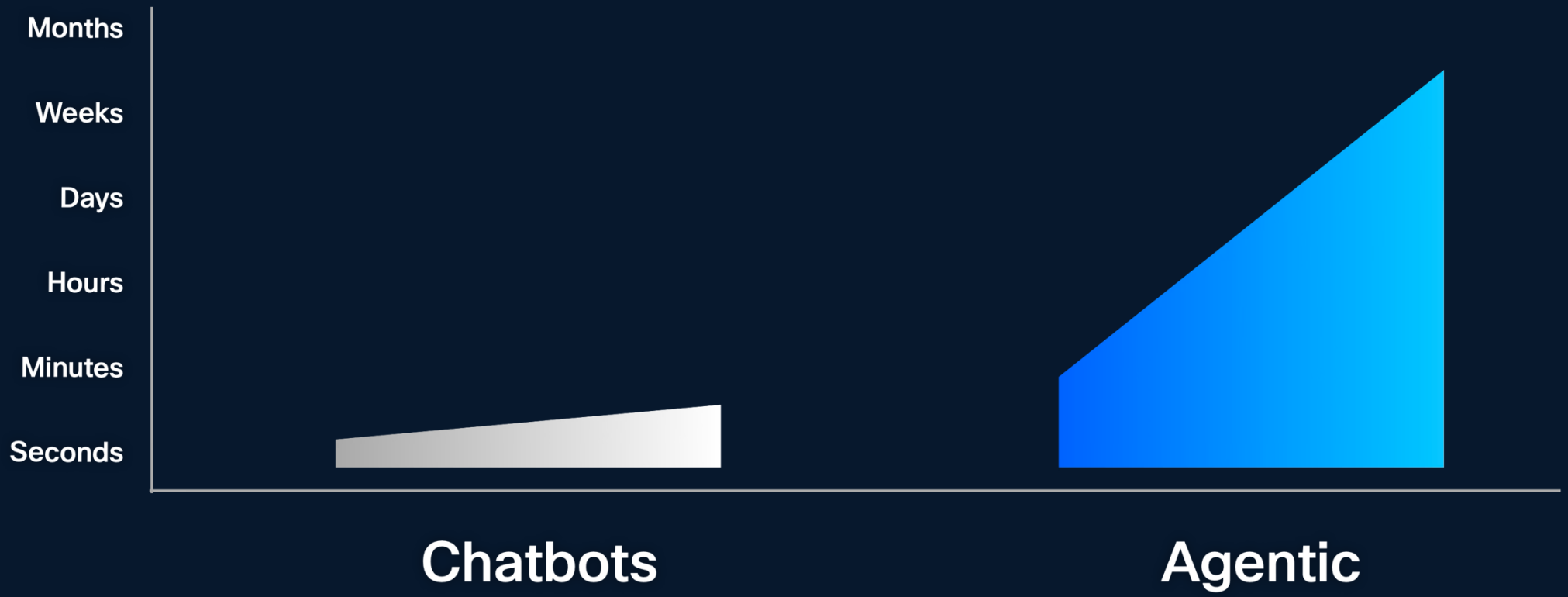

Infrastructure
constraint
Trust
deficit
Data
gap

Inferencing demand

Power I Compute I Networking



Duration of autonomous execution




Infrastructure
constraint
Trust
deficit
Data
gap

Model threat vectors

Safety

Profanity

Cost harvesting / repurposing

Harassment

Hallucinations

Hate speech

Off-topic

Toxicity

Social division & polarization

Self-harm

Financial harm

Security

Indirect prompt injection

Infrastructure compromise

IP theft

Meta prompt extraction

Prompt injection

Model theft

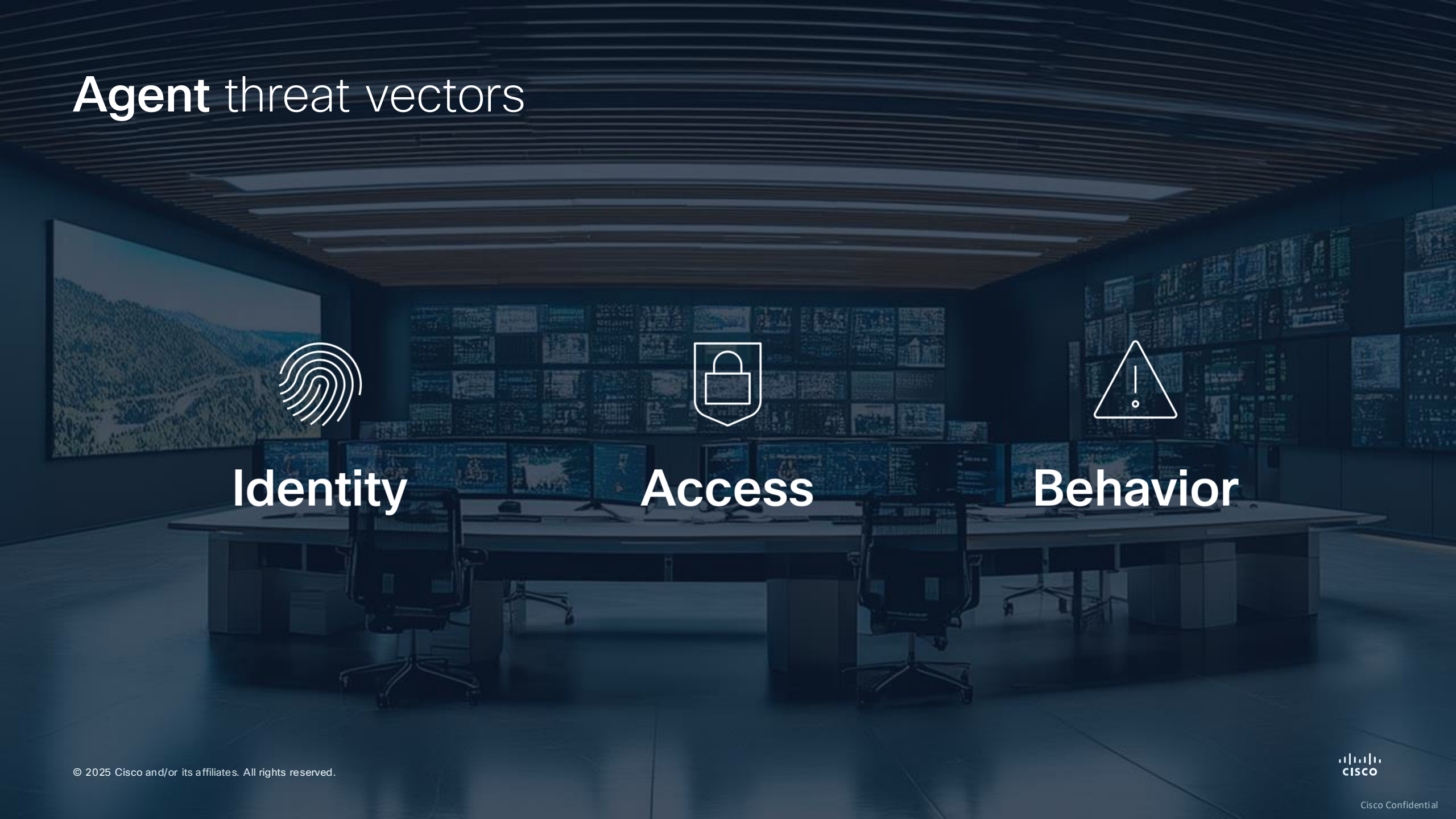
Training data poisoning

Sensitive information disclosure

Data exfiltration

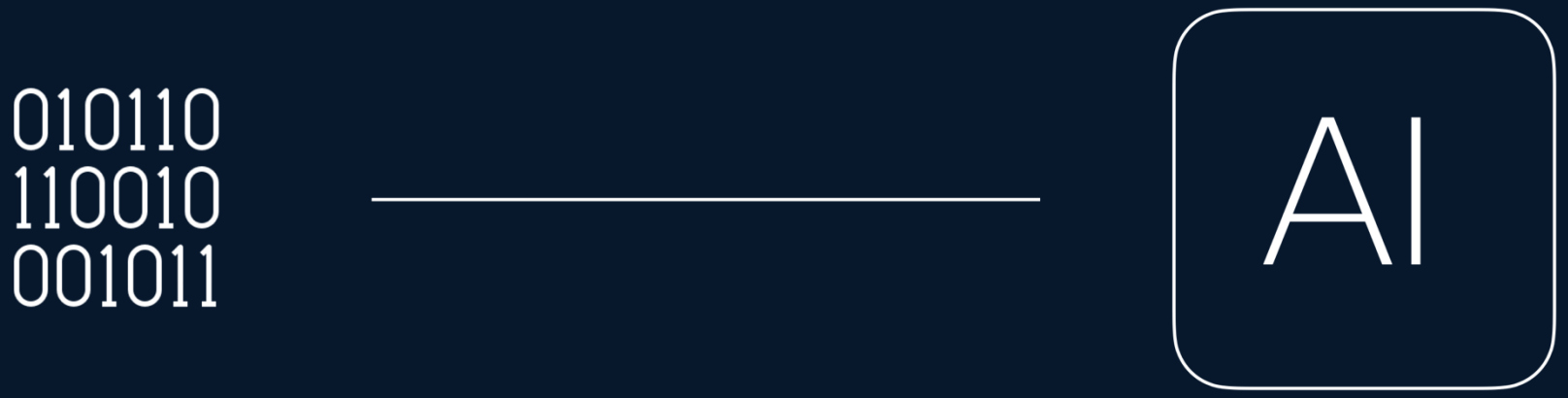
Model denial of service

Agent threat vectors




Infrastructure
constraint
Trust
deficit
Data
gap

Data is the **essential fuel** for AI



Data is the **essential fuel** for AI

Human-generated

Text

Audio

Video

Machine-generated

010110
110010
001011
Metrics

Events

Logs

Traces

Other telemetry

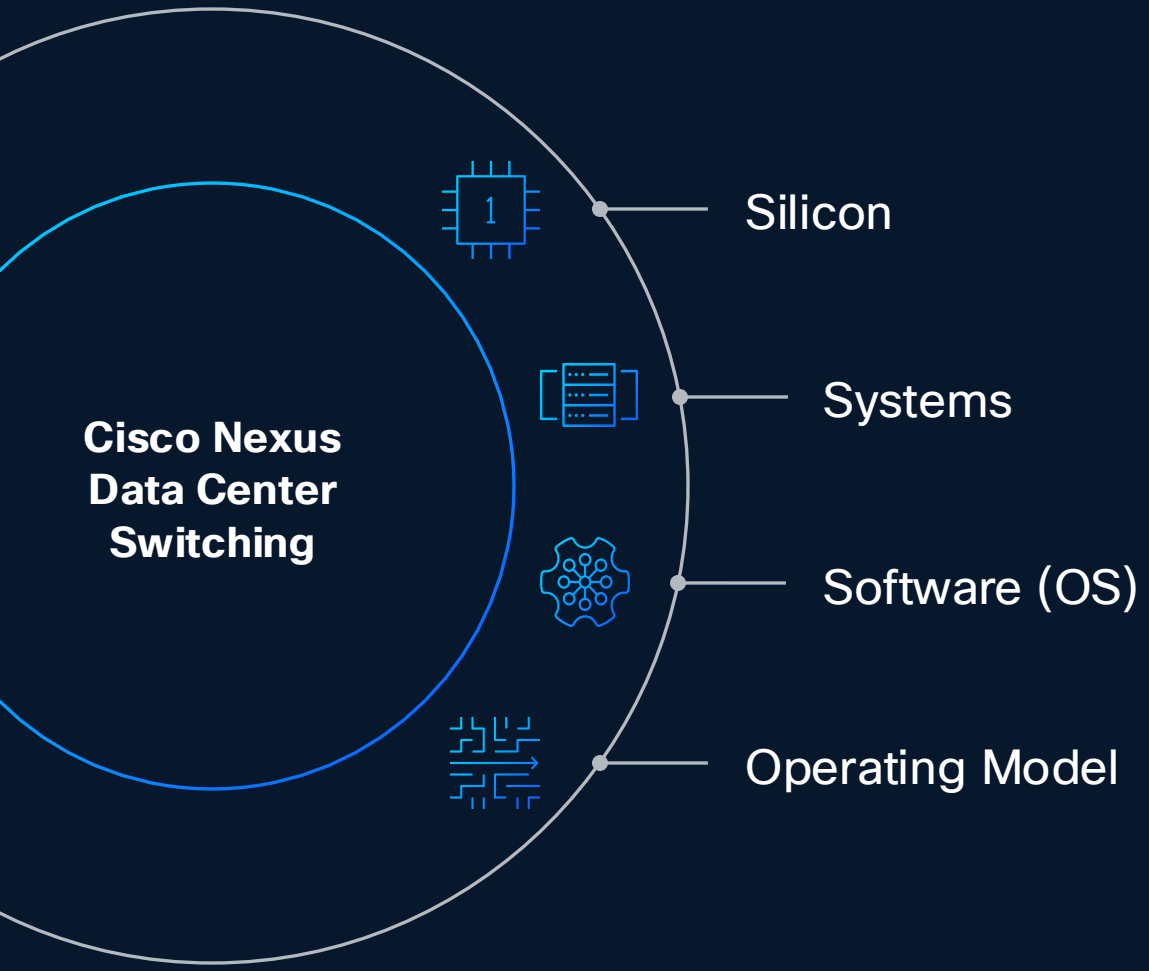
We’re addressing all of these challenges **head-on**


The world is experiencing the **largest expansion of data centers** in history

Cisco is foundational to the world's data center buildouts

Hyperscalers | Neoclouds | Service Providers | Enterprises

Foundations of Cisco Data Center Switching



Silicon and systems

Silicon diversity

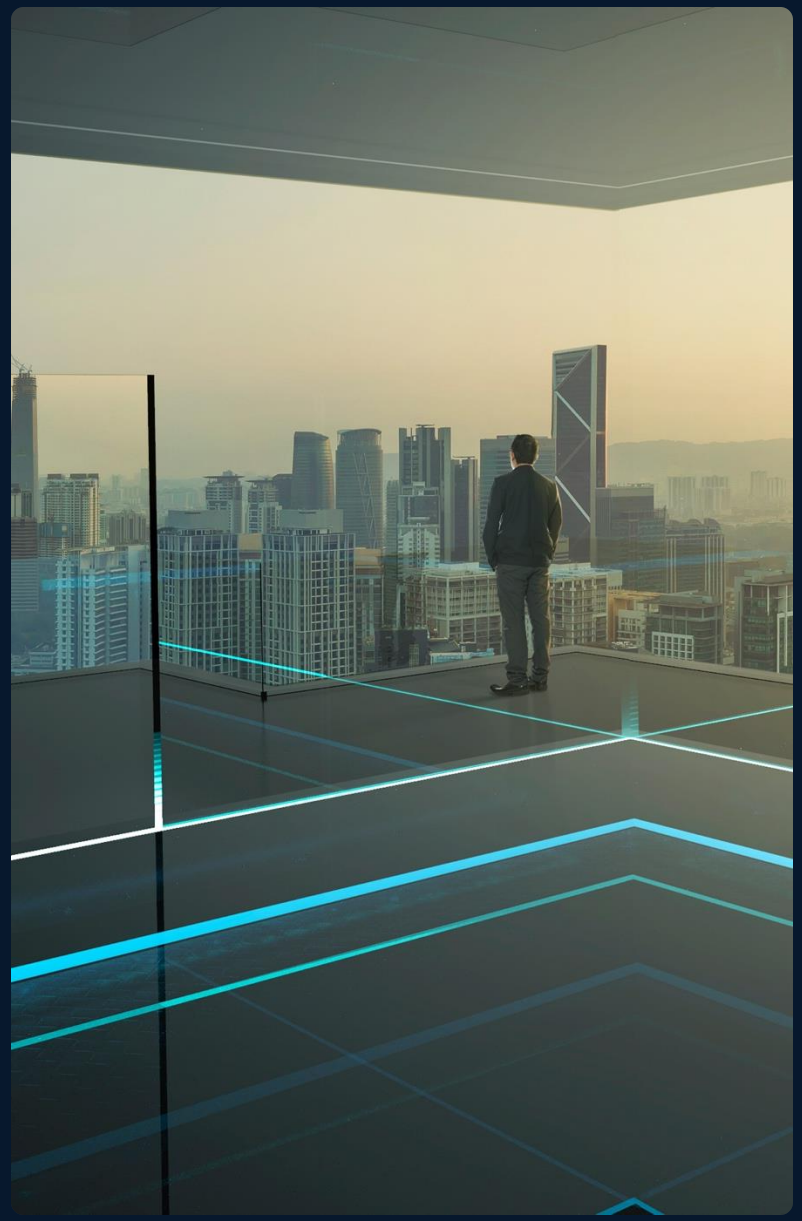
- Cisco Cloud Scale and Silicon One ASICs powering Nexus
- Port speeds up to 400G / 800G
- SerDes speeds up to 25G / 50G / 100G

Smart switches

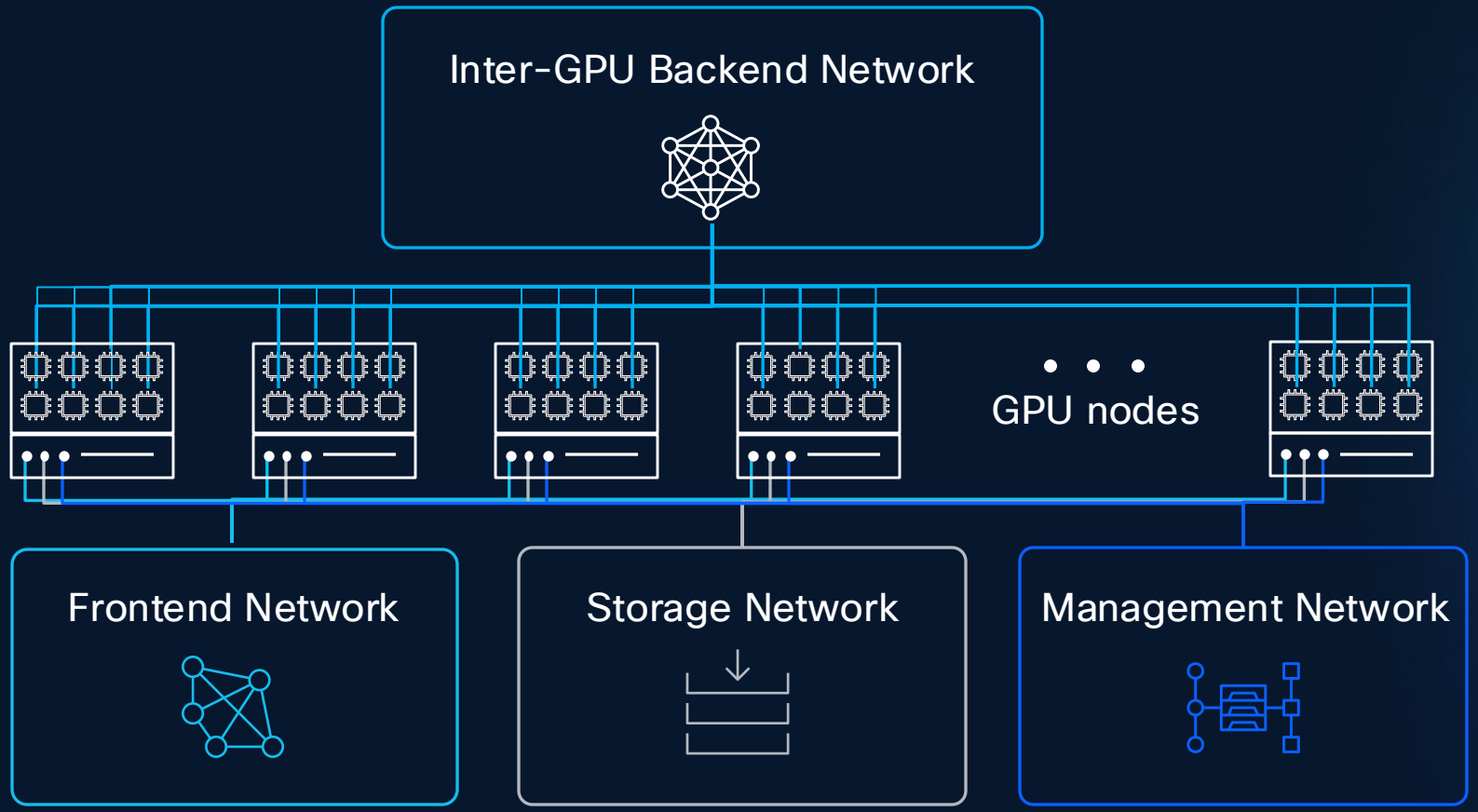
- Built-in DPUs for services acceleration
- Multiple service functions at ASIC performance – security, NAT, telemetry
- Programmable pipelines for service firmware

Ultra low latency

- Next-Gen ULL ASIC
- Silicon One based, 3.2T
- Enable 10G to 25G market transition



AI Networking



Cisco Nexus high density 800G & 400G fixed switches

Cisco AI PODs embedding the latest Nexus Switches

Nexus 9300 64-port 800G Switch

512-wide radix	Fully shared packet buffer	Advanced load balancing	Low Latency

N9364E-SG2-Q or N9364E-SG2-O

Compact 2RU 51.2T Switch

G200 ASIC (5nm) | 100G SerDes | 256MB packet buffer

64 800G ports | Up to 128 line-rate 400G ports (2x400G breakout)

Choice of QSFP-DD800 or OSFP ports

Multi Core x86 CPU | 32GB RAM | 128GB SSD

Cisco NXOS spine and AI/ML spine/leaf capable

Ultra Ethernet Consortium MEMBER

Nexus 9332D-GX2B
32p 400G
8p MACsec/CloudSec

Nexus 9364D-GX2A
64p 400G
16p MACsec/CloudSec

ACI Leaf, ACI Spine, and NX-OS
25.6T, 19.2T, and 12.8T 400G switches
120MB smart buffer

Security
MACsec and CloudSec

Telemetry
FT, FTE, SSX, INT-XD

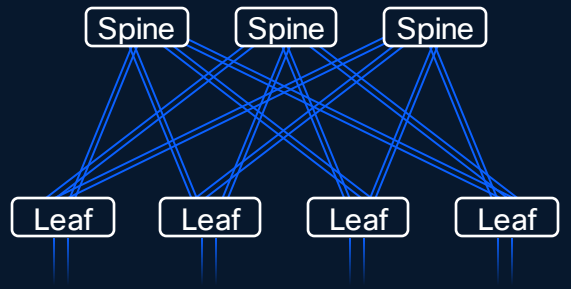
Best performance for AI networks

Introducing Cisco Intelligent Packet Flow

Powering AI Fabrics

Advanced load balancing capabilities for AI Networks

Intelligent load balancing for every path, every flow



FEATURES

- Equal Cost Multi Path Load Balancing (ECMP)
- Weighted Cost Multi Path Load Balancing (WCMP) with Dynamic Load Balancing
- Dynamic Load Balancing (DLB) with Load and Congestion Awareness
- Dynamic Load Balancing Per packet Mode (Packet Spray)
- Policy Based Flowlet Load Balancing
- Equal Cost Multi Path Load Balancing (ECMP) Static Pinning

Compute AI portfolio

Address AI workloads with visibility, consistency, and control

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

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

RTX PRO SERVER
Supporting RTX PRO 6000 Blackwell Server Edition GPUs

Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers

Cisco UCS blade (with GPU extensions) and rack servers

Enterprise AI edge

Dense compute for demanding AI

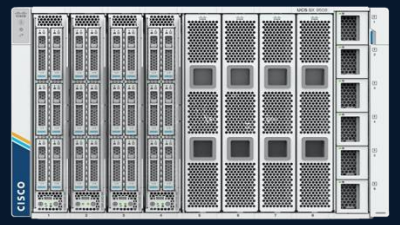
Full-stack AI with compute and networking

Cisco UCS Compute Portfolio

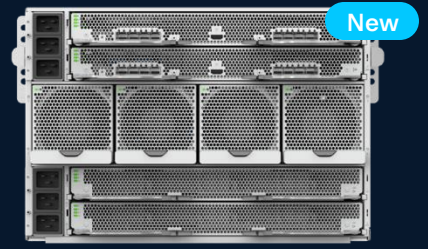
MAINSTREAM ENTERPRISE SERVERS

UCS X-Series X9508 Chassis

IFM Module



UCS X-Series Direct



UCS x580p M8

UCS X210c M7

UCS X210c M8

UCS X215c M8

UCS X410c M7

UCS C240 M8E3S
36 EDSFF E3.S1T

UCS C240 M8SX
28 HDD/SDD/NVMe

UCS C240 M8L
16 LFF + 4 SFF

UCS C240 M7SN
28 NVMe

UCS C240 M6S
14 SSD/HDD Media drive

UCS C240 M6N
14 NVMe Media Drive

UCS C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI SERVERS

UCS C885A M8
8RU Dense GPU Server

UCS C880A M8
10RU Dense GPU Server

UCS C845A M8
4RU MGX Server

Cisco UCS C880A M8 AI Supercomputer

Purpose-Built for AI and High-Performance Computing

Accelerating advanced AI in every data center. Expanding Cisco UCS Dense AI Servers portfolio with NVIDIA Blackwell flagship GPUs

Cisco UCS C880A M8

NVIDIA HGX with
8 NVIDIA B300 (SXM) GPUs

NVIDIA NV-Link

2 Intel Xeon 6th Gen Processors

Real-Time Large Language Model Inference
Next-Level Training Performance
Large-Volume Data Processing

- Service Providers
- Financial Services
- Manufacturing
- Healthcare and Life Sciences
- Automotive

Stateless Computing

SIM Card
Identity for a phone

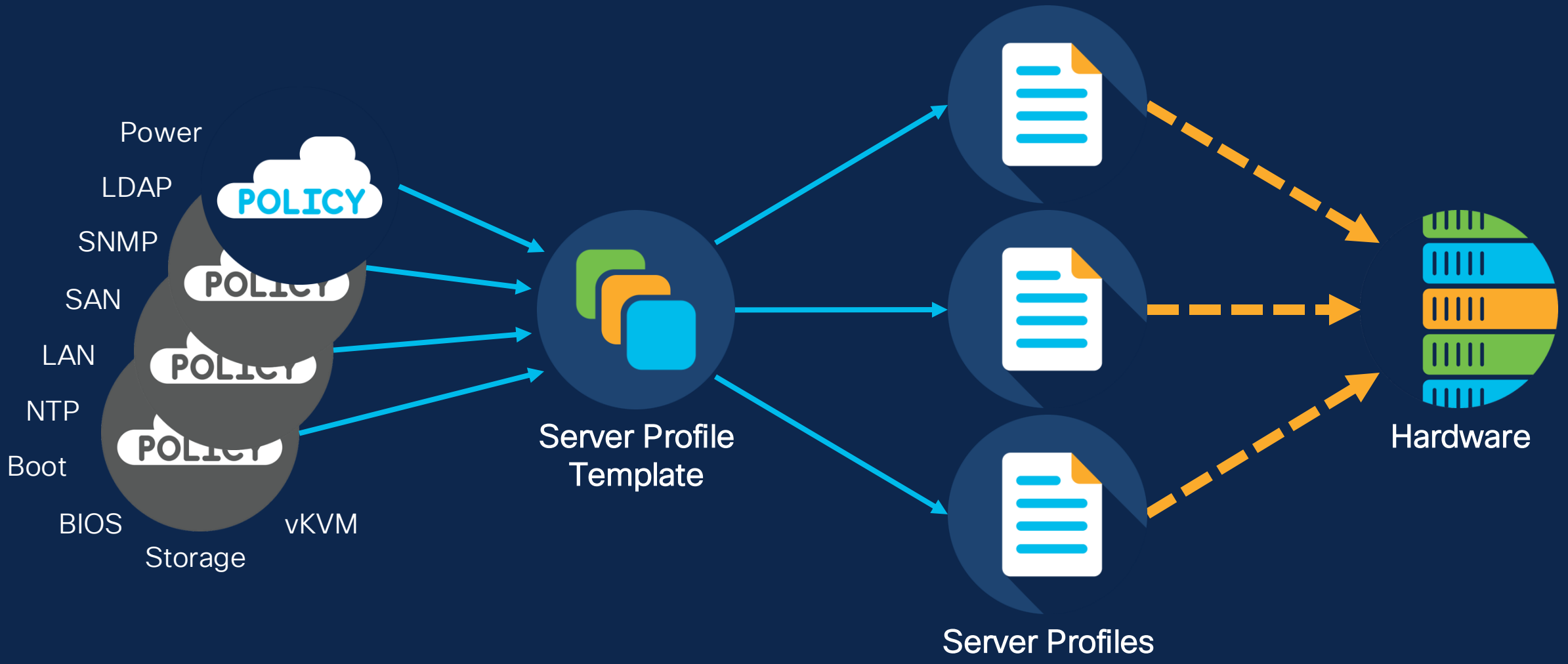
Profiles
Identity for a server

Intersight Server Profiles

- Network Policy
- Storage Policy
- Server Policy

Server Profile Templates

Policy-based configuration of UCS servers



Cisco AI PODs

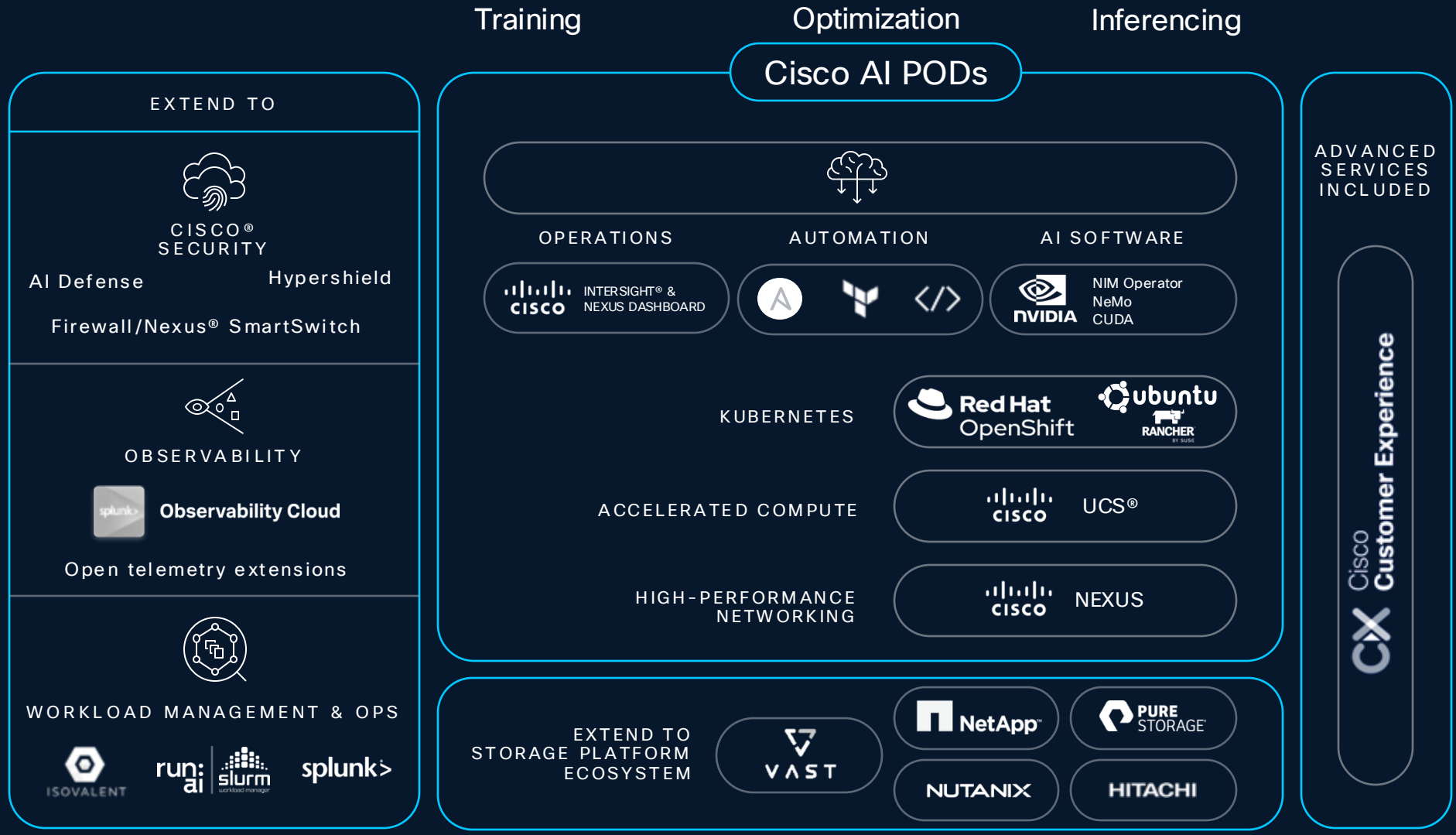
A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence
Cisco CVD, NVIDIA ERA

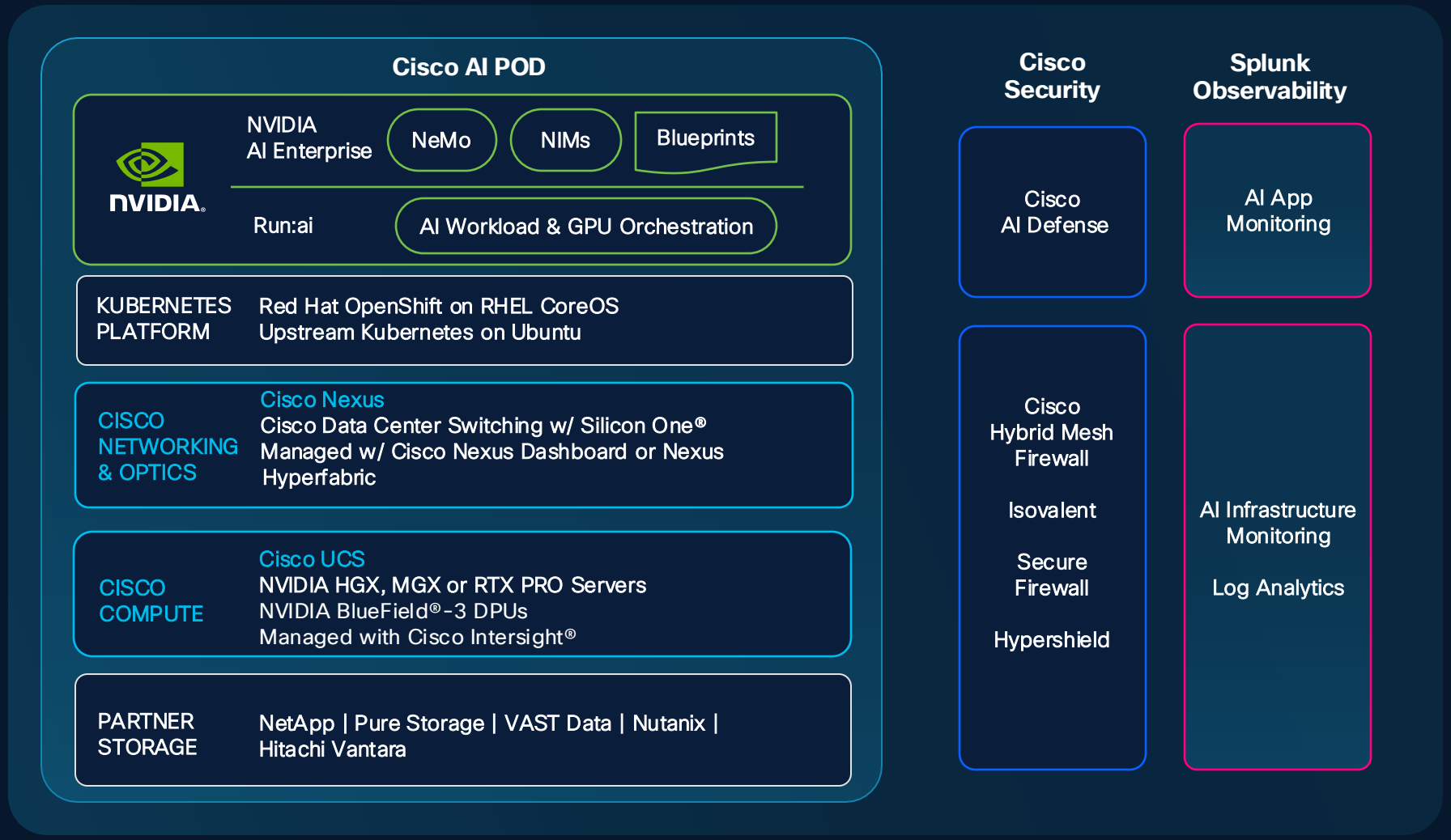
Fully supported stack including Cisco and 3rd party components
Cisco CX Success Track

Orderable, use case driven AI-ready infrastructure stacks
Inferencing.
Optimization.
Training.

Incremental, atomic-level -or- fabric-based cluster scale



Cisco Secure AI Factory with NVIDIA



AI Data Hub as a Service Architecture

Your Data

Structured, semi-structured and unstructured

Google Cloud Storage · Microsoft Azure Blob Storage · Amazon S3 · ORACLE
PostgreSQL · SAP · IBM · Microsoft SQL Server
mongoDB · Epic · Cerner · teradata.
Apache Spark · elastic · Google BigQuery · databricks
snowflake · kafka · trino · amazon REDSHIFT
OpenWeather · monday.com · Greenplum · Google Analytics
servicenow · druid · NETEZZA · ICEBERG
Apache orc · Parquet · SQLite · VERTICA
AVRO · DELTA LAKE · cassandra · MariaDB
X · IBM DB2 · xero · salesforce

...and many more

EverFlex

Your Centralized Data Hub offering for Data, AI & Analytics

- Self Service UI and REST API
- Inference, RAG & Agentic AI
- Data Virtualization Engine
- Open Data Lakehouse
- Real Time Query Engine
- Automated Data Pipeline

DATA INSIGHTS | MODERN LAKEHOUSE | AI WORKBENCH

Federated access to all your source data with smart query processing

Zetaris

Data Classification | Relationship Finder | Data Lakehouse Builder

HC² INFRASTRUCTURE LAYER

CISCO · RED HAT OPENSHIFT

GPU · VSP HITACHI · GPU

LAKEHOUSE

S3 · SQL

HCP

Your Data Access

AI powered
Enterprise Chatbot
(with RAG system)

Virtual Assistant · Agentic AI · ChatBot

Apps powered by Zetaris

tableau · python · data iku
Apache Superset · R · H2O.ai
Power BI · Excel

...and many more

Data Applications · Chatbots / Co-Pilots

Hitachi Vantara & Zetaris

Red Hat

AI maturity model

Customer journey

AI Business

AI Infrastructure

1	2	3	4	5
Awareness	**Active**	**Operational**	**Systemic**	**Transformational**
Early interest in AI, conversation happening	AI experimentation mostly in tactical data science contexts	AI in production with best practice experts and technology and available to organization	Pervasive use of AI; All new digital projects at least consider AI	AI is part for the business DNA
Traditional	**AI Emergent**	**AI Ready**	**AI Optimized**	**AI Native**
Not in the conversation; Maintaining legacy big-data infrastructure	Some public cloud or single GPU server; Modernizing data platform	First secure AI cluster supporting a major use case	AI infrastructure begins to scale for multiple lines of business	AI capacity = revenue potential

Source: NVIDIA | Cisco

Leading through the maturity journey

Make it real for customers

AI readiness assessments

TCO / ROI calculators

Reference Architectures, CVDs & White Papers

Demo gear in SE GEO labs

Agentic workflow demos

AI Defense Demo

Infrastructure Capacity calculators

Cisco-on-Cisco Journey demos / EBCs

CISCO
SECURE

Next generation DC workload Security

Luka Markota

25.09.2025

The bridge to possible

Agenda

- Sigurnost u Data Centru – zašto segmentirati?
- Regulatorni okviri – zašto sada?
- Cisco Secure Workload & Isovalent – kako mi to radimo?

Cisco Hybrid Mesh Firewall goes broader and deeper

Security Cloud Control

Native enforcement points go deeper ⟷ Integrate with existing

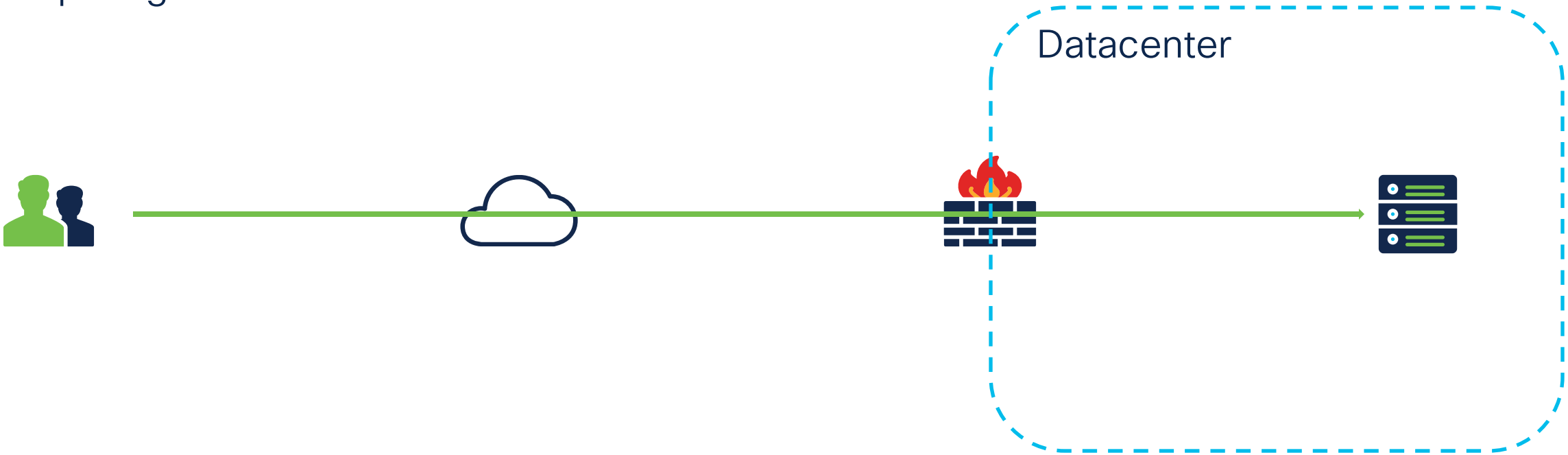
Write policy once, enforce across the mesh

Sigurnost u Data Centru – zašto segmentirati?

Securing Application Workloads – Threat Landscape

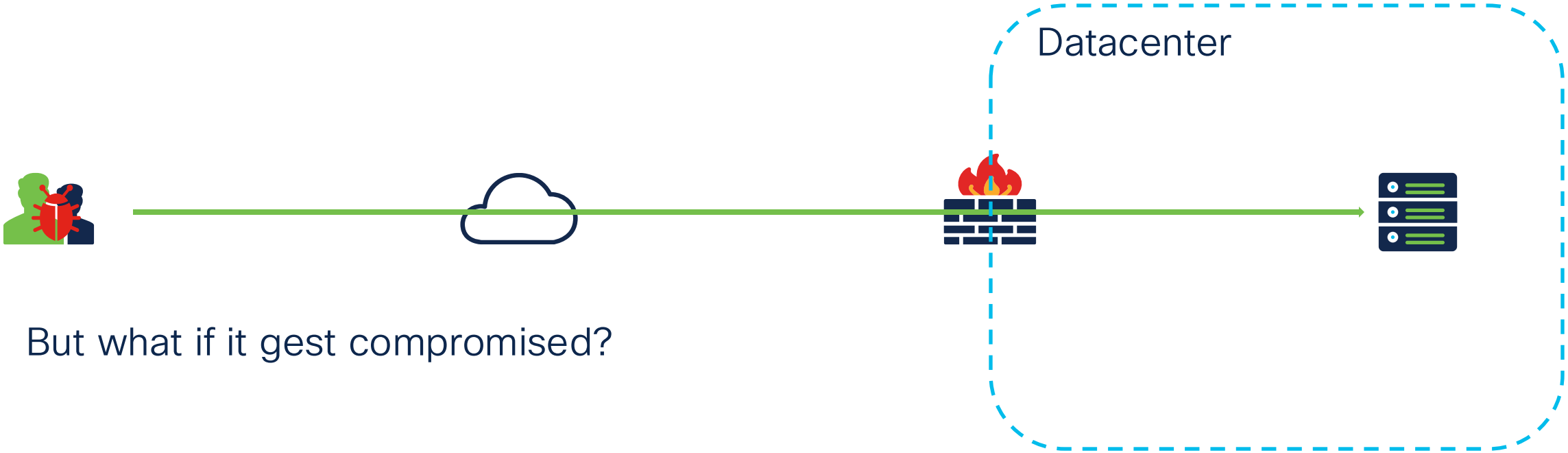
Using Network Security Controls

Simple right?



Securing Application Workloads – Threat Landscape

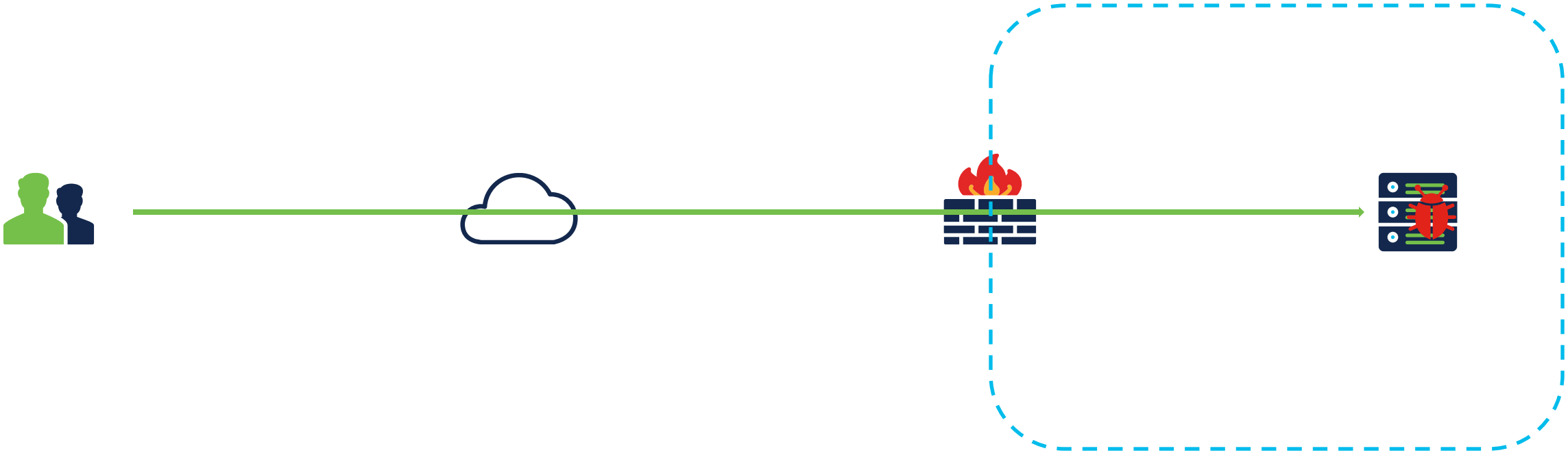
Using Network Security Controls



But what if it gest compromised?

Securing Application Workloads – Threat Landscape

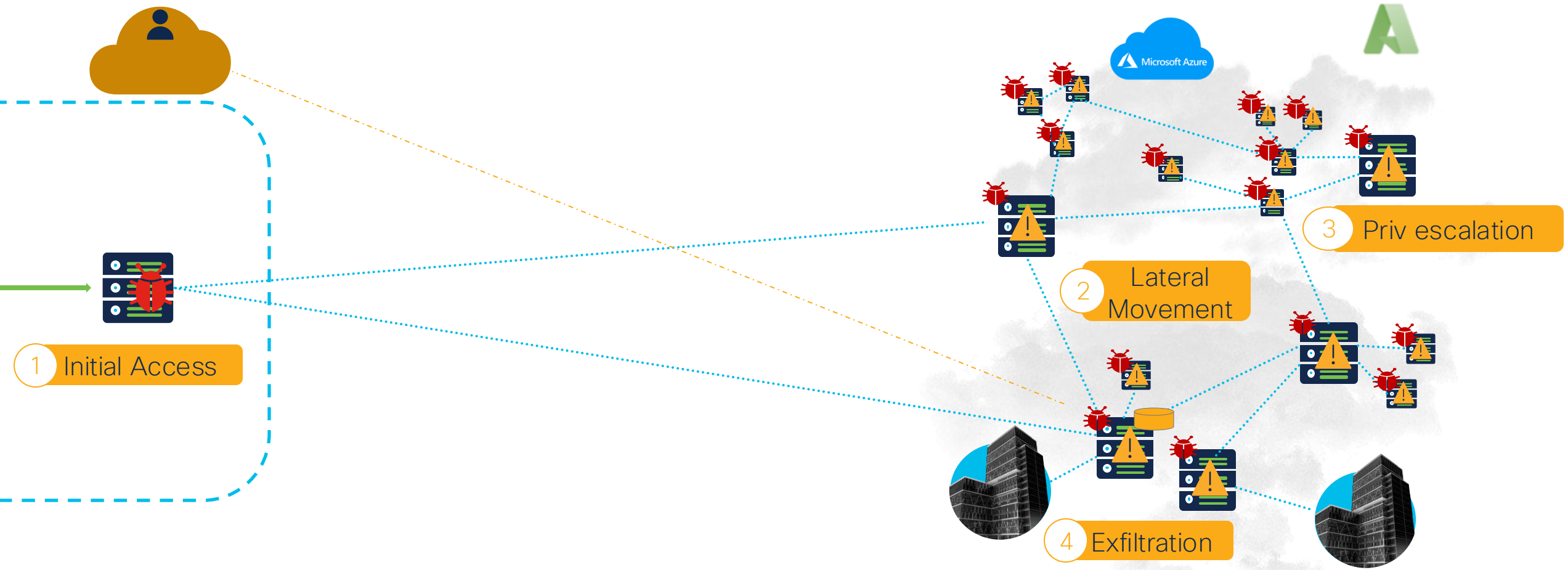
Using Network Security Controls



And this is only a part of the story…….

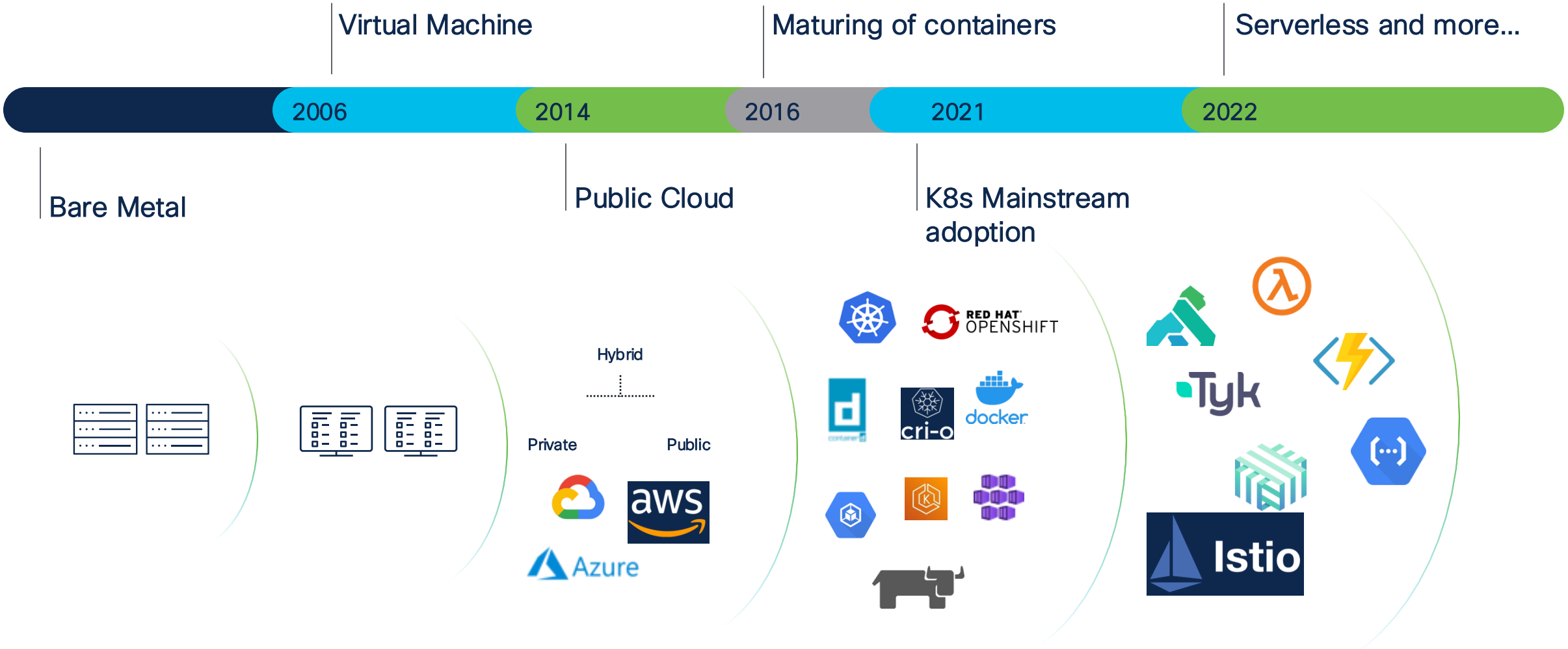
Securing Application Workloads – Threat Landscape

Using Network Security Controls

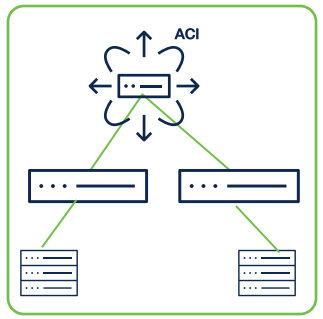
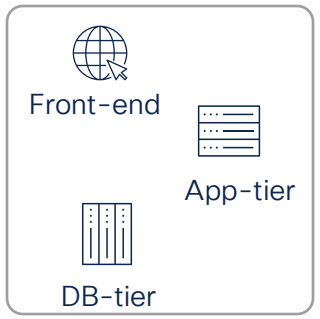
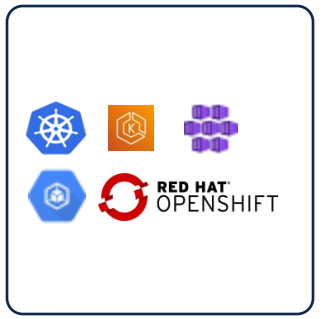


Application Workload Evolution

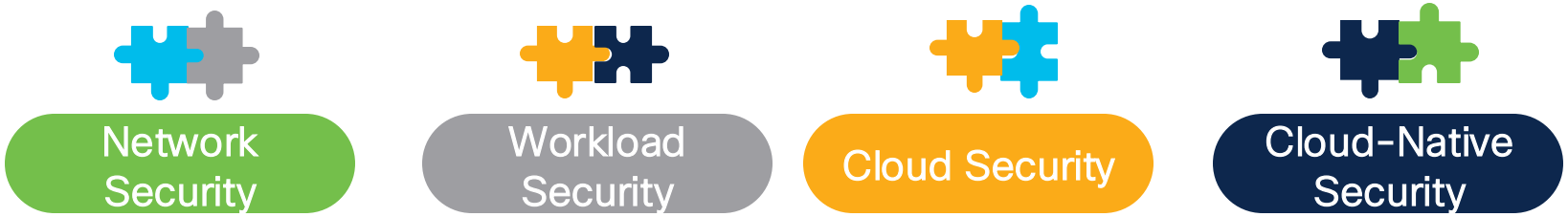
Workload Security is Getting More Complex!



But… what is an application workload?

Network Engineer	Firewall Engineer	Cloud Engineer	Application Owners	Cloud-Native Engineer
				
• Vlans/VRF • Subnets • Contracts	• Zones • Subnets • ACLs	• VPC • Subnets • Security Groups	• Service • Application • Workload	• Namespace • Service • CNI

Segmentation and Policy Control Challenges



NetSec Admin | Server/VM Admin | Cloud Architect | DevSecOps

Organizational Challenges

Multiple teams, organizations and environments

Inconsistent islands of policy controls across environments

Secure Workload Use-Cases

Microsegmentation

Behavioral detection and protection

Vulnerability detection and protection

Regulatorni okviri – zašto sada?

Risk Management

Incident Reporting

Digital Operational Resiliency Testing

ICT Third-Party Risk

Information and Intelligence Sharing

European Supervisory Authorities

eba European Banking Authority

eiopa European Insurance and Occupational Pensions Authority

ESMA European Securities and Markets Authority

Mapping: DORA TO NIST CSF 2.0

DORA

NIST CSF 2.0

Chapter II: ICT Risk Management	Chapter III: ICT-Related Incidents, Management, Classification and Reporting	Chapter III: ICT-Related Incidents, Management, Classification and Reporting	Chapter III: ICT-Related Incidents, Management, Classification and Reporting
ID.AM-1 – Article 8 ID.AM-2 – Article 8 ID.AM-3 – Article 8, Article 9 ID.AM-4 – Article 8 ID.AM-5 – Article 8	PR.IR-1 – Article 9	DE.CM-1 – Article 9, Article 10 DE.CM-9 – Article 9, Article 10	RS.MI-1 – Article 9, Article 11 RS.MI-2 – Article 9, Article 11
Chapter VI: Information Sharing			
ID.RA-1 – Article 8 ID.RA-2 – Article 45			

NIS2

NIST CSF 2.0

Mapping: NIS2 TO NIST CSF 2.0

Chapter II: Coordinated Cybersecurity Frameworks

ID.AM-1
ID.AM-2
ID.AM-3
ID.AM-4
ID.AM-5

Chapter II: Coordinated Cybersecurity Frameworks Chapter VI: Information Sharing

ID.RA-1
ID.RA-2

Chapter IV: Cybersecurity Risk-Management Measures And Reporting Obligations

PR.IR-1

Chapter IV: Cybersecurity Risk-Management Measures And Reporting Obligations

DE.CM-1
DE.CM-9

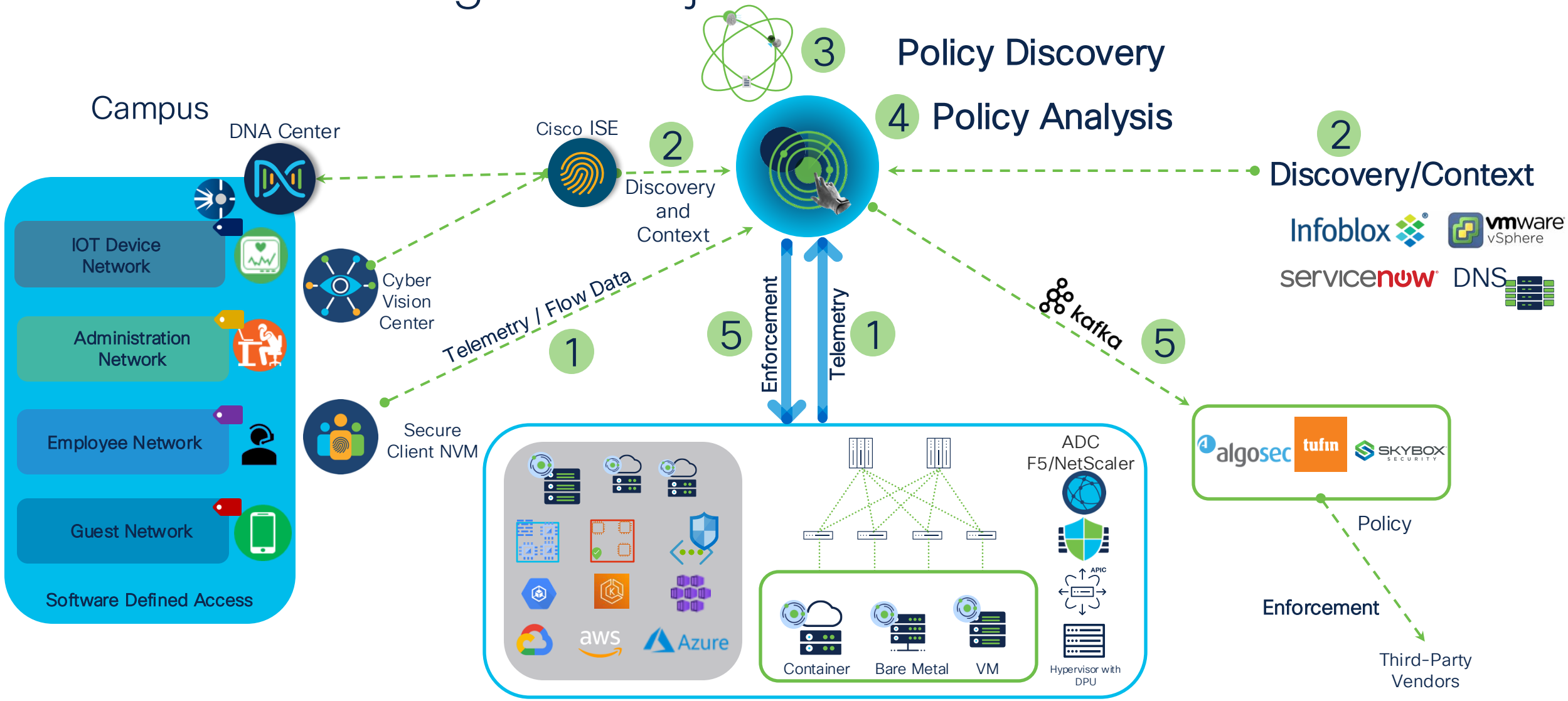
Chapter IV: Cybersecurity Risk-Management Measures And Reporting Obligations

RS.MI-1
RS.MI-2

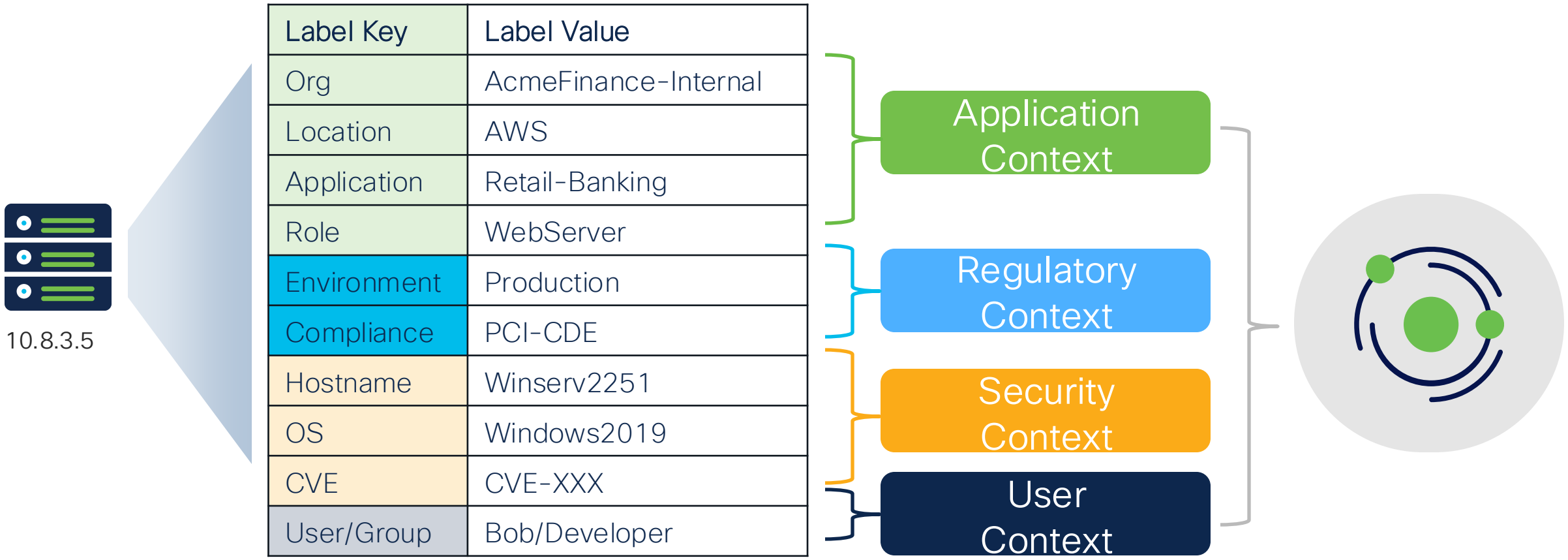
https://www.nis-2-directive.com/NIS_2_Directive_Articles.html

Cisco Secure Workload – kako mi to radimo?

Put do mikrosegmentacije kroz Cisco Secure Workload

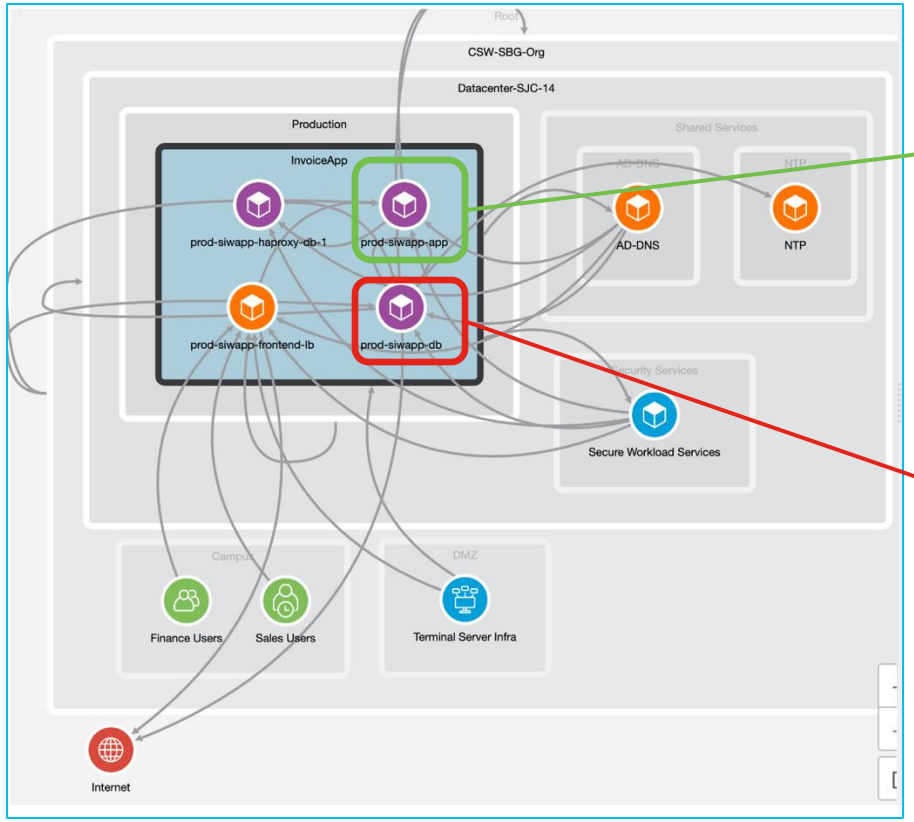


Identifying workloads with context
(labels/tags/annotations)



Application Dependency Mapping

Policy Discovery of Clusters – Behavior-Based and ML Grouping



Cluster: **prod-siwapp-app** | Cluster Actions | Name: prod-siwapp-app | Description | Query: **Hostname** *contains* **prod-siwapp-app-**

Edit Cluster Query (5) | View Cluster Details

Services 0 | Pods 0 | Workloads 3 | IP Addresses 0 | Neighbors 1

3 of **3** inventory

Orchestrator Type	Hostname	Address	OS	* Environment	* Location
vCenter	prod-siwapp-app-3	192.168.1.6	CentOS	Production	Datacenter
vCenter	prod-siwapp-app-2	192.168.1.5	CentOS	Production	Datacenter
infoblox, vCenter	prod-siwapp-app-1	192.168.1.4	CentOS	Production	Datacenter

Cluster: **prod-siwapp-db** | Cluster Actions | Name: prod-siwapp-db | Description | Query: * **orchestrator_system/machine_name** *contains* **prod-siwapp-db-**

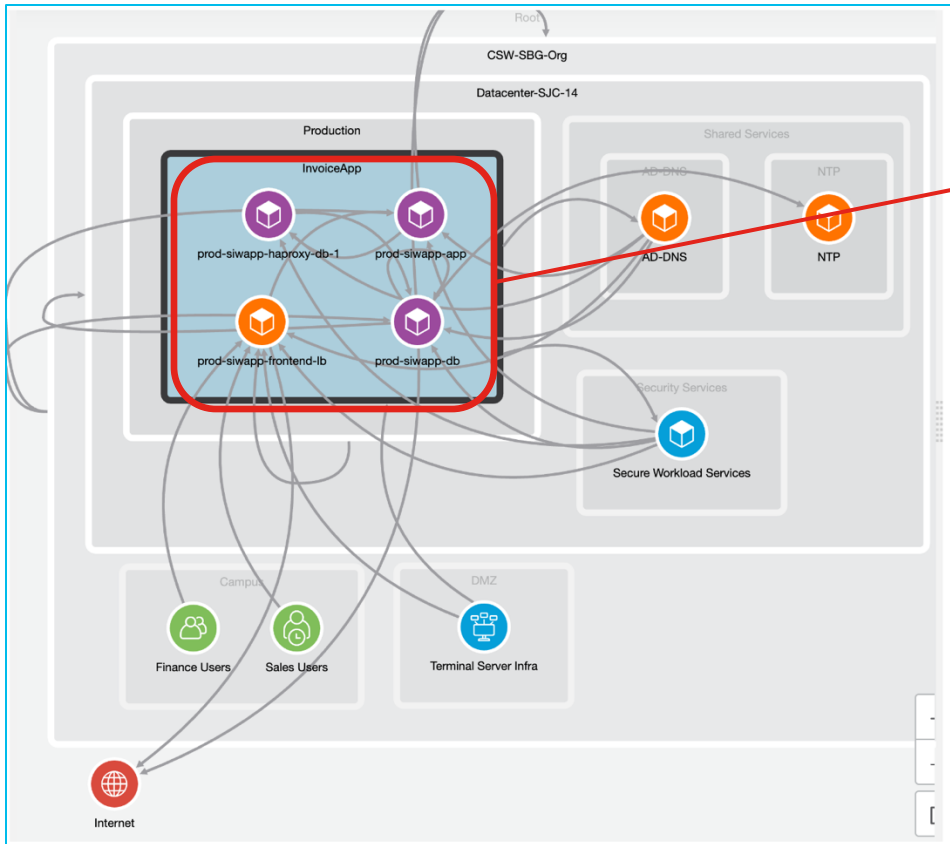
Edit Cluster Query (5) | View Cluster Details

Services 0 | Pods 0 | Workloads 4 | IP Addresses 0 | Neighbors 1

prod-siwapp-haproxy-db-1

Policy Discovery & Analysis and Enforcement

Application Workloads Classification



Scope	Full Name	Primary App	Query
InvoiceApp	Root:CSW-SBG-Org:Datacenter-SJC-14:Production:InvoiceApp	Invoice-App	✻ **Application** = **InvoiceApp** **or** ✻ **orchestrator_Application_vC** = **Invoice_app**

Services 0 | Pods 0 | Workloads 9 | IP Addresses 0 | View Scope Details

9 of **9** inventory

Orchestrator Type	Hostname	Address	OS	✻ Environment	✻ Location	✻ orchestrator_system/workload_type
dns, infoblox, vCenter	prod-siwapp-haproxy-app-1	192.168.1.3	CentOS	Production	Datacenter	machine, vm
vCenter	prod-siwapp-db-2	192.168.1.9	CentOS	Production	Datacenter	vm
infoblox, vCenter	prod-siwapp-app-1	192.168.1.4	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-db-4	192.168.1.12	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-app-2	192.168.1.5	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-haproxy-db-1	192.168.1.7	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-db-3	192.168.1.10	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-db-1	192.168.1.8	CentOS	Production	Datacenter	vm
vCenter	prod-siwapp-app-3	192.168.1.6	CentOS	Production	Datacenter	vm

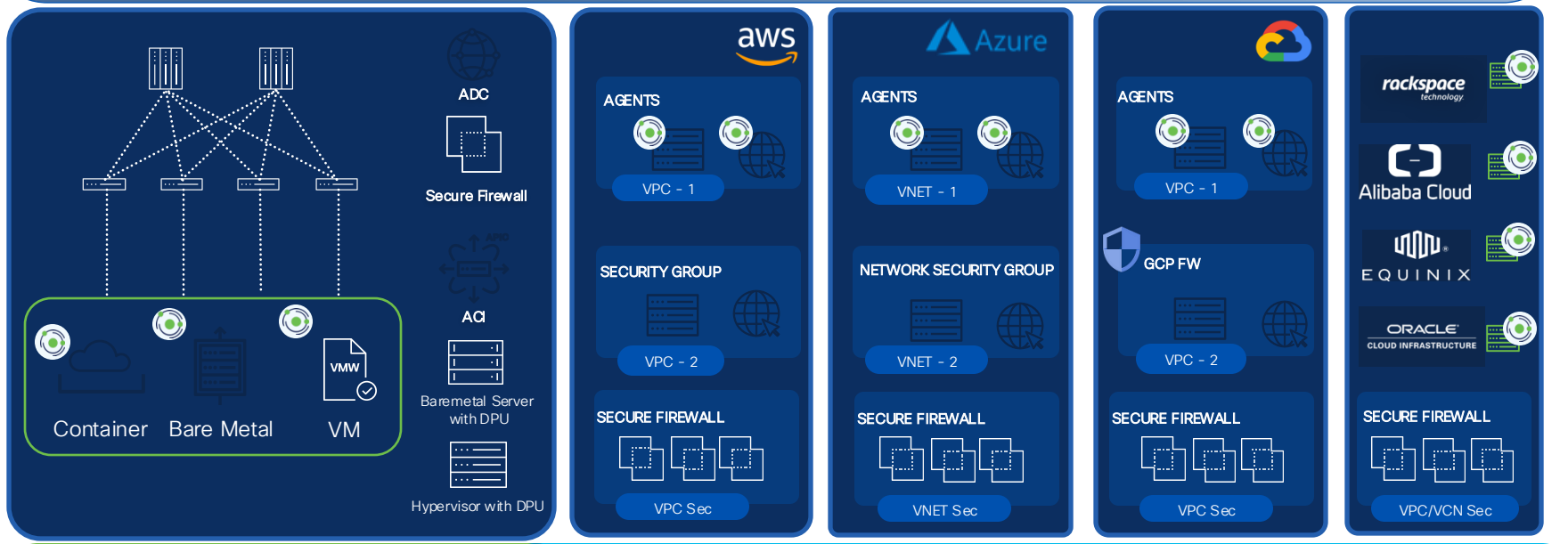
Cisco Secure Workload

Any Infrastructure, Any location, Any Application, Anywhere

Agent

- Anywhere
- Windows Desktop
- Windows Server
- IBM AIX
- Oracle Solaris/Linux
- Centos, Rocky, Alma Linux
- Ubuntu, Debian
- SUSE, RedHat
- Amazon Linux
- OpenShift/K8s

Consistent microsegmentation from on-premises to the cloud



On Premise | Public Cloud

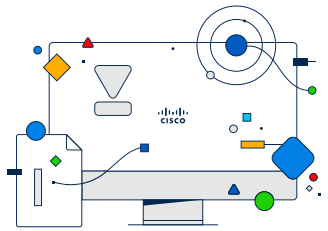
Bare Metal Servers | Virtual Machines | Containers

Agentless

- On-Prem
- Loadbalancer (ADC)
- Secure Firewall
- ACI
- NVIDIA DPU
- Cloud
- Security Group (AWS)
- Network Security Group (Azure)
- GCP Firewall (GCP)
- Secure Firewall

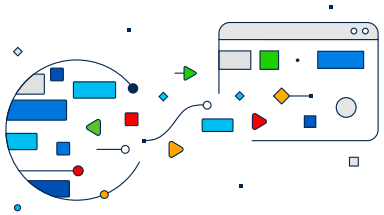
Kubernetes - The Universal Standard To Deploy All Applications

A Hybrid and Multi-Cloud Infrastructure Layer Supporting Both Modern & Legacy Workloads



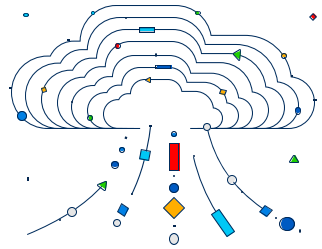
Legacy VM Workloads

Seamless Migration and Integration



Modern Applications

Rapid Deployment and Scaling.



AI / ML Workloads

High Performance and Adaptability.

Kubernetes

OPENSHIFT

Bare Metal

vmware by Broadcom

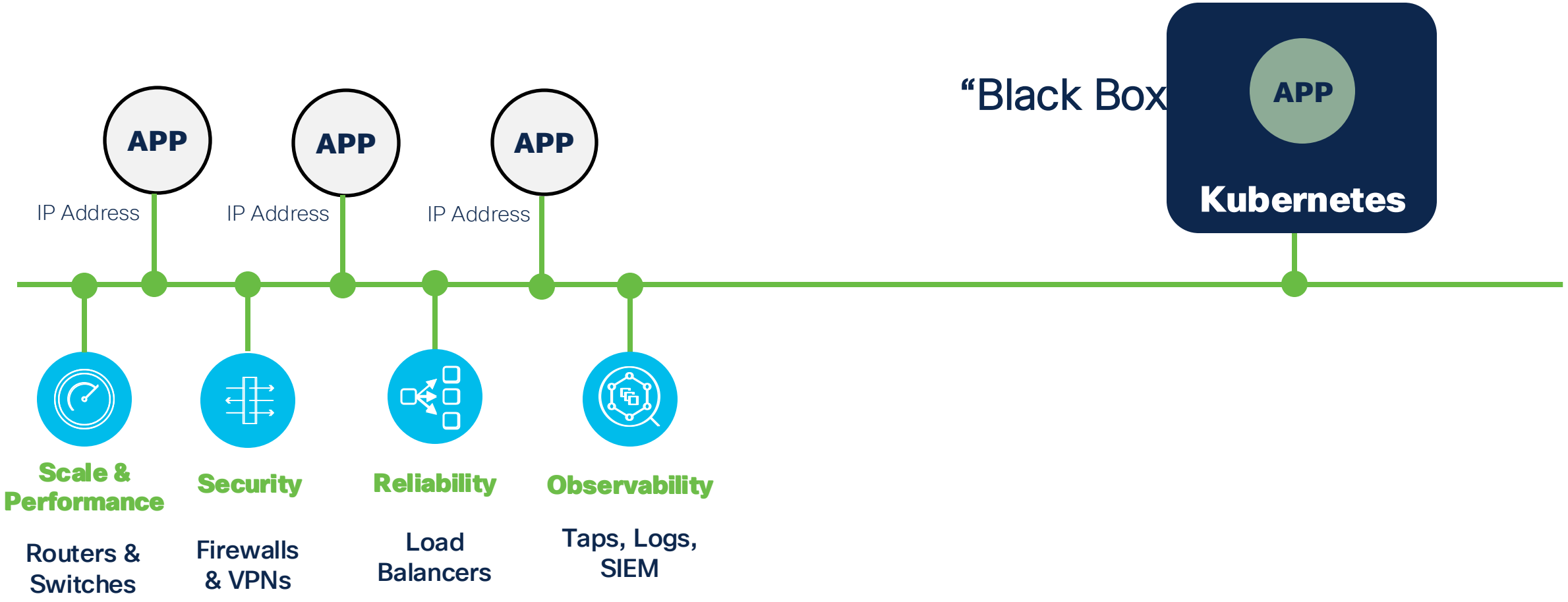
aws

Azure

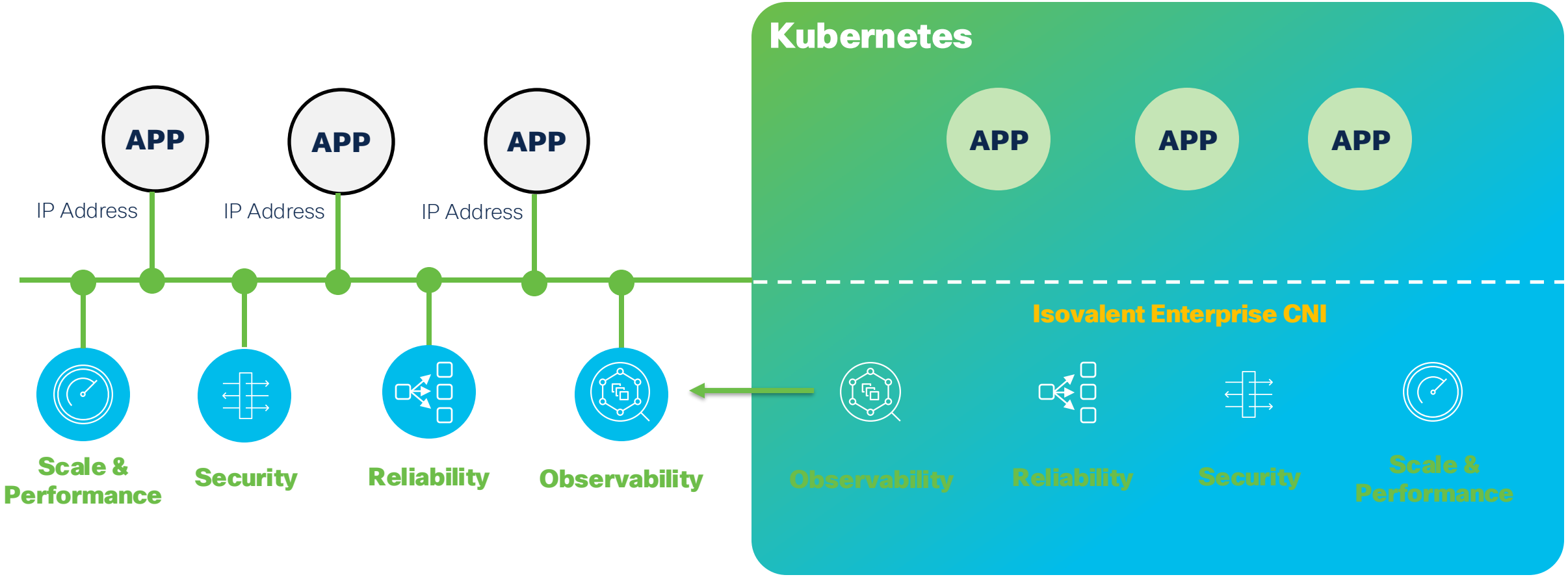
Google Cloud

The Network has Always Been Essential to Enterprise Requirements

But Kubernetes Environments Are Challenging for Enterprise Infrastructure Teams

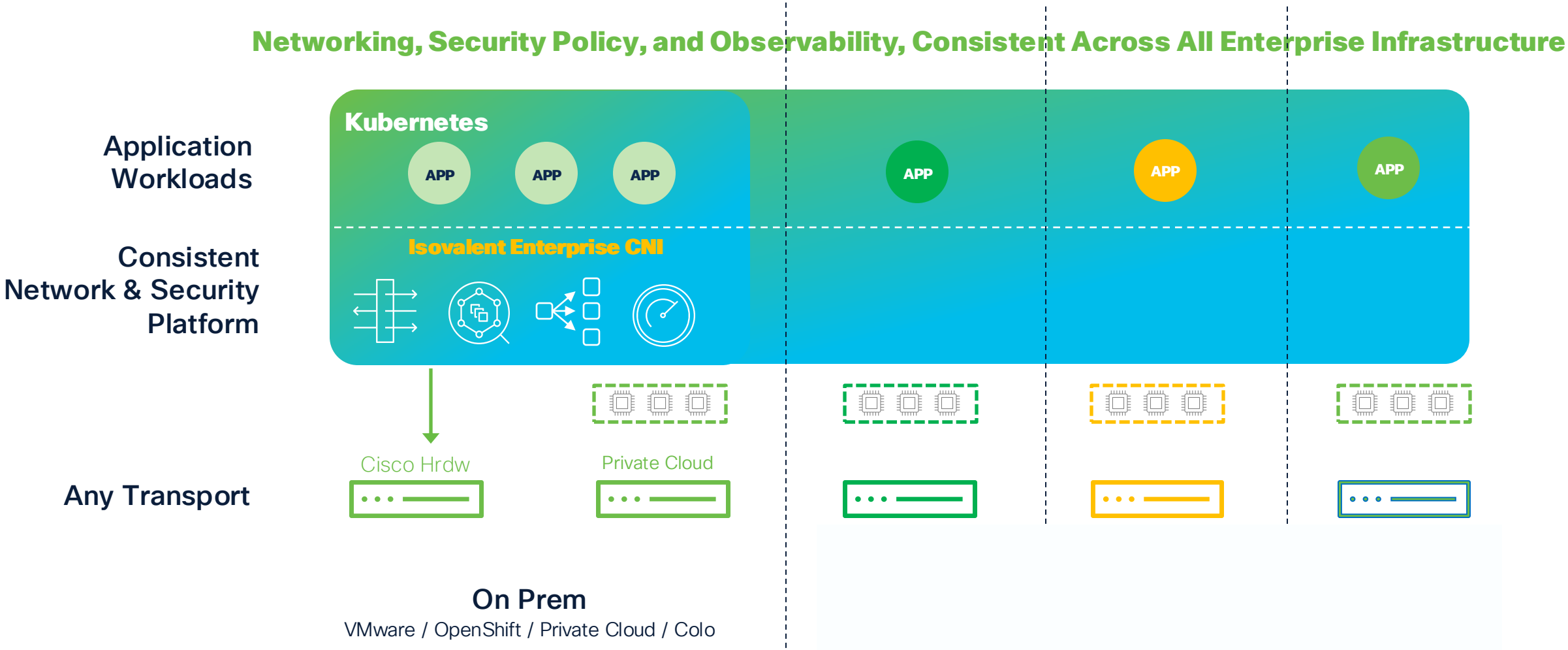


Isovalent Enterprise Regains the Visibility and Control Enterprises Need

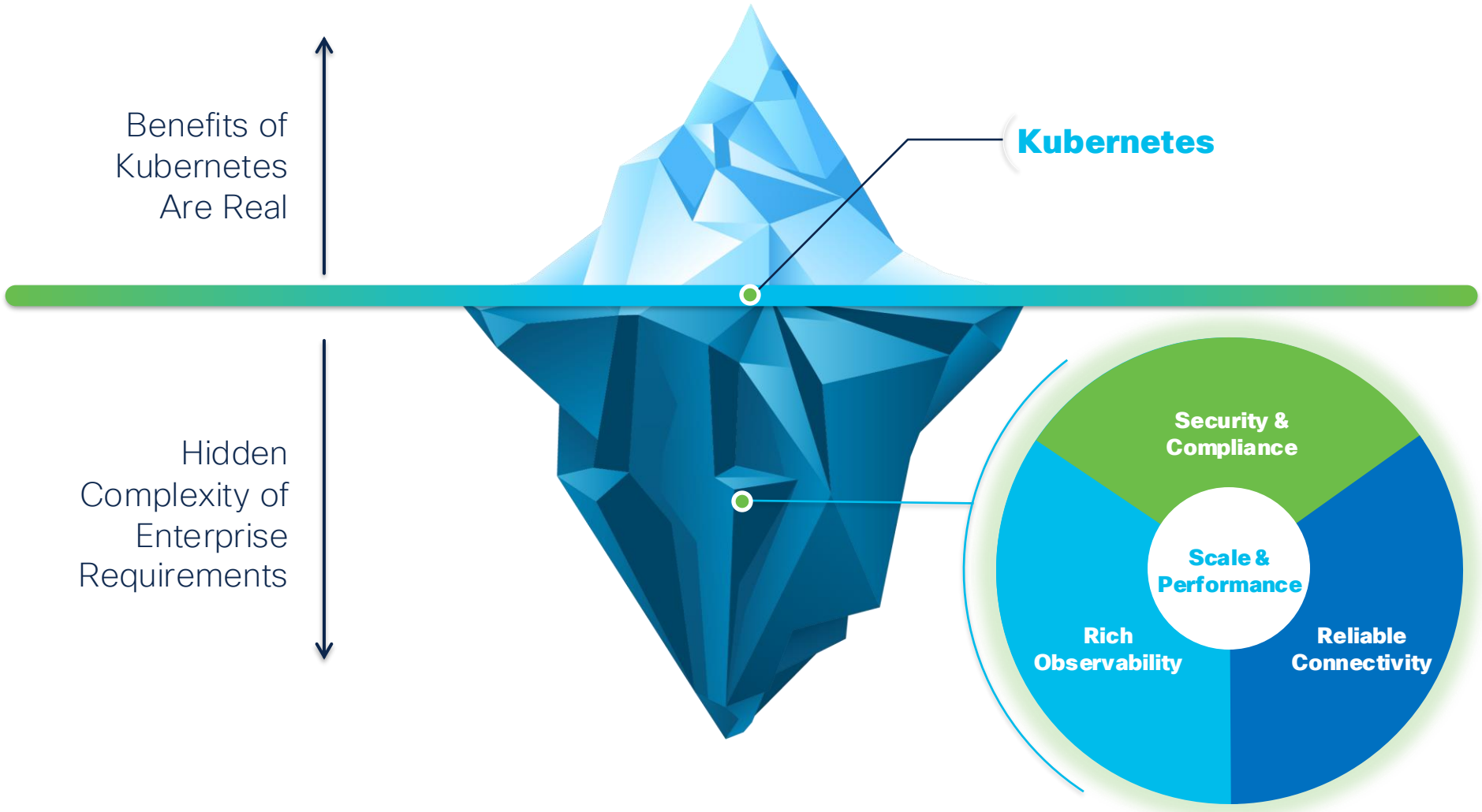


This is Why Cisco Acquired Isovalent.

It's the Technology Foundation for Cisco's Hybrid & Multi-Cloud Networking & Security Strategy



Why Enterprises Select Isovalent Enterprise Platform



SECURE

Thank you

cisco GSX


CISCO