Don’t AI Networks Design Themselves?

CCDE and AI Infrastructure

Mark Holm – CCDE Exam Program Manager
CCIE #34763, CCDE #20160020

Zig Zsiga – Principal Architect
CCIE #44883, CCDE #20160032

Rick Bauer –Technical Solutions Architect
CCIE #9482, CCDE #20140008

Cisco Webex App

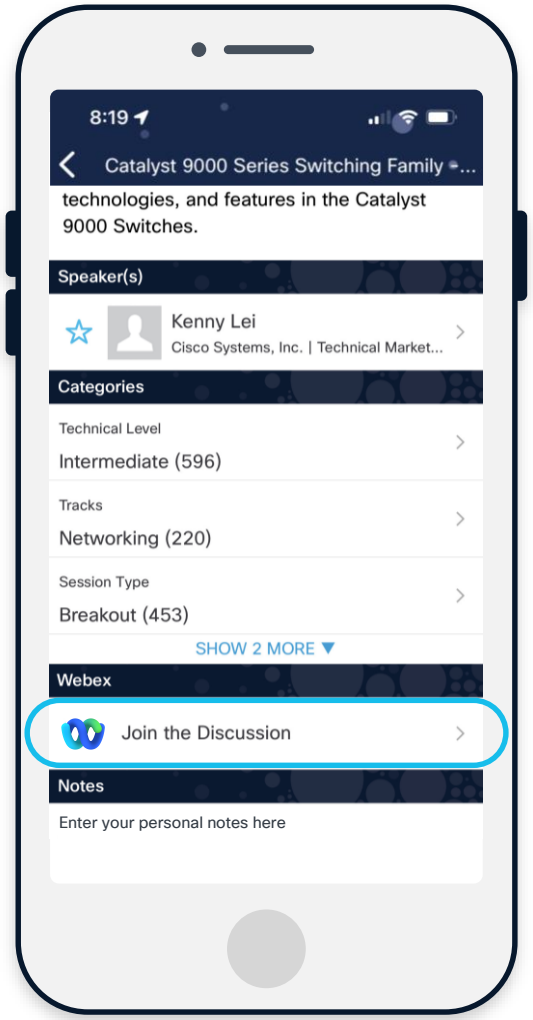
Questions?

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

How

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

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



Agenda

01 Introduction

02 CCDE primer

03 Making design decisions

04 AI network design use cases

05 How this ties to the CCDE

06 Q&A

Introduction

What CCDE is and what it isn’t

What CCDE is

- Focuses on which technologies fit where, and why.
- 99.9% vendor-agnostic *)
- Translate both business and technical requirements into designs

What CCDE is not

- Implementation and/or configuration, troubleshooting or operations
- Data sheets
- Product specifics
- Software release specifics
- A business test

Session introduction

- This is not a deep-dive session on AI infrastructure network design.
- We will talk design, but in context of the CCDE AI Infrastructure elective.
- CCDE basics are not covered in the session
  - We expect that you already know about the CCDE.
  - This specifically targets the CCDE AI Infrastructure elective.
- This is not about detailed AI model knowledge but how to support them in the network.

- Most of the session will be about a design use case.
  - There will be requirements and constraints (these are listed in the reference section slides).
  - We will guide you through the design choices and rationales.
  - Think of it as a live performance of a CCDE exam scenario.

- Find additional information and references at the end of the session presentation.

Definition of AI-enabled networks

- What are AI-enabled networks?
- How to they compare to traditional data centers?

- Could AI be used to build the design?

- Yes, but...
  - When designing networks you must consider a broad set of information.
  - One must also (often) read between the lines to gather what's needed.

- What are the goals and the customer requirements and expectations?

CCDE AI Infrastructure

How does this tie to the CCDE?

Good news!

The use cases we have covered today is structured very similar to what you will find in the CCDE.

If you feel comfortable (and like) doing information gathering, analysis, and make design choices based on all of it, you are already on your way!

Relating this to the CCDE

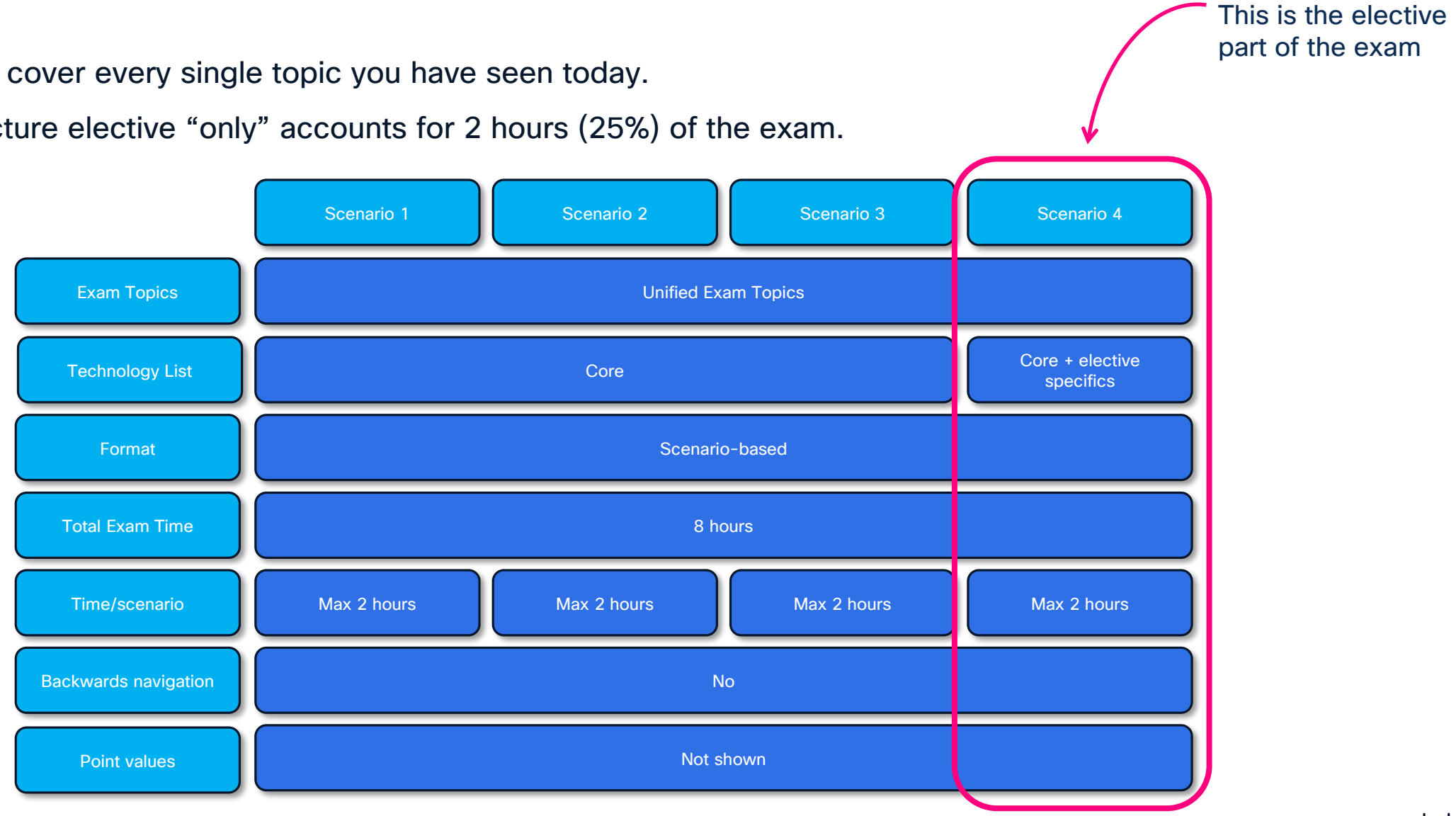
- Much effort is put into making the CCDE exam realistic and true to life.
- What you have seen here will be similar in the CCDE Practical exam.
- You will be given a lot of information – like what you have seen today.
- Aspects and considerations discussed today are used in the exam too.
- Think and act the same way as we have done throughout the session.

- CCDE is not necessarily about best practices.
- Don’t overrule the facts given to you (fighting the test)

- But don’t expect a 1:1 between this session and the exam.

Feeling a bit overwhelmed?

- The exam won't cover every single topic you have seen today.
- The AI Infrastructure elective "only" accounts for 2 hours (25%) of the exam.



Feeling a bit overwhelmed?

- That's understandable – it is a lot to think about.

- The topics we've discussed today are all valid and "fair game".
- ...but isn't a definitive list.
- Any specific delivery of the exam can contain other topics.
- Everything decision you make in the exam is about high-level design.
- No CLI, no configuration snippets, troubleshooting, or APIs to memorize.

- Keep an open mind and take your time reading and digesting the information.
- All the information you need to answer the questions will be there.

Making design decisions

Making network design decisions

- Most of the time, there are several ways to achieve the goal.
- Which approach is correct?

- Are there bad arguments for making a certain decision?
- If based on collected facts, not really.
- ...but was it made considering all the provided information?

- We will be covering a lot of ground today.

Constraints in network design

- Whether we like to admit it or not, there's always going to be constraints.
- Constraints come in many shapes and forms.
- Some are technical, some are business, others are vaguer.
- They will have an impact on the final design.

- As a designer, one of your most important tasks job is to identify them.
- Ensure that you talk to all relevant stakeholders.
  - Executives, operational staff, on-site staff.
- Do not make assumptions or take shortcuts.

Financial constraints

- There’s a saying in the design world: Money always win! (And yes, it is true!)

- Identify the needed features and functionality.
- Avoid gold plating.
- Regardless of how cool the design is, someone must pay for it.

- Consider available CAPEX and OPEX.
- Is there a reasonable ROI?
- Is the TCO within the expectations?

Business constraints

- Business requirements will influence the outcome.
- If the final outcome does not meet the business' needs, the design has failed.

- Consider the physical environment – is there room?
- What are the cooling and power requirements? Can they be met?
  - Hint: GPUs tend to like lots of power and generate a good amount of heat.
- Is high-performance required full-time?
- Protecting data going in and out using solutions like DLP.
- Is the staff comfortable with the solution? Do they need training?
- Are sufficient services attached to meet the defined RPO or RTO?
- Are there compliance aspects involved?
  - For example, data governance, data validation, encryption of data at rest, data sovereignty, data classifications or local/global regulations.

Network constraints and requirements

- Networking: Routing protocols, overlays, topology choice, IP/IPv6 or dual-stack,
- Data transport
  - Link speeds
  - SFP/QSFP vs. DAC
  - Single- or multi-site
  - Link multiplexing, CWDM, DWDM
- Greenfield or brownfield deployment?
- How to attach to the existing network?
- Storage: data access latency, replication strategy, backup.
- Redundancy and resiliency.

- Moving AI workloads through use of automation.

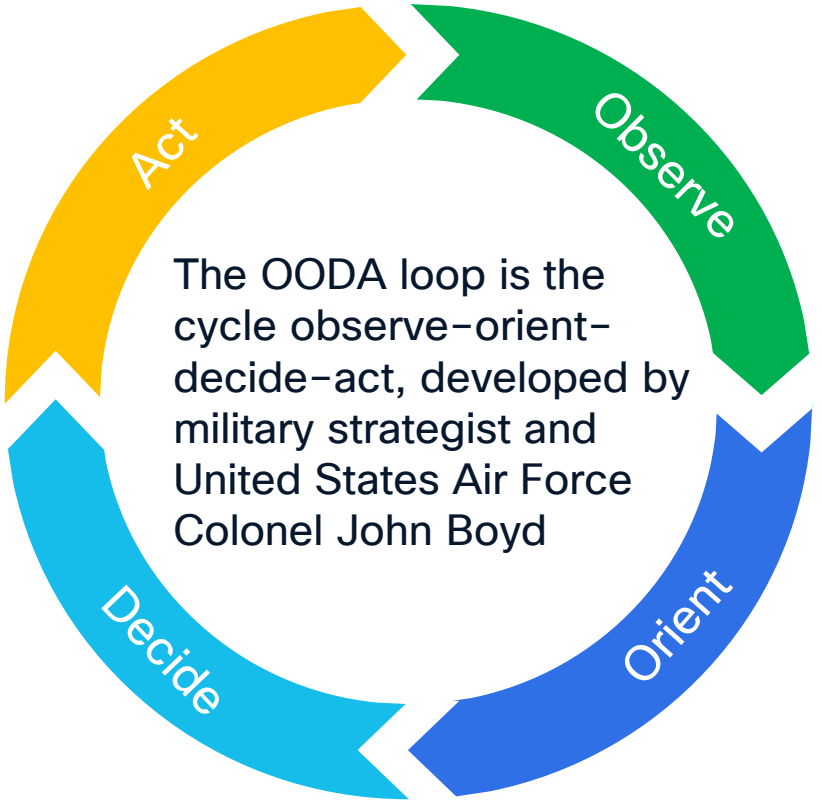
Security and scalability

- Data security and integrity validation.
- Validation of data input to the models.
- Solution security
  - Protecting against rogue use, protecting the back ends, web application firewalls/NGFWs, etc.
- Must elastic scaling using cloud be available?
- Must the data be placed close to the users for faster response times?

AI Infrastructure design use cases

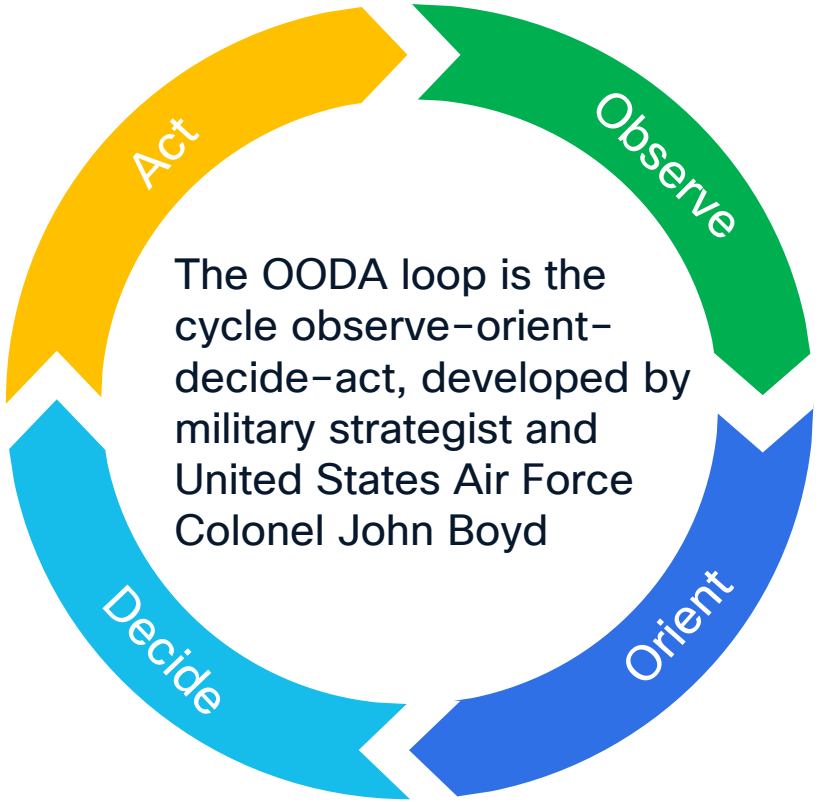
Designing AI networks

- There are significant differences in technical requirements depending on use case.
- As a designer, it's your task to find the best fit.
  - Ask questions.
  - Analyze information.
  - Ask some more questions.
  - Decide.
  - Rinse and repeat.
- Are there any givens?
  - Greenfield / brownfield deployment
  - Single-site or multi-site?
- Latency is key when training a model.
- Understand the AI workload life-cycle (training, inferencing, delivering)
- On-prem, cloud, or hybrid
  - Cost likely to be a huge factor here



Designing AI networks

- As you progress, determine if/when there are trade-offs
- Sometimes a decision can limit the technical capabilities, for instance a decision not to buy a license.
- Someone else could make a decision for you.
  - Things got too expensive.
  - Budget is not there.
  - Staff training not an option.
  - ...and so on
- Accept the “new reality” and continue your work.
- Keep in mind: Money is king!
- So is the people aspect (staff not familiar with a technology).
- Do not make decisions based on personal preference alone.
- Make decisions based on facts you have been given.
  - Apply experience and technical knowledge to catch “This won’t work” situations.



AI Workload Requirements

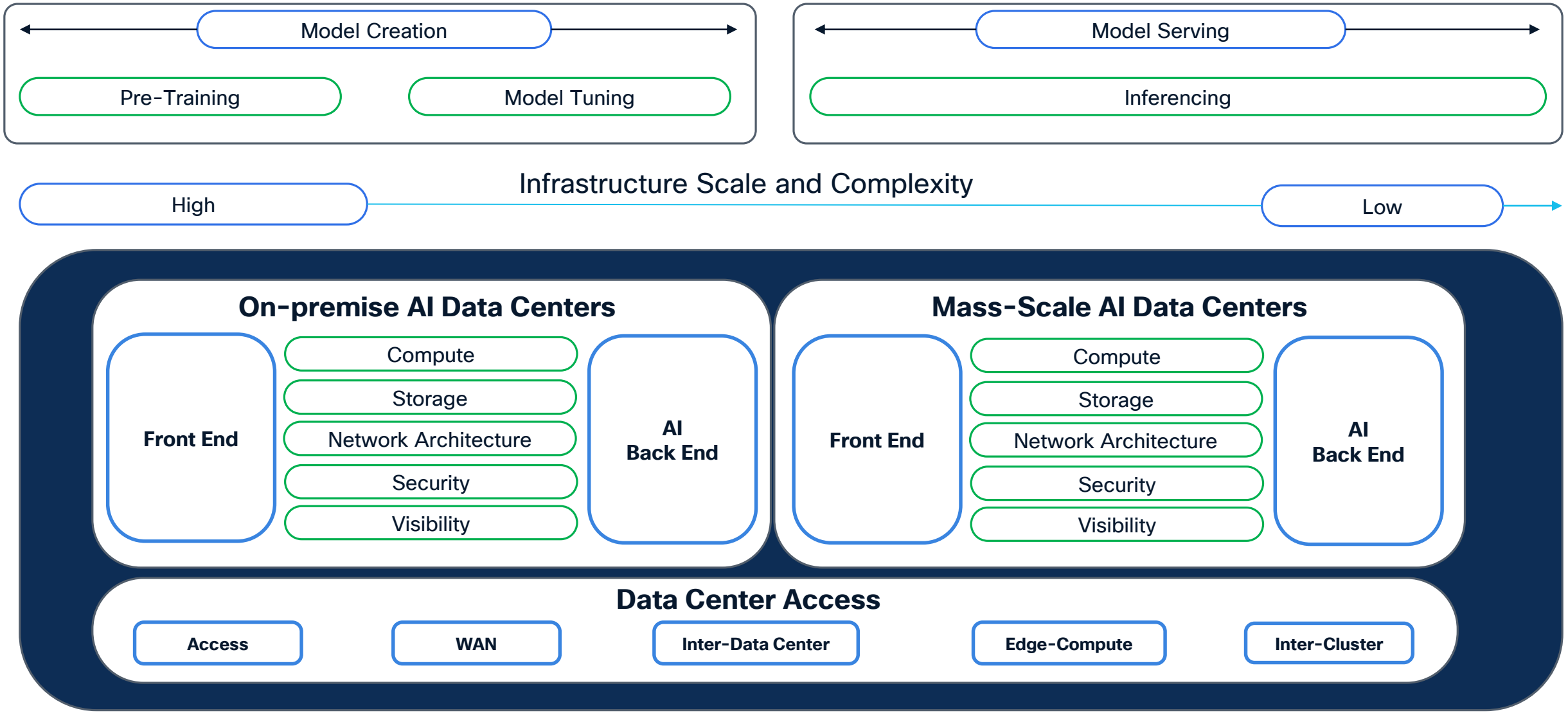
Technical Requirements

- Zero Loss
- Low Latency
  - Round Trip Time (RTT) 4.5 microseconds (µs)
- High Bandwidth
- Optimal Path Selection
- Congestion Avoidance and Notification
- Flow Control
  - few large flows – worst flow matters!
- Quality of Service (QoS)
- Security
- Observability (Performance, Quality, Availability)
- Job Completion Time (JCT)
- Storage
- Data Center (Power, Space, Cooling)
- Scalable and Predictable

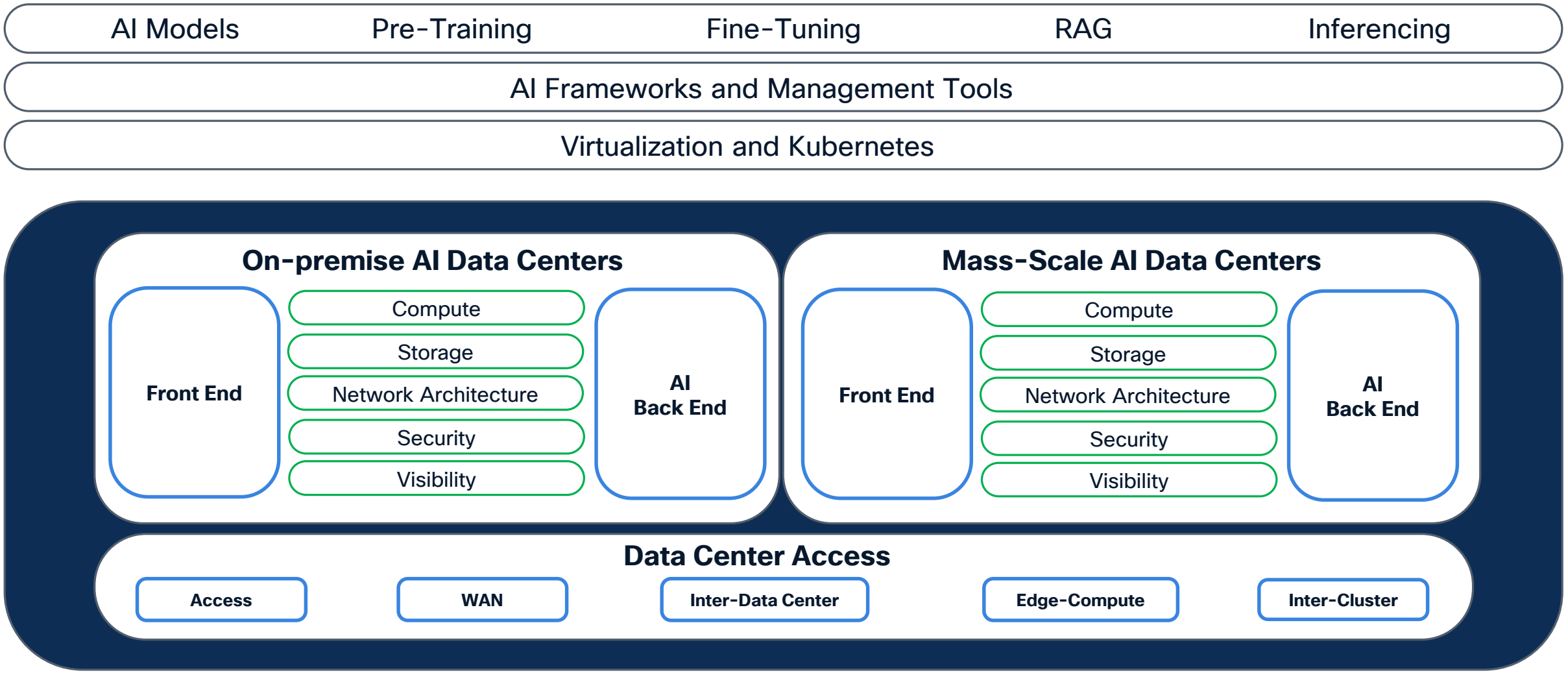
Other Requirements

- Data Concerns
  - Classification
  - Sovereignty
  - Data Lifecycle
  - Intellectual Property (IP)
  - Data Loss
  - Data Quality
- Money (CAPEX / OPEX)
- People (Staff, Training, Skillset, etc)
- Support (Operational Sustainability)
- Agile or Lean IT

Understanding the AI Model Lifecycle



AI Modeling and Infrastructure Requirements



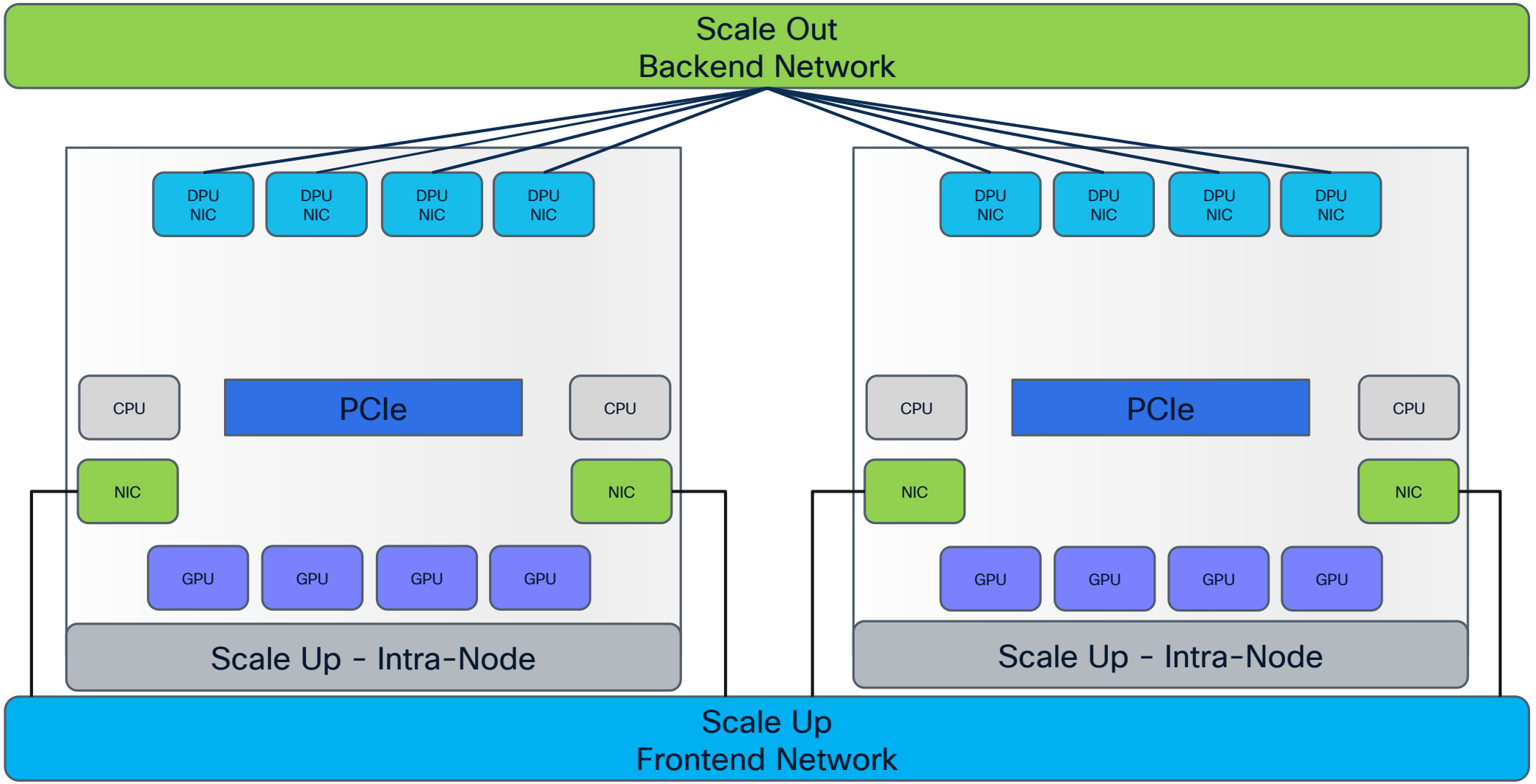
InfiniBand vs RoCEv2

	InfiniBand	RoCEv2
Round Trip	2µs	5µs
Forwarding model	Local ID	IP
Flow Control	Credit Based	PFC/ECN/DCQN
Load Balancing	Per-Packet	ECMP DLB

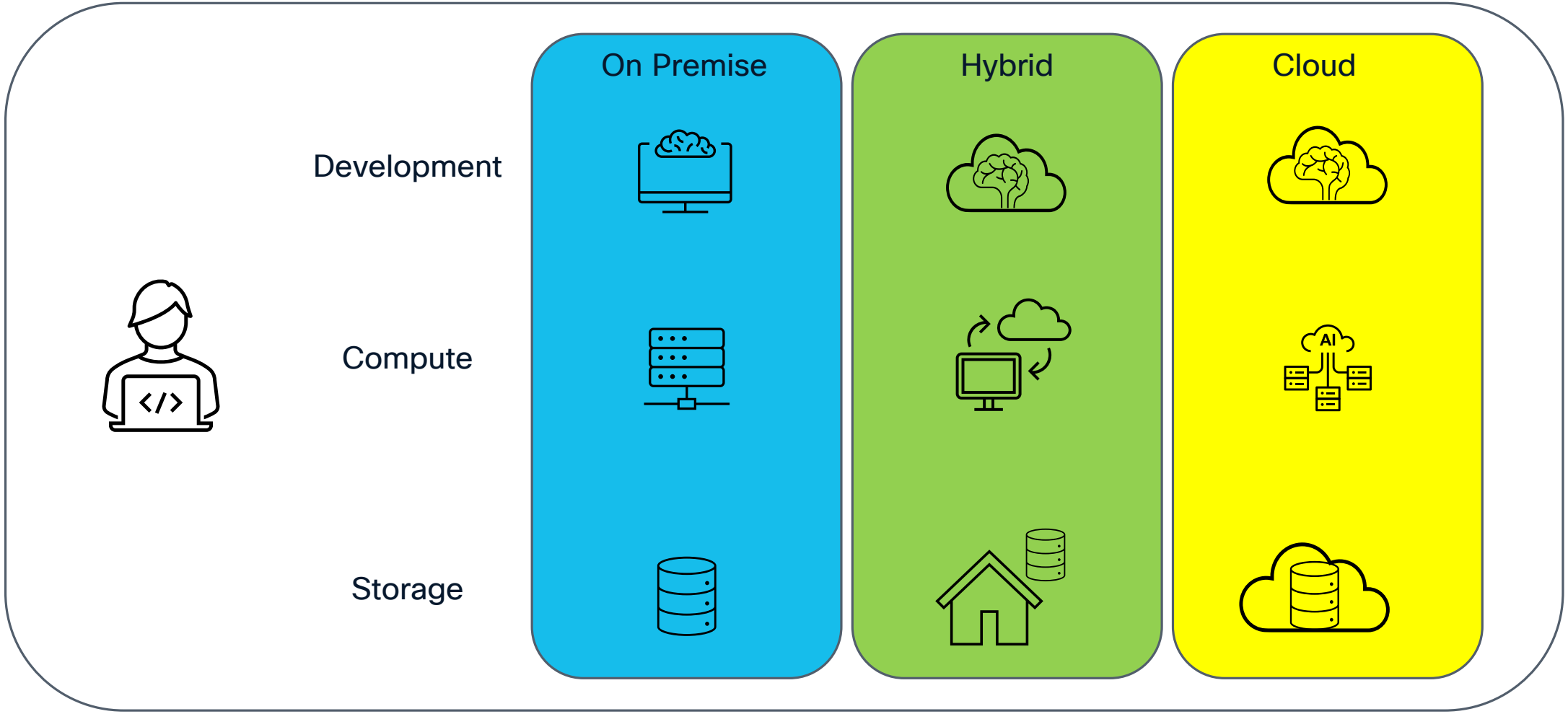
Massively Scalable DC Fabric

- Topologies supported – Two-Tier, Three-Tier, and Rail
- Layer 3 Design
- Routing Protocol - EBGP peering over Point-to-Point links and different ASNs at each Tier
  - Link-State protocols can be used but there are challenges (flooding, update propagation, dynamic topology detection)
- ECMP with Dynamic Load Balancing (DLB)
- IPv4 and IPv6 support (IPv4 NLRI with IPv6 Next-hop) RFC-5549
- RDMA Support
  - RoCEv2 – Quality of Service (QoS), Priority Flow Control (PFC), Explicit Congestion Notification (ECN)
- Continuous Operation
  - Maintenance Mode (Graceful Insertion and Removal (GIR) or equivalent)
  - Nonstop Forwarding (NSF)
  - Graceful Restart (GR)
  - Hitless Upgrades

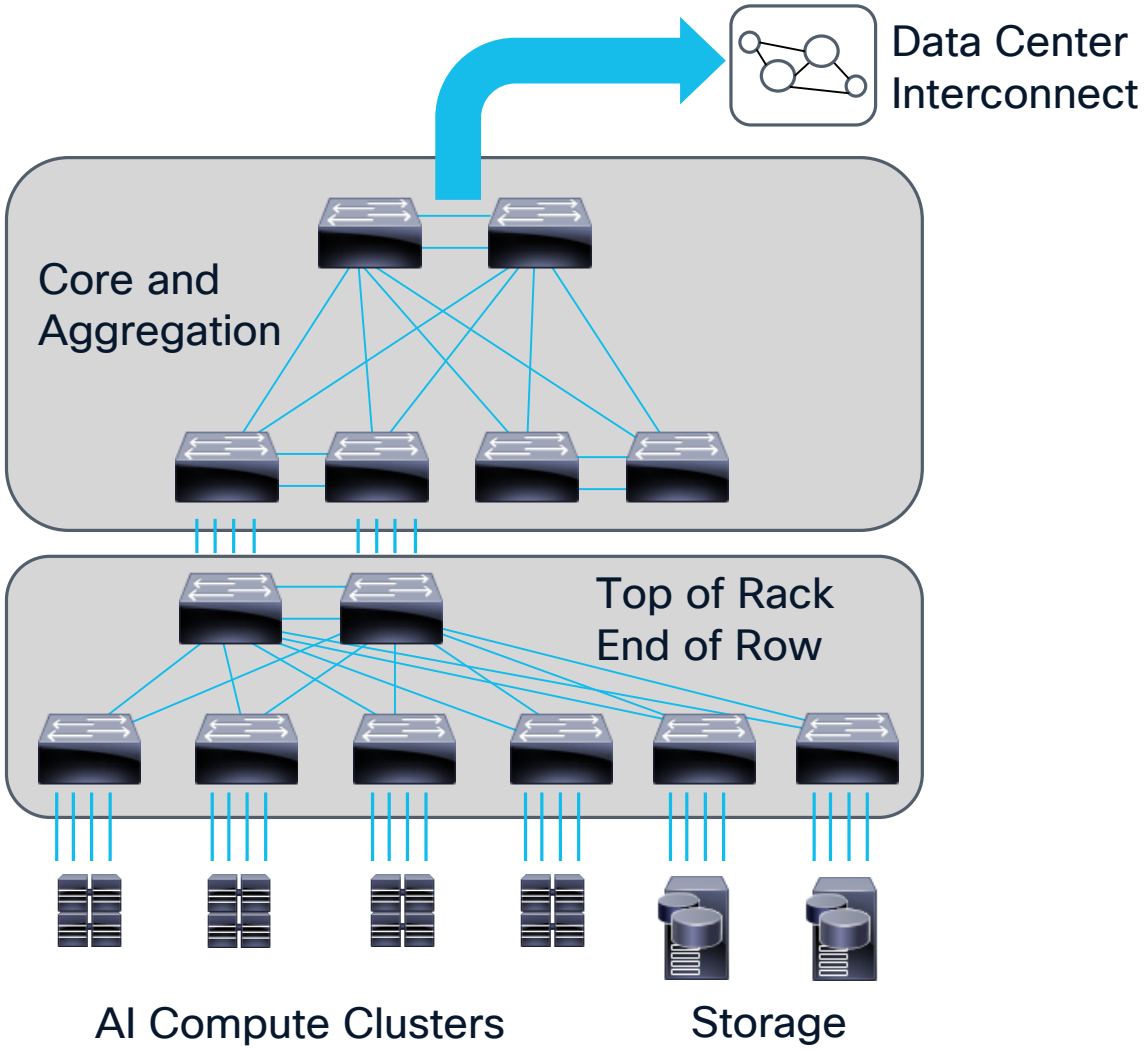
xPU Server Diagram

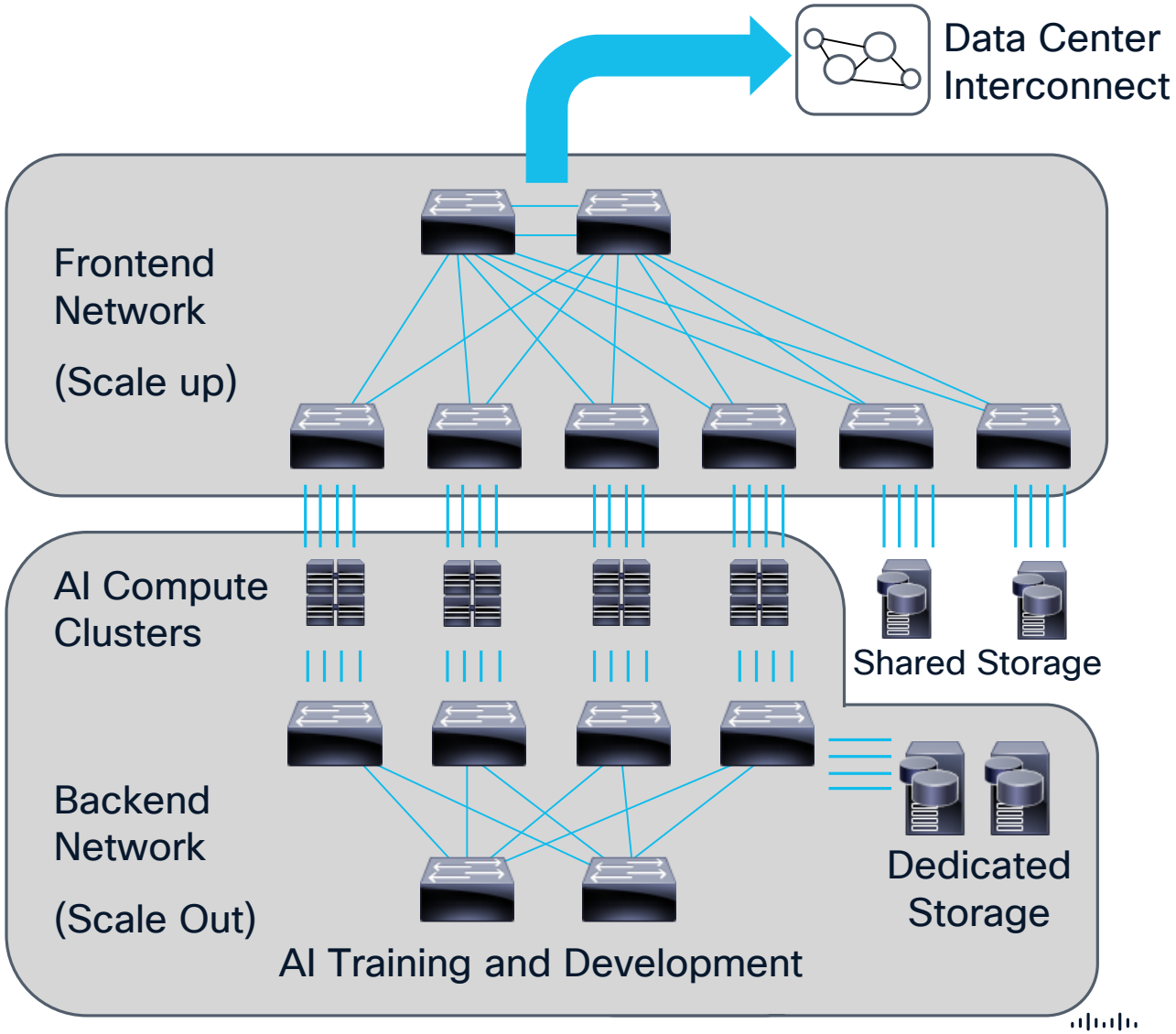


Where do I put all this stuff?

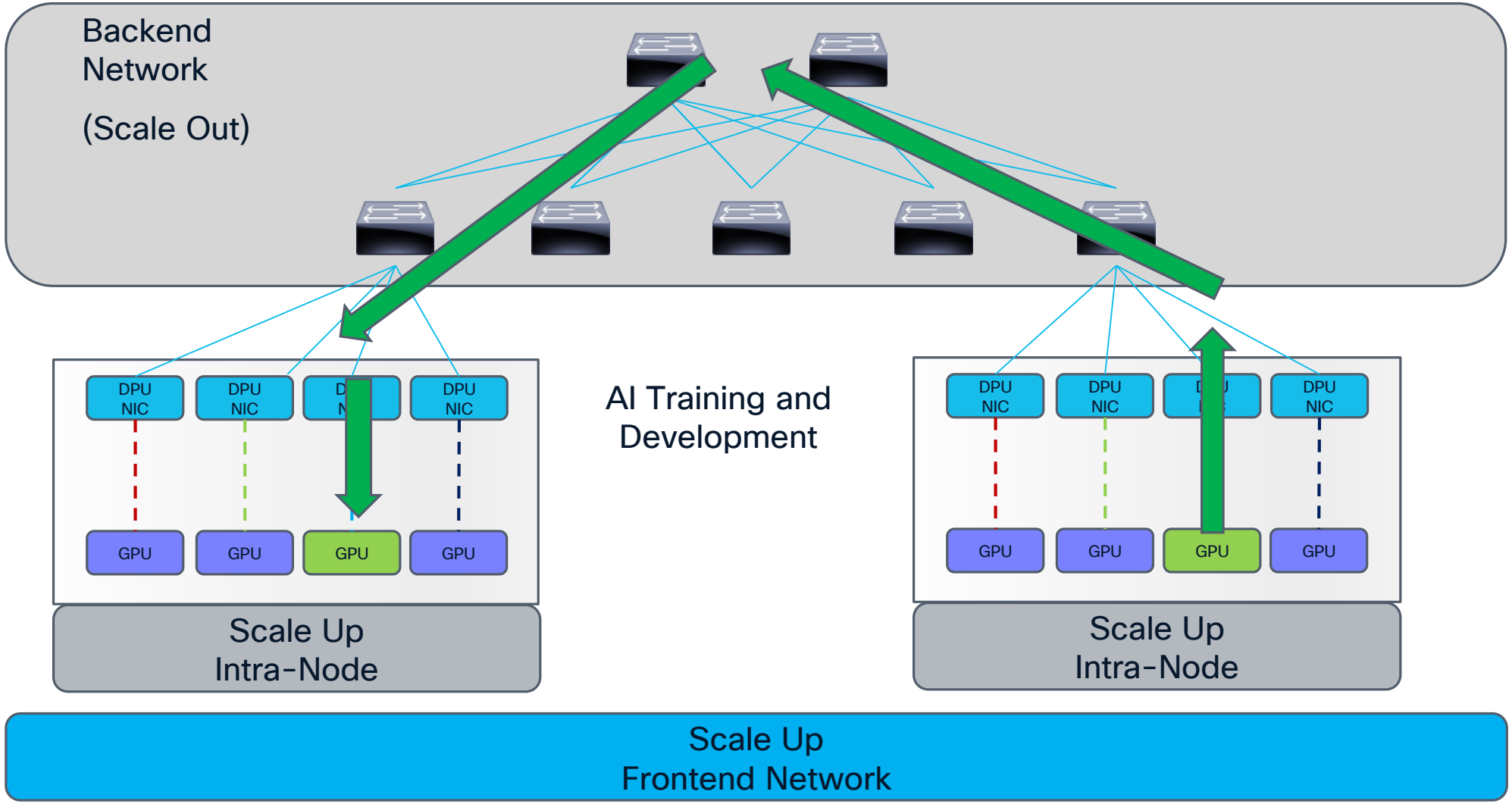


On Premise – Brownfield Design or Greenfield?

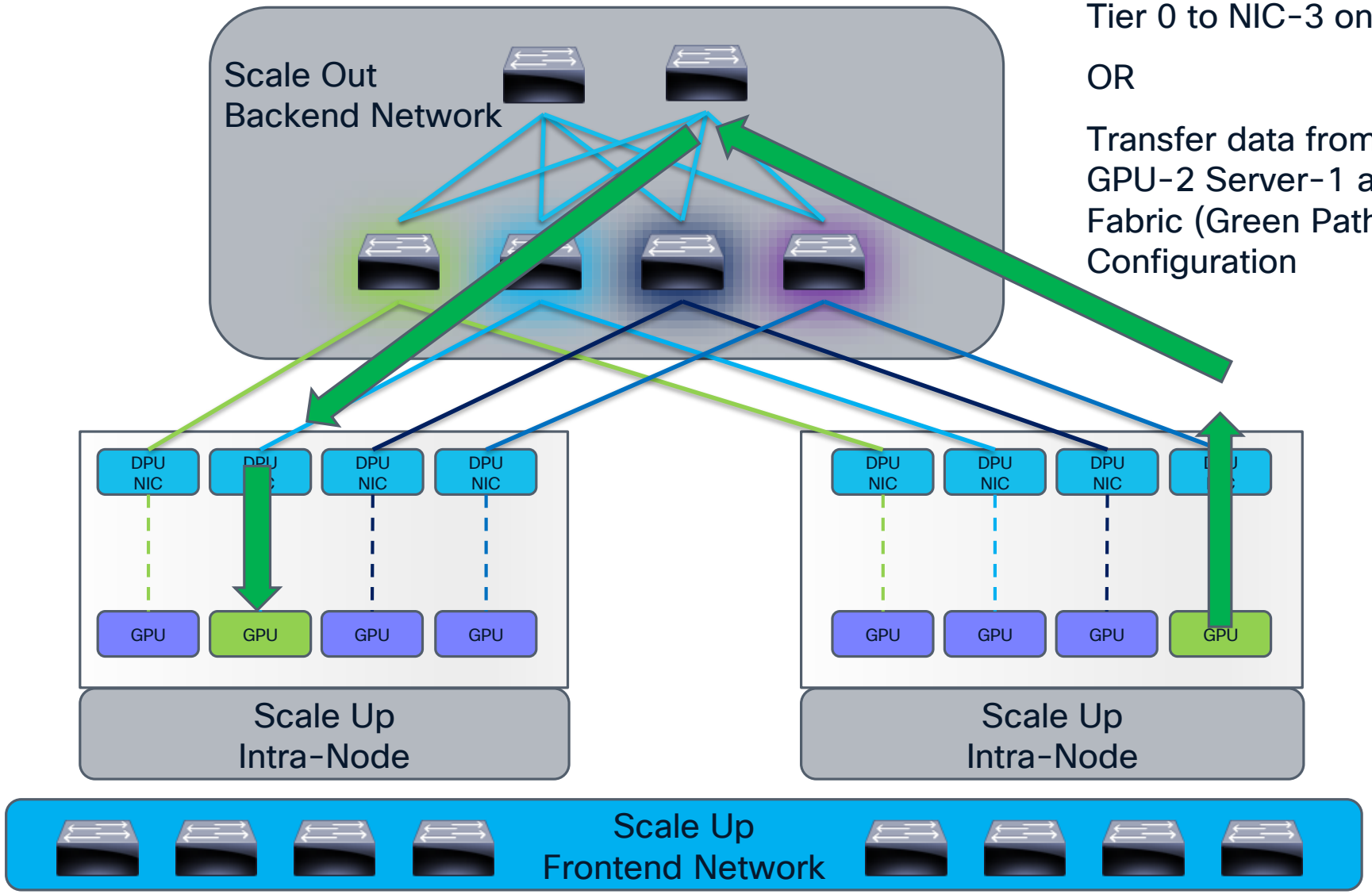




Massively Scalable Fabric Topology (Backend)



Rail Topology Massively Scalable Fabric Topology (Backend)



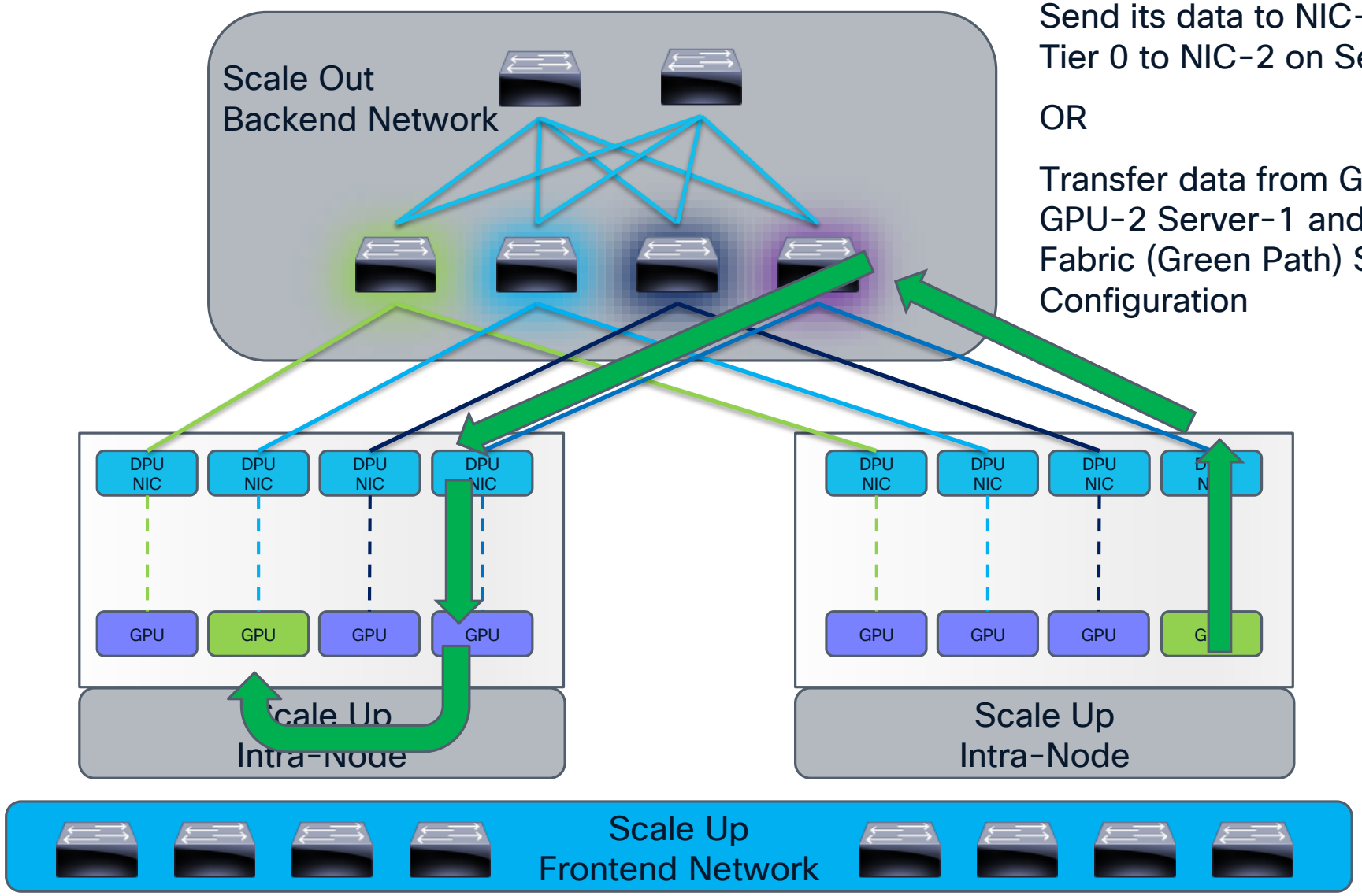
To communicate with GPU-2 on Server-1 Server-2 GPU-4 can;

Send its data to NIC-4 which will traverse Tier 0 to NIC-3 on Server-2 (Green Path)

OR

Transfer data from GPU-4 Server-2 to GPU-2 Server-1 and cross the Intra-Node Fabric (Green Path) Server Based Configuration

Rail Topology Massively Scalable Fabric Topology (Backend)



To communicate with GPU-2 on Server-1 Server-2 GPU-4 can;

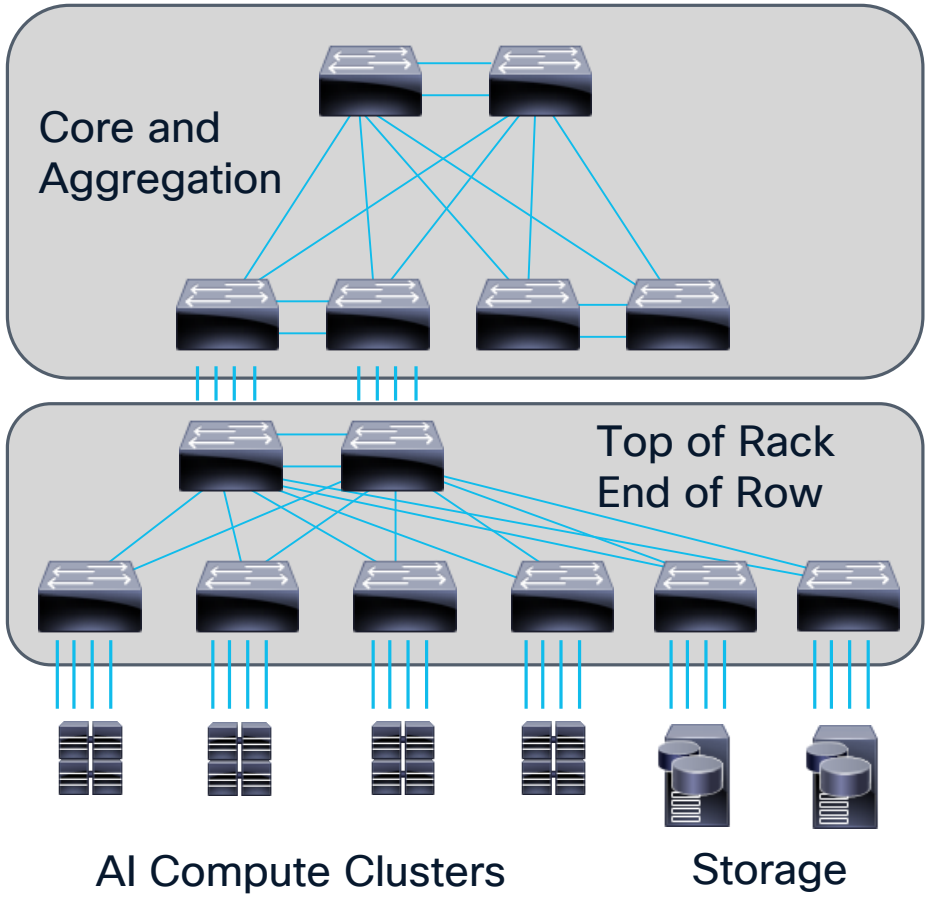
Send its data to NIC-4 which will traverse Tier 0 to NIC-2 on Server-2 (Green Path)

OR

Transfer data from GPU-4 Server-2 to GPU-2 Server-1 and cross the Intra-Node Fabric (Green Path) Server Based Configuration

Scenario

Retrofit Network Design



Technical Requirements

- Backend Training and Inferencing – GPU to GPU clustering RMA network
- Low Latency 4.5 microsecond (µs) RTT
- High bandwidth 400/800Gig
- High Reliability
- Scale up Network – GPU connectivity tensor parallelism
- Scale out Network – connect 10,000 GPU together UEC domain
- Servers saturate at line rate (400Gig) simultaneously
  - Traffic is synchronized
  - Few Flows per host but very large and sensitive to delay and loss Slowest flow matters
  - Slowest flow matters
- Data Plane tuning – QoS, ECN, PFC, and MDT to monitor
- Network Time – PTP/NTP
- Non-disruptive Operations

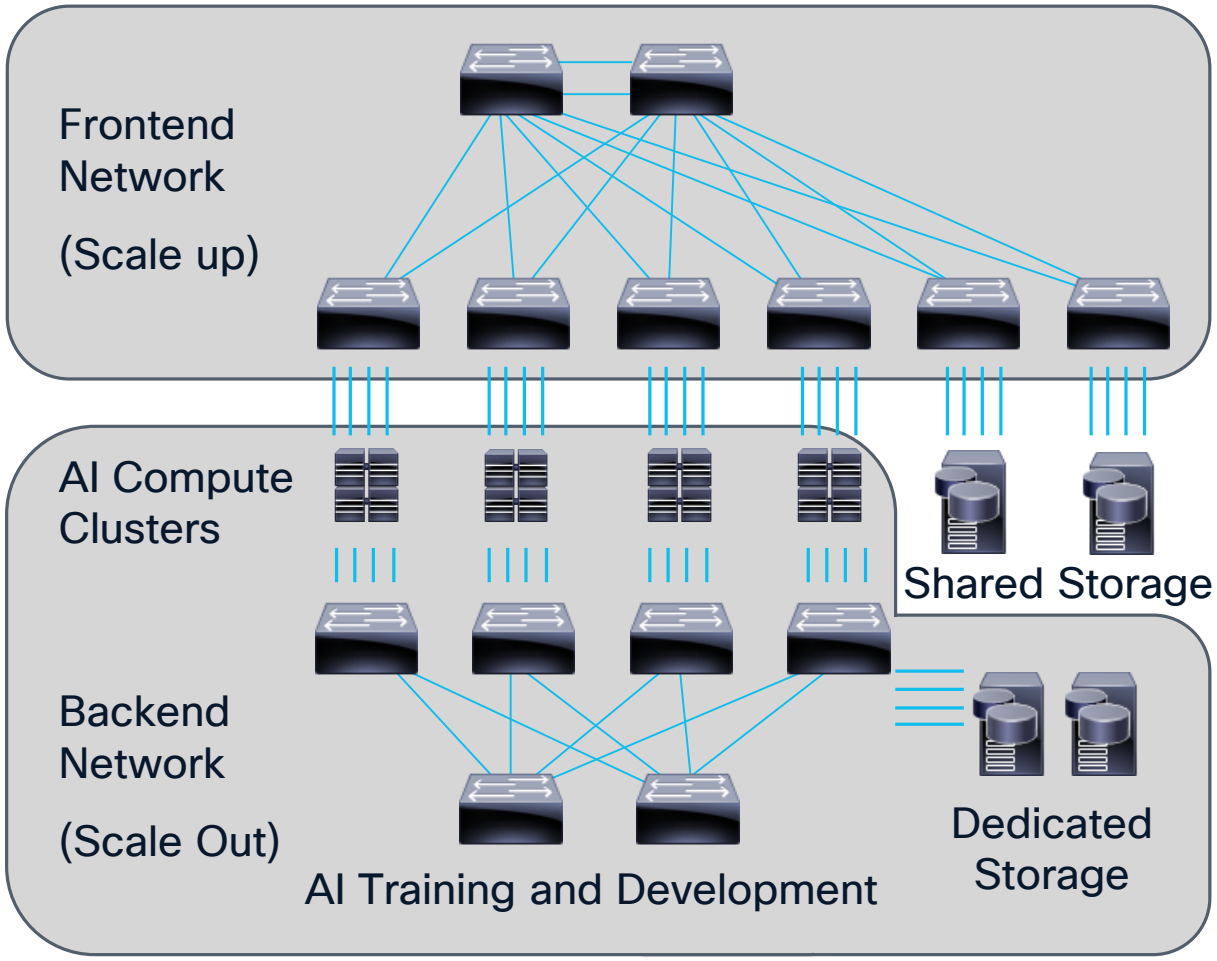
Design Benefits

- Low cost to implement
- Little or No changes to operations
- Tribal Knowledge

Design Challenges

- Requires Spanning-Tree for loop prevention
- Slow Convergence – Layer 2 – SPT Layer 3 – Routing Protocols (Timers and BFD)
- First Hop Redundancy Protocol required (HSRP/VRRP/GLBP)
- Load Balancing – L2 SPT or L3 ECMP
- Congestion Avoidance – (TCP Windowing and Slow Start)
- Network Queueing and Scheduling
- Loss Detection at the end host (TCP Windowing and Slow Start)
- Large Broadcast and failure Domains
- Configuration Complexity (20+ Protocols)
- Unique configurations per device

Massively Scalable Fabric Design



Technical Requirements

- Backend Training and Inferencing – GPU to GPU clustering RMA network
- Low Latency 4.5 microsecond (µs) RTT
- High bandwidth 400/800Gig
- High Reliability
- Scale up Network – GPU connectivity tensor parallelism
- Scale out Network – connect 10,000 GPU together UEC domain
- Servers saturate at line rate (400Gig) simultaneously
  - Traffic is synchronized and bursty
  - Few Flows per host but very large and sensitive to delay and loss Slowest flow matters
  - Slowest flow will affect JCT
- Data Plane tuning – QoS, ECN, PFC, and MDT to monitor
- Network Time – PTP/NTP
- Non-disruptive Operations

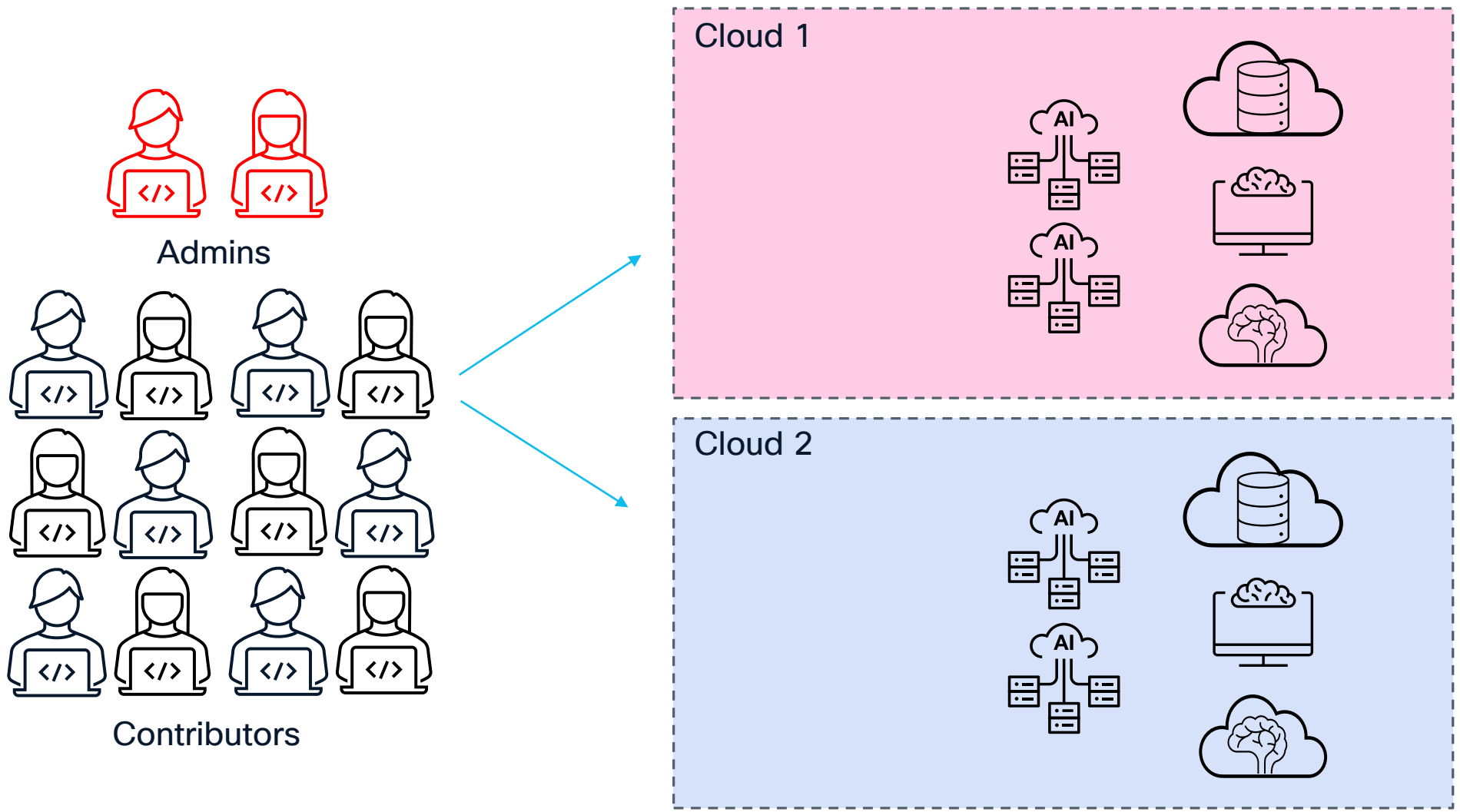
Design Benefits

- Scheduled Fabric Load Balancing and flow reassembly in hardware
- RoCEv2 – RDMA from the network
- Explicit Congestion Notification ECN (WRED & AFD Approximate Fair Drop)
- Priority Flow Control PFC
- ECMP DLB – dynamically adjust path selection based on current conditions

Design Challenges

- Cost
- Operational Change
- DC Impact – Power, cooling, space, etc.

AI Cloud Design



Q&A

Complete Your Session Evaluations

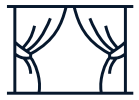
Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

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

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

Continue your education



Visit the Cisco Showcase for related demos

Book your one-on-one Meet the Engineer meeting

Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact us at: Webex

Resources

- RoCE networks for distributed AI training at scale
- NVIDIA AI Enterprise: Software Reference Architecture
- Cisco Data Center Networking Blueprint for AI/ML Applications
- Benefits of Remote Direct Memory Access over Routed Fabrics
- Introduction to RoCEv2 Network

Thank you

cisco Live !


CISCO

Reference slides

What is CCDE?

- CCDE assesses design principles to deliver robust network designs
  - Cost-effective, scalable, fast, available, secure and manageable networks.
  - But mainly, network designs which meet the given requirements!
- Analyzing and optimizing existing and new environments.
- Design choices based on business and technical requirements.
- Cost/benefit analysis.
- Keeping financial concepts in mind: CAPEX, OPEX, RCO, ROI.
- Ensuring operational goals (RPO, RTO) are met.

- Most importantly, asking the “Why?” questions.

CCDE is all about high-level design

Examples of HLD topics

- Topology decisions.
- Identifying protocols/features to implement or enable.
- Build high-level implementation or migration plans.
- Adjusting network architecture/design to support business changes.
- Balancing cost and “get-the-latest-and-greatest gear”.
- Merging networks (acquisitions).
- Divestiture (selling off).
- Resiliency/redundancy levels.
- It is not a business test – you need deep technical knowledge.
  - You just apply your knowledge in a different way.

Reference slides

- Slides in this section are included for your reference and not covered in the session presentation.
- Currently work-in-progress

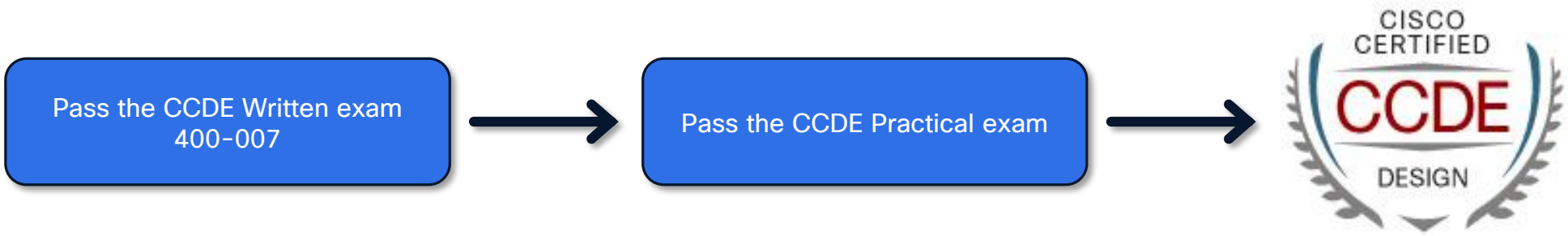
Starting your journey

- Must-read documents
  - Unified Exam Topics
  - Technology List(s)
- Unified Exam Topics
  - Describes the tasks you are expected to be able to do
  - Defines the domain weightings
- Technology lists
  - Contains the technologies you can encounter in the exam
  - The Core technology list is shared between Written and Practical

Getting the CCDE

The path to getting your License to Design

- No formal pre-requisites.
  - 4-7 years of relevant job experience is recommended.
- CCIE is not expected nor required.



- Passing CCDE Written earns you Cisco Certified Design Specialist
- Written exam offered at Pearson VUE testing centers.
- Practical exam is administered in-house.
  - Offered at Cisco testing centers and at mobile events.
  - Offered on selected dates – check CLN for dates.

CCDE Written Exam

- Location: Pearson VUE Testing Centers
- Duration: 2 hours
- Exam code: 400-007
- 90-110 questions
  - Multiple Choice
  - Drag & Drop
  - Scored & non-scored items
  - No “skip question”, no “go back”

- Pre-requisites: None, though the following is recommended:
  - Thorough understanding of networking design principles
  - In-depth understanding of the exam topics and technologies
- Focus on design aspects of the technologies.

CCDE Practical exam structure

- There are four modules, each two hours in length:
  - 3 Core modules
  - 1 elective module (you choose this)
- Core modules are randomly assigned.
- The elective is a specialization module.
- Some topics can be found in both Core and elective
  - More detailed and comprehensive knowledge is expected in the elective.
  - xWDM is an example of a topic covered in both core and Large-Scale Networks

CCDE Practical exam

- Each scenario has a set of resources, such as:
  - Introduction
  - Company background, strategies, history etc.
  - Topology drawings
  - On-going correspondence, emails, chats, etc.
- Resources won’t go away once you have received them.
- All are easily accessible in a dedicated section in exam interface.
- New resources can (and will) be introduced with a question.

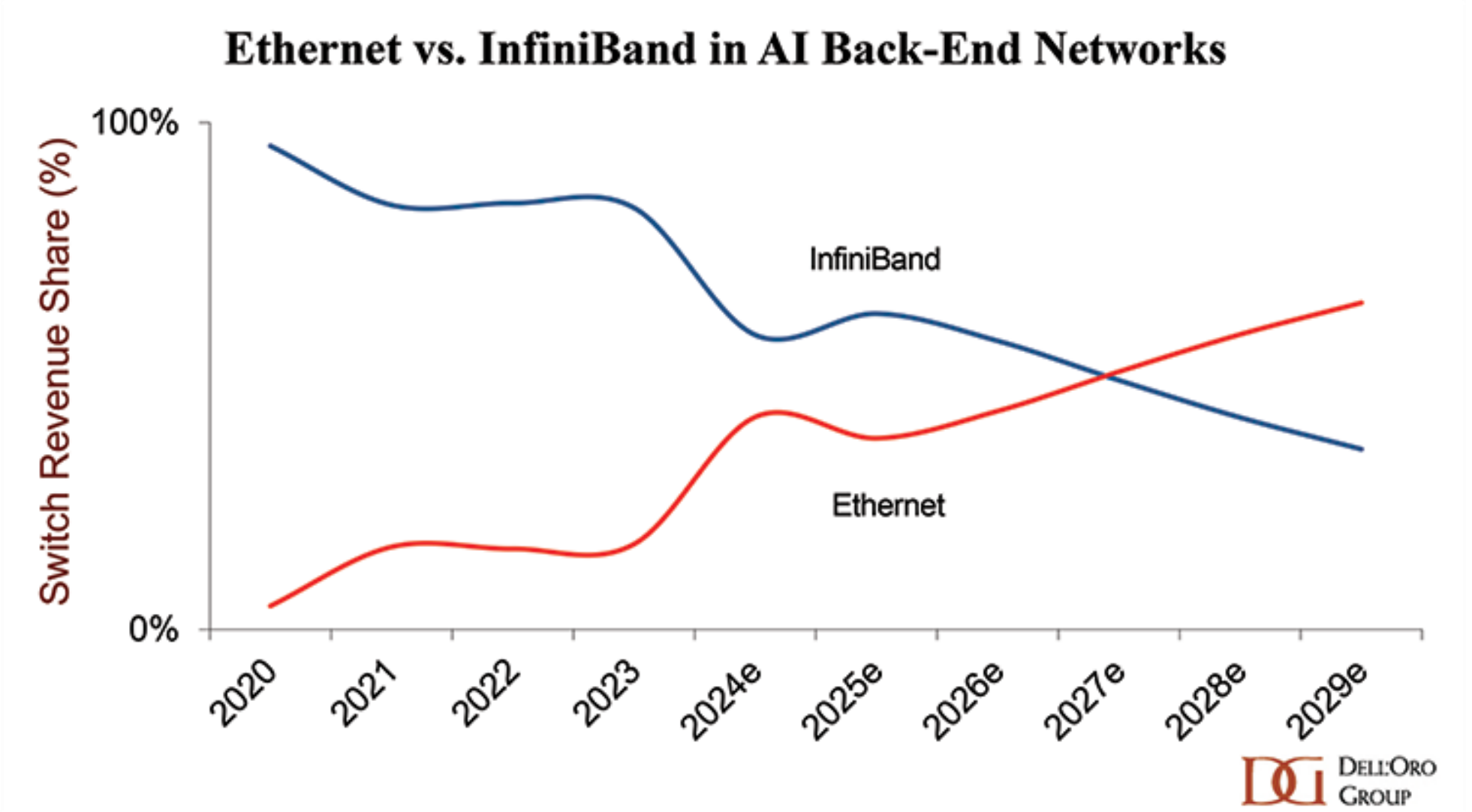
What to consider while gathering information

- Business objective
- Project objective
- Current design
- Skillset(s) required/needed
- Business constraints
- Technical constraints
- Other constraints
- Special situations
- Questions answered
- Changes in direction/solution
- Technology selection
- Other things that could be relevant (make a mental note)

Design use cases

- A Practical exam scenario typically focuses on one or two use-cases.
- This helps setting the stage and the mindset you must assume.
- The provided exhibits will indicate the use-case.
- Pretend you are sitting at the meetings with the company.
- Try to understand why they are moving in this or that direction.
- Are there any consequences of making this choice?

The future of the Backend Network is Ethernet



Hard facts of AI network design

- These topics heavily impact the efficiency of the backend network
  - Loss
  - Congestion avoidance
  - Average Link Utilization / Max Link Utilization
  - Job Completion Time JCT
  - Entropy – ensure all paths are utilized and congested links avoided
  - Serialization delay
  - Round Trip Time / Port-to-port-delay
  - Load-balancing
  - 10µs microsecond RTT 100-300m Length 400/800Gig
  - Scale up network - GPU connectivity tensor parallelism
  - Scale out network - connect 10,000 GPU together
  - DAC or fiber (yes, there is a difference!)
  - (Sounds like the good old CCIE studying days using T1/E1s, right)
  - Monitoring
  - Configuration
  - Operational Sustainability (Patching) Non-disruptive