

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC



AI/ML for Network Engineers

Patrick LeMaistre

Solutions Engineer

Prapanch Ramamoorthy

Principal Engineer

Dave Zacks
BRKENT-2209

Distinguished Engineer

CISCO *Live!*

#CiscoLiveAPJC



AI/ML for Network Engineers

Patrick LeMaistre

Solutions Engineer

Prapanch Ramamoorthy

Principal Engineer

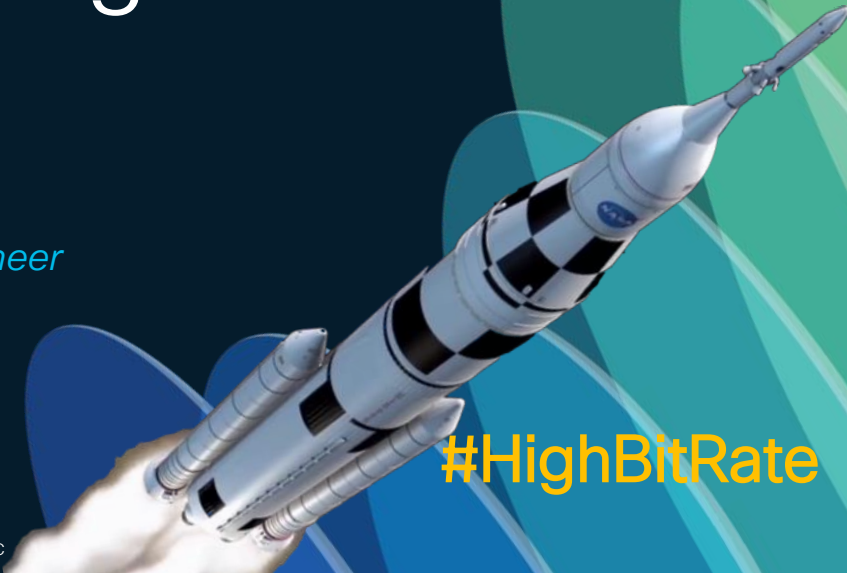
Dave Zacks
BRKENT-2209

Distinguished Engineer

CISCO *Live!*

#CiscoLiveAPJC

#HighBitRate

A detailed illustration of a Space Shuttle launching, angled upwards from the bottom right towards the top right. The shuttle is white with black and white checkered patterns on the side of the orbiter. It is surrounded by large, overlapping, semi-transparent green and blue shapes that resemble stylized leaves or petals, creating a dynamic, organic background.

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 November 15, 2024.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKENT-2209>



Agenda

- Introduction

AI/ML, Generative AI,
Neural Networks, Transformers ...

- What is Cisco doing with AI/ML?

- Using AI/ML in Solutions
- Building Networks for AI/ML



- On-Prem AI Solutions

- Open Source LLMs w/ UCS and Nexus
- Demo Time!



- Leveraging LLMs for Networking

- LLMs Deeper Dive
- Working with LLMs
- Demo Time!



- Wrap-up

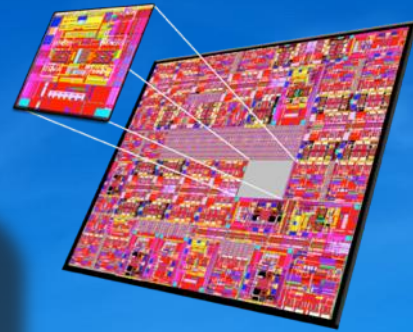
by way of Introduction ...

I am a **Distinguished Engineer** in the Network Transformation team, and have been with Cisco for 25 years.

I work primarily with large, high-performance Enterprise network architectures, designs, and systems. I have over 39 years of experience with designing, implementing, and supporting solutions with many diverse network technologies.

I have a strong background in, and focus on, customer requirements, and integrating these into the products and solutions Cisco builds.

I have a special interest in **Flexible Hardware, Fabrics, Assurance and ML/AI**.



Dave Zacks
Distinguished Engineer
dzacks@cisco.com

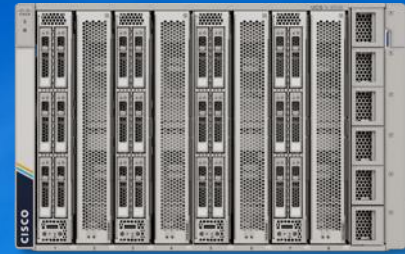


by way of Introduction ...

I am a **Compute Solutions Engineer** in Cisco's Cloud Infrastructure and Software Group. I have been with Cisco for 25 years focused on the areas of **AI, high performance computing, cloud, mass-scale infrastructure and mobility**.

I am the author of the AI Toolkit for UCS X-Series and C-Series which can be found on GitHub at: www.github.com/pl247/ai-toolkit

Prior to Cisco, I was co-founder and Vice President of Network Operations for Escape Communications, and hold a patent pertaining to Data Center Ethernet.



Patrick LeMaistre
Solutions Engineer
plemaist@cisco.com

by way of Introduction ...

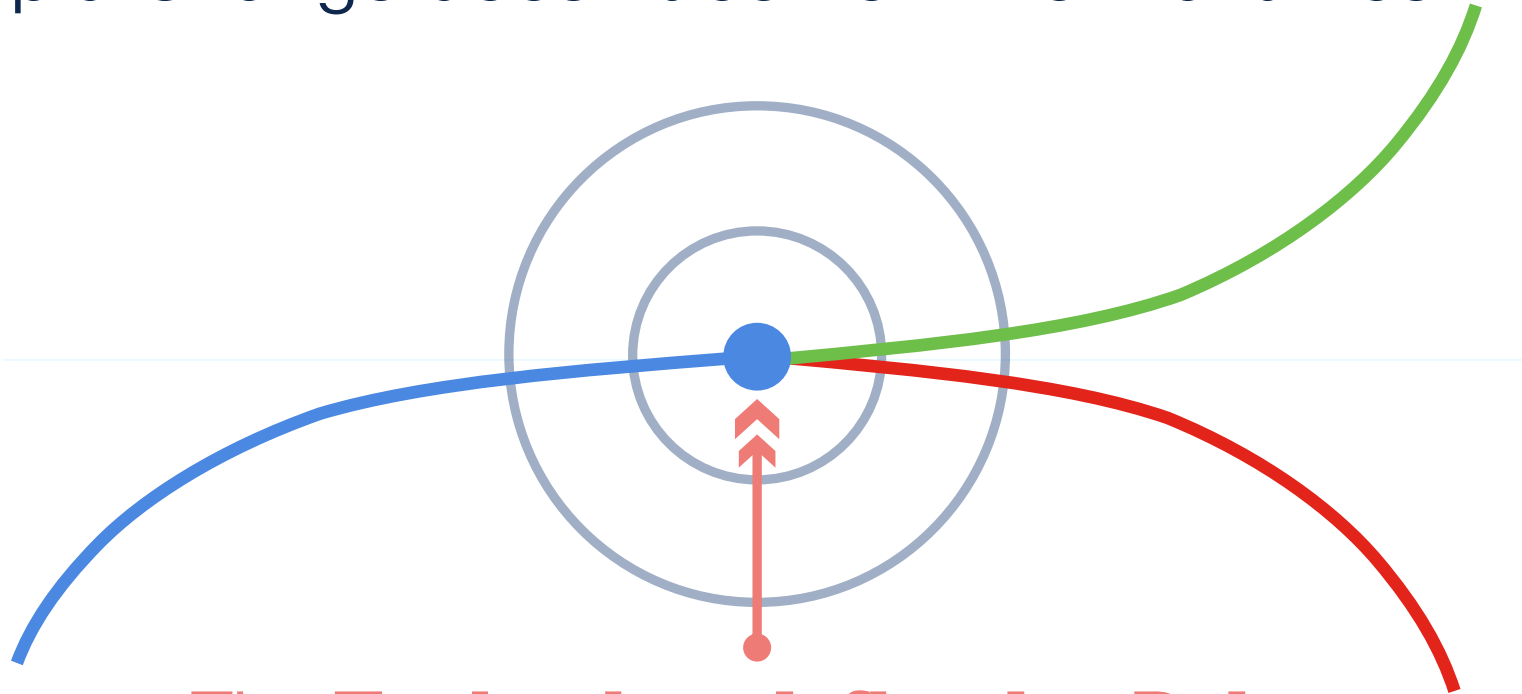
I am a Principal Engineer in Cisco TAC. I am 36 years old, with 15 years experience in **Cybersecurity** and a strong focus in the security and networking spaces. Presently focused on developing **Agentic workflows** to help improve customer interactions.

I am a Problem Solver, a **Serviceability** architect, a Teacher, and an Innovator, in addition to being a husband and father (2 boys). I spend my spare time volunteering with organizations in the cybersecurity space.

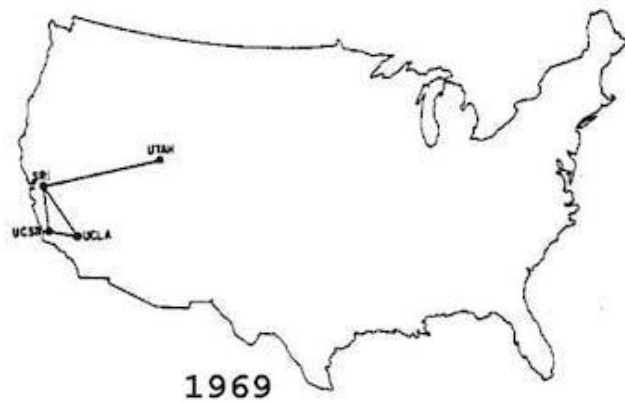
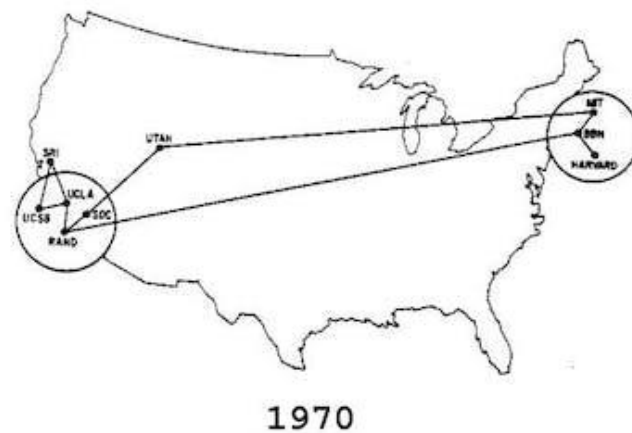
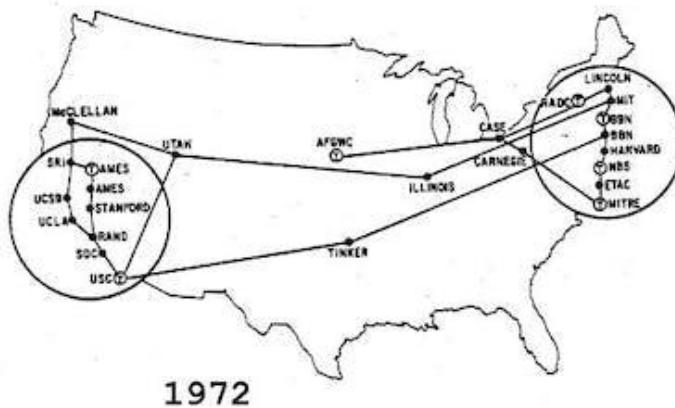
Prapanch Ramamoorthy
Principal Engineer
praprma@cisco.com

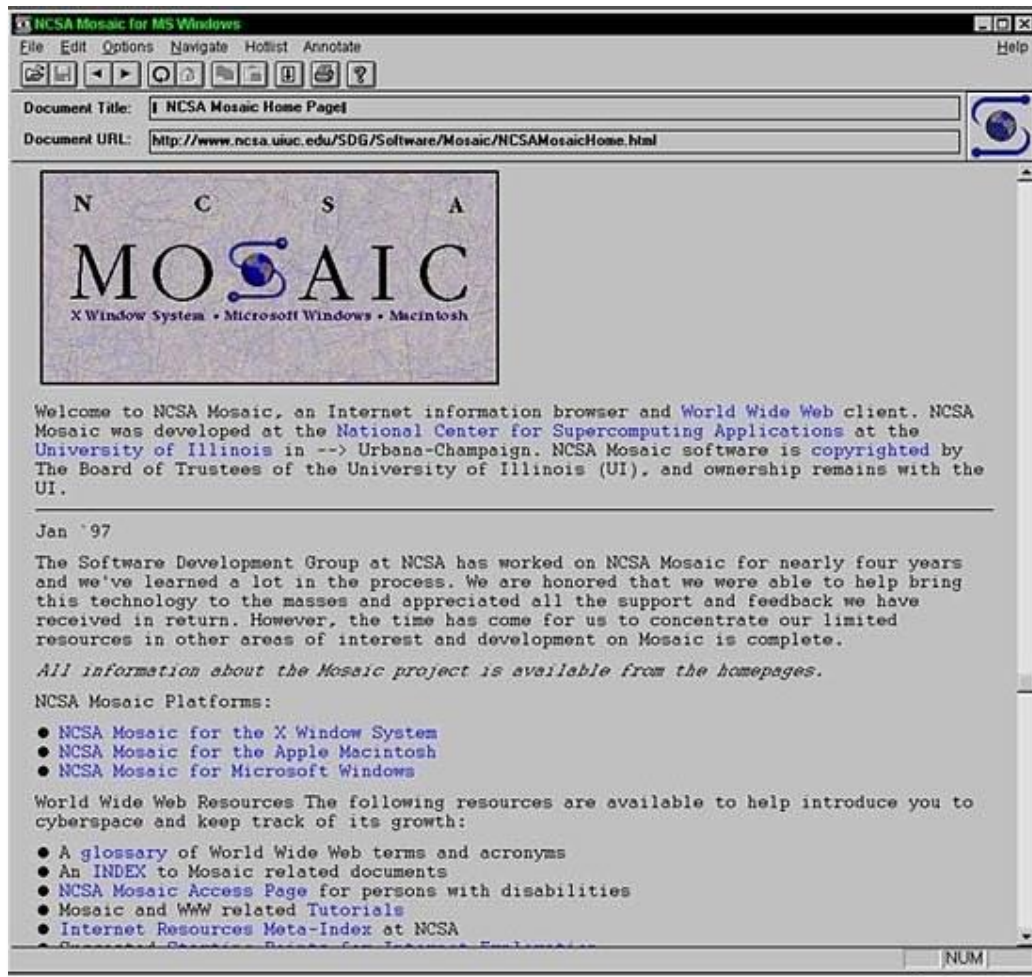


Rapid Change doesn't come in normal times

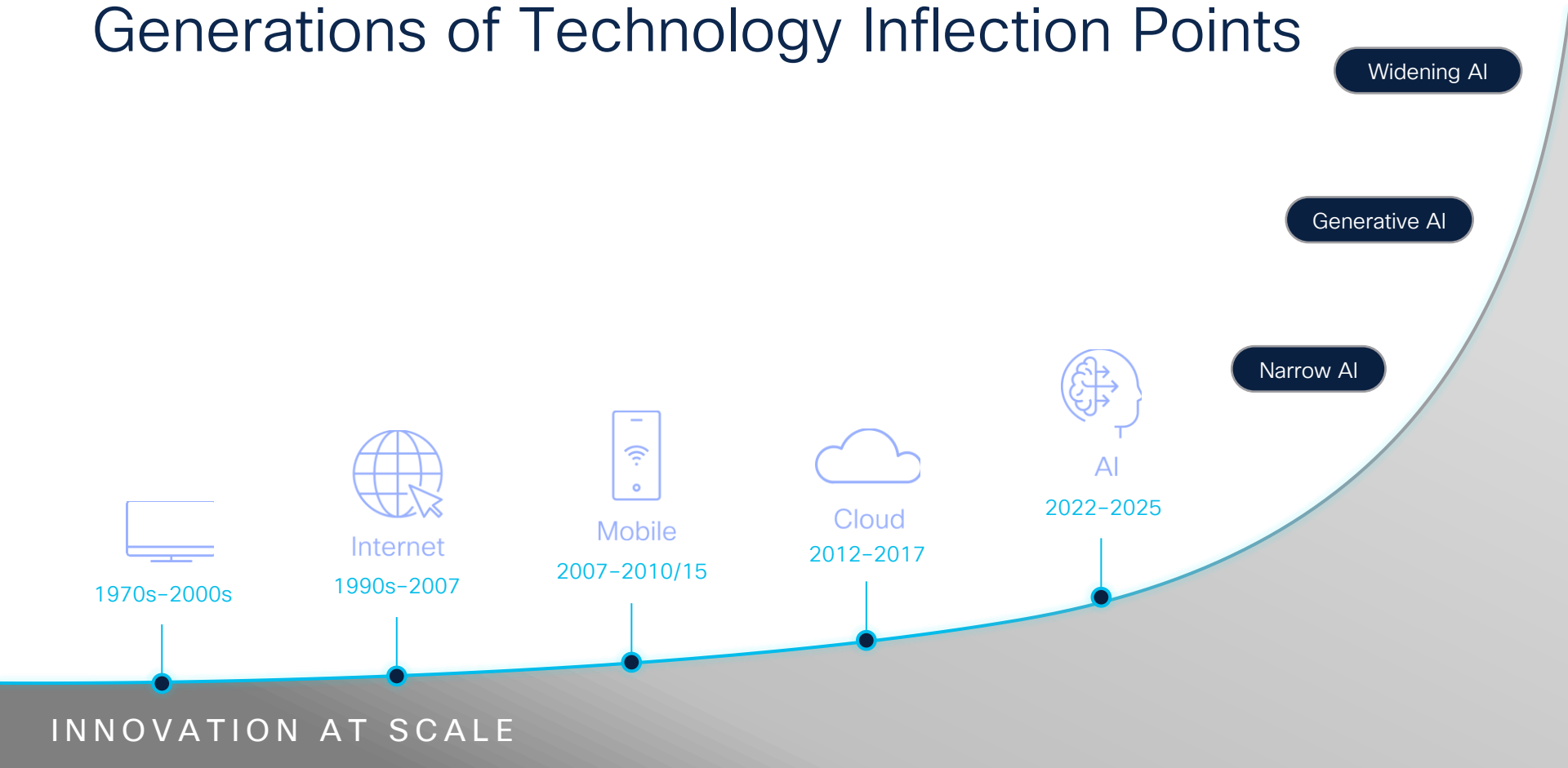


The **Technology Inflection Point**





Generations of Technology Inflection Points



AI encompasses the whole world
of ML and Deep Learning

An AI technology where the
rules are not set in the
program, but are **learned** while
the program is used

Artificial Intelligence

Machine Learning

Deep Learning

Generative AI

Language
Models

A form of ML that uses
Neural Learning Networks

A powerful mechanism
that allow neural
networks to
**learn language and
generate content**

A Snapshot of GenAI Capabilities ...

Casual / Fun



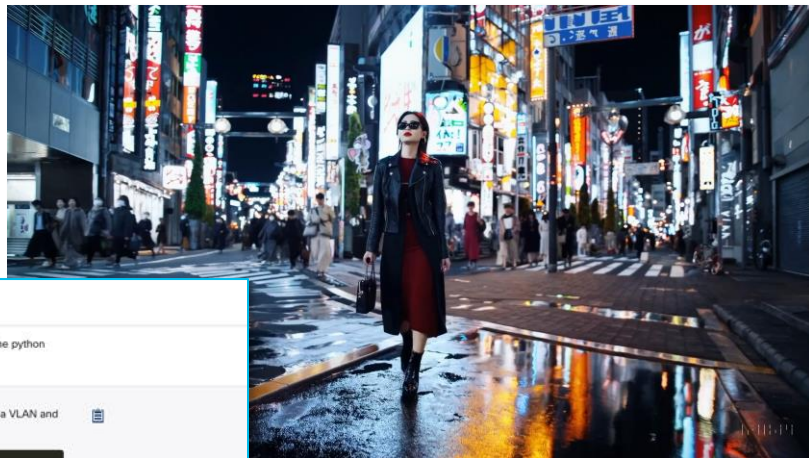
tell me a joke about the summer olympics in the style of Albert Einstein



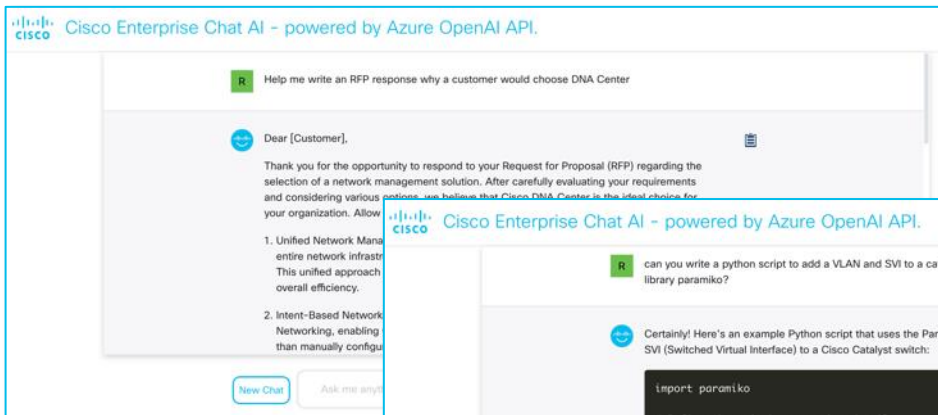
"Why did the photon refuse to compete in the Summer Olympics?"

Because it didn't want to be weighed down by mass and preferred to move at the speed of light, where time dilation makes every event feel like a lifetime!"

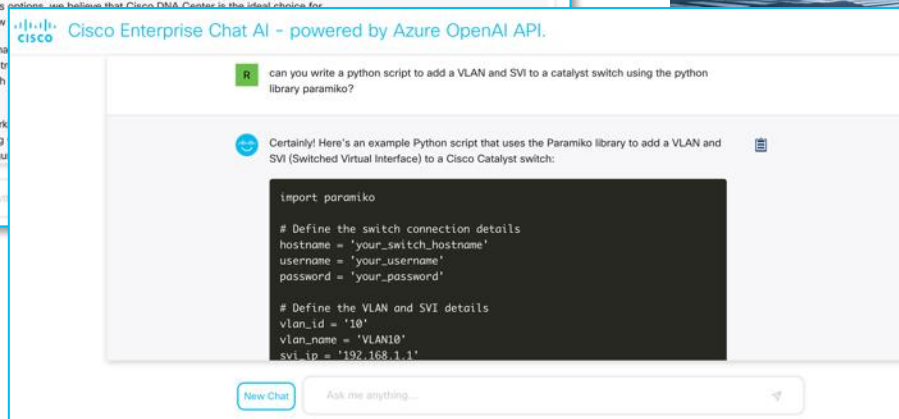
Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.



MULTIMEDIA –
Text, Pictures, Video, ...



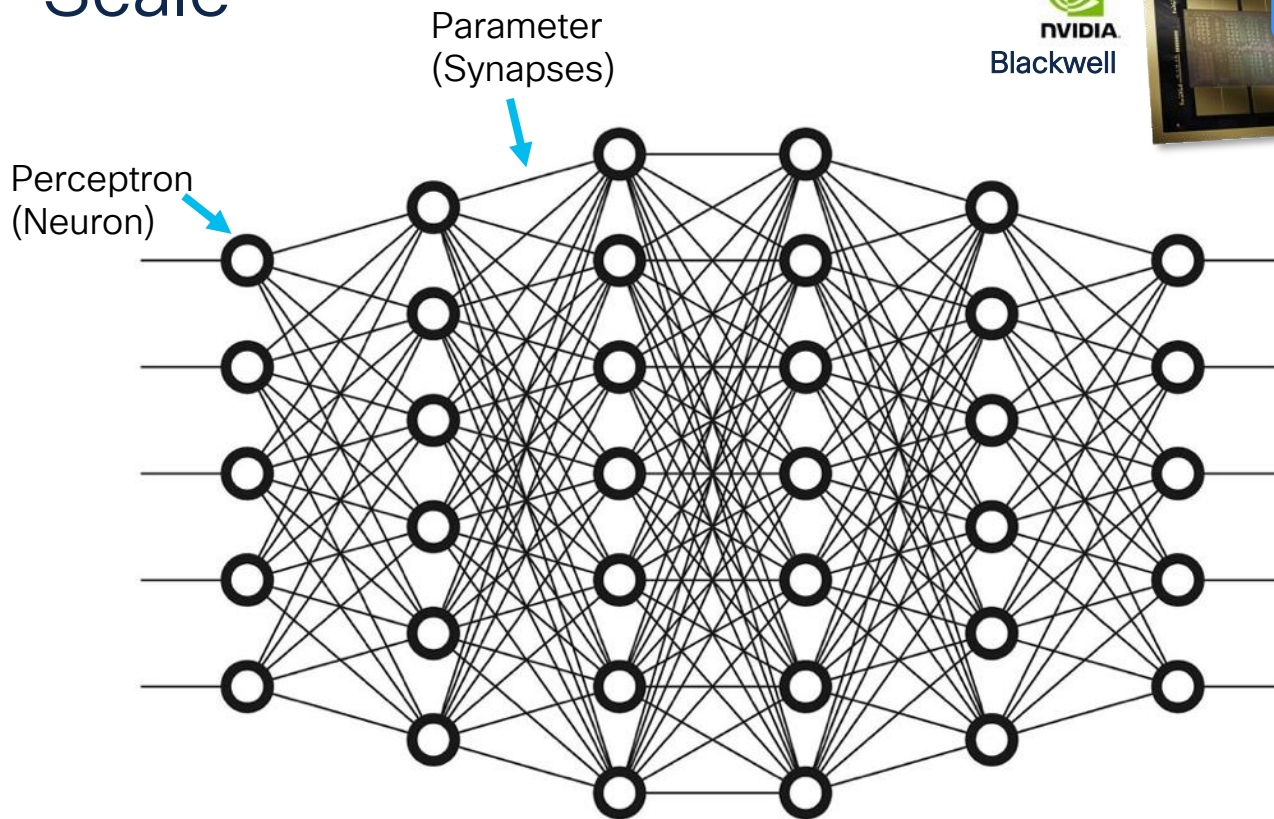
BUSINESS –
Answering RFPs,
Writing code ...



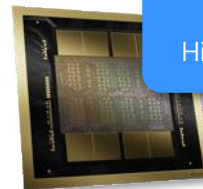
cisco *Live!*

Why is this happening now?

Scale



NVIDIA
Blackwell



Advances in Silicon –
High-density, High-performance GPUs

NVIDIA Flagship Accelerator Specification Comparison			
	B200	H100	A100 (80GB)
FP32 CUDA Cores	A Whole Lot	16896	6912
Tensor Cores	As Many As Possible	528	432
Boost Clock	To The Moon	1.98GHz	1.41GHz
Memory Clock	8Gbps HBM3E	5.23Gbps HBM3	3.2Gbps HBM2e
Memory Bus Width	2x 4096-bit	5120-bit	5120-bit
Memory Bandwidth	8TB/sec	3.35TB/sec	2TB/sec
VRAM	192GB (2x 96GB)	80GB	80GB



Geoffrey Hinton – the “Godfather” of Deep Learning

cisco Live!

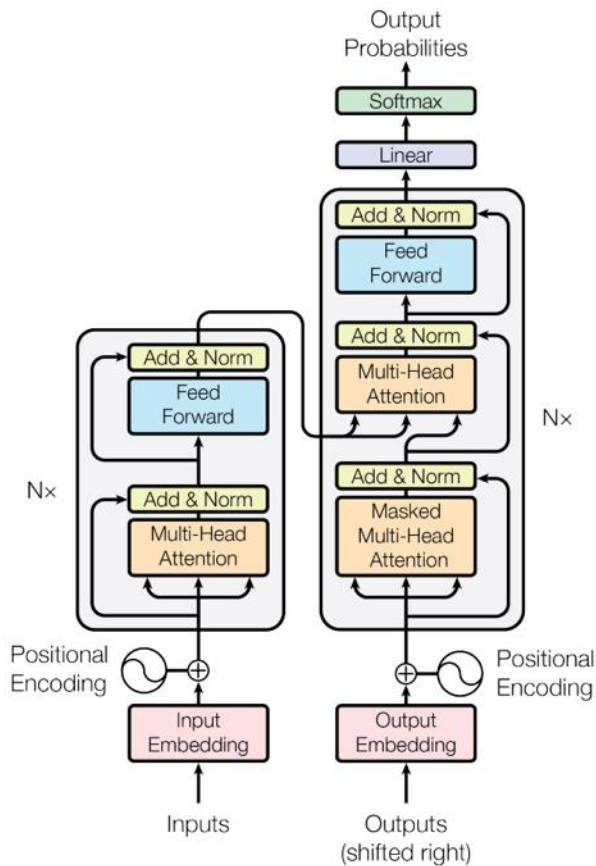


Figure 1: The Transformer - model architecture.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

[arXiv:1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL]

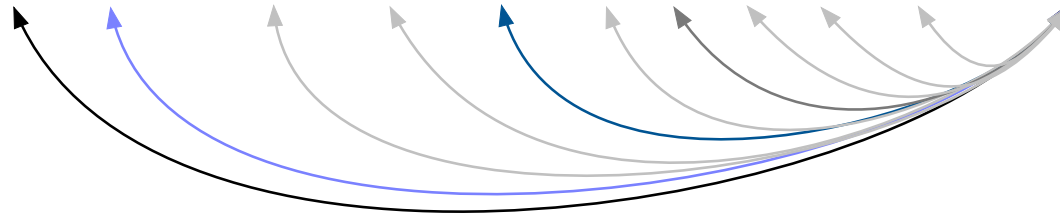
Attention Mechanism – Overview

You have no problem interpreting “bank” in the following sentence:

“I swam across the river to get to the other bank.”

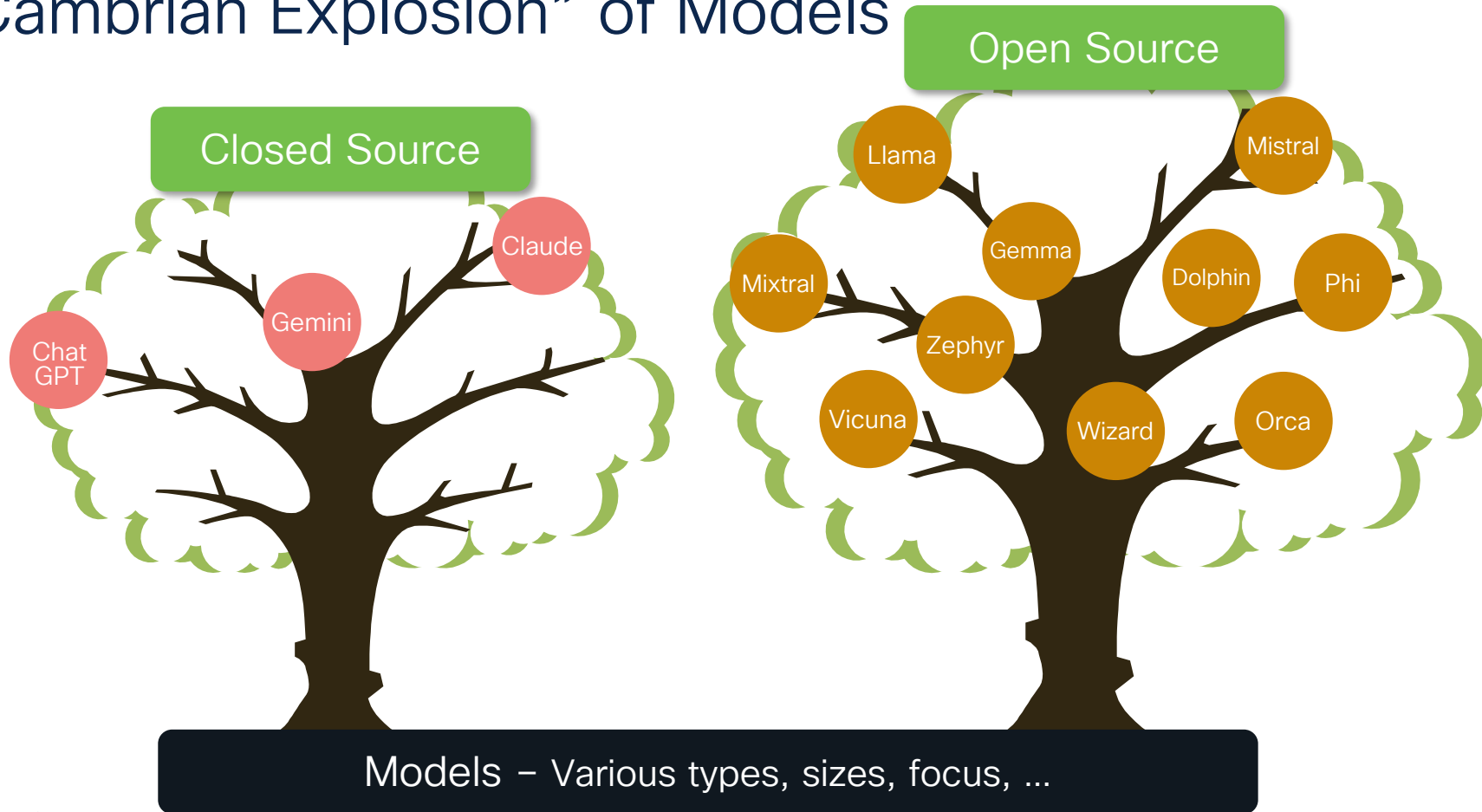
A machine needs some help...

I swam across the river to get to the other bank.

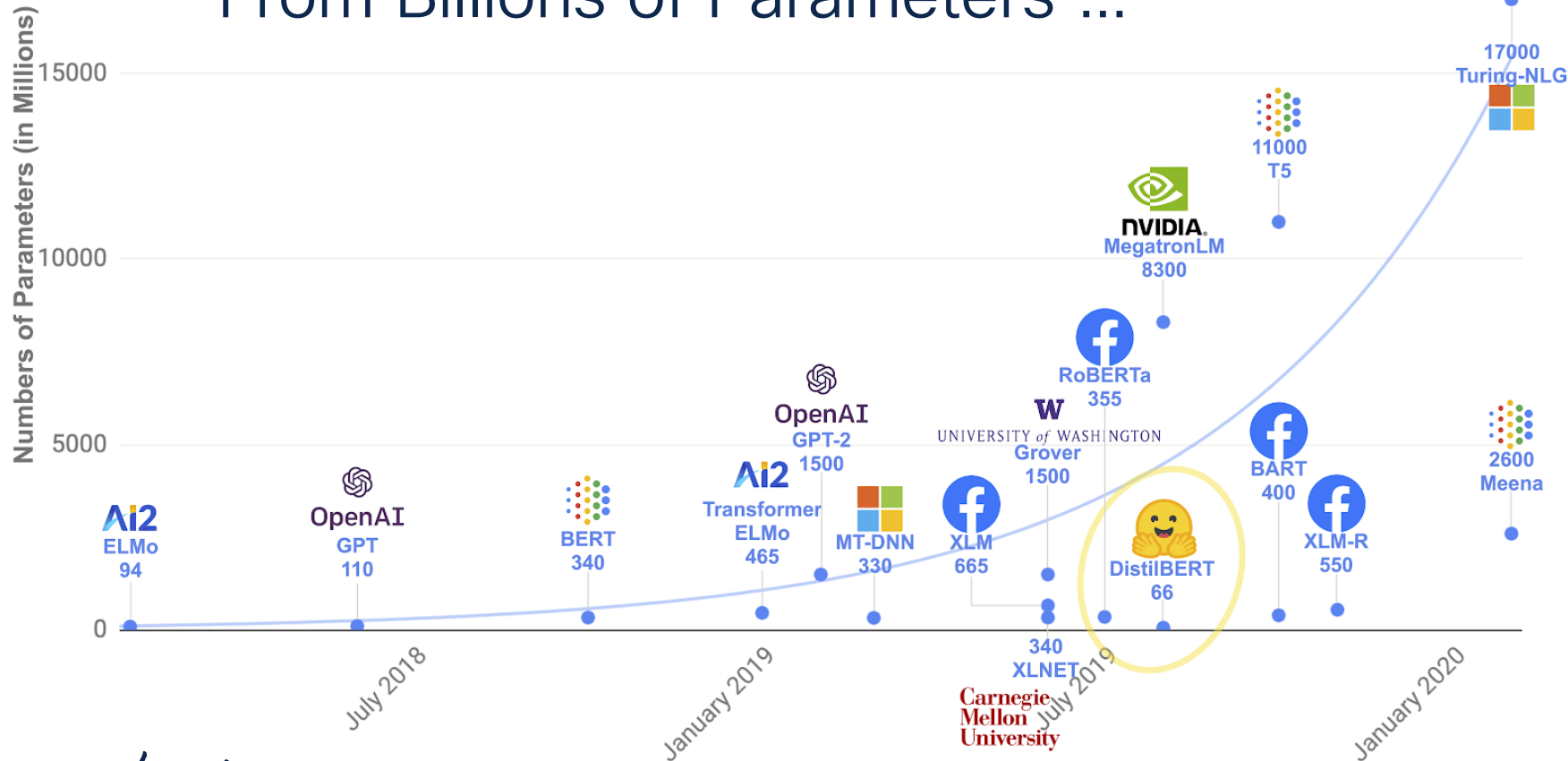


The goal of the attention mechanism is to add
contextual information to words in a sentence.

“Cambrian Explosion” of Models



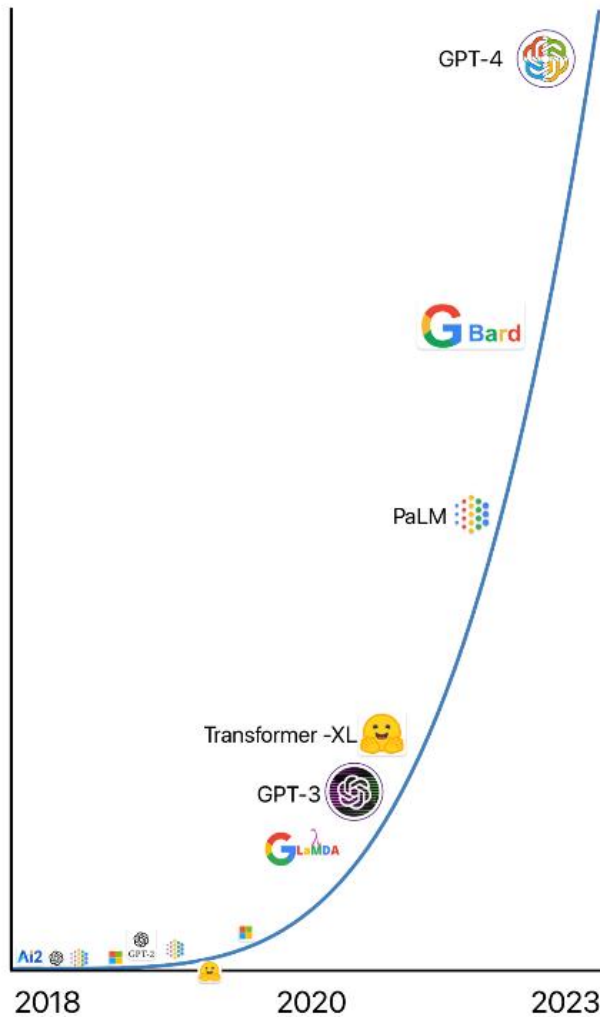
From Billions of Parameters ...



To Trillions ...

1,000,000,000,000

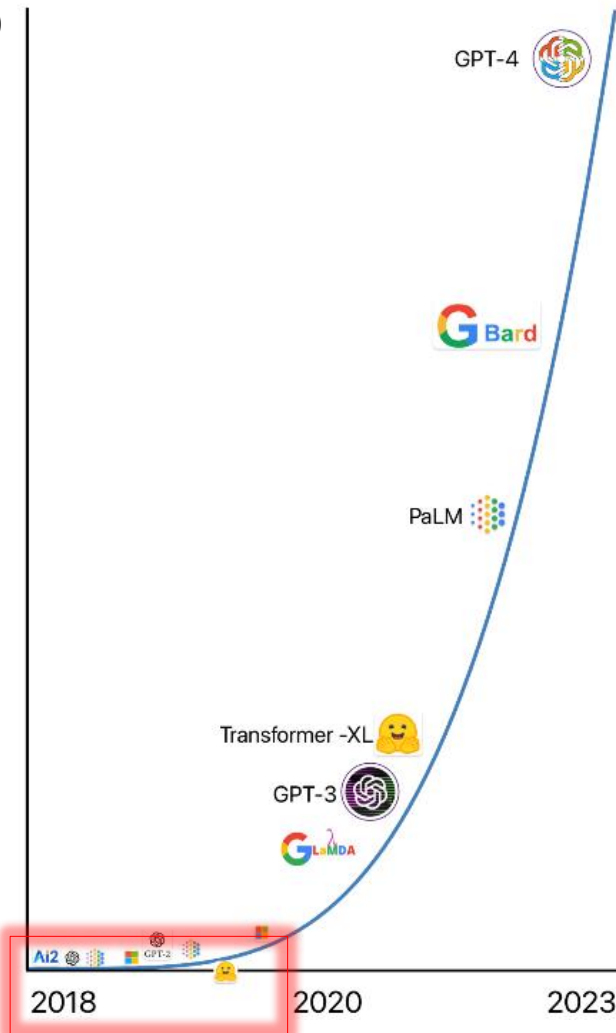
500,000,000,000



To Trillions ...

1,000,000,000,000

500,000,000,000



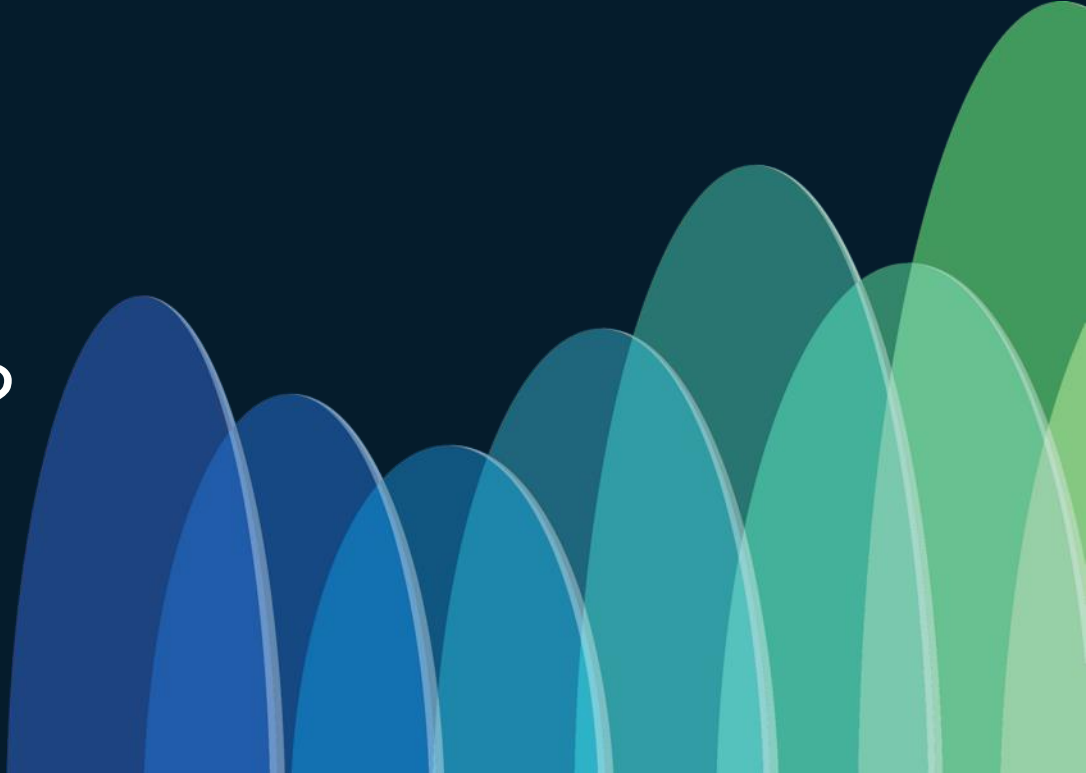
FUN FACT!

The human brain contains
86 billion neurons, and over
100 trillion synaptic connections



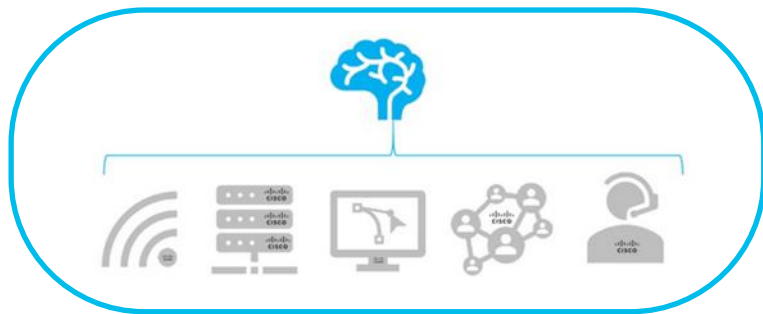
Overview –

Where is Cisco Using AI/ML in Products Today?

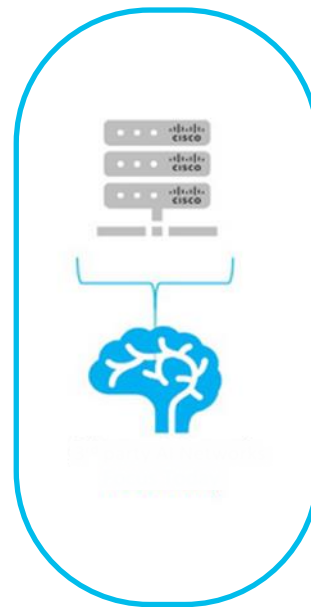


Artificial Intelligence and Cisco

AI in Cisco – AI to improve products

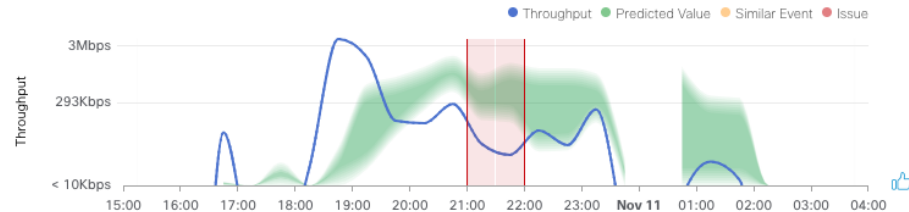


AI on Cisco – Products to improve AI

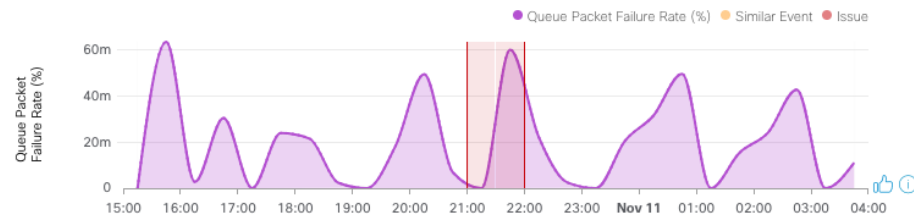
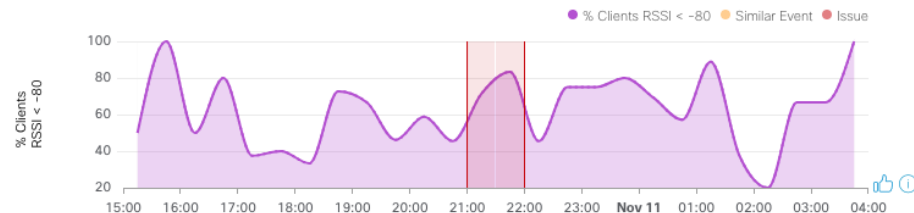


AI-Powered Wireless Analytics

AI-Driven Anomaly Detection
Find + Root Cause Complex Issues



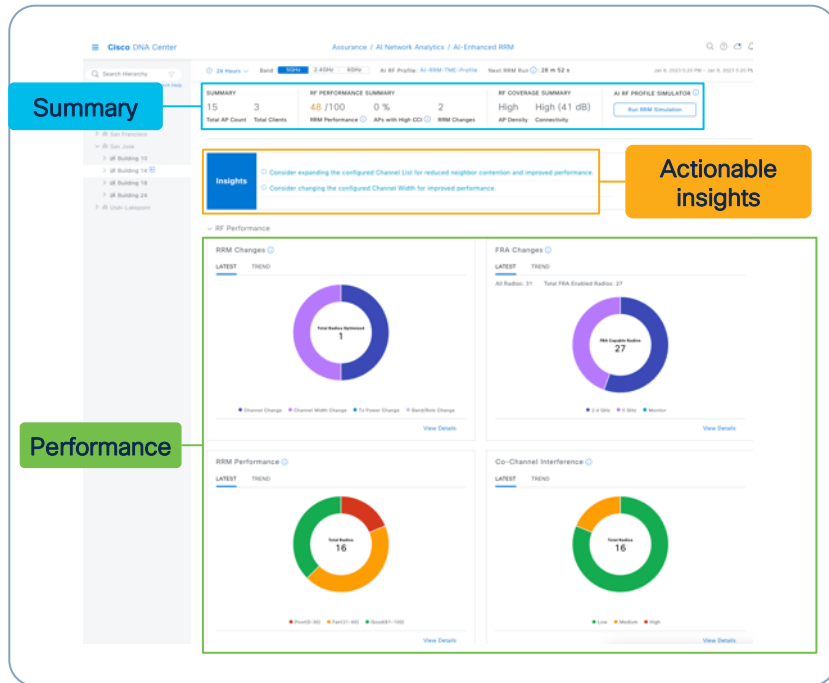
Probable network causes



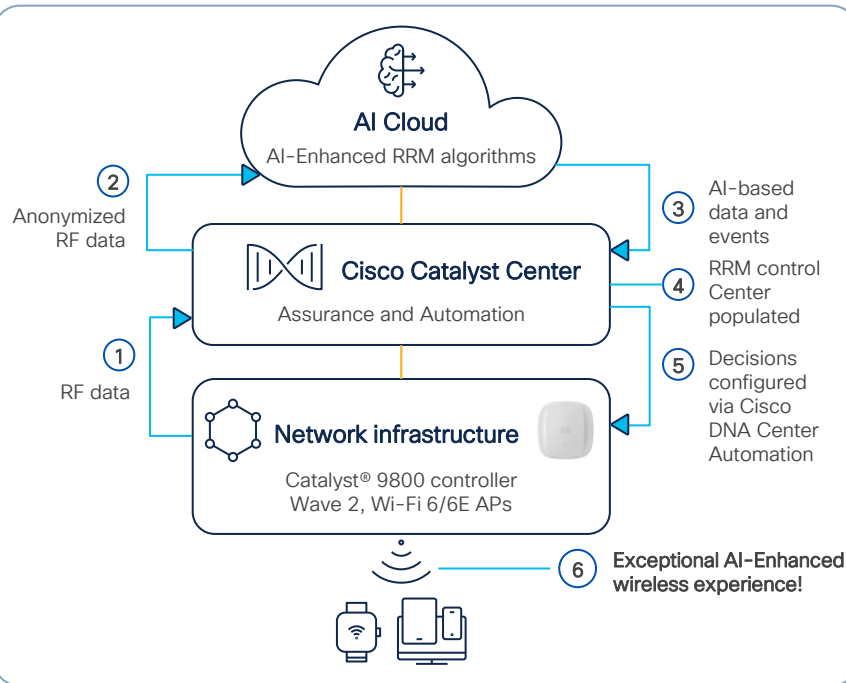
AI-Enhanced RRM

Catalyst AI-Driven RRM solution

Deep RF visibility & advanced control



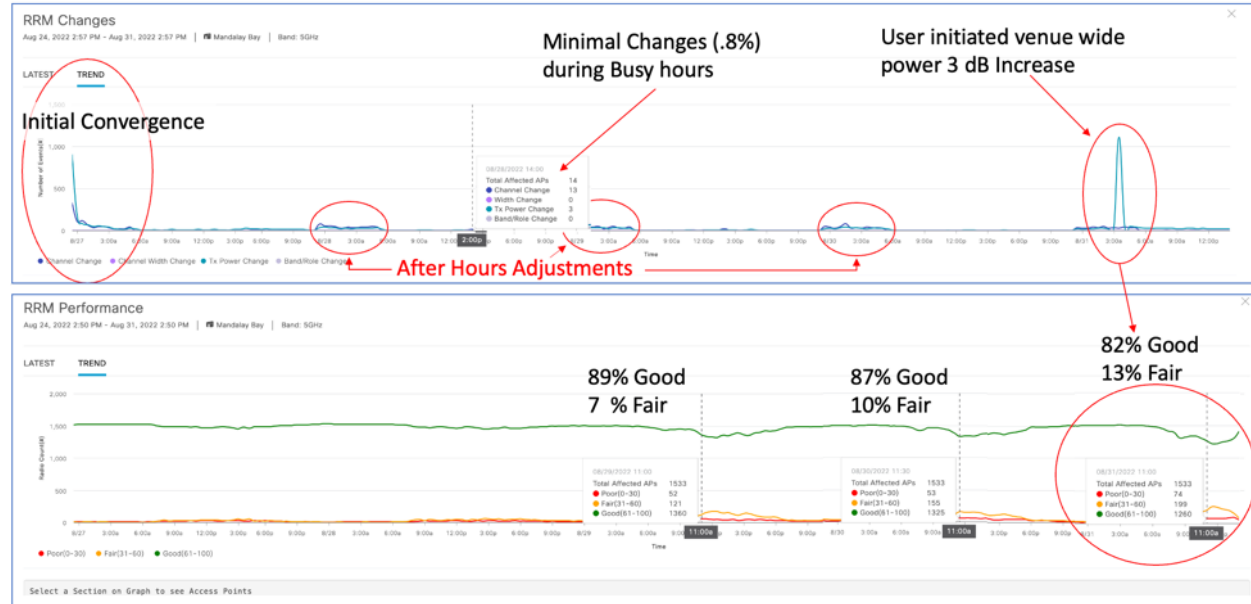
Proactive optimizations for all deployment sizes



AI-Enhanced RRM in Action

Cisco Impact 2022

- Initial Convergence
~3 Hours
- Changes made at Night
- Health stayed above 85%
(very good with load)
- Manual Changes made last day, easy to spot the decrease in efficiency



Predictive Internet

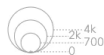
SLA Violations Across the World
and how much Predictive Networks can help

0
CUSTOMERS

0
COUNTRIES

2
CITIES

 Cisco
redictive
etworks



CISCO *Live!*

Cisco Live Melbourne (Dec 2023)

AI Policy Assistant

Policy Assistant

- Reduces security policy complexity
- Optimizes policy for efficacy & efficiency

The screenshot shows the Cisco Defense Orchestrator (DO) interface. The main panel displays the 'ACP - Production' configuration page. The top navigation bar includes 'Return to Access Control Policy' and 'ACP - Production'. The main content area shows a table of rules with columns for Name, Action, Source, and Destination. The table lists four rules: 1 External (Block), 2 Internal (Allow), 3 Block Malwares (Block), and 4 Block Torrent (Block). The AI Assistant chat interface is visible on the right side of the screen, showing a message from the AI Assistant asking for clarification on rules with low hit counts.

AI Assistant

I'm trying to determine what rules are poorly written by analyzing the number of hits. Can you show me the rules with the least number of hits in the last 30 days?

AI Assistant 8:12 AM

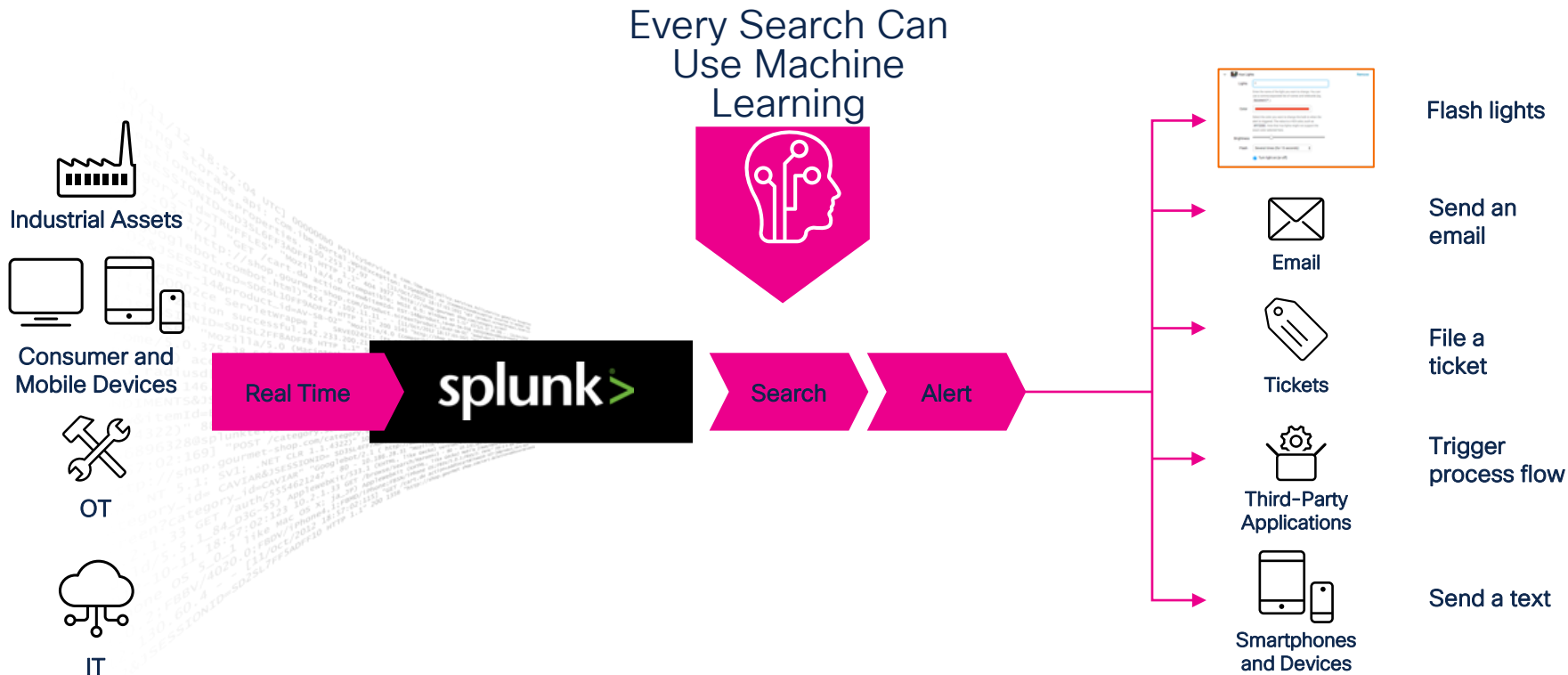
There are five access control rules with low hit counts in the last 30 days.

View details

Regenerate

Ask a question or request, or type "?" for suggestions

Splunk Machine Learning Tool Kit (MLTK)

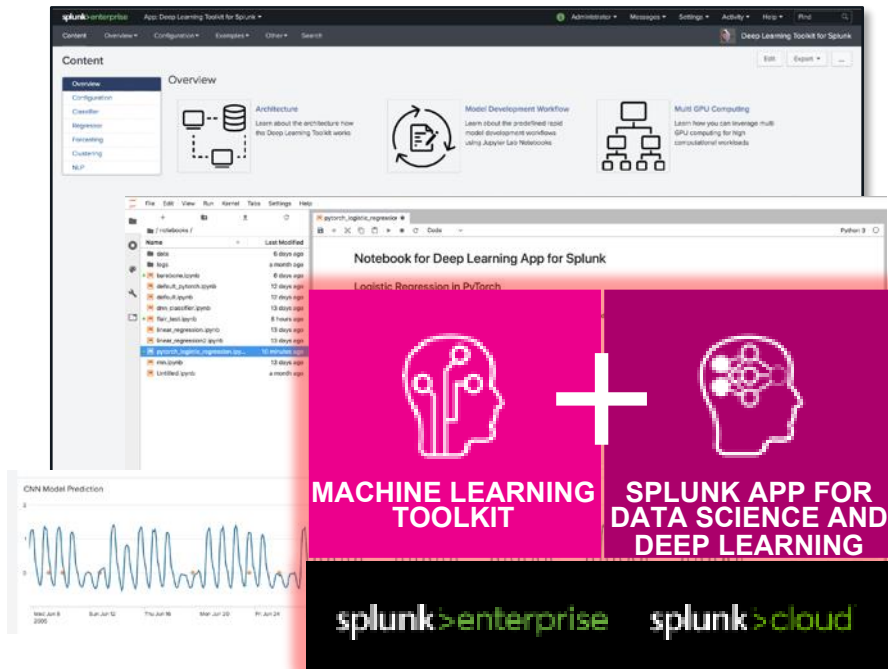


Splunk App for Data Science and Deep Learning (DSDL)

Extension for MLTK to operationalize advanced custom AI / ML use cases



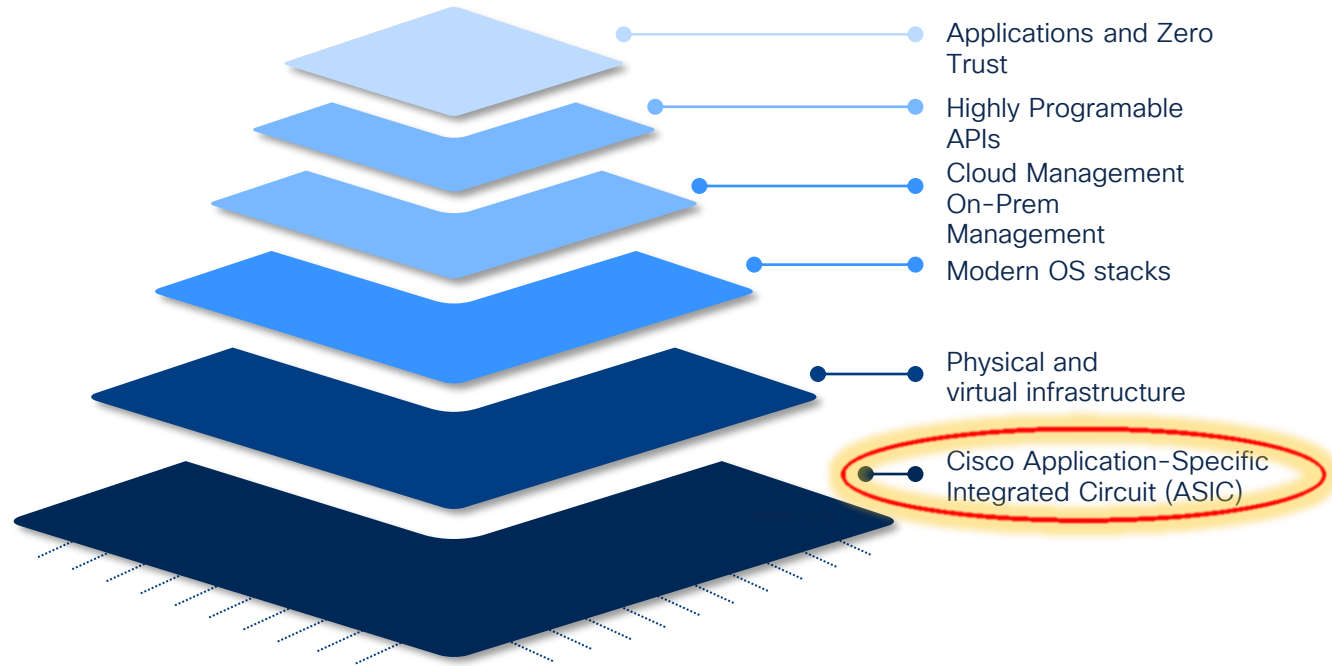
- **Built for Data Scientists**
- **Code Examples:** Guided model building, testing, and deployment of deep learning frameworks
- **Container Management:** Containers & Models can be productionized for scalability & optimization of resources, e.g. CPU & GPU
- **Frameworks:** PyTorch, Tensorflow, SpaCy, DASK, Rapids, Spark, Jupyter Notebooks & Tensorboard, ...
- **Extensible to operationalize any use case**
- **Open Source:**
New frameworks & Python Libraries freely available for integration into DSDL via github
- **Acceleration:** Support for GPU accelerated machine learning and inference pipelines



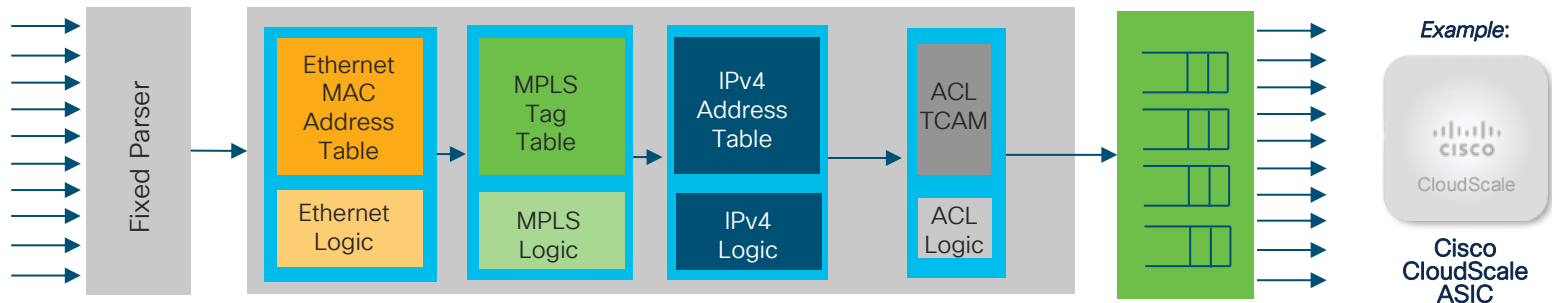
Building Networks for AI/ML Workloads



Cisco Silicon Foundational Elements



Fixed vs. Programmable Packet Processing



Fixed Pipeline: features and functionality are baked-in at design time

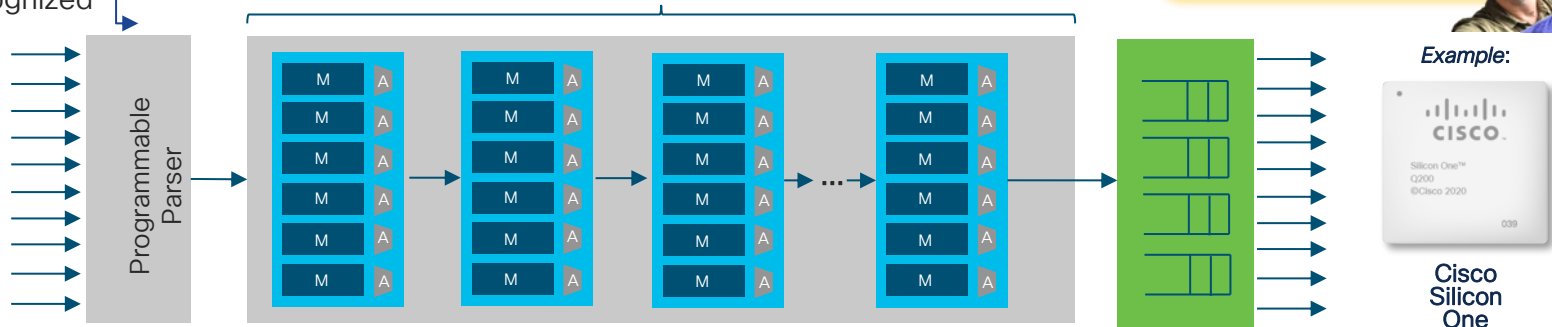
Want to know more about
Cisco Silicon for AI?

BRKARC-2095
Cisco Live US 2024



You declare which
headers are recognized

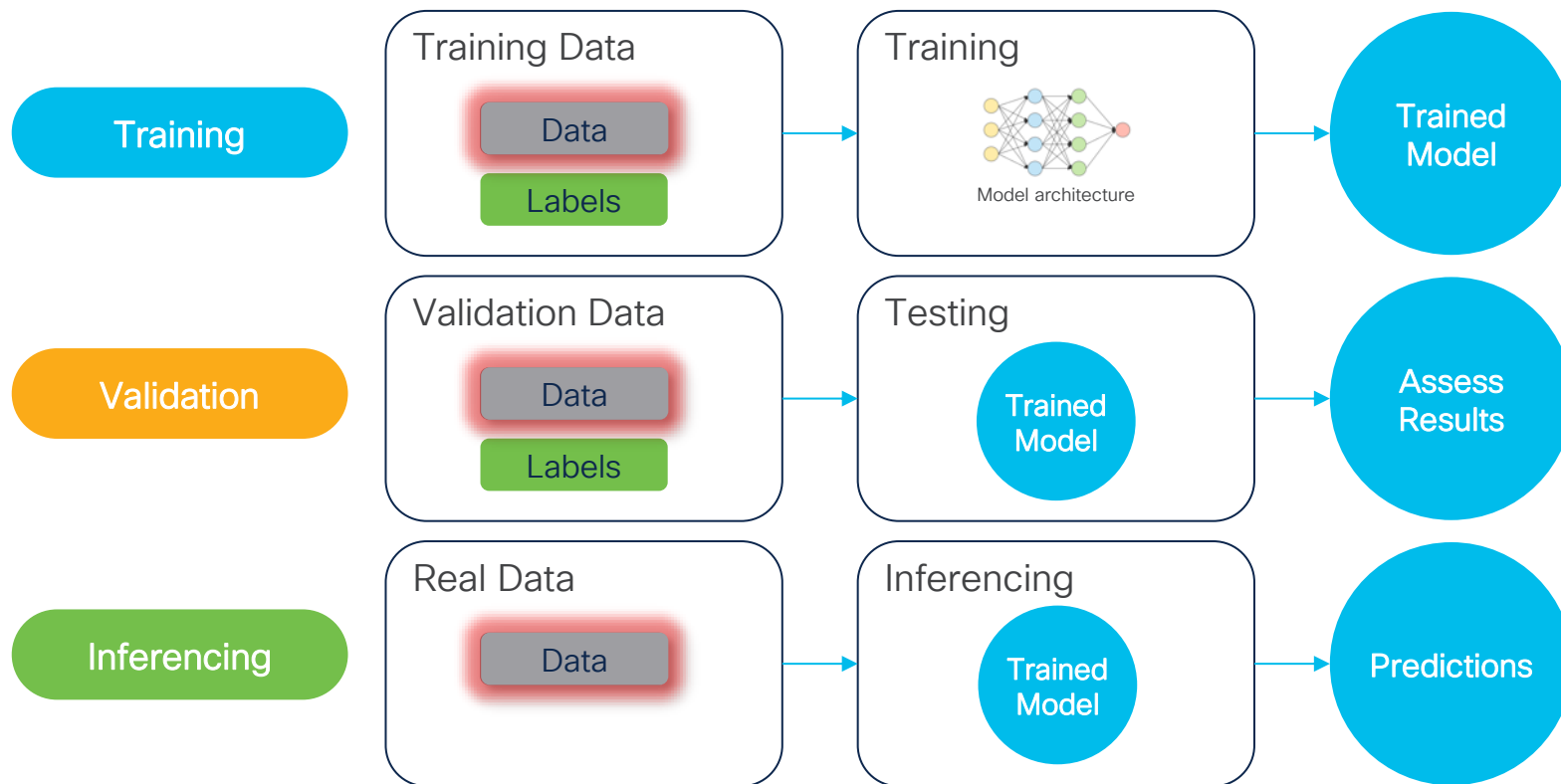
You declare what tables are needed and how packets are processed



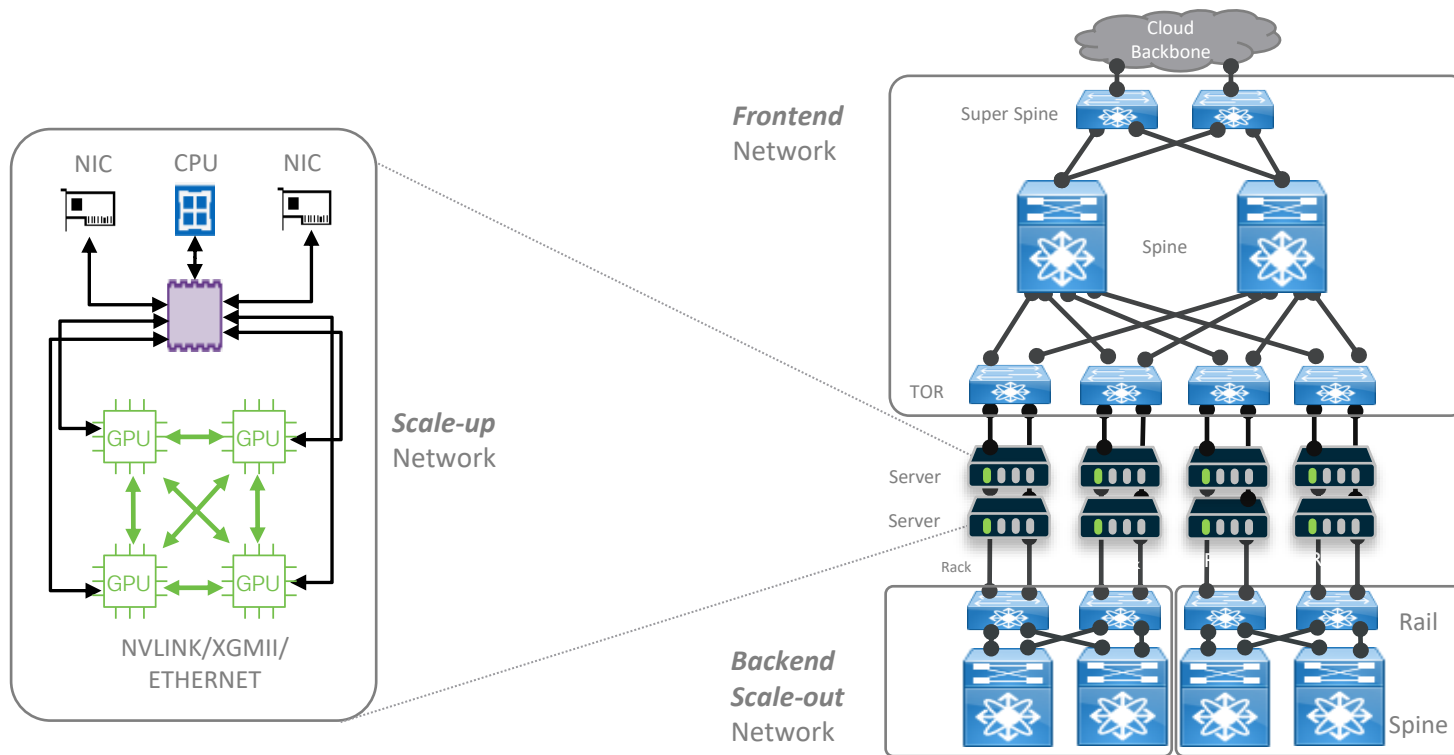
Programmable Pipeline: all stages identical, customer-defined match-action logic

cisco Live!

AI - Stages

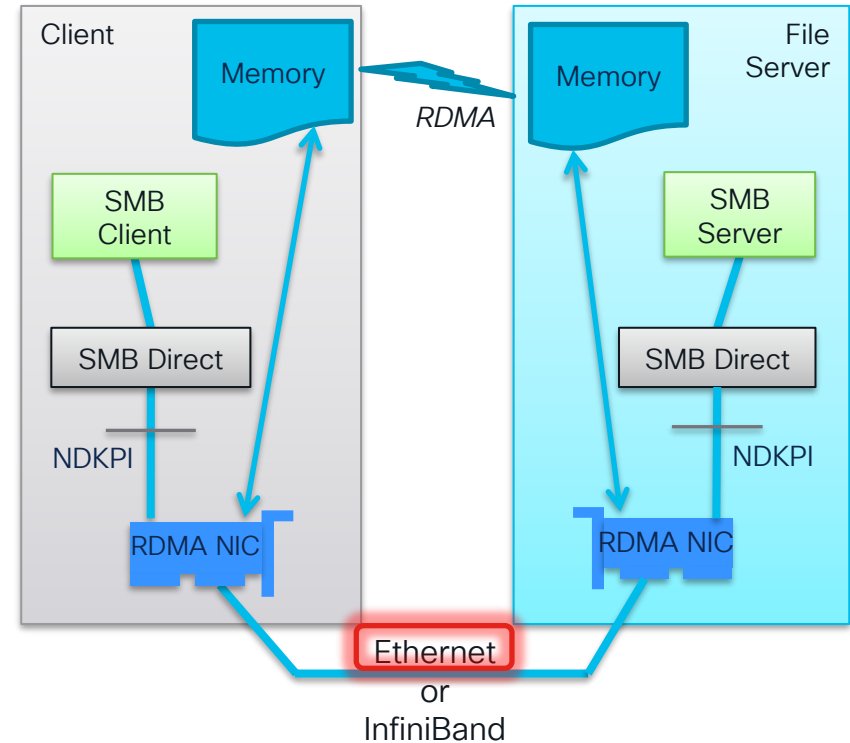


AI – Network Type Fundamentals

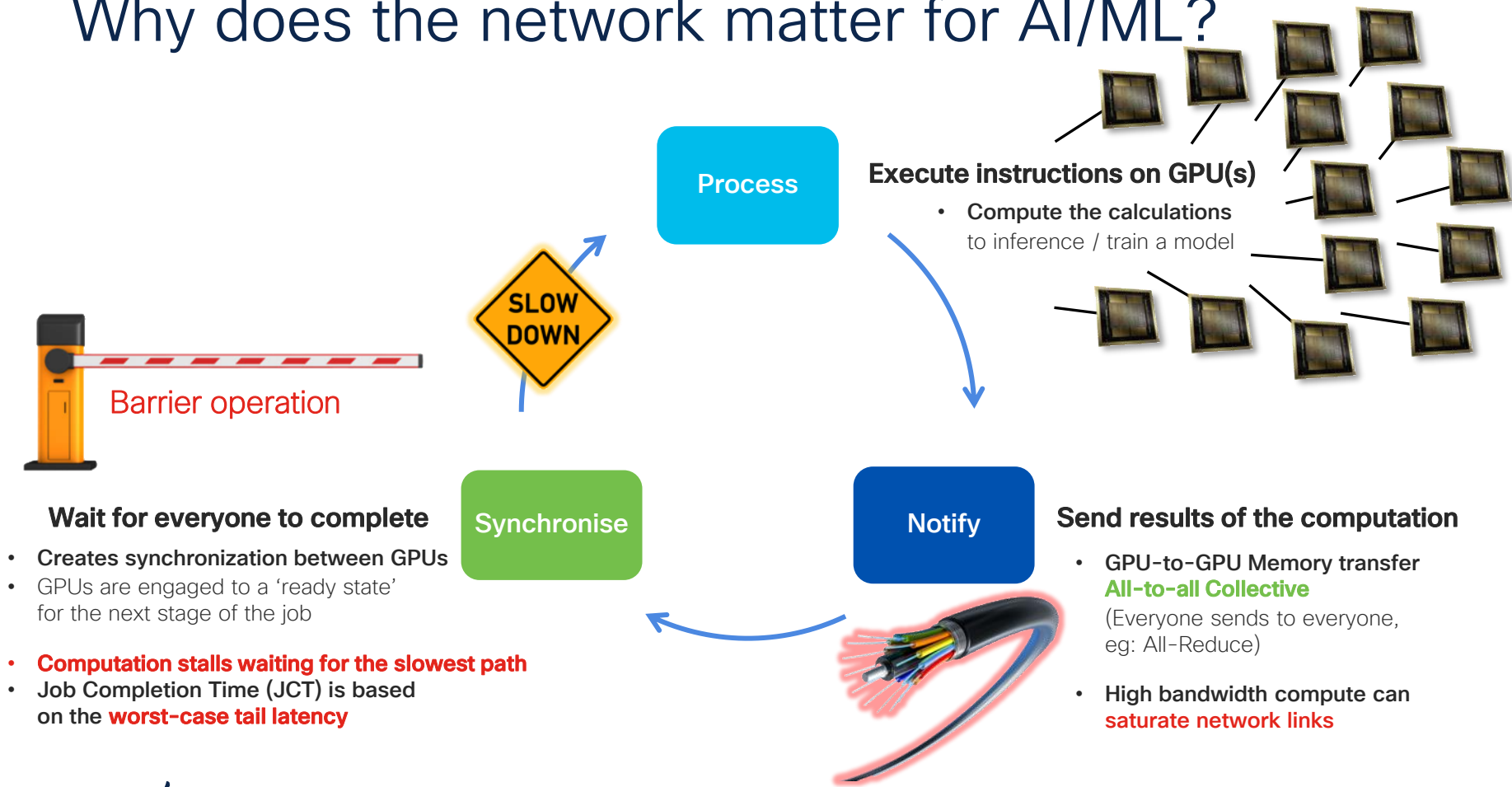


RDMA – Remote Direct Memory Access

- Allows application software to communicate directly with the hardware (RDMA NIC)
- Bypasses OS stack
- RDMA delivers, low latency, high throughput, zero copy capabilities
- RDMA Hardware Technologies
 - **RoCEv2: RDMA over Converged Ethernet**
 - iWARP: RDMA over TCP/IP
 - Infiniband



Why does the network matter for AI/ML?



Data Center Quantized Congestion Notification

- IP ECN (Explicit Congestion Notification) or PFC (Priority Flow Control) alone cannot provide a valid Congestion Management framework
- The **two of them together** provide the desired result of having lossless RDMA communications across Ethernet networks (this is called **DCQCN**)
- The requirements are:
 - Ethernet devices compatible with both techniques
 - Proper configurations applied

-
- Elephant v. Mice flows
 - AFD and Smart Buffers

ROCEV2
WITHOUT CONGESTION
MANAGEMENT



ROCEV2
WITH ECN



ROCEV2
WITH PFC



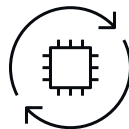
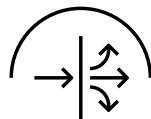
ROCEV2 WITH
ECN AND PFC



RoCEv2 as a Scale-out Transport Protocol

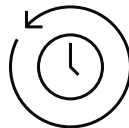
Good, but can be improved upon

PFC Requires Intense Buffering



Victim flows, congestion trees, PFC storms and deadlocks

Go-back-N retransmission



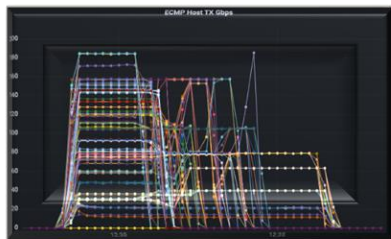
Congestion control relies on pause and timeout

Perfect Load Balancing and Congestion Control Exists Today...

...*inside* distributed systems by Cisco's perfectly load-balanced, congestion-free fabrics.

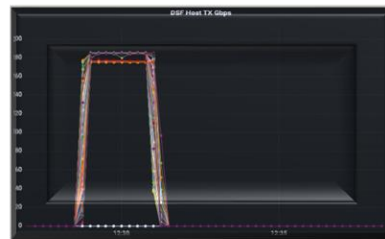
Today's hardware-based output-queued crediting mechanism used in our class-leading modular chassis can be applied at the network level with **Cisco Scheduled Ethernet**, removing the network barrier to job completion time

Traditional ECMP Ethernet Traffic
Pattern by Host



Tail Latency = Idle GPUs

Scheduled Ethernet
Enabled Traffic by Host

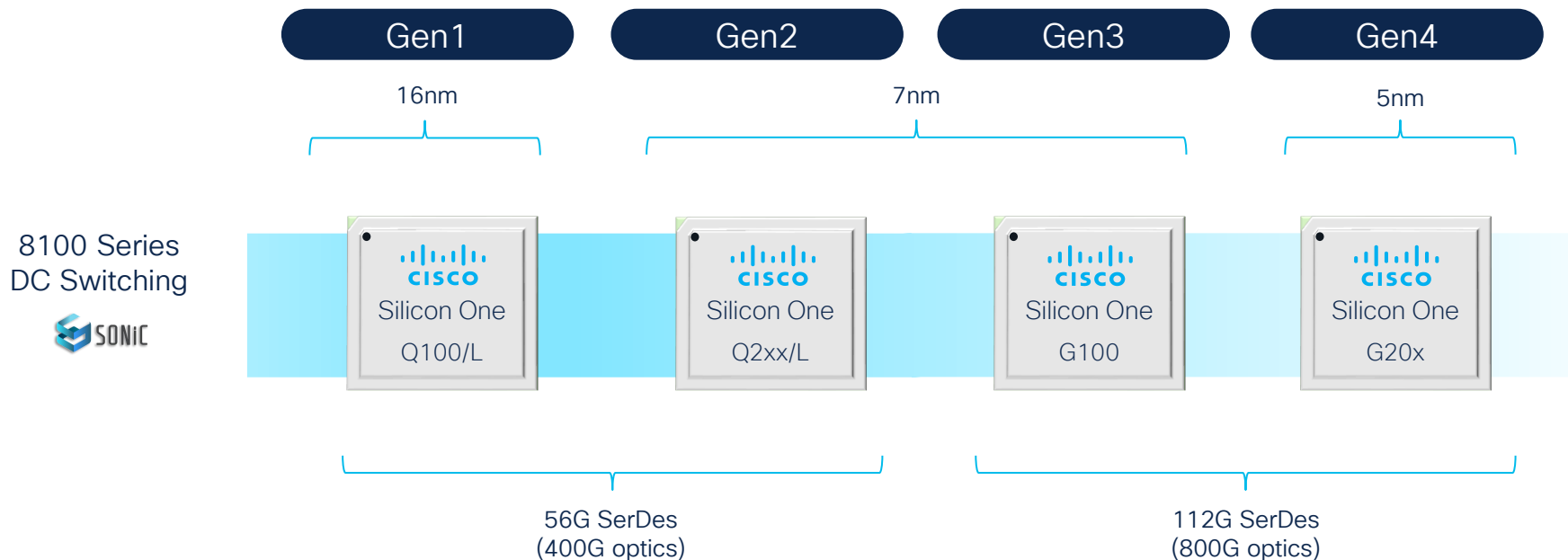


Synchronizing Begins Sooner
Job Completes 2x Faster!



**Cisco Live US 2024– check out BRKNWT-3406
for a deeper dive into Scheduled Ethernet!**

Cisco Silicon One - Evolution



Cisco Silicon One G200

Uniquely efficient and Optimized for AI/ML

One architecture

A simpler and easier network to maintain



High Performance

2x higher performance than G100



Sustainability via technology

2x more power efficient than G100



Ultra-low latency

2x Lower Latency than G100



Optimal network design

512-wide radix enables flatter, more efficient networks



Fully shared packet buffer

Optimal fairness, burst performance, JCT



51.2 Tbps



Advanced 112 Gbps SerDes

Cisco designed next-generation ADC SerDes
Support for Optics, 4-meter DAC, LDO and CPO

Advanced load balancing

Non-correlated WECMP avoids hash polarization
Congestion-aware stateful load balancing
Congestion-aware packet spraying



Link failure avoidance

HW based traffic link failure redistribution optimizes real-world large-scale deployments



Programmable processor

Deterministic ultra-low latency processor with run to completion for ultimate flexibility

435B+

Lookups per second

Enables advanced features like SRv6 uSID



Deep visibility & Analytics

In-band telemetry including emerging protocols
Hardware analyzers enable post event debuggability

cisco Live!

Cisco 8122-64EH



64 x 800G

Shipping Q3 2024



Silicon One

G200

Hardware Summary

Single 51.2T G200 ASIC (5nm)
256 MB SRAM packet buffer

Quad Core x86 CPU
64 GB DRAM

RS-232, and 10 GbE control plane expansion.
1 GbE Management Ports
2xSFP+ PIE (Punt-Inject Engine) , 1xUSB 2.0

4 Fans, 1+1 PSU Redundancy
Port Side Intake airflow

2kW AC & 3kW AC & DC PSUs

(H) 3.45 x (W) 17.3 x (D) 24.7 in.
(H) 8.76 x (W) 43.95 x (D) 67.4 cm
63.5 lbs (29.5 kg)

- 51.2T G200 Optimized for high-radix DC and AI applications
- 2x400 GbE per cage
- ETC 800 GbE support
- Lowest power consumption per-bit switch
- 512 x 100G – full 512 interface network radix
- Exceptional Serdes performance for powering LPO optics



AI Ethernet Fabric Options

	1	2	2 _u	3	
	Ethernet	Enhanced Ethernet		Ultra Ethernet	Scheduled Ethernet
Load Balance	Stateless ECMP	Stateful Flow/Flowlet	Spray & Re-order in SmartNIC	Endpoint Controlled adaptive packet spraying	Spray & Re-order in leaf
Fabric Congestion Management	Congestion Reaction with ECN/PFC	Congestion Reaction with congestion score to adjust distribution		Network influenced Congestion Management	Congestion Avoidance
Link Failure	Software	Hardware			
Job Completion Time	Good	Better		Even Better	Best
Coupling between NIC and Fabric	No		Yes		No
Place in Network	Frontend, Backend			Backend	Frontend, Backend

Effectiveness IS dependent on Traffic Characteristics

Effectiveness IS NOT dependent on Traffic Characteristics

Ultra Ethernet – Vision

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

**THE NEW ERA
NEEDS A
NEW
NETWORK**

*Ultra***Ethernet**

As **performant** as a
supercomputing interconnect

As **ubiquitous** and
cost-effective as Ethernet

As **scalable** as a cloud data
center

Want to Know More?

BRKNWT-2407 – Cisco Live US 2024
Available in Cisco Live Online Library



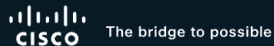
Ramesh Sivakolundu

Director, Technical Marketing



Scott Carter

Leader, Product Management



Building Ethernet AI Fabrics with Silicon One

Designs, Vision, and Challenges

Ramesh Sivakolundu
Scott Carter
BRKNWT-2407

CISCO *Live!*

#CiscoLive

CISCO *Live!*

Why Networking is Relevant to AI Deployments

LLMs are orders of magnitude more intensive than DLRM



Deep Learning Recommendation Models

Search, Feed ranking. Ads & content recommendation

Inference needs a few Gigaflops for 100ms TTFT

Narrower scope, domain specific

Training: ~100 Gigaflop/ sentence



Large Language Models

Intricacies of human language

Inference needs 10s of Petaflops for 1 sec TTFT

Generate intelligent, creative responses

Training : ~1 Petaflop/ sentence

An Improved user experience means a *faster time to first token*, making *distributed inference an imperative*

On-Prem AI Solutions

Open Source LLMs with Cisco
UCS and Nexus

CISCO *Live!*



Patrick

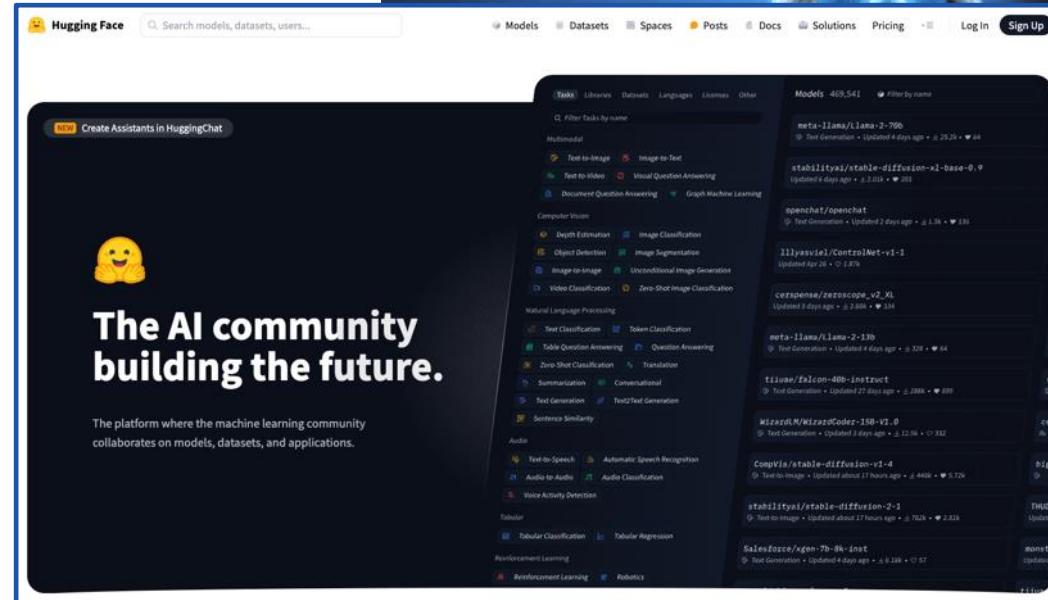
On-Prem AI Solutions:

Business Challenges of Proprietary GenAI Systems

1. **Cost:** Cloud-based GenAI solutions are becoming a lucrative source of recurring revenue for OpenAI, Google, Anthropic, etc.
2. **Privacy and Security:** Commercially available LLMs are something of a “black box” for users. The users have no control over how they were trained, bias, etc.
3. **Training Gap:** Models are usually months or years out of date. Requires Retrieval Augmented Generation (RAG)
4. **Fine Tuning:** Cloud-based LLMs are “foundational models” and lack fine tuning for certain use cases

Meta has a different Approach – open source AI LLaMA (Large Language Model Meta AI)

- LLaMA (Large Language Model Meta AI) is an LLM released by Meta AI in February 2023
- Mar 3, 2023 (a week after it's release) LLaMA's model was leaked to the open source community
- Unlike other LLMs, Meta LLaMA is now available to the Open Source community



Open Source AI and Responsible AI Use



Cisco believes in **data privacy**



Cisco has a **Responsible AI (RAI) framework** and **controls** around transparency, fairness, accountability, privacy, security and reliability



With the emergence of Generative AI, we are **evolving our RAI framework and controls** to surface new variants of issues, such as (but not limited to):

- False content
- Hallucinations
- Bias
- Unanticipated output
- Harmful output
- Deep fakes

Please keep in mind that **open source AI tools and capabilities** may or may not have been checked against issues such as these (or others).

As such, **please use any such tools at your own risk, and in accordance with any applicable AI policies from your organizations.**



“Working closely with Cisco, we’re making
it easier than ever for enterprises to
obtain the infrastructure they need to
benefit from AI.”

Jensen Huang
Founder and CEO, NVIDIA

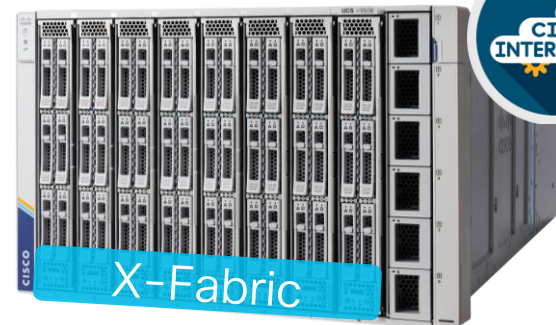


Nexus AI Networking

cisco *Live!*



UCS C885A – 8 x GPU Server

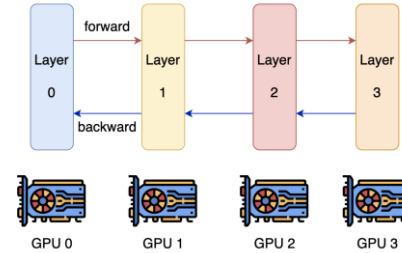


UCS-X Series – Modular Chassis

Splitting AI Model Across GPUs

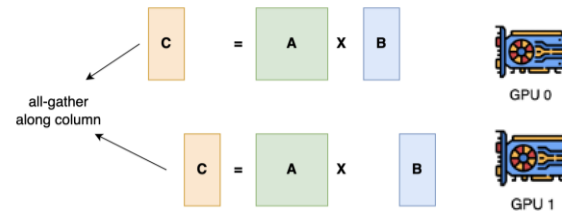
Pipeline Parallelism

Different layers to different GPUs.
Usually on GPU on different hosts
(lower bit rate)



Tensor Parallelism

Layer computations sent to different GPUs.
Best for multi-GPU single host
(higher bit rate)



Demo Time!



www.github.com/pl247/ai-toolkit

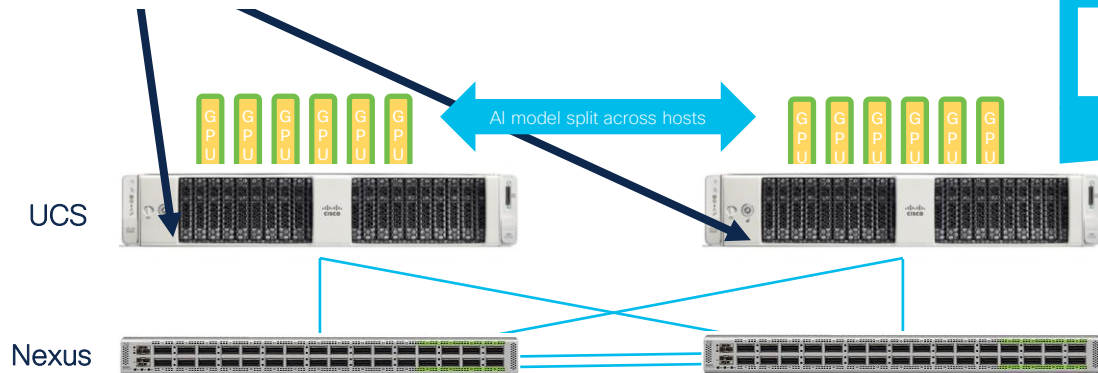
Demo

Multi-GPU Multi-Host



Intersight

- Deploy Server Profile
- OS (ubuntu) & AI Tools



cisco *Live!*

OI

Open WebUI

LLM

LLM Server



Hugging Face

Models



RAY

Clustering Framework



AI
Enterprise

GPU Driver & CUDA



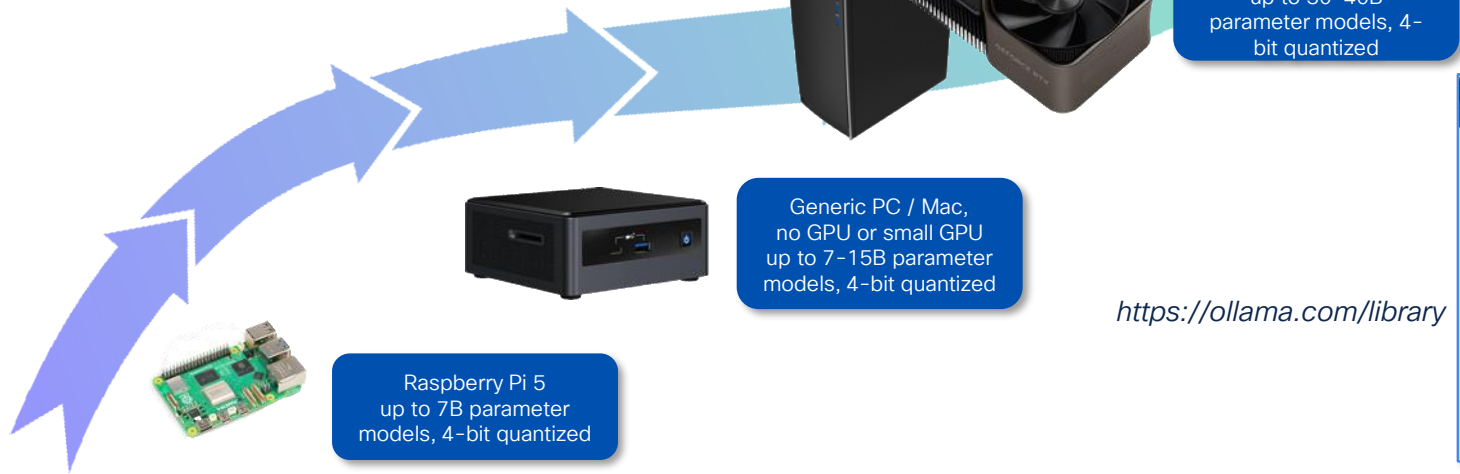
Ubuntu

Operating System

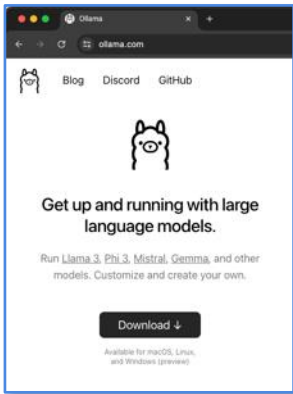
And Keep in Mind ...

You can experiment with all of the AI capabilities we are showing, and get your hands dirty, almost no matter what hardware you can get your hands on ...

Model quantization is the process of reducing the precision of a model's parameters from floating-point to lower bit-width representations (such as 8-bit integers) to decrease its memory footprint and computational requirements while aiming to maintain accuracy.



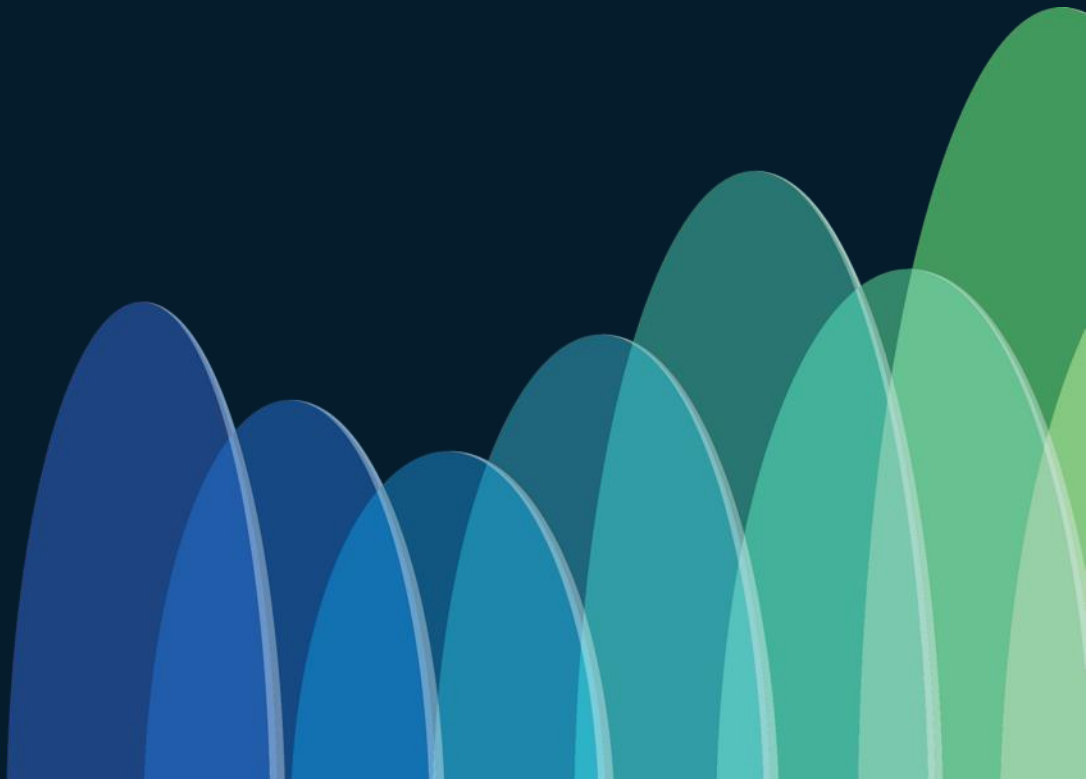
<https://ollama.com/library>



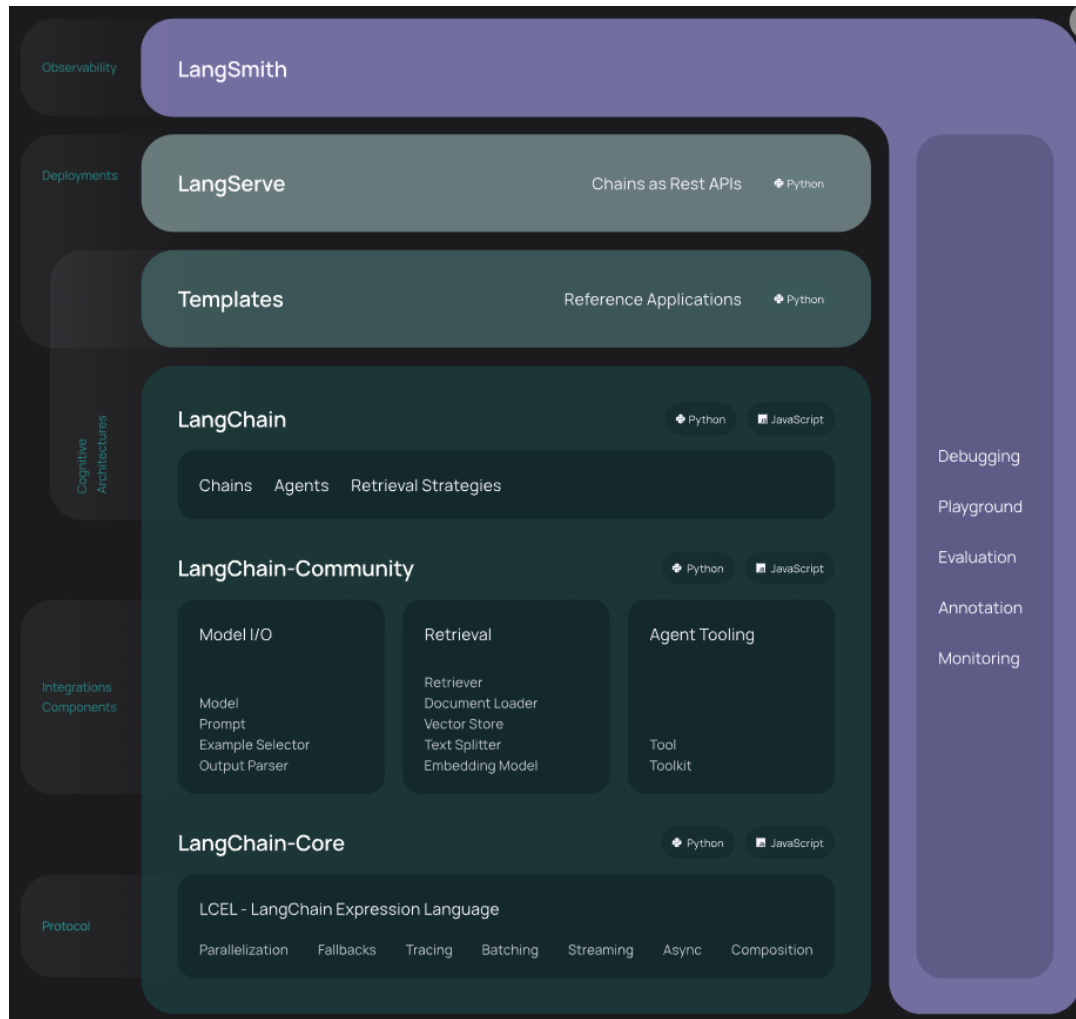
Large Language Models (LLMs), GPTs, and more ... and their use in Networking



Langchain



The LangChain Ecosystem



An Open-Source Framework for AI Development

- Python or JavaScript
- Highly abstracted
- Easy
- Accessible

Integrated openAI

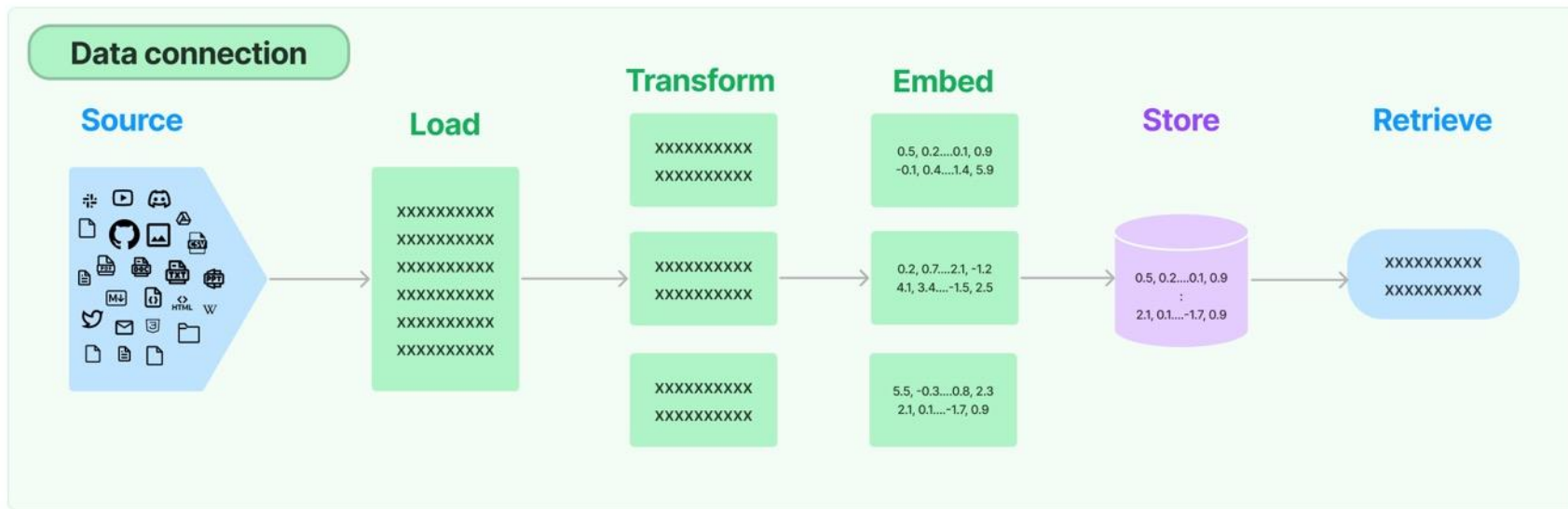
Options for *private LLMs*

Evolving quickly and daily

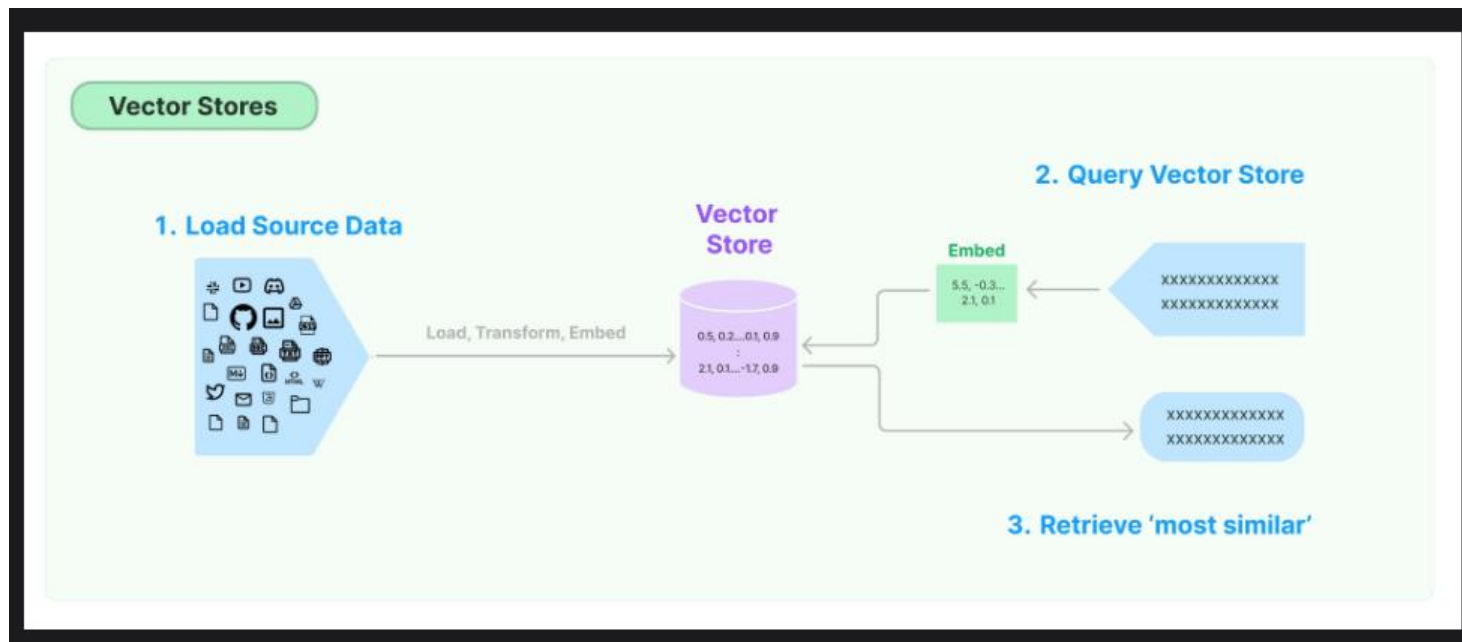
“Like developing on wet cement”

LCEL (Langchain Expression Language)

A High-Level LangChain Abstraction



Vector Store



```

#Load OpenAI API Key
os.environ["OPENAI_API_KEY"] = openai.api_key

#Setup Loader - in this case a PDF Loader
loader = PyMuPDFLoader("cisco-annual-report-2022.pdf")

# Load and split the pdf into pages
pages = loader.load_and_split()

# setup a text splitter
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=250,
    chunk_overlap=20,
    length_function=len,
)

```

```

# split the pages into smaller chunks called docs
docs = text_splitter.split_documents(pages)

# transform to embeddings
embeddings = OpenAIEmbeddings()

# setup and store docs and embeddings into ChromaDB
vectordb = Chroma.from_documents(docs, embedding=embeddings,
| | | | | | | | | persist_directory=".")

#Make the database persisten
vectordb.persist()

# setup memory so it remembers previous questions and answers
memory = ConversationBufferMemory(memory_key="chat_history", return_messages=True)

```

```

# perform the Conversational Retrieval Chain
qa = ConversationalRetrievalChain.from_llm(OpenAI(temperature=0.5) , vectordb.as_retriever(), memory=memory)

#Run the question
question = "Who is the senior leadership team at Cisco?"
result = qa.run(question)

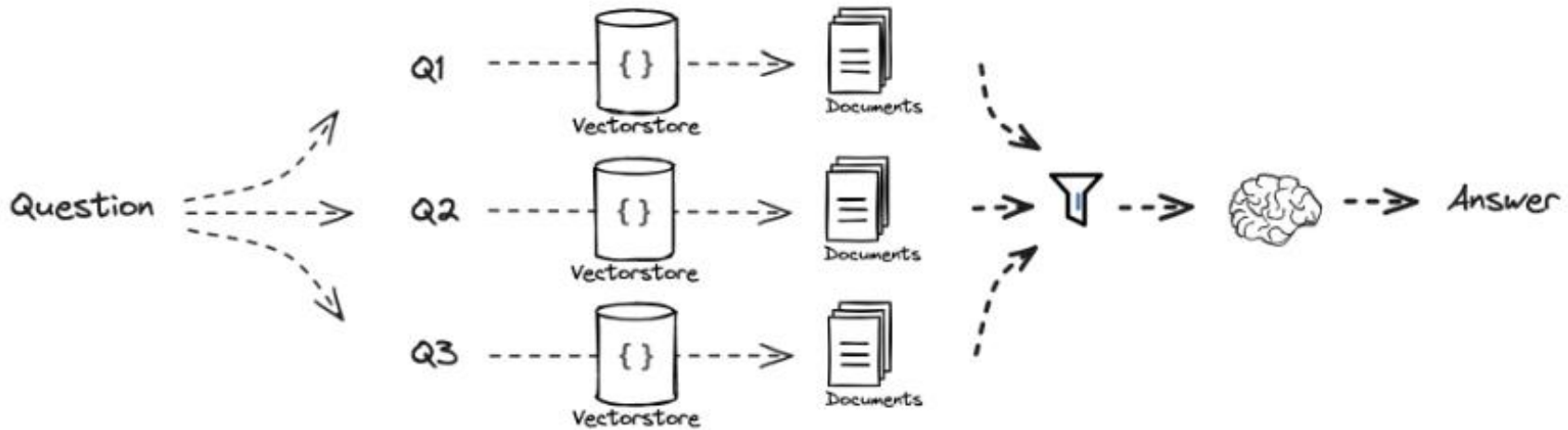
#Print the results to the screen
print(result)

```

Advanced RAG – RAG-Fusion

Part 6: RAG-Fusion

Flow:



[rag-from-scratch/rag_from_scratch_5_to_9.ipynb](#) at main · langchain-ai/rag-from-scratch · GitHub

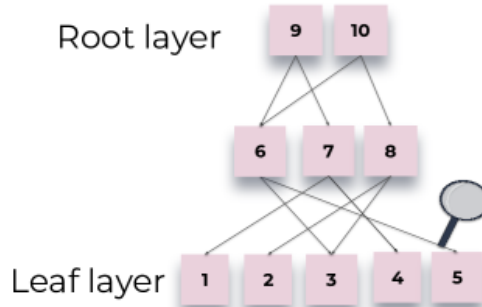
The drawbacks and flaws of RAG approaches

[RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval \(arxiv.org\)](https://arxiv.org/abs/2310.07551)

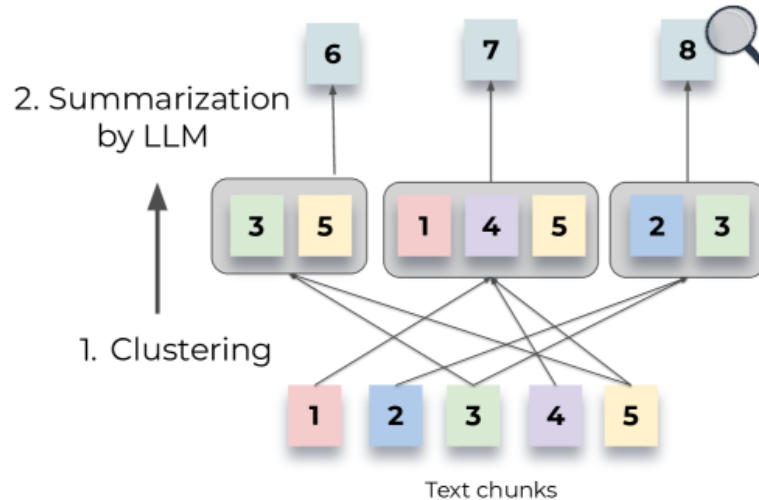
Nevertheless, existing retrieval-augmented approaches also have flaws. The one we tackle is that most existing methods retrieve only a few short, contiguous text chunks, which limits their ability to represent and leverage large-scale discourse structure. This is particularly relevant for thematic questions that require integrating knowledge from multiple parts of a text, such as understanding an entire book, as in the NarrativeQA dataset (Kočiský et al., 2018). Consider the fairy tale of Cinderella, and the question “How did Cinderella reach her happy ending?”. The top- k retrieved short contiguous texts will not contain enough context to answer the question.

Look familiar? Like a Well Designed Data Center!

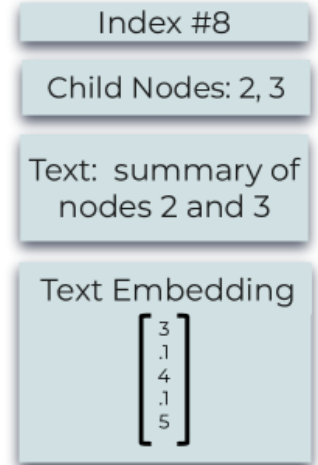
RAPTOR Tree



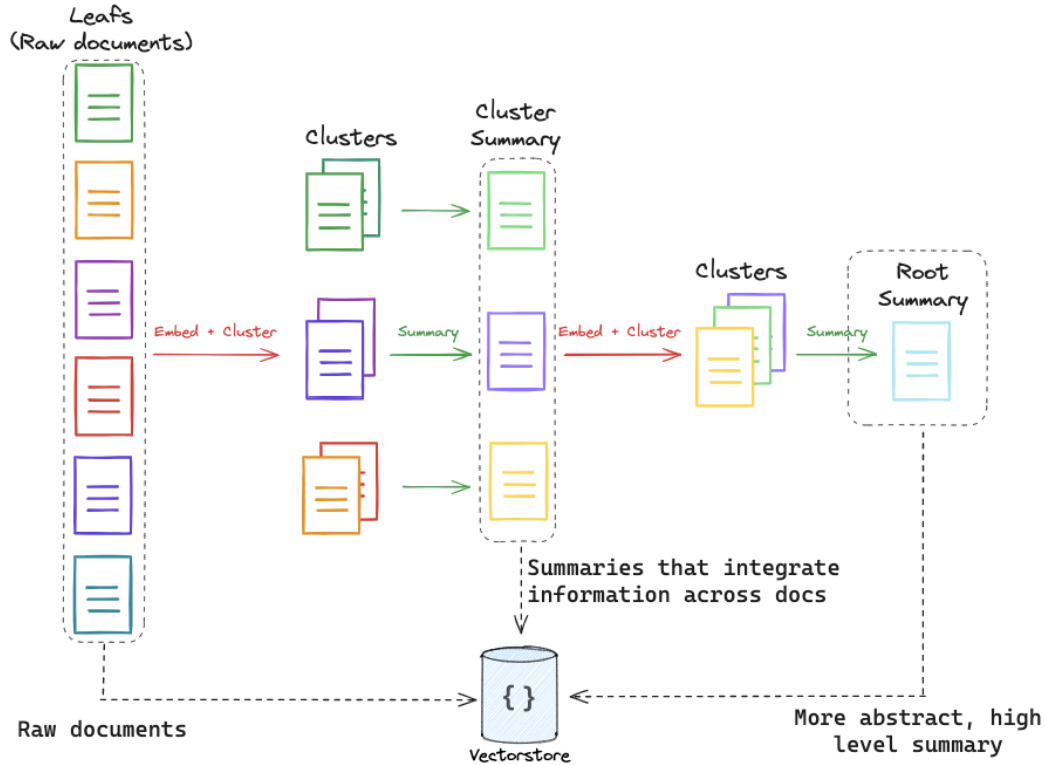
Formation of one tree layer



Contents of a node



RAPTOR (Document tree at varying levels of abstraction)



[langchain/cookbook/RAPTOR.ipynb at master · langchain-ai/langchain · GitHub](https://github.com/langchain-ai/langchain/blob/master/cookbook/RAPTOR.ipynb)

Open Source: Free; Private; Local



**Get up and running with large
language models.**

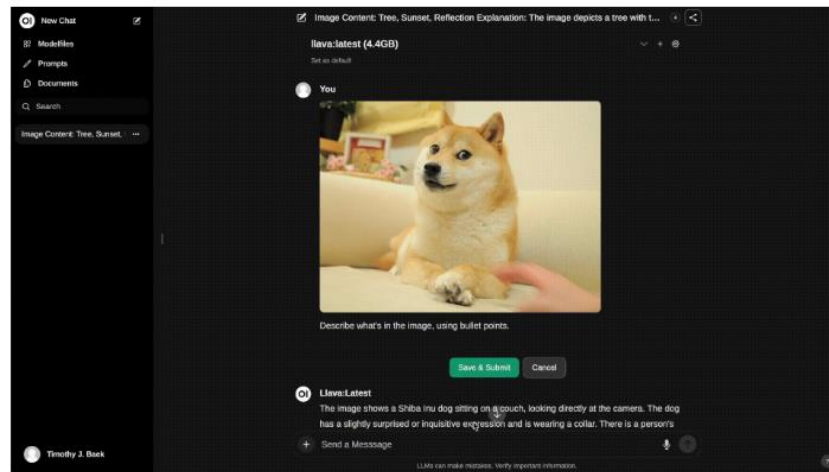
Run [Llama 2](#), [Code Llama](#), and other models.

Customize and create your own.

Open WebUI

Open WebUI

Open WebUI is an extensible, feature-rich, and user-friendly self-hosted WebUI designed to operate entirely offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs.



Pick Your Model(s)

 **Models**

Filter by name...

Most popular

llama2

Llama 2 is a collection of foundation language models ranging from 7B to 70B parameters.

↓ 1.1M Pulls ↓ 102 Tags ⌚ Updated 2 months ago

gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

↓ 642.8K Pulls ↓ 102 Tags ⌚ Updated 3 days ago

mistral

The 7B model released by Mistral AI, updated to version 0.2.

↓ 524.8K Pulls ↓ 68 Tags ⌚ Updated 2 weeks ago

codellama

gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

↓ 642.8K Pulls ⌚ Updated 3 days ago

latest

↓ 102 Tags

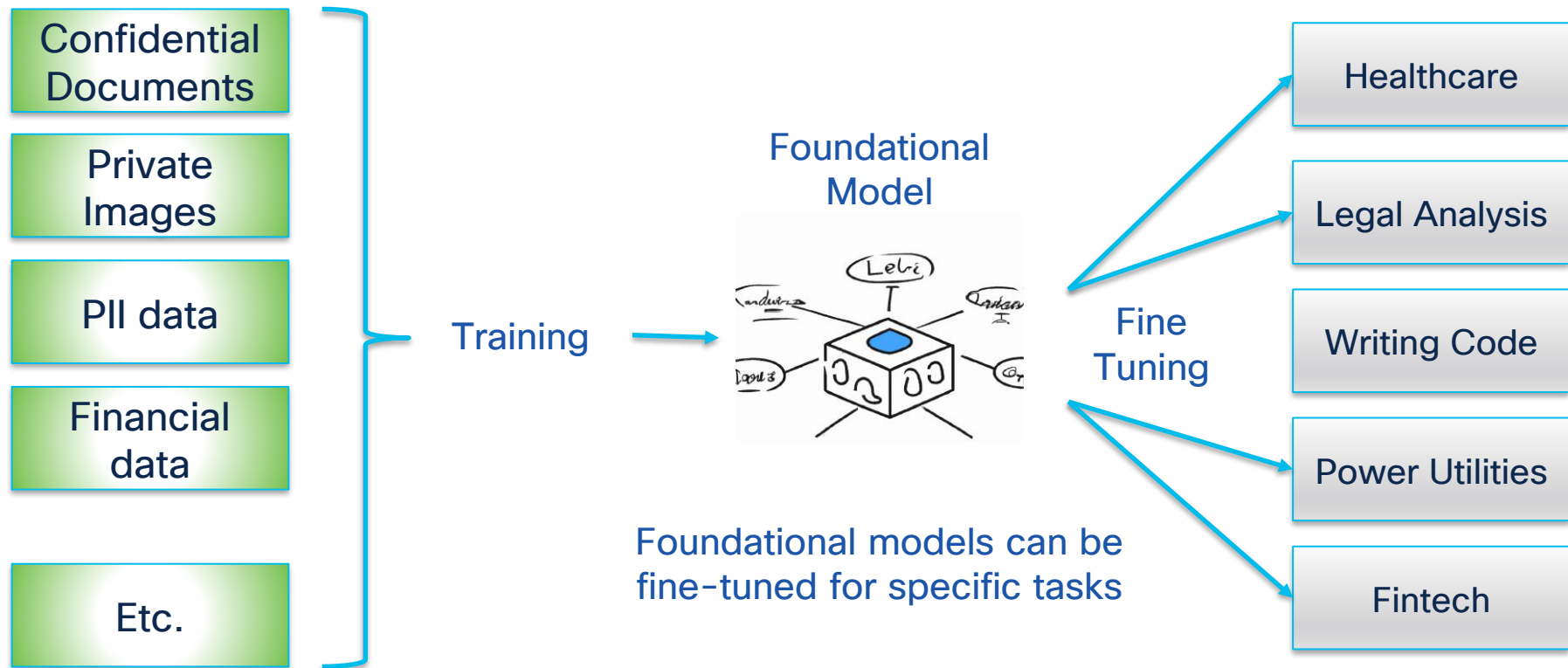
ollama run gemma

Updated 6 days ago

a72c7f4d0a15 · 5.0GB

model	family gemma · parameters 9B · quantization 4-bit	5.0GB
license	Gemma Terms of Use Last modified: February 21, 2024 By using, re...	8.4kB
template	<start_of_turn>user {{ if .System }}{{ .System }} {{ end }}{{ .P...	136B
params	{"penalize_newline":false,"repeat_penalty":1,"stop":["<start_of_...	109B

Foundational vs. Fine Tuned Models



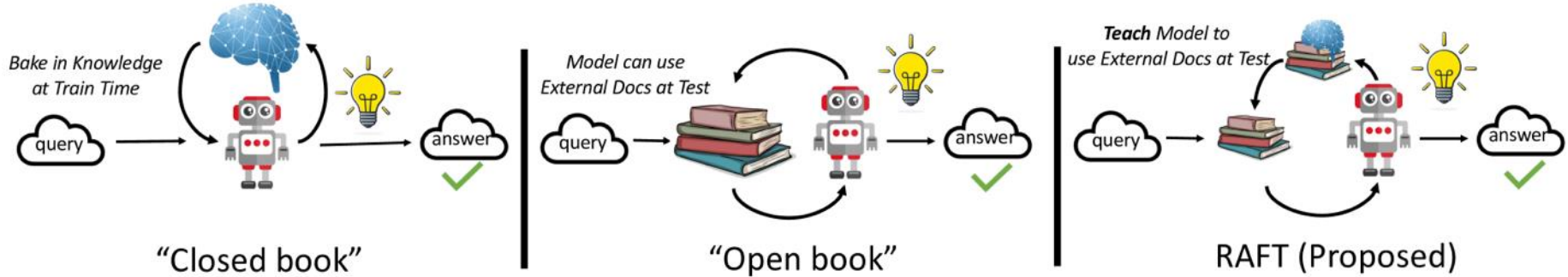
Fine Tuning

Fine-tuning is the process of taking a pretrained machine learning model and further training it on a smaller, targeted, domain-specific, data set. The aim of fine-tuning is to maintain the original capabilities of a pretrained model

1. **Setup:** Loading our dataset and filtering down to one domain to fine-tune on.
2. **Data preparation:** Preparing your data for fine-tuning by creating training and validation examples, and uploading them to the `Files` endpoint.
3. **Fine-tuning:** Creating your fine-tuned model.
4. **Inference:** Using your fine-tuned model for inference on new inputs.

https://github.com/openai/openai-cookbook/blob/main/examples/How_to_finetune_chat_models.ipynb

RAFT (Retrieval Augmented Fine Tuning)



[RAFT: Adapting Language Model to Domain Specific RAG \(arxiv.org\)](https://arxiv.org)

[RAFT \(Retrieval Augmented Fine-tuning\): A new way to teach LLMs \(Large Language Models\) to be better at RAG \(Retrieval Augmented Generation\) \(microsoft.com\)](https://microsoft.com)

Demo Time!



Learn how to build your own

Room 216
Wednesday, 13
November
@15.00



Unlock the Future:

Introduction to GenAI and Retrieval Augmented Generation (RAG) techniques for your networks

Patrick LeMaistre

Solutions Engineer

Prapanch Ramamoorthy

Principal Engineer

Dave Zacks

Distinguished Engineer

BRKAPP-2562

CISCO *Live!*

#CiscoLiveAPJC

CISCO *Live!*

Keynote Deep Dives

Wednesday
10:30am – 11:30am



Experiences Amplified:
How AI Can Fuel
Better Employee
and Customer
Experiences

Level 1
Room 106



Smart, Secure,
Seamless:
Transforming
Experiences with
Next-Generation
Networking

Level 2
Room 204



Harness a Bold
New Era:
Transform Data
Centre and
Service Provider
Connectivity

Level 2
Room 203



Securing User to
Application and
Everything in
Between

Level 2
Melbourne Room 2



Unlocking
Digital Resilience
through Unified
Observability

The HUB
Centre Stage

Complete Your Session Evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a **Cisco Live T-Shirt**.



Complete your surveys in the **Cisco Live mobile app**.



Continue your education



- Visit the Cisco Stand for related demos
- Book your one-on-one Meet the Expert 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



Thank you



#CiscoLiveAPJC

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC