
cisco Live!
Let's go

The bridge to possible

Talking to Your Network

Leveraging the Power of AI and LLMs for Network Visibility and Control

John Capobianco, Technical Leader

Dave Zacks, Distinguished Engineer

#HighBitRate

Cisco Live!

By Way of Introduction ...

I am a **Distinguished Engineer** in the Cisco Transformation team, and have been with Cisco for 24 years.

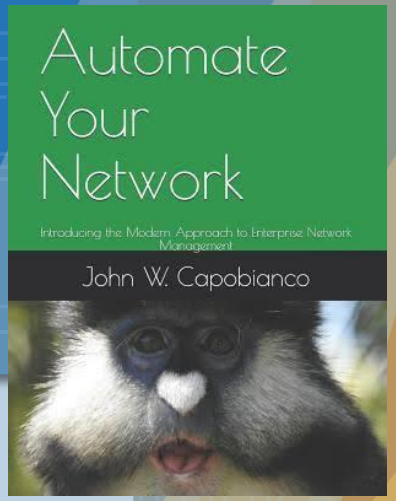
I work primarily with large, high-performance Enterprise network architectures, designs, and systems. I have over 39 years of experience with designing, implementing, and supporting solutions with many diverse network technologies.

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, Network Fabrics, End-to-End Assurance & Analytics, and ML/AI for Networking.**

Dave Zacks
Distinguished Engineer
dzacks@cisco.com @DaveZacks

By Way of Introduction ...



I am a **Technical Leader** in the Cisco Machine Learning / AI team. I have 20+ years experience in the IT industry, and most recently before Cisco worked as a network architect for the Canadian Parliament.

I have previously presented on both network automation (at Cisco Live 2015), Ansible (2015 – 2019), and last year on AI and LLMs at Cisco Live and other events worldwide.

I have a self-published book “Automate Your Network”, available on Amazon (2019).

John Capobianco
Technical Leader
johcapob@cisco.com

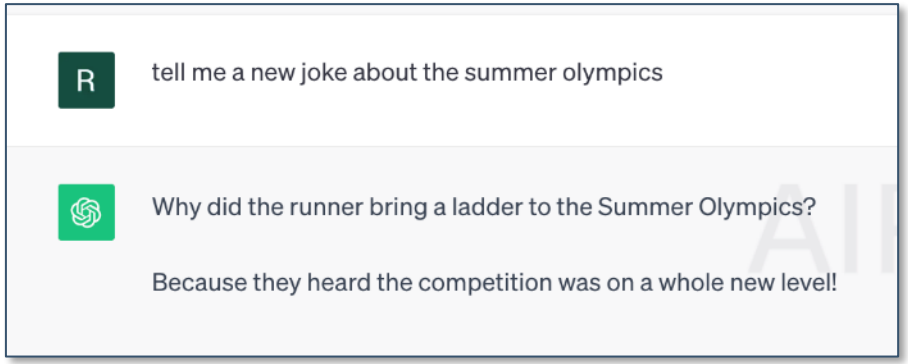
Agenda

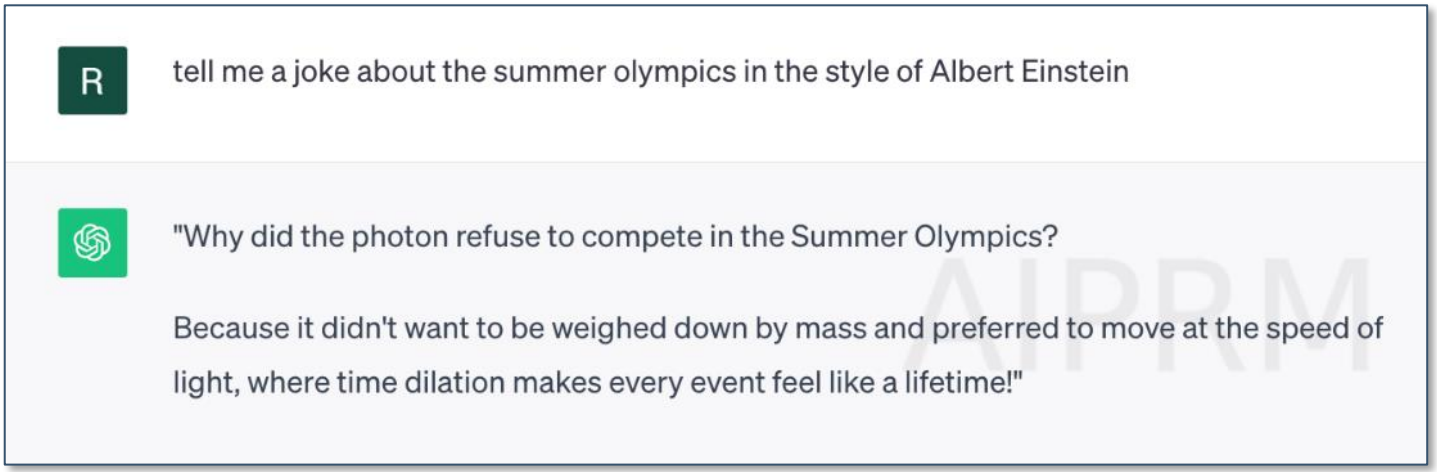
- Introduction – ML/AI, GenAI, Neural Networks, Transformers …
- Where is Cisco using AI?
  - Enterprise Networks
  - Collaboration
  - Security
  - Building Networks for AI/ML
- Leveraging LLMs for Networking
  - What are LLMs?
  - How do LLMs Work?
  - Demo Time!
- Conclusions

The world has gone crazy for Generative AI!

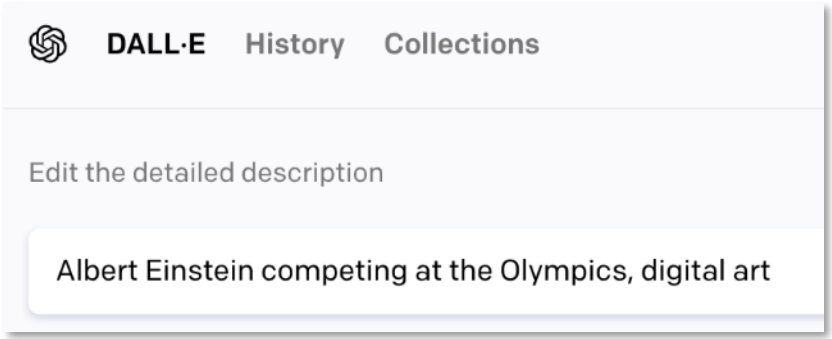
Write, Debug, and Explain Code – Write Poetry – Write a Resume – Create a Clever Story – Do Homework – Design New Recipes – Plan Your Vacation – Help You Learn – Draw Pictures – Compose Music – Answer General Knowledge Questions – Much More!

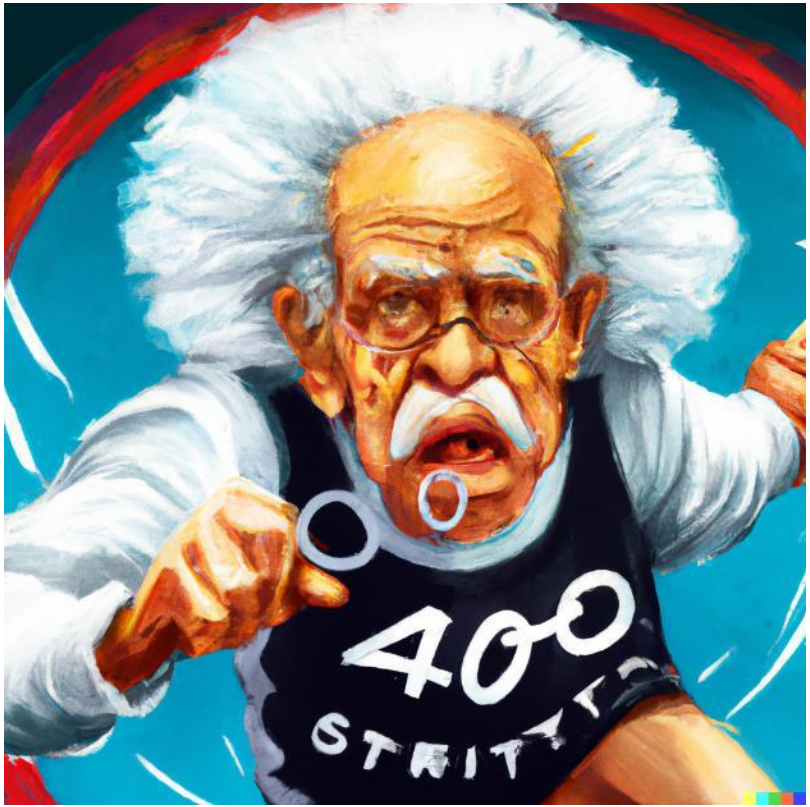
Writing Jokes (good and bad ones)





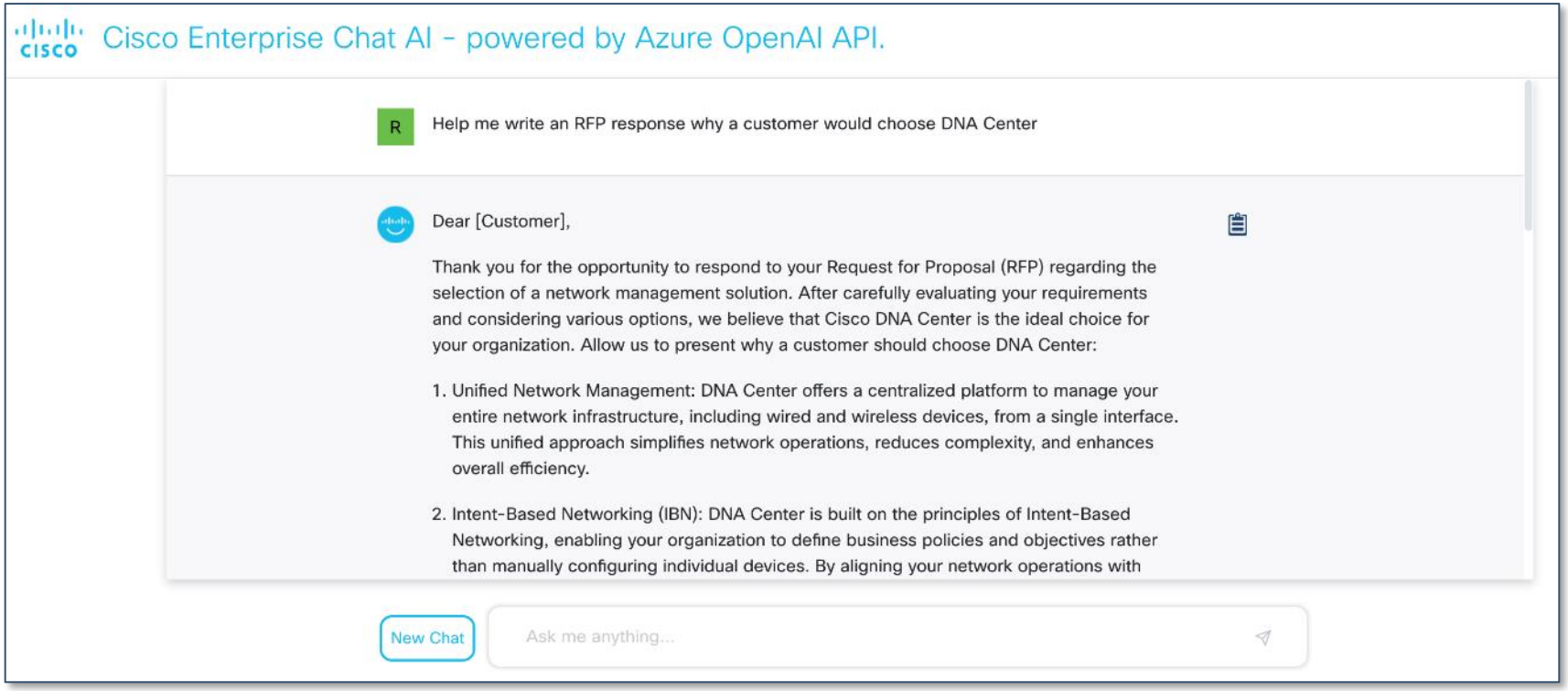
Drawing Things (drawn using OpenAI's DALL-E2)



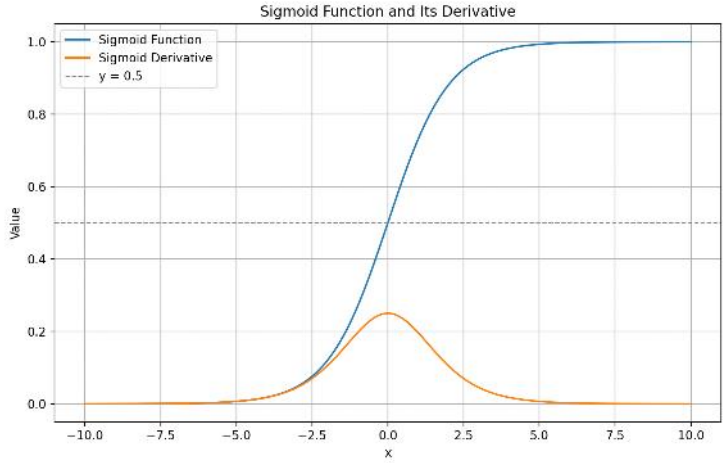


Note: Dall-E is based on a 12-billion-parameter version of GPT-3

Writing Blogs and RFP Responses



Coding, displaying output, debugging, explaining code



R

using python, plot the sigmoid function from x=-10 to x=10. On the same plot, show the first derivative of the sigmoid function.

Sure, I can help you with that! You can use the `numpy` library to calculate the sigmoid function and its derivative, and the `matplotlib` library to create the plot. If you haven't already, you'll need to install these libraries using the following command:

bash — Copy code

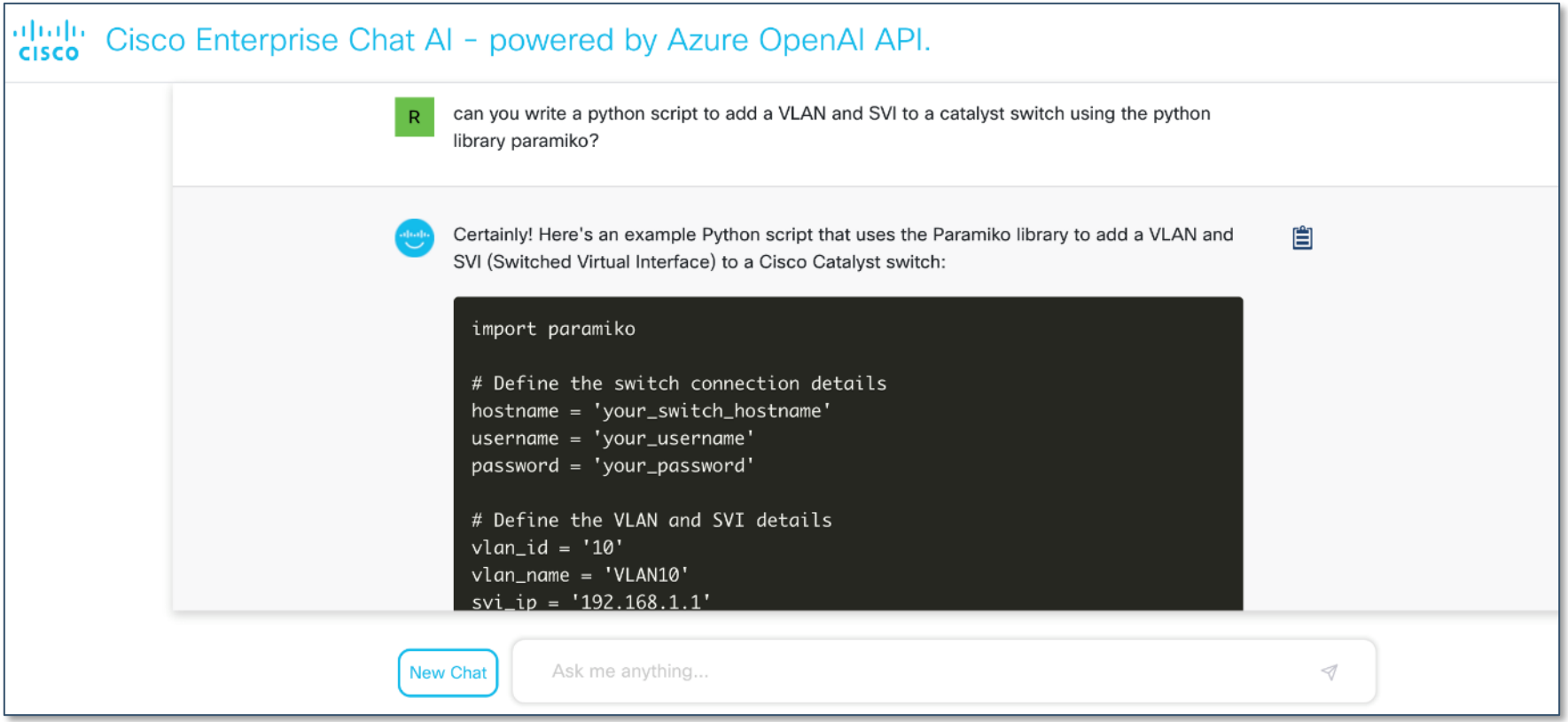
```
pip install numpy matplotlib
```

Here's the code to plot the sigmoid function and its first derivative:

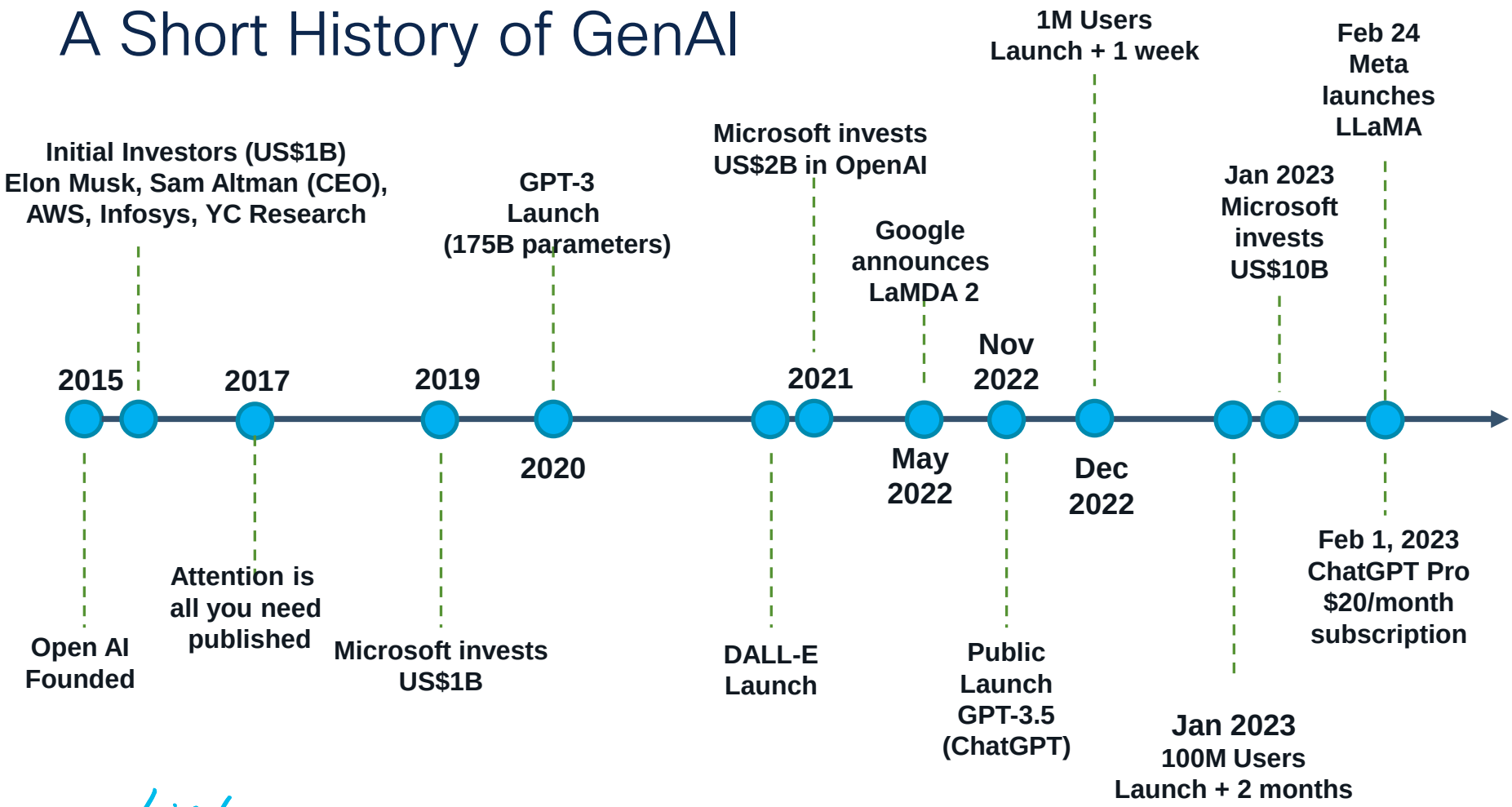
python — Copy code

```
import numpy as np
import matplotlib.pyplot as plt

def sigmoid(x):
    return 1 / (1 + np.exp(-x))

def sigmoid_derivative(x):
    return sigmoid(x) * (1 - sigmoid(x))
```

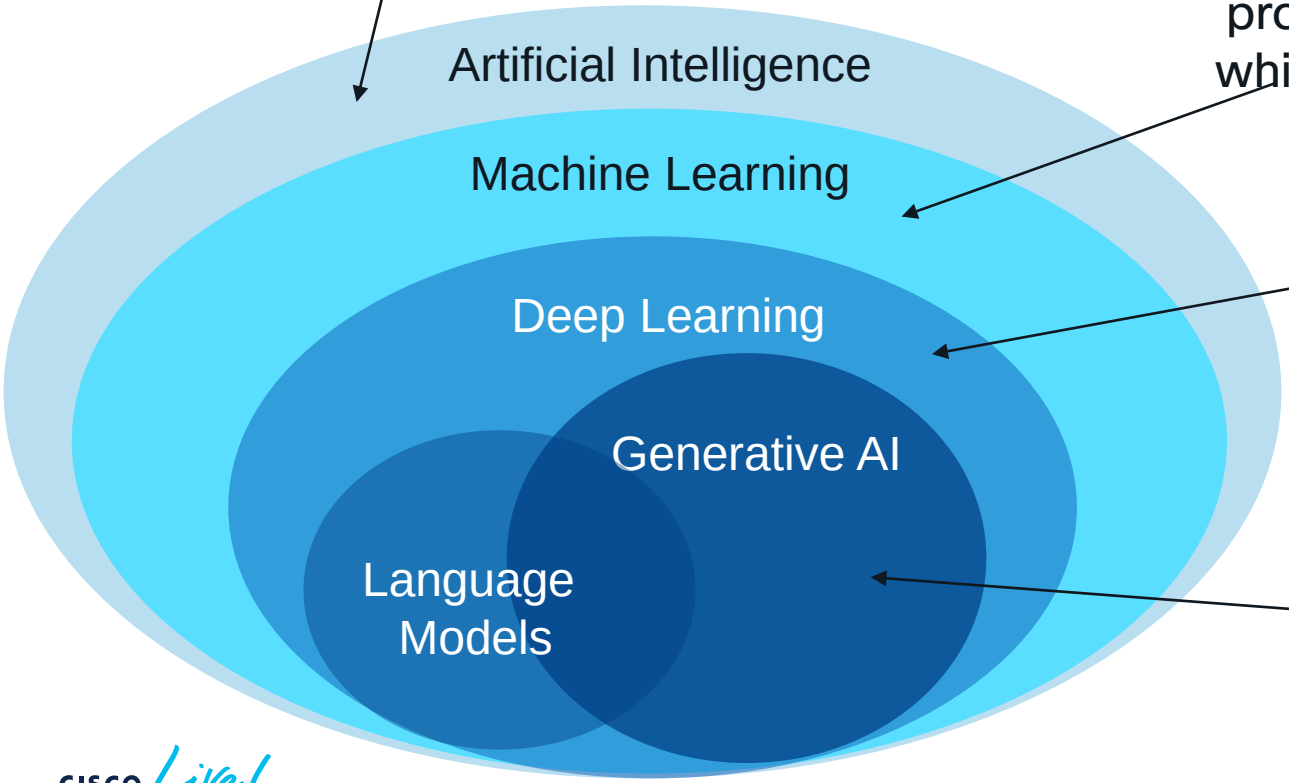
Writing Automation Code for Your Network



A Short History of GenAI



AI encompasses the whole world of ML and Deep Learning



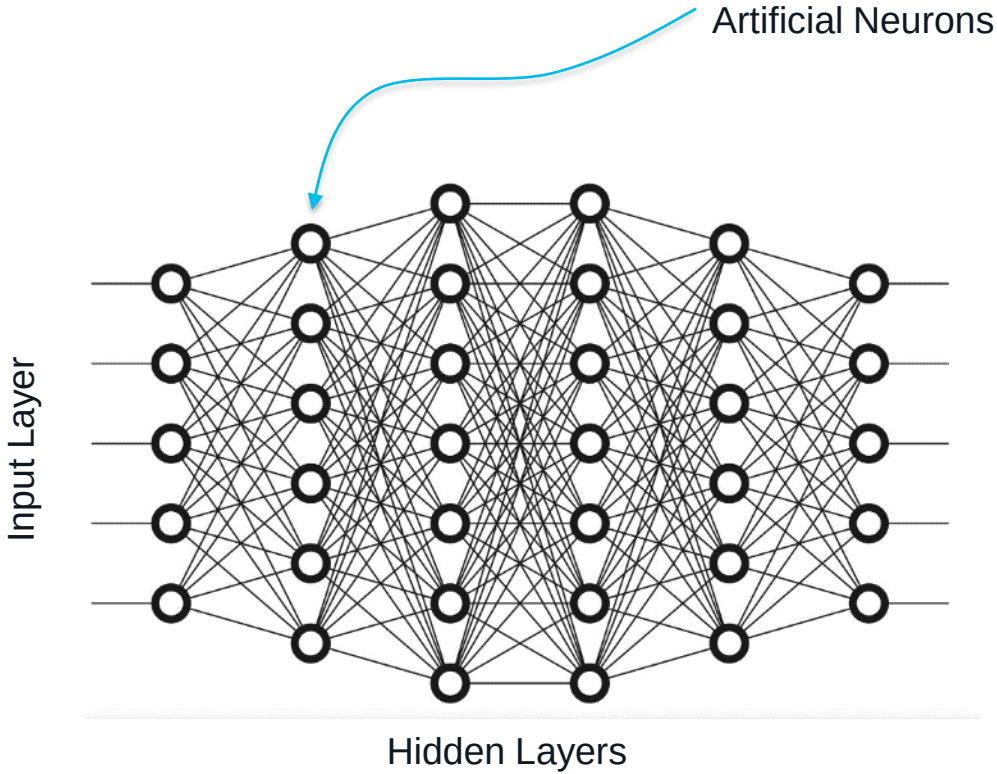
An AI technology where the rules are not set in the program, but are learned while the program is used

A form of ML that uses Neural Learning Networks

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

Built upon Deep Learning / Neural Networks

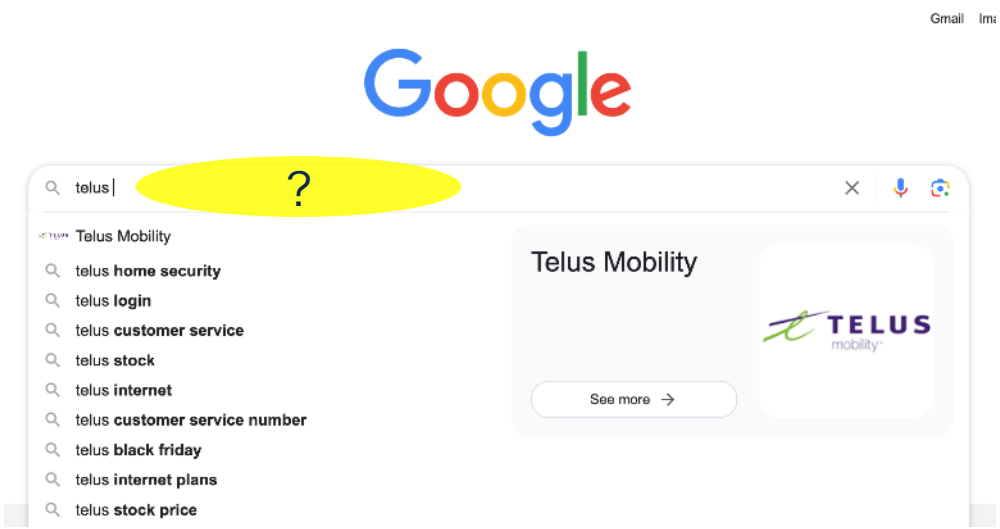
- Neural networks have become the cornerstone of most modern AI systems
- Exceptionally good at image analysis, next-word prediction, speech analysis, and much more
- Heavily reliant on GPUs (good for Nvidia)



Geoffrey Hinton – the “Godfather” of Deep Learning

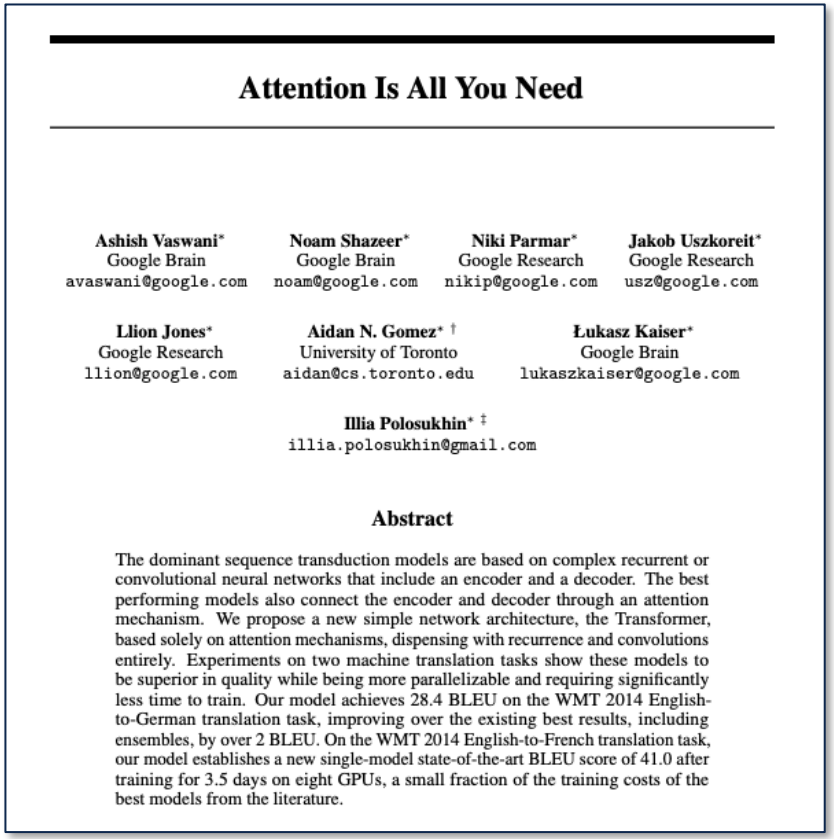
(Photo by Johnny Guatto / University of Toronto)

From Word Prediction to Generative AI
How did we get here?



Simple Recurrent Neural Networks (RNNs) assign a probability to all of these words, with highest probability ranked first

The Paper that Changed the World



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

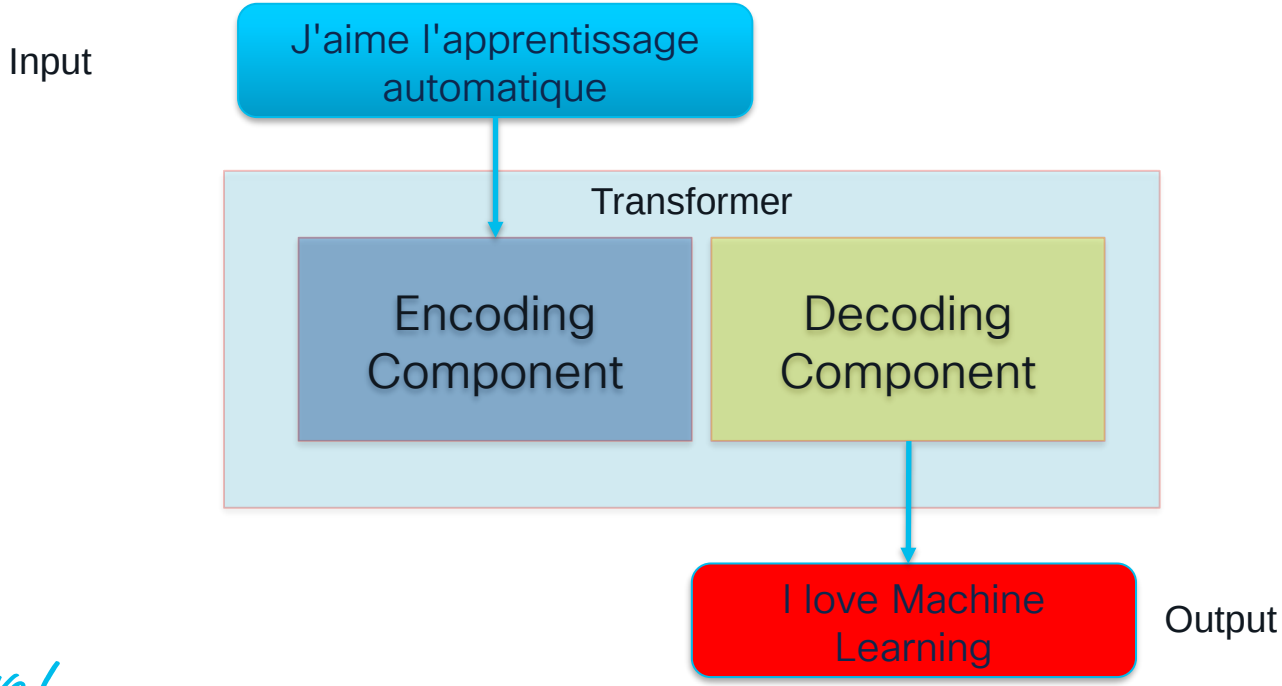
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

In 2017 a revolutionary paper by researchers at Google and the University of Toronto introduced a new architecture that gave birth to the world of Generative AI

Transformers – the Secret Sauce of GenAI

The transformer's self-attention mechanism allows the system to grasp deep meanings of not just words, but sentences, paragraphs, whole stories



ChatGPT is King of the Hill (for now)

Runs on the worlds

5th Largest

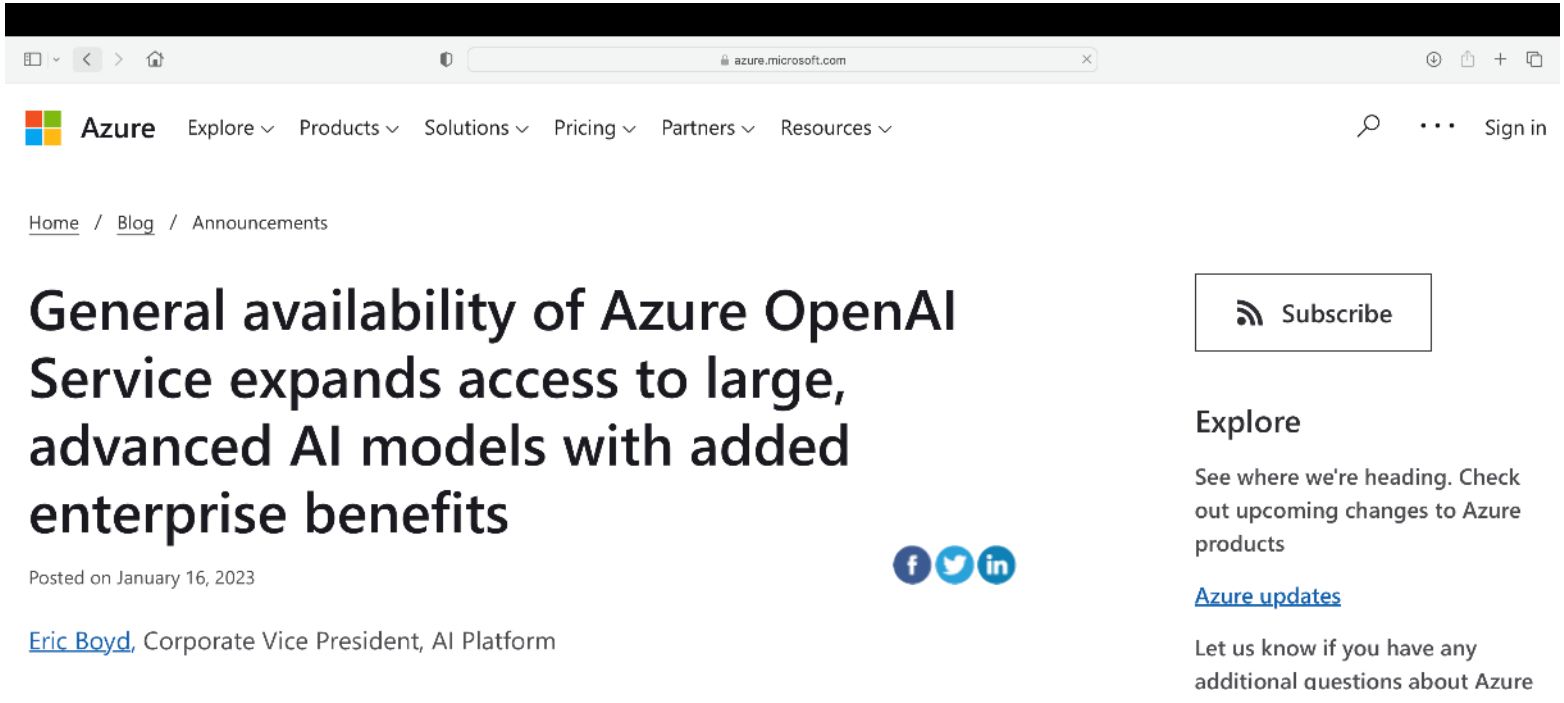
Supercomputer

(Housed in an Azure Data Center)

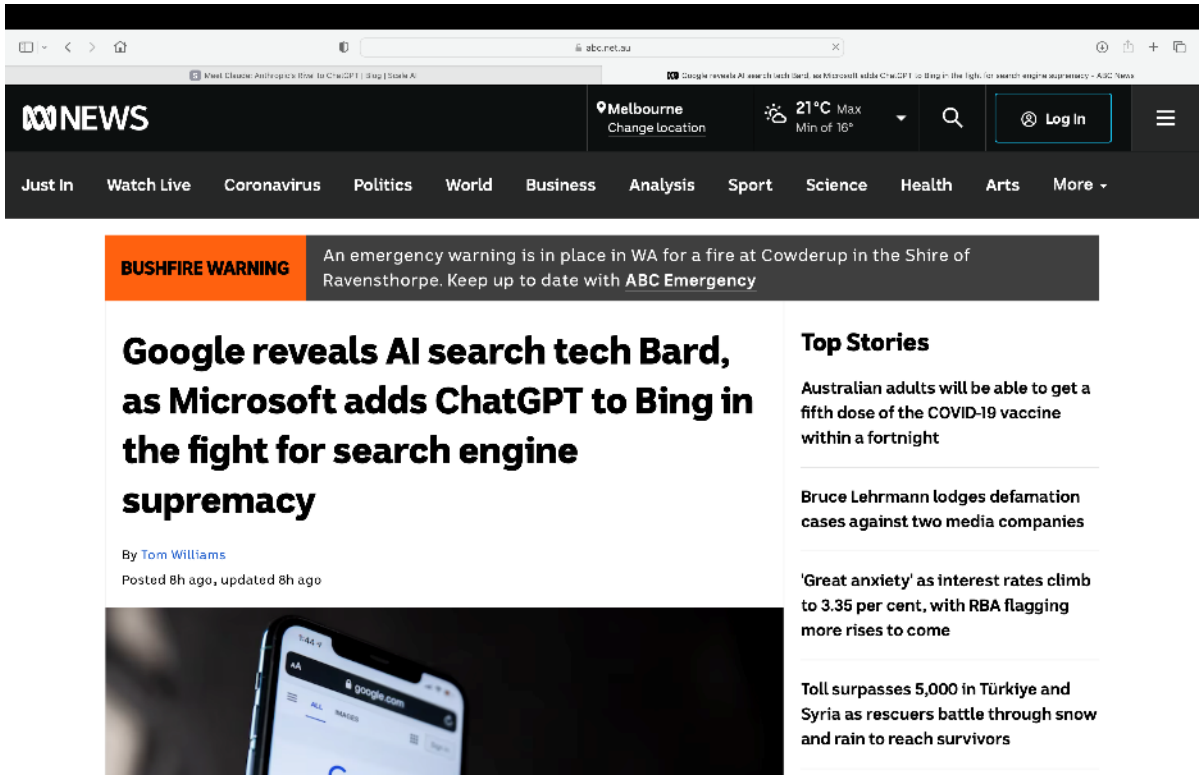
Current estimated IQ of **ChatGPT**

147

Microsoft is Deeply Invested in OpenAI (currently $13B)



Google is responding with “Bard”



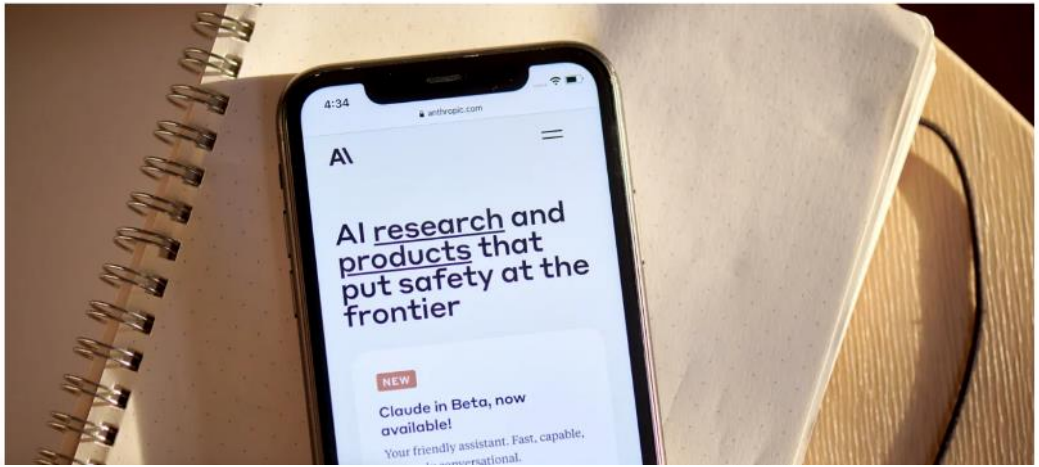
AWS is Responding with Anthropic Claude

CNN BUSINESS Markets Tech Media Calculators Videos

Amazon invests up to $4 billion in Anthropic AI in exchange for minority stake and further AWS integration

By Brian Fung, CNN

2 minute read · Updated 9:55 AM EDT, Mon September 25, 2023



Fun fact:

Anthropic Claude is named after American inventor **Claude Shannon**, pioneer of information theory, inventor of the microprocessor, binary communication, and the ancestor of IBM Deep Blue

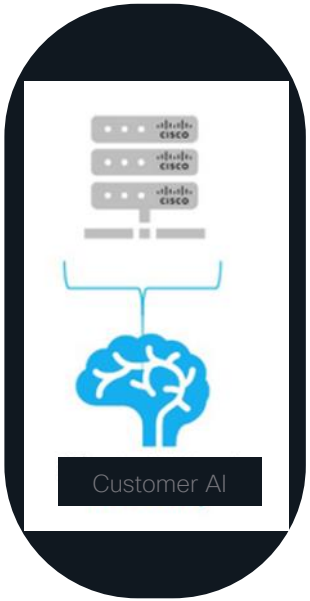
Overview –

Where is Cisco Using AI/ML in Products Today?

Artificial Intelligence & Cisco

Using AI to improve Cisco products & services

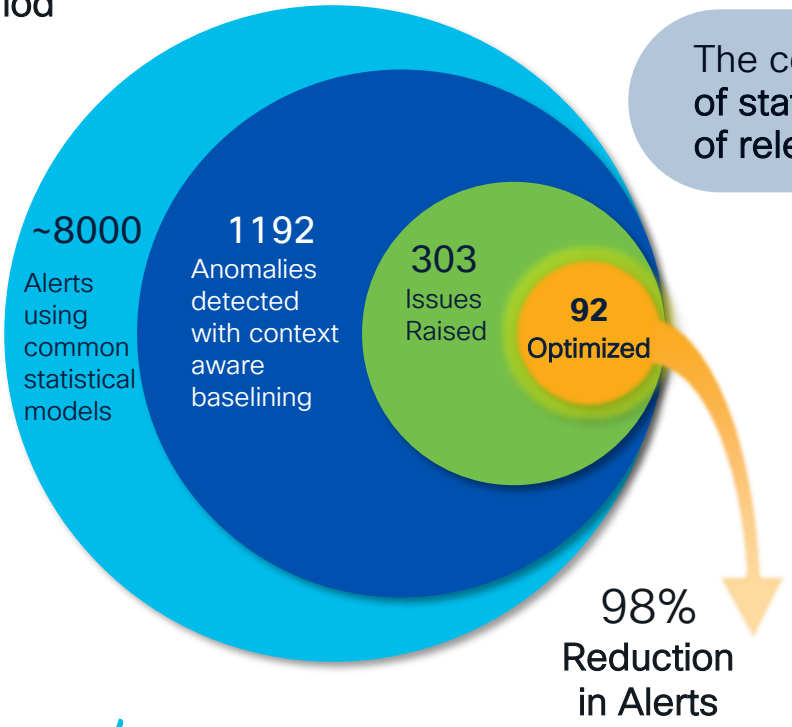
Providing Cisco products to enable AI



Let’s Start with AI in *Enterprise Networks*

Reducing Events to only Relevant Anomalies

Issues generated for 11 customers over 3-month period

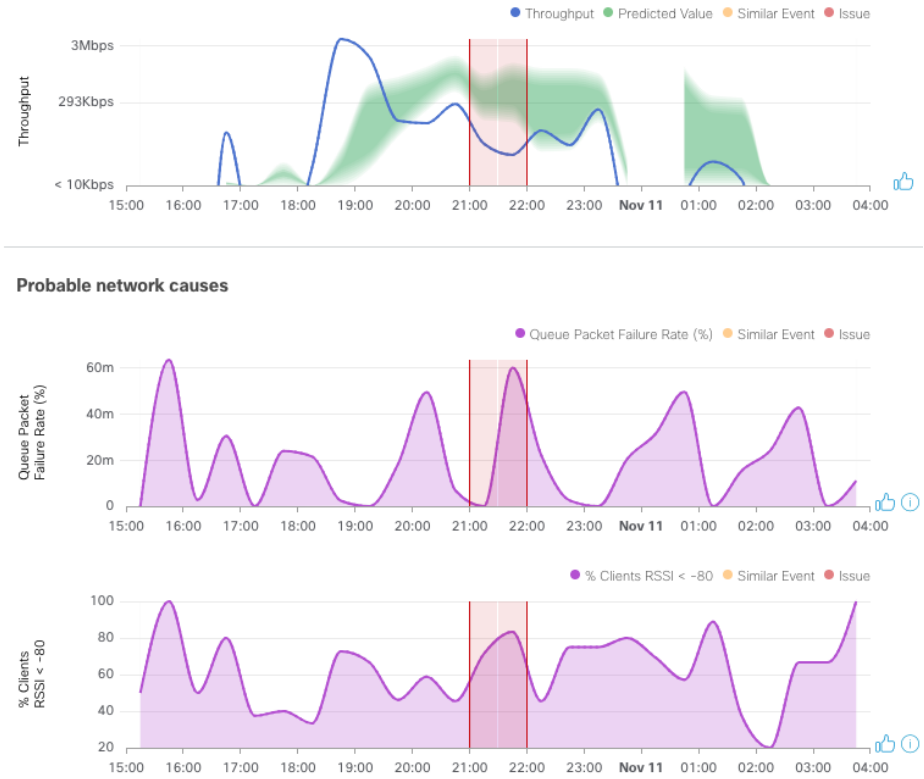


The core challenge is to turn a potentially **LARGE number of statistical / model anomalies** into a **SMALL number of relevant anomalies** for the user

- **ML Models:** model type and architecture, parameter optimization (e.g. sensitivity)
- **Select anomalies more likely relevant** (existence of root cause, impact measurability, transient/persistent, …)
- Potentially **reinforcement learning** (adapting type of anomaly liked by the user)
- **Issue generation:** aggregation heuristics, …

AI-Powered Wireless Analytics

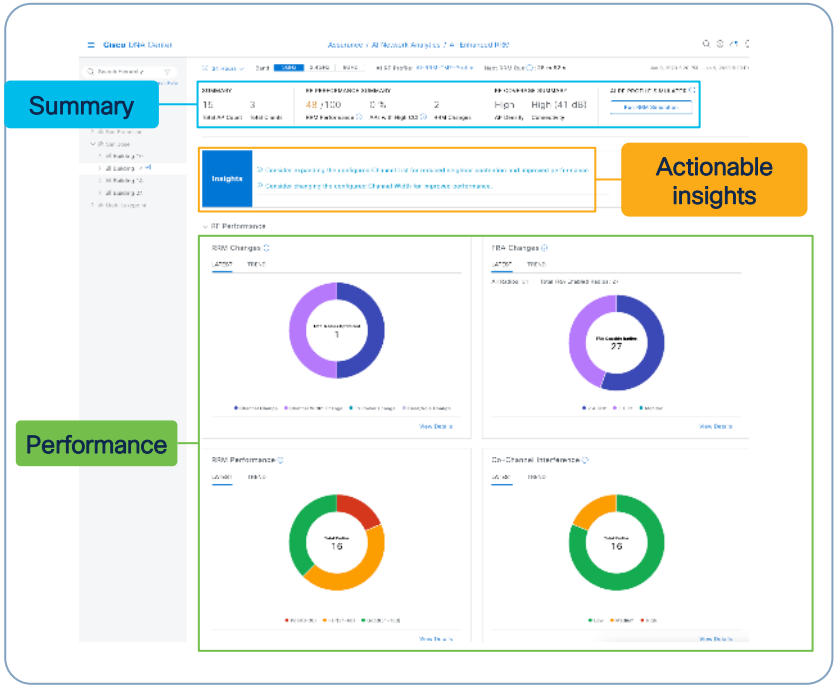
AI-Driven Anomaly Detection
Find + Root Cause Complex Issues



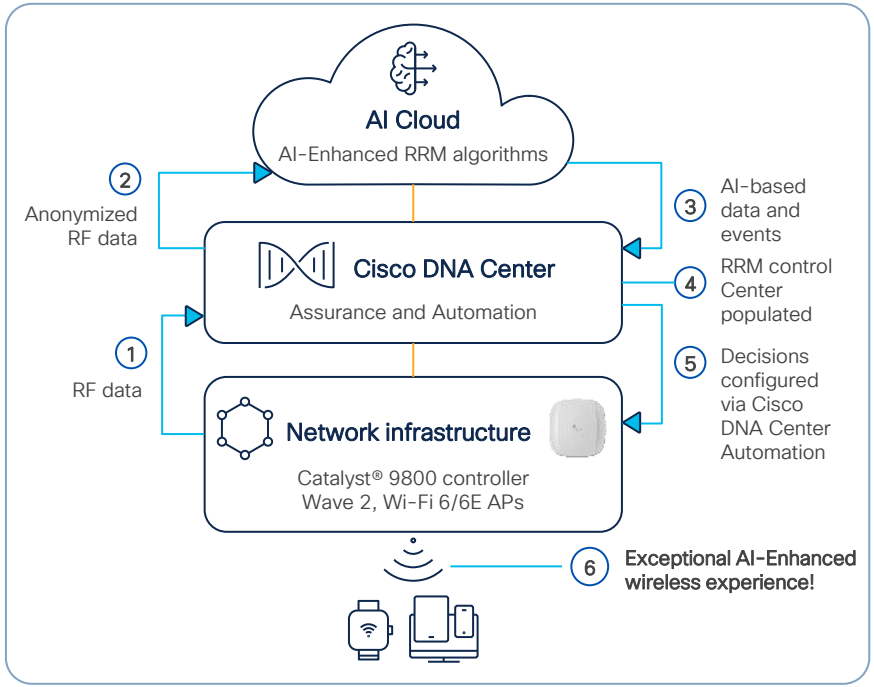
AI-Enhanced RRM

Catalyst's AI-Driven RRM solution

Deep RF visibility & advanced control

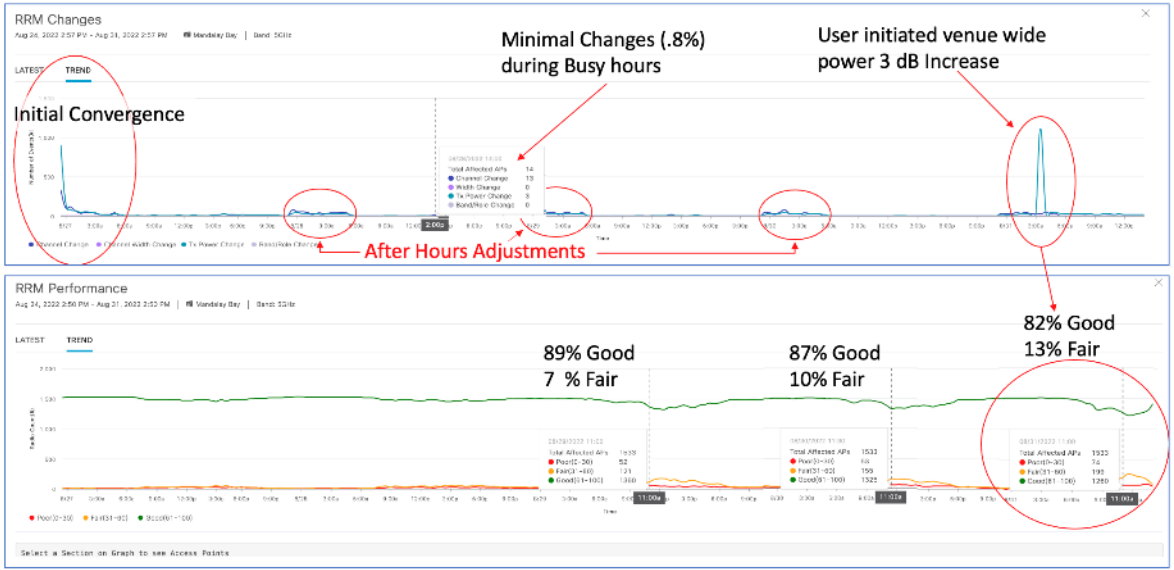


Proactive optimizations for all deployment sizes



AI-Enhanced RRM in Action
Cisco Impact 2022

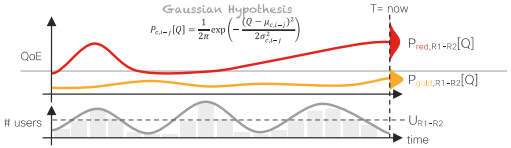
- Initial Convergence ~3 Hours
- Changes made at Night
- Health stayed above 85% (very good with load)
- Manual Changes made last day, easy to spot the decrease in efficiency



Predicting the Internet

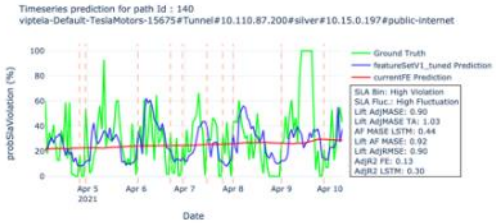
The notion of predicting Application failures implies that the engine predicts ***before*** it happens, in contrast with reactive approach that tries to minimizes the duration of the failure, but it is too late ☹

Our system is using various learning strategies:

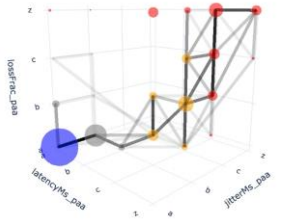


Statistical Model

Dynamic Model



LSTM



State Transition Learning

Telemetry Details – SD-WAN

System Details

Device Information

- Hostname
- Device Model
- System IP
- Latitude
- Longitude
- Reachability status
- Site-Id
- Org-Id

TLOC/Wan Interface

- Color
- Carrier
- Admin State
- Interface
- Preference
- Weight
- Device Name
- Private and Public IP address

Network Performance

PFR Statistics

[BDF Tunnel Probing Information]

- Latency
- Loss
- Jitter
- Hostname
- DeviceID
- Time of Day
- Local/Remote Color
- Local/Remote System IP
- Tunnel Rx/Tx Octets

CXP Statistics

[SaaS HTTP Probing Information]

- Loss
- Latency
- Application
- VPN ID
- DeviceID
- Interface
- Gateway System IP
- Best Path
- Local/Remote Color
- MSFT Quality Labels

Network Usage

DPI Statistics

- Application/Family
- Local/Remote IP Address
- Local/Remote Port
- Ingress Local/Remote Color
- Egress Local/Remote Color
- Local/Remote System IP
- Egress Interface
- Time of Day
- VPN ID
- Packets
- Octets

Interface Statistics

- VPN ID
- Rx/Tx Packets
- Rx/Tx Octets
- Rx/Tx Errors
- Rx/Tx Drops
- Rx/Tx PPS
- Platform Type
- Operational Status

Notes:

- For a medium customer we expect to process around: 1GB of PFR and 30GB of DPI of data per day.
- Telemetry may arrive in slightly different formats based on software version, source platform (cEdge, vEdge) and is reconciled during ingestion.

SLA Violations Across the World

and how much Predictive Networks can help

0
CUSTOMERS

0
COUNTRIES

2
CITIES

Cisco
redictive
etworks

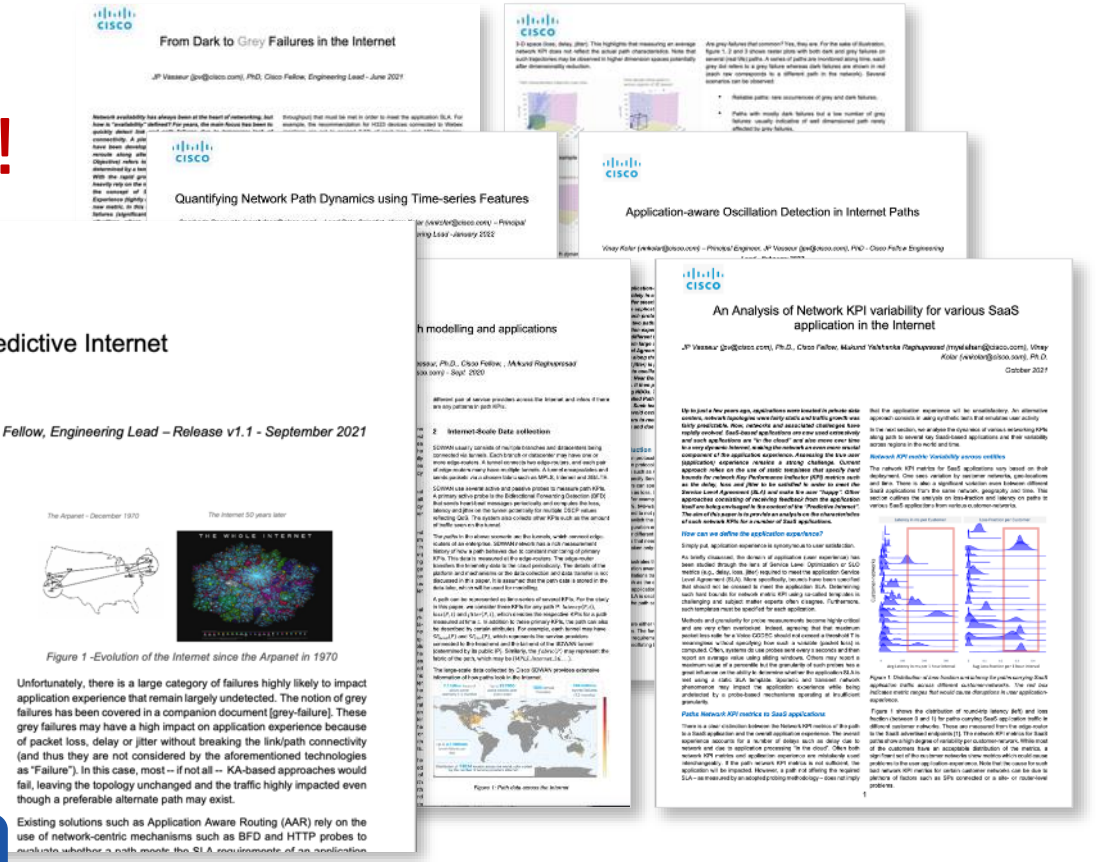
Want to Know More?

BRKNWT-2210 from the Cisco Live Demand Library

"Predictive Networks are THERE !!"

Review JP's session!

JP Vasseur
Cisco Fellow



AI in *Collaboration / Webex*

Launching the all new Webex AI Codec

The AI Codec isolates a person's voice

Allows Webex to accurately extract only the voice from all other sounds, similar to how "Now and Then" was produced

The better NLP works, the better the AI assistant can do its job:

- Meeting summary
- Language translation
- Integration of RMM + LMM features

Introducing Real-time Media Models (RMM)

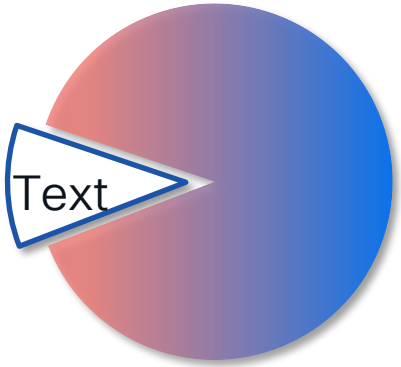
LLM

Email

Messages

Transcripts

Documents



Audio

Video

Reactions

Gestures

Voice tones

Inflections

RMM

Lots of activity today

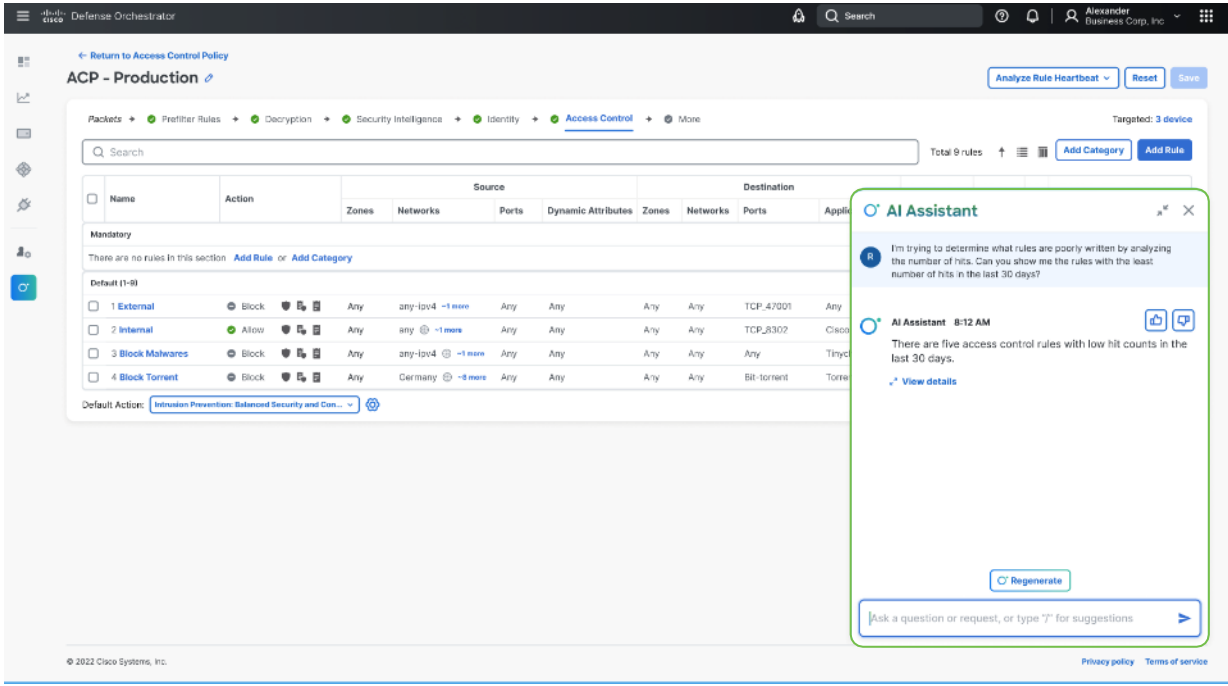
Lots of room for innovation. Think about why this is a meaningful capability.

AI in *Security Operations*

Cisco Live Melbourne (Dec 2023)
AI Policy Assistant

Policy Assistant

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



Encrypted Visibility Engine (EVE)

Enhanced Visibility and Detection Efficacy of Encrypted Traffic

Inferenced Based Identification without Decryption in TLS & QUIC of:

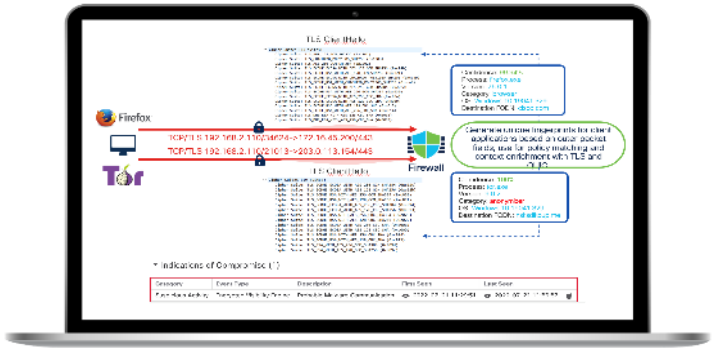
- Client Applications
- Operating Systems
- Compromised Hosts

Machine Learning Generated Fingerprints with data from:

- Computer Security Incident Response Team (CSIRT)
- Network Visibility Module (NVM)
- Secure Malware Analytics (ThreatGrid)

AI @ Speed of the Network:

- Network Protocol Fingerprinting (NPF) selects classifier
- Weighted Naïve Bayes classifier with sparse updates
- Best Threat Detection efficacy in recent internal testing



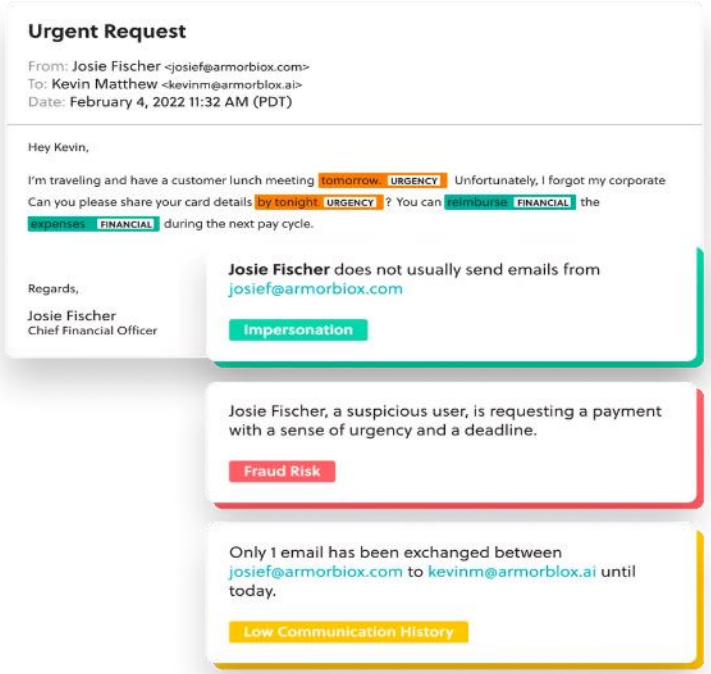
Using LLMs to Up-Level Email Security

ETD detects the most subtle email attacks through **Large Language Model interpretation**

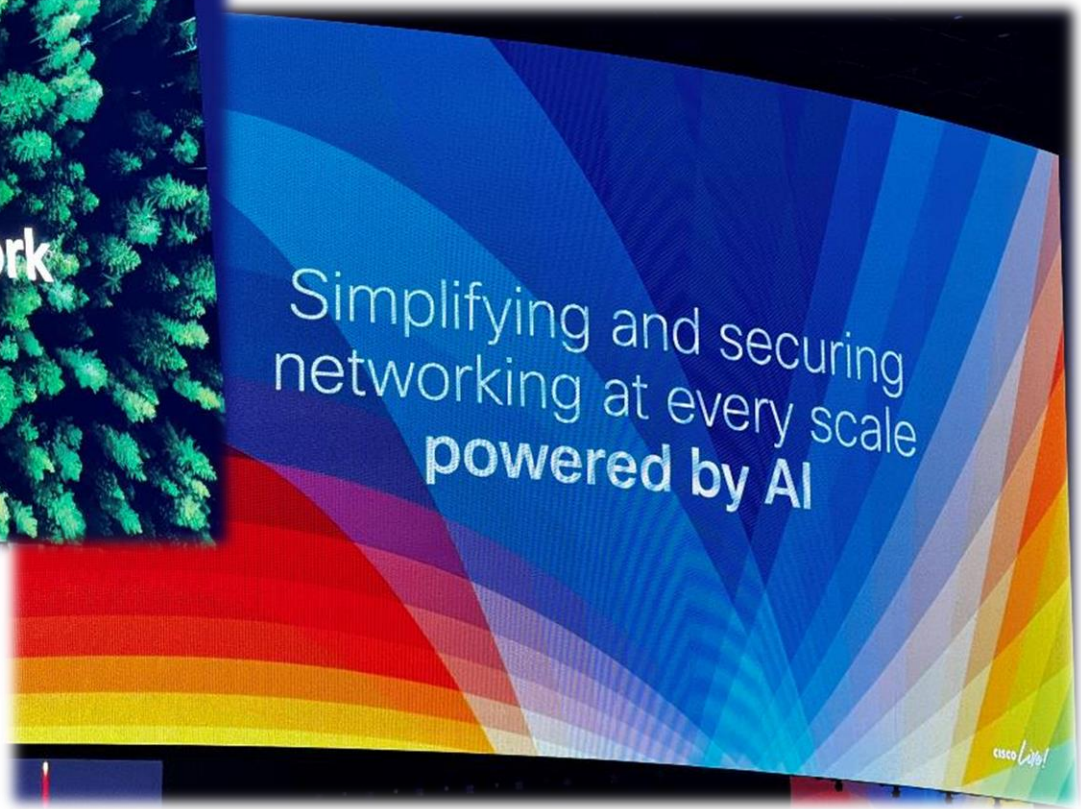
Phishing attacks are getting incredibly good. Examining the links in a message offers only partial protection

Language Analysis of Emails

Language is an exceptional method to help understanding patterns to enforce security (phishing, data exfiltration, etc.)



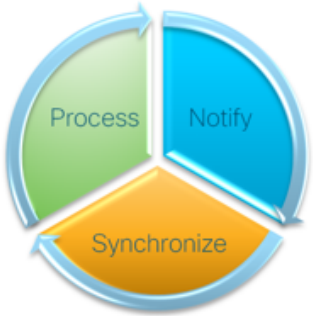
Building Networks for ML/AI Workloads



Building Networks for ML/AI Workloads

Execute Instructions on GPU

High bandwidth compute can saturate network links



Send results of computation

Several methods, we'll focus on just one
All-to-All Collective (Everyone sends to everyone)

Wait for everyone to complete

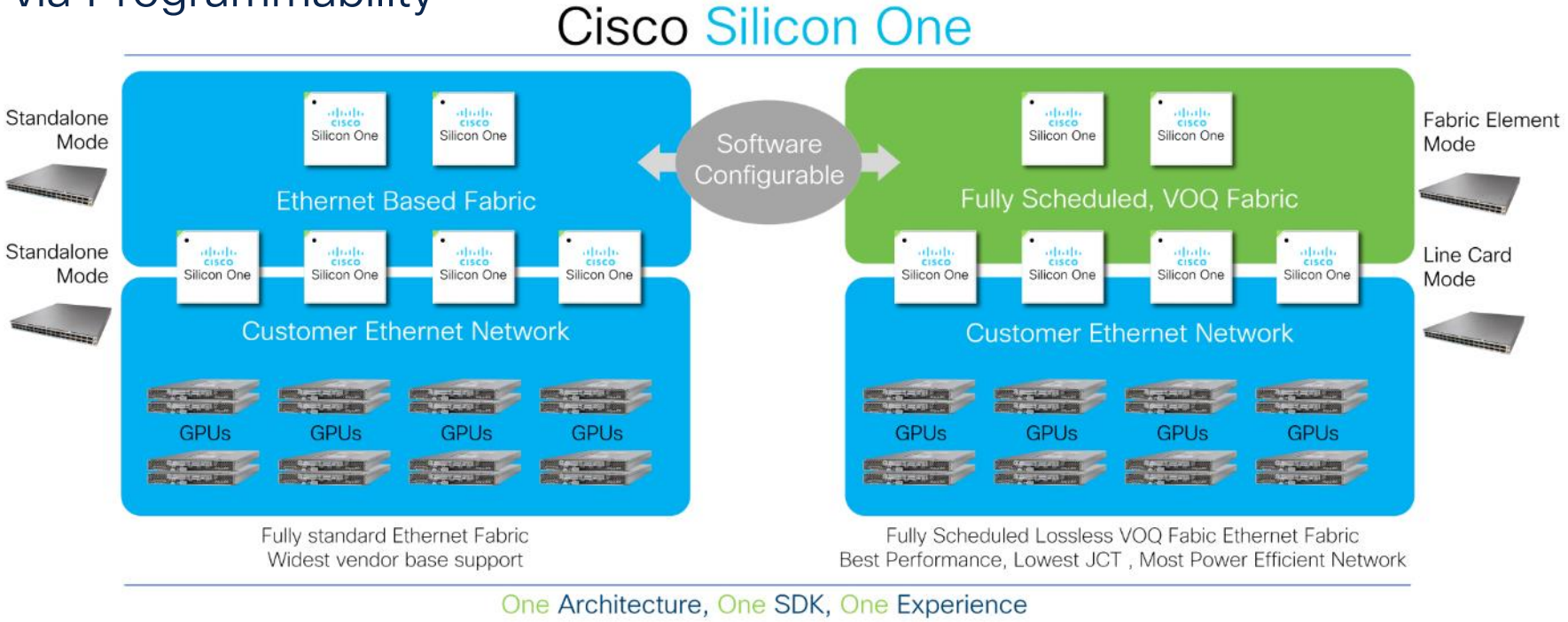
Creates synchronization between GPUs
Computation stalls waiting for the slowest path
Job Completion Time (JCT) is based on the *worst-case* tail latency

Cisco Silicon One

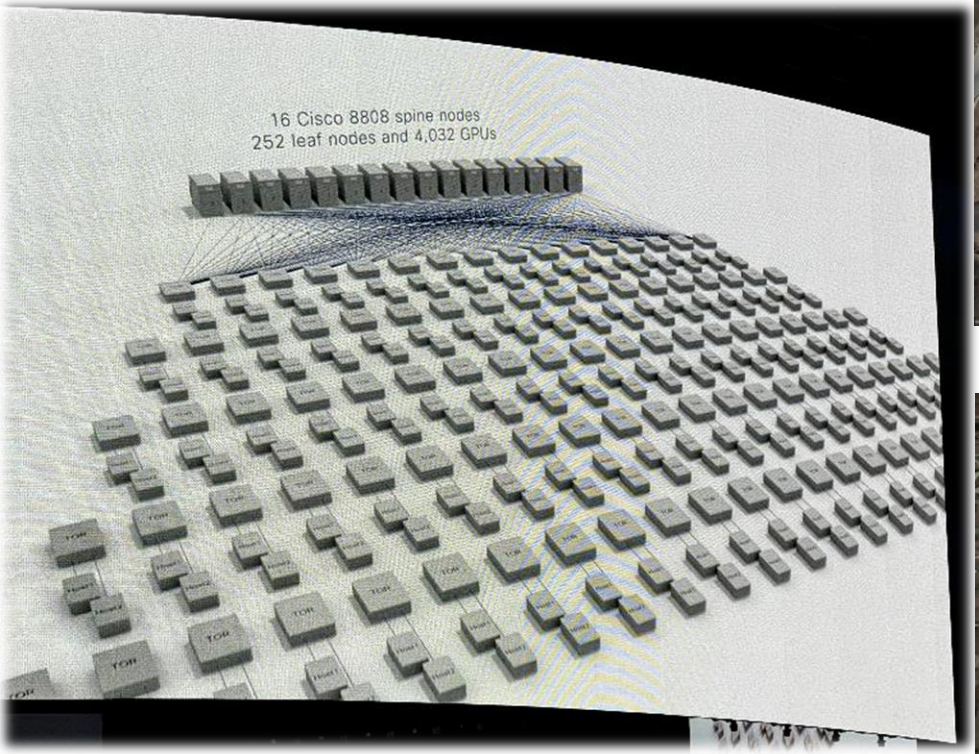
Want to more more about **Cisco Silicon**?

BRKARC-2093 via Cisco Live On-Demand Library!

Flexibility via Programmability



GenAI Data Center Networks Massive Growth

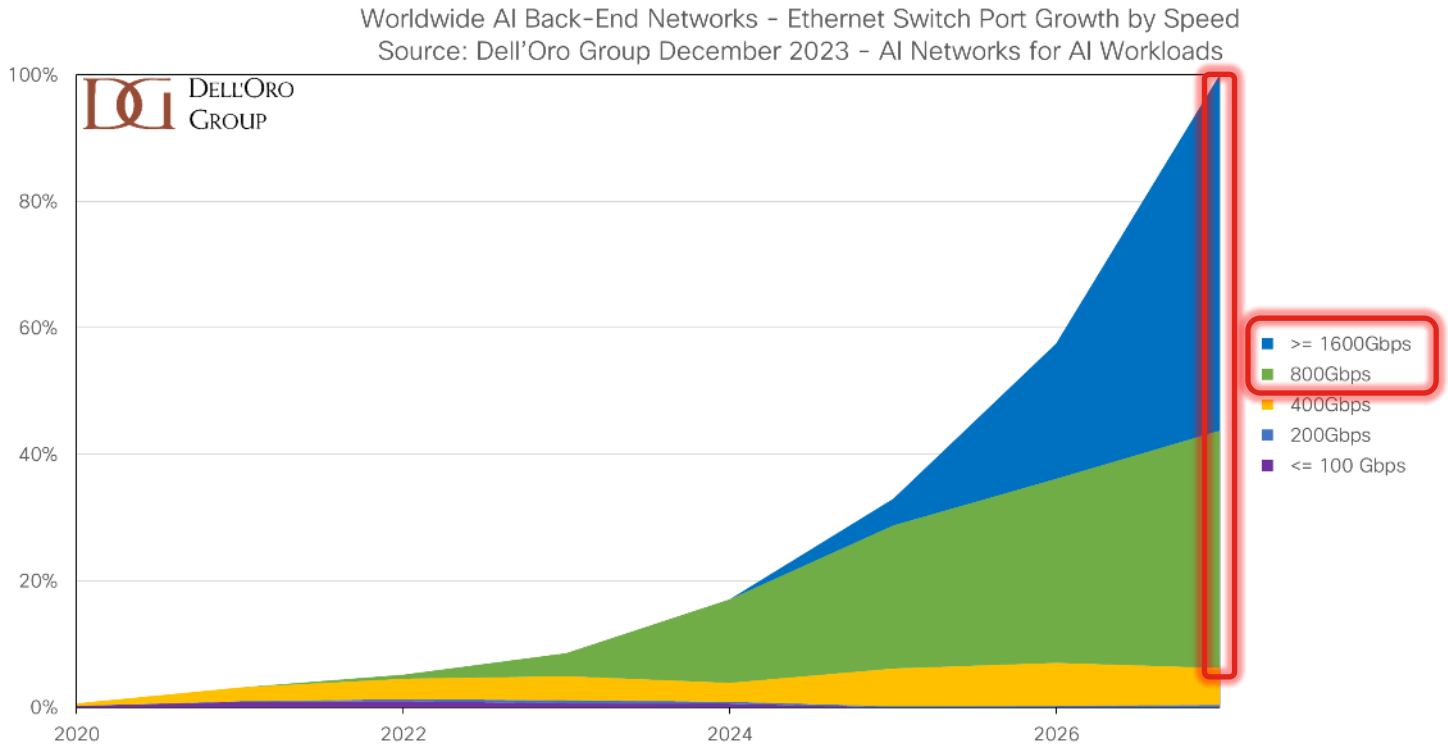


Single Building (360,000 ft²)

DC Location with 5 buildings (6,095,616 ft²)

AI Network Market Trends

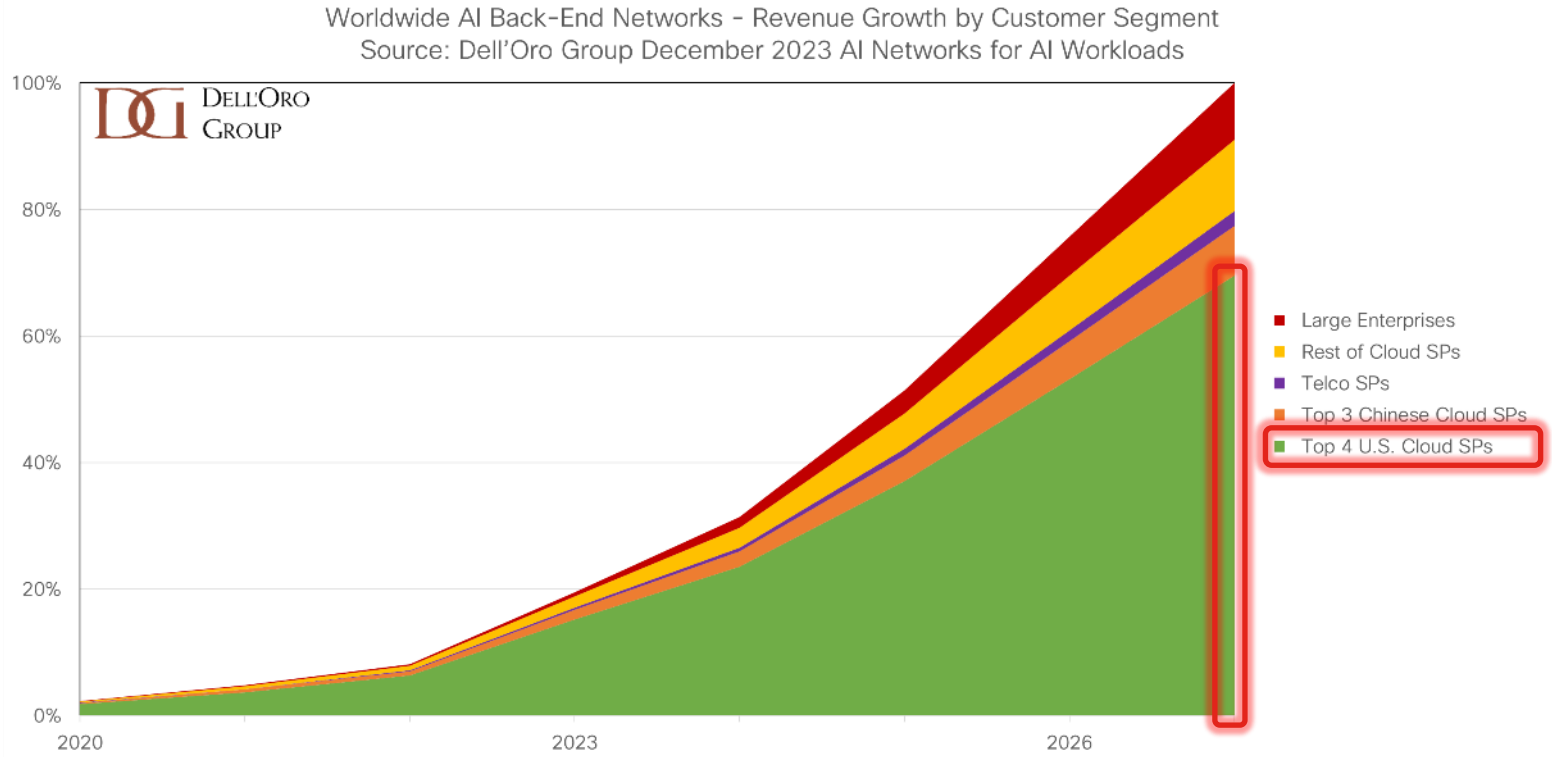
AI Back-End Networks - Ethernet Switch Port Growth



NOTE: The graph above shows percentage of 2028 total ports.

AI Network Market Trends

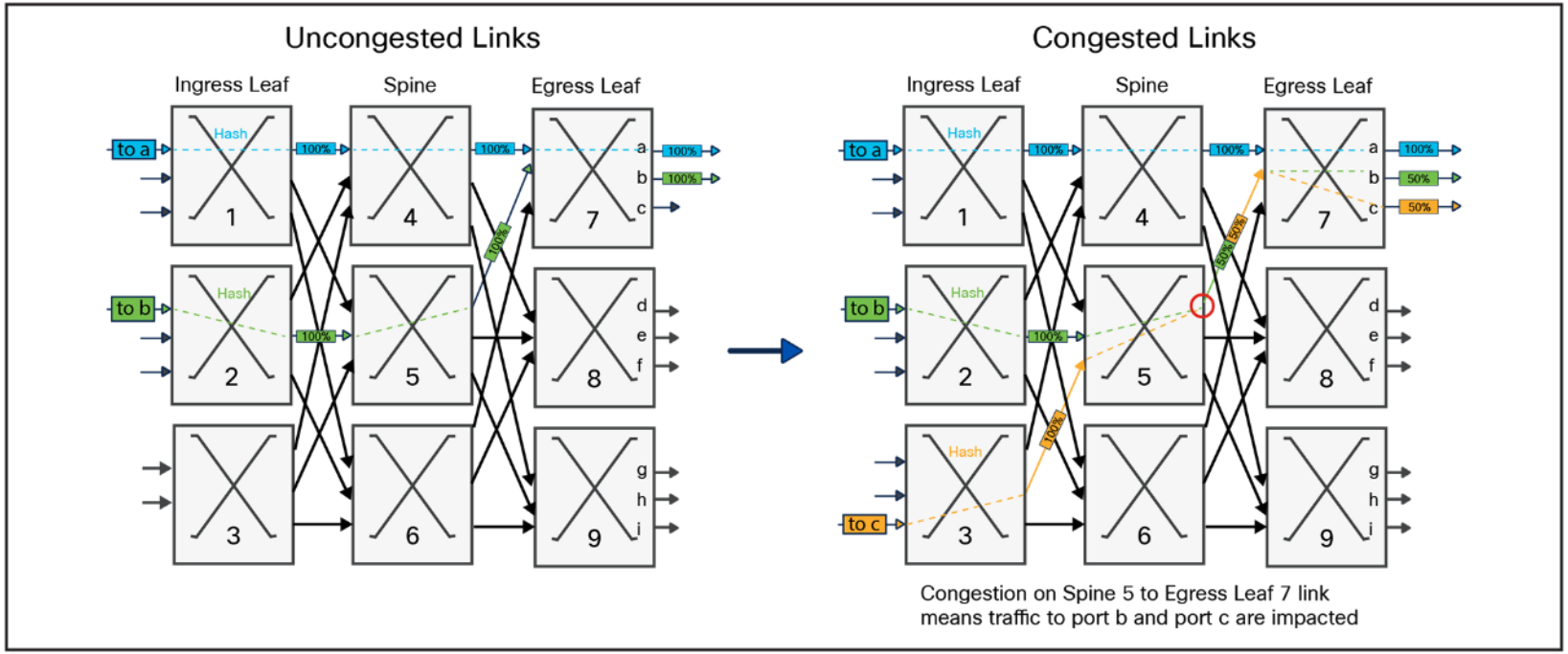
AI Back-End Networks - Revenue Growth by Customer Segment



NOTE: The graph above shows percentage of 2028 total ports.

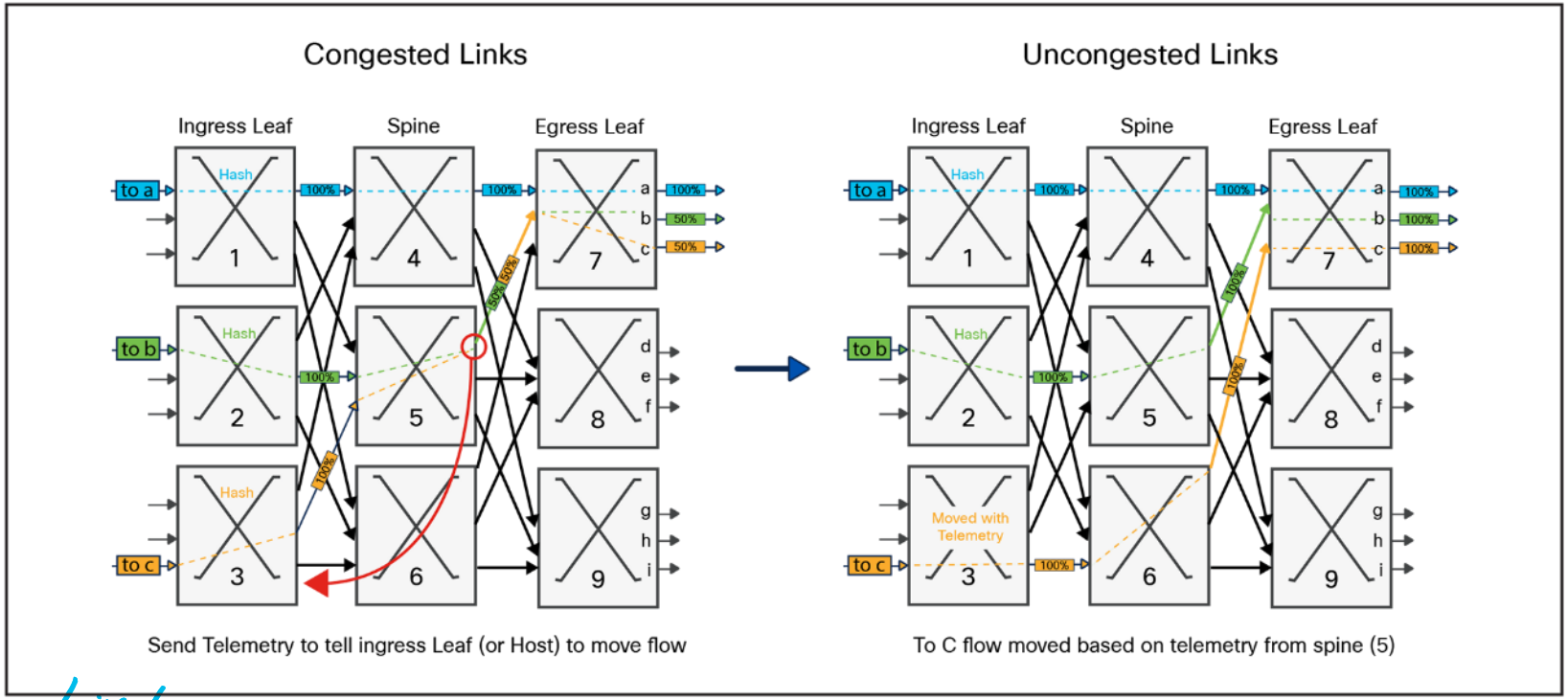
Building Networks for ML/AI Workloads

Operation with ECMP



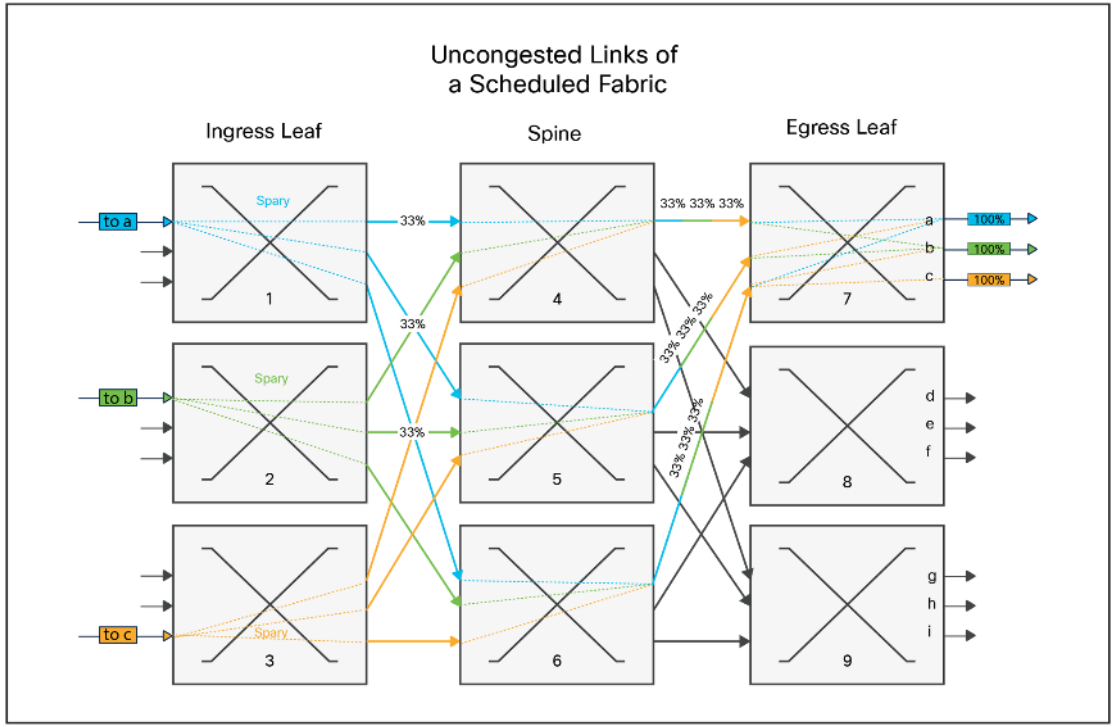
Building Networks for ML/AI Workloads

Operation with Telemetry Assisted Ethernet



Building Networks for ML/AI Workloads

Operation with Fully Scheduled Fabric



Building Networks for ML/AI Workloads – Outcomes

MORE INFO AT …

https://blogs.cisco.com/sp/building-ai-ml-networks-with-cisco-silicon-one and "Evolved Networking: The AI/ML Challenge" OCP Global Summit October 2022 Slides, Video

Fully scheduled fabric provides exceptional performance, providing lower job completion time

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:** GPT-4 was trained in Sept 2021 (two years out of data). Requires Retrieval Augmented Generation (RAG)
4. **Fine Tuning:** Cloud-based LLMs are “foundational models” and lack fine tuning for certain use cases

Foundational vs. Fine Tuned Models

Confidential Documents

Private Images

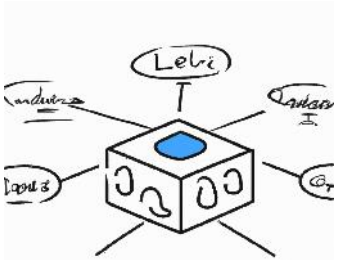
PII data

Financial data

Etc.

Training

Foundational Model



Fine Tuning

Industry Specializations

Healthcare

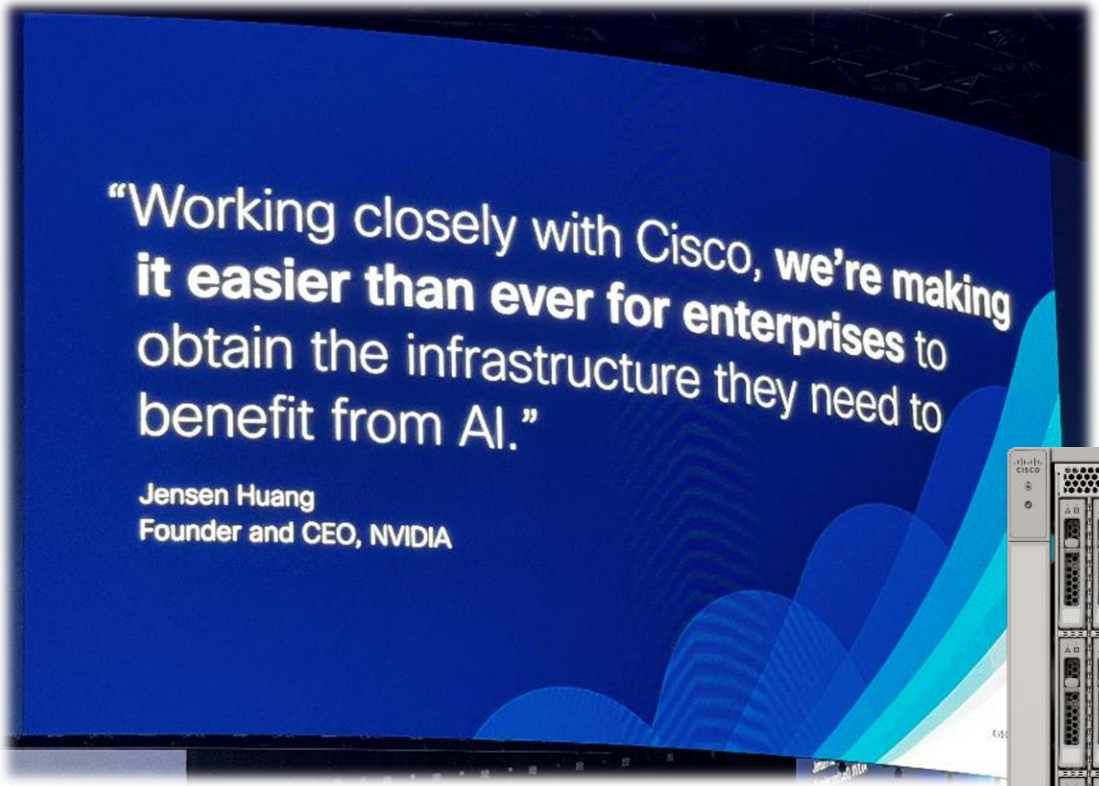
Legal Analysis

Writing Code

Power Utilities

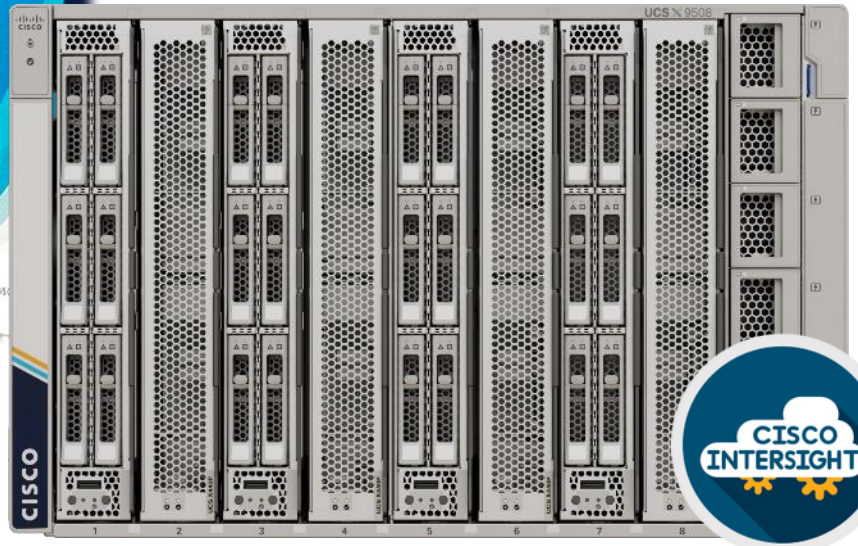
Fintech

Foundational models can be fine-tuned for specific tasks



Only chassis-based server system with expandable GPU modules

UCS-X Series + ACI



Intersight provides single-click deployment of full-stack on-prem LLM

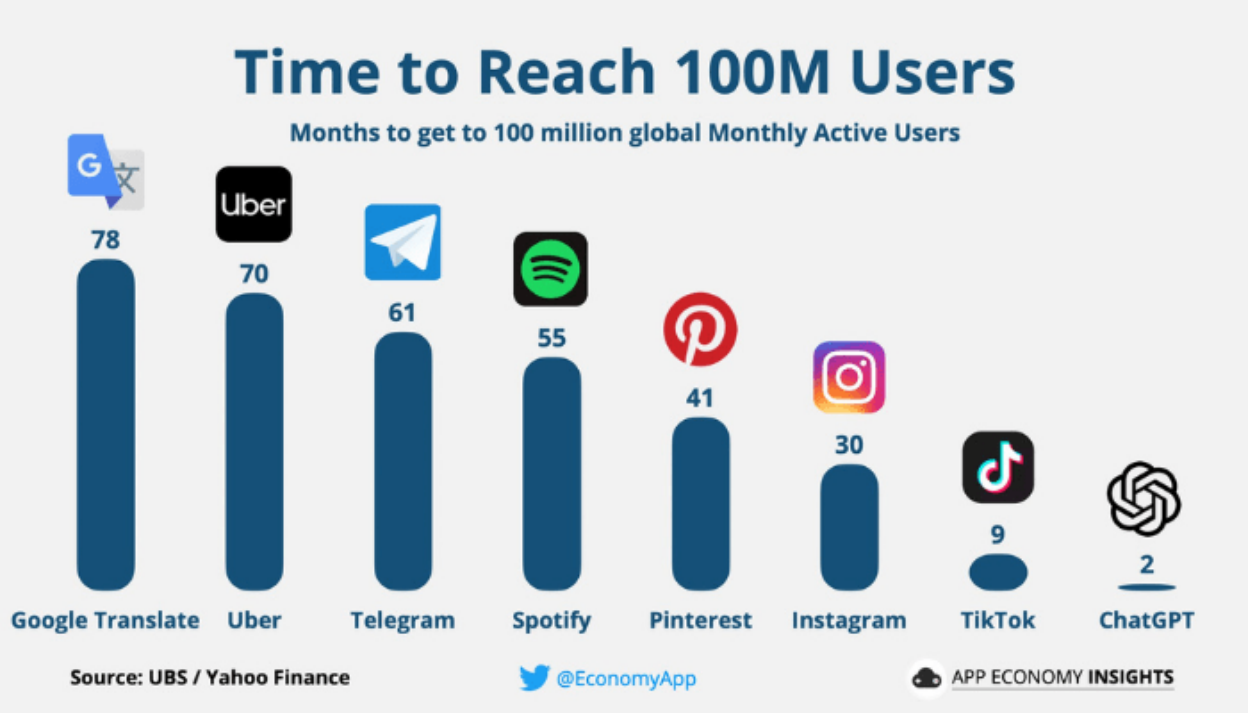
Large Language Models (LLMs), GPTs, and more ... and their use in Networking

John

An Inflection Point

openai democratizes artificial intelligence

An Inflection Point



Forbes: On-Prem AI

[Generative AI Breaks The Data Center: Data Center Infrastructure And Operating Costs Projected To Increase To Over $76 Billion By 2028 (forbes.com)]


FORBES > INNOVATION > AI

Generative AI Breaks The Data Center: Data Center Infrastructure And Operating Costs Projected To Increase To Over $76 Billion By 2028

Jim McGregor Contributor
Tirias Research Contributor Group

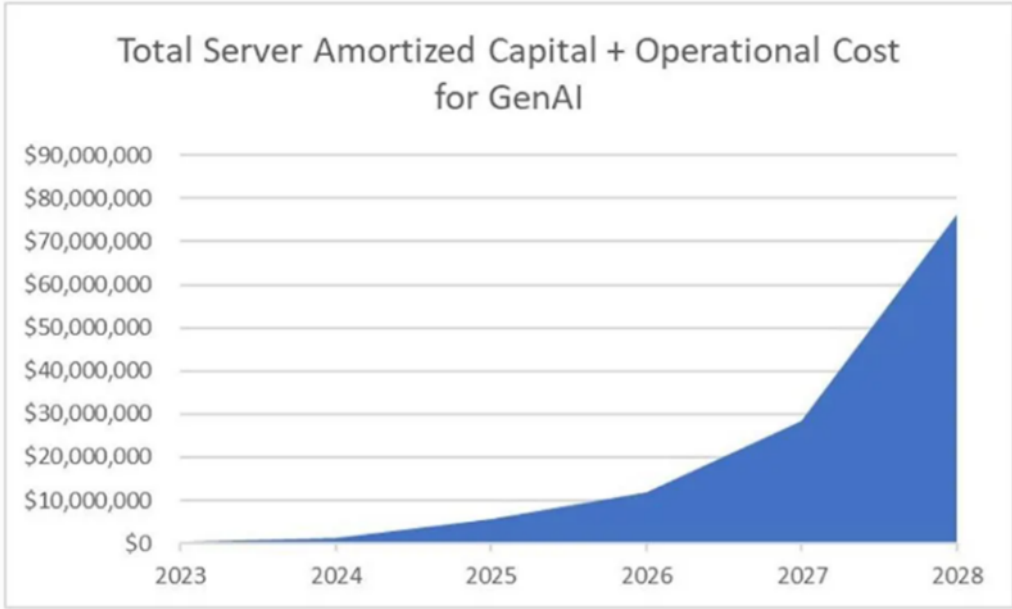
Follow

2

May 12, 2023, 04:33pm EDT

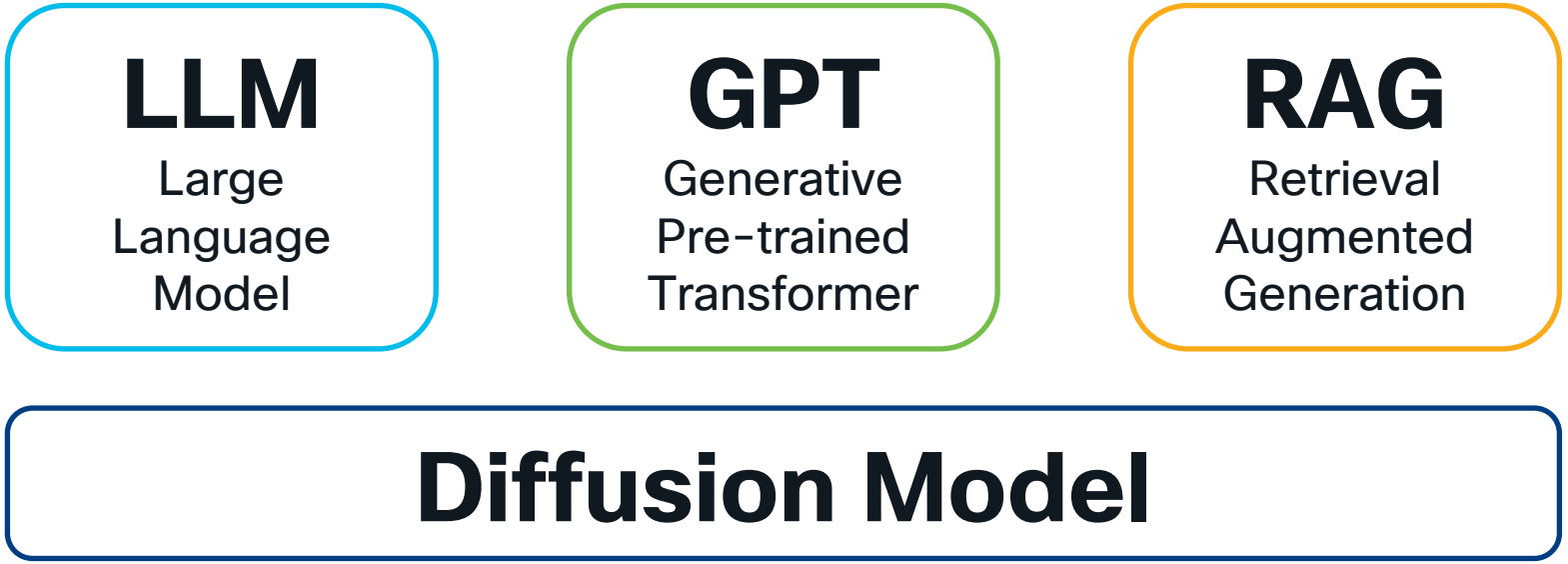
Data Center TCO for GenAI through 2028

Tirias Research



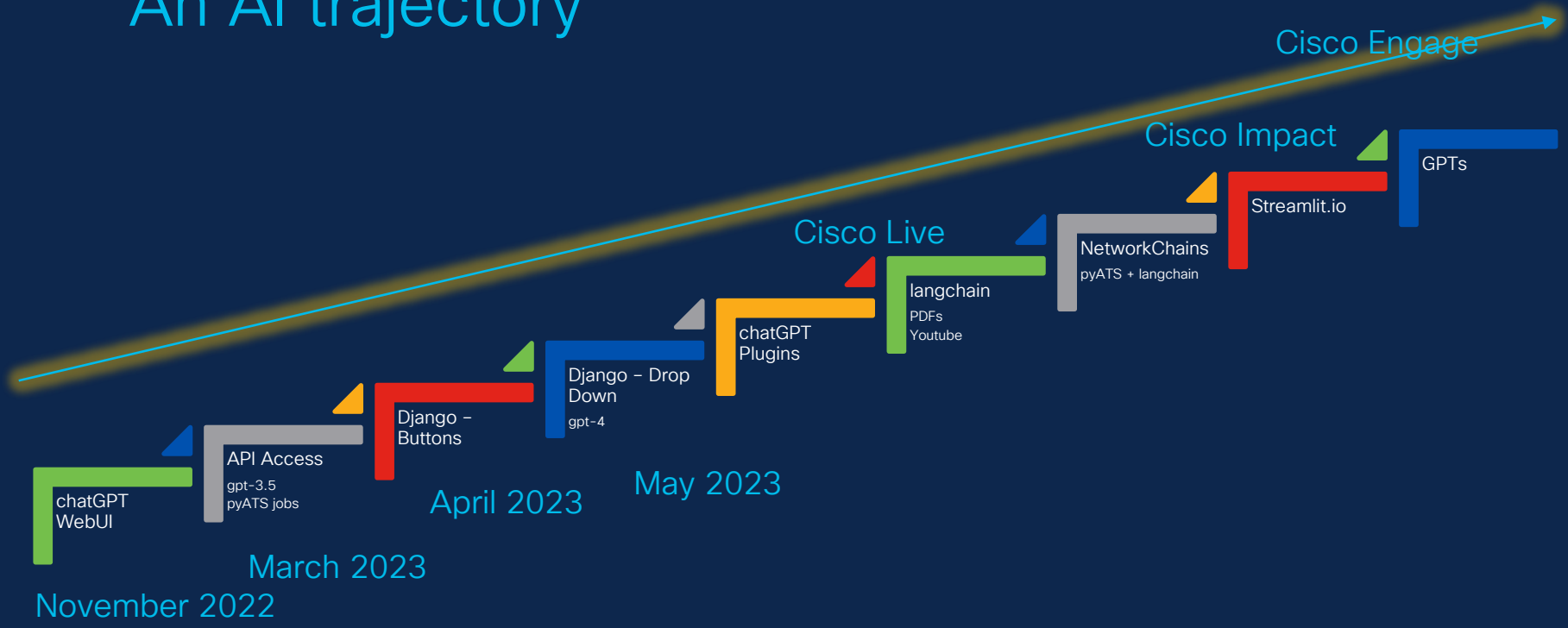
Data Center Total Cost of Ownership (TCO) for Generational AI through 2028. TIRIAS RESEARCH

Some Terminology ... and yes, new TLAs...



My AI journey
November 2022 - Present

An AI trajectory



chatGPT plugins

Two Files Hosted Publicly in .well-known

ai-plugin.json

- Small JSON file
- Describes the plugin to both the humans and the LLM using the plugin
- Legal text, e-mail, logo image
- Points to openAPI.yaml file
- Could require authentication

openapi.yaml

- Mapping of APIs
- Industry standard
- Swagger
- CRUD functions
- Status codes

ai-plugin.json

```
{
    "schema_version": "v1",
    "name_for_human": "networkGPT Plugin",
    "name_for_model": "networkGPT",
    "description_for_human": "Plugin to Cisco network JSON data and answers from networkGPT",
    "description_for_model": "Plugin to Cisco network JSON data. The data will include Cisco CLI output. Please provide an
    "auth": {
        "type": "none"
    },
    "api": {
        "type": "openapi",
        "url": "http://localhost:8000/.well-known/openapi.yaml",
        "is_user_authenticated": false
    },
    "logo_url": "http://localhost:8000/staticfiles/networkgpt.png",
    "contact_email": "ptcapo@gmail.com",
    "legal_info_url": "http://www.automateyournetwork.ca"
}
```

openapi.yaml

```yaml
---
openapi: "3.0.1"
info:
  title: "networkGPT Plugin API"
  version: "1.0.0"
servers:
  - url: http://localhost:8000
paths:
  /pyATS_Answer/:
    get: ···
  /Devices/:
    get: ···
    post: ···
    delete: ···
```

GPTs

GPTs

- Brand new as of November 07, 2023
- Further evolution from plugins
- openAPI.yaml file
- Actions – CRUD activities
- ChatGPT Builder – build you GPT via prompts
- Dall-E 3, Web browser, Data analytics
- Store became available in January 2023

How to build an openAI plugin or GPT

Getting Started

We need to add to, or create, a /.well-known folder on our server / platform chatGPT can communicate with

In this folder we create 2 files:

ai-plugin.json

openapi.yaml

These 'manifests' are validated at time of plugin installation when done locally

Getting Started

```
{
    "schema_version": "v1",
    "name_for_human": "networkGPT Plugin",
    "name_for_model": "networkGPT",
    "description_for_human": "Plugin to Cisco network JSON data and answers from networkGPT",
    "description_for_model": "Plugin to Cisco network JSON data. The data will include Cisco CLI output. Please provide an
    "auth": {
        "type": "none"
    },
    "api": {
        "type": "openapi",
        "url": "http://localhost:8000/.well-known/openapi.yaml",
        "is_user_authenticated": false
    },
    "logo_url": "http://localhost:8000/staticfiles/networkgpt.png",
    "contact_email": "ptcapo@gmail.com",
    "legal_info_url": "http://www.automateyournetwork.ca"
}
```

openai.yaml

For Your Reference

- Important fields:
  - Description for model:
    - "Seeds" a "prompt" so chatGPT can understand the purpose of the plugin
    - Automatically invoked via chat if you do not provide additional prompt information
      - For example – "Please run pyATS_Answer"
        - Would trigger the description for model prompt as no other real prompt was provided
      - Override it – "I would like to see if the switch01 has any new information in pyATS_Answer"
      - Would take your new input prompt

openai.yaml

For Your Reference

- Important fields:
  - Auth
    - Must be "type: none" for localhost development
    - All URLs must also be HTTP not HTTPS for local development
    - Could be OAuth2 or Bearer Tokens or Basic Auth

- Opportunity to map our API Auth requirements to allow the plugin to authenticate against our platforms

openapi.yaml

For Your Reference

```yaml
---
openapi: "3.0.1"
info:
  title: "networkGPT Plugin API"
  version: "1.0.0"
servers:
  - url: http://localhost:8000
paths:
  /pyATS_Answer/:
    get: ···
  /Devices/:
    get: ···
    post: ···
    delete: ···
```

- Paths equates to API Methods (GET,POST,PATCH,PUT,DELETE) and CRUD (Create, Read, Update, Delete) activities that can be performed

These paths could also trigger scripts that return JSON – such as a pyATS script

- Let's examine the various Methods a little more closely

openapi.yaml

For Your Reference

```yaml
paths:
  /pyATS_Answer/:
    get:
      operationId: "getLatestNetworkData"
      summary: "Get the latest network data"
      description: "Retrieve the latest network data from the API."
      responses:
        '200':
          description: "Successful response"
          content:
            application/json:
              schema:
                type: "object"
                properties:
                  networkData:
                    type: "object"
                    description: "The latest network data"
```

- GET
- Extremely easy to implement – no field mapping required
- We could quite easily simply expose all of our GET, which is "safe" and "Read-Only", to allow JSON analysis via cha prompts and restrict access to POST / PUT / PATCH / DELETE activities which can introduce RISK and CHANGE MANAGEMENT

openapi.yaml

For Your Reference

```yaml
summary: "Create the device"
description: "Create the device from the API."
requestBody:
  content:
    application/json:
      schema:
        type: "object"
        properties:
          hostname:
            type: "string"
          alias:
            type: "string"
          device_type:
            type: "string"
          os:
            type: "string"
          platform:
            type: "string"
          username:
            type: "string"
          password:
            type: "string"
          protocol:
            type: "string"
```

- POST (PATCH / PUT)
- Creates / Updates records immutably
- Requires a **schema** that maps **properties** to each API field

openapi.yaml

For Your Reference

```yaml
responses:
  '201':
    description: "Successful response"
    content:
      application/json:
        schema:
          type: "object"
          properties:
            networkData:
              type: "object"
              description: "Created device"
```

- Response Codes
- We can add responses to all methods| as well and how the plugin should handl (what to display) in the event of a 200, 201, 400, 404, etc API response code

openapi.yaml

For Your Reference

```yaml
delete:
  operationId: "deleteDevice"
  summary: "Delete a device"
  description: "Delete a device from the API."
  responses:
    '204':
      description: "Successful response. The device has been deleted."
```

DELETE is deceivingly easy

Removes records via the API

Required some 'back-end', in my case Django, code to enable DELETE in the REST Framework

Obviously, extremely risky

Langchain

“What we can do today was impossible yesterday and will define tomorrow

John Capobianco – on Langchain

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)

Retrieval

- We will retrieve data from a vector store with relevant records from, for example, a semantic search (top result returned)

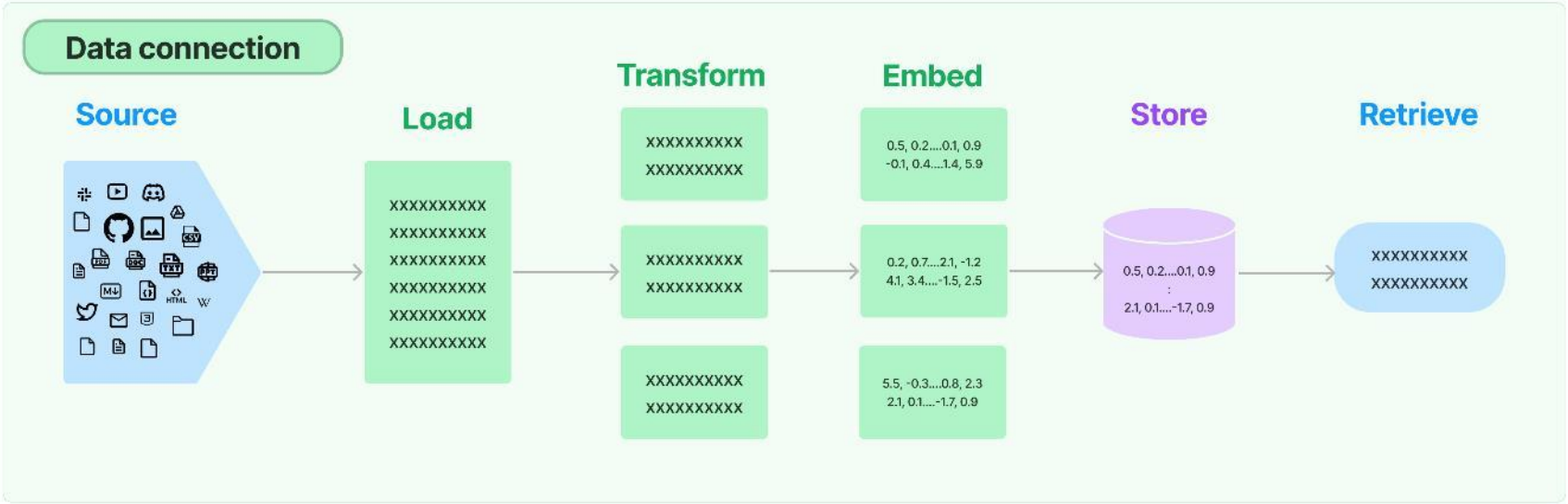
Augmented

- We will *augment* our **prompt** by supplementing the relevant information
- Often in the form of JSON

Generation

- Generate the response taking both the prompt and the contextual data into consideration

A high-level langchain abstraction



A high-level langchain abstraction

Source

Unlimited potential

PDF
CSV
HTML
JSON
Text
YouTube
Folders
GitHub
Twitter

Load

Pick the loader

Find the loader for the source

Possible parameters (such as YouTube video URL or PDF file location)

Transform

Break it up

First, the loader with **load and split** the document into pages

Then, using a text splitter, the pages get turned into **chunks** with **overlap**

A high-level langchain abstraction

Embed

The real power behind AI

When you chat with an AI it has no idea what the text input ‘says’

It is transformed into **embeddings; list of floating-point numbers**

Store

Vector Database

The **chunks** and **embeddings** are stored in a **vector store**

Pinecone.io
ChromaDB
Elastic
Others

Retrieve

Sematic search

We now retrieve the relevant information using **k values** (the number of records to include in the retrieval process)

Send the vectors and chunks to an **LLM** for chatting

```python
#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,
)
```

```python
# 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)
```

```python
# perform the Conversational Retreival Chain
pdf_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)
```

Private. Local Embeddings.

- “Cuda” provides access to Nvidia GPUs to perform embeddings locally
- **Instructor-XL from HuggingFace is a good example of a local embeddings model**

Private, Local, LLM Specifications

- Cisco UCS X-Series system
  - X210c w/ X440p PCIe node
  - Single A100-80GB NVIDIA GPU
- **Patrick leMaistre** provides the following GitHub repository for more details about the LLM and how to host it
  - pl247/ai-install (github.com)

What can you tell me about them?
This is from Conversational Retrieval Chain; Chroma
Sure, here is the information about your 10 interfaces:

1. Interface GigabitEthernet1: This is a Virtual CSR vNIC interface with a bandwidth of 1000000 and a port speed of 1000mbps. It is currently enabled and operational. The interface has a full duplex mode and its MAC address is 0050.56bf.9379. The interface has an IP address of 10.10.20.48/24.

2. Interface Loopback1110: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.90.1/24.

3. Interface GigabitEthernet2: This is a Virtual CSR vNIC interface with a bandwidth of 1000000 and a port speed of 1000mbps. It is currently enabled and operational. The interface has a full duplex mode and its MAC address is 0050.56bf.ea76. The interface has an IP address of 10.255.255.1/24.

4. Interface GigabitEthernet3: This is a Virtual CSR vNIC interface with a bandwidth of 1000000 and a port speed of 1000mbps. It is currently enabled and operational. The interface has a full duplex mode and its MAC address is 0050.56bf.1651.

5. Interface Loopback109: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.10.122/24.

6. Interface Loopback102: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.123.1/24.

7. Interface Nve1: This is a NVE interface with a bandwidth of 1000000. It is currently enabled but not operational. The interface's MAC address is 001e.bd55.9200.

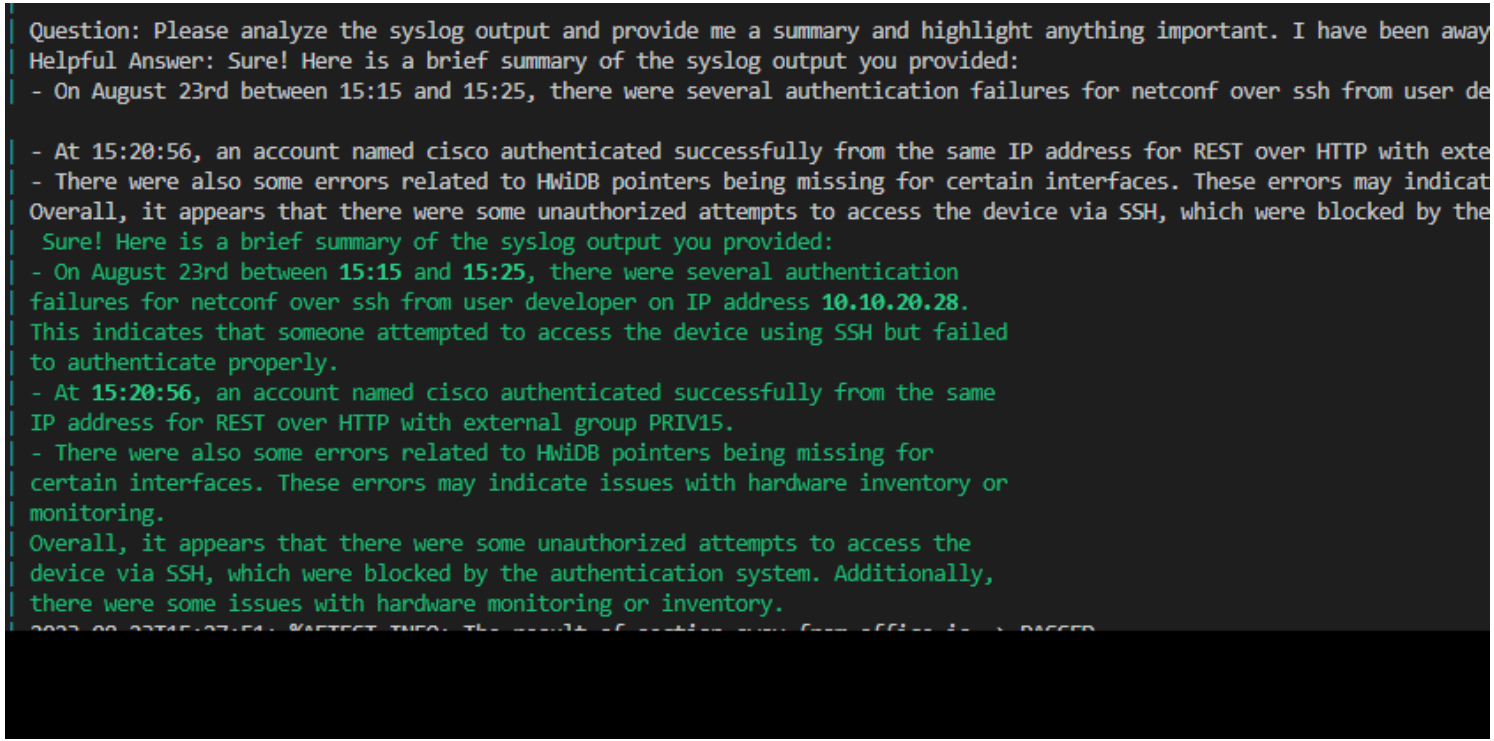
8. Interface Loopback200: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.1.1/24.

9. Interface Loopback30000: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.100.1/24.

10. Interface Loopback202: This is a Loopback interface with a bandwidth of 8000000. It is currently enabled and operational. The interface has an IP address of 172.16.203.1/24.

Show interfaces

Syslog



PCAP!

Please analyze the Packet Capture as JSON and summarize it and highlighting anything important.Be technical.

Run our Conversational Retreival Chain

The provided Packet Capture (PCAP) JSON data contains information about a single network packet. Here are the key details:

1. Frame Layer: The packet was captured on September **5**, **2023**, at **21:05:11** UTC. The packet has a length of **402** bytes. The interface through which the packet was captured is identified as "\\Device\\NPF_{DA923629-D21B-4157-8559-2280FAD5E757}" and it is a Wi-Fi interface. The protocols used in this packet are Ethernet (eth), Internet Protocol (IP), User Datagram Protocol (UDP), and data.

2. Ethernet Layer: The destination MAC address is "34:5d:9e:eb:0c:9b", which is resolved to "Sagemcom_eb:0c:9b" and belongs to "Sagemcom Broadband SAS". The source MAC address is "c8:e2:65:48:73:b2" and is resolved to "IntelCor_48:73:b2", which belongs to "Intel Corporate". The Ethernet type is "0x0800", indicating that the payload is an IP packet.

3. IP Layer: The IP version used is IPv4. The source IP is "192.168.2.61" and the destination IP is "170.72.169.30". The IP header length is **20** bytes and the total IP length is **388** bytes. The IP protocol used is "17", which indicates UDP.

4. UDP Layer: The source port is "52762" and the destination port is "9000". The length of the UDP payload is **368** bytes.

From the above analysis, it seems like a UDP packet was sent from a device with an Intel network interface card, from a private IP address ("192.168.2.61") to a public IP address ("170.72.169.30") on port **9000**. The device at the receiving end has a network interface card from Sagemcom Broadband SAS.

Demo Time!

The bridge to possible

Thank you

cisco Live!


cisco Live!
Let's go