AI Factories:
La fucina per Opportunità e Sfide

Valerio Rizzo, PhD | EMEA AI Technical Lead

“

AI eats HPC– Attack of the Killer Tensor Cores

Prof. Torsten Hofler
ISC 2023

Why Accelerators, Programming and Tools Converged

Workloads moving to low-precision computing
- 30X Energy consumption going from FP8 to FP64
- 30X Compute speed up going from FP64 to FP8

Support for quantization and sparsity
- Vector scaling and zero points
- Structured (N:M) and arbitrary (>50%) sparsity

Data Movement Optimization
- Smaller data-type will manifest latency issue
    - Dealing with Local and sparse connectivity
    - Dealing with skyrocketing cost of network technologies

One language to rule them all
- Python Ecosystem includes AI and DS tools and libraries

*ISC 2023, Torsten Hoefler on Inheriting Excellence: High-Performance Computing at a Crossroads

Symbiotic Mutualistic Relationship

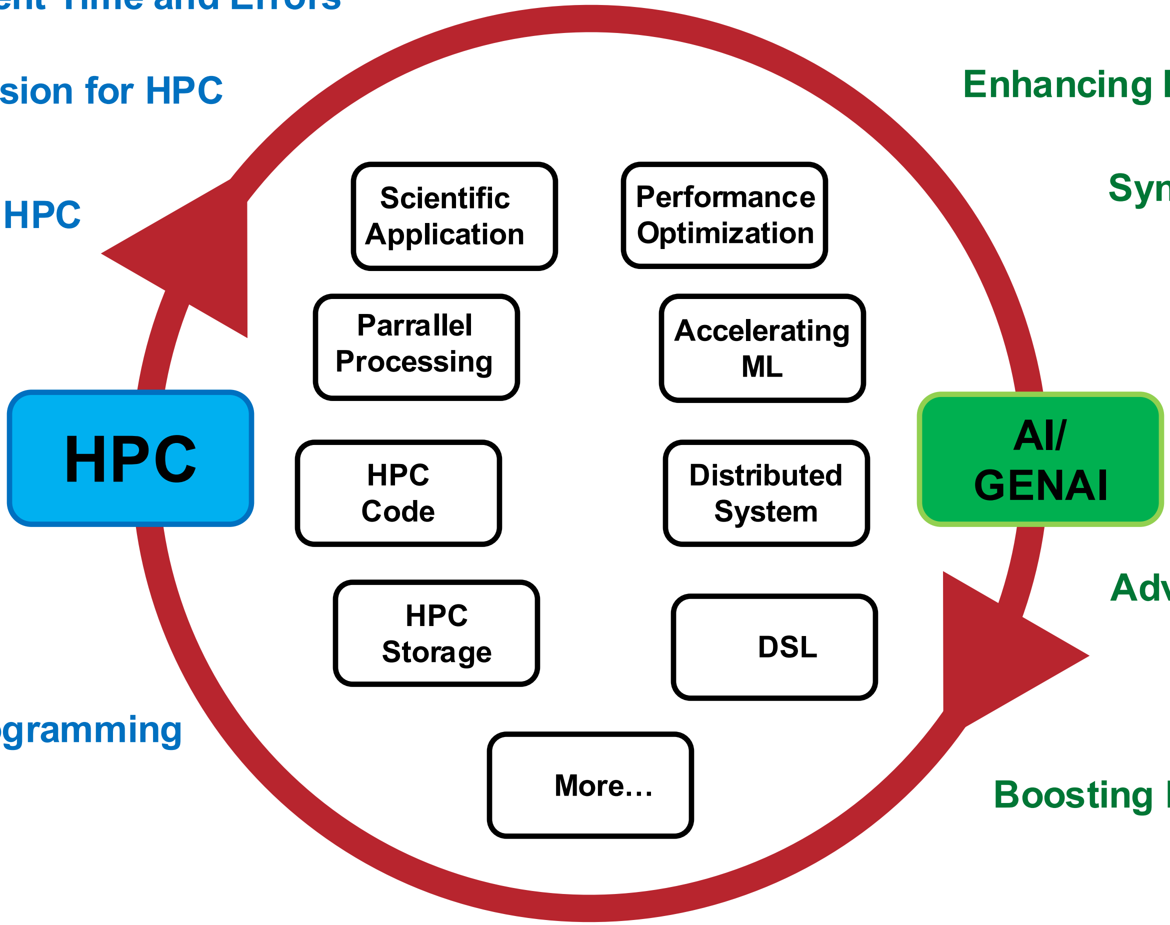
Reduction in Development Time and Errors

Multimodel Learning and Fusion for HPC

Effient Parrallel Code Generation for HPC

Code Representation for HPC

Facilitation of Natural Language Programming



Enhancing Model Size and Complexity

Synthetic Data Generation

Advancing AI/GenAI Training Efficency

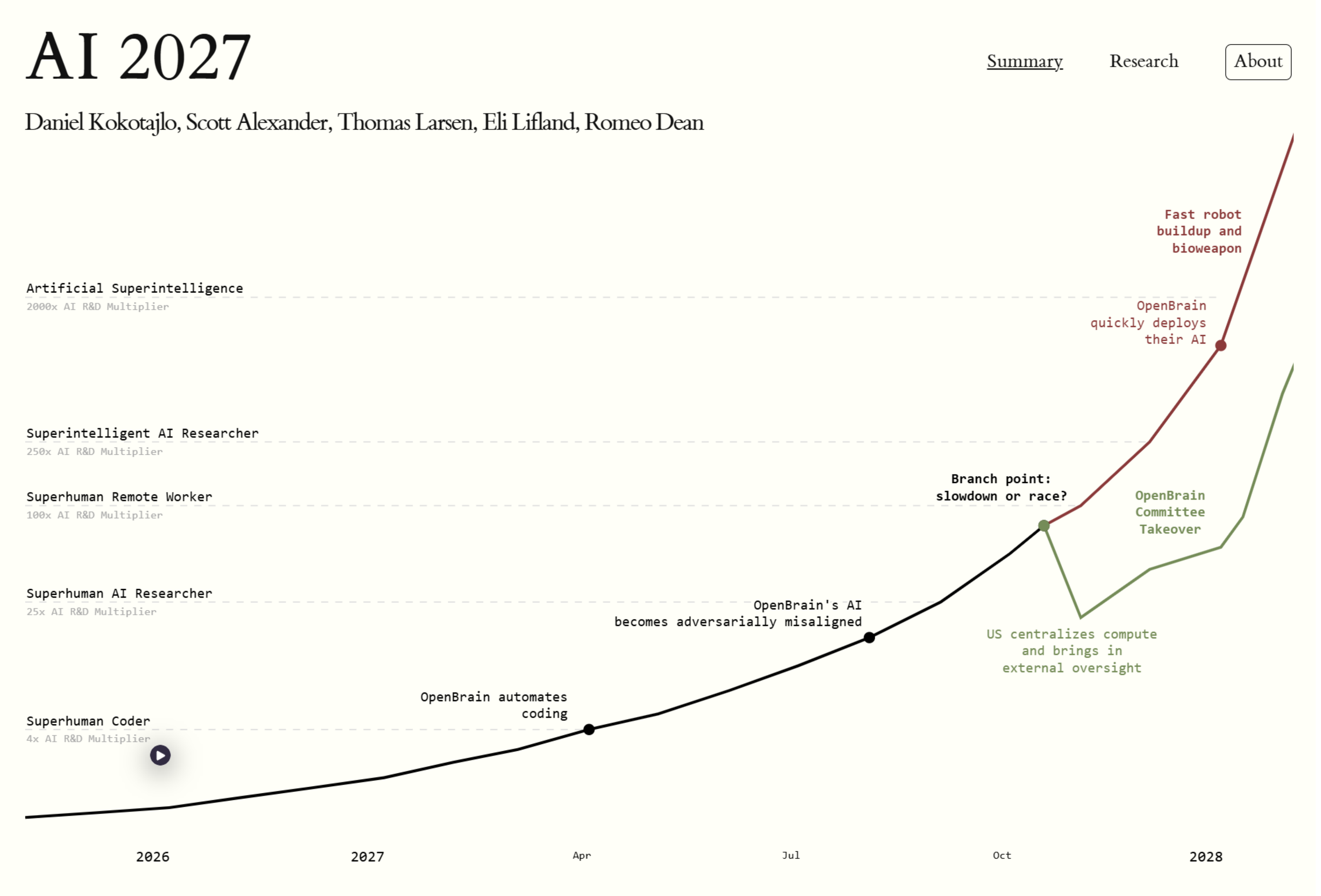
Boosting Latency and Throughput for RT App

Chen, Le, et al. "Position Paper: The Landscape and Challenges of HPC Research and LLMs." arXiv preprint arXiv:2402.02018 (2024).

In the past, we wrote the software and ran it on computers. In the future, the computer is going to generate the tokens for the software

Jensen Huang
CEO, Nvidia

AI of Tomorrow



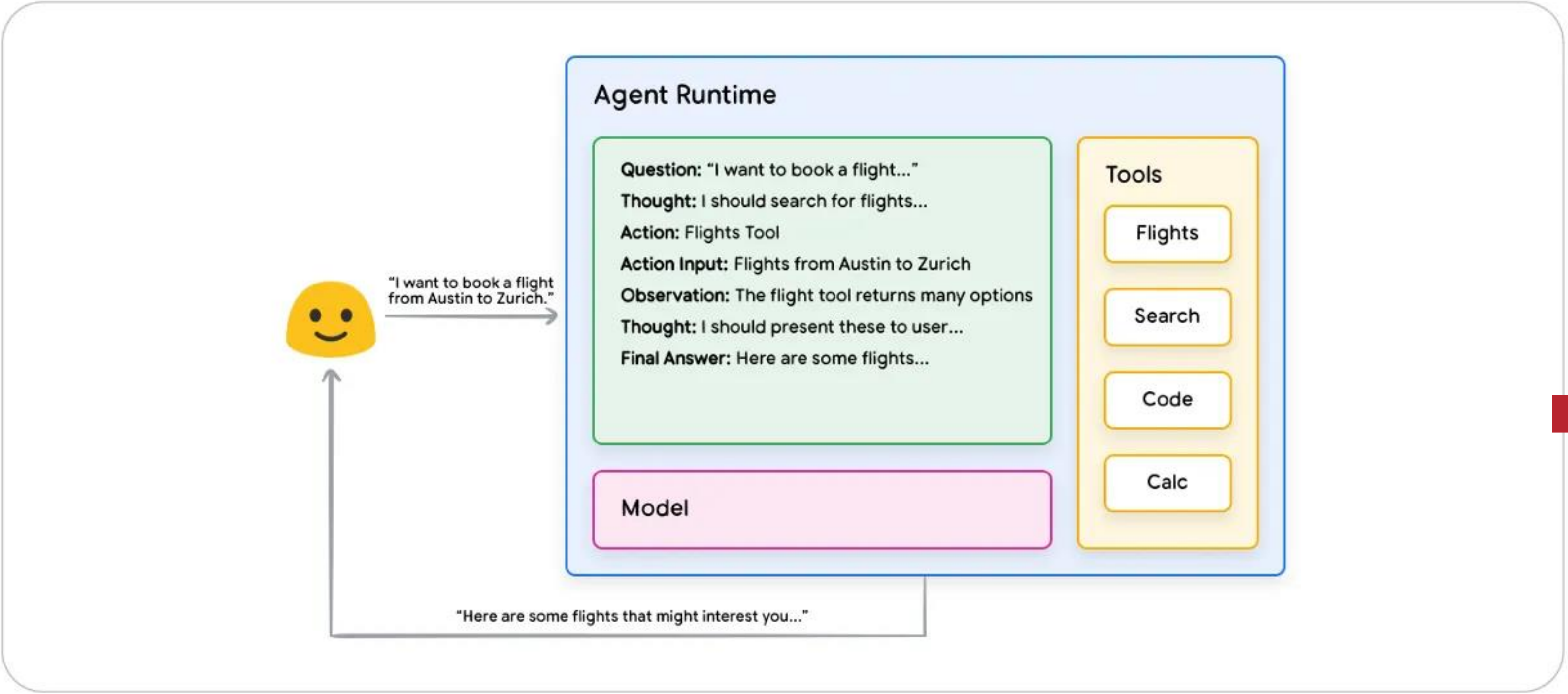
https://ai-2027.com/

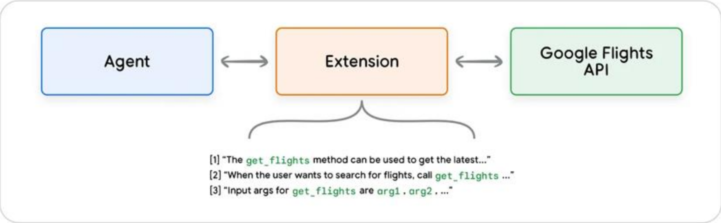
“

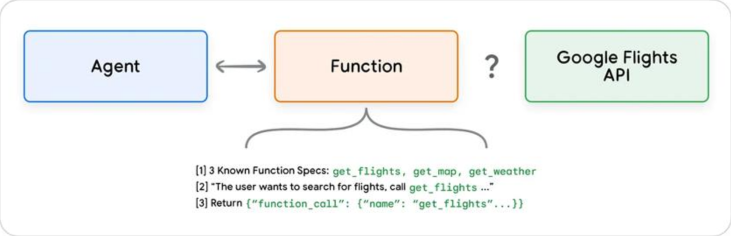
AI agent workflows will drive massive AI progress this year. This is an important trend, and I urge everyone who works in AI to pay attention to it.

Andrew Ng – Forbes “Agents Are The Future Of AI. Where Are The Startup Opportunities?“

AI Agents

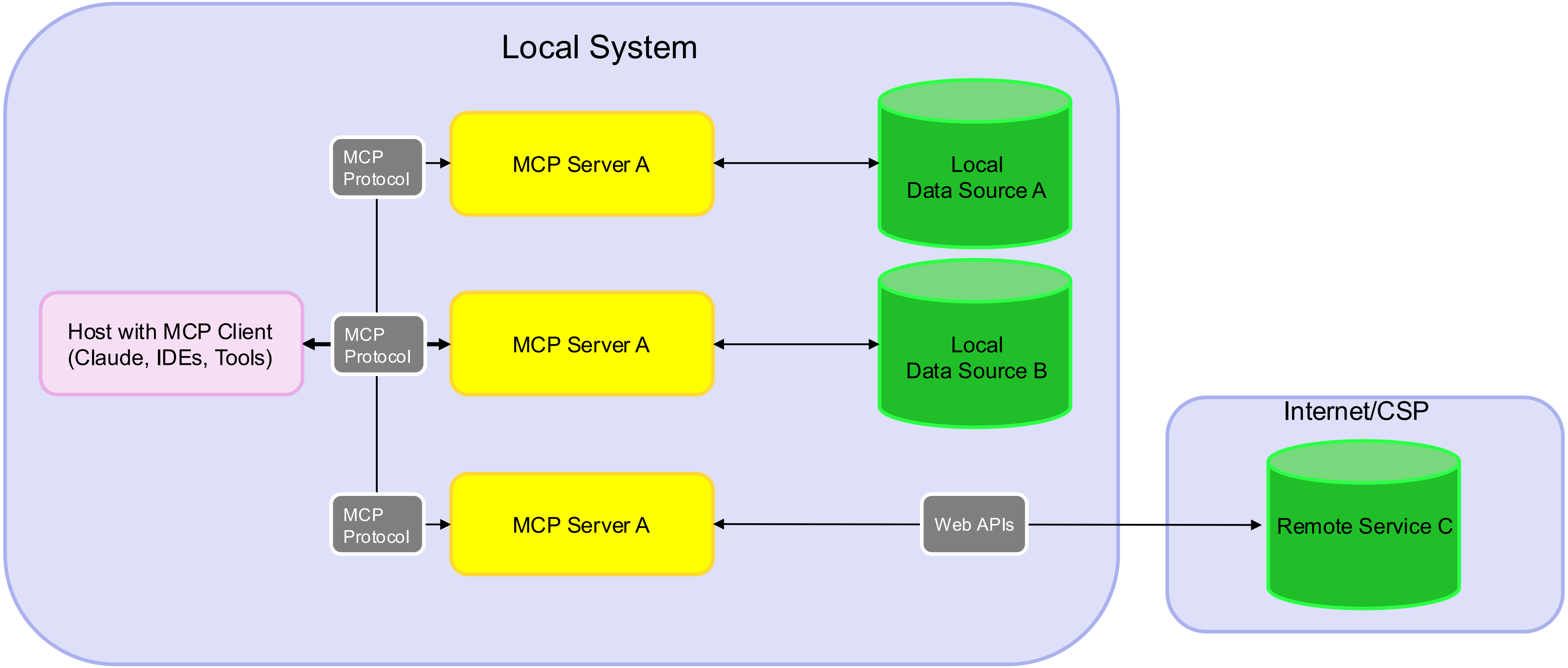






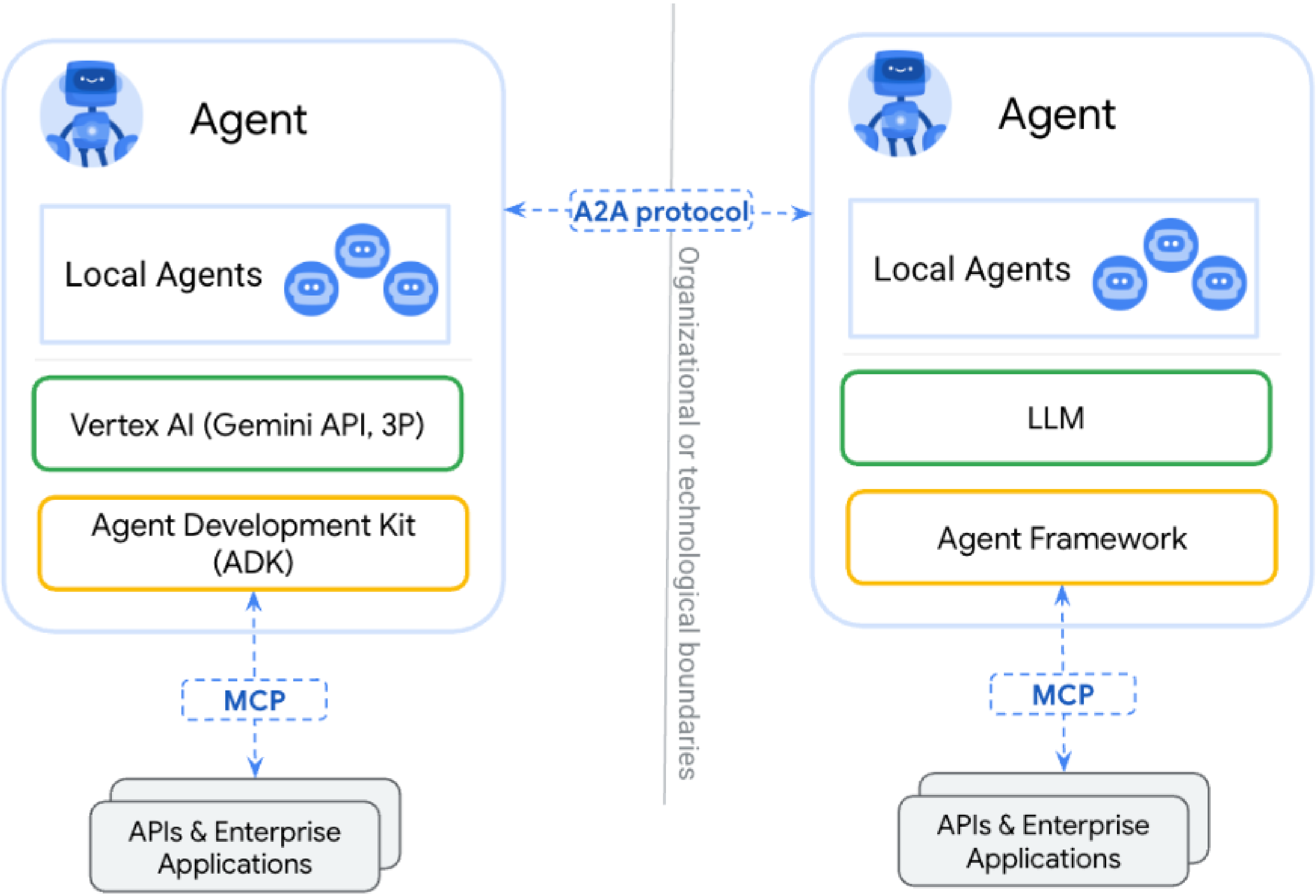

```python
import os
from langgraph.prebuilt import create_react_agent
from langchain_core.tools import tool
from langchain_community.utilities import SerpAPIWrapper
from langchain_community.tools import GooglePlacesTool

# Setting up API keys
os.environ["SERPAPI_API_KEY"] = "XXXXX"
os.environ["GPLACES_API_KEY"] = "XXXXX"

# Define the search tool using SerpAPI
@tool
def search(query: str):
    """Use the SerpAPI to run a Google Search."""
    search = SerpAPIWrapper()
    return search.run(query)

# Define the places tool using Google Places API
@tool
def places(query: str):
    """Use the Google Places API to run a Google Places Query."""
    places = GooglePlacesTool()
    return places.run(query)

# Initialize the model
model = ChatVertexAI(model="gemini-1.5-flash-001")

# List of tools available to the agent
tools = [search, places]

# Query to ask the agent
query = "Who did the Texas Longhorns play in football last week? " \
        "What is the address of the other team's stadium?"

# Create the agent using the model and tools
agent = create_react_agent(model, tools)

# Input message structure
input = {"messages": [("human", query)]}

# Process the agent's response in stream mode
for s in agent.stream(input, stream_mode="values"):
    message = s["messages"][-1]
    if isinstance(message, tuple):
        print(message)
    else:
        message.pretty_print()
```

Model Context Protocol



Agent-2-Agent Protocol

Agentic AI will be the new Stack?



Agentic AI in Particle Physics

The AutoFLUKA case

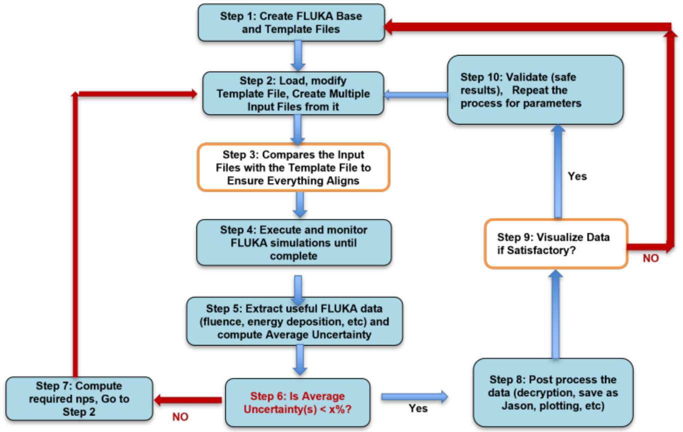


Figure 1: Schematic of a hypothetical FLUKA workflow adopted for the automation. Steps 3 and 9 are highlighted because they require a human in the loop to verify the FLUKA-Fortran code syntax and to judge the accuracy of the results respectively.

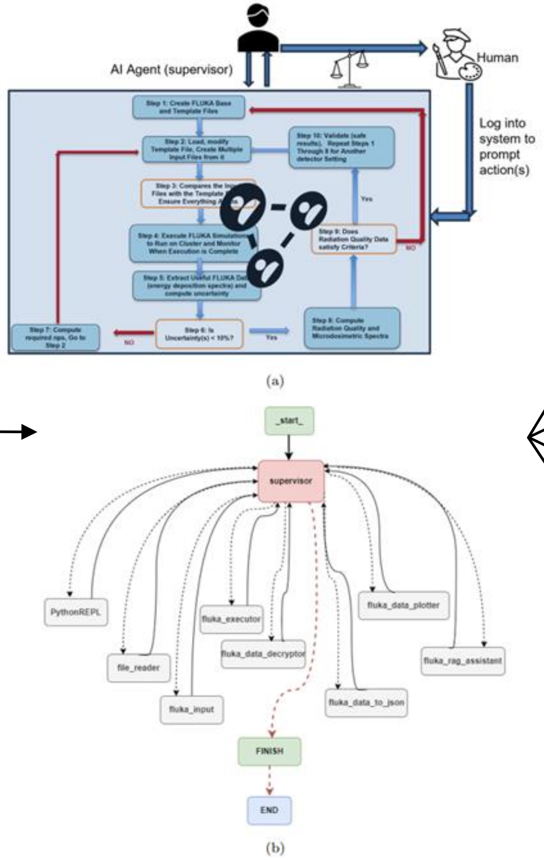


Figure 4: Schematic view of the multi-agent workflow, showing the supervisor AI agent at the top, coordinating actions of different agents within the blocks. (b) –Graph visualization of the multi-agent workflow. Notice that each agent takes actions from as well as reports back to the agent supervisor until the task is marked as complete, after which the FINISH + END sequence is triggered. The human in the loop is to log into the system to initiate the action.

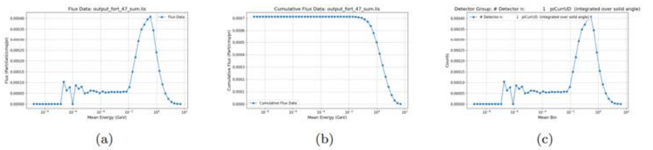


Figure 6: Plots generated during the workflow; (a)- Spectral flux extracted from the **"_sum.lis"** output file; (b)- cumulative flux also extracted from the **"_sum.lis"** output file; (c)- Spectral flux extracted from the bin-wise **"_tab.lis"** output file. Notice that this is identical to the **"_sum.lis"** file data and was done for verification purposes.

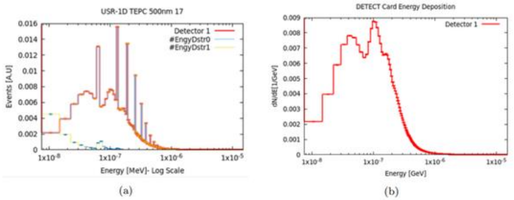


Figure 7: (a) -Energy deposition spectrum in a low-density Tissue Equivalent gas recorded by DETECT card in FLUKA showing unwanted spikes due to wrong settings in the physics cards; (b) The same code with the inclusion of EMFCUT (PROD-CUT activation), DELTARAY, MULSOPT (with single scattering activation) and removal of PART-THRES and EMFFIX cards as recommended yielded the correct results. AutoFLUKA was able to reproduce these recommendations even more concisely

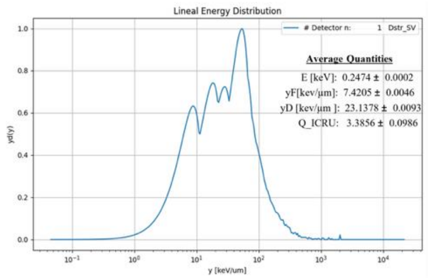


Figure 9: Dose distribution of the lineal energy generated by the microdosimetric spectra tool. E represents the Average energy deposited in the Detector's sensitive volume which simulates the Tissue site, yF represents the frequency-mean of the lineal energy while yD represents the dose-mean of the lineal energy.

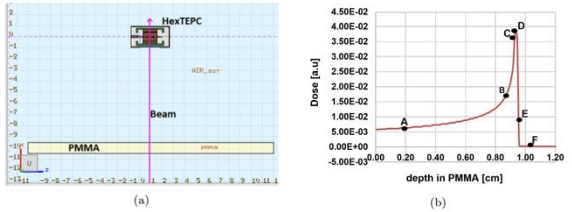


Figure 8: a) - Irradiation scenario in FLUKA for the Design of and optimization of HexTEPC ; (b) Different irradiation positions across the Bragg peak for a 62 MeV/u carbon ion beam passing through Polymethyl methacrylate or PMMA (Ndum et al., 2024).

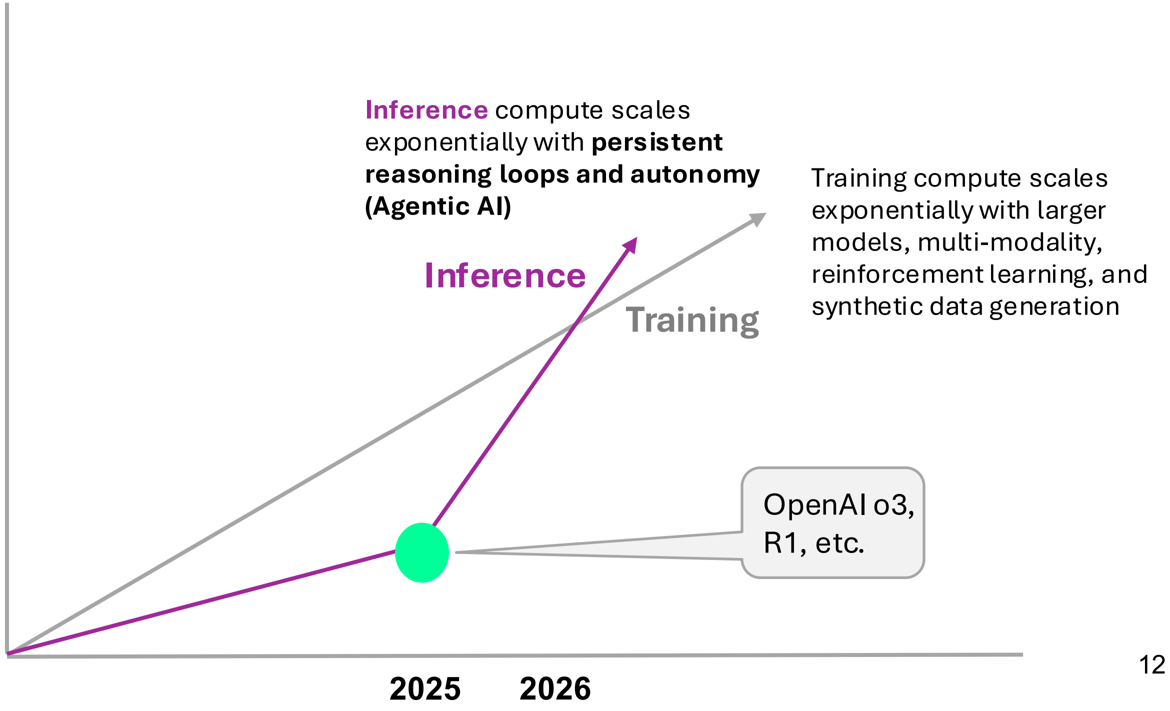
Exponential Demand for Compute

The amount of computation we need at this point because of **agentic AI**, because of **reasoning**, is easily **100 times more** than we thought we needed this time last year.

— Jensen Huang, GTC 2025 Keynote

AI is going through an inflection point

- "By 2028, over 80% of AI accelerators will be used for inference, up from 40% in 2023." **Gartner, (2024)**
- "Inference is expected to become the dominant workload by 2030." **McKinsey (2024)**
- "As the focus of AI shifts from training to inference, edge computing will be required to address the need for reduced latency and enhanced performance." **IDC (2023)**



AI Factories

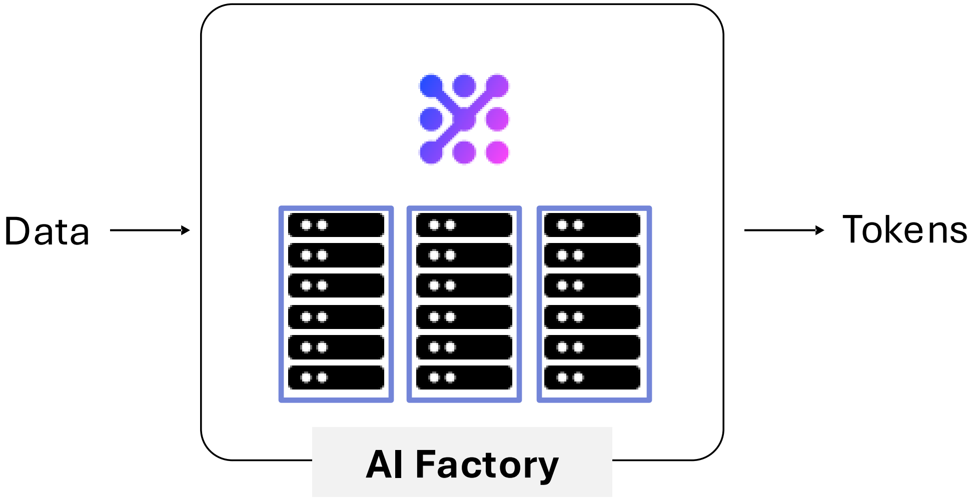
New Class of Data Centers

Production of tokens for digital intelligence

Raw Material: *Electricity + Data*

Product: *Tokens*

Production metric = $\dfrac{(Token\ /\ sec)}{W}$



Infrastructural Challenges
- Hardware -

INNOVATION

The Silent Burden Of AI: Unveiling The Hidden Environmental Costs Of Data Centers By 2030

By Yusuf Sar, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Aug 16, 2024, 07:45am EDT

- Microsoft plans to invest $100 billion over the next five years
- Google plans to invest an additional $100 billion in expanding its data
- Amazon is also heavily investing in the same range as Google and Microsoft.

Illustrative Examples of the Impact:

- Power consumption is projected to increase to 1743 TWh by 2030 (from 524 TWh)
  → 828.925 million tons of CO_2
- Additional 73.931 million tons of CO_2 for Server Production

Key considerations:

- Sustainable Power Integration
- Hardware Lifecycle Management
- Transparent Emissions Reporting
- Innovation In Energy Efficiency

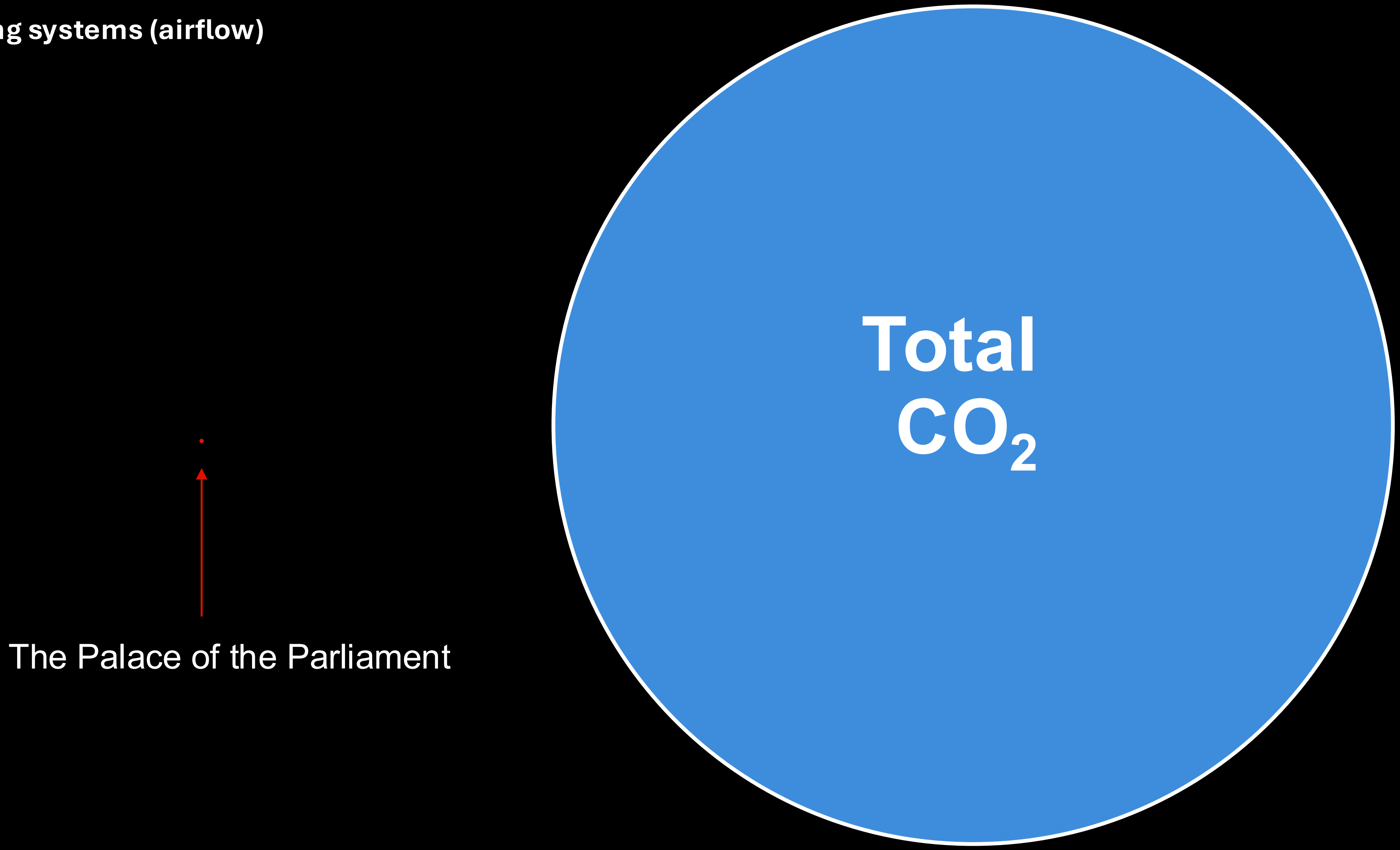
How to visualize 828Mt?

The Palace of the Parliament is the heaviest building in the world, weighing about

4Mt

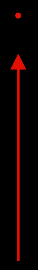
The Cost of AI

traditional cooling systems (airflow)

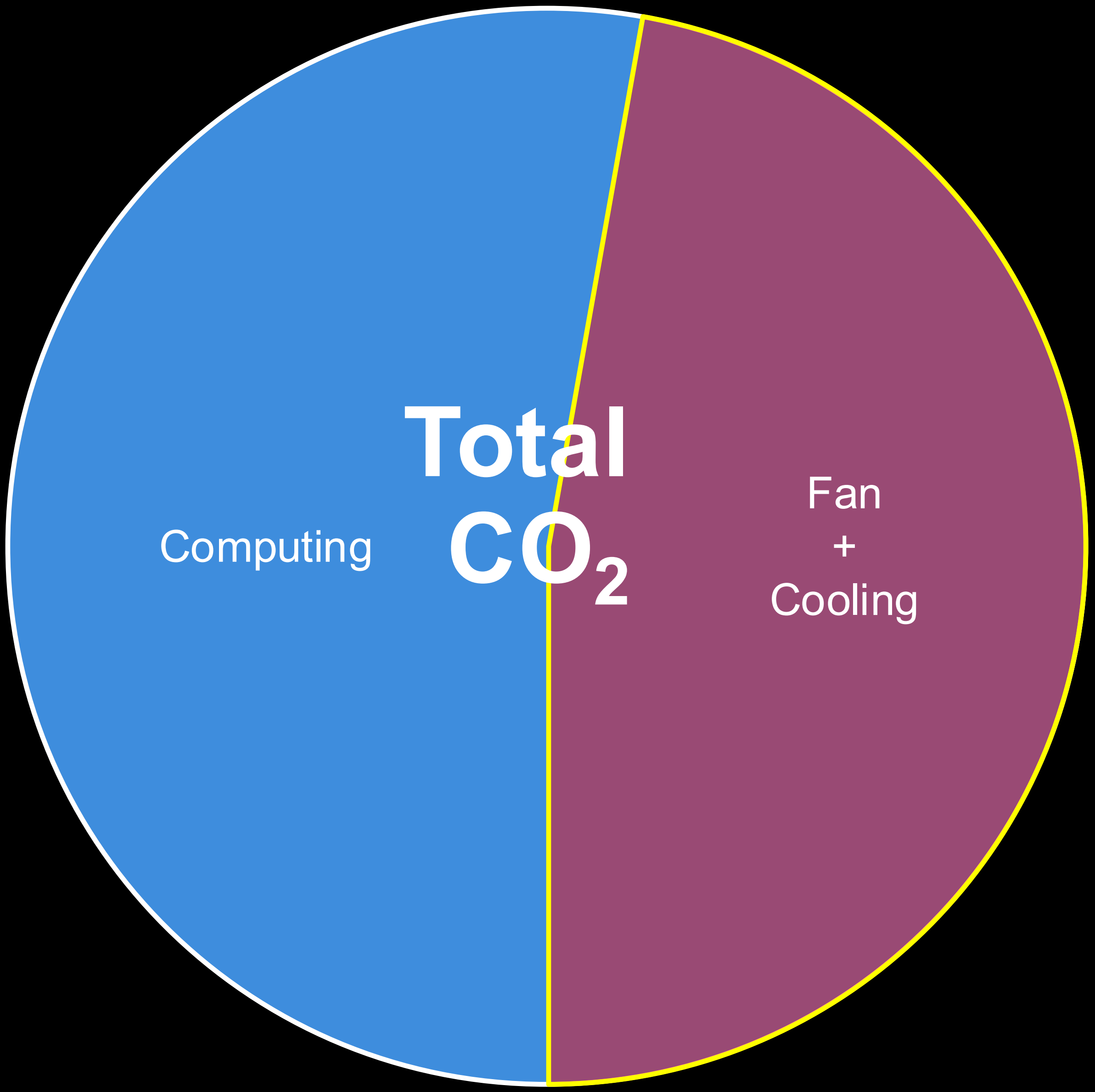


Lenovo

The Cost of AI



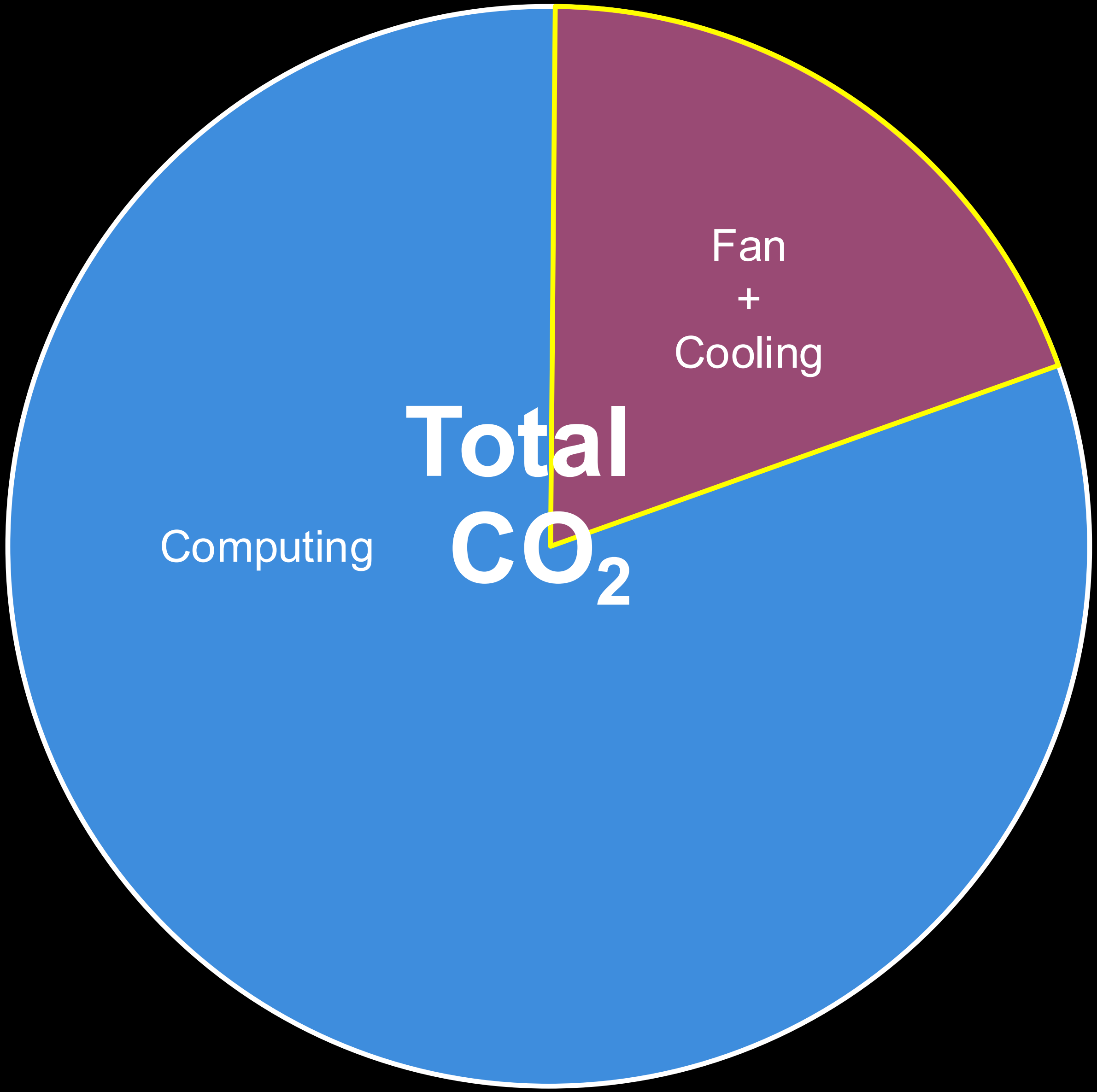
The Palace of the Parliament



The Cost of AI

Neptune DWC systems (liquid)

The Palace of the Parliament

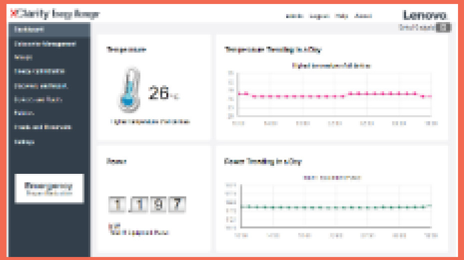


Smarter Engineers in Energy Efficiency – a Lifetime Payback

Energy Management Tooling

Measure, manage, optimize power

xClarity Energy Manager (LXEM)

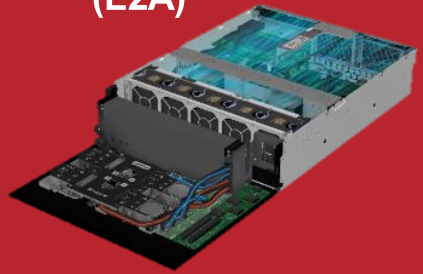


Energy Aware Runtime (EAR)

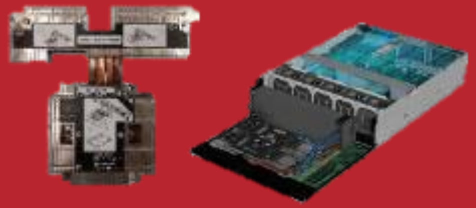
Heat Mitigation Innovation

Lower fan speed, less power

Liquid to Air Exchanger (L2A)



Thermal Transfer Module (TTM)



Efficient Component Selection

Same work, less power

Titanium Power Supplies

Idle Power State Controls

Dynamically optimize the frequency and power control

Low Loss Materials

Liquid Cooling

Rethink how cooling gets done

Rear Door Heat Exchanger

Neptune™

Liquid Cooled Systems

Lenovo Sustainable Solutions

Lenovo Neptune™

Up to 40% reduction in power costs resulting from a 3.5x improvement in thermal efficiencies vs. air cooled

Factory Integrated Racks

Saving 3.5 million pounds of cardboard and 1.8 million pounds of plastic over 5 years

More Sustainable Packaging

Including the use of 90%+ recycled foam and bags made from 30% ocean bound plastic

Lenovo TruScale

Avoids over-provisioning, reducing energy consumption for a lower carbon footprint

CO2 Offset Services

Carbon offset credits fund projects, including reforestation, renewable energy, and solar

Lenovo Asset Recovery

15 years experience in asset recycling and more than 1M+ assets properly disposed

MLPerf Race

7+
Rounds Submissions

20+
Winnings Benchmarks

125+
MLCommons Members and Affiliates

ML Commons

50+
Benchmarked Platforms

2020
Started Membership

6
Benchmark Suites

56,000+
MLPerf Performance Results to-date

Better AI for Everyone

Building trusted, safe, and efficient AI requires better systems for measurement and accountability. MLCommons' collective engineering with industry and academia continually measures and improves the accuracy, safety, speed, and efficiency of AI technologies.

- Reduced complexity, quicker deployments and time to value
- Provides a standard 'ruler' for AI workload performance
- Helps improve AI workload performance through software improvements
- Guide our customers to the best solutions

Infrastructural Challenges

- Software -

Critical to Making LLM Work:
GPU Efficiency Solve the Bottleneck of LLM Scalability

- **Very few people can build cluster with >10,000 GPUs**
  - Complex communication bottlenecks makes GPU with low MFU.
  - Loss convergence becomes more difficult in ten thousands of GPU distribute training cluster.
  - Larger cluster brings higher failure rate.

- **Frequent GPU failures****
  - more than 20 times per month in a 1000 GPU cluster, and GPU failures are random.
  - Single GPU failure brings down the entire cluster of 10,000GPUs for 1-2 hours.

- **Low MFU* wastes 60+% of the money**
  - It is very hard to improve MFU of GPU, usually<40% in big cluster.
  - 50%+ of the computation doesn't matter in training (not help convergence).

- **Experiments compete for GPUs with model training**
  - LLM training requires thousands of GPUs over several months.
  - Experiments require clusters of different sizes for days to weeks.
  - Static allocation of GPUs results in suboptimality & under utilization.

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	4	131,072	16	16M	380	38%

Table 4 **Scaling configurations and MFU for each stage of Llama 3 405B pre-training.** See text and Figure 5 for descriptions of each type of parallelism.

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 **Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training.** About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

*MFU: Model Flops Utilization
**The Llama 3 Herd of Models Llama Team, AI @ Meta

Hybrid AI Factory with Lenovo & NVIDIA

Validated applications & solutions

60+ AI Applications & Platforms

Lenovo **AI Innovators**

165+ Enterprise Solutions

AI Professional Services

Discovery | Advisory | Fast Start | Deploy & Scale | Managed

Data Management

Metadata store, Data Lake & Governance.

ONTAP · CLOUDERA · dataiku

LLMOps + System Orchestration

AI Platforms & Data Science Tools

NVIDIA AI Enterprise

NVIDIA Dynamo

LiCO · run:ai · mlflow · cnvrg.io · FlexAI · inai

Data Lake
Structure, Unstructured, Vector DBs, Graph DBs Knowledge Bases, IoT/Sensors

Data Tools
Governance, Labeling, Cleansing, Data Fabric

Frameworks
Voice, Vision, LLM

Pre-trained AI Models
ML, Voice, Vision, LLM

Platforms

Reference Architectures

- vmware — **Private AI Platform**
- NetApp — **AIPOD™**
- docker · kubernetes — **Open-source Private AI**
- NUTANIX — **GPT-in-a-Box™**
- Red Hat — **OpenShift AI**

Storage, Network & Compute

Lenovo ThinkSystem · Lenovo ThinkAgile

80+ AI Ready Platforms

NetApp · IBM Spectrum Scale · WEKA · ddn

OVX 4-8 GPU · **HGX 4 GPU** · **HGX 8 GPU**

NVIDIA

AI Starter Kits for Lenovo Hybrid AI Platform

Simple solutions to speed enterprise AI deployments

Rapid AI adoption with pre-validated, pre-configured solutions

Simplified support with end-to-end Lenovo solution

Flexible expansion to scale with a growing business

	Small	Medium	Large	XL
Compute	SR675 V3	SR675 V3	SR675 V3 (2)	SR675 V3 (4)
Storage	DG5200	DG5200	DM7200F	DM7200F
GPUs	L40s (4)	L40s (8)	L40s (16)	L40s (32)
Networking	SN3700	SN4600	SN4600	SN4600

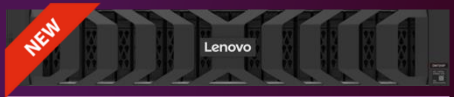
Use Cases:

- ❖ RAG
- ❖ Inferencing
- ❖ Model Fine Tuning

The New 2025 Lenovo Data Storage Solutions Portfolio

21 New ThinkSystem & ThinkAgile Products

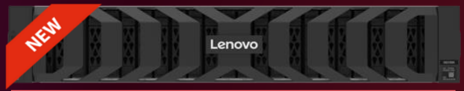
Mid-Range Flash		Entry Flash/Hybrid	Software-Defined	SMB FC Switch



ThinkSystem DM Series

High Performance, Unified AI Data

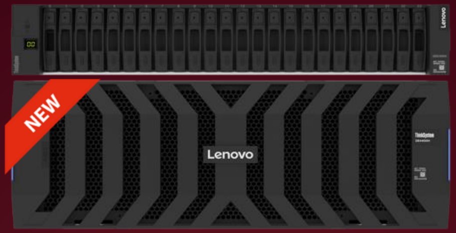
- ThinkSystem DM7200F
- ThinkSystem DM5200F
- ThinkSystem DM5200H
- ThinkSystem DM3200F



ThinkSystem DG Series

Flash Performance at HDD Economics

- ThinkSystem DG7200
- ThinkSystem DG5200



ThinkSystem DE Series

Simple, Channel Friendly Entry-level Storage

- ThinkSystem DE4800F (2U24)
- ThinkSystem DE4800H (2U12)
- ThinkSystem DE4800H (2U24)
- ThinkSystem DE4800H (4U60)
- ThinkSystem DE4200H (2U12)
- ThinkSystem DE4200H (2U24)

ThinkAgile V4 Series

Accelerated Roll-Out & Simplified Lifecycle

ThinkAgile HX Series
- ThinkAgile HX630 V4
- ThinkAgile HX650 V4
- ThinkAgile HX650 V4 Storage

ThinkAgile VX Series
- ThinkAgile VX630 V4
- ThinkAgile VX650 V4
- ThinkAgile VX650 V4-SAP

ThinkAgile MX Series
- ThinkAgile MX630 V4
- ThinkAgile MX650 V4

ThinkSystem DB Series

Gen 7 Entry-level Switch for SMB

- ThinkSystem DB710S

ISG Common Bezel

Smarter
technology
for all

Lenovo

thanks.