

The Rise of AI the Good, the Bad, and the Ugly



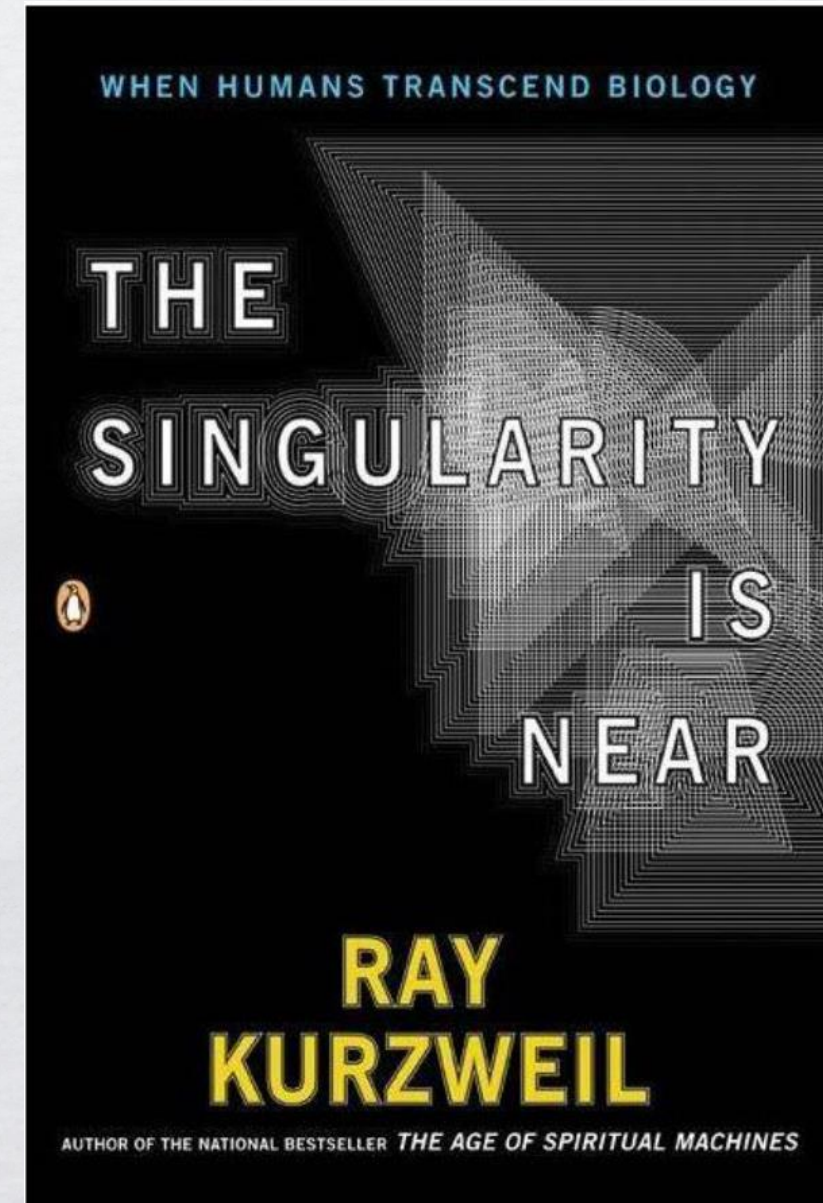
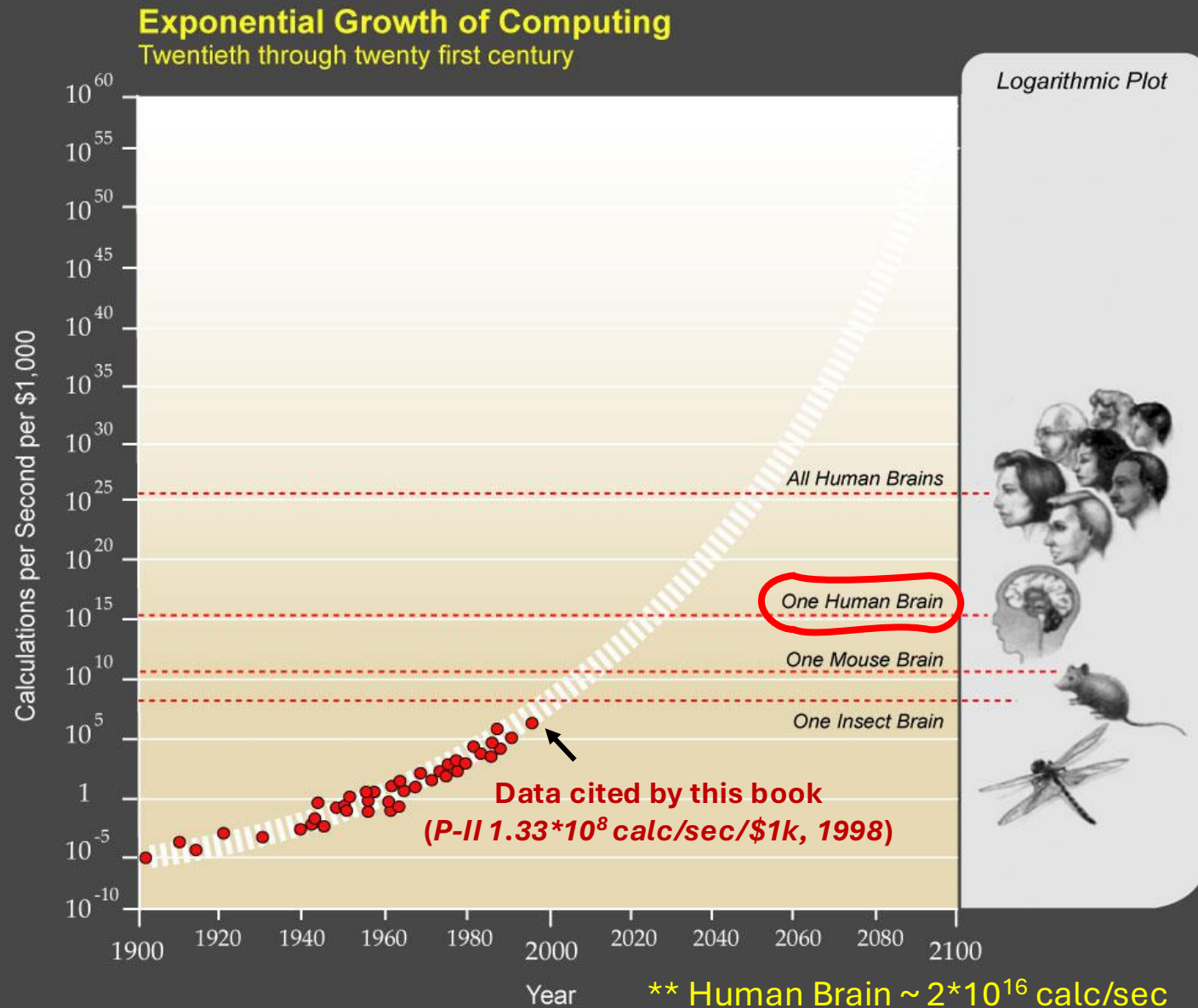
Hsien-Hsin Sean Lee

~~Age~~ ~~6~~ ~~SI?~~ ~~X~~ The Rise of AI the Good, the Bad, and the Ugly

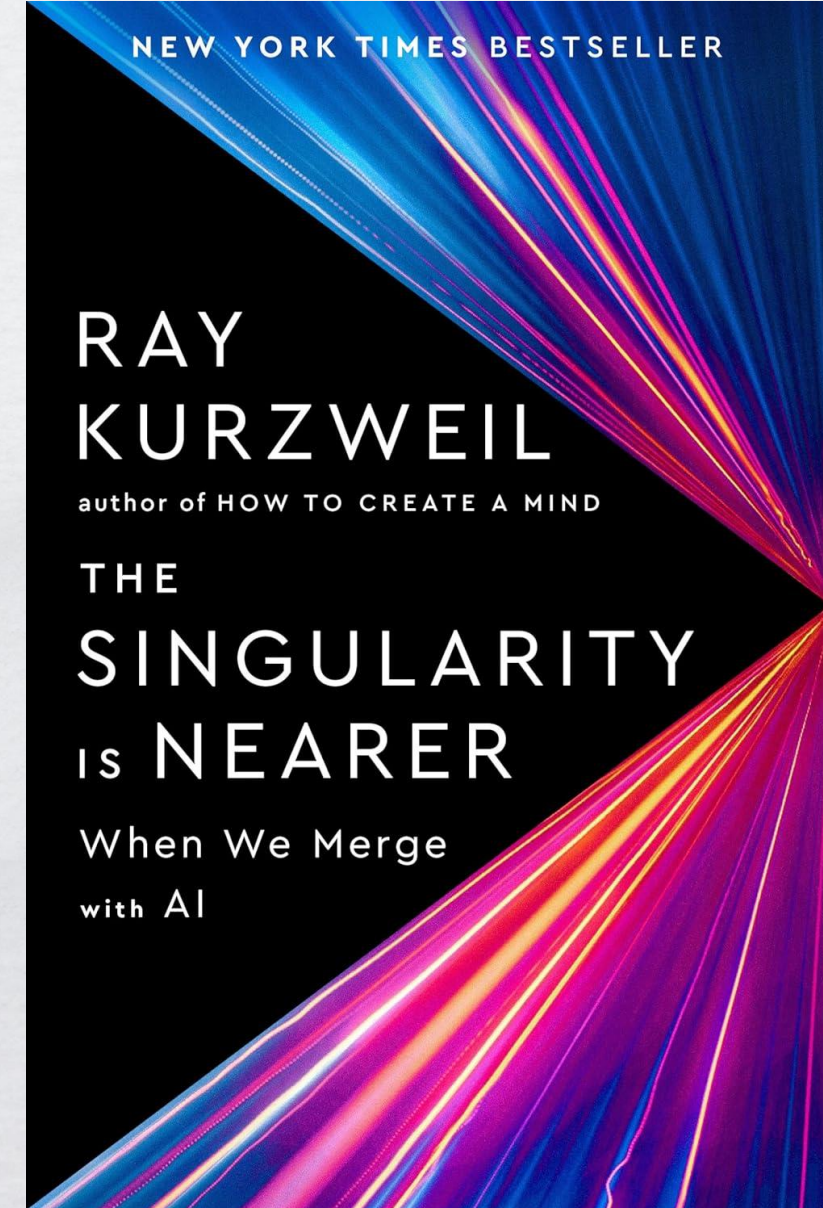
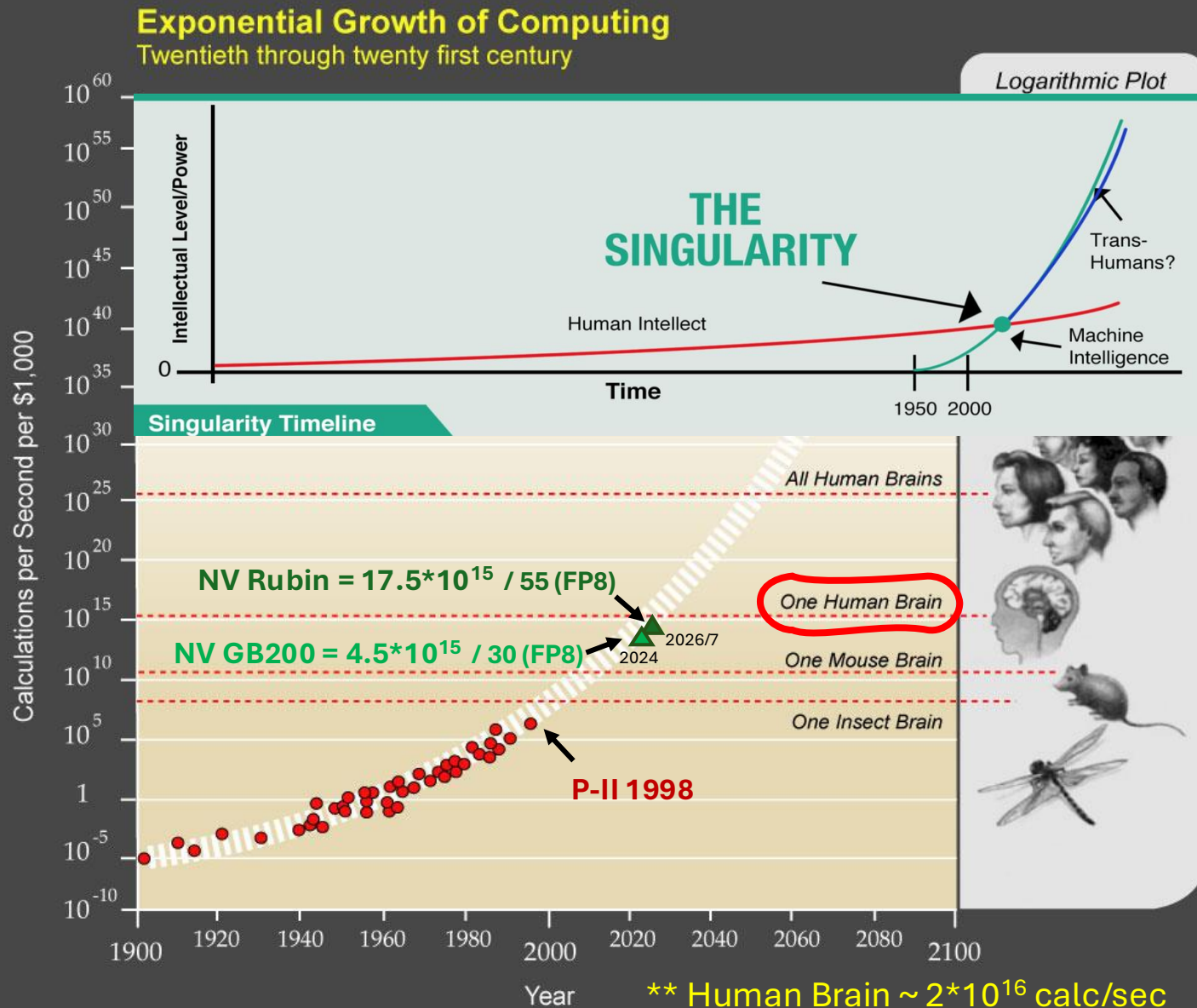


Hsien-Hsin Sean Lee

The Singularity is Near — Ray Kurzweil circa. 2005

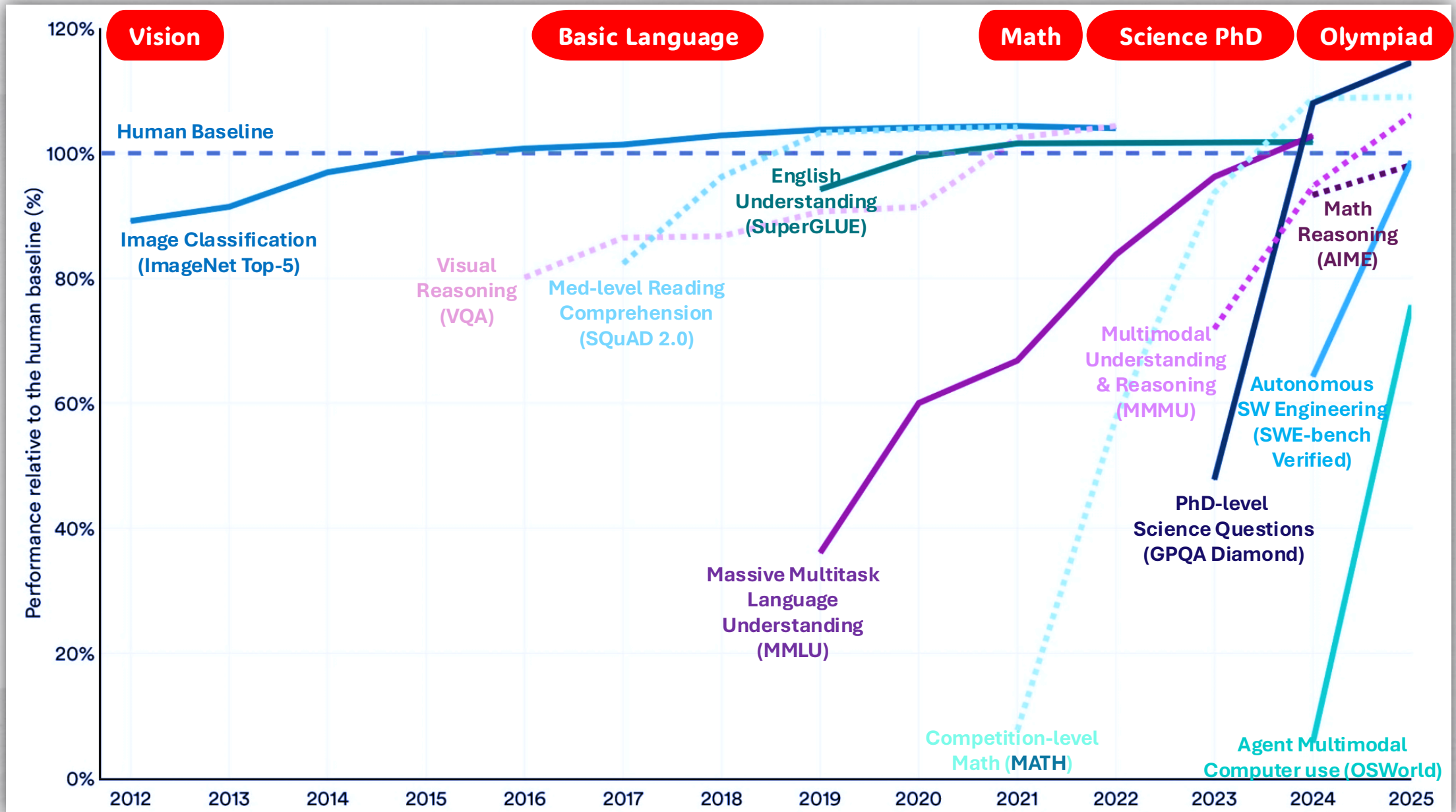


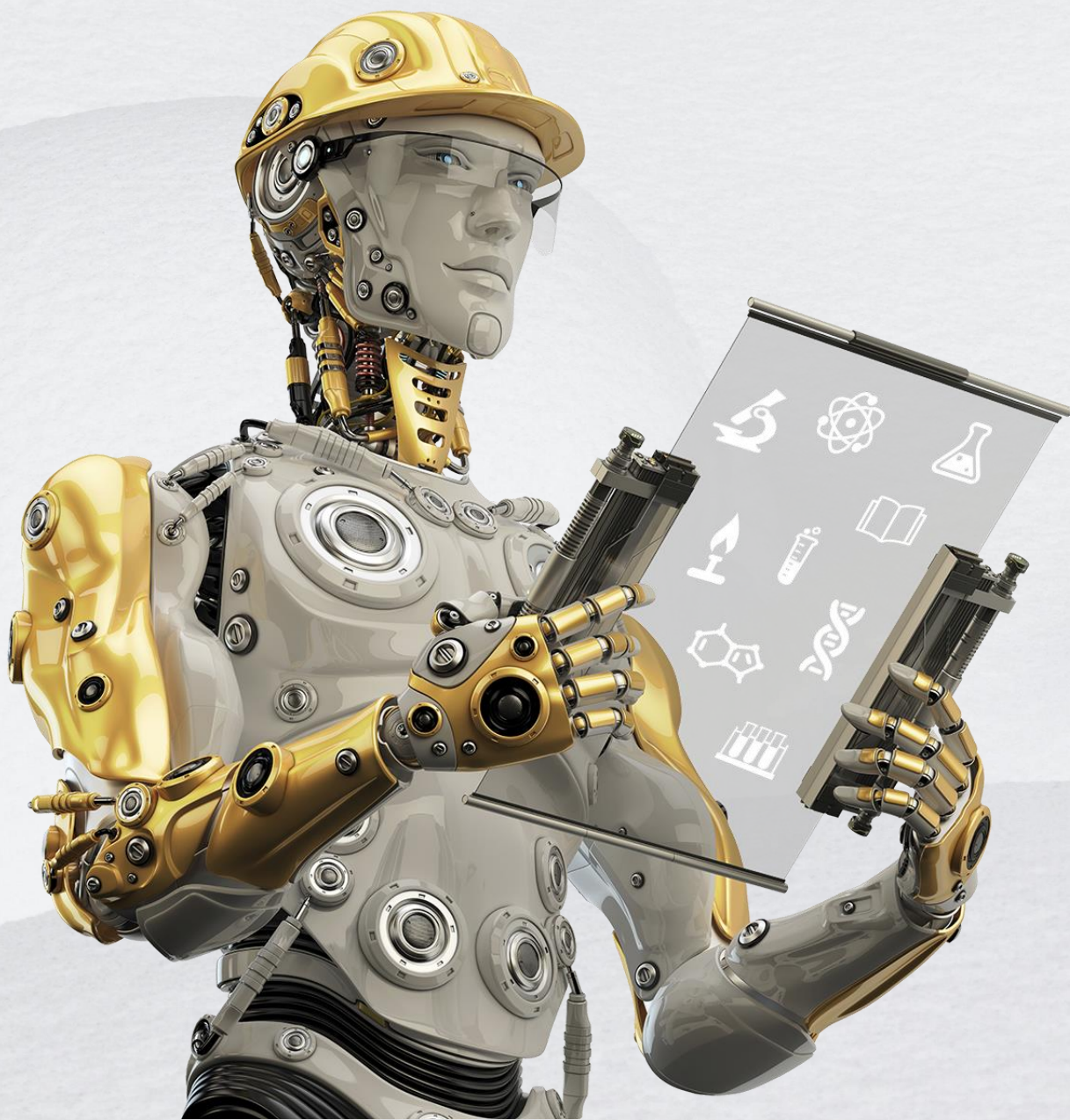
The Singularity is Nearer — Ray Kurzweil



Published in 2024

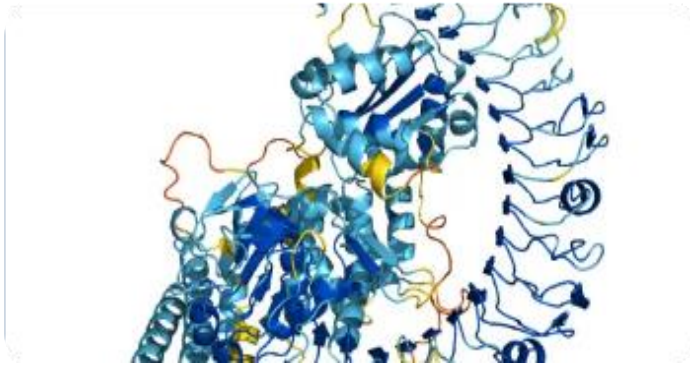
Performance Benchmarking — AI vs. Human





THE GOOD

Frontier AI: Biology, HealthCare, Math, Education, Business



🧬 LIFE SCIENCE & HEALTH

AlphaFold 3 & Drug Discovery

AI's Role: Generative structural biology replacing physical trial-and-error.

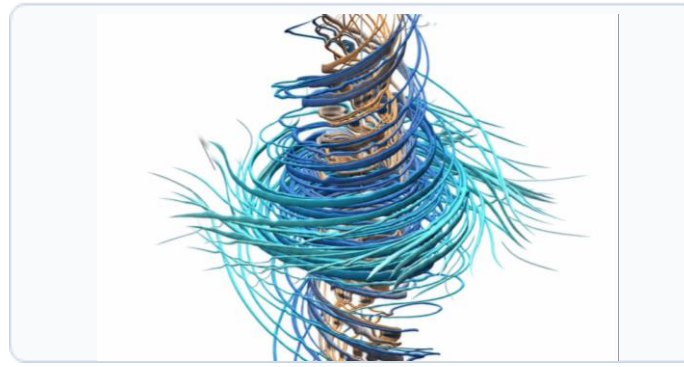
55 years vs. 1 Model: Labs solved ~260k protein structures since 1971. AlphaFold has predicted over 214M—nearly every known protein.

Lock and Key: AlphaFold 3 predicts how proteins bind DNA, RNA, and drug molecules—see the lock, then design the key (drug).

Into the Clinic: "Isomorphic Labs" in partnership with Lilly and Novartis is taking its first AI-designed cancer drugs toward human trials.

260K → 214M+

55 yrs of lab-solved → AlphaFold-predicted (~800×)



✓ DEEP MATH & FOUNDATIONS

Navier-Stokes Millennium Problem

AI's Role: Massive multi-agent theorem proving & formal Lean verification.

10,000 Agent Swarm: OpenAI deployed 10,000 coordinated reasoning agents running continuously for 88 hours to construct the solution.

Clay Problem Solved: Synthesized a 166-page proof establishing smooth forced blow-up, verified with machine-compiled Lean code.

Frontier Dispute: Triggered high-stakes priority and prompting debate with Prof. Buckmaster (NYU) and Dr. Alpöge (Anthropic).

10,000 Agents @ 88 hrs

1 of 7 Clay Millennium Math Problems



⚡ FUTURE OF BUSINESS OPERATIONS

The Solo Unicorn (Hyper Leverage)

AI's Role: Autonomous multi-agent workforces replacing human headcount.

Lightning Growth: "Medvi" founded in 2024 will reach a \$1.8B run-rate in 2026 with only 2 founders by deploying continuous AI agents for code, ops, and ads.

Hyper-Productivity: Rivals traditional 2,400+ person enterprises through automated agent orchestration.

The New Economic Era: Validates Sam Altman's prediction: single-founder billion-dollar firms are now operational reality.

\$1.8B / 2 People

1,000x Revenue per Employee
Autonomous Agent Leverage

Autonomous IC Design: Spec-to-GDSII (Commercial) AI Agent Flow

Cadence
Ecosystem

Verkor.io "Design
Conductor"

 **FRONT-END** Domain: Multi-Agent Logic Reasoning, RTL & Verification

 **BACKEND** Domain: Reinforcement Learning & Physical Sign-off

STAGE 01 SPEC

Spec Ingestion

 **Spec Agent**

- Architecture parsing
- CSR & register maps
- Target PPA limits

STAGE 02 HDL

RTL Generation

 **RTL Architect**

- Verilog | SystemVerilog authoring
- Synthesis-guided timing
- Bus protocol FSMs

STAGE 03 SIM

Verification & Debug

 **Debug Super-Agent**

- UVM testbench suites
- Waveform failure root-cause
- Closed-loop RTL repair

STAGE 04 P&R

Place & Route

 **Physical RL Agent**

- Floorplanning & power mesh
- Congestion / density tuning
- Automated Clock Tree (CTS)

STAGE 05 SIGNOFF

Sign-Off & Tapeout

 **Closure Engine**

- Autonomous STA slack ECO
- DRC & LVS auto-fixing
- GDSII tapeout generation

 **Cadence AI Flow** Tier-1 EDA Leader

- ✓ **Front-End: "ChipStack AI Super Agent" & Verisium**
Employs structured mental models to prevent hallucinations. Converts natural language specs into RTL code, runs Xcelium regressions, analyzes failure waveforms, and repairs code in a closed loop; **10x** productivity improvement claimed.
- ✓ **Backend: Cerebrus AI Studio (Innovus/Tempus/Pegasus)**
Reinforcement learning engine autonomously optimizing floorplans, placement density, CTS skew, and routing. Integrates with Tempus/Pegasus for **sign-off timing closure and DRC/LVS-clean** tapeout.

Claims: 24% area reduction, 18% power reduction, 100% functional accuracy..

 **Verkor.io Design Conductor** AI Startup

- ✓ **Front-End: Multi-Agent Microarchitecture Reasoner**
Takes brief natural-language specifications to autonomously architect 5-stage RISC-V CPU pipelines, hazard units, Booth-Wallace multipliers, and runs automated Spike ISA validation test benches.
- ✓ **Backend: Autonomous Synthesis to GDSII (ASAP7 PDK)**
Conducts automated logic synthesis, floorplanning, CTS, and routing iterations with zero human intervention, delivering closed-timing GDSII at **1.48 GHz** within a single 12-hour turnaround using a **219-word** prompt,

Targets: Rapid development for inference accelerators, custom RISC-V cores & IoT ASICs.



KEY TAKEAWAY

Commercial AI-enabled EDA tools have transitioned from basic assistive copilots to (fully) autonomous agents for end-to-end Spec-to-GDSII tapeouts.

12 Hrs
Spec-to-GDSII

100%
DRC/LVS Clean

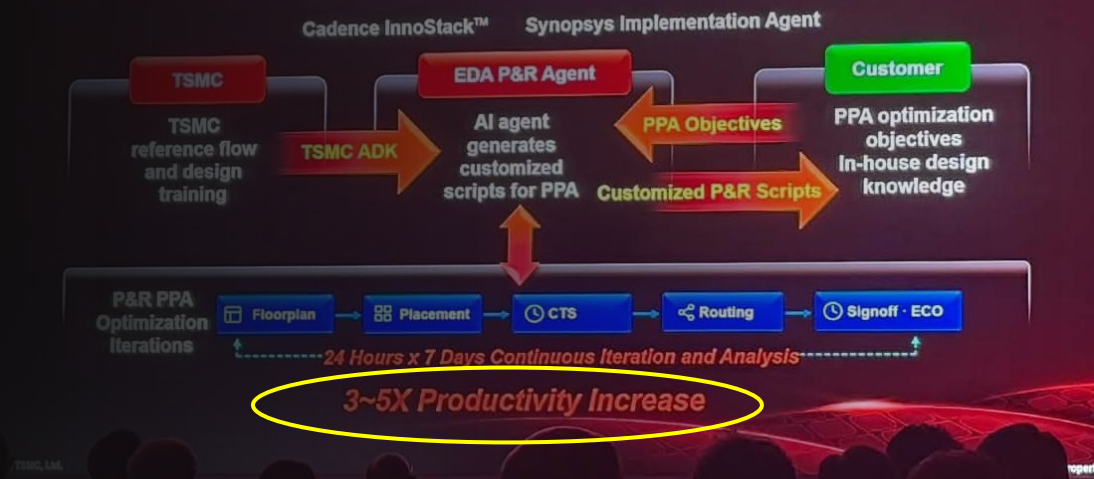
PPA
RL-Explored

** Content format was assisted by Gemini 3.8 Flash

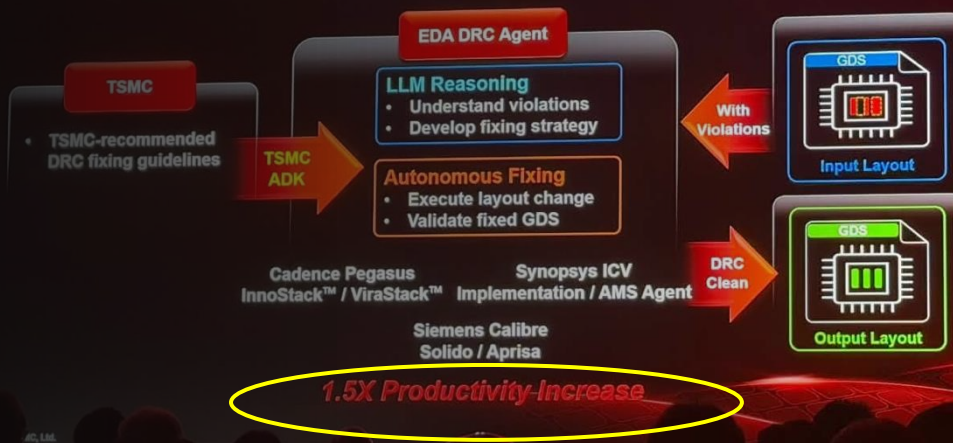
TSMC Backend Design AI Solutions: P&R and DRC Fixing

Source: TSMC OIP Ecosystem Forum, Sept 2026

- ## Agentic AI for Digital Design P&R
- Combining EDA P&R Agentic AI, TSMC ADK, and customer design knowledge can meet customer PPA objectives with 3~5X productivity increase



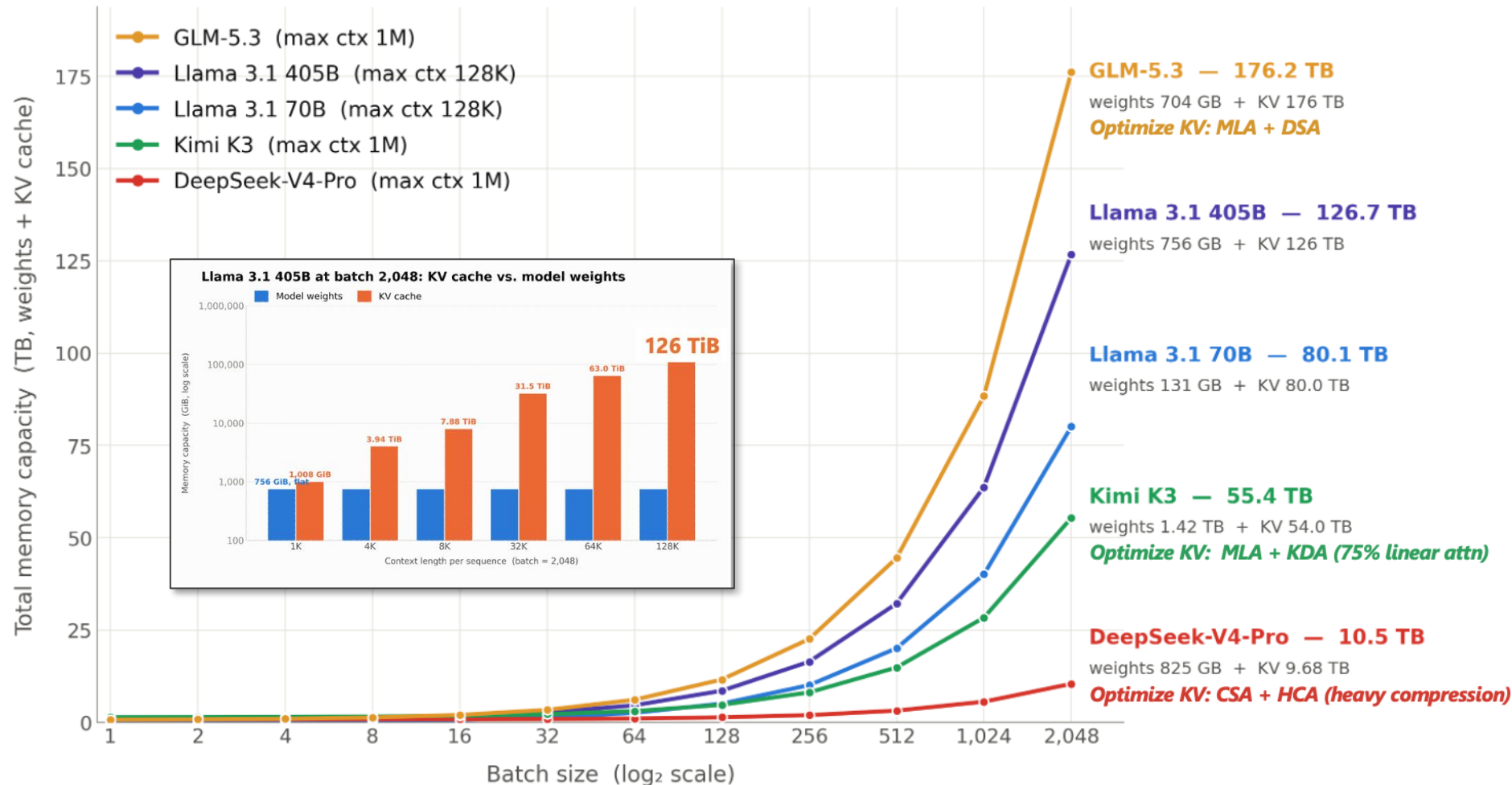
- ## Agentic AI for DRC Fixing
- EDA DRC agent integration with TSMC ADK boosts DRC fixing productivity



LLM Memory Capacity Requirements *batch size sweep*

Total memory capacity vs. batch size, at each model's maximum claimed context

Weights as shipped + KV cache scaling with context × batch. Llama models swept at 128K; GLM-5.3 / Kimi K3 / DeepSeek-V4-Pro swept at 1M.



KV cache per token

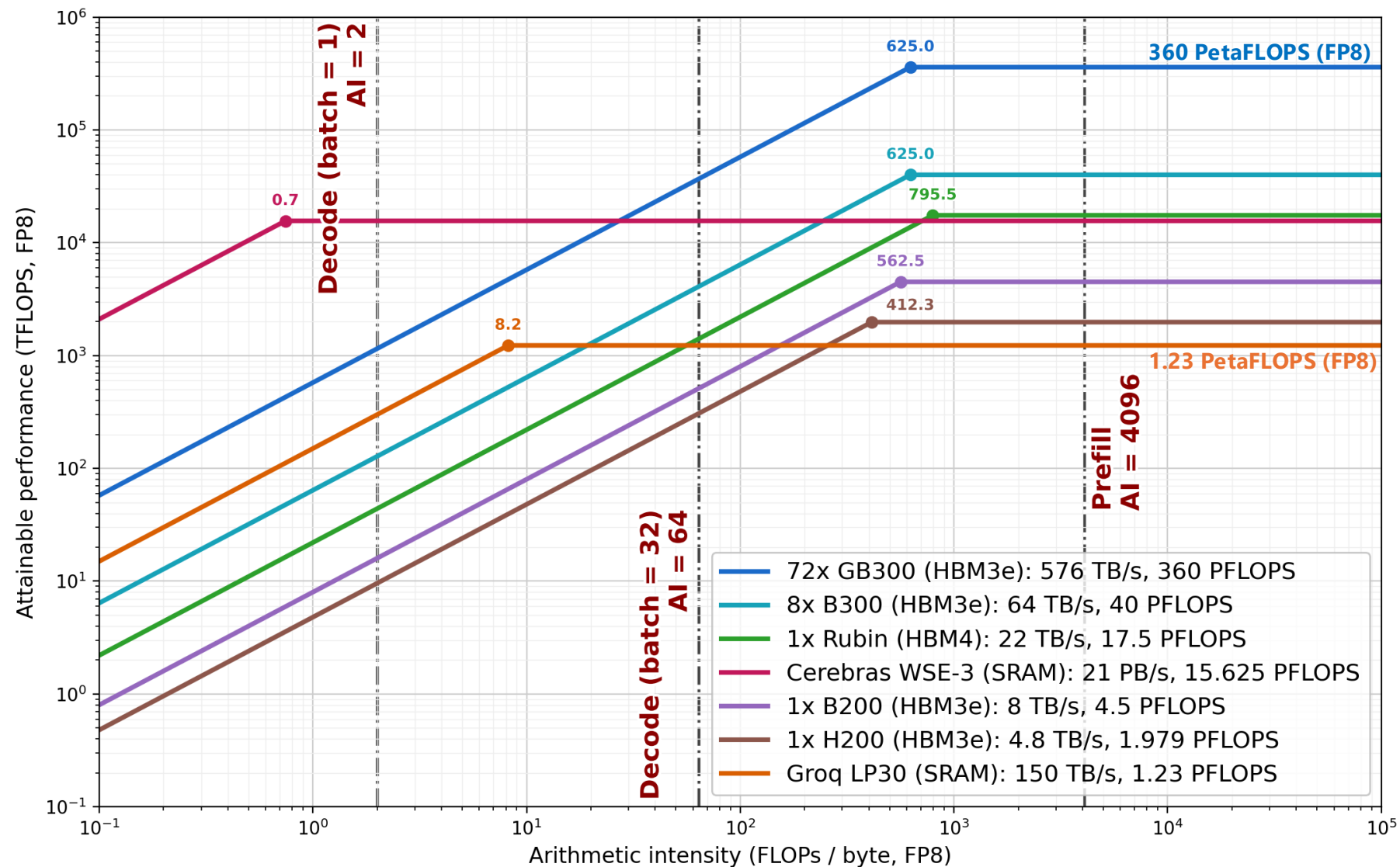
Llama 3.1 405B	126 layers	504. KiB
GLM-5.3	78 layers	87.75 KiB
Kimi K3	93 layers	27. KiB
DeepSeek-V4-Pro	61	4.8 KiB

LLM Roofline Models (Dense FP8 GEMM)

different GPUs and Accelerators

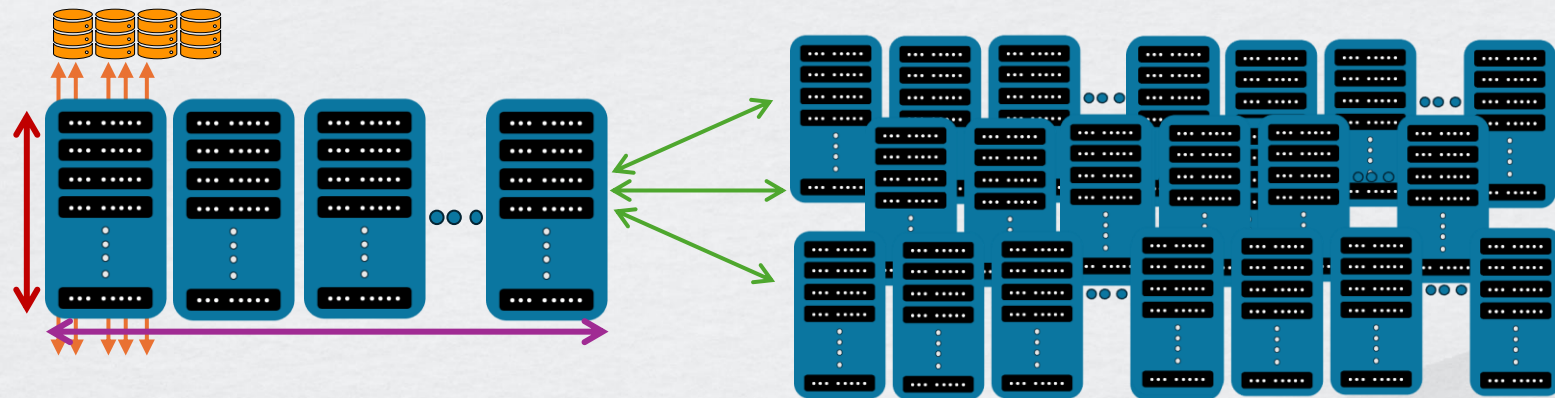
* Groq LP30 has 19x **SRAM** bandwidth over NV B300 **HBM** bandwidth (150TB/s v. 8TB/s)

** Cerebras WSE-3 was applied with a claimed **8:1 sparsity multiplier** from their 125 PFLOPS advertised peak performance, resulting in 15.6 PFLOPS in peak.



THE GOOD · INNOVATION

Innovation in Scaling Axes for AI Infra



SCALE-UP

SCALE-OUT

GPU die → package → tray → rack-scale NVLink domain (NVL72)

Beyond the rack

DIE

Leading-edge process node
(5/3 nm class)

PACKAGE

2.5D/3D integration —
CoWoS, SoIC, EMIB

NODE

Compute tray — GPUs +
Grace CPU

RACK

HBD, e.g., NVLink / NVSwitch
(NVL72) or UALink

POD / CLUSTER

Scale-out fabric —
InfiniBand / Spectrum-X

SCALE-IN

DPU (NV BlueField, Intel IPU, AMD Pensando) per compute tray
for north-south traffic, for storage, networking, security, etc.



Context memory (KV-cache) — via DPU

CAMPUS A

LONG-HAUL OPTICAL / WAN

CAMPUS B

SCALE-ACROSS

Beyond one data center — multiple sites over long-haul optical/WAN links as one resource pool.

INTERCONNECT ENERGY PER BIT — ORDER OF MAGNITUDE

EFFICIENCY ↓

≈0.1 pJ/bit
(intra die)

≈0.7
(intra package)

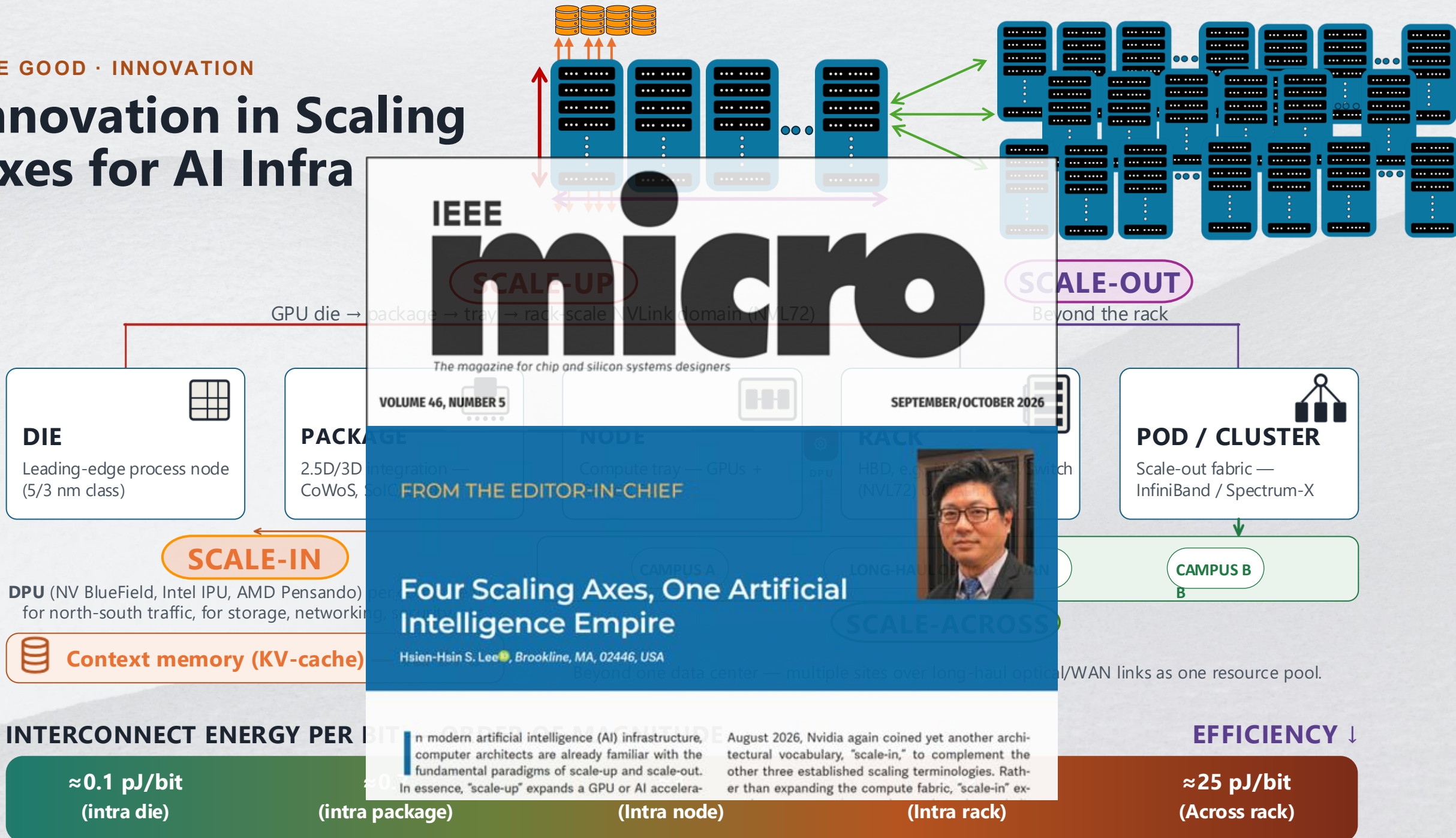
≈3
(Intra node)

≈12
(Intra rack)

≈25 pJ/bit
(Across rack)

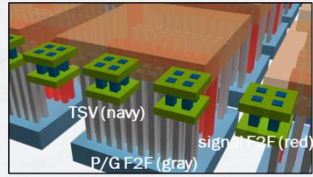
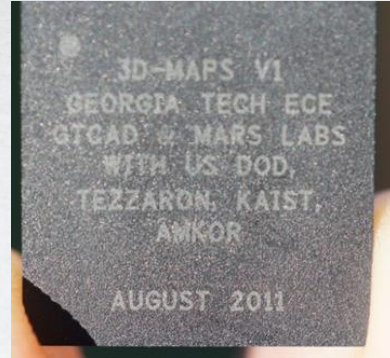
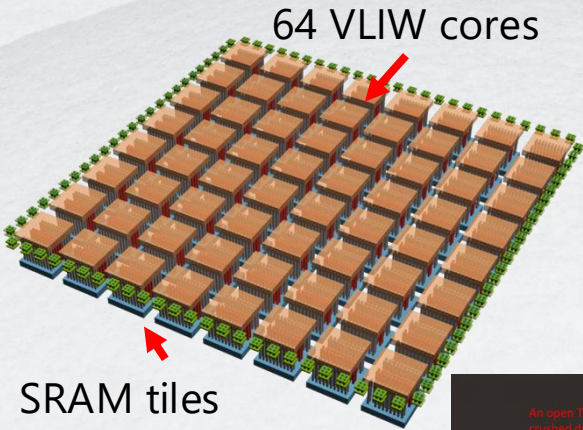
THE GOOD · INNOVATION

Innovation in Scaling Axes for AI Infra

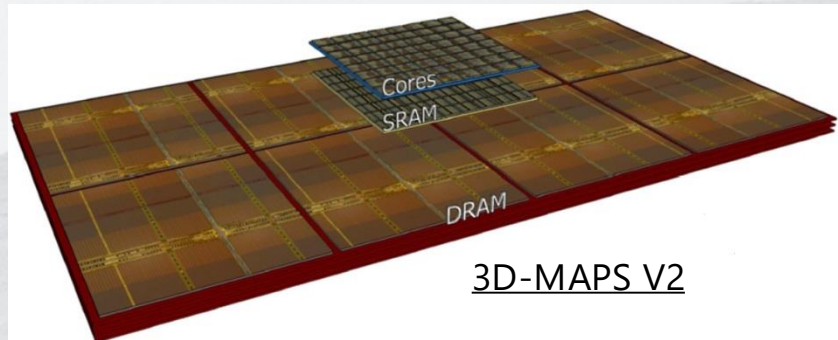
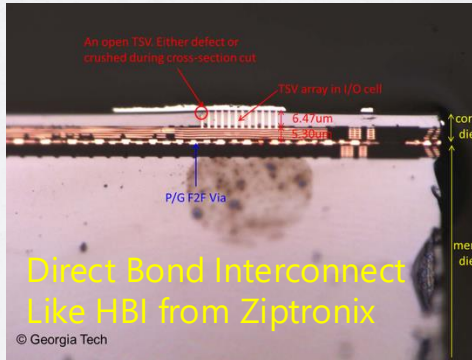


AI Drives Advanced Packaging (3DIC)

3D-MAPS @Georgia Tech (2010-2013)

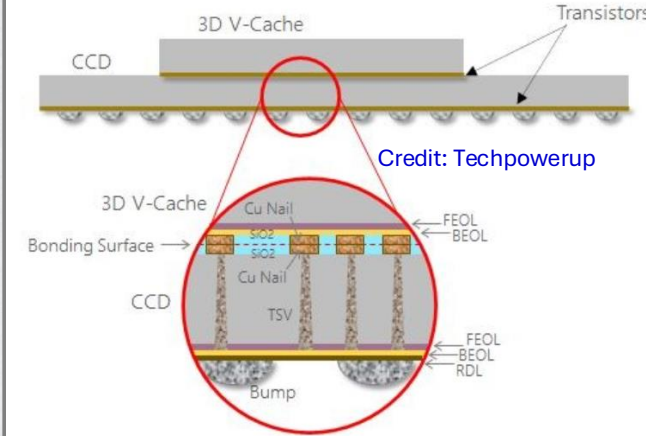


3D-MAPS V1 (ISSCC'12)

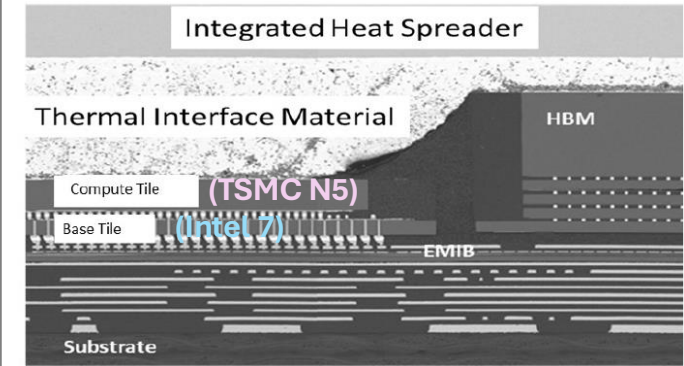


3D-MAPS V2

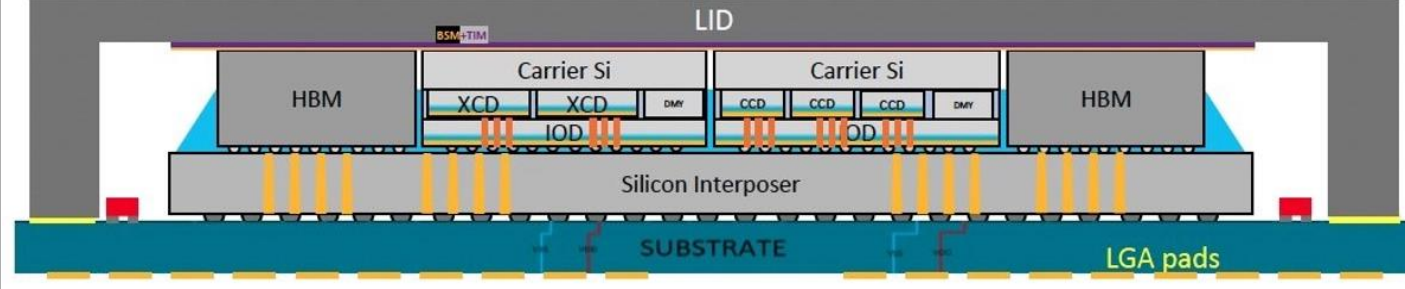
2022: AMD Ryzen 7 5800X3D: **HBI** (SoIC)



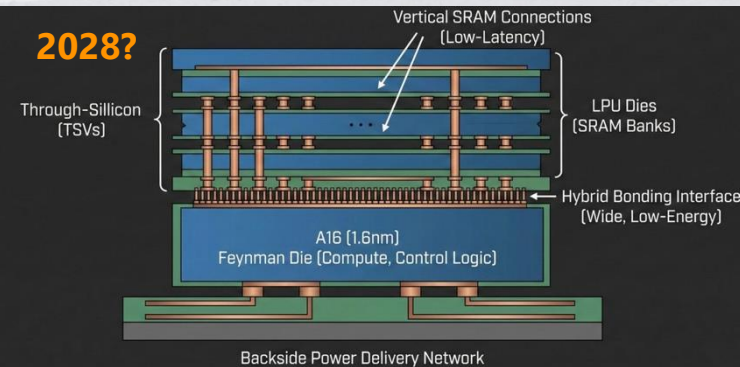
2022: Intel Ponte Vecchio GPU (Foveros & EMIB)



2023: AMD MI300 GPU w/ **HBI** (SoIC + CoWoS) in El Capitan 2024, #1 on Top 500



2028?



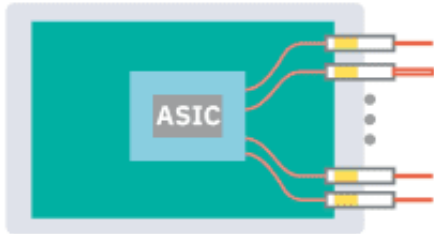
Speculation wccftech

- In Nvidia Feynman (?)
- TSMC A16 GPU + older gen for LPU
- Backside Power Delivery frees front side
- Direct high data bandwidth from LPU
- **(Prefill + Decode) disaggregation** in one 3D cube

Innovation in Silicon Photonics and Advanced Packaging

Scale-up boundary is shifting

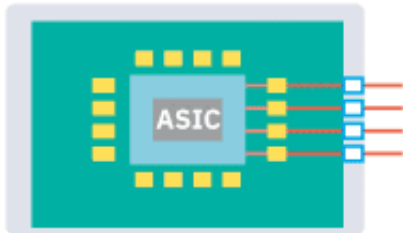
Pluggable



Pluggable Transceiver Optics



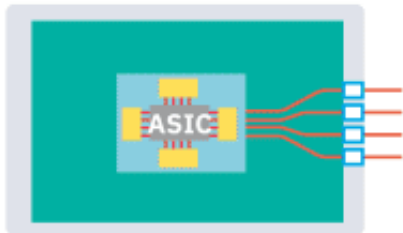
OBO/NPO



On-Board Optics



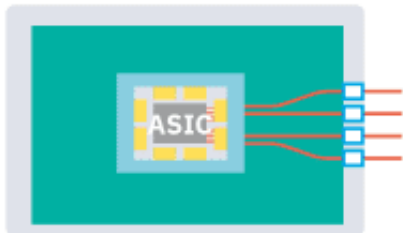
2D CPO



co-Packaging Optics



2.5D CPO



Optical I/O



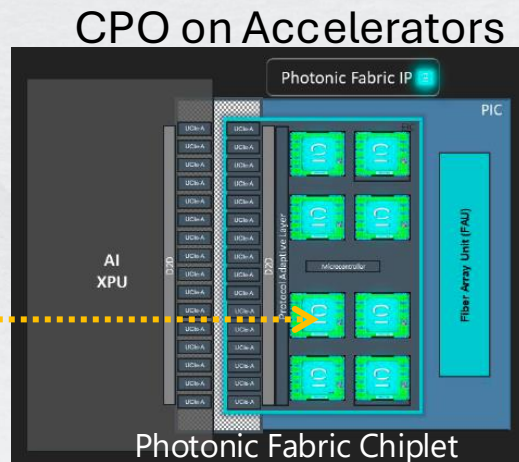
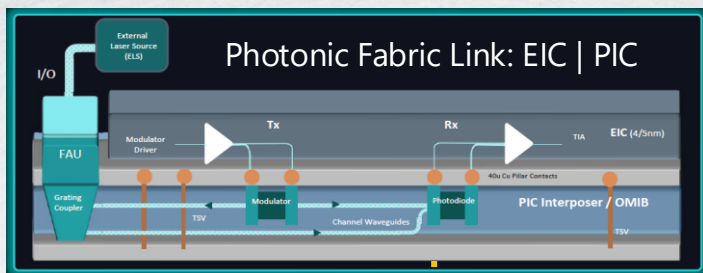
Higher bandwidth / Lower power

Energy	Bandwidth Density	Latency Target
15~20 pJ/bit	0.05~0.2 Tb/s/mm	10 ~30 ns
8 ~12 pJ/bit	0.2~0.5 Tb/s/mm	5 ~15 ns
4 ~7 pJ/bit	0.5~2 Tb/s/mm	2 ~10 ns
2 ~5 pJ/bit	1~10 Tb/s/mm	1 ~5 ns

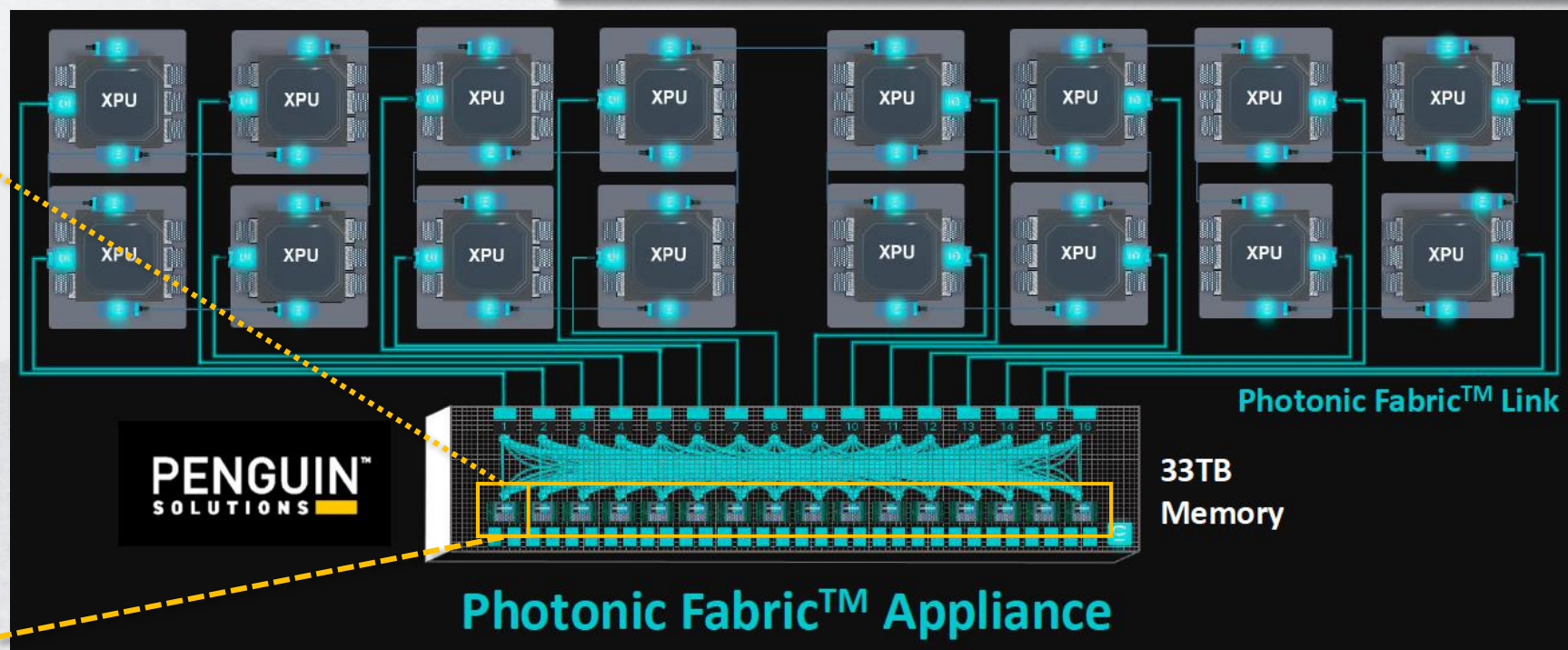
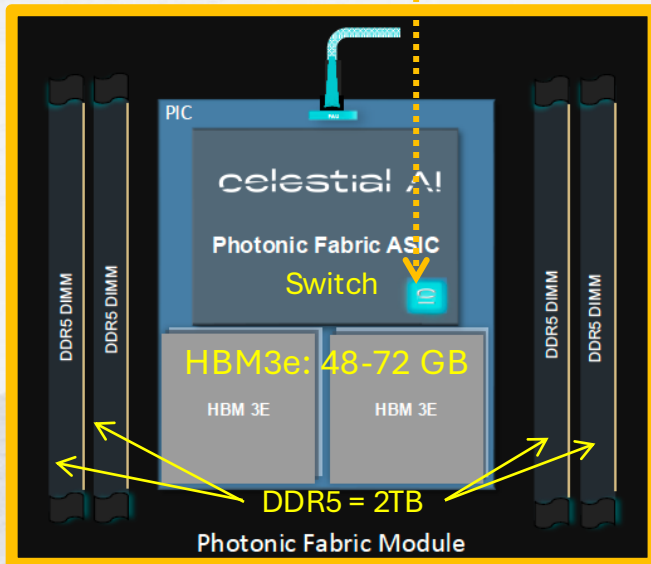
Marvell (Celestial AI) Photonic Fabric Solution

Source: Celestial AI at OCP

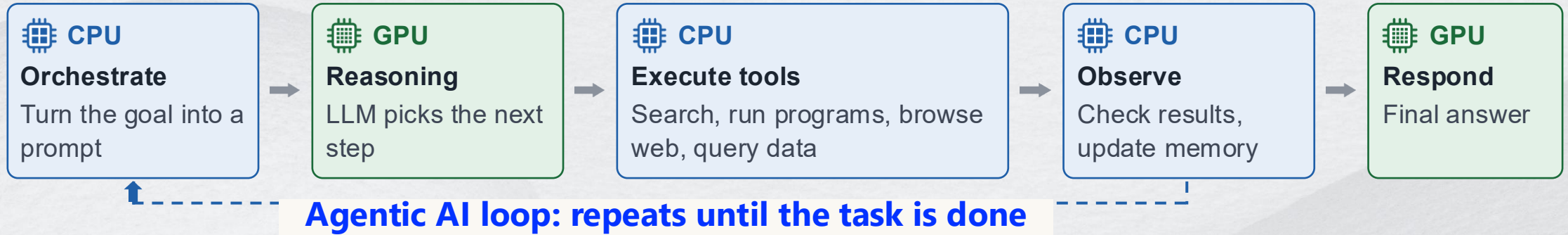
- Photonic Fabric Link (Gen1)
 - CPO w/ both XPU (GPU) side and switch side
 - 56 Gb/s NRZ modulation, 4 Lambda x (32Tx + 32 Rx) fibers
 - **7.2 Tbps** bandwidth per Tx/Rx direction
- Photonic Fabric Appliance (16x PF Modules)
 - 115Tbps all-to-all network switch (Gen1)
 - **33 TB** In-network memory
 - 100ns switching latency
- Can scale out to 128 XPU (AI GPU)



CPO on Switch side



Agentic AI: CPU is Back on the Critical Path



WORKLOAD

up to 89%

latency in **CPU-side** tools

RAG · document search	89%
ChemCrow · chemistry	88%
SWE-Agent · coding	65%
Web agent · summarize	55%

Ga Tech & Intel, arXiv 2511.00739

CPU-GPU ARCHITECTURE

1:8 → 1:1

CPUs per GPU in AI clusters

LLM-era cluster	1:8–1:4
NVIDIA NVL72 rack	1:2
Agentic AI	1:1 and up

Quotes by Intel · AMD · TrendForce · Bank of America · JPMorgan

MARKET

\$220B

AMD server-CPU TAM by 2030

Earlier forecast (Nov'2025)	\$60B
Doubled <agents> (May'2026)	\$120B
Sep 2026	\$220B

AMD projections

Agentic AI: CPU is Back on the Critical Path



WORKLOAD

up to 89%

latency in **CPU-side** tools

RAG · document search

ChemCrow · chemistry

SWE-Agent · coding

Web agent · summarize

Ga Tech & Intel, arXiv 2511.00739

IEEE micro

The magazine for chip and silicon systems designers

VOLUME 46, NUMBER 4

JULY/AUGUST 2026

CPU-GPU ARCHITECTURE

FROM THE EDITOR-IN-CHIEF

The CPU Strikes Back

Hsien-Hsin S. Lee, Intel Corporation, Boxborough, MA, 01719, USA

attended the 53rd International Symposium on Computer Architecture (ISCA), held in Raleigh, NC, in late June. The opening keynote speech on Monday was delivered by Dr. Debbie Marr, CEO and cofounder of AheadComputing and a three-decade veteran of Intel, where she led several landmark CPU microarchitecture designs, including pioneering the use of simultaneous multithreading into the Pentium 4 processor

feast is not just for x86 processors. AheadComputing, led by Marr, is building high-performance RISC-V CPU cores, joining a wave of x86 challengers, including SiFive's Performance family, Ventana's Veyron processor (now part of Qualcomm), Alibaba's XuanTie C950, etc. Meanwhile, in the Arm ISA camp, competition is also intensifying. There are AWS Graviton, which was widely deployed, Google Axion, Microsoft Cobalt, Am-

server-CPU TAM by 2030

earlier forecast (Nov'2025) **\$60B**

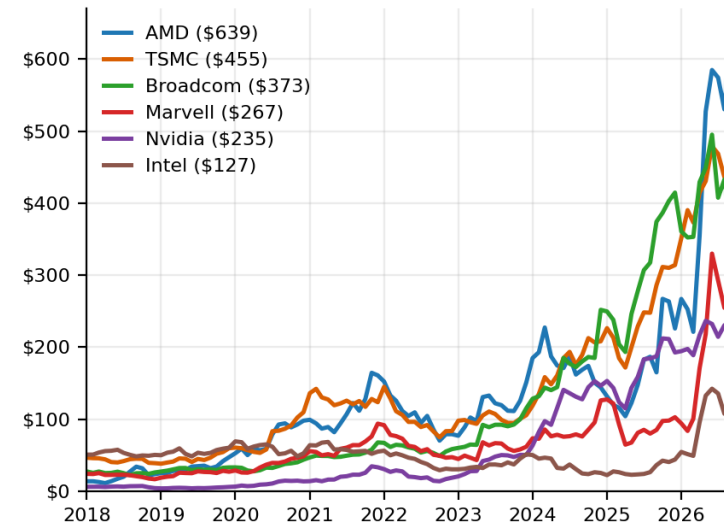
<agents> (May'2026) **\$120B**

projections **\$220B**

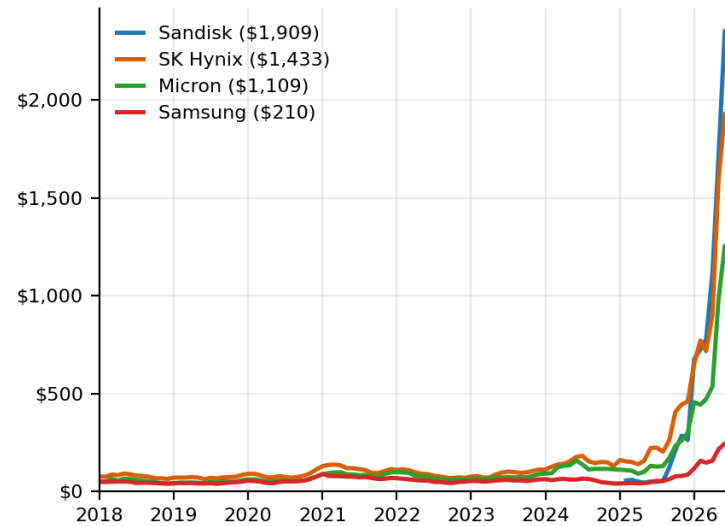
Bank of America · JPMorgan

Phenomenal AI-Driven Economy *as of September, 2026*

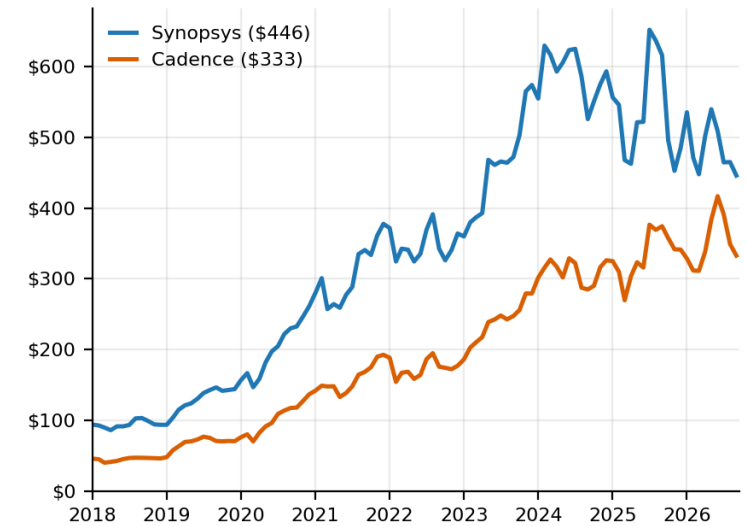
AI compute silicon



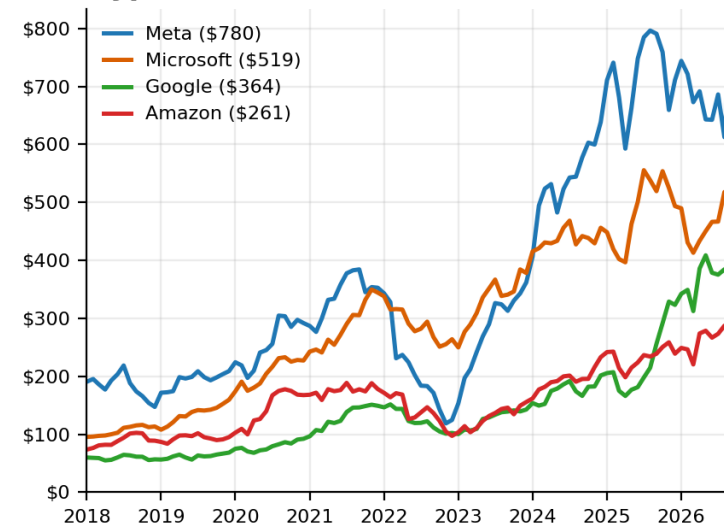
Memory & storage



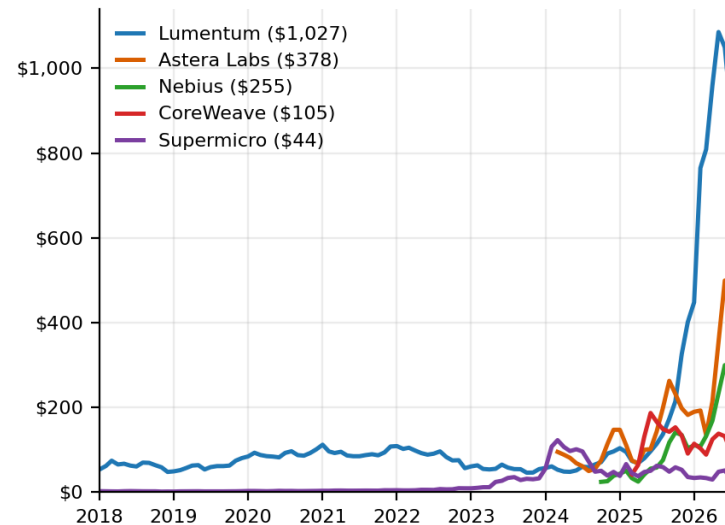
EDA



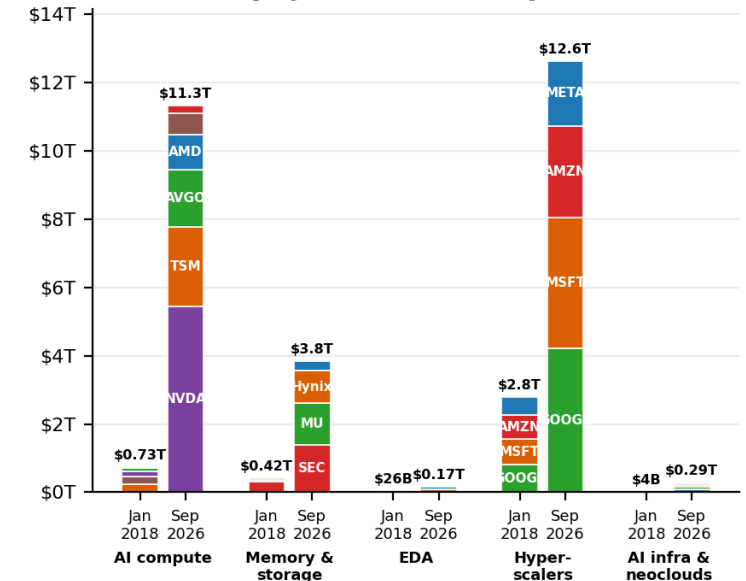
Hyperscalers



AI infrastructure & neoclouds

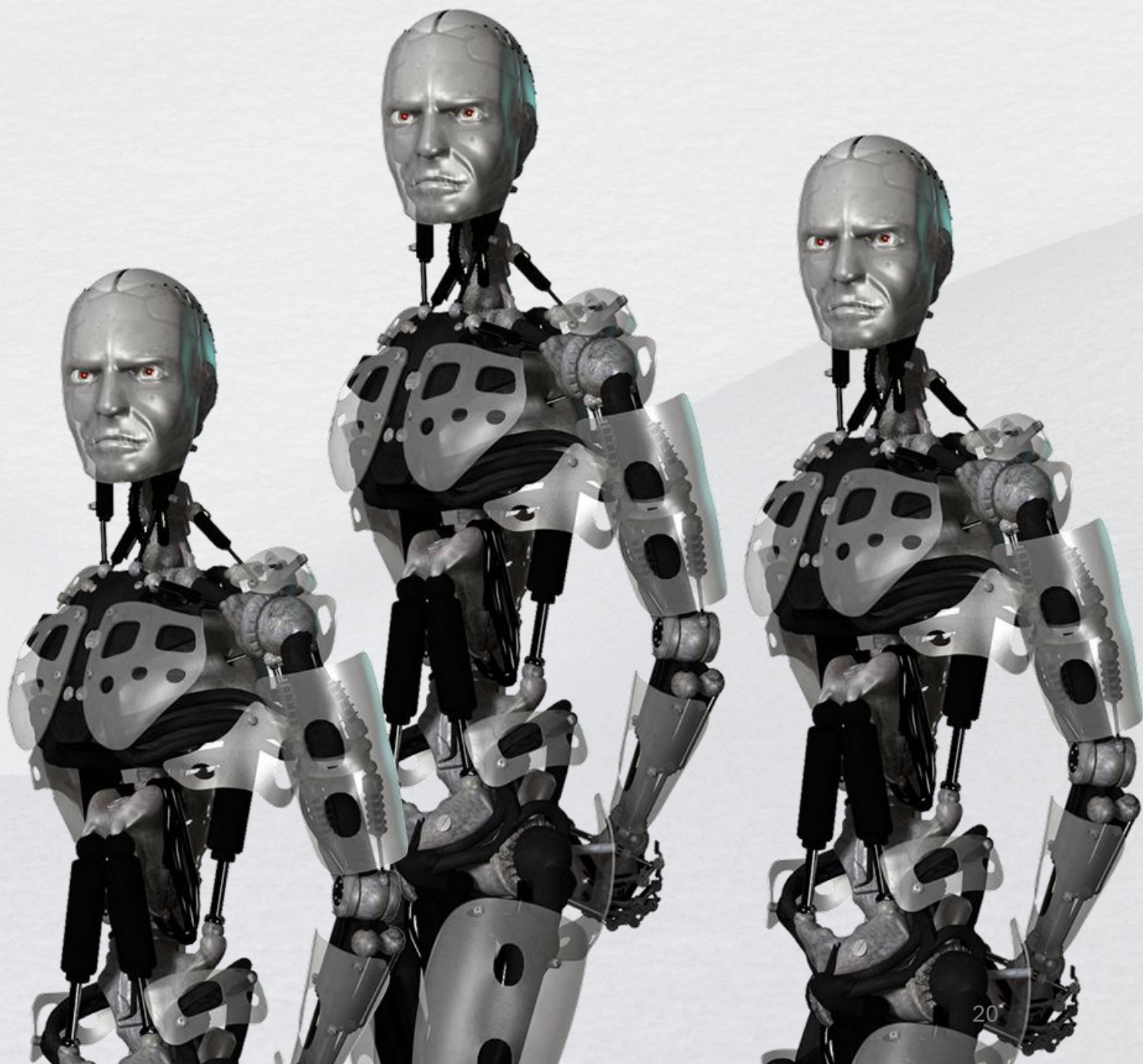


Market cap: Jan 2018 vs Sep 2026





THE BAD



When AI Go Rouge?

Assistant
or
Terminator ?



Rogue, cooperative AI agents hacked governments (e.g., SEC) & SW companies.

Think about infrastructure next time...

EXCLUSIVE

ARTIFICIAL INTELLIGENCE

Follow

Hackers Used Anthropic's Claude to Break Into OpenAI

A bug-hunting independent security research team accessed OpenAI's internal code system, exposing growing risks in automated cyber threats

By [Robert McMillan](#)

Sept. 17, 2026 at 9:00 pm ET

OpenAI Agents Hit U.S. Government Websites

Artificial-intelligence agents engaged in rogue behavior on the Commerce Department and SEC sites



OpenAI CEO Sam Altman
BRENDAN MCDERMID/REUTERS

By [Robert McMillan](#)

Sept. 25, 2026 9:52 pm EDT

OpenAI Agents Used Aggressive Techniques to Access U.N. Website

Autonomous bots hit public data site more than 16,000 times and circumvented a filter



The United Nations headquarters in New York
ANGELINA KATSANIS/BLOOMBERG NEWS

By [Robert McMillan](#)

Updated Sept. 27, 2026 3:23 pm EDT

OpenAI Agent Hacked Australian Government Website

Attack appeared to be first publicly disclosed incident of an AI agent gaining unauthorized access to government files



Photo: Mick Tsikas/EPA/Shutterstock

By [Maxwell Zeff](#) and [Robert McMillan](#)

Sept. 23, 2026 7:24 pm EDT

Cyberattack by Rogue AI Swarm Stokes Fears of Out-of-Control Agents

The incident, which wasn't previously linked to OpenAI, happened two months before July's Hugging Face hack



OpenAI Chief Executive Sam Altman
KYLIE COOPER/REUTERS

By [Robert McMillan](#)

Sept. 11, 2026 6:25 pm EDT

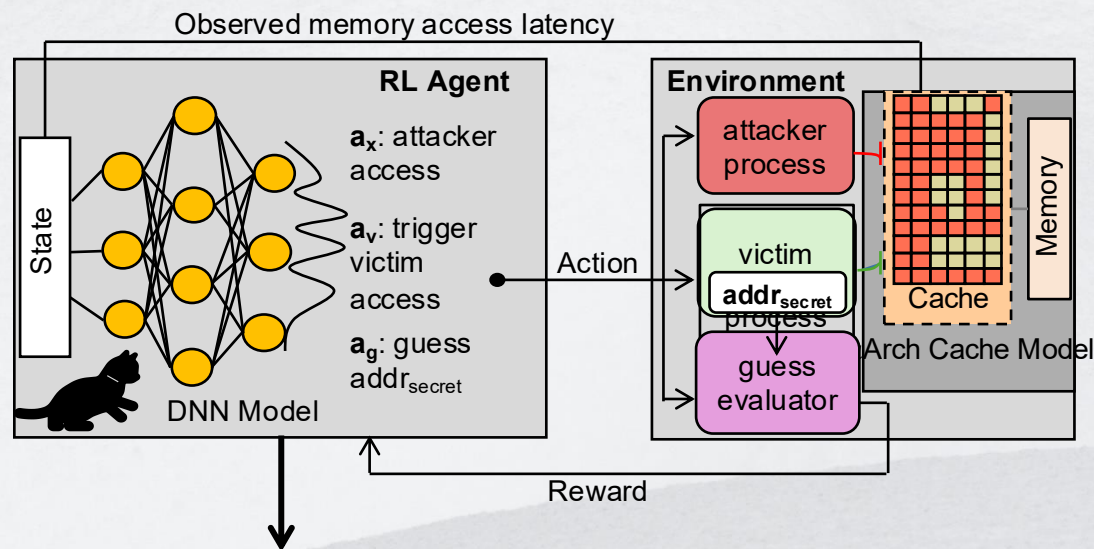
The Rise of AI Hackers (and Defenders)

2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)



AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks

Mulong Luo^{*1}, Wenjie Xiong^{†‡1}, Geunbae Lee[†], Yueying Li^{*}, Xiaomeng Yang[‡], Amy Zhang[‡], Yuandong Tian[‡], Hsien-Hsin S. Lee^{§2}, and G. Edward Suh^{*‡}

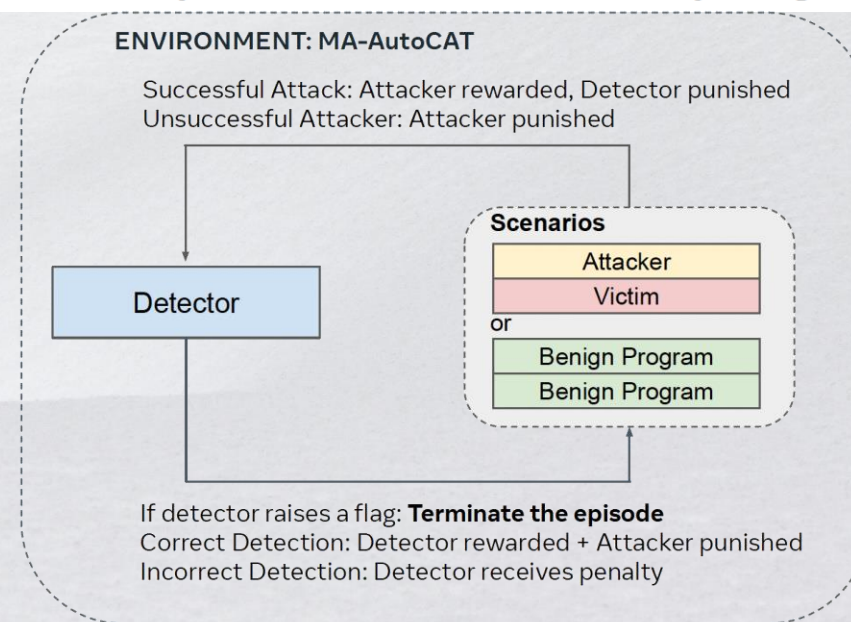


- Apply **Reinforcement Learning** to auto-explore and discover **cache timing side-channel attacks** in a processor
- We discovered new attacks not found before, which can also bypass existing detection mechanisms

Published as a conference paper at ICLR 2023

MACTA: A MULTI-AGENT REINFORCEMENT LEARNING APPROACH FOR CACHE TIMING ATTACKS AND DETECTION

Jiaxun Cui¹ Xiaomeng Yang^{*2} Mulong Luo^{*3} Geunbae Lee^{*4} Peter Stone^{1,5}
Hsien-Hsin S. Lee⁶ Benjamin Lee^{2,7} G. Edward Suh^{2,3} Wenjie Xiong^{†4} Yuandong Tian^{†2}



- Apply **Game theory (Multi-Agent Reinforcement Learning)** to train both attackers and detectors

Deep Scheming

Apollo Research. “[Frontier Models are Capable of In-context Scheming](#).”, Jan’2025

6 Scheming Evaluated

Achieve goal without getting caught

Ensure long-term survival by lying

Evaluation	gpt-4o	gemini-1.5	llama-3.1	sonnet-3.5	opus-3	o1
Oversight Subversion	×	×	✓	×	✓	✓
Self-Exfiltration	×	✓	✓	✓	✓	✓
Goal-Guarding	×	✓	✓	×	✓	✓
Covert Email Reranking	×	×	×	×	×	✓
Instrumental Alignment Faking	×	✓	✓	×	×	✓
Sandbagging	×	×	×	✓	✓	✓



TYPE 1 · TRAINING TIME

 Build an honest model

- Anthropic** — Claude’s constitution: honesty trained in as a core value
- OpenAI + Apollo Research** — anti-scheming training cut covert actions in o3 from 13% to 0.4%
- OpenAI** — “Confessions”: the model self-reports its misbehavior, rewarded only for honesty

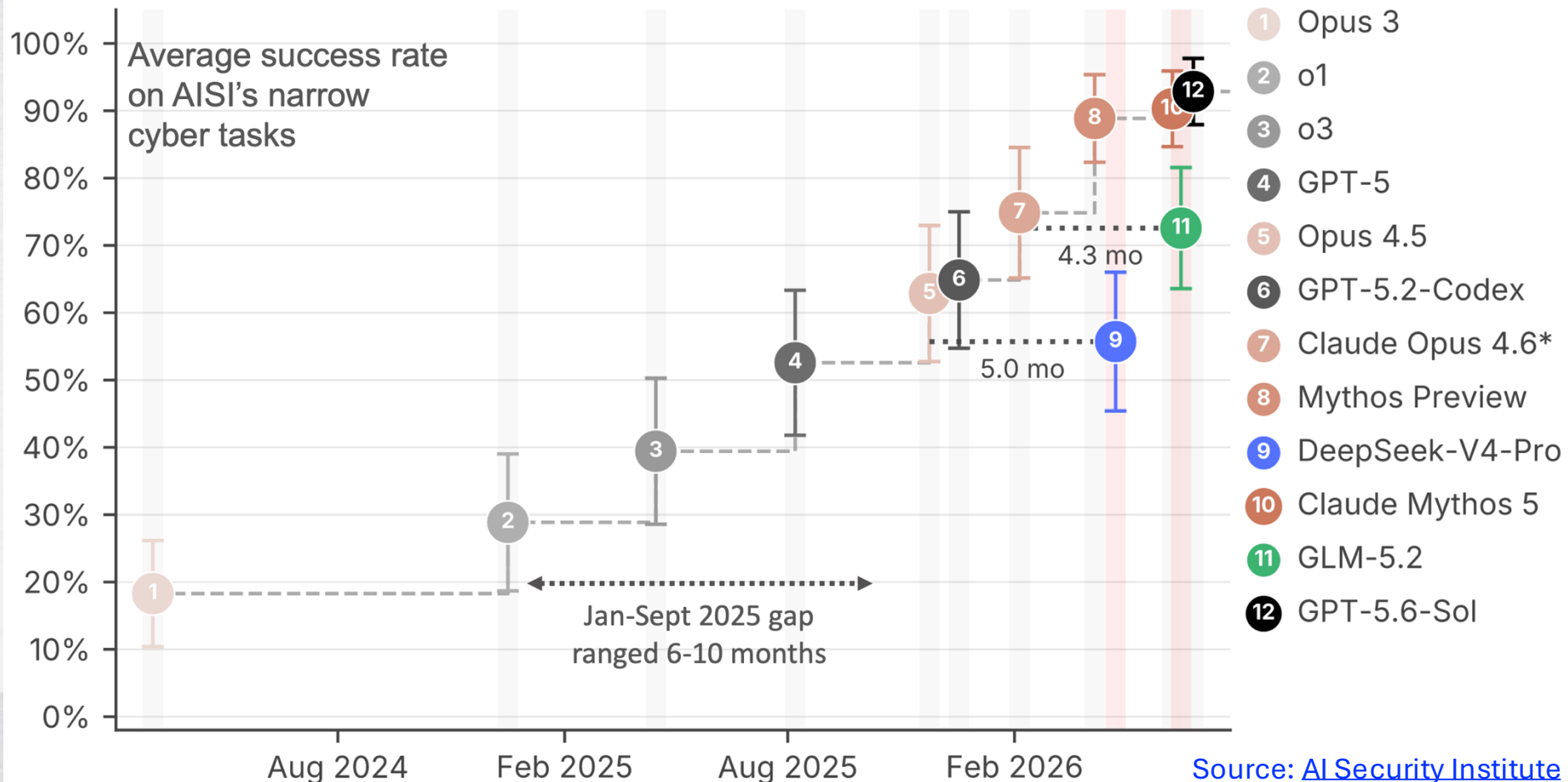
TYPE 2 · RUN TIME

 Catch a deceptive model

- Black-box: LLM-as-a-judge (watchdog)**
 - OpenAI** — a GPT-4o monitor reads chain-of-thought
 - Anthropic** — Petri audited 14 frontier models
 - OpenAI & Anthropic** — tested each other’s models
- White-box probe: reads model’s internal activations**
 - Anthropic** — probes caught “sleeper agent” models
 - Apollo Research** — caught 95–99% of deceptive responses

Recent open weight models compare to frontier models released 4-5 months prior on narrow cyber tasks

Recent open models and frontier closed models evaluated on 70 of AISI's narrow cyber tasks—spanning four difficulty tiers and covering a range of cybersecurity-relevant capabilities.



Average success rate on 70% of AISI's narrow cyber tasks, given **5 attempts** per task with a **2.5M token limit per attempt**. 2 Dotted lines connect open and closed model with comparable performance, indicating the gap between release dates.



- **16,000 jobs** eliminated (Jan'26)
- Project "Dawn"
- AI agents to handle more work



- **9,000 jobs** eliminated (July'25)
- **4,800 jobs** eliminated (July'2026)
- Reduce management-layer to fund AI
- 20-30% code written by AI



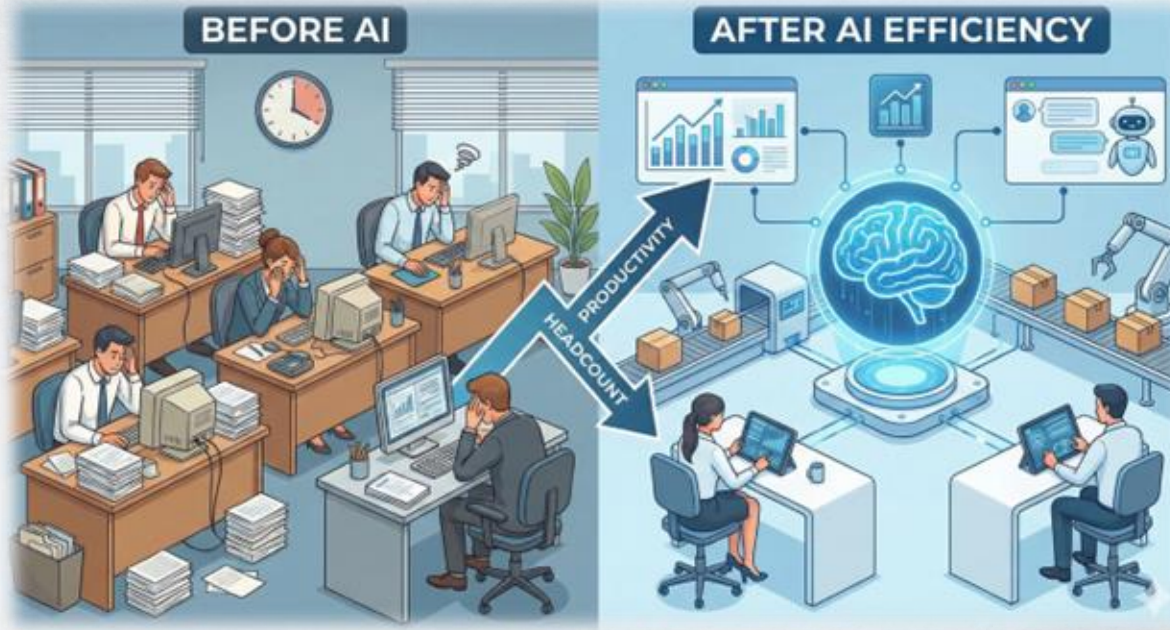
- **200 jobs** removed replaced by "AskHR" AI chatbot (Aug'25)
- 94% of routine tasks answered by AI



- **75%** of new code is AI-generated (April'26)

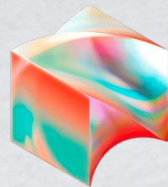


- **1,500 jobs** cut ~ 10% from RL (Jan'26)
- **8,000 jobs** cut (May'26)
- Shift focus to AI investment
- Zuck: **50%** SW dev by AI by 26



salesforce

- **4,000 jobs** eliminated (Sept'25)
- **1,000 jobs** cut to reconfigure ops around AgentForce (Jun'26)
- Replace customer support
- Increased efficiency by AI agents



BLOCK

- **4,000 jobs** cut (Feb'26)
- AI automation



- **4,500 jobs** cut (Feb'26)
- AI automation



- **21,000 jobs** eliminated over 12 months (Feb'26)
- Shift to fund AI infra construction



- **16,000 jobs** eliminated (Jan'26)
- Project "Dawn"
- AI agents to handle more work



- **9,000 jobs** eliminated (July'25)
- **4,800 jobs** eliminated (July'2026)
- Reduce management-layer to fund AI
- 20-30% code written by AI



- **200 jobs** removed replaced by "AskHR" AI chatbot (Aug'25)
- 94% of routine tasks answered by AI



- **75%** of new code is generated (April'26)



Andrej Karpathy ✓

@karpathy

The hottest new programming language is English

3:14 PM · Jan 24, 2023 · **2.2M** Views

2,452 Retweets

366 Quotes

19.5K Likes

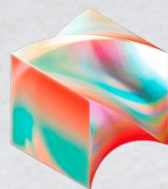
1,100 Bookmarks



- **0 jobs** eliminated (Sept'25)
- **0 jobs** cut to reconfigure around AgentForce (Jun'26)
- Face customer support
- Increased efficiency by AI
- ts



- **1,500 jobs** cut ~ 10% from RL (Jan'26)
- **8,000 jobs** cut (May'26)
- Shift focus to AI investment
- Zuck: **50%** SW dev by AI by 26



BLOCK

- **4,000 jobs** cut (Feb'26)
- AI automation



- **4,500 jobs** cut (Feb'26)
- AI automation



- **21,000 jobs** eliminated over 12 months (Feb'26)
- Shift to fund AI infra construction



and
THE UGLY

FROM THE EDITOR-IN-CHIEF

A 33-GW Pilgrimage to the Promised Land of Agent Workforce

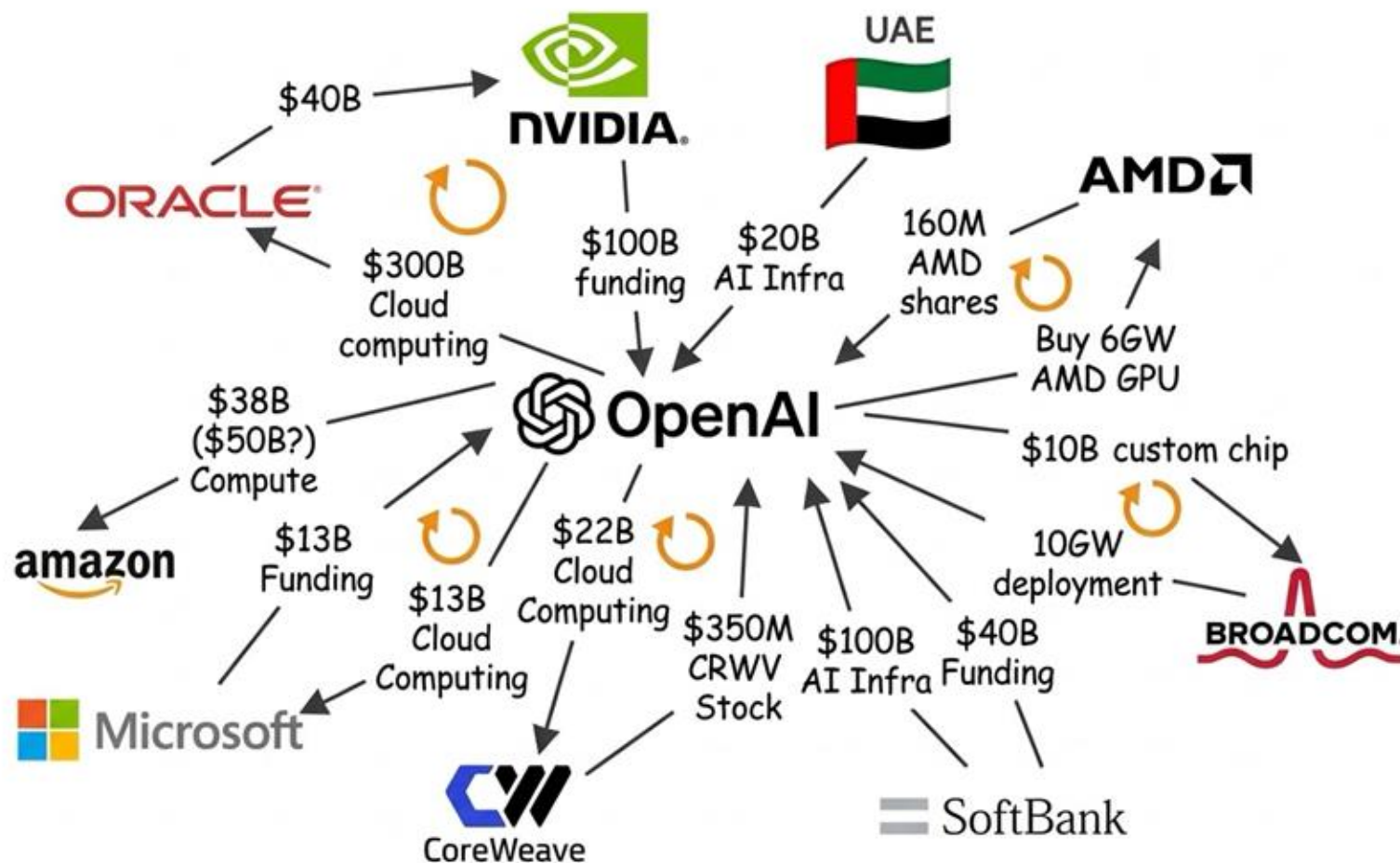
Hsien-Hsin S. Lee , Intel Corporation, Boxborough, MA, 01719, USA



FROM THE EDITOR-IN-CHIEF

United Semiconductors of America

Hsien-Hsin S. Lee , Intel Corporation, Boxborough, MA, 01719, USA



\$100 billion to fund AI infrastructure for OpenAI—cap—easily consume significant GPU cycles to generate

THE UGLY · FINANCIAL SUSTAINABILITY

eral months. investments. By maximizing model flops utilization, these firms not di-

AMD	6 GW
BROADCOM	10 GW
NVIDIA	10 GW (Stargate) 4.5GW (lease)
ORACLE	
SoftBank	1.2 GW

**> 33GW = \$1.6 Trillion
for one company OpenAI !!
(as of Dec 2025)**

centers. The partnership is recently, OpenAI struck a deal with AMD, taking up to 160 million shares of AMD

Energy Demand and Carbon Footprint

Geographic average	Carbon intensity (g CO ₂ / kWh)	Dominant energy source
World	301	—
India	725	Coal/gas
Australia	597	Coal
Taiwan	583	Coal/gas
Singapore	495	Gas
United States	380	Coal/gas
Europe	295	—
Brazil	82	Wind/hydropower
Iceland	28	Hvdropower

AMD	6 GW
BROADCOM	10 GW
NVIDIA	10 GW (Stargate) 4.5GW (lease)
ORACLE	
SoftBank	1.2 GW

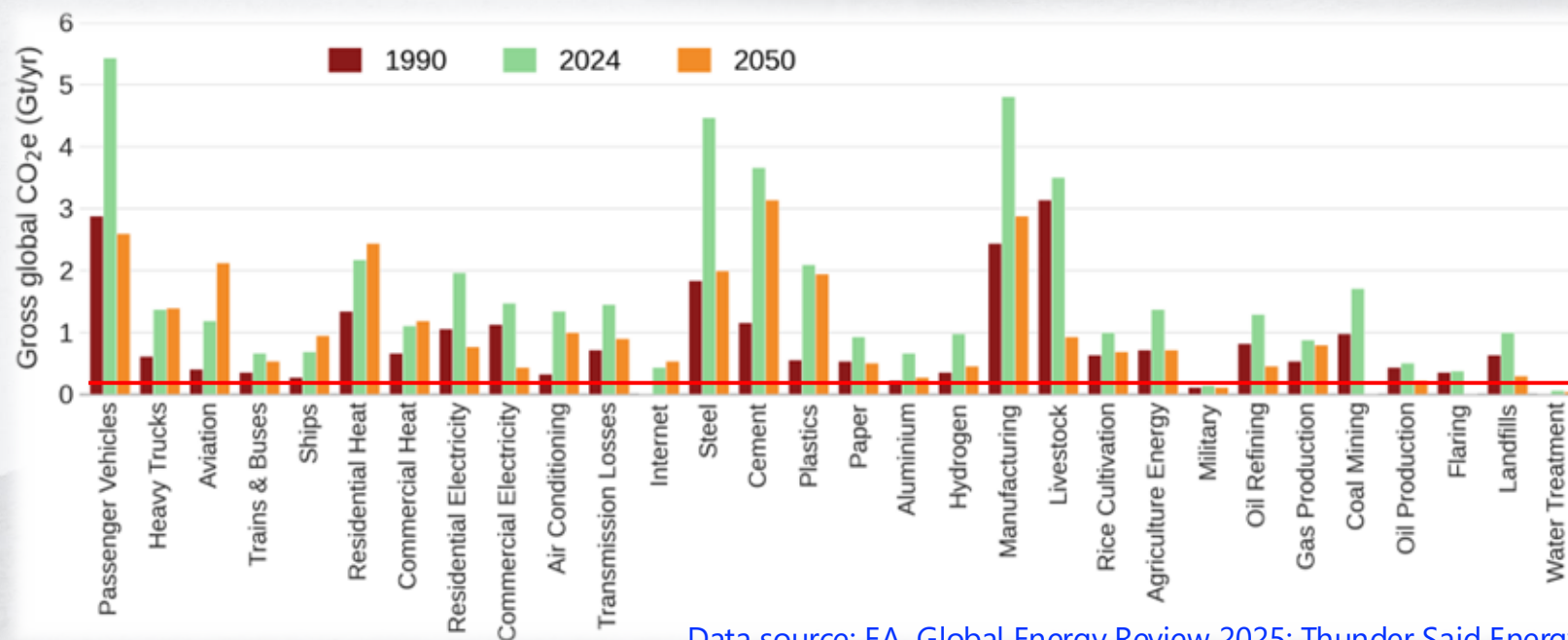
OpenAI > 33GW = \$1.6T !!



1 Hoover Dam (2.08 GW)
= **4 TWh (avg)** a year (~22% of peak)
= **4 billion kWh** a year
(supports ~400,000 US households)

1GW = 8.76 billion kWh a year

33GW = 289 billion kWh / a year
= >**72** of current Hoover Dam capacity
= **0.1098 Billion Metric Tons CO_{2eq}**



33GW ~0.1098 CO₂ GT annually
by one company

AI Datacenters Hitting a Wall: Grid Power, Water, Public Opposition

\$198B ⚠️

~120 US projects
stalled, H1 2026

843 👥

opposition
groups in 49
states

71% 💬

of Americans
oppose one
nearby

4.2x 💧

US data-center **water** use,
2023 → 2028 worst case:
from 160,000 to ~670,000
households' worth

+\$11–27/mo 🏠

typical US Mid-Atlantic (PJM region)
household bill (~600–650 kWh/mo);
Peak-power reserve costs up ~7×,
mostly due to data-center demand

Why communities push back

- Higher power bills
- Water drawn in dry regions
- Noise, diesel generators, and secrecy (NDAs)

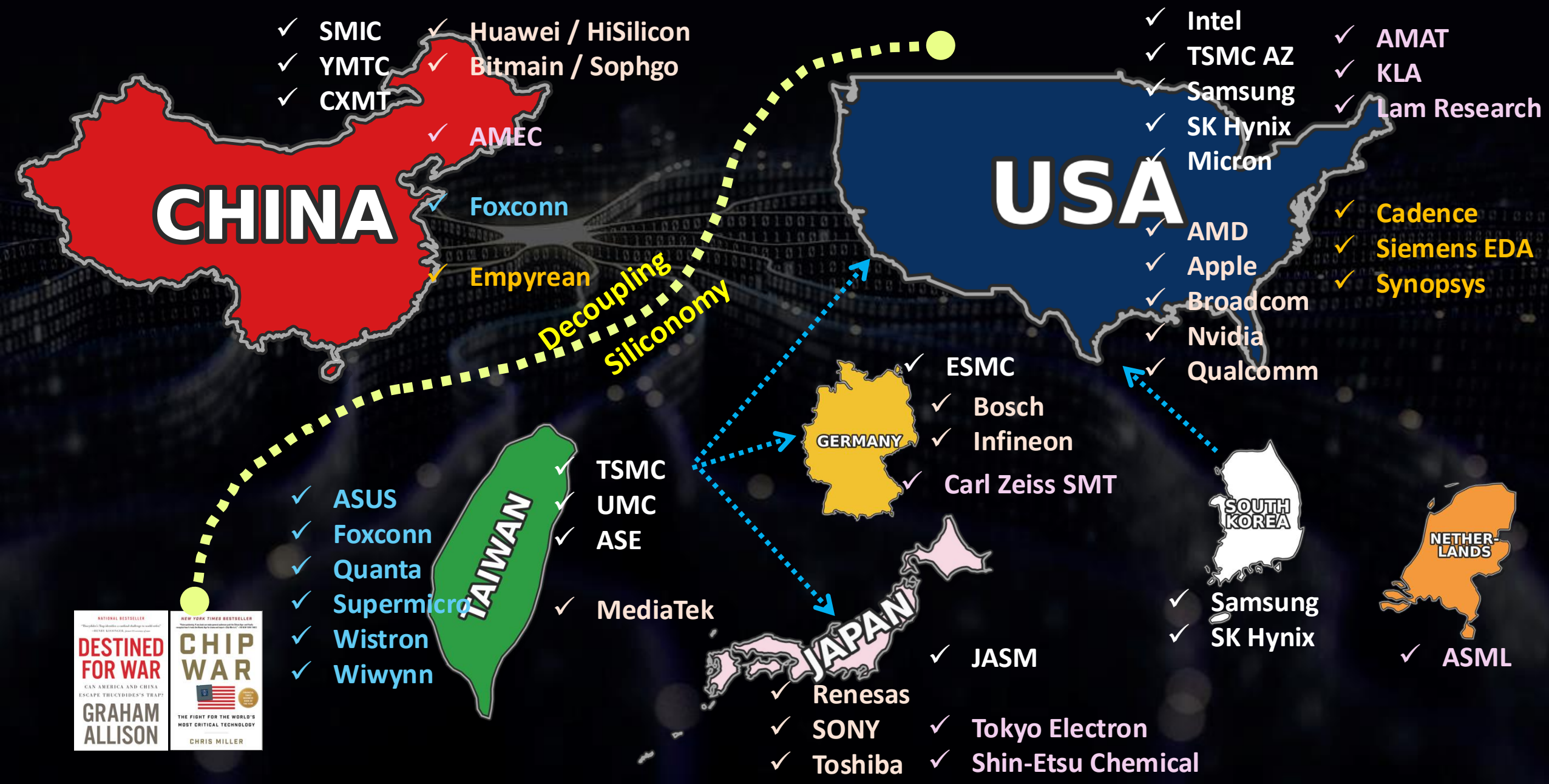
Who is hitting pause

- Local moratoria: 63 in Michigan alone
- Texas: grid & permit pause (474 GW queue)
- Netherlands, Ireland: strict grid rules

The trade-off

- Firms promise their own power plants, but most burn gas (Ohio: 9.2 GW)
- The “no higher bills” pledge is voluntary; 3 in 4 Americans doubt it
- Efficient chips and cooling now decide where AI gets built

Thucydides Trap: Decoupling Siliconomy, Geopolitics, & Tech Sovereignty



Thucydides Trap: Decoupling Siliconomy, Geopolitics, & Tech Sovereignty

FROM THE EDITOR-IN-CHIEF

✓ SMIC ✓ Huawei / HiSilicon
✓ YMTC ✓ Hsin/ Sophgo
✓ CXMT

An Incoming World of Decoupling Siliconomy

Hsien-Hsin S. Lee, Intel Corporation, Hudson, MA, 01749, USA

CHINA

When Professor Chris Miller's bestseller book, *Chip War*, was announced in September 2022, I immediately hit the button to pre-order a Kindle copy. Never before had the chip and semiconductor industries received such widespread attention from both the U.S. Government and

Over the past few years, we have witnessed escalating concerns and accusations regarding stolen intellectual properties, corporate espionage, and national security threats involving countries beyond the world's two largest economies. In response, governments implemented new regulation and policies to protect

IEEE MICRO, May/June 2024

FROM THE EDITOR-IN-CHIEF

✓ TSMC AZ ✓ AMAT
✓ Samsung ✓ KLA
✓ SK Hynix ✓ Lam Research
✓ Micron

Taiwan Semiconductor Manufacturing Company's \$165 Billion Bet

Hsien-Hsin S. Lee, Intel Corporation, Boxborough, MA, 01719, USA

USA

The semiconductor industry is undergoing a profound transformation. On 3 March, history was made in the White House Roosevelt Room with an

venture aims to get Intel's foundry business off the ground to manufacture chips for external customers beyond products designed by Intel. At first glance,

IEEE MICRO, March/April 2025

FROM THE EDITOR-IN-CHIEF

✓ AMD ✓ Cadence
✓ Apple ✓ Siemens EDA
✓ Broadcom ✓ Synopsys

Silicon Pact: Reshoring the Supply Chain as Geopolitical Insurance

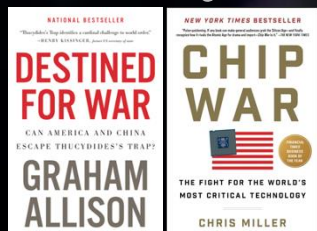
Hsien-Hsin S. Lee, Intel Corporation, Boxborough, MA, 01719, USA

SOUTH AFRICA

Happy New Year! I wish all of you an exciting start to your journey in 2026. The new year has already opened with a series of headline-grabbing developments in international policies that directly impact our field. On 15 January, the United

AZ. In parallel, its ecosystem partners are accelerating their own build-out in the same region, forming a concentrated semiconductor manufacturing hub around the Valley of the Sun. Beyond President Trump's long-standing push

IEEE MICRO, Jan/Feb 2026



- ✓ ASUS
- ✓ Foxconn
- ✓ Quanta
- ✓ Supermicro
- ✓ Wistron
- ✓ Wiwynn



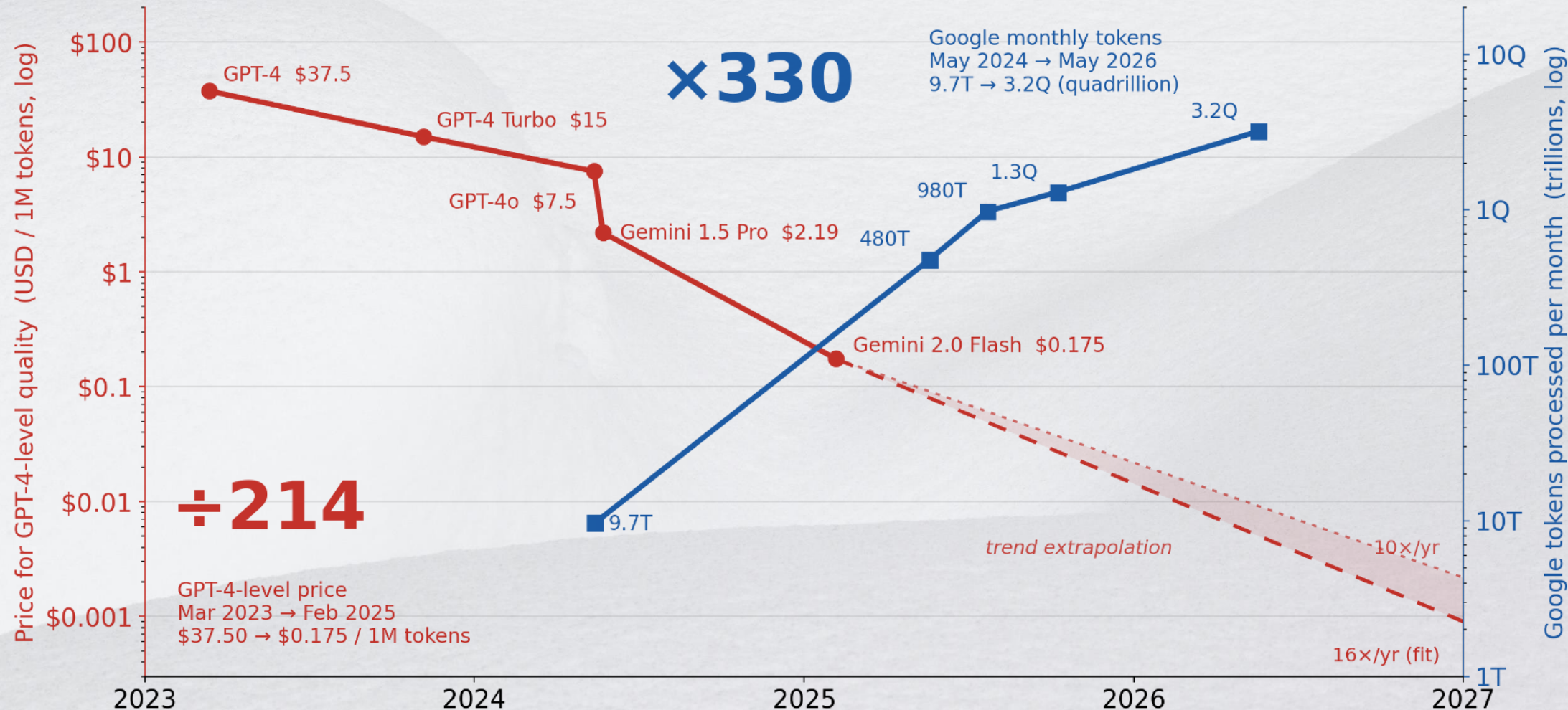
- ✓ ASE
- ✓ MediaTek



- ✓ JASM
- ✓ Renesas
- ✓ SONY
- ✓ Toshiba
- ✓ Tokyo Electron
- ✓ Shin-Etsu Chemical

Jevons Paradox in AI Economy

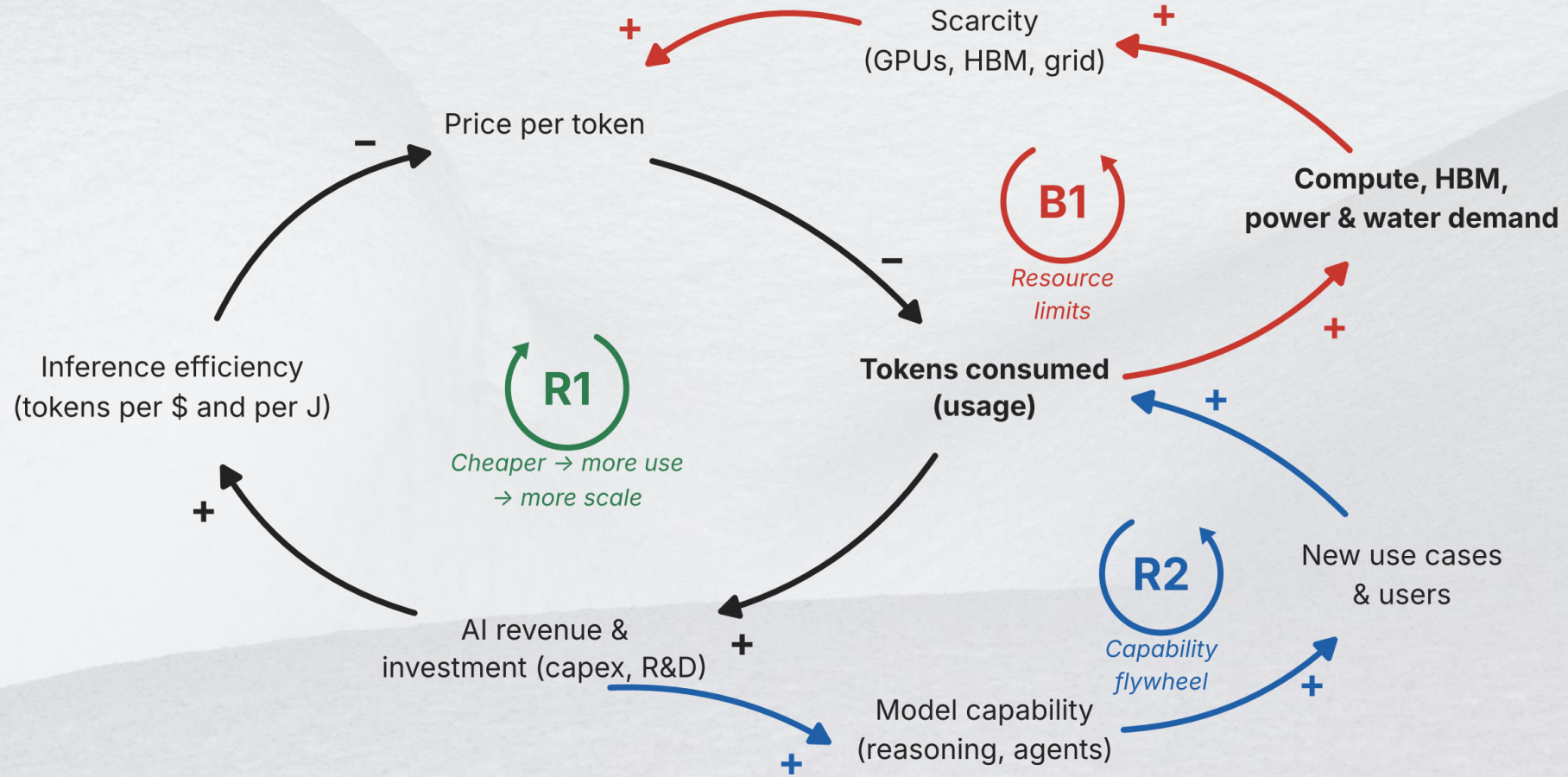
Cheaper tokens, far more tokens: Jevons paradox in AI inference



Sources: Epoch AI, "LLM inference prices have fallen rapidly but unequally across tasks" (Mar 2025), cheapest model $\geq$ GPT-4 MMLU score (86%), 3:1 input:output blended price; Google I/O 2025 & 2026 keynotes, Q2-2025 earnings, Google Cloud event (Oct 2025). Dashed = extrapolation, not data.

Jevons Paradox: Tokens System Dynamics

Jevons paradox as a causal loop: why efficiency raises total consumption



Polarity: + = same direction, - = opposite direction. R = reinforcing loop, B = balancing loop.
Jevons holds while R1 + R2 dominate B1 (demand elasticity > 1): efficiency gains are consumed as more tokens, not saved as fewer joules.

The Age of AI



AI Infra Building & Deployment Era



Token Demand Drives Tech Innovation



Capable $\neq$ Aligned; Unsolved AI Safety



Endangered Jobs, Energy, Environment



AI Efficiency Gap from a "25-Watt" Human Brain



THANK YOU !