The Gigawatt Gamble

The Economics, Impact, and an Alternative Path for Global AI Inference Deployment

Thomas Sohmers
CEO, Positron AI
thomas@positron.ai

POSITRON

Introduction

Thomas Sohmers
CEO, Positron AI
thomas@positron.ai

CEO & Co-Founder

Director of Technology Strategy

Principal Hardware Architect

CEO & Co-Founder

Student & Entrepreneur

Forbes 30 under 30
2013 Thiel Fellow
MIT Researcher

Three things we will cover

1

Are we in an AI bubble?
A brief economics discussion

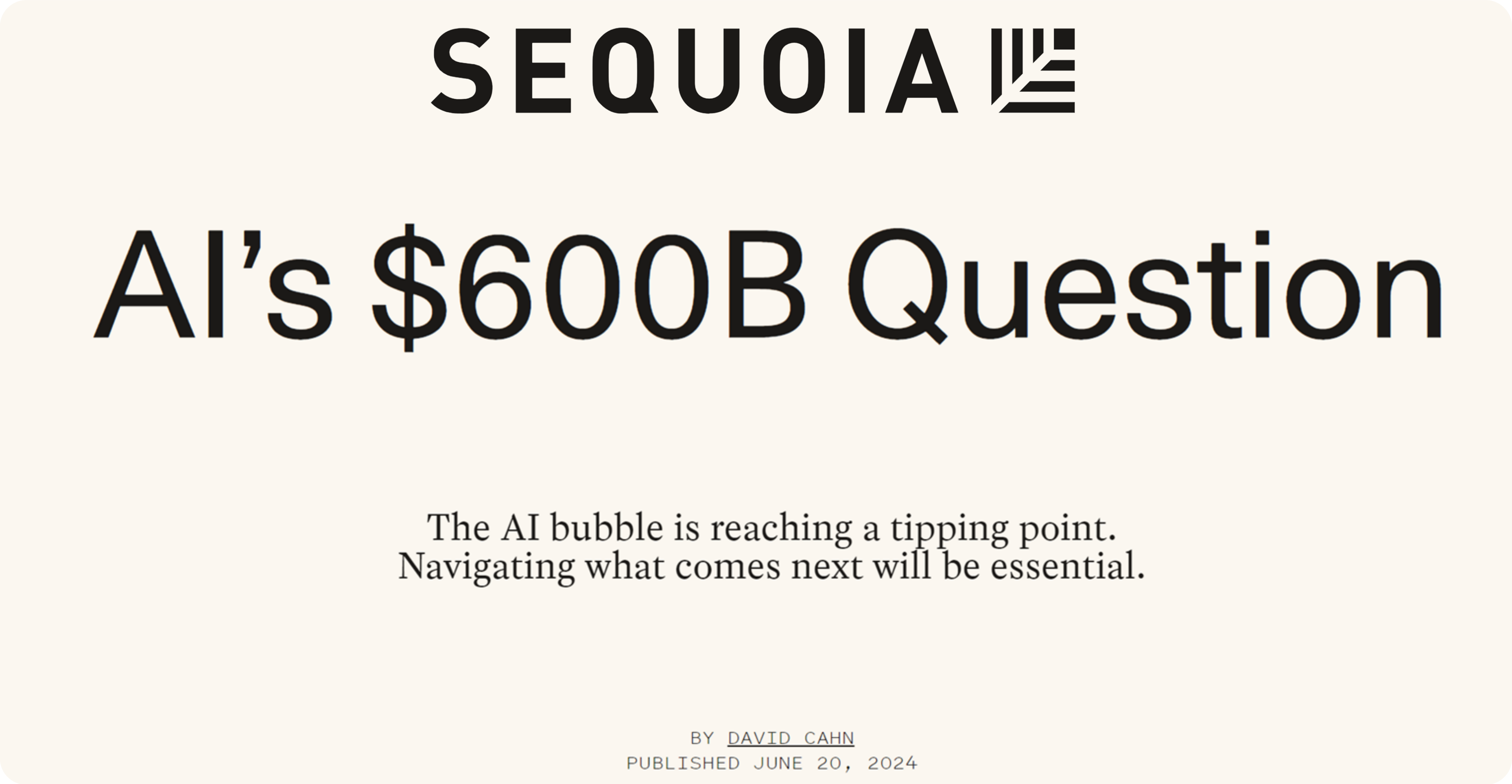
2

Understanding the coming economic upheaval

3

What are the paths to the future?

Bubble and Economies



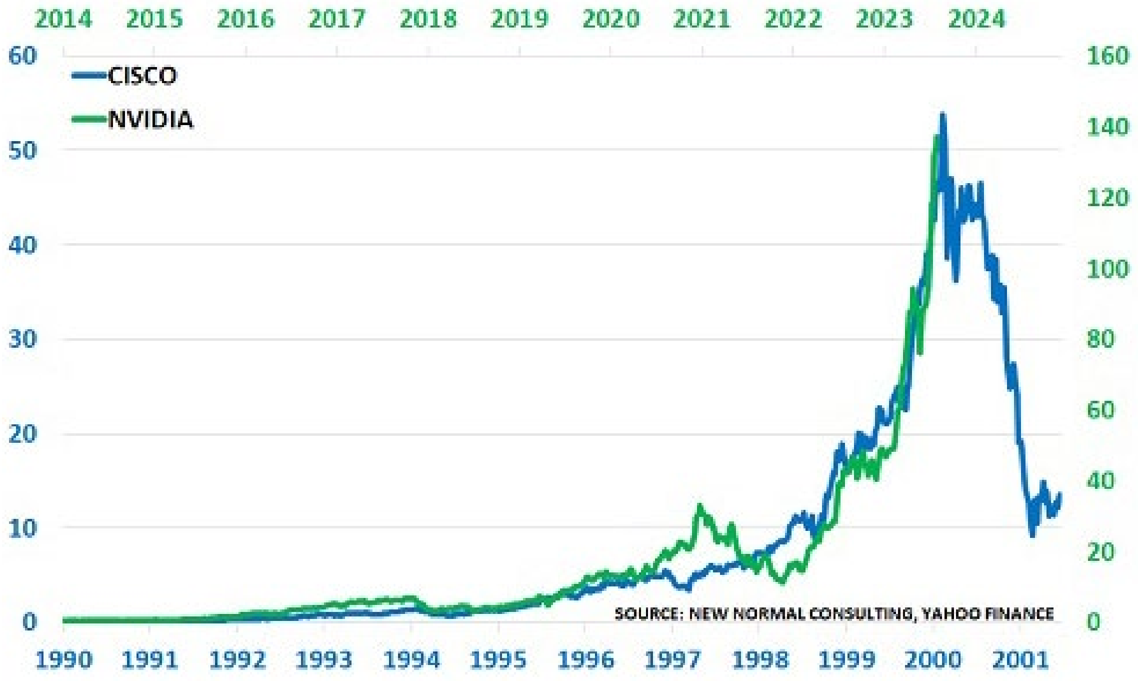
SEQUOIA

AI's $600B Question

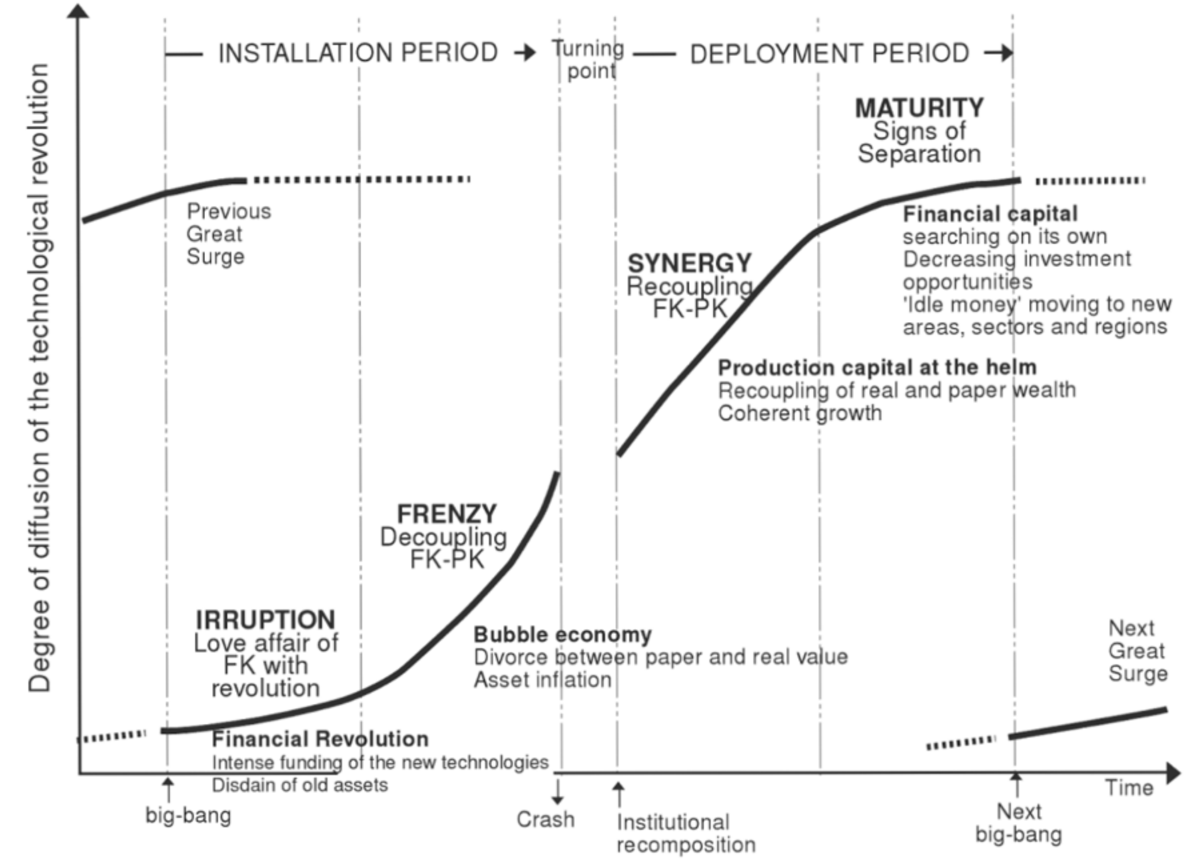
The AI bubble is reaching a tipping point.
Navigating what comes next will be essential.

BY DAVID CAHN
PUBLISHED JUNE 20, 2024

Are we in a bubble?



SOURCE: NEW NORMAL CONSULTING, YAHOO FINANCE


Degree of diffusion of the technological revolution
Time
big-bang
Crash
Institutional recomposition
Next big-bang
Previous Great Surge
Next Great Surge

INSTALLATION PERIOD → Turning point — DEPLOYMENT PERIOD →

IRRUPTION
Love affair of FK with revolution

FRENZY
Decoupling FK-PK

SYNERGY
Recoupling FK-PK

MATURITY
Signs of Separation

Financial Revolution
Intense funding of the new technologies
Disdain of old assets

Bubble economy
Divorce between paper and real value
Asset inflation

Production capital at the helm
Recoupling of real and paper wealth
Coherent growth

Financial capital
searching on its own
Decreasing investment opportunities
'Idle money' moving to new areas, sectors and regions

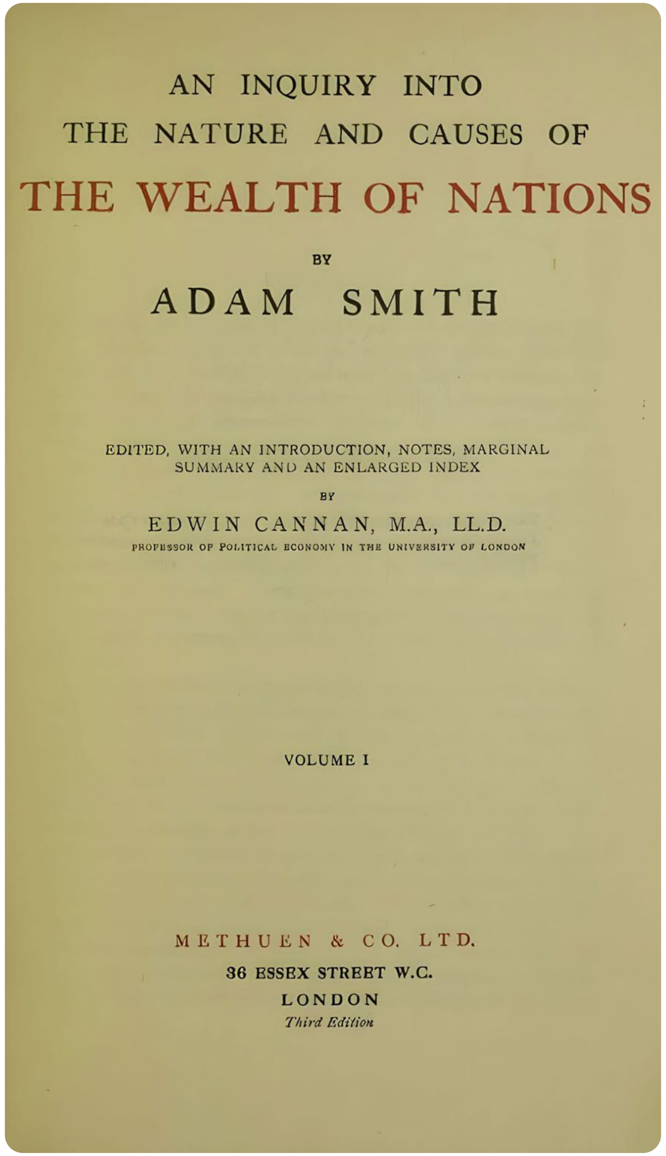
My thesis

The economic value of AI isn't transient or a ramp to to a point of required capacity, like the build out of the internet and wired, and wireless connectivity, or the adoption social media. These technology trends had predictable end-limits.

The value of AI fundamentally changes the fundamentals on which the global economy is structured: land, capital & labor. AI will turn labor into a near limitless resource injecting an exponential factor in the basic equations of wealth creation. This in turn will consume capital at unprecedented levels, and will rapidly test the limits of the land (e.g. power & water).

This is not a bubble; it's the leading indicator of the exponential potential of 'free' labor.

(pssst…and the hyperscalers know this.)



AN INQUIRY INTO
THE NATURE AND CAUSES OF

THE WEALTH OF NATIONS

BY

ADAM SMITH

EDITED, WITH AN INTRODUCTION, NOTES, MARGINAL
SUMMARY AND AN ENLARGED INDEX

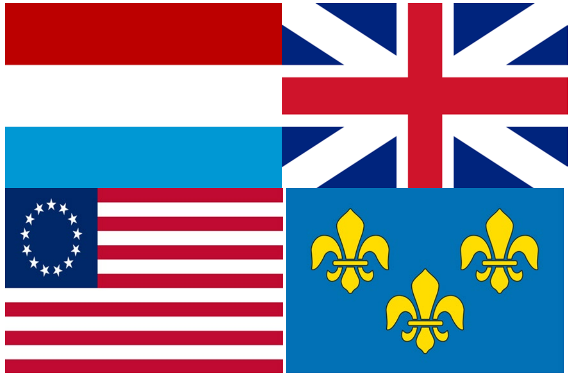
BY

EDWIN CANNAN, M.A., LL.D.
PROFESSOR OF POLITICAL ECONOMY IN THE UNIVERSITY OF LONDON

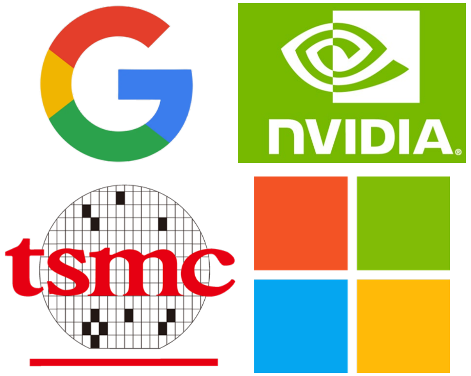
VOLUME I

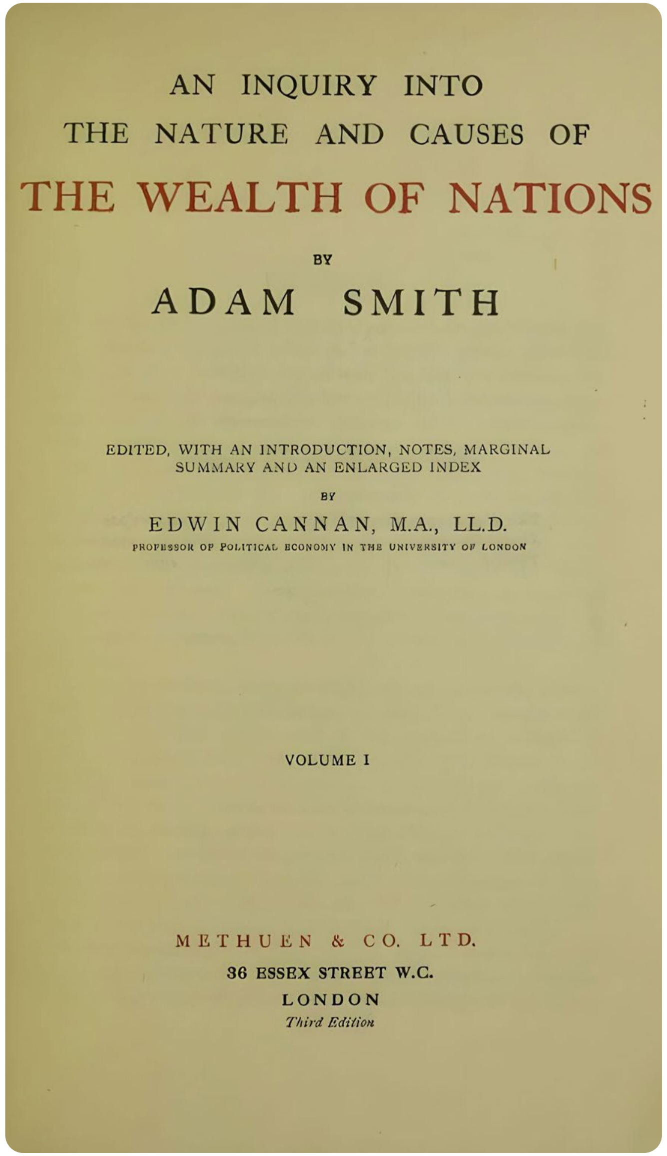
METHUEN & CO. LTD.
36 ESSEX STREET W.C.
LONDON
Third Edition

1776



2026





AN INQUIRY INTO
THE NATURE AND CAUSES OF
THE WEALTH OF NATIONS

BY

ADAM SMITH

EDITED, WITH AN INTRODUCTION, NOTES, MARGINAL
SUMMARY AND AN ENLARGED INDEX

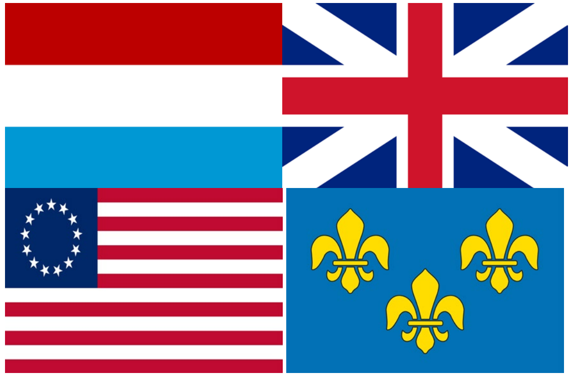
BY

EDWIN CANNAN, M.A., LL.D.
PROFESSOR OF POLITICAL ECONOMY IN THE UNIVERSITY OF LONDON

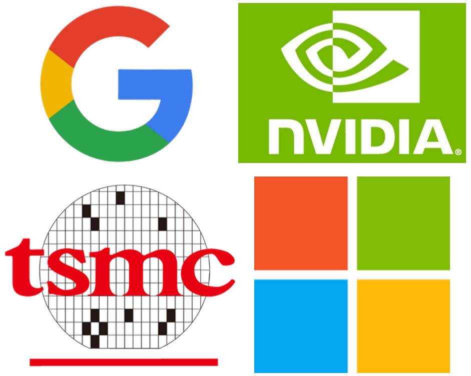
VOLUME I

METHUEN & CO. LTD.
36 ESSEX STREET W.C.
LONDON
Third Edition

1776



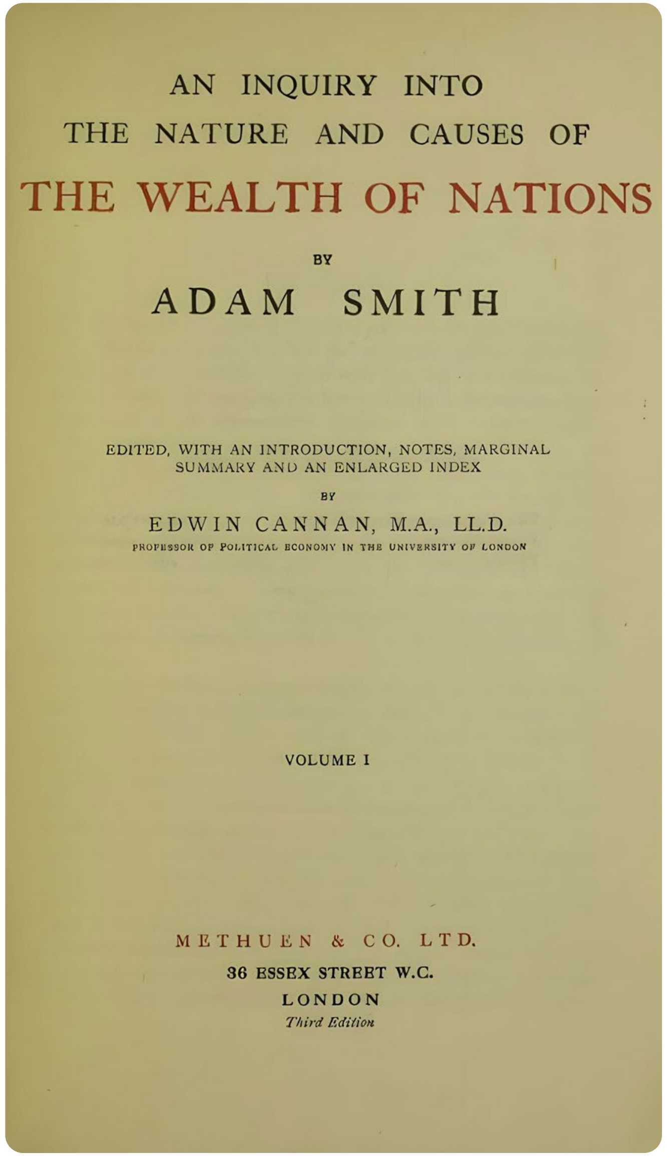
2026



Farms

Land

Data Centers



AN INQUIRY INTO
THE NATURE AND CAUSES OF
THE WEALTH OF NATIONS

BY

ADAM SMITH

EDITED, WITH AN INTRODUCTION, NOTES, MARGINAL
SUMMARY AND AN ENLARGED INDEX

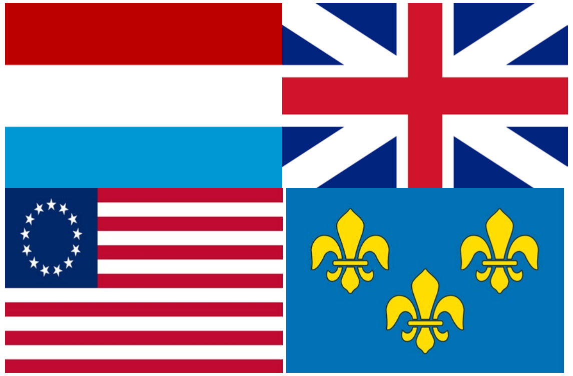
BY

EDWIN CANNAN, M.A., LL.D.
PROFESSOR OF POLITICAL ECONOMY IN THE UNIVERSITY OF LONDON

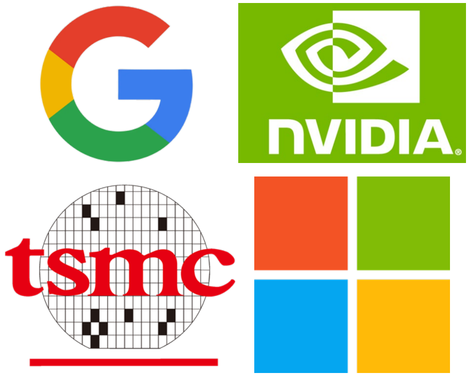
VOLUME I

METHUEN & CO. LTD.
36 ESSEX STREET W.C.
LONDON
Third Edition

1776

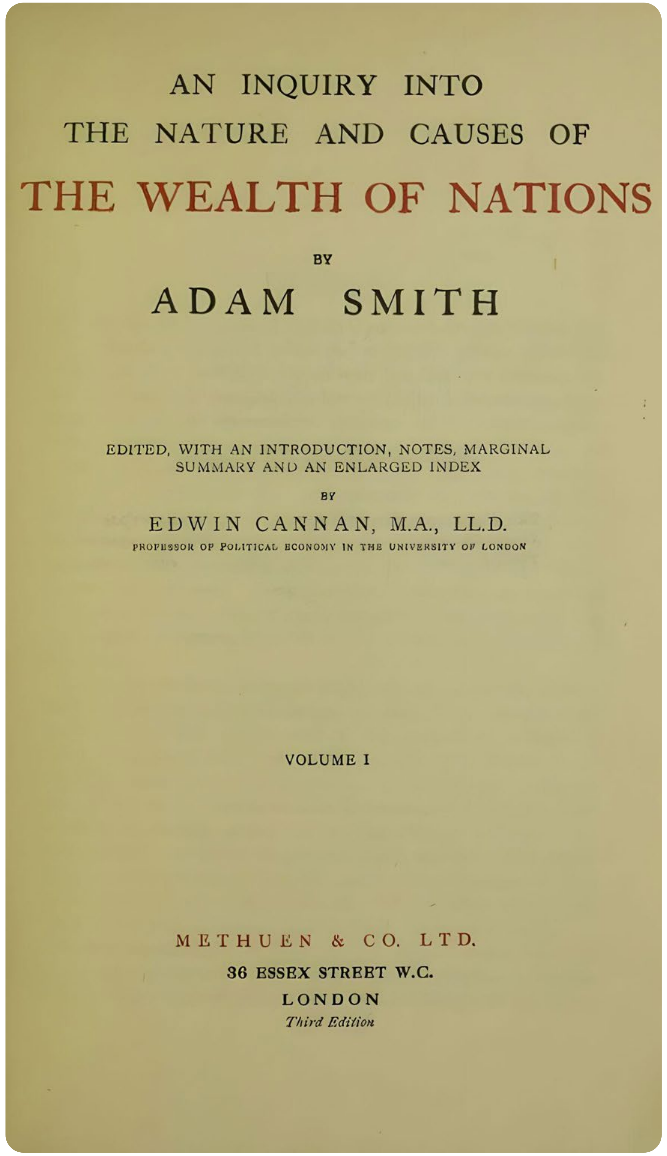


2026



Farms **Land** Data Centers

Humans **Labor** Machines



AN INQUIRY INTO
THE NATURE AND CAUSES OF
THE WEALTH OF NATIONS

BY

ADAM SMITH

EDITED, WITH AN INTRODUCTION, NOTES, MARGINAL
SUMMARY AND AN ENLARGED INDEX

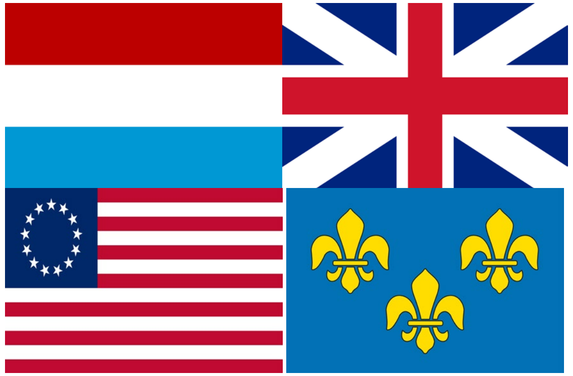
BY

EDWIN CANNAN, M.A., LL.D.
PROFESSOR OF POLITICAL ECONOMY IN THE UNIVERSITY OF LONDON

VOLUME I

METHUEN & CO. LTD.
36 ESSEX STREET W.C.
LONDON
Third Edition

1776

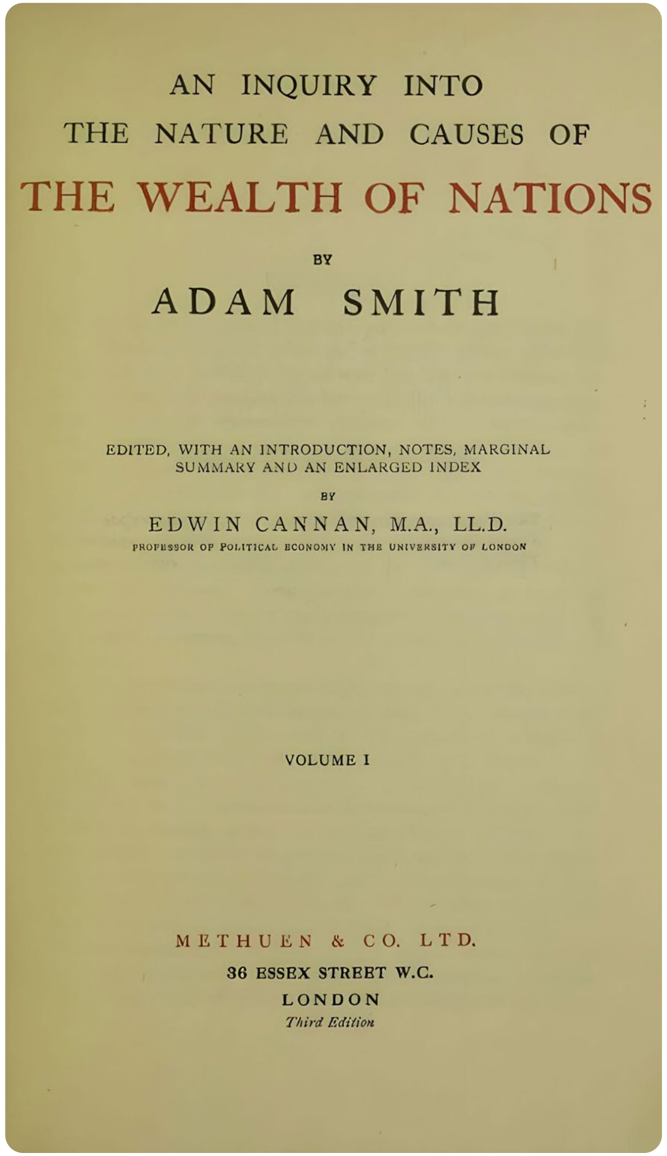


2026

NVIDIA

tsmc

Farms	**Land**	Data Centers
Humans	**Labor**	Machines
Steam Engine	**Capital**	GPUs



AN INQUIRY INTO
THE NATURE AND CAUSES OF
THE WEALTH OF NATIONS

BY

ADAM SMITH

EDITED, WITH AN INTRODUCTION, NOTES, MARGINAL
SUMMARY AND AN ENLARGED INDEX

BY

EDWIN CANNAN, M.A., LL.D.
PROFESSOR OF POLITICAL ECONOMY IN THE UNIVERSITY OF LONDON

VOLUME I

METHUEN & CO. LTD.
36 ESSEX STREET W.C.
LONDON
Third Edition

The Factors of Production: Land

In 1776, land's value was what could be directly farmed from it, leading to expansion.

Like land expansion, Data Centers are the embodiment of new wealth creation.

New Data centers are being built now but also have limits to expansion, particularly access to power and water.

500MW Google Datacenter (source: Semianalysis)

TSMC Arizona Fab 21 (source: TSMC)

550MW Desert Sunlight Solar Farm (source: TIME)

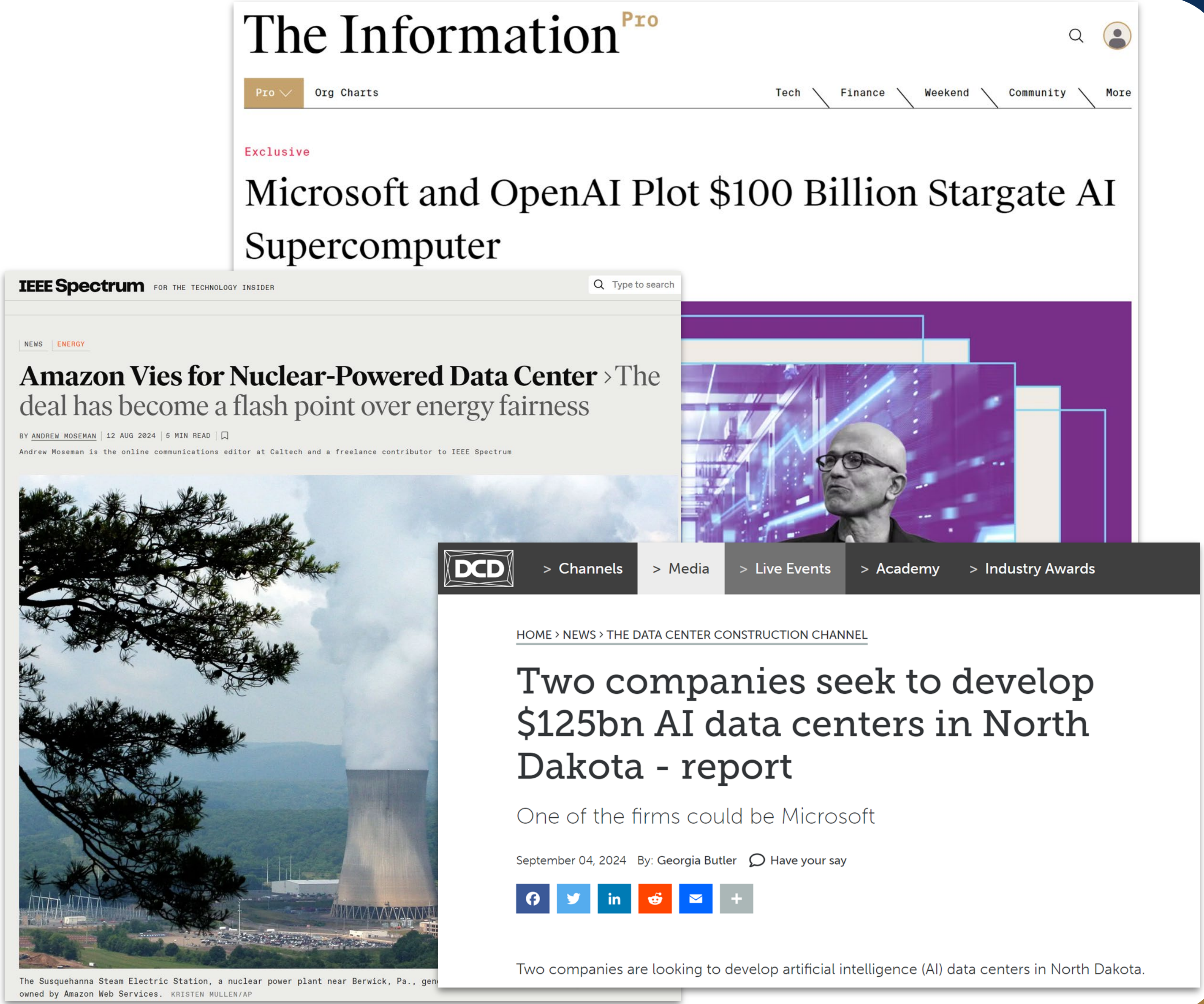
The Factors of Production: Capital

AI Infrastructure spend is expected to hit over **$200B** ***per year*** by 2030, with total spend between 2024 and 2030 approaching **$1T.**

The Manhattan Project, between 1942 and 1946 spent **~$32B** in 2024 dollars

The Interstate Highway System, between 1956 and 1992, spent **~$600B** in 2024 dollars.

And for some reason analysts expect direct ROI…



The Information Pro

Exclusive

Microsoft and OpenAI Plot $100 Billion Stargate AI Supercomputer

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

NEWS | ENERGY

Amazon Vies for Nuclear-Powered Data Center › The deal has become a flash point over energy fairness

BY ANDREW MOSEMAN | 12 AUG 2024 | 5 MIN READ

Andrew Moseman is the online communications editor at Caltech and a freelance contributor to IEEE Spectrum

The Susquehanna Steam Electric Station, a nuclear power plant near Berwick, Pa., gen owned by Amazon Web Services. KRISTEN MULLEN/AP

HOME › NEWS › THE DATA CENTER CONSTRUCTION CHANNEL

Two companies seek to develop $125bn AI data centers in North Dakota - report

One of the firms could be Microsoft

September 04, 2024 By: Georgia Butler Have your say

Two companies are looking to develop artificial intelligence (AI) data centers in North Dakota.

The Factors of Production: Labor

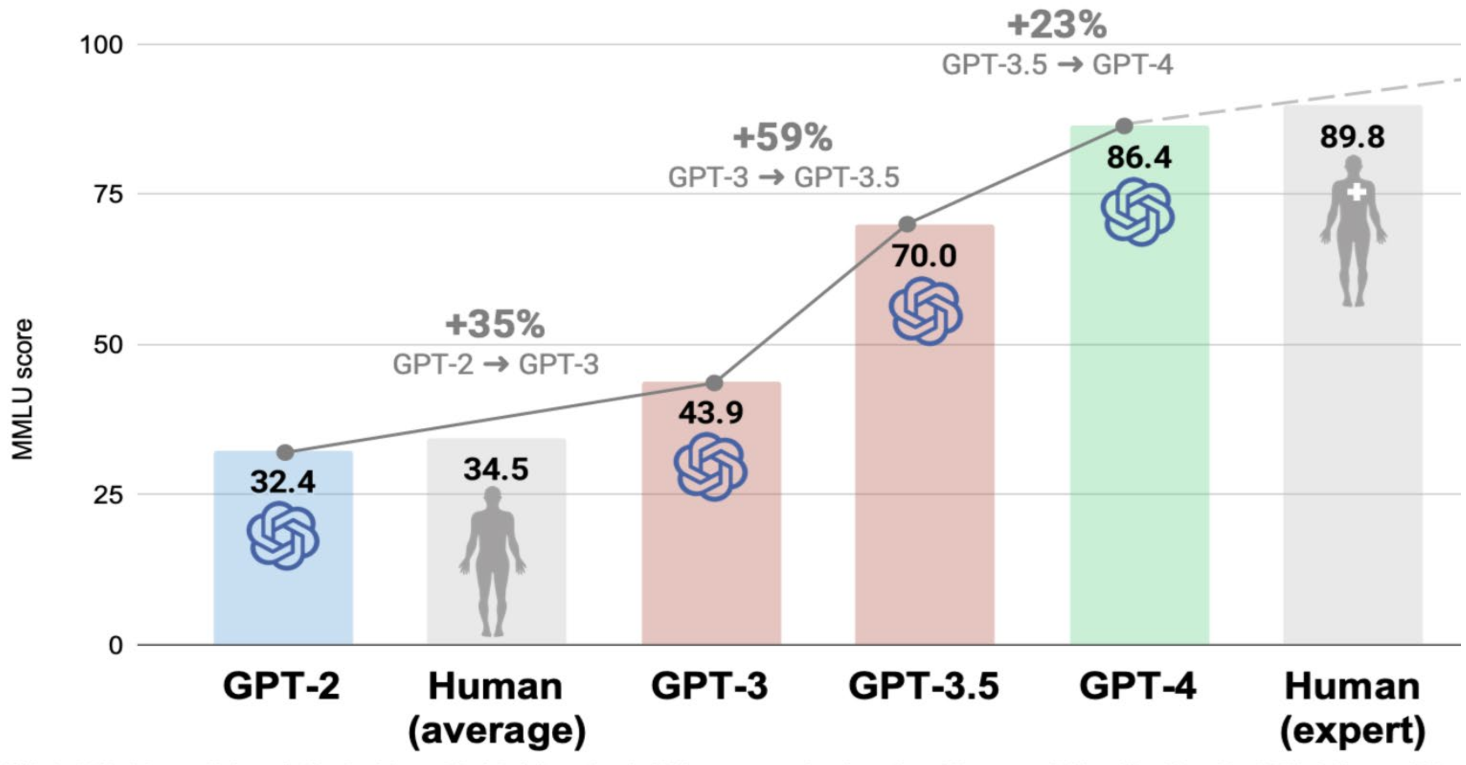


Image Source: LifeArchitect.ai/gpt4-5

Business will very soon have the ability to to spin up 100K expert employees on demand.

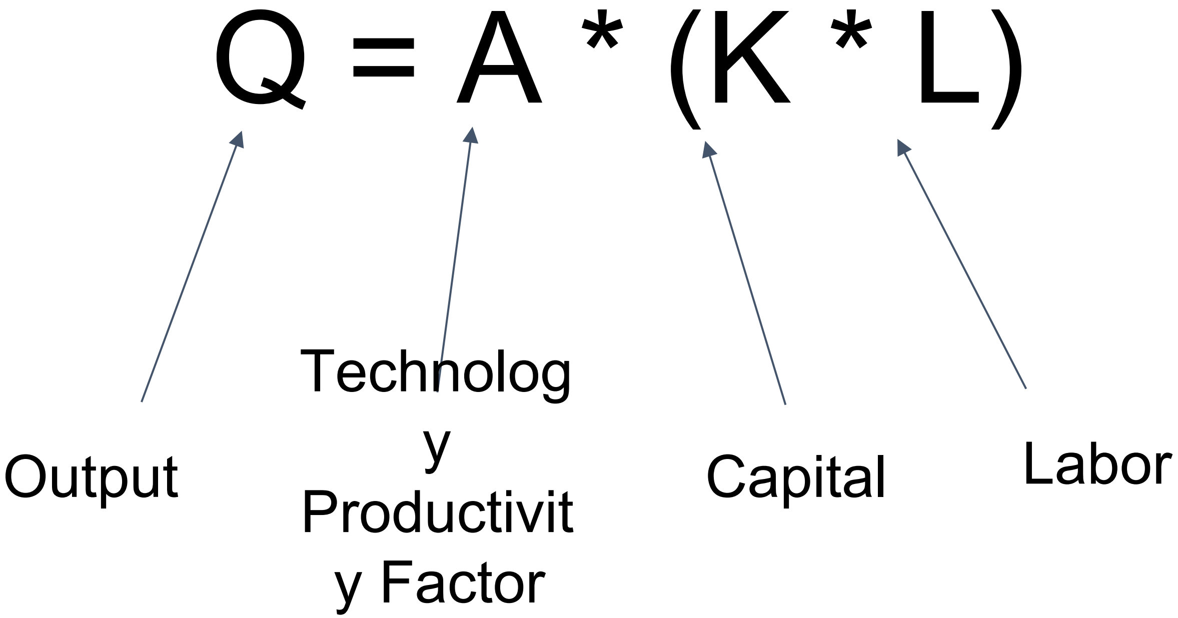
It can't be just be the biggest and wealthiest companies who can do this!

"Labour was the first price, the original purchase-money that was paid for all things. It was not by gold or by silver, but by labour, that all wealth of the world was originally purchased." –Adam Smith

Economics Lesson

(Simplified) Cobb-Douglas Production Function:

Output (or Wealth) is created using two main factors: Capital (machines, investments) and Labor.



$$Q = A * (K * L)$$

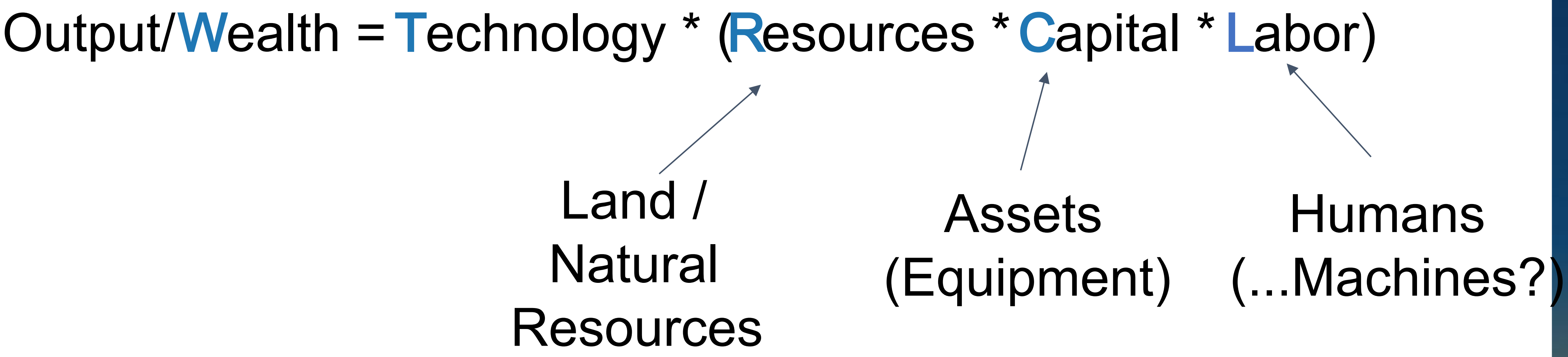
Output

Technology Productivity Factor

Capital

Labor

Our own variant



Output/Wealth = Technology * (Resources * Capital * Labor)

Land / Natural Resources

Assets (Equipment)

Humans (...Machines?)

$$W = T * (R * C * L)$$

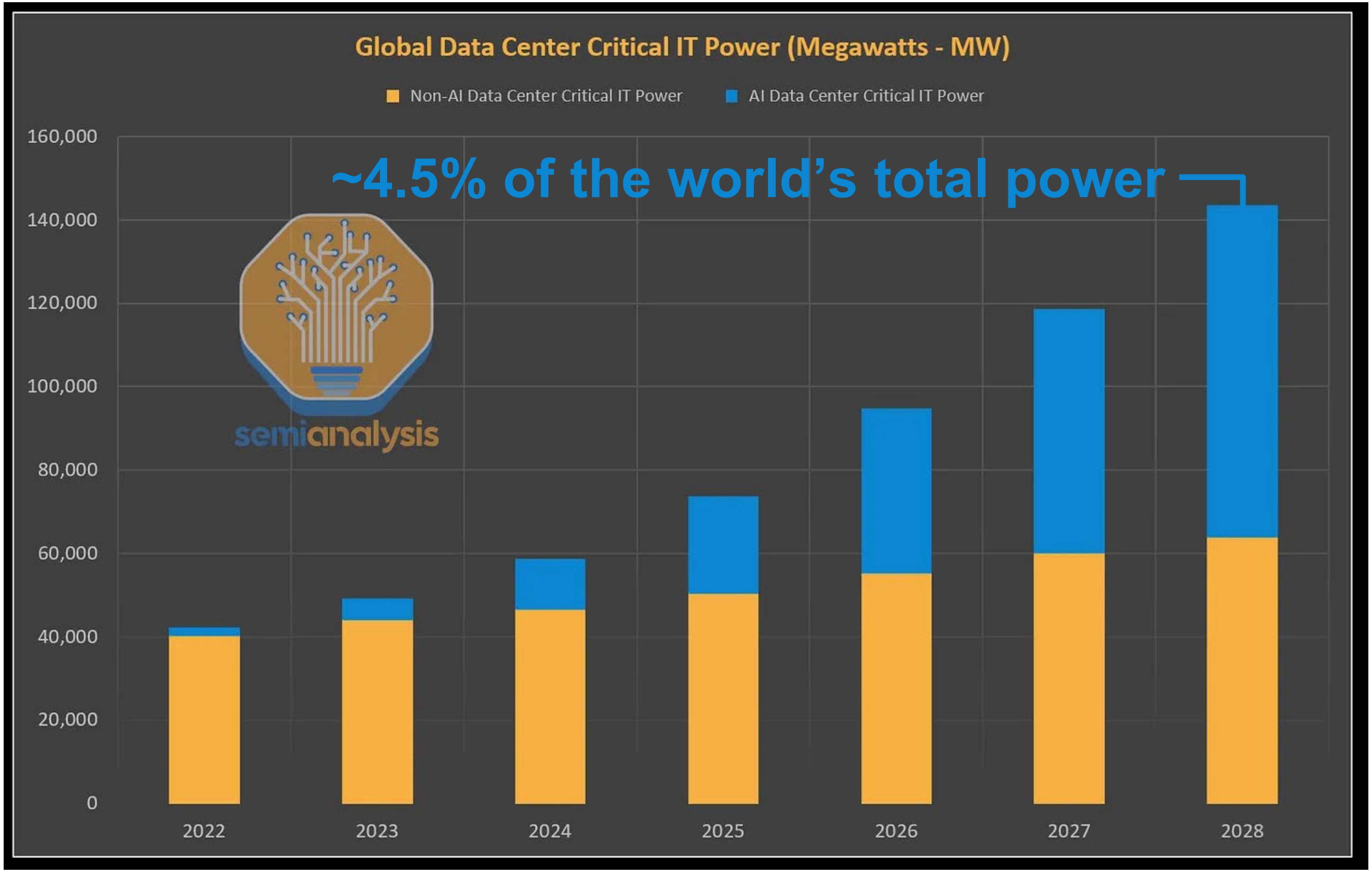
The Previous Industrial Revolutions

1st Industrial Revolution	2nd Industrial Revolution	3rd Industrial Revolution	4th Industrial Revolution
Early Mechanized Production + Steam Power	Electrification + Mass Transportation + Long Distance Communication	Digitization + Fiat Currency	Emergence of Artificial Intelligence
1780-1880	1880-1980	1980-2020	2020→
Increase in labor through specialization + mechanization	Increase in resources through mass extraction	Use of capital re-aligns to the value of information. Labour shifts to knowledge.	Labour becomes exponential, effectively limitless.
$W = T_1 * (1R * 1C * 2L)$	$W = T_2 * (3R * 2C * 1L)$	$W = T_3 * (3R * 3C * 2L)$	$W = T_4 * (xR * xC * xL^n)$

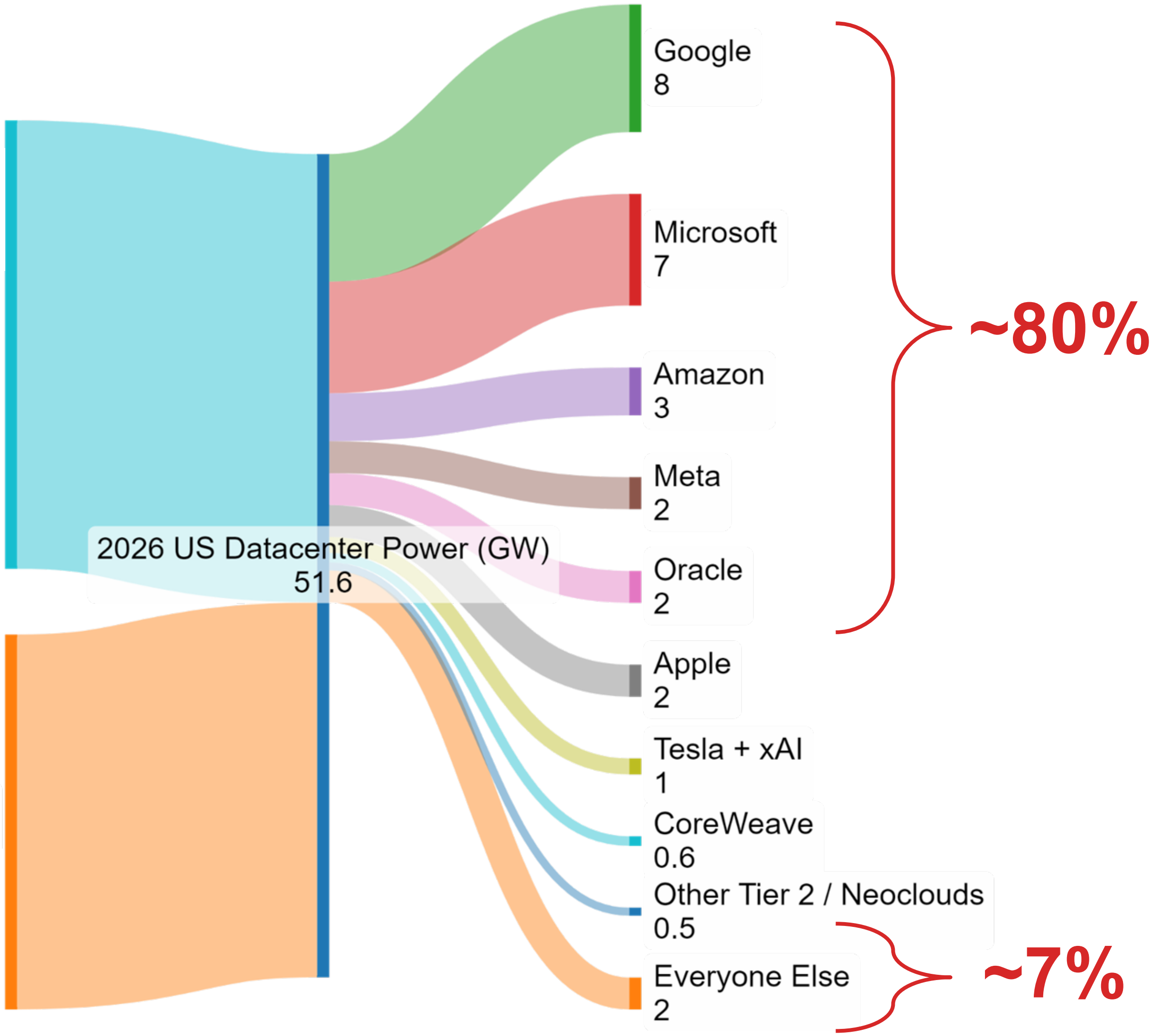
1st order effects we can see in front of us **today:**

Power consumption

AI based skilled labour;
AI writing code


Global Data Center Critical IT Power (Megawatts - MW)
Non-AI Data Center Critical IT Power
AI Data Center Critical IT Power
~4.5% of the world's total power
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
2022
2023
2024
2025
2026
2027
2028
semianalysis

The Consolidation Problem



Data source: Semianalysis

The Unpriced Externalities

Global Datacenter Deployments

	2024	2030
H100 Equivalents Deployed	2.25M	135M
Gigawatts Deployed	8.5	144
TWatt/hour Consumed	73	1160
Million Metric tons of CO_2	51	810

2030 Projected Power Usage Equivalent

160 million homes
~total number of homes in US+CAN

91 billion gallons of gas
~8 months of US gasoline usage

400 billion gallons of water for cooling
equivalent to 4 million households

Why it is totally justified

Llama 3.1 405B can currently replace junior engineers ***and/or*** **augment senior engineers at ~30 tokens per second on 8xH100s**

Generation: 30 tokens/sec = ~2.5 million tokens/day = 75 million tokens/month = 900 million tokens a year

Assume 10:1 input to output with $3/Million Tokens and it only costs a company **~$30,000/yr for a "Llama employee"**

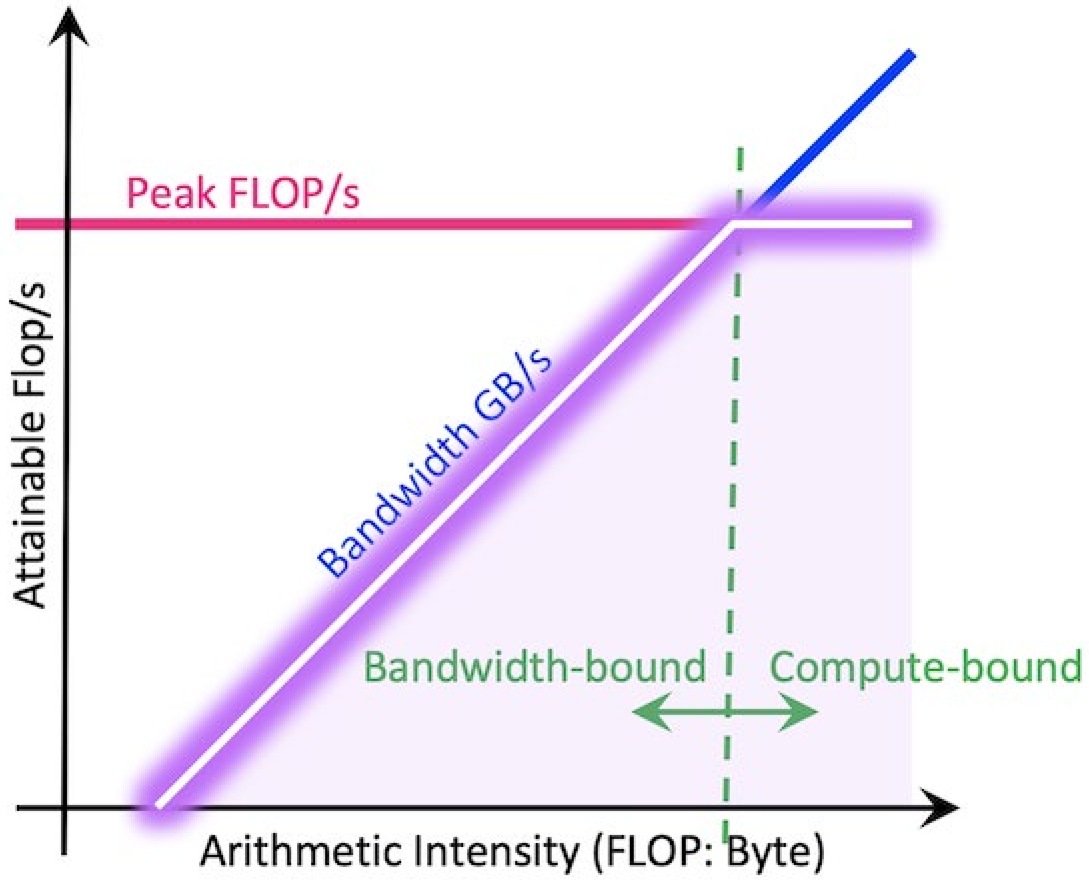
So what is

POSITRON

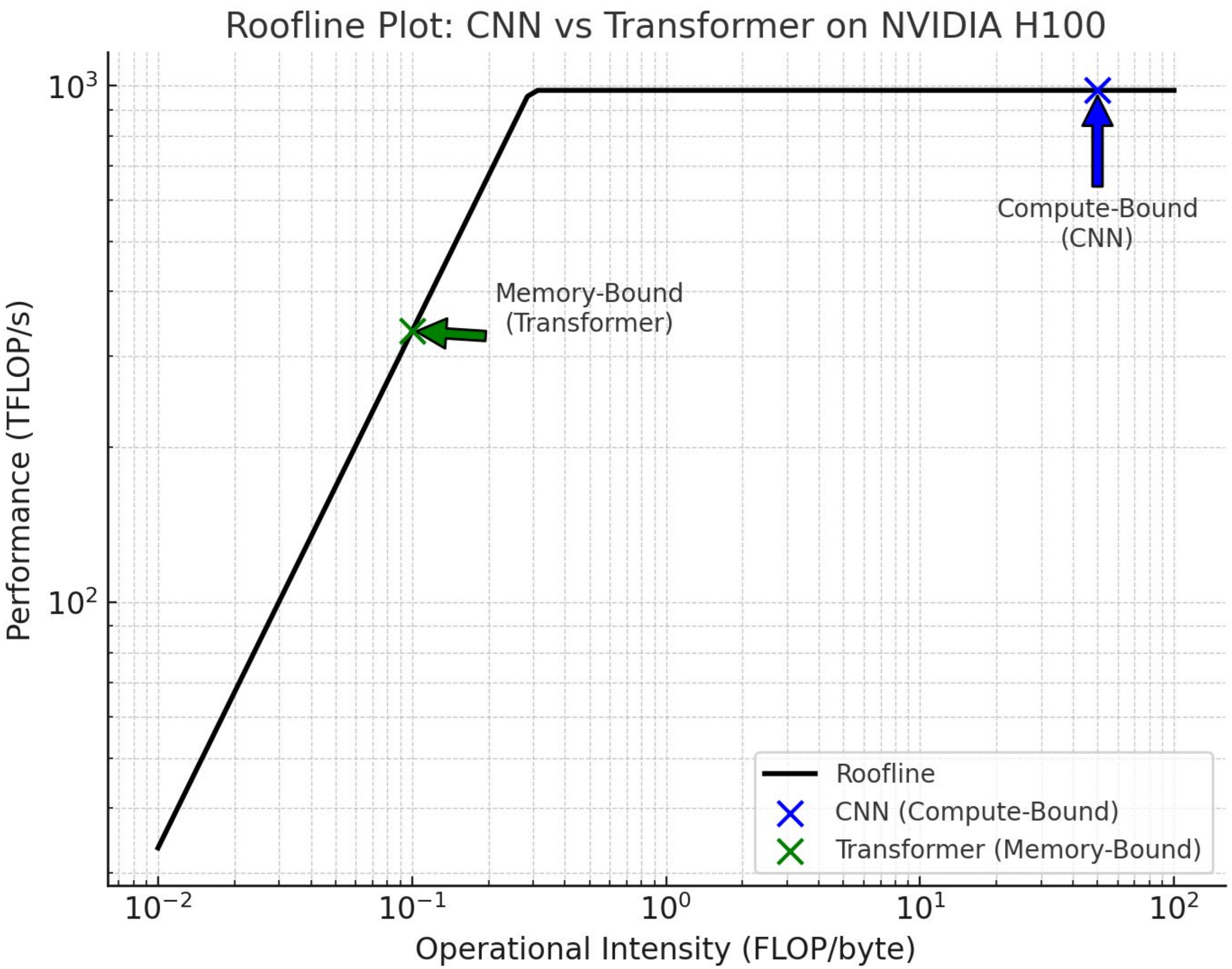
Positron Facts

- Founded April 2023
- Goal = Change the underlying economics of applied AI, starting immediately with making inference affordable to more people
- $12M in seed funding
- **18 months from idea to shipping production hardware**
- 21 employees
- We’re hiring

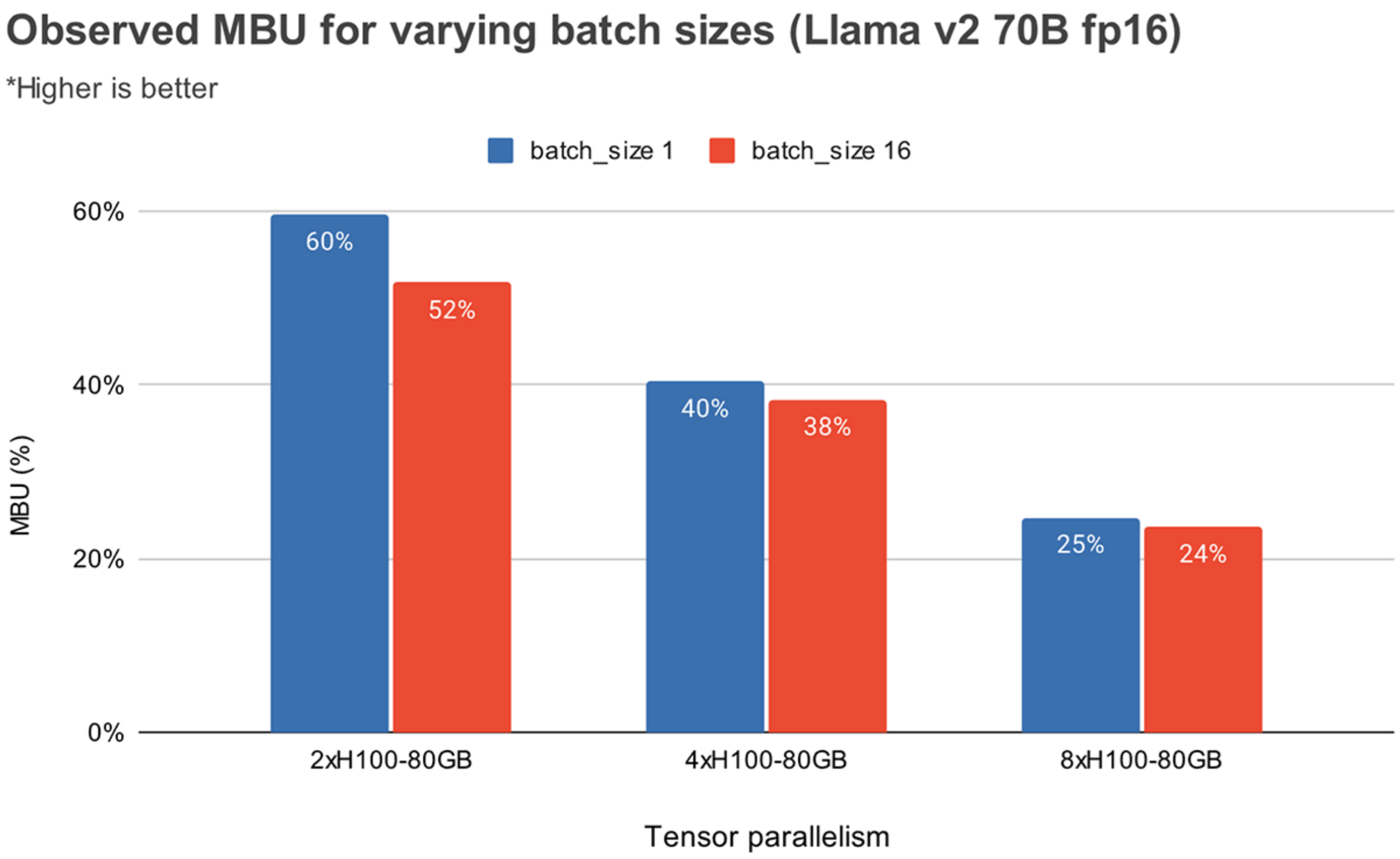
Arithmetic Intensity



What makes Transformers different?

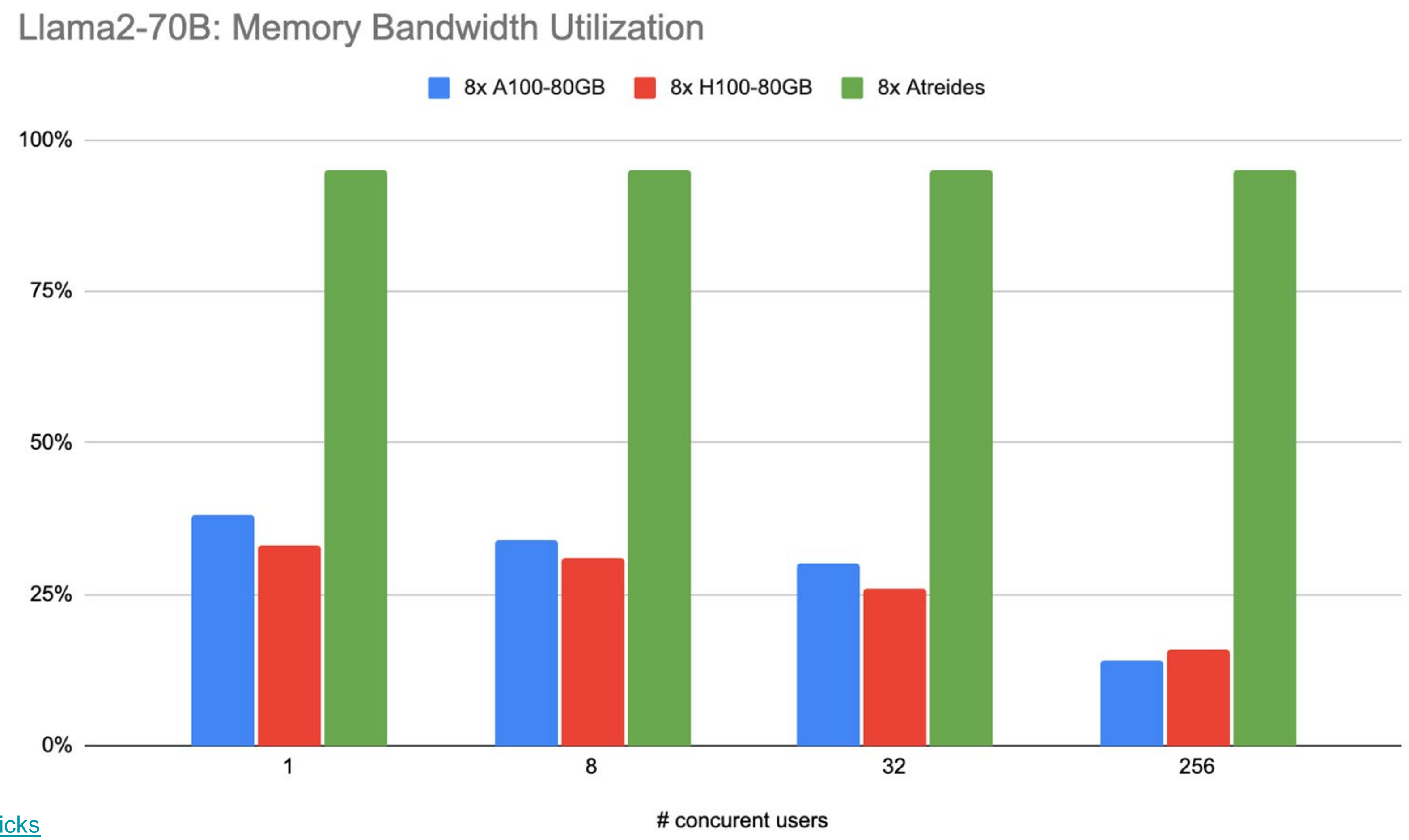


Positron high level memory story



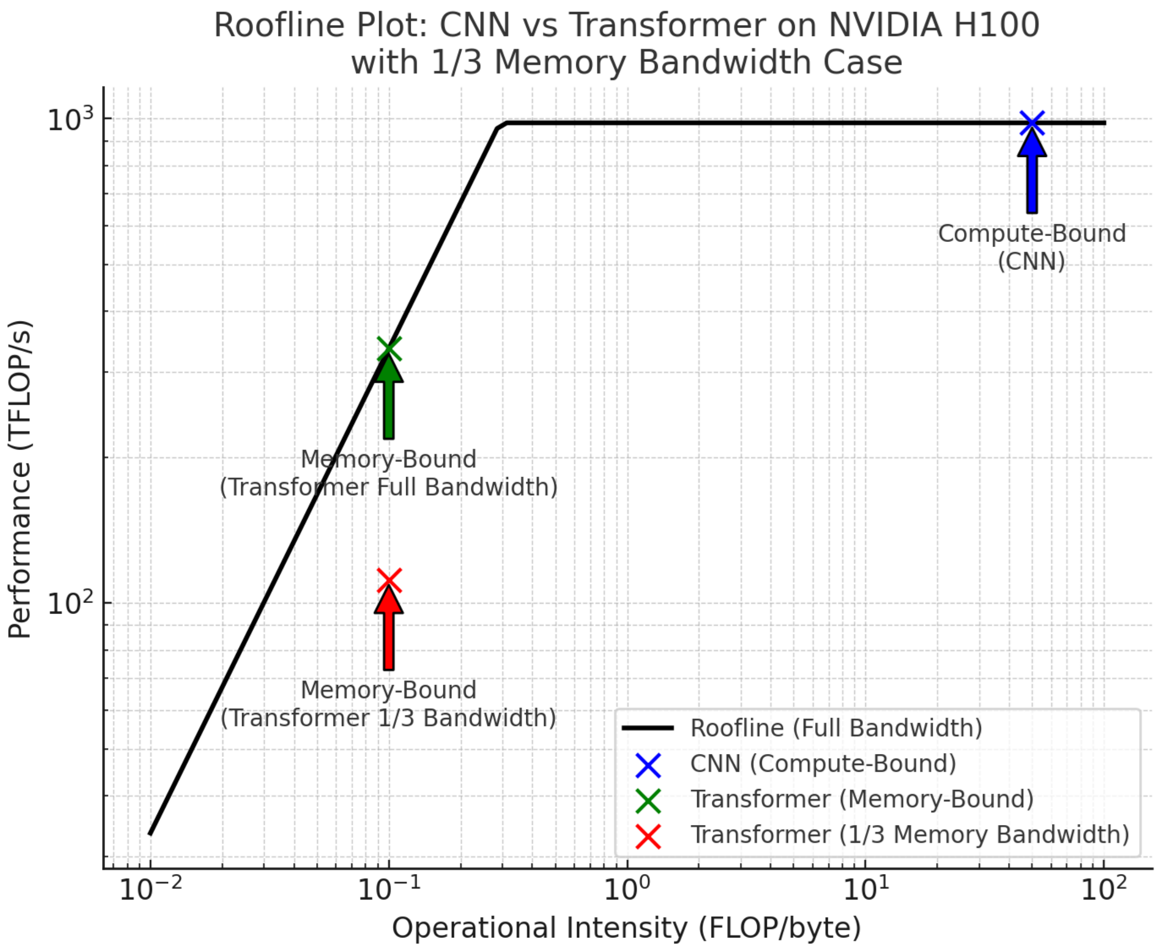
NVIDIA Source data: Databricks

Positron high level memory story

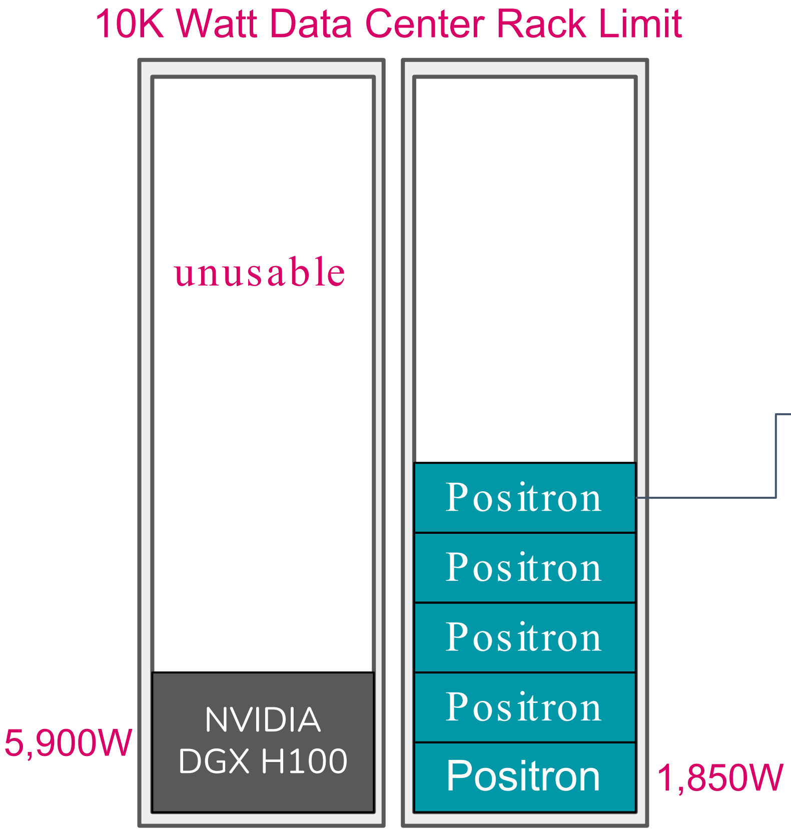


NVIDIA Source data: Databricks

Positron high level memory story



More density per watt = more economic value



Atlas Server

(~= 1 SW engineer of applied AI)

~6x more applied AI per 10KW footprint

~½ the cost

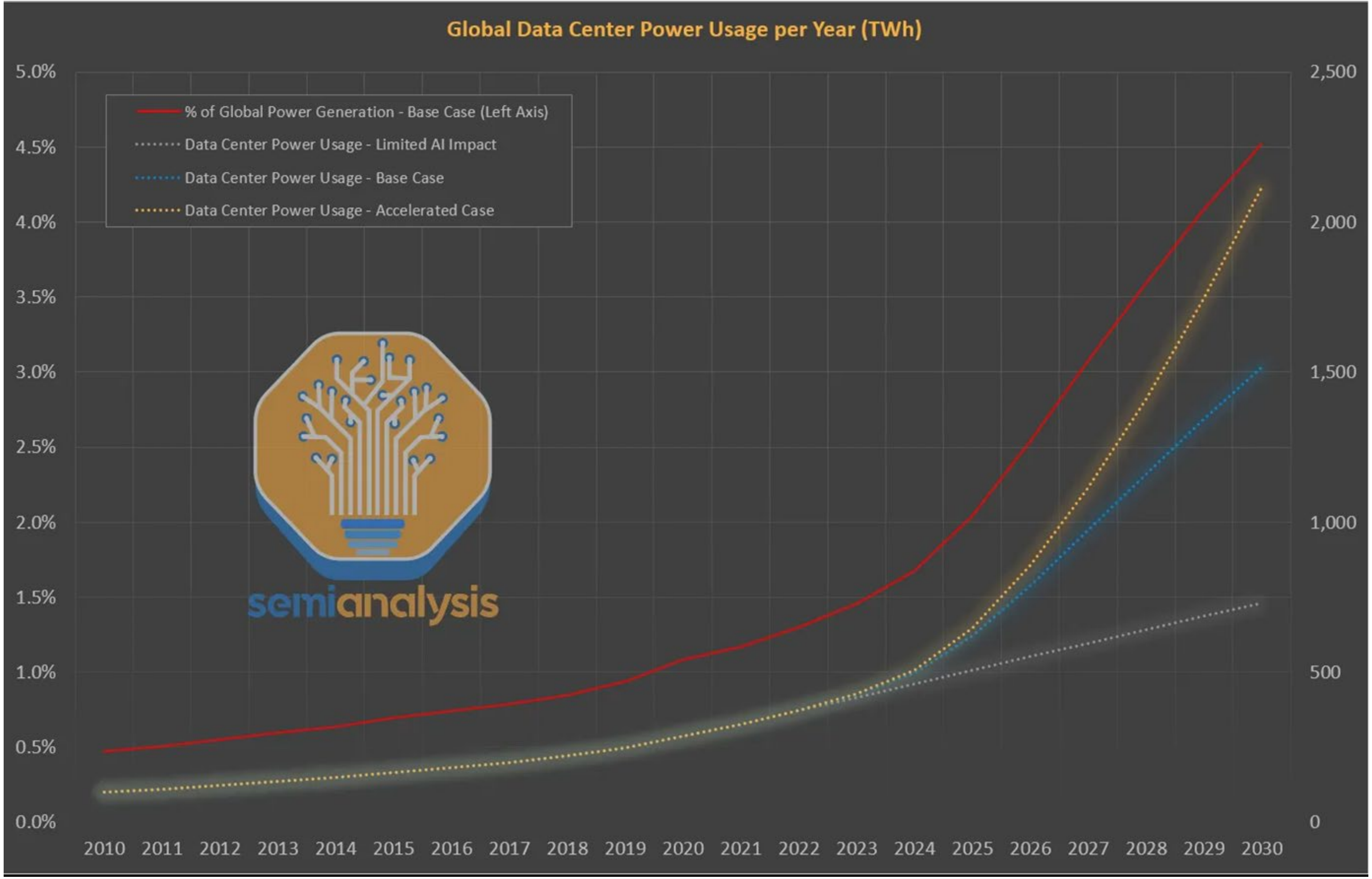
Positron shipping product

Special Thanks to our investors…

POSITRON

Come talk to us at Booth 35/36
and
Product Demo Stage @ 4:15 today

Backup


Global Data Center Power Usage per Year (TWh)
% of Global Power Generation - Base Case (Left Axis)
Data Center Power Usage - Limited AI Impact
Data Center Power Usage - Base Case
Data Center Power Usage - Accelerated Case
5.0%
4.5%
4.0%
3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
2,500
2,000
1,500
1,000
500
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
semianalysis

Edit Master title style

- Edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

The Previous Industrial Revolutions

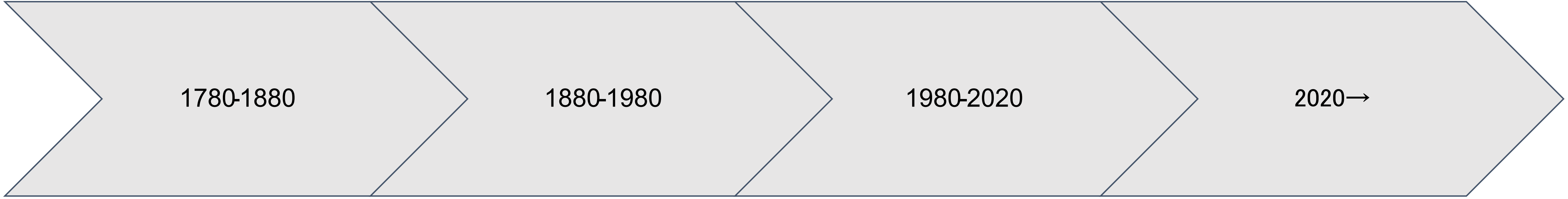
First Industrial Revolution

(Early Mechanized Production + Steam Power)

Second Industrial Revolution

(Mass Transportation + Long Distance Communication)

land limits reached; size, environmental etc, no more being created, resources constraints over time



1780-1880

1880-1980

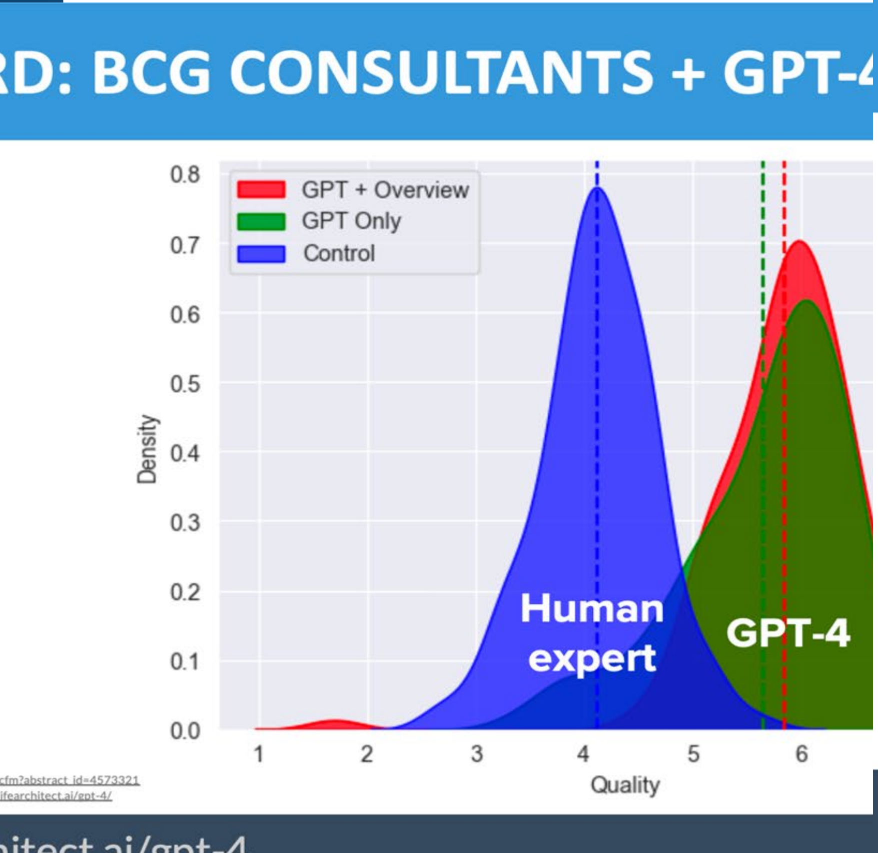
1980-2020

2020→

Land	+
Labor	+++
Capital	++

Labor goes expenetial tipping point where labour can grow indefinitely

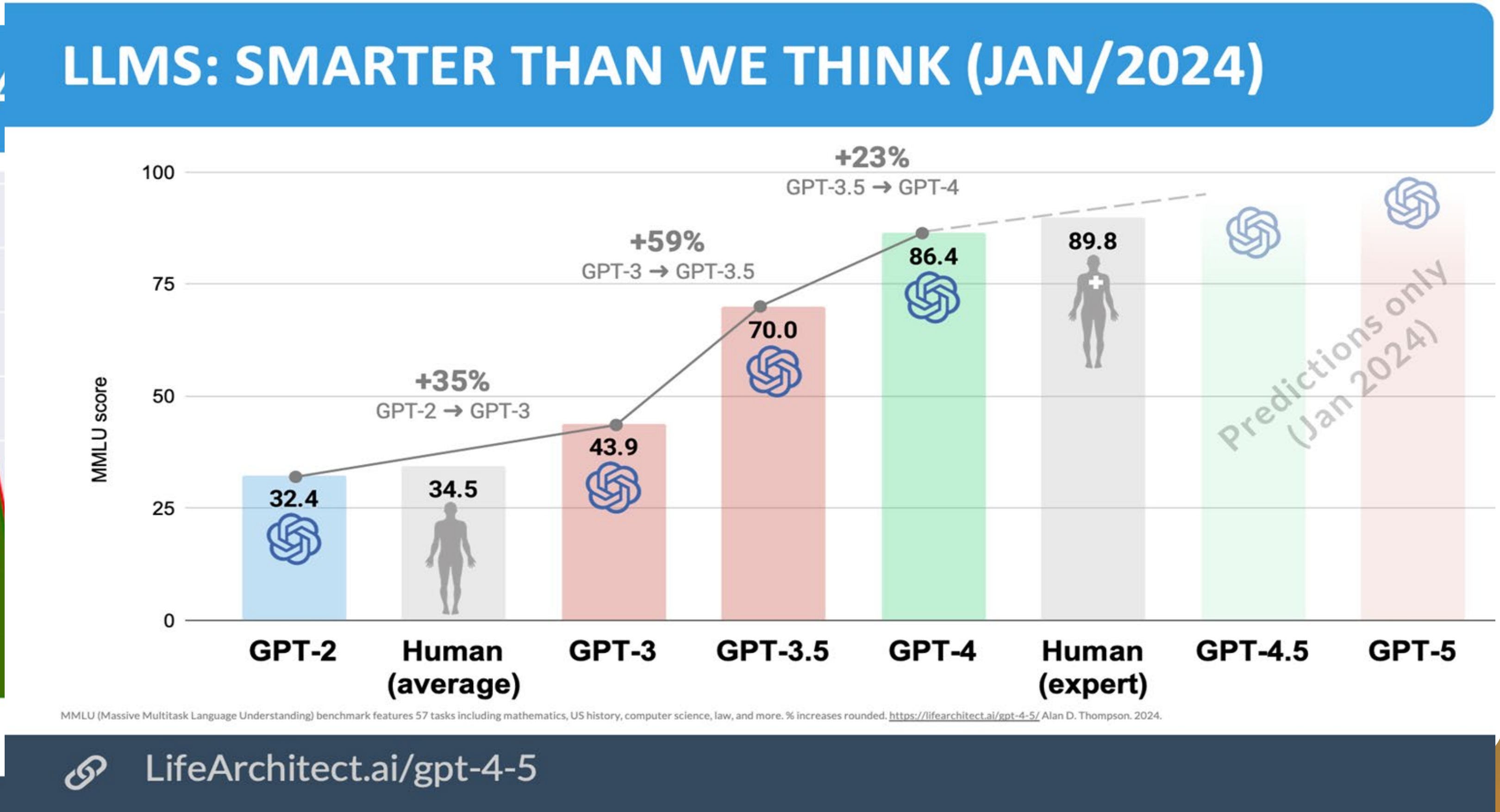
biggest capital outlay so far



RD: BCG CONSULTANTS + GPT-4

cfm?abstract_id=4573321
lifearchitect.ai/gpt-4/

itect.ai/gpt-4



LLMS: SMARTER THAN WE THINK (JAN/2024)

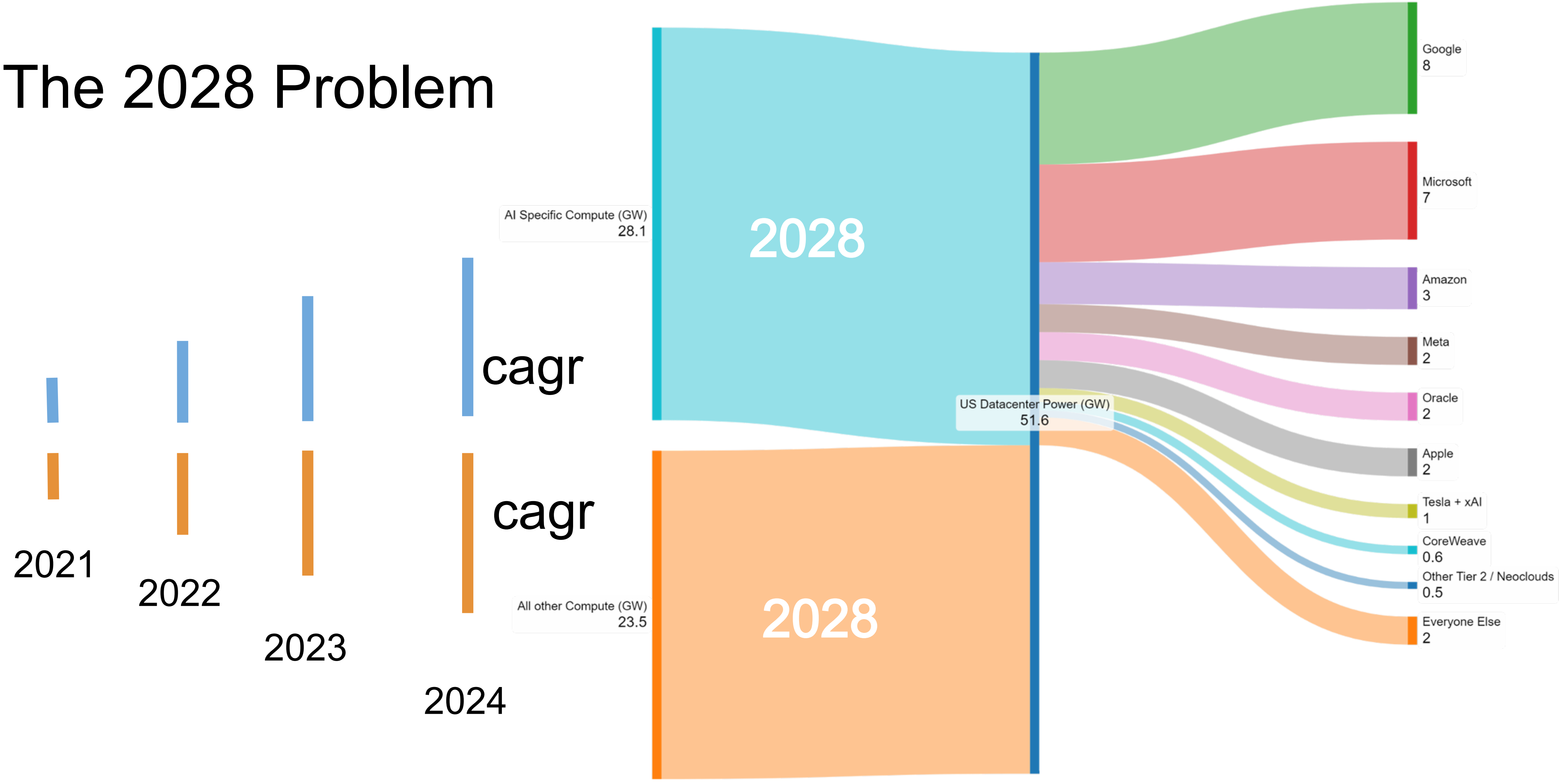
MMLU (Massive Multitask Language Understanding) benchmark features 57 tasks including mathematics, US history, computer science, law, and more. % increases rounded. https://lifearchitect.ai/gpt-4-5/ Alan D. Thompson. 2024.

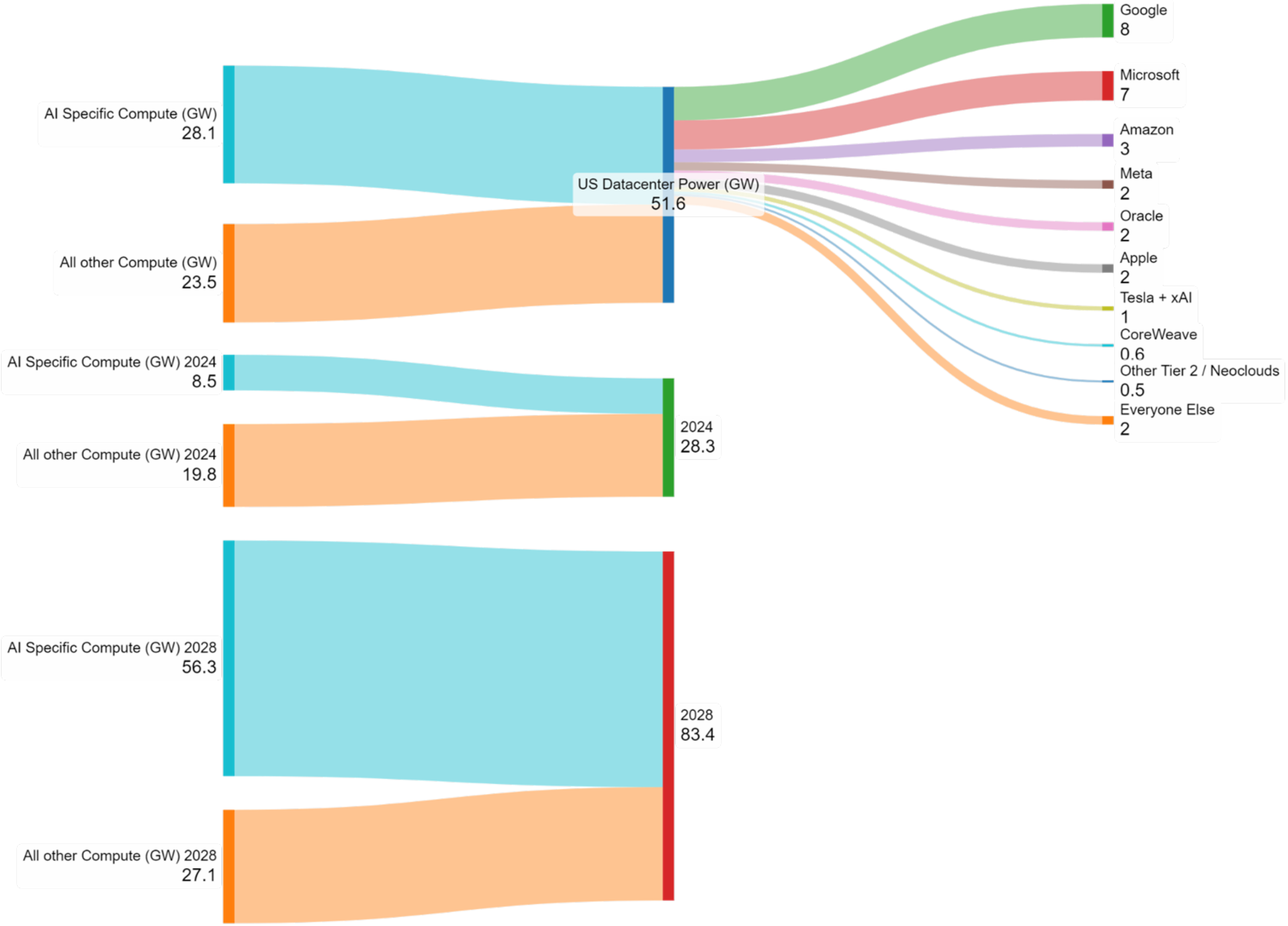
LifeArchitect.ai/gpt-4-5

AI HARDWARE & EDGE AI SUMMIT

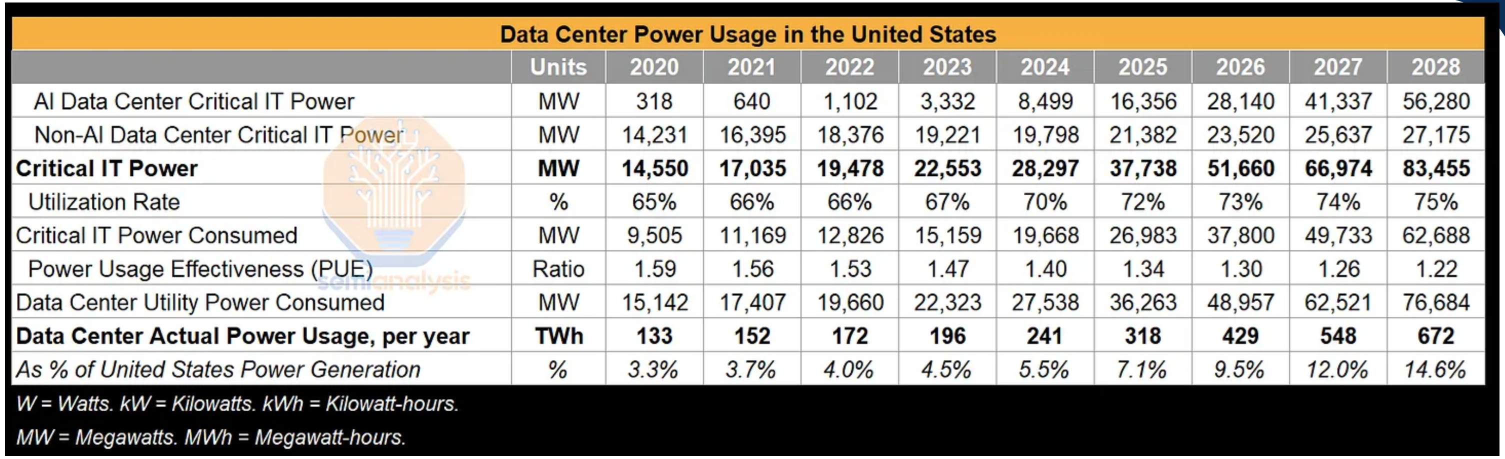
AI Hardware & Systems
@aiandsystems

The 2028 Problem




AI Specific Compute (GW)
28.1
All other Compute (GW)
23.5
AI Specific Compute (GW) 2024
8.5
All other Compute (GW) 2024
19.8
AI Specific Compute (GW) 2028
56.3
All other Compute (GW) 2028
27.1
US Datacenter Power (GW)
51.6
2024
28.3
2028
83.4
Google
8
Microsoft
7
Amazon
3
Meta
2
Oracle
2
Apple
2
Tesla + xAI
1
CoreWeave
0.6
Other Tier 2 / Neoclouds
0.5
Everyone Else
2

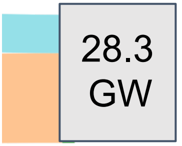
The 2028 Problem



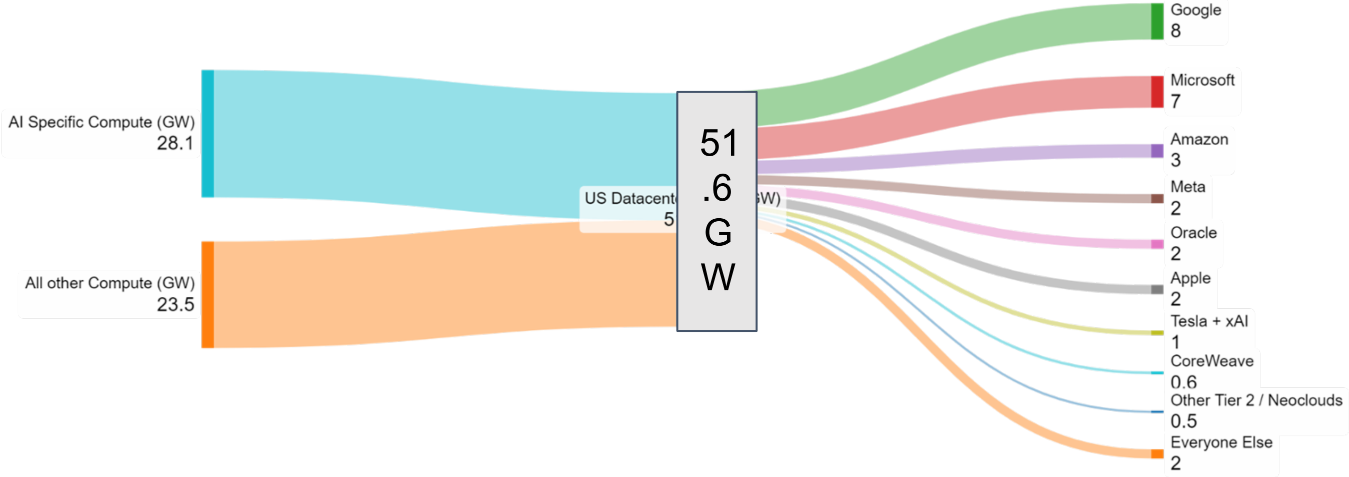
Data Center Power Usage in the United States	Units	2020	2021	2022	2023	2024	2025	2026	2027	2028
AI Data Center Critical IT Power	MW	318	640	1,102	3,332	8,499	16,356	28,140	41,337	56,280
Non-AI Data Center Critical IT Power	MW	14,231	16,395	18,376	19,221	19,798	21,382	23,520	25,637	27,175
Critical IT Power	**MW**	**14,550**	**17,035**	**19,478**	**22,553**	**28,297**	**37,738**	**51,660**	**66,974**	**83,455**
Utilization Rate	%	65%	66%	66%	67%	70%	72%	73%	74%	75%
Critical IT Power Consumed	MW	9,505	11,169	12,826	15,159	19,668	26,983	37,800	49,733	62,688
Power Usage Effectiveness (PUE)	Ratio	1.59	1.56	1.53	1.47	1.40	1.34	1.30	1.26	1.22
Data Center Utility Power Consumed	MW	15,142	17,407	19,660	22,323	27,538	36,263	48,957	62,521	76,684
Data Center Actual Power Usage, per year	**TWh**	**133**	**152**	**172**	**196**	**241**	**318**	**429**	**548**	**672**
As % of United States Power Generation	*%*	*3.3%*	*3.7%*	*4.0%*	*4.5%*	*5.5%*	*7.1%*	*9.5%*	*12.0%*	*14.6%*

W = Watts. kW = Kilowatts. kWh = Kilowatt-hours.

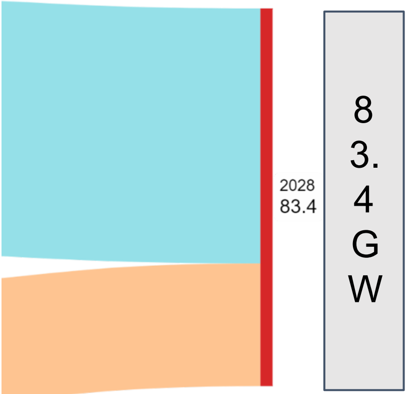
MW = Megawatts. MWh = Megawatt-hours.



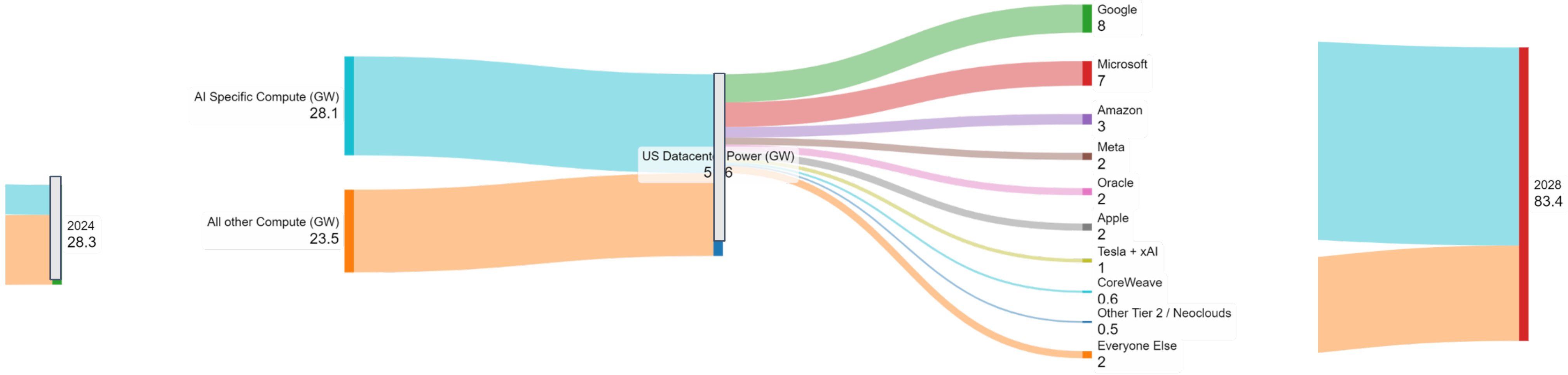
2024



2026



2028


2024
28.3
AI Specific Compute (GW)
28.1
All other Compute (GW)
23.5
US Datacent Power (GW)
5 6
Google
8
Microsoft
7
Amazon
3
Meta
2
Oracle
2
Apple
2
Tesla + xAI
1
CoreWeave
0.6
Other Tier 2 / Neoclouds
0.5
Everyone Else
2
2028
83.4