peak

WATTSNEXT WEBINAR

THE RACE TO WIN AI

POWER, GEOPOLITICS
AND THE DC DATA CENTER

The GenAI Big Bang

Expansion of the data center universe

Frank Berry
Analyst
IT Brand Pulse

IT BRAND PULSE™

AI Primer

Generative AI Solar System

ITBRAND PULSE

Power

NVIDIA

Data Centers

Models

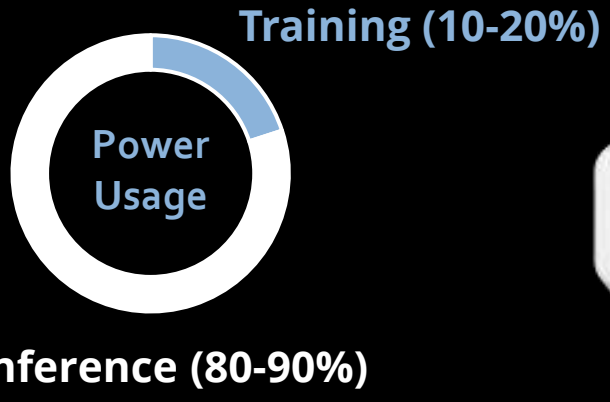
Consumer, Horizontal & Vertical Business Apps & 30B Agents by 2030

Billions of Users

Model Training vs. Inference

Training

Rare, expensive, and intensive process (like teaching)



Inference

Frequent, and conversational process (like working)

Analyzing vast datasets (pre-training) and refining it via human feedback (RLHF).

Responding to prompts with conversation, content, code, summarization, language translation.

Feeding the AI millions of images of roads, signs, and pedestrians to learn to recognize them.

The car's system instantly recognizing a stop sign in real-time while driving.

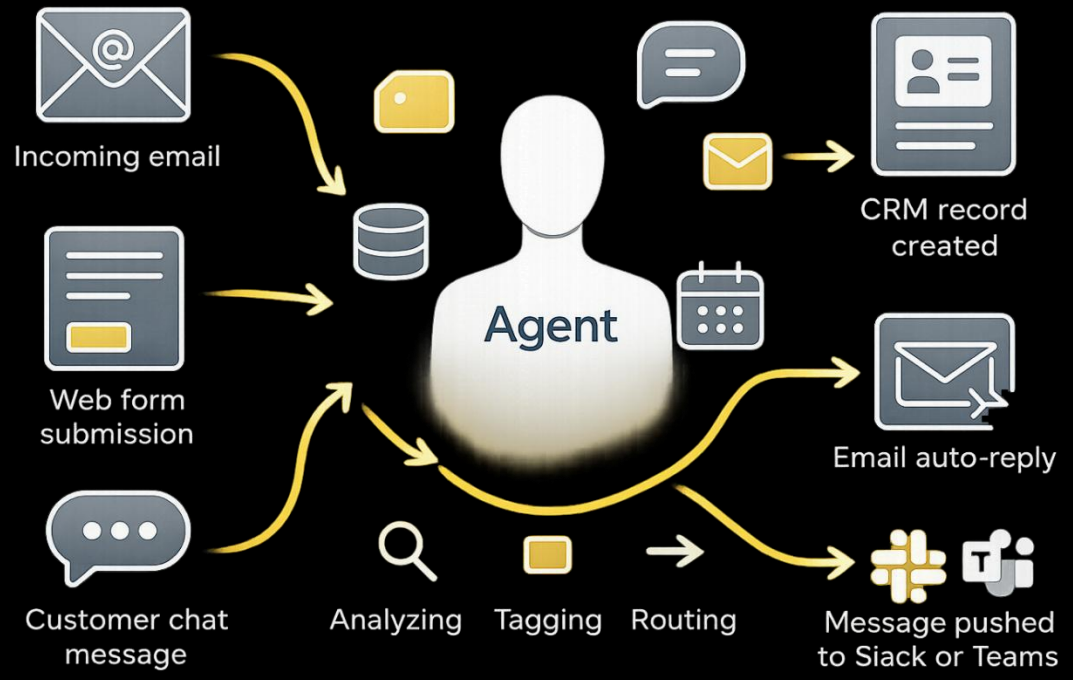
Apps and Agents

AI Apps

Tools that assist with specific tasks. Requires human prompt to initiate every action.

AI Agents (AI Workers)

Systems using planning and external tools to automate multi-step work without human intervention.



AI Currency: Tokens

Token costs for an API that charges $0.002 per 1,000 tokens

Prompt Length	Response Length	Total Tokens	Estimated Cost
50 tokens	150 tokens	200 tokens	$0.0004
200 tokens	400 tokens	600 tokens	$0.0012
1,000 tokens	1,000 tokens	2,000 tokens	$0.0040

Units of data processed by AI models (prompt or response)

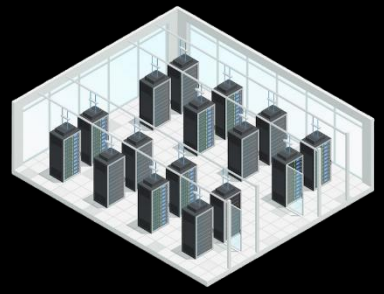
1 token = 75 words (or 1.33 tokens per word).

Models like GPT-4o support up to 128,000 tokens (approx. 300 pages) in 1 context window

Organizations are now managing “tokens per user”

AI Power Consumption

Data Center Site (Stargate)

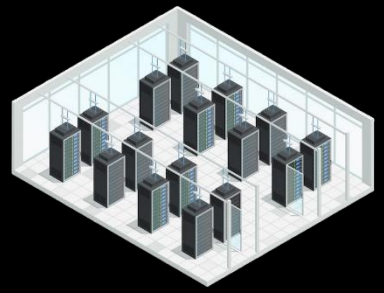


7,000MW

Homes

7,000,00

Data Center Site (Before AI)



100MW

Homes

100,000

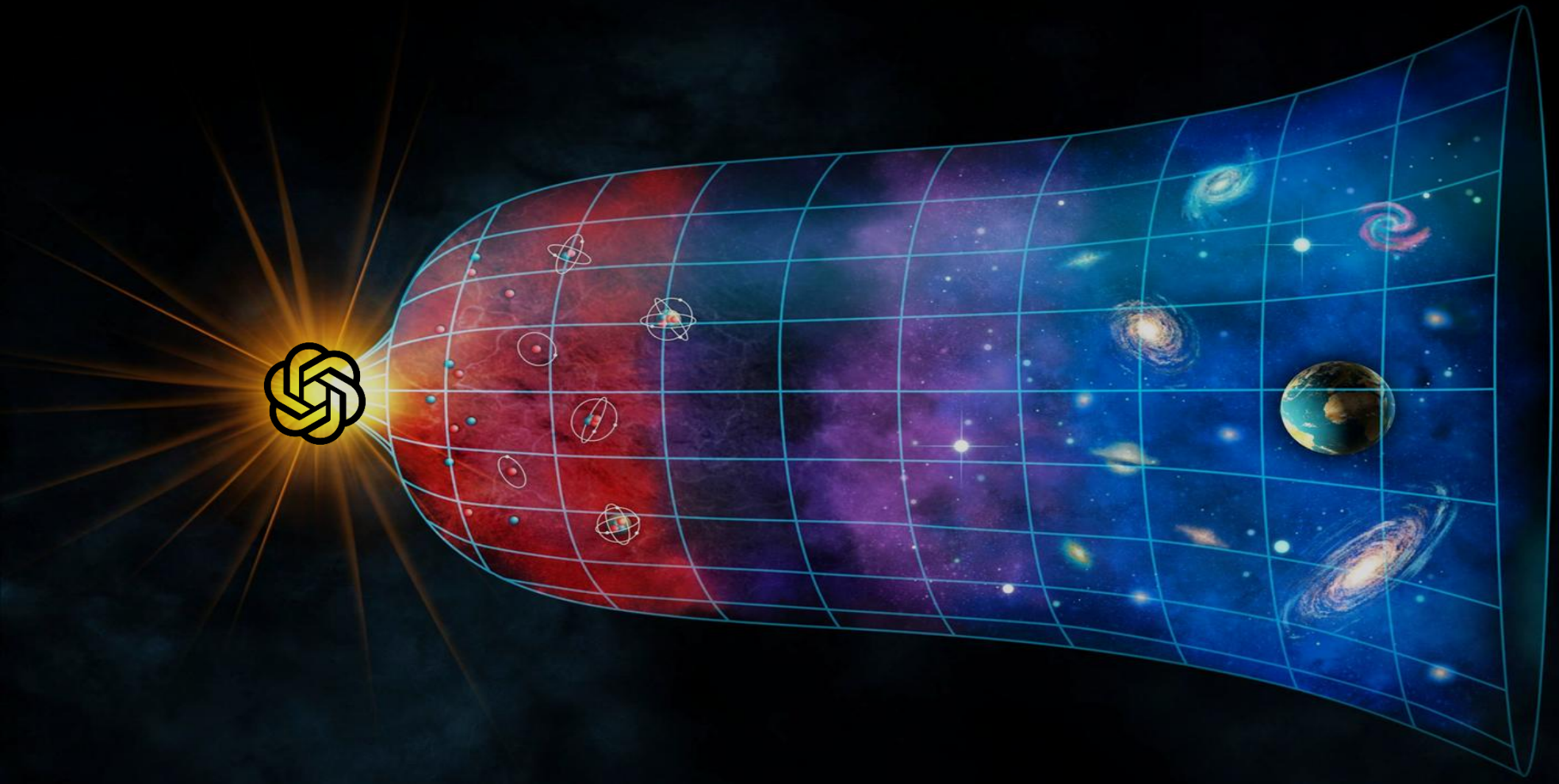

ITBRAND
PULSE
The GenAI Big Bang

"AI, like most transformative technologies, grows gradually, then arrives suddenly."

Reid Hoffman, cofounder of LinkedIn

The Big Bang of Generative AI

Nov 20, 2022



Users

Models

Agents

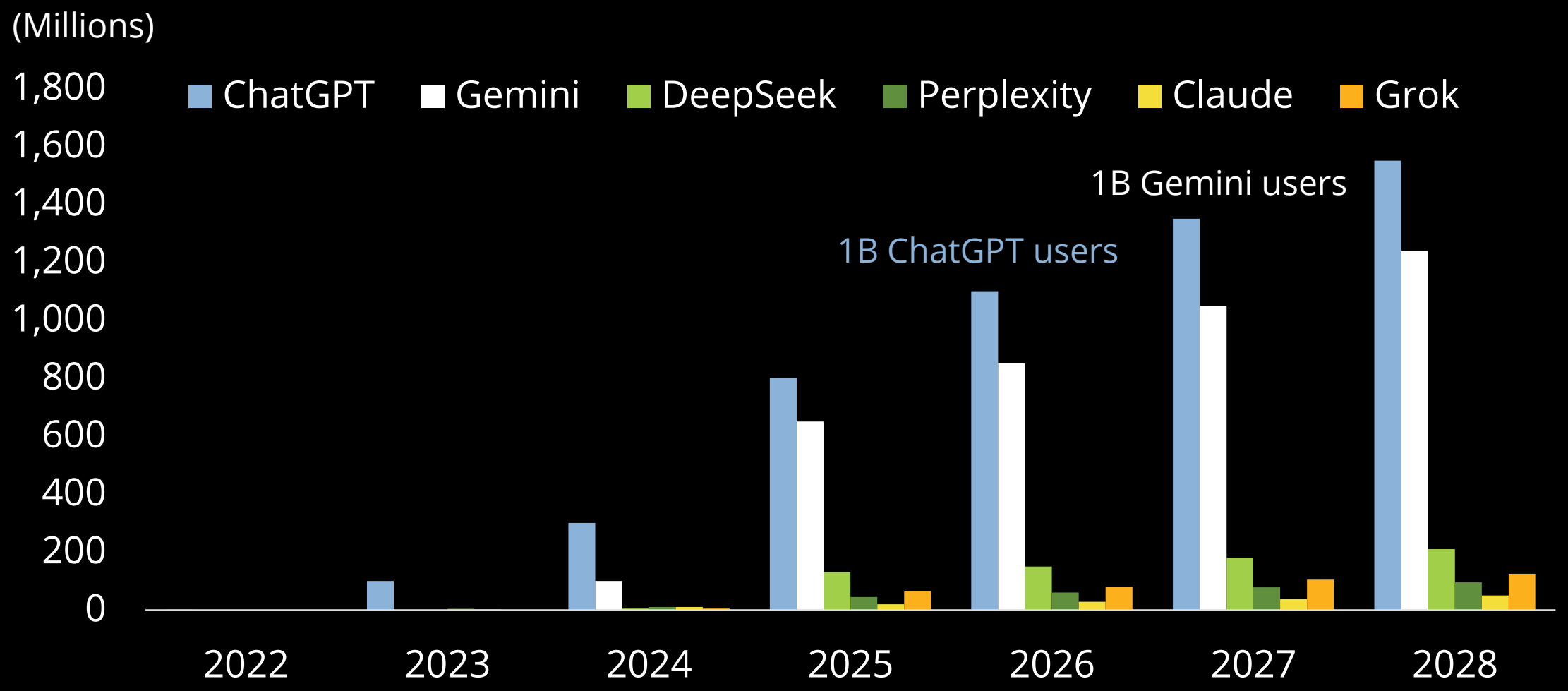
Intelligence

Compute

Power

Investment

1,000x Growth by 2026



Short Term DC Power Usage in U.S.

Almost 4x from 2022 to 2028 for Critical IT Power

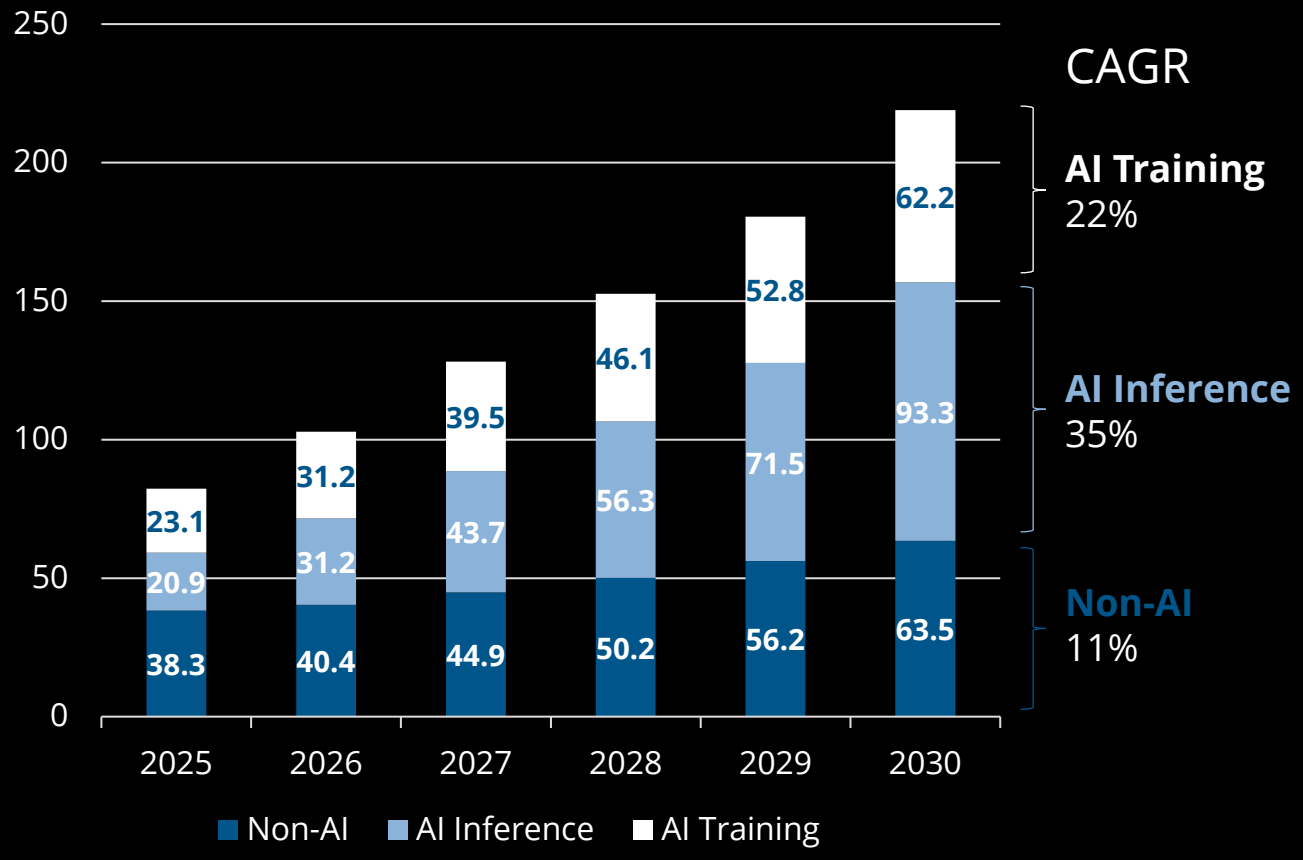
Data Center Power Usage in the United States										
	Units	2020	2021	2022	2023	2024	2025	2026	2027	2028
AI Data Center Critical IT Power	MW	318	640	1,102	3,332	8,499	16,356	28,140	41,337	56,280
Non-AI Data Center Critical IT Power	MW	14,231	16,395	18,376	19,221	19,798	21,382	23,520	25,637	27,175
Critical 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.

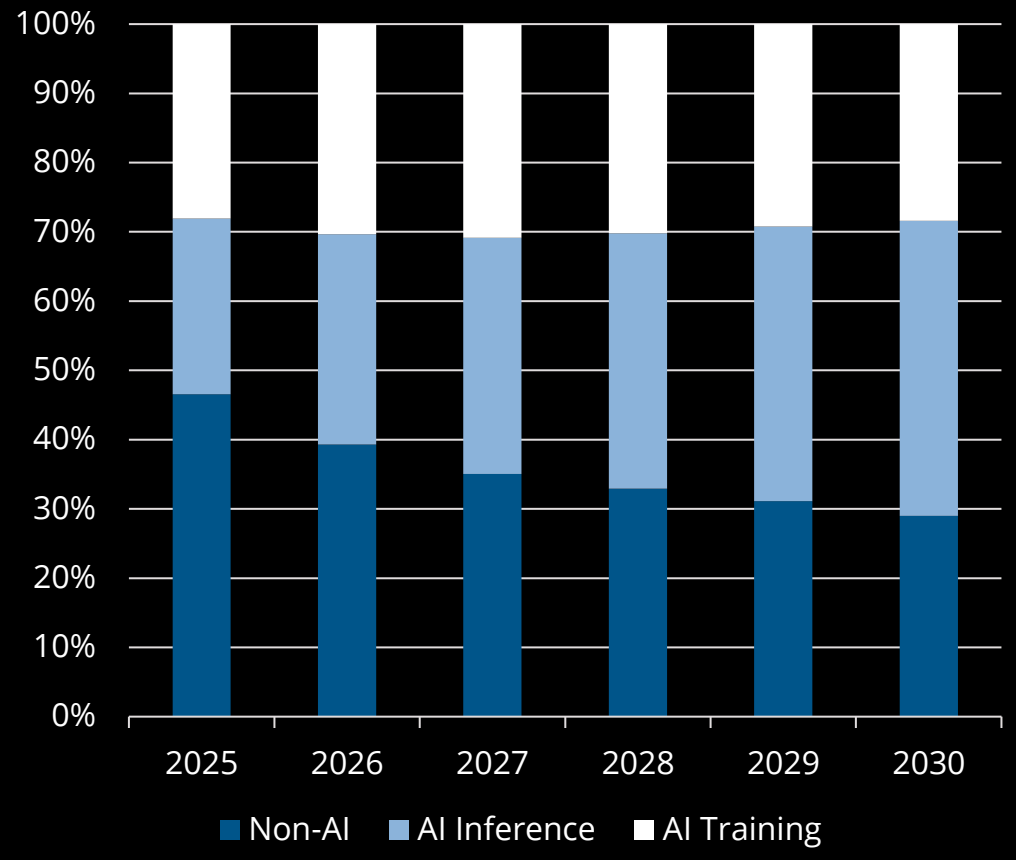
MW = Megawatts. MWh = Megawatt-hours.

Global DC Power Demand

Global Data Center Demand (GW)



Global Data Center Demand (%)



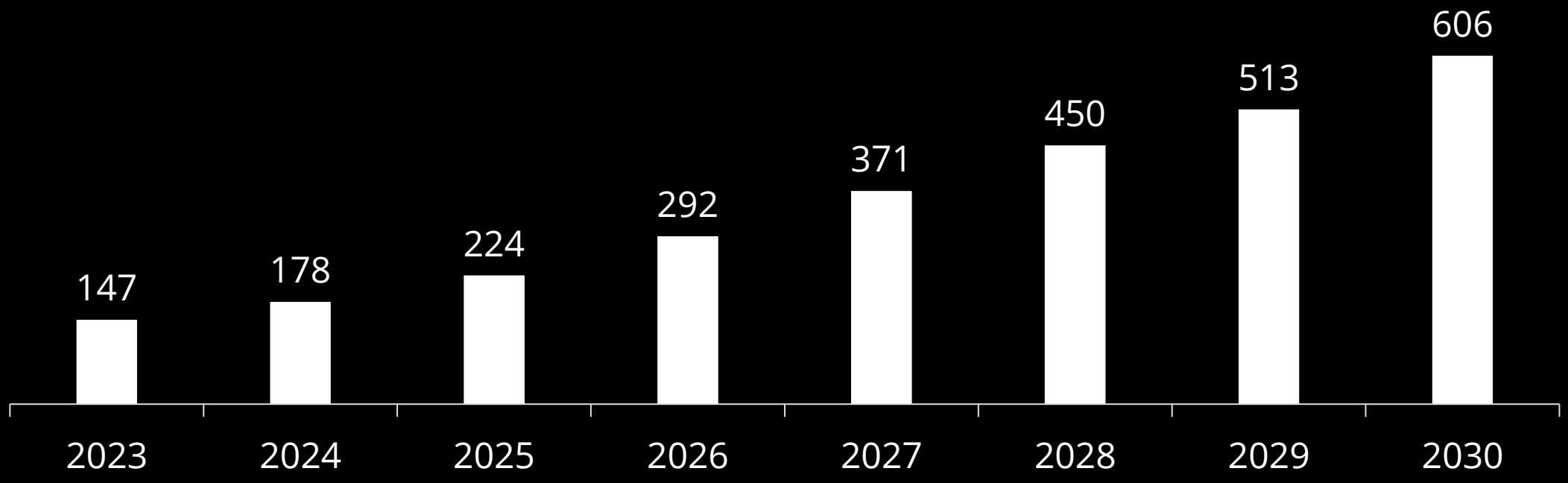
Source: McKinsey Data Center Demand Model

Terawatt-hours of Demand

Share of total US power demand

3.7%	4.3%	5.2%	6.5%	8.0%	9.3%	10.3%	11.7%

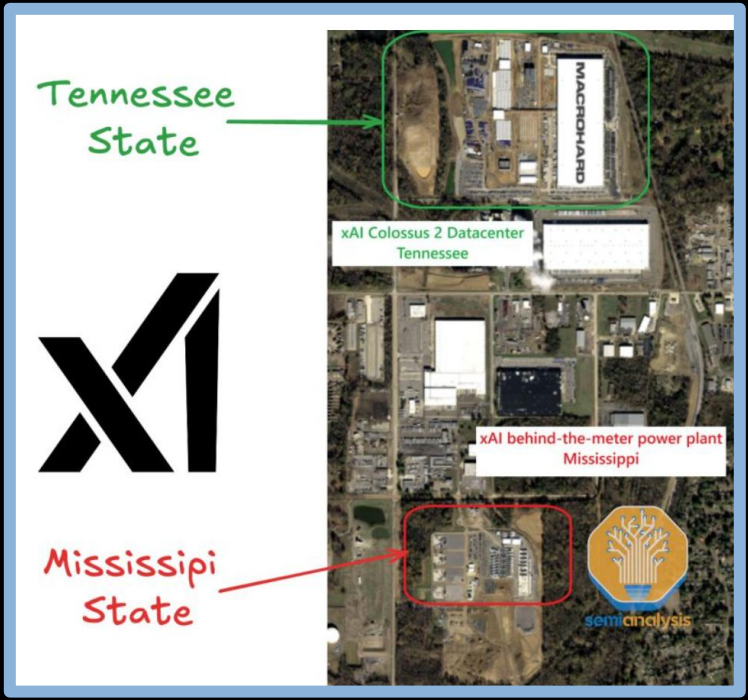
US data center energy consumption (Terawatt hours of demand)



https://www.mckinsey.com/industries/private-capital/our-insights/how-data-centers-and-the-energy-sector-can-sate-ais-hunger-for-power

Govts. & Grids Too Slow

Data Center Operators Securing Power on Their Own



ENERGY • META

Meta orders 10 gas-fired power plants for its Hyperion AI campus in rural Louisiana—more than triple the initial plan

By Jordan Blum
Editor, Energy

March 27, 2026, 5:50 PM ET

Add us on G

Microsoft deal propels Three Mile Island restart, with key permits still needed

By Reuters

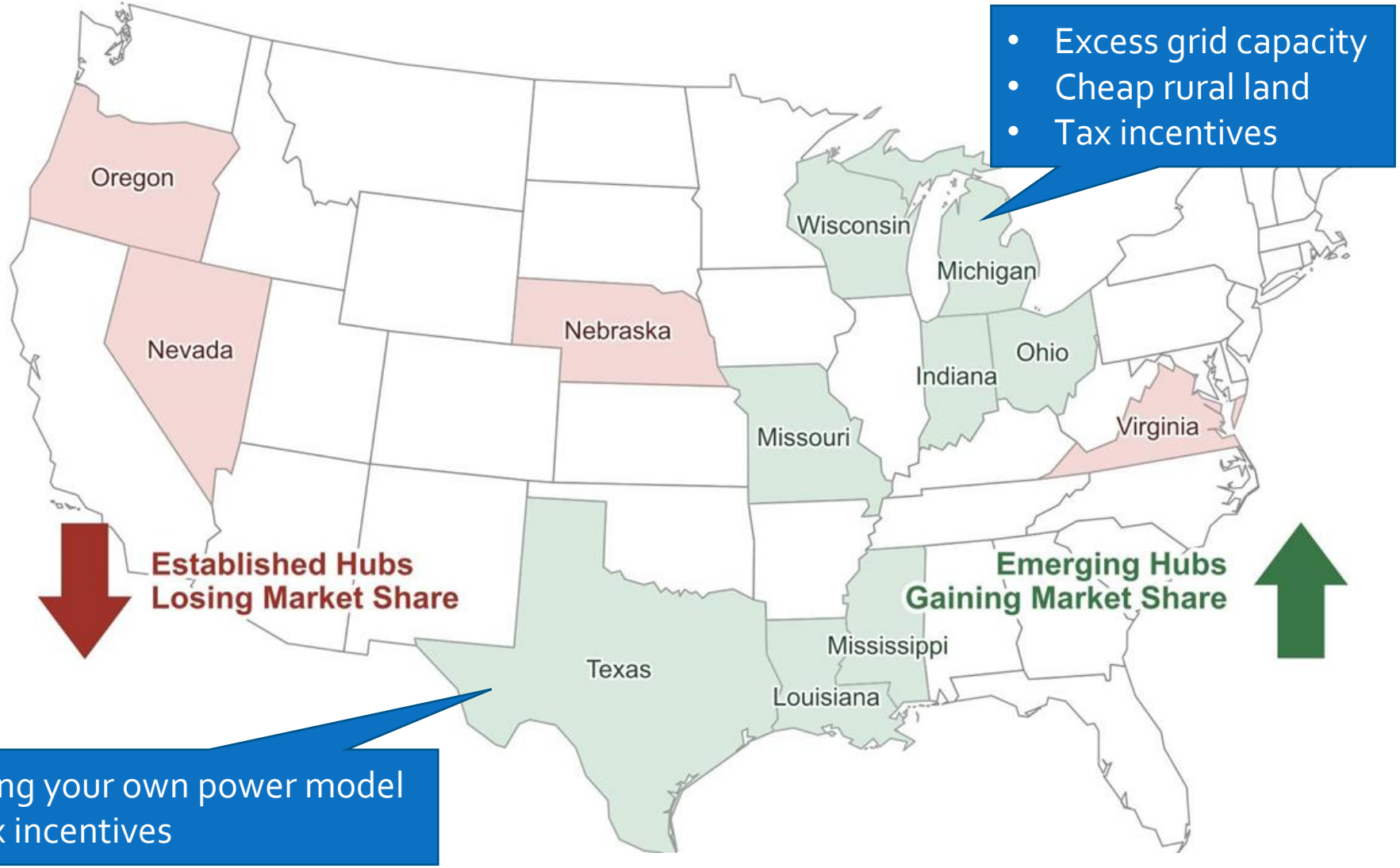
September 21, 2024 9:58 AM MST · Updated September 21, 2024

AI changing the US hub landscape

US Data Center Developments

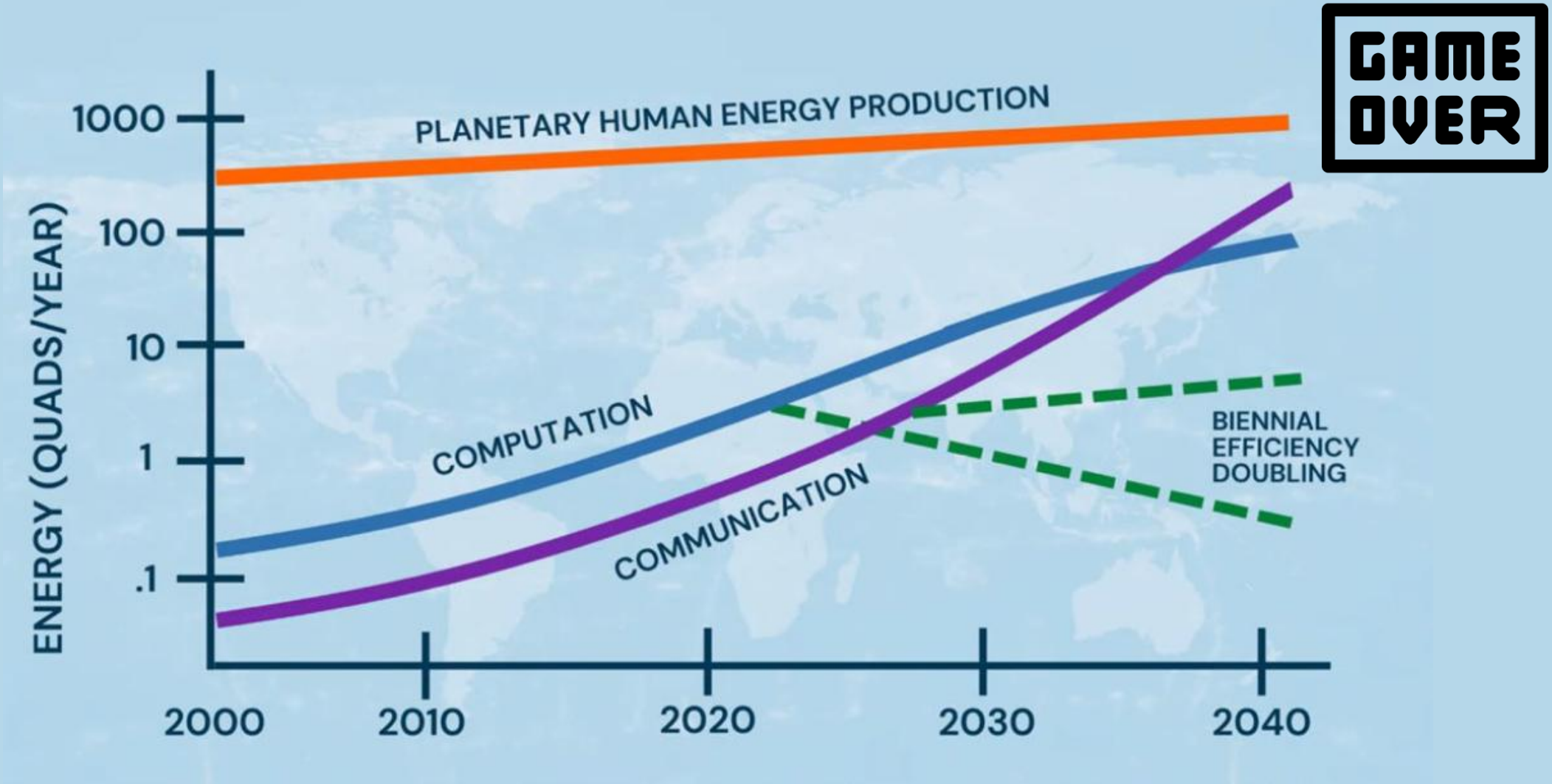
Market share changes

Source: Synergy Research



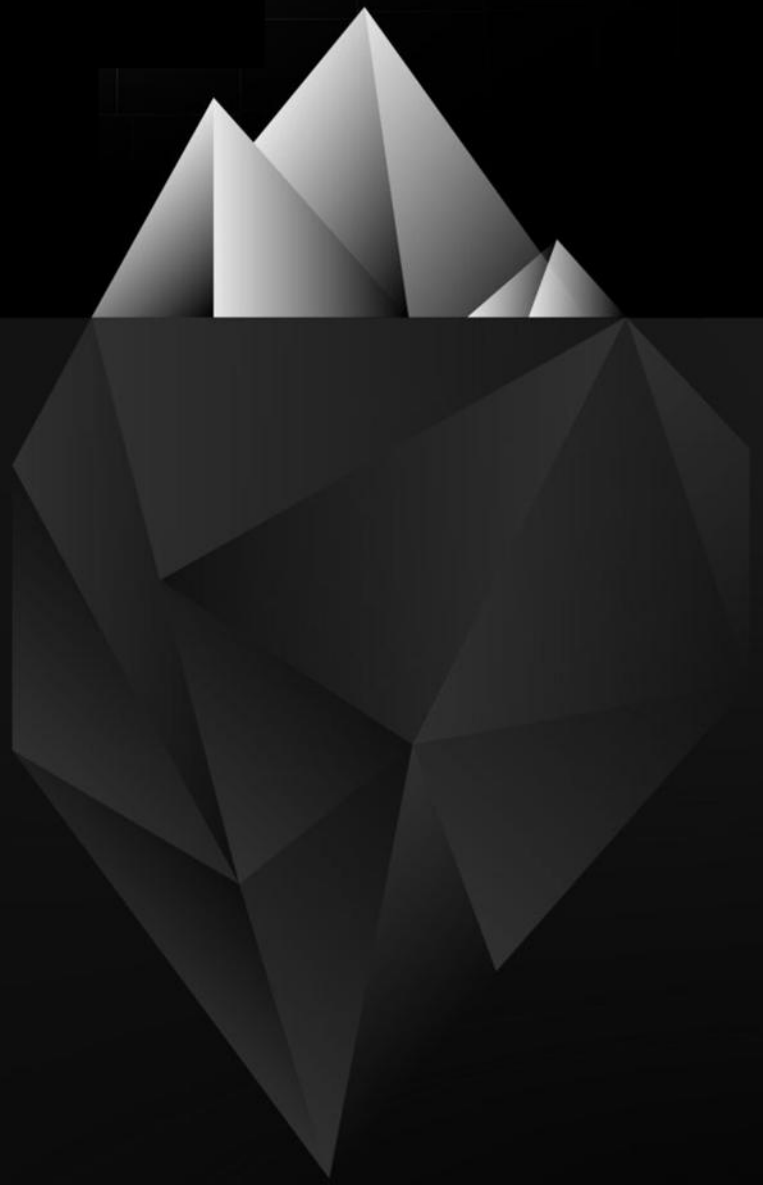

ITBRAND
PULSE
The Efficiency Imperative

Get More Efficient or Run Out of Power



Source: Semiconductor Research Corporation (SRC): https://ees2.slac.stanford.edu/sites/default/files/2023-11/Kaarsberg%20Welcome%20July%2019%202023.pdf

The Inference Iceberg



Peak Chip Specifications

Cost per GPU per hour
FLOPS per dollar

Extreme codesign across networking, memory, storage, software and power

Cost per Token
Tokens per Watt

Source: NVIDIA, https://blogs.nvidia.com/blog/lowest-token-cost-ai-factories/

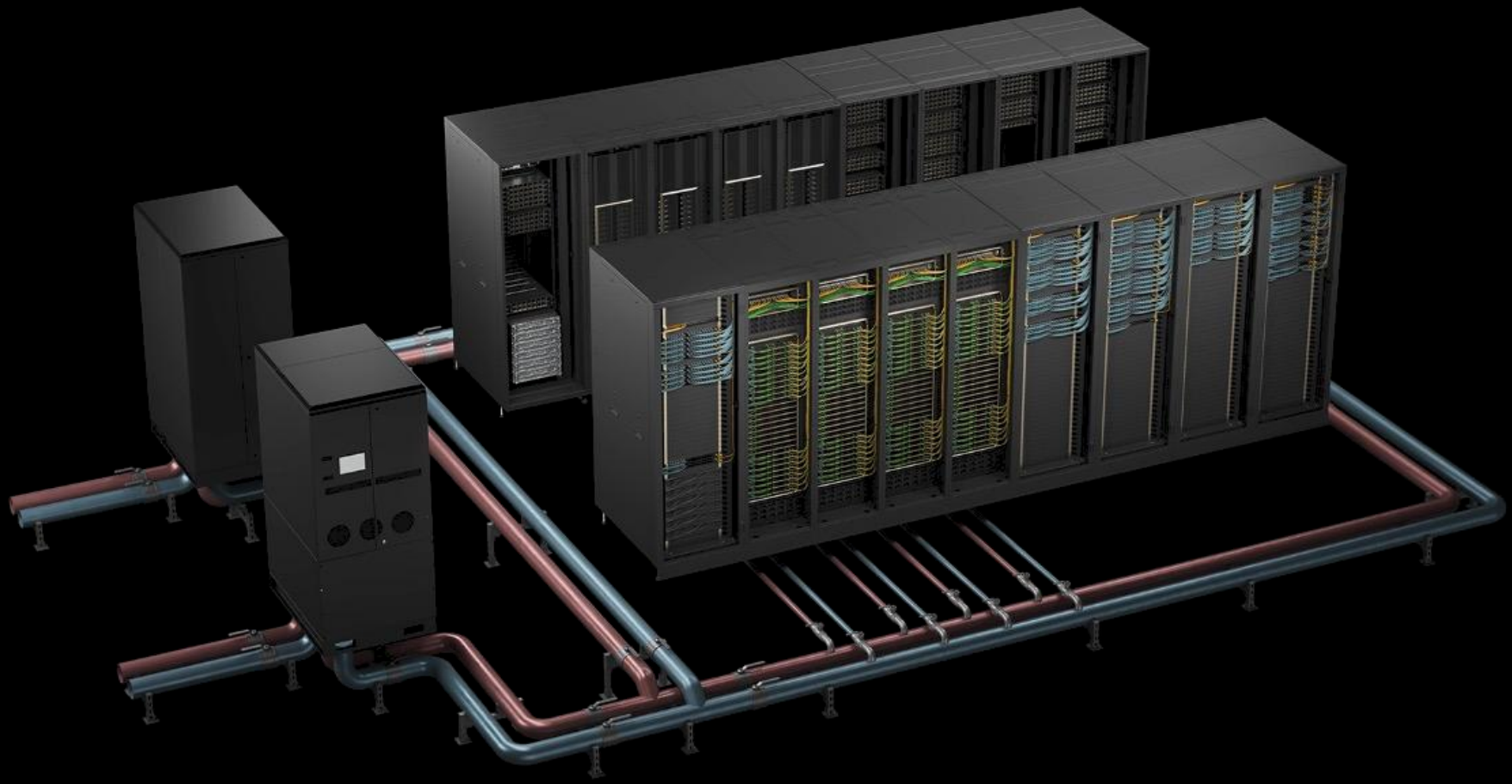
Top 6 ways of improving efficiency

#1 Liquid Cooling

Cooling consumes 30-40% of DC power

Liquid cooling cuts temps by 65%

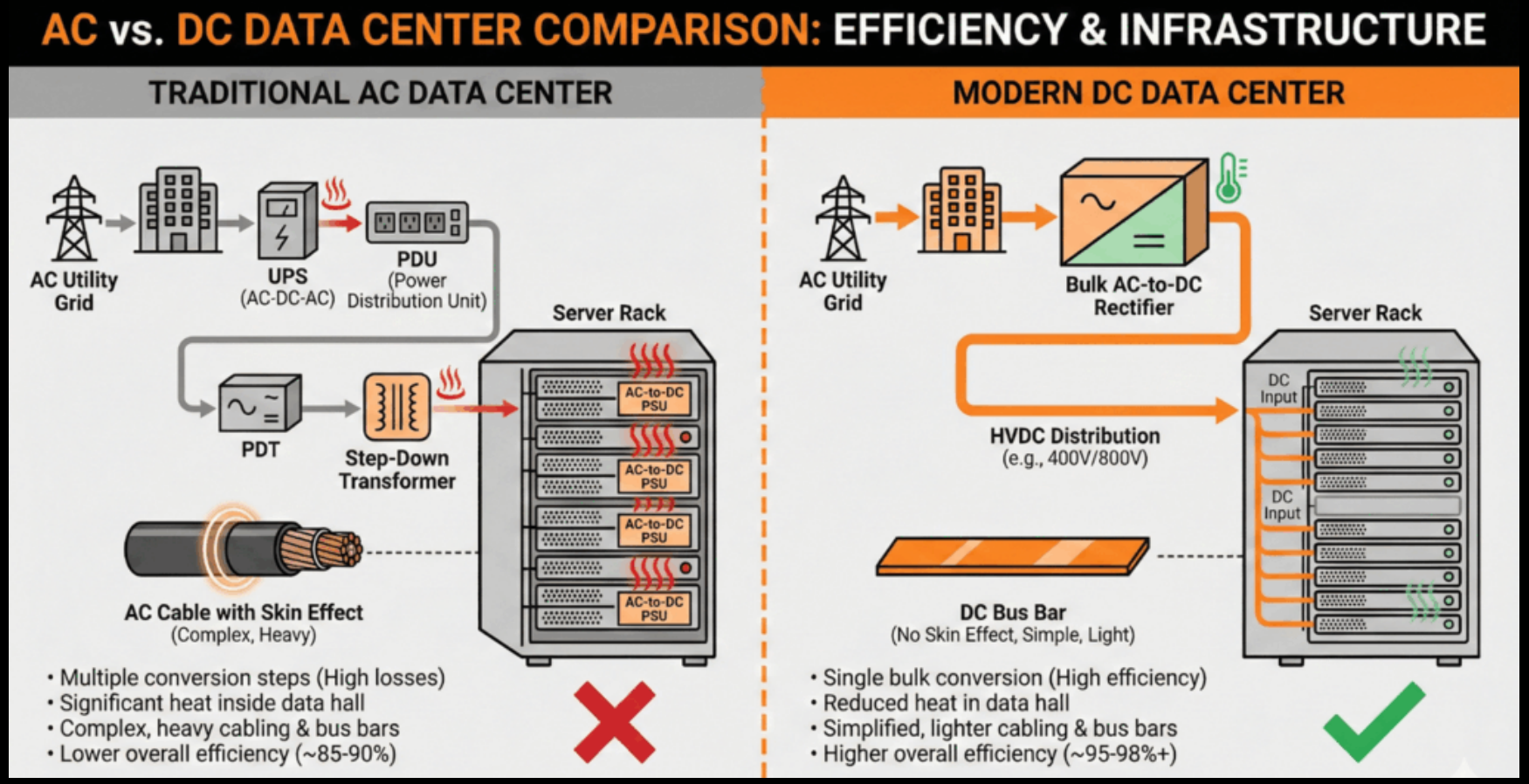
~25–40% of total efficiency gains



Top 6 ways of improving efficiency

#2 Power delivery improvements (UPS, power conversion)

(10-25% of total efficiency gains)



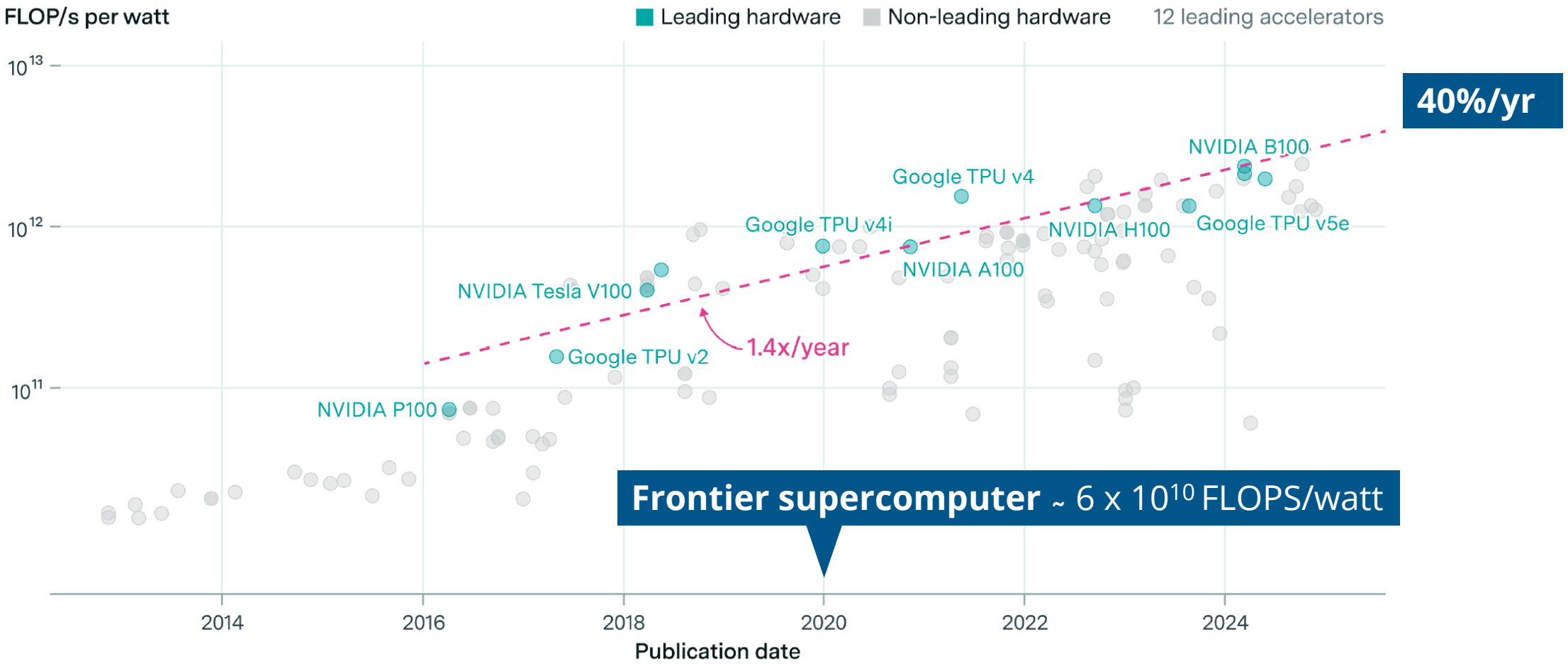
Source: The Fusion Report: https://thefusionreport.substack.com/p/why-data-centers-are-switching-to

Top 6 ways of improving efficiency

#3 More efficient chips

(25-35% of total efficiency gains)

Human brain ~5 x 10^{16} FLOPS/watt



Source: Epoch AI: https://epoch.ai/data-insights/ml-hardware-energy-efficiency

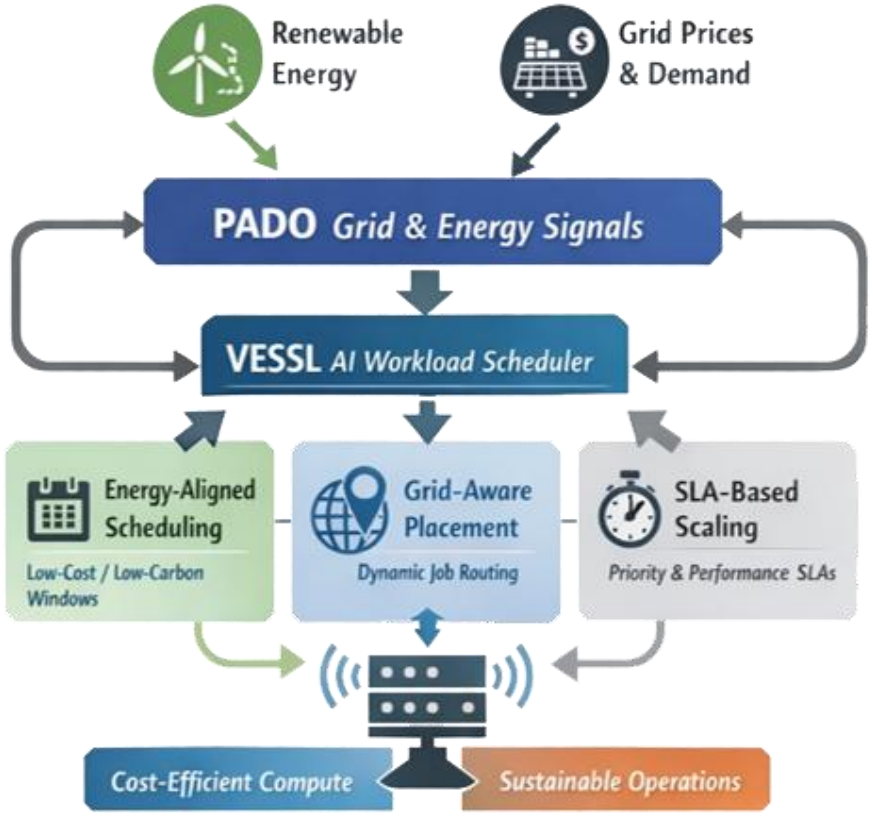
Top 6 ways of improving efficiency

#4 AI-driven workload optimization

(10-20% of total efficiency gains)

Orchestration Stack

Energy-aware AI workflow



PADO & VESSL

Systems dynamically adjust workloads, cooling, and power in real time.

AI is being used to make AI infrastructure more efficient.

Source: it Brand Pulse: https://itbrandpulse.com/industry-first-the-arrival-of-energy-oriented-mlops-and-the-convergence-of-power-and-compute-orchestration/

Top 6 ways of improving efficiency

#5 Renewable and alternative energy integration

(5-10% of total efficiency gains)





Co-locating with solar, wind, hydro reduces grid losses and emissions

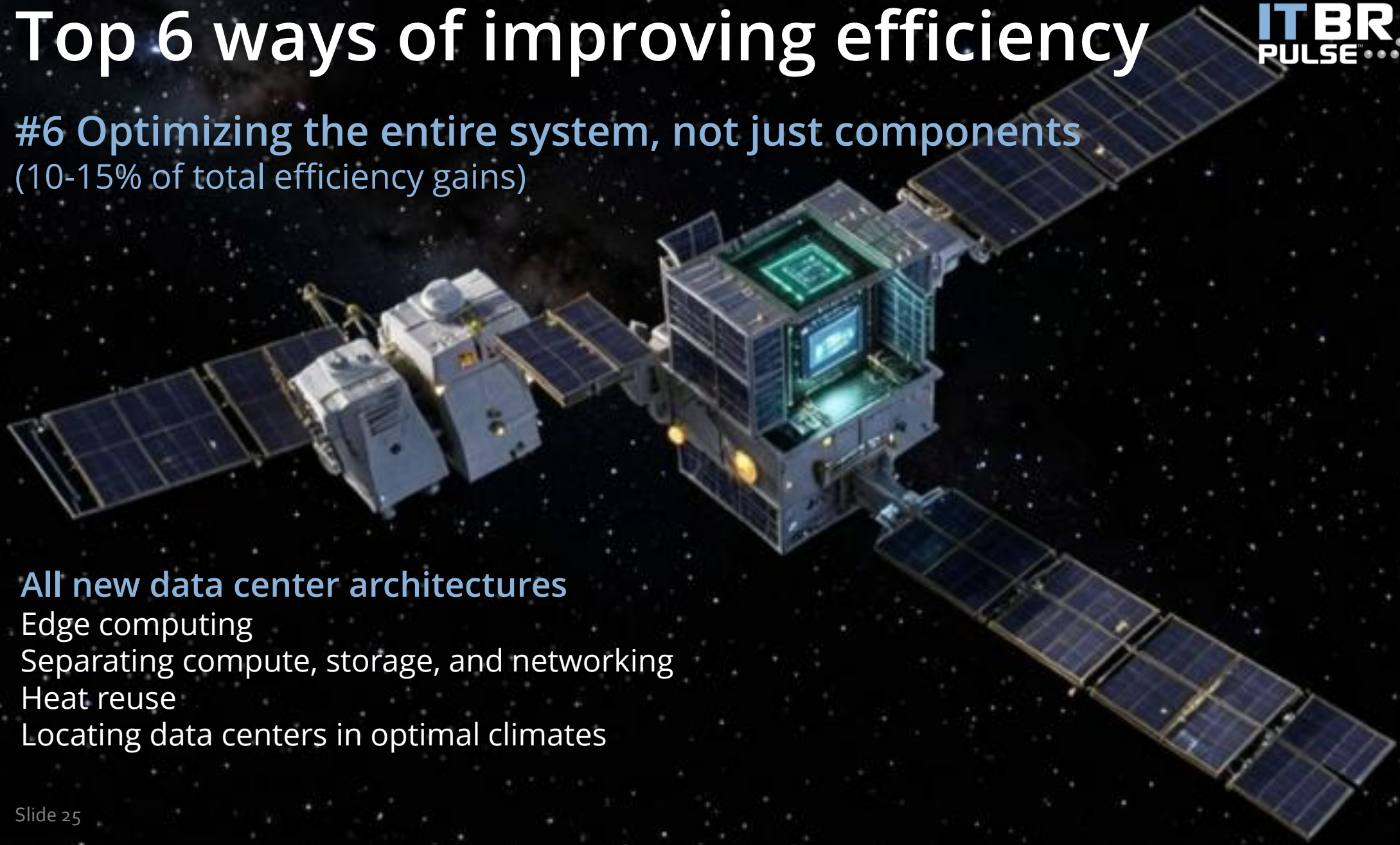
Small modular nuclear reactors (SMRs) and even space-based solar concepts

Energy-aware scheduling - run workloads when power is cheapest/cleanest

Top 6 ways of improving efficiency

#6 Optimizing the entire system, not just components

(10-15% of total efficiency gains)

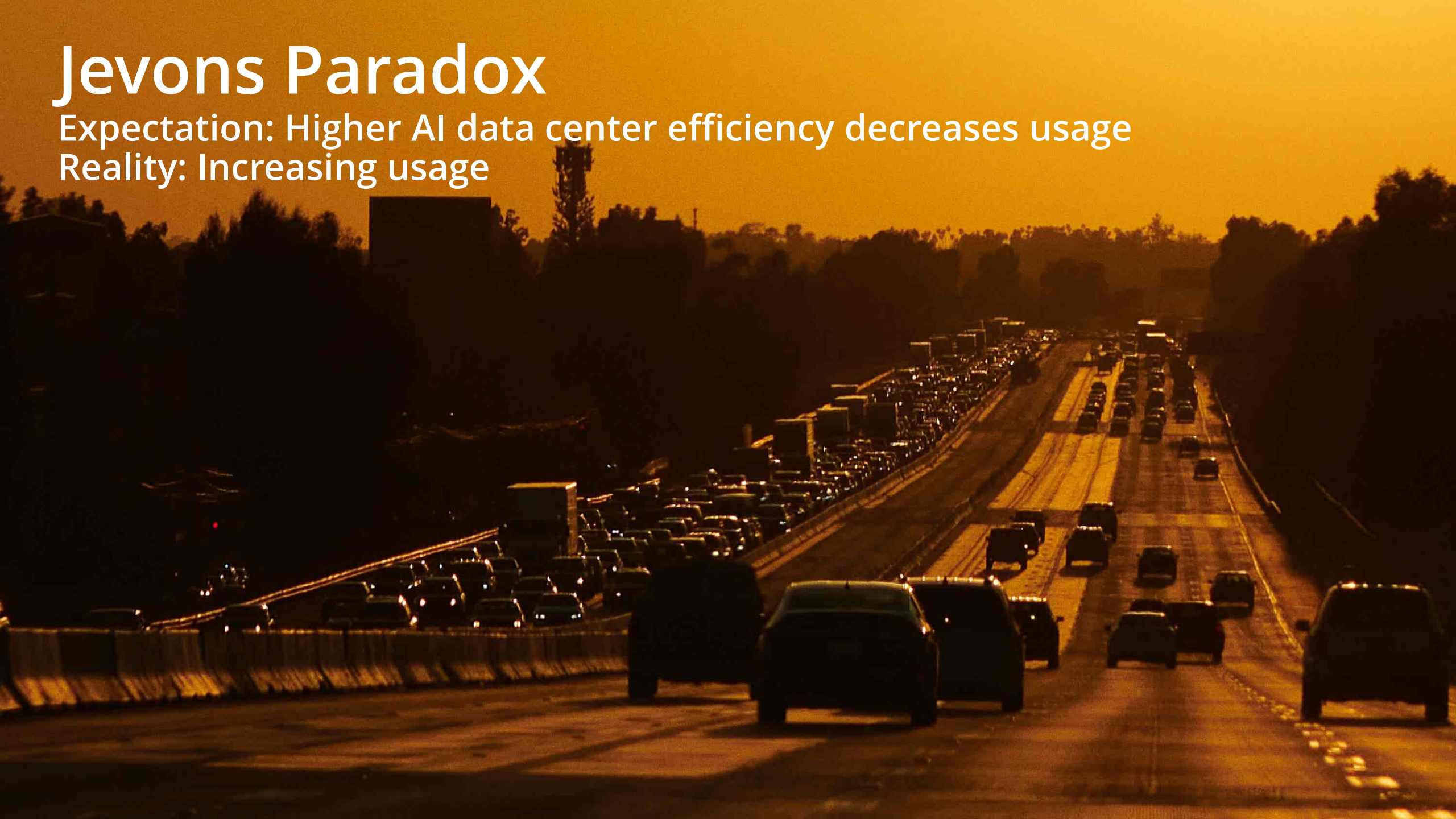


All new data center architectures

Edge computing
Separating compute, storage, and networking
Heat reuse
Locating data centers in optimal climates

Jevons Paradox

Expectation: Higher AI data center efficiency decreases usage
Reality: Increasing usage




IT BRAND
PULSE™