How Silicon Valley and Energy Companies Must Both Win for AI to Scale

Jieul Jang, Senior Director @ Hanwha Qcells

Executive Summary

AI's growth depends equally on efficiency gains and renewable deployment

The Efficiency Reality

AI chips deliver **2–3x efficiency gains** per generation, but demand grows **10x faster**

The Energy Scale

In the States, **176 TWh today → 400–450 TWh by 2028**

The Speed Solution

Solar + storage: 18 months vs 7–10 years for gas

The Partnership Imperative

Silicon Valley + Energy sector partnership prevents AI power wall

The AI Energy Reality

One AI Server Rack = 38 Tesla Charges Every Day

120kW

Power Consumption[1]

(GB200 NVL72 Rack)

Drains Tesla Model Y in > 1 hour[2]

38

Daily Tesla Charges

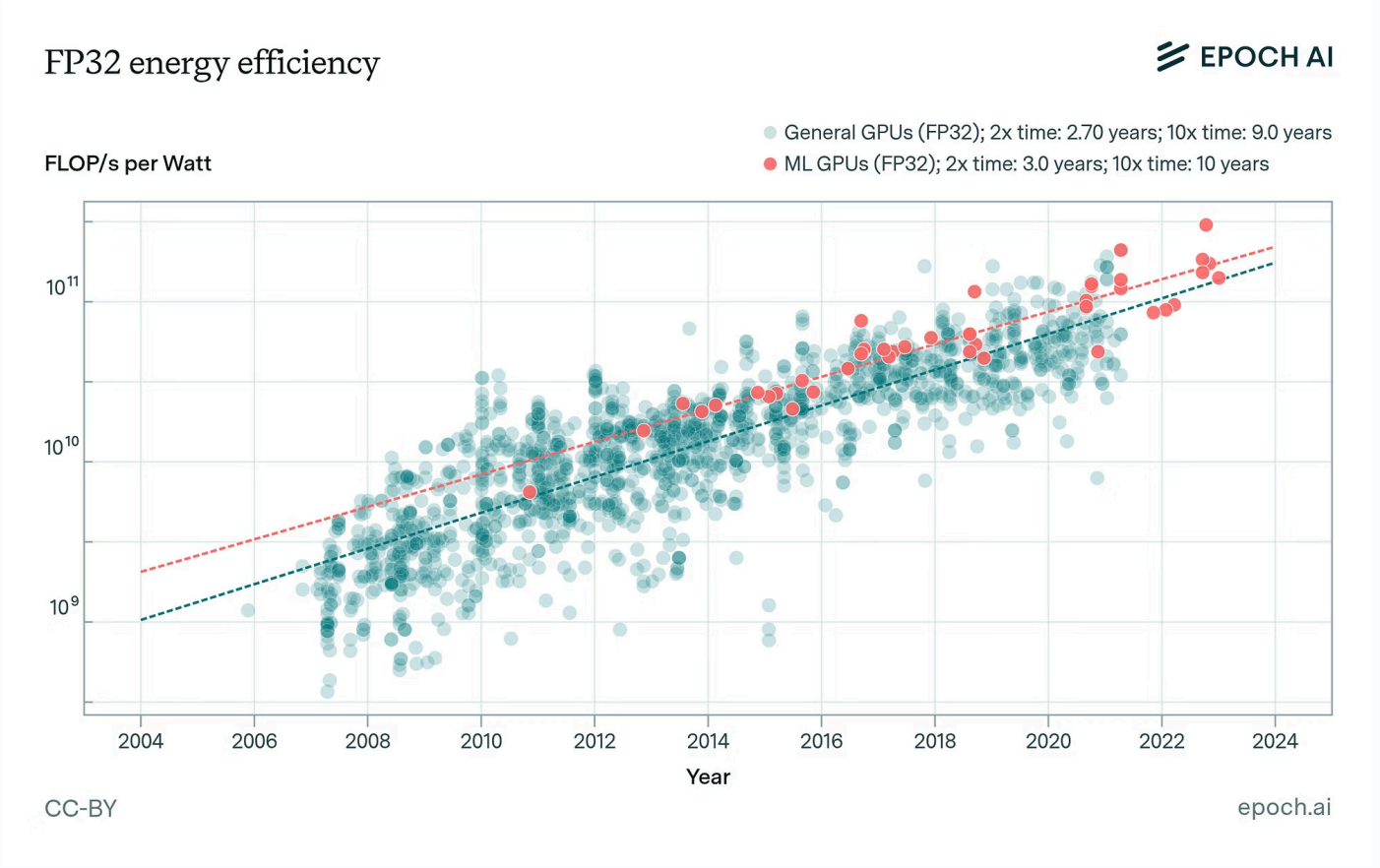
Driving 12,000 miles everyday

Yet, this is one of the most energy-efficient AI server systems

The Efficiency Gains

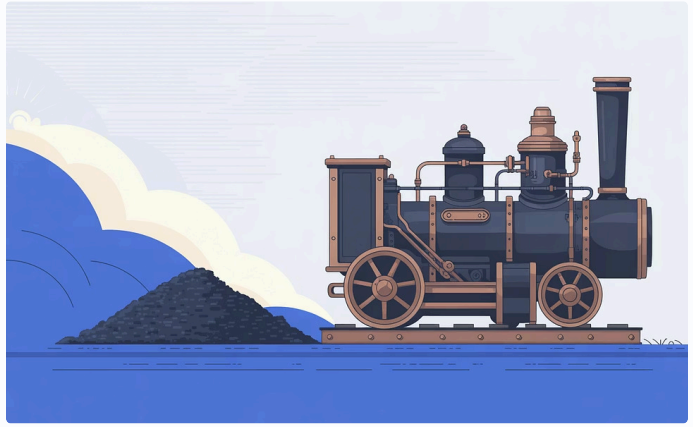
AI Chips: ~70x More Efficient Since 2008

Using ~1.5% of the energy for the same work compared to 15 years ago.[1]



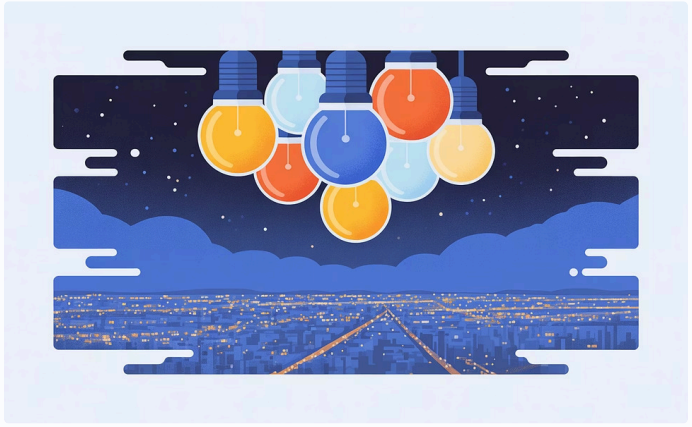
The Jevons Paradox

Why Better Efficiency Leads to MORE Energy Use, Not Less



Coal Engines 1860s

Efficiency doubled → Coal use increased 10x



LEDs 2000s

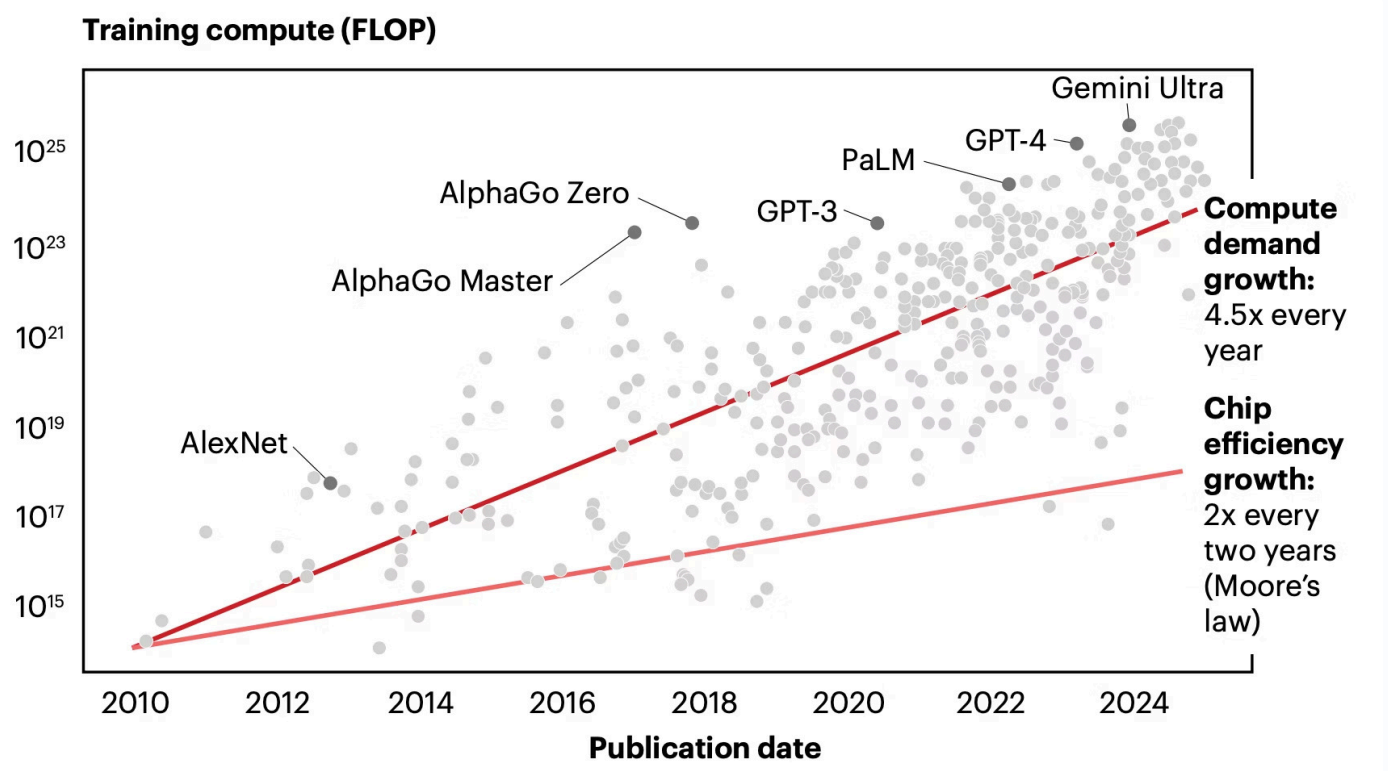
10x more efficient → Lighting energy use still grew

AI Today

98% more efficient → more total energy usage

The Brutal Math

Efficiency Growing 2x, But Demand Growing +10x[1]



Notes: Chip efficiency growth not shown to exact scale, with the rate of growth intended to be illustrative; FLOP=floating point operations, which are the number of calculations a system performs
Source: Epoch AI

What This Means

Even WITH Breakthrough Efficiency, We Need at least 2x Grid Capacity for AI by 2028

No Efficiency Gains

+10x energy need
(impossible to meet)

Current Gains

3x energy need
(crisis scenario)

Breakthrough Gains

2x energy need
(still massive challenge)

Current State - The Starting Line

US data centers use 176 TWh as of 2023
= Larger than New York State's entire electricity use
= 2x South Korea's residential electricity use

4.4% of U.S. electricity - this is AFTER a decade of efficiency gains

Models Keep Growing

Efficiency Fuels Bigger and More Complex Models, Not Lower Power

Parameter Scaling

From **GPT-3's 175 billion** parameters to **GPT-4's estimated 1.7 trillion**, models have rapidly expanded. While future parameter growth may plateau, other factors drive demand.

Multimodal Capabilities

Modern AI processes not just text, but also **images, video, and audio**, multiplying the computational load for each interaction.

The Demand Explosion

From 0 to 700+ Million Users in less than 3 Years[1]

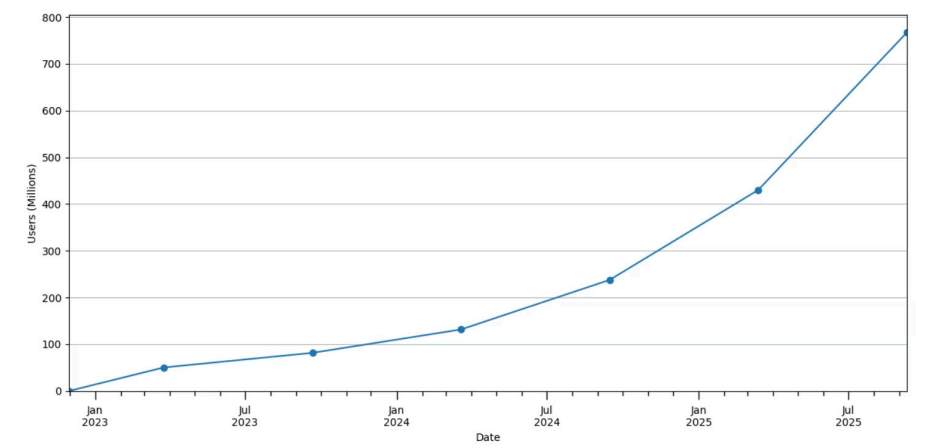


Figure 3: Weekly active ChatGPT users on consumer plans (Free, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–September 2025.

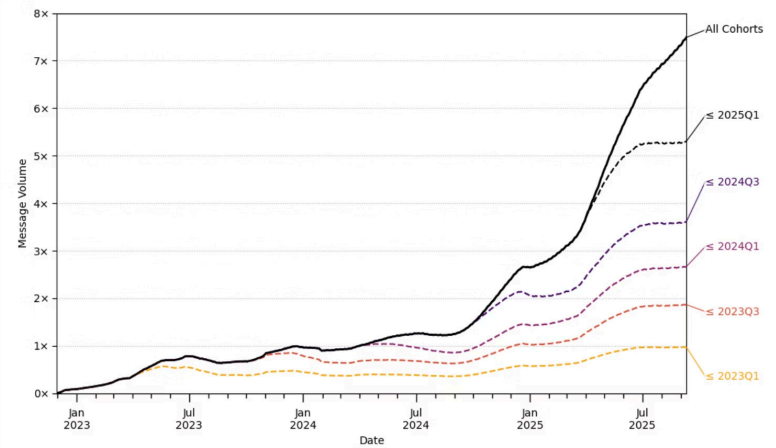


Figure 4: Daily message volumes from ChatGPT consumer plans (Free, Plus, Pro), split by sign-up date of the requesting user. Reported values are moving averages of the past 90 days. Y-axis is an index normalized to the reported value for "All Cohorts" at the end of Q1 2024 (April 1, 2024).

Inference usage jumped ~6x in one year

Mega Data Centers = City-Scale Demand

AI demand is arriving faster than ever

Meta Hyperion

5 GW of planned capacity, equivalent to powering **5 million homes** in a major city like New York.

OpenAI Stargate

An ambitious **$500 billion** project targeting **10 GW** total power, matching **New York City's summer peak demand**.

PG&E Grid Requests

Facing **3.5 GW** in new data center requests by 2029, including **1.6 GW for San José alone**.

Rack Math

A single **5 GW** data center translates to roughly **40,000 NVIDIA Blackwell racks**, each a massive power consumer.

The Tech Front - What's Actually Happening

Silicon Valley IS Delivering Efficiency Gains

Nvidia Blackwell

2.5x performance/watt improvement over previous generation[1]

Google TPU

3x improvement in AI training efficiency over 4 years[2]

Liquid Cooling

~40% power usage efficiency improvement in data center operations[3]

Edge Computing

20-30% reduction in central load for certain workloads

Real progress, just not enough to offset 10x demand growth

Time-Shifting

Running AI Training When the Sun Shines

40%

Workload Shifting

Google shifts this percentage of workloads to high-renewable periods

20%

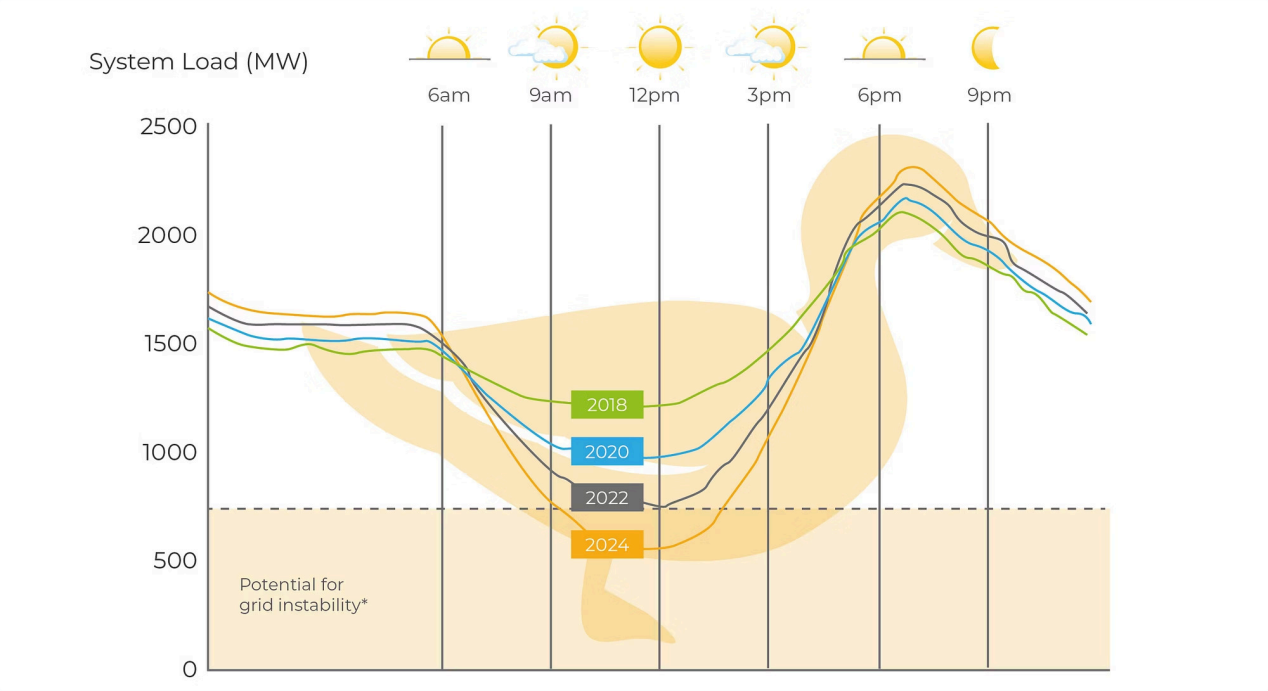
Energy Reduction

Achievable reduction with no new capacity investment required

But this strategy only works for training, not real-time inference

The Duck Curve Problem Remains

AI Inference Runs 24/7, Solar Doesn't



Peak Solar Generation

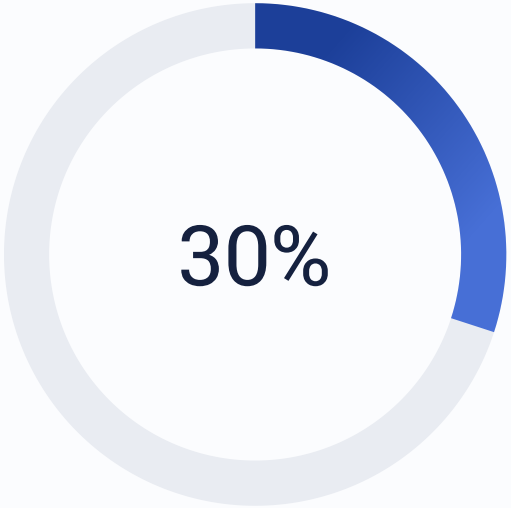
12:00 PM - Maximum renewable output during midday hours

Peak AI Inference Demand

6:00-10:00 PM - When users are most active, solar is unavailable

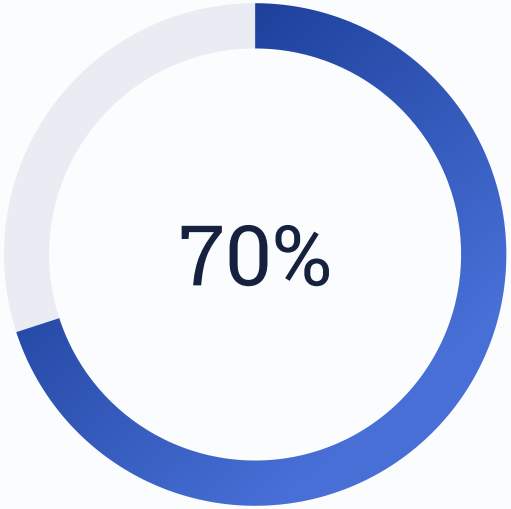
Why Distributed Isn't Enough

Edge Computing Helps but Doesn't Solve the Problem



Maximum Load Reduction

Potential reduction with aggressive edge deployment



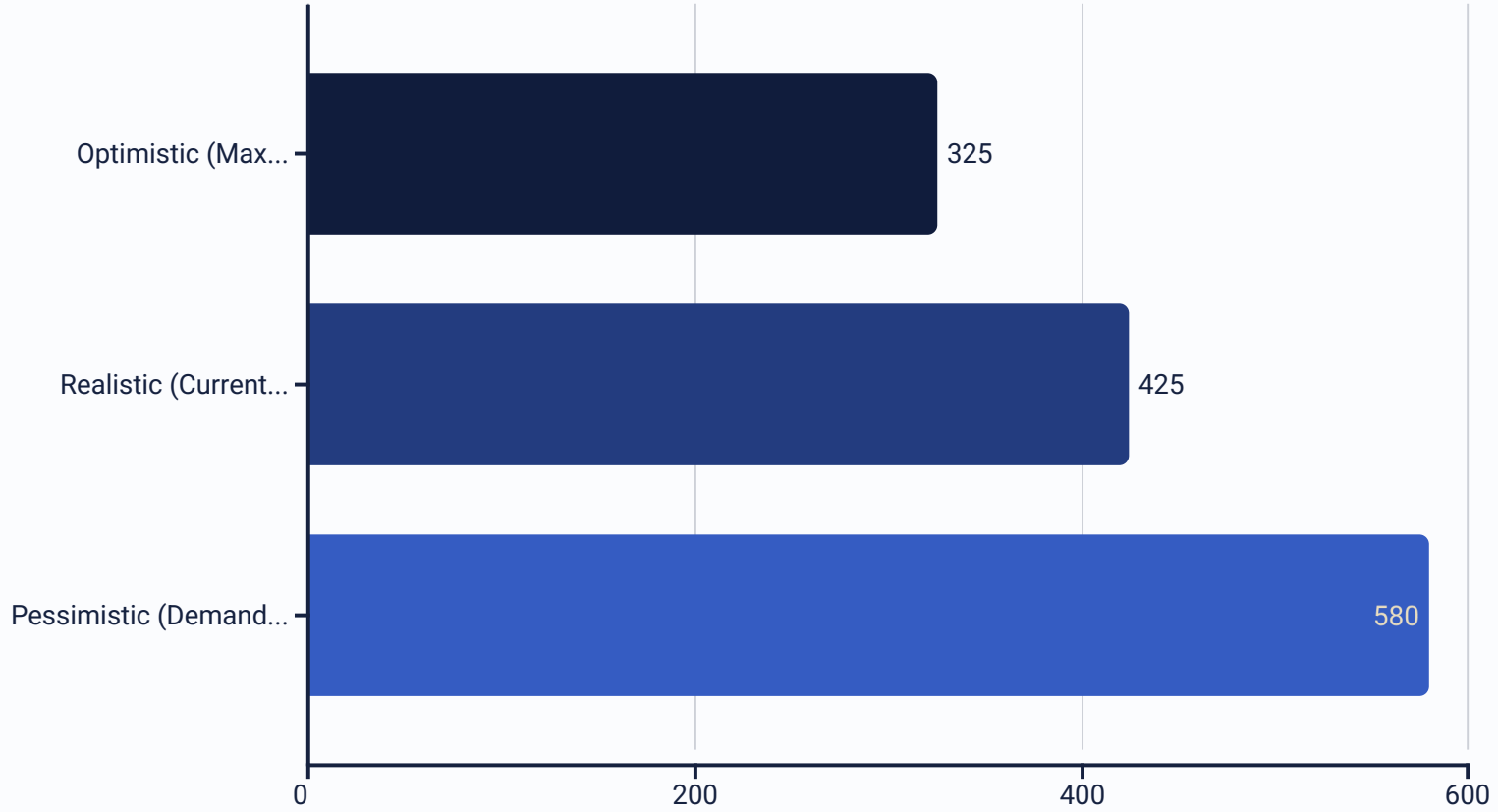
Still Centralized

Training and large model inference remain in hyperscale data centers

Edge computing reduces but doesn't eliminate the fundamental challenge

The 2028 Reality

Most Realistic Scenario: 400-450 TWh = ~10% of total U.S. electricity



Even the optimistic scenario equals adding **Texas-sized electricity demand** to the grid

The Infrastructure Reality

Tech Moves in Quarters, Energy Moves in Decades

0.5

New AI Model

6 months development cycle

2

New Chip Generation

2 years from design to production

1.5

New Solar Farm

18 months from decision to operation

8.5

New Gas Plant

7-10 years total timeline

10

New Nuclear Plant

10+ years development and construction

Solar's Speed Advantage

18 Months from Decision to Electrons

Permitting Phase

3-6 months for regulatory approvals and site preparation

Construction Phase

9-12 months for installation and commissioning

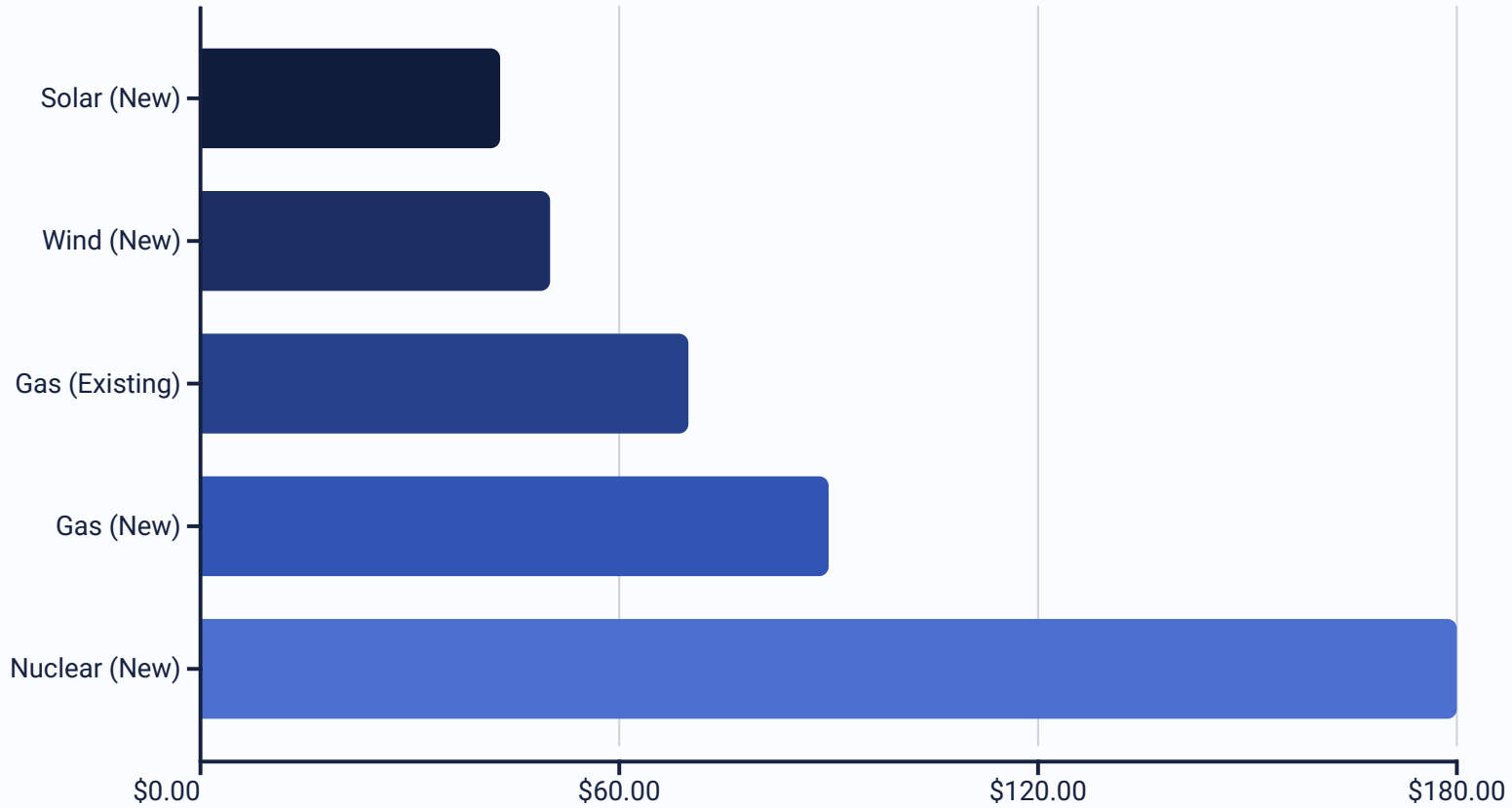
Grid Connection

3 months for interconnection and testing (if interconnection is secured)

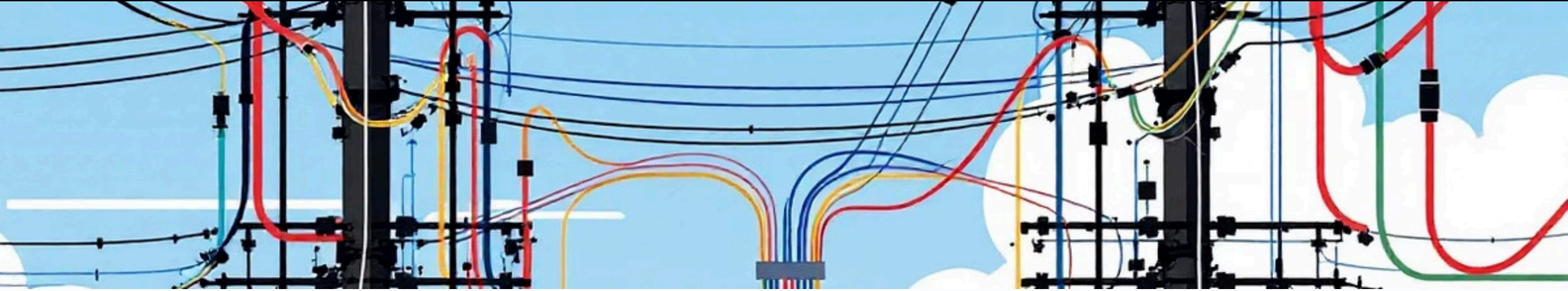
The ONLY option that matches AI's timeline demands

The Cost Reality

Cheaper to Build New Solar Than Run Existing Gas



Economics and speed align perfectly for renewable deployment



The Interconnection Bottleneck

AI doesn’t just need power plants — it needs plugs.

2,600 GW

Stuck in Queue

New energy projects awaiting connection, nearly double U.S. capacity.

5 Years

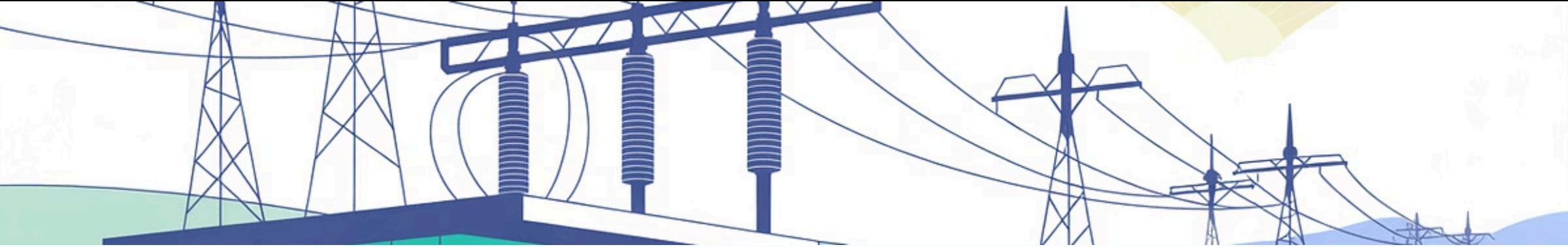
Average Wait Time

Projects face crippling delays to connect to the grid.

70%

Projects Withdrawn

High attrition due to prolonged interconnection bottlenecks.



Grid Enhancement Reality

$180B Needed for Transmission Alone

Transmission Lines

$180B for new high-voltage connections

Substation Upgrades

$50B for capacity and smart grid integration

Smart Grid Technology

$30B for advanced control and monitoring systems

Even with massive renewable deployment, the grid itself needs fundamental upgrades to handle distributed generation and AI's concentrated loads.

Unlocking the Grid — Private-Sector Levers

If You Can’t Change the Rules, Change Your Strategy

Site Smarter

Retiring fossil plants, HV hubs (230-500 kV), hydro/nuclear regions

Design Smarter

Co-locate storage, hybrid solar+storage

Build Smarter

Share upgrades with neighbors, co-fund grid tech

Operate Smarter

Time-shift training, acquire stalled queue positions

The Hyperscaler Strategy

Buy Everything: Renewable + Nuclear + Efficiency

Amazon Web Services

34 GW renewables contracted + nuclear SMR investments for comprehensive portfolio approach

Microsoft

23 GW renewables + Three Mile Island nuclear restart partnership

Google

24/7 carbon-free energy by 2030 across all operations and data centers

Meta

18 GW renewables contracted + on-site generation pilots

The Innovation Sprint

VC Money Flooding Into Efficiency & Generation

$7.6B

AI Chips[1]

Investment in next-generation processors and AI accelerators

$3.3B

Energy Storage[2]

Long-duration batteries and grid-scale storage solutions

$2.1B

Energy Management[3]

Efficiency, EV charging, and clean energy tools

$4.5B

Solar[4]

Solar generation and solutions to optimize generation

Two-front war requires two-front investment across the entire stack

Three Uncomfortable Truths

What Silicon Valley Needs to Accept

Efficiency Won't Save Us

Jevons Paradox means efficiency gains increase total consumption, not reduce it

Interconnection is the choke point

2,600 GW stuck in queues, 5–7 year waits, >70% dropouts

Other Options Too Slow

10+ years for nuclear, 7-10 years for gas doesn't match AI's 2-3 year growth cycles

Only massive renewable deployment can scale at the required speed

For Entrepreneurs - The Real Opportunities

Build for the Efficiency-Demand Gap

1

12+ Hour Storage Solutions

Technologies that can bridge the overnight gap when solar isn't available

2

Grid Flexibility Software

AI-driven systems that optimize renewable energy distribution and consumption

3

Hybrid Renewable Controllers

Smart systems managing solar, wind, and storage as integrated units

4

Advanced Cooling Technology

Solutions for 200kW+ server racks that dramatically reduce cooling energy

For Corporate Leaders - Strategic Imperatives

Lock in Energy Now to Win the AI Race

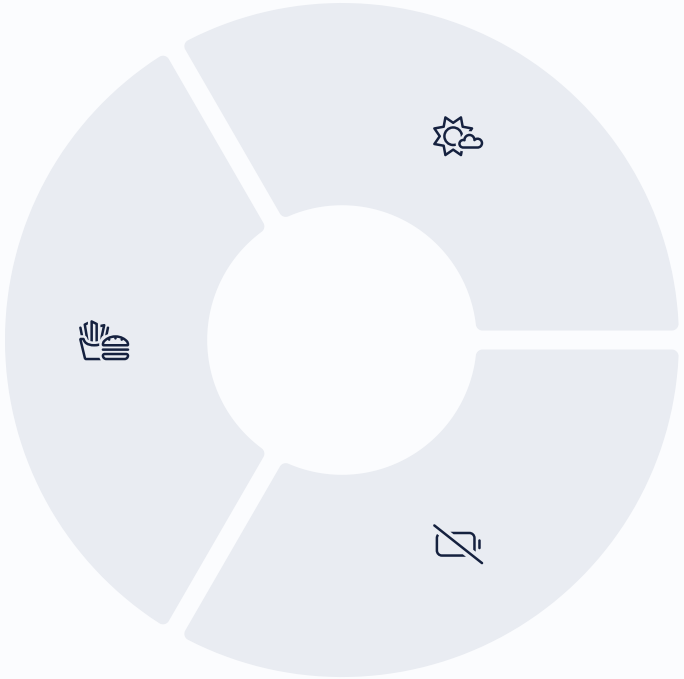
- **Secure PPAs**
  Lock in long-term renewable power purchase agreements before prices rise
- **On-site Generation**
  Evaluate distributed solar and storage for critical data centers
- **Co-location Strategy**
  Plan data center locations near renewable energy sources
- **Time-shifting**
  Implement workload scheduling to match renewable availability

For Investors - The Double Play

Bet on BOTH Efficiency and Capacity

Efficiency Tech

Chips, cooling, and optimization S/W



Renewable Generation

Solar, wind, and deployment companies

Enabling Technology

Storage, grid tech, and integration

Single bets on either efficiency or capacity will fail. The winning strategy requires balanced exposure to both sides of the energy equation.

The Race We Must Win Together

Tech is doing its part. Energy must match.

Tech

Deliver 3x efficiency by 2030 ✓ (on track)

Energy

Deploy 3x capacity by 2030 ⚠ (behind)

The companies solving BOTH sides win the decade.

Thank you