SURVIVING THE SURGE

AI Infrastructure at Scale Without Breaking the Planet

AI does not fail because of lack of ambition. It fails when physical limits are ignored. Scaling AI without breaking the planet is not only possible—it is becoming a competitive advantage. Success in 2026 belongs to operators who solve physics, not those who optimize spreadsheets.

Version: 3.1 — January 2026 Update

By: Helder Lira | HML Services

EXECUTIVE SUMMARY

Artificial intelligence is scaling faster than the physical systems that support it. Compute capacity, model complexity, and inference demand continue to accelerate, while power generation, grid connections, water availability, and permitting processes move at infrastructure speed. The central constraint on AI growth in 2026 is no longer capital or silicon availability, but **time-to-power**: the ability to secure electricity, cooling, land, permits, and social license fast enough to deploy infrastructure responsibly.

2,300 GW
US interconnection queues (up 53% from 1,500 GW in 2023)

4% → 12%
US electricity consumption by 2030 (tripled in 6 years)

399B gallons
Texas data centre water usage by 2030 (6.6% of state's total)

THE SCALE PROBLEM HAS ACCELERATED

The Crisis Is Quantified (January 2026)

Metric	Previous (Dec 2025)	Current (Jan 2026)	Change
Grid Interconnection Queue	1,500 GW	2,300 GW	+53%
PJM Timeline	5-7 years	8+ years	Doubled
US Data Centre Power	2-3% of electricity	4% of electricity	+33%
Project Success Rate	~30%	20% (2000-2018)	-33%

In practice, electricity access has become the schedule driver.

Critical Markets Under Strain:

- Northern Virginia: 7-year interconnection delays, grid moratoriums emerging
- Texas: Utility-imposed transparency requirements on duplicate requests
- Phoenix: Extreme water stress + 12% grid reserve margin = dual constraint
- Singapore: 6% reserve margin + extreme water stress + tropical climate

WATER: FROM QUIET CONSTRAINT TO REGULATORY BARRIER

Water is now a regulatory barrier requiring specific technical solutions for project approval.

Regulatory Environment Shift (2025-2026)

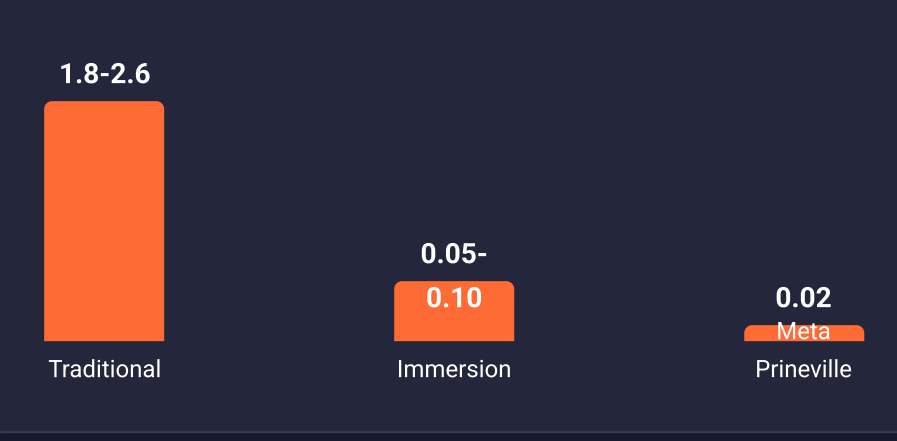
December 2025 - Emerging
Water identified as parallel bottleneck

January 2026 - Enforced
Water is now a regulatory barrier

Active Regulations:

- EU Energy Efficiency Directive: Now mandatory and enforced (first deadline passed September 2024)
- Texas SB 7: Active legislation addressing 399 billion gallons by 2030
- Santa Clara County: Recycled water required for new builds
- Johor, Malaysia: 30% rejection rate documented

Water Consumption Comparison (L/kWh):



THE A+ PATHWAY PROJECTION

The A+ Pathway is not a single technology solution. It is a system-level approach built on five reinforcing pillars.

Appropriate Siting
Select locations based on power availability, grid strength, water resilience, and permitting risk

Grid-Aware Architecture
Design facilities that align with grid constraints and support demand shaping

Power Strategy
Treat energy procurement as a core competency with PPAs, storage, and utility partnerships

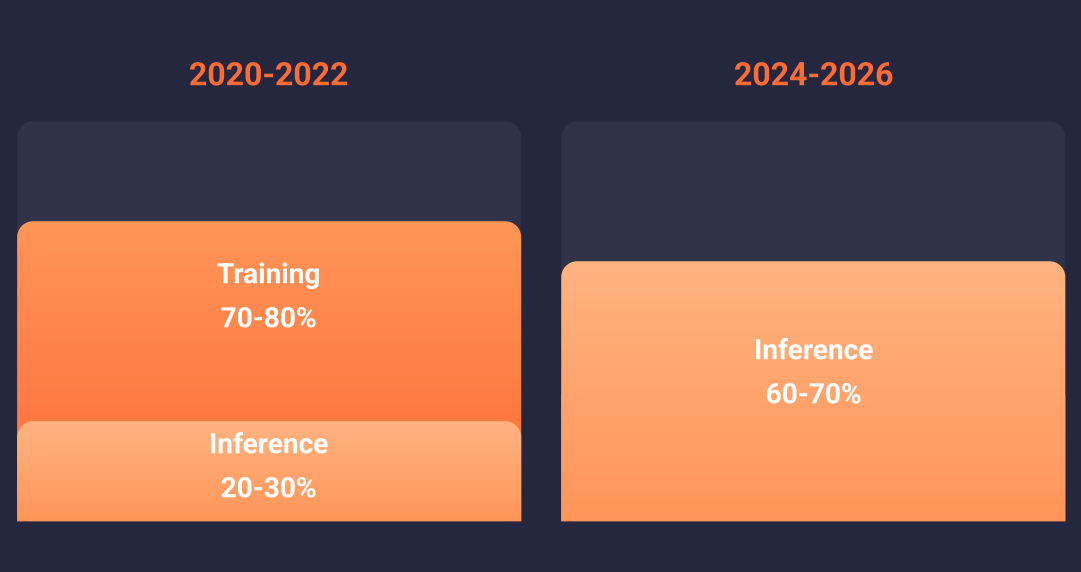
Liquid & Low-Impact Cooling
Adopt cooling strategies matched to climate, water availability, and density

Unified Measurement
Anchor performance claims to recognised KPIs with consistent boundaries

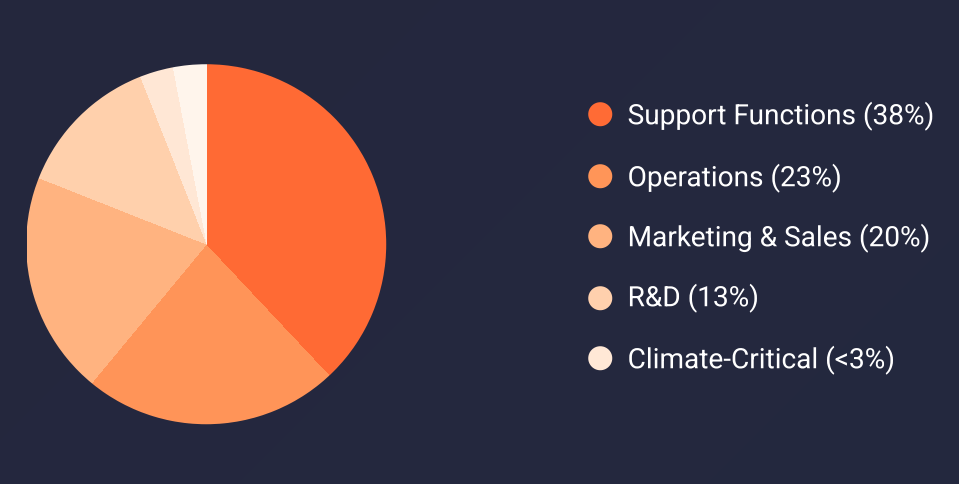
THE MISALLOCATION CRISIS

We are in an AI arms race. But what are we racing toward?

THE WORKLOAD SHIFT: TRAINING → INFERENCE



ENTERPRISE AI BUDGET ALLOCATION (BCG 2025)



THE SCALE OF CONSUMER AI

- ChatGPT: 800M weekly active users (Sept 2025)
- Sora downloads: 12M in first weeks
- Query energy impact: 10x traditional search (2.9 vs 0.3 watt-hours)

THE REALITY: We are allocating the vast majority of AI compute resources to consumer engagement, content generation, and advertising optimization—while climate-critical applications receive a fraction of available resources.
AI COULD mitigate 5-10% of global GHG emissions by 2030 (Google/BCG analysis) **BUT ONLY if we allocate the compute to let it.**

THE CARBON MATH

~500 tonnes CO_2e

Single large language model training (equivalent to: 5 cars × 200,000 km lifetime)

~15 million tonnes CO_2e

Global AI training (2024)

122% CAGR

Inference (ChatGPT queries) growth rate

Energy Impact:

ChatGPT query uses ~10x electricity of traditional Google search (2.9 watt-hours vs 0.3 watt-hours)

GRID INTERCONNECTION CRISIS

In just one month, grid access has transitioned from risk to primary constraint. Projects without secured interconnection are effectively stalled indefinitely.

2,300 GW queued, 8+ year timelines

Primary schedule driver

- PJM Timeline: 5-7 years → 8+ years (doubled)
- Project Success Rate: ~30% → 20% (2000-2018) (-33%)

CASE STUDIES

Google Mayes County
AI-controlled HVAC + 24/7 wind PPA

1.06	155	0.2
PUE	g/kWh CO_2e	L/kWh water

Meta Prineville
Geothermal + dry cooling

1.09	0	0.02
PUE	g/kWh CO_2e	L/kWh water

IMPLEMENTATION ROADMAP

0-6 months	Publish open-carbon ledger v1.0 with real-time CO_2e per job <2% error
6-18 months	Commission 5 MW immersion edge pods with PUE <1.05, water <0.2 L/kWh
18-36 months	Secure 24/7 clean PPAs covering 50% load with hourly matching coefficient >90%
36-60 months	Deploy SMR-anchored 1 GW AI campus with carbon intensity <50 g/kWh

SOCIAL LICENSE TO OPERATE

ELECTRICITY BILL IMPACTS
$18/month Maryland, $16/month Ohio (PJM 2025-26)

PROJECTED CONSUMER IMPACT
8% average US increase by 2030, potentially ≥25% in Northern Virginia

WATER COMPETITION
Texas data centres consuming 6.6% of state's total water use by 2030

LOCAL GRID STRAIN
Single AI campus drawing 200+ MW in regions with <15% reserve margin

SOCIAL LICENSE QUANTIFIED
Opposition is no longer NIMBYism—it is economically rational response to quantified local impacts. Transparent engagement is now an approval prerequisite.

FROM VISION TO DELIVERY

Constraint	Status	Impact
Grid interconnection	2,300 GW queued, 8+ year timelines	Primary schedule driver
Water regulation	Mandatory reporting, allocation limits	Permitting blocker
AI power growth	4% → 12% US electricity by 2030	Consumer bill increases
Social license	Quantified community impacts	Project opposition

The A+ Pathway offers a practical route forward—grounded in engineering reality rather than aspiration.

WHAT CHANGED: DECEMBER 2025 → JANUARY 2026

Grid Interconnection Crisis Deepened
Queue volume: 2,300 GW (up 53%)
PJM timelines: 8+ years (doubled)
Only 20% of 2000-2018 projects reached operation

Water Regulations Transitioned
From Emerging to Enforced
EU EED reporting now mandatory
Texas SB 7 addressing 399B gallons by 2030

AI Power Consumption Exceeded Projections
Current US consumption: 4% of national electricity
Trajectory: 12% by 2030 (tripled in 6 years)
Goldman Sachs: 165% increase by 2030

Behind-the-Meter Solutions Shifted
From Optional to Required
>45 GW projected shortfall requiring alternative power strategies
BYOP accelerating in Northern Virginia, Texas

Social License Quantified
PJM $9.3B capacity increase: Direct residential bill impacts
Texas water consumption: 6.6% of state total by 2030
Consumer electricity bills: Projected 25%+ increases in high-demand markets

AI does not fail because of lack of ambition. It fails when physical limits are ignored.
Scaling AI without breaking the planet is not only possible—it is becoming a competitive advantage.
The organisations that lead the next phase will be those that understand power systems, water constraints, grids, and communities as deeply as they understand models and code.
Success in 2026 belongs to operators who solve physics, not those who optimize spreadsheets.

APPENDIX — DATA-SOURCE REFERENCE PACK

1. AI Compute & Energy Growth

- IEA (2024). Data Centres and AI Special Report — projects 3–5x electricity demand growth by 2030.
- Uptime Institute (2023–2024). Global Data Centre Survey — increasing HPC energy density, cooling limits.
- Deloitte (Nov 2025). "More compute for AI, not less" — inference workloads 60-70% of total compute by 2026.
- AllAboutAI (Dec 2025). AI Environment Statistics 2026 — inference now represents 60-70% of AI energy consumption.
- McKinsey (Dec 2025). "The next big shifts in AI workloads and hyperscaler strategies" — training demands up to 1 MW per rack.

2. Water Use in Data Centres

- Lawrence Berkeley National Laboratory (LBNL), 2022. Water use intensity range: 1.8–2.6 L/kWh for evaporative cooling.
- Arizona and California Water Authority curtailment cases (2020–2023): multiple directives limiting industrial water use.
- Santa Clara County (2024). Recycled water requirements for new data center builds — now mandatory.
- Johor, Malaysia (2024). 30% data center project rejection rate documented due to water constraints.
- Texas Water Development Board (2024). Data centres projected to consume 399 billion gallons by 2030.

3. Immersion Cooling Performance

- Submer (2023), Asperitas (2023), Icetope (2024) — PUE 1.03–1.10 in controlled immersion deployments.
- WUE reported between 0.05–0.10 L/kWh.
- Meta Prineville (2024). Geothermal + dry cooling achieving 0.02 L/kWh WUE with zero carbon intensity.

4. Carbon Forecasting & Compliance

- IEA Net Zero Pathways (2023–2024). Data centre emissions projected to reach 2–3% of global CO_2e by 2030.
- EU CBAM documentation (2023). Expansion to electricity and carbon-intensive digital imports.
- EU Energy Efficiency Directive (2024). Mandatory reporting now enforced.
- Google/BCG Report (2023). AI has potential to mitigate 5-10% of global GHG emissions by 2030.

5. Power Queue & Interconnection Delays

- US DOE Interconnection Queue Report (2024): 2,300 GW in queues; average interconnection time now 8+ years.
- PJM (2025). Timeline extended from 5-7 years to 8+ years (doubled).
- EPRI modelling indicates that 10–20 GW of upcoming queue load is associated with hyperscale AI deployments.

6. AI-Driven Efficiency Layer

- Google DeepMind (2016–2023). ML-driven cooling optimisation delivering 30–40% energy reduction.
- Meta internal research (2023). AI-assisted fan and thermal control sequencing yields 15–25% efficiency gains.

7. SMR Deployment Timelines

- IAEA SMR Roadmap (2024): Commercial deployments unlikely before early 2030s.
- NuScale project cancellation (2023) illustrates regulatory and financial hurdles.

8. Workload Distribution & Market Shift

- Bloomberg Intelligence (Nov 2024). "Generative AI 2025 Assessing Opportunities" — training infrastructure 35% of market by 2032, inference 52%.
- Bain Technology Report (2025). AI compute demand growing at 2x Moore's Law rate; global requirements could reach 200 GW.
- IDC (Nov 2025). AI infrastructure spending reached $82B in Q2 2025, 166% YoY increase.

9. Consumer AI & Application Growth

- Andreessen Horowitz (Dec 2025). "State of Consumer AI 2025" — ChatGPT 800M weekly users (Sept 2025), Sora 12M downloads.
- BCG (2025). Enterprise AI budget allocation: Support functions 38%, Operations 23%, Marketing/Sales 20%, R&D 13%, climate applications <3%.
- OpenAI metrics (Sept 2025). ChatGPT queries consume 10× electricity vs. traditional Google search.
- McKinsey (2025). 78% of organizations report using AI in at least one business function, up from 55% in 2023.

10. Climate AI Potential vs. Allocation Reality

- Google/BCG Report (2023). AI mitigated 2.4M metric tons CO_2e through Google Maps fuel-efficient routing alone.
- BCG Enterprise Budget Analysis (2025). Climate-critical applications receive <3% of enterprise AI compute budgets.
- Current allocation reality: 85%+ of inference compute serves consumer engagement vs. <15% for scientific/climate applications.

11. Social License & Community Impact

- PJM capacity auction (2025-26). $9.3B increase driving $18/month residential bill impact in Maryland, $16/month in Ohio.
- Goldman Sachs (2024). US electricity bills projected to increase 25%+ in highest-demand markets by 2030.
- IEA Energy and AI Report (2024). Global data center electricity use expected to grow from 415 TWh (2024) to 945 TWh (2030).
- AllAboutAI (2025). United States leads global AI electricity consumption with 200+ TWh annually (65% of global share).

Email: info@hml-services.com | Website: www.hmlservices.biz