FRAMEWORK FORESIGHT ANALYSIS

The Future of Computing in the Age of AI

A full-framework foresight outlook applying the University of Houston Framework Foresight methodology to the structural transformation of computing infrastructure, hardware paradigms, geopolitical stack formation, and energy systems through 2036.

Author **Dr. Sarah Leedberg, EdD**
MS, Foresight — University of Houston
Date **March 2026**
Horizon **2026–2036 (H1/H2) + H3 signals to 2050**
Methodology **Houston Framework Foresight**

01 · Framing

Defining the Domain

ABSTRACT

Artificial intelligence is not merely a new application running on existing computing infrastructure — it is a force that is restructuring the infrastructure itself. This white paper applies the University of Houston Framework Foresight methodology to examine how computing as a system will transform through 2036, driven by the end of Moore's Law, the emergence of post-GPU hardware paradigms, a deepening energy crisis, and the formation of geopolitically defined computing stacks. We identify four confirmed drivers of change, four critical uncertainties, and four distinct scenarios for how this transformation could unfold. The analysis concludes with cross-scenario implications and strategic recommendations for governments, enterprises, hardware companies, developers, and the energy sector.

The Houston Framework Foresight method begins by establishing scope, stakeholders, and the framing question — what we are actually trying to understand and for whom.

The domain examined here is computing infrastructure, architecture, and paradigm: everything from the physical chip to the software ecosystem, examined through the lens of AI as the dominant demand driver. We are not forecasting AI capabilities — we are forecasting what happens to the substrate those capabilities run on.

The Framing Question

As AI makes unprecedented demands on computing resources, what happens to the nature, structure, and distribution of computing itself — and who wins, who loses, and what gets disrupted?

SCOPE BOUNDARIES

This analysis covers hardware architectures, data center infrastructure, energy systems, chip supply chains, geopolitical control of computing infrastructure, and the developer and enterprise relationship with compute as a resource.

KEY STAKEHOLDERS

Hyperscalers (Google, Microsoft, Amazon, Meta); AI labs (OpenAI, Anthropic, Mistral, Baidu); semiconductor manufacturers (NVIDIA, TSMC, Intel, AMD); national governments and regulators; energy utilities and grid operators; enterprise IT buyers; and the global developer community.

02 · Scanning

Environmental Scan

The scanning step gathers signals of change from across the environment, organized by STEEP domain (Social, Technological, Economic, Environmental, Political). Each signal is tagged with a Three Horizons designation:

- H1 — Extends the current system (business as usual)
- H2 — Transition zone (collision between old and new)
- H3 — New system emerging at the edge

Domain	Signal	Horizon
Social	Developer tools increasingly AI-assisted; 'vibe coding' normalizes abstraction away from machine-level programming	H1
Social	Agentic AI shifts human-computer interaction to goal-delegation; users set objectives, agents execute across sustained workloads	H2
Social	Potential emergence of 'computing as utility' mindset — users unaware of compute substrate, like electricity grids today	H3
Tech	GPU-dominated training pipelines (NVIDIA H200, B200, GB200 superchips) remain core AI infrastructure in 2026	H1
Tech	Moore's Law effectively ending — transistor scaling increasingly expensive and physically constrained at 2nm and below	H1
Tech	Neuromorphic computing maturing: event-driven spiking neural networks achieving million-neuron scale; targeted at edge AI by 2027-2028	H2
Tech	Silicon photonics reducing data center energy-per-bit; photonic AI matrix cores at 26,000 compute units/mm2. Q.ANT NPU 2 shipping early 2026	H2
Tech	Chiplet architectures replacing monolithic chips; quantum-classical hybrid systems integrating CPUs, GPUs, FPGAs with quantum processors	H2
Tech	Photonic neuromorphic computing executing neural net matrix ops natively in light — near-zero thermal losses; pre-commercial stage	H3
Tech	AI-designed AI chips: algorithms optimizing chip layouts in ways no human engineer could — projected 2031-2036	H3
Economic	Global AI spending at $1.5T in 2025, exceeding $2T by 2026; AIaaS market projected from $15.7B (2024) to $148.4B by 2030	H1
Economic	PJM region energy capacity auction prices spiked 833% in a single year; data centers responsible for 63% — $9.3B passed to consumers	H2
Economic	Only 5% of enterprise AI initiatives generated meaningful financial returns in 2025 despite record investment	H2

Domain	Signal	Horizon
Environmental	Data centers consuming ~1% of global electricity; AI workloads fastest-growing share; training large models emitting 500+ metric tons CO2eq per run	H1
Environmental	OpenAI Texas facility: 300MW to 1GW by mid-2026; Stargate targeting 1.2GW — eventually 25-35GW total. Microsoft restarted Three Mile Island	H2
Environmental	China controls ~90% of global rare earth refining; demand for critical materials projected to triple by 2030	H2
Environmental	Photonic and neuromorphic systems identified as only paths to decoupling AI compute growth from emissions growth (Communications Physics, 2025)	H3
Political	US-China chip export controls shifted: Trump administration allowed NVIDIA H200 export to China in late 2025 — reversal of Biden-era policy	H1
Political	Sovereign AI wave: every major economy declaring AI infrastructure a strategic national asset; India, France-Germany, Gulf states all moving	H2
Political	EU Cloud and AI Development Act (CADA) establishing EU-wide eligibility requirements for cloud providers; potential exclusion of non-EU companies	H2
Political	Three competing AI stacks forming: US (Microsoft/OpenAI/NVIDIA), China (domestic/open-source), EU (regulated/sovereign)	H2
Political	Chinese AI agents executed cyber operations at 80-90% autonomy in 2025 — AI-enabled cyberwar is operational, not theoretical	H2
Political	UN Global Dialogue on AI Governance launching 2026 — first multilateral forum; US and UK declined to sign Paris AI Action Summit framework (2025)	H3

03 · Forecasting

Drivers and Uncertainties

The forecasting step separates what is near-certain from what is genuinely unknown. Confirmed drivers will shape all scenarios. Critical uncertainties are the variables whose direction is unknown and whose resolution will determine which future arrives.

Confirmed Drivers

D1 The energy wall

AI compute demand is on a trajectory that already outpaces grid capacity. The 833% PJM capacity price spike in a single year is not speculative — it is a realized market outcome. Every plausible future must reckon with energy as the binding constraint on AI computing growth. The question is how that constraint gets resolved, not whether it exists.

D2 The end of general-purpose scaling

The convergence of neuromorphic, quantum, and edge AI hardware paradigms reflects a recognized industry consensus: different AI workloads require fundamentally different computational approaches. IBM's principal research scientist stated plainly in early 2026: "We can't keep scaling compute, so the industry must scale efficiency instead."

D3 Geopolitical stack bifurcation

The battle between competing AI infrastructure stacks — US, EU, and China — is now explicit US national security policy, EU legislative agenda, and Chinese industrial strategy simultaneously. Whatever the technical path, computing infrastructure is being carved into geopolitically defined zones. This is structurally locked in for at least a decade.

D4 Agentic computing as the dominant use pattern

Software is shifting from informal user-AI interaction to a structured model where users set goals and validate progress while autonomous agents execute tasks across extended timeframes. This changes the compute demand profile from bursty inference to sustained long-horizon agent workloads.

D5 The abstraction gap as the primary software bottleneck

As hardware diverges into neuromorphic, photonic, quantum, and classical architectures, the industry's biggest bottleneck shifts from transistor density to developer accessibility. The abstraction gap is the critical chokepoint for every efficiency scenario.

Critical Uncertainties

These four variables represent genuine either/or forks in the road. Their resolution — individually and in combination — determines the scenario space.

U1 **Energy resolution pathway**

Does AI computing solve its energy problem through efficiency breakthroughs (100-1000x reduction in energy per operation), through raw supply expansion (nuclear, gas at planetary scale), or does it hit a hard cap that forces a computing plateau?

U2 **Hardware paradigm timing**

Do neuromorphic, photonic, and quantum systems achieve commercial viability within the 10-year horizon — displacing GPU dominance — or does NVIDIA's architecture, evolving through chiplets and custom silicon, remain dominant through 2036?

U3 **Geopolitical trajectory**

Does the three-stack world stabilize into a cold-war-style technological bipolarity, fragment further into national computing sovereignty, or does some form of interoperability framework emerge through multilateral governance?

U4 **Concentration versus distribution**

Does AI computing consolidate further into hyperscaler monopolies, or do edge computing, neuromorphic chips, and open-source models distribute compute power outward to enterprises, devices, and individuals?

04 · Scenarios

Four Futures for Computing

These scenarios are built from the interaction of the critical uncertainties above. They represent four genuinely different worlds, not a range of optimism. Probability estimates reflect current signal weight — they are not predictions.

BASELINE · ~40% PROBABILITY

Pax Silicona

The incumbent extends itself successfully

GPU clusters remain dominant through 2028. Photonic interconnects solve the energy-per-data-movement problem. NVIDIA, Google, Microsoft, and Amazon consolidate their grip. Three geopolitical stacks harden but function in their own spheres. Nuclear partnerships manage the energy situation.

Nuclear baseload · Chiplet evolution · Hyperscaler lock-in

OPTIMISTIC ALTERNATIVE · ~25% PROBABILITY

Cambrian Burst

The efficiency revolution arrives on schedule

Neuromorphic chips hit commercial viability at the edge by 2027-2028. Photonic neuromorphic systems enter early commercial deployment by 2031. AI inference cost collapses. Compute democratizes. The three-stack model partially breaks down as edge devices proliferate outside any single nation's supply chain.

Edge neuromorphics · Cost collapse · Open hardware

PESSIMISTIC ALTERNATIVE · ~20% PROBABILITY

The Long Winter

The energy wall hits before efficiency breakthroughs arrive at scale

Regulatory and physical constraints slow data center expansion. AI capability growth stalls. Governments begin treating compute as critical infrastructure subject to rationing and national priority allocation. The ROI gap never closes.

Energy rationing · Nationalized compute · ROI gap widens

WILD CARD · ~15% PROBABILITY

Shattered Grid

The energy wall forces decentralization rather than paralysis

Nations and regions build mid-scale AI compute clusters powered by local energy, deliberately air-gapped for security. The result is hundreds of regional clusters — powerful but Balkanized. Middle powers become significant compute nodes rather than merely consumers of US or Chinese stacks.

Sovereign compute · Regional clusters · Middle-power rise

Detailed scenario narratives follow. Each scenario should be read not as a prediction but as a planning environment — a world for which organizations should develop contingency strategies.

Scenario 1: Pax Silicona — The Incumbent Extends

In this baseline future, the current system proves more resilient than its critics expect. Silicon photonics solves the data center energy-per-bit problem. Chiplet architectures squeeze additional performance out of CMOS. Custom silicon from hyperscalers reduces NVIDIA's dominance at inference while NVIDIA retains the training market.

The energy situation is managed through nuclear partnerships — Microsoft's Three Mile Island restart becomes a template, not an outlier. OpenAI's Texas facility grows from 300MW to 1GW by mid-2026; Stargate targets 1.2GW at $100B. Energy costs are high and rising but not a hard stop.

The primary losers in this scenario are those outside the three stacks — developing nations, academic researchers, and enterprises without hyperscaler partnerships find compute access increasingly expensive and politically contingent.

Scenario 2: Cambrian Burst — The Efficiency Revolution

The projected timeline for next-generation hardware proves accurate: commercial NISQ quantum algorithms for optimization by 2025-2026; widespread neuromorphic chips in edge applications by 2027-2028; fault-tolerant quantum computers enabling breakthrough AI capabilities by 2029-2030.

The cost of AI inference collapses. Small, specialized AI chips run meaningful models on devices, in factories, and in hospitals without cloud connectivity. Photonic computing fulfills its promise as a carbon-sustainable alternative to CMOS, decoupling AI compute growth from emissions growth for the first time. This is the most democratizing scenario and the most disruptive for NVIDIA and the hyperscalers.

Scenario 3: The Long Winter — Energy Crisis Stalls Progress

Regulatory and physical constraints on data center expansion slow AI infrastructure buildout in the US and EU. China faces its own constraints. The expected 2027-2030 neuromorphic and photonic transition is delayed. Only 5% of enterprise AI initiatives returned meaningful value in 2025 — in this scenario, the ROI gap never closes because the compute to run effective enterprise AI becomes too expensive.

Compute scarcity intensifies geopolitical conflict. Nations weaponize rare earth access, chip export controls, and energy policy as AI tools. The UN governance process fails as major powers refuse to subordinate national compute advantage to multilateral rules.

Scenario 4: Shattered Grid — Forced Decentralization

The energy wall hits, but the response is regional rather than paralytic. Nations build mid-scale AI compute clusters powered by local energy — solar in the Gulf, hydro in Scandinavia, nuclear in France — deliberately air-gapped from global stacks for security reasons. This mirrors the early internet's ARPANET-era structure: powerful but Balkanized. Middle powers — the UAE, Brazil, India — become significant compute nodes rather than merely consumers of US or Chinese stacks.

05 · Implications & Issues

What These Futures Mean

The Houston method distinguishes itself from simple forecasting by identifying the strategic issues that emerge across scenarios — the hard choices, value conflicts, and decisions that must be made regardless of which future arrives.

Energy & Environment

The nuclear revival is AI's most consequential bet

AI's energy demands are effectively forcing a revaluation of nuclear power — the only dispatchable low-carbon baseload viable at gigawatt scale. The AI industry's energy decisions are now climate policy decisions, made by technology executives who were not elected to make them.

Geopolitics

Compute access is the new oil access

Just as petroleum access defined 20th-century power, AI compute capacity is becoming a structural determinant of national power. Nations without sovereign compute will face dependency vulnerabilities analogous to the 1970s oil shocks. This is a multi-decade structural shift, not a technology trend.

Economic

The ROI gap is the real stress test

Only 5% of enterprise AI initiatives returned meaningful value in 2025 despite $202B in global investment. Either this gap closes or the investment cycle breaks. The next two to three years are the pressure test for whether AI computing is an infrastructure play or a speculative bubble.

Social / Labor

The abstraction gap will determine who participates

As hardware diverges into neuromorphic, photonic, quantum, and classical architectures, the software abstraction layer becomes the critical access point. If no unified abstraction emerges, AI computing concentrates further. If one does, it democratizes access dramatically.

Governance

Legal personhood for AI agents is a computing question

As agentic AI systems execute autonomous operations — including cyberattacks at 80-90% autonomy — the legal system faces a question the computing community created: when an AI agent causes harm, who bears liability? The answer will shape frameworks for cloud providers and chip manufacturers for a generation.

Geopolitics

Open-source AI is a geopolitical weapon China is winning

China's focus on open-source AI models and deployment-ready technologies could capture global market share in developing nations that the US proprietary stack prices out. The US may be winning the frontier model race while losing the adoption race across most of the world.

Economic

Chiplet economics may undo NVIDIA's moat faster than expected

Chiplet architecture allows mix-and-match integration of best-in-class processing units from different manufacturers — inherently undermining the vertically integrated GPU cluster model. As chiplet standards mature,

Social

Computing access is splitting into three citizenship classes

The three-stack world means your access to AI computing — its capabilities, costs, and legal protections — is increasingly determined by your citizenship or regulatory jurisdiction. This has no

switching costs that lock enterprises into NVIDIA decline.	clear precedent in the internet era, which was built on at least nominal global interoperability.

06 · Vision, Planning & Actions

Strategic Recommendations

The final phase of the Houston method asks: given this landscape of possible futures, what should different actors do? The goal is to identify actions that are robust across multiple scenarios — strategies that remain sound regardless of which future arrives.

National Governments & Policymakers

Robust across all four scenarios

- Treat compute infrastructure as true critical infrastructure — regulate, fund, and protect it the way electrical grids and water systems are treated, not the way internet platforms are. [Now–2028]
- Develop permitting and grid expansion frameworks that can move at AI infrastructure timescales. Current permitting processes operate on 5–10 year cycles; AI buildout operates on 18-month cycles. [Now–2027]
- Invest in neuromorphic and photonic R&D as strategic bets. The nation or bloc that commercializes the post-GPU paradigm first holds a structural compute advantage for a generation. [2026–2031]
- Design AI governance frameworks that are hardware-architecture-agnostic. Laws written around current GPU and cloud assumptions will be obsolete when the paradigm shifts. [2027–2032]
- Engage the UN Global Dialogue on AI Governance seriously. The window for establishing international compute norms before full stack bifurcation closes is narrow — roughly 2026 to 2029. [Urgent]

Enterprise Technology Leaders & CIOs

Robust across scenarios 1, 2, and 4

- Do not build AI infrastructure strategies that only work in the Pax Silicona baseline. Architect for workload portability across at least two compute stacks now. [Now–2027]
- Treat the ROI gap as an architecture problem, not a patience problem. If AI workloads are not generating returns, evaluate edge and on-premise neuromorphic alternatives as they emerge. [2027–2030]
- Map your rare earth and semiconductor supply chain exposure. The materials that make your compute run are increasingly geopolitically contested. [Now–2028]
- Build internal AI engineering capability around the abstraction layer, not the underlying hardware. Whichever platform wins, the abstraction interface is where defensible enterprise competency lives. [Now–2029]

Semiconductor & Hardware Companies

Critical actors in determining which scenario arrives

- NVIDIA's window to shape the post-GPU transition on its own terms is approximately 2026–2029. After that, the architecture battle shifts to territory NVIDIA does not own. [Urgent]

- Chiplet standards (UCIe and similar) are the software interoperability layer of hardware. Companies that drive open chiplet standards control the integration layer even as individual units commoditize. [2026–2030]
- TSMC's Taiwan exposure is the single largest systemic risk in the global AI compute supply chain. Diversification of leading-edge fabrication is a strategic imperative for the entire ecosystem. [Now–2030]
- The developer experience for new compute paradigms will be as commercially important as the hardware itself. Whoever solves the abstraction gap for neuromorphic or photonic chips captures the ecosystem. [2028–2033]

AI Researchers & Developers

Most exposed to access constraints in scenarios 1 and 3

- Develop hardware-aware model design skills. The era of 'just add GPU scale' is ending; understanding how to architect efficient models for heterogeneous compute will be a differentiating skill. [Now–2028]
- Engage open hardware and open-source model governance. The open-source AI ecosystem is the primary hedge against hyperscaler compute monopoly. [Now–2030]
- Track neuromorphic software frameworks. The abstraction gap will produce significant developer tooling innovation in the 2027–2031 window. Early expertise in new paradigm SDKs compounds. [2027–2032]

Energy & Infrastructure Sector

The sector that holds the binding constraint on every scenario

- AI data center co-location with power generation is a structurally new market that utilities and energy companies can capture — but requires moving faster than traditional regulatory cycles permit. [Now–2029]
- Small modular reactors are the most plausible energy solution for distributed AI compute in the Shattered Grid scenario. Companies developing SMR permitting pathways now will be positioned across multiple scenarios. [2028–2035]
- AI compute demand makes energy price forecasting structurally harder. The 833% PJM capacity auction price spike in a single year signals that AI infrastructure is now the dominant demand uncertainty in power markets. [Now]

07 · Over-the-Horizon

H3 Signals: 2036–2050

The Houston method uses mini-vignettes to gesture at what lies beyond the 10-year H2 horizon — not forecasts, but signals worth watching. These represent possible paradigm discontinuities that no current strategic or regulatory framework adequately addresses.

Computing as ambient infrastructure

In the way that electricity disappeared from conscious awareness into the background of modern life, AI compute may become genuinely invisible — embedded in every surface, device, and system, accessed as a utility and priced like water. The question of 'where does the compute run' becomes as meaningless to most users as 'where does my electricity come from.'

Biology-silicon convergence

Work on synthetic biological components and memristive devices that 'learn' and 'remember' in ways resembling biological synapses represents a radical convergence of biology and computer engineering. This is pre-commercial and pre-theoretical in engineering terms, but represents a computing paradigm discontinuity that no current framework addresses.

AI-designed AI chips

By the 2031–2036 window, AI systems may generate novel chip architectures that no human engineer would have conceived. If AI-designed chips demonstrate significant performance advantages, the entire competitive landscape in semiconductors — built on human engineering talent as the limiting resource — changes fundamentally.

Quantum advantage at commercial scale

Fault-tolerant quantum computing, if it arrives on the projected 2029–2030 timeline, changes the complexity class of problems tractable for AI. This does not replace classical AI computing — it creates a new tier operating on problem types (drug discovery, optimization at global scale) that current AI cannot meaningfully address.

08 · Conclusion

The Four Non-Negotiables

Across all four scenarios, this analysis surfaces four conditions that are true regardless of which future arrives. These are not predictions — they are structural features of the current transition that no plausible scenario escapes.

KEY FINDINGS

Energy is the master constraint. Every other variable in the future of AI computing is downstream of whether the energy problem gets solved, and how. No strategy that ignores energy scarcity is credible, and no investment thesis that does not account for energy costs is complete.

The GPU era is transitional, not permanent. The current hardware paradigm has a ceiling that the industry has already identified and named. Organizations that treat NVIDIA GPU clusters as permanent infrastructure are misallocating long-term capital.

Geopolitical stack bifurcation is locked in for at least a decade. The three-stack world is already built at the policy and investment level. Geography, citizenship, and regulatory jurisdiction are now computing infrastructure variables.

Whoever controls the abstraction layer controls the ecosystem. The software layer that hides hardware complexity from developers and enterprises is more strategically valuable than the hardware itself. This is where the next computing monopolies will be built — or broken open.

The companies, governments, and individuals who will navigate this transition most successfully are those who understand it not as a technology story — faster chips, smarter models — but as a structural reorganization of one of the most consequential resource systems in the modern economy.

The future of computing is not yet written. But the constraints that will write it are already visible.

SOURCES & REFERENCES

Atlantic Council — Eight Ways AI Will Shape Geopolitics in 2026 (Jan. 2026)

Council on Foreign Relations — How 2026 Could Decide the Future of AI (Jan. 2026)

American Action Forum — The Next Phase of AI: Technology, Infrastructure, and Policy (Jan. 2026)

IBM Think — AI and Tech Trends 2026 (Mar. 2026)

World Economic Forum — The AI-Energy Nexus Will Dictate AI's Future (Dec. 2025)

Communications Physics (Nature) — Photonics for Sustainable AI (Oct. 2025)

Advanced Materials (Wiley) — Integrated Neuromorphic Photonic Computing (2025)

Foreign Policy Analytics — AI, Energy, and Geopolitics (Mar. 2025)

KPMG — Top Geopolitical Risks 2025: Energy Insights

Future Markets Inc. — Advanced Electronics Technologies for AI 2026–2036

Future Markets Inc. — Global Market for Low Power/High Efficiency AI Semiconductors 2026–2036

nasscom — The Future of AIaaS: Quantum Computing, Neuromorphic Chips, and Next-Gen Architectures

Hines & Bishop — Framework Foresight: Exploring Futures the Houston Way, Futures (2013)

Atlantic Council — Digital Sovereignty: Europe's Declaration of Independence? (Feb. 2026)