strategy&
Part of the PwC network

Meeting the compute imperative in the Gulf countries

The Gulf's AI journey is entering a transformative phase. AI is shifting from experimentation to scale, agentic systems are proliferating, and Gulf nations are building AI ecosystems featuring homegrown champions and local models, such as KSA's Allam,[1] UAE's Falcon,[2] and Qatar's Fanar.[3] This phase of AI development requires a critical enabler: sovereign compute processing power at scale. Now, Gulf nations must assess their future compute needs and act decisively to secure compute capacity.

AI is poised to become part of daily life in the Gulf. In healthcare for example, AI agents will review patient test results, suggest diagnoses, generate treatment plans, and manage patient journeys based on individual needs – updating records, scheduling follow-ups, and coordinating medication orders.

AI models are the heart of this transformation – and sustaining them requires vast processing power. The greatest demand for compute power will come from running models for real-time, low-latency inference. Compute will also be needed to build and train local models, and to fine-tune global models for regional and sector-specific applications.

To forecast how much compute power Gulf nations will require in the near term, we analyzed the volume of data that AI models will process — measured in model tokens, or small fragments of text, images, or other inputs – across training, fine-tuning and inference. We then translated this demand into the processing required (FLOPS, or floating-point operations per second) and the hardware needed to deliver it through GPUs. Our analysis reveals that the region will need 400,000 – 500,000 GPUs by 2028. This level of demand is significant: for context, xAI's supercomputer Colossus currently uses 200,000 GPUs with a trajectory towards 1,000,000.[4]

Beyond the near term, the next wave of AI – real-world models that integrate digital intelligence with physical data and enable robots to operate in dynamic environments – will likely accelerate compute demand even further.

Data sovereignty mandates require that much of the Gulf's compute demand – across government, regulated industries, and critical infrastructure – be met locally. Accordingly, Gulf countries are making bold moves to secure compute sovereignty. Initiatives such as KSA's HUMAIN AI data center and UAE's Stargate project aim to build hyperscale capacity on domestic soil. With these efforts, the region is on track not only to meet local GPU demand, but also export compute power.

Risks to demand fulfillment remain, however. GPUs needed for the Gulf's infrastructure initiatives will be shipped gradually over years, and access can be interrupted by export controls and geopolitical shifts. Technology cycles move fast and each refresh renews dependence on external

> This phase of AI development requires a critical enabler: sovereign compute processing power at scale. Now, Gulf nations must assess their future compute needs and act decisively to secure compute capacity.

supply chains. Compute demand remains uncertain and volatile, and rapid changes in AI adoption could create overcapacity or shortages. Rising energy demand from AI workloads could strain energy systems without coordinated planning. Critical government and security workloads can be crowded out by commercial and export demand without careful governance.

Gulf governments can mitigate these risks and address the compute sovereignty imperative through the following multi-layered strategies.

1 Governments can secure supply chains by working with the private sector to maintain GPU reserves, diversify global vendors, and deepen partnerships with chipmakers through long term contracts and strategic investments. The UK government is partnering with multiple U.S. tech firms to expand supercomputing capacity.[5] In the Gulf, HUMAIN[6] and G42[7] – in collaboration with their governments – continue to develop partnerships with chipmakers such as NVIDIA, AMD, and Cerebras to enhance access resilience.

2 Governments can strengthen local compute ecosystem capabilities — from semiconductor design and manufacturing to data-center setup and operations — through investment programs, targeted incentives, and regulatory support. Examples in the semiconductor space include the U.S. CHIPS and Science Act[8] to boost domestic chip production, and KSA's Alat[9] and National Semiconductor Hub[10] initiatives to build semiconductor design and manufacturing capabilities.

3 Governments can improve infrastructure efficiency and access resilience through regional cooperation. A Gulf-wide compute infrastructure initiative, such as the European High-Performance Computing Joint Undertaking, could optimize investments, improve utilization, and enhance the Gulf's collective bargaining power in global supply chains.[11]

4 Governments can safeguard critical workloads during compute demand peaks through rigorous capacity allocation governance. High-performance computing initiatives in Japan and Europe illustrate how access tiers and allocation rules preserve capacity for essential and emergency applications.[12]

5 Governments can institutionalize continued compute capacity forecasting that is informed by compute usage observatories and integrated with energy planning. This would enable more effective capacity management and alignment between compute infrastructure, energy systems, and national priorities.

Gulf countries have a successful history of anticipating bottlenecks in essential arenas such as food and water security. Now, they must address the compute imperative to seize control of their AI journeys and evolve into global AI leaders.

[1] H. Marshad, "Arabic AI Breakthrough" Daleel, Oct. 5, 2025.
[2] "UAE launches Arabic language AI model as Gulf race gathers pace" Reuters, May 21, 2025.
[3] "Welcome to Fanar: Arabic Generative AI Platform" Fanar, accessed Dec. 16, 2025.
[4] "Colossus" xAI, accessed Nov. 7, 2025.
[5] "Britain boosts computing power in $1.3 billion AI drive" Reuters, July 17, 2025.
[6] "HUMAIN Expands Strategic Partnership with NVIDIA" HUMAIN, accessed Dec 16, 2025.
[7] "G42 Receives U.S. Approval for Advanced AI Chip Exports" G42, Nov. 20, 2025.
[8] "The CHIPS Act: What it means for the semiconductor ecosystem" PwC, accessed Dec. 16, 2025.
[9] "HRH Crown Prince Launches 'Alat' to Contribute in Making Saudi Arabia a Global Hub for Electronics and Advanced Industries" SPA, Feb. 1, 2024.
[10] M. Al-Barakati "Saudi Arabia launches 'National Semiconductor Hub' to drive industry localization" Arab News, June 5, 2024.
[11] "The European High Performance Computing Joint Undertaking".
[12] "Overview of HPCI" The Research Organization for Information Science and Technology; "The European High Performance Computing Joint Undertaking."

Hani Zein
Partner
hani.zein@strategyand.pwc.com

Fawaz Bou Alwan
Partner
fawaz.boualwan@strategyand.pwc.com

Ali Ghaddar
Principal
ali.ghaddar@strategyand.pwc.com

Mahsa Ettefagh
Manager
mahsa.ettefagh@strategyand.pwc.com

www.strategyand.pwc.com/me

© 2026 PwC. All rights reserved. PwC refers to the US member firm or one of its subsidiaries or affiliates and may sometimes refer to the PwC network. Each member firm is a separate legal entity. Please see www.pwc.com/structure for further details. This content is for general information purposes only and should not be used as a substitute for consultation with professional advisors.