Submer



NOVEMBER 2026

GCC AI & HPC Infrastructure Report

Powering Sovereign Compute: Overcoming Power, Heat, and Water Limits at Scale

SUBMER.COM/MEA

Table of Contents

[00]	Methodology, Scope & Limitations	4
	Scope	4
	Estimation Approach	4
[01]	Executive Summary	5
	Key Takeaways	6
[02]	Market Overview & Sizing	7
[2.1]	Established Market Baselines	8
[2.2]	TAM / SAM / SOM : Modelled Estimates, 2026	9
[03]	Industry Trends & Drivers	11
[3.1]	The Catalyst: Blackwell-Class Compute Arrives in the Gulf	12
[3.2]	Secured Grid Power: The Hard Currency of AI	13
[3.3]	Resilient High-Density Thermal Management	14
[3.4]	Local Operational Sovereignty	15
[3.5]	Water-Free Sustainability	16
[3.6]	Delivery Models and Partner Ecosystems	17

[04] Competitive Landscape 18

[4.1] Compute and Silicon Layer 19

[4.2] GCC Data Centre Developers and Operators 20

[4.3] Thermal and Cooling Stack 21

[05] Strategic Opportunities & Risks 23

[5.1] Barriers to Entry 24

[5.2] Risks 25

[5.3] Untapped Growth Opportunities 26

[06] 2026-2030 Future Outlook 27

[6.1] Expected Trajectory 28

[6.2] Recommendations 29

For IT and Infrastructure Leaders: CIOs, CTOs, and Data Centre Directors

For Government and Policy Stakeholders

For Sustainability and ESG Leaders

For Investors

[07] Where Submer Fits 30

[A] References 33

Methodology, Scope & Limitations

This report combines publicly available market data, government and operator announcements, and analyst forecasts published through Q2 2026. It is designed for C-level technology, infrastructure, policy, and sustainability leaders who need a practical view of the GCC AI and high-performance-computing infrastructure market.

Scope

Geography

The report covers the six GCC member states: Saudi Arabia, the United Arab Emirates, Qatar, Kuwait, Bahrain, and Oman. The analysis places greater weight on Saudi Arabia and the UAE, which currently account for most committed and operational AI infrastructure capacity in the region.

Market definition

The report focuses on AI- and HPC-oriented data centre infrastructure, including compute infrastructure, GPU and accelerator platforms, high-density facilities, and the power and thermal systems required to support them. Adjacent areas such as storage, networking, and cloud services are referenced where they influence infrastructure demand.

Time horizon

The report uses 2026 as the base year and provides a forward view to 2030.

Estimation Approach

Where market size is stated, this report separates figures reported directly by third-party analysts from figures modelled specifically for this report.

Modelled estimates use a build-up methodology. Committed IT-load capacity, measured in MW, is converted into an addressable infrastructure-spend value using public capital-intensity benchmarks. That figure is then filtered to identify the share of the opportunity relevant to cooling and high-density infrastructure.

All modelled figures include explicit assumptions. Ranges are used in preference to false-precision point estimates.

GCC AI infrastructure sizing remains uncertain. Much of the announced capacity is sovereign, project-based, and not yet commissioned. Chip-export approvals, energisation timelines, and power-delivery schedules remain key constraints.

Third-party analysts also apply different market definitions. Regional data centre market estimates for 2024-2025 vary from roughly US$2.4bn to US$3.5bn, depending on the scope included.

All forward-looking figures in this report should therefore be read as decision-support ranges, not audited accounts.

Executive Summary

The GCC is moving from cloud adoption to sovereign AI infrastructure at scale. National strategy, more than ordinary enterprise IT demand, is driving the shift.

AI capability now sits alongside energy security, digital sovereignty, and economic diversification as national infrastructure. Sovereign wealth funds and national champions are compressing what would normally be a decade-long build-out into a much shorter investment cycle.

Three factors define the opportunity.

First, capacity.

The GCC data centre project pipeline now exceeds US$93 billion across more than 174 tracked projects, according to MEED/GlobalData, May 2026 [1]. FTI Consulting expects GCC data centre capacity to increase from approximately 1GW today to 3.3GW by 2030, a projection independently echoed by PwC Middle East [11][20].

Second, compute.

In November 2025 the United States authorised exports of up to 35,000 NVIDIA GB300-class Blackwell chips each to Saudi Arabia's HUMAIN and the UAE's G42 [2]. In July 2026 it went further for the UAE, reclassifying it into the highest trust tier of its export control system, so approved UAE entities including G42 and Core42 can now source advanced NVIDIA and AMD chips without individual licences and without volume caps [19]. Saudi access continues under the earlier framework. HUMAIN also plans to deploy up to 600,000 NVIDIA GPUs across Saudi Arabia and the United States over the next three years under its expanded NVIDIA partnership [4].

Third, execution.

These systems need to be powered and cooled in 40°C-plus ambient conditions, in water-scarce economies, and at rack densities that exceed the practical limits of traditional air cooling.

Capital and chips are no longer the hard part; the scarce resource now is accountable capability across advisory, design, build, integration, and operations.

Ambition, capital, and demand are largely in place; the binding constraint now is delivery capability. Large-scale AI and HPC infrastructure in the GCC will depend on reliable power, high-density thermal design, low-water operation, integration depth, and the ability to run complex facilities at scale.

For CIOs, CTOs, data centre operators, sovereign infrastructure sponsors, and investors, the implication is clear: infrastructure decisions made in this cycle will determine whether future AI platforms can scale economically, reliably, and sustainably.



Key takeaways

01

National strategy is leading the market, not conventional demand.

Investment is anchored to Saudi Vision 2030, UAE Vision 2031, the UAE Digital Economy Strategy 2031, and sovereign-fund balance sheets. This gives the build-out stronger funding visibility than a typical enterprise IT cycle, while increasing exposure to policy, power, and geopolitical risk.

02

High-density AI has reached the limits of legacy cooling.

AI racks now draw 30kW to 150kW and rising. A single GB300 NVL72 rack draws roughly 132 to 135kW at rack TDP and up to about 155kW at peak, and is liquid-cooled by design [21]. Air cooling remains suitable for general-purpose workloads, but it is no longer sufficient for high-density AI clusters. Cooling architecture will be a core infrastructure decision over the next five years.

03

Sustainability is becoming a licence-to-operate issue.

In a region dependent on desalinated water, cooling-water consumption is a strategic constraint [11]. Mandatory water-use-effectiveness standards have not yet arrived, but operators that design for low-water, closed-loop, and immersion-based approaches now will hold a structural cost and siting advantage as the market scales.

04

The advisory gap is a commercial opportunity.

Sovereign and enterprise buyers are procuring large-scale AI infrastructure faster than they can build the internal expertise to specify, integrate, and operate it. Capital and chips are no longer the hard part; the scarce resource now is accountable capability across advisory, design, build, integration, and operations.

For decision-makers, the message is clear: the GCC is building sovereign AI infrastructure at scale and pace. Operators and partners that treat thermal performance and water efficiency as foundations for compute density will be best positioned to lead.

The region's hot, arid, power-rich, and water-constrained profile makes it one of the most demanding environments for high-density infrastructure. It also makes the GCC a high-value reference market for companies that can deliver reliable AI infrastructure outcomes under these conditions.

Submer

[02]

Market Overview & Sizing

GCC AI & HPC Infrastructure Report

Market Overview & Sizing

The GCC AI and HPC infrastructure opportunity is estimated using a defined addressable-market view, with assumptions clearly separated from third-party market baselines. Regional data remains fragmented across operator announcements, sovereign commitments, and analyst methodologies, so the analysis distinguishes external market references from the modelled TAM, SAM, and SOM used for cooling and high-density infrastructure planning.

2.1. Established Market Baselines

The market model is anchored in the following third-party figures:

METRIC	VALUE	SOURCE / NOTE
GCC data centre market, overall	US$2.39bn, 2025 → US$21.67bn, 2034	IMARC (2026 edition); 26.78% CAGR, 2026–2034 [22]
GCC data centre market, alternate definition	US$3.48bn, 2024 → US$9.49bn, 2030	Arizton via ResearchAndMarkets (Mar 2025); 18.2% CAGR [23]
GCC project pipeline	US$93bn+ across 174+ projects	MEED / GlobalData, May 2026 [1]
UAE investment pipeline	US$46.1bn	Coherent Market Insights, 2026
Saudi investment pipeline	US$35.3bn	Coherent Market Insights, 2026
Saudi installed IT load and pipeline	222MW, Q1 2025 + ~760MW planned by 2030	S&P Global, December 2025 [10]
UAE near-term capacity moves	Khazna 1GW plan; 200MW with Microsoft; Stargate UAE 1GW	DCD / Reuters, October–November 2025 [8] [12]
Global immersion-cooling market	US$5.72bn, 2026 → US$13.33bn, 2031	Mordor Intelligence, January 2026; 18.44% CAGR [5]
Global liquid-cooling market	US$6.77bn, 2026 → US$18.79bn, 2031	Mordor Intelligence, January 2026; 22.65% CAGR [6]



Reading the range

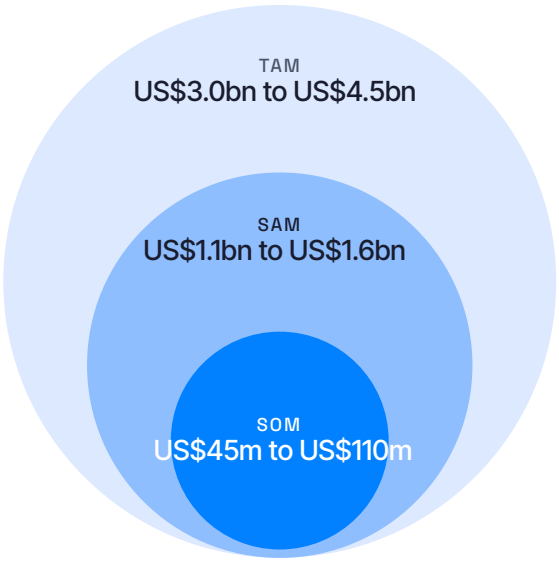
Published estimates vary because sources define the market differently. Lower values usually reflect narrower facility revenue, while pipeline figures capture announced project value and provide a broader view of committed capital.

For infrastructure planning, the US$93bn committed pipeline is the most useful reference point because it reflects capital already assigned to named projects, even if the timing of delivery remains uncertain.

2.2. TAM / SAM / SOM: Modelled Estimates, 2026

The model starts with committed AI- and HPC-relevant IT load across the GCC, where public commitments point to roughly 3.0GW to 3.5GW of AI-oriented capacity in build or firm planning for delivery by 2028-2030. This includes HUMAIN's approximately 500MW phase-one programme, DataVolt/NEOM's 1.5GW campus, Khazna/G42 and Microsoft's 200MW expansion, Stargate infrastructure, and projects in Qatar, Oman, and Bahrain. The capacity view is then converted into infrastructure spend using AI-grade data centre benchmarks of around US$9m to US$12m per MW of IT load, compared with roughly US$7m per MW for legacy facilities; the model uses US$10m per MW as a midpoint. For liquid-cooled AI facilities, thermal management and the high-density mechanical and integration scope addressable by specialist providers is assumed to represent approximately 12% to 18% of built cost, with 15% used as the model midpoint. This is conservative against published benchmarks: Turner & Townsend's Data Centre Cost Index 2025 puts the full mechanical package at 33% of facility cost in liquid-cooled AI builds, up from 22% in air-cooled designs [24], while McKinsey allocates around 25% to 30% of AI facility capex to combined power and cooling equipment [25].

GCC AI and HPC infrastructure: 2026 modelled estimates
Addressable market layers, US dollars

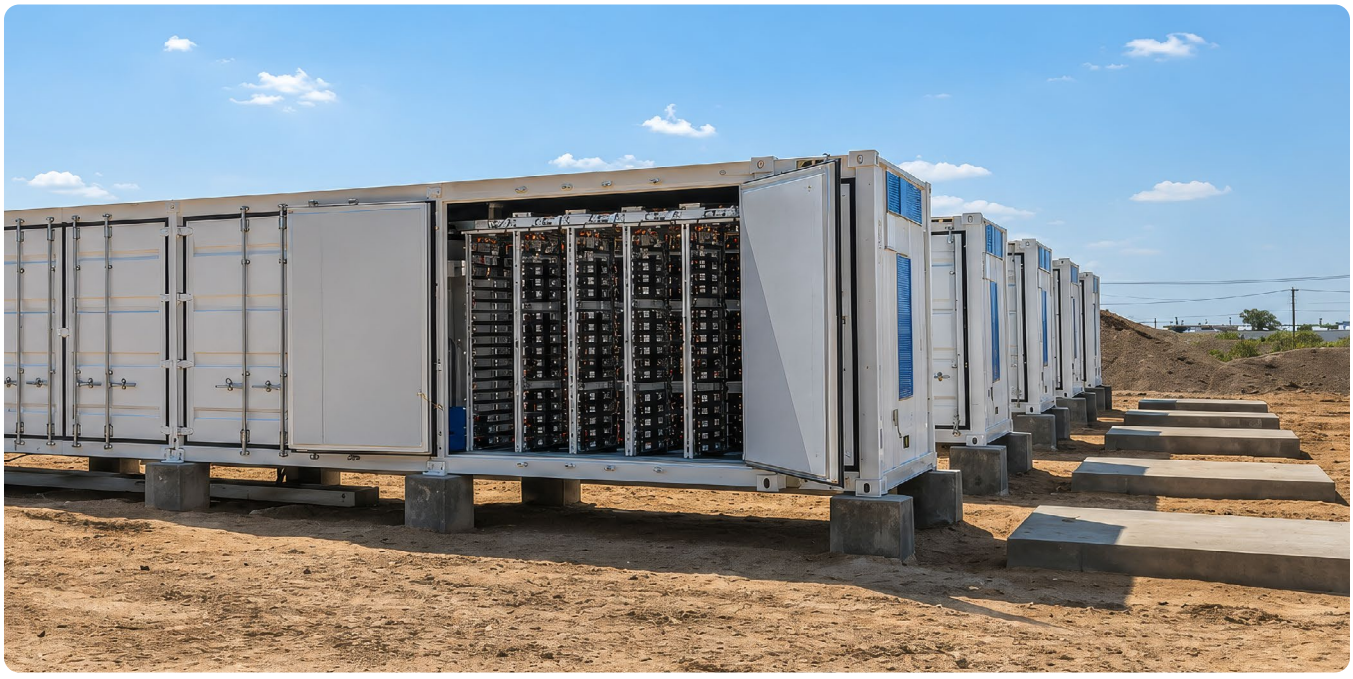


Total Addressable Market (TAM)
All GCC AI and HPC infrastructure spend reaching procurement in the 2026 cycle

Serviceable Available Market (SAM)
Cooling and high-density scope addressable by a liquid cooling specialist

Serviceable Obtainable Market (SOM)
Realistically obtainable 2026 revenue for a focused specialist

Source: Submer MEA modelled estimates. Circles are not drawn to scale.



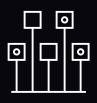
LAYER	2026 VALUE, MODELLED	DEFINITION AND ASSUMPTIONS
Total Addressable Market (TAM)	US$3.0bn – US$4.5bn	Total GCC AI/HPC infrastructure spend reaching procurement in the 2026 cycle, including facility, power, and thermal scope across all six GCC states and all cooling architectures.
Serviceable Available Market (SAM)	US$1.1bn – US$1.6bn	Serviceable share addressable by an advise-design-build-operate liquid cooling and high-density specialist. Includes the cooling and high-density mechanical/integration share of liquid-cooled AI builds in KSA, UAE, and leading Qatar/Oman projects.
Serviceable Obtainable Market (SOM)	US$45m – US$110m	Realistically obtainable 2026 revenue for a focused specialist executing 2-4 lighthouse engagements plus advisory and integration scopes. Represents approximately 4%-7% of SAM, consistent with an early mover building reference projects rather than pursuing commodity volume.

The TAM, SAM, and SOM therefore represent the 2026 addressable slice of a multi-year build programme, based on capacity reaching procurement or construction during the 2026 cycle, estimated at approximately 25% to 30% of the committed multi-year pipeline.

TAM provides the scale of the opportunity, but SAM is the more useful planning metric. Under this model, US$1.1bn to US$1.6bn of 2026-cycle spend is addressable by an integrated cooling and high-density infrastructure specialist.

The difference between SAM and obtainable revenue is not a demand issue. It reflects delivery credibility. Buyers in this market will not commit on technology claims alone; they will require proven deployment capability, local presence, operational maturity, and accountability across the full infrastructure lifecycle.

These figures are modelled estimates prepared for decision support, not third-party audited values. Sensitivity is concentrated in three variables:



Chip-delivery and export-licence timing, which determines how much capacity can be energised in 2026.



Liquid-cooling attach rate, which is expected to rise sharply but is not yet contractually fixed across all projects.



Capital intensity per MW, which varies by site conditions, renewable-power integration, grid readiness, and environmental constraints.

A ±30% sensitivity band is therefore appropriate around each figure.

Submer

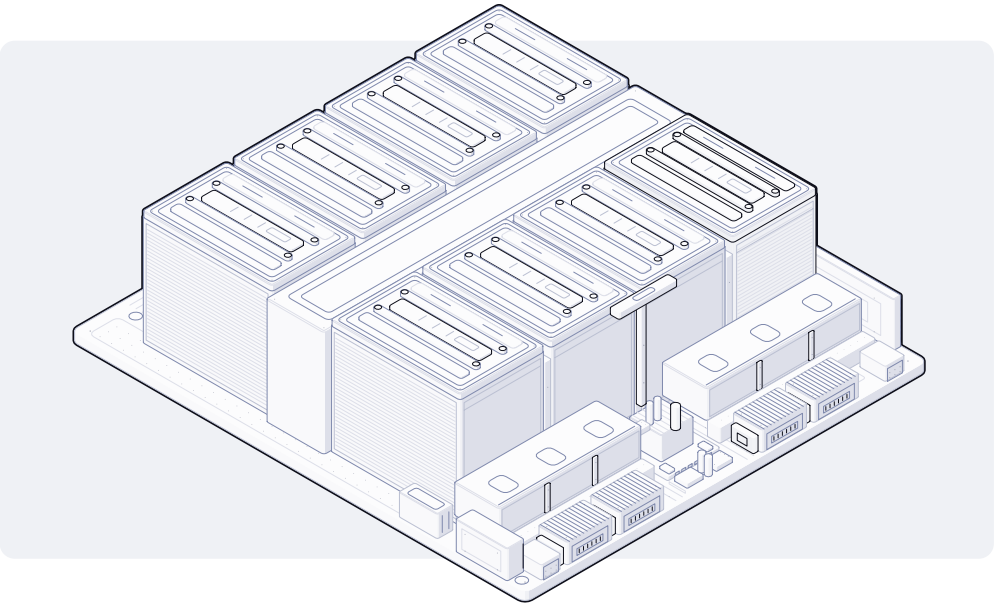
[03]

Industry Trends & Drivers

GCC AI & HPC Infrastructure Report

Industry Trends & Drivers

Five foundations will determine which GCC AI infrastructure deployments succeed over the next three to five years: secured grid power, resilient high-density thermal management, local operational sovereignty, water-free sustainability, and strong delivery models built on partner ecosystems. Each has a direct impact on facility design, project timing, capital allocation, and operating economics. This section examines each foundation, beginning with the compute catalyst that is stressing all five at once.



3.1 The Catalyst: Blackwell-Class Compute Arrives in the Gulf

The defining 2025-2026 development was the US Commerce Department's November 2025 authorisation of advanced NVIDIA Blackwell exports to the Gulf, reversing the earlier Biden-era restrictions [2]. Saudi Arabia's HUMAIN and the UAE's G42 were each cleared for up to 35,000 GB300-class chips. In July 2026 the US moved the UAE into the highest tier of its export control system, removing per-shipment licences and volume caps for approved UAE entities such as G42 and Core42 [19]. Saudi access continues under the November 2025 framework, so the two markets now operate under different export regimes.

HUMAIN's expanded NVIDIA partnership targets up to 600,000 GPUs over three years across Saudi Arabia and the United States [4]. The first phase, an 18,000-GPU GB300 Grace Blackwell supercomputer, is already underway [3] [17].

Up to 600,000 GPUs. HUMAIN's expanded NVIDIA partnership over three years.

Competition is increasing behind NVIDIA. AMD reported US$5.8bn in data centre revenue in Q1 2026, up 57% year on year [13], and HUMAIN signed a parallel AMD supply deal in May 2025 as part of a US$10bn programme [16]. Hyperscaler custom silicon (Google TPU, AWS Trainium, Microsoft Maia, Meta MTIA) is the fastest-growing accelerator category. For facility planners, the implication is hardware heterogeneity: sites must support NVIDIA, AMD, and custom ASICs, each with different power and thermal profiles, which favours vendor-neutral infrastructure over single-vendor reference designs.

3.2 Secured Grid Power: The Hard Currency of AI

Every megawatt spent pushing cold air through a room is a megawatt not running compute.

Power has become the hard currency of AI. The conversation used to be about how fast a shell could be built. Now it is about how many megawatts can be drawn from the grid, and how quickly.

The scale shift is best expressed in absolute terms. A decade ago a 30MW data centre was considered large; today a 200MW facility is considered normal, and gigawatt-class campuses are in construction [27]. In the GCC, the first 200MW of the Stargate UAE cluster is expected online in 2026 within a 1GW first phase [12], Microsoft and Khazna are adding 200MW before end-2026 [7][8], and AWS's US$5.3bn Saudi region launches in 2026 [14]. Each of these is a claim on grid capacity, and securing that capacity now takes longer than building the physical site: substations, transmission upgrades, and commissioning gate delivery regardless of capital availability.

This is where cooling stops being a purely thermal question and becomes an economic one. Every megawatt spent pushing cold air through a room is a megawatt not running compute. Advanced liquid cooling cuts facility overhead load, which means an operator extracts materially more AI capacity from the exact same grid allocation [26]. In a region where grid access is the binding constraint, thermal efficiency is as much a compute-capacity lever as a sustainability measure. Section 3.3 examines the technologies that deliver it.



3.3 Resilient High-Density Thermal Management

As rack densities move beyond 30kW and toward 150kW, traditional air cooling has reached its practical limit for high-density AI halls. Air cooling remains appropriate for general-purpose workloads, but not for the density profile leading AI platforms now require.

Two liquid cooling technologies dominate, with different cost structures:

Direct-to-chip (DTC) led 2025 liquid-cooling adoption, with 42.85% of liquid-cooling revenue [6]. Its advantage is integration speed, especially in existing data halls.

Immersion is the fastest-growing segment, forecast at a 26.6% CAGR through 2031 [5][6]. It supports densities above 100kW per rack, can achieve PUE levels of 1.02 to 1.03, and materially reduces fan energy.

COST DIMENSION	DIRECT-TO-CHIP	IMMERSION
ENTRY CAPEX	Lower; retrofit-friendly in existing halls	Higher; roughly 2 to 3 times air-cooled baseline [5]
WHAT DRIVES COST	Cold plates per server SKU, CDUs, secondary loop, plus retained air handling for residual heat (roughly 10% of load on GB300-class systems [21])	Tanks, dielectric fluid, and integration; near-total heat capture to liquid
OPEX OPPORTUNITY	Meaningful fan and chiller reduction versus air	Largest energy saving: fan energy largely eliminated, lowest achievable PUE [5]
REVENUE-SIDE EFFECT	Fastest path to AI-ready capacity in brownfield sites	Highest density per square metre; immersion-ready colocation commands 20 to 40% rack premiums [5]
WATER PROFILE	Depends on secondary loop design	Closed-loop, zero direct water consumption achievable

A single GB300 NVL72 rack draws up to about 155kW at peak and is liquid-cooled by design.

The practical pattern is hybrid: operators match DTC, immersion, and rear-door exchangers to workload within one facility. The capex premium of liquid resolves through opex over the facility life, because at AI scale energy is the dominant lifetime cost line, and industry cost indices already show mechanical scope rising from 22% to 33% of facility cost in liquid-cooled AI builds [24]. A full comparative TCO model is beyond the scope of this report.

Validation is no longer the question. The hyperscalers have moved liquid cooling into production at scale: Microsoft is deploying rack-scale direct-to-chip cold plate cooling alongside its latest GPU servers [30], and the region's flagship GB300 deployments are liquid-cooled as shipped [12][21]. OCP and ASHRAE reference designs continue to reduce adoption friction for risk-averse buyers.



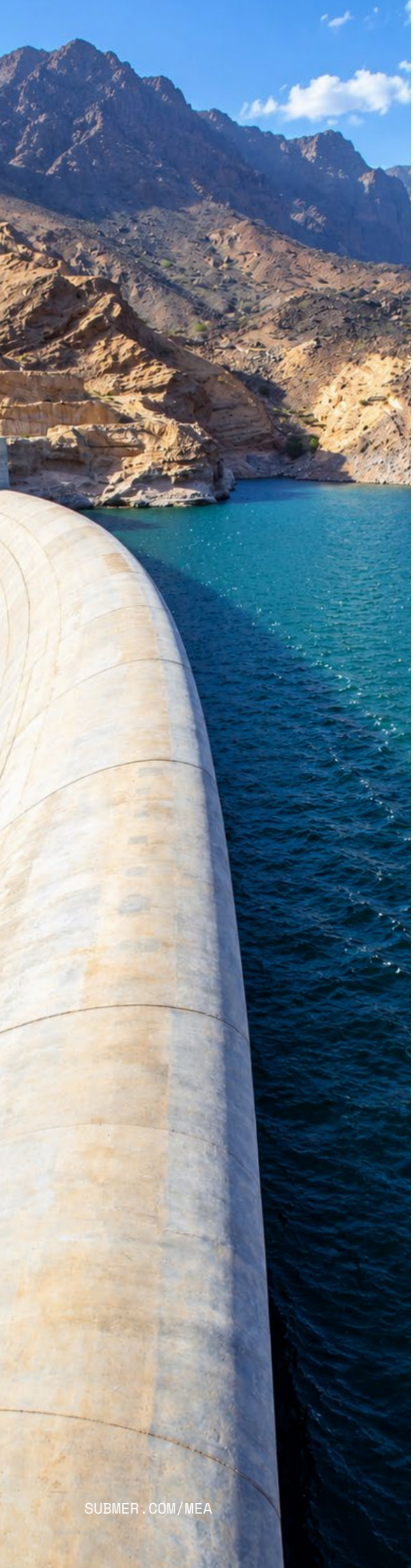
3.4 Local Operational Sovereignty

In the GCC, AI infrastructure is a state-backed priority, and the definition of sovereignty is maturing. Storing data inside a border is basic compliance. True AI sovereignty is operational independence: the ability to build, operate, and maintain critical AI infrastructure locally, which requires reliable power, resilient facilities, technology partnerships, supply-chain access, and the engineering expertise to keep national platforms running without external dependency.

The national commitments reflect this. In Saudi Arabia, 66 of Vision 2030's 96 strategic objectives are tied to data and AI [9], the Kingdom announced US$14.9bn in AI investment at LEAP 2025 within the US$100bn Project Transcendence programme [16], and HUMAIN sits under the Public Investment Fund. The UAE appointed the world's first AI Minister in 2017, established MGX as a dedicated AI-infrastructure investment vehicle, and set a 20% of GDP digital-economy target for 2031. Across the region, governments are building national AI platforms from the ground up, not merely buying cloud capacity.

Storing data inside a border is basic compliance. True AI sovereignty is operational independence.

The practical implication for buyers: if sovereignty is a capability rather than a location, then the first procurement step is a capability assessment. Sovereign buyers need a trusted advisor who understands what operational independence actually requires across power, thermal, integration, and operations, before hardware is specified. This is also why demand is concentrated among a small number of sovereign-backed entities (HUMAIN, G42, PIF, MGX, NEOM) whose procurement decisions shape the whole market.



3.5 Water-Free Sustainability

Water is the harder physical constraint. Saudi data centres reportedly consumed around 15 billion litres of water in 2024 [10][11]. A single 1GW facility can require 1.5 million to 5 million m^3 of water per year, with traditional evaporative cooling pushing that five to ten times higher [12]. These figures are analyst estimates rather than metered disclosures and should be read as directional, but the underlying exposure is real: in economies dependent on energy-intensive desalination, cooling water is a strategic cost, not just an environmental one.

The important point is that this constraint is avoidable by design. Closed-loop liquid cooling paired with non-evaporative secondary cooling achieves zero direct water consumption even under extreme heat loads, and dry or hybrid designs delivering very low water use are already the expected norm in new Saudi builds [12]. Regional operators increasingly expect zero-direct-water thermal management to shift from a preference to a practical requirement for site approvals as national sustainability programmes mature [26], though no binding GCC-wide standard exists today.

That last point matters for how this section should be read. Here, the pressure is commercial, not regulatory. There is currently no GCC mandate limiting PUE or water use. What exists instead is a TCO race: facilities that waste less energy and consume no potable water carry structurally lower operating costs, face fewer siting frictions, and align with the Saudi Green Initiative and Circular Carbon Economy frameworks that anchor national investment narratives. Against data centre hubs in other regions, the GCC's combination of low-cost power and greenfield, water-free design is a genuine cost advantage. Operators who build water-intensive facilities today are locking in a structurally higher cost base, one that will weigh on them as the market matures, whether or not a regulator ever intervenes.

One regional nuance is worth noting: waste-heat reuse, common in European efficiency narratives, has limited relevance in a 40°C-plus climate. The credible GCC sustainability story centres on zero direct water consumption, closed-loop efficiency, and reduced desalination burden.

Closed-loop liquid cooling achieves zero direct water consumption even under extreme heat loads.

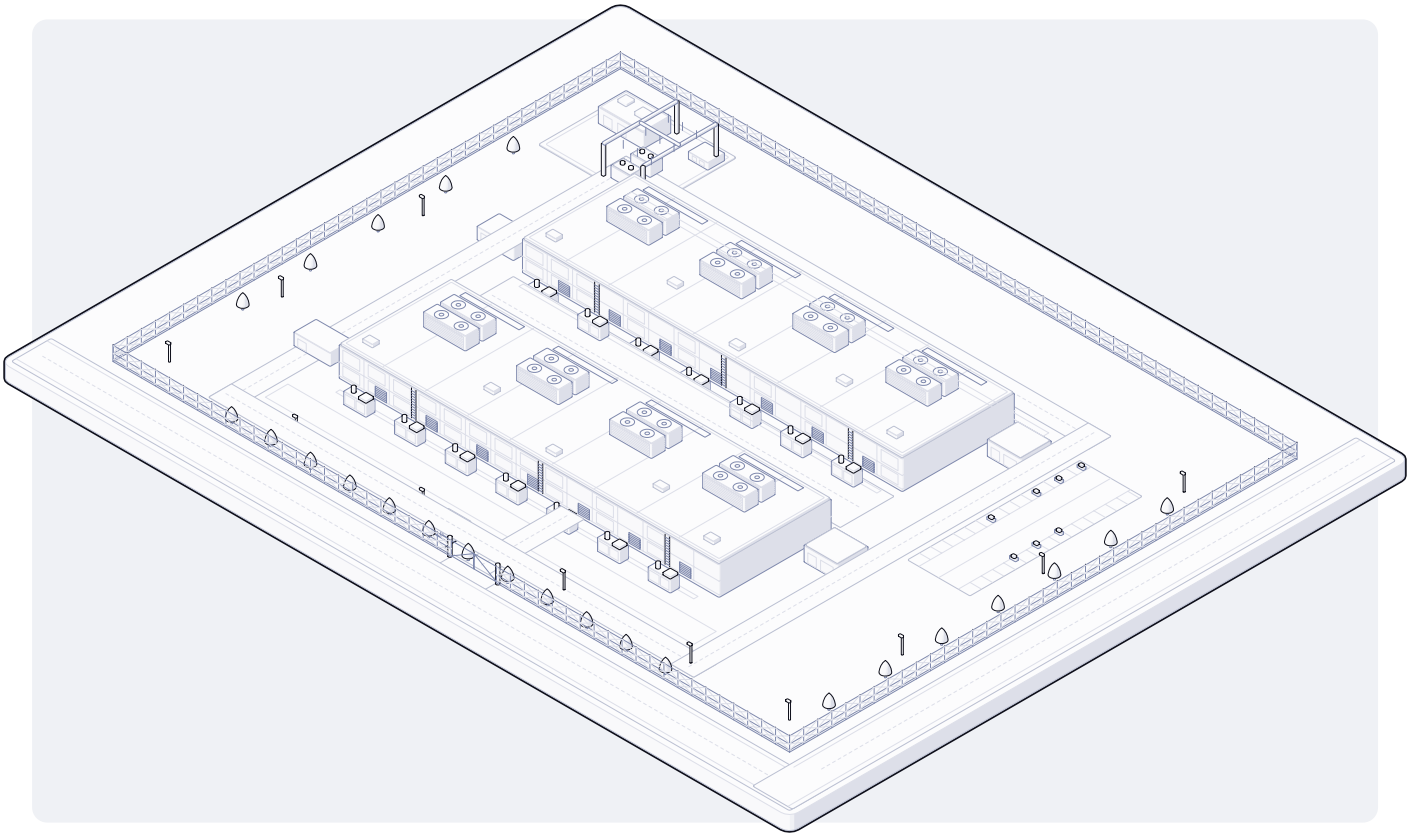
3.6 Delivery Models and Partner Ecosystems

The final foundation is how infrastructure gets delivered. Buying compute, networking, and thermal systems from five different vendors leaves operators spending more time troubleshooting interface points than running workloads [26]. AI infrastructure has become too complex for components to be considered separately, which is why sovereign buyers increasingly prefer single-accountability models where thermal, mechanical, and IT layers are designed as one system.

Two delivery shifts follow. First, modularity: pre-engineered, factory-validated building blocks compress deployment from years to months and let operators start with the capacity they need while keeping a clear path to expansion. Second, ecosystems: no single company owns every layer of the AI stack, so delivery quality now depends on tight coordination between chipmakers, OEM server manufacturers, software orchestrators, and mechanical infrastructure teams; validated partner integrations reduce deployment risk.

AI infrastructure has become too complex for components to be considered separately.

The hyperscaler overlay reinforces this. Microsoft's US$15.2bn UAE commitment [7], its Khazna expansion [8], and AWS's Saudi region [14] create hybrid topologies in which sovereign data stays on national platforms while burst and specialised workloads run on hyperscaler capacity. For the region's buyers, the winning posture is an integrated delivery partner operating inside a strong ecosystem, not a collection of component contracts.



Submer

[04]

Competitive Landscape

GCC AI & HPC Infrastructure Report

Competitive Landscape

The GCC AI infrastructure market is a layered value chain, not a single competitive arena. Compute vendors, data centre operators, cooling specialists, and integration partners compete on different dimensions, while buyers increasingly need these layers to work as one system.

4.1 Compute and Silicon Layer

PLAYER	CORE OFFERING	GCC POSITION	STRATEGIC NOTE
NVIDIA	GB300 Grace Blackwell, InfiniBand, Omniverse	Dominant; cleared for KSA and UAE	Estimated 80%+ accelerator revenue share; CUDA lock-in; anchor of HUMAIN and Stargate [2][3]
AMD	Instinct MI350 / MI355, ROCm	Credible second source; HUMAIN supply deal [13][16]	Open-software positioning; strongest in inference and cost-arbitrage use cases
Intel	Gaudi accelerators, Xeon CPUs	Limited DC accelerator share, below 1%	Cost-entry positioning; NVIDIA partnership has reopened relevance
Qualcomm	AI200 (2026) and AI250 (2027) rack-scale inference accelerators, Hexagon NPU heritage	HUMAIN anchor customer: 200MW deployment from 2026 [28]	Inference-first economics: 768GB LPDDR per card, liquid-cooled 160kW racks; evidence of training vs inference specialisation at the silicon level
Google TPU	Ironwood (TPU v7), GA April 2026; eighth generation splits into TPU 8t (training) and TPU 8i (inference)	Indirect via Google Cloud regions; Anthropic anchor at up to 1 million chips	First TPU positioned inference-first; 9,216-chip pods at 42.5 exaFLOPS; the most credible custom-silicon challenge to NVIDIA [29]
Hyperscaler ASICs	AWS Trainium/Inferentia, Microsoft Maia, Meta MTIA	Indirect exposure through cloud regions	Fastest-growing accelerator category, around 45% YoY; increases thermal heterogeneity



4.2 GCC Data Centre Developers and Operators

OPERATOR	BACKING	FLAGSHIP / SCALE	POSITIONING
HUMAIN	Saudi PIF	500MW programme; up to 600,000 GPUs [3][4]	Full AI value chain; sovereign anchor
DataVolt	Vision Invest, KSA	NEOM Oxagon US$5bn campus, net-zero [31]	Sustainable AI factory; renewable-powered
Khazna	G42 / MGX, UAE	UAE's largest operator; 1GW expansion plan; 200MW with Microsoft [8]	Hyperscale colocation; Stargate infrastructure layer
G42	Abu Dhabi	Stargate UAE 1GW; first 200MW in 2026 within 5GW campus [12]	Sovereign AI and cloud; Microsoft-linked
Others	Various	Equinix, Gulf Data Hub, Moro Hub, Edgnex, Ezditek, ICS Arabia	Colocation, edge, enterprise, and emerging local platforms

4.3 Thermal and Cooling Stack

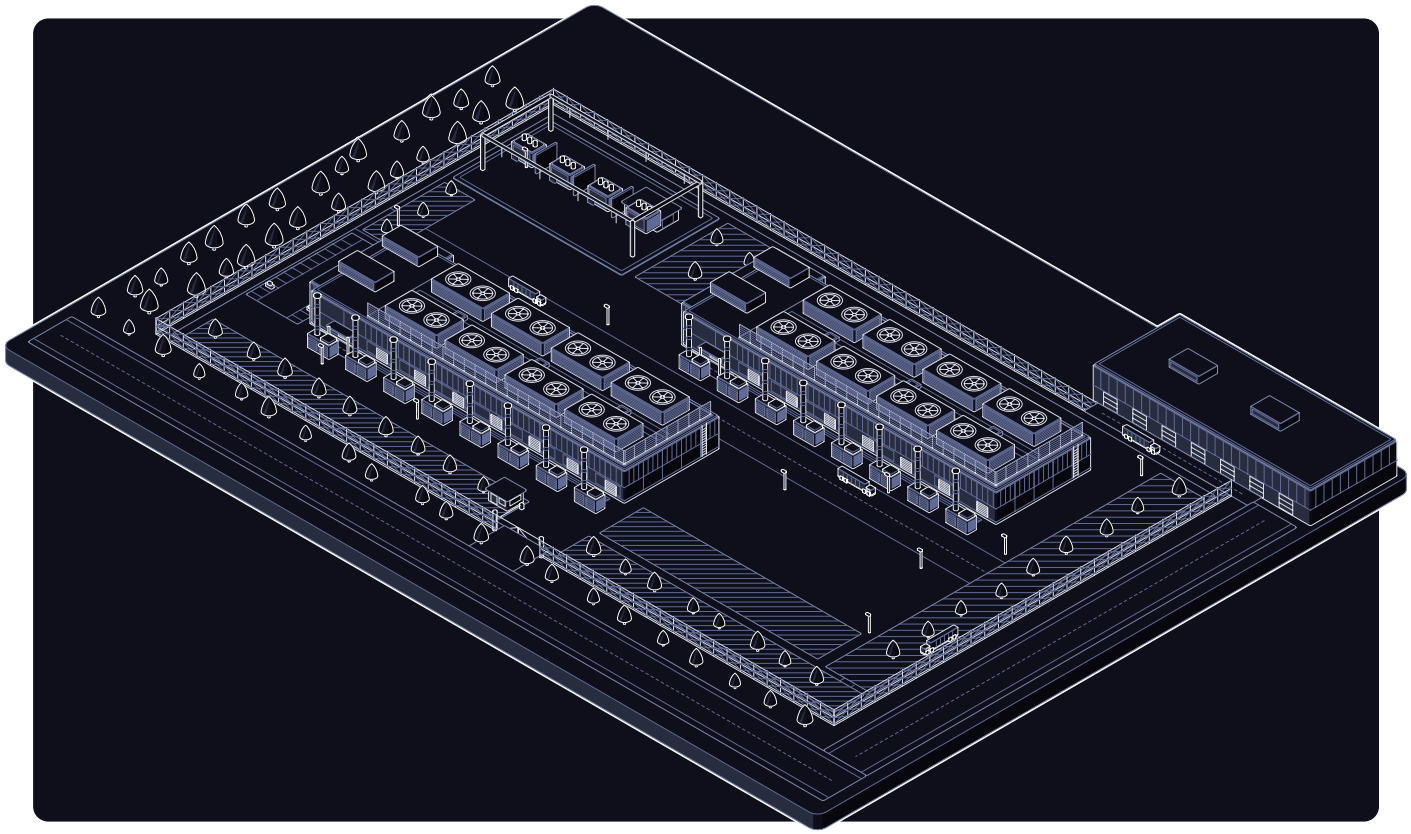
Thermal design is becoming one of the main determinants of facility economics. The comparison below assesses the main named players against the criteria that matter most for GCC AI deployment.

CAPABILITY	SUBMER	GRC	LIQUIDSTACK	ICEOTOPE	VERTIV / SCHNEIDER
Cooling portfolio breadth	Delivers both immersion and direct-to-chip liquid cooling at MW scale, backed by in-house system design	Single-phase immersion systems	Immersion (two-phase heritage) plus DTC	Chassis-level precision immersion	DTC, rear-door and hybrid air/liquid; full thermal portfolio
Density and efficiency	>100kW/rack; PUE 1.02-1.03, the lowest in this set	>100kW/rack; low PUE	>100kW/rack; low PUE	Chassis-level; low PUE	Competitive at DTC densities
Delivery and integration capability	End-to-end: advise, design, build, operate under single accountability, from chip level to live operations	Cooling systems supply; integration via partners	Systems supply; integration via partners	Technology and OEM partner model	Equipment and services at facility scale; integration typically via EPC contractors
Beyond-cooling stack	Full-stack: modular DC design-build, integrated AI compute, sovereign AI cloud and edge, access to 8GW+ land and power pipelines	Cooling-focused	Cooling-focused	Cooling-focused	Power, thermal and facility management portfolio
GCC / MEA presence	Saudi subsidiary, dedicated MEA leadership team	Project activity in the region [5]	Limited regional footprint	Limited regional footprint	Established regional channels and installed base
Water profile	Closed-loop, zero direct water consumption under extreme heat loads	Closed-loop immersion	Closed-loop immersion	Closed-loop immersion	Depends on architecture selected

In a multi-vendor hall, a row overheats and the finger-pointing begins... When one partner is answerable for thermal, mechanical, and IT as a single system, there is no interface to hide behind.

The pricing models indicate the market is separating into two commercial models. In the component model, vendors sell cooling products and leave integration risk with the customer, contractor, or facility operator. In the integrated model, the partner is accountable for the outcome across cooling, power, IT integration, commissioning, and operations.

For GCC sovereign buyers, the integrated model is increasingly stronger. These buyers are procuring at scale, often on compressed timelines, and may not yet have deep internal teams for liquid cooling and high-density integration. Single-accountability is best understood through the failure mode it removes. In a multi-vendor hall, a row overheats and the finger-pointing begins: the cooling vendor blames the server manufacturer, the manufacturer blames the network layout, and the operator is left owning a problem that nobody sold them. When one partner is answerable for thermal, mechanical, and IT as a single system, there is no interface to hide behind. The result is lower delivery risk, cleaner design decisions, and a shorter, more predictable path from specification to operational capacity.



[05]

Strategic Opportunities & Risks

GCC AI & HPC Infrastructure Report

Strategic Opportunities & Risks

5.1 Barriers to Entry

Capital and relationships

Flagship projects are sovereign-backed and relationship-led. Access is shaped by PIF, MGX, G42, NEOM, and national champions, not only by formal tender processes.

Power and grid access

Substation and transmission lead times can slow delivery even when capital is secured. Grid readiness is one of the clearest gating factors for large-scale AI campus deployment.

Specialist talent scarcity

Liquid-cooling design, fluid management, high-density integration, commissioning, and operations expertise remain scarce globally and are still limited in the GCC. This raises the entry barrier and strengthens the position of experienced specialists.

Localisation requirements

Data-localisation rules and in-country value expectations, including Saudi local-content requirements, favour companies with real local presence and proven regional delivery capability.



5.2 Risks

RISK	DRIVER	MITIGATION LENS
Geopolitical / export	Chip-export licences may tighten if US policy changes	Use vendor-neutral, multi-silicon cooling designs to reduce single-supply exposure
Water / sustainability	Desalination dependence and likely future WUE expectations	Prioritise closed-loop and liquid designs that anticipate regulation and provide improved competitiveness
Power delivery	Grid and substation lead times	Use modular, phased, high-efficiency builds to reduce avoidable MW overhead
Integration and security	Heterogeneous hardware and sovereign data sensitivity	Adopt single-accountability integration and sovereign-by-design architectures
Over-build / demand timing	Announced capacity may outpace energised demand	Phase commissioning against confirmed workload ramp and energisation milestones
Standards fragmentation	Proprietary, non-interoperable systems	Align with OCP and ASHRAE guidance where possible to reduce lock-in





5.3 Untapped growth opportunities

Cooling retrofit and greenfield design.
As liquid cooling becomes the default for AI workloads, every announced AI facility becomes a potential cooling and integration opportunity. This is the largest near-term addressable pool.

Advisory-first engagement.
Governments, sovereign funds, and national champions need AI and HPC infrastructure planning before hardware procurement. Partners involved early can shape specifications, delivery models, and accountability structures.

Centres of excellence and experience centres.
Hands-on environments help stakeholders test high-density technologies, reduce adoption risk, and build commercial confidence before committing to full-scale deployment.

Edge and inference build-out.
5G-driven edge capacity through operators such as stc, Mobily, and Zain, combined with the shift toward inference workloads, creates demand for compact liquid-cooled deployments beyond hyperscale campuses.

MEA export compute.
Gulf sovereign compute is increasingly positioned to serve the wider Middle East and Africa, including the 400M+ Arabic-speaking market through models such as ALLAM. This extends the opportunity beyond domestic workloads.

Integration services.
The strongest commercial position is built on accountable delivery across cooling, power, server integration, commissioning, and operations, well beyond any single component.

Submer

[06]

2026-2030 Future Outlook

GCC AI & HPC Infrastructure Report

2026-2030 Future Outlook

6.1 Expected Trajectory

2026

2026-2027: First AI capacity comes online.

Phase-one GB300 clusters from HUMAIN, Stargate, and Khazna/Microsoft are expected to enter production. Liquid cooling becomes the default requirement for new AI builds. Power readiness and chip-licence timing, more than demand, will determine the pace of delivery.

2027

2027-2028: Scale and sustainability converge.

DataVolt/NEOM's net-zero phase one, targeting 300MW, is expected to move toward operation. Immersion adoption accelerates as rack densities pass 100kW. If GCC WUE or efficiency guidance becomes binding during this period, operators that designed ahead of regulation will have an advantage.

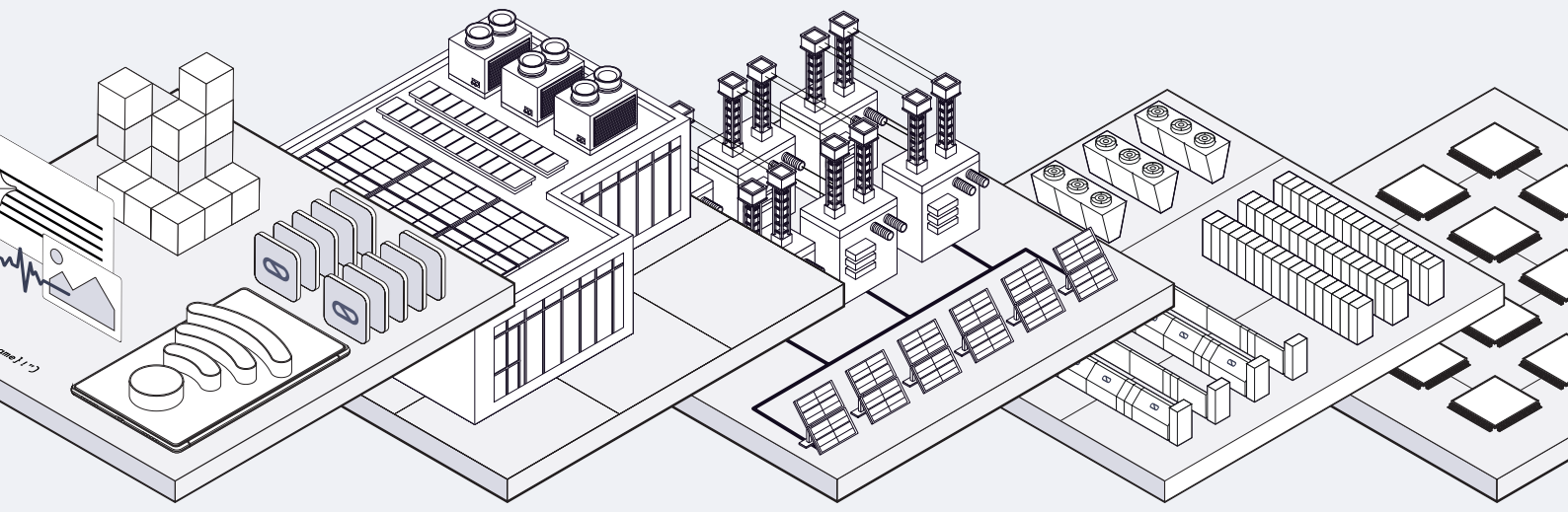
2028

2028-2030: GCC matures into a regional compute hub.

The GCC consolidates its position as a sovereign-AI and export-compute hub for MEA. Multi-GW campuses operate at scale, hybrid cooling architectures become standard, and value shifts from initial build-out toward operations, optimisation, and lifecycle services.

2030

Globally, the liquid-cooling market is expected to reach US$18.79bn by 2031, while immersion cooling is expected to reach US$13.33bn [5][6]. The GCC is likely to outpace these global curves because its climate, density ambitions, and water constraints make advanced cooling an operating requirement, not an option.



6.2 Recommendations

For IT and Infrastructure Leaders: CIOs, CTOs, and Data Centre Directors

Treat the accelerator roadmap as the thermal roadmap.

Design now for more than 100kW per rack. Retrofitting air-cooled facilities later will be more expensive, slower, and operationally disruptive.

Design for vendor neutrality.

Heterogeneous silicon is the base case. Facilities should support NVIDIA, AMD, and custom ASIC platforms without tying the cooling architecture to one hardware reference.

Buy outcomes, not boxes.

Prioritise partners that can take accountability across cooling, power, integration, commissioning, and operations.

For Government and Policy Stakeholders

Set efficiency standards early.

Clear guidance now will reduce the risk of water-intensive facilities that become expensive to retrofit later. It will also help align private capital with Vision 2030, UAE Vision 2031, and regional sustainability objectives.

Incentivise closed-loop and liquid designs.

Permitting, procurement, and financing can move the market toward lower water and energy designs aligned with the Saudi Green Initiative and Circular Carbon Economy frameworks.

For Investors

Look beyond hardware margin.

The more durable margin pool is likely to sit in lifecycle services and operations, especially as the market enters the 2028-plus operate-and-optimise phase.

Diligence the real constraints.

Power-delivery timelines, water strategy, export-licence exposure, local delivery capability, and operational readiness matter more than headline capacity announcements.

For Sustainability and ESG Leaders

Anchor the narrative in competitiveness.

In the GCC, credibility comes from being able to compete with other regions that are data centre powerhouses. Moreover, resource efficiency means reducing desalination burden and improving infrastructure design.

Instrument from day one.

Real-time granular data from the grey space down to the GPU must be integrated. This goes far beyond PUE and WUE and into the inner workings of everything from power distribution to the heat dissipation. Key stakeholders such as clients are coming to expect this level of measurement.



[07]

Where Submer Fits

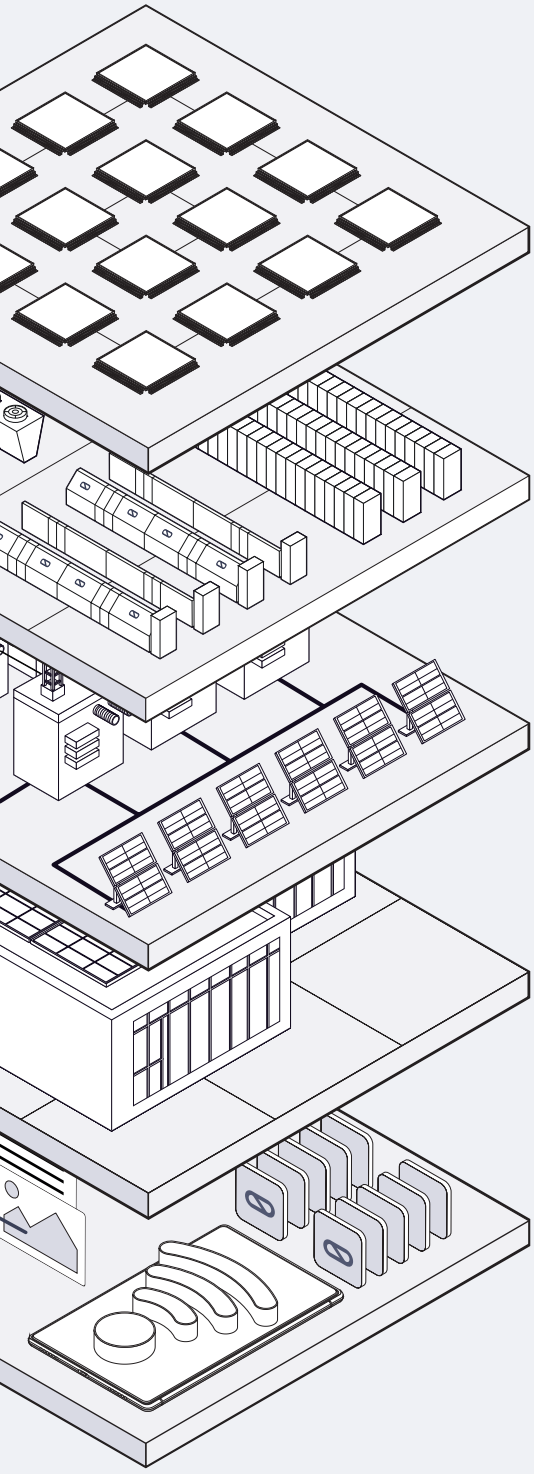
GCC AI & HPC Infrastructure Report

Where Submer Fits

The market conclusion is clear: the GCC's constraint is less about capital or access to chips than about the ability to deliver infrastructure that can power, cool, integrate, and operate high-density AI compute reliably in hot, water-scarce conditions. Submer's relevance sits in that delivery gap.

Submer aligns with the direction of the GCC market: densities above 100kW per rack, PUE levels down to 1.02, closed-loop low-water operation, and suitability for hot and arid environments.

Most of the commercial value lies around the cooling system, in accountable delivery across advisory, design, build, and operations, aligned with the infrastructure model GCC buyers are increasingly seeking.



Submer MEA
Trusted Advisor Across the Full Stack

Single-Accountability from Design to Operations (Advise → Design → Build → Operate).
Submer acts as an end-to-end infrastructure partner: a single design authority with integration governance across an otherwise multi-contractor delivery. One engineering vision and one accountable team carry the project from initial layout to final deployment, which eliminates friction between trades, cuts commissioning time, and ensures the facility hits its designed capacity. This is the single-accountability model GCC sovereign buyers increasingly prefer.

Thermal expertise at AI density.
Submer brings more than a decade of liquid-cooling deployment across both direct-to-chip and immersion, supporting densities above 100kW per rack with very low PUE and closed-loop, low-water operation suited to Gulf conditions.

Regional presence.
A Saudi subsidiary, UAE team, and dedicated MEA leadership team position Submer as a regional design-build and sustainability partner aligned with Saudi Vision 2030 and UAE Vision 2031.

Vertical integration.
Through its Group companies and strategic partnerships, Submer is active across every layer of AI data centre infrastructure: cooling, power, modular design and build, compute integration, and cloud and edge operations. This allows hyperscalers, hardware providers, and sovereign-compute initiatives to source turnkey, sustainable AI campuses from a single platform.

Advisory-first, lighthouse-led approach.
The commercial strategy is built around flagship reference projects, a regional centre of excellence, advisory engagement with governments and funds, and full IT integration services.

For decision-makers, the implication is consistent with the wider market analysis: GCC AI infrastructure winners will be those that treat thermal performance, water efficiency, and lifecycle accountability as core conditions for scaling compute density.

Submer's MEA proposition is positioned around that requirement: practical advisory, high-density cooling expertise, regional execution, and accountable delivery across the infrastructure lifecycle.

References

[1] MEED / GlobalData, The GCC Data Centre Projects Market 2026, May 2026. https://www.globaldata.com/store/report/gcc-data-centre-projects-market-analysis/

[2] US Department of Commerce, Statement on UAE and Saudi Chip Exports, 19 November 2025. https://www.commerce.gov/news/press-releases/2025/11/statement-uae-and-saudi-chip-exports

[3] NVIDIA Newsroom, HUMAIN and NVIDIA Announce Strategic Partnership to Build AI Factories of the Future in Saudi Arabia, May 2025. https://nvidianews.nvidia.com/news/humain-and-nvidia-announce-strategic-partnership-to-build-ai-factories-of-the-future-in-saudi-arabia

[4] Middle East AI News, HUMAIN to deploy 600,000 NVIDIA GPUs across KSA, USA, November 2025. https://www.middleeastainews.com/p/humain-to-deploy-600000-nvidia-gpus

[5] Mordor Intelligence, Data Center Immersion Cooling Market 2026–2031, January 2026. https://www.mordorintelligence.com/industry-reports/immersion-cooling-market-in-data-centers

[6] Mordor Intelligence, Data Center Liquid Cooling Market 2026–2031, January 2026. https://www.mordorintelligence.com/industry-reports/data-center-liquid-cooling-market

[7] Microsoft, Microsoft's $15.2 billion USD investment in the UAE, 3 November 2025. https://blogs.microsoft.com/on-the-issues/2025/11/03/microsofts-15-2-billion-usd-investment-in-the-uae/

[8] Data Center Dynamics, Microsoft and G42 expand data center capacity plans in UAE by 200MW, November 2025. https://www.datacenterdynamics.com/en/news/microsoft-and-g42-expand-data-center-capacity-plans-in-uae-by-200mw/

[9] Saudi Data and AI Authority, SDAIA and Vision 2030. https://sdaia.gov.sa/en/SDAIA/SdaiaStrategies/Pages/sdaiaAnd2030Vision.aspx

[10] S&P Global, The Saudi Arabia data center market: a catalyst for economic innovation, December 2025. https://www.spglobal.com/en/research-insights/special-reports/look-forward/data-center-frontiers/saudi-arabia-data-center-market

[11] Arab News, How Gulf states can develop data centers without straining scarce water resources, September 2025. https://www.arabnews.com/node/2612606/saudi-arabia

[12] Mobile Europe, Middle East datacenter ambitions face a water squeeze, July 2025. https://www.mobileeurope.co.uk/middle-east-data-centre-ambitions-face-a-water-squeeze/

[13] AMD Investor Relations, AMD Reports First Quarter 2026 Financial Results, 5 May 2026. https://ir.amd.com/news-events/press-releases/detail/1284/amd-reports-first-quarter-2026-financial-result

[14] AWS, AWS to launch first cloud infrastructure region in the Kingdom of Saudi Arabia by December 2026 (US$5.3bn planned investment, first announced March 2024). https://www.aboutamazon.com/news/aws/aws-cloud-region-saudi-arabia

[15] Rest of World, The Gulf's trillion-dollar AI dream faces a water crisis, 2025. https://restofworld.org/2025/gulf-ai-water-crisis/

[16] Mordor Intelligence, Saudi Arabia AI-Optimised Data Center Market, November 2025. https://www.mordorintelligence.com/industry-reports/saudi-arabia-artificial-intelligence-ai-data-center-market

[17] Data Center Dynamics, Saudi Arabian AI venture Humain buys 18,000 Nvidia GB300 chips, March 2026. https://www.datacenterdynamics.com/en/news/saudi-arabian-ai-venture-humain-buys-18000-nvidia-gb300-chips-several-hundred-thousand-more-on-the-way/

[18] MEED, Data centres present compelling opportunities, May 2026. https://www.meed.com/data-centres-present-compelling-opportunities

[19] Reuters, US makes it easier to export NVIDIA AI chips and military equipment to the UAE, 10 July 2026. https://www.reuters.com/world/middle-east/us-makes-it-easier-export-certain-military-items-ai-chips-commercial-satellites-2026-07-10/

[20] PwC Middle East, Unlocking the data centre opportunity in the Middle East, 2025. Corroborates the 1GW to 3.3GW GCC capacity projection. https://www.pwc.com/m1/en/media-centre/articles/unlocking-the-data-centre-opportunity-in-the-middle-east.html

[21] HPE, QuickSpecs: NVIDIA GB300 NVL72 by HPE (132kW rack TDP; ~155kW), 2026. https://www.hpe.com/us/en/collaterals/collateral.a50009244enw.html ; corroborated by Lenovo Press product guide LP2357 (135kW TDP, 155kW peak). https://lenovopress.lenovo.com/lp2357-lenovo-nvidia-gb300-nvl72-rack-scale-ai

[22] IMARC Group, GCC Data Center Market Report by Type, Infrastructure, IT Equipment, Tier Type, and Country, 2026–2034. https://www.imarcgroup.com/gcc-data-center-market

[23] Arizton (via ResearchAndMarkets), GCC Data Center Market: Investment Analysis & Growth Opportunities 2025–2030, March 2025. https://www.arizton.com/market-reports/gcc-data-center-market-investment-analysis-report

[24] Turner & Townsend, Data Centre Cost Index 2025. https://www.turnerandtownsend.com/insights/

[25] McKinsey & Company, The Cost of Compute: A $7 Trillion Race to Scale Data Centers, April 2025. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers ; see also McKinsey Global Institute, Colocation Data Centers: The Infrastructure Race Behind AI, June 2026. https://www.mckinsey.com/mgi/our-research/colocation-data-centers-the-infrastructure-race-behind-ai

[26] CXO DX, Redefining the AI Data Centre, interview with Khalid Aljamed, General Manager MEA, Submer, July 2026 edition, pp. 17-20. https://issuu.com/leapmediasolutions/docs/cxo_dx_july_2026

[27] McKinsey & Company analysis, reported in Network World, Data center spending to top $1 trillion by 2029, 2025. https://www.networkworld.com/article/3820973/data-center-spending-to-top-1-trillion-by-2029-as-ai-transforms-infrastructure.html

[28] Qualcomm Newsroom, Qualcomm Unveils AI200 and AI250: Redefining Rack-Scale Data Center Inference Performance for the AI Era, 27 October 2025; deployment details via Data Center Dynamics. https://www.datacenterdynamics.com/en/news/qualcomm-launches-ai200-and-ai250-chip-offering-targeting-inferencing-workloads-at-rack-scale/

[29] Google, Ironwood: The First Google TPU for the Age of Inference, Google Cloud Next 2025; technical documentation at Google Cloud. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/ironwood-tpu-age-of-inference/ and https://docs.cloud.google.com/tpu/docs/tpu7x

[30] Microsoft, Microsoft quantifies environmental impacts of datacenter cooling in new Nature study; Microsoft states it is deploying rack-scale cold plate (direct-to-chip) cooling and is not using immersion in datacenter operations, September 2025. https://news.microsoft.com/source/features/sustainability/microsoft-quantifies-environmental-impacts-of-datacenter-cooling-from-cradle-to-grave-in-new-nature-study/

[31] NEOM / Saudi Press Agency, NEOM and DataVolt launch sustainable AI project in Oxagon with a US$5bn first-phase investment; 1.5GW net-zero campus, operational from 2028, February 2025. https://www.spa.gov.sa/en/N2259124

Connected intelligence for AI infrastructure

Prepared by Submer MEA, Market Intelligence and Strategy. Figures labelled as modelled are Submer estimates based on the stated assumptions and should be read as decision-support ranges, not audited values. © 2026 Submer.

Author: Sami Alfaraj, Field CTO, Submer MEA