

AI Global
Alliance

WORLD
ECONOMIC
FORUM

In collaboration with Bain & Company

AI Infrastructure in the Age of Sovereignty: Requirements, Strategies and a Trusted Framework for Digital Embassies

WHITE PAPER

MAY 2026



Contents

Foreword	3
Executive summary	4
1 The evolving artificial intelligence infrastructure landscape	5
1.1 Dynamics shaping artificial intelligence infrastructure	5
1.2 The foundation of AI infrastructure	6
1.3 AI infrastructure choices that build resilience	8
2 Designing AI infrastructure strategies	9
2.1 Defining the AI sovereignty spectrum and its reference strategies	9
2.2 Examining local requirements of reference strategies	11
2.3 Deriving an economy-specific AI infrastructure strategy	16
3 Expanding AI infrastructure through digital embassies	17
3.1 Key challenges to overcome to realize digital embassies	17
3.2 Key benefits of trusted setups	18
3.3 A global framework for innovative and trusted digital embassies	19
Conclusion: Key considerations for AI ecosystem actors	23
Appendix	24
Contributors	26
Endnotes	28

Disclaimer

This document is published by the World Economic Forum as a contribution to a project, insight area or interaction. The findings, interpretations and conclusions expressed herein are a result of a collaborative process facilitated and endorsed by the World Economic Forum but whose results do not necessarily represent the views of the World Economic Forum, nor the entirety of its Members, Partners or other stakeholders.

© 2026 World Economic Forum. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, including photocopying and recording, or by any information storage and retrieval system.

Foreword



Cathy Li
Head, Centre for AI
Excellence; Member,
Executive Committee,
World Economic Forum



Florian Mueller
Senior Partner and Head,
AI, Insights & Solutions for
Europe, Middle East and
Africa, Bain & Company

Artificial intelligence (AI) infrastructure is rapidly becoming one of the most consequential strategic assets of the digital age. What was once viewed as a technical enabler is now a central determinant of national resilience and economic competitiveness.

Cumulative investment in AI-dedicated infrastructure exceeded \$600 billion between 2010 and 2024 and is projected to surpass \$400 billion annually by 2030.¹ These investments matter for reasons that extend well beyond technological advancements. Growing reliance on increasingly complex financing structures and greater use of debt mean that disruptions or underperformance can have spillover effects across the wider financial system. At the same time, AI infrastructure is becoming increasingly integral to the delivery of public services. Out of 33 Organisation for Economic Co-operation and Development (OECD) countries surveyed, 85% have adopted data-sharing systems and 73% have implemented digital identity solutions within public infrastructure.² These investments raise the stakes for cybersecurity, continuity and the protection of sensitive data and networks. Rising geopolitical tensions further heighten long-term dependency risks. Export controls, technology restrictions and fragmented supply chains make infrastructure choices increasingly difficult – and costly – to reverse.

Against this backdrop, the motivation to pursue AI sovereignty is stronger than ever. Yet for most economies, self-sufficiency is unrealistic given the scale, complexity and concentration of AI infrastructure required. Therefore, AI sovereignty must be pursued through strategic interdependence – built through deliberate choices about where to rely on trusted international partners under enforceable safeguards and where to retain domestic control. These decisions are becoming harder to make. Rapid technological advances, capital-intensive investments and

binding resource constraints – from energy and land to hardware availability – are reshaping the AI infrastructure landscape. Decisions made today will shape an economy's AI infrastructure capacity – and its resilience and competitiveness – for decades to come.

This paper, in collaboration with Bain & Company, is part of the World Economic Forum's AI Global Alliance's work on AI competitiveness. It builds on the Forum's previous thought leadership in this series: [Blueprint for Intelligent Economies: AI Competitiveness through Regional Collaboration](#) and [Rethinking AI Sovereignty: Pathways to Competitiveness through Strategic Investments](#). This paper describes AI infrastructure as a core element of sovereign AI ecosystems and AI competitiveness. It offers economies a practical foundation for making urgent strategic choices by deconstructing the evolving AI infrastructure landscape and examining how economies can design their sovereign AI infrastructure strategies by balancing international collaboration with domestic control.

Notably, the paper explores digital embassies as one option to extend access to AI infrastructure beyond national borders. Trust is a central challenge in shared AI infrastructure. The Forum and a global community of stakeholders have co-designed a framework to reduce uncertainty and strengthen trust in digital embassies. The Global Framework for Innovative and Trusted Digital Embassies, launched in this paper, is a blueprint for building resilient, future-ready digital embassies.

Above all else, this paper is a call to action for economies to make deliberate, strategy-led infrastructure choices. We encourage policy-makers to convene AI ecosystem stakeholders across the public and private sectors and join us. Together, we can advance AI competitiveness and strengthen global resilience.


Executive summary


Artificial intelligence sovereignty ambitions are rising alongside the stakes. Today's infrastructure choices will shape economic resilience and competitiveness for decades.


Artificial intelligence (AI) infrastructure³ has moved to the centre of global debates on competitiveness. Geopolitical tensions, the digitization of public services and intricate financing structures have made AI infrastructure a strategic concern – rather than a purely technological one. As demand for compute and data storage accelerates and architectures evolve, the AI infrastructure landscape grows increasingly complex. Economies must urgently prioritize long-term resilience in their infrastructure choices. That requires a clear, shared understanding of what constitutes AI infrastructure and what constrains it in practice.

1 The AI infrastructure landscape

The AI infrastructure landscape comprises three building blocks, each corresponding to a different state of data:

 **Compute** (“data in use”) refers to computing power and processing capacity used to train and deploy AI models and applications.

 **Connectivity** (“data in motion”) describes the digital network infrastructure that links data centres, edge or endpoint resources and end users.

 **Data storage** (“data at rest”) refers to the systems and facilities that securely hold and manage data at scale.

To develop these building blocks at scale, economies rely on a set of non-negotiable technical (energy, water, land, hardware, cybersecurity) and institutional (policy, talent, capital/financing) prerequisites.

2 Designing AI infrastructure strategies

AI sovereignty is becoming increasingly relevant in shaping AI infrastructure strategies. Economies can pursue AI sovereignty by balancing international collaboration with domestic control. This paper introduces an AI sovereignty spectrum, along which

economies can select a strategic direction (i.e. higher or lower interdependence) and define their AI infrastructure strategy in line with local capabilities.

To inform this decision-making, this paper describes two reference strategies at opposite ends of the AI sovereignty spectrum:

1. **Trusted international partnerships** – where sovereignty is achieved through shared infrastructure and governance with reliable partners
2. **Extensive domestic ownership** – anchored in local control of AI infrastructure

For each reference strategy, this paper outlines the AI infrastructure building blocks and technical prerequisites that must be locally available for the strategy to succeed. These reference strategies – trusted international partnerships and extensive domestic ownership – are not binary choices. Most economies pursue hybrid strategies that combine trusted partnerships with selective domestic ownership.

3 Creating trust in shared arrangements: a framework for digital embassies

Establishing and enforcing trust is essential in hybrid strategies, where critical AI infrastructure capabilities depend on external partners. The Forum, together with a global community of stakeholders, has developed a Global Framework for Innovative and Trusted Digital Embassies to help strengthen trust between economies and support resilient digital embassy agreements. When governed effectively, digital embassies can serve as a credible option for extending sovereign AI infrastructure beyond national borders.

Resilient AI infrastructure strategies require ecosystem alignment: policy-makers, investors and other ecosystem actors should align on a clear strategic direction and future-ready system design. Meanwhile, they should jointly manage binding resource constraints, enable trusted partnerships across borders and engineer resilience across cybersecurity, operations and finance.

1

The evolving artificial intelligence infrastructure landscape

To make sound strategic choices, economies must first understand the building blocks and prerequisites of artificial intelligence infrastructure.

1.1 Dynamics shaping artificial intelligence infrastructure

“ The AI infrastructure landscape is not only complex but also evolving rapidly.

Artificial intelligence (AI) infrastructure has become increasingly complex. Global demand for compute and data storage is accelerating rapidly, and the pace of technological change is making strategic choices more urgent.

The AI infrastructure landscape is evolving within a set of interconnected dynamics:

- 1 **Workloads are moving outward.** As AI applications shift from pilots to everyday use, inference demand is expected to grow far faster than training, pushing compute closer to users and sensitive data.⁴ Edge and on-device deployments are accelerating to enable real-time applications, such as autonomous systems and smart cities, and to meet regulatory compliance when data cannot move freely (e.g. by avoiding sharing sensitive data on centralized clouds⁵). Consequently, interoperable data architectures that enable portability and controlled sharing are becoming increasingly important.
- 2 **Physical constraints are becoming binding.** Power, cooling, land and hardware constraints are increasingly shaping **what** and **where** AI infrastructure can be built, reflecting the broader “AI-energy nexus”⁶ and its related impacts. These constraints are driving novel approaches, such as subsea data centres that use seawater for cooling, as well as a stronger focus on resource efficiency, including photonic computing⁷ (i.e. using light to perform computations rather than electricity) and optical interconnects that can deliver roughly tenfold gains in energy efficiency.⁸

- 3 **The frontier continues to scale.** Even as inference becomes more distributed, frontier training and large-scale simulation workloads⁹ are increasingly run on exascale-class systems to gain speed and precision. For example, France’s Alice Recoque supercomputer is scheduled to enter production in 2026.¹⁰ In parallel, storage and networking solutions are evolving to handle larger datasets and surges in AI-driven traffic.¹¹

- 4 **The risk baseline is rising.** As AI infrastructure becomes more distributed and system-critical, security is shifting towards privacy-preserving and resilience-first architectures. Federated learning, for example, enables model training across devices and organizations without moving raw data, embedding privacy by design instead (e.g. via smartphone-based learning¹²). At the same time, economies are hardening connectivity to ensure continuity and control through domestically governed satellite systems and quantum-secure networks. Europe has initiatives aimed at both: the Infrastructure for Resilience, Interconnectivity and Security by Satellite (IRIS²) constellation¹³ and the European Quantum Communication Infrastructure (EuroQCI) initiative.¹⁴

These dynamics underscore that the AI infrastructure landscape is not only complex but also evolving rapidly. Economies should shape their AI infrastructure choices around flexibility and future readiness to ensure resilience. Doing so requires a clear understanding of AI infrastructure’s building blocks, as well as the prerequisites for deploying them at scale.

1.2 The foundation of AI infrastructure

Building blocks of AI infrastructure

The AI infrastructure landscape is comprised of three building blocks:¹⁵ compute, connectivity and data storage. Given that AI fundamentally depends

on data – often described as the “fuel” of AI¹⁶ – it is useful to frame each building block around states of data. Figure 1 describes these in detail.

Each building block can be further disaggregated into a set of architectural configurations, as shown in Figure 2.

FIGURE 1 The three building blocks of AI infrastructure are framed around states of data







Building blocks	State of data	Description
 Compute	Data in use (e.g. being accessed, processed or modified)	Computing power and processing capacity required to train and deploy AI models and applications
 Connectivity	Data in motion (e.g. moving between systems, devices or networks)	Digital network infrastructure that links data centres for compute and storage, edge or endpoint resources and end users
 Data storage	Data at rest (e.g. stored on servers or the cloud)	Systems and facilities for securely storing and managing data at scale, enabling its use for AI training and inference

FIGURE 2 The building blocks of AI infrastructure, organized by architectural configuration

Building blocks	Architectural configurations	Description
 Compute Data in use	Centralized compute	AI processing performed in large-scale data centre or cloud regions (e.g. hyperscale data centres, large-scale inference clusters, AI supercomputers)
	Distributed compute	AI workloads spread across locations, often on smaller data centres or at network edges (e.g. micro data centres, edge inference clusters, multi-access edge computing)
	Endpoint compute	AI tasks run directly on local devices such as smartphones or internet of things (IoT) sensors (e.g. on-device AI models, inference accelerators, in-vehicle compute units)
 Connectivity Data in motion	Physical networks	Wired network infrastructure carrying data within and between data centres and other computing or storage sites (e.g. terrestrial fibre buildouts, subsea cables)
	Access networks	Last-mile and wireless infrastructure connecting users and devices to the internet or cloud services (e.g. 5G/6G mobile networks, satellite systems)
 Data storage Data at rest	Centralized storage	Data stored in a small number of designated locations (e.g. cloud platforms, data lakes, central backup vaults)
	Distributed storage	Data stored across locations, often on smaller data centres or at network edges (e.g. federated storage architectures, distributed storage networks, multi-site replication)

Prerequisites of AI infrastructure

Economies must consider the non-negotiable prerequisites required to develop, sustain and scale each building block. While technical prerequisites determine whether AI infrastructure is physically feasible at scale, institutional prerequisites determine whether such projects can be delivered and sustained over time.

- 1 Technical prerequisites** include access to energy, water, land, hardware and robust cybersecurity capabilities. They encompass both the physical and digital elements required to deploy and operate AI infrastructure. This paper focuses primarily on technical prerequisites, as they represent the most binding and widely shared feasibility constraints across economies.
- 2 Institutional prerequisites** include policy frameworks, skilled talent and access to capital and financing. These prerequisites enable the governance and economic conditions required to develop and sustain AI infrastructure over time. While institutional prerequisites remain essential, they are more context-specific and are only summarized at a high level in this paper.

Investigation of technical prerequisites and mitigation approaches

This section highlights key constraints for each technical prerequisite, as well as potential mitigation approaches.

⚡ Energy: AI infrastructure – especially data centres and accelerator clusters – requires large volumes of continuous power, and demand is growing faster than many grids can expand. The International Energy Agency (IEA) estimates that global data centre electricity consumption reached approximately 415 terawatt hours (TWh) in 2024 and could rise to 1,200 TWh by 2035 under current AI growth trajectories.¹⁷ A surge of this magnitude could strain generation and transmission capacity and complicate decarbonization objectives. Economies are responding by pairing renewable energy with firm, dispatchable and storable low-carbon sources (e.g. nuclear, hydro, geothermal) to meet round-the-clock demand. In parallel, some are exploring non-grid and distributed energy solutions – such as fuel-cell-based systems – to bypass grid constraints and reduce conversion losses.

🌊 Water: In many regions, cooling requirements make water a binding constraint for AI infrastructure. The largest AI-focused facilities may draw up to 5 million gallons (approximately 19 million litres) of water per day – comparable to the daily needs of a city of roughly 50,000 people.¹⁸ That places significant pressure on local water supplies. Mitigation approaches typically combine planning and technology, including watershed-aware site selection and increased use of recycled or non-potable water for cooling.

🏠 Land: AI-ready data centres require large sites with access to high-capacity power and fibre connectivity. A single AI training facility may require a minimum of 200 acres of land,¹⁹ which substantially narrows the number of viable locations and makes zoning, permitting and community acceptance decisive bottlenecks. Economies can reduce these barriers by pre-zoning suitable sites and streamlining approval processes. This approach was adopted by the UK's AI Growth Zones initiative,²⁰ which aims to fast-track permitting for data centre buildouts.

🔒 Hardware: Access to high-performance AI chips and critical supply chain steps remains concentrated among a small number of firms and geographies, creating significant supply chain risk. For instance, graphics processing unit (GPU) design is dominated by NVIDIA, based in the US,²¹ extreme ultraviolet lithography machines are produced exclusively by ASML in the Netherlands,²² and more than 90% of advanced foundry GPUs are produced by the Taiwan Semiconductor Manufacturing Company.²³ Mitigation strategies should therefore emphasize diversifying supply sources, building trusted alliances and selectively localizing parts of the value chain. Several economies are incentivizing domestic fabrication capacity, including through initiatives such as the US CHIPS and Science Act,²⁴ the European Chips Act²⁵ and India's Semiconductor Fabs Scheme.²⁶

🔒 Cybersecurity: As AI infrastructure becomes more distributed and critical to economic systems and starts hosting agentic workloads, it presents a larger attack surface and becomes a more attractive target. These risks are further amplified by AI-enabled threats, such as AI-enabled social engineering tools.^{27,28} Effective mitigation requires a layered cybersecurity posture, including strengthened national cyber capabilities.^{29,30} Economies can use AI to monitor critical networks – such as energy grids,³¹ hospitals³² and financial systems³³ – in real time and at scale. Where agentic workloads are deployed, agent-specific safeguards (e.g. containment boundaries) can be implemented. Strong governance mechanisms, including legal and institutional coordination frameworks, should complement these technical measures.

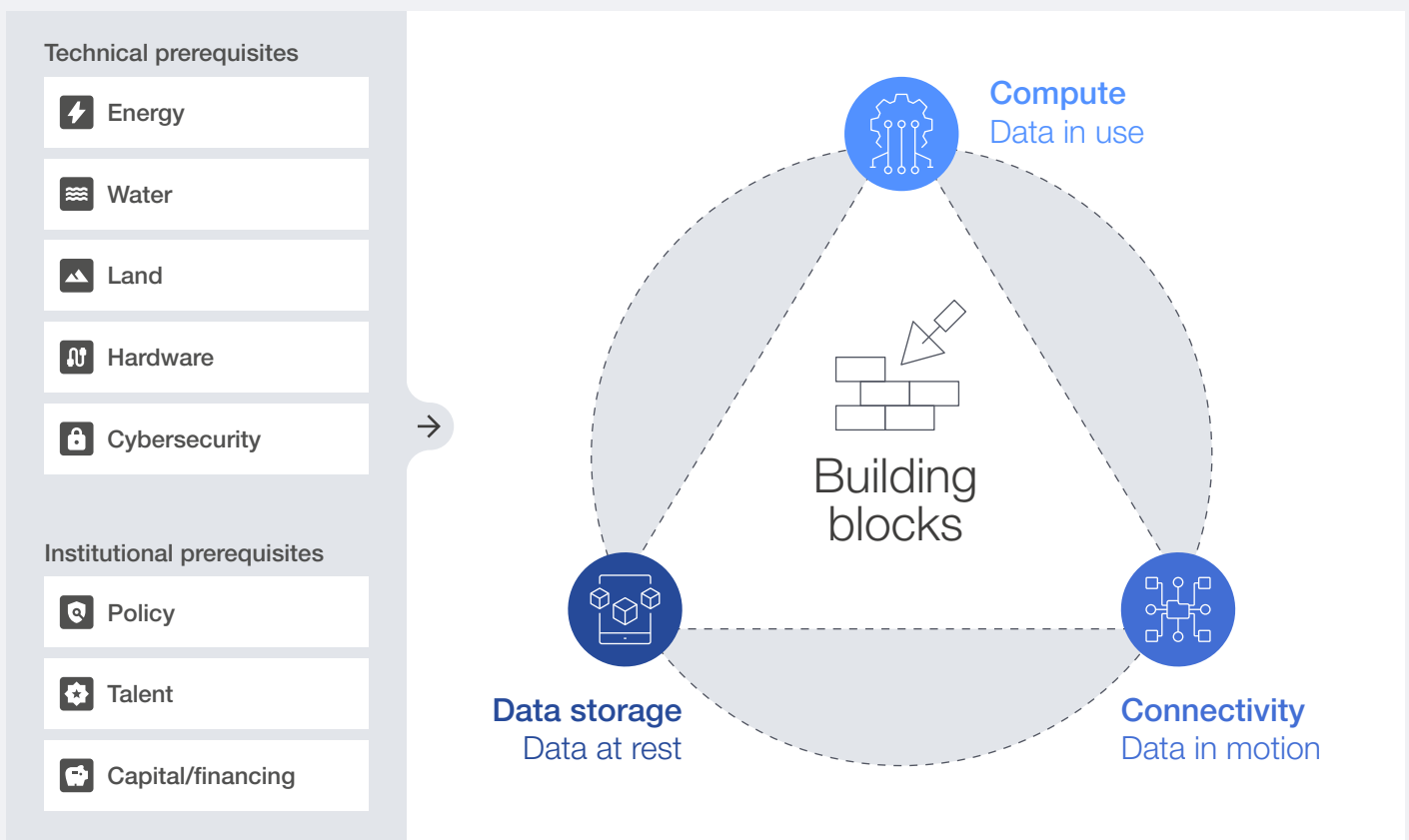


1.3 AI infrastructure choices that build resilience

Developing AI infrastructure requires coordination across the building blocks. Economies must also consider the technical and institutional prerequisites that determine whether these building blocks can be deployed at scale (see Figure 3). These choices are becoming increasingly complex, driven by a rapidly evolving AI infrastructure landscape and amplified by growing geopolitical and geoeconomic fragmentation.

As AI infrastructure becomes increasingly long-lived and system-critical, architectural decisions can create durable dependencies – across hardware supply chains, cloud and platform ecosystems, and cross-border data and connectivity routes – that are difficult to reverse. In this context, AI infrastructure choices shape not only an economy’s capacity but also its long-term resilience – while AI sovereignty agendas become even more relevant in AI infrastructure design.

FIGURE 3 The AI infrastructure landscape comprises three building blocks and their corresponding prerequisites



2

Designing AI infrastructure strategies

Economies can design AI infrastructure strategies by balancing international collaboration with domestic control in line with their local capabilities.

AI sovereignty agendas are gaining momentum amid growing geopolitical fragmentation, and AI infrastructure strategies are becoming a primary lever for building resilience and safeguarding sovereignty outcomes.

As outlined in *Rethinking AI Sovereignty: Pathways to Competitiveness through Strategic Investments*, AI sovereignty refers to the ability of economies to shape, deploy and govern AI ecosystems in accordance with their own values, while ensuring strategic and operational control,

flexibility and, ultimately, resilience through a combination of localized investment and trusted international collaboration.³⁴

A decision compass for sovereign AI infrastructure strategies

There is no single AI infrastructure model to achieve sovereignty. This chapter examines how economies can design AI infrastructure strategies based on their local capabilities and the degree of interdependence they are willing to accept.

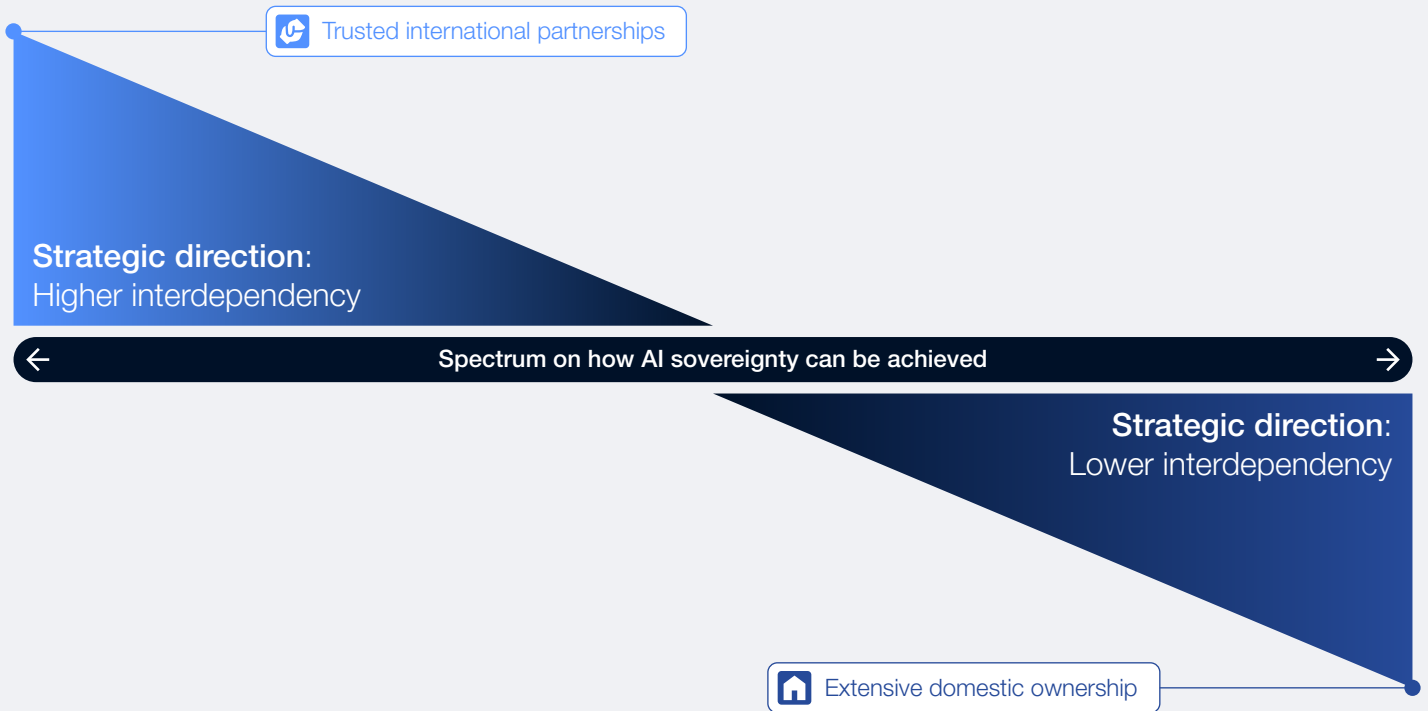


2.1 Defining the AI sovereignty spectrum and its reference strategies

Economies can pursue AI sovereignty through a range of strategies that fall along a spectrum of interdependence. Interdependence refers to reliance on foreign infrastructure, commercial partnerships and shared governance arrangements. Economies may pursue high strategic interdependence on one end of the continuum or low strategic interdependence on the other (see Figure 4).

This paper describes two reference strategies – one at each end of the AI sovereignty spectrum – for illustrative purposes. However, these reference strategies are not binary choices. In practice, most economies blend elements of both, balancing international collaboration with domestic control.

FIGURE 4 | AI sovereignty can be achieved along a spectrum of high to low interdependence



“ Economies can pursue AI sovereignty through a range of strategies that fall along a spectrum of interdependence.

Reference strategy:
trusted international partnerships

On one end of the spectrum, AI sovereignty is achieved through trusted cooperation with other economies or foreign companies – for example, by accessing high-compute facilities in a host economy. This strategy requires shared governance; an economy must secure AI compute and other infrastructure capabilities while retaining legal control and enforceable technical safeguards over critical workloads.

Several emerging infrastructure arrangements follow this model, including:

- **Digital embassies:** Sovereign capacity is delivered through secure spaces; an economy’s own laws govern its data, even when hosted abroad.
- **Pooled, multi-economy capacity:** The European High-Performance Computing Joint Undertaking, for example, allocates access to shared supercomputing infrastructure across participating European economies.³⁵

A partnership-based approach is best suited for economies with limited resources, those seeking to scale quickly or those pursuing resilience through cross-border data redundancy. For example, Estonia has secured data storage through a trusted international partner to ensure continuity of core state functions in the event of a domestic disruption. In 2017, Estonia signed

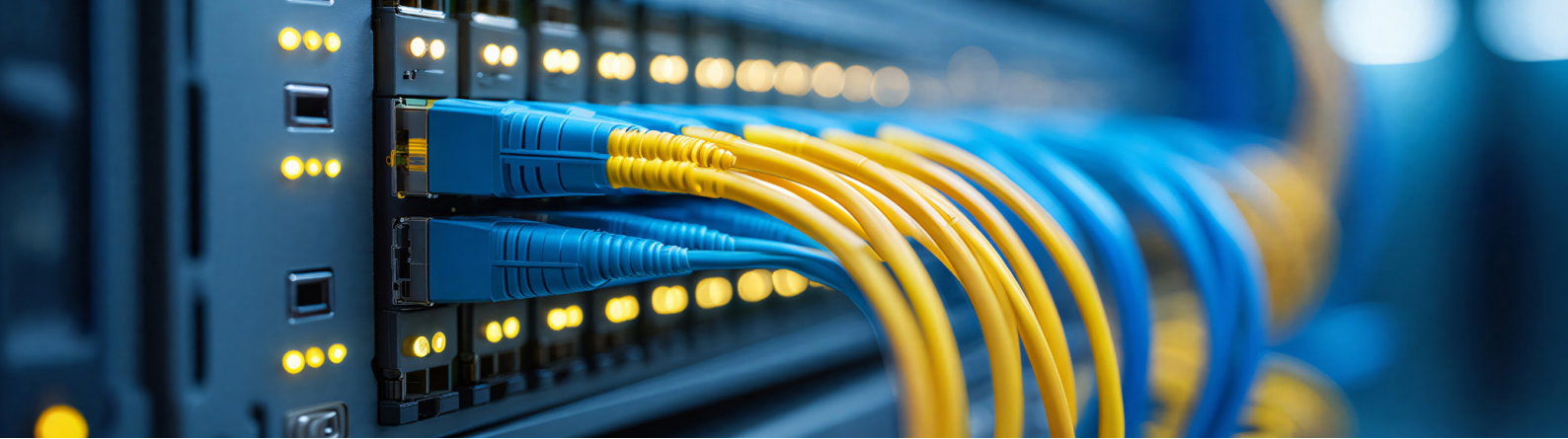
a bilateral agreement allowing government data and essential registries to be hosted in a secure facility in Luxembourg, under Estonian control.³⁶

Reference strategy:
extensive domestic ownership

The opposite end of the AI sovereignty spectrum is defined by domestic ownership and control. An economy builds and operates sovereign AI compute and data infrastructure on domestic soil under full local control, harnessing domestic providers.

In practice, this strategy often translates into government or nationally controlled infrastructure platforms and domestic hosting for sensitive datasets. It is most compelling for economies that have the domestic resources, budget and capabilities for domestic anchoring and that are looking to maximize assurance and control over sensitive workloads, such as defence.

Rethinking AI Sovereignty: Pathways to Competitiveness through Strategic Investments documented that few economies can pursue a “full stack” approach given the scale of investment required. Only a small number of economies – primarily China and the US – can come close to achieving extensive domestic ownership. For example, China has anchored its compute capacity with domestic providers, and its cloud infrastructure is operated largely by local firms. Alibaba Cloud, Huawei Cloud and Tencent Cloud jointly hold roughly 70% of cloud infrastructure market share.³⁷



2.2 Examining local requirements of reference strategies

To design their AI infrastructure strategy, economies should start by choosing a strategic direction along the AI sovereignty spectrum – i.e. whether to move towards higher or lower interdependence. Economies can use the two reference strategies at either end of the spectrum as anchor points for decision-making.

This paper sets out, for each reference strategy, the AI infrastructure building blocks that need to be available locally (requirement 1) as well as









the technical prerequisites that must be in place to implement each building block in practice (requirement 2).

Requirements for the trusted international partnerships strategy

Requirement 1:

Locally required AI infrastructure building blocks

FIGURE 5 The trusted international partnerships strategy requires local ownership of connectivity, but not compute or data storage

Building blocks	Locally required AI infrastructure building blocks
 Compute	 Not required at scale
 Connectivity	 Required at scale
Physical networks	 Required at scale to transport cross-border data traffic, particularly large volumes from cloud solutions, due to higher and more stable bandwidth capacity
Access networks	 Required at scale to enable local access to remote or shared infrastructure, particularly for edge systems – given the need for ubiquitous coverage and ultra-low-latency connectivity
 Data storage	 Not required at scale

 Required at scale*  Not required at scale*

Note: *At scale refers to capacity sufficient to provide high national coverage of domestic users and institutions continuously.

Implications:

- The trusted international partnerships strategy is anchored in local connectivity. Economies ensure robust domestic network infrastructure while securing compute and data storage through trusted international partners.
- Economies cannot substitute architectural configurations for connectivity; both

physical networks and access networks are required locally.

- Even when an economy relies on trusted partners for compute and data storage (under enforceable legal and technical controls), it may still pursue select local compute and data storage as a domestic complement – for example, to reduce latency or support sensitive use cases.

Requirement 2:

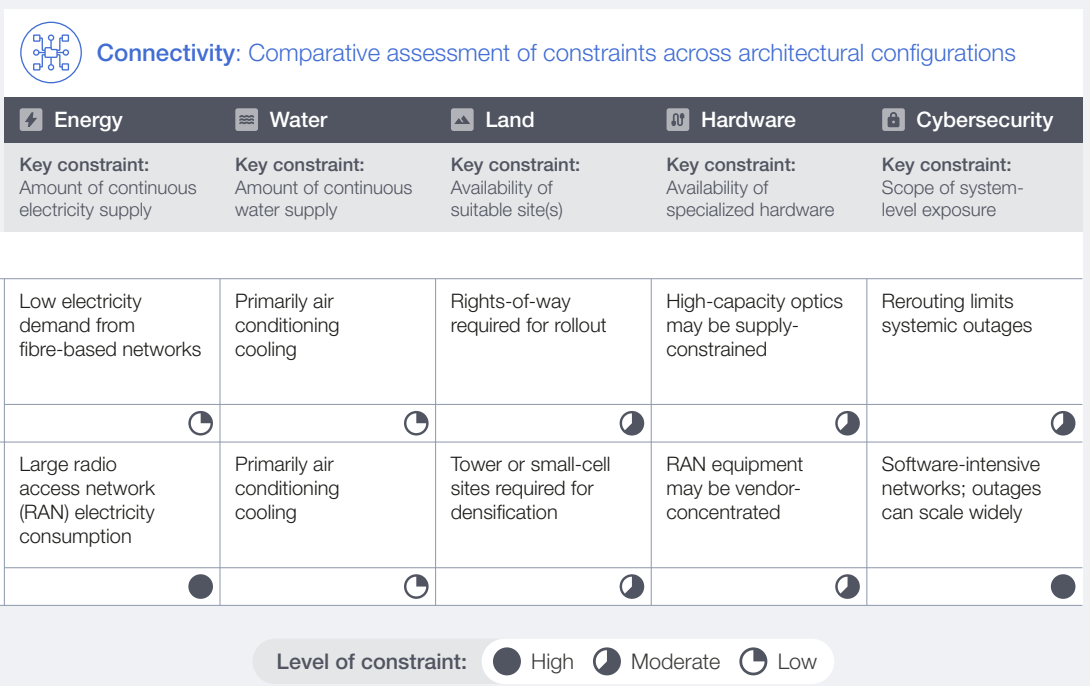
Technical prerequisites for developing AI infrastructure building blocks locally

Connectivity is typically less constrained overall by technical prerequisites than large-scale compute or data storage and is therefore faster and less capital-intensive to scale. Connectivity generally requires lower absolute levels of energy, water, land and specialized hardware and can often be expanded incrementally, reducing implementation risk. In contrast, compute and data storage – particularly in centralized configurations – require continuous high-load electricity, cooling (often with significant water use), large, permitted sites and scarce specialized hardware.

Economies pursuing a trusted international partnerships strategy should develop both physical networks (e.g. fibre backbones and submarine cables) and access networks (e.g. mobile radio networks and fixed last-mile connectivity). However, these layers face materially different technical constraints at the local level. Disaggregating connectivity helps economies identify which prerequisites are most likely to constrain progress and where targeted interventions are required.

Figure 6 compares the technical prerequisites across physical and access networks. The comparison does not imply a choice between connectivity layers; rather, it helps inform prioritization and design.

FIGURE 6 Comparative assessment of technical prerequisites across connectivity



Implications:

- Physical networks are moderately constrained by land, hardware and cybersecurity; energy and water are typically even less binding (see the Appendix for more detail).
- Access networks are most constrained by system-wide energy requirements and cybersecurity exposure (see the Appendix for more detail).
- This asymmetry suggests that economies could treat connectivity as two distinct implementation challenges: for physical networks, focus on securing land permits, sourcing high-capacity network hardware and building redundancy for failover; for access networks, prioritize managing energy consumption and ensuring robust cybersecurity measures at scale.

Key takeaways for economies considering a trusted international partnerships strategy:

- 1 Economies following this strategy require locally available connectivity, encompassing both physical networks

and access networks. Compute and data storage can be sourced through trusted partners under enforceable controls.

















- 2 The prerequisites for compute and data storage are more demanding than those for connectivity overall. As a result, a trusted international partnership strategy can be faster to implement.
- 3 When implementing physical networks and access networks, economies should consider their distinct implementation challenges.
- 4 Increased reliance on cross-border connectivity and partner performance heightens exposure to external risk, underscoring the need for strong polices, robust governance and resilience planning.

Requirements for the extensive domestic ownership strategy

Requirement 1:

Locally required AI infrastructure building blocks

FIGURE 7 **The extensive domestic ownership strategy requires local ownership across compute, connectivity and data storage**

Building blocks	Locally required AI infrastructure building blocks
 Compute	 Required at scale
Centralized compute	 Required at scale to train frontier models and run inference for national platforms or defence-grade AI solutions, as it enables economies of scale for large model training and supports stronger technical and operational controls
Distributed compute	 Supplementary option**
Endpoint compute	 Not required at scale
 Connectivity	 Required at scale
Physical networks	 Required at scale to transport data traffic, particularly large volumes from cloud solutions, due to higher and more stable bandwidth capacity
Access networks	 Required at scale to enable access to domestic infrastructure, particularly for edge systems, given the need for ubiquitous coverage and ultra-low-latency connectivity
 Data storage	 Required at scale
Centralized storage	 Required at scale for efficiency in governance of large national datasets
Distributed storage	 Supplementary option***
 Required at scale*  Supplementary option  Not required at scale*	

Note: *At scale refers to capacity sufficient to provide high national coverage of domestic users and institutions continuously. **Technically possible as sole sovereign compute option in specific contexts, such as inference-heavy, latency-sensitive or federated workloads. ***Technically possible as sole data storage option in specific contexts, such as federated, multi-site or edge storage.

Implications:

- An economy must have all three AI infrastructure building blocks available locally to pursue an extensive domestic ownership strategy.
- However, economies can decide which architectural configurations to provide locally for compute and data storage. While centralized compute and data storage are required locally, economies can assess whether to add distributed options based on priorities such as resilience and latency. Endpoint compute remains optional at scale, but it is required for certain applications and can support privacy and offline continuity.

Requirement 2:

Technical prerequisites for developing AI infrastructure building blocks locally

An extensive domestic ownership strategy requires economies to develop compute, connectivity and data storage locally. Doing so entails meeting a significant set of technical prerequisites – particularly

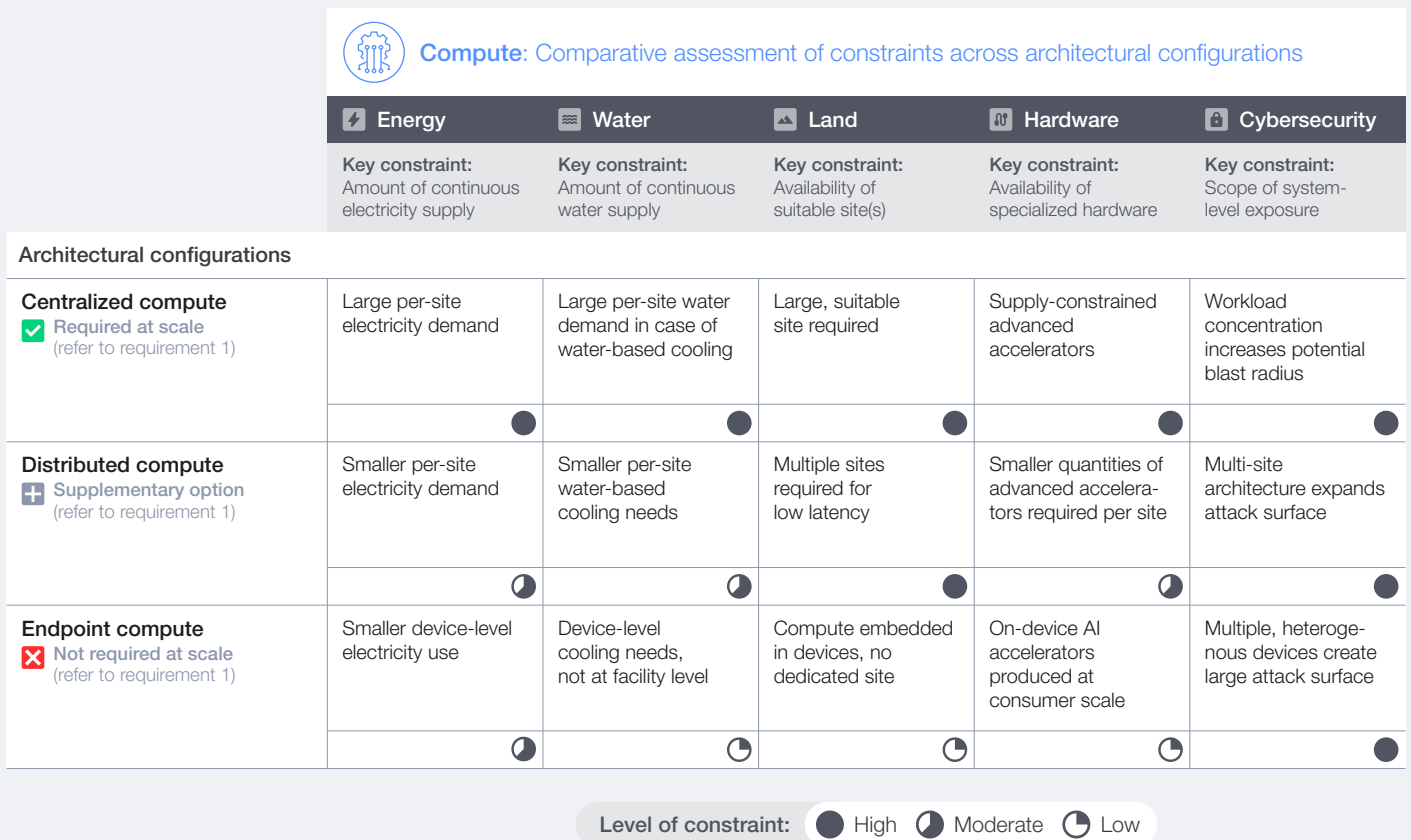
for the local buildout of large-scale, centralized compute and data storage infrastructure. However, economies retain architectural choices – primarily within compute and data storage (e.g. distributed versus endpoint compute). Local technical prerequisites can guide these decisions.

There are key differences in technical prerequisites across architectural configurations for compute and data storage. Understanding these options can help economies make informed design and investment decisions.

Figure 8 compares the technical prerequisites across centralized, distributed and endpoint compute. Figure 9 does the same across centralized and distributed data storage. These comparisons do not imply a choice among the architectural options for compute and data storage respectively; rather, they help inform prioritization and design.

Connectivity prerequisites are not reassessed here, as the considerations outlined for the trusted international partnerships strategy apply equally in this context.

FIGURE 8 Comparative assessment of technical prerequisites across compute

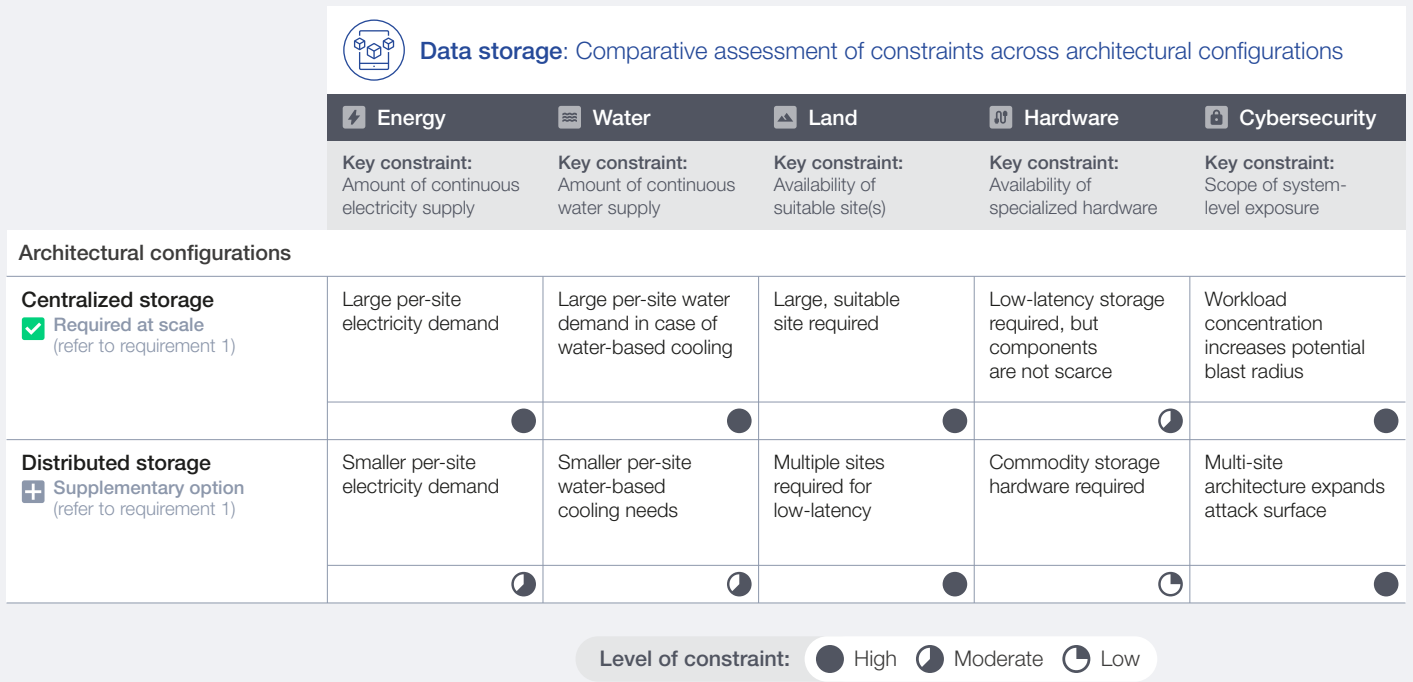


Implications:

- Centralized compute is heavily constrained across all major technical prerequisites. Distributed compute, by contrast, is comparatively less constrained. Deploying multiple, smaller compute facilities reduces pressure on aggregated power and water supply and specialized hardware needs. This makes distributed compute an attractive supplement (see the Appendix for more detail).

- Economies should consider a tiered architecture model when designing compute infrastructure. In this model, only the most sensitive or demanding workloads should be placed in centralized environments, while others should be deployed in distributed ones. All inference workloads are already shifting towards distributed infrastructure to achieve lower latency and localize sensitive data.³⁸

FIGURE 9 Comparative assessment of technical prerequisites across data storage



Implications:

- As with compute, technical prerequisites tightly constrain centralized data storage infrastructure, while distributed storage architectures are comparatively less constrained due to their modular design (see the Appendix for more detail).
- This asymmetry supports a tiered approach to data storage, in which the most sensitive datasets are anchored in tightly controlled, centralized environments. Distributed storage solutions can complement this model to improve security and resilience. Strong encryption, access controls and continuous monitoring are essential to ensure data sovereignty in such environments.

Key takeaways for economies considering an extensive domestic ownership strategy:

- 1 An extensive domestic ownership strategy requires a comprehensive set of AI infrastructure building blocks built and operated locally. This includes at least one domestic compute capability and one domestic data storage capability, complemented by robust physical and access connectivity.
- 2 Technical prerequisites for compute and data storage represent the most binding constraints of the domestic ownership strategy.
- 3 Because centralized compute and data storage options place particularly high demands on technical prerequisites, economies should design tiered architectures. The most sensitive workloads should be anchored in tightly controlled domestic environments, with complementary distributed options to enhance resilience.

2.3 Deriving an economy-specific AI infrastructure strategy

Having reviewed the requirements associated with each reference strategy, economies can develop an initial view of their preferred strategic direction along the AI sovereignty spectrum – whether closer to trusted international partnerships (higher interdependency) or extensive domestic ownership (lower interdependency). Subsequently, they can start translating this strategic direction into an economy-specific AI infrastructure strategy.

In practice, few economies sit at either extreme of the AI sovereignty spectrum. Most adopt hybrid strategies, combining trusted partnerships with foreign countries or commercial entities with selective domestic ownership to balance control, affordability, speed and security.

For instance, economies may meet their sovereign infrastructure needs by partnering with trusted hyperscalers for elastic capacity or scale, while

using legal and technical safeguards to keep sensitive workloads under national control. Singapore, for example, combines strong domestic governance with extensive use of trusted hyperscalers through its Government on Commercial Cloud (GCC) model. More than 70% of Singapore's eligible government systems are already on GCC.³⁹ At the same time, Singapore anchors sensitive capabilities domestically through its National Supercomputing Centre,⁴⁰ which received SGD 270 million (Singaporean dollars) in 2024 to build a next-generation supercomputer.⁴¹

Hybrid approaches underscore the central challenge of how trust is established and enforced when critical AI infrastructure capabilities depend on external partners. Digital embassies are one mechanism that can provide shared access to compute and data storage; under trusted agreements, they ensure sovereign control.

Expanding AI infrastructure through digital embassies

With trust, digital embassies can enable access to shared AI infrastructure. A global framework can help build it.

“ Digital embassies have resurfaced as a viable sovereign infrastructure option within the broader landscape of national AI infrastructure strategies.

As demand for compute and data storage accelerates, many economies – particularly emerging ones – find it challenging to meet technical and institutional prerequisites (e.g. land, energy, capital) for building and scaling high-assurance domestic AI infrastructure in short time periods. Digital embassies offer a mechanism to extend access to sovereign AI infrastructure beyond national borders. They enable the creation of secure digital spaces in a host country where a guest country’s (or organization’s) data, workloads storage and compute remain governed under its own jurisdiction.

The concept was inspired by the need to ensure continuity of critical digital services during national disruptions. Estonia pioneered the concept of hosting government data and essential registries in a secure, legally protected facility outside its borders.⁴² It signed a bilateral agreement with Luxembourg in 2017 following large-scale cyberattacks in 2007, recognizing its dependence on digital public infrastructure.⁴³ This arrangement demonstrated that national digital operations could be protected and restored if domestic systems were compromised – and that sovereign control over digital operations could extend extraterritorially. In 2021, Monaco entered into a similar agreement

with Luxembourg.⁴⁴ These cases established the first practical precedents for digital embassies.

Today, digital embassies have resurfaced as a viable sovereign infrastructure option within the broader landscape of national AI infrastructure strategies. They expand the traditional notion of digital sovereignty – once tied to national borders and physical territories – by demonstrating how trusted shared arrangements can preserve control and governance over data and digital operations even when hosted abroad. They complement – rather than replace – other sovereign AI infrastructure options, such as national clouds, trusted shared compute arrangements and federated data infrastructures.

However, establishing and scaling such a model is complex. Each economy must assess a wide range of economic, geopolitical, legal, technical and operational risks involved. For example, an economy should assess the financial viability of a digital embassy as it would any other AI infrastructure investment and consider geopolitical risks (e.g. armed conflicts, sanctions) and other crisis scenarios (e.g. cyberattacks, natural disasters) in its overall evaluations.

3.1 Key challenges to overcome to realize digital embassies

The digital embassy landscape is evolving. Newer models extend the concept beyond secure data storage to include arrangements that enable workload storage and compute for non-sensitive data to benefit from home-country legal protections. Notable examples include Saudi Arabia’s draft *Global AI Hub Law*⁴⁵ and Bahrain’s host law approach to offering cloud computing services to foreign parties.⁴⁶

As digital embassies expand in scope and start to involve a broader set of actors, their design

becomes more intricate. Storing, managing and regulating data and digital operations on an extraterritorial basis raises several trust-related challenges:

- **Political and diplomatic feasibility challenges:** Establishing a digital embassy typically requires bilateral or host country authorization, which depends on sustained political will, mutual confidence and diplomatic alignment. With no common reference framework for key considerations

or to anchor expectations, negotiations remain time-intensive and difficult to conclude.

- **Governance and sovereign assurance challenges:** Understandably, governments require high legal and policy certainty to entrust sensitive digital functions to infrastructure outside their borders. Currently, digital embassy arrangements have differing legal bases (e.g. bilateral agreements) and varying interpretations of what constitutes sovereign control. Without shared guidelines around legal protections, immunities and oversight, policy-makers and lawmakers find it difficult to assess reliability or compare options across contexts.

- **Technical and operational maturity challenges:** Policy-makers currently lack clarity on how technical control would be maintained, how data and workloads could be ported or exited, or how ongoing operations would remain aligned with sovereign expectations over time. Without common reference points for secure architectures, interoperability and operational assurance, digital embassies remain difficult to position as a durable component of sovereign infrastructure.

These challenges can be further amplified due to structural imbalances between host and guest countries. Resolving them can unlock several clear benefits.



3.2 Key benefits of trusted setups

When centred around trust, digital embassies offer four distinct advantages:

- 1 **Ensuring continuity when domestic infrastructure is at risk:** Digital embassies enable economies to access secure, sovereign-governed digital infrastructure when local facilities face geopolitical, cybersecurity or natural disaster risks (e.g. Estonia's treaty-based premises in Luxembourg).
- 2 **Expanding sovereign capacity when domestic infrastructure is insufficient:** Digital embassies offer a practical way to secure immediate sovereign capacity abroad – temporarily or long-term – when scaling locally is slow or cost-prohibitive (e.g. due to limited land or energy). They can help bridge the growing global AI divide by unlocking new economic opportunities, especially for emerging economies facing rising digital demand.
- 3 **Providing an alternative path to retaining control over critical data and services:** While many economies rely on data localization requirements to safeguard sovereignty, digital embassies offer another route: data and operations hosted abroad remain under home-country law to maintain control and accountability. For example, Bahrain allows data hosted in its cloud computing centres to remain under Swiss law.⁴⁷
- 4 **Enabling more efficient and demand-aligned infrastructure planning:** Digital embassies allow governments to access sovereign-governed capacity without committing to large, upfront investments. By scaling capacity in line with actual demand, they help reduce stranded assets, improve capital efficiency and contribute to a more stable global infrastructure financing ecosystem – an increasingly important consideration given shared dependencies.

3.3 A global framework for innovative and trusted digital embassies

Establishing a digital embassy is inherently complex. It requires political will, legal clarity and a deep understanding of the regulatory, technical and operational challenges associated with extending sovereign digital capacity beyond national borders. While there can be no one-size-fits-all model, a global framework can provide a baseline for establishing and operationalizing digital embassies upon which economies can build. It can outline the main models of digital embassies and their legal and operational implications, as well as define the minimum safeguards and design choices needed to ensure trust, continuity and interoperability across jurisdictions.

As an impartial and international multistakeholder platform, the Forum convened a global community to co-design such a practical framework for innovative and trusted digital embassies. The framework, launched in this paper, intends to build confidence in shared data storage and compute options. It is designed to support rather than replace robust bilateral agreements by providing a principled foundation for negotiations. Moreover, the framework helps create a more level playing field among economies interested in establishing digital embassies by outlining considerations that are mutually beneficial for all.

The intent to draft this framework was first announced at the Forum's Annual Meeting in Davos in 2026.⁴⁸

Guiding principles

The framework was designed around the following principles:

- Legal robustness, including clarity on access rights, data disclosure, jurisdiction, privacy laws and dispute resolution
- Interoperability
- High-assurance operations and safety evaluation
- Resilience by design, including permission to use end-to-end encryption without intermediary keys
- Exit portability (i.e. ensuring data and workloads can move without lock-in)

Key considerations

Digital embassies succeed only if they are sovereign – both in design and practice. This requires building and sustaining long-term trust – between nations, and between nations and commercial partners. Figure 10 outlines five key dimensions along which trust must be established.

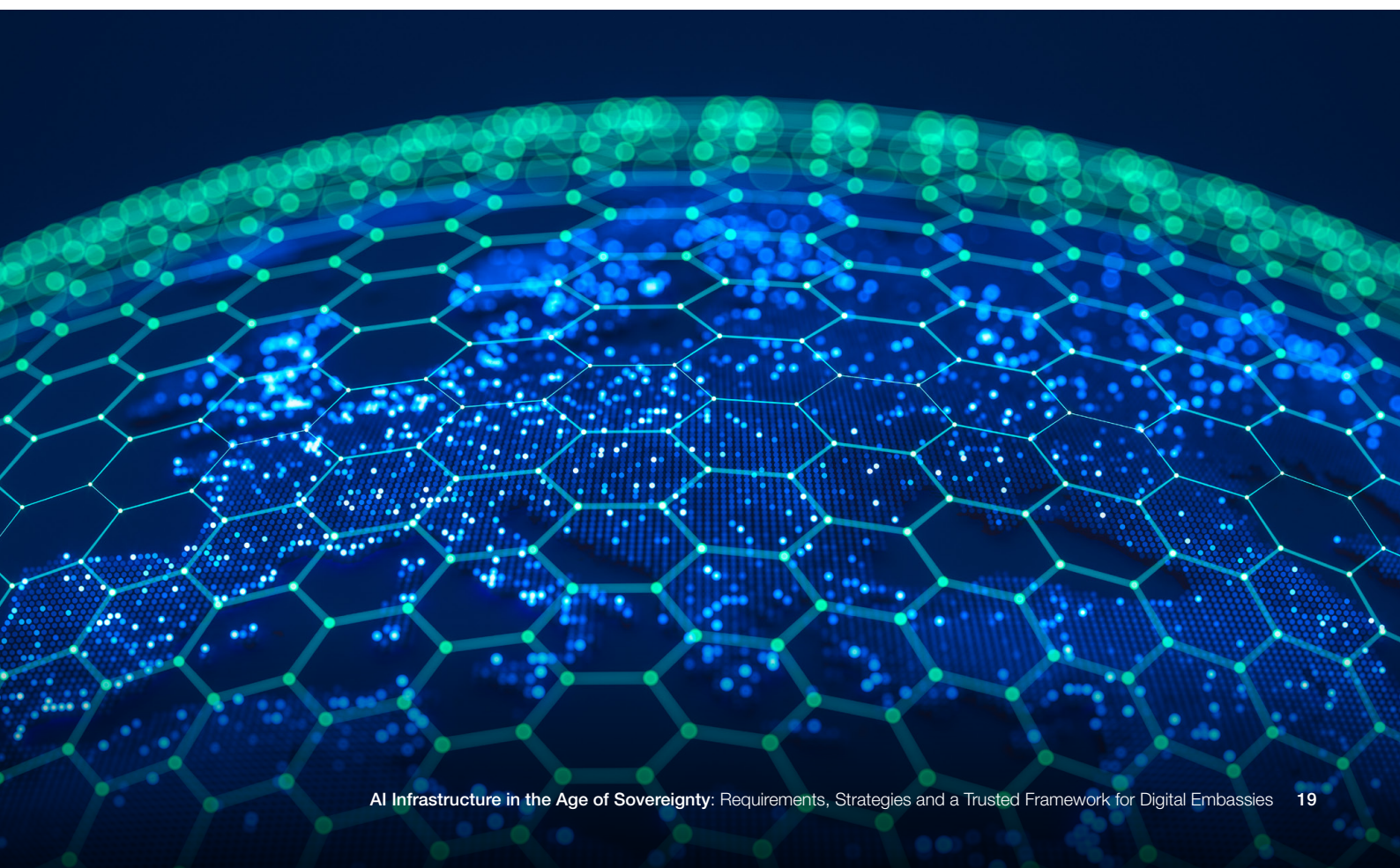


FIGURE 10 | For digital embassies to succeed, countries must build and maintain trust across five key dimensions



1 Political commitment

As governments make AI a national priority, political commitment becomes essential to the credibility and longevity of digital embassy arrangements. National leadership must develop the capabilities needed to negotiate, implement and steward these agreements while ensuring whole-of-government alignment so decisions outlast political cycles and inspire confidence among partners and investors. One way to signal this commitment is through multi-year initiatives that anchor long-term digital and AI infrastructure strategies. Success also depends on identifying the right partners. Governments may begin piloting arrangements with trusted political allies.

Watch-outs: Insufficient or inconsistent political support – whether due to shifting priorities, limited resourcing or fragmented coordination – can undermine long-term confidence in digital embassy arrangements. Additionally, failing to identify the right partner upfront may lead to complications during negotiations.

2 Legal basis and scope of agreement

Establishing a digital embassy should begin with a formal agreement that defines how a nation's data and digital operations will be governed extraterritorially while hosted abroad. Depending on national priorities and data sensitivity, countries may pursue this in different ways.


A **bilateral, treaty-based sovereign premises** – such as the agreement between Estonia and Luxembourg – offers strong inviolability and diplomatic-style protections. Alternatively, a **host-statute designation**, where national law enables foreign entities to operate under home-country legal protections within the host’s territory, can be implemented more rapidly while still providing a clear legal basis for maintaining sovereignty. In Bahrain, for instance, the government enacted a domestic law permitting foreign jurisdictions to apply their own laws to designated cloud environments, subject to case-by-case approvals from individual providers.

Saudi Arabia’s draft *Global AI Hub Law*⁴⁹ introduces a further evolution of this model. It outlines three types of legally defined hubs for hosting and processing foreign data and services:

1. **Private hubs** are dedicated environments operating exclusively under the guest country’s laws.
2. **Extended hubs** are areas where an operator or its users may host workloads under guest-country law.
3. **Virtual hubs** are shared environments where foreign workloads operate under foreign-state legal authority.

Each hub type has tailored legal and operational requirements, including the need for bilateral agreements with guest governments or regulatory approvals, depending on the model.

In any model, clarity on immunities, dispute resolution, access rights, data disclosures, privacy laws and commercial use are essential to ensure the digital embassy operates as a trusted extension of sovereign authority.


 **Watch-outs:** Weak or incomplete laws can create loopholes in extraterritorial hosting arrangements that can be exploited and harm national reputations.

3 Data management

Clear rules defining how data is classified, protected and accessed can enable governments to retain meaningful control while supporting safe and responsible shared operations. Select examples include:

- **Data classification and handling:** Governments should establish clear classification policies defining which categories of data may be hosted in a digital embassy environment and under what conditions. These policies should outline handling requirements across sensitivity levels to ensure protections are proportionate and consistently applied.

- **Data residency and jurisdiction:** Agreements should specify the residency expectations for different data types and clarify how home-country law governs data stored or processed within the digital embassy. This provides assurance that data remains subject to national legal protections, even when hosted extraterritorially.
- **Data access and oversight:** Clear role-based permissions, strong authentication and least-privilege principles should ensure that only authorized officials can view or modify sensitive datasets. Governing arrangements should strictly define – and log – any access and data disclosure requests by the host state or third parties. Institutionalized, independent oversight (e.g. through a joint commission) should review telemetry, audit results and material changes on a regular cadence.

 **Watch-outs:** Unclear rules on classification, residency, access or cross-border data movement can introduce uncertainty and erode trust. Establishing transparent, comprehensive data management principles is critical to enabling the secure, sovereign use of digital embassies.

4 Technical policies and safeguards

Architecture is where sovereignty is operationalized. Technical safeguards should ensure that sensitive workloads are isolated, interoperability is preserved and economies retain long-term strategic choice in how and where their digital operations run. Select examples include:

- **Isolation and segmentation:** Technical architects should implement strong physical and logical separation of sovereign workloads, reinforced through domain segmentation and confidential computing for sensitive data and functions. Data must remain secure and protected across the entire environment.
- **Interoperability and exit:** Procuring authorities should require open profiles and APIs, ensure portability of data and priority workloads where feasible and maintain documented migration or exit playbooks so transitions across environments and vendors can be executed safely and predictably.
- **Cybersecurity foundations:** Core technical safeguards should include robust encryption standards, secure connectivity pathways and designs that minimize attack surfaces while enabling economies to maintain control over sensitive digital functions.



“ Success will depend on clear legal foundations, strong technical safeguards and careful consideration of geopolitical and operational realities.

👁️ **Watch-outs:** Over-reliance on a single provider or proprietary stack, or excessively bespoke integrations, can limit long-term flexibility and introduce unintended dependencies. In addition, geopolitical complexities or potential diplomatic tensions can affect the continuity of access across borders, making credible portability and alternative pathways essential to maintaining sovereign control.

5 Operational rules

Credibility is maintained through consistent, transparent and well-governed operations. Operational rules define how incidents are handled and how performance is verified over time. These rules should allow governments to demonstrate that a digital embassy functions as a trusted extension of sovereign infrastructure. Select examples include:

- **Preparedness and continuity:** Operators should maintain tested playbooks for incident response and escalation, ensuring readiness for energy, network, cyber or supply chain disruptions. Operational teams should periodically conduct exercises – such as failover or portability drills – to validate continuity plans and build shared confidence among stakeholders.
- **Assurance and safety:** Organizations should commission independent audits, certifications and appropriate AI safety evaluations aligned with workload sensitivity. Summary findings, along with trend metrics such as uptime, incident resolution

times and workload portability, should be transparently reported to reinforce accountability.

👁️ **Watch-outs:** Weak observability can defeat strong paper protections. Continuous monitoring should be treated as a baseline practice to prevent this. Operational resilience must also account for geopolitical or diplomatic shifts that could influence coordination or escalation pathways – making transparent processes and well-rehearsed responses essential. Clear, time-bound protections may also be established to ensure host country demand surges don’t constrain access.

Digital embassies offer a compelling option for economies seeking to extend sovereign AI infrastructure beyond their borders while preserving continuity, control and trust. As the landscape evolves to include a broader range of models and actors, success will depend on clear legal foundations, strong technical safeguards and careful consideration of geopolitical and operational realities.

These complexities underscore the need for a baseline of shared principles to help governments and partners navigate extraterritorial data governance with confidence. As an impartial, global, multistakeholder platform, the Forum has developed a practical framework to reduce uncertainty, strengthen trust and mitigate the amplification of structural imbalances across economies while scaling digital embassy models. In doing so, the framework supports the establishment and operation of resilient, future-ready digital embassies.

Conclusion:

Key considerations for AI ecosystem actors

AI infrastructure choices are increasingly shaping long-term competitiveness and resilience, making AI infrastructure strategies a national priority. Policy-makers should align closely with key ecosystem actors – including investors, energy and connectivity providers, and AI infrastructure developers and operators – to deliver on this agenda.

Policy-makers, investors and other ecosystem actors must jointly address two broad considerations, outlined below. Both are grounded in the AI infrastructure building blocks and the non-negotiable prerequisites that constrain them.

1 Strategic positioning and system design

Key partners for policy-makers and investors: AI infrastructure developers

- **Define the strategic direction and segment workloads accordingly:** The AI sovereignty spectrum highlights two reference strategies – trusted international partnerships and extensive domestic ownership – but most economies will adopt hybrid strategies. Policy-makers should determine which workloads and datasets require the highest level of assurance and continuity (and should therefore be managed domestically) and which can safely rely on external capacity under enforceable legal and technical safeguards.
- **Plan for shifting demand and architecture to avoid stranded assets:** Governments should continuously reevaluate how demand is evolving (e.g. training versus inference, public sector versus commercial workloads and priority sectors) and how architectural configurations are shifting (e.g. centralized versus distributed, or cloud versus edge). These factors should be embedded into modular, phased buildouts. Such designs should have explicit upgrade paths that promote future readiness, mitigate obsolescence risk and ensure capacity scales with actual demand rather than past assumptions.

2 Strategic safeguards and resilience mechanisms

Key partners for policy-makers and investors: Energy and connectivity providers and AI infrastructure operators

- **Treat energy, water and land as first-order design constraints:** Compute and storage strategies are only credible if integrated with national resource planning. This requires early coordination across permitting bodies, utilities and local authorities; clear efficiency and sustainability standards; and realistic timelines for grid and connectivity upgrades. When these constraints are binding, economies can use partnerships, site selection and architectural decisions to expand feasible options without compromising security or sustainability objectives.
- **Operationalize trust and strategic flexibility in partnerships – especially across borders:** Trusted partnerships, including digital embassies, can accelerate access to sovereign-governed capacity when domestic buildout is slow or constrained. Trust requires robust legal frameworks covering scope, jurisdiction, access rights and dispute resolution; technical safeguards such as isolation, encryption and secure connectivity; operational assurance through audits, monitoring and independent oversight; and credible interoperability and exit portability to avoid new lock-in risks.
- **Engineer resilience across cybersecurity, operations and finance:** Cybersecurity requirements should scale with the potential attack surface created by architectural choices and include clear accountability mechanisms across government, operators and vendors. Continuity planning should be tested through failover exercises and portability drills. Policy-makers should also work with investors and stakeholders to strengthen bankability, align financing structures with technology refresh cycles and reduce the risk of delays or stranded assets.

As economies move forward, they must prioritize disciplined execution. AI infrastructure strategies must translate into investible roadmaps that align buildouts, resource planning and risk controls. Ultimately, success will depend on whether these strategies remain future-ready and resilient in practice.

Appendix

Definitions

Key constraints:

- **Energy:** Amount of continuous electricity supply required to support expected load operations
- **Water:** Amount of continuous water supply required for cooling and heat rejection
- **Land:** Availability of suitable site(s) with physical space and enabling conditions required (e.g. zoning, rights-of-way, grid connectivity)
- **Hardware:** Availability of specialized hardware (e.g. advanced accelerators, radio units, high-performance storage components) that may be supply-constrained (e.g. concentrated among a small number of foreign suppliers)
- **Cybersecurity:** Scope of system-level exposure that must be prevented or contained, reflecting both the attack surface (i.e. number of potential entry points) and the potential blast radius (i.e. scope of impact once compromised)

Constraint levels

- **Low:** Unlikely to constrain deployment at relevant scale
- **Moderate:** May constrain deployment at relevant scale depending on conditions or context
- **High:** Likely to constrain deployment at relevant scale in most conditions or contexts

Assessments

The following assessments provide directional comparisons across architectural configurations within each building block. The assessments are based on consolidated research:

Connectivity

Energy:

- **Physical networks (low):** Fibre-based fixed networks generally use less electricity to carry the same amount of data compared to other technologies,⁵⁰ resulting in lower energy demand at scale.
- **Access networks (high):** Last-mile and wireless networks typically consume more electricity, with radio access network driving most use.⁵¹

Water:

- **Physical networks (low):** Backbone fibre requires minimal water; cooling is primarily handled through air conditioning at equipment sites.⁵²
- **Access networks (low):** Base-station cooling primarily relies on air conditioning.⁵³

Land:

- **Physical networks (moderate):** No single-site dependencies, but rights-of-way permits and duct access across routes are typically required for rollout.⁵⁴
- **Access networks (moderate):** No single-site dependencies, but many suitable tower or small-cell sites are typically required for network densification (e.g. with appropriate zoning).⁵⁵

Hardware:

- **Physical networks (moderate):** Fibre backbones are typically standardized,⁵⁶ commoditized⁵⁷ inputs. However, hyperscale AI networks may depend on supply-constrained, ultra-high-capacity optical modules.⁵⁸
- **Access networks (moderate):** Generally not scarce, but large-scale deployment of radio and baseband hardware may be affected by vendor concentration.⁵⁹

Cybersecurity:

- **Physical networks (moderate):** System-wide outages are possible, but traffic rerouting can reduce risk when routes are diverse.⁶⁰
- **Access networks (high):** Software-intensive networks have numerous remote access paths; major flaws or inadequate patching can create widespread impact.⁶¹

Compute

Energy:

- **Centralized compute (high):** Large per-site electricity demand for continuous operations.⁶²
- **Distributed compute (moderate):** Smaller per-site electricity demand, although small-site inefficiencies and redundancy can raise overall use.⁶³
- **Endpoint compute (low):** Device-level electricity draw, resulting in moderate system-level electricity use across multiple devices – often constrained by device battery⁶⁴ rather than continuous grid supply.

Water:

- **Centralized compute (high):** Large per-site water demand when sites rely on water-based heat rejection (e.g. cooling towers).⁶⁵
- **Distributed compute (moderate):** Smaller per-site water demand, though less efficient thermal designs can raise overall cooling needs.⁶⁶
- **Endpoint compute (low):** Constrained by device thermal limits;⁶⁷ no facility cooling water requirements.

Land:

- **Centralized compute (high):** Requires a large, suitable, permitted site with access to grid power and fibre.⁶⁸
- **Distributed compute (moderate):** Many suitable, permitted locations required for low-latency applications.⁶⁹

- **Endpoint compute (low):** No dedicated site required; compute is embedded in devices.⁷⁰

Hardware:

- **Centralized compute (high):** Requires supply-constrained advanced accelerators at scale (e.g. GPUs).⁷¹
- **Distributed compute (moderate):** Often uses advanced accelerators (although in smaller quantities per site), along with more off-the-shelf components.⁷²
- **Endpoint compute (low):** On-device accelerators (e.g. neural processing units) are increasingly common and produced at consumer scale.^{73,74}

Cybersecurity:

- **Centralized compute (high):** Central platforms concentrate workloads and traffic within data centres, potentially enabling lateral movement⁷⁵ and increasing blast radius.
- **Distributed compute (high):** Distributed architectures expand the attack surface.⁷⁶
- **Endpoint compute (high):** Weak device security⁷⁷ and a large number of endpoints can facilitate the spread of attacks.⁷⁸

Data storage

Energy, water, land and cybersecurity constraints for centralized and distributed storage are similar to those assessed for centralized and distributed compute, respectively.

Hardware:

- **Centralized storage (moderate):** Often requires low-latency storage⁷⁹ (e.g. flash), but components are generally not scarce.⁸⁰
- **Distributed storage (low):** Typically runs on commodity hardware⁸¹ and requires smaller quantities per site.

Contributors

Lead authors

Samira Gazzane

Policy Lead, Future-Ready Economies,
World Economic Forum

Sita Gleichauf

Fellow, World Economic Forum;
Senior Manager, Bain & Company

Mariana Justo Pereira

Fellow, World Economic Forum;
Consultant, Bain & Company

Shreya Sahay

Fellow, World Economic Forum;
Consultant, Bain & Company

Philipp Sautner

Partner and Head, AI,
Insights & Solutions for Germany,
Bain & Company

Francesca Zanolla

Global Lead, AI Strategic Integration,
World Economic Forum

World Economic Forum

Maria Basso

Head, AI Applications and Impact

Teysir Bedretdin

Specialist, AI Governance
and International Collaboration

Agustina Callegari

Initiatives Lead, Technology Governance,
Safety and International Cooperation

Tarik Fayad

Middle East and North Africa
Lead, AI Strategic Integration

Ariella Inglese

Community and Events Specialist,
AI Governance and International Collaboration

Karla Yee Amezaga

Initiatives Lead, AI and Data Governance

Bain & Company

Aron Philipp

Senior Manager

Acknowledgements

Basma AlBuhairan

Managing Director, Centre for the
Fourth Industrial Revolution Saudi Arabia

Anuraag Bahl

Vice-President, Product, Palantir Technologies

Thomas Bohné

Founder and Head, Cyber-Human Lab,
University of Cambridge

Erik Brynjolfsson

Director, Digital Economy Lab,
Stanford University

Simon Chesterman

Senior Director, AI Governance,
AI Singapore, National University of Singapore

Stephanie Cohen

Chief Strategy Officer, Cloudflare

Ali Dalloul

Group Chief Strategy Officer, G42

Naima Al Falasi

Senior Vice-President, AI Strategy and
Transformation, Mubadala Investment Company

Rebecca Finlay

Chief Executive Officer, Partnership on AI

Olaf J. Groth

Professional Faculty, UC Berkeley
Haas School of Business

Hiroki Habuka

Research Professor, Graduate School
of Law, Kyoto University

Peter Hallinan

Director, Responsible AI,
Amazon Web Services

Ian Hodgkinson

Professor of Strategy, Loughborough University

Carl Holshouser

Vice-President, Global Government Affairs, CoreWeave

Hu Guodong

Researcher, CCID, China

Tom Jackson

Professor of Information and Knowledge Management, Loughborough University

Nathan Jokel

Senior Vice-President, Corporate Strategy and Alliances, Cisco Systems

Amit Joshi

Professor of AI, Analytics and Marketing Strategy, International Institute for Management Development (IMD) Business School

Sean Kask

Chief AI Strategy Officer, SAP

Ann Marie Lavigne

Vice-President, Strategic Initiatives, Snowflake

Harrison Lung

Group Chief Strategy Officer, e&

Pallavi Mahajan

Chief Technology and AI Officer, Nokia

Derek Manky

Chief Security Strategist and Global Vice-President, Threat Intelligence, Fortinet

Chiara Marcati

Chief AI Advisory and Business Officer, AI 71

Adrian Marcellus

Chief Executive Officer, MYCentre4IR

James O'Day

Managing Director and Head of Innovation, CVC Capital Partners

Farheen Rahimtoola

Executive Director, JP Morgan Chase

Francesca Rossi

IBM Fellow and Global Leader for Responsible AI and AI Governance, IBM

Crystal Rugege

Managing Director, Centre for the Fourth Industrial Revolution Rwanda

Jim Ryan

Senior Vice-President and Chief Strategy Officer, Liberty Global

Anne-Lise Thieblemont

Vice-President, Government Affairs, Qualcomm

Eser Tireli

Managing Director, Data Science and AI, CVC Capital Partners

Dustin Todd

Vice-President and Head of Government Affairs, Synopsys

Bhushan Trivedi

Assistant Vice-President, Indian Investment Promotion Agency, India

Andrew Wells

Chief Data and AI Officer, NTT Data North America

Thomas Wolf

Co-Founder and Chief Science Officer, Hugging Face

Hala Zeine

Chief Strategy Officer, ServiceNow

Kai Zenner

Head of Office and Digital Policy Adviser for Member of the European Parliament Axel Voss, European Parliament

Production**Louis Chaplin**

Editor, Studio Miko

Laurence Denmark

Creative Director, Studio Miko

Jay Kelly

Designer, Studio Miko

Endnotes

1. World Economic Forum. (2026). *Rethinking AI Sovereignty: Pathways to Competitiveness through Strategic Investments*. <https://www.weforum.org/publications/rethinking-ai-sovereignty/>.
2. Organisation for Economic Co-operation and Development (OECD). (2025). *Government at a Glance 2025: Digital public infrastructure*. https://www.oecd.org/en/publications/government-at-a-glance-2025_0efd0bcd-en/full-report/digital-public-infrastructure_1cee4220.html.
3. For the purposes of this paper, AI infrastructure refers to the compute, connectivity, and data storage systems that enable the development and deployment of AI models and applications. This definition is detailed in the first chapter of this paper.
4. Global Data Center Hub. (2025). *Where Sovereignty Meets Speed: The Rise of Sovereign Clouds and Edge Data Centers*. <https://www.globaldatacenterhub.com/p/where-sovereignty-meets-speedthe>.
5. Bowen, E. (2024). *Distributed Storage Explained: Revolutionizing Data*. Telnix. <https://telnix.com/resources/what-is-distributed-storage>.
6. Smart, L. & Hsu, S. (2025). *The AI-energy nexus will determine AI's impact. We must account for it better*. World Economic Forum. <https://www.weforum.org/stories/2025/12/ai-energy-nexus-ai-future/>.
7. Ben-Shlomi, R. (2025). *How photonic computing can move from promise to commercialization*. World Economic Forum. <https://www.weforum.org/stories/2025/08/photonic-computing-promise-commercialization/>.
8. Bourne, J. (2025). *Photonics powering AI data centres: The latest innovations*. Microelectronics UK. <https://microelectronicsuk.com/blog1/photonics-powering-ai-data-centres-latest-innovations>.
9. DDN. (2025). *Why HPC Is Your Path to AI*. <https://www.ddn.com/resources/whitepapers/why-hpc-is-your-path-to-ai/>.
10. European High-Performance Computing Joint Undertaking. (2025). *Contract Signed for Alice Recoque, Europe's New Exascale Supercomputer*. https://www.eurohpc-ju.europa.eu/contract-signed-alice-recoque-europes-new-exascale-supercomputer-2025-11-18_en.
11. Ben-Shlomi, R. (2025). *How photonic computing can move from promise to commercialization*. World Economic Forum. <https://www.weforum.org/stories/2025/08/photonic-computing-promise-commercialization/>.
12. Caballar, R. & Stryker, C. (n.d.). *What is federated learning?* IBM. <https://www.ibm.com/think/topics/federated-learning>.
13. Langkamp, K. (2025). *Deutsche Telekom strengthens Europe's connectivity: Participation in the IRIS² satellite project*. Deutsche Telekom. <https://www.telekom.com/en/media/media-information/archive/telekom-backs-eu-satellite-project-iris2-1093312>.
14. European Commission. (n.d.). *European Quantum Communication Infrastructure - EuroQCI*. <https://digital-strategy.ec.europa.eu/en/policies/european-quantum-communication-infrastructure-euroqci>.
15. Global Partnership on Artificial Intelligence (GPAI). (2024). *AI for Net Zero: Assessing Readiness for AI*. https://wp.oecd.ai/app/uploads/2025/05/AI-for-net-zero_assessing-readiness-for-AI.pdf.
16. Benioff, M. (2016). *The AI revolution is coming fast. But without a revolution in trust, it will fail*. World Economic Forum. <https://www.weforum.org/stories/2016/08/the-digital-revolution-is-here-but-without-a-revolution-in-trust-it-will-fail/>.
17. International Energy Agency (IEA). (n.d.). *Energy demand from AI*. <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>.
18. Gorey, J. (2025). *Data Drain: The Land and Water Impacts of the AI Boom*. Lincoln Institute of Land Policy. <https://www.lincolninst.edu/publications/land-lines-magazine/articles/land-water-impacts-data-centers/>.
19. Bowen, M. (2025). *Land and location: Considerations in an explosive data centre market*. Intelligent Data Centres. <https://www.intelligentdatacentres.com/2025/02/07/land-and-location-considerations-in-an-explosive-data-centre-market/>.
20. UK Department for Science, Innovation and Technology. (2025). *AI Growth Zones*. <https://www.gov.uk/government/collections/ai-growth-zones>.
21. Hajdari, U. (2025). *Nvidia shares slip as AI accelerator race shifts interest to Google chips*. Euronews. <https://www.euronews.com/business/2025/11/25/nvidia-shares-slip-as-ai-accelerator-race-shifts-interest-to-google-chips>.
22. Timings, J. (2022). *Busting ASML myths*. ASML. <https://www.asml.com/en/news/stories/2022/busting-asml-myths>.
23. The Economist. (2025). *The world's biggest chipmaker needs to move beyond Taiwan*. <https://www.economist.com/briefing/2025/08/21/the-worlds-biggest-chipmaker-needs-to-move-beyond-taiwan>.
24. TrendForce. (2025). *EU & Japan Grant Major Semiconductor Subsidies: Infineon, TSMC in Focus*. <https://www.trendforce.com/news/2025/02/25/news-eu-japan-grant-major-semiconductor-subsidies-infineon-tsmc-in-focus/>.
25. Parvini, S. (2025). *What changes to the CHIPS act could mean for AI growth and consumers*. Associated Press. <https://apnews.com/article/trump-semiconductors-chips-act-3592f1ed8b8cd4f2145cfa8a4985046c>.
26. Salve, P. (2025). *India is betting \$18 billion to build a chip powerhouse. Here's what it means*. CNBC. <https://www.cnbc.com/2025/09/23/india-is-betting-18-billion-to-build-a-chip-powerhouse-heres-what-it-means.html>.

27. Federal Bureau of Investigation (FBI) San Francisco Division. (2024). *FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence*. <https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence>.
28. UK National Cyber Security Centre (NCSC). (2024). *The near-term impact of AI on the cyber threat*. <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>.
29. National Security Agency. (n.d.). *Artificial Intelligence Security Center (AISC)*. <https://www.nsa.gov/AISC/>.
30. UK National Cyber Security Centre (NCSC). (2025). *Impact of AI on cyber threat from now to 2027*. <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>.
31. Reuters. (2025). *Eletrobras partners with C3 AI to modernize Brazil's power grid*. <https://finance.yahoo.com/news/eletrobras-partners-c3-ai-modernize-110351192.html>.
32. UK Department of Health and Social Care. (2025). *World-first AI system to warn of NHS patient safety concerns*. <https://www.gov.uk/government/news/world-first-ai-system-to-warn-of-nhs-patient-safety-concerns>
33. Michaels, D. (2023). *SEC Says It's Using AI to Surveil Markets and Assist Investigations*. Wall Street Journal. <https://www.wsj.com/livecoverage/stock-market-today-dow-jones-09-12-2023/card/sec-says-it-s-using-ai-to-surveil-markets-and-assist-investigations-21lvQMfrrjlgkkS4lyY>.
34. World Economic Forum. (2026). *Rethinking AI Sovereignty: Pathways to Competitiveness through Strategic Investment*. <https://www.weforum.org/publications/rethinking-ai-sovereignty/>.
35. LUMI AI Factory. (n.d.). *About us: European High-Performance Computing Joint Undertaking (EuroHPC JU)*. <https://lumi-ai-factory.eu/about-us/eurohpc-ju/>.
36. e-Estonia. (n.d.). *Data Embassy*. <https://e-estonia.com/solutions/e-governance/data-embassy/>.
37. Mordor Intelligence. (2025). *China Cloud Computing Market Size & Share Analysis - Growth Trends and Forecast (2026 - 2031)*. <https://www.mordorintelligence.com/industry-reports/china-cloud-computing-market>.
38. Global Data Center Hub. (2025). *Where Sovereignty Meets Speed: The Rise of Sovereign Clouds and Edge Data Centers*. <https://www.globaldatacenterhub.com/p/where-sovereignty-meets-speedthe>.
39. Government Technology Agency of Singapore. (2025). *Government on Commercial Cloud (GCC)*. <https://www.tech.gov.sg/products-and-services/for-government-agencies/software-development/government-on-commercial-cloud/>.
40. National Supercomputing Centre (NSCC) Singapore. (n.d.). *Overview*. <https://www.nsc.sg/overview/>.
41. National Supercomputing Centre (NSCC) Singapore. (2024). *New S\$270 Million Grant to Boost National Supercomputing Infrastructure, HPC Capabilities and Talent Development*. <https://www.nsc.sg/wp-content/uploads/2024/10/Embargoed-till-25-Oct-2024-4pm-Media-Release-for-NSCCs-ASPIRE-2A-2A-Launch.pdf>.
42. e-Estonia. (n.d.). *Data Embassy*. <https://e-estonia.com/solutions/e-governance/data-embassy/>.
43. Savouroux, E. (2025). *A world first: Estonia opens a 'data embassy' in Luxembourg*. Blue Europe. <https://www.blue-europe.eu/analysis-en/short-analysis/a-world-first-estonia-opens-a-data-embassy-in-luxembourg/>.
44. Luxembourg Government. (2024). *E-embassies in Luxembourg*. <https://luxembourg.public.lu/en/invest/innovation/e-embassies-in-luxembourg.html>.
45. Saudi National Competitiveness Center (Istitlaa). (n.d.). *Global AI Hub Law*. <https://istitlaa.ncc.gov.sa/en/Transportation/citc/globalailaw/Pages/default.aspx>.
46. Bahrain Economic Development Board (EDB). (n.d.). *Legislative Decree No. 56 of 2018 In Respect of Providing Cloud Computing Services to Foreign Parties*. <https://bahrainbusinesslaws.com/laws/Law-of-Providing-Cloud-Computing-Services-to-Foreign-Parties>.
47. Bahrain News Agency. (2022). *Bahrain grants Switzerland jurisdiction over entities data hosted in cloud computing centers*. <https://www.bna.bh/en/BahraingrantsSwitzerlandjurisdictionoverentitiesdatahostedincloudcomputingcenters.aspx?cms=q8FmFJgiscL2fwlzON1%2BDuAcP8AwUPbSHF3A4DJwnDc%3D>.
48. World Economic Forum. (2026). *Digital Embassies for Sovereign AI*. <https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2026/sessions/digital-embassies-for-sovereign-ai/>.
49. Saudi National Competitiveness Center (Istitlaa). (n.d.). *Global AI Hub Law*. <https://istitlaa.ncc.gov.sa/en/Transportation/citc/globalailaw/Pages/default.aspx>.
50. European Commission, Broadband Competence Offices (BCO) Network Support Facility. (2020). *Fibre is the most energy efficient broadband technology*. <https://digital-strategy.ec.europa.eu/en/library/fibre-most-energy-efficient-broadband-technology>.
51. Frield, D., Glynn, P., Isidoro, L. & Lobato, J. (2023). *Reducing Energy Consumption in Mobile Networks*. *NEC Technical Journal*, vol. 17, no. 1, pp. 23-28. <https://www.nec.com/en/global/techrep/journal/g23/n01/pdf/230104.pdf>.
52. BeyondTech. (2015). *How will fiber optics save the world?*. <https://beyondtech.us/blogs/beyond-blog/how-fiber-optic-save-the-world>.
53. Tark Thermal Solutions. (2025). *Cooling for Mobile Base Stations and Cell Towers*. https://www.mouser.com/pdfDocs/TTS_Application-Note_CoolingforMobileBaseStations1.pdf.
54. Organisation for Economic Co-operation and Development (OECD). (2008). *Public Rights of Way for Fibre Deployment to the Home*. https://www.oecd.org/en/publications/public-rights-of-way-for-fibre-deployment-to-the-home_230502835656.html.

55. Walker, W. (2024). *The future of real estate is digital: How data centers and 5G are shaping the next generation of infrastructure*. Walker & Dunlop. <https://www.walkeranddunlop.com/insights/future-real-estate-digital>.
56. International Telecommunication Union (ITU). (2024). *Characteristics of a single-mode optical fibre and cable [Recommendation ITU-T G.652 (08/2024)]*. https://www.itu.int/rec/dologin_pub.asp?id=T-REC-G.652-202408-1%21%21PDF-E&lang=e&type=items.
57. Federal Reserve Bank of St. Louis (FRED). (2025). *Producer Price Index by Commodity: Metals and Metal Products: Fiber Optic Cable*. <https://fred.stlouisfed.org/series/WPU10260333>.
58. TrendForce. (2025). *AI Data Centers Ignite a Laser Shortage Wave; Nvidia's Strategic Lock-In Reshapes the Global Laser Supply Chain, Says TrendForce*. <https://www.trendforce.com/presscenter/news/20251208-12823.html>.
59. Pongratz, S. (2025). *What to Expect from RAN in 2025*. Dell'Oro Group. <https://www.delloro.com/what-to-expect-from-ran-in-2025/>.
60. US Government Accountability Office. (2021). *CISA Should Assess the Effectiveness of its Actions to Support the Communications Sector [Report GAO-22-104462]*. <https://www.gao.gov/assets/gao-22-104462.pdf>.
61. Network and Information Systems (NIS) Cooperation Group. (2019). *EU coordinated risk assessment of the cybersecurity of 5G networks*. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62132.
62. Mukherjee, S. & Raitano, L. (2025). *Explainer-Keeping cool: heat a key challenge for data centers and AI*. Yahoo Tech. <https://tech.yahoo.com/ai/articles/explainer-keeping-cool-heat-key-125831274.html>.
63. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2020). *Edge Computing: Considerations for Reliable Operation*. https://www.ashrae.org/file%20library/technical%20resources/bookstore/tb_edgecomputing_sep2020.pdf.
64. Talluri, R. (2025). *The AI Power Drain: Why Battery Limitations Threaten the Future of Mobile AI*. Enovix. <https://www.enovix.com/the-ai-power-drain-why-battery-limitations-threaten-the-future-of-mobile-ai/>.
65. US Department of Energy. (n.d.). *Cooling Water Efficiency Opportunities for Federal Data Centers*. <https://www.energy.gov/femp/cooling-water-efficiency-opportunities-federal-data-centers>.
66. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2020). *Edge Computing: Considerations for Reliable Operation*. https://www.ashrae.org/file%20library/technical%20resources/bookstore/tb_edgecomputing_sep2020.pdf.
67. Nikfar, N. (2024). *On-device AI and its thermal implications*. Hot Chips Symposium 2024. <https://hc2024.hotchips.org/assets/program/tutorials/HC2024.T2.Qualcomm.NaderNikfar.final-0824.pdf>.
68. Johnston, J. (2026). *2026: Data centers look to rural areas*. Agri-View. https://agupdate.com/agriview/markets/crop/article_a4cc13d4-904a-436e-b892-9a1e196ebbd8.html.
69. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2020). *Edge Computing: Considerations for Reliable Operation*. https://www.ashrae.org/file%20library/technical%20resources/bookstore/tb_edgecomputing_sep2020.pdf.
70. Nikfar, N. (2024). *On-device AI and its thermal implications*. Hot Chips Symposium 2024. <https://hc2024.hotchips.org/assets/program/tutorials/HC2024.T2.Qualcomm.NaderNikfar.final-0824.pdf>.
71. Seetharaman, D. & Dotan, T. (2023). *The AI Boom Runs on Chips, but It Can't Get Enough*. *Wall Street Journal*. <https://www.wsj.com/tech/ai/the-ai-boom-runs-on-chips-but-it-cant-get-enough-9f76f554>.
72. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2020). *Edge Computing: Considerations for Reliable Operation*. https://www.ashrae.org/file%20library/technical%20resources/bookstore/tb_edgecomputing_sep2020.pdf.
73. Schuman, E. (2024). *PCs with NPUs tweaked for AI now account for one of every five PCs shipped, says Canalis*. Computerworld. <https://www.computerworld.com/article/3606904/pcs-with-npus-tweaked-for-ai-now-account-for-one-of-every-five-pcs-shipped-says-canalis.html>.
74. International Data Corporation (IDC). (2024). *The Rise of Gen AI Smartphones*. <https://www.idc.com/resource-center/blog/the-rise-of-gen-ai-smartphones/>.
75. Chandramouli, R. (2022). *Guide to a Secure Enterprise Network Landscape [NIST Special Publication (SP) 800-215]*. National Institute of Standards and Technology (NIST). <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-215.pdf>.
76. Ibid.
77. Cybersecurity and Infrastructure Security Agency (CISA). (2017). *Heightened DDoS Threat Posed by Mirai and Other Botnets*. <https://www.cisa.gov/news-events/alerts/2016/10/14/heightened-ddos-threat-posed-mirai-and-other-botnets>.
78. US Government Accountability Office. (2021). *CISA Should Assess the Effectiveness of its Actions to Support the Communications Sector [Report GAO-22-104462]*. <https://www.gao.gov/assets/gao-22-104462.pdf>.
79. Cayton, P. (n.d.). *Accelerating AI Workloads with NVMe® over Fabrics (NVMe-oF™) Technology*. NVM Express. <https://nvmexpress.org/accelerating-ai-workloads-with-nvme-over-fabrics-nvme-of-technology/>.
80. TrendForce. (2024). *Increased Production and Weakened Demand to Drive NAND Flash Prices Down 3–8% in 4Q24, Says TrendForce*. <https://www.trendforce.com/presscenter/news/20241015-12327.html>.
81. De Vente, P. (2024). *The pros and cons of using distributed storage*. Quest. <https://blog.quest.com/the-pros-and-cons-of-using-distributed-storage/>.



COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org