इलेक्ट्रॉनिकी एवं सूचना प्रौद्योगिकी मंत्रालय
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY
सत्यमेव जयते

AI IMPACT SUMMIT
भारत 2026 INDIA

सर्वजन हिताय | सर्वजन सुखाय
WELFARE FOR ALL | HAPPINESS OF ALL

INDIAai

Playbook for Advancing Resilient AI Infrastructure

Resilience, Innovation and Efficiency Working Group

Dalberg

Playbook for Advancing Resilient AI Infrastructure

Resilience, Innovation and Efficiency Working Group

Table of Contents

Foreword

Artificial intelligence is rapidly becoming an integral part of modern economies, shaping how data is processed, services are delivered, and innovation is pursued across sectors. As countries move from piloting AI applications to deploying them at scale, the resilience and sustainability of the infrastructure underpinning these systems will play a decisive role in determining the long-term trajectory of AI adoption. The India AI Impact Summit 2026 provides an important platform to advance global understanding of how AI can scale in ways that are both high-performing and resource-conscious, in line with the guiding Sutras of People, Planet, and Progress.

The Resilience, Innovation, and Efficiency Working Group, led by India with France as co-chair, has examined the growing interdependence between AI systems and the physical infrastructure that supports them. As demand for compute expands, pressures on energy systems, water resources, and critical infrastructure are becoming more pronounced, underscoring the need for coordinated planning and a systems perspective. Addressing these challenges is essential not only for advancing AI innovation, but also for ensuring that its benefits can be realised without creating new constraints for development.

The Advancing Resilient AI Infrastructure Playbook reflects the insights of participating countries, international organisations, and experts engaged through the Working Group. It highlights how a systems approach spanning the AI stack, energy systems, and policy frameworks can help countries navigate emerging constraints while enabling responsible scale. By bringing together practical solutions, enabling policy considerations, and a Country Reference Framework, the Playbook offers a structured lens to support decision-making across diverse national contexts.

India's experience demonstrates the importance of building strong digital and compute foundations alongside investments in energy-efficient and inclusive innovation. Through initiatives under the IndiaAI Mission, we are working to expand access to compute, develop efficient and indigenous AI models, and promote responsible AI development that supports economic growth while remaining aligned with sustainability considerations. These efforts reflect our broader conviction that resilient infrastructure is central to ensuring that AI can serve as a driver of long-term development.

We express our sincere appreciation to Mr. Pankaj Agarwal, Secretary, Ministry of Power and Chair of the Resilience, Innovation, and Efficiency Working Group, to the Government of France for its partnership as co-chair, to all participating countries and organisations for their valuable contributions, and to Dalberg for its support as knowledge partner in shaping the analytical foundations of this Playbook. Their collaboration has been instrumental in ensuring that this effort reflects diverse perspectives and practical insights.

This Playbook underscores a simple but important principle, namely that the choices made today in how AI systems are designed, powered, and deployed will shape their impact for decades to come. By aligning innovation with resilience and resource efficiency, we can ensure that AI strengthens economies, supports sustainable growth, and delivers meaningful benefits for societies around the world.

Shri S. Krishnan, IAS
Secretary,
Ministry of Electronics and Information Technology
Government of India

The convergence of AI expansion and the energy transition presents one of the defining challenges of our time. Data centres have reached electricity consumption levels comparable to those of entire nations, and the rapid growth of AI workloads is further intensifying this demand.

This need not become a structural constraint; it can instead be a catalyst for innovation.

India's experience in managing one of the world's largest power systems while pursuing ambitious clean energy goals has taught us that constraints, when confronted deliberately, drive smarter solutions. The challenge before us is to ensure that the rapid expansion of AI infrastructure strengthens and accelerates the energy transition, rather than placing additional strain on power systems or reversing hard-won gains.

This playbook demonstrates that such alignment is achievable. Through coordinated action across AI design, infrastructure planning, and grid integration, countries can reduce energy intensity, align compute with clean power availability, and build systems that adapt to local realities. The solutions span efficient model design, demand-responsive operations, renewable energy integration, and waste heat recovery.

Critically, the approaches presented here acknowledge that countries face different starting conditions and will pursue different priorities – some accelerating clean energy investment through AI demand, others optimising within existing constraints, and still others strategically leveraging renewable energy abundance.

What matters is that we act with foresight. The India AI Impact Summit provides a platform to advance international cooperation, share knowledge, and pilot solutions that ensure AI infrastructure becomes an asset to energy systems, not a liability.

The choices we make today will echo for decades. Let us choose resilience, efficiency, and responsibility.

Shri Pankaj Agarwal, IAS
Secretary,
Ministry of Power
Government of India

Executive Summary

Artificial Intelligence (AI) is a powerful lever for countries to drive economic development, inclusion, and public value. AI is being adopted at an unprecedented speed across countries, reshaping how services are delivered and how innovation is pursued.

AI also holds significant potential to accelerate climate action across sectors. Emerging applications are already supporting improved forecasting of renewable energy supply, more efficient management of energy and industrial systems, enhanced monitoring of emissions and land-use change, and faster climate risk assessment and disaster response. As countries build the foundations for responsible and resilient AI, there is a parallel opportunity to harness AI as a tool for mitigation, adaptation, and more effective delivery of climate and development priorities.

Demand for energy, water, and land increase as adoption scales. This puts growing pressure on local systems and infrastructure. In advanced economies, data centres are often clustered in a small number of locations, intensifying competition for local electricity and water resources. In developing economies, AI adoption is unfolding alongside efforts to expand on developmental priorities, such as reliable electricity for health, education, and public services, bringing new considerations into how limited resources are planned and allocated. Together, these dynamics present an opportunity for countries to explore how AI can be developed and deployed in ways that are resilient and resource-conscious, allowing its benefits to scale while remaining aligned with broader development objectives. This Playbook examines promising innovations across the AI value chain and energy systems and offers a suite of strategic pathways and innovation bundles for countries to explore given their operating context and needs.

Today, two interlinked constraints limit resilient AI growth: 1) limited availability of clean and reliable energy, and 2) mismatch between data centre siting and local resource availability. In the near term, pathways for meeting incremental electricity demand from AI-driven data centres will differ across countries, and it will be important to avoid long-term carbon lock-in through early clean energy and grid planning. The intermittent nature of clean energy such as solar and wind are insufficient to meet the required reliability and uptime of data centres. Moreover, the pace of data centre expansion is faster than the development of reliable clean energy infrastructure. At the same time, siting decisions for new data centres typically prioritise connectivity, network reliability, and access to end users, resulting in data centres concentrated in locations where water resources are already constrained. Together, these two constraints underscore the need for a systems approach to building resilient AI – one that requires coordinated action across the AI stack, energy systems, and land and water resources.

But this moment of constraint is also one of opportunity: a systems approach anchored in building resilient AI can help address these constraints and set countries up for long-term success in leveraging AI to meet national development goals. Coordinated choices across the AI stack (AI use cases, model design, and data centre siting and operations), energy systems, and grid integration can lead to meaningful reductions in energy use, water consumption, and emissions. Individually, these levers can deliver substantial gains, but the largest benefits come when solutions reinforce and build on each other. To complement these innovations, enabling policies and standards play a critical role in guiding transparency, performance guidance, and long-term planning.

Countries are likely to follow different paths in building resilient AI, shaped by their national priorities, contexts, AI ambitions and the resources available to support it. Variances in access to clean and reliable energy, alongside the availability of other enabling resources such as investment capital, land, water systems, and institutional capacity, influence how resilient AI infrastructure can scale in practice. Recognising this diversity, the Playbook outlines a Country Reference Framework that provides indicative guidance on how governments and stakeholders can think through their strategic pathways to resilient and resource-conscious AI. The framework highlights a set of illustrative country archetypes:

- AI-Energy Accelerators, that can align growth of AI infrastructure with accelerated investment in clean, firm power and grid readiness so that AI demand supports energy systems development;
- AI Scalers, that can focus on improving efficiency across the AI stack to scale domestic AI capabilities within existing infrastructure constraints;
- Clean Power Players, that can leverage low-carbon power to host AI infrastructure and provide clean compute capacity to wider markets; and
- AI Foundation Builders, that can capture AI's productivity and service-delivery benefits through efficient, fit-for-purpose models and services, while progressively investing in the renewable energy and digital infrastructure needed to support domestic AI capacity over time.

Ultimately, the choices made today will shape AI's trajectory for decades. AI's growth trajectory and associated economic and social impact will be shaped by the deliberate design and deployment choices made today. Building resilient AI is about aligning it with energy readiness, infrastructure capacity, and long-term development priorities. As countries pursue different pathways based on their ambitions and starting conditions, there is a shared opportunity to ensure that AI strengthens essential systems, supports inclusive economic outcomes, and manages pressures on energy, water, and land. With coordinated, systems-level action and continued learning, resilient AI can help unlock transformative outcomes for people, planet, and progress.

The insights and illustrative examples in this Playbook are informed by inputs received through consultations with countries and international organisation, including responses submitted via a Request for Information (RFI) process.

Background

The **India AI Impact Summit 2026** marks a pivotal shift in the global AI discourse, from ambition to impact. Guided by the Sutras of People, Planet, and Progress, and structured through seven thematic priorities, the Summit convenes governments, international organisations, industry, academia, and civil society to advance practical, development-oriented outcomes that ensure AI delivers inclusive, sustainable, and measurable impact.

Within this framework, the **Resilience, Innovation & Efficiency Working Group** was constituted to address a critical frontier question: how can AI systems scale in ways that are high-performing yet resource-efficient, economically viable, and aligned with real-world development constraints? The Working Group has focused on advancing intelligent systems that are resilient by design, capable of operating under energy, water, infrastructure, and institutional constraints , while remaining accessible, scalable, and inclusive.

Through structured deliberations, including expert consultations, the Working Group surfaced a convergence of perspectives: AI's rapid expansion is placing growing demands on resources, including energy systems, water resources, and critical infrastructure. In response, the Working Group coalesced around the need for a systems approach, one that aligns AI growth with resource efficiency, infrastructure planning, and long-term development priorities. The **Advancing Resilient AI Infrastructure Playbook** is a **knowledge output** of this process. It reflects the collective insights, practical experiences, and policy considerations surfaced through the Working Group's engagement and written inputs.

Rather than prescribing a one-size-fits-all model, the Playbook offers a coordinated, resource-conscious framework to guide countries in scaling AI responsibly. It provides a structured lens across the AI stack, energy systems, and policy architecture, alongside a Country Reference Framework that recognises diverse starting conditions and strategic pathways.

As one of the outcome of the Resilience, Innovation & Efficiency Working Group under the India AI Impact Summit 2026, this Playbook is intended to serve as a practical guide for policymakers, industry, and system planners and in doing so, it advances the Summit's broader commitment of delivering tangible progress for people, planet, and future generations.

Introduction

Artificial Intelligence (AI) represents a pivotal opportunity to drive economic growth, productivity, and innovation across sectors. When deployed effectively, AI has the potential to expand access to services, strengthen enterprises, and support more inclusive development pathways. Across sectors, AI is already being embedded into essential functions such as energy management, logistics and supply chains, financial services, healthcare delivery, and public administration, improving decision-making, lowering transaction costs, and enabling systems to operate more efficiently at scale. As adoption accelerates globally, AI is increasingly becoming a foundational layer of economic and digital infrastructure, shaping how data is processed, resources are allocated, and services are delivered across entire ecosystems.

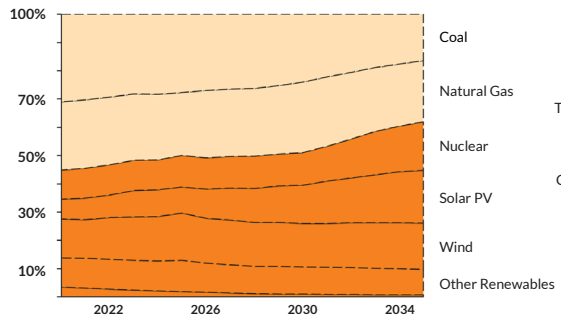
The development and use of AI are resource-intensive and its widespread integration into the global economy is reshaping the demands placed on underlying systems. Data centres rely on large volumes of electricity to power computing equipment and maintain continuous operations, while water is primarily used for cooling servers and managing the heat generated by high-intensity AI workloads. A single AI-focused data centre consumes as much electricity as 100,000 households, with the largest facilities currently under construction projected to consume up to twenty times that amount.[1] Looking ahead, data centre electricity consumption is set to more than double to around 945 TWh by 2030.[2] Additionally, large data centres can consume up to 5 million gallons of water per day, equivalent to the water use of a town populated by 10,000 to 50,000 people.[3]

Two binding constraints are creating complex, interconnected challenges to the continued growth of AI systems. These two constraints are: 1) the limited availability of clean, reliable power and 2) the mismatch between data centre siting and local resource availability. In the near term, much of the incremental energy demand from AI-driven data centres is likely to be met by fossil fuels, as variable renewables alone cannot yet provide the reliability and uptime these facilities require, and clean energy expansion is not keeping pace with data centre growth. At the same time, data centre location decisions are typically driven by connectivity and proximity to users rather than availability of clean energy and grid capacity. Further, AI innovation and infrastructure are increasingly gravitating towards geographies with abundant energy, capital, and digital capacity, raising barriers to entry for countries, firms, and innovators operating under tighter resource constraints. The compute, connectivity, and power requirements of many AI solutions make their adoption challenging in low-resource and low-connectivity contexts. However, in such contexts, AI's potential economic and social returns are often highest, including for small and medium-sized enterprises and local service providers.

Figure 1: Data centre expansion outpaces the development of clean energy infrastructure[4]

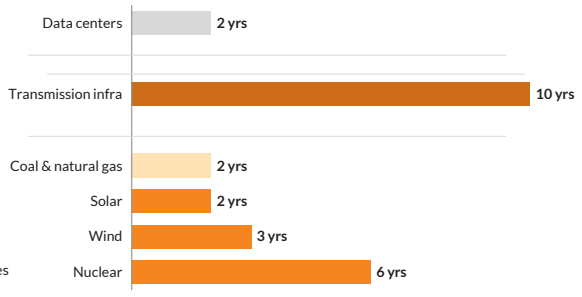
A meaningful shift to renewables is still far away, with fossil fuels supplying 40-50% of global electricity for data centers.

Fig: Historical and forecasted energy mix for global data centre electricity, 2020–2034[1]



Data centres and generation projects can be built quickly but the critical bottleneck is establishing network connectivity

Fig: Comparison of typical timelines for grid projects vs. data center construction[2]



Together, these challenges weaken the return on investment across a wide range of AI applications, from commercial and industrial deployments to public-interest and "AI for Good" development use cases. This makes it harder for many solutions to move beyond pilots and achieve durable, commercially viable scale, as limited access to reliable power, cooling, connectivity, and supporting infrastructure raises operating costs and constrains where solutions can be deployed. Where these resources are scarce or expensive, applications become harder to run consistently and scale beyond initial pilots, making durable, commercially viable expansion difficult.

Addressing these constraints requires a systems approach that recognises the interdependencies between AI workloads, energy supply, infrastructure planning, and regulatory choices, and supports long-term resilience. Decisions around AI deployment, data centre siting, power generation, and grid capacity are often made in silos, even though their impacts on resource efficiency are tightly linked. Without coordinated planning, efforts to scale AI risk shifting pressure points across systems rather than resolving them, thereby undermining reliability, increasing costs, and limiting the flexibility of AI infrastructure.

A systems approach can help to design and manage these elements together to strengthen resilience and allow AI to scale in a way that is both reliable and adaptable across different contexts.

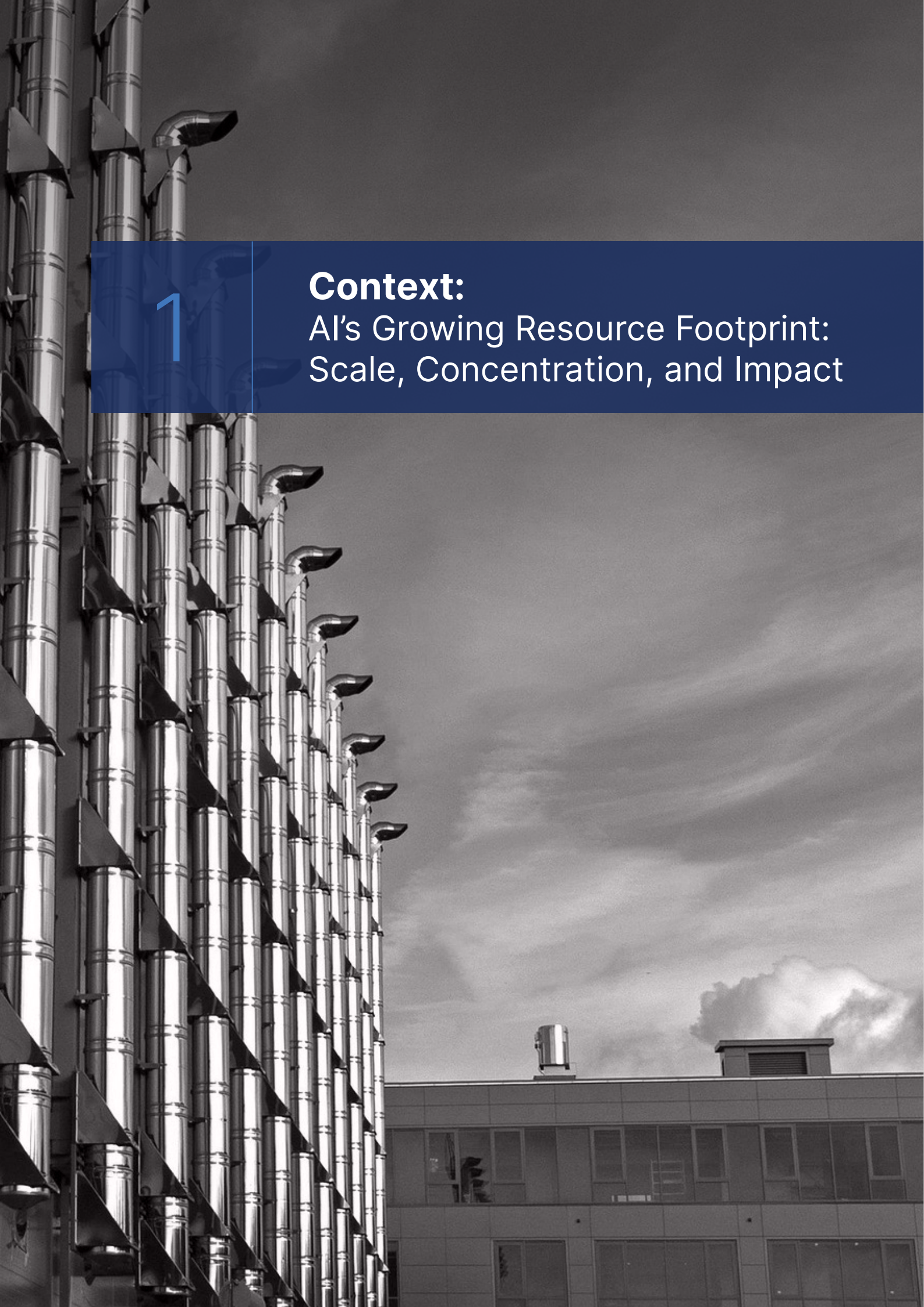
At the core of this approach is resilience. In the context of AI, resilience refers to the ability of systems, spanning compute, energy, infrastructure, and institutions, to absorb rapid growth, adapt to changing conditions, and continue delivering reliable value over time. Resilient AI systems are not only powerful, but also efficient, adaptable, and capable of operating across diverse contexts without locking economies into fragile or inflexible pathways. Embedding resilience enables AI to scale while maintaining system stability, protecting essential services, and preserving space for broader development priorities.

This playbook aims to challenge the emerging narrative that AI advancement and progress in AI-driven economic gains must come at a substantial environmental cost. It explores the current momentum as a strategic inflection point to shape how AI systems and infrastructure are designed, deployed, and integrated so they can scale reliably under real-world constraints.

This Playbook examines how the AI value chain can be made more resilient and efficient by drawing on emerging evidence, technical innovations, and strategic approaches from across geographies. It focuses on three core questions:

- How will AI scale, and what will this mean for the reliability and resilience of energy systems?
- What technological solutions and enabling conditions can advance resource-conscious and resilient AI?
- What strategic pathways are available to governments, the private sector, and the broader international community to advance resilient AI?

The expansion of AI infrastructure will shape economic outcomes, livelihoods, and societal well-being in the decades ahead. A clear understanding of real-world constraints, existing solutions, and the potential impact are critical to making good decisions and avoiding potentially false trade-offs. Addressing this challenge requires deliberate actions: improving efficiency and resilience across the AI value chain, increasing investment in clean and reliable energy, and putting in place necessary policies and enabling conditions to effectively prioritise resilient AI. With the right approach, AI can become efficient enough to scale, inclusive enough to serve, and ethical enough to endure.



1 Context:
AI's Growing Resource Footprint: Scale, Concentration, and Impact

As AI adoption moves from experimentation to economy-wide deployment, the question facing policymakers and industry is no longer whether AI will scale, but how. AI is increasingly embedded into core economic functions, supporting real-time decision-making, automating complex workflows, and enabling continuous optimisation across sectors. Unlike earlier waves of digitalisation, AI systems place simultaneous demands on compute intensity, latency, reliability, and availability. AI workloads must often operate continuously, respond in real time, and scale rapidly as usage grows. As a result, AI growth is tightly coupled with the capacity of underlying physical systems, particularly electricity, cooling, data transmission, and land. This shift from episodic use of compute to persistent, always-on demand fundamentally alters the profile of digital infrastructure required to support AI-enabled systems.

Rapid growth of data centre footprint

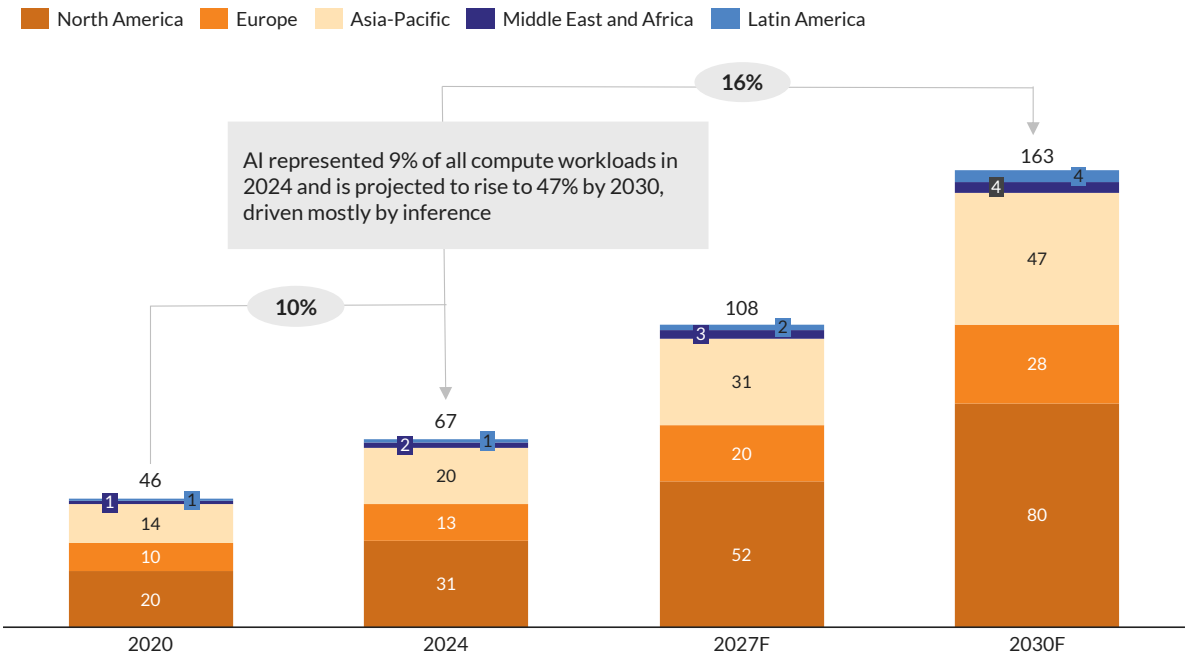
These pressures are most visible in the rapid expansion of data centre infrastructure, the physical backbone of AI deployment. Data centres concentrate compute, storage, and networking capacity, translating AI adoption into tangible demand for electricity, water, and land. The exponential growth in AI workloads significantly increases data centre capacity requirements that existing power and energy systems may not always be designed to absorb. The mismatch between the rapid pace of technological advancement and the longer development timelines of power and energy infrastructure further exacerbates the challenge.

There are different types of data centres, and distinguishing between them is critical to understanding how AI-driven infrastructure is evolving and where pressures on energy and resources are likely to emerge. Data centres vary not only in scale, but also in function and workload profile. Very large facilities, typically above 250 MW, are generally hyperscale data centres operated by large technology firms. Smaller facilities below this threshold are more commonly co-location or enterprise data centres serving a wider range of users and applications.[5,6,7] Importantly, scale alone is not a reliable proxy for AI intensity: not all hyperscale data centres are dedicated to AI workloads, and many support a mix of cloud services, storage, and conventional computing alongside AI.[8] Recognising these distinctions is essential for accurately assessing infrastructure needs, energy demand, and resilience implications, and for avoiding one-size-fits-all assumptions in policy and planning.

AI-driven demand for data centres comes from three main sources. First, training advanced and frontier models requires extremely high-intensity compute loads concentrated over short periods, often involving thousands of specialised chips running continuously for weeks (for example, GPT-4 is estimated to have taken ~15 weeks and consumed ~42.4 GWh of electricity).[9,10] Second, once models are deployed, inference demand can contribute to sustained increases in everyday computing, particularly when AI features are integrated at scale into mainstream digital services and always-on applications. Third, beyond core training and inference, broader AI-adjacent activity is expanding the overall data and compute footprint of digital systems, increasing requirements for storage, transmission, edge computing, and new data-heavy use cases such as real-time analytics, industrial IoT, and generative video. Together, these dynamics place growing pressure on energy and digital infrastructure systems.

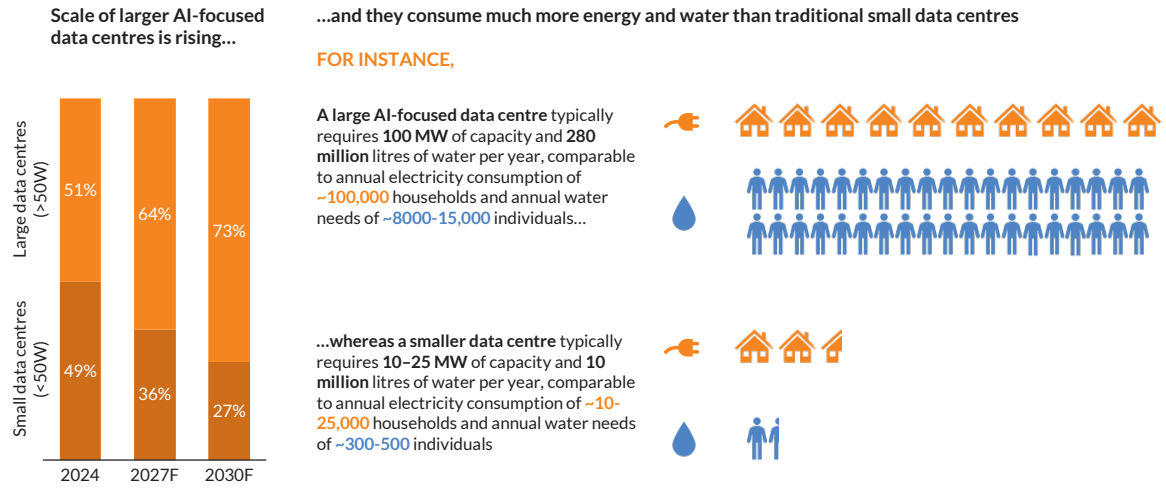
Together, these trends are driving rapid growth in global data centre capacity. Capacity is expected to expand by approximately 16% on average over the next five years, compared to around 10% growth over the previous five years, with roughly 80% of capacity remaining concentrated in advanced economies.[11]

Figure 2: Global data centre capacity by region (Gigawatts)[12]



Recent growth has been driven primarily by large, AI-focused facilities designed to support intensive training and inference workloads. While these facilities are often more efficient per unit of compute than traditional data centres, their scale results in much higher absolute electricity and water use. A typical large AI-focused data centre consumes 100 MW of capacity and 280 million litres of water per year, equivalent to the annual electricity consumption of ~100,000 households, compared to 10 million litres of water per year and ~10,000-25,000 households for smaller data centres (see Fig.3).

Figure 3: As the share of larger AI-focused data centres rises, the overall water and electricity use will rise[13,14,15,16,17]



Globally, data centres used 415 TWh of electricity globally in 2024, and according to estimates, this is set to rise more than 2x to approximately 945 TWh by 2030. AI-related demand is expected to make up 35–50% of that total (compared to 5-15% in 2024).[18] Water use for cooling data centres is increasing even more rapidly: between 2024 and 2028, AI-focused data centres' water consumption is expected to increase 11x from approximately 95 billion litres per year to around 1,068 billion litres annually. This is the equivalent of the basic annual water needs of roughly 30–60 million people, or the size of Spain's population.[19,20,21] Water needs are expected to further intensify as data centres are built in regions with average temperatures above 27°C, where more water is required for cooling. Some estimates suggest rising temperatures can impact water requirements of two-thirds of data centre hubs by 2040.[22]

A small global share, big local consequences

Data centres remain a small share of global electricity and emissions; however, these figures hide a more nuanced country and regional story. At the global level, data centres currently account for ~1.5% of total electricity consumption and are projected to reach 2–4% by 2030. Reflecting this increase in energy use, data centre emissions are expected to rise from around 0.5% of total global emissions today to approximately 1.0–1.4% by 2030.[23]

In advanced economies, where data centres are largely concentrated today, clustered data centre hubs account for a high share of electricity demand.[24] Clustering intensifies energy demand and infrastructure pressure in a small number of locations, worsening local grid constraints, land use, and water impacts when compared to more geographically dispersed pattern of data centre development. In some Global North countries, data centres can make up to 10-20% of the nation's electricity consumption, with demand heavily concentrated in cities serving as data infrastructure hubs. In some high-density hubs, data centres are also becoming a major driver of water demand, potentially accounting for up to 90% of industrial water use by 2030.[25]

In developing economies, data centres account for a much smaller share of national electricity consumption and are expected to remain so over the next five years. For example, in some countries in South-East Asia, data centres consumed less than 1% of total electricity in 2024 and are estimated to remain below 3% through 2030.[26] Even in this context, data centres have significant local impacts and can place significant pressure on regional power grids, water resources, and land use. This is particularly significant in urban and industrial hubs where data centre capacity is concentrated. In these contexts, data centre electricity demand must be carefully integrated alongside ongoing priorities, such as efforts to improve reliability of supply for households and small businesses. However, the nature of these constraints can also drive demand for more resilient technologies. The demand for new data centres in developing countries creates opportunities to embed modern, more efficient technologies and hardware from the outset. This has the potential to lower energy and water intensity relative to retrofitting existing data centres, enabling resilient and resource-efficient AI and data centre scale-up from the outset.

2 The Challenge: Binding Constraints to Resilient AI Development

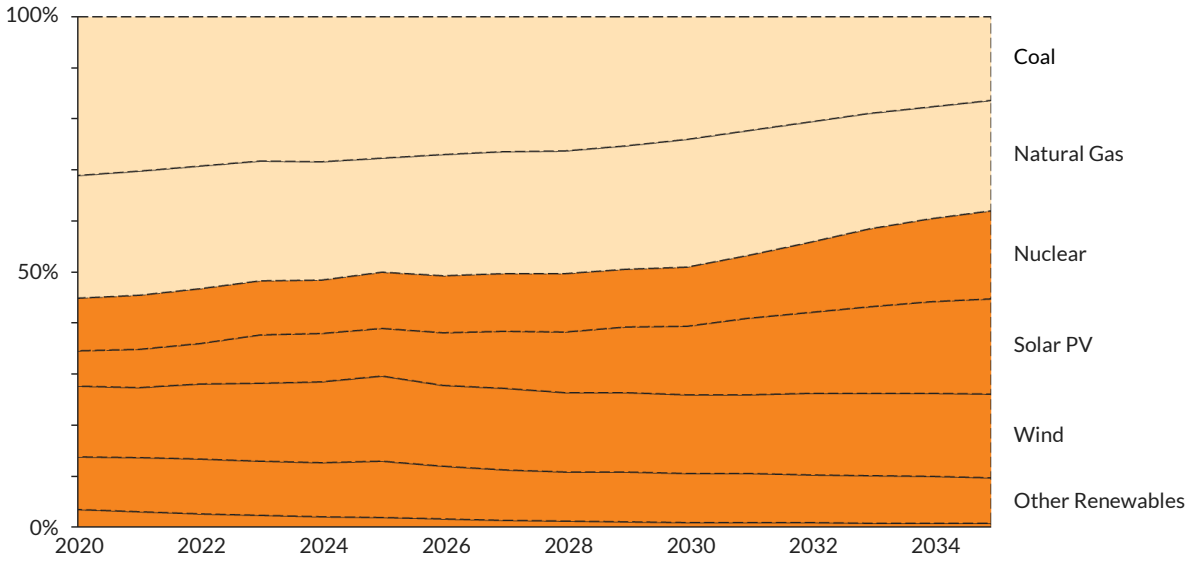
Given the impacts outlined above, it is critical to build resilient and resource-conscious AI systems and infrastructure that maximise impact using fewer resources. However, there are two binding constraints: (i) the limited availability of clean, reliable power to meet additional demand, and (ii) a growing mismatch between data centre siting and local resource availability. These constraints are complex and interconnected, and addressing them requires taking a systems view of how data centre development intersects with energy systems and land and water resources.[27] Without rapid, coordinated action to align data centre expansion with cleaner, firmer energy supply, water- and energy-efficient design, and stronger local planning, countries risk locking in long-term consequences to people and the planet.

Constraint 1: Limited availability of clean, reliable power to meet additional demand

At the global level, electricity used by data centres is currently sourced in roughly equal parts from clean sources (renewables and nuclear) and non-clean sources (coal, oil, and natural gas) and is projected to remain largely unchanged over the next five years. This trajectory underscores the structural challenge of expanding electricity supply at the speed and scale required by AI, while maintaining energy security and managing system costs.

Transition pathways differ significantly across regions, shaped by existing energy mixes, infrastructure readiness, and national priorities. Countries with large data centre markets continue to meet most of the data centre electricity demand with natural gas and coal. Europe is expected to meet most of its incremental electricity demand with renewables and nuclear power, while already relying heavily on low-carbon sources to meet existing electricity needs. However, this accelerated shift has involved higher system costs and a continued reliance on fossil fuels during periods of low renewables output or peak demand. Additionally, countries in Asia are likely to remain more coal-reliant in the near term even as renewables expand over time.

Figure 4: Historical and forecasted energy mix for data centre electricity globally[28]



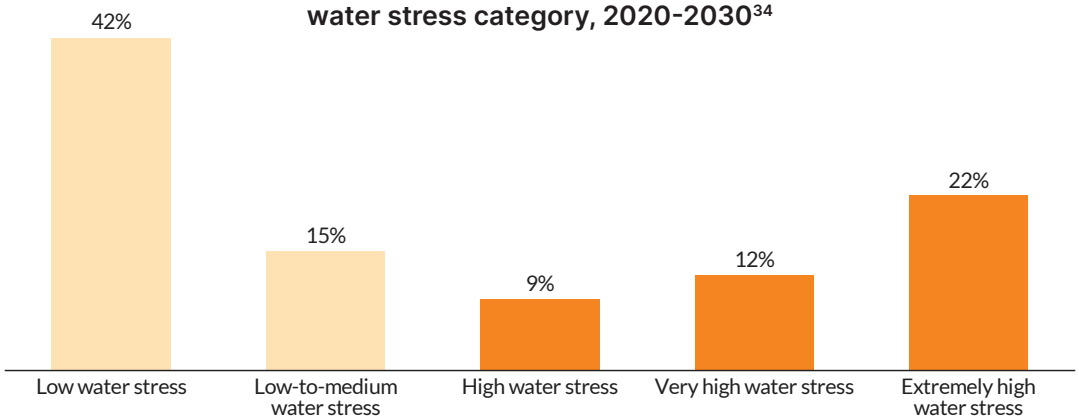
Two factors explain the persistence of this carbon heavy energy mix:

1. **AI-related data centres require extremely high reliability grids (99.999% uptime), which is most easily achieved through fossil fuels.** While existing electricity grids can generally deliver this level of reliability, they remain heavily dependent on fossil fuels. This is because clean energy sources such as solar and wind are intermittent, making it challenging to consistently meet the continuous uptime requirements of data centres without additional system support. In countries where grids are unable to meet this reliability requirement, data centres often rely on on-site generation that is typically based on diesel and is thus carbon-intensive.[29,30]
2. **The pace of data centre expansion is faster than the development of reliable clean energy infrastructure.** Planning and constructing a hyperscale data centre typically takes between 0.5 and 3 years, whereas transmission build-out and grid interconnection capacity required to deliver wind or solar energy to data centres can take a decade or more.[31,32,33] In addition, data centres are not always the primary or anchor customers for new clean energy projects, limiting their ability to directly drive the deployment of low-carbon power. In practice, this has often led operators to contract available renewable options based on access or convenience, such as wind power, which may not be the most cost-effective or system-optimal choice to meet data centre demand.

Constraint 2: Mismatch between data centre siting and local resource availability

Demand for data centres is increasingly concentrated in locations where water resources are already constrained, highlighting the need to balance AI infrastructure expansion with local environmental considerations. Data centre siting decisions have typically prioritised network connectivity and latency, market access and proximity to the end users, reliable power availability, and economic incentives. This has resulted in a concentrated development of data centres near metropolitan or urban hubs, which already face high resource demand due to dense populations and competing urban uses. Globally, 43% of data centres operate in high water stress areas.[34,35] As AI workloads scale and facilities become larger and more resource-intensive, this traditional siting logic is increasingly out of step with resilience and system realities. AI-focused data centres require substantially more electricity and water for operations and cooling, particularly in warmer climates, turning land and water from secondary considerations into binding local constraints.

Figure 5: Proportion of global data centers under each water stress category, 2020-2030[34]



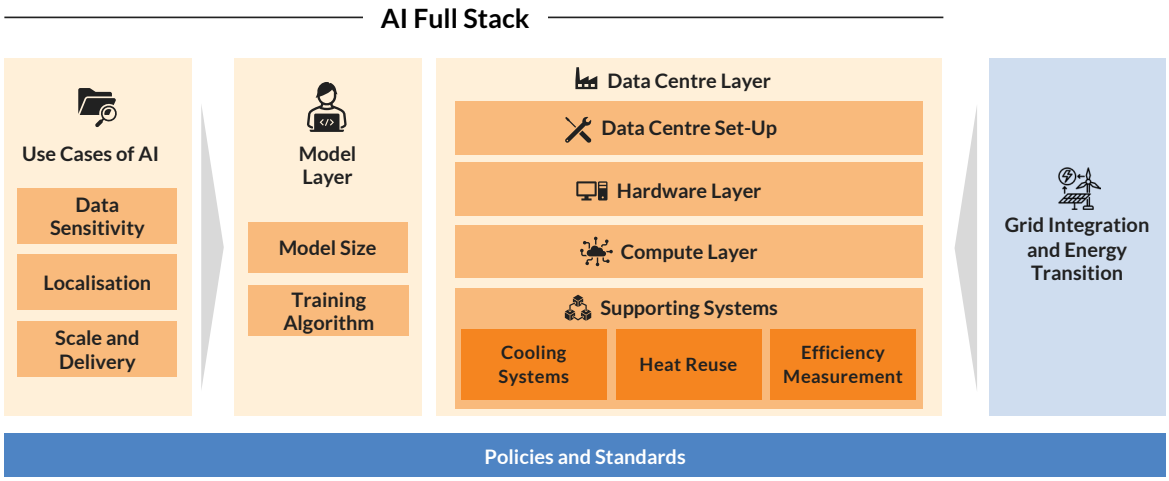
Together, these two constraints underscore the need for a systems approach to building resilient AI, one that requires coordinated action across the AI stack, energy systems, and land and water resources.[36] Limited clean and firm energy readiness at a reasonable cost reflects how AI demand interacts with power generation, grids, and energy markets, while misaligned data centre siting exposes pressures on land use and water availability. While individual solutions can deliver meaningful gains, they have the greatest impact when implemented together, since challenges span the AI technology stack, energy systems, and the policies and practices governing the use of land and water resources.

It is therefore critical to take a systems lens that weighs and considers the full spectrum of levers required to implement effective, integrated solutions. This includes improving efficiency across models, hardware, compute, and data centre operations, aligning new facilities with locations that have cleaner power and greater resource availability, and accelerating investment in clean, firm energy supply. These measures can help ensure that the benefits of AI-driven growth are shared by all, without creating long-term costs for energy systems, communities, or the environment.

3 Solution Set: Adopting a Systems Approach to Advancing Resilient AI

The rapid global expansion of AI has been matched by the emergence of powerful solutions that directly address the binding constraints outlined in the previous section. These solutions are intended to enable AI to scale efficiently without constraining its vast economic potential. As highlighted above, doing so requires a systems approach, with concurrent solutions across the AI stack, energy systems and public policy.

Figure 6: A systems map outlining the AI Stack, Energy Systems, and Policies and Standards



A systems approach mobilises multiple stakeholders to coordinate investments towards building resilient AI, rather than driving isolated efforts. This includes making efficient design and deployment choices across:

1. **AI Use Cases:** defining the purpose of AI solutions and their reliability, latency, and data requirements to shape downstream decisions including the type and intensity of compute required. This helps ensure that compute and energy use are scaled to meet a robust assessment of current and future requirements.
2. **Model Layer:** focusing on compact, task-specific models that are efficiently trained can deliver comparable or even superior performance to larger general-purpose counterparts. These approaches directly help reduce compute intensity, electricity use, associated energy and water demand, and help tackle the rising energy footprint of AI.
3. **Data Centre Layer (including set-up, hardware, compute, and supporting operations):** determining where data centres are located and how facilities are designed and operated drives the energy, water, and emissions footprint of AI infrastructure. This enables resilient, energy-efficient, and resource-aware scaling of AI.
4. **Grid Integration and Energy Transition:** assessing how data centres interact with the power systems through demand responsiveness and clean power sourcing determines whether AI demand accelerates or strains the energy transition. This helps prevent locking-in energy to fossil fuels and promotes non-fossil generation.
5. **Policies and Standards:** creating necessary framework and conditions needed to align AI growth with energy and development priorities by setting guidance on transparency, performance, coordinated planning, and shared learning.

Solutions Compendium across the AI Stack and Energy Systems

There is a host of proven and emerging solutions that are being implemented globally across the AI stack and associated energy systems.[37,38] The following compendium of such solutions highlights how stakeholders across the AI ecosystem can together advance resilient and resource-conscious AI systems and infrastructure, while meeting performance and reliability needs.

1. AI Use Cases

A use case lens should be the first step for any AI build and deployment decision by policymakers and private sector stakeholders.

#1 Tie AI development to national economic and social outcomes.

Policymakers could start by identifying priority AI use cases that align with their national strategies, such as improved public services, R&D and innovation, private-sector productivity, high-performance computing (HPC), or inference at scale for citizen or enterprise workflows.[39] For example, **Norway's National Strategy for Artificial Intelligence** requires public agencies to consider the potential of AI innovations and value creation particularly in areas where the country has strong business and research communities such as health, energy and maritime.[40] **Vietnam plans to establish a National Data Centre by 2030** that will integrate all national and sectoral databases for public service delivery, data-driven policymaking, ensuring cybersecurity, etc., effectively reserving a portion of the country's energy, land, and financial resources for sovereign digital infrastructure.[41] **The Canadian Sovereign AI Compute Strategy** has committed $2 billion to expand domestic compute (including public supercomputing) and strengthen data/IP protection to enable efficient, resilient innovation across research and industry.[42] **The UAE's AI in the Ring Index** provides for performance evaluation that goes beyond technical accuracy to assess contextual and cultural alignment of AI tools. Further, **UAE's Generative AI Guidelines and the AI Adoption Guideline in Government Services,** supports entities in making informed decisions on use cases, data handling, human oversight, and operational deployment.[43] Since 2018, **France has implemented a national strategy for artificial intelligence, structured around research, training, and the economic dissemination of AI technologies.** This strategy has mobilized €2.6 billion to date. The last phase focuses efforts on deploying AI in sectors with the highest potential for productivity gains and strategic sovereignty.

2. Model Layer

Choices by AI developers in the model layer, such as model size and training approach directly shape compute and energy requirements.[44]

#2 Proactively define service-level requirements.

Private sector organisations and government entities can define reliability and latency needs, data storage, compute intensity, data characteristics like sensitivity and degree of localisation, and privacy and governance requirements. This enables proactive workload optimisation and more appropriate model and infrastructure choices later, reducing unnecessary compute, helping to reduce energy use, emissions, and costs. For example, **OpenEuroLLM is a €37.4 million collaborative project** between leading AI startups and research organisations to develop a foundational AI model with defined parameters to maximise relevance to Europe.[45] By identifying the scope of the model proactively, it reduces the need for more extensive and powerful frontier LLM models with projected training costs of over €1 billion by 2027.[46] **BloombergGPT is a specialised LLM for financial services tasks** and trained on a mix of proprietary financial and general-purpose data. By scoping the model for financial tasks, Bloomberg reduced the model size to 50 billion parameters, roughly a third of frontier models like GPT-3.

#3 Build fit-for-purpose small models.

Small language models (SLMs) and workflows that are optimized for specific tasks, industries, or languages, have significantly reduced model size, computational intensity and energy demands, making them 10-30xcheaper to operate than oversized general-purpose models. UNESCO's research indicates that tailored, small models can reduce energy use by up to 90%.[47] For example,

the **UAE AI Charter and Generative AI Guidelines** encourage fit-for-purpose and proportional AI deployment. This supports informed use-case selection and helps avoid unnecessarily complex or compute-intensive solutions.[48] Government entities in Egypt use **compressed Arabic NLP models** deployed on national infrastructure to automate public service delivery workflows.[49] This significantly reduces compute costs compared to large, general-purpose models. **Myanmar** is currently pursuing a strategy focused on localized, lightweight, and resilient solutions.[50]

#4 Develop edge-deployable models.

AI models that are trained in a data center but runs inference on consumer devices like smartphones or laptops rather than in the cloud, avoid high energy consumption from continuous cloud compute and data transfer. For example, Mexico is using edge computing to reduce data traffic to the cloud and, therefore, decrease the energy consumption of transmission.[51] AI-enabled **smart agriculture and healthcare diagnostic models deployed in Egypt** are lightweight models supporting agricultural decision-making and imaging diagnostics. They run on existing hardware, improving service quality in resource-constrained settings.[52] Similarly, **PlantVillage's Nuru app or Farmline's Darli AI** embeds edge AI directly onto farmers' mobile devices, enabling real-time weather updates and offline crop disease diagnosis without needing continuous connectivity to the internet.[53,54] **Qure AI, an Indian healthcare company,** builds models to interpret medical imaging that are deployable locally in hospitals. Local deployment reduces reliance on continuous network connectivity, particularly valuable in low-bandwidth settings; and reduces turnaround times by 40%, with typical processing times of under 20 seconds per scan.[55]

#5 Optimise models through smart training and model compression.

Smarter training techniques such as knowledge distillation and model compression (pruning and quantisation) reduce the compute intensity of models while retaining accuracy. Knowledge distillation reduces energy use by training a smaller 'student' model to replicate the behaviour and reasoning of a larger, more computationally intensive 'teacher' mode.[56] Pruning removes unnecessary elements in neural networks, reducing computational complexity, while quantisation reduces the numerical precision of computations to improve efficiency and speed. Accenture Labs found that training models on just 70% of the full dataset led to less than a 1% reduction in accuracy, while reducing energy consumption by 47%.[57] For example, **Canada's National AI Institutes** are demonstrating compute-efficient AI practices at the model application layer: The Vector Institute's FairSenseAI reports substantial emissions reductions from optimization during training and inference while integrating responsible AI risk tools, illustrating how energy efficient frameworks can be embedded in practical systems.[58] **DistilBERT,** a distilled version of Google's BERT produced ~47% less CO_2 during training and reduced the model's size by 40%, while preserving 97% accuracy and running 60% faster.[59] In China, Baidu made its Ernie AI models open-source in 2025, with internal performance benchmarks matching leading AI models.[60] These include the Ernie Slim and Lite models that are compressed and distilled models for increased efficiency.[61] Baidu also publishes the PaddleSlim, an open source model compression toolkit using pruning and quantisation strategies to reduce model sizes up to 75%.[62]

3. Data Centre Layer

3.a. Data Centre Set-Up:

The data centre layer translates compute requirements into physical infrastructure decisions, including where data centres are sited and how they are built and operated.

#6 Site and co-locate data centres based on resource availability.

Operators can locate data centres in greater proximity to clean energy sources and low water stress regions. They can also set-up localised energy infrastructure such as captive power generation stations based on non-fossil fuel sources. This lowers operational emissions and reduces likelihood of water risk (e.g., shortages, higher costs, community pushback) over a data centre's lifetime. For example, **Finland's National Roadmap for Data Centres** prioritises data centres that are located close to power generation, supporting their energy needs and reducing the need for new grid build.[63] **Brazil's**

Scala is building a ~4.75 GW green data centre campus in Rio Grande do Sul, strategically sited near abundant renewable energy and backed by a memorandum of understanding with a clean energy provider.[64] **Ontario's proposed grid connection approval** regime establishes a gating mechanism that prioritizes data centres delivering economic value and domestic data housing, aligning siting/timing with system adequacy and clean power availability.[65]

#7 Consider micro and modular data centres.

These are smaller, modular edge facilities located near end users or data sources that deliver inference workloads for use cases requiring low latency or local data processing, reducing the need to build large-scale, resource-intensive data centres. For example, **Kenyan utility KenGen** has set up a 52kW modular data centre, powered by renewable energy batteries, that has the potential to be scaled based on local needs allowing compute in energy-constrained contexts.[66] **EdgeUno**, a network infrastructure and edge cloud company, in Latin America has set up edge data centres in more than 47 locations including Chile, Argentina, Costa Rica, etc to deliver low-latency services closer to users while regional facilities handle batch workloads.[67]

#8 Employ low-carbon principles in construction.

Low-carbon concrete and steel, reused or recycled materials, and optimised structural systems and layouts ensure efficiency during data centre construction.[68,69] This reduces embodied carbon in new builds, cutting lifecycle emissions beyond operational energy efficiency. For example, **France has set up a task force** to identify brownfield sites that can be converted and whose existing buildings can be reused as data centre.[70] In **Mexico**, LEED Gold/Silver certification requires the infrastructure to meet international standards for energy efficiency and sustainable materials from the beginning.[71] **Meta** has re-engineered its data centre design and materials strategy by eliminating concrete in non-essential applications, adopting low-carbon concrete mixes with fly ash and slag that cut emissions by up to ~20% below regional baselines.[72] India's **Vigyanlabs** developed specialised bricks that require 70% less cement and lower cooling costs by 60% in its micro data centres.[73,74]

3.b. Hardware Layer:

Adopting efficient, task-appropriate architectures and embedding circularity across hardware lifecycles help manage rapid obsolescence, reduce e-waste and the overall footprint of AI infrastructure.

#9 Use energy-efficient AI accelerators for compute.

Efficient, task-appropriate hardware like Neural Processing Units (NPUs), Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs) is designed to perform parallel computations more efficiently than general-purpose processors.[75,76] When upfront capital investment is limited, cloud-based NPUs/TPUs are useful. However, current innovations are early-stage and better suited for specialised tasks, while GPUs remain more versatile for a variety of demanding tasks.[77] Lower energy use per computation reduces heat generation and cooling requirements, thereby lowering overall electricity and water demand. For example, **Google's Tensor Processing Units (TPUs)** are custom-designed AI accelerators optimised for training large deep learning models.[78] They deliver better results compared to traditional CPUs, while being more energy efficient, and outperform GPUs in tasks tailored to the TPU architecture, but do not perform as well when comparing across broader applications.[79]

#10 Cap power utilization of hardware.

Power capping limits how much electricity servers and accelerators are allowed to draw even during heavy compute loads and several cloud and server companies offer built-in options to turn on 'power saving' mode. This lowers energy consumption and cooling requirements, while preserving performance. For example, **MIT Lincoln Laboratory** researchers have developed power-capping approaches that reduce energy use by 12–15% while increasing time-to-result by only ~3%.[80] **NVIDIA** is designing new GPUs with built-in power limits that allow data centres to cap power draw with minimal performance impact.[81]

#11 Employ circularity principles in hardware management.

Design hardware with longer lifespans and modular parts, refurbish and redeploy components, and partner with recyclers to recover valuable materials when devices reach end of life to reduce hardware turnover and e-waste from data centres. Refurbishing and reusing components could reduce e-waste generation and embodied emissions by up to 86%. Many leading companies are beginning to collect, refurbish, and reuse computing hardware, supporting circularity and reducing e-waste.[82]

3.c. Compute Layer:

By shifting workloads to specific time slots and locations where clean electricity is available, cloud and CPU-based platforms can run AI training and inference using less carbon-intensive energy.

#12 Dynamically schedule workloads based on clean energy availability.

Smarter workload management shifts non-critical AI computations across time periods or geographies to align with periods of better grid stability and higher renewable energy availability, enabling cheaper and greater use of clean power.[83] This increases the share of renewables use, reduces emissions and grid congestion and can lower overall energy use and costs. For example, **GreenPow's** proprietary scheduling algorithms embedded in its cloud infrastructure enables AI workloads to run at times and places with the lowest emissions leading to emissions reduction of up to 60% per task compared to traditional cloud execution, and 30% reduction in energy costs especially during green peak hours.[84] Leading tech companies are also utilising carbon intelligent computing platforms along with hourly grid carbon forecasts from Electricity Maps to align their data centre operations with the availability of low carbon electricity across their global fleet to reduce overall footprint without impacting service reliability.[85]

#13 Virtualise servers to enable cloud computing.

Virtualisation and consolidation allow multiple virtual servers to run on a single physical server, consolidating workloads and improving server utilisation by lowering the requirement for physical energy-intensive data centres. Cloud data centres are custom-designed facilities optimised for existing hardware usage and energy efficiency resulting in 1.4x–2x lower energy consumption compared to on-premises computing.[86] For example, in **Canada**, NRCan's Best Practice Guide for Canadian Data Centres provides Canada-specific guidance for virtualization/consolidation to systematically lower energy and water use. It includes practical tips based on international best practices, all tailored for the Canadian context.[87] Through large-scale virtualisation, **Citigroup** had consolidated 40,000 servers, improving server utilisation rate from 5-10% to 40-50%, delivering immediate energy savings while reducing data centre land requirements.[88,89] They currently use IBM's LinuxONE servers that turn on unused cores supporting thousands of workloads in the footprint of a single system.[90]

#14 Use CPU-first inference platforms.

Use CPU-based inference platforms that accelerate AI output generation and reduce latency, reducing dependence on resource-intensive GPUs. This is done by optimising the runtime execution of models on CPUs and improving token-by-token generation efficiency while preserving the model's original weights and accuracy. This reduces high upfront and operational costs, and energy demands for cooling, enabling AI to be deployed in resource constrained environments. For example, **Kompact AI's** proprietary algorithms have already adapted AI models from big-tech companies to be run on CPUs, reducing fixed costs by up to 50% compared to GPUs and decreasing energy usage cost by 66% annually from lower compute intensity and cooling needs.[91] **Intel** is developing CPUs with embedded accelerators to serve as cost-effective and powerful alternatives to GPUs.[92]

3.d. Supporting Systems Layer:

The tertiary operations of a data centre can be made more resource-efficient by leveraging better cooling systems, reusing excess heat and measuring energy efficiency. [93]

#15 Design thermal-aware, efficient cooling systems.

Operators employ innovative liquid cooling mechanisms like direct-to-chip cooling and immersion cooling where high rack densities and local climate conditions justify it. This improves cooling efficiency and lowers energy demand for thermal management, which tends to use 40% of a data centre's total energy consumption.[94,95] These systems can improve PUE by 1.5-1.7x compared to traditional air-cooling systems.[96] For example, **Microsoft** is currently piloting chip-level **zero-water evaporated cooling technologies** in the US, which use a closed loop system for circulation without requiring fresh water supply, improving WUE metrics by 39%.[97,98] **ST Telemedia Global Data Centres** in Singapore partnered with Phaidra to deploy AI-powered cooling systems that save energy between 10-30%.[99] **Mexico**, following an investment announcement to develop a campus with six data centers in Querétaro, is considering adopting innovations like waterless cooling, where, unlike traditional data centers, this will use closed-loop air cooling systems given that they are being developed in water stressed regions.[100]

#16 Recover waste-heat.

Heat produced as a by-product in data centres can be recovered and directed through pipes to internal or external energy systems like industrial facilities, buildings or district heating grids. This helps displace traditional fossil-based heating and lowers system-wide emissions beyond the data centre. For example, **UK is integrating waste heat recovery** into future infrastructure planning, ensuring data centres contribute to energy efficiency and affordability.[101] **Bahnhof in Sweden** pumps the excess heat to nearby district buildings.[102] **The Tallaght District Heating Scheme in Ireland** captures waste heat from a nearby Amazon data centre and distributes it to connected buildings, delivering nearly 3,700 MWh of heat and saving over 1,100 tonnes of CO_2 to date.[103,104,105]

#17 Measure and track energy efficiency of operations.

Data centre infrastructure management (DCIM) tools can monitor, measure and control key metrics including power usage effectiveness (PUE), carbon usage effectiveness (CUE), and water usage effectiveness (WUE), which track energy use, carbon emissions, and water use of data centres respectively.[106] This enables **benchmarking and continuous optimisation** to reduce energy, carbon, and water intensity over time. For example, **UK's Water Delivery Taskforce** brings together the Department of Environment, Food and Rural Affairs, regulators and water industry representatives to optimise water usage, wastewater and drainage infrastructure in data centres.[107] **France's ELAN** law has set energy performance targets for the largest data centres (greater than 10,000 m^2 IT room floor area) to have a PUE of 1.20 in 2030.[108]

4. Grid Integration and Energy Transition Layer

Sourcing power from cleaner energy sources and smarter integration with existing grid systems (that optimise when and how data centres consume electricity) enables scaling AI alongside green energy transition.

#18 Smart grid integration through demand-responsive operations.

Demand-response and demand-flexibility programmes allow data centres to temporarily reduce load during periods of peak demand, easing strain on systems that still rely primarily on fossil fuels. This reduces emissions by shifting flexible compute to when the grid is cleaner and more stable. For example, **Google's 24/7 carbon-free energy (CFE) ambition** includes demand-side solutions that shift non-urgent compute tasks, such as YouTube video processing or ML workloads, to periods when the grid is less constrained and less emissions-intensive.[109] In **Canada**, utility-level instruments like demand response are providing mechanisms to manage peak loads and improve operational efficiency for energy intensive computing.[110] **Verrus**, a data centre company, enables data centres to rapidly curtail their energy load by up to 100% during grid stress events through grid-aware controls, allowing a switch to on-site generators and battery energy storage systems.[111,112]

#19 Scale renewables and low-carbon energy over time; in the near term, bridge gaps with captive renewables and storage to meet on-site demand where clean grid power is unavailable.

Virtual power purchase agreements (VPPAs), unbundled renewable energy certificates (RECs) and 24/7 carbon-free energy (CFE) matching approaches are incentives that can enable investments in clean energy generation.[113,114] Renewable energy certificates (RECs) document the consumption of renewable energy allowing data centre operators to own and trade the associated environmental benefits, whereas carbon-free energy (CFE) refers to matching every hour of a data centre's electricity consumption with an equivalent volume of carbon-free energy generation within the same local or regional electricity grid. In the near term, operators can use captive renewables for on-site generation through microgrids, and batteries for storage (when onsite renewables are not generating enough power).[115] For example, **data centres in New Zealand** are backed by **long-term Power Purchase Agreements (PPAs)** with renewable generators, which help underpin new clean-energy investment while providing operators with predictable, low-emissions electricity over the long term.[116] **Ireland's Commission of Regulation Utilities** (CRU) requires new data centres to provide onsite or local generation and/or storage capacity to match the requested data centre demand capacity.[117] **CtrlS Datacenters in India** has constructed GreenVolt 1, a 125MW captive solar farm in Nagpur, supplying 60% of the energy needs of its 116MW Mumbai data centre campus.[118] In Malaysia, **Google and TotalEnergies** signed a 21-year **power purchase agreement** (PPA) under which TotalEnergies will supply 1 TWh of certified renewable electricity from the Citra Energies solar plant to support Google's data centre operations in the region.[119]

5. Guiding Principles for Policies and Standards

#20 Anchor national AI policies in resource-efficiency and resilience.

Define a national AI strategy that guides full stack development in line with resource availability and development priorities. For example, **India's National Strategy for Artificial Intelligence** considers AI as a tool for inclusive growth across priority sectors like agriculture, healthcare, education and focuses on AI adoption, creation of large foundational datasets and industry collaborations while incorporating ethical and privacy guardrails.[120] Outlining principles to encourage fit-for-purpose models prevents over-sizing AI systems and offering incentives to de-risk resilient AI innovations (e.g., providing government-backed grants or tax credits to help firms develop efficient AI solutions), reinforces demand-shaping measures like procurement guidance and user-facing transparency and encourages resilient and phased AI expansion.[121] For example, **France's National AI Strategy** has been rolled out in three phases, focusing on enhancing research capabilities, enabling deployment in priority economic sectors, and accelerating compute infrastructure (including public computing capabilities). Sustainability is a core component of this strategy implemented through various dimensions: low-carbon electricity generation, data center efficiency regulations, local pilot projects using sustainable AI to accelerate the ecological transition, or voluntary guidelines (e.g., for responsible AI public procurement).[122] **The UAE AI Readiness Assessment** is used by government entities and their Chief AI Officers to assess institutional maturity, prioritise AI investments, sequence adoption pathways, and allocate resources more efficiently. This approach supports informed decision-making and helps ensure that AI systems are deployed where organisational readiness, data availability, and governance capacity are strongest.[123] **New Zealand's Strategy for AI: Investing with Confidence** highlights the need to adopt AI solutions for local challenges instead of developing capital and resource-intensive foundational models.[124] **Germany's** Federal Environment Ministry de-risks innovation and generates shared evidence by funding **"AI lighthouse projects for the environment, climate, nature and resources,"** that address environmental challenges while serving as models for resource-conscious digitalization.[125]

#21 Align energy transition goals with AI infrastructure growth.

Forecasting AI sector growth and integrating energy demands with grid planning strategies and low carbon or renewable energy build-out

can guide data centre growth in grid-favourable, clean-energy zones through spatial coupling. For example, **UK's AI Growth Zone programme** creates designated sites with enhanced access to power and streamlined planning processes. These zones are designed to deliver large-scale compute capacity through priority grid connections and reduced electricity costs.[126] **US Department of Energy's recommendations for Powering Artificial Intelligence and Data Centre Infrastructure** calls for assessing AI load growth and integrating it with power planning through grid peak load management and on-site energy generation and storage.[127] **Canada's provinces** are taking differentiated approaches to managing the growth of AI data centres while protecting grid reliability and affordability.[128] **Ontario** is introducing an approval regime that prioritises grid connections for data centres delivering clear economic and social value and aligning with clean power availability, moving away from automatic access for large loads. **Québec,** through Hydro-Québec, provides transparent on-grid and off-grid rate structures that shape siting and operational efficiency, particularly in remote areas. **Alberta** has adopted a phased approval process for connecting large AI loads, balancing its ambition to attract data centre investment with the need to maintain affordable and reliable utilities.[129] To support AI infrastructure growth, **France** is leveraging its low-carbon electricity, pairing it with pre-identified sites that meet specific criteria (including environmental ones) and offering accelerated grid connections for high-power projects.

#22 Guide and monitor resource efficiency requirements.

Creating standardised benchmarks to measure, compare, and improve energy, water and emissions footprint of AI infrastructure and rewarding data centres that meet these expectations, reduces strain on local resources while demonstrating new resource-efficiency techniques. For example, **Germany's Energy Efficiency Law (2023)** imposes binding requirements on data centres, including minimum efficiency thresholds (e.g. PUE targets for new facilities), the use of 100% renewable electricity, and provisions for waste heat utilization.[130] Under **Singapore's Infocomm Media Development Authority's (IMDA's) Green Data Centre Roadmap,** the government awards approvals based on meeting strict efficiency and flexibility requirements (e.g., efficient liquid cooling with goals to reduce WUE from the median 2.2m3/MWh to 2.0 m3/MWh).[131]

#23 Encourage transparency of AI's resource impact.

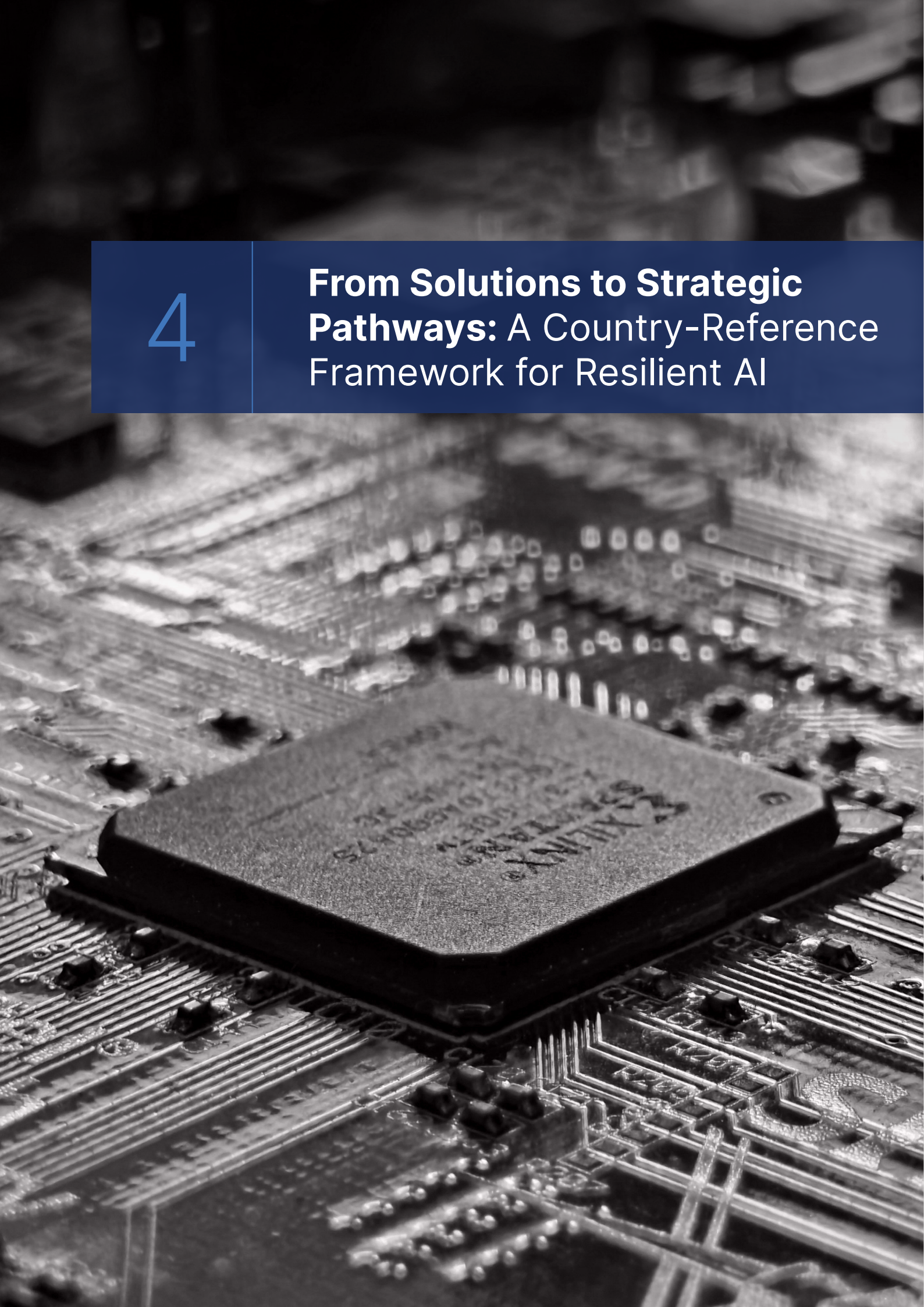
Establishing standardised non-binding methods and interoperable frameworks to measure and report energy use and carbon emissions, including net impact reporting frameworks for AI companies and a public registry of national data centres with reported climate metrics, reduces fragmentation and enables collaboration on best practices to quantify, compare, and reduce environmental footprint of AI systems.[132,133] For example, the **EU Artificial Intelligence Act** requires providers of general-purpose AI models to maintain technical documentation of the model, including a breakdown of energy consumption or estimation based on computational resources used.[134] The **European Union's Data Centre Sustainability Rating Scheme** requires data centres, of capacity greater than 500 kW across the EU, to annually disclose standardised data on energy use, efficiency, water consumption, waste heat reuse and sustainability performance.[135] **France** published in 2024 the first General Framework for Frugal AI, with a methodology to evaluate the environmental impact of AI and a list of good practices.[136]

#24 Enable shared learning and knowledge dissemination.

Creating structured collaboration across governments, industry, and academia can advance AI and climate research and innovations, codify early lessons and shared evidence, and support the development of efficient data centres for AI deployment. This can lower learning costs, speed up diffusion of effective approaches, and support alignment across diverse country contexts. Innovation and experimentation should be pursued in ways that strengthen resilience, resource efficiency, and domestic security. For example, **UK's AI Research Resource (AIRR)** provides free access to AI-supercomputers for academics, public sector organisations and SMEs, supporting innovation and research across

the UK. It has received a £1 billion investment to expand national AI infrastructure by at least twentyfold by 2030.[137] **France's Priority Research Programme and Equipment on AI** provides €73 millions over six years in government funding to resilient AI projects innovating efficient training and inference on specialised hardware and design and deployment of frugal, resource-efficient AI models.[138] **New Zealand's AI Research Platform** developed by the Institute for Advanced Technology provides $70 million over 7 years to create new knowledge-intensive firms to lift the country's economic productivity.[139] The **Mexican government** is encouraging the creation of **National Artificial Intelligence Laboratory,** which will support the development of models that improve public services and train AI specialists.[140]

4 From Solutions to Strategic Pathways: A Country-Reference Framework for Resilient AI



The pace and form of AI deployment are increasingly shaped by countries' underlying energy system constraints and resource endowments, rather than by technological capability alone. As AI and data centre demand grow rapidly, differences in access to clean and reliable power, capital, labour, land, and water (among other factors) are creating divergent national pathways for scaling AI. These structural conditions influence not only how quickly AI can scale, but also what kinds of AI infrastructure, deployment models, and policy choices are viable in different contexts.

Building on this, the process of identifying a country's most salient resilient AI solutions, and the strategic pathways that follow, may be guided by three core principles:[141]

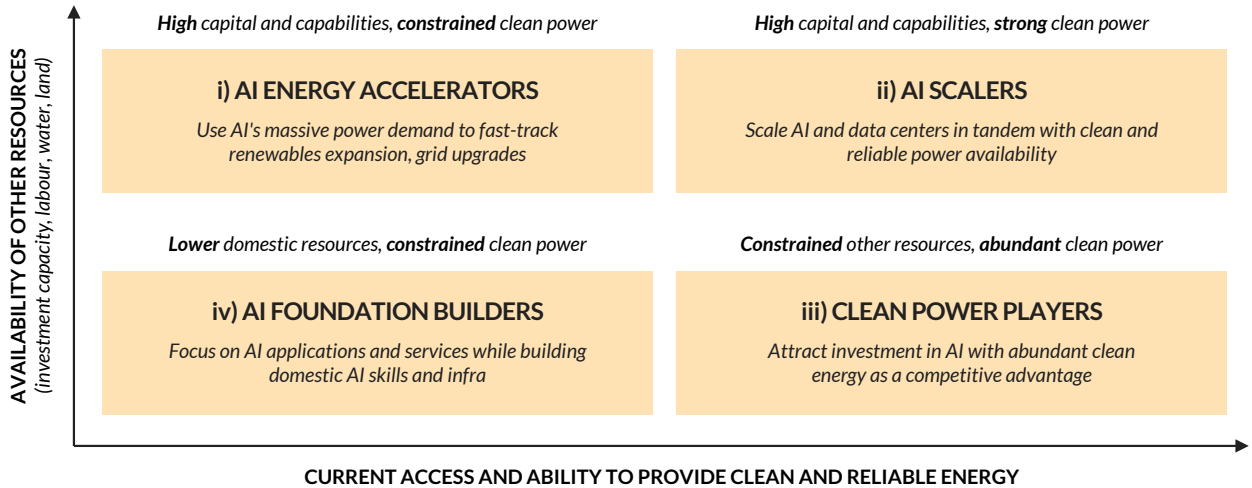
1. **Prioritize Value Creation**
   AI deployment should deliver meaningful economic and social value relative to the resources it consumes. One practical implication is that fit-for-purpose models can be as strategically important as scaling compute infrastructure.

2. **Ground decisions in local country context**
   Local contexts and development priorities will determine the prioritisation given to different solutions even though each would help unlock scale with sustainability. These include constraints in energy systems, land and water availability, financial resources, and institutional capacity.

3. **Maximize efficiency and flexibility before expansion**
   Optimise current capacity utilization to the extent current solutions allow. It is critical to have an accurate assessment of incremental requirements before making significant investments.

The framework that follows applies these principles and situates countries across two critical dimensions: their capacity to supply clean and reliable power, and the availability of complementary resources such as capital, skilled labour, land, and water. Together, these dimensions give rise to four broad country archetypes, each associated with a distinct strategic role in the AI–energy transition.

Figure 7: Country-reference framework[143]



The four archetypes are indicative and are meant to offer a structured way to think through trade-offs and policy options. Country contexts vary widely and they may face risks and opportunities that transcend the boundaries of this framework. The framework is thus intended as an indicative guide that countries can refer to, based on their specific development objectives, resource endowments and institutional capacities.

Rather than prescribing a single optimal pathway, the framework and the deep dives in the following sections highlight how different starting conditions create different opportunities, risks, and strategic pathways for aligning AI growth with energy systems, development priorities, and long-term resilience goals.

(i) AI-Energy Accelerators

Archetype Characteristics:

This archetype sits at an important inflection point, with high AI ambition coupled with capital, skills, and industrial depth. However, structural constraints in grid capacity, or firm renewable power availability mean that AI and data centre infrastructure must scale within tight energy boundaries, making system coordination and phased AI sequencing central to their scaling journey.

Potential Risks

- **Long lead times for clean and firm power expansion:** Renewable generation, storage, transmission upgrades, and firming capacity require multi-year planning, permitting, and capital deployment cycles which can take time, creating near-term pressure to meet AI and data centre demand through existing, carbon-intensive, power sources.
- **Binding competition for limited grid capacity:** Large AI loads may place additional demands on grids. In some contexts, this could require governments and system operators to carefully consider the equitable distribution and allocation of reliable, clean power across developmental priorities.
- **Risk of early infrastructure lock-in:** Without clear policy direction, early AI infrastructure investments may increase the likelihood of pathways that are more carbon-intensive and potentially challenging or costly to adjust over time.

Key opportunities:

These constraints may position AI–Energy Accelerators to **use AI and data centre demand as a strategic lever to strengthen the business case for clean energy investments,** grid upgrades, and firming capacity. Energy scarcity also creates strong incentives to **prioritise efficiency-first AI deployment,** including flexible workloads, modular infrastructure, and early adoption of low-cost, resource-efficient solutions. Beyond supply-side impacts, these countries can also lead in demand-side innovation, including flexible and carbon-aware compute and alternative infrastructure models that decouple AI value creation from ever-rising energy use. The need to align AI growth with energy system constraints can also **strengthen cross-sector planning and coordination,** positioning these countries to shape emerging global norms for resilient AI infrastructure. Over time, this creates scope to develop **exportable technical, operational, and policy expertise** relevant to other energy-constrained contexts.

Potential strategic pathways:

Advance AI growth without displacing critical development priorities

- Introduce tiered reliability and access frameworks, prioritising firm power for critical sectors while requiring AI loads to support grid stability through demand response and workload shifting.
- Define priority AI use cases (e.g. public services, critical infrastructure) for preferential access, while constraining peak-time access for non-urgent training or batch workloads.
- Promote an efficiency-first AI deployment model, encouraging compact, fit-for-purpose models and optimisation measures (such as pruning and quantization) as standard practice for government and enterprise use.

Embed resilience as a core principle in AI policy and governance

- Develop standardised non-binding transparency and disclosure frameworks for energy use, emissions, and water impacts, supported by verification mechanisms and evolving performance benchmarks.
- Consider linking incentives to demonstrated efficiency outcomes by AI models and data centres.
- Codify lessons learned into reusable templates and planning blueprints for energy-efficient, clean-powered AI infrastructure, enabling replication domestically and in energy-constrained contexts.

Prioritise energy transition along with AI scale-up

- Treat AI and data centre demand as an anchor load to de-risk renewable generation, storage, firming capacity, and grid upgrades through long-term commitments such as PPAs/VPPAs, co-investment, or capacity contracts.
- Release grid capacity in phases, prioritising projects that commit to measurable efficiency and standards or serve national/strategic priorities.
- Bridge near-term clean energy gaps through defined transition pathways that combine interim mechanisms (e.g. PPAs, unbundled RECs, captive renewables and storage) with clear milestones toward low-carbon supply.

Design AI infrastructure to preserve low-carbon transition pathways

- Consider phased or modular data centre build-outs, aligned with existing grid capacity and planned upgrades
- Prioritise energy-efficient hardware, higher utilisation through virtualisation, and circularity in hardware lifecycle management to limit incremental power demand as compute scales.
- Enable carbon-conscious compute through time- and geo-shifting of non-urgent workloads based on grid carbon intensity.
- Strengthen efficient data centre operations through performance metrics (PUE, WUE, CUE), advanced cooling technologies, and waste heat reuse where feasible.

For AI–Energy Accelerators, the central challenge is how to scale AI in ways that strengthen energy systems, safeguard broader development priorities, and avoid long-term carbon lock-in. Doing so will require deliberate sequencing, clear policy signals, and governance choices that prioritise efficiency and flexibility in the near term to preserve low-carbon options over time. In navigating this balance, these countries can use AI growth as a testing ground for aligning compute expansion with energy transition pathways, generating lessons that are likely to be relevant well beyond their own contexts.

(ii) AI Scalers

Archetype characteristics:

Countries in the AI Scalers segment are characterised by relatively strong underlying conditions for large-scale AI development, including comparatively higher access to clean, reliable power, along with other critical inputs such as land, water, capital, skills, and institutional capacity. With the right choices, these countries can emerge as global hubs for resilient end-to-end AI deployment and innovation.

Potential risks:

- **AI demand outpacing energy system readiness:** Even in relatively well-prepared systems, AI and data centre demand can scale faster than new clean generation, storage and grid upgrades can be planned, permitted, and delivered.
- **Localised environmental and system stress:** Large, concentrated AI loads can create pressure on land and water resources and exacerbate grid congestion in specific regions. In some cases, this can increase short- to medium-term reliance on fossil-based power.
- **Inefficient AI deployment choices:** Without explicit efficiency guardrails, AI systems may be deployed with unnecessarily high energy, water, and emissions intensity due to suboptimal choices in model design, hardware selection, and operational practices.

Key opportunities:

Existing resource availability creates an opportunity for AI Scalers to **pace AI growth in congruence with strengthening energy systems, embedding resilience into AI design and deployment.** Embedding efficiency and resource-conscious expectations into national AI policies and frameworks can enable strong infrastructure, institutional capacity and **proactive coordination between AI developers, electrical utilities, regulators, and planners.** Leveraging technical and financial resources can enable early leadership in resilient-by-design AI systems, spanning model efficiency, and energy-aware data centre operations. Over time, these countries can generate standards, tools, and reuseable blueprints for resilient and resource-conscious AI that are transferable both domestically and internationally.

Potential strategic pathways:

Shape AI scale-up to support long-term transition to low-carbon pathways

- Forecast national AI and data centre demand and align these with clean energy generation and grid upgrade plans such as transmission/substation upgrades and clean power firming.
- Ensure data centres pair onsite or near-site renewable energy with firming capacity (e.g., storage or firm clean supply), and where feasible encourage tighter temporal matching.
- Ensure spatial coupling by enabling data centre growth in grid-favourable 'clean first' industrial zones with proven grid headroom and proximity to abundant renewables, while limiting heavy builds in urban regions.

Embed efficiency and resilience objectives in national AI strategies from the outset

- Anchor AI growth in national use case prioritisation (e.g., high-public-value services, sensitivity needs), keeping model and infrastructure choices proportional to actual needs and reducing always-on, oversized deployments.
- Follow a portfolio approach to develop a balanced mix of larger all-purpose and small task-specific models depending on use cases.
- Adopt an efficiency lens in data centres including energy-efficient accelerators, early integration of circularity principles across hardware lifecycles, efficient compute scheduling and systematic tracking of energy, carbon, and water performance to support continuous improvement.

Enable resource efficient AI systems through timely investments

- De-risk frontier innovations by co-investing or providing grants for efficient AI solutions that are technically proven. State enterprises can fund R&D projects that drive efficiency and use of renewables.
- Designate specific research facilities as living labs where data centre operations are measured using standardised metrics, creating shared evidence base and allowing new technologies to be tested in a controlled, real-world environment.

Establish and scale global benchmarks and best practices for efficient AI

- Codify best practices, build templates and create a public, reusable blueprint for efficient data centres covering facility design, hardware and compute efficiency.
- Consider setting up a structured training process for data centre operators and AI developers, creating a learning community of practitioners to diffuse the latest knowledge and drive collective enforcement of quality.

For AI Scalers, the central challenge is not whether AI can scale, but how it scales. The task ahead is to ensure that rapid AI expansion reinforces clean energy transitions, resource efficiency, and long-term system resilience. Doing so will require deliberate pacing, strong coordination across sectors, and a shift from viewing resource-efficiency as a constraint to treating it as a source of strategic advantage. If successfully navigated, these countries can establish a globally applicable model for resilient AI at scale, shaping norms, practices, and expectations beyond their own borders.

(iii) Clean Power Players

Archetype characteristics:

Countries in the Clean Power Players segment have abundant clean electricity from strong renewable endowments but face constraints in capital, land, skilled labour, and institutional capacity needed for large-scale AI infrastructure. Stable clean energy systems can be leveraged to participate selectively in the global AI value chain, including through exporting compute and advancing domestic socio-economic priorities.

Potential risks:

- **Concentration of resource impacts:** Without careful demand modelling and infrastructure planning (siting, expansion, and operating standards), AI infrastructure can cluster in specific geographies, creating land and water stress hotspots that are difficult to reverse.
- **Lock-in from long-term contracts and assets:** Long-lived infrastructure and contractual arrangements with foreign firms can lock countries into suboptimal uses of scare resources and limit future policy flexibility.

Key opportunities:

Despite these constraints, Clean Power Players are well positioned to **use clean electricity as a source of strategic leverage in global AI markets.** Potential AI growth can be **aligned with domestic value creation without creating strain** on existing resource capacity. Binding constraints in other resources create a strong rationale to **prioritise efficiency-first, modular, and distributed infrastructure models and full stack resilience** that limits environmental impact while preserving flexibility.

Potential strategic pathways:

Ensure hosting big 'export compute' data centres offers domestic benefits

- Fast-track permitting to 'clean zones' with available capacity and provide financial incentives for projects that enable local benefits such as job creation, co-development of domestic tools.
- Create a registry capturing the location, capacity, efficiency measures, renewable-energy use, & expansion plans of data centres and require operators to report energy and resource efficiency metrics.
- Consider performance-based expansion pathways for large data centres, linking additional MW allocations and renewals to evidence of local value creation (e.g., jobs, training, supplier development) and resilience.
- Build the foundational AI stack, with clean, standardised domestic data systems across sectors, prioritising small models and enabling affordable compute access, and local talent pipelines.

Sequence and spatially plan infrastructure growth to avoid excessive long-term strain on national resources

- Screen sites for proximity to clean generation/firming and low water stress (including access to non-potable sources) and keep large compute campuses where grid and water impacts are manageable while deploying latency-critical inference through modular data centres and consumer-grade edge nodes.
- Invest in firming up renewable energy through battery, pumped hydro, compressed air, or thermal storage to make existing solar and wind capacity dispatchable for 24/7 compute workloads.[143]
- Expansion plans should demonstrate maintenance of PUE/WUE performance and not exceed local capacity. Ensure the use of low-impact construction practices and 'design for circularity'.
- Prioritise energy-efficient accelerators like NPUs/FPGAs/TPUs), use power-capping on GPUs to limit peak draw, and boost utilisation through dynamic scheduling to cleaner/less-constrained hours.
- Adopt climate-appropriate cooling (e.g., direct-to-chip liquid cooling or immersion cooling for high-density racks), prioritise closed-loop/low-water designs, and set strict WUE expectations.
- Reduce e-waste by extending equipment life through refurbishment, reuse, and encourage recovery/resale pathways.

For Clean Power Players, participation in the AI value chain hinges on restraint as much as ambition; scaling compute in ways that remain aligned with grid capacity, clean energy availability, and broader development needs. Countries will also need to manage complementary constraints around land, talent, capital, and connectivity to maximize value creation. Achieving this balance will require explicit resource envelopes, strong conditionality on domestic value creation, and long-term spatial and expansion strategies that avoid irreversible lock-in. If managed well, these countries can demonstrate how compute exports can coexist with national development priorities, offering a replicable model for resource-constrained, clean-energy-rich contexts.

(iv) AI Foundation Builders

Archetype characteristics:

Countries in the AI Foundation Builders segment face binding constraints across both energy systems and other complementary resources such as capital, skills, connectivity, and institutional capacity. The greatest potential value from AI in these contexts lies in the adoption of AI-enabled services that support existing development priorities. By prioritising service-level deployment and keeping compute, energy, and infrastructure needs near pre-AI levels, AI Foundation Builders can realise near-term gains without straining constrained systems while building the skills, institutions, and evidence base for deeper future participation.

Potential risks:

- **Limited governance and procurement readiness:** In early-stage AI ecosystems, where procurement capacity and regulatory frameworks for governing AI are still developing, reliance on external AI service providers can increase exposure to unfavourable contractual terms, vendor lock-in, and reduced control over pricing, data governance, service continuity, and future flexibility
- **Global models can be misaligned to local needs:** AI services designed for global markets may not fully reflect local contexts, user needs, or delivery realities, limiting their effectiveness or uptake.

Key opportunities:

AI Foundation Builders are building from a relatively blank slate, making it **easier to embed resilience from the outset across the stack,** from use cases, model choice, energy procurement, and operations. Tight resource constraints can drive innovation in prioritising high-impact services, using efficient models, and enabling low-power, low-bandwidth AI deployment. This **foundation-first approach** offers a way to translate AI adoption into durable development outcomes without overextending scarce resources.

Potential strategic pathways:

Invest in building foundational capacity for AI growth

- Build the capacity of public institutions to procure, negotiate, and manage AI services effectively, including contract terms related to data use, pricing, service continuity, interoperability, and exit options. Develop clear guidelines and templates to reduce asymmetries with service providers, protect sovereignty and public value, and preserve flexibility to adapt or scale over time.
- Increase incremental clean energy supply and firm up priority nodes by upgrading non-fossil captive installations through renewable storage systems, and on-site solar generation and battery capacity. Additionally, consider financing these investments through concessional climate finance (e.g., Green Climate Fund, International Development Association) at favourable interest rates.
- Build the foundational AI stack, with clean, standardised domestic data systems across sectors, small models, affordable compute access, and local talent pipelines, so countries can deploy high-impact applications while steadily strengthening domestic capability. Establish a secure data exchange/interoperability layer that enables standardised sharing, used to fine-tune or train fit-for-purpose models.

Take early actions to position countries to leapfrog in AI capability over the medium term

- Prioritise limited context-specific use cases where AI meaningfully improves outcomes, considering metrics like impact potential, data availability and operational fit based on existing connecting and power.
- Start building fit-for-purpose compressed models aligned to use case latency, data sensitivity and local contexts using pruning, quantisation and distillation techniques.
- Strengthen affordable, reliable energy access for anchor institutions like schools & hospitals and address barriers like device cost, skills, safety. Investing in digital access allows AI-enabled solutions to start reaching people as domestic AI capacity scales.
- Ensure heavy training takes place on external cloud/regional High-Performance Computing (HPC); deploy inference locally through micro data centres and edge nodes powered by on-site/ near site captive renewable energy and batteries.
- Align AI policy with energy transition by making clean energy and climate impact a first-order design principle. Prioritise low-emissions models, and compute via siting and infrastructure requirements, efficiency standards, load shifting, reporting, and incentives.

For AI Foundation Builders, engagement with AI is shaped by strategic sequencing and institutional strengthening. The focus is on deploying AI in ways that advance priority development outcomes, while building the energy, digital, and governance foundations needed for long-term capability. By following a demand-led pathway focused on high-impact use cases, strong governance, and incremental capability building, these countries can avoid premature external lock-in, retain flexibility to deepen their role in the AI ecosystem and demonstrate how AI adoption can advance development goals under tight resource constraints.

5 Conclusion: A Shared Agenda for People, Planet, and Progress

The global expansion of AI presents a unique opportunity to shape a new generation of digital infrastructure, one that is resilient, efficient, and aligned with broader development goals. As AI becomes more deeply embedded across economies, countries have a growing set of choices and tools to ensure that its scale reinforces system reliability, supports resilient growth**, and delivers lasting value.**

Resilience and resource efficiency are increasingly shaping what effective AI scale looks like in practice. Technological advances across the AI stack demonstrate that it is possible to deliver meaningful performance and impact while using fewer resources, operating with greater flexibility, and adapting to local conditions. When AI is designed with these principles in mind, it strengthens the systems it depends on, from energy and infrastructur**e to institutions and markets.**

A central opportunity going ahead lies in adopting a systems approach to AI growth. By recognising the interconnections between AI workloads, energy systems, infrastructure planning and policy frameworks, countries can proactively align decisions across sectors and deploy solutions effectively across the AI stack. This requires (i) deliberate early efforts to set clear strategic priorities for AI, (ii) mechanisms to systematically measure and stress-test AI's real-world impacts during deployment, and (iii) adaptive regulations that support real-world use while ensuring accountability and flexibility, especially, as these technologies continue to evolve. Together, this coordination enables AI to scale in ways that are responsive to local conditions while preserving the ability to adapt as tec**hnologies and markets mature.**

This approach will naturally take different forms across countries, reflecting variation in energy systems, resource endowments, and institutional capacity. Some countries will play a leading role in hosting and advancing large-scale AI infrastructure, demonstrating how efficiency and resilience can be embedded at the frontier, while some will focus on leveraging AI demand to strengthen energy systems or selectively participating in the AI value chain through targeted, fit-for-purpose deployments. Many will prioritise accessing AI-enabled services rather than building extensive infrastructure, shaping demand for efficient, transpar**ent, and resilient solutions.**

The international community will need to come together to coordinate efforts as countries pursue different pathways and roles, develop new insights, establish standards, and support development of resilient AI globally. Important questions remain about how AI's expansion will play out across different development and resource contexts, given how rapidly AI is evolving. Countries are still learning how to align AI deployment with national priorities, design standards that support efficiency and competitiveness, and improve visibility into how AI infrastructure is built and operated. The international community has a critical role in accelerating this collective transition. By creating platforms to share lessons, tools, and best practices, supporting collaborative research and innovation, including on the aggregated impacts of emerging solutions, and advancing common standards and metrics for transparency and reporting, global cooperation can lower barriers and shorten learning curves for all countries. Mobilising finance, technical assistance, and capacity-building will further enable nations to pursue pathways suited to their own contexts while c**ontributing to shared progress. Done well, this collective effort can ensure that AI evolves not as a fragmented or uneven force, but as a resilient and resource-efficient foundation for economic growth, one that genuinely serves a shared agenda o**f **people, planet, and progress.**

In this context, the India AI Impact Summit, and in particular the Working Group on Resilience, Innovation, and Efficiency, seeks to advance these objectives by strengthening international cooperation on resilient AI development. The Working Group focuses on advancing resource-conscious and resilient AI development by promoting best practices, enabling knowledge-sharing, and supporting real-world piloting of resilient AI solutions. By convening governments, industry, and experts around common priorities, the Summit provides an opportunity to build consensus on how AI infrastructure can scale responsibly, transparently, and equitably, ensuring that AI's benefits are realised across diverse development contexts, and that global progress is both tangible and inclusive.

Annexure

Glossary

24/7 Carbon Free Energy (CFE) Matching	Matching every hour of a data centre's electricity consumption with an equivalent volume of carbon-free energy generation within the same local or regional electricity grid, ensuring clean energy supply aligns continuously with demand.
AI Architecture	It refers to the structural design of an AI system, defining how its components and layers are organised and how data flows through them to balance performance, efficiency, and deployment constraints.
AI Stack	It refers to the full set of layers required to design, build, and operate AI systems. This includes AI use cases, model layer, and data centre layer (hardware, compute, and supporting operations like cooling).
Batch Processing Workloads	AI processing that handles large volumes of data together in scheduled or periodic jobs rather than in real time. Batch workloads are commonly used for model training, large-scale data analysis, or offline predictions where immediate results are not required.
Captive renewables	It refers to renewable projects established by companies or industrial facilities primarily to meet their own energy needs, as opposed to conventional renewable solutions that sell power to the grid.
Carbon Usage Effectiveness (CUE)	CUE is a measure of a data centre's carbon efficiency. It is calculated by dividing the CO2 emissions caused by total data centre energy by the energy consumption of IT computing equipment. A lower CUE value means less CO_2 is emitted per IT work done.
Circularity	It refers to practices that optimise resource use and minimise waste across the entire AI life cycle, emphasising sustainability and economic efficiency.
Clustered Data Centre Hubs	Clustered data centre hubs are geographical areas with high concentrations of interconnected data centres, leveraging shared infrastructure for massive data processing, storage, and AI, offering benefits like reduced latency, resilience, and powering digital economies, though straining local power grids.
Compressed Air Energy Storage (CAES)	It refers to compressing air into underground caverns or purpose-built tanks when power is abundant and releasing to drive turbines when needed. This costs lower than batteries for long duration (8+ hours) but requires suitable geology or large above-ground tanks.
Consumer-Grade Devices	It refers to devices designed for everyday users rather than professionals or large organisations. They are typically more affordable, easier to use, and less powerful or customisable than professional or enterprise-grade products.
Edge computing	Edge computing processes data near its source rather than in centralised cloud data centres to reduce latency, bandwidth use, and downtime.
Field-Programmable Gate Arrays (FPGAs)	They are reconfigurable chips that can be customised for specific AI models, reducing the need for additional chips.
Grid-Favourable Sites	Those locations that are close to existing transmission lines or substations, therefore minimising the need for extensive new power infrastructure.

High-Performance Compute (HPC)	It is a technology that uses clusters of powerful processors that work in parallel to process massive, multidimensional data sets and solve complex problems at extremely high speeds.
Industrial internet of things	Industrial IoT is the use of connected sensors, devices, and machines in industrial settings to collect and analyse data for automation, efficiency, and predictive maintenance.
Latency	The time delay between when an input is received by an AI system and when the output is produced. Low latency is critical for real-time applications such as video analysis, robotics, or decision support systems.
Model parameters	They are the internal configuration variables of an AI or machine learning model which control how it processes data and makes predictions. They are the variables that the model learns during training.
Natural Language Processing (NLP)	It is a subset of AI that uses machine learning to teach computers how to understand and work with human language, like text and questions.
Neural Processing Units (NPUs)	They are dedicated AI chips that mimic the functioning of the human brain and have energy-efficient parallel processing capabilities that reduce unnecessary processing.
Power Purchase Agreements (PPAs)	PPAs are multi-year contracts to directly purchase power from renewable energy producers at a fixed price via the grid, providing price stability for buyers while giving renewable energy developers the long-term revenue certainty needed to finance projects.
Power Usage Effectiveness (PUE)	PUE is a measure of data centre energy efficiency. It is calculated by dividing total data centre energy consumption by the amount of energy used specifically for computing tasks. The nearer the quotient to 1, the greater the efficiency, thereby reducing operational costs.
Pumped Hydro Storage	It refers to pumping water uphill into a reservoir when renewables are abundant and releasing it downhill through turbines when needed. It is the cheapest, long-duration (12+ hours) storage technique, but requires specific geography (elevation change, water availability) and large upfront capital.
Renewable Energy Firming Capacity	It denotes the requisite ability within an electrical power system to provide dispatchable, stable energy generation or load management that directly counteracts the inherent intermittency and variability of solar and wind resources.
Thermal Storage	It refers to storing heat (in molten salt, hot water, etc) during the day and using it at night to generate electricity (concentrated solar power with storage) or directly for heating/cooling in buildings. This technique is suitable for solar-heavy grids with storage duration of typically 6–15 hours.
Unbundled Renewable Energy Certificates (RECs)	Unbundled renewable energy certificates (RECs), also known as guarantees of origin or energy attribute certificates (EACs), document the consumption of renewable energy allowing data centre operators to own the associated environmental benefits.
Utilities	Utilities are regulated providers of essential infrastructure services such as electricity, water, heating, that generate, transmit, and distribute these services reliably and at scale to support everyday life and economic activity.
Water Usage Effectiveness (WUE)	WUE is a measure of the efficiency of water usage for cooling in a data centre. It is calculated by dividing total annual water use (litres) by the IT equipment's annual energy consumption (kWh). A lower WUE indicates better water efficiency.

Detailed Solution Set

AI Use Cases

A use case lens should be the first step for any AI build and deployment decision. Aligning on a use case before making investments across the rest of the AI stack is important as it shapes the model size, training approach, and therefore how much computing power and efficiency is required. This keeps AI development focused on clear outcomes and ensures resource are efficiently utilised.

Policymakers could start by identifying priority AI use cases, e.g., improved public services, R&D and innovation, private-sector productivity, high-performance computing (HPC), or inference at scale for citizen or enterprise workflows.[144] E.g., Norway's National Strategy for Artificial Intelligence requires public agencies to consider the potential of AI innovations and value creation particularly in areas where the country has strong business and research communities such as health, energy and maritime. India's Handbook of AI, ML, VR, AR and Robotics Solutions and Roadmap for its Adoption in Electric Utilities outlines power sector use cases such as predictive maintenance of power systems, AI-enabled energy storage systems optimisation and renewable energy forecasting.[145] From these priority use cases, four sets of requirements can guide downstream AI stack decisions:

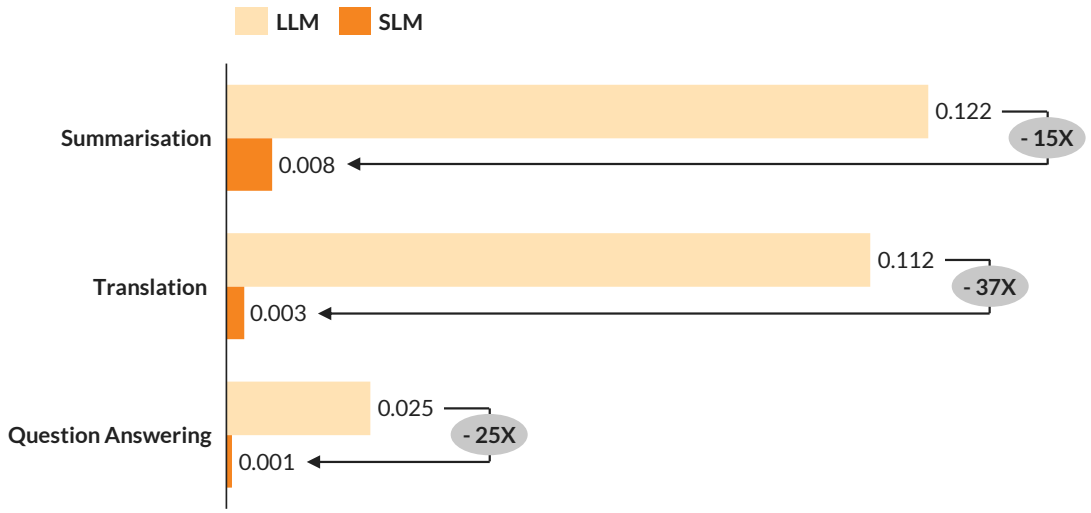
- **Reliability and Latency Needs:** Define the service level required by AI models, i.e, whether applications must be always-on and low latency such as public services for citizens, fraud detection requiring higher uptimes through resilient infrastructure or whether delays and occasional downtime are tolerable such as in research experiments or periodic analytics allowing more flexibility and lower cost.
- **Storage vs Compute Requirement:** Understand whether data centres are primarily used for AI compute workloads, general compute or for other tasks like storing large datasets and archives.
- **Compute Intensity:** Understand whether compute demand is primarily for AI training or inference workloads to determine the frequency and volume of energy required. Also understand whether there are occasional spikes in demand or if it remains steady through the day.
- **Data characteristics and governance requirements:** Identify the level of data sensitivity, localisation requirements and privacy regulations governing AI applications to determine the need for fit-for-purpose models and whether compute should be carried out onshore or outsourced. Use cases that rely on context-specific, highly localised datasets, like agriculture or public administration, often require tailored models and local fine-tuning, while use cases involving highly sensitive data or personally identifiable information, such as AI tools in banking or defence, often require strict data governance or sovereign compute. E.g., Vietnam plans to establish a National Data Centre by 2030 that will integrate all national and sectoral databases for public service delivery, data-driven policymaking, ensuring cybersecurity, etc., effectively reserving a portion of the country's energy, land, and financial resources for sovereign digital infrastructure.[146]

Model Layer

Choices by AI developers in the model layer such as model size and training approach directly shape compute and energy requirements. [147]

Building compact fit-for-purpose models: Fit-for-purpose models perform a narrow range of repetitive tasks and materially reduce computational intensity and carbon emissions. Rather than defaulting to large, general-purpose models, organisations/countries can achieve efficiency gains by designing targeted small language models and workflows optimised for specific tasks, industries, or languages. According to UNESCO, latest research indicates that small models tailored to a specific task can reduce energy use by up to 90%.[148] Small AI is seen as an emerging tool for innovation in low-resource settings, such as AI tutors in Nigeria or diagnostic apps for nurses in South Sudan, that run on minimal data and power and on consumer-grade devices like laptops or smartphones, reducing need for computational power and large-scale data centres.[149,150]

Figure 8: Difference in energy consumption by LLMs vs SLMs by task (Energy kWh/1000 inferences)[152]



Case deep dive: Fit for purpose, small agro-AI tools[152]

Context and solution:

Fit-for-purpose, small AI models are effective in low-resource settings with connectivity, cost, and data availability constraints. Instead of optimising for general-purposes, smaller models are designed to perform strongly on defined tasks, requiring significantly lower compute and can be deployed through simpler infrastructure and lower-bandwidth channels. This enables wider access to AI tools across the Global South. Detailed below are two examples of lightweight agro-AI tools: Agrosavia and Darli AI.

Agrosavia, the Colombian Agricultural Research Corporation, has developed AI-enabled irrigation tools for coffee farms, embedding predictive models and local agronomic data to help optimise irrigation timing and volumes in response to micro-climatic conditions. These models are specifically calibrated for coffee systems, reducing the need for heavy general-purpose compute and energy-intensive analytics, while delivering actionable insights that improve resource efficiency.

Farmline, an agri-tech company in Ghana, developed Darli AI, an interactive voice response tool for farmers in Sub-Saharan Africa, Asia and South America. Available in 27 languages, and accessible via WhatsApp, text or phone, the tool focuses on targeted, task-specific services such as real-time weather updates, crop disease detection support, and best practices in regenerative farming.

Enablers for success:

1. Access to localised, smaller, high-quality data sets for on micro-climatic conditions, crop-specific agronomy and regional farming practices enables models to remain small without compromising on accuracy and specificity.
2. Supportive local institutions and partnerships with farming communities enable access to decades of agricultural research and stakeholder networks.
3. Continuous feedback and learning loops through field testing and ongoing user input allows models to be iteratively improved and is critical for maintaining effectiveness as farmer needs and climate conditions evolve.

4. Infrastructure-appropriate design and delivery channels ensures accessibility for end-users by aligning deployment of solutions with constraints on devices and connectivity, including mediums available to them such as IVR, SMS, and WhatsApp. [153]

Impact:

Small agro-AI models reduce the reliance on large, centralised data centres due to lower compute requirement. In turn, the resulting tools improve farming practices, water-use efficiency and enhance climate resilience across the agricultural value chain. AI-driven irrigation tools and systems alone can reduce water use by 30-50%, while improving crop yields by 20-30%.[154] Moreover, with its lower compute needs, these tools can thus run in limited connectivity areas.

Summary takeaway:

1. Fit-for-purpose models and smaller, compact models can deliver the same performance as generalised LLMs for specific tasks when designed around local realities of connectivity and device constraints, and supported by appropriate systems building with local partnerships and access to localised datasets.
2. Sustained successful delivery within communities requires strong on-ground relationships and mechanisms to incorporate feedback and iterate on the end AI-enabled solution.

- **Model optimisation:** The compute intensity of AI models can be reduced through smarter model training and model compression. Smarter training lower training energy use by relying on smaller datasets or more efficient approaches such as knowledge distillation. Accenture Labs found that training models on just 70% of the full dataset led to less than a 1% reduction in accuracy, while reducing energy consumption by 47%.[155] Knowledge distillation trains a smaller 'student' model to replicate the behaviour and reasoning of a larger, more computationally intensive 'teacher' model.[156] Model compression techniques such as pruning or quantisation can then be applied to reduce inference energy requirement. Pruning reduces unnecessary elements in neural networks, thereby reducing computational complexity and quantisation reduces the numerical precision of computations, conserving ~33% energy without significantly compromising on accuracy.[157]

Case deep dive: Carbon reduction through smart training methods[158]

Context and solution:

As AI adoption scales, carbon footprint of model training has emerged as a critical constraint. As a result, model efficiency has become a critical lever for reducing carbon emissions without sacrificing performance. Smarter training and compression techniques reduce energy use during training and deployment, lowering overall compute intensity. DistilBERT, an NLP model developed by Hugging Face, is a leading example of this.[159]

It was created using knowledge distillation which allows a smaller model ("student") to mimic a larger, more complex one ("teacher"). In this case, the "teacher" was BERT (Bidirectional Encoder Representations from Transformers), a powerful but energy-intensive language model developed by Google. DistilBERT keeps BERT's core language understanding while removing excess complexity, resulting in a lighter, faster model with comparable performance.

Enablers for success:

1. Access to a strong foundational "teacher" model (BERT) provides a high-performance reference point, enabling optimisation efforts to focus on efficiency gains rather than rebuilding the model from scratch, reducing training-related emissions.
2. Clear, efficiency-first design goals to shape training decisions, for example reducing parameters, inference latency, and compute cost. This ensures optimisation is focused on energy efficiency and reducing compute-intensity, rather than outperforming previous models.
3. Open-source access to distillation and compression techniques enables models to continually be iterated upon by a global community and reduces duplicate compression efforts that drive emissions.

Impact:

Through this training method, DistilBERT manages to reduce the BERT model's size by 40%, while preserving 97% of its capacity to understand language and running 60% faster. DistilBERT also produced ~47% less CO_2 during training than BERT, demonstrating substantial reduction in carbon footprint without sacrificing model performance.

Summary takeaway:

Smarter training and model compression enable energy-efficient training and inference by leveraging existing foundational models that can transfer mature capability to newer "student" models. By treating efficiency as a first-order model design requirement, developers can reduce computational intensity and associated emissions without sacrificing accuracy.

- **Edge-Deployable :** For use cases requiring low latency or local data processing (e.g., computer-vision models for video analytics, AI-driven industrial control and robotics, or decision systems supporting transport, energy, or public safety), AI models can be designed to run directly on end-user devices.[160] In practice, this means deploying lightweight, optimised models that perform inference locally on consumer-grade hardware, such as smartphones, laptops, or embedded devices with GPUs or NPUs. In regions such as Asia-Pacific, large cloud data centres are typically used for model training and batch processing, while real-time inference is handled by models running on local devices. Shifting latency-critical inference to edge-deployable models reduces data transfer and network congestion, while also limiting reliance on large, centralised data centres in areas constrained by energy, land, or water availability.

Case deep dive: Edge AI[161,162]

Context and solution:

Cloud-centric models and cloud-based inference requires reliable and sustained data transmission to centralised data centres. This creates significant challenges of high latency and unreliable performance in low-resource and connectivity constrained regions. Edge AI provides an effective and resilient alternative, by allowing AI models to run inference tasks on local devices instead of transmitting data to the cloud for computation.

An example of this is PlantVillage's Nuru app which embeds edge AI directly onto Kenyan farmers' mobile devices, to diagnose crop pest and disease from leaf photos, even in offline environments. The app uses machine learning to recognise symptoms of diseases and delivers actionable management advice locally

on the device. Nuru uses a deep learning object detection model developed at Penn State University in partnership with the International Institute of Tropical Agriculture and the United Nations Food and Agricultural Organisation, trained on the world's largest open access library of crop health knowledge.

Enablers for success:

1. Continued development of architectures and optimisation approaches that reduce model size and increase efficiency, enabling companies to build solutions that are deployable on low-powered devices.[163]
2. Open-access to high-quality, domain-specific datasets (e.g., crop disease images) enables targeted model training for accuracy, avoids duplication of data collection, and lowers barrier of entry for companies to develop specialised tools.
3. Partnerships between universities, international agencies, private sector companies, and local institutions combine technical expertise and advancements with local knowledge, supporting effective training, validation and dissemination of edge tools.
4. Procurement processes that specify design requirements such as offline functionality, low-powered device compatibility can advance scaled edge AI services and drive innovation.

Impact:

While models are still trained in the cloud, shifting inference processes locally substantially reduces data transmission and cloud compute requirements. Reducing compute and data usage leads to lower energy demands, while improving latency and reliability in low-connectivity settings. Tools like Nuru are then able to scale deployment and accessibility of reliable and resilient AI assistance to rural communities in Sub-Saharan Africa.

Summary takeaway:

Edge-deployable AI models are effective in low-connectivity contexts where real-time support is required. Separating cloud-based training from local, on-device inference reduces energy use and data centre infrastructure demands while enabling real-time, context-specific AI support at the edge. Therefore, these edge models can significantly improve accessibility, reliability, and resilience of AI deployment.

Data Centre Layer:

The data centre layer translates compute requirements into physical infrastructure decisions, including where data centres are sited and how they are built and operated. Design choices around location, hardware, compute efficiency and supporting cooling and power systems play a decisive role in managing the environmental footprint of AI at scale.[164]

1.1. Data centre set-up:

It is critical to be mindful of where data centres are located and how they are built to ensure they do not strain existing power and water resources in their local community and contribute to circularity.

- **Siting decision based on clean energy and water availability:** Data centre siting decisions should prioritise proximity to clean energy sources and low water stress regions. Locating facilities near renewable generation and storage reduces carbon intensity, while water-stress screening ensures data centres are built in regions with low water scarcity and access to non-potable sources.[165] Innovative approaches, such as closed-loop systems in Scotland that use proximity to wastewater management

for cooling, demonstrate how siting choices can improve efficiency outcomes.[166] Governments also encourage mindful siting decisions such as Finland's National Roadmap for Data Centres prioritises data centres that support the power system by being in grid-favourable sites close to power generation hence reducing the need for new grid build. North Holland's Data Centre Strategy restricts the development of new data centres to designated industrial zones to mitigate concentration in urban areas and promote balanced geographical distribution.[167]

Case deep dive: World's largest solar co-located data centre[168,169]

Context and solution:

While efficiencies in operating data centres can reduce overall energy use, the carbon footprint of data centres is ultimately determined by the energy mix powering the data centre. Data centres co-located with renewable generation sites like solar or wind farms are better able to leverage low-carbon energy sources. Moro Hub, a subsidiary of the digital arm of the Dubai Electricity and Water Authority (DEWA), tackles AI's rising emissions by co-locating its green data centre directly at the Mohammed bin Rashid Al Maktoum Solar Park, the world's largest single-site solar park, ensuring the data centre is entirely powered by renewable energy.

The initiative is delivered in collaboration with technology partners like Huawei, Microsoft, Dell Technologies, SAP, and Intel. Advanced data centre designs, AI-enabled energy management, and data centre infrastructure management (DICM) systems are used to optimise cooling and operational efficiency, thus reducing resource wastage and overall carbon footprint.

Enablers for success:

1. Strong government commitment to and clear alignment of dual goals of sovereign AI and energy transition, ensuring long-term stable policy support and concessional land and capital to incentivise resilient infrastructure projects.
2. Strong partnerships between renewable energy authorities, digital infrastructure developers, and technology partners to improve knowledge sharing on grid developments and data centre demands.
3. Streamlined zoning and development processes for renewable projects to leverage geographical advantages of high renewable energy sites, increasing opportunities to channel low-carbon energy into digital growth.

Impact:

The result of this project is a resilient digital infrastructure that significantly reduces emissions while supporting high-performance cloud, AI, IoT, and cybersecurity services. With planned capacity exceeding 100 MW, this project will serve as the world's largest solar-powered data centre, with the potential to reduce over 10,500 tons of CO_2 annually.

Summary takeaway:

The strategic co-location of data centres with renewable infrastructure is best used for large, always-on compute loads in regions with strong clean energy potential. Low-carbon energy procurement from nearby facilities is cost-effective and maintains high-performance compute, leading to scaling of resilient AI. Integrating energy system planning with digital infrastructure development enables governments to progress dual goals of digital infrastructure and renewable energy growth.

- **Micro and modular data centres:** Micro and modular data centres are small, self-contained compute facilities that can be deployed quickly near users or data sources. These AI-optimised modular data centres integrate all necessary infrastructure, including power, cooling, remote monitoring, fire suppression, and racks that support the latest GPUs, into a single, portable unit and can sometimes be stacked when compute needs increase. In practice, this involves running trained models on localised AI compute (e.g., GPU- or NPU-enabled edge nodes) sited at telecom exchanges, factories, hospitals, or transport hubs. These small-scale data centres can reduce systems-level energy use and emissions by avoiding data movement and enabling local processing. For example, Kenyan utility KenGen has set up a 52kW modular data centre, powered by renewable energy batteries, that has the potential to be scaled based on local needs allowing compute in energy-constrained contexts.[170] EdgeUno, a network infrastructure and edge cloud company, in Latin America has set up edge data centres in more than 47 locations including Chile, Argentina, Costa Rica, etc to deliver low-latency services closer to users while regional facilities handle batch workloads.[171]

Case deep dive: Vigyan Labs' "AI-in-a-box" solutions for micro data centres[172]

Context and solution:

As AI adoption increases, traditional large-scale data centres create pressure on energy systems, water resources, and environmental impact from infrastructure development. This exacerbates issues in countries with already constrained electricity grids or water scarce settings. Smaller micro data centre designs reduce environmental impact by integrating with green construction principles and operational plans optimised with energy reduction in mind.

Vigyan Labs developed FEMTO, an "AI-in-a-Box" micro data centre that reimagines enterprise AI infrastructure as a modular, self-contained unit. Rather than requiring dedicated facilities, FEMTO delivers a complete data center, including compute, storage, and AI capabilities, in a compact 4U form factor (just 7 inches of rack height) that can be deployed in standard office environments. The system provides 4000 AI TOPS processing power and supports enterprise-grade LLMs directly on-premises, ensuring data sovereignty and eliminating dependency on cloud providers. Its plug-and-play design enables setup in minutes, with edge-ready architecture capable of handling demanding AI workloads while maintaining silent operation and minimal space requirements.

Enablers for success:

1. Policy and regulations that support and incentivise innovations in smaller data centres, resulting in greater energy and water efficiency.
2. Right-sizing AI needs to forecast compute demand to determine whether micro data centres are sufficiently powerful compared to traditional, large-scale data centres.
3. Incentives for CPU-first approaches that prioritise cost and energy efficiency, CPU-based hardware, AI-optimised operations, and rooftop solar, all to minimize power consumption.

Impact:

FEMTO's AI-in-a-Box design delivers complete data center capabilities in just 4U of rack space, enabling organizations to deploy enterprise-grade LLMs on-premises with 99.9% uptime while maintaining full data sovereignty. By eliminating the need for dedicated facilities and enabling deployment in standard office environments, the solution makes high-performance AI infrastructure accessible to organizations and geographies traditionally excluded from AI deployment.

Summary takeaway:

Right-sized, AI-optimised micro data centres are well suited to resource-constrained and distributed environments where hyperscale facilities are impractical or too energy intensive. By combining renewable energy, efficient CPU-based hardware, and local deployment in compact, modular facilities, this approach helps deliver scalable AI infrastructure that dramatically reduces energy use while supporting digital growth.

- **Reducing building and construction footprint:** Energy use and emissions can be lowered through more efficient building design and construction practices—from lighting and equipment choices to construction methods. Minimising construction waste, increasing material recycling, and ensuring responsible waste management can further support circularity.[173,174]

Case deep dive: Low-carbon concrete in data centre construction[175]

Context and solution:

Embodied carbon from construction materials, especially concrete, is a major and often overlooked source of emissions in data centre buildouts. Applying green design principles in the construction phase, such as using alternative construction materials, can reduce carbon footprint from the set-up phase.

To address this, Meta has re-engineered its data centre design and materials selection strategy to reduce the carbon footprint of concrete used in new set-ups. This includes eliminating concrete in non-essential applications, and adopting low-carbon concrete mixes with fly ash and slag that cut emissions to below regional baselines. The innovation extends to AI-optimised concrete formulations, developed with the University of Illinois Urbana-Champaign and industry partners, that accelerate discovery of mixes with improved sustainability, strength, and cure performance, reducing embodied carbon in lab settings.

Meta has integrated these approaches into data centre projects such as its Rosemount (Minnesota) and DeKalb (Illinois) campuses and collaborates with consortiums like iMasons Climate Accord to scale adoption of low-carbon concrete industry-wide.

Enablers for success:

1. Infrastructure developers to integrate carbon emissions targets into building design processes, which incentivises developing and implementing resilient solutions (e.g., low-cement usage and low-carbon mixes) as a first-order objective.
2. Procurement processes for new data centres that reward low-emissions design proposals, ensuring the uptake of solutions as digital infrastructure grows.
3. Partnerships with universities, research labs, and industry consortiums that can accelerate development and uptake of low-carbon cement mixes, and ensure access and supply for alternative materials.

Impact:

Design-streamlining and low-carbon concrete mixes can reduce the carbon footprint of concrete by over 30%, compared to previous designs, and by up to 20% below regional industry baselines respectively.

Summary takeaway:

Low-carbon concrete strategies are most impactful in large-scale data centre buildouts where construction emissions are significant to climate impacts. Integrating emissions into design and procurement processes help inform resilient supply-chain decisions from the outset. Resilient data centre construction at scale requires pairing material innovation with practical deployment pathways and access to suppliers of alternative materials to achieve low-carbon construction at scale.

1.2. Hardware Layer:

It is important to consider hardware choices carefully, since they directly influence power consumption, cooling requirements, and the overall footprint of AI infrastructure. This includes adopting efficient, task-appropriate architectures where suitable and embedding circularity across hardware lifecycles to manage rapid obsolescence and reduce e-waste.

- **Energy-efficient AI accelerators:** As AI workloads grow, innovations in specialised AI hardware are making it possible to efficiently handle parallel computations with far lower energy use. Neural Processing Units (NPUs) are dedicated AI chips that mimic the functioning of the human brain and have energy-efficient parallel processing capabilities that reduce unnecessary processing.[176] Field-Programmable Gate Arrays (FPGAs) are reconfigurable chips that can be customised for specific AI models, reducing the need for additional chips.[177] For example, Google's Tensor Processing Units (TPUs) are custom-designed AI accelerators optimised for training large deep learning models.[178]

- **Capping energy consumption by hardware:** Energy-efficient hardware techniques cap power usage to reduce overall energy consumption while preserving performance and lowering cooling requirements. For example, MIT Lincoln Laboratory researchers have developed power-capping approaches that reduce energy use by 12–15% while increasing time-to-result by only ~3%, and NVIDIA GPUs are designed with built-in power limits that allow data centres to cap power draw with minimal performance impact.[179,180]

- **Incorporating circularity in hardware management:** As computing devices typically have lifespans of two to five years and are frequently replaced with the most up-to-date versions, it is important to inculcate circularity for resilience & sustainability. Strategies such as designing hardware with longer lifecycles that is easy to upgrade and recycle, alongside refurbishing and reusing components, could reduce e-waste generation by up to 86%. Companies are beginning to recover, resell, and reuse hardware through collaborative platforms as part of a broader circular economy.[181]

1.3. Compute Layer:

It is critical to be mindful of how AI workloads are executed across training and inference, as operational choices a can significantly influence energy demand and the ability to align compute with cleaner power. This includes improving utilisation and efficiency through workload optimisation (e.g. scheduling flexible tasks across time or locations), cloud compute and CPU-first model deployment platforms.

- **Dynamic workload scheduling for green compute:** Improving utilisation through smarter workload management by shifting AI computations across time periods or geographies to align with periods of grid stability or peak renewable energy availability.[182] Non-critical AI workloads are scheduled to time periods with better grid stability or assigned to data centres in geographical zones with greater renewable capacity, enabling cheaper and higher use of clean power.[183]

Case deep dive: Green compute through carbon forecasting[184,185]

Context and solution:

Always-on data centres run AI workloads 24/7, irrespective of compute urgency, even during carbon-intense grid hours, requiring fossil fuel plants to operate longer to meet the constant energy requirements of data centres and increasing emissions. To tackle this massive footprint, Google utilises a carbon-intelligent computing platform that aligns its data centre operations with the availability of low-carbon electricity: working harder when the sun shines and winds blow and shifting flexible compute to times and places where clean energy is abundant.

This system uses hourly grid carbon forecasts from Electricity Maps together with internal demand predictions to reschedule non-urgent compute tasks (like photo indexing or video processing), enabling workloads to shift across its global data centre fleet to where electricity emissions are lowest to reduce overall footprint without impacting service reliability.

Enablers for success:

1. Hyperscalers and large private companies can distribute data infrastructure across regions globally to enable flexibility in distributing workloads in response to clean energy demands.
2. Embed flexibility into data centre workload design to more efficiently prioritise tasks by urgency and to better enable load shifting around core services.
3. Increase availability and access to hourly grid carbon-intensity data from electricity providers across regions to equip data centre operators with real-time grid data to inform smart compute scheduling platforms.

Impact:

This carbon-aware load shifting within and across data centres reduces grid-level CO2 emissions. Targeting non-urgent tasks with lower impacts on users maintains service reliability, while increasing resilience & sustainability.

Summary takeaway:

Carbon-aware compute scheduling is best used in large, multi-site data centre fleets with flexible workloads. Awareness of carbon-intensity of grid electricity and internal classification of urgent and non-urgent compute tasks enables data centre operators to optimise workloads against clean energy availability, reducing overall carbon footprint.

- **Cloud based computing through virtualisation:** Traditional data centres often run servers at just 10-15% utilisation. Shifting computing workloads to the cloud through virtualisation reduces emissions by allowing multiple virtual servers to run on a single physical server, consolidating workloads and improving resource usage. Through virtualisation and consolidation, major organisations like Citigroup have improved utilisation to around 50%, delivering immediate energy savings while reducing facility requirements.[186] Cloud data centres are custom-designed facilities optimised for existing hardware usage and energy efficiency resulting in 1.4x–2x lower energy consumption compared to on-premise computing.[187]

Case deep dive: Carbon-aware cloud computing[188]

Context and solution:

Cloud computing consolidates workloads through virtualisation, reducing emissions by increasing individual server utilisation and lowering number of physical machines required. While these consume less electricity in data centres compared to traditional servers, it still draws from carbon-heavy grids, leading to significant emissions. Carbon-aware cloud computing builds on the efficiency gains of virtualisation with optimised timing of consolidated workloads.

Cloud computing platforms like GreenPow embed proprietary scheduling algorithms into its cloud infrastructure to decide when and where AI workloads should run for the lowest emissions while still meeting latency and data-sovereignty constraints.

The system continuously evaluates renewable availability at data centre locations, regional carbon-intensity signals, time-of-day electricity prices, and workload constraints, then automatically routes compute tasks to locations or hours that produce the lowest carbon footprint with maximum uptime, ensuring virtualised workloads are executed when electricity is least emissions-intensive.

Enablers for success:

1. Multi-region cloud infrastructure with diverse grid energy-mix enables routing workloads to lower-emission locations while being mindful of latency and data sovereignty requirements.
2. Continuous access to granular data on renewable availability, regional carbon-intensity signals and time-of-day electricity prices supports real-time optimisation of workload placement.

Impact:

This routing enabled an emission reduction of up to 60% per task compared to traditional cloud execution, and ~30% reduction in energy costs especially during green peak hours.

Summary takeaway:

Cloud computing platforms can consider incorporating carbon-aware workload shifting in accordance with low-carbon periods and geographies by evaluating renewable availability and energy prices hourly. This approach enables reduction in emissions and energy costs of compute.

- **CPU-first inference platforms:** Specialised AI-hardware like GPUs are often power-hungry and costly to procure. A CPU-first inference architecture offers an alternative pathway by optimising AI models to run efficiently on widely available CPU servers without compromising accuracy.

1.4. Supporting systems (i.e., Cooling, Heat Reuse):

Outside computing and hardware efficiency, the tertiary operations of a data centre can be made more resource-efficient by leveraging better cooling systems, reusing excess heat and measuring energy efficiency.

- **Efficient cooling systems and thermal aware design:** Liquid cooling can be employed where high rack densities and local climate conditions justify it. Direct-to-chip liquid cooling involves circulating a coolant directly through microchannels or cold plates attached to the heat-generating components while immersion cooling submerges servers or electronic components in a dielectric fluid that ab-

sorbs and transfers heat away.[189] These systems can improve PUE by 1.5-1.7x compared to traditional air-cooling systems.[190,191] Microsoft is currently piloting chip-level zero-water evaporated cooling technologies in the US which reduce wastage and improve WUE metrics by 39%.[192] In this closed loop system, water is filled once during construction and continually circulated between the servers and chillers to dissipate heat without requiring a fresh water supply. Another innovative cooling system is used by Google in its Hamina data centre leveraging seawater from the nearby Bay of Finland that is pumped from the sea and transported through underground granite tunnels. The cold water from the sea is used for cooling servers and then is released back to the sea, reaching its original temperature again.[193]

Case deep dive: AI for water-conscious AI[194]

Context and solution:

Data centre cooling is a major driver of energy and water consumption. Improving cooling efficiency without compromising uptime is therefore a key lever for reducing operational emissions. Increasing real-time responses to environmental conditions of data centres, such as temperature and humidity, can increase optimise cooling needs.

ST Telemedia Global Data Centres has partnered with AI specialist Phaidra to pilot an AI-driven autonomous control system across its Singapore facilities. The system continuously analyses thousands of real-time sensor trends to optimise cooling performance and energy efficiency, particularly within hybrid set-ups combining air and liquid cooling technologies, as thermal and workload conditions change.

Enablers for success:

1. Innovations and development of high-quality sensors to provide real-time information on temperature, flow rates, and other environmental conditions that can be integrated into control systems
2. Strong collaboration between data centre operators and AI specialists ensure deployment of innovation solutions with robust guardrails and iterative tuning.
3. Development of hybrid cooling systems provide control systems with multiple levers for optimisation and provides AI and autonomous systems to optimise for efficiency depending on different environmental factors

Impact:

Initial estimates from the pilot suggest possible cooling energy savings of ~10%, with potential to reach up to 30% as the AI learns more operational data.

Summary takeaway:

Cooling optimisation can be deployed in data centres with significant cooling energy demands and in areas of increasing water stress. By intelligently adjusting cooling operations in response to workload and environmental conditions, AI-enabled systems maintain reliability for high-performance computing workloads, while reducing water stress levels.

- **Waste-heat reuse:** The excess heat dissipated in the data centres can be recycled to internal or external energy systems like industrial facilities, building or district heating grids for utilisation, improving overall system efficiency. This has been achieved in-practice by Bahnhof in Sweden to pump the excess heat to nearby district buildings.[195] Another example is of the UK government providing 65 million pounds to five innovative green heating projects that used waste heat from nearby data centres to warm ~10,000 homes.

Case deep dive: Innovative heat reuse[196,197]

Context and solution:

While data centres consume high levels of energy, they also generate large volumes of low-grade waste heat that is vented and lost, leading to significant system inefficiencies. This can create opportunity for wasted heat supply to respond to unmet clean heat demands or contribute to reducing grid demands.

The Tallaght District Heating Scheme addresses a core challenge in Ireland's energy system: the country still relies heavily on imported fossil fuels to meet most of its heating demand. To confront this, a pioneering collaboration was formed between South Dublin County Council, its energy agency Codema, and Amazon Web Services (providing waste heat). This network captures waste heat from a nearby Amazon data centre and uses advanced heat-pump and insulated pipe network technology to distribute it to connected buildings. This scheme converts a commonly wasted by-product into a reliable, low-carbon heat source for public use.

Enablers for success:

1. Public sector leadership and stewardship of strategic collaborations between power utilities, government agencies and private players aligns planning, regulation and investment, de-risking first-of-a-kind district heating schemes.
2. Physical co-location of the data centre with residential or commercial areas with dense heat demand clusters makes waste-heat recovery viable and minimises distribution losses.
3. Strategic planning for stable heat offtake to civic buildings and social housing de-risks investment in waste-heat recovery initiatives and creates opportunity for greater crowding in over time.

Impact:

Since its launch in 2023, TDHS has supplied heat to public assets including South Dublin County Hall, Tallaght County Library, the TU Dublin Tallaght campus, and will soon service 133 affordable apartments, cumulatively delivering nearly 3,700 MWh of heat and saving over 1,100 tonnes of CO_2 to date.[198] This innovative heating approach makes TDHS the only district heating project in Ireland or the UK using data centre waste heat.

Summary takeaway:

Waste-heat reuse becomes scalable when a stable, high-grade heat source (such as a data centre) is paired with proven infrastructure (heat pumps plus insulated networks) and a clear anchor-load strategy to guarantee demand. This approach cuts fossil fuel reliance in heating and repurposes wasted heat byproduct to improve overall energy system resilience.

- **Measuring and tracking energy efficiency:** Robust measurement of energy used, and emissions generated is essential to improving data centre efficiency. Key metrics include power usage effectiveness (PUE), carbon usage effectiveness (CUE), and water usage effectiveness (WUE), which track energy, carbon, and water intensity respectively.[199] Continuous monitoring of these metrics enables operators to benchmark performance, identify inefficiencies, and quantify the impact of design and operational improvements over time. E.g., Amsterdam Duurzaam Digital Policy emphasises energy efficiency, circular economy principles, and residual heat utilisation, mandating that new data centres achieve a PUE of 1.2 or less.

Case deep dive: Data centre infrastructure management (DCIM) to track and measure emissions[200,201]

Context and solution:

As data centres become more energy and water intensive, the tracking and measurement of efficiency alone is insufficient if not operationalised. Metrics such as PUE and WUE provide benchmarks, operators need real-time and granular data to identify inefficiencies and make appropriate adjustments. DCIM's address this gap be creating an integrated platform for monitoring and tracking, while also controlling power and cooling needs of data centres.

Providers such as Sunbird can optimise cooling operations by controlling airflow and temperature against the compute and workload demands. Identifying overcooling or other inefficiencies, the DCIM system is able to translate data in operational changes.

Enablers for success:

1. Standard efficiency metrics can streamline development of management systems and enabling more targeted controls to reach efficiency benchmarks.
2. Regulatory measures that reward transparency and increased efficiencies that incentivises operationalising efficiency gains, rather than keeping as a one-off compliance threshold

Impact:

Enabling precise monitoring and automated optimisation of cooling and power systems, DCIM platforms report cases of up to 90% reductions in cooling power requirements. These improvements directly reduce energy use, operational costs, and decreases overall emissions in line with regulatory standards.

Summary takeaway:

Robust measurement is key to data centre resilience, both for regulatory compliance, but as operational data for real-time monitoring and control systems. DCIM tools are not only able to track efficiency metrics but improve efficiencies with automated power and cooling adjustments.

Grid Integration and Energy Transition

Resilient energy infrastructure is a key enabler of the AI ecosystem as computing and inference needs scale rapidly. This includes both sourcing power from cleaner energy sources and smarter integration with existing grid systems, optimising when and how data centres consume electricity while accelerating the transition to low-carbon supply.

- **Smart grid integration through demand-responsive operations:** Demand-response programmes allows data centres to temporarily reduce load (by up to 20%) during periods of peak demand, easing strain on systems that are still primarily powered by fossil fuels.[202] E.g., Google's 24/7 carbon-free energy (CFE) ambition includes demand-side solutions that shift non-urgent compute tasks, such as YouTube video processing or ML workloads, to periods when the grid is less constrained and less emissions-intensive.[203]

Case deep dive: Data centres as grid flexible assets[204,205]

Context and solution:

Data centres are traditionally seen as passive, high-demand power consumers, contributing to grid stress and delaying new interconnections amid rising AI workloads and electrification. However, a share of data centre demand is flexible and can be modulated in response to grid conditions to support electricity grids rather than stress it.

Verrus is redefining this model by designing data centres as active grid assets capable of rapid demand response and flexibility. In collaboration with the National Renewable Energy Laboratory (NREL), Verrus successfully demonstrated its proprietary power flow management, battery energy storage systems (BESS), and grid-aware controls on a 70 MW test platform. The enables the data centre to not only shift workloads in response to grid conditions, but to operate independently in periods of high grid stress.

Enablers for success:

1. Collaborative co-designing of electrical and AI infrastructure ensures data flow to ensure grids take advantage of data centre workload flexibility, while also reassuring data centre operators of grid activity data to forecast reliability.
2. Collaboration with independent research institutions such as NREL validates performance at scale, de-risking the solution among utilities, regulators and investors.
3. Foster development of battery storage technology to increase scale and position of data centres as grid assets.

Impact:

The solution enables data centres to curtail load up to 100% within one minute during grid stress scenarios, transition seamlessly to battery-powered islanded operation, and provide grid support without compromising compute operations. This technology allows data centres to carry out demand response operations, helping grid operators balance peak loads.

Summary takeaway:

Grid-flexible data centres are particularly essential in regions facing grid stress. They work by combining battery storage, power flow control and grid-aware controls to modulate demand. This is essential to turn data centres from passive, high-demand loads into active grid assets that support reliability.

- **Scale renewables over time; in the near term, bridge gaps with captive renewables and storage to meet on-site demand where clean grid power is unavailable:** Renewables, primarily wind, solar PV, and hydro, currently supply about 27% of global data centre electricity consumption, a share projected to rise to 50% by 2030 supported by power purchase agreements (PPAs), unbundled RECs and emerging 24/7 CFE matching approaches that better align consumption with low-carbon supply.[206,207] Hyperscalers are signing long-term clean-power contracts, investing directly in generation assets, and financing early-stage nuclear and geothermal projects.[208] Amazon, for example, has finance more than 500 solar and wind projects globally, making it the world's largest corporate buyer of renewable energy in 2024.[209] Additionally, in regions where grids remain carbon-heavy or clean energy is constrained, data centres can rely on on-site/ near-site clean captive power. E.g., Ireland's Commission of Regulation Utilities (CRU) requires new data centres to provide onsite or local generation and/or storage capacity to match the requested data centre demand capacity.[210]

Case deep dive: PPAs as an enabler of low-carbon AI growth[211]

Context and solution:

Countries with carbon-intensive electricity grids are unable to deliver renewable power to match the growth of digital markets and the required data centre expansion. Power purchase agreements (PPAs) can bridge the gap by enabling data centre operators to directly contribute to renewable energy development to generate stable and resilient energy instead of competing with limited existing supply.

In Malaysia, Google and TotalEnergies signed a 21-year PPA under which TotalEnergies will supply 1 TWh of certified renewable electricity from the Citra Energies solar plant to support Google's data centre operations in the region. The agreement sits within Malaysia's Corporate Green Power Programme (CGPP), which is designed to enable direct corporate–developer agreements and emphasises additionality by creating new renewable generation for corporate demand rather than reallocating existing supply. This enables Google to scale its capacity with a low-carbon approach without waiting for decarbonisation of the existing grid.

Enablers for success:

1. Government policies and programmes that foster new corporate-development agreements and ensuring defined regulatory mechanisms for agreement to be operationalised.
2. Defined national roadmap for digital growth and renewable energy development can provide private sector with direction on solutioning that responds to growth demands.
3. Stable and mature business environment attractive for hyperscalers to expand operations in, while also lowering risk for hyperscalers and energy developers to enter.
4. Abundant renewal energy sites for data centre developers and energy providers to site renewable energy projects and co-locate data centre nearby.

Impact:

The PPA provides Google with long-term price and supply certainty and a credible pathway to align expansion with its 24/7 carbon-free energy ambitions, while TotalEnergies secures a long-duration offtake that de-risks investment. The arrangement demonstrates how tailored power solutions enable hyperscalers to scale operations, even in emerging markets, while advancing renewable energy development goals.

Summary takeaway:

Power purchase agreements lock-in long-term offtake between data centre operators and renewable developers, providing price and supply certainty to reduce carbon emissions in a cost-effective manner. Therefore, PPAs can align data centre expansion with low-carbon pathways by enabling timely investment in new renewable capacity, rather than relying solely on slower grid upgrades.

Level of Adoption of Resilient and Resource-Conscious AI Solutions

Adoption of the technical and operational solutions outlined above varies greatly, primarily driven by differences in investment capacity of stakeholders, grid and infrastructure readiness, and enabling policy and regulatory environments. These solutions are assessed and categorised from low to high adoption, based on synthesis of real-world implementation, industry and policy reports, academic literature, and press releases, which together provide insight into the degree of uptake by companies and governments. The mapping is intended to indicate broad direction and relative levels of adoption, not to give a precise or exhaustive quantitative assessment. Additionally, the mapping does not comment on the technological maturity of solutions.

Figure 9: Illustrative spectrum of resilient AI solutions[210]

Level of adoption of sustainable AI solutions

LOW → HIGH

Some solutions are currently emerging, limited to theoretical evidence and small-scale pilots in certain geographies

- **Circular hardware management** platforms for AI remain nascent, and only 22% of AI e-waste is being formally collected and recycled[1].
- **Low-carbon construction** is not widely adopted due to inconsistent carbon-aware procurement policies and only 14% of data centres worldwide are LEED-certified[2,3,4].
- **Dynamic workload scheduling** is in early-stages as few operators have access to grid carbon-intensity or large fleets to shift compute[5].
- **Waste-heat recovery** is largely limited to few district heating networks in Nordic and other temperate countries[6].
- **Smart grid integration** is in early-stages as most data centres have high uptime requirements for always-on compute[7].

Many solutions are scaling beyond pilots, but efficiency gains remain concentrated amongst early-movers

- **AI use-cases (government and private sector)** are being increasingly included in national AI strategies and tailored model development[8].
- **Fit-for-purpose small models** are being recognized for their accuracy but large, general-purpose models dominate AI development[9].
- **Micro and modular data centres** are projected to grow 1.6x faster than traditional data centres through 2030, but their adoption beyond low-resource contexts remains limited[10].
- **Power capping** in new accelerators is gradually being introduced but limited to avoid performance trade-offs.
- **Energy-efficient AI accelerators** are being developed but adoption is low with GPUs accounting for 60% revenue share in 2024 due to their greater versatility[11].
- **Resource-aware data centre-siting** decisions are largely driven by energy prices or government requirements, not water or clean power considerations[12].
- **Measuring and tracking energy efficiency** is limited and not disclosed due to few reporting and performance mandates with a survey finding that 76% of data centres track PUE while 43% track water usage[13].

Commercially proven solutions are being implemented by small companies, hyperscalers, and across geographies

- **Efficient model optimisation and compression** innovations are developing from both industry and research organisations, with increasing adoption particularly in low-resource environments[14].
- **Edge deployable models** are becoming rapidly more mainstream, driven by demand for real-time AI and energy efficiency; forecasts suggest by 2030, 89% of systems sold will be edge systems[15].
- **Cloud computing through virtualization** is widely adopted and invested, with companies investing significantly in cloud infrastructure and countries like Malaysia developing national AI and cloud computing policies[16,17].
- **Efficient cooling systems and thermal aware design** are codified in data centre best practices and regulations like the EU Code of Conduct for Data Centres, Singapore's Green Data Centre Roadmap, and to be included in the forthcoming African Code of Conduct for data centres[18,19,20].
- **Scaling renewables over time** is well-established practice across both hyperscalers and small private sector companies; while PPAs are currently concentrated in Europe, they are emerging in Africa and Asia[21,22,23,24].

Endnotes

1 IEA, Energy and AI, 2025
2 IEA, Energy and AI, 2025
3 Environmental and Energy Study Institute, Data Centers and Water Consumption, 2025
4 IEA, Energy and AI, 2025; GridLab, Practical Guidance and Considerations for Large Load Interconnections, 2025
5 This playbook defines small data centres having <50MW capacity, large data centres having 50-250 MW capacity and very large, hyperscale data centres having >250MW capacity.
6 IBM, Hyperscale vs. colocation: Go big or go rent?
7 Viavi Solutions, What is a Hyperscale Data Centre?
8 Equinix, What is a Hyperscale Data Center?, 2024
9 IEA, Energy and AI, 2025
10 Google, How to scale AI training to up to tens of thousands of Cloud TPU chips with Multislice, 2023
11 Bain, Data Centre Model, 2025
12 Bain, Data Centre Model, 2025
13 This playbook defines small data centres having <50MW capacity, large data centres having 50-250 MW capacity and very large, hyperscale data centres having >250MW capacity.
14 Bain, Data Centre Model, 2025
15 Water needs are calculated based on WHO's minimum standards of 50—100 litres per person per day
16 Uptime Institute, Water is Local, 2025
17 IEA, Energy and AI, 2025
18 IEA, Energy and AI, 2025
19 IEA, Energy and AI, 2025
20 Morgan Stanley, AI's Growing Thirst for Water, 2025
21 At WHO minimum standards of 50–100 litres per person per day
22 Rest of World, We mapped the world's hottest data centers, 2-25
23 IEA, Energy and AI, 2025
24 Clustered data centre hubs are geographical areas with high concentrations of interconnected data centres, leveraging shared infrastructure for massive data processing, storage, and AI, offering benefits like reduced latency, resilience, and powering digital economies, though straining local power grids.
25 McKinsey, Keeping cool in the data age, 2025
26 Ember Energy, Aligning ASEAN's digital growth with energy transition goals, 2025
27 Labour and skills are excluded here as they are not environmental resources and therefore do not form part of the resilience function considered in this section. These factors are addressed in the Strategic Pathways to Resilient AI section, which outlines the opportunities and pathways available to countries (based on their specific contexts) to participate in building AI systems.
28 IEA, Energy and AI, 2025
29 IEA, Energy and AI, 2025
30 STTelemedia, State-of-Art Features of Data centres in Mumbai and Mahape
31 USEIA, Cost and Performance Characteristics of New Generating, 2023
32 RMI, Powering the Data-Center Boom, 2024
33 GridLab, Practical Guidance and Considerations for Large Load Interconnections, 2025
34 S&P Global, Beneath the surface: Water stress in data centers, 2025
35 Data Centre Dynamics, AI data center growth deepens water security concerns in high-stress states – report, 2025
IEA, Energy and AI, 2025
36 AI Stack refers to the full set of layers required to design, build, and operate AI systems. This includes AI use cases, model layer, and data centre layer (hardware, compute, and supporting operations like cooling).
37 Refer to the annexure for detailed evidence on the solutions outlined in the table
38 Policies and standards are guiding principles that can inform and shape solutions, and are not fixed or exhaustive solutions alone
39 HPC is a technology that uses clusters of powerful processors that work in parallel to process massive, multidimensional data sets and solve complex problems at extremely high speeds.
40 Government of Norway, National Strategy for Artificial Intelligence, 2020
41 Voice of Vietnam, Vietnam launches National Data Centre to drive digital transformation, 2025
42 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
43 Sourced from UAE's inputs received in the Request for Information form shared with Working Group members
44 Model size is often reflected as model parameters which are the internal configuration variables of an AI or machine learning model which control how it processes data and makes predictions. They are the variables that the model learns during training.
45 OECD, Governing with Artificial Intelligence, 2025
46 Cottier et al., The rising costs of training frontier AI models, 2024
47 UNESCO, AI Large Language Models: new report shows small changes can reduce energy use by 90%, 2025
48 Sourced from UAE's inputs received in the Request for Information form shared with Working Group members
49 Sourced from Egypt's inputs received in the Request for Information form shared with Working Group members

50 Sourced from Myanmar's inputs received in the Request for Information form shared with Working Group members
51 Sourced from Mexico's inputs received in the Request for Information form shared with Working Group members
52 Sourced from Egypt's inputs received in the Request for Information form shared with Working Group members
53 European Data Portal, PlantVillage Nuru
54 World Economic Forum, This start-up is helping millions of farmers across Africa in 27 languages, 2024
55 QureAI, Product
56 IBM, What is Knowledge Distillation?, 2025
57 Accenture, Making generative AI greener, 2025
58 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
59 Liu, V. & Yin, Y., Green AI: exploring carbon footprints, mitigation strategies, and trade offs in large language model training, 2024
60 Reuters, China's Baidu to make latest Ernie AI model open-source as competition heats up, 2025
61 PRNewswire, Baidu Presents a Suite of Toolkits and Models to Supercharge AI Creativity at Create 2024, 2024
62 PaddlePaddle, Model Compression-Document-PaddlePaddle Deep Learning Platform, 2025
63 Government of Finland, National Roadmap for Data Centres, 2025
64 Tony Blair Institute for Climate Change, Powering AI in the Global South, 2024
65 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
66 Data Center Dynamics, KenGen installs BESS unit at modular data center in Nairobi, Kenya, 2025
67 EdgeUno, Data Centers
68 IBM, What is a green data centre?, 2025
69 ERM, Circular Building & Construction Supply Chains for Data Centers: A strategy to tackle environmental impact, 2025
70 Sourced from France's inputs received in the Request for Information form shared with Working Group members
71 Sourced from Mexico's inputs received in the Request for Information form shared with Working Group members
72 Meta, Advancing Low Carbon Concrete in our Data Centers, 2024
73 Vigyanlabs, Vigyanlabs, 2025
74 The Print, Vigyanlabs' CEO highlights AI's role in reducing energy consumption: "We have saved 2 terabits globally", 2025
75 Neural Processing Units (NPUs) are dedicated AI chips that mimic the functioning of the human brain and have energy-efficient parallel processing capabilities that reduce unnecessary processing
76 Field-Programmable Gate Arrays (FPGAs) are reconfigurable chips that can be customised for specific AI models, reducing the need for additional chips
77 IBM, What's the Difference Between AI accelerators and GPUs?
78 Google, Accelerate AI development with Google Cloud TPUs
79 Datacamp, Understanding TPUs vs GPUs in AI: A Comprehensive Guide, 2024
80 MIT News, New tools are available to help reduce the energy that AI models devour, 2023
81 Nvidia, How New GB300 NVL72 Features Provide Steady Power for AI, 2025
82 MIT Technology Review, AI will add to the e-waste problem. Here's what we can do about it, 2024
83 PennState Institute of Energy and the Environment, Why AI uses so much energy - and what we can do about it, 2025
84 GreenPow, Can AI Make Your Cloud Greener? Unpacking the Future of Carbon-Aware Automation, 2025
85 Google, Our data centers now work harder when the sun shines and wind blows, 2020
86 Google, Good News About the Carbon Footprint of Machine Learning Training, 2022
87 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
88 Hanwha Data Centres, Data Center Energy Efficiency Best Practices, 2025
89 Data Centre Dynamics, Citi consolidates from 70 data centers to 20, 2012
90 IBM, New IBM LinuxONE Servers Help Reduce Energy Consumption as Clients Increasingly Make Sustainability a Business Priority, 2022
91 Kompact AI, Kompact AI, 2025
92 Intel, Intel Unveils Leadership AI and Networking Solutions with Xeon 6 Processors, 2025
93 Tertiary operations refers to the supporting, non-compute infrastructure that enables the data centre to run, such as cooling and thermal management, facility-level energy and water use, heat recovery and reuse
94 Direct-to-chip liquid cooling involves circulating a coolant directly through microchannels or cold plates attached to the heat-generating components while immersion cooling submerges servers or electronic components in a dielectric fluid that absorbs and transfers heat away
95 World Economic Forum, 6 ways data centres can cut their emissions - without compromising the AI boom, 2025
96 McKinsey and Company, Keeping cool in the data age, 2025
97 In this closed loop system, water is filled once during construction and continually circulated between the servers and chillers to dissipate heat without requiring a fresh water supply
98 Microsoft, Sustainable by design: Next generation datacenters consume zero water for cooling, 2024
99 ST Telemedia Global Data Centres, ST Telemedia Global Data Centres Collaborates with Phaidra to Optimise Data Centre Cooling and Operations with AI, 2024
100 Sourced from Mexico's inputs received in the Request for Information form shared with Working Group members
101 Sourced from UK's inputs received in the Request for Information form shared with Working Group members
102 KPMG, Going green: data centres, 2025
103 Sustainable Energy Authority of Ireland, Tallaght District Heating Scheme, 2024
104 Codema Dublin's Energy Agency, Tallaght District Heating Scheme, 2024
105 June 2024 (latest data available)
106 PUE is calculated by dividing total data centre energy consumption by the amount of energy used specifically for computing tasks. The nearer the quotient to 1, the greater the efficiency, thereby reducing operational costs. Similarly, CUE measures carbon emissions associated with energy use and WUE determines the gallons of water used per kWh of IT load.

107 Sourced from UK's inputs received in the Request for Information form shared with Working Group members
108 Sourced from France's inputs received in the Request for Information form shared with Working Group members
109 Google, How we're making data centers more flexible to benefit power grids, 2025
110 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
111 Latitude Media, Verrus successfully demos its flexible data center technology, 2025
112 National Renewable Energy Laboratory, Vulcan Test Platform: Demonstrating the Data Center as a Flexible Grid Asset, 2025
113 PPAs are multi-year contracts to directly purchase power from producers at a fixed price via the grid, providing price stability for buyers while giving renewable energy developers the long-term revenue certainty needed to finance projects whereas unbundled energy attribute certificates (EACs), also known as guarantees of origin or renewable energy certificates (RECs), document the consumption of renewable energy allowing data centre operators to own the associated environmental benefits.
114 Matching every hour of a data centre's electricity consumption with an equivalent volume of carbon-free energy generation within the same local or regional electricity grid, ensuring clean energy supply aligns continuously with demand
115 Captive renewables refer to renewable projects established by companies or industrial facilities primarily to meet their own energy needs, as opposed to conventional renewable solutions that sell power to the grid.
116 Sourced from New Zealand's inputs received in the Request for Information form shared with Working Group members
117 CRU, New Electricity Connection Policy for Data Centres, 2025
118 Data Centre Dynamics, CtrlS completes 125MW solar farm to power Mumbai campus, 2025
119 DataCentre Magazine, How Google and TotalEnergies Secure Power for Data Centres, 2025
120 Niti Aayog, National Strategy for Artificial Intelligence, 2018
121 Sourced from Germany's inputs received in the Request for Information form shared with Working Group members
122 Sourced from France's inputs received in the Request for Information form shared with Working Group members
123 Sourced from UAE's inputs received in the Request for Information form shared with Working Group members
124 Sourced from New Zealand's inputs received in the Request for Information form shared with Working Group members
125 Sourced from Germany's inputs received in the Request for Information form shared with Working Group members
126 Sourced from UK's inputs received in the Request for Information form shared with Working Group members
127 US Department of Energy, Recommendations on Powering Artificial Intelligence and Data Center Infrastructure, 2024
128 Sourced from Canada's inputs received in the Request for Information form shared with Working Group members
129 Utilities are regulated providers of essential infrastructure services such as electricity, water, heating, that generate, transmit, and distribute these services reliably and at scale to support everyday life and economic activity.
130 Sourced from Germany's inputs received in the Request for Information form shared with Working Group members
131 Infocomm Media Development Authority, Driving a Greener Digital Future: SINGAPORE'S GREEN DATA CENTRE ROADMAP, 2024
132 This includes the reporting of scope 2 and 3 emissions
133 UNEP, AI End-to-End: The environmental impact of the full AI life cycle needs to be comprehensively assessed, 2024
134 EU Artificial Intelligence Act, Article 53: Obligations for Providers of General-Purpose AI Models, 2025
135 European Commission, Commission adopts EU-wide scheme for rating sustainability of data centres, 2024
136 AFNOR 2314
137 Sourced from UK's inputs received in the Request for Information form shared with Working Group members
138 Sourced from France's inputs received in the Request for Information form shared with Working Group members
139 Sourced from New Zealand's inputs received in the Request for Information form shared with Working Group members
140 Sourced from Mexico's inputs received in the Request for Information form shared with Working Group members
141 Principle 3: Maximise efficiency and flexibility such as shifting lower-priority AI tasks across times/geographies helps better determine how much additional capacity is needed. While this may come at the expense of speed, it ensures new builds are meeting true demand.
142 Assumption: This framework assumes that all countries are interested and will be partaking in the AI value chain.
143 1. Pumped Hydro Storage refers to pumping water uphill into a reservoir when renewables are abundant and releasing it downhill through turbines when needed. Cheapest long-duration storage (12+ hours), but requires specific geography (elevation change, water availability) and large upfront capital; 2. Compressed Air Energy Storage (CAES) refers to compressing air into underground caverns or purpose-built tanks when power is abundant and releasing to drive turbines when needed. Lower cost than batteries for long duration (8+ hours) but requires suitable geology or large above-ground tanks; 3. Thermal Storage refers to storing heat (in molten salt, hot water, etc) during the day and using it at night to generate electricity (concentrated solar power with storage) or directly for heating/cooling in buildings. Good for solar-heavy grids; duration typically 6–15 hours.
144 HPC is a technology that uses clusters of powerful processors that work in parallel to process massive, multidimensional data sets and solve complex problems at extremely high speeds.
145 India Smart Grid Forum, Handbook of AI, ML, VR, AR and Robotics Solutions and Roadmap for it's Adoption in Electric Utilities, 2025
146 Voice of Vietnam, Vietnam launches National Data Centre to drive digital transformation, 2025
147 Model size is often reflected as model parameters which are the internal configuration variables of an AI or machine learning model which control how it processes data and makes predictions. They are the variables that the model learns during training.
148 UNESCO, AI Large Language Models: new report shows small changes can reduce energy use by 90%, 2025
149 Consumer-grade means a product is designed for everyday users rather than professionals or large organisations. It is typically more affordable, easier to use, and less powerful or customisable than professional or enterprise-grade products.
150 World Bank, Digital Progress and Trends Report, 2025
151 UNESCO, Smarter, smaller, stronger: resource-efficient generative AI & the future of digital transformation, 2025
152 World Bank, Digital Progress and Trends Report, 2025

153 IVR (Interactive Voice Response) is an automated phone system that lets callers interact with a computer system using voice commands or keypad presses to get information, make requests, or route calls without a live agent.
154 Oguztürk, AI-driven irrigation systems for sustainable water management, 2025
155 Accenture, Making generative AI greener, 2025
156 IBM, What is Knowledge Distillation?, 2025
157 UNESCO, Smarter, smaller, stronger: resource-efficient generative AI & the future of digital transformation, 2025
158 Liu, V. & Yin, Y., Green AI: exploring carbon footprints, mitigation strategies, and trade offs in large language model training, 2024
159 Natural Language Processing (NLP) is a subset of AI that uses machine learning to teach computers how to understand and work with human language, like text and questions
160 The time delay between when an input is received by an AI system and when the output is produced. Low latency is critical for real-time applications such as video analysis, robotics, or decision support systems.
161 European Data Portal, PlantVillage Nuru
162 World Bank, Digital Progress and Trends Report, 2025
163 AI architecture refers to the structural design of an AI system, defining how its components and layers are organised and how data flows through them to balance performance, efficiency, and deployment constraints.
164 Circularity refers to practices that optimise resource use and minimise waste across the entire AI life cycle, emphasising sustainability and economic efficiency.
165 IBM, The future of AI and energy efficiency, 2025
166 BBC, Scottish data centres powering AI already using enough water to fill 27 million bottles a year, 2025
167 Those locations that are close to existing transmission lines or substations, therefore minimising the need for extensive new power infrastructure.
168 Govt of Dubai, Moro's Green Data Center Receives the Future Fit Seal, 2025
169 Huawei, Moro Builds the MEA Largest Tier III Sustainable Data Center Empowered by Huawei ModularDC with SmartLi UPS Solution, 2025
170 Data Center Dynamics, KenGen installs BESS unit at modular data center in Nairobi, Kenya, 2025
171 EdgeUno, Data Centers
172 Vigyan Labs, FEMTO
173 IBM, What is a green data centre?, 2025
174 Circularity refers to practices that optimise resource use and minimise waste across the entire AI life cycle, emphasising sustainability and economic efficiency.
175 Meta, Advancing Low Carbon Concrete in our Data Centers, 2024
176 Lenovo, Understanding Neural Processing Units (NPUs), 2025
177 Cornell Chronicle, AI hardware reimagined for lower energy use, 2025
178 Google, Accelerate AI development with Google Cloud TPUs
179 MIT News, New tools are available to help reduce the energy that AI models devour, 2023
180 Nvidia, How New GB300 NVL72 Features Provide Steady Power for AI, 2025
181 MIT Technology Review, AI will add to the e-waste problem. Here's what we can do about it, 2024
182 PennState Institute of Energy and the Environment, Why AI uses so much energy - and what we can do about it, 2025
183 Expert interviews
184 Electricity Maps, Google data centers shift their computations to cleaner times and locations, 2025
185 Google, Our data centers now work harder when the sun shines and wind blows, 2020
186 Hanwha Data Centres, Data Center Energy Efficiency Best Practices, 2025
187 Google, Good News About the Carbon Footprint of Machine Learning Training, 2022
188 GreenPow, Can AI Make Your Cloud Greener? Unpacking the Future of Carbon-Aware Automation, 2025
189 World Economic Forum, 6 ways data centres can cut their emissions - without compromising the AI boom, 2025
190 PUE is calculated by dividing total data centre energy consumption by the amount of energy used specifically for computing tasks. The nearer the quotient to 1, the greater the efficiency, thereby reducing operational costs.
191 McKinsey and Company, Keeping cool in the data age, 2025
192 Microsoft, Sustainable by design: Next generation datacenters consume zero water for cooling, 2024
193 Google Datacentres, Hamina, Finland, 2024
194 ST Telemedia Global Data Centres, ST Telemedia Global Data Centres Collaborates with Phaidra to Optimise Data Centre Cooling and Operations with AI, 2024
195 KPMG, Going green: data centres, 2025
196 Sustainable Energy Authority of Ireland, Tallaght District Heating Scheme, 2024
197 Codema Dublin's Energy Agency, Tallaght District Heating Scheme, 2024
198 June 2024 (latest data available)
199 PUE is calculated by dividing total data centre energy consumption by the amount of energy used specifically for computing tasks. The nearer the quotient to 1, the greater the efficiency, thereby reducing operational costs. Similarly, CUE measures carbon emissions associated with energy use and WUE determines the gallons of water used per kWh of IT load.
200 Earth Overshoot Day, EcoDataCenter, the world's first climate-positive data center, 2025
201 Scheider Electric, Innovating climate positive results: Ecodatacenter, 2025
202 Harvard Business Review, How Data Centers Can Support Energy Resiliency While Managing AI Demand, 2025
203 Google, How we're making data centers more flexible to benefit power grids, 2025
204 Latitude Media, Verrus successfully demos its flexible data center technology, 2025
205 National Renewable Energy Laboratory, Vulcan Test Platform: Demonstrating the Data Center as a Flexible Grid Asset, 2025
206 IEA, Energy and AI – Energy Supply for AI, 2025

207 PPAs are multi-year contracts to directly purchase power from producers at a fixed price via the grid, providing price stability for buyers while giving renewable energy developers the long-term revenue certainty needed to finance projects whereas unbundled energy attribute certificates (EACs), also known as guarantees of origin or renewable energy certificates (RECs), document the consumption of renewable energy allowing data centre operators to own the associated environmental benefits.

208 World Economic Forum, How AI can accelerate the energy transition, rather than compete with it, 2025

209 Amazon, Amazon is the largest corporate purchaser of renewable energy globally for the fifth year in a row, 2025

210 CRU, New Electricity Connection Policy for Data Centres, 2025

211 DataCentre Magazine, How Google and TotalEnergies Secure Power for Data Centres, 2025

212 Sources: 1. MIT Technology Review, AI will add to the e-waste problem. Here's what we can do about it, 2024, 2. LEED (Leadership in Energy and Environmental Design) certification is the world's most widely used green building rating system, 3. US Green Building Council, Applying LEED to data center projects, 2025, 4. Brookings, The future of data centers, 2025, 5. Green Web Foundation, Why we joined the Real Time Cloud Carbon Footprint Working Group, 2024, 6. The Energy Mix, Finland, Sweden Warm Up to Data Centre District Heat Amid Lingering Sustainability Concerns, 2025, 7. Harvard Business Review, How Data Centers Can Support Energy Resiliency While Managing AI Demand, 2025, 8. Oxford Insights, Government AI Readiness Index 2025, 9. Stanford, AI Index Report 2025, 10. BusinessWire, Micro Mobile Data Center Global Business Research Report 2025, 11. Mordor Intelligence, AI Accelerators Market Size & Share Analysis - Growth Trends and Forecast, 2025, 12. New York Times, Their Water Taps Ran Dry When Meta Built Next Door, 2025, 13. Uptime Institute, Global Data Center Survey 2024, 14. Leanne et al., Exploring Model Compression Techniques for Efficient Inference of Large Language Models, 2025. 15. Intel, Edge AI: Paving the Way for Intelligent, Resilient Deployments with a Systems Mindset, 2025, 16. IBM, Cloud investments soar as AI advances, 2025. 17. Government of Malaysia Ministry of Digital, Launch Of Malaysia's National Cloud Computing Policy (NCCP), 2025, 18. European Commission, 2025 Best Practice Guidelines for the EU Code of Conduct on Data Centre Energy Efficiency, 2025, 19. Infocomm, Driving a Greener Digital Future, 2024, 20. Yale Clean Energy Forum, How Hyperscalers Are Powering Their Data Centers, 2025, 21. Africa Data Centres Association, The African Code of Conduct: leading Africa towards sustainable data centre growth - African Actors of Data Center Association (ADCA), 2025, 22. IEA, Data Centres and Data Transmission Networks, 2025, 23. Intelligent Data Centres, Africa Data Centres and Distributed Power Africa work together to reach sustainable development goals, 2023, 24. Asian Insiders, Clean Energy Remains Crucial for Asian Data Centre Development , 2025

Disclaimer and Use Restriction Notice

This document has been prepared in connection with the India AI Impact Summit by the Government of India, with Dalberg Advisors acting as knowledge partner. It is intended solely to support deliberations and related processes under the Summit.

The country examples, case studies, and illustrative insights contained herein are based, in part or in full, on inputs received through the Summit's Request for Information (RFI) and related consultations with Working Group members, participating institutions, and other relevant stakeholders. The information provided by such contributors has not been independently verified. Accordingly, responsibility for the accuracy, completeness, and reliability of any data, metrics, or representations remains with the respective submitting entities. Neither the Government of India nor its knowledge partners make any representation or warranty, express or implied, as to the accuracy, validity, or completeness of such information, and they shall not be liable for any errors, omissions, or subsequent changes in the underlying data.

Nothing in this document shall be construed as legal, policy, technical, or professional advice. The contents are for informational and consultative purposes only and do not constitute official positions or commitments of the Government of India unless expressly stated.