



SHARON
- AI -

WRITE YOUR OWN RULES

The AI Capacity Crunch

Volume 1: Why GPU access has become a boardroom priority



GPU access isn't a 2027 problem. It's here. It's now.

For the last 2-years GPU access been the hidden handcuffs inhibiting business innovation. It's a combination of issues – from data sovereignty and compute access to commercial cost and competing priorities.

We call it the 'The AI Capacity Crunch'

It's now a burning issue for CIOs, CTOs and CEOs across Australia. Which is why Sharon AI asked our AI experts to write this two-part guide. To help Australian business leaders solve this dilemma.

In volume 1 we'll answer the key question every board member and team leader should be asking right now:

01. Why is AI now in the infrastructure phase?
02. Why is GPU access now a boardroom issue?
03. What's the opportunity hiding in plain sight?
04. What's the cost of waiting?
05. What is the 'AI Scale Gap'?
06. How do you manage misaligned executives?
07. Are you ready for 'The AI Capacity Crunch'?

Important note:

If you get valuable insights from Volume 1 of 'The AI Capacity Crunch' then request a copy of Volume 2: 'How to build a successful AI Factory for your Australian business' by emailing sales@sharonai.com

The new constraint on AI ambition

AI has moved from experimentation to strategic priority. The question is no longer whether AI matters, it's whether your organisation can get the infrastructure to act on it.

Boards, executives and business units are asking how artificial intelligence can lift productivity, accelerate innovation, reduce cost, improve customer experience and unlock new revenue. The ambition is well understood however there is a constraint that many organisations are only beginning to confront.

AI does not scale on ambition alone. It scales on infrastructure.

Enterprise AI depends on access to high-performance compute, GPU capacity, secure data environments, scalable storage, fast networking, and the operational capability to support training, inference and mission-critical workloads. None of that is automatic, and in 2026 none of it can be assumed.

Capital is moving because capacity is scarce

The signals are now unmistakable. In June 2026, Broadcom, Apollo and Blackstone announced a strategic platform to fund more than 20 gigawatts of AI compute through 2028, beginning with a US\$35 billion commitment to expand Anthropic's capacity by more than one gigawatt from mid-2026.¹ Private capital is moving at this scale for one reason: the infrastructure required to meet AI demand is genuinely supply-constrained.

When the largest AI developers are securing capacity years in advance, it is a clear indicator of how tight the market has become.

Australia sits inside this same dynamic. The federal National AI Plan names sovereign digital infrastructure a strategic priority⁵, and new AI-ready, sovereign capacity is coming online – but it is finite, and it is being contracted ahead of demand. The organisations that secure access early are better placed to move faster, test more confidently, scale more reliably, and avoid being constrained when capacity tightens further.

Sharon AI is perfectly positioned to help

We are already helping Australian enterprises secure that access now – with sovereign, enterprise-grade GPU infrastructure purpose-built for this moment.

“

**AI strategy without
infrastructure access
is a roadmap without a
road.**

02. WHY IS GPU ACCESS NOW A BOARDROOM ISSUE?

AI ambition is ubiquitous. AI capacity is not.

The challenge is no longer whether AI matters. It is whether the organisation has the infrastructure to deploy it reliably, scale it confidently, and secure it on its own terms.

Most organisations now understand the potential of AI. Far fewer have confronted the physical reality required to deploy it. AI adoption is placing sustained pressure on compute, data centres, power, storage and networking. That pressure is reshaping the market in ways which reach directly into enterprise planning.

The bottleneck has moved from chips to power

For the last few years, the constraint on AI infrastructure was hardware availability. That constraint hasn't gone away, and it's now compounded by a new one. The binding limit is increasingly access to the power and equipment that delivers it. Grid interconnection queues stretch for years. Lead times for high-voltage transformers and switchgear have stretched from months to multiple years.⁴ Gartner has projected power shortages could restrict 40% of AI data centres by 2027³.

That's a constraint that cannot be solved within a single budget cycle.

The rise of purpose-built AI clouds

In response, a new category of infrastructure has emerged. NVIDIA describes AI clouds as; "a growing ecosystem of purpose-built clouds, co-designed with full-stack AI infrastructure to meet surging demand from enterprises, startups and nations seeking new vendors and regional capacity."² These platforms are engineered around the realities of modern AI workloads – high-density GPUs, high-bandwidth networking and sovereign deployment. They are not retrofitted from general-purpose computing.

Australia is inside the same equation

This is not a distant, offshore story. Australia ranks among the most attractive destinations globally for data centre investment. The federal National AI Plan has elevated sovereign digital infrastructure to a strategic priority, with formal government expectations now shaping how new capacity is developed.^{5,6}

Purpose-built, AI-ready and sovereign facilities are coming online through 2026 and 2027. Importantly, capacity is being contracted ahead of demand, which means availability is a function of timing as much as budget.

What does it mean for each leader?

For the CIO	For the CTO	For the CFO
AI capacity decides readiness	AI capacity enables or hinders	AI capacity defines opportunity
AI infrastructure is now part of enterprise readiness – including governance, resilience, integration, security and compliance.	AI infrastructure access determines what can actually be built, and how fast.	Infrastructure shapes the economics of AI: cost, return, capital allocation and the financial consequences of delayed execution.

Why it matters

A power-and-equipment constraint behaves differently from a chip constraint. You cannot buy your way out of it at short notice, because the limiting factor sits in multi-year infrastructure, not in a purchase order. That is what turns capacity into a planning decision rather than a procurement one.

Sharon AI is perfectly positioned to help

Our sovereign infrastructure is a purpose-built, AI-ready facility. It's available right now for Australian enterprises that cannot afford to wait.

GPU Capacity is an emerging competitive advantage.

Access to GPU-powered high-performance compute is no longer just a technical requirement. It is becoming a strategic advantage for organisations that want to move faster from AI concept to commercial value.

GPU access governs how quickly an organisation can test models, deploy AI applications, support inference workloads, process large datasets and respond to new market opportunities. It is the difference between an AI strategy that can be executed and one that remains just an ambition.

The risk is subtle, because it does not look like failure.

An organisation can have strong use cases, clean data, executive sponsorship and an approved budget - and still be unable to move at the speed the opportunity requires, simply because it cannot reliably access the necessary compute. When that happens, programs stall quietly, and the cost is competitive rather than visible.

GPU Capacity is now a strategic differentiator.

For the CIO	For the CTO	For the CFO
Enables enterprise	Decides technical limits	Defines opportunity costs
The right infrastructure helps business units access approved, secure and governed AI capabilities exactly when they need them.	Decides how much teams can access the performance and capacity required for model development, optimisation, deployment and scaling.	Avoids slower product development and missed productivity gains. Also means the organisation doesn't depend on capacity it doesn't control during peak demand.

Delayed access creates delayed business value

Waiting to secure AI infrastructure creates hidden costs: delayed pilots, delayed production, delayed revenue impact, delayed productivity gains, and rising exposure to future price and capacity constraints.

The cost of AI infrastructure is usually framed as a technology expense. That framing is incomplete. The more important number is the cost of inaction – the value that does not arrive while an organisation waits. Unlike the cost of compute, the cost of waiting rarely appears on an invoice, which is exactly why it is so often underweighted.

The real cost of delay

- AI projects held in pilot stage, unable to reach production
- Business units sourcing their own unapproved tools to keep moving
- Rising shadow AI risk as workarounds proliferate outside governance
- Compute costs that climb as demand tightens supply
- Longer time to model deployment and slower innovation cycles
- Reduced ability to act on data the organisation already holds
- Lost first-mover advantage in markets where speed compounds

What does it mean for each leader?

For the CIO	For the CTO	For the CFO
Dangerous fragmentation	Expensive trade-offs	Higher total cost of ownership
Delayed infrastructure access pushes teams toward fragmented, ungoverned AI usage – the precise outcome enterprise governance exists to prevent.	Delayed compute slows experimentation, architecture decisions and production deployment, and forces premature trade-offs that are expensive to unwind.	Delayed execution reduces ROI, extends payback periods and raises TCO as the organisation buys capacity in a tighter market.

The cost of GPU access is visible. The cost of waiting is not. So, by the time the enterprise aligns behind an opportunity, the opportunity is long gone.

The gap between AI experiments and enterprise-scale AI

Many organisations can experiment with AI. Far fewer are ready to scale it into secure, reliable, repeatable production environments.

A proof of concept can often be delivered with limited infrastructure and a generous appetite for manual effort. Production AI is a different proposition. It has to run reliably, securely and economically. It has to deliver every day, against real data and real obligations. That requires a foundation most experiments never touch.

Here's what does AI production really requires:

- Reliable GPU capacity, not opportunistic access
- Secure data access and clear data boundaries
- Scalable storage and high-performance networking
- Monitoring, governance and cost controls
- Repeatable model deployment pathways
- Compliance, auditability and operational support

Understanding the 'AI Scale Gap'

Each stage of AI maturity asks something different of the business – and something different of the infrastructure underneath it. The gap most organisations fall into sits between pilot and production, where the infrastructure that proved the idea cannot carry it at scale.

What does it mean for each leader?

Stage	What the business wants	What infrastructure must support
Experimentation	Test AI ideas quickly	Flexible, on-demand compute access
Pilot	Validate a business use case	Secure data and controlled environments
→ THE AI SCALE GAP ←		
Production	Deploy at scale	Reliable GPU capacity, governance, monitoring
Optimisation	Improve performance and ROI	Cost visibility, workload tuning, ongoing capacity

Ultimately, experimentation tolerates improvisation. Production does not.

Organisations that plan for the production and optimisation needs before they arrive are the ones who cross the gap without stalling.

What does it mean for each leader?

For the CIO	For the CTO	For the CFO
Confidence in progress	Fit for purpose architecture	Investment is measurable
Meeting enterprise security and integration standards.	Delivering the right architecture and performance profile.	Tying infrastructure spend to measurable outcomes from day one.

CIOs, CTOs and CFOs must decide together

AI infrastructure is no longer owned by technology alone. It now sits at the intersection of enterprise risk, technical performance and financial return.

For most of the cloud era, infrastructure decisions could be delegated. AI changes that. The choices now being made about compute access carry consequences that no single function can own in isolation.

AI infrastructure decisions now shape:

- Business strategy and competitive positioning
- Technology architecture and speed to market
- Risk management and security posture
- Capital allocation and operational efficiency
- Data governance and regulatory exposure

That breadth is why AI infrastructure needs a shared decision framework, owned across the executive team rather than handed down through it.

One key question for each leader

For the CIO	For the CTO	For the CFO
Can we govern this securely and responsibly?	Can we build and scale what the business needs?	Can we justify the investment and manage the risk?

All three need to meet to decide AI infrastructure strategy

None of these questions can be answered well in isolation. A decision that satisfies one and ignores the others is the decision most likely to fail. It's governed but unaffordable. It's fast but ungoverned. It's cheap but not scalable. Without alignment, the strategy falls apart. That's not a courtesy, it's the control that holds it together.

06. ARE YOU READY FOR 'THE AI CAPACITY CRUNCH'?

Six simple questions. One business-critical answer.

AI infrastructure is becoming one of the most important strategic decisions facing executive teams. The question is no longer simply whether your organisation has an AI strategy. The question is, can you access the compute capacity to execute it?

Start with a quick self-assessment

01 Do you have a defined AI roadmap for the next 12 months?

☐ Yes ☐ No

02 Have you identified which use cases require GPU or high-performance computing?

☐ Yes ☐ No

03 Do you know where your AI workloads will run and under who's jurisdiction?

☐ Yes ☐ No

04 Have you assessed your data sovereignty and security requirements?

☐ Yes ☐ No

05 Do you have visibility over the cost of delaying AI execution?

☐ Yes ☐ No

06 Are your CIO, CTO and CFO aligned on AI infrastructure priorities?

☐ Yes ☐ No

If you're in any doubt, book an AI Infrastructure Readiness audit.

Speak to a Sharon AI expert one-on-one. Together we can design your organisation's AI roadmap, design around your GPU computing requirements and ensure your infrastructure readiness. Email us at sales@sharonai.com

Figures are accurate as at the date of each source.

1. Broadcom Inc. "Broadcom, Apollo, and Blackstone Establish Landmark Strategic Platform to Accelerate More Than 20 Gigawatts of Global AI Deployments." 9 June 2026. <https://investors.broadcom.com/news-releases/news-release-details/broadcom-apollo-and-blackstone-establish-landmark-strategic>
2. NVIDIA. "NVIDIA AI Cloud Ecosystem Expands Worldwide to Meet Global AI Compute Demand." NVIDIA Blog, 2026. <https://blogs.nvidia.com/blog/ai-cloud-ecosystem/>
3. Gartner. "Gartner Predicts Power Shortages Will Restrict 40% of AI Data Centers By 2027." Press release, 12 November 2024. <https://www.gartner.com/en/newsroom/press-releases/2024-11-12-gartner-predicts-power-shortages-will-restrict-40-percent-of-ai-data-centers-by-20270>
4. POWER Magazine. "Transformers in 2026: Shortage, Scramble, or Self-Inflicted Crisis?" January 2026, citing Wood Mackenzie Q2 2025 transformer lead-time survey. <https://www.powermag.com/transformers-in-2026-shortage-scramble-or-self-inflicted-crisis/>
5. Australian Government, Department of Industry, Science and Resources. National AI Plan – "Capture the opportunities" (data-centre investment ranking per Knight Frank, 2025). Released 2 December 2025. <https://www.industry.gov.au/publications/national-ai-plan/capture-opportunities>
6. Australian Government, Department of Industry, Science and Resources. "Expectations of Data Centres and AI Infrastructure Developers." 23 March 2026. <https://www.industry.gov.au/publications/expectations-data-centres-and-ai-infrastructure-developers>

Be ready for 'The AI Capacity Crunch'.

Join us and write your own rules.

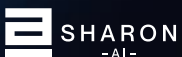
 sales@sharonai.com

 www.sharonai.com

About Sharon AI

Sharon AI is a High-Performance Computing company focused on Artificial Intelligence and Cloud GPU Compute Infrastructure.

Our cloud GPU platform and compute infrastructure is accelerating the build of AI factories and sovereign AI solutions, powering the next wave of accelerated computing adoption.



WRITE YOUR OWN RULES