World Wide Technology | SHARON -AI-

The New Rules of Transformation

An Infrastructure Leader's Playbook for the AI Factory Era

Co-authored by World Wide Technology and Sharon AI



About World Wide Technology and Sharon AI

World Wide Technology

World Wide Technology (WWT) is a global technology solutions provider helping organisations make a new world happen by turning ambition into real-world outcomes. Founded in 1990, WWT brings together strategy, deep technical expertise and world-class technology partnerships to help public- and private-sector organisations design, build and scale intelligent AI, digital, cybersecurity, cloud and infrastructure solutions. Through its Advanced Technology Center (ATC) — a collaborative ecosystem featuring state-of-the-art hardware and software — WWT enables clients and partners to conceptualise, test and validate innovative technology and then deploy solutions at scale using its global integration and distribution capabilities. With more than 14,000 team members and over 60 locations worldwide, WWT's culture — grounded in core values and leadership philosophies — has been recognised by Fortune and Great Place to Work® for its commitment to innovation, trust and creating a great place to work for all. WWT serves clients in nearly 200 countries across six continents, and through Softchoice, a World Wide Technology company, it supports U.S. commercial and SMB markets and all of Canada.

wwt.com

SHARON -AI-

Sharon AI, Australia's leading Neocloud, is a high-performance computing company focused on Artificial Intelligence and Cloud GPU Compute Infrastructure. Sharon AI's cloud GPU platform and compute infrastructure is accelerating the build of AI factories and sovereign AI solutions, powering the next wave of AI training, inference, and mission-critical workloads. Sharon AI is actively advancing a footprint across the Asia-Pacific region to meet the accelerating demand for sovereign AI infrastructure across one of the world's fastest-growing technology markets. Sharon AI is headquartered in Sydney, Australia.

sharonai.com

The rules have changed

The AI inflection point has arrived. In the space of eighteen months, AI has moved from a line item on the IT agenda to the agenda itself and how that happened matters as much as the fact that it did. For most organisations, the first phase of AI adoption looked like a series of contained experiments: a GPU cluster here, a proof of concept there, a handful of data scientists running workloads in a cloud environment that no one expected to scale.

That phase is over.

78% of organisations intend to increase AI investment in the next fiscal year (Deloitte, 2025). Yet only 7% have fully scaled AI across their operations and just 6% report meaningful EBIT impact (McKinsey, November 2025). The constraint is not ambition – every organisation in that gap wants to scale. The constraint is the infrastructure they are trying to scale on.

Production-grade AI has fundamentally different requirements from the environments most organisations built for experimentation. It is not an upgrade to existing data centre strategy. It is not a larger version of a workload. It is a new category of infrastructure, with its own architecture principles, its own operational requirements, and its own set of decisions that, once made, are difficult to reverse. I&O leaders are being asked to make those decisions now, often without the frameworks or precedents that have guided every previous generation of technology investment.

This playbook is written for those leaders. It addresses two things directly: a decision framework for determining where AI workloads should run, and a reference architecture for building the infrastructure to run them.

Why AI programmes stall before they scale

The gap between a working AI prototype and a production AI system is wider than most organisations anticipate. Prototypes are built to demonstrate capability. Production systems are built to deliver it, repeatedly, at scale, under operational and regulatory constraints that proof of concepts rarely design for.

The most common failure mode is not technical. Organisations stall because they try to scale AI workloads through infrastructure that was never designed to carry them. General-purpose cloud environments, legacy compute clusters and shared networking architecture can support experimentation. They cannot reliably support the data throughput, latency requirements and governance demands of production AI at enterprise scale. The result is a pattern that has become familiar across the ANZ market: a successful pilot, a decision to scale, and then months of unplanned infrastructure remediation before anything of organisational value.

A second failure mode is sequential decision-making. Many organisations approach AI infrastructure the way they approached previous technology generations: evaluate, select, procure, deploy, repeat. That sequence assumes each layer of the stack can be optimised independently. In an AI Factory environment, it cannot. Compute, networking, storage and software are interdependent in ways that punish siloed procurement. A GPU cluster under-supported by storage throughput will not perform to specification. A high-performance storage environment on the wrong network fabric will not deliver its rated throughput. The performance of the system is determined by how the components work together, not by the specification of any individual part.

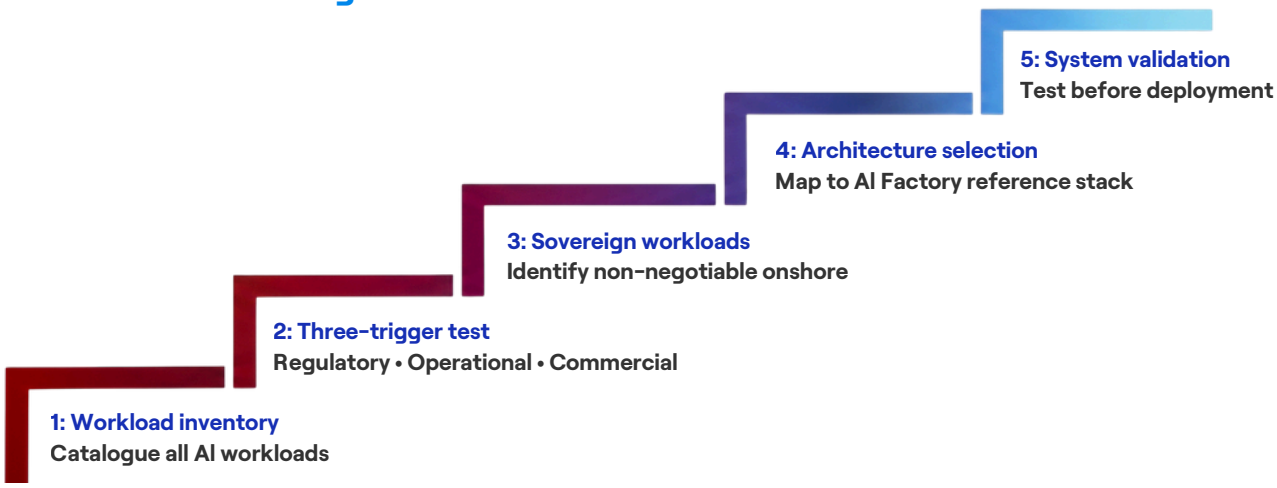
> The organisations moving fastest are those who have shifted from evaluating technology in isolation to validating the entire stack, before it reaches the production environment.

Part 1: The AI workload decision framework

Where AI workloads run is one of the most consequential infrastructure decisions an I&O leader will face. The default assumption of the previous decade, that public cloud is the appropriate home for most enterprise workloads, is under active review across every critical sector of the ANZ economy.

Three converging forces are driving that shift: **regulatory pressure, operational requirements and commercial logic.** This framework gives you a structured way to make the placement call, workload by workload, across all three.

The framework at a glance



Step 1: Catalogue AI workloads

Key to the framework is establishing a complete inventory of your current and planned AI workloads. This estate-mapping step aligns with WWT's AI Technology Capability framework, which uses workload inventory as the entry gate for placing organisations on an AI Factory or pre-Factory adoption pathway. For each workload, capture:

- Workload type (training, inference, fine-tuning, data preparation)
- Data classification (what data the workload touches and its sensitivity)
- Latency and performance requirements (real-time, near-real-time, or batch)
- Expected volume and growth trajectory over three to five years
- Current hosting environment and any known constraints

Step 2: Apply the three-trigger test

For each workload, work through the three triggers.
A single **YES** to a regulatory question will often override all other considerations.

Trigger 1: Regulatory	
Does this workload process data subject to Australian or NZ data residency requirements?	YES → Dedicated / onshore infrastructure required
Does inference or model training involve personally identifiable, financial or health data?	YES → Map against applicable residency obligations before proceeding
Does your organisation operate under APRA, IRAP, or equivalent critical infrastructure regulation?	YES → Sovereign, on-premises or co-located infrastructure strongly indicated
Are you subject to sector-specific AI governance requirements that mandate local processing?	YES → Engage legal and compliance before finalising architecture

Gartner predicts that by 2028, 65% of governments will enforce data sovereignty rules restricting cross-border AI use. Organisations yet to map their workloads against these obligations are carrying risk they have not fully priced.

Trigger 2: Operational	
Does this workload have latency requirements that cloud environments cannot guarantee?	YES → Dedicated infrastructure eliminates that variability
Is it embedded in operational processes where downtime or variability has direct business impact?	YES → Onshore deployment removes dependency on external connectivity
Does the workload depend on connectivity between your operational environment and AI compute?	YES → Co-location or on premises removes that single point of failure
Have you experienced performance degradation from cloud variability in comparable workloads?	YES → Strong indicator that production workloads require dedicated architecture

Production AI systems require consistent, predictable performance. Latency variability acceptable for many enterprise workloads becomes a material issue when AI is embedded in operational processes requiring real-time or near-real-time inference. The dependency on network connectivity between an organisation's operational environment and its AI compute is a single point of failure that many risk teams have not yet fully assessed.

Trigger 3: Commercial	
Is this workload projected to run at significant volume for a known period of time?	YES → Commercial case for dedicated infrastructure is positive over time
Have you modelled total cost of ownership for dedicated vs. cloud over a realistic horizon?	YES → TCO modelling consistently favours dedicated at production scale
Are inference costs already material and growing as you move toward production scale?	YES → A signal that cloud economics at scale are no longer optimal
Would a 40–60% reduction in inference cost (achievable on dedicated infrastructure) materially improve unit economics?	YES → Organisations on dedicated infrastructure are achieving these reductions (WWT, 2026)

Commercial considerations may not lead to a binary decision. A hybrid architecture, in which certain workloads remain in cloud environments while production inference and sensitive training workloads move to dedicated infrastructure, is the right answer for most organisations.
The framework below maps trigger outcomes to placement recommendations.

Step 3: Determine placement

Each row below maps to a combination of Step 2 trigger outcomes. The framework is conservative on regulatory exposure: a single regulatory **YES** sends the workload to a sovereign architecture regardless of how the operational and commercial triggers score, because regulatory risk cannot be priced away by commercial savings. Treat the placements below as the starting point for the I&O, compliance and architecture conversation, not as the final call.

Note: One or more **YES** within a trigger constitutes a positive outcome for that trigger.

Trigger outcome	Recommended placement	Action
All three triggers positive	Dedicated / onshore	**Proceed to Part 2** to evaluate the AI Factory architecture
Regulatory trigger positive (regardless of others)	Dedicated / onshore **required**	Non-negotiable – **proceed to Part 2:** AI Factory architecture
Operational + commercial positive; regulatory negative	Hybrid - with this workload onshore	**Architect for hybrid** – prioritise low-latency / high-availability / high-performance workloads on dedicated
Commercial only positive	Cloud for now	**Reassess at scale** – monitor TCO trajectory
No triggers positive	Public cloud	**Retain in cloud** – revisit as workload volume grows

Part 2: The AI Factory reference architecture

An AI Factory is a purpose-built infrastructure environment designed to support the full AI production lifecycle: data ingestion and preparation, model training and fine-tuning, inference at scale, and the operational monitoring required to maintain model performance over time.

Unlike a general-purpose data centre, it is designed around the specific workload characteristics of AI: high-throughput data movement, GPU-intensive compute, low-latency inference and continuous operational monitoring.

The AI Factory is also the recommended architecture for sovereign and regulated workloads, not only for scale: its purpose-built design directly addresses the data residency, throughput and operational continuity requirements that public-cloud environments cannot reliably guarantee. Workloads arriving here from Step 3 on regulatory grounds belong in this architecture for the same reasons that production-scale workloads do.

The reference architecture validated by Sharon AI in production is built on four integrated components. Each is selected for its performance in AI workloads and its ability to function as part of a unified system rather than a collection of independent parts.

The same four-component stack has been validated in WWT's Advanced Technology Center (ATC) and exercised through the AI Proving Ground, giving I&O leaders **two reinforcing proof points: production validation by Sharon AI and pre-deployment validation by WWT.**

See reference architecture components on next page.

Component	Partner	Why it matters at production scale
Compute and AI network fabric	**NVIDIA**	GPU architecture directly determines model training speed and inference performance. NVIDIA's CUDA ecosystem underpins most production AI frameworks. The NVIDIA networking stack, spanning Spectrum-X Ethernet fabric and Quantum InfiniBand switching completes the picture: purpose-built interconnect that sustains GPU utilisation at scale and eliminates the network as the performance bottleneck.
Enterprise security and validated reference design	**Cisco**	Cisco participates in the Sharon AI stack through the Secure AI Factory with NVIDIA validated reference design, co-engineered to meet the compliance and access-control requirements of regulated industries. The AI Defense layer adds security controls that apply to AI traffic flows and model interactions. WWT is the first partner to deploy AI Defense on this architecture in production.
Storage layer	**VAST Data**	Storage is frequently the overlooked constraint and the first bottleneck in AI deployments. Training runs require rapid access to large datasets. Inference at scale requires low-latency retrieval. VAST's architecture is purpose-built for these requirements: high throughput and fast access without the complexity of legacy high-performance storage.
Physical infrastructure	**NEXTDC**	Carrier-neutral, sovereign data centre facilities across Australia. For organisations with data residency requirements, NEXTDC's footprint provides assurance that workloads remain within Australian jurisdiction throughout the AI production lifecycle. High-density power and cooling infrastructure support the compute intensity of production AI.

The components in detail

NVIDIA: Compute and AI network fabric

The choice of GPU architecture matters. Not all AI workloads have the same compute profile, and the gap between a correctly specified GPU environment and an incorrectly specified one is significant in both performance and cost terms. NVIDIA's ecosystem provides the software stack, including CUDA, that underpins most production AI frameworks, reducing integration friction and simplifying operational management of the compute environment.

What is less commonly understood is that NVIDIA's networking portfolio is an equal part of the production AI story. The GPU is only as effective as the fabric that feeds it. NVIDIA's network components are engineered specifically for the traffic patterns AI workloads generate: high-volume, latency-sensitive, all-to-all communication between accelerators that no general-purpose network architecture was designed to carry.

Category	Component	Primary use case
Switches	**Quantum / Spectrum-X**	InfiniBand lossless switching (Quantum) and AI Ethernet fabric (Spectrum-X) - purpose-built for GPU-to-GPU all-to-all traffic
Adapters	**ConnectX-8 SuperNIC**	RDMA-capable network adapter providing high-bandwidth, low-latency connectivity between accelerators and the fabric
Management	**DOCA / NetQ**	Unified telemetry, observability, and programmability layer across the full NVIDIA network fabric

Quantum: InfiniBand for large-scale training

Unlike Ethernet, which manages congestion after the fact, InfiniBand is a lossless fabric by design: it guarantees packet delivery without retransmission and delivers the sub-microsecond latency that keeps GPU synchronisation intact across large clusters.

NVIDIA Quantum switches are the production implementation of that standard for AI workloads - high-radix switching architecture that scales to tens of thousands of GPUs in a single fabric domain without the latency degradation that limits conventional network designs at scale.

For organisations running distributed model training at serious scale, Quantum InfiniBand is the fabric the world's largest AI clusters are built on, and the benchmark against which every alternative is measured.

Spectrum-X: Ethernet AI fabric
While standard ethernet was engineered for general enterprise traffic - bursty, tolerant of variability, and forgiving of occasional packet loss, AI workloads are none of those things. Spectrum-X is built specifically to eliminate that failure mode. Its Multipath Reliable Connection technology detects congestion in microseconds and dynamically reroutes traffic before blocking occurs, sustaining GPU utilisation at the levels production AI economics require.

For organisations running Ethernet infrastructure - the majority of enterprise environments - Spectrum-X delivers InfiniBand-class performance consistency without requiring a wholesale fabric replacement.

Cisco: Enterprise security and validated reference design

Cisco's role in the Sharon AI validated stack is focused where it matters most for regulated-industry organisations: security and compliance assurance. The Secure AI Factory with NVIDIA validated reference design, co-developed by Cisco and NVIDIA, provides a pre-tested architecture blueprint that meets the access-control, segmentation, and audit requirements of sectors including financial services, government, and healthcare. Cisco AI Defense adds a security layer specifically for AI traffic: monitoring model interactions, enforcing access policies, and providing the audit trail that compliance frameworks require.

WWT is the first partner to have deployed AI Defense on this architecture in a production environment, which means I&O leaders in the ANZ market have a **proven path to compliant AI infrastructure** rather than a reference that has not yet been tested in the field.

NEXTDC: Physical infrastructure

Physical infrastructure is where sovereignty commitments are either kept or compromised. A data residency policy is only as enforceable as the facility it relies on - and most general-purpose data centres were not built for the power density, cooling capacity or jurisdictional certainty that production AI requires. NEXTDC's facilities are engineered for high-density compute, with the carrier-neutral connectivity that removes single points of failure from the network path.

NEXTDC **operates exclusively** within Australia, the sovereignty assurance it provides is structural rather than contractual - jurisdictional certainty is built into the architecture, not negotiated into a service agreement.

VAST Data: Storage layer

The storage challenge in AI environments is not just capacity - it is the combination of throughput, concurrency and latency that traditional architectures were never asked to sustain simultaneously. Legacy high-performance storage typically trades one of these properties against another: high throughput at the cost of latency, or low latency at limited scale. VAST's architecture is built around a disaggregated model that delivers all three together, without the operational complexity that made previous generations of HPC storage a specialised discipline in their own right.

For I&O teams that do not have a dedicated storage engineering capability, that simplification is as significant as the performance.

Why integration is the critical variable

AI infrastructure is now a strategic decision requiring architectural rigour (WWT) and Australian operational presence (Sharon AI).

While selecting the right components is necessary, making them perform as a system is what determines production outcomes. AI workloads are interdependent in ways that punish siloed procurement: a GPU cluster undersupported by storage throughput will not perform to specification; a high-performance storage environment on the wrong network fabric will not deliver its rated throughput.

World Wide Technology operationalises the Sharon AI reference architecture stack. WWT validates the full system before it reaches the customer environment, compressing deployment timelines and reducing the technical risk that comes with first-of-kind implementations. I&O leaders get a single, accountable path from architecture decision to live deployment. Components arrive on site already tested and validated as a system.

> For I&O leaders in the ANZ market, where data sovereignty, supply chain reliability and operational resilience sit at the top of every agenda, the difference between a reference architecture and a running factory comes down to how that integration work is sequenced and executed.

Part 3: Your next move

The organisations that will lead in the AI Factory era are not those with the largest budgets. They are the ones making the right decisions in the right sequence. Here is that sequence:

Step 1: Complete your workload inventory
Use the estate mapping criteria in Part 1. If you do not have a current picture of what AI workloads you are running or planning, start here before any infrastructure decision.

Step 2: Apply the three-trigger framework
Run every candidate workload through the regulatory, operational and commercial triggers. Document the outcome for each. This becomes your placement map and your business case for infrastructure investment.

Step 3: Identify your sovereign workloads
Any workload that triggers a regulatory **YES** requires a dedicated, sovereign architecture. These are non-negotiable and should be prioritised in your infrastructure roadmap regardless of commercial trigger outcomes.

Step 4: Validate architecture against your workload profile
Use the reference architecture in Part 2 as your baseline. The Sharon AI validated stack, NVIDIA, Cisco, VAST and NEXTDC, has been tested against production AI workloads in critical-industry environments. Validate it against your specific requirements before committing to procurement.

Step 5: Test the full system before you deploy
Do not procure components and integrate on-site. Validate the full stack in a production-equivalent environment first - for the Sharon AI stack, that environment is WWT's AI Proving Ground at the Advanced Technology Center, with production-equivalent labs across compute, networking, storage and workload management. This is what compresses deployment timelines and prevents the costly remediation that derails most AI infrastructure programmes between pilot and production.

World Wide Technology | SHARON -AI-

The window for getting this right is **narrower** than it appears. Infrastructure decisions **made now**, on the placement, architecture and integration questions this playbook is built around - will shape operational capability for the next five to seven years.

The organisations that treat those decisions with the rigour they deserve and move with partners who have already validated the path will not just be better positioned.

They will be structurally harder to displace.



Make a new world happen.

World Wide Technology
wwt.com/contact-us

Sharon AI
sharonai.com/contact