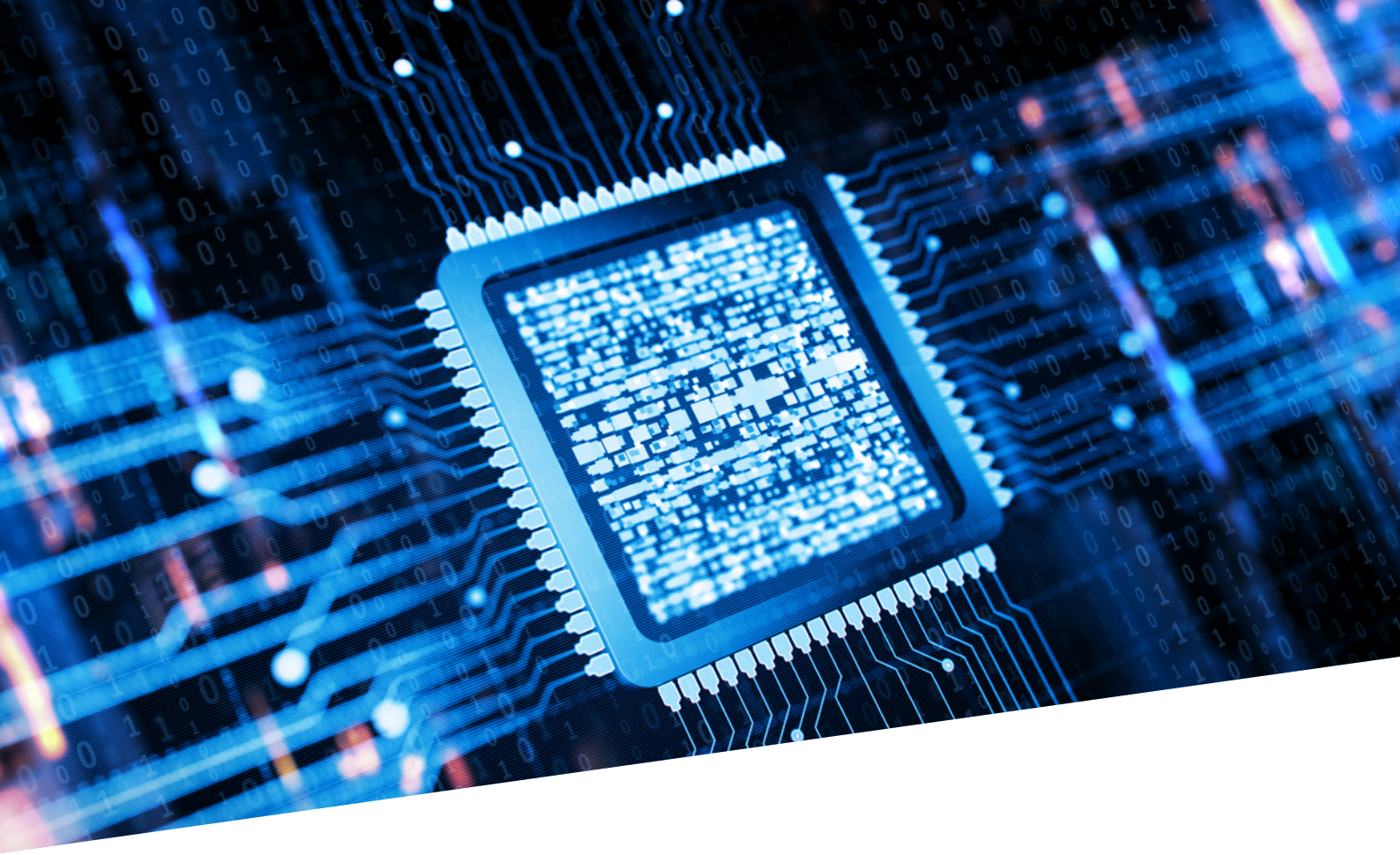


Who really controls the GPUs?

Cyber, operational and export-control risk and its effect on GPU-as-a-Service bankability

Gilbert + Tobin

GPU-as-a-Service is becoming a key route to AI compute at scale, but ownership of the chips says little about who controls the infrastructure around them. Cyber, operational and export-control risk follows that control more closely than legal title, making that mismatch increasingly material to providers, financiers, data-centre operators and AI customers.

The business model in plain English

A GPU-as-a-Service (**GPUaaS**) provider aggregates high-performance GPU (**Graphic Processing Unit**) servers into clusters and sells that computing capacity for AI (**artificial intelligence**) training and inference. Beneath that simple proposition sit materially different operating models.

The provider may own, lease or finance the GPUs, house them in a third-party data centre, and rely on original equipment manufacturer (**OEM**) firmware and software it did not develop. Its service may range from dedicated 'bare metal' servers to a managed cluster or integrated AI platform, and its customer may in turn provide the compute to affiliates or downstream users.

Four things that are easy to conflate therefore come apart: who owns the physical asset, who owns the software and intellectual property, who technically controls the operating environment and who is contractually and legally answerable for the service. From a cyber risk and regulatory perspective, the useful diligence question is not 'who owns the GPUs?' but 'who controls each layer required to deliver the AI compute capacity to the customer, and who bears the consequences when a layer fails, is compromised or is used unlawfully?'

Bare metal describes the hardware, not the management

GPUaaS offerings range from customer-controlled bare metal through provider-managed clusters to integrated AI platforms. Those labels are useful shorthand, but they are not a control map.

A customer can receive dedicated physical servers while the provider still controls firmware, the operating-system image, cluster scheduling, monitoring, remediation and the management interface. In another superficially similar offering, the customer may hold far greater administrative control. 'Bare metal' therefore describes the hardware substrate, not necessarily who manages it.

The commercial consequence is simple: the cyber and operational control boundary must be mapped, not inferred from the product name.

Three risks that travel together

Cybersecurity

Privileged control is distributed. The customer may control its AI application but not the server-management systems beneath it; the provider may promise availability while relying on OEM firmware and a third-party data centre it cannot itself remediate. Generic 'shared responsibility' language is therefore inadequate. Diligence should identify who controls patching, privileged access, monitoring, isolation, recovery and incident response at each layer, and whether the provider's cyber maturity matches the layer(s) it manages. As the provider moves up the stack, it assumes more security-critical functions and should demonstrate correspondingly mature governance, change and vulnerability management, privileged-access controls, continuous monitoring, tested response and appropriately scoped independent assurance. Applicable critical-infrastructure obligations sit alongside, rather than being displaced by the contract.

Cyber insurance

GPU contracts and financing documents often require borrowers to maintain cybersecurity insurance covering first- and third-party damages for cyber breaches and outages, so the policy wording and carve-outs should be carefully examined. Updates to model carve-out language for grey zone cyber-attacks without commensurate coverage in force majeure definitions may result in the GPUaaS provider being unable to claim under an insurance policy and facing damages for missed service-level agreements (**SLAs**) without recourse to force majeure. The precise scope of cyber and business interruption insurance, risk of claims being denied and back-to-back contractual obligation mapping are critically important in GPUaaS financing and operational agreements.

Operational resilience

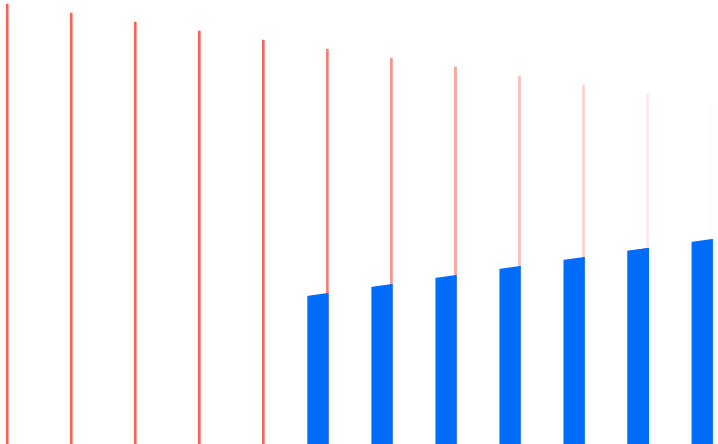
A GPU cluster is not a collection of interchangeable chips: compute, networking, storage, power, cooling and management software must operate together. Failure of a comparatively inexpensive component can idle very valuable capacity, and tightly synchronised training can be disrupted by failed nodes. Yet the provider may promise an end-to-end service without equivalent control or remedies against the OEM or data centre operator. Back-to-back service levels, replacement rights, telemetry, spare capacity, step-in and transition rights and concentration risk therefore become the core commercial terms.

Cyber policies, procedures and certifications

The diligence implication is not simply to ask whether the provider holds ISO 27001 or SOC 2 credentials. It is to test whether the scope of that assurance, and the provider's people, processes and technical controls, match the service actually being sold. A certification that does not cover the relevant cluster, management plane or material supplier may say little about the risk being financed or outsourced.

Export controls and prohibited use

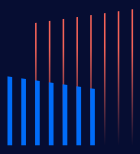
GPUaaS separates the location of the hardware from the identity and location of the person using its compute: a GPU can remain in an Australian or other permitted data centre while being accessed remotely from elsewhere. US controls on advanced computing items and prohibited end users and end uses therefore matter to the model, as does Bureau of Industry and Security (**BIS**) guidance on diversion risk. The proposed Remote Access Security Act would give BIS express authority over certain remote access to items controlled under the Export Administration Regulations (**EAR**); it has passed the House but remains proposed legislation. Australian controls on intangible supplies of controlled technology may also be relevant to particular arrangements. The applicable controls and fault standards must be assessed transaction by transaction.



Diligence should therefore go beyond know-your-customer (**KYC**) and geolocation. Providers should address downstream users and resale, end-use representations, access controls, relevant usage anomalies and red-flag escalation, with the technical ability to restrict or suspend access when diversion or prohibited use is identified. This need not mean inspecting confidential models or data. It means designing proportionate identity, logging, monitoring and enforcement controls around remote access to high-end compute. As providers move further up the stack and gain greater visibility and control over customer access, regulatory expectations that they use those capabilities to prevent misuse are also likely to rise.

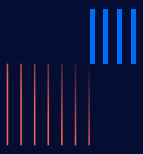


Why this matters commercially



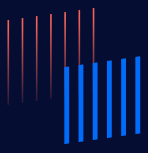
Providers and sponsors

The provider sits at the centre of operational dependencies it does not fully control while making contractual commitments to customers and financiers across those dependencies. Its enterprise value depends on access to scarce GPUs, a reliable revenue stream from creditworthy customers, continued regulatory ability to supply and operate compute, reliable data-centre capacity and power, robust OEM support, effective cyber controls and its ability to manage customer and end-user risk.



Hyperscalers and frontier AI customers

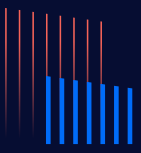
Outsourcing the infrastructure does not outsource dependency on it. Large training programmes may rely on thousands of GPUs operating together, with little equivalent replacement capacity available at short notice. Customers therefore need visibility into the critical upstream dependencies behind the provider's SLA, the provider's remedies if those dependencies fail, and the cyber or export-control events that could interrupt access to business-critical compute.



Financiers and due diligence

A lender is financing both hardware and the cash flows that hardware is expected to generate. Both depend on an operational, cyber and regulatory ecosystem the borrower may not fully control. Bankability therefore turns not only on customer credit and residual asset value, but on whether the borrower has adequate cyber and export-control controls, insurance, OEM support and contractual protection across the dependency chain. Financing terms may translate those risks into conditions precedent, covenants and default triggers around uptime, insurance, material contracts, cyber incidents, sanctions and export controls.

For sponsors and advisers, these are therefore credit fundamentals rather than peripheral technology issues. Weak controls or contractual gaps can affect leverage, covenant protection, conditions to funding and ultimately access to capital. Diligence should extend beyond title, security, customer credit and power supply to the control stack and dependencies that determine whether the financed GPUs can lawfully and reliably generate revenue.



Data-centre operators and OEMs

These parties often control critical layers, including physical security, power, cooling, operational technology (**OT**) networks and, in some cases, rack-level access, without any direct contractual relationship with the AI customer or end user. OEMs retain ownership of firmware and software IP and may restrict or withdraw support where export-control concerns arise. Their cooperation is essential to the bankability of the financing: tripartite or collateral access agreements are routinely required by lenders to ensure that the financier can step in, access and realise the GPU assets if the borrower defaults, notwithstanding the data-centre operator's separate contractual relationship with the borrower. In addition, data-centre operators typically exclude all consequential loss, including loss of revenue, and do not monitor customer equipment, meaning a facility-level incident can idle GPU capacity worth hundreds of millions while the operator's liability is capped at 12 months of fees and the borrower has no direct recovery path for lost customer revenue.

Data-centre operators and OEMs should therefore understand the obligations being constructed around their services, including availability commitments, incident response cooperation, maintenance and warranty support for the full term of the financing, and export-control cooperation, because those obligations increasingly form part of the conditions precedent and ongoing covenants in the credit documentation.

Four key questions to ask

01

Who controls what?

Map each material physical and software layer: who has privileged access, and who can patch, change, isolate and recover it?

02

What happens when it fails?

Test dependencies, service levels, incident response, upstream remedies, replacement capacity and recovery rights.

03

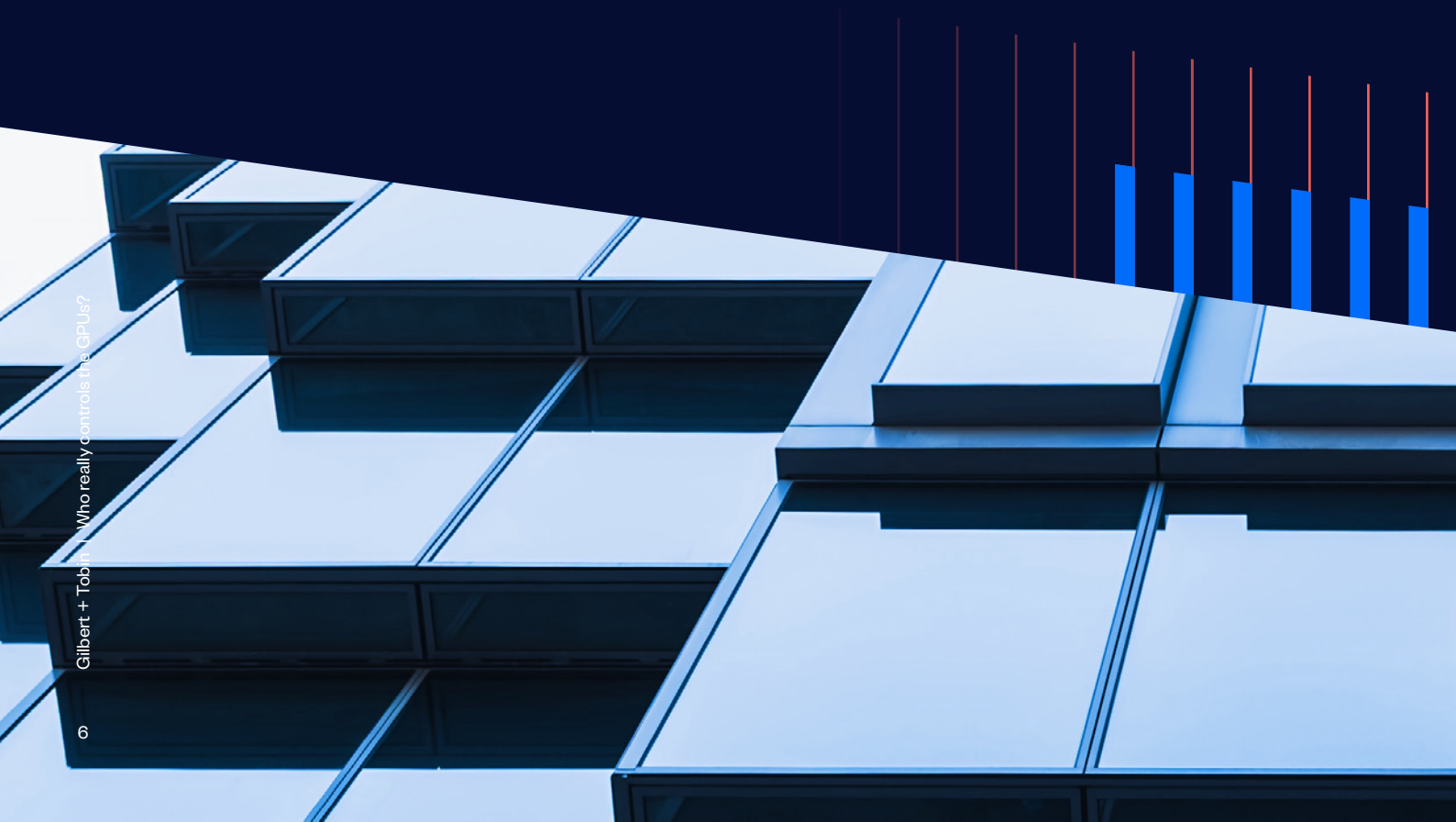
Who can use the compute – and for what?

Look beyond KYC to beneficial ownership, downstream users, access location, resale, intended end use, diversion controls, red-flag escalation and the provider's ability to suspend prohibited access.

04

Does the risk framework support bankability?

Test whether cyber controls, export-control compliance, insurance and contractual protections match the risks on which funding, covenants and default thresholds depend.



Conclusion

In GPU-as-a-Service, owning the chip is where the risk analysis starts, not where it ends. Bankability depends on whether the borrower can manage cyber, operational and export-control risks around the asset well enough for the GPUs to perform, the revenues to flow and the collateral to retain value. The contracts, controls and financing terms should follow the parties that actually control those risks.



Ross Phillipson
Partner
+61 448 535 330
RPhillipson@gtlaw.com.au



James Frixou
Special Counsel
+61 448 101 311
JFrixou@gtlaw.com.au





This publication is for information purposes only and does not constitute legal advice. If you want legal advice, you must seek specific advice tailored to your circumstances and you should not rely on this publication as a substitute for obtaining legal advice. The content is general information only, and it should be viewed as current at the time of first publication. © Gilbert + Tobin 2026.