Dedicated servers or cloud instances for AI workloads

As AI models grow larger and more complex, the compute infrastructure that runs them faces new demands. Training a single large language model or running inference on a high-traffic recommendation engine requires sustained throughput, low-latency memory access, and reliable interconnects between accelerators. Teams building these systems must decide where to run their workloads: on dedicated servers installed in-house or in colocation, or on cloud instances provisioned on demand. Each approach offers different trade-offs in cost, control, and scalability.

Understanding the compute options

Dedicated AI servers are physical machines configured with multiple AI accelerators, high-bandwidth memory, and fast interconnects. They live in a server rack that you own or lease, and you manage the hardware lifecycle yourself. Cloud instances, by contrast, are virtualised slices of a shared infrastructure. You rent them by the hour or by the month and scale their resources through an API. Both can handle model training and model inference, but the experience of using them differs significantly.

When dedicated servers make sense

Dedicated servers excel when workloads are predictable and sustained. If your team trains the same family of models day after day, or runs inference on a steady stream of requests, owning the hardware can reduce per-unit costs over time. You choose the exact number of accelerators, the memory bandwidth, and the interconnect topology. There are no noisy neighbours consuming the same memory channel or PCIe bus. Performance is consistent because the hardware is yours alone.

Another advantage is full control over the software stack. You can install custom kernels, tune the operating system for low-latency processing, and pin processes to specific cores and accelerators. For research teams experimenting with novel model architectures or unusual precision formats, this flexibility matters. You are not constrained by a cloud provider's supported framework versions or driver policies.

Dedicated servers also simplify data governance. Sensitive data never leaves your premises or your chosen colocation facility. For regulated industries such as healthcare or finance, this can be a hard requirement. You control physical access, network segmentation, and audit trails.



The downside is upfront capital expenditure and ongoing operational overhead. Hardware depreciates, and the pace of accelerator innovation means a server that feels fast today may be mid-range in two years. You must also handle cooling, power efficiency, and redundancy yourself. A single server rack can draw tens of kilowatts, so power efficiency becomes a critical part of total cost ownership. Scaling up means buying and installing more hardware, which takes weeks or months.

When cloud instances fit better

Cloud instances shine for variable or unpredictable workloads. If your training jobs come in bursts, or your inference traffic spikes at certain times of day, cloud instances let you match capacity to demand without idle hardware. You can spin up a hundred instances for a weekend training run and then tear them down. The billing model shifts from capital expense to operational expense, which many organisations prefer for experimental projects.

Cloud providers also offer a wide range of instance types, from CPU-only nodes to machines with multiple high-end accelerators. You can mix and match within the same cluster, using different accelerators for different stages of a pipeline. This flexibility is useful when you are prototyping a new model and do not yet know its hardware requirements. You can test on smaller instances and then scale up once the architecture is stable.

Another benefit is access to specialised hardware without a long-term commitment. If a new AI accelerator generation arrives, you can start using it in the cloud almost immediately. There is no need to buy, ship, and install physical units. For teams that want to evaluate multiple

accelerator types before committing to a hardware standard, cloud instances provide a low-risk testing ground.

The trade-off is less control. You share the physical host with other customers, so performance can vary if the provider oversubscribes resources. Interconnect latency between accelerators may be higher than in a tightly coupled dedicated server, especially if your workload spans multiple cloud nodes. And your data travels over the provider's network, which may raise compliance concerns.

Memory bandwidth and interconnect considerations

Both options depend heavily on memory bandwidth and interconnect performance. During model training, accelerators need to read and write large tensors quickly. A server with high memory bandwidth reduces the time spent waiting on data transfers. In a dedicated server, you can choose components that balance memory bandwidth with cost. In a cloud instance, you are limited to whatever the provider has configured.



Interconnect latency matters when you distribute a model across multiple accelerators. A training run that uses eight accelerators on one server typically communicates through NVLink or similar high-speed links. Spreading those accelerators across two or more cloud instances adds network hops, which can slow down training significantly. For models that fit on a single server, the difference is small. For very large models that require dozens or hundreds of accelerators, the interconnect topology of the cloud provider becomes a major factor.

Total cost ownership beyond the sticker price

Comparing total cost ownership between dedicated servers and cloud instances requires looking beyond the hourly rate. A dedicated server's cost includes the hardware purchase, power, cooling, facility space, and staff time for maintenance. A cloud instance's cost includes the hourly fee plus data egress charges, storage fees, and the premium for high-performance instances. Over a three-year horizon, a heavily used dedicated server can be cheaper than an equivalent cloud instance running continuously. For lightly used workloads, the cloud almost always wins on cost.

Power efficiency is part of that equation. A dedicated server running at 30% utilisation wastes most of its power budget on idle accelerators. Cloud instances let you power down resources when they are not needed, though you still pay for the time they are active. Some cloud providers now offer instances with power management features that reduce energy use during low demand, further improving efficiency.

Hybrid approaches

Many organisations run a hybrid setup. They use dedicated servers for steady, high-volume training and inference, and cloud instances for overflow traffic, experimentation, or disaster recovery. A common pattern is to maintain a base cluster of dedicated servers that handles 70% of the load, then burst into cloud instances during peak periods. This balances the cost efficiency of dedicated hardware with the flexibility of cloud scaling.

Another hybrid model uses dedicated servers for model training, where consistency and interconnects matter most, and cloud instances for model inference, where latency requirements vary and traffic patterns fluctuate. Some teams also run development and testing on cloud instances, then deploy the final model to dedicated servers for production.

Making the decision

Start by characterising your workload. Measure the average utilisation, peak demand, and growth rate. Estimate how long you plan to run the same hardware generation. If your workload is steady and you can commit to a three- to five-year hardware lifecycle, dedicated servers are worth a close look. If your workload is sporadic or you expect rapid changes in model architecture, cloud instances give you more freedom to adapt.



Also consider your team's expertise. Managing dedicated servers requires skills in hardware troubleshooting, cooling design, and capacity planning. Cloud instances require skills in orchestration, cost tracking, and security policy. Neither is inherently easier, but the skill sets are different. A team with strong hardware experience may find dedicated servers more comfortable, while a team steeped in DevOps may prefer cloud instances.

For readers exploring these options, resources like amd.com provide information on the processors, accelerators, and platforms that power both dedicated and cloud-based AI infrastructure. The same silicon often appears in both settings, so understanding the hardware capabilities helps you evaluate either path.

Finally, revisit the decision periodically. The hardware landscape changes quickly. A new generation of accelerators with higher memory bandwidth or improved power efficiency can shift the cost equation. Cloud providers also update their instance types and pricing. What made sense last year may not make sense next year. Build a process to reassess your compute strategy every 12 to 18 months.