

AI Workload Checklist

A practical intake guide to match AI workloads with the right GPU class, deployment model, data controls and support plan.

Training

Inference

Image / Video

Private AI

Use before requesting GPU capacity or private cluster sizing.

Define the workload

Complete this section first. It helps decide whether the customer needs a single node, reserved capacity, or a private cluster.

Workload Type

- ☐ Training / fine-tuning
- ☐ Batch inference
- ☐ Real-time inference API
- ☐ Image generation
- ☐ Video generation
- ☐ 3D rendering / simulation

Stage & Timeline

- ☐ Proof of concept
- ☐ Pilot with users
- ☐ Production deployment
- ☐ Urgent one-time run
- ☐ Recurring monthly workload
- ☐ Expected start date known

Output Goal

- ☐ Lower latency
- ☐ Higher throughput
- ☐ More VRAM
- ☐ Better media quality
- ☐ Cost control
- ☐ Private environment

Model and data checklist

Model Profile

- ☐ Model family identified
- ☐ Parameter size estimated
- ☐ Context length known
- ☐ Precision target known
- ☐ Framework known: PyTorch / TensorFlow / TensorRT
- ☐ Pretrained weights available

Data Profile

- ☐ Dataset size estimated
- ☐ Input type: text / image / video / 3D
- ☐ Daily data volume estimated
- ☐ Storage retention period known
- ☐ Dataset upload method known
- ☐ Backup requirement defined

Data Sensitivity

- ☐ Public / non-sensitive data
- ☐ Customer confidential data
- ☐ Personal data present
- ☐ Regulated data present
- ☐ Private cluster required
- ☐ Access approval process needed

Compute sizing checklist

Use these items to estimate GPU class, node size, storage, networking and operating support.

01

GPU Memory**VRAM per GPU required**

Model size, batch size, context length and media resolution

02

GPU Count**Single GPU or multi-GPU**

Parallel training, multi-user serving or batch generation

03

Storage**Fast NVMe / object storage**

Dataset size, checkpoints, generated media and logs

04

Networking**Standard or high-throughput**

Multi-node training, large uploads and private connectivity

05

Runtime**Container or managed stack**

CUDA, drivers, frameworks, libraries and workflow tools

GPU class fit guide

Data center-class options only. Final availability depends on region, configuration and confirmed customer requirement.

RTX PRO 6000 Blackwell SE

Universal AI, visual computing, rendering, inference and video workloads.

B200 / GB200 Blackwell

Large training, high-throughput inference and rack-scale private AI deployments.

H200 / H100

LLM training, HPC, high-memory inference and mature production AI.

L40S

Generative AI, graphics, rendering, virtual workstations and video acceleration.

L4

Energy-efficient inference, video AI, visual computing and cost-sensitive serving.

A100

Stable training, analytics, HPC and legacy production workloads.

Deployment and security checklist

Deployment Model

- ☐ Dedicated GPU node
- ☐ Reserved capacity block
- ☐ Private AI cluster
- ☐ Managed AI stack
- ☐ AI Studio backend
- ☐ Customer-managed container

Access & Controls

- ☐ Named users defined
- ☐ SSH / VPN access required
- ☐ Role-based access needed
- ☐ Private network required
- ☐ Audit trail required
- ☐ Data deletion process defined

Operations

- ☐ Monitoring required
- ☐ Support SLA needed
- ☐ Backup required
- ☐ Uptime target known
- ☐ Scaling path planned
- ☐ Billing owner assigned

Commercial readiness

01

Capacity Window

Start date, end date, expected GPU hours and peak usage period.

03

Success Metric

Latency, throughput, image/video output, training time or cost per run.

02

Budget Control

Monthly budget range, pilot cap, approval process and billing contact.

04

Expansion Trigger

When to move from pilot to reserved capacity or private cluster.

Tip: ask for sizing only after workload, data, runtime and commercial owner are clear.

Ready for a GPU capacity assessment?

Send these details to SilicaDrive so the team can match your workload to the right GPU class and deployment model.

Minimum information to submit

- ☐ Workload type and target output
- ☐ GPU type preference, if known
- ☐ Number of users or jobs
- ☐ Data size and sensitivity
- ☐ Timeline and budget range

Website: silicadrive.com
Sales: sales@silicadrive.com
Support: support@silicadrive.com

GPU Cloud | Bare Metal | Private AI Clusters | AI Studio

Use this checklist before buying, reserving or deploying GPU capacity.