

PRIVATE AI COMPUTE & SERVICES

Deploy and run AI on dedicated compute.

Your models. Your data. The support you need.

Run your AI services without buying and operating your own GPU servers.

Dedicated compute for your Private AI deployment

For system integrators, software providers and enterprise AI teams. Deliver your AI applications on our dedicated compute, with development and validation support when needed.

Dedicated resources

An agreed compute allocation for your workload.

Operational support

Compute operations and technical support, as agreed.

Flexible configurations

Start with a reference system or discuss different needs.

Run your AI. Develop and validate as needed.

01

Managed Private AI

Run your AI services. Use Pilot first if validation is needed.

For: Teams deploying Private AI services.

You: Deliver applications and manage data and workflows.

RiverSoft: Compute operations; optional software, model deployment and integration.

Setup fee plus recurring service fee

02

Private AI Workspace

Develop on dedicated compute, then validate or deploy when ready.

For: AI developers and technical SI teams.

You: Develop, test and deploy your models.

RiverSoft: Dedicated compute, basic setup and remote access.

Weekly or monthly compute access

03

Private AI Pilot

Test on a reference system before choosing a deployment configuration.

For: Teams validating deployment needs.

You: Provide workloads, data and success criteria.

RiverSoft: Compute access, scoped setup and evaluation.

Project fee with agreed deliverables

Compute matched to your workload

Start with an immediately available reference configuration, or talk to us about other GPUs, larger systems and multi-node requirements.

Immediately available reference configurations

GPU / SERVER PLATFORM	MEMORY PER GPU	MEMORY BANDWIDTH PER GPU	TOTAL GPU MEMORY
1 × NVIDIA L40S Altos BrainSphere R385 F6	48 GB GDDR6 ECC	864 GB/s	48 GB
1 × NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition Altos BrainSphere R385 F6	96 GB GDDR7 ECC	1,792 GB/s	96 GB
2 × NVIDIA RTX PRO 6000 Blackwell Server Edition Altos BrainSphere R680 F7	96 GB GDDR7 ECC	1,597 GB/s	192 GB

Higher-capacity configurations to discuss

Illustrative 8-GPU system references for larger models, training and inference.

GPU CONFIGURATION	MEMORY PER GPU	TOTAL GPU MEMORY	GPU FABRIC
8 × NVIDIA B200 Blackwell	180 GB HBM3e	1,440 GB	5th-gen NVLink
8 × NVIDIA B300 Blackwell Ultra	288 GB HBM3e	2,304 GB	5th-gen NVLink

B200 / B300 figures use NVIDIA DGX 8-GPU systems as sizing references. System platform, sourcing, availability and lead time are confirmed in the proposal.

CPU, RAM, storage, networking and software are scoped to your needs. Total GPU memory is the sum across GPUs; usable model capacity depends on software and model parallelism.

Specification references: L40S · RTX PRO Max-Q · RTX PRO Server · DGX B200 · DGX B300

Your requirements define the configuration.

Need a different GPU or more compute? Tell us what you want to run.
We'll work with you on the hardware, service scope and timeline.

www.riversoft.com.tw