RDP AI Single-Node Servers

GPU-Accelerated Inference & Development Servers

Private GenAI Copilot (RAG + Internal Knowledge) | Computer Vision Inference | AI Dev & Shared Sandbox

6 AI Server SKUs — Use-case first → Customer picks Core or Pro → Enterprise AI infrastructure starts

---

MAKE IN INDIA

Dual-socket server CPUs + AI-class GPUs — Single-node servers for inference, copilots, and team AI development environments

Single-Node. AI-Ready. Enterprise-Grade.

RDP AI Single-Node Servers bring GPU-accelerated AI infrastructure to enterprise departments — enabling private GenAI copilots, computer vision inference, and shared AI development environments without multi-node complexity.

That is why we have launched 6 AI Server SKUs across 3 enterprise use cases — covering Core and Pro configurations with dual-socket CPUs, AI-class GPUs, and 2U/4U rack-mount form factors.

Built for AI infrastructure:

✓ Single-node simplicity (No cluster overhead — single server deployments for departmental AI)
✓ GPU-accelerated inference (LLM copilots, computer vision, shared AI development)
✓ Enterprise-grade reliability (Dual PSU, remote management, ECC RAM, hot-swap drives)

Where AI Servers fit best

Private copilots: RAG over internal docs, policy Q&A, secure department AI assistants

CV inference: Production vision pipelines, multi-camera analytics, inspection systems

Shared AI dev: Team GPU sandbox, notebooks, experiments, evaluation environments

Departmental AI: Single-server deployments for 50-200 user AI services

Why single-node matters

Simpler deployment: No cluster orchestration or multi-node complexity

Lower TCO: Single server footprint with full GPU acceleration

Faster time-to-value: Deploy and start serving AI workloads immediately

Departmental scale: Right-sized for team/department AI needs

→ **Core tier: Mid-range GPU for pilots and evaluation | Pro tier: High-end GPU for production and scale**

USE CASE 1: PRIVATE GENAI COPILOT (RAG + INTERNAL KNOWLEDGE)

GenAI Core + GenAI Pro

For internal copilots, RAG over policies/SOPs, secure document Q&A, 50–200 users (depending on concurrency)

Who needs this?

Organizations deploying private GenAI copilots, RAG systems over internal documentation, secure Q&A services, department-specific AI assistants, and teams requiring on-premise LLM inference

What makes it different?

LLM inference-class GPUs with high VRAM, large RAM for context windows, fast NVMe for vector databases, network bandwidth for multi-user access, and configurations optimized for RAG pipelines

GenAI Core (Single-Node | 2U)

Model No. 911261

LLM Inference Mid | 2U Rack

CPU: Dual-socket server CPU class

RAM: 512GB (Up to 1TB)

Storage: 2×1.92TB NVMe (OS/Cache) + 4×3.84TB NVMe (Data)

GPU: LLM Inference GPU class (Mid)

Network: 2×25GbE (or 2×10GbE)

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for internal copilots, RAG over policies/SOPs, secure document Q&A, 50–200 users (depending on concurrency)

GenAI Pro (Single-Node | 2U/4U)

Model No. 912261

LLM Inference High | 2U/4U Rack

CPU: Dual-socket server CPU class

RAM: 1TB

Storage: 2×3.84TB NVMe (OS/Cache) + 6×3.84TB NVMe (Data)

GPU: LLM Inference GPU class (High / max VRAM option)

Network: 2×25/100GbE

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for higher concurrency copilots, multi-department usage, longer context workflows and always-on private AI services

Contact us: sales@rdp.in | www.rdp.in/contactus

USE CASE 2: COMPUTER VISION INFERENCE SERVER (PRODUCTION AI FOR CAMERAS)

Vision Inference Core + Vision Inference Pro

For production vision inference, multi-camera pipelines, inspection/safety analytics and stable low-latency deployments

Who needs this?

Manufacturing quality control, surveillance systems, retail analytics, smart city deployments, safety/inspection automation, and teams deploying production computer vision at scale

What makes it different?

Vision inference GPUs with high VRAM for multi-stream processing, fast storage for video datasets, low-latency network configuration, and optimized for 24/7 production inference workloads

Vision Inference Core (Single-Node | 2U)

Model No. 921261

Vision Inference Mid | 2U Rack

CPU: Dual-socket server CPU class

RAM: 256–512GB

Storage: 2×1.92TB NVMe (OS) + 4×3.84TB NVMe (Hot data)

GPU: Vision inference GPU class (Mid, higher VRAM preferred)

Network: 2×25GbE

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for production vision inference, multi-camera pipelines, inspection/safety analytics and stable low-latency deployments

Vision Inference Pro (Single-Node | 2U/4U)

Model No. 922261

Vision Inference High | 2U/4U Rack

CPU: Dual-socket server CPU class

RAM: 512GB–1TB

Storage: 2×3.84TB NVMe (OS) + 6×3.84TB NVMe (Hot data)

GPU: Vision inference GPU class (High / max VRAM option)

Network: 2×25/100GbE

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for high-throughput vision, more streams, higher resolution workloads and mission-critical inference services

Contact us: sales@rdp.in | www.rdp.in/contactus

USE CASE 3: AI DEV & SHARED SANDBOX SERVER (TEAM WORKSPACE)

AI Dev Sandbox Core + AI Dev Sandbox Pro

For shared GPU sandbox for teams, notebooks, experiments, evaluation pipelines and controlled internal AI environments

Who needs this?

AI development teams, data science groups, ML engineers sharing GPU resources, organizations building internal AI platforms, and teams running Jupyter notebooks and experiment pipelines

What makes it different?

General AI GPU for diverse workloads, large RAM for multi-user environments, ample project storage, network bandwidth for remote access, and configurations optimized for shared team usage

AI Dev Sandbox Core (Single-Node | 2U)

Model No. 931261

General AI GPU Mid | 2U Rack

CPU: Dual-socket server CPU class

RAM: 512GB

Storage: 2×1.92TB NVMe (OS) + 4×3.84TB NVMe (Projects)

GPU: General AI GPU class (Mid)

Network: 2×25GbE

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for shared GPU sandbox for teams, notebooks, experiments, evaluation pipelines and controlled internal AI environments

AI Dev Sandbox Pro (Single-Node | 2U/4U)

Model No. 932261

General AI GPU High | 2U/4U Rack

CPU: Dual-socket server CPU class

RAM: 1TB

Storage: 2×3.84TB NVMe (OS) + 6×3.84TB NVMe (Projects/Data)

GPU: General AI GPU class (High)

Network: 2×25/100GbE

Remote Mgmt: IPMI/iDRAC class | Rails included

Best for larger teams, parallel experiments, heavier datasets, multi-project usage and high headroom internal AI platform starter

Contact us: sales@rdp.in | www.rdp.in/contactus

AI Single-Node Servers Scope & Boundaries

AI Single-Node Servers Scope: Departmental AI Infrastructure

These 6 AI Single-Node Server SKUs are designed for single-server deployments — private copilots, computer vision inference, and shared AI development environments. **This is the scope of AI Single-Node Servers.**

If customer needs personal workstation GPU computing

→ **We have other business verticals who will take care of this**

- GPU Workstations: NVIDIA RTX-class tower workstations
- Desktop form factors for individual AI developers
- AI development, computer vision, and creator workflows
- Personal GPU computing for single-user scenarios

For these requirements, we'll connect you with our GPU Workstation vertical.

If customer needs multi-node clusters or datacenter AI

→ **We have other business verticals who will take care of this**

- Multi-node AI clusters: Distributed training and inference
- Pod/rack-scale deployments: 4+ node configurations
- Datacenter AI infrastructure: Full stack orchestration
- High-scale AI services: Beyond single-server capacity

For these requirements, we'll connect you with BU4 Data Center & AI Infrastructure.

Clear understanding:

AI Single-Node Servers (this portfolio): Single-server AI infrastructure for inference, copilots, and shared development
Beyond Single-Node: Personal workstations → GPU Workstation vertical | Multi-node clusters → BU4 Datacenter vertical

MAKE IN INDIA

Built for India.
Ready for AI Infrastructure.

RDP AI Single-Node Servers bring enterprise-grade GPU-accelerated infrastructure to Indian organizations — enabling private copilots, computer vision, and shared AI development while supporting Make in India procurement priorities.

6
AI Server SKUs
Across 3 Use Cases

Single-Node
Simplicity
No Cluster Overhead

100%
Made in India
Quality Assured

24/7
Enterprise Support
SLA Committed

RDP Technologies Limited
Most Affordable, High Quality, On-Time Support

Contact Sales
sales@rdp.in
www.rdp.in/contactus

Tell us your AI use case + user count + deployment needs

We will recommend the right AI Single-Node Server configuration for your organization.

Email: sales@rdp.in

www.rdp.in/contactus

Use Case 1: Private GenAI Copilot

GenAI Core: 512GB RAM, LLM GPU Mid, 2U

GenAI Pro: 1TB RAM, LLM GPU High, 2U/4U

Perfect for RAG copilots, internal docs Q&A, 50-200 users

Use Case 2: Computer Vision Inference

Vision Inference Core: 256-512GB, Vision GPU Mid, 2U

Vision Inference Pro: 512GB-1TB, Vision GPU High, 2U/4U

Perfect for multi-camera analytics, inspection systems, production CV

Use Case 3: AI Dev & Shared Sandbox

AI Dev Sandbox Core: 512GB, General AI GPU Mid, 2U

AI Dev Sandbox Pro: 1TB, General AI GPU High, 2U/4U

Perfect for team GPU sandbox, notebooks, experiments, shared AI platform

---