RDP AI Infrastructure

Enterprise AI Factory Building Blocks

GenAI Inference | Vision AI | AI Training | AI Pods

25 AI SKUs — Use-case first → CIO picks blocks → AI infrastructure starts

---

MAKE IN INDIA

NVIDIA GPUs & Intel Xeon / AMD EPYC processors — AI Pods, GPU Servers, Storage, Fabric, and Operations for enterprise AI factories

AI Building Blocks. Use-Case First. Enterprise-Ready.

RDP AI Infrastructure brings standardized, reliable building blocks to enterprise AI deployments — enabling GenAI inference, vision analytics, and model training with predictable performance and lifecycle management.

That is why we have launched 25 AI SKUs across 3 AI use cases — covering AI Pods, GPU Servers, Storage, Fabric, and Operations.

Built for enterprise AI outcomes:

✓ Use-case AI building blocks (Pick use case → Select blocks → Start AI journey)
✓ 80/20 flexibility model (80% standardized + 20% workload-specific flexibility)
✓ Predictable lifecycle (Acceptance packs + SLA framework for long-term confidence)

Where AI Infrastructure fits best

GenAI Inference: Private copilots, RAG systems, embeddings, enterprise inference services

Vision AI: Video analytics, factory safety, quality inspection, surveillance intelligence

AI Training: Fine-tuning, training runs, batch AI pipelines, model iteration

AI Pods: Complete rack-scale systems ready for CIO-level deployment decisions

Why building-block approach matters

✓ **Reduce chaos:** No more custom designs per tender — quote-ready SKUs with fixed configs

✓ **Predictable performance:** Fixed GPU/CPU/storage tiers eliminate guesswork

✓ **Lifecycle confidence:** Acceptance packs, burn-in, SLA framework included

✓ **Make in India:** Local assembly, support, and procurement compliance

Next: See AI-POD options (GenAI, Vision, Training) → Pick one → Then see which GPU/Storage/Fabric blocks are inside

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

AI INFRASTRUCTURE — RACK-SCALE SYSTEMS

AI-POD — Complete AI Factory in 1 Rack

"Pick your AI use case → Quote one AI-POD model → Get complete rack system"

What is an AI-POD?

A complete 1-rack AI system with GPU compute, hot storage, lossless fabric, acceptance pack, and ops runbooks. CIO picks one model number based on use case.

80/20 Flexibility Model

80% fixed (chassis, GPUs, storage class, fabric tier) + 20% flexible (CPU brand choice, minor storage/network adjustments without changing model identity).

AI-POD-7101 — GenAI Inference Pod

Model No. AI-POD-7101

1 Rack | GenAI Inference

Use case: Private copilots, RAG, embeddings, enterprise inference

Compute: 3× AI-GPU-72201 (4-GPU H100 80GB nodes)

Storage: AI-STG-6301B (Hot NVMe HA, ~120TB usable)

Fabric: AI-FAB-6401 (100GbE lossless HA pair)

Ops: AI-OPS-6501 (acceptance) + 6502 (runbooks) recommended

Target: CIOs, GCCs, BFSI, manufacturing HQ needing private AI copilots

AI-POD-7102 — Vision AI Pod

Model No. AI-POD-7102

1 Rack | Vision Analytics

Use case: Video analytics, factory safety, quality inspection, surveillance

Compute: 4× AI-GPU-72101 (4-GPU L40S 48GB nodes)

Storage: AI-STG-6302A (Capacity tier, ~750TB usable) + optional 6301A for indexing

Fabric: AI-FAB-6401 (100GbE lossless HA pair)

Ops: AI-OPS-6501 + 6502 recommended

Target: Manufacturing plants, warehouses, campuses, smart city integrators

AI-POD-7103 — Training / Fine-tuning Pod

Model No. AI-POD-7103

1 Rack | AI Training

Use case: Fine-tuning, training runs, batch AI pipelines, model iteration

Compute: 2× AI-GPU-75101 (HGX H100 8×80GB nodes)

Storage: AI-STG-6301B (hot tier) + AI-STG-6302A (dataset lake) recommended

Fabric: AI-FAB-6402 (200GbE lossless HA pair)

Ops: AI-OPS-6501 + 6502 + optional 6503 for managed monitoring

Target: ISVs, product engineering teams, research units, large enterprises

Each AI-POD includes rack, PDUs, management switch, cabling plan, and acceptance pack

AI-GPU SHELF — 4-GPU INFERENCE ENTRY (48GB VRAM CLASS)

AI-GPU-72101 / 72102

"RAG, copilots, embeddings, departmental inference, PoCs"

Purpose

Entry-level inference node for RAG systems, corporate copilots, embeddings, and departmental AI workloads requiring 48GB VRAM per GPU.

Fits into Pods

Works standalone OR inside AI-POD-7101 (starter) / AI-POD-7102 (Vision Pod) — default node choice for vision analytics deployments.

AI-GPU-72101 — 4-GPU Inference Node (Entry | Intel)

Model No. AI-GPU-72101

Intel Platform

Chassis: 4U GPU server, redundant hot-swap fans

CPU: 2× Intel Xeon Gold 6430 (32C each)

Memory: 512GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA L40S 48GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 2× 3.84TB U.2 NVMe

Network: 1× ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~2.5-3.5kW node power)

Recommended for Intel-preferred accounts, standardized inference workloads

AI-GPU-72102 — 4-GPU Inference Node (Entry | AMD)

Model No. AI-GPU-72102

AMD Platform

Chassis: 4U GPU server, redundant hot-swap fans

CPU: 2× AMD EPYC 9354 (32C each)

Memory: 512GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA L40S 48GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 2× 3.84TB U.2 NVMe

Network: 1× ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~2.5-3.5kW node power)

Recommended for AMD-preferred accounts, cost-optimized inference

Price-only decision summary: 4× L40S 48GB, 512GB RAM, 2×100GbE, 2×NVMe scratch

AI-GPU SHELF — 4-GPU INFERENCE CORE (80GB VRAM CLASS)

AI-GPU-72201 / 72202

"Enterprise concurrency, heavier RAG, multi-tenant inference, higher VRAM workloads"

Purpose

Core inference node for enterprise concurrency, heavier RAG workloads, multi-tenant inference services, and higher VRAM requirements per GPU.

Fits into Pods

Default node for AI-POD-7101 (GenAI Inference Pod) / Core node for AI-POD-7102 when higher inference capacity is needed.

AI-GPU-72201 — 4-GPU Inference Node (Core | Intel)

Model No. AI-GPU-72201

Intel Platform

Chassis: 4U GPU server, redundant fans, redundant PSUs

CPU: 2× Intel Xeon Gold 6454S (32C each)

Memory: 768GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA H100 80GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe

Network: ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~3.0-4.5kW node power)

Recommended for mission-critical inference, enterprise-scale RAG deployments

AI-GPU-72202 — 4-GPU Inference Node (Core | AMD)

Model No. AI-GPU-72202

AMD Platform

Chassis: 4U GPU server, redundant fans, redundant PSUs

CPU: 2× AMD EPYC 9454 (48C each)

Memory: 768GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA H100 80GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe

Network: ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~3.0-4.5kW node power)

Recommended for high-core-count workloads, AMD-preferred accounts

Price-only decision summary: 4× H100 80GB, 768GB RAM, 2×100GbE, 4×NVMe scratch

AI-GPU SHELF — 8-GPU TRAINING ENTRY (80GB TRAINING CLASS)

AI-GPU-75101 / 75102

"Fine-tuning, training runs, batch AI pipelines, ISVs/R&D teams starting training"

Purpose

Entry training node with HGX H100 (NVLink/NVSwitch) for fine-tuning, training runs, batch AI pipelines, and R&D teams starting AI model training.

Fits into Pods

Starter node for AI-POD-7103 (Training Pod) — delivers 8-GPU NVLink connectivity for efficient multi-GPU training workloads.

AI-GPU-75101 — 8-GPU Training Node (Entry | Intel)

Model No. AI-GPU-75101

Intel Platform

Chassis: 8U HGX-class 8-GPU platform, service rails

CPU: 2× Intel Xeon Platinum 8480+ (56C each)

Memory: 1TB DDR5 ECC (expandable platform dependent)

GPU: NVIDIA HGX H100 (SXM) — 8× 80GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe (high write endurance)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W 80+ Titanium PSU (~8-12kW node power)

Recommended for training workloads requiring maximum GPU interconnect bandwidth

AI-GPU-75102 — 8-GPU Training Node (Entry | AMD)

Model No. AI-GPU-75102

AMD Platform

Chassis: 8U HGX-class 8-GPU platform, service rails

CPU: 2× AMD EPYC 9654 (96C each)

Memory: 1TB DDR5 ECC (expandable platform dependent)

GPU: NVIDIA HGX H100 (SXM) — 8× 80GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe (high write endurance)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W 80+ Titanium PSU (~8-12kW node power)

Recommended for high-core-count CPU requirements, AMD-preferred training deployments

Price-only decision summary: HGX H100 8×80GB, 1TB RAM, 2×200GbE, 4×NVMe scratch

AI-GPU SHELF — 8-GPU TRAINING PRO (HIGH-MEMORY TRAINING CLASS)

AI-GPU-75301 / 75302

"Heavier fine-tuning, larger models, longer sustained training, faster checkpointing"

Purpose

Pro training node with HGX H200 (141GB per GPU) for heavier fine-tuning, larger models, longer sustained training runs, and faster checkpoint operations.

Fits into Pods

Pro/core node for AI-POD-7103 (Training Pod) when higher GPU memory capacity is required for larger model training workloads.

AI-GPU-75301 — 8-GPU Training Node (Pro | Intel)

Model No. AI-GPU-75301

Intel Platform

Chassis: 8U HGX-class platform

CPU: 2× Intel Xeon Platinum 8480+ (56C each)

Memory: 2TB DDR5 ECC (16×128GB)

GPU: NVIDIA HGX H200 (SXM) — 8× 141GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 8× 3.84TB U.2 NVMe (high endurance for checkpoint bursts)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W Titanium PSU (~10-14kW node power)

Recommended for largest model training, maximum GPU memory capacity requirements

AI-GPU-75302 — 8-GPU Training Node (Pro | AMD)

Model No. AI-GPU-75302

AMD Platform

Chassis: 8U HGX-class platform

CPU: 2× AMD EPYC 9654 (96C each)

Memory: 2TB DDR5 ECC (16×128GB)

GPU: NVIDIA HGX H200 (SXM) — 8× 141GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 8× 3.84TB U.2 NVMe (high endurance for checkpoint bursts)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W Titanium PSU (~10-14kW node power)

Recommended for high-core-count + high-memory training workloads

Price-only decision summary: HGX H200 8×141GB, 2TB RAM, 2×200GbE, 8×NVMe scratch

AI-STG SHELF — HOT NVME HA STORAGE (PERFORMANCE TIER)

AI-STG-6301A / 6301B

"Vector DB, hot corp, checkpoints, active datasets"

Purpose

Hot-tier NVMe storage with dual HA controllers for RAG/vector databases, model cache, checkpointing, and active AI datasets requiring low latency and high IOPS.

Fits into Pods

6301A: Entry hot tier for AI-POD-7101/7102 starter deployments
6301B: Core hot tier for AI-POD-7101 (larger GenAI) and default for AI-POD-7103 (Training)

AI-STG-6301A — Hot NVMe HA Storage (Entry)

Model No. AI-STG-6301A

Entry Hot Tier

Form factor: 2U, 24× U.2 NVMe bays (front)

Controllers: Dual active-active HA, hot-swappable

Media: 24× 3.84TB U.2 NVMe (enterprise), ~92TB raw

Usable: ~60-70TB (depends on RAID/EC + hot spare policy)

Connectivity: 8× 100GbE ports total (QSFP28)

Protocols: NFSv3/v4, NVMe/TCP, iSCSI

Features: Thin provisioning, snapshots, QoS, HA failover

Power: Dual redundant PSUs (~800W-1.6kW)

Default hot tier for smaller GenAI inference pods, optional fast index for Vision

AI-STG-6301B — Hot NVMe HA Storage (Core)

Model No. AI-STG-6301B

Core Hot Tier

Form factor: 2U, 24× U.2 NVMe bays (front)

Controllers: Dual active-active HA, hot-swappable

Media: 24× 7.68TB U.2 NVMe (enterprise), ~184TB raw

Usable: ~120-140TB (depends on RAID/EC policy)

Connectivity: 8× 100GbE ports total (QSFP28)

Protocols: NFSv3/v4, NVMe/TCP, iSCSI

Features: Thin provisioning, snapshots, QoS, HA failover

Power: Dual redundant PSUs (~800W-1.6kW)

Core hot tier for larger GenAI pods, default for AI-POD-7103 training checkpoints

Price-only decision summary: 6301A = ~60-70TB usable | 6301B = ~120-140TB usable

AI-STG SHELF — CAPACITY + THROUGHPUT STORAGE (RETENTION TIER)

AI-STG-6302A / 6302B

"Dataset lake, vision retention, long retention + replay"

Purpose

Capacity-tier storage with SSD cache for vision/video retention, AI dataset lakes, long-term retention, and high-throughput sequential workloads.

Fits into Pods

6302A: Default retention tier for AI-POD-7102 (Vision), optional dataset lake for AI-POD-7103
6302B: Core retention tier for larger deployments, recommended dataset lake for large training pods

AI-STG-6302A — Capacity Storage (Entry)

Model No. AI-STG-6302A

Entry Capacity Tier

Form factor: 4U, 60× LFF bays (3.5")

Controllers: Dual HA controllers, hot-swappable

Media: 60× 18TB NL-SAS HDD + 4× 3.84TB SSD cache, ~1080TB raw

Usable: ~750-850TB (depends on RAID6/EC policy + hot spares)

Connectivity: 8× 25GbE (SFP28) or 4× 100GbE (QSFP28) option

Protocols: NFSv3/v4, SMB (optional), S3 (optional)

Features: Snapshots, quotas, tiering, replication-ready

Power: Dual redundant PSUs (~1.0-2.5kW)

Default retention tier for vision analytics, dataset lake for smaller training deployments

AI-STG-6302B — Capacity Storage (Core)

Model No. AI-STG-6302B

Core Capacity Tier

Form factor: 4U base + 4U expansion (dual paths)

Controllers: Dual HA controllers, hot-swappable

Media: 120× 18TB NL-SAS (base+expansion) + SSD cache, ~2160TB raw

Usable: ~1.5-1.7PB (policy dependent)

Connectivity: 8× 25GbE or 4× 100GbE option

Protocols: NFSv3/v4, SMB (optional), S3 (optional)

Features: Snapshots, quotas, tiering, replication-ready

Power: Dual redundant PSUs per shelf

Core retention tier for multi-site vision, recommended dataset lake for large training pods

Price-only decision summary: 6302A = ~750-850TB usable | 6302B = ~1.5-1.7PB usable

AI-GPU SHELF — 4-GPU INFERENCE CORE (80GB VRAM CLASS)

AI-GPU-72201 / 72202

"Enterprise concurrency, heavier RAG, multi-tenant inference, higher VRAM workloads"

Purpose

Core inference node for enterprise concurrency, heavier RAG workloads, multi-tenant inference services, and higher VRAM requirements per GPU.

Fits into Pods

Default node for AI-POD-7101 (GenAI Inference Pod) / Core node for AI-POD-7102 when higher inference capacity is needed.

AI-GPU-72201 — 4-GPU Inference Node (Core | Intel)

Model No. AI-GPU-72201

Intel Platform

Chassis: 4U GPU server, redundant fans, redundant PSUs

CPU: 2× Intel Xeon Gold 6454S (32C each)

Memory: 768GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA H100 80GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe

Network: ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~3.0-4.5kW node power)

Recommended for mission-critical inference, enterprise-scale RAG deployments

AI-GPU-72202 — 4-GPU Inference Node (Core | AMD)

Model No. AI-GPU-72202

AMD Platform

Chassis: 4U GPU server, redundant fans, redundant PSUs

CPU: 2× AMD EPYC 9454 (48C each)

Memory: 768GB DDR5 ECC (expandable to 2TB+)

GPU: 4× NVIDIA H100 80GB (PCIe Gen5 x16)

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe

Network: ConnectX-6 Dx dual-port 100GbE

Power: 2× 3000W 80+ Titanium PSU (~3.0-4.5kW node power)

Recommended for high-core-count workloads, AMD-preferred accounts

Price-only decision summary: 4× H100 80GB, 768GB RAM, 2×100GbE, 4×NVMe scratch

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

AI-GPU SHELF — 8-GPU TRAINING ENTRY (80GB TRAINING CLASS)

AI-GPU-75101 / 75102

"Fine-tuning, training runs, batch AI pipelines, ISVs/R&D teams starting training"

Purpose

Entry training node with HGX H100 (NVLink/NVSwitch) for fine-tuning, training runs, batch AI pipelines, and R&D teams starting AI model training.

Fits into Pods

Starter node for AI-POD-7103 (Training Pod) — delivers 8-GPU NVLink connectivity for efficient multi-GPU training workloads.

AI-GPU-75101 — 8-GPU Training Node (Entry | Intel)

Model No. AI-GPU-75101

Intel Platform

Chassis: 8U HGX-class 8-GPU platform, service rails

CPU: 2× Intel Xeon Platinum 8480+ (56C each)

Memory: 1TB DDR5 ECC (expandable platform dependent)

GPU: NVIDIA HGX H100 (SXM) — 8× 80GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe (high write endurance)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W 80+ Titanium PSU (~8-12kW node power)

Recommended for training workloads requiring maximum GPU interconnect bandwidth

AI-GPU-75102 — 8-GPU Training Node (Entry | AMD)

Model No. AI-GPU-75102

AMD Platform

Chassis: 8U HGX-class 8-GPU platform, service rails

CPU: 2× AMD EPYC 9654 (96C each)

Memory: 1TB DDR5 ECC (expandable platform dependent)

GPU: NVIDIA HGX H100 (SXM) — 8× 80GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 4× 3.84TB U.2 NVMe (high write endurance)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W 80+ Titanium PSU (~8-12kW node power)

Recommended for high-core-count CPU requirements, AMD-preferred training deployments

Price-only decision summary: HGX H100 8×80GB, 1TB RAM, 2×200GbE, 4×NVMe scratch

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

AI-GPU SHELF — 8-GPU TRAINING PRO (HIGH-MEMORY TRAINING CLASS)

AI-GPU-75301 / 75302

"Heavier fine-tuning, larger models, longer sustained training, faster checkpointing"

Purpose

Pro training node with HGX H200 (141GB per GPU) for heavier fine-tuning, larger models, longer sustained training runs, and faster checkpoint operations.

Fits into Pods

Pro/core node for AI-POD-7103 (Training Pod) when higher GPU memory capacity is required for larger model training workloads.

AI-GPU-75301 — 8-GPU Training Node (Pro | Intel)

Model No. AI-GPU-75301

Intel Platform

Chassis: 8U HGX-class platform

CPU: 2× Intel Xeon Platinum 8480+ (56C each)

Memory: 2TB DDR5 ECC (16×128GB)

GPU: NVIDIA HGX H200 (SXM) — 8× 141GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 8× 3.84TB U.2 NVMe (high endurance for checkpoint bursts)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W Titanium PSU (~10-14kW node power)

Recommended for largest model training, maximum GPU memory capacity requirements

AI-GPU-75302 — 8-GPU Training Node (Pro | AMD)

Model No. AI-GPU-75302

AMD Platform

Chassis: 8U HGX-class platform

CPU: 2× AMD EPYC 9654 (96C each)

Memory: 2TB DDR5 ECC (16×128GB)

GPU: NVIDIA HGX H200 (SXM) — 8× 141GB with NVLink/NVSwitch

Boot Storage: 2× 1.92TB SSD RAID-1

Scratch: 8× 3.84TB U.2 NVMe (high endurance for checkpoint bursts)

Network: ConnectX-7 dual-port 200GbE

Power: 4× 3000W Titanium PSU (~10-14kW node power)

Recommended for high-core-count + high-memory training workloads

Price-only decision summary: HGX H200 8×141GB, 2TB RAM, 2×200GbE, 8×NVMe scratch

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

AI-STG SHELF — HOT NVME HA STORAGE (PERFORMANCE TIER)

AI-STG-6301A / 6301B

"Vector DB, hot corp, checkpoints, active datasets"

Purpose

Hot-tier NVMe storage with dual HA controllers for RAG/vector databases, model cache, checkpointing, and active AI datasets requiring low latency and high IOPS.

Fits into Pods

6301A: Entry hot tier for AI-POD-7101/7102 starter deployments
6301B: Core hot tier for AI-POD-7101 (larger GenAI) and default for AI-POD-7103 (Training)

AI-STG-6301A — Hot NVMe HA Storage (Entry)

Model No. AI-STG-6301A

Entry Hot Tier

Form factor: 2U, 24× U.2 NVMe bays (front)

Controllers: Dual active-active HA, hot-swappable

Media: 24× 3.84TB U.2 NVMe (enterprise), ~92TB raw

Usable: ~60-70TB (depends on RAID/EC + hot spare policy)

Connectivity: 8× 100GbE ports total (QSFP28)

Protocols: NFSv3/v4, NVMe/TCP, iSCSI

Features: Thin provisioning, snapshots, QoS, HA failover

Power: Dual redundant PSUs (~800W-1.6kW)

Default hot tier for smaller GenAI inference pods, optional fast index for Vision

AI-STG-6301B — Hot NVMe HA Storage (Core)

Model No. AI-STG-6301B

Core Hot Tier

Form factor: 2U, 24× U.2 NVMe bays (front)

Controllers: Dual active-active HA, hot-swappable

Media: 24× 7.68TB U.2 NVMe (enterprise), ~184TB raw

Usable: ~120-140TB (depends on RAID/EC policy)

Connectivity: 8× 100GbE ports total (QSFP28)

Protocols: NFSv3/v4, NVMe/TCP, iSCSI

Features: Thin provisioning, snapshots, QoS, HA failover

Power: Dual redundant PSUs (~800W-1.6kW)

Core hot tier for larger GenAI pods, default for AI-POD-7103 training checkpoints

Price-only decision summary: 6301A = ~60-70TB usable | 6301B = ~120-140TB usable

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

AI-STG SHELF — CAPACITY + THROUGHPUT STORAGE (RETENTION TIER)

AI-STG-6302A / 6302B

"Dataset lake, vision retention, long retention + replay"

Purpose

Capacity-tier storage with SSD cache for vision/video retention, AI dataset lakes, long-term retention, and high-throughput sequential workloads.

Fits into Pods

6302A: Default retention tier for AI-POD-7102 (Vision), optional dataset lake for AI-POD-7103
6302B: Core retention tier for larger deployments, recommended dataset lake for large training pods

AI-STG-6302A — Capacity Storage (Entry)

Model No. AI-STG-6302A

Entry Capacity Tier

Form factor: 4U, 60× LFF bays (3.5")

Controllers: Dual HA controllers, hot-swappable

Media: 60× 18TB NL-SAS HDD + 4× 3.84TB SSD cache, ~1080TB raw

Usable: ~750-850TB (depends on RAID6/EC policy + hot spares)

Connectivity: 8× 25GbE (SFP28) or 4× 100GbE (QSFP28) option

Protocols: NFSv3/v4, SMB (optional), S3 (optional)

Features: Snapshots, quotas, tiering, replication-ready

Power: Dual redundant PSUs (~1.0-2.5kW)

Default retention tier for vision analytics, dataset lake for smaller training deployments

AI-STG-6302B — Capacity Storage (Core)

Model No. AI-STG-6302B

Core Capacity Tier

Form factor: 4U base + 4U expansion (dual paths)

Controllers: Dual HA controllers, hot-swappable

Media: 120× 18TB NL-SAS (base+expansion) + SSD cache, ~2160TB raw

Usable: ~1.5-1.7PB (policy dependent)

Connectivity: 8× 25GbE or 4× 100GbE option

Protocols: NFSv3/v4, SMB (optional), S3 (optional)

Features: Snapshots, quotas, tiering, replication-ready

Power: Dual redundant PSUs per shelf

Core retention tier for multi-site vision, recommended dataset lake for large training pods

Price-only decision summary: 6302A = ~750-850TB usable | 6302B = ~1.5-1.7PB usable

AI INFRASTRUCTURE — LOSSLESS NETWORK FABRIC

AI-FAB — Lossless Fabric Options

"High-performance, lossless networking for AI workloads"

Why Lossless Fabric Matters

AI workloads generate massive east-west traffic. Lossless Ethernet with PFC/ECN ensures zero packet drops and maximum GPU utilization.

What's Included

Each AI-FAB includes: 2× HA switches, lossless tuning, QoS templates, telemetry, and complete cabling plan.

AI-FAB-6401 — 100GbE Fabric

Model No. AI-FAB-6401

Pod-Scale

- 2× ToR switches (HA pair, MLAG/vPC)
- 32×100GbE QSFP28 per switch
- Lossless profiles (PFC/ECN), QoS
- Cabling for 4× GPU nodes + storage
- Telemetry: port health, errors

Default for AI-POD-7101 (GenAI) and AI-POD-7102 (Vision)

AI-FAB-6402 — 200GbE Fabric

Model No. AI-FAB-6402

Training-Scale

- 2× ToR switches (HA pair, MLAG/vPC)
- 32×200GbE QSFP112 per switch
- Lossless profiles, QoS for checkpoints
- Cabling for 2× 8-GPU training nodes
- Higher east-west throughput

Default for AI-POD-7103 (Training) — optimized for training workloads

AI-FAB-6403 — 400GbE Spine

Model No. AI-FAB-6403

Multi-Rack

- 2× Spine switches (400GbE)
- Integrates with 6401/6402 as leaf
- Use: 2+ rack deployments
- Future-proof scaling

Optional for multi-rack AI factory expansion (3-10 racks)

AI-FAB delivers predictable performance — network doesn't bottleneck expensive GPU investments

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

AI INFRASTRUCTURE — OPERATIONS

AI-OPS-6501 — Acceptance Pack

"Go-live confidence with comprehensive validation"

Purpose

Comprehensive validation to ensure production-ready infrastructure before go-live.

Mandatory for AI-PODs

Included by default. Recommended for standalone AI-GPU deployments.

Rack + GPU Validation

- ✓ Power feed verification (A + B)
- ✓ PDU load balancing
- ✓ Thermal mapping
- ✓ GPU stress testing (multi-hour)
- ✓ PCIe Gen5 x16 checks
- ✓ NVLink/NVSwitch (HGX)

Storage + Fabric

- ✓ Throughput baseline
- ✓ Latency testing (P50/P95/P99)
- ✓ HA failover test
- ✓ Lossless profile (PFC/ECN)
- ✓ Link tests (no errors)
- ✓ MLAG/vPC redundancy

Inventory & Docs

- ✓ Serial number inventory
- ✓ MAC address mapping
- ✓ Firmware baselines
- ✓ Network topology (as-built)

Deliverables

- ✓ FAT/SAT acceptance report
- ✓ Performance baseline sheet
- ✓ Handover pack + credentials
- ✓ Known-good firmware list

Reduces post-deployment escalations by 80%+ through comprehensive validation

AI INFRASTRUCTURE — DAY-2 OPERATIONS

AI-OPS-6502 — Day-2 Runbook Pack

"Run your AI factory without chaos"

Purpose

Monitoring baselines, alerts, patch procedures, and RMA workflows for confident operations.

Recommended for AI-PODs

Strongly recommended. Reduces operational escalations by 60%+.

Monitoring Baselines

GPU: Temp, power, util, ECC

Storage: NVMe wear, IOPS, latency

Fabric: Link errors, PFC, bandwidth

System: Fans, PSU, BMC

Alert Thresholds

Critical: GPU fail, storage failover

Warning: High temps, NVMe wear

Info: Firmware drift, predictive

Escalation: L1/L2/L3 contacts

Firmware & Patching

✓ Known-good versions (GPU, BMC, BIOS)

✓ Compatibility matrix (driver/CUDA)

✓ Quarterly patch windows

✓ Rollback procedures

Spares + RMA

✓ Recommended spares list

✓ GPU RMA procedures

✓ Storage RMA workflows

✓ Response SLAs (4hr/8hr/NBD)

Deliverables

✓ Runbook PDF (100+ pages) ✓ Monitoring checklist ✓ Escalation contacts

AI-OPS-6503 / 6504 — Managed Services

"Optional managed monitoring and health checks"

AI-OPS-6503 — Managed Monitoring

Model No. AI-OPS-6503

Optional

Purpose: Light managed service

Monthly health dashboard

Proactive monitoring + alerts

Patch planning + support

Incident triage (within SLA)

Quarterly reviews

Pricing: 8-12% annual

For CIOs wanting ops peace of mind without internal GPU expertise

AI-OPS-6504 — Annual Health Check

Model No. AI-OPS-6504

Optional

Purpose: Yearly preventive maintenance

Full diagnostics (GPU, storage, fabric)

Firmware baseline refresh

Performance drift assessment

Thermal/power check

Capacity planning review

Pricing: 3-5% annual

For long-term operations protecting multi-crore investments

AI-OPS Decision Summary

AI-OPS Model	Type	When to Choose
AI-OPS-6501 Acceptance	Mandatory	Every AI-POD (included)
AI-OPS-6502 Runbooks	Recommended	All AI-PODs for ops confidence
AI-OPS-6503 Monitoring	Optional	Single accountability needed
AI-OPS-6504 Health Check	Optional	Long-term, large investments

AI Infrastructure Decision Guide

Which AI-POD should you pick?

Customer Intent	Pick This AI-POD	What's Inside
Private AI copilots / RAG / Enterprise inference	**AI-POD-7101** GenAI Inference Pod	3× AI-GPU-72201 (H100 80GB) + AI-STG-6301B + AI-FAB-6401 + AI-OPS-6501/6502
Vision analytics / Camera pipelines / Safety & QC	**AI-POD-7102** Vision AI Pod	4× AI-GPU-72101 (L40S 48GB) + AI-STG-6302A (+6301A opt) + AI-FAB-6401 + AI-OPS-6501/6502
Fine-tuning / Training / Batch AI	**AI-POD-7103** Training Pod	2× AI-GPU-75101 (HGX H100 8×80GB) + AI-STG-6301B (+6302A opt) + AI-FAB-6402 + AI-OPS-6501/6502

Which AI-GPU node should you pick standalone?

Workload Type	GPU Memory Needed	Pick This Model
Entry inference (RAG, copilots, embeddings)	48GB VRAM class	**AI-GPU-72101/72102** (4× L40S 48GB)
Core inference (enterprise concurrency, heavier RAG)	80GB VRAM class	**AI-GPU-72201/72202** (4× H100 80GB)
Training entry (fine-tuning, batch AI)	80GB + NVLink required	**AI-GPU-75101/75102** (HGX H100 8×80GB)
Training pro (larger models, sustained training)	141GB + NVLink required	**AI-GPU-75301/75302** (HGX H200 8×141GB)

80/20 Flexibility Model

✓ 80% fixed: Chassis, GPU class, storage tier, fabric class (model number stays same)
✓ 20% flexible: CPU brand choice (Intel/AMD), minor storage/network adjustments
✓ GPU availability: If named GPU is constrained, we quote equivalent tier without changing model

MAKE IN INDIA

Built for India.
Ready for Enterprise AI.

RDP AI Infrastructure brings standardized, enterprise-grade AI building blocks to Indian organizations — enabling GenAI inference, vision analytics, and model training while supporting Make in India procurement priorities.

25
AI Infrastructure SKUs
Across 3 AI Use Cases

80/20
Flexibility Model
Standardized + Flexible

100%
Made in India
Quality Assured

24/7
Enterprise Support
SLA Committed

RDP Technologies Limited
Enterprise AI Infrastructure — Made in India

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

Tell us your AI use case + workload + scale requirements

We will recommend the right AI building blocks and configuration for your enterprise AI factory.

Email: sales@rdp.in

www.rdp.in/contactus

Use Case 1: GenAI Inference

AI-POD: 7101

GPU Nodes: AI-GPU-72201/72202 (4× H100 80GB)

Storage: AI-STG-6301B (Hot NVMe HA)

Fabric: AI-FAB-6401 (100GbE)

Perfect for private copilots, RAG, enterprise inference services

Use Case 2: Vision AI

AI-POD: 7102

GPU Nodes: AI-GPU-72101/72102 (4× L40S 48GB)

Storage: AI-STG-6302A (Capacity tier)

Fabric: AI-FAB-6401 (100GbE)

Perfect for video analytics, factory safety, quality inspection

Use Case 3: AI Training

AI-POD: 7103

GPU Nodes: AI-GPU-75101/75102 (HGX H100 8×80GB)

Storage: AI-STG-6301B + 6302A

Fabric: AI-FAB-6402 (200GbE)

Perfect for fine-tuning, training runs, batch AI pipelines

RDP Technologies Limited | Nacharam, Hyderabad | Make in India IT Hardware OEM
25 AI SKUs covering 3 AI Pods + 8 GPU Servers + 4 Storage Systems + 3 Fabric Options + 4 Ops Packs