NVIDIA RTX PRO 6000 Blackwell 96GB

Complete Technical Documentation Pack for GPU Supercluster Implementation

---

1. GPU Technical Specifications

Core Architecture Details

- **GPU Architecture:** Blackwell (5nm TSMC process)
- **GPU Chip**: GB202
- **Die Size**: 750 mm²
- **Transistor Count**: 92.2 billion
- **CUDA Cores**: 24,064
- **Tensor Cores**: 752 (5th Generation)
- **RT Cores**: 188 (4th Generation)
- **Texture Units**: 752
- **ROPs**: 192
- **Base Clock**: 1,590 MHz
- **Boost Clock**: 2,617 MHz

Memory Specifications

- **Memory Type:** GDDR7 with ECC
- **Memory Capacity:** 96 GB
- **Memory Interface**: 512-bit
- **Memory Speed**: 28 Gbps effective
- **Memory Bandwidth:** 1,792 GB/s
- **L2 Cache**: 128 MB

Performance Metrics

- **FP32 Performance**: 125 TFLOPS
- **FP16 Performance**: 250 TFLOPS (500 with sparsity)
- **FP8 Performance**: 2,000 TOPS
- **FP4 Performance**: 4,000 TOPS (with sparsity)
- **AI Processing Power**: Up to 4,000 TOPS

- **Ray Tracing**: 2X performance vs previous gen

Power and Thermal

- **Total Board Power (TBP)**: 600W
- **Power Connector**: 1× 16-pin (12V-2×6)
- **Cooling Design**: Double-flow-through
- **Operating Temperature**: 0°C to 50°C
- **Form Factor**: Dual-slot
- **Dimensions**: 304mm × 137mm × 40mm

Connectivity

- **PCIe Interface**: PCIe 5.0 x16
- **Display Outputs**: 4× DisplayPort 2.1b
- **Maximum Resolution**: 16K at 60Hz, 8K at 240Hz
- **Multi-GPU**: No NVLink support (PCIe only)

Advanced Features

- **MIG Support**: Up to 4 instances
- **vGPU Support**: Yes, with vGPU 19.0
- **DLSS 4**: Multi Frame Generation
- **AV1 Encode/Decode**: Yes
- **Neural Shaders**: Integrated

Launch Details

- **Announcement Date**: March 18, 2025
- **Launch Price**: $8,565 USD
- **Availability**: Q2 2025

---

2. RTX PRO 6000 Blackwell Variants

Three Editions Available

1. Workstation Edition (600W)

- Active cooling with double-flow-through design
- For single-GPU desktop workstations
- Maximum performance configuration

2. Max-Q Workstation Edition (300W)

- Lower power for multi-GPU workstations
- Supports up to 4 GPUs per system
- Better for power-constrained environments

3. Server Edition (Configurable up to 600W)

- Passive cooling (no fans)
- Requires server chassis airflow
- Designed for datacenter deployment

---

3. Reference Documents and Downloads

Official NVIDIA Documentation

1. RTX PRO 6000 Blackwell Workstation Edition Datasheet (PDF)
   - Complete technical specifications
   - Performance benchmarks
   - Power and thermal guidelines
2. NVIDIA RTX PRO Blackwell Architecture Whitepaper (PDF)
   - Deep dive into Blackwell architecture
   - Neural rendering capabilities
   - 5th Gen Tensor Cores
   - MIG configurations
3. RTX PRO 6000 Max-Q Datasheet (PDF)
   - 300W variant specifications
   - Multi-GPU configurations

Product Pages

- RTX PRO 6000 Blackwell Workstation Edition
- RTX PRO 6000 Blackwell Server Edition
- RTX PRO 6000 Blackwell Family Overview

---

4. Server Platform Compatibility

Dell Technologies

PowerEdge R7725 (2U)

- **GPU Support**: 2× RTX PRO 6000 Server Edition
- **Processors**: Dual AMD EPYC
- **Announced**: SIGGRAPH 2025
- **Availability**: Late 2025

PowerEdge XE Series

- Various configurations supporting RTX PRO 6000
- Air and liquid cooling options

SuperMicro Solutions

20+ Systems Supporting RTX PRO 6000:

SYS-212GB-NR (MGX-based)

- **Form Factor**: 2U
- **GPU Support**: Up to 4× RTX PRO 6000
- **Design**: Single-socket for edge deployment
- **Cooling**: Air-cooled

AS-8125GS-TNHR

- **GPU Support**: 8× GPU capable
- **Processors**: Dual AMD EPYC 9004
- **Cooling**: Liquid cooling required for 8× 600W

4U GPU Systems

- Up to 8× RTX PRO 6000 Server Edition
- NVIDIA-Certified configurations
- Support for BlueField-3 and ConnectX-7

Lenovo ThinkSystem

ThinkSystem SR780a V3

- **GPU Support**: RTX PRO 6000 Server Edition
- **Form Factor**: Various (2U, 4U, 5U)

- **Cooling**: Neptune liquid cooling optional
- **Lenovo Product Guide**

HPE Solutions

ProLiant DL385 Gen11

- **GPU Support**: Up to 2× RTX PRO 6000
- **Form Factor**: 2U
- **Availability**: September 2, 2025

ProLiant DL380a Gen12

- **GPU Support**: Up to 8× RTX PRO 6000
- **Form Factor**: 4U
- **Features**: iLO 7 with Silicon Root of Trust

Additional Vendors

- Cisco UCS platforms
- ASUS GPU servers
- GIGABYTE G-Series
- Quanta Cloud Technology (QCT)
- Wistron

---

5. Networking for RTX PRO 6000 Clusters

No NVLink Support

Critical: RTX PRO 6000 Blackwell does NOT support NVLink, requiring network-centric scaling strategies

Recommended NICs

NVIDIA ConnectX-8 SuperNIC

- 800 Gb/s capability
- PCIe Gen5/Gen6 support
- Integrated DPU functions
- GPUDirect RDMA

NVIDIA ConnectX-7

- 400 Gb/s dual-port
- Current production standard
- OSFP connectors
- RoCEv2 optimized

RoCEv2 Configuration

```yaml
Network Settings:
  MTU: 9000 bytes (jumbo frames)
  DSCP: 48 for lossless traffic
  ECN: Enabled
  PFC: Priority 3
  Congestion Control: DCQCN
  Buffer: 50% for lossless class
  RDMA: GPUDirect enabled
```

Switch Infrastructure

For 10,000 GPU Clusters:

- NVIDIA Spectrum-4 switches (800 Gb/s)
- 2-layer fat-tree topology
- 1:1 oversubscription ratio
- 400 Gb/s per GPU minimum

For 50,000+ GPU Clusters:

- 3-layer Clos or Dragonfly+
- Multiple network planes
- 800 Gb/s per GPU recommended
- Optical circuit switching

---

6. Cluster Architecture Without NVLink

Pod Design Guidelines

Small Pod (512-1024 GPUs)

```
Configuration:
- 64-128 servers (8 GPUs each)
- 16-32 racks
- 2× spine switches per pod
- 400 GbE per GPU
- Single failure domain
```

Medium Pod (2048-4096 GPUs)

```
Configuration:
- 256-512 servers
- 64-128 racks
- Multi-tier switching
- 800 GbE per GPU
- Zone isolation
```

Scaling Strategies

1. **Network-centric design** due to no NVLink
2. **Data parallelism** preferred over model parallelism
3. **High-bandwidth networking** critical (400-800 Gb/s)
4. **Distributed training frameworks** required
5. **Storage performance** becomes bottleneck

---

7. Power and Cooling Infrastructure

Power Requirements at Scale

1,000 RTX PRO 6000 GPUs (600W)

- GPU Power: 600 kW
- Server Overhead: 100 kW
- Networking: 50 kW
- **Total IT Load**: 750 kW
- **With PUE 1.25**: 937.5 kW

10,000 RTX PRO 6000 GPUs

- GPU Power: 6 MW
- Server Overhead: 1 MW
- Networking: 500 kW

- **Total IT Load**: 7.5 MW
- **With PUE 1.25**: 9.375 MW

50,000 RTX PRO 6000 GPUs

- GPU Power: 30 MW
- Server Overhead: 5 MW
- Networking: 2.5 MW
- **Total IT Load**: 37.5 MW
- **With PUE 1.25**: 46.875 MW

Cooling Specifications

Heat Output: 2,047 BTU/hr per GPU (600W)

Air Cooling:

- Workstation Edition: Active double-flow-through
- Server Edition: Passive, requires chassis airflow
- Maximum 4-8 GPUs per rack air-cooled

Liquid Cooling (Recommended for scale):

- Direct-to-chip for higher density
- Supports 16+ GPUs per rack
- Required for full 600W operation in servers

Double-Flow-Through Cooling Design

- Two separate airflow paths
- Optimized for 600W sustained operation
- Eliminates thermal throttling
- Quiet operation under load

---

8. Storage Integration

Recommended Solutions

- **VAST Data**: Universal storage platform
- **WEKA**: GPU-optimized parallel filesystem
- **DDN AI400X2**: Purpose-built for AI
- **Ceph**: Open-source option

Performance Requirements

- **Per GPU**: 2-5 GB/s bandwidth
- **Latency**: <100 microseconds
- **Capacity**: 200-500 GB active dataset per GPU
- **Protocol**: NFS over RDMA or GPUDirect Storage

---

9. Software Stack

Driver Requirements

- **Driver Version**: R550 or newer
- **CUDA Toolkit**: 12.4 or later
- **vGPU Software**: 19.0 or later

MIG Configuration

```bash
# Enable MIG mode
nvidia-smi -mig 1

# Create MIG instances (up to 4)
nvidia-smi mig -cgi 19,19,19,19 -C

# Note: Some users report MIG issues in current drivers
```

Framework Support

- PyTorch 2.3+
- TensorFlow 2.15+
- JAX with CUDA 12.4
- RAPIDS for data science

---

10. Pricing and TCO

Component List Pricing

- **RTX PRO 6000 Blackwell**: $8,565
- **ConnectX-7 400GbE**: $3,000-5,000
- **ConnectX-8 800GbE**: $8,000-12,000

- **400GbE Switch**: $50,000-150,000
- **Server (8-GPU)**: $50,000-100,000

10,000 GPU TCO (3-Year)

CapEx:

- GPUs: $85.65M
- Servers: $70M
- Networking: $40M
- Storage: $20M
- Infrastructure: $30M
- **Total**: ~$245M

OpEx (Annual):

- Power (9.4 MW): $7M
- Cooling: $1.5M
- Staff (15 FTE): $3M
- Maintenance: $12M
- **Total**: ~$23.5M/year

3-Year Total: ~$315M

---

11. RTX PRO Server Configurations

2U Mainstream Servers (New)

- 2× RTX PRO 6000 GPUs
- Most popular form factor
- Available from all major OEMs
- Ideal for edge/branch deployment

4U High-Density

- 4-8× RTX PRO 6000 GPUs
- Liquid cooling recommended
- Current availability

MGX Platform

- NVIDIA reference design
- 2-4 GPUs in compact form
- Edge-optimized

---

12. Deployment Best Practices

Critical Considerations

1. **No NVLink** - must architect around PCIe limitations
2. **600W power** - double that of RTX 6000 Ada
3. **Liquid cooling** recommended for >4 GPU systems
4. **PCIe 5.0** required for full bandwidth
5. **MIG support** may have driver issues initially

When to Choose RTX PRO 6000 Blackwell

✅ **Ideal for:**

- Large model inference (96GB memory)
- Graphics + AI workloads
- Enterprise virtual desktops
- Cost-sensitive vs H100/H200
- Edge AI deployments

❌ **Not ideal for:**

- Workloads requiring NVLink
- Power-constrained facilities
- Small-scale deployments (<100 GPUs)
- Pure training workloads (consider H100/H200)

Key Takeaways

1. **96GB GDDR7 memory** - Largest professional GPU memory
2. **600W TDP** - Requires robust power/cooling
3. **No NVLink** - Network becomes critical bottleneck
4. **$8,565 launch price** - Premium positioning

5. **Blackwell architecture** - Latest generation
6. **MIG support** - Up to 4 instances (driver dependent)
7. **March 2025 launch** - Early adoption phase
8. **Wide OEM support** - All major vendors onboard

For organizations requiring maximum memory capacity and willing to invest in proper power/cooling infrastructure, the RTX PRO 6000 Blackwell offers compelling capabilities, especially for inference and mixed AI/graphics workloads.