Finalized GPU Server Package Versions (LLM Infrastructure)

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Component / Package | Version | Purpose |
| System/GPU Stack | NVIDIA Driver | 550.54.14 | Compatible with H200 GPUs |
| System/GPU Stack | CUDA Toolkit | 12.3.2 | GPU computing support for DL frameworks |
| System/GPU Stack | cuDNN | 8.9.5 | Deep learning primitives for CUDA |
| System/GPU Stack | NCCL | 2.19.4 | Multi-GPU communication layer |
| AI Frameworks | PyTorch | 2.2.2+cu123 | Primary framework for LLM training/inference |
| AI Frameworks | TensorFlow | 2.15.0 | Optional framework (for non-HF workloads) |
| AI Frameworks | HuggingFace Transformers | 4.40.0 | LLM models and inference |
| AI Frameworks | Accelerate | 0.28.0 | Easy multi-GPU setup for training |
| AI Frameworks | DeepSpeed / Megatron-LM | latest | For distributed LLM training (optional) |
| Model Ops | MLflow | 2.12.1 | Experiment tracking, model registry |
| Model Ops | Weights & Biases (wandb) | 0.16.6 | Optional - visual experiment tracking |
| Model Ops | TorchServe / TF Serving | 0.9.1 / 2.15.0 | Model serving backends |
| Orchestration | Kubernetes (via OCP) | 1.27+ | Workload scheduling (OpenShift 4.14+) |
| Orchestration | Kubeflow | 1.8.0 | ML pipeline orchestration, model serving |
| Orchestration | Run:AI CLI / Agent | 2.15.0 | GPU scheduling, quotas, and isolation |
| Orchestration | Argo Workflows | 3.5.7 | Pipeline automation (used under Kubeflow) |
| Orchestration | Airflow (Optional) | 2.9.0 | Task orchestration for pipelines |
| Storage | NetApp Trident CSI | 24.02.0 | OpenShift + NetApp volume provisioning |
| Storage | S3 CLI / MinIO Client | latest | Access object storage |
| Monitoring | Prometheus + DCGM Exporter | 2.48.1 / 3.3.2 | GPU + infra monitoring |
| Monitoring | Grafana | 10.3.1 | Dashboarding |
| Monitoring | Run:AI UI + Metrics Plugin | latest | Cluster & GPU metrics via Run:AI |
| Security & Dev | Trivy | latest | Image vulnerability scanning |
| Security & Dev | JupyterLab | 4.1.5 | Notebook interface |
| Security & Dev | VS Code Server | latest | Web IDE for debugging and dev |
| Security & Dev | Helm | 3.14.x | Kubernetes chart deployments |
| Security & Dev | Kubectl | 1.27.x | Kubernetes CLI |