COREDGE

Dflare AI Datasheet

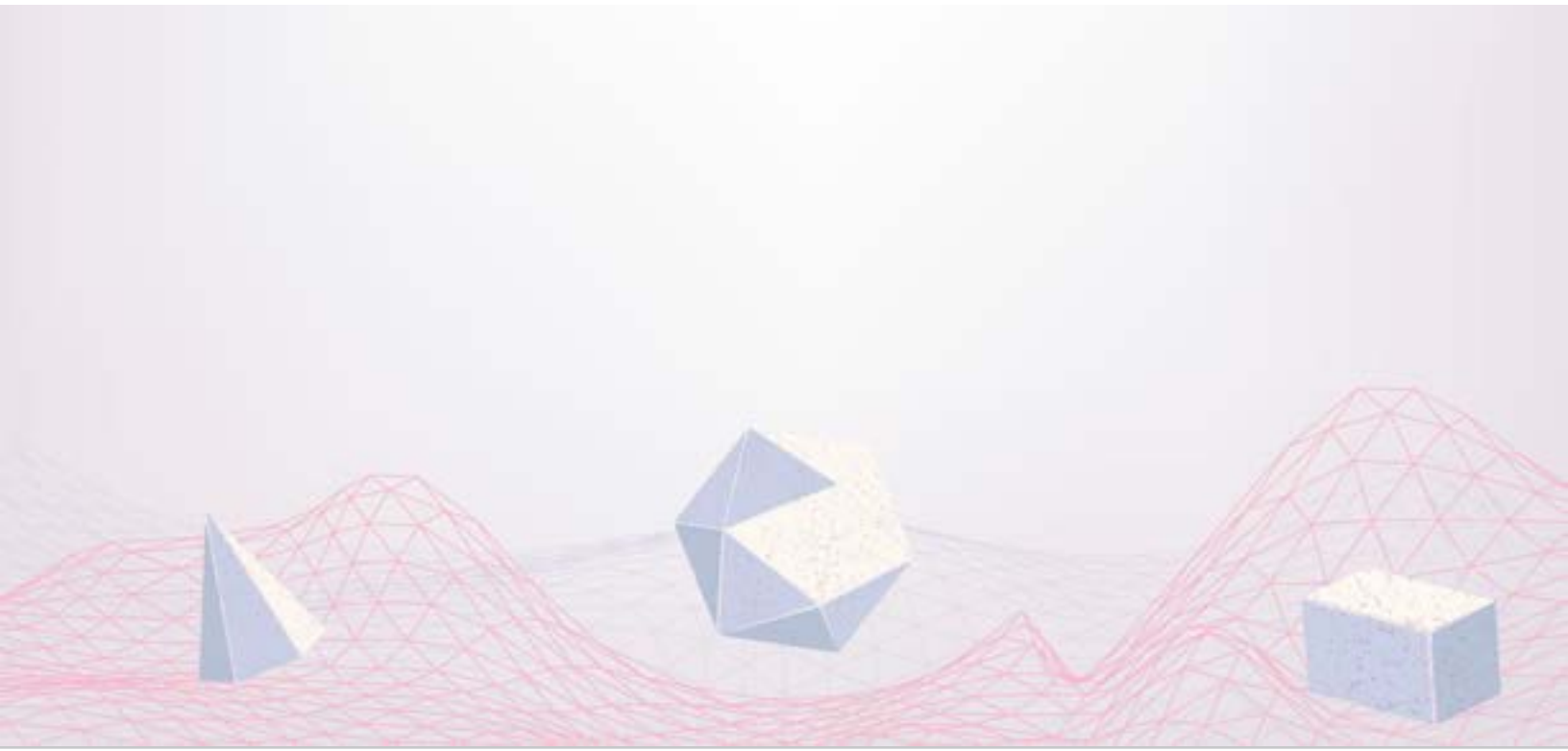
Datasheet

Transforming bare metal GPU servers into production-ready, multi-tenant compute environments

Delivering the raw power of dedicated hardware with the operational simplicity of a managed cloud service.

Version 1.0

Contents



Modern AI workloads require massive GPU scale, high-throughput data pipelines, strict multi-tenant isolation, and support for both cloud-native and HPC workloads. Traditional cloud and on-prem systems fail to deliver all four simultaneously. Traditional cloud providers impose virtualization overhead and noisy-neighbor effects that degrade training throughput, while on-premises HPC clusters lack the self-service provisioning and lifecycle management that modern AI teams require.

Dflare AI was purpose-built to address this gap — providing unified GPU infrastructure that combines bare metal performance, hardware-enforced isolation, and full lifecycle automation.

Platform Overview

Dflare AI is an enterprise GPU infrastructure platform designed to deliver bare metal performance with cloud-like usability. The platform is composed of three primary layers:

Access Layer — Portal UI, REST APIs, CLI

Control Plane — Workflow Orchestrator, Cluster Manager, Network Manager, Identity & Access, Monitoring & Metering

Data Plane — GPU nodes, Kubernetes / Slurm workloads, InfiniBand fabric, Parallel filesystem

ML Platform — GPU notebooks, distributed training, LLM inference, fine-tuning, experiment tracking, dataset management

Dflare AI efficiently manages bare metal GPU nodes across any infrastructure, offering a unified experience for application and infrastructure lifecycle management. The platform caters to enterprises, AI labs, managed service providers, and government organizations, assisting in GPU provisioning, workload orchestration, security, and billing.

Use Cases

1

Large-Scale Model Training

Multi-node, multi-GPU distributed training across hundreds of GPUs with RDMA-based InfiniBand interconnect. Scale AI training from 8 to 10,000+ GPUs without architectural changes, achieving near-linear scaling.

2

Unified Kubernetes + Slurm Orchestration

Run both containerized and Slurm workloads on the same bare metal infrastructure with unified networking, storage, security, and billing — managed through a single control plane.

3

GPU-as-a-Service for Multi-Tenant Environments

Offer dedicated bare metal GPU clusters to multiple tenants with hardware-level isolation at InfiniBand switch (partition key), filesystem (access control map), and network fabric (VRF/VXLAN).

Key Features

Bare Metal GPU Performance

Direct GPU access without virtualization overhead. Hardware-level BIOS and OS tuning pre-applied via golden images. Near-native efficiency (99%+ of bare metal GPU peak). Direct GPU access without virtualization overhead. Hardware-level BIOS and OS tuning pre-applied via golden images. Near-native efficiency (99%+ of bare metal GPU peak). Standard GPU slicing via NVIDIA MIG (Multi-Instance GPU) profiles enables efficient resource utilization.

Observability

GPU utilization, cluster health, and job performance monitoring. Metrics collectors, time-series database, and dashboards for real-time observability across compute, network, and storage layers.

Dual-Fabric Network Architecture

Ethernet (VXLAN/EVPN/BGP) for control plane and tenant VPCs. InfiniBand for RDMA-based high-performance GPU-to-GPU and GPU-to-storage communication with Partition Keys for tenant isolation.

Zero Trust Security

Defense-in-depth security model. OAuth2/OIDC identity, RBAC + ABAC authorization, VRF + VLAN + PKey network isolation, storage ACLs, and TLS 1.2+ / mTLS encrypted service-to-service transport.

Central IAM and Multi-Tenancy Controls

Integrates with enterprise identity providers (OIDC/OAuth2). Issues short-lived JWTs, enforces RBAC/ABAC at every API boundary, and isolates tenants via dedicated realms and scoped tokens for enhanced multi-tenancy support.

Unified Kubernetes + Slurm

Single control plane manages Kubernetes and Slurm. Kubernetes, powered by CKP (Coredge Kubernetes Platform), handles containerized workloads via device plugins. Supports CNCF Certified Kubernetes versions (1.29 - 1.35). Slurm handles batch workloads with GPU-aware scheduling using GRES. Both share the same underlying nodes, storage, and network.

ML Platform

Integrated machine learning environment with GPU notebooks, distributed training, LLM inference with OpenAI-compatible APIs, model fine-tuning, experiment tracking with MLflow, and dataset management.

Key Benefits

Automated Lifecycle Management

From bare metal power-on to production cluster — fully automated. No SSH, no manual configuration. Provision > operate > monitor > bill.

Near-Bare-Metal Performance

Eliminate virtualization overhead — achieve near-bare-metal performance with cloud-like operational simplicity. Zero hypervisor, direct device access.

Hardware-Level Tenant Isolation

Isolation at InfiniBand switch hardware (partition key), filesystem (access control map), and network fabric (VRF/VXLAN). Enforce strict multi-tenancy through hardware and software controls.

Comprehensive Observability

Real-time observability across compute, network, and storage layers enables proactive capacity management and rapid incident response.

Scalable Multi-Cluster Management

Scale AI training from 8 to 10,000+ GPUs without architectural changes. Horizontally scalable GPU nodes with leaf-spine fabric expansion.

Zero Trust Security

Defense-in-depth security model based on zero-trust principles ensures a high level of security for the managed GPU infrastructure.

Get in touch with us

https://coredge.io info@coredge.io