TECHNICAL ARCHITECTURE

SilicaDrive GPU Cloud Technical Architecture

A clean reference architecture for GPU cloud, bare-metal nodes, AI Studio, inference APIs, developer workspaces and private enterprise clusters.

GPU Cloud | Bare Metal | AI Studio | Private AI

REFERENCE STACK

Layered platform architecture

Each layer has one clear role: customer experience, AI services, orchestration, execution, and platform telemetry.

Customer Portal

Login, plans, credits, invoices, tickets and usage visibility

AI Studio & Inference APIs

Generation tools, model endpoints and customer-facing APIs

Developer Workspace

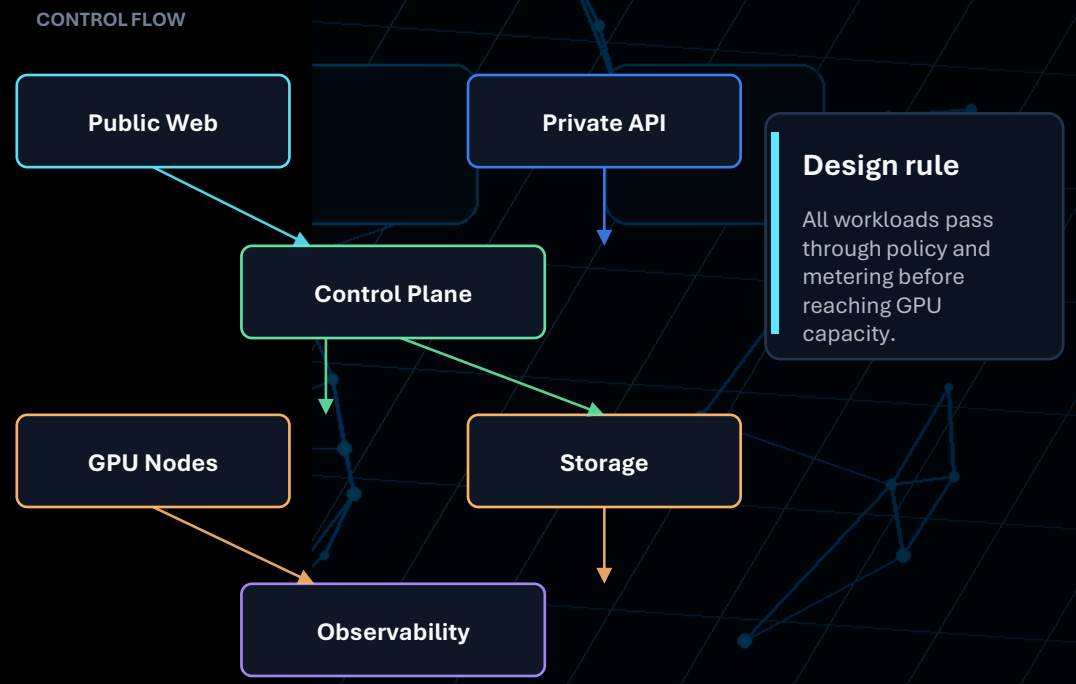
Containers, jobs, notebooks, datasets and environment templates

GPU Control Plane

Scheduling, quotas, metering, policy checks and capacity allocation

Bare-Metal GPU Fabric

Dedicated nodes, storage, network, power and cooling



CORE MODULES

What the platform needs to run well

The architecture separates product features from control-plane responsibilities and infrastructure operations.

Portal

Accounts, plans, billing, support and usage views

API Gateway

Authentication, rate limits and request routing

Policy Engine

Safety rules, quotas, credit checks and access limits

Scheduler

Job queue, capacity selection and worker assignment

Storage

Datasets, outputs, logs, checkpoints and backups

Telemetry

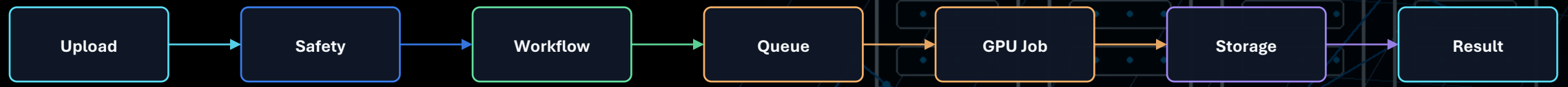
Metrics, events, audit logs, alerts and customer visibility

Result: a platform where customers see simple AI tools while SilicaDrive controls security, metering, GPU utilization and reliability behind the scenes.

AI STUDIO

Image and video generation pipeline

A production-ready AI Studio needs safe input handling, workflow selection, GPU job execution, and traceable result delivery.



SUPPORTING SERVICES

Asset Library

Customer uploads, reference images, generated files and metadata.

Workflow Library

Text-to-image, image-to-video and video workflow templates.

Model Registry

Approved models, versions, memory profiles and quality tiers.

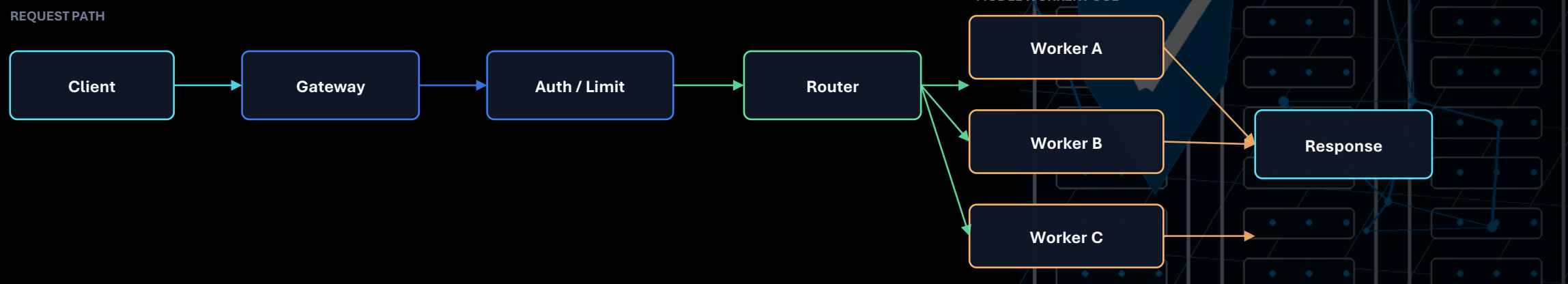
Credit Metering

Per-job cost, runtime events, output history and billing record.

INFERENCE API

Low-latency model serving design

Inference endpoints should isolate tenants, keep hot model pools ready, and meter every request without exposing GPU nodes.



Controls

API keys, per-plan limits, request validation and abuse rules.

Serving

Warm model pools, batching windows and fallback queues.

Reliability

Health checks, retries, timeouts and graceful degradation.

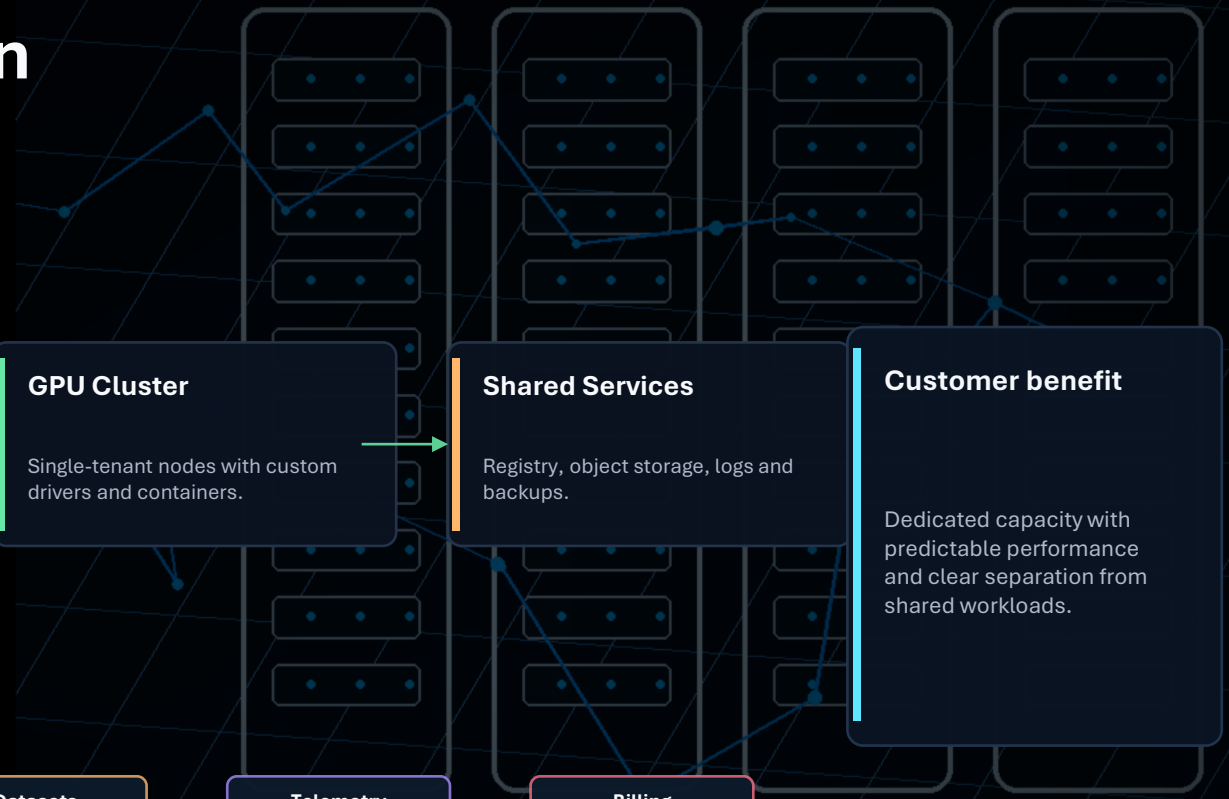
Metering

Request, token, image, video, runtime and bandwidth events.

PRIVATE CLUSTERS

Dedicated GPU capacity design

Enterprise customers can run isolated workloads on dedicated GPU nodes with controlled access and managed support.



ENTERPRISE ACCESS

Customer VPC / VPN

Private access, allowlist and secure routing.

Jump Host

SSH keys, admin access and audit trail.

GPU Cluster

Single-tenant nodes with custom drivers and containers.

Shared Services

Registry, object storage, logs and backups.

Customer benefit

Dedicated capacity with predictable performance and clear separation from shared workloads.

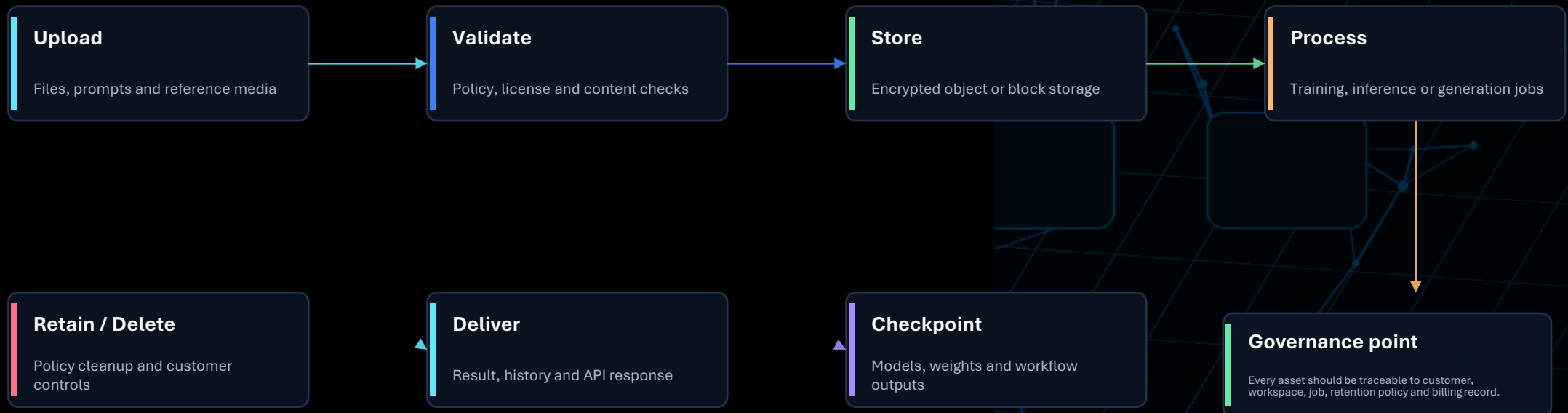
OPERATING CONTROLS

- Access Policy
- Secrets
- Images
- Datasets
- Telemetry
- Billing

DATA & MODEL LIFECYCLE

Governed movement of assets

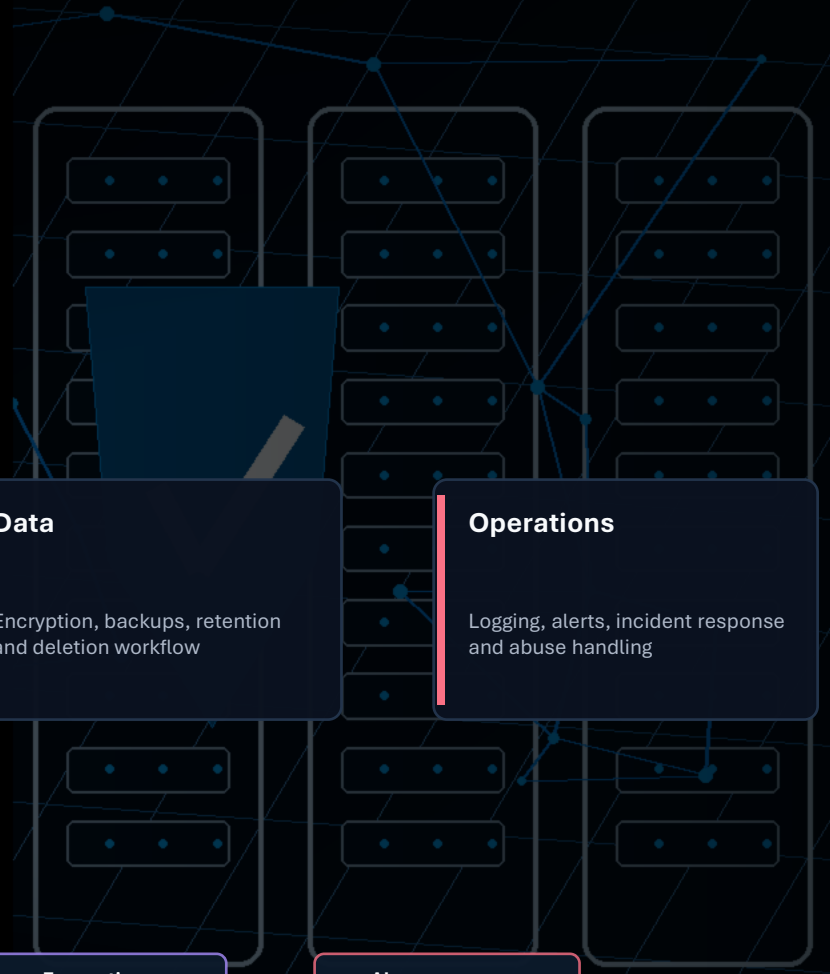
Uploads, datasets, checkpoints, generated outputs and logs need clear ownership, encryption, retention and deletion controls.



SECURITY ARCHITECTURE

Defense-in-depth controls

Security should be layered across identity, network boundaries, compute isolation, data protection and operational response.



Identity

MFA, API keys, roles and workspace boundaries

Network

Segmentation, firewall rules, WAF and private endpoints

Compute

Tenant isolation, hardened images and controlled drivers

Data

Encryption, backups, retention and deletion workflow

Operations

Logging, alerts, incident response and abuse handling

DEFAULT PRINCIPLES

- Least privilege
- Tenant isolation
- Auditability
- Secure defaults
- Encryption
- Abuse response

Operating note

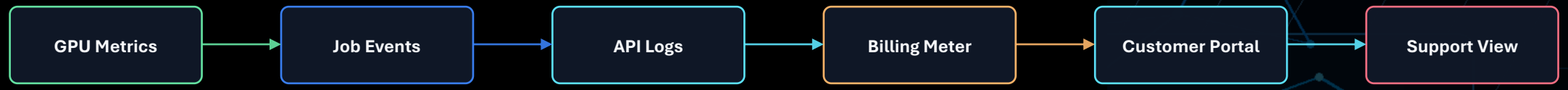
Keep portal, API gateway and approved inference endpoints separate from internal GPU management networks.

OBSERVABILITY & BILLING

Operational telemetry converted into customer visibility

Monitoring feeds support, capacity planning, SLA tracking, abuse detection, usage history and credit-based billing.

EVENT PIPELINE



Track

GPU hours, VRAM, storage, bandwidth, jobs and API requests.

Interpret

Capacity planning, customer transparency and anomaly detection.

Alert

Errors, saturation, suspicious traffic and abnormal spend.

Show

Usage history, credits, invoices, logs and status.

IMPLEMENTATION ROADMAP

Build, secure and scale the platform

A practical sequence for moving from customer portal and AI Studio into inference APIs and dedicated enterprise clusters.

01

Portal & credits

Login, plans, credits, invoices, support and usage visibility.

02

AI Studio MVP

Text-to-image, image-to-video, job queue and output library.

03

Inference APIs

Authenticated endpoints, model pools, rate limits and billing events.

04

Private clusters

Dedicated GPU nodes, enterprise access controls and managed support.

Website: silicadrive.com Sales: sales@silicadrive.com Support: support@silicadrive.com