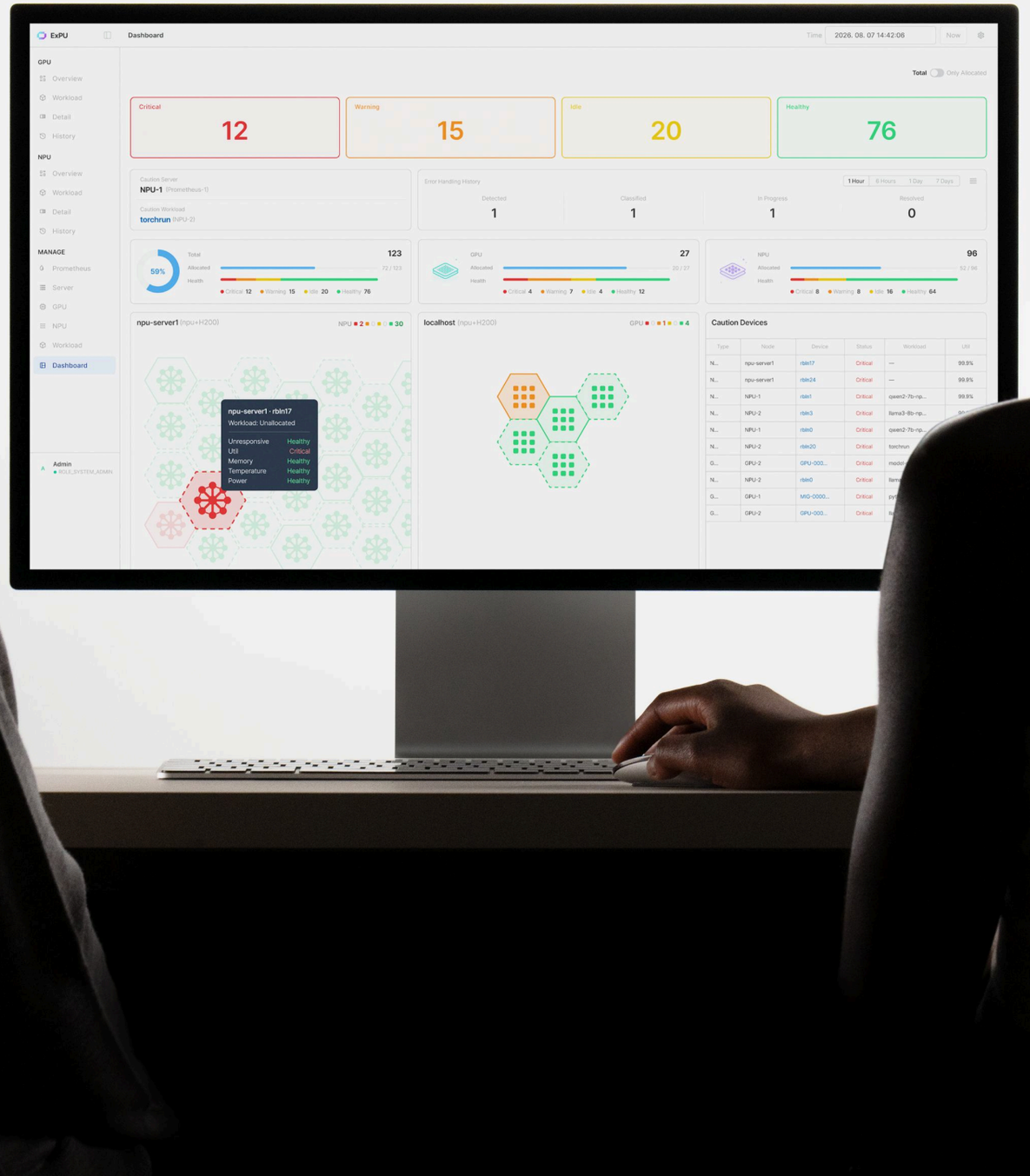
ExPU

Beyond monitoring, GPU and NPU resources managed as one

From allocation to reclaim, without complex commands

As more companies adopt AI, accelerators like GPUs and NPUs are multiplying fast. But when each device needs its own commands and tools, it is hard to tell where capacity sits idle and where it is overloaded.

ExPU unifies resource status and workload occupancy across heterogeneous accelerators, so checking, allocating and reclaiming all happen in one place.



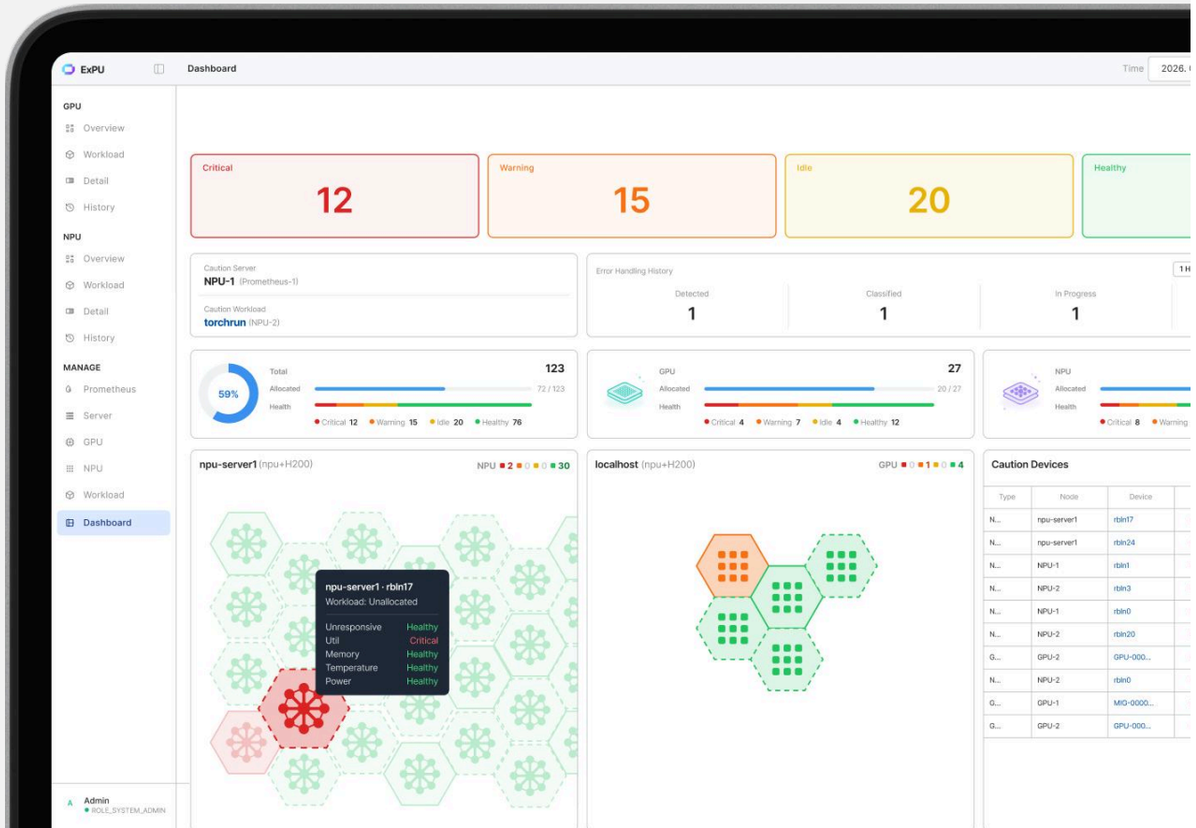
ExPU

Why ExPU stands out

AI accelerator resource management, from monitoring to action

ExPU is a unified resource management platform for AI accelerators. It monitors GPUs and NPUs together and allocates, partitions and reclaims per workload.

Without looking up complex commands, operators see resource status, anomalies and workload occupancy on one screen and act right away.



Global Standard

Proven at scale

EXEM covers the full stack, from databases to cloud, AI and big data, and grows with 1,100+ customers across 29 countries.

Countries served
29+ countries

Cumulative customers
1,100+ customers

Product Highlights

Key capabilities

Unified control of GPU and NPU in one platform

Manage mixed accelerator fleets under one scheme and spot devices that drift from normal patterns early.

Two-way analysis of workloads and devices

Map devices held per workload and workloads per device, and visualize utilization, temperature and power together.

Resource management from monitoring to action

Link allocation and reclaim with clicks, not commands, and return idle-but-held resources to the available pool at once.

Transparent detection from codified expertise

Field diagnostic know-how becomes rules that watch 24/7 and leave an auditable trail.

Find and reclaim resources held without being used

Detect occupancy left behind with no workload attached and return it to the pool instantly.

GPU partitioning by click, not by CLI

Set MIG, MPS and Time-Slicing modes on screen and check running processes.

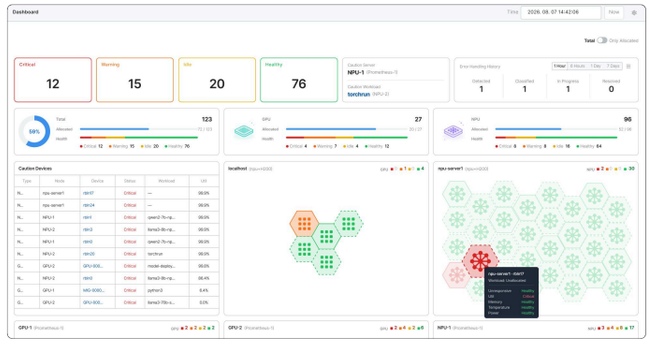
Unified monitoring

Every accelerator in the cluster on one screen: monitor live and get ahead of failures.

1 Unified live status

Cluster-wide uptime and allocation

- Check GPU and NPU status and allocation on a single screen
- Monitor all cluster resources in real time
- Catch anomalies early: temperature, power and unresponsive devices



2 Workload-level tracking

Track the accelerators each workload holds

- See per-workload usage and hold time
- Compare resource use across workloads
- Spot resource-heavy workloads fast

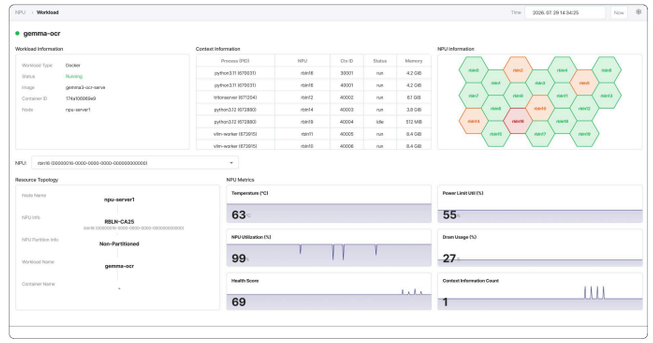
Resource analysis

Analyze usage from both the workload and the device side to find waste and decide when to reclaim.

1 Two-way workload mapping

Map devices held per workload and workloads running per device

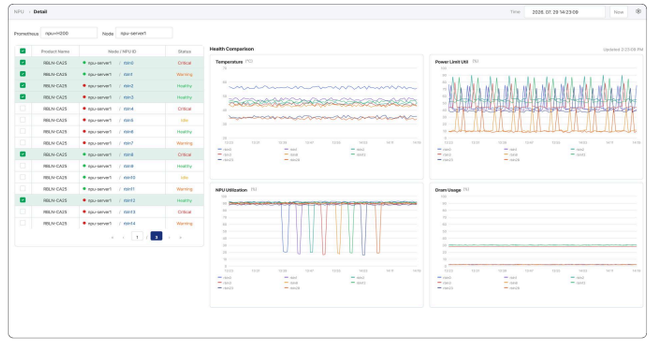
- See which workload uses which device, live
- See workloads running on each device, live



2 Utilization, temperature, power

Key KPIs as heatmaps and line charts

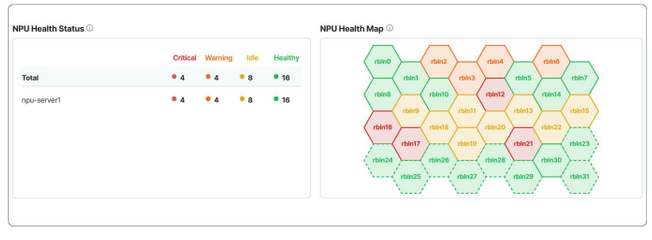
- Utilization, temperature and power in one view
- Device state alongside workload compute state



3 Four-level resource diagnosis

Critical, Warning, Idle and Healthy classes

- Act on failures before they land
- Identify abandoned (Idle) workloads
- Device uptime plus evidence for reclaim calls



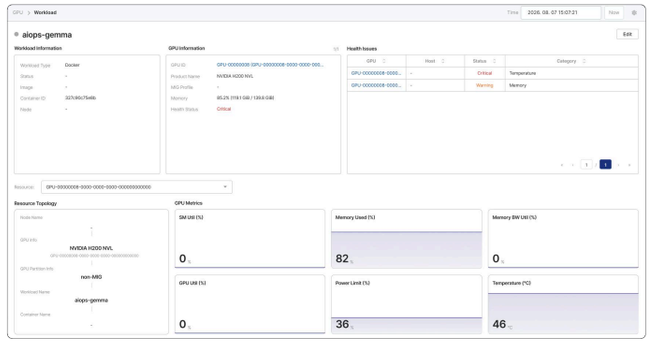
Resource control

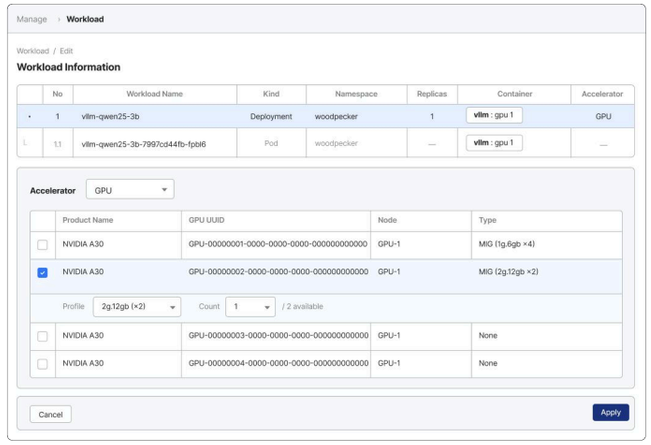
Act on resources from the screen where you found them, handling allocation and reclaim without switching tools.

1 Act on resources in a click

Applied safely via standard container APIs

- One workflow from detection to control
- Review the target workload, then apply at once
- Request changes on screen, no separate tool

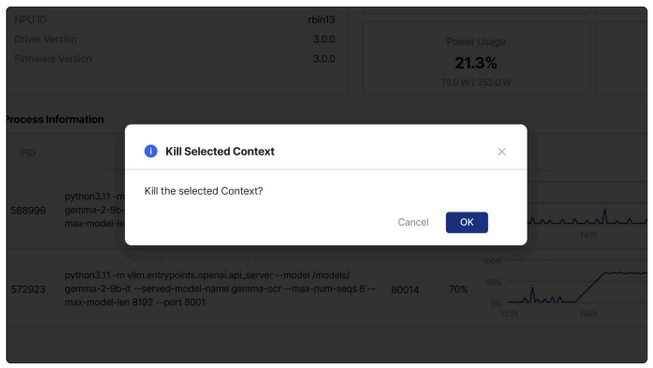




2 Reclaim ghost occupancy

Clear occupancy with no workload attached

- Spot resources held but unused on screen and reclaim them at once
- Cut idle holds and lift effective utilization



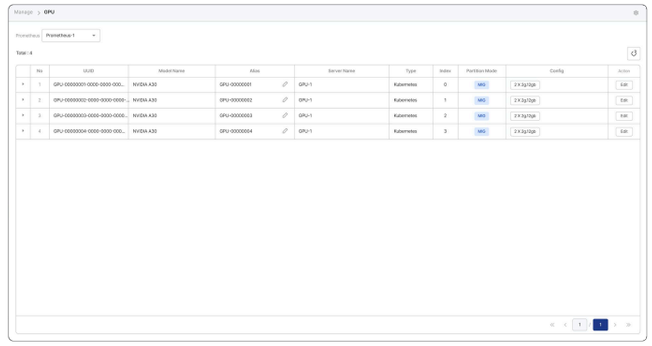
GPU & NPU features

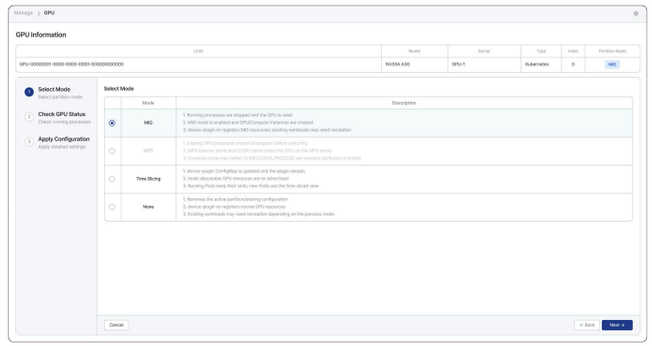
Set GPU partitioning on screen without complex CLI, and catch the errors that matter the moment they occur.

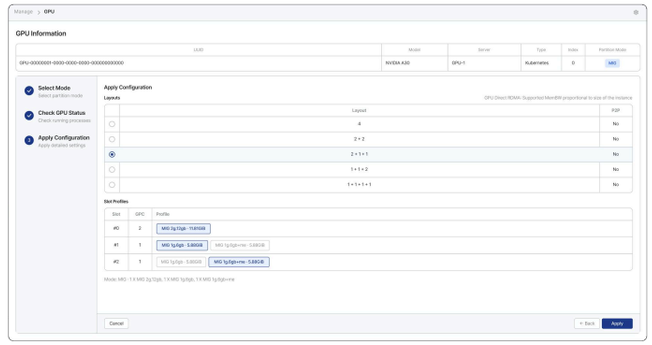
1 Partitioning setup GPU

MIG, MPS, Time-Slicing set by click, not CLI

- Click the GPU to partition and pick a mode
- Check running processes
- Apply per-mode settings instantly



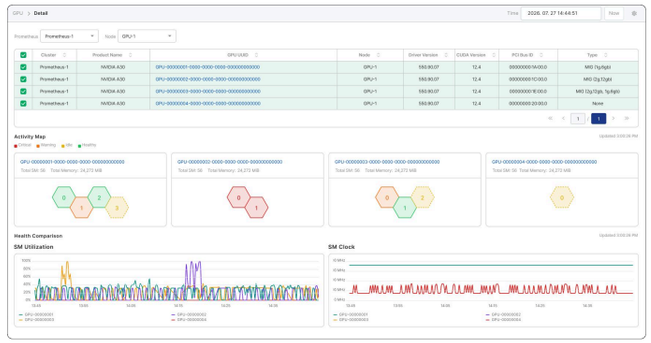




2 Partitioning hexagon view GPU

Partition modes visualized as a hexagon view

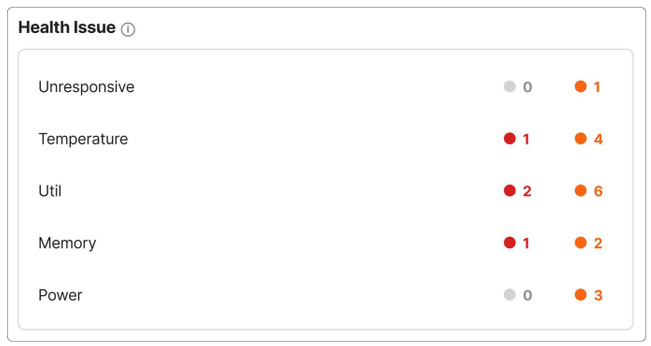
- MIG, MPS, Time-Slicing distribution at a glance
- Compare GPU-specific metrics per device



3 Rule-based anomaly detection

One-off critical errors vs sustained metrics

- Unresponsive is caught the moment it occurs
- Temperature and power checked for duration
- Identify devices outside normal range early



4 Baseline deviation detection

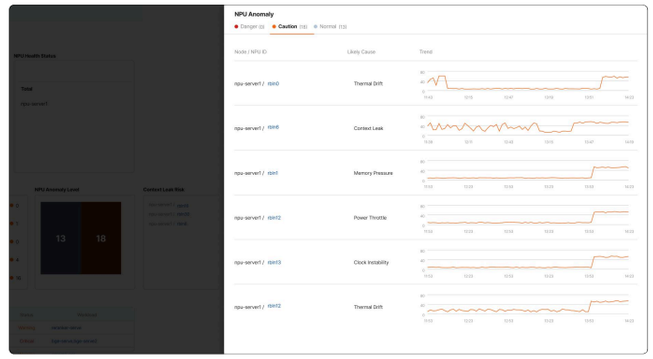
Learns each device's normal, flags drift early

- Statistical shifts from average usage patterns
- Detects baseline drift with no fixed thresholds

5 Failure cause candidates NPU

Abnormal occupancy after a workload ends

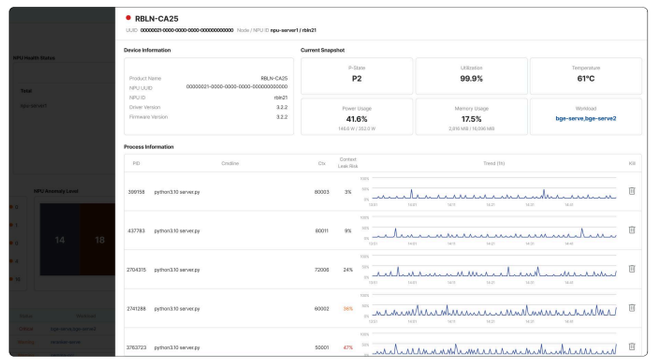
- Find leftover holds and mark them for reclaim
- Track leak signals such as memory exhaustion ETA and process memory



6 Context Leak risk analysis NPU

Metric correlation flags devices to reclaim

- Identify risk devices from leak evidence
- Act before failure, lift infrastructure efficiency



Architecture

How ExPU is deployed

1 Monitored servers

Exporter on every accelerator server

- GPU servers: NVIDIA Driver, ExPU Exporter
- NPU servers: RBLN Driver, ExPU Exporter
- To scale out, install the Exporter and register it as a Prometheus target

2 ExPU server

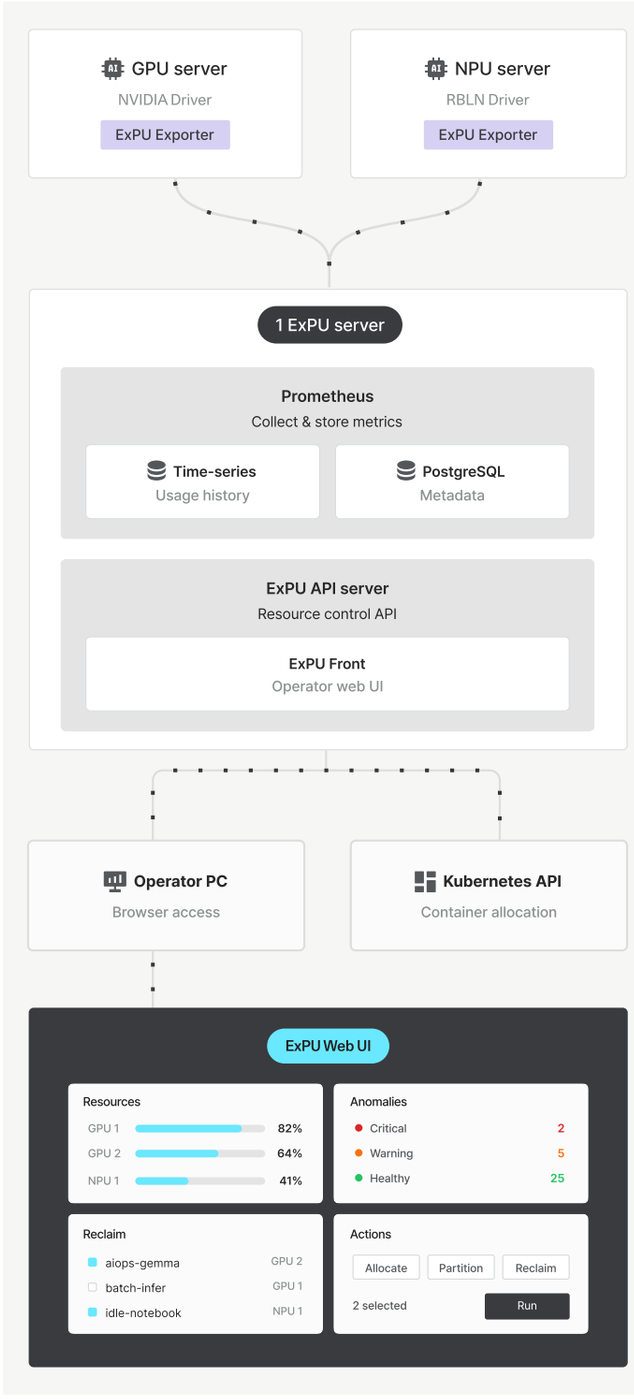
Collection, storage, API, UI in one

- Prometheus stores metrics as time series
- PostgreSQL stores resource config and metadata
- ExPU API server and Front on the same host

3 Access and integration

Browser access, API integration

- Operators reach the Web UI from a browser
- Kubernetes API for Pod/Container allocation
- Operational data stays inside your network



Platform Specs

Supported environments

Web browser

Browser : Chrome, Edge
Resolution : 1920×1080 / 1440×900 min

ExPU monitoring server

OS : Linux (Rocky/RHEL, CentOS, etc.)
CPU : 2 Core (min) / 4 Core (recommended)
RAM : 8GB (min) / 16GB (recommended)
DISK : SSD 300GB or more
- OS 100GB
- Metrics 200GB or more

SW : Docker, Prometheus, PostgreSQL, Kubernetes
NODE : 1 server
Placement : avoid servers under heavy GPU/NPU load

Monitored servers

Collected : GPU/NPU status, utilization, temperature, memory, processes, Pod/Container mapping
Installed : ExPU Exporter, vendor Exporter
Exporter footprint
- CPU : about 1 Core
- RAM : about 2GB
- DISK : 10GB or more

NODE : every server with a GPU or NPU

* Values without a separate note are the minimum required to run ExPU.
Recommended specs vary with the scale of the managed fleet and the metric retention period.

www.ex-em.com

Data Everywhere, Make it Matter

40, Magokjungang 8-ro 5-gil, Gangseo-gu, Seoul, Republic of Korea
TEL 02-6203-6300 **FAX** 02-6203-6301

exem