

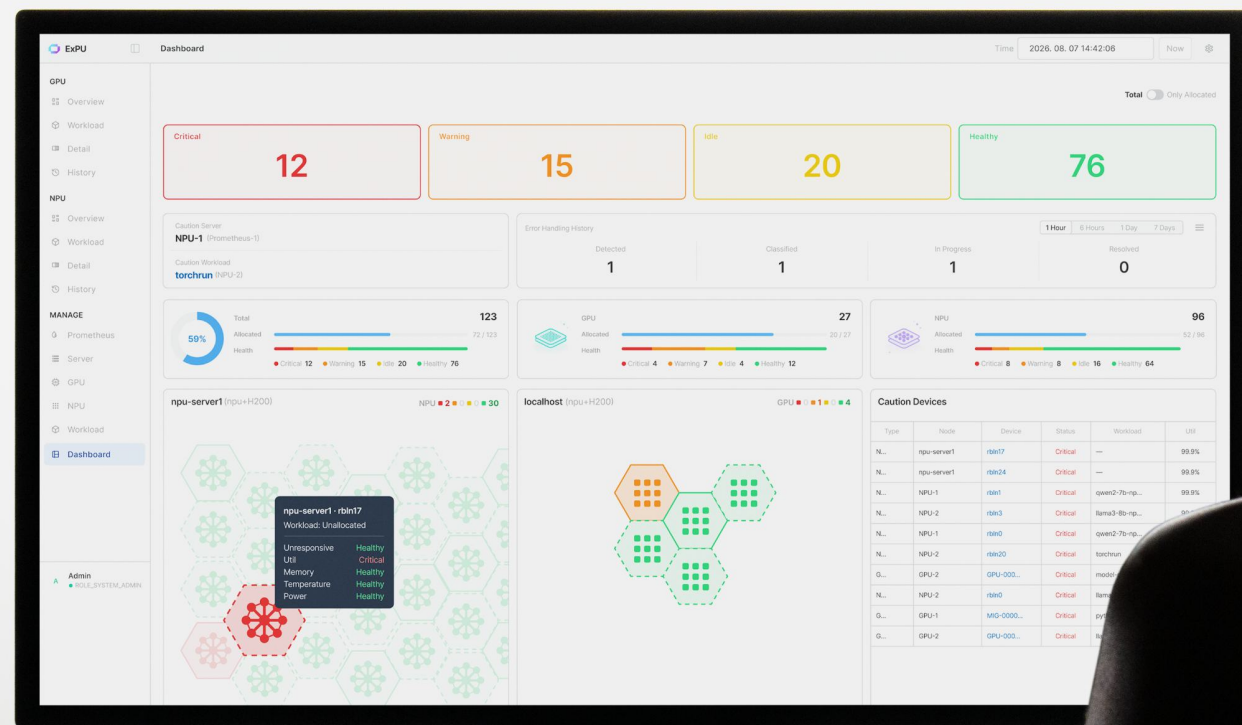


모니터링을 넘어 GPU와 NPU 자원을 하나로 통합 관리하다

복잡한 명령어 없이, 클릭만으로 할당부터 회수까지

AI를 도입하는 기업이 늘면서 GPU와 NPU 같은 가속기도 빠르게 늘고 있습니다. 하지만 장비마다 다른 명령어와 도구로 상태를 확인하다 보면 어디에 자원이 남고 어디가 과부하인지 파악하기 어렵습니다.

ExPU는 이기종 가속기의 자원 현황과 워크로드 점유 상태를 하나의 플랫폼에서 통합해, 확인부터 할당과 회수까지 이어서 처리합니다.



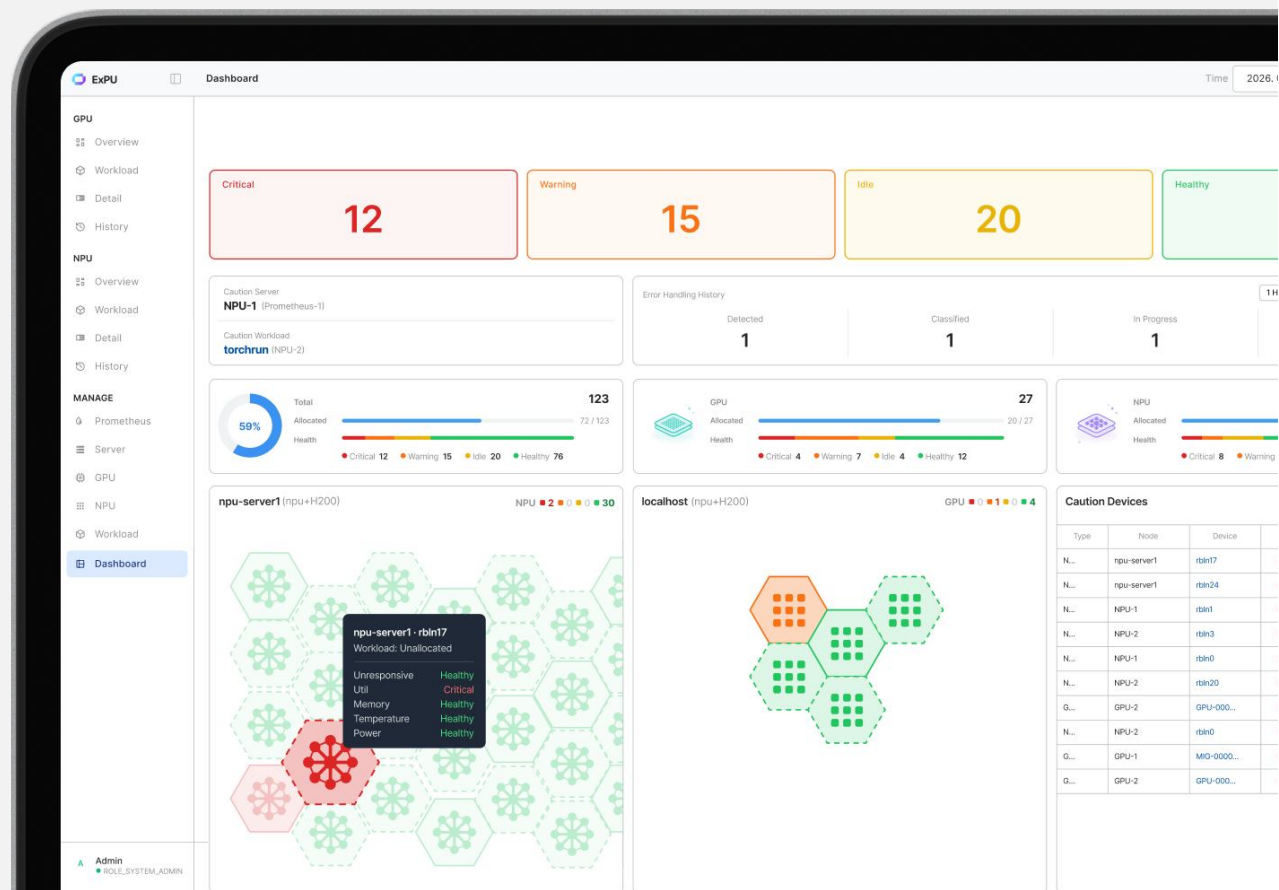


엑스피유가 특별한 이유

모니터링에서 조치까지 이어지는 AI 가속기 자원 관리

ExPU는 GPU와 NPU 등 이기종 가속기를 통합 모니터링하고, 워크로드 단위로 자원을 할당, 분할, 회수하는 AI 가속기 통합 자원 관리 플랫폼입니다.

사용자는 복잡한 명령어를 확인하지 않고 한 화면에서 자원 현황과 이상 징후, 워크로드 점유 상태를 파악하고 곧바로 조치할 수 있습니다.



Global Standard

검증된 기술력

엑셀은 데이터베이스를 비롯한 시스템 전 구간에서 클라우드, AI, 빅데이터까지. 전세계 29개국, 1,100개 이상의 고객사와 함께 성장하고 있습니다.

서비스 제공 국가

29개국 +

누적 고객사 수

1,100개사 +



Product Highlights

제품 특징점

GPU와 NPU를 하나의 플랫폼에서 통합 관제

종류가 다른 가속기가 혼재되어도 하나의 체계로 관제하고, 정상 패턴에서 벗어난 디바이스를 조기에 식별합니다.

워크로드와 디바이스를 양방향으로 분석

워크로드별 점유 디바이스와 디바이스별 실행 워크로드를 매핑하고, 이용률, 온도, 전력을 함께 시각화합니다.

모니터링에서 조치까지 이어지는 자원 관리

복잡한 명령어 없이 클릭만으로 자원 할당과 회수까지 연결합니다. 점유만 남은 자원도 찾아 즉시 회수해 사용 가능한 자원으로 되돌립니다.

진단 노하우를 물로 코드화한 투명한 이상 탐지

현장의 진단 노하우를 물로 전환해 24시간 감시하고, 판정 근거를 남겨 추적하고 감사할 수 있습니다.

쓰지 않는데 잡혀 있는 자원을 찾아 회수

연결된 워크로드 없이 남은 점유를 클릭으로 탐지해, 사용 가능한 자원 풀로 즉시 되돌립니다.

CLI 없이 클릭으로 GPU 파티셔닝 설정

MIG, MPS, Time-Slicing 모드를 화면에서 즉시 설정하고, 실행 중 프로세스를 확인합니다.

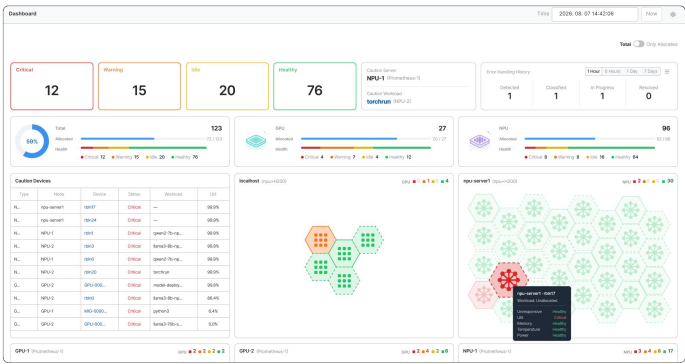
실시간 통합 모니터링

클러스터 전체 가속기의 가동 상태와 자원 할당 현황을 한 화면에 통합해 실시간으로 모니터링하고 선제적으로 장애에 대응합니다.

1 실시간 통합 현황

클러스터 전체 가속기의 가동 상태와 할당 현황을 통합

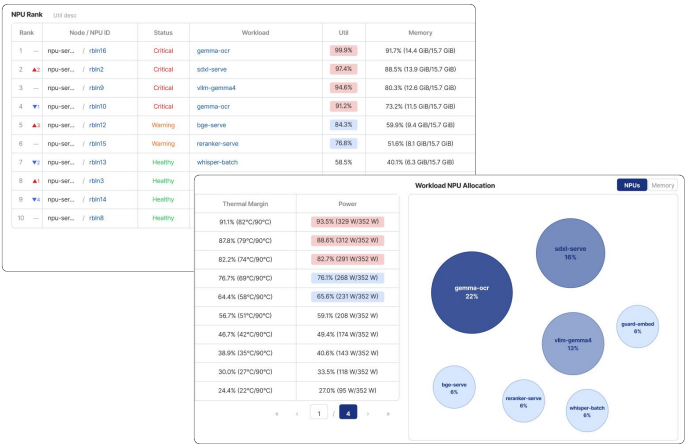
- GPU, NPU 가동 상태와 자원 할당 현황을 한 화면에서 확인
- 클러스터 전체 자원을 실시간으로 모니터링
- 온도와 전력 이상, 응답 불가 장비 등 이상 징후를 조기에 포착



2 워크로드 단위 추적

워크로드별로 어떤 가속기를 점유하는지 추적

- 워크로드별 가속기 사용량과 점유 시간을 확인
- 여러 워크로드의 자원 사용을 한눈에 비교
- 자원을 많이 쓰는 워크로드를 빠르게 식별



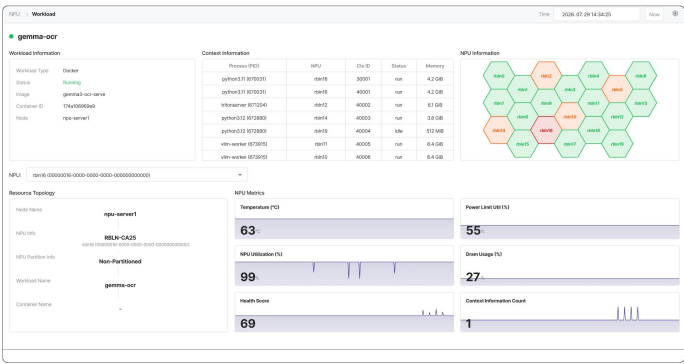
자원 분석

워크로드와 디바이스 양쪽 관점에서 자원 사용을 분석해, 비효율 자원과 회수 시점을 판단합니다.

1 워크로드와 디바이스 양방향 매핑

워크로드별 점유 디바이스와 디바이스별 실행 워크로드를 매핑

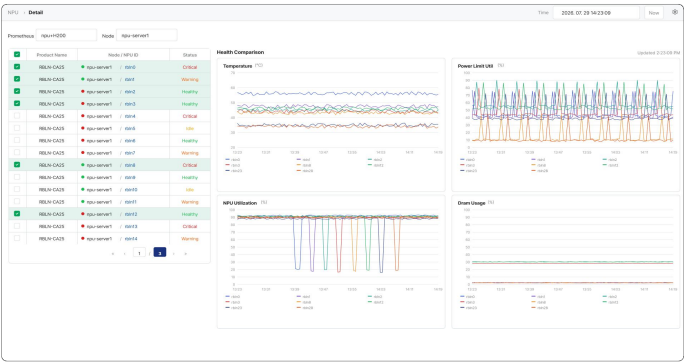
- 어떤 워크로드가 어떤 장비를 쓰는지 실시간으로 확인
- 디바이스에서 실행 중인 워크로드를 실시간으로 확인



2 이용률, 온도, 전력 시각화

관점별 핵심 KPI를 히트맵과 꺾은선으로 시각화

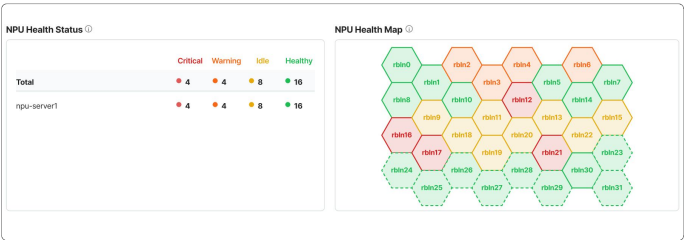
- 이용률, 온도, 전력을 한 화면에서 확인
- 디바이스 물리 상태와 워크로드 연산 상태를 함께 확인



3 자원 상태 4단계 진단

Critical, Warning, Idle, Healthy로 분류해 선제 대응

- 상태 분류로 장애에 선제적으로 대응
- 방치된 워크로드(Idles)를 식별
- 디바이스 가동률과 회수 판단 근거 제공

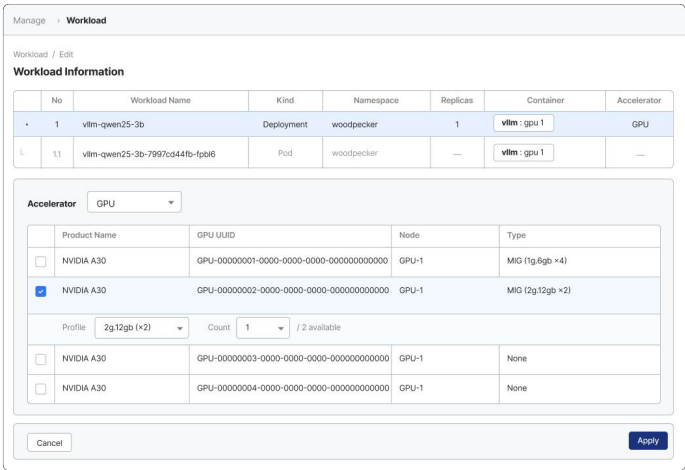
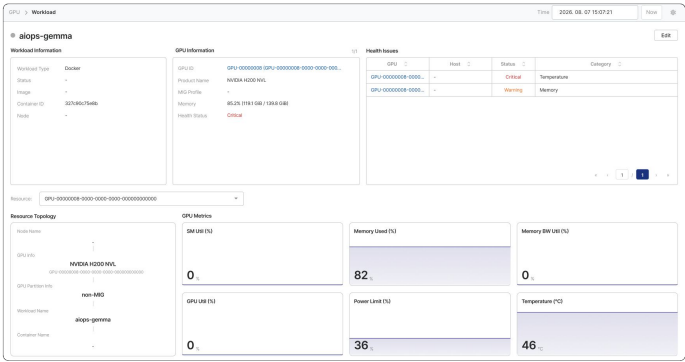


📌 자원 제어

확정한 화면에서 곧바로 자원을 조치해, 별도 도구로 이동하지 않고 할당과 회수를 처리합니다.

1 클릭만으로 자원 조치

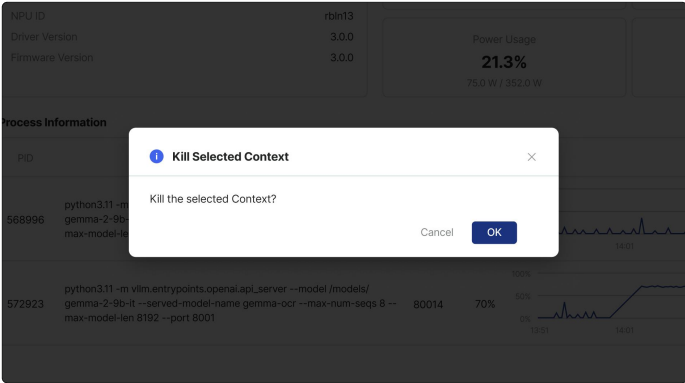
- 컨테이너 환경의 표준 API로 안전하게 적용
- 이상 징후 감지부터 자원 제어까지 단일 워크플로우
- 변경할 워크로드의 현황을 확인하고 즉시 적용
- 별도 도구로 이동하지 않고 화면에서 자원 조정 요청



2 유령 점유 회수

연결된 워크로드 없이 남은 점유를 정리

- 쓰지 않고 점유만 된 자원을 화면에서 확인하고 즉시 회수
- 방치된 점유를 줄여 실효 가동률을 높임



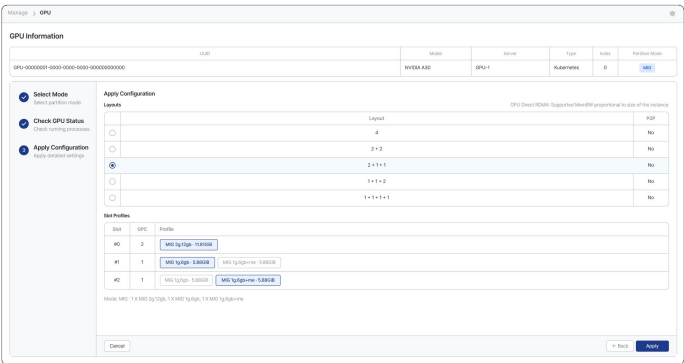
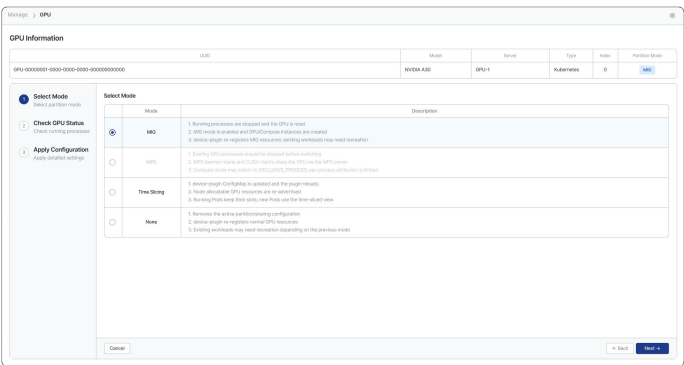
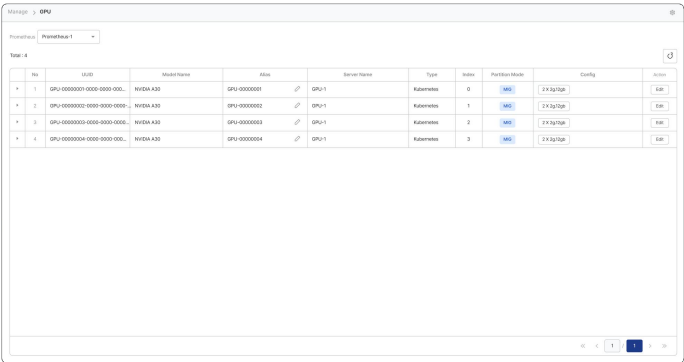
⚙️ GPU와 NPU 특화 기능

복잡한 CLI 없이 화면에서 GPU 파티셔닝을 설정하고, 발생 자체가 중요한 오류를 감지합니다.

1 파티셔닝 설정 GPU

CLI 없이 클릭으로 MIG, MPS, Time-Slicing 설정

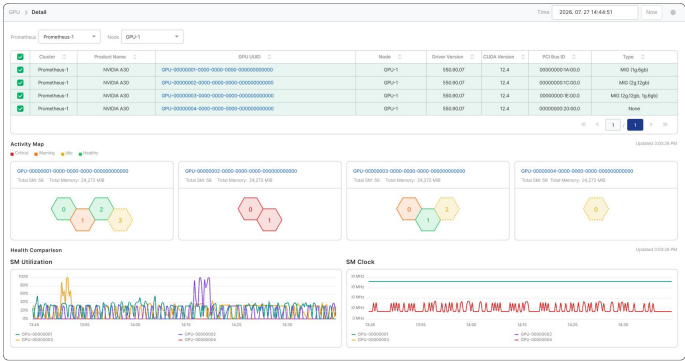
- 파티셔닝할 GPU를 클릭해 모드 선택
- 실행 중 프로세스를 확인
- 모드별 세부 설정을 즉시 적용



2 파티셔닝 현황 hexagon 뷰 GPU

파티셔닝 모드 현황을 hexagon 뷰로 가시화

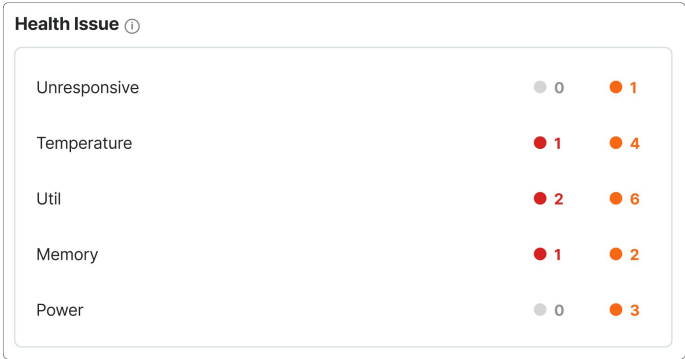
- MIG, MPS, Time-Slicing 분배 현황을 한눈에 파악
- GPU 특화 지표를 디바이스별로 비교



3 룰 기반 이상 탐지

발생 자체가 핵심인 오류와 지속 지표를 구분해 감시

- Unresponsive는 발생 자체를 즉시 감지
- 온도와 전력은 일정 시간 지속 여부를 확인
- 정상 범위를 벗어난 디바이스를 조기에 식별



4 Baseline 이탈 탐지

장비별 정상 기준을 스스로 확인해 이상을 조기 탐지

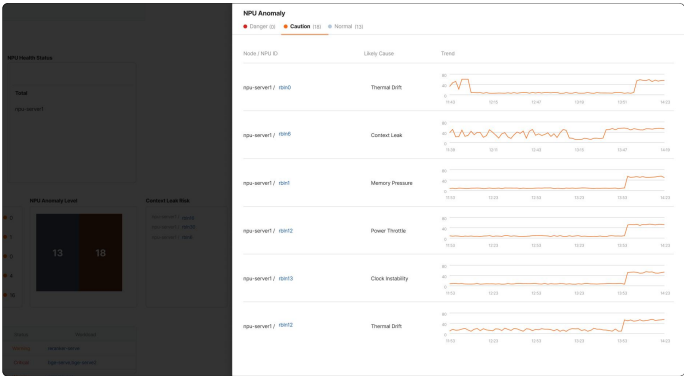
- 평균 사용 패턴 기반 통계적 비정상 변화 포착
- 고정 임계치 없이 Baseline 이탈을 자동 감지



5 장애 원인 후보 분석 NPU

워크로드 종료 후 남은 비정상 점유를 탐지

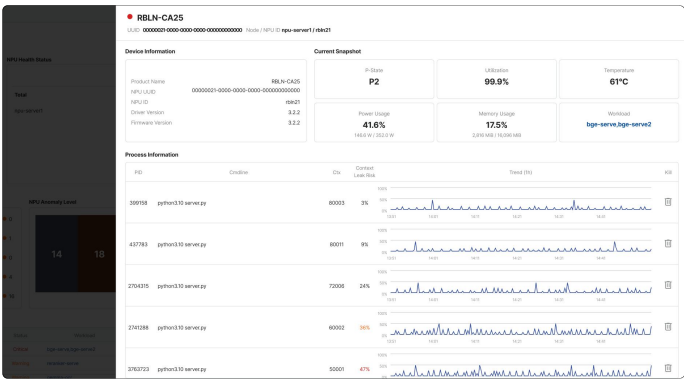
- 종료 후에도 남은 점유 상태를 탐지해 회수 대상 식별
- 메모리 고갈 임박 ETA와 프로세스 메모리 등 누수 지표 추적



6 Context Leak 위험 장비 분석 NPU

메트릭 상관관계로 회수 대상 장비 판정

- 누수의 직접 증거를 기반으로 위험 장비 식별
- 장애로 이어지기 전에 대응해 인프라 효율을 높임



Architecture

구조도

1

모니터링 대상 서버

- 가속기를 장착한 서버마다 Exporter 설치
- GPU 서버에 NVIDIA Driver와 ExPU Exporter 설치
 - NPU 서버에 RBLN Driver와 ExPU Exporter 설치
 - 서버 확장 시 Exporter 설치 및 Prometheus 수집 대상 등록만으로 모니터링 시작

2

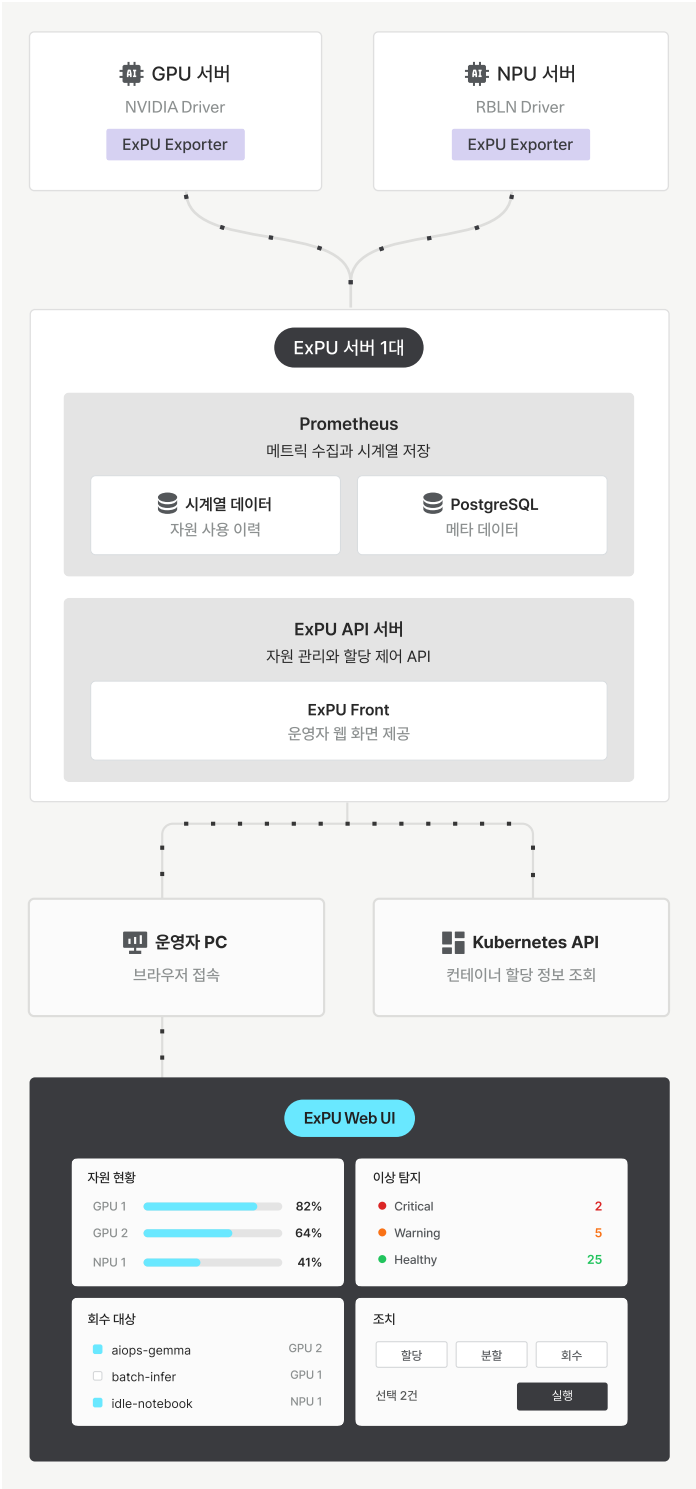
ExPU 서버

- 수집과 저장, API와 화면을 한 서버에 구성
- Prometheus가 메트릭을 수집해 시계열로 저장
 - PostgreSQL이 자원 구성 정보와 메타 데이터를 저장
 - ExPU API 서버와 Front를 같은 서버에 배치

3

접속과 연계

- 설치 없이 브라우저로 접속하고 API로 연계
- 운영자는 별도 설치 없이 브라우저로 Web UI 접속
 - Kubernetes API로 Pod과 Container 할당 정보 조회
 - 운영 데이터는 고객 내부망 안에서만 처리



Platform Specs

지원 스펙 · 환경

웹 브라우저

지원 브라우저 : Chrome, Edge
권장 해상도 : 1920×1080 / 최소 해상도: 1440×900

ExPU 통합 모니터링 서버

OS : Linux (Rocky/RHEL, CentOS 등)
CPU : 2 Core (최소) / 4 Core (권장)
RAM : 8GB (최소) / 16GB (권장)
DISK : SSD 300GB 이상
- OS 100GB
- Metrics 200GB 이상
SW : Docker, Prometheus, PostgreSQL, Kubernetes
NODE : 1대
배치 : NPU 사용량이 높은 서버는 피해 설치

모니터링 대상 서버

수집 : GPU/NPU 상태, 사용률, 온도, 메모리, 프로세스, Pod/Container 매핑
설치 : ExPU Exporter, 제조사별 Exporter
Exporter 구동 자원
- CPU : 1 Core 내외
- RAM : 2GB 내외
- DISK : 10GB 이상
NODE : GPU/NPU 장착 서버 전체

* 별도 표기가 없는 값은 ExPU 운영에 필요한 최소 사양입니다. 관제 대상 규모와 지표 보관 기간에 따라 권장 사양은 달라질 수 있습니다.

우리는 데이터를 통해 철학하고 혁신합니다