NEWSYS · AI SERVER BUILD CASE

4U 랙마운트에서 수냉 없이 GPU 2장을 66°C로 유지하는 방법

맞춤 에어플로우 덕트로 해결한 NVIDIA L4 ×2 AI 서버 제작 사례 · 100% 부하 20분 실측

100% 부하 정상상태	덕트 적용 개선 폭	열 스로틀	연속 부하 검증
66.0 °C	**-21.7 °C**	**0 회**	**20 분**
적용 전 87.7 °C	GPU1 정상상태 기준	클럭 1,245~1,290 MHz 유지	GPU 2장 동시 100%

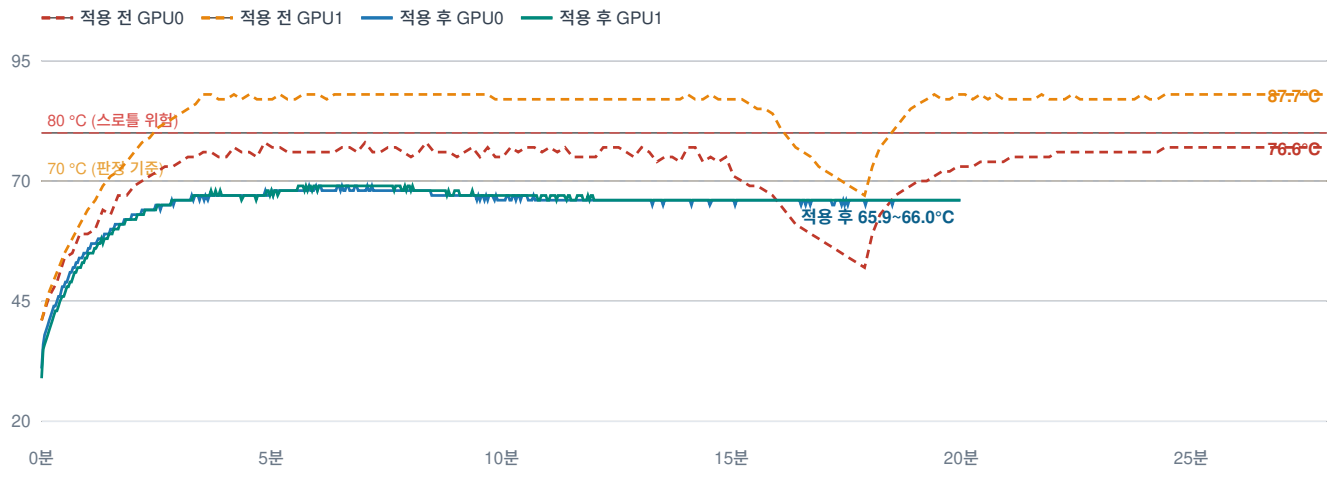
한 줄 요약 — 팬이 없는 패시브 GPU(NVIDIA L4)를 일반 4U 랙마운트 케이스에 넣으면 케이스 기류가 GPU 흡기면을 비켜가면서 온도가 88 °C까지 치솟습니다. 수냉으로 가지 않고, **케이스 내부에 맞춤 에어플로우 덕트를 설계·제작** 해 기류를 GPU 히트싱크로 강제 유도하는 것만으로 100% 부하에서 66 °C, 열 스로틀 0회를 달성했습니다.

1. 제작 요청 (고객 요구사항)

요구사항	내용
폼팩터	일반 4U 랙마운트 케이스 사용 — 특수 섀시·개조 섀시 불가
GPU 구성	NVIDIA L4 ×2 (24 GB, 72 W, **팬 없는 패시브 카드**)
냉각 방식	**수냉 배제** — 누수 리스크, 납품 후 유지보수 부담, 표준 워터블록 확보 난이도
운용 조건	납품 후 장시간 연속 추론 운전, 원격 관리
제약	표준 부품 위주, 납기 내 제작, 현장에서 팬/필터 교체 가능해야 함
검증 조건	출하 전 100% 부하 연속 시험 + 전후 데이터 제시

2. 문제점 — 왜 일반 4U에서 온도가 터지는가

- **패시브 카드의 숙명** — L4는 자체 팬이 없어 섀시 기류에 100% 의존합니다. 카드 흡기면에 충분한 정압이 걸리지 않으면 히트싱크가 열을 못 뺍니다.
- **기류 단락(short-circuit)** — 일반 4U 케이스는 전면 팬과 GPU 존 사이에 빈 공간이 많아, 공기가 GPU를 비켜 상단·측면으로 빠집니다.
- **케이블·브래킷 간섭** — PCIe 전원/라이저 케이블이 흡기면을 가려 국부적으로 기류가 죽습니다.
- **측정 결과** — 제작 전 동일 조건 부하에서 **GPU1 정상상태 87.7 °C(최고 88 °C)**, 28분 중 **약 23분을 80 °C 이상**에서 보냈고 GPU0도 76.6 °C로 판정 기준(70 °C)을 넘었습니다.
- **수냉은 답이 아니었습니다** — L4급 72 W 패시브 카드는 워터블록 선택지가 사실상 없고, 커스텀 제작 시 누수·유지보수 리스크가 납품 후 고객에게 전가됩니다.



[그림 1] 부하 시작을 0분으로 정렬한 온도 곡선 — 점선(제작 전)과 실선(덕트 적용 후)

3. 해결 — 맞춤 에어플로우 덕트 설계

1 기류 측정
전면 팬 풍량과 GPU 흡기면 실측, 카드 간격·슬롯 위치 기준으로 덕트 형상 설계

2 덕트 제작
GPU 흡기면과 전면 팬 존을 밀폐 연결하는 맞춤 덕트 제작, 리크(틈새) 최소화

3 기류 정리
PCIe 전원·라이저 케이블을 기류 경로 밖으로 재배치, 슬롯 브래킷 간섭 제거

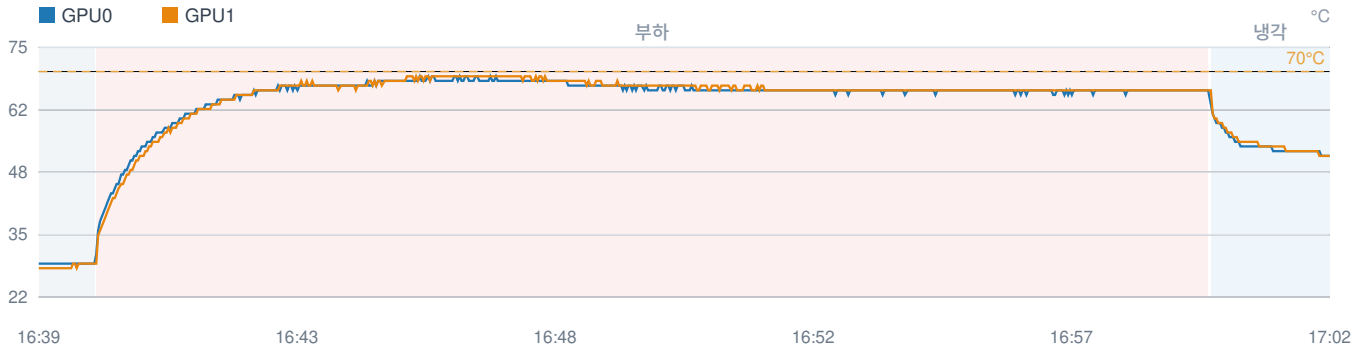
4 부하 검증
100% 부하 20분 연속 시험, 온도·전력·클럭·스로틀 사유 실측 기록

- **핵심 원리** — 차가운 외기를 GPU 히트싱크 **바로 앞까지** 가두어 보내고, 배기는 케이스 후면으로 분리. 같은 팬으로도 유효 풍량이 크게 달라집니다.
- **수냉 대체 관점** — 배관·펌프·냉각수·누수 센서가 필요 없고, 고장 시 팬 1개 교체로 끝납니다. 납품 후 유지보수 난이도가 수냉과 비교할 수 없을 만큼 낮습니다.
- **비용** — 워터블록·펌프·라디에이터 구성 대비 부품 수가 적고, 표준 4U 케이스를 그대로 사용합니다.

4. 검증 결과

구분	GPU	정상상태	최고	사용률	전력	개선
제작 전	GPU0	76.6 °C	78 °C	60 %	60 W	—
	GPU1	87.7 °C	88 °C	90 %	55 W	
덕트 적용 후 100% 부하 20분	GPU0	65.9 °C	69 °C	100 %	72 W	▼ 10.7 °C
	GPU1	66.0 °C	69 °C	100 %	72 W	▼ 21.7 °C

정상상태 = 부하 구간 마지막 20% 평균. 제작 전 측정은 10초 간격·부하 패턴이 달라 절대값 비교에는 한계가 있으나, 적용 후는 양쪽 100%에서 더 낮은 온도를 유지했습니다.



[그림 2] 100% 부하 20분 — 3분 내 66 °C 평형 도달 후 상승 없음, 부하 종료와 동시에 즉시 하강

5. 개선점 및 다음 단계

구분	내용
덕트 표준화	케이스·GPU 조합별 덕트 규격화(1U/2U/4U, 단일/듀얼 카드), 3D 프린팅 목업 → 판금·사출 양산 공정 전환
소음·진동	정압형 팬 선정과 팬 커브 튜닝, 덕트 진동 감쇠(가스켓) 적용 — 소음 실측 데이터 확보 예정
유지보수성	툴리스 탈착 덕트, 전면 필터 접근성 개선, 팬 단위 교체 구조
모니터링	흡기 온도 센서 추가 → ΔT 기준 관리, 80 °C 초과 알림·주기 리포트 자동화(현재 운용 중)
적용 확장	A2·H100 PCIe 등 패시브 카드, 8 GPU 구성, 수냉 하이브리드(고밀도 구성 시)
한계 명시	공랭 최적화로 달성한 결과이며 수냉 대비 절대 상한이 높은 방식은 아닙니다. 400 W급 이상 카드는 별도 설계가 필요합니다.

6. 홈페이지 · 블로그 게시용 텍스트

제목안

- 4U 랙마운트에서 수냉 없이 GPU 2장을 66°C로 — 에어플로우 덕트 제작기
- L4 88°C → 66°C, 바꾼 건 팬이 아니라 '공기의 길'이었습니다
- 수냉 대신 덕트 — 패시브 GPU 서버를 공랭으로 완성한 방법

요약 (2문장)

팬이 없는 NVIDIA L4 2장을 일반 4U 랙마운트 케이스에 탑재한 AI 서버에서, 케이스 내부 기류를 GPU 흡기면으로 강제 유도하는 맞춤 에어플로우 덕트를 제작했습니다. 100% 부하 20분 시험 결과 정상상태 온도가 87.7 °C에서 66.0 °C로 내려가고 열 스로틀은 0회로 사라졌습니다.

본문 (초안)

AI 서버를 만들 때 가장 많이 받는 요청은 "일반 4U 케이스에 GPU를 넣어 달라"입니다. 문제는 데이터센터용 GPU인 NVIDIA L4가 **팬이 없는 패시브 카드**라는 점입니다. 자체 팬이 없으니 케이스 기류가 카드 흡기면을 제대로 지나가야 하는데, 일반 4U 케이스는 전면 팬과 GPU 사이에 빈 공간이 많아 공기가 GPU를 비켜갑니다. 실제로 제작 전 측정에서 GPU1은 28분 중 23분을 80 °C 이상에서 보냈고 최고 88 °C까지 올라갔습니다.

선택지는 두 가지였습니다. 수냉으로 가거나, 공기를 다시 설계하거나. L4급 72 W 카드는 표준 워터블록이 사실상 없고 커스텀 제작 시 누수·유지보수 리스크가 납품 후 고객에게 남습니다. 그래서 **수냉 대신 덕트**를 택했습니다. 전면 팬 존과 GPU 흡기면을 밀폐 연결하는 맞춤 덕트를 제작하고, 케이블과 브래킷을 기류 경로 밖으로 정리했습니다.

결과는 100% 부하 20분 시험에서 정상상태 66.0 °C, 열 스로틀 0회, 클럭 1,245~1,290 MHz 유지였습니다. 적용 전과 비교하면 GPU1 기준 약 21.7 °C가 내려간 수치입니다. 같은 팬, 같은 케이스에서 '공기의 길'만 바꾼 결과입니다.

수냉은 분명 고성능 선택지입니다. 하지만 72 W급 패시브 카드 2장을 안정적으로 돌리는 것이 목적이라면, 공랭 최적화가 유지보수·비용·납기 모든 면에서 합리적입니다. 뉴시스는 케이스와 GPU 조합별 덕트 규격화와 출하 전 100% 부하 검증 리포트를 함께 제공합니다.

태그

AI서버제작 GPU서버 NVIDIA L4 에어플로우덕트 랙마운트서버 공랭설계 수냉대체 서버냉각 ComfyUI서버 뉴시스

홈페이지 카드 문구

4U 공랭 AI 서버 (L4 ×2)
맞춤 에어플로우 덕트로 100% 부하 66 °C·열 스로틀 0회. 수냉 없이 유지보수 부담을 낮춘 구성.
· 20분 연속 부하 검증 · 전후 데이터 리포트 제공

데이터 근거: 100% 부하 20분 연속 시험(2초 간격 nvidia-smi 폴링), 제작 전 측정 2026-10-29 / 제작 후 2026-11-03
뉴시스(NEWsys) · 대표 박재우 · jwpark@newsyst.com · 010-3136-7623 · 생성 2026-11-03 17:16