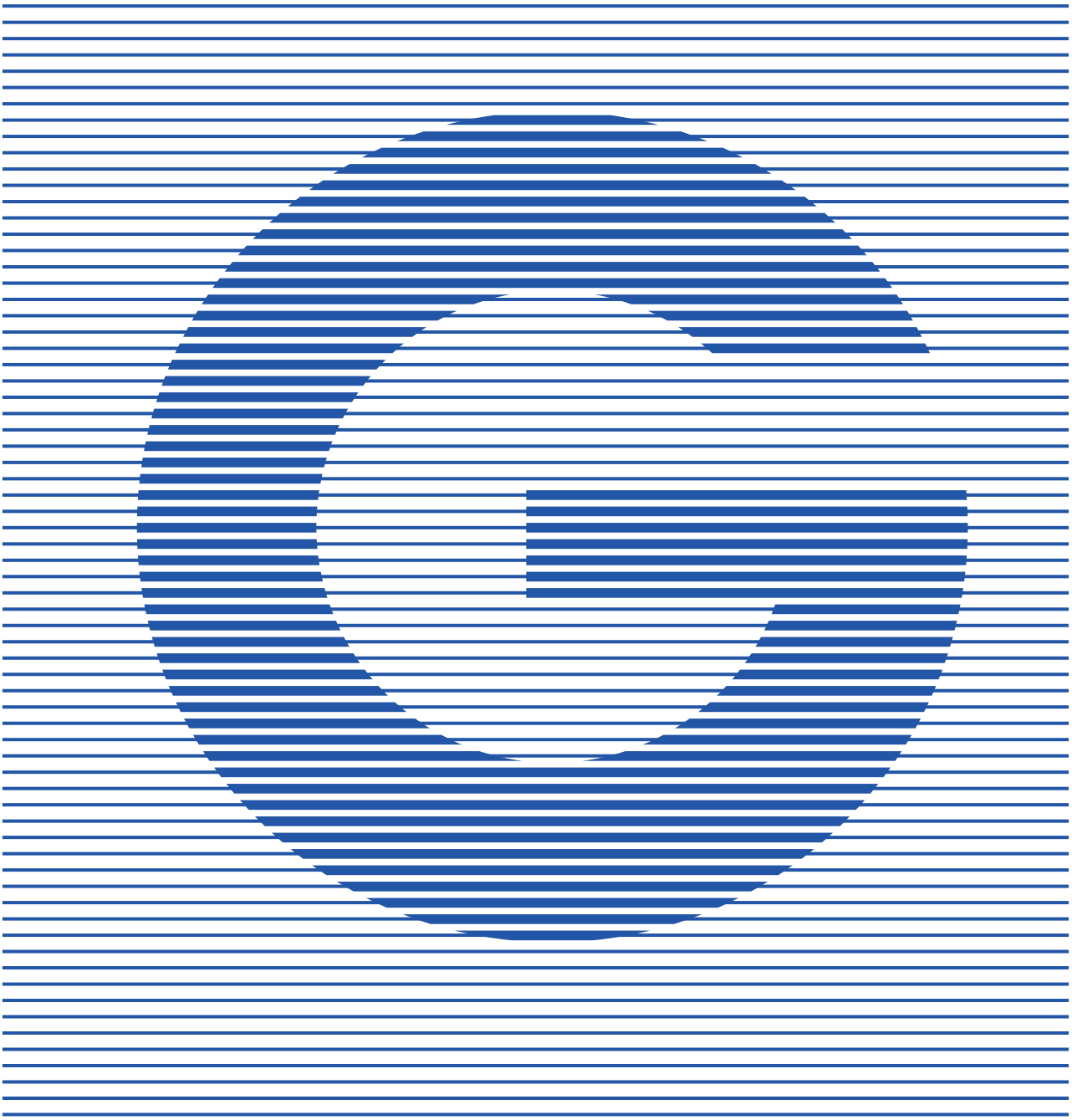
deep gadget



AI 인프라도 기본에서부터.

AI infrastructure starts from the basics.

deepgadget.com

MANYCORESOFT

deep gadget

dg = deep learning serving Gadget

deep gadget의 새로운 로고는 모든 냉각의 기본인 skived heat sink를 Gadget의 첫 글자에 형상화한 것으로,
가장 기본이 되는 것부터 철저하고 정밀하게 디자인하며 최고의 컴퓨팅 파워를 만들어가는 딥가젯의 철학과 기술력을 상징합니다.

목차

AI 인프라도
기본에서부터.

매니코어소프트 소개

ManyCore + Soft!

매니코어소프트는 H/W와 S/W 전체에 걸친 최적화 역량으로
초거대 AI를 비롯한 다양한 산업 분야에
최고의 컴퓨팅 파워를 제공하는 회사입니다.

매니코어소프트는 서울대학교 멀티코어 컴퓨팅 연구실(현 천둥 연구실)**에서 출발한 고성능 컴퓨팅 전문 기업입니다. 매니코어는 다수의 작업을 효율적으로 처리하기 위해 2~8개의 멀티코어를 넘어 수백~수천개의 코어가 집적된 아키텍처를 일컫습니다.**

고성능 컴퓨팅, 즉 HPC(High Performance Computing)**를 구현하기 위해서는 하드웨어 기술과 소프트웨어 역량이 모두 중요합니다. '매니코어소프트'는 '매니코어'와 '소프트'의 합성어로, 두 역량을 모두 갖춘 HPC 전문가 집단을 의미합니다. 특히 AI 산업 분야 전체 기술 스택을 지원할 수 있는 Full stack AI Company임을 강조합니다.**

매니코어소프트는 고성능 GPU 액체냉각 서버 설계 및 제작부터, 대규모 AI 인프라 컨설팅 및 구축, 관리까지 전 영역에 걸쳐 서비스를 제공합니다. HPC 분야, 특히 가속기(GPU 및 FPGA) 등을 사용하는 분야에 서울대 연구실의 세계적인 연구성과를 바탕으로 한 기술력을 보유하고 있으며, 이를 기반으로 다양한 산업의 선도 기업 및 기관들과 지속적으로 교류·협업하고 있습니다.

10년의 시간동안 매니코어소프트는 하드웨어와 소프트웨어 전체에 걸친 최적화 역량을 철저히 다져왔습니다. 현재 매니코어소프트는 누구와도 비교할 수 없는 노하우로 초거대 AI를 비롯한 다양한 산업 분야에 최고의 컴퓨팅 파워를 제공합니다. 또한, AI와 GPU 그 이상을 준비하며 미래를 향해 나아가고 있습니다.

Services

서버 제조

Product

GPU 서버

액체냉각식 GPU 서버

deep gadget

슈퍼컴퓨터 수냉 기술

Domain Knowledge

ManyCoreSoft

HPC 인프라 구축

Product

AI/ML

On-premise
AI 클라우드 컴퓨팅

GPU 클러스터
구축 및 가상화 기술

Domain Knowledge

ManyCoreSoft

Background

오픈 소스 OpenCL 프로그래밍 환경 SnuCL 개발

- 멀티 노드의 여러 다양한 종류의 가속기를 한 노드에 있는 것처럼 추상화
- 이종 클러스터 시스템에서 고성능과 프로그래밍 용이성을 동시에 구현
- 세계 70개국 이상의 학교, 기업, 연구소에서 다운로드하여 사용 중
- 컴퓨터 분야 최상위 학술대회에서 논문(9편)과 튜토리얼(9회)로 소개

국내 최초의 GPU 슈퍼컴퓨터 '천둥' 자체 설계 및 구축

- 저비용, 저전력에 초점을 둔 설계 적용
- TOP500 리스트 277위, Green500 리스트 32위
- TOP50 리스트 중 7번째로 전력 효율이 높은 아키텍처로 선정
- 컨슈머 GPU를 고밀도로 장착하고 액체 냉각 방식으로 냉각한 세계 최초의 슈퍼컴퓨터
- 소프트웨어 최적화 기술로 범용 하드웨어에서 단일 노드 기준 성능 세계 1위 달성

Leadership

CEO

박정호

· CEO & Co-Founder of ManyCoreSoft
· Head of Research & Co-Founder of Moreh
· PhD in EECS and a BS in Computer Science & Engineering from Seoul National Univ.

Research Interests

· Parallelization and optimization of applications on heterogeneous clusters
· Design and implementation of hyperscale AI models and infrastructures

Advisor

이재진 교수

• Professor in Dept. of CSE, Dean of Graduate School of Data Science at Seoul National Univ.
• Leader of the Thunder Research Group at SNU
• PhD in CS from UIUC, MS in CS from Stanford University
• IEEE fellow

Research Interests

• Programming systems of heterogeneous machines
• Parallelization and optimization of deep learning models and frameworks

MCS History

2012. 07	MANYCORESOFT 설립
2012. 10	이종슈퍼컴퓨터 '천둥' 개발 및 구축 1/50 비용으로 국내 최초 TOP500 list 227th 달성 (에너지 효율 Global 7th)
2013. 07	㈜코스콤 컨설팅 계약 체결 IB 업계 최초 실시간 업무처리 솔루션
2014.	HPE, Intel, AMD 등과의 협업을 포함한 다양한 파트너십, 기술 개발 및 인프라 프로젝트
2015.	수냉 시스템, 'SC2015 Emerging Technology 10선' 중 하나로 선정
2016.	GPU 수냉시스템 특허 등록 '패킷 처리 장치 및 방법(10-1954306)' '컴퓨터용 수냉식 냉각장치 및 그 구동방법 (10-2118786)'
2017.	OpenCL 커널 코드 자동화 및 병렬 처리 표준화 시스템에서의 기술 성취
2018.	금융 서비스용 머신러닝 솔루션 개발, SK 하이닉스 운영가속화 하드웨어 시스템 설계 기여
2018. 12	하드웨어 제조 공장 설립 / DEEPGadget 출시
2019. 10	TCB 기술신용평가 T3 / 벤처기업 인증
2020. 08	KT cloud 대규모 GPU cluster 구축
2021. 08	Moreh와 하드웨어 인프라 포괄적 협력 협약
2022.	GPU 연 3000개 이상 설치 달성
2023.	고객사 100곳 돌파
2024 & 2025.	5세대 서버 dg5 & HW Solution dg-tp 출시 텐스토렌트와의 협업을 포함한 다양한 파트너십

교육 사업

매니코어소프트는 가속기(Accelerator) 컴퓨팅의 중요성에 대한 인식을 높이고, 이를 활용할 수 있는 전문 프로그래밍 인력을 양성하고자 2013년부터 '가속기 프로그래밍 학교'를 서울대학교 천둥연구실과 함께 개최하고 있습니다.

- **2013년부터 매년 여름과 겨울에 4박 5일간 진행**
- **병렬 컴퓨팅 및 GPU 구조, GPU 프로그래밍 (CUDA/OpenCL) 및 최적화 기법 등 교육**
- **대학원생 및 기업체의 연구원 등 대상**

< 2013 겨울학교 >

< 2023 여름학교 >

Why liquid cooling?

" 컴퓨터의 미래는 액체냉각이다. "

< Jensen Huang, **CEO of NVIDIA** >

과거 액체냉각 라디에이터가 없던 시절의 공랭 내연기관 자동차는 30분 운행하고 엔진 냉각을 위해서 보닛을 열고 2시간가량 세워두었습니다. 액체냉각 라디에이터가 개발되고 엔진을 직접 냉각함으로 5 ~ 6시간, 그 이상의 엔진 가동이 가능해졌습니다.

컴퓨터도 마찬가지입니다. 과거 대부분의 고성능 서버는 대당 합산 전력 1kW 이상이 흔하지 않았으며 공랭으로도 코어 냉각이 가능했습니다. 하지만 현재는 AI 학습 / 추론 / 연산 / 렌더링의 다양한 목적으로 서버 1대당 3kW~6kW, 그 이상을 필요로 합니다. 그리고 늘어나는 성능에 비례하여 발열은 지속적으로 증가하는 추세입니다.

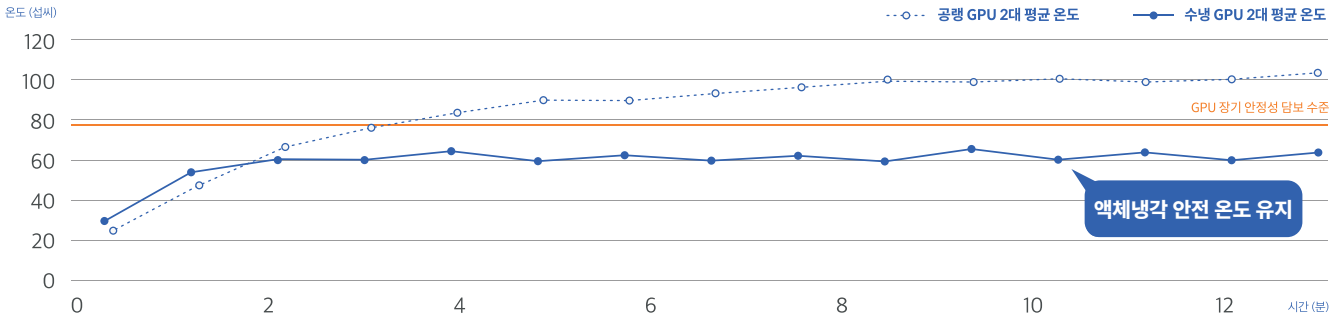
그 결과 **현재 일반 공랭식 서버에 GPU를 여러개 장착하면 GPU 성능이 최대 50%까지 떨어집니다.** 또한, 시스템이 불안정하게 동작하며, CPU, GPU, NPU의 자발적 성능 제한(throttling)이 발생합니다. 지속적인 고열은 제품 내구성에 영향을 미쳐 제품 수명이 단축됩니다.

결과적으로 사용자는 고비용의 서버 장비를 구입하고도 활용할 수 없는 문제를 겪습니다. 실제로 개인 사용자 뿐만 아니라 대기업의 최신식 IDC(Internet Data Center)에서도 관련 문제로 운영에 어려움을 겪는 사례들이 보고됩니다.

	Air (20°C)	Water (20°C)
Thermal conductivity (열전도성) [J/(m*K*s)]	0.026	0.598
Volumetric heat capacity (열용량) [J/(m^3*K*s)]	1213	4174472
Thermal inertia (열관성) [J/(m^2*K*s)]	5.09	1579.98

물은 공기보다 300배 이상 높은 열관성을 가져, 더 많은 열을 보유하고 잘 배출합니다.
현재 유일한 해결책은 물을 이용한 액체냉각 방식의 냉각입니다.

액체냉각 vs. 공랭 GPU 발열 및 성능 감소율



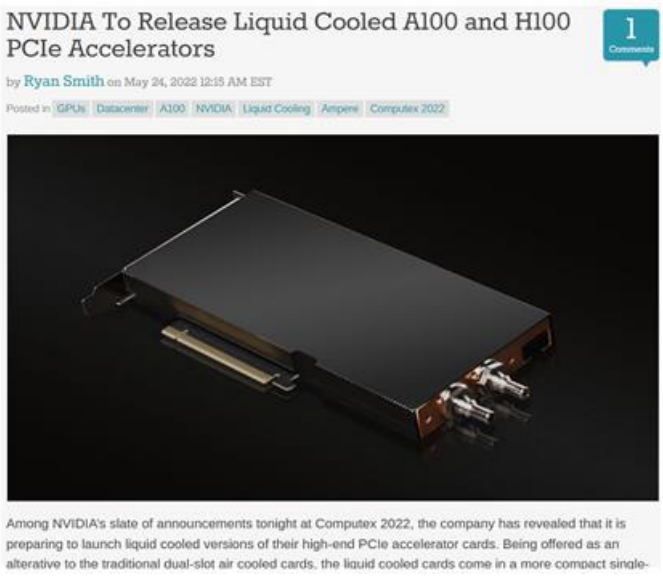
최대 성능 대비 감소율

공랭 : 하락	97%	95%	94%	92%	90%	89%
액체냉각 : 유지	98%	98%	98%	98%	98%	98%

액체냉각을 활용하면...

- 실내 온도에 거의 영향 받지 않고 시스템을 냉각할 수 있습니다.
- 냉각 전력 소비 감소로 에너지 효율이 향상됩니다.
- 높은 밀도로 컴퓨팅 장치를 장착할 수 있습니다.

이미 NVIDIA를 비롯한 AI 인프라 업체들은 Direct Liquid cooling을 적극적으로 도입하고 있습니다. (2022~)



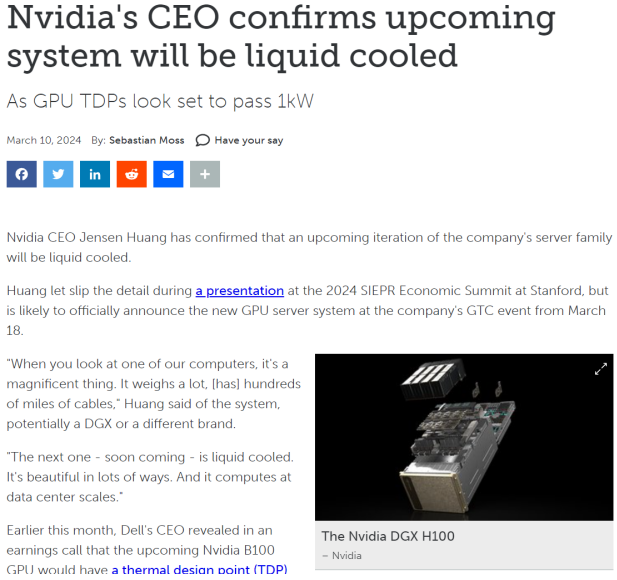
NVIDIA To Release Liquid Cooled A100 and H100 PCIe Accelerators

1 Comments

by Ryan Smith on May 24, 2022 12:15 AM EST

Posted in GPUs Datacenter A100 NVIDIA Liquid Cooling Ampere Computex 2022

Among NVIDIA's slate of announcements tonight at Computex 2022, the company has revealed that it is preparing to launch liquid cooled versions of their high-end PCIe accelerator cards. Being offered as an alterative to the traditional dual-slot air cooled cards, the liquid cooled cards come in a more compact single-



Nvidia's CEO confirms upcoming system will be liquid cooled

As GPU TDPs look set to pass 1kW

March 10, 2024 By: Sebastian Moss Have your say

Nvidia CEO Jensen Huang has confirmed that an upcoming iteration of the company's server family will be liquid cooled.

Huang let slip the detail during a presentation at the 2024 SIEPR Economic Summit at Stanford, but is likely to officially announce the new GPU server system at the company's GTC event from March 18.

"When you look at one of our computers, it's a magnificent thing. It weighs a lot, [has] hundreds of miles of cables," Huang said of the system, potentially a DGX or a different brand.

"The next one - soon coming - is liquid cooled. It's beautiful in lots of ways. And it computes at data center scales."

Earlier this month, Dell's CEO revealed in an earnings call that the upcoming Nvidia B100 GPU would have a thermal design point (TDP)

The Nvidia DGX H100
– Nvidia

deep gadget만의 액체냉각 기술

슈퍼컴퓨터 연구 개발진 X 도요타자동차 냉각 부분 엔지니어

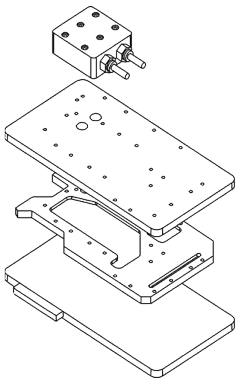
deep gadget은 흔한 커스텀 수냉이 아닙니다. 서울대 연구실의 슈퍼컴퓨터 개발·구축 성과, 도요타자동차 본사 20년 경력 엔지니어의 설계 그리고 10년 간의 R&D가 합쳐진 정교한 차세대 냉각 시스템입니다.

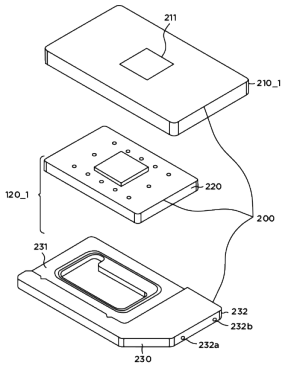
- 냉각 기술 특허 및 H/W 설계 역량 보유
- 냉각판(Cold plate)를 자체 설계하여 NVIDIA AI GPU 외에도 Gaming GPU, CPU, NPU 등 다양한 AI 가속기 및 Infiniband NIC 액체 냉각 지원
- 최신 H/W 구성요소들을 가장 빠르게 장착 (CXL, NMC, PIM 등의 차세대 계산 및 메모리/스토리지 등)

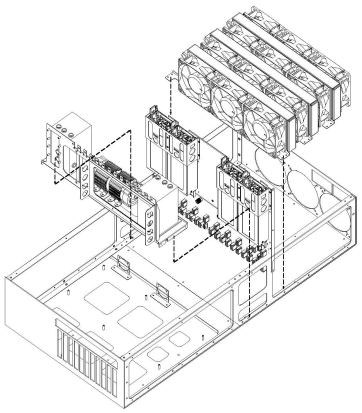
특허 및 실용신안

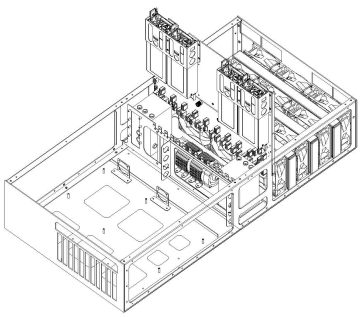
컴퓨터용 수냉식 냉각장치 및 그 구동방법
(COOLING DEVICE USING WATER
FOR COMPUTER
AND DRIVING METHOD THEREOF)
등록번호 : 10-2118786

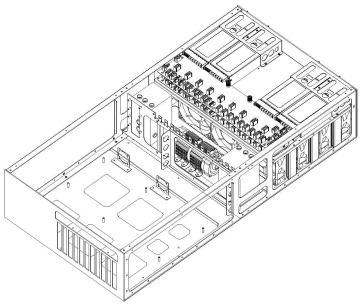
고밀도 GPU 액체 냉각을 위한 냉각판
등록번호 : 20-0477833, 20-0479465











별도 장치가 전혀 필요 없는 빌트인 액체냉각

열원 Direct Liquid Cooling 시스템 + 최첨단 수로 설계 = 압도적인 냉각 퍼포먼스

deep gadget 하나만으로 A100 GPU가 16장까지 쾌적하게 동작합니다.

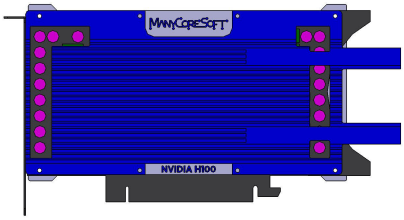
- 실온(30도 이상)에서도 사용 가능, 별도 항온항습이 필요 없음
- Chiller 등 외장 장치 및 배관이 전혀 필요하지 않음
- 인프라 구축 및 관리 비용이 절감되며, 에너지 효율이 높습니다.

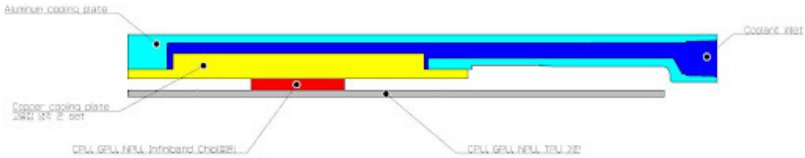
열원 Direct Liquid Cooling 시스템

1. 열 전도가 높은 구리 냉각판을 열원에 직접 부착함으로 열을 빠르게 전도
2. 라디에이터 & 쿨링 팬으로 식혀진 냉각액으로 구리 냉각판의 고밀집 핀을 통과

위 두가지를 동시에 행함으로 GPU 냉각 극대화를 가능하게 하는 시스템

※열 전달 : 열원(Chip) → Liquid → Radiator → Air

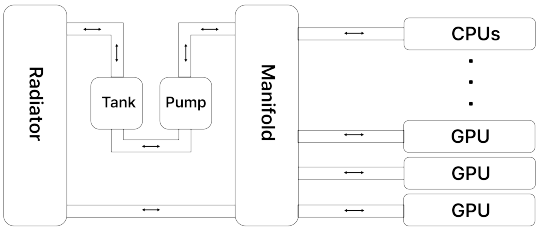




1:1 병렬 수로 설계

각 부품을 독립적으로 냉각해 최고의 퍼포먼스를 제공합니다.

(타 수냉: 모든 부품을 직렬 연결해 열이 누적되는 비효율적 시스템)

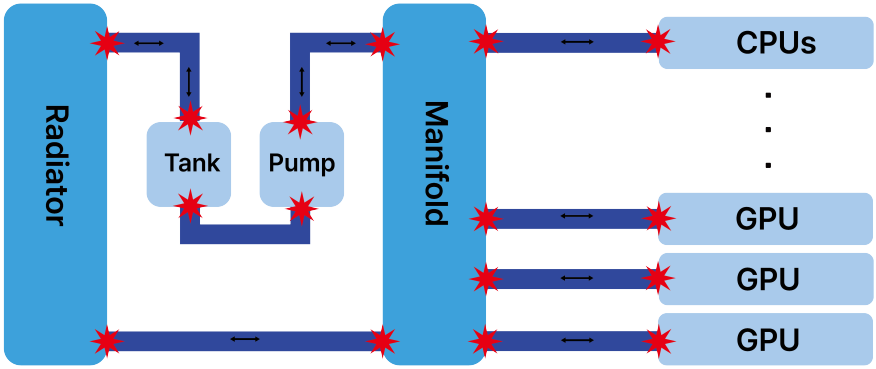


안심할 수 있는 5가지 이유

01. 10년 간 누수 피해 0건

deep gadget은 판매 역사 상 제품 과실의 누수 피해가 단 한 건도 없습니다.
10년 동안의 기술력과 신뢰로 안전을 보증합니다.

02. 설계단 : 완전 밀폐 설계



① 특수 접착제

부품 간 나사로 연결된 모든 Leak point에 특수 접착제를 도포하여 완전 밀폐 구조로 충격에도 안정적

② 호스, 호스 피팅, 호스 클립

공차 설계로 선정한 호스 부품을 사용하여 빈틈없는 구조

③ 퀵 커넥터

유압, 유량의 설정값 대비 3배 이상 출력에서도 밀폐를 보장하는 퀵 커넥터 사용으로
CPU, GPU ↔ Manifold 간 위험을 제거(서버 운용 중에도 안전하게 탈장착 가능)

④ 랙 선반 완비

자연재해의 침수, 누수의 경우에도 2차 피해를 완벽 방지합니다.
랙 선반은 dg4F 규격보다 큰 사이즈이며서버 내부에 보관된 냉각액의 약 2배 용량입니다.
(w x d x h : 43cm x 89.5cm x 2.5cm)
랙 선반 또한 제조 과정에서 테스트를 거치며 부식을 방지하기 위해서 방청 및
방수 도료로 마감하였습니다.

**※ dg cooling의 완전 밀폐 구조는 안전할 뿐만 아니라,
냉각액을 99.99% 보존하여 냉각액을 주기적으로 관리할 필요 없이 반영구적 사용이 가능합니다.**

03. 제조단 : 4회에 걸쳐 총 144시간 테스트 실행

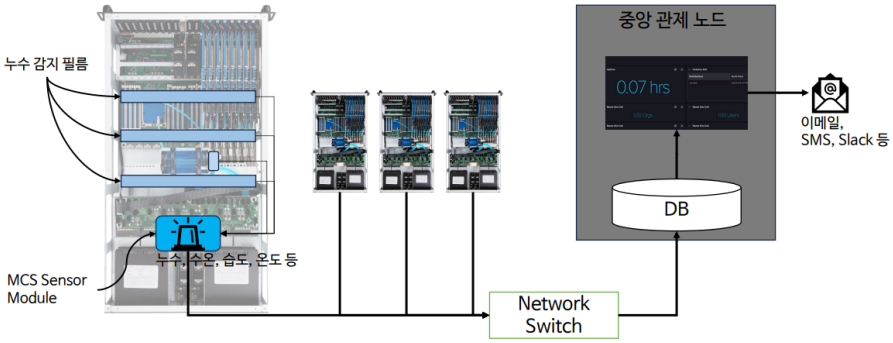
① CPU/GPU 냉각판 장착 후 24시간 테스트

② Leak point에 특수 접착제 도포 및 호스 연결 완료 후 24시간 테스트

③ 냉각액 2차 테스트 24시간 실행

④ 출고 전 72시간 burn-in 테스트로 냉각 성능 및 안정성 최종 점검

04. 제품 출고시 : 관제 시스템



05. 출고 후 : 든든한 품질보증

제품 구입 후 업계 최장 기간인 3년 품질 보증을 제공합니다. (자세한 범위와 방식은 28p에서 확인 가능합니다.)

또한, 자연재해의 침수/누수의 경우에 최대한 예방할 수 있도록 가이드를 드립니다.

냉각액 정보

전 제품에 최고 품질의 냉각액을 사용하여 냉각효율이 높으며 부식, 동파, 박테리아, 녹조 등 변형에서 안전합니다.

dg4F 기준 전체 용량(L) : 1.3L

***냉각액 성분**

· 증류수 70~75%

· Propylene Glycol(프로필렌글리콜) 25~30%

· Potassium Phosphate Dibasic(디포타슘포스페이트) 1%이하

· Sodium molybdate(몰리브덴산나트륨) 1%이하

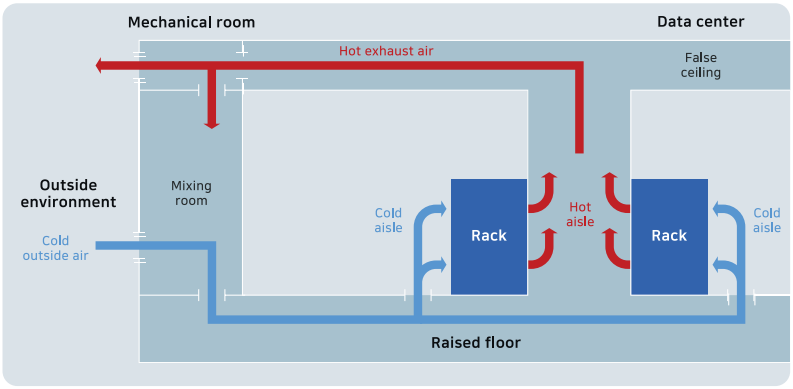
· Meta-toluic Acid(m-톨루일산) 1%이하

Electrical Conductivity	2500
Freezing Point	-15°C(5F)
Specific Gravity @20°C	1.03
UV Reactive	Blue
Viscosity @20°C (cP)	2.3

dg 냉각으로 대규모 AI 클러스터 비용 절감하기.

deep gadget은 Free-Cooling 가능!

- 압도적 성능의 빌트인 액체냉각으로 별도 장치가 전혀 필요 없음
- 실온(30°C) 이상의 고온 환경에서도 정상적으로 운영 가능함
- 국내 기후에서 deep gadget과 외기만을 이용하여 획기적인 에너지 효율 달성 가능
- PUE를 1.1 이하로 쉽게 낮출 수 있음



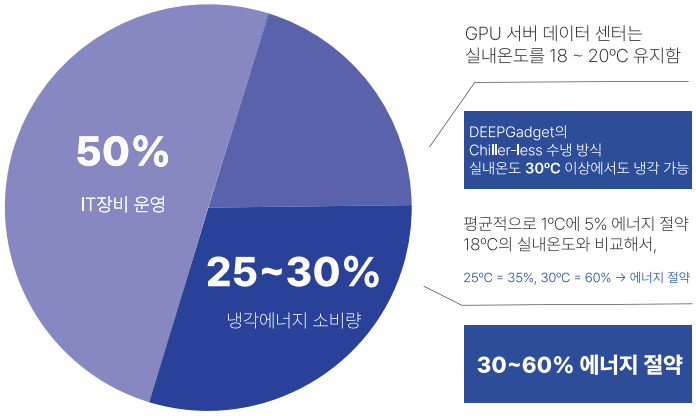
*PUE(Power usage effectiveness) = 냉각효율성

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

데이터센터 268곳의 평균 PUE = 1.8

*Avgerinou, Maria, Paolo Bertoldi, and Luca Castellazzi. "Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency." Energies 10.10 (2017): 1470.

데이터센터 에너지 절약 시나리오



데이터 센터 에너지 절약
deep gadget

GPU 서버 데이터 센터는
실내온도를 18 ~ 20°C 유지함

DEEPGadget의
Chiller-less 수냉 방식
실내온도 **30°C** 이상에서도 냉각 가능

평균적으로 1°C에 5% 에너지 절약
18°C의 실내온도와 비교해서,

25°C = 35%, 30°C = 60% → 에너지 절약

30~60% 에너지 절약

*데이터센터의 온도 1도마다
4.3~9.8% 에너지를 줄일 수 있다는 연구 결과

Table 7
Energy consumption per unit area at the different temperature set points.

Month	Energy consumption per unit area at 24 °C (kWh/m²)	Energy consumption per unit area at 25 °C (kWh/m²)	Energy consumption per unit area at 26 °C (kWh/m²)
January	2.328	2.116	1.923
February	2.104	1.898	1.722
March	2.334	2.120	1.930
April	2.183	2.009	1.849
May	1.949	1.793	1.654
June	1.887	1.747	.625
July	1.869	1.781	1.705
August	1.863	1.723	1.601
September	1.834	1.705	1.594
October	1.930	1.783	1.667
November	2.003	1.847	1.715
December	2.314	2.127	1.974

* Iyengar, Madhusudan, et al. "Server liquid cooling with chiller-less data center design to enable significant energy savings." 2012 28th annual IEEE semiconductor thermal measurement and management symposium (SEMI-THERM). IEEE, 2012.

* The results shown that the percentage of energy saving was 4.3-9.8% for every 1°C rise in temperature set points.
Nan Wang, Jiangfeng Zhang, Xiaohua Xia, Energy consumption of air conditioners at different temperature set points, Energy and Buildings, Volume 65, 2013, Pages 412-418

비용, 에너지, CO_2감소 예상 시나리오

탄소중립 경영<ESG>

- ▶ 운영비(에너지 비용) 절감
  - 연간 3.94억원
  - 5년 19.7억원
- ▶ CO_2 감소
  - 연간 1,295 t 감소
  - 5년 6,478 t 감소

- ▶ 운영비(에너지 비용) 절감
  - 연간 6.6억원
  - 5년 32.9억원
- ▶ CO_2 감소
  - 연간 2,159 t 감소
  - 5년 10,797 t 감소

항목	기존 공냉 환경	dg 액체냉각[30도]	dg 액체냉각 + Free-Cooling
IT Power Consumption	1,000 kW	1,000 kW	1,000 kW
PUE	1.60	1.30	1.10
Total Power Consumption	1,600 kW	1,300 kW	1,100 kW
Annual Energy	14,016,000 kWh	11,388,000 kWh	9,636,000 kWh
Annual Cost	21 억원	17.1 억원	14.5 억원
Annual CO_2 Footprint	6,910 t	5,614 t	4,751 t
Cooling Power Consumption	500 kW	200 kW	≈0 kW

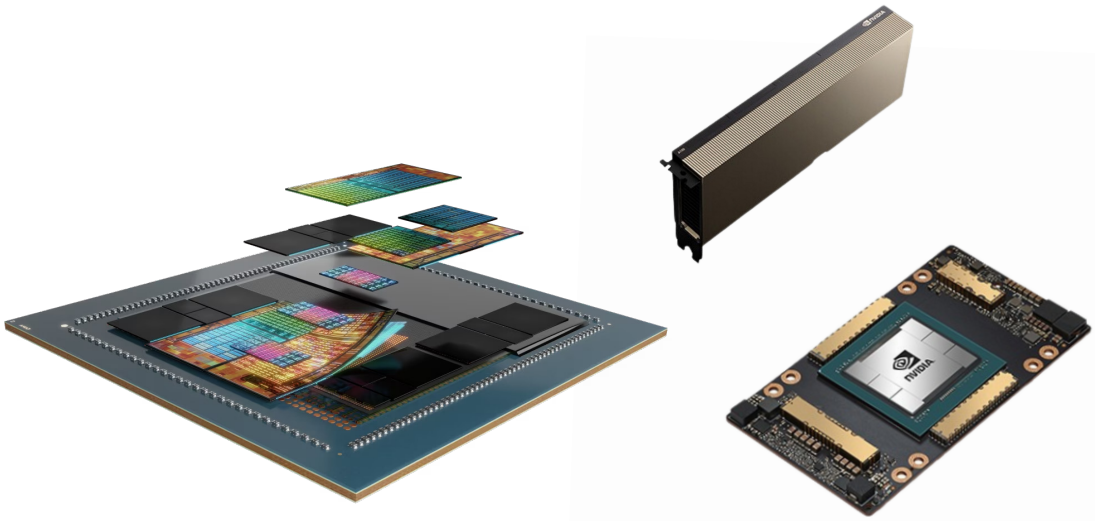
냉각 방식 종합 비교

항목	일반 공랭식	일반 Chiller 수냉식	Immersion	dg 액체냉각 + Free-Cooling
PUE	1.6내외	1.2내외	1.1내외	1.05내외
설비비용	고비용	고비용	고비용	저비용
냉각 성능	중간	높음	높음	높음
서버 관리	중간	어려움	매우 어려움	중간
상면 비용	중간	높음	매우 높음	낮음
밀집도	중간	낮음	매우 낮음	높음
소음	높음	적음	적음	적음

*데이터센터를 위한 랙 단위 상품 별도 문의

기본을 지키는 서버가
가장 강력한 컴퓨팅 파워를 만든다.

빌트인 dg 액체냉각과 air flow 설계



beyond GPU 를 비롯한 30+ 다양한 가속기 지원

deep gadget

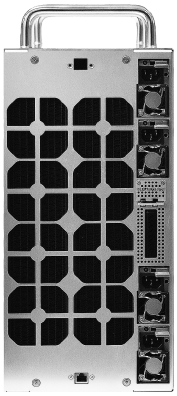
모든 소프트웨어가 준비된 채로.

dg = AI Gadget : AI 연구개발에 필요한 SW를 기본 탑재하여 바로 딥러닝을 수행할 수 있습니다.

- Ubuntu/RHEL/Rocky Linux/Windows
- NVIDIA/AMD/Tenstorrent/Furiosa 드라이버, Infiniband/GbE 드라이버
- CUDA, cuDNN, NCCL, cuBLAS, TensorRT, ROCM 등 런타임 및 라이브러리
- PyTorch, Tensorflow, DeepSpeed, Horovod, vLLM 등 DL 프레임워크 및 3rd party 솔루션
- MPI, Anaconda, Docker 등 개발 도구

서버실에서도 사무실에서도 랙마운트 이상의 성능.

dg5 Workstation



❶ Rackmount 성능 x Workstation의 간편함

두가지 타입으로 모두 사용할 수 있는 서버로, 필요에 따라 서버실 랙에 장착하거나 사무실에서 쓸 수 있습니다.

환경과 공간에 관계없이 최고의 작업을 진행해보세요.

❷ 몸집은 다운, 냉각은 업그레이드

최첨단 수로와 냉각판 설계로 같은 공간 크기 대비 강력하고 안정적인 냉각 성능을 자랑합니다.

❸ 어떤 경우라도 안정적인 전원 공급

4개의 Redundent 파워로 언제나 안정적이고 파워풀하게.

❹ 센서와 디스플레이 탑재로 관리가 편해집니다.

내부 동작 및 데이터를 감지하는 5개 센서와, 이를 표시하는 디스플레이를 기본으로 탑재하여

누구나 쉽게 관리할 수 있습니다. (모니터링 시스템 별도 판매 예정)

- **빌트인 dg 액체냉각**
- **7개의 PCIe Gen5 × 16슬롯 지원**
  - · PCIe 스위치가 없는 직접 연결 방식
- **고성능 GPU 최대 7개 냉각 가능**
  - · 2개의 펌프, 4개 라디에이터
  - · 16개 쿨링팬의 강력한 냉각
- **최대 50dB로 일반 사무실 수준 정숙성**

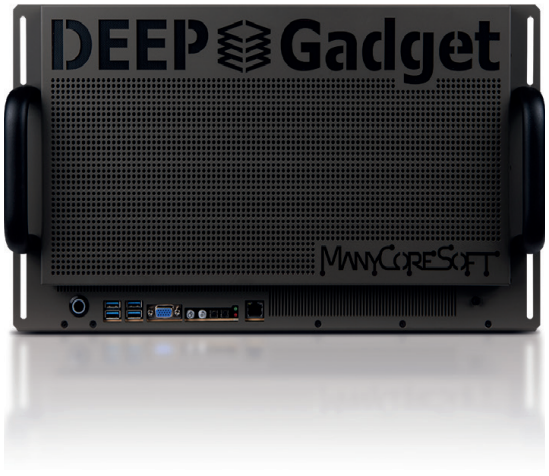
- **언제 어디서든 강력한 컴퓨팅 파워를 원하는 고객**
  - · 데이터센터, 기업, 연구기관, 개인 작업자 등
  - · AI 모델 학습/추론
  - · AI, Rendering, Encoding 등 각종 GPU farm 구축
  - · 다중 GPU 이미지 작업 (인코딩/렌더링/CAD 등)

dg4 Flagship

압도적 성능의 플래그십 서버

- **빌트인 dg 액체냉각**

- **Dual EYPC/Xeon CPU 지원**
- **PCIe 스위치가 없는 직접 연결 방식**
  - 19개의 PCIe Gen4×8 슬롯 제공
- **고성능 GPU 최대 16개 냉각 가능**
  - 4개의 펌프로 높은 유량
  - 6개 라디에이터
  - 18개 쿨링팬의 강력한 냉각
- **최대한의 GPU 병렬 처리를 원하는 고객**
  - 대규모 AI 모델 학습/추론
  - 과학 계산 및 시뮬레이션



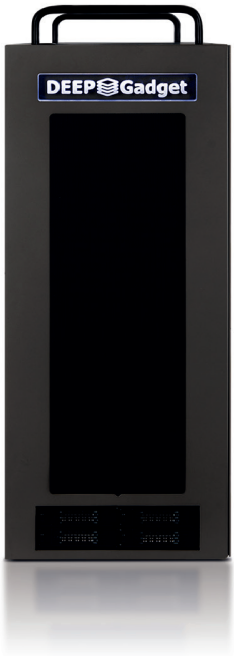
dg4 Rackmount

컴퓨팅 파워의 새로운 기본.

- **빌트인 dg 액체냉각**

- **Dual EYPC/Xeon CPU 지원**
- **PCIe 스위치가 없는 직접 연결 방식**
  - 9개의 PCIe Gen4×16 슬롯 제공
- **고성능 GPU 최대 10개 냉각 가능**
  - 2개의 펌프로 높은 유량
  - 3개 라디에이터
  - 9개 쿨링팬의 강력한 냉각
- **높은 GPU-CPU 통신 성능을 원하는 고객**
  - 분산 병렬 AI 모델 학습
  - AI, Rendering, Encoding 등 각종 GPU farm 구축

H/W Product



dg4 Workstation

강력하게, 조용하게, 콤팩트하게.

- **빌트인 dg 액체냉각**

- **고성능 Threadripper Pro Workstation CPU 지원**
- **PCIe 스위치가 없는 직접 연결 방식**
  - 7개의 PCIe Gen4×16 슬롯 제공
- **고성능 GPU 최대 7개 냉각 가능**
  - 2개의 펌프로 높은 유량
  - 2개 라디에이터
  - 7개 쿨링팬의 강력한 냉각
- **최대 50dB 이하로 일반 사무실 수준 정숙성**
- **오피스에서 대규모 GPU 작업을 원하는 고객**
  - 중소 규모 AI 연구
  - 다중 GPU 이미지 작업 (인코딩/렌더링/CAD 등)

beyond GPU

NPU를 활용한 AI 서빙 시스템

- **초거대언어모델 LLM 서빙 가젯 with Tenstorrent**
- **이미지 관련 AI 서빙 가젯 with FuriosaAI**

GPU vs. NPU

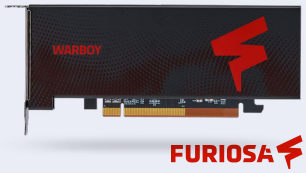
- **일반적인 AI 학습 및 추론에는 GPU, 특화된 곳은 NPU**
- **NPU는 GPU보다 에너지 효율이 높고, 경제적입니다.**

LLM 서빙 가젯

Model : dg-LLM-n300
NPU: Tenstorrent Wormhole n300 x 16(최대)
NPU Memory: 384 GB
LLM Performance: 4,192 TOPS (FP8)
TDP: 5.6 kW

Tenstorrent AI card

Vision AI 서빙 가젯



Model: dg-VISION-WB
NPU: Furiosa AI WARBOY x 16(최대)
NPU Memory: 256 GB
Vision Performance: 1,024 TOPS (INT8)
TDP: 1.8 kW

매니코어소프트는 세계적인 NPU 제조사와 협력하여 GPU 다음 시대를 준비하고 있습니다.

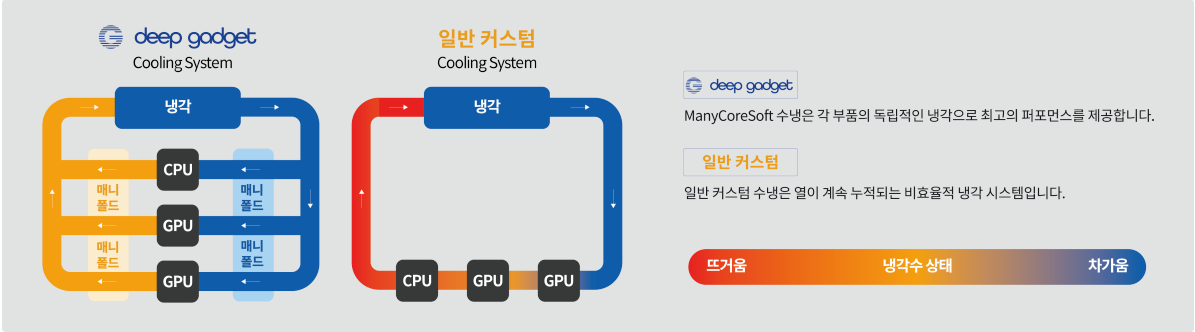
dg-Transplant

GPU에서 발생하는 발열이 고민이신가요?
비싸게 구입한 서버를 사용하지 못하고 계신가요?

dg-Transplant Solution을 이용하면
보유하고 계신 타사 GPU서버의 주요 구성품을
딥가젯 액체냉각으로 변경 장착하실 수 있습니다.

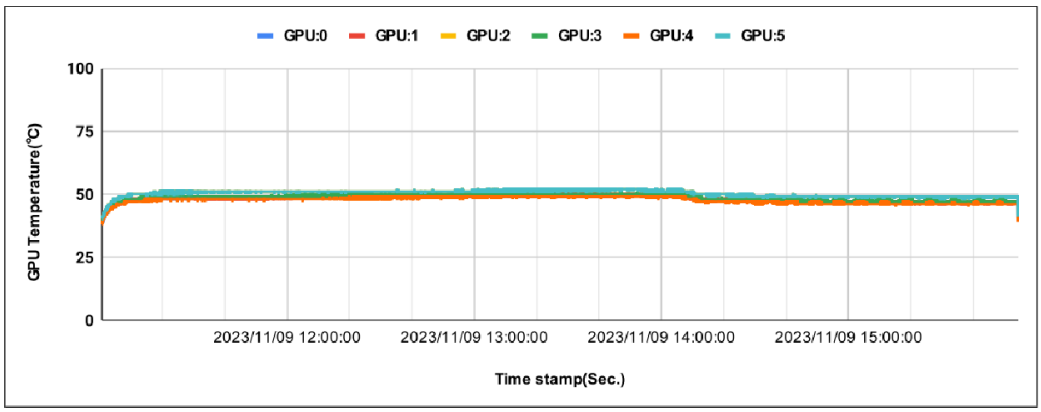
현존 모든 코어/메모리의 발열을 빠르게 해소하는 dg 액체냉각을 어디든 적용해보세요.

· 제조사 서버 가격의 약 10%로 고가의 서버 100% 성능 활용 가능
· 완전 밀폐 시스템으로 관리 걱정 불필요
· 3년 품질 보증
· (예정) 각종 센서로 서버 상태 모니터링 & 알람 서비스 제공



dg4R-4090 6대 Full Load 테스트(5 Hours)

Throttleing threshold: 90℃ Max Peak: 52 ℃



H/W One pager

	GPU				NPU	
						
Model	**dg5W**	**dg4F**	**dg4R**	**dg4W**	**dg-LLM-n300 (Tenstorrent)**	**dg-VISION-WB (FurisoaAI)**
Type	Rackmount/ Worksation	9U Rackmount	6U Rackmount	Workstation	Rackmount/ Worksation	Rackmount/ Worksation
CPU	AMD Ryzen™ Threadripper™ PRO 5955WX Intel® Xeon® Silver 4314	2 x AMD EPYC™ 7003Series Processors 2×3rd Generation Intel® Xeon® Scalable Processors	2 x AMD EPYC™ 7003Series Processors 2 × 3rd Generation Intel® Xeon® Scalable Processors	AMD Ryzen™ Threadripper PRO 7000 WX-Series Processors 5th Generation Intel® Xeon® Scalable Processors	AMD EPYC™ 7003Series Processors 3rd Generation Intel® Xeon® Scalable Processors	AMD EPYC™ 7003Series Processors 3rd Generation Intel® Xeon® Scalable Processors
GPU	Max 7ea	Max 16ea	Max 12ea	Max 7ea	Max 16ea	Max 16ea
Memory	최대 512GB DDR4-3200 ECC	최대 2TB DDR4-3200 ECC	최대 2TB DDR4-3200 ECC	최대 512GB DDR4-3200 ECC	최대 1TB	최대 512GB
M.2 NVMe SSD	2 Slots	1 Slots	1 Slots	2 Slots	1 Slots	1 Slots
PSU	4 × 2,500W 이하 Hot swappable	4 × 2,500W 이하 Hot swappable	4 × 2,500W 이하 Hot swappable	2 × 2000W Dual Power	4 × 2,500W	4 × 1,200W
Hot Swap Bay	4ea	18ea	8ea	4ea	18ea	8ea
WIFI	WIFI-6			WIFI-6		

가속기 지원 라인업	
대규모 학습용	NVIDIA H100, NVIDIA A100, AMD MI300X, AMD MI250, AMD L40S 등
추론과 소규모 학습용	AMD M210, NVIDIA RTX 6000 Ada, NVIDIA RTX 4090, NVIDIA RTX 4080 등
추론 전용	Tenstorrent Wormhole n300 NPU
Vision 전용	Furiosa WARBOY NPU

dg Solution	
dg-Transplant	DGX Station V100-4, DGX Station A100-4, DGX Station A100-8, DGX Station H100-8, HGX-A100-8, HGX-H100-8 등

*새로운 architecture가 지속적으로 추가되고 있습니다.

*새로운 지원 모델 및 서비스가 지속적으로 추가되고 있습니다.

S/W Full Stack One pager

Our Service

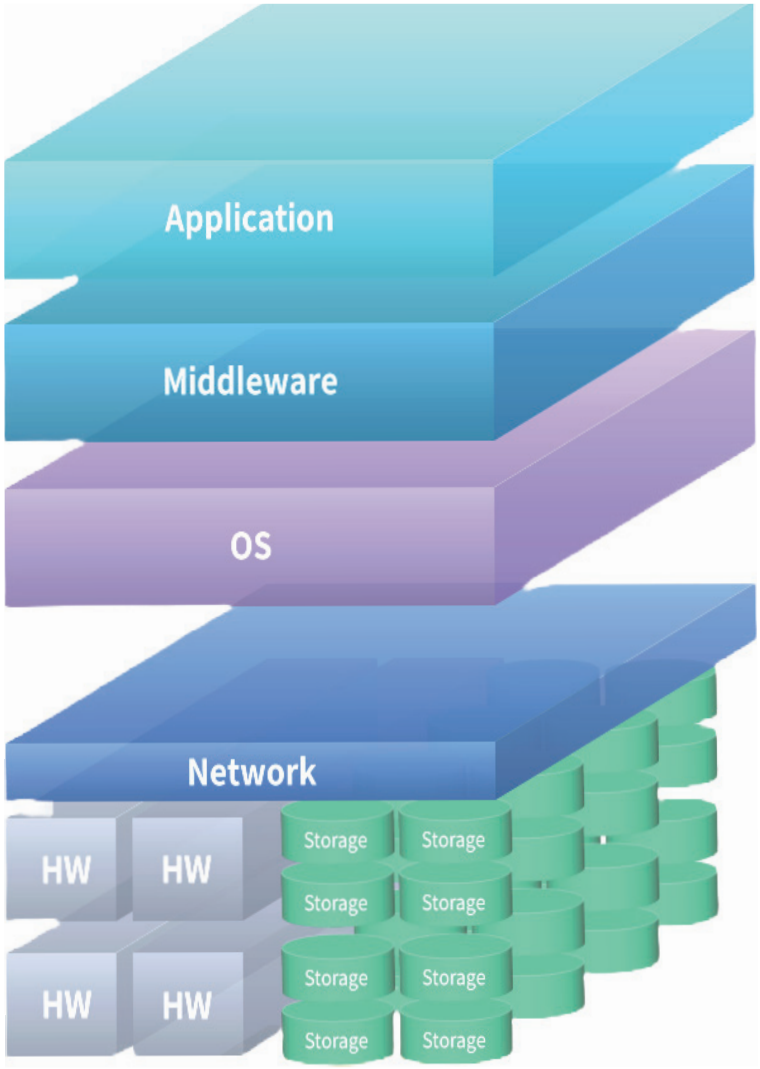
컨설팅부터 구축, 관리, 평가에 이르기까지 다양한 환경에서 HPC 구축에 대한 차별화된 역량을 보유하고 있습니다.

Consulting

- Application에 최적화된 클러스터 구성
- On-premise cluster 성능 최적화 컨설팅

System Intergration & Maintenance

- Application
  - ·GPGPU Programming 최적화
  - ·HPC, AI, Bioinformatics application 최적화
- Middleware
  - ·Cluster Scheduler(Slurm 등)
  - ·Cluster Monitoring Tool
  - ·Kubernetes, OpenStack 등
- OS
  - ·OS Tuning
  - ·Authentication SW(LDAP, NIS) 지원
- Network
  - ·고속 InfiniBand, Ethernet Network 구축
  - ·Fat-tree 구성, 이중화 지원 등
  - ·MPI, UCX 지원 및 최적화
- Storage
  - ·고속 병렬 스토리지 지원
  - ·Lustre, Ceph, HPE GLFS, HPE QUMULO 지원



Evaluation

- HPL, MLPerf 벤치마크

Software Products

01. Storage : MCSxHPE

Hewlett Packard Enterprise

1. GLFS(Green Lake File Storage)

1) AI 모델 훈련, 서빙, 클라우드 서비스를 위한 스토리지 솔루션
2) HDD 없는 All flash로 구성 가능
3) RDMA(GPUDirect Storage) 지원을 통한 고성능 AI 워크로드 지원
4) 하드웨어 + 소프트웨어 솔루션을 함께 제공

·하드웨어 : Alletra Storage MP

· Compute enclosure, Switch, Storage Enclosure 3가지로 구성

· scale-up/scale-out이 용이: 필요에 따라 용량을 늘리거나 (D-node 증설) 연산 성능을 높일 수 있음(C-node 증설)

· 소프트웨어

· Docker 컨테이너 기반 솔루션

· 다양한 I/O 성능 최적화 완료

*Compute enclosure(C-node): 유저레벨 서비스에 필요한 연산을 담당 (API 서버, RDA(Remote Direct Access), 클라우드 등)
*Switch : C-node와 D-node 간 고속 데이터전송 네트워크 지원 (NVMe fabric, infiniband 등)
*Storage Enclosure(D-node) : 고속 데이터 전송, 저장, 읽기 쓰기 연산을 담당

2. QUMULO

1) 대용량 AI 처리를 위한 간편하고 경제적인 NAS 솔루션
2) 다양한 파일공유 프로토콜 지원 (NFS, SMB, REST등)
3) HDD/SSD를 혼합하여 구성 가능
4) SSD Cache 기능 지원

02. gadgetini : dg 클러스터 통합 관리 솔루션 (개발 중)

· dg 클러스터 통합 관제

· dg 액체냉각 시스템 정보(온도, 습도, 유속, 누수, 유량 등) 제공

· 시스템 정보(CPU, GPU, 메모리, 네트워크 등) 제공

· 스토리지 정보(사용량, 대역폭 등) 제공

· Slack, gmail 등 알람 기능 지원

· 자원 및 작업 관리

· K8s 기반의 컨테이너 단위 자원 격리 및 가상화 지원

· Slurm 기반의 GPU 작업관리 지원

· 스토리지 자원 관리 및 가상화 지원

· Workload automation 지원

· Batch 작업 생성, 스케줄링, 자원 할당, 가상화

· MLOps/AIOps/DataOps 기능 제공 (추가 솔루션)

· 원격 기술지원

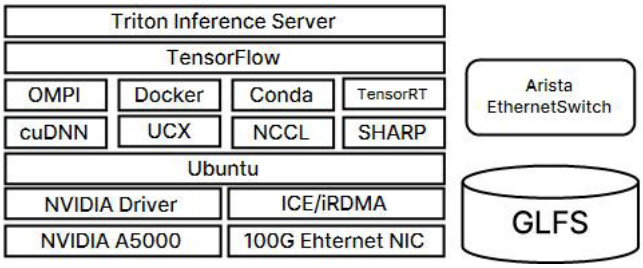
· 클러스터 idle 자원을 활용한 공유 컴퓨팅 (추가 솔루션)

· 공유 등록된 idle 상태의 자원을 다른 워크로드에 대여하여 컴퓨팅 자원 활용률 극대화

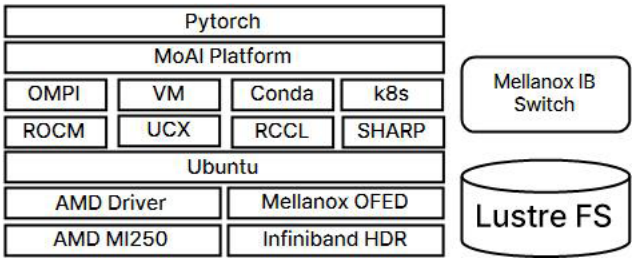
구축 사례

01. AI 인프라 구축 사례

1) 대규모 GPU 클러스터 빌드
2) 3000개 이상의 GPU 설치
3) 500대 이상의 서버를 구축 및 운영
4) IB 스위치 64개, IB 1,690개 이상 구성



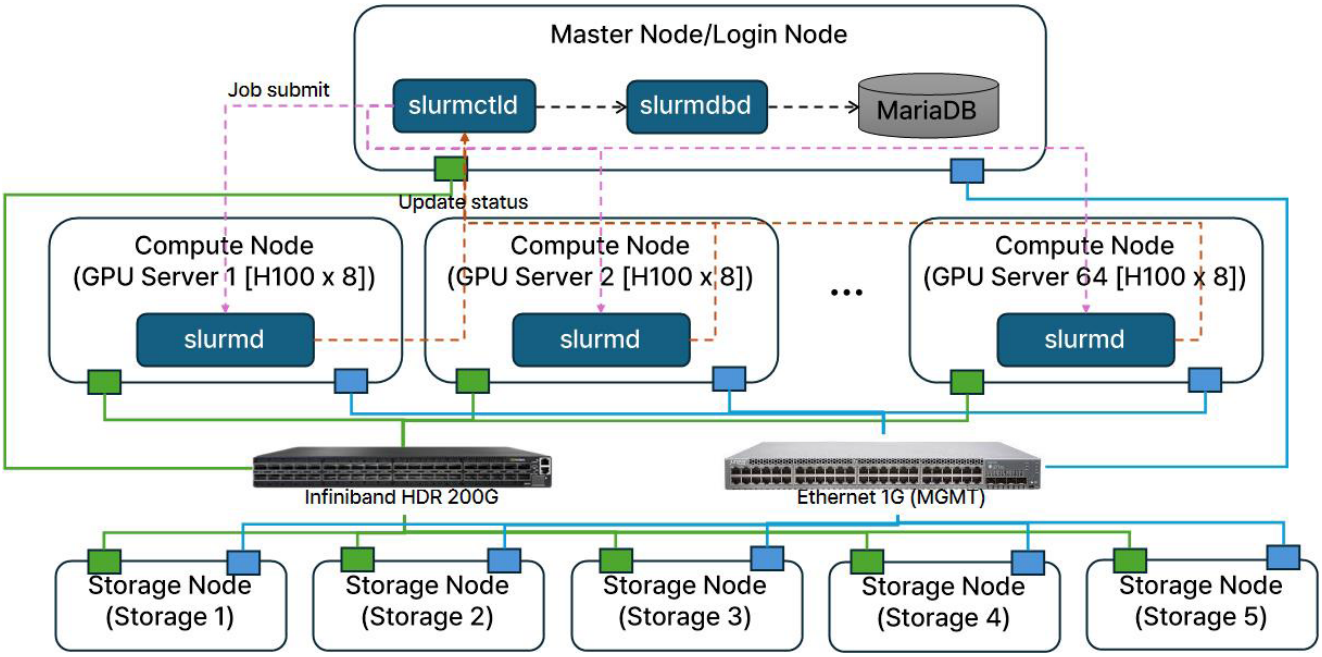
AI Inference System with NVIDIA GPUs



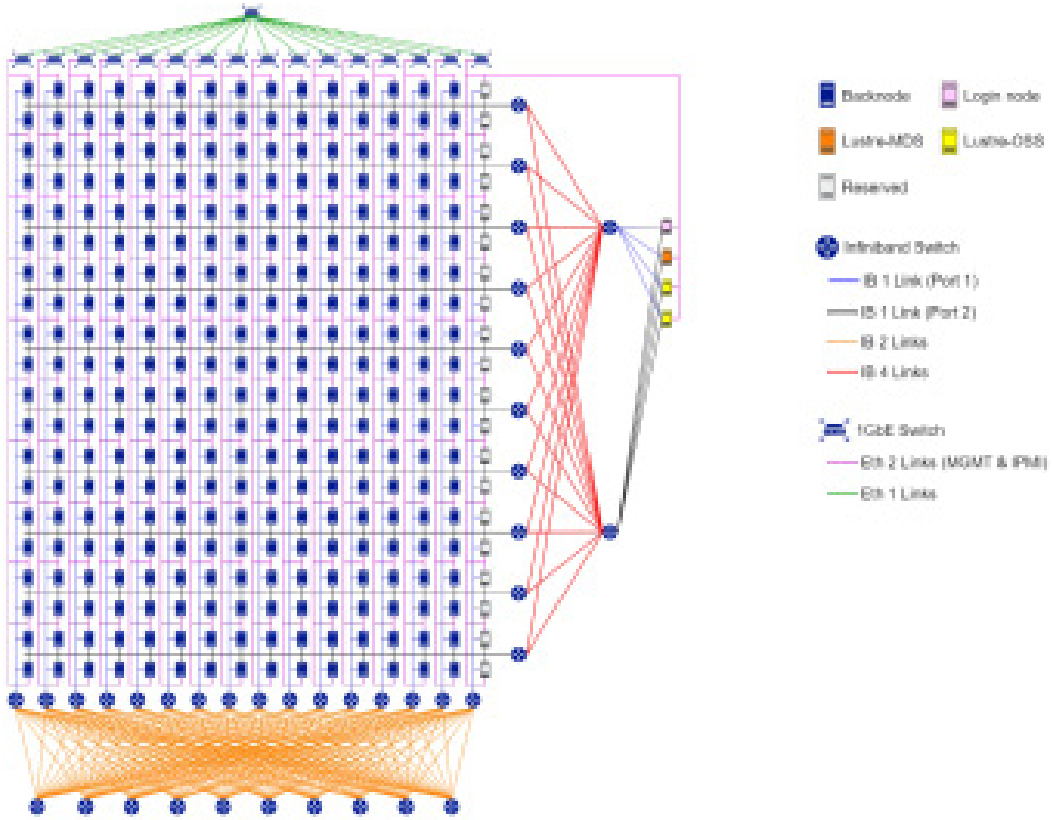
AI Training System with AMD GPUs

02. SLURM 구축 사례

· 국내 최대 중공업 기업
· 'L' SI 업체

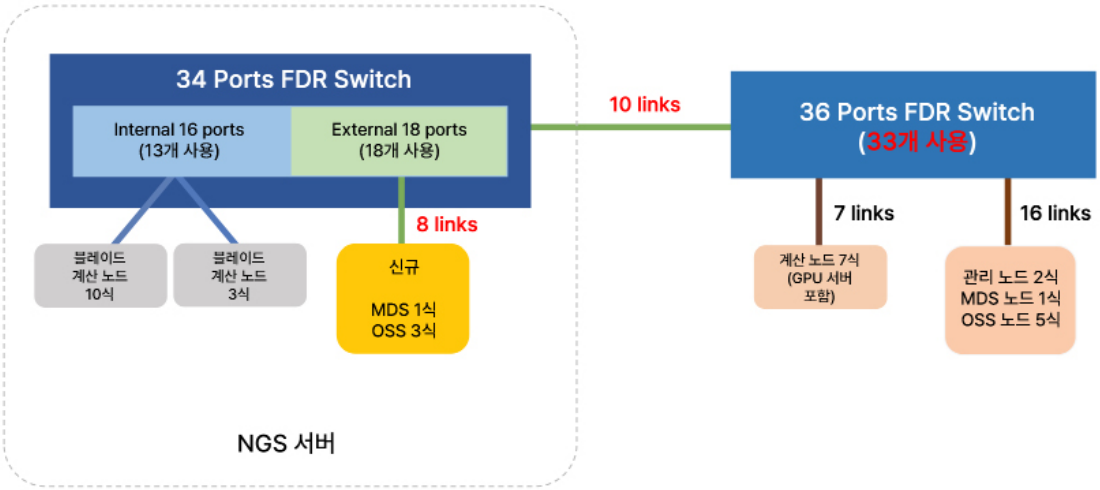


· 국내 최대 AI 클라우드



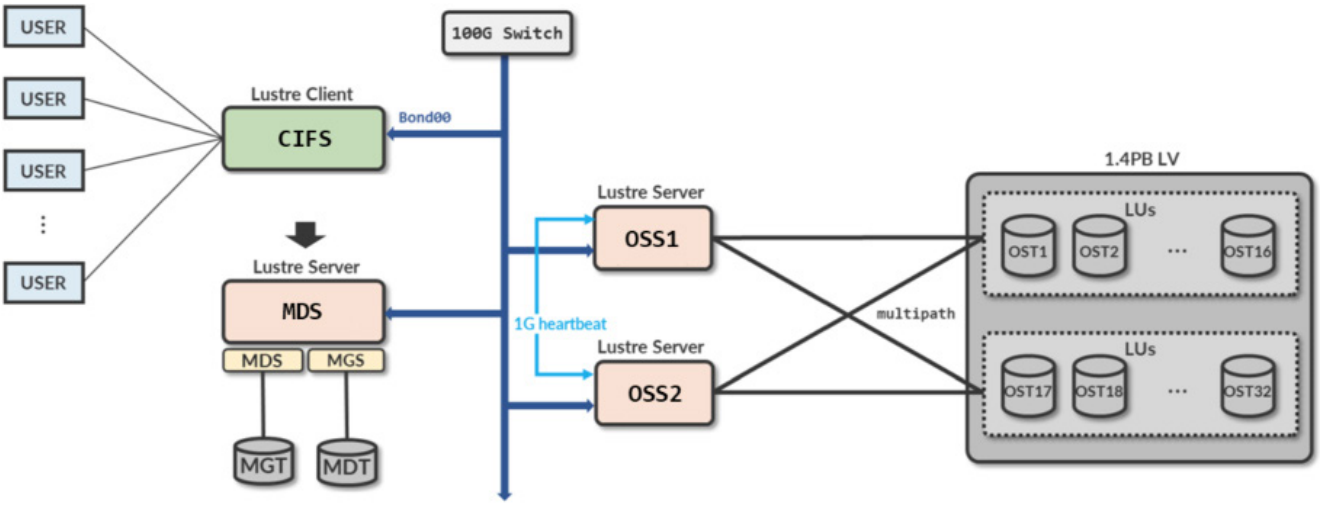
03. 고속 병렬 스토리지 Lustre File System 구축 사례

· 국립대학교 병원



· 컨텐츠 제작 상장 기업

1) 1.4PB 규모 분산 스토리지 클러스터
2) CIFS 서버에서 SAMBA로 Windows 유저 접속환경 제공
3) 엔드유저 I/O Throughput: 10Gbps



04. 클라우드 및 컨테이너 기반 클러스터 구축 사례

· 국내 최대 AI 클라우드

1) 300노드 규모 클라우드 서비스용 클러스터 구축
2) Docker container 기반 자원 할당

· AI 스타트업 데이터센터

1) 120노드 규모 내부 개발용 클러스터 구축
2) Kubernetes 기반 자원할당 및 스케줄링

· 국립대학교

1) 연구용 클러스터
2) Slurm 기반 자원할당 및 스케줄링

Support

든든한 3년 간의 품질 관리 서비스

매니코어소프트의 모든 제품 구입 시 업계 최장 기간인 3년간 품질 보증, SW 복구 서비스, HPC 기술 지원 서비스를 제공합니다.

3년 HW 수리

3년 SW 복구 서비스

3년 HPC 기술지원

· **HW 수리**

구입 후 3년 내에 발생하는 HW 장애에 대해 수리해드립니다.
직원이 방문하여 직접 서버를 회수하며, 수리 완료시 직접 가져다 드립니다.

보증 범위 : Chassis, dg 액체냉각 시스템, Power Supply, CPU, GPU, Memory, Infiniband NIC, Motherboard, Storage(내부 데이터 제외)

· **SW 복구 서비스**

SW 장애발생 시, 출고 시 설치했던 SW를 복구해드리는 서비스를 지원합니다.
직원이 방문하여 직접 서버를 회수하며, 설치 완료 시 직접 가져다 드립니다.

· **HPC 기술지원**

HPC 기술을 바탕으로 GPU 클러스터 및 클라우드 기술을 종합적으로 활용하여,
고객의 필요에 맞춘 최적의 HPC 솔루션 및 AI 환경을 제공합니다.
매니코어소프트의 기술지원과 함께 안정적인 운영을 경험해보세요.

kubernetes PyTorch slurm workload manager TensorFlow NVIDIA CUDA AMD ROCm

*우발적인 손상의 경우 본인 부담금이 부과될 수 있습니다.

Partners & Contact

· **본사 및 연구소** 서울특별시 관악구 관악로 1 서울대학교 138동 308호

· **메인 오피스** 서울특별시 구로구 디지털로27가길 27 (구로동) 9층 매니코어소프트 (08375)

· **E-mail** contact@manycoresoft.co.kr

· **Tel** 070-4443-6660

deepgadget.com

Official Partners

WAVEFIVE
All about AI Computing


deep gadget