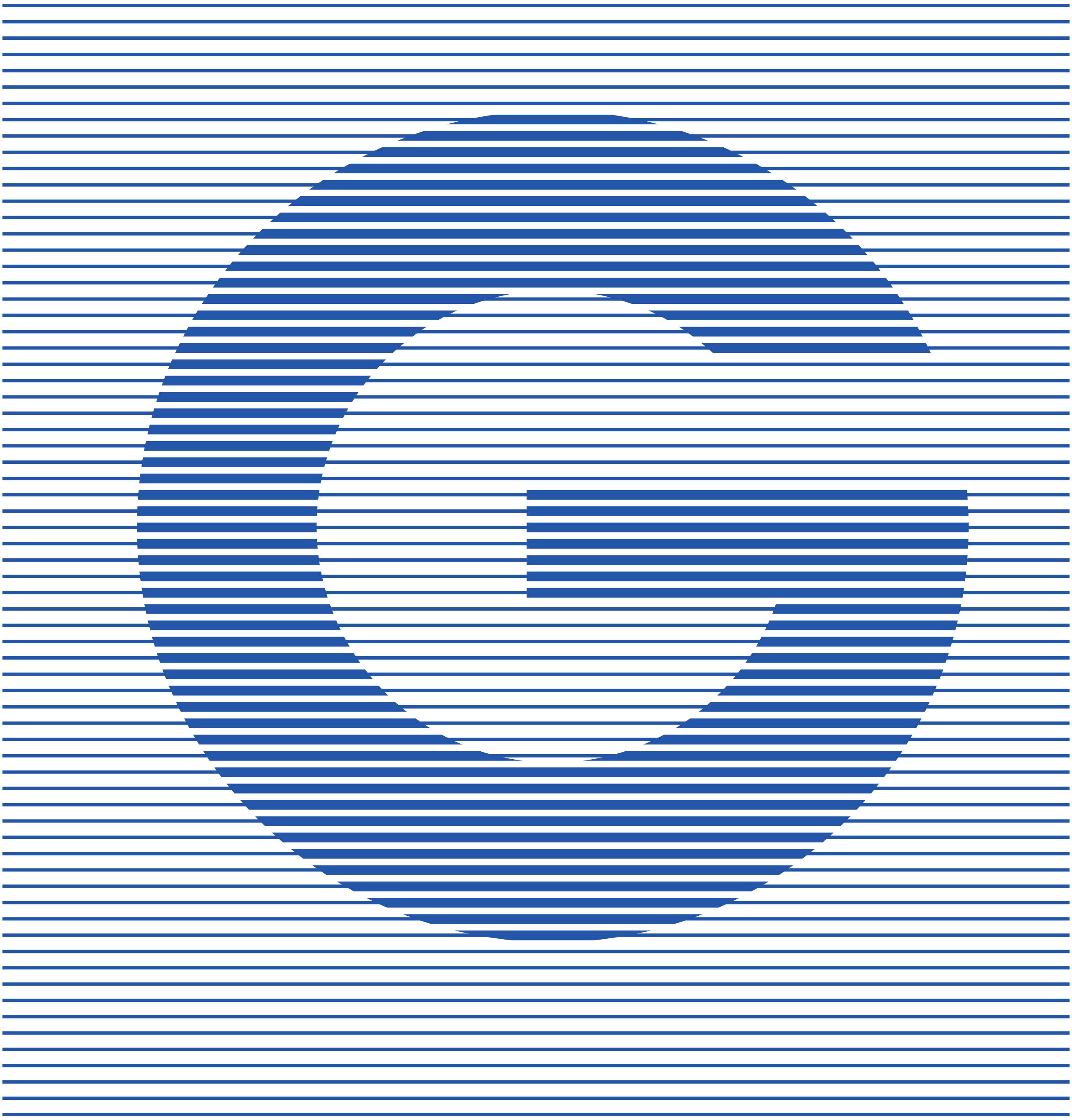
deep gadget



Every great thing starts from the basics.

So does AI infrastructure.

deepgadget.com

MANYCORESOFT

deep gadget

dg = deep learning serving Gadget

The new logo of Deep Gadget embodies the skived heat sink, a fundamental element of all cooling, sculpted into the first letter of "Gadget." This symbolizes the philosophy and technical prowess of Deep Gadget, dedicated to designing rigorously and precisely from the very basics to create the ultimate computing power.

AI infrastructure starts from the basics.

‣ **Company**
ManyCoreSoft Introduction 04
Services & Background 05
Leadership 06
MCS History & Educational Initiative 07
‣ **Why liquid cooling?** 08
‣ **dg® liquid cooling**
Exclusive dg® Liquid Cooling Technology 10
5 reasons to trust us with Confidence 12
‣ **Enterprise**
For Large-Scale AI Clusters, 14
Save costs with dg®
‣ **H/W Product** 16
dg5 Workstation 18
dg4 Series & beyond GPU(NPU) 19
‣ **H/W Solution : dg-Transplant** 21
‣ **H/W One pager** 22
‣ **S/W Service**
S/W Full Stack One pager 23
S/W Products 24
Implementation Cases 25
‣ **Support** 28
‣ **Partners & Contact** 29

Introduction to ManyCoreSoft

ManyCore + Soft!

ManyCoreSoft leverages optimization capabilities across **both Hardware and Software** domains to deliver **unparalleled computing power** across various industries, including Hyperscale AI.

ManyCoreSoft, founded by members of Seoul National University's Multicore Computing Laboratory (now known as Thunder Lab), is a high-performance computing specialist. "Many-cores" refer to architectures integrating hundreds to thousands of cores beyond 2-8 multicore processors, efficiently handling numerous tasks.

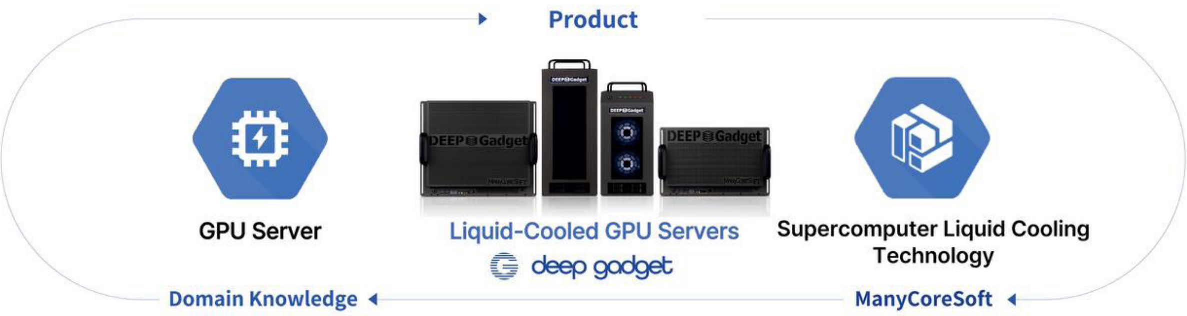
In achieving High Performance Computing (HPC), both hardware and software capabilities are crucial. "ManyCoreSoft" is a portmanteau of "Many-cores" and "Software," symbolizing expertise in both aspects, offering comprehensive support as a full-stack AI Company across the entire AI industry technology stack.

ManyCoreSoft provides services spanning from high-performance GPU liquid-cooled server design and manufacturing to large-scale AI infrastructure consulting, deployment, and management. Leveraging the globally recognized research achievements of Seoul National University's laboratory, particularly in HPC fields utilizing accelerators such as GPUs and FPGAs, we continuously collaborate with leading companies and institutions.

Over the past decade, ManyCoreSoft has meticulously honed optimization capabilities across hardware and software. Presently, armed with unparalleled expertise, we provide superior computing power across various industries, including Hyperscale AI. Furthermore, we prepare for the future beyond AI and GPUs, advancing toward tomorrow's challenges.

Services

Server Manufacturing



Product

GPU Server

Liquid-Cooled GPU Servers

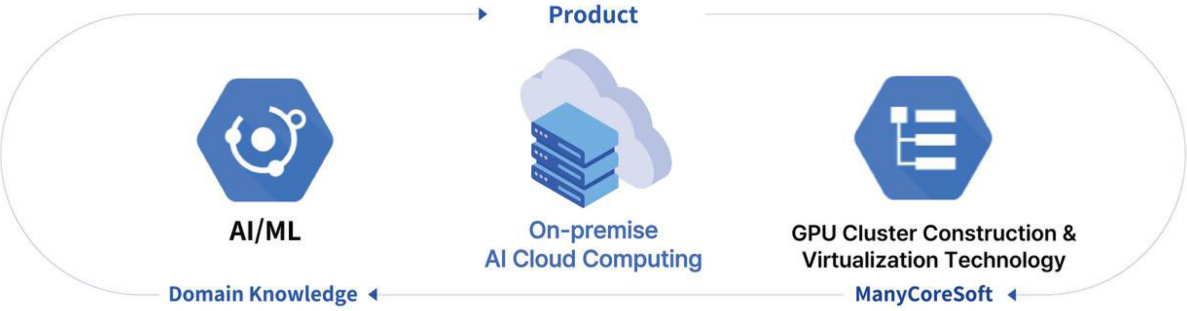
deep gadget

Supercomputer Liquid Cooling Technology

Domain Knowledge

ManyCoreSoft

HPC Infrastructure Construction



Product

AI/ML

On-premise AI Cloud Computing

GPU Cluster Construction & Virtualization Technology

Domain Knowledge

ManyCoreSoft

Background

Development of the open-source OpenCL programming environment SnuCL

· Abstraction of diverse accelerators across multiple nodes as if they were on a single node

· Attains high performance and ease of programming in heterogeneous cluster systems

· Utilized by schools, companies, and research institutes in over 70 countries worldwide

· Presented in 9 papers and 9 tutorials at top-tier academic conferences

Development of domestically designed GPU supercomputer 'Thunder'

· Focuses on cost-effective, energy-efficient design

· Ranked 277th on TOP500 and 32nd on Green500 lists

· Recognized as 7th most power-efficient on TOP50 list

· The world's first liquid-cooled supercomputer with consumer GPUs

· Achieved top performance on single-node universal hardware through software optimization

Leadership

CEO

Jungho Park

· 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

prof. Jaejin Lee

- 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	Foundation of ManyCoreSoft
2012. 10	Developed and built the heterogeneous supercomputer 'Thunder' at 1/50th the cost, achieving a national first and ranking 227th on the TOP500 list (7th globally in energy efficiency)
2013. 07	Consulting contract with Koscom for the first real-time business processing solution in the IB industry
2014.	Various partnerships including HPE, Intel, AMD for tech development and infrastructure projects
2015.	Liquid cooling system selected as 'SC2015 Emerging Technology 10'
2016.	Registered 2 patents for GPU liquid cooling (Patent registration number: 10-1954306, 10-2118786)
2017.	Achieved advancements in OpenCL kernel code automation and parallel processing standardization
2018.	Development of machine learnin solutions for Financial services, Contributed to the design of accelerated hardware for SK Hynix
2018. 12	Established a hardware plant / Launched 'deep gadget'
2019. 10	T3 credit ratings by TCB / Venture enterprise certification
2020. 08	Construction of large-scale GPU clusters in the KT Cloud Platform
2021. 08	Comprehensive cooperation agreement with Moreh for hardware infrastructure
2022.	Installed over 3000 GPUs annually
2023.	Acquisition of more than 100 customers
2024.	Release of the dg5 server & HW Solution dg-tp / Featuring partnerships including collaboration with Tenstorrent.

Educational Initiative

Since 2013, ManyCoreSoft has been organizing the "Accelerator Programming School" together with Seoul National University's Thunder Research Lab, aiming to raise awareness of the importance of Accelerator Computing and cultivate specialized programming talents capable of harnessing its

- **Since 2013, annual 4-night, 5-day sessions held in summer and winter**

- **Curriculum includes parallel computing, GPU architecture, GPU programming (CUDA/ OpenCL), and optimization techniques**

- **For Graduate students and Researchers from academia and industry.**

2013 Accelerator Programming 겨울학교
일시: 2월 18일 ~ 22일 장소: 대교HRD센터 주최: 서울대학교 매니코어프로그래밍연구단 주관: 매니코어소프트(주)

< 2013 Winter School >

< 2023 Summer School >

Why liquid cooling?

" The future of computers lies in Liquid Cooling. "

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

In the past, air-cooled internal combustion engine cars without liquid cooling radiators had to be parked with the hood open for about 2 hours after just 30 minutes of driving to cool the engine. However, with the development of liquid cooling radiators that cool the engine directly, it became possible to run the engine for 5 to 6 hours or even longer.

The same applies to computers. In the past, most high-performance servers rarely exceeded a total power consumption of 1kW per unit, and core cooling was possible with air cooling. However, today, for various purposes such as AI training, inference, computation, and rendering, each server requires 3kW to 6kW or more. As performance increases, heat generation continues to rise proportionally.

As a result, **installing multiple GPUs on a typical air-cooled server can cause a decline in GPU performance by over 10-20%.** This can lead to system instability and voluntary performance throttling of CPU, GPU, and NPU. Continuous high heat also affects product durability, resulting in a shortened lifespan.

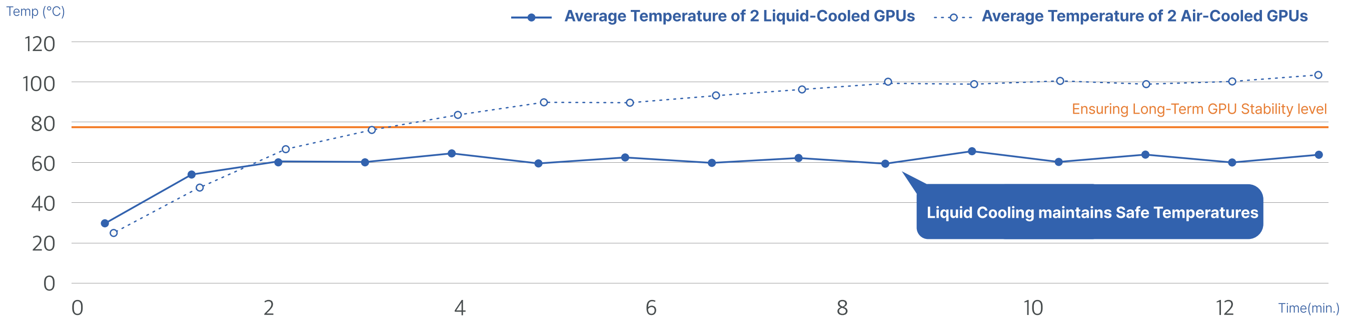
As a result, they face the challenge of purchasing high-cost server equipment that they cannot effectively utilize. This issue is reported not only among individual users but also in state-of-the-art IDCs (Internet Data Centers) operated by major corporations, where operational difficulties arise due to related issues.

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

The thermal conductivity of water exceeds that of air by over 300 times, allowing it to absorb and dissipate more heat effectively. Currently, the only solution to address this issue is Liquid Cooling using water.

Liquid Cooling vs. Air Cooling:

GPU Heat and Performance Degradation Rate



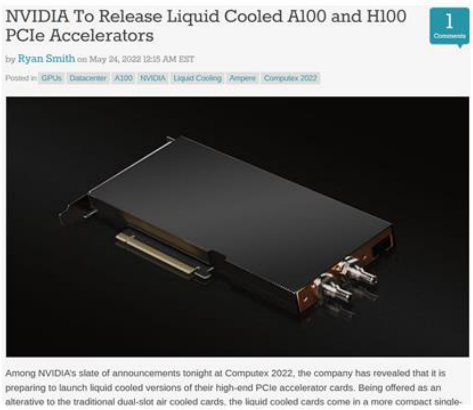
Degradation Rate relative to Maximum Performance

Air Cooling : Decline	97%	95%	94%	92%	90%	89%
Liquid Cooling : Maintain	98%	98%	98%	98%	98%	98%

With Liquid Cooling, you can..

- **Keep systems cool with minimal impact from indoor temperatures.**
- **Improve energy efficiency by reducing cooling power consumption.**
- **Install high-density computing devices with ease.**

Since 2022, infrastructure companies including NVIDIA have been actively adopting **Direct Liquid Cooling.**



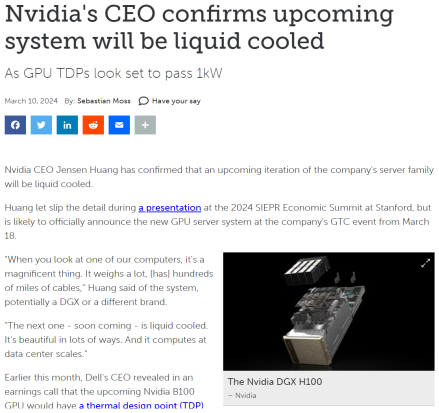
NVIDIA To Release Liquid Cooled A100 and H100 PCIe Accelerators

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

Exclusive dg® Liquid Cooling Technology

Supercomputer R&D Experts X Toyota Automotive Cooling Engineer

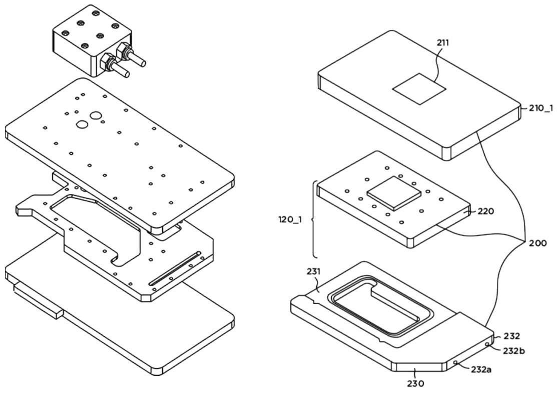
deep gadget® isn't just another custom liquid cooling system. It's a state-of-the-art, next-generation cooling solution that integrates the supercomputer expertise from Seoul National University's research lab, the design skills of a Toyota headquarters engineer with 20 years of experience, and a decade of intensive R&D.

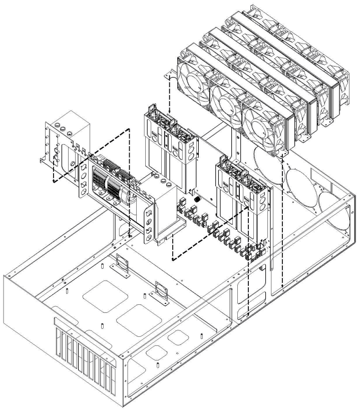
- Patented cooling technology and advanced hardware design expertise.
- Custom-designed cold plates for liquid cooling a wide range of AI accelerators, including NVIDIA AI GPUs, gaming GPUs, CPUs, NPUs, and Infiniband NICs.
- Fast integration of cutting-edge hardware components like CXL, NMC, and PIM for next-gen computing and memory/storage solutions.

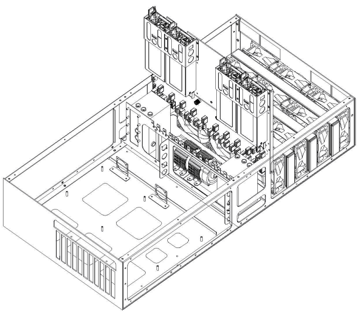
Patents:

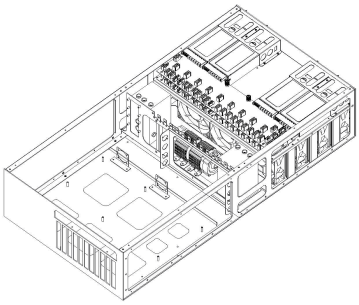
- COOLING DEVICE USING WATER FOR COMPUTER AND DRIVING METHOD THEREOF(컴퓨터용 수냉식 냉각장치 및 그 구동방법) Patent No : 10-2118786

- Cooling Plates for High-Density GPU Liquid Cooling(고밀도 GPU 액체 냉각을 위한 냉각판) Patent No: 20-0477833, 20-0479465









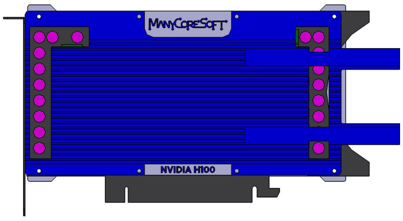
Built-in Liquid Cooling without the Need for Additional devices

Direct Liquid Cooling + Cutting-Edge Channel Design = Unrivaled Cooling Performance

With just one deep gadget®, up to 16 A100 GPUs can operate seamlessly.

- **Usable even at room temperature (30°C and above), no separate temperature or humidity control required**
- No need for external devices like chillers or piping
- Cost savings in infrastructure setup and management, high energy efficiency

Direct Liquid Cooling System

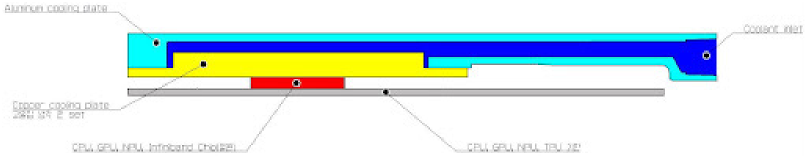


1. Directly attaches high thermal conductivity copper cooling plates to heat sources for rapid heat conduction

2. Simultaneously passes high-density fins of the copper cooling plates through a cooling liquid cooled by radiators and cooling fans

These two simultaneous processes maximize GPU cooling

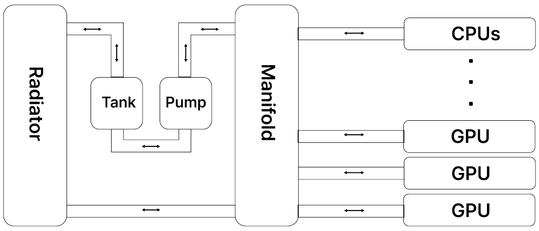
※ Heat Conduction : Heat source(Chip) → Liquid → Radiator → Air



1:1 Parallel Channel Design

Independently cools each component to deliver optimal performance.

(Other liquid cooling systems: serially connect all components, leading to inefficient heat accumulation)

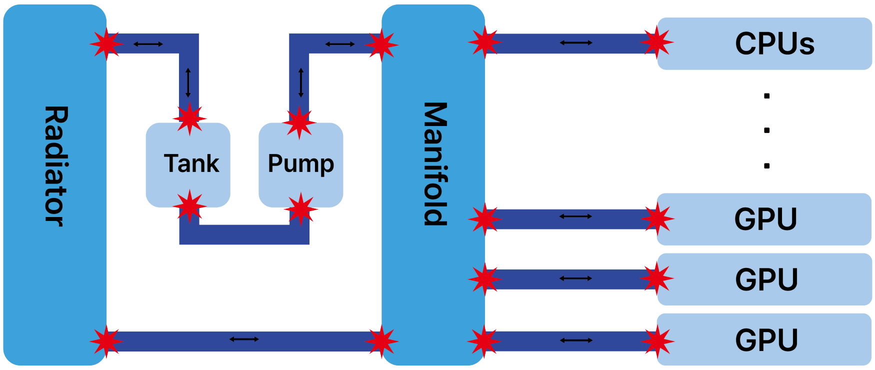


5 reasons to trust us with confidence

01. Zero Incidents of Leakage for 10 Consecutive Years

deep gadget® has a flawless sales history with zero incidents of product leakage.
With a decade of expertise and trust, we guarantee safety.

02. Design : Fully Sealed



① Special Adhesive All leak points, connected by screws between components, are coated with special adhesive for a fully sealed structure ensuring stability even under impact.

② Hose, Fitting, Clip Utilizing hose components selected through meticulous design, ensuring a seamless structure.

③ Quick Connector The quick connectors guarantee sealing even with outputs exceeding 3 times the set values for hydraulic flow, preventing risks between CPU, GPU, and the manifold (safe detachment/attachment during server operation).

④ Rack Shelf Equipped in the event of flooding or leaks, our system ensures complete prevention of secondary damage. The rack shelf, larger than the dg4F standard size, holds approximately double the cooling liquid capacity of the server (43cm x 89.5cm x 2.5cm). It undergoes testing during manufacturing and is finished with waterproof and corrosion-resistant coatings to prevent corrosion.

※ dg® Cooling's fully sealed structure ensures safety and preserves cooling liquid at a 99.99% rate, enabling virtually permanent use without the need for coolant management.

03. Manufacturing : 144 hours of testing over 4 cycles

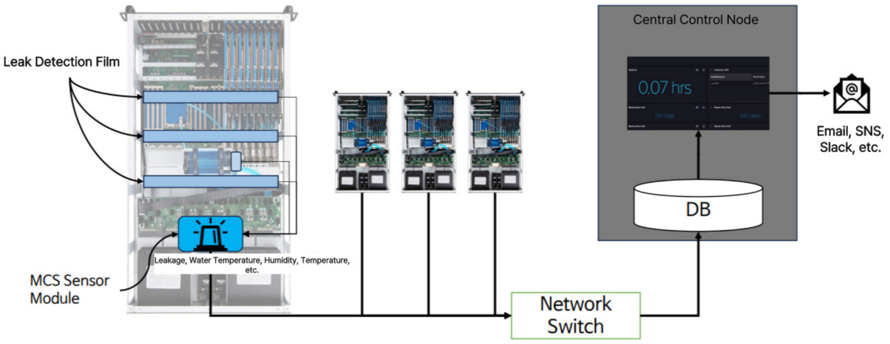
① Rigorous 24-hour test after attaching CPU/GPU cooling plates

② Comprehensive 24-hour test after applying special adhesive to leak points and completing hose connections

③ Thorough 24-hour secondary test of the cooling liquid

④ Extensive 72-hour burn-in test before delivery to ensure unmatched cooling performance and stability

04. Upon Shipment : Control System



05. Post-Shipment: Robust Quality Assurance

We offer an unparalleled 3-year quality guarantee with every purchase, setting the industry standard for reliability. (See page 28 for detailed coverage and terms.) Additionally, our expert guidance helps you safeguard against natural disasters like flooding and leaks, ensuring your investment remains secure.

Coolant information

Our product employs top-quality coolant, ensuring high cooling efficiency while safeguarding against corrosion, freezing, bacteria, algae, and other contaminants. Total capacity based on dg4F standards: 1.3 liters.

***Coolant Composition**

·Distilled Water: 70~75%

· Propylene Glycol: 25~30%

· Potassium Phosphate Dibasic: ≤ 1%

· Sodium molybdate: ≤ 1%

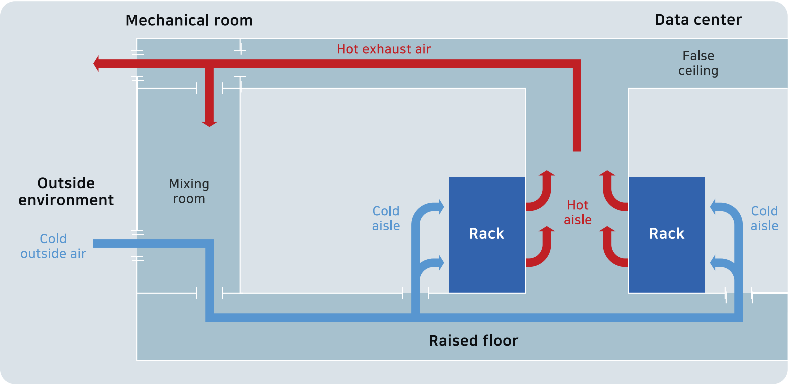
· Meta-toluic Acid: ≤ 1%

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

For Large-Scale AI Clusters, Save costs with dg®.

Built-in deep gadget® system offers Free-Cooling!

- Built-in liquid cooling boasts overwhelming performance, eliminating the need for additional devices.
- Capable of operating smoothly even in high-temperature environments exceeding 30°C.
- In domestic climates, Remarkable energy efficiency can be achieved with only deep gadget and outdoor air
- Easily lower PUE to 1.1 or below.



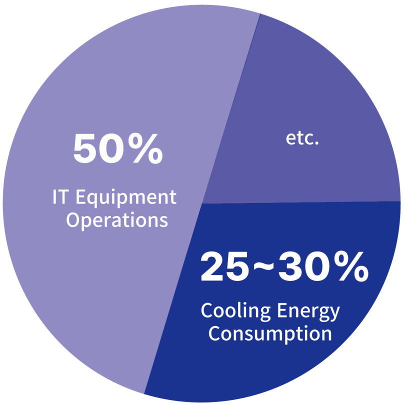
*PUE = Power usage effectiveness)

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

Datacenter 268 locations average 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.

Data Center Energy Saving Scenario



Data Center Energy Savings
deep gadget

In GPU server data centers, indoor temperatures are maintained between 18 and 20°C.

deep gadget®'s Chiller-less liquid cooling system can cool indoor temperatures even at 30°C or higher.

On average, a 1°C decrease saves 5% of Energy Consumption.

Comparing to an indoor temperature of 18 degree°C,

25°C = 35%, 30°C = 60% → Energy Savings

30~60% Energy Savings

*A study found that for every 1-degree temperature in the data center, energy consumption can be reduced by 4.3% to 9.8%.

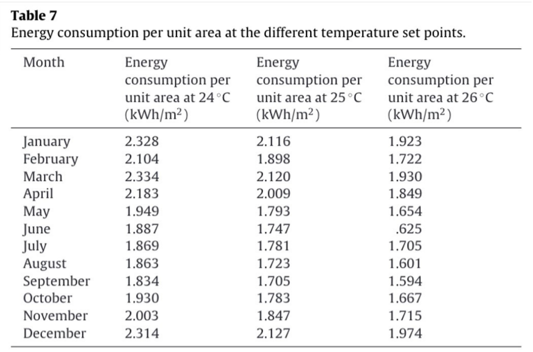


Table 7
Energy consumption per unit area at the different temperature set points.

Month	Energy consumption per unit area at 24 °C (kWh/m^2)	Energy consumption per unit area at 25 °C (kWh/m^2)	Energy consumption per unit area at 26 °C (kWh/m^2)
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

Estimated Scenarios for Cost, Energy, and CO_2 Reduction

Environmental, Social and corporate Governance

▸ **Energy Cost Savings**
- $0.28 million Annually
- $1.42 million over 5 years

▸ **CO_2 Reduction**
- 1,295t Annually
- 6,478t over 5 years

▸ **Energy Cost Savings**
- $0.47 million Annually
- $2.38 million over 5 years

▸ **CO_2 Reduction**
- 2,159t Annually
- 10,797t over 5 years

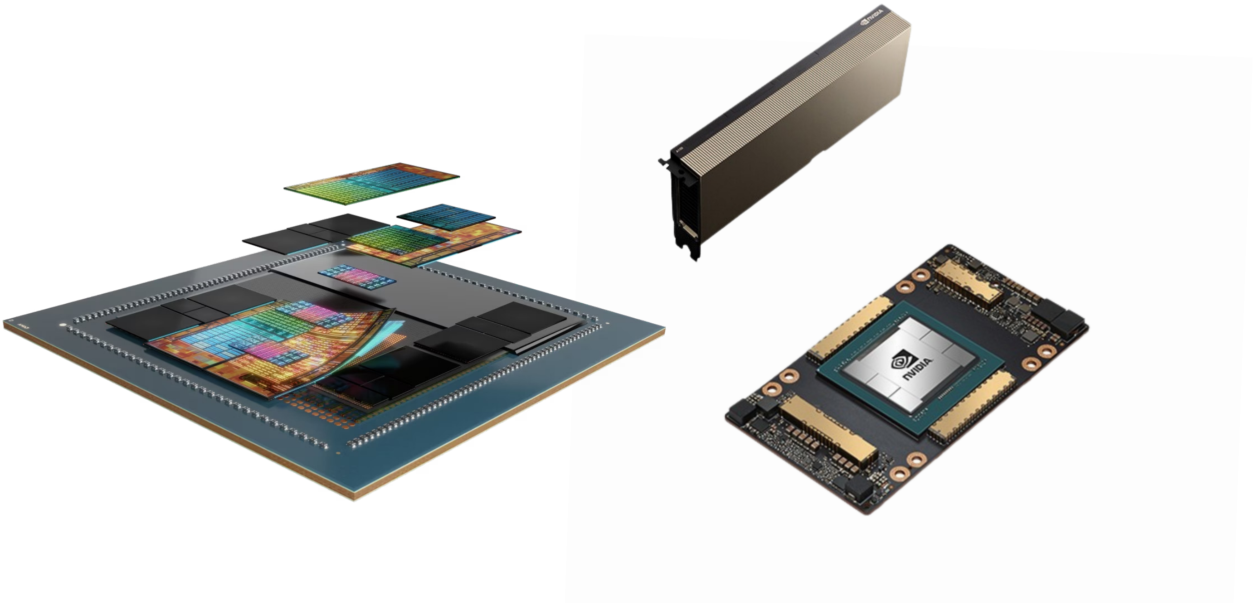
항목	Conventional Air Cooling	dg liquid cooling[30°C]	dg liquid cooling + Free-Cooling
IT Power Consumption	1,000 kW	1,000 kW	1,000 kW
PUE	1.60	1.30	1.10
Total Power Consumption	1,600 kW	1,300 kW	1,100 kW
Annual Energy	14,016,000 kWh	11,388,000 kWh	9,636,000 kWh
Annual Cost	$1.59 million	$1.25 million	$1.05 million
Annual CO_2 Footprint	6,910 t	5,614 t	4,751 t
Cooling Power Consumption	500 kW	200 kW	≈0 kW

Comprehensive Comparison of Cooling Methods

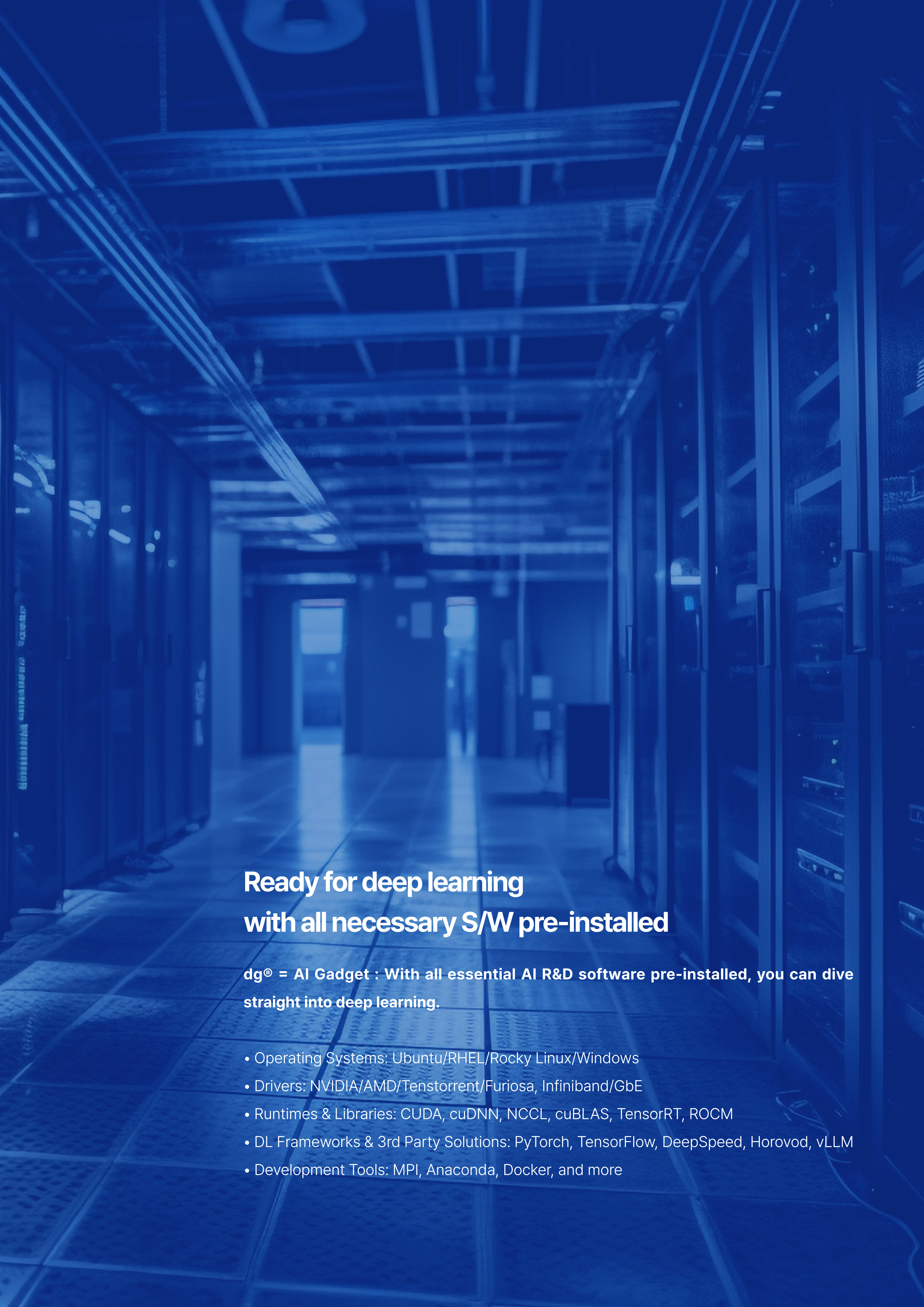
Category	Air Cooling	Water Cooling with Chiller	Immersion	dg® Liquid Cooling + Free-Cooling
PUE	Around 1.6	Around 1.2	Around 1.1	Around 1.05
Equipment Costs	High Cost	High Cost	High Cost	Low Cost
Cooling Performance	Moderate	High	High	High
Server Management	Moderate	Difficult	Very Difficult	Moderate
Overhead Costs	Moderate	High	Very High	Very Low
Density	Moderate	Low	Very Low	High
Noise	High	Low	Low	Low

THE BEST computing power starts from THE BASICS.

Built-in dg® Liquid Cooling and Air Flow design



Support for 30+ accelerators, including Beyond GPUs



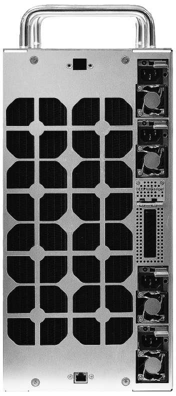
Ready for deep learning with all necessary S/W pre-installed

dg® = AI Gadget : With all essential AI R&D software pre-installed, you can dive straight into deep learning.

- Operating Systems: Ubuntu/RHEL/Rocky Linux/Windows
- Drivers: NVIDIA/AMD/Tenstorrent/Furiosa, Infiniband/GbE
- Runtimes & Libraries: CUDA, cuDNN, NCCL, cuBLAS, TensorRT, ROCM
- DL Frameworks & 3rd Party Solutions: PyTorch, TensorFlow, DeepSpeed, Horovod, vLLM
- Development Tools: MPI, Anaconda, Docker, and more

Performance Beyond Rackmounts : From Data Centers to Labs

dg5 Workstation®



❶ Rackmount Performance Meets Workstation Convenience

Experience the best of both worlds with our versatile server that can be rack-mounted in data centers or used conveniently in your office. No matter the environment or space, achieve top-tier performance effortlessly.

❷ Compact Size, Enhanced Cooling

State-of-the-art channel and cooling plate design delivers powerful and stable cooling performance within the same compact footprint.

❸ Reliable Power Supply in Any Situation

Equipped with 4 redundant power units, our server ensures consistent, powerful, and stable operation at all times.

❹ Simplified Management with Built-In Sensors and Display

Easily manage your system with 5 integrated sensors that detect internal operations and data, all displayed on a built-in screen. A separate monitoring system will also be available for purchase.

- **Built-In dg® Liquid Cooling**
- **Supports 7 x PCIe Gen5 × 16 Slots**
  - · Direct connection with no PCIe switch
- **Cooling up to 7 high-performance GPUs**
  - · 2 Pumpss, 4 Radiators
  - · 16 Cooling Fans: powerful performance
- **Quiet operation :** up to 50dB, suitable for standard office environment

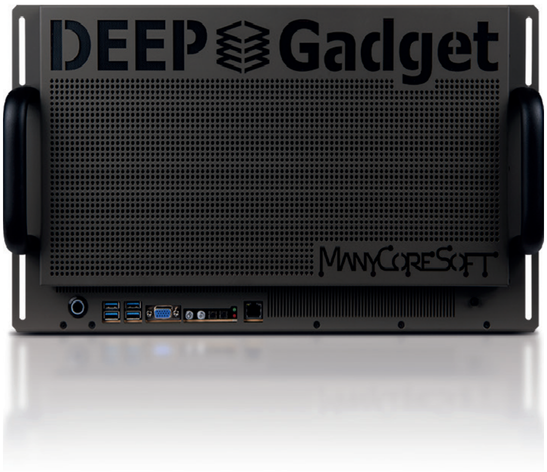
- **For Customers Needing Powerful Computing Anywhere, Anytime**
  - · Data centers, enterprises, research institutions, individual professionals
  - · AI model training/inference
  - · GPU farms for AI, rendering, encoding, and more
  - · Multi-GPU image processing (encoding, rendering, CAD, etc.)

dg4 Flagship®

Flagship server with overwhelming performance

- **Built-In dg Liquid Cooling**

- **Supports Dual EPYC/Xeon CPUs**
- **Direct Connection with No PCIe Switch**
  · 19 PCIe Gen4×8 Slots
- **Efficiently Cools up to 16 High-Performance GPUs**
  · 4 High-Flow Pumps
  · 6 Radiators
  · 18 Cooling Fans
- **Perfect for Maximum GPU Parallel Processing:**
  · Ideal for Large-Scale AI Model Training/Inference
  · Optimized for Scientific Computation and Simulation



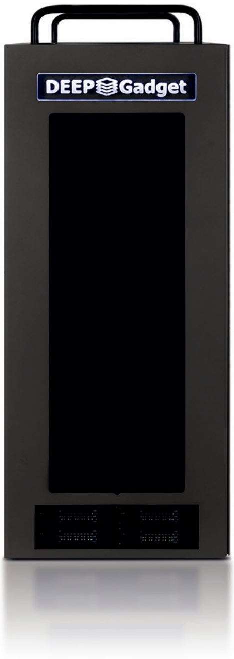
dg4 Rackmount®

The new basic of computing power.

- **Built-In dg Liquid Cooling**

- **Supports Dual EPYC/Xeon CPUs**
- **Direct Connection with No PCIe Switch**
  · 9 PCIe Gen4×16 Slots
- **Efficiently Cools up to 10 High-Performance GPUs**
  · 2 High-Flow Pumps
  · 3 Radiators
  · 9 Cooling Fans
- **Optimized for High GPU-CPU Communication:**
  · Distributed Parallel AI Model Training
  · Ideal for AI, Rendering, and Encoding

H/W Product



dg4 Workstation®

Powerfully, Quietly, Compactly.

- Built-In dg Liquid Cooling

- Supports High-Performance Threadripper Pro Workstation CPU
- Direct Connection with No PCIe Switch
  - 7 PCIe Gen4×16 Slots
- Efficiently Cools up to 7 High-Performance GPUs
  - 2 High-Flow Pumps
  - 2 Radiators
  - 7 Cooling Fans
- Quiet Operation at Below 50dB : suitable for typical office
- Perfect for Large-Scale GPU Tasks in the Office
  - Small to Medium-Scale AI Research
  - Multi-GPU Image Processing (Encoding/Rendering/CAD)

beyond GPU

GPU vs. NPU

· GPU: General AI trading and Inference, NPU: Specialized Applications

· NPUs are more energy-efficient and economical compared to GPUs

AI Serving System Utilizing NPU Technology

· LLM Serving Gadget with Tenstorrent

· Image AI Serving Gadget with FuriosaAI

LLM Serving Gadget

tenstorrent

Model : dg-LLM-n300

NPU: Tenstorrent Wormhole n300 x 16(Max)

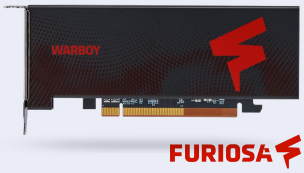
NPU Memory: 384 GB

LLM Performance: 4,192 TOPS (FP8)

TDP: 5.6 kW

Tenstorrent AI card

Vision AI Serving Gadget



FURIOSA

Model: dg-VISION-WB

NPU: Furiosa AI WARBOY x 16(Max)

NPU Memory: 256 GB

Vision Performance: 1,024 TOPS (INT8)

TDP: 1.8 kW

ManyCoreSoft is partnering with leading NPU manufacturers to usher in the next era beyond GPUs.

dg-Transplant®

Are you struggling with the heat generated by your GPUs?
Unable to utilize your expensive servers to their full potential?

With the **dg-Transplant Solution**, you can intergrade dg liquid cooling to your existing third-party GPU servers for optimal performance.

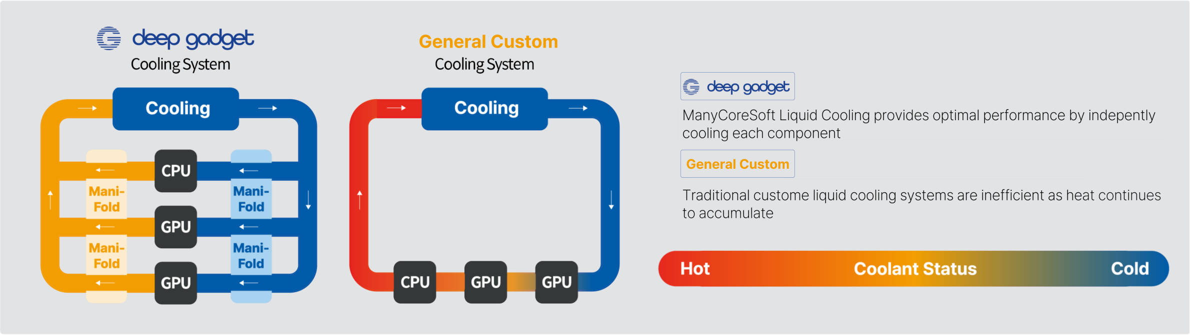
Experience unmatched efficiency, swiftly dissipating heat from all cores and memory in any system:

· Utilize 100% performance of your high-end servers at approximately 10% of the cost of new servers.

· Worry-free maintenance with our completely sealed design

· 3-Year Quality Guarantee

· (Upcoming Feature) Monitor server status and receive alerts with our advanced sensor integration

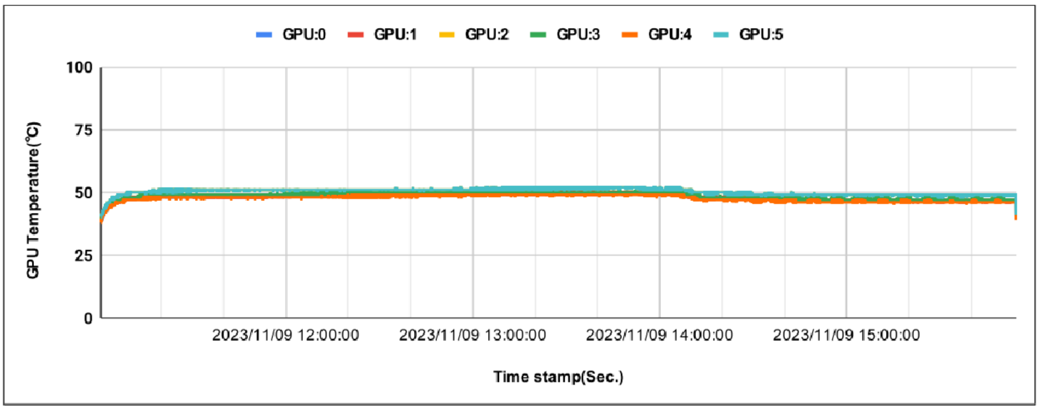


ManyCoreSoft Liquid Cooling provides optimal performance by indepently cooling each component

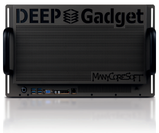
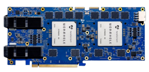
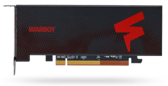
Traditional custome liquid cooling systems are inefficient as heat continues to accumulate

dg4R-4090 6 Units Full Load Test (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	Max 512GB DDR4-3200 ECC	Max 2TB DDR4-3200 ECC	Max 2TB DDR4-3200 ECC	Max 512GB DDR4-3200 ECC	Max 1TB	Max 512GB
M.2 NVMe SSD	2 Slots	1 Slots	1 Slots	2 Slots	1 Slots	1 Slots
PSU	4 × 2,500W ≤ Hot swappable	4 × 2,500W ≤ Hot swappable	4 × 2,500W ≤ 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		

Accelerator Support Lineup	
For Large-Scale Training	NVIDIA H100, NVIDIA A100, AMD MI300X, AMD MI250, AMD L40S etc.
For Inference and Small-Scale Training	AMD M210, NVIDIA RTX 6000 Ada, NVIDIA RTX 4090, NVIDIA RTX 4080 etc.
For Inference Only	Tenstorrent Wormhole n300 NPU
Vision-Specific	Furiosa WARBOY NPU

*New architectures are continuously being added.

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 etc.

*New support models and services are continuously being added.

S/W Full Stack One pager

Our Service

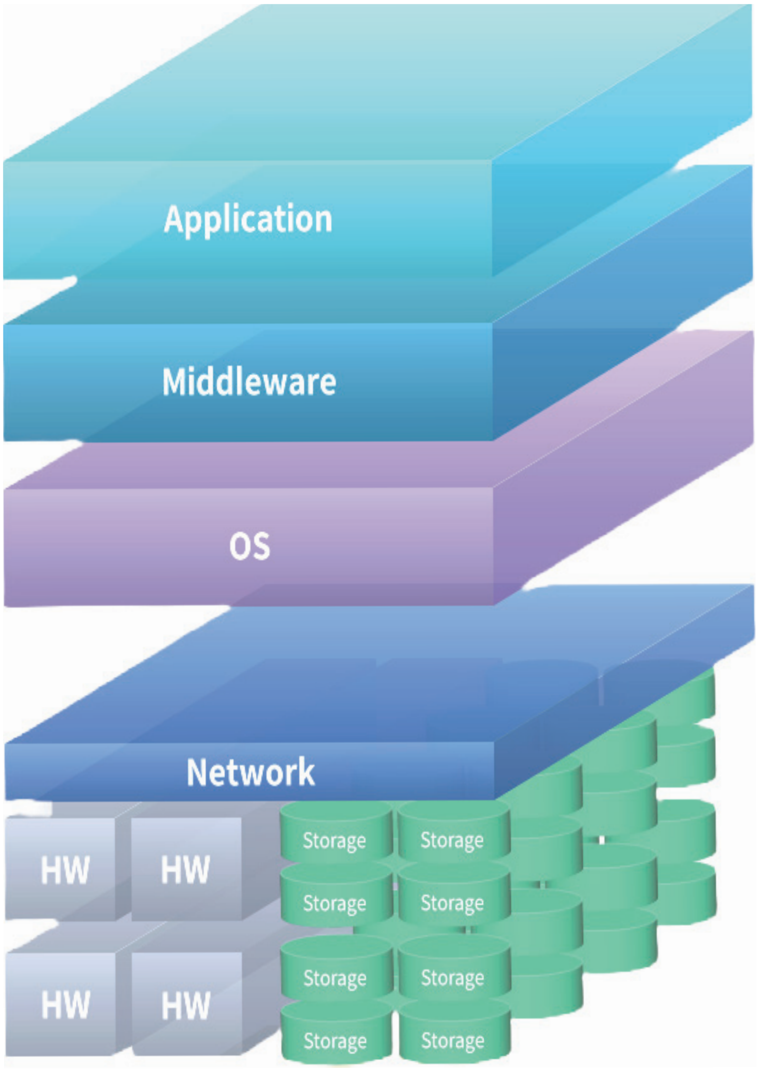
We specialize in HPC systems, providing consulting, implementation, management, and evaluation in a variety of environments.

Consulting

- Application-Optimized Cluster Configuration
- On-Premise Cluster Performance Optimization Consulting

System Intergration & Maintenance

- Application
  - GPGPU Programming Optimization
  - HPC, AI, Bioinformatics Application Optimization
- Middleware
  - Cluster Scheduler (Slurm, etc.)
  - Cluster Monitoring Tool
  - Kubernetes, OpenStack, etc.
- OS
  - OS Tuning
  - Authentication Software (LDAP, NIS) Support
- Network
  - High-Speed InfiniBand, Ethernet Network Setup
  - Fat-Tree Configuration, Redundancy Support
  - MPI, UCX Support and Optimization
- Storage
  - High-Speed Parallel Storage Support
  - Lustre, Ceph, HPE GLFS, HPE QUMULO Support



Evaluation

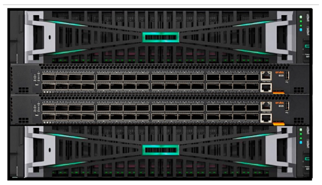
- HPL, MLPerf Benchmarking

Software Products

01. Storage : MCSxHPE

Hewlett Packard
Enterprise

1. GLFS(Green Lake File Storage)



1) Storage solutions for AI model training, serving, and cloud services

2) Configurable as all-flash without HDDs

3)Supports high-performance AI workloads with RDMA (GPU Direct Storage)

4) Comprehensive hardware and software solutions provided

- **Hardware :** Alletra Storage MP
- **Components:** Compute Enclosure, Switch, Storage Enclosure
- **Easily scale up or scale out as needed:**

Increase capacity by adding D-node,

Enhance computational performance by adding C-nodes

- **Software:**
  - Docker container-based solution
  - Optimized for various I/O performance needs

*Compute enclosure(C-node): Handles computation required for user-level services (API server, RDA (Remote Direct Access), cloud, etc.)
*Switch : Supports high-speed data transfer network between C-nodes and D-nodes (NVMe fabric, InfiniBand, etc.)
*Storage Enclosure(D-node) : Manages high-speed data transfer, storage, read/write operations

2. QUMULO

1) Simple and cost-effective NAS solution for large-scale AI processing

2) Supports various file sharing protocols (NFS, SMB, REST, etc.)

3) Configurable with a mix of HDDs and SSDs

4) Supports SSD cache functionality

02. Gadgetini: Cluster Unified Management Solution

- **dg Cluster Unified Management**
  - Provides dg Liquid Cooling System Info (temperature, humidity, flow rate, leakage, flow volume, etc.)
  - Provides System Info (CPU, GPU, memory, network, etc.)
  - Provides Storage Info (usage, bandwidth, etc.)
  - Supports Alerts via Slack, Gmail, etc.
- **Resource and Task Management**
  - Supports Kubernetes-based Container Resource Isolation and Virtualization
  - Supports Slurm-based GPU Task Management
- **Supports Workload Automation**
  - Batch job creation, scheduling, resource allocation, and virtualization
  - Provides MLOps/AIOps/DataOps functionality (additional solution)
- **Remote technical support**
- **Shared computing using cluster idle resources (additional solution)**
  - Maximizes computing resource utilization by renting registered idle resources to other workloads

Implementation Cases

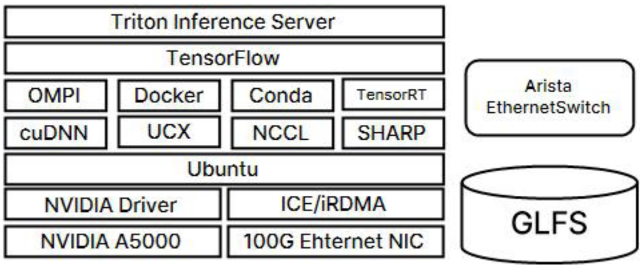
01. AI Infrastructure Implementation Cases

1) Large-Scale GPU Cluster Build

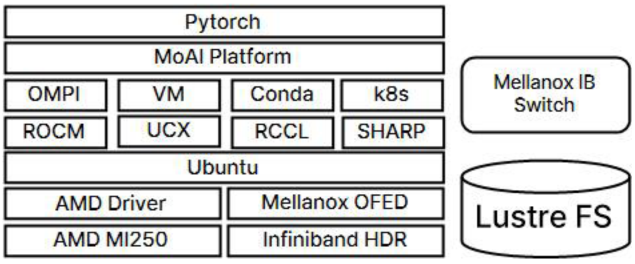
2) Installation of Over 3000 GPUs

3) Deployment and Operation of Over 500 Servers

4) Configuration of 64 IB Switches and Over 1690 IBs



AI Inference System with NVIDIA GPUs

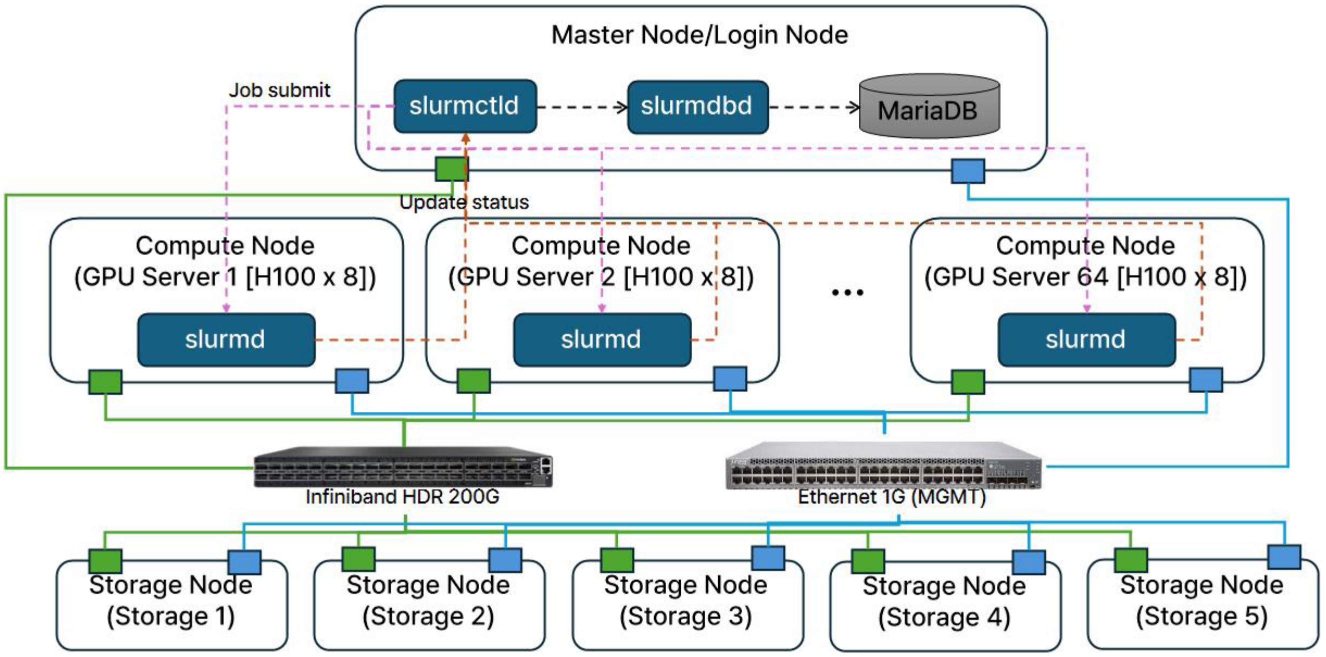


AI Training System with AMD GPUs

02. SLURM Deployment Cases

· Leading Heavy Industry Corporation in South Korea

· 'L' Systems Integration Company



Reliabe 3-Year Quality Management Service

With every purchase of ManyCoreSoft products, we provide industry-leading services including a 3-year quality guarantee, software recovery service, and HPC technical support.

3 years of Technical support service for HW repairs

3 years of SW recovery service

3 years of HPC technical support services

· **HW repairs**

We provide repair services for hardware failures occurring within 3 years of purchase. Our staff will visit to collect the server for repair, and once repaired, deliver it back to you.

Warranty covers:: Chassis, dg Liquid Cooling System, Power Supply, CPU, GPU, Memory, Infiniband NIC, Motherboard, Storage(excluding internal data).

· **SW recovery service**

In case of software failures, we offer a service to restore the software installed at the time of delivery. Our staff will visit to collect the server for restoration, and once the installation is complete, deliver it back to you.

· **HPC Technical Support**

Leverage our expertise in HPC technology to integrate GPU clusters and cloud technologies comprehensively, providing tailored HPC solutions and AI environments that meet your specific needs. Experience stable operations with ManyCoreSoft's technical support.

kubernetes PyTorch slurm workload manager TensorFlow NVIDIA CUDA AMD ROCm

*In the case of accidental damage, a deductible may be applied.

With 150+ Customers

kt
MOREH
KB국민은행
HYOSUNG TNS
ETRI
KSOE
한국조선해양
KCB
Next&Bio
IBK기업은행
KIMS 재료연구소
서울대학교병원
연세대학교 의료원
대웅제약
ILDONG
한국해양과학기술원

and more.

Partners & Contact

· **Headquarters & Research Institute**	138-308, 1, Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea (Post No. 08826)
· **Main Office**	1108, 11 Digital-ro 33-gil, Guro-gu, Seoul (Post No. 08380)
· **E-mail**	contact@manycoresoft.co.kr

deepgadget.com

Official Partners

WAVEFIVE
All about AI Computing

deep gadget