TEN

Company Introduction

Make AI accessible for all.

Company Name	TEN Inc.
CEO	Sejin Oh
Established	July 3, 2020
Address	No. 1203, Hyeonik Building, Taehaeran-ro, Yeoksam-dong, Gangnam-gu, Seoul, South Korea
Main Services	MLOps solution, ‘AI Pub’ AI infrastructure consulting service ‘RA:X’
Employees	22 employees
Website & Social Media	https://ten1010.io/ https://www.youtube.com/@ten1300 https://www.linkedin.com/company/ten1010/

Since launching our services in 2020, we have been selected for various projects and awards, recognized for our strong competitiveness and potentials.

2020

02 Started business by launching the AI Pub service
07 Established TEN Inc.

2021

03 Joined NVIDIA Inception Program
04 Received seed funding from Kingsley Ventures
05 Official partnership with NetApp
06 Acquired Venture Business Certification Selected for N&UP Program[1]
07 Selected for TIPS[2] Startup Growth and Technology Development Project
08 Established company research center
Received K-Startup Award in the "AI Service" category at the Korea Leading Company Awards
11 Attended NVIDIA GTC[3] selected as Top Startup from Korea
Selected as service provider for AI Voucher Project

2022

01 Received Award of Excellence on 2021 Global Scale Up IR Day of Global Business Partnership Program
03 Selected for AI Voucher Project
07 Received K-Startup Award in the "MLOps Platform" category at the Korea Leading Company Awards
08 Selected for Initial Startup Package of Korea Institute of Startup & Entrepreneurship Development
11 Certified as Company with Technological Excellence by KoData (T4 Level)
12 Received pre-series A round funding
(Ascendo Ventures, Quantum Ventures Korea, Korea Credit Guarantee Fund)

2023

01 Registered 5 patents related to resource management (KOR)
(10-2488614, 10-2488615, 10-2488618, 10-2488619, 10-2488620)
02 Hosted the First AI Workshop[4]
07 Signed NetApp Preferred Partnership
08 Signed IBM Partnership
09 Signed Redhat Partnership
11 Completed project on high-density GPU-based cluster computing system for deep learning computation for Korea Agency for Defense Development
12 AI Pub : GS Certification Grade 1 from TTA
'MOST Promising Korean Tech Company 2023' awarded by CIO Review

2024

01 Industry-Academic Cooperation Agreement on AI Infrastructure with Sungkyunkwan University SAL · COMPASS LAB
04 AI Alliance MOU with Kolon Benit
05 Awarded the 19th Digital Innovation Grand Prize for IT

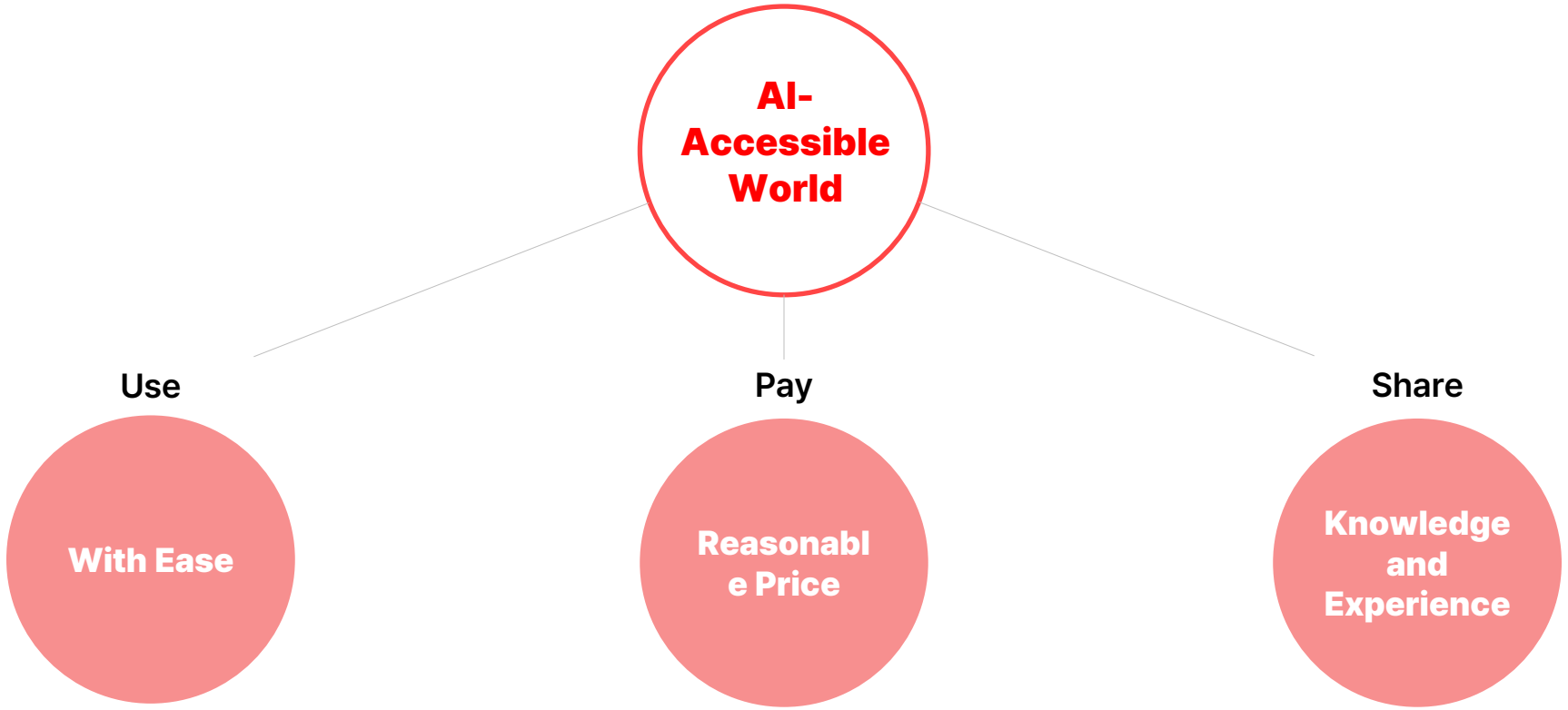
1) N&UP Program: Joint project between NVIDIA and Ministry of SMEs and Startups aimed at supporting businesses to enter the global market
2) TIPS: Tech Incubator Program for Startup
3) NVIDIA GTC: NVIDIA GPU Technology Conference

TEN is

committed to making AI more accessible[Mission] **for all**

by creating tools that lower the barrier to technology.[Vision]

In a world where AI is more accessible, everyone can create value and enjoy the benefits of using AI.
TEN offers affordable AI tools that are easy to use to make AI more accessible to all.
We envision a world where experience and knowledge related to AI is shared throughout the entire society.



Workshop (Feb 2023)
Workshop on building AI infrastructure based on NVIDIA GPU and MLOps software

Webinar 1 (April 2022)
Easier management of AI services and data on MLOps platform in Kubernetes environment

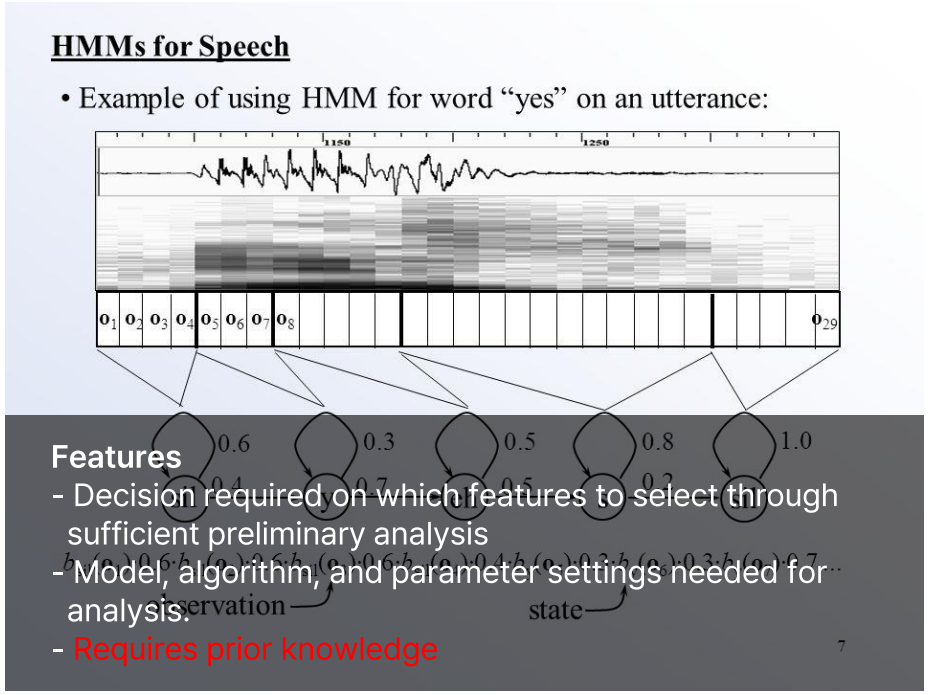
Webinar 2 (April 2022)
Building NVIDIA HyperScale AI infrastructure environment and Kubernetes-based AI development and operation environment

How are you preparing for AI in the era of deep learning?

We are living in the era of deep learning, where AI is at the center of attention of all members of society, from individuals to businesses. In the very near future, everyone will be able to convert their knowledge into data and automate it through training.

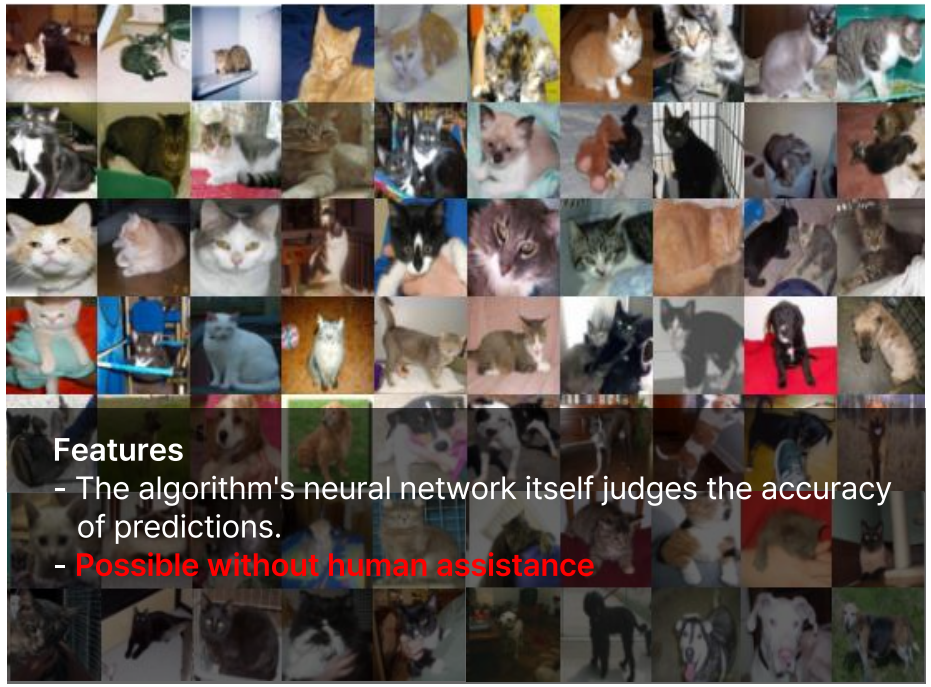
Unlike the times of statistical pattern recognition and physical phenomenon modeling, today, the ideas of the individual creating the AI have become ever more important.

Statistical Pattern Recognition



Features
- Decision required on which features to select through sufficient preliminary analysis
- Model, algorithm, and parameter settings needed for analysis.
- Requires prior knowledge

Deep Learning

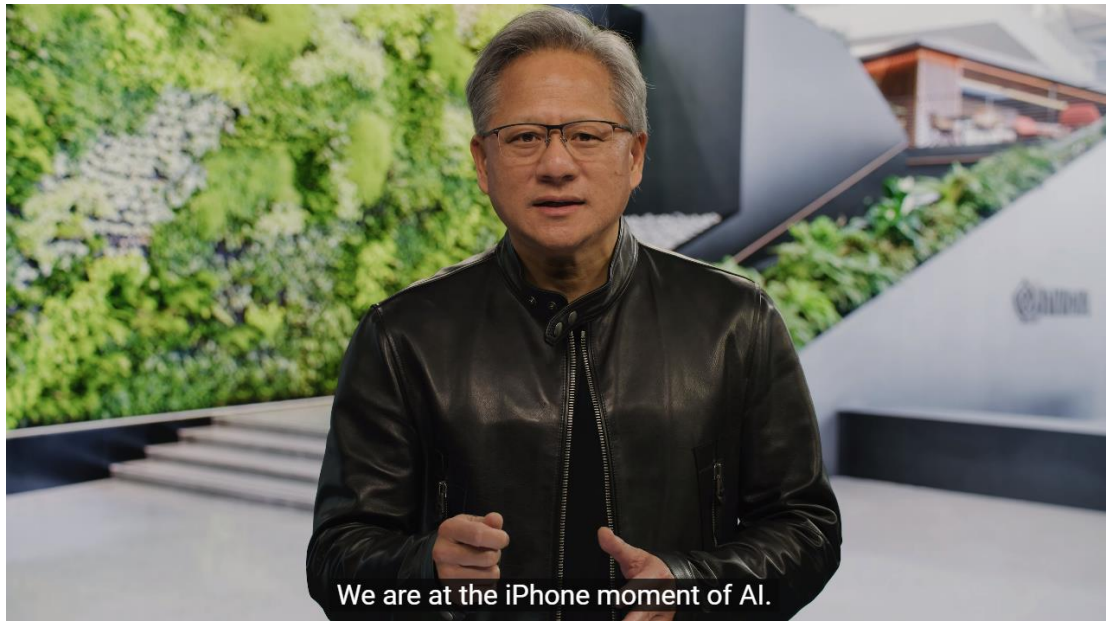


Features
- The algorithm's neural network itself judges the accuracy of predictions.
- Possible without human assistance

Building a Cat Detector using Convolutional Neural Networks — TensorFlow for Hackers (Part III) | by Venelin Valkov | Medium

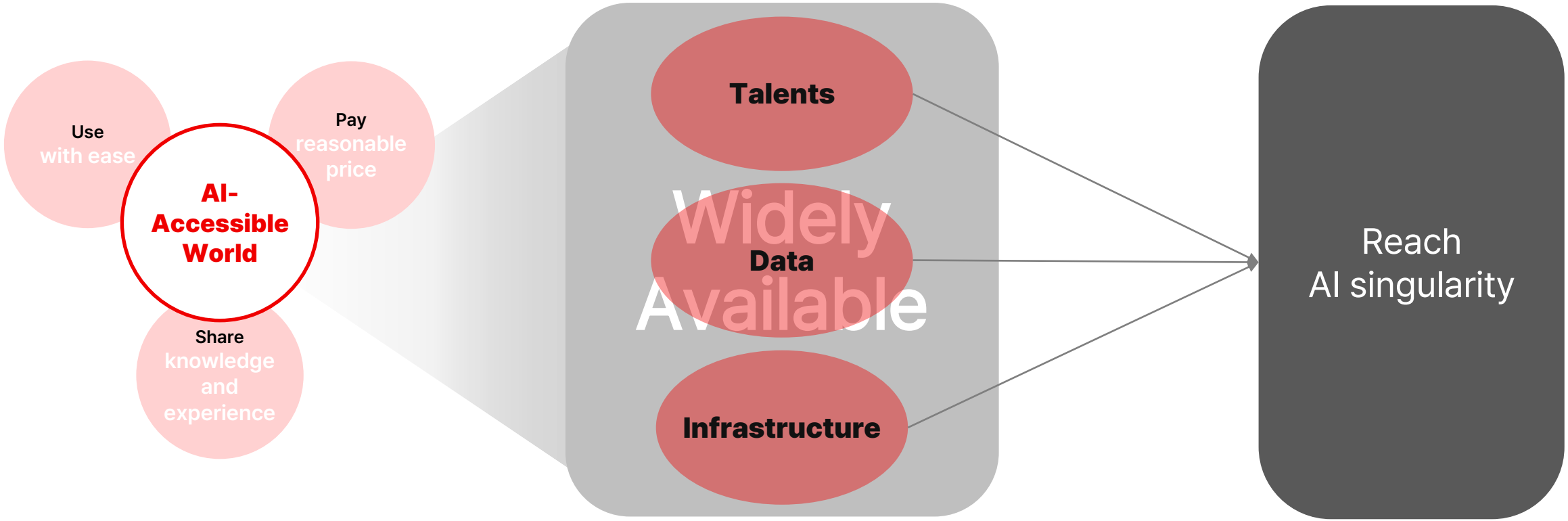
“We are at the iPhone moment of AI.” - Jensen Huang

Startups are competing to create innovative products and business models, while existing companies explore ways to respond.



GTC 2023, Jensen Huang (NVIDIA CEO)

We are currently in the process of reaching AI singularity.
More than ever before, these times require talents, data, and infrastructure to make AI more widely available.



We are gradually achieving the conditions required to enable the general public, rather than just certain groups of experts, to build their own AI and train them using their knowledge converted into data.

However, the issue of cost still remains when it comes to infrastructure, and it cannot be solved with time and effort alone.

An era where you can create AI without a PhD in machine learning

Anyone can convert their knowledge into data and train AI

Infrastructure = Cost

Since the emergence of deep learning, the performance of AI has been predominantly determined by computing power. Unlike traditional statistical pattern recognition, computing power has a significant impact on the performance for deep learning.

To enhance performance, significant investment is required.
That is why resource efficiency will become increasingly important going forward.

Statistical Pattern Recognition

Performance ∝ Computing power

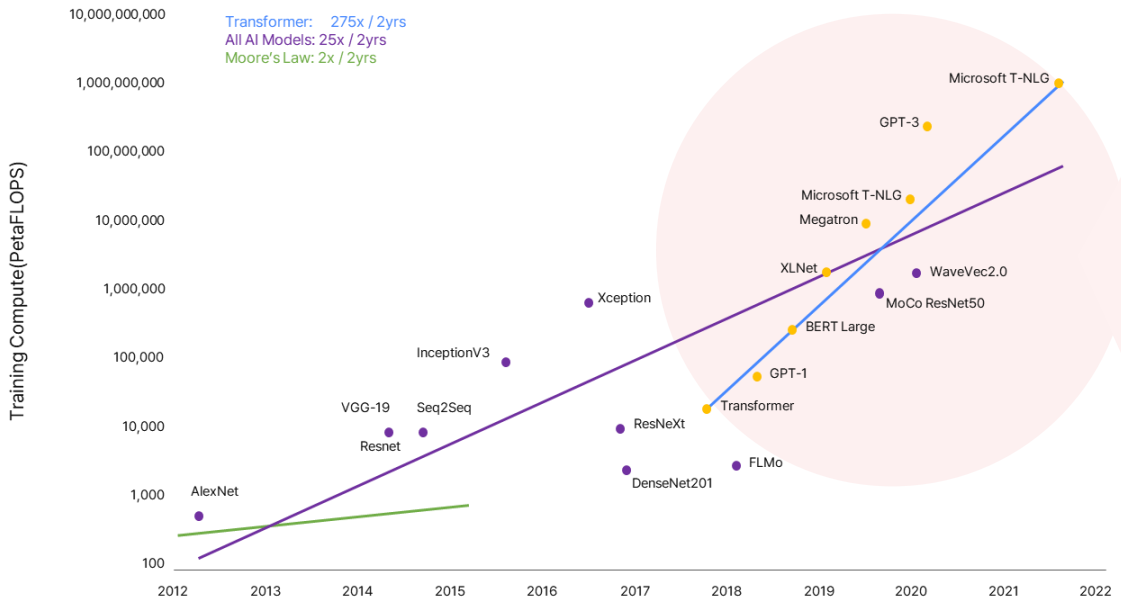
Weak proportional relationship

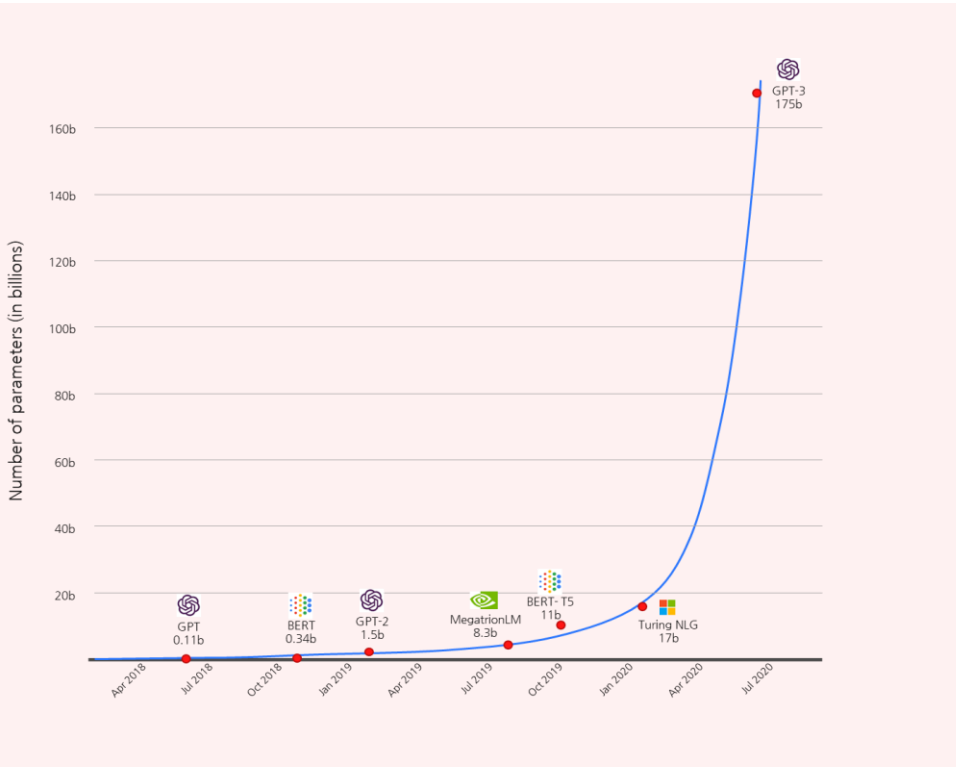
Deep Learning

Strong proportional relationship

Companies of all sizes and industries are striving to build high-performance server infrastructure.
The linear scale clearly shows the exponential increase in the required computing resource.
This is a clear indication of the imminent AI singularity and the importance of infrastructure.

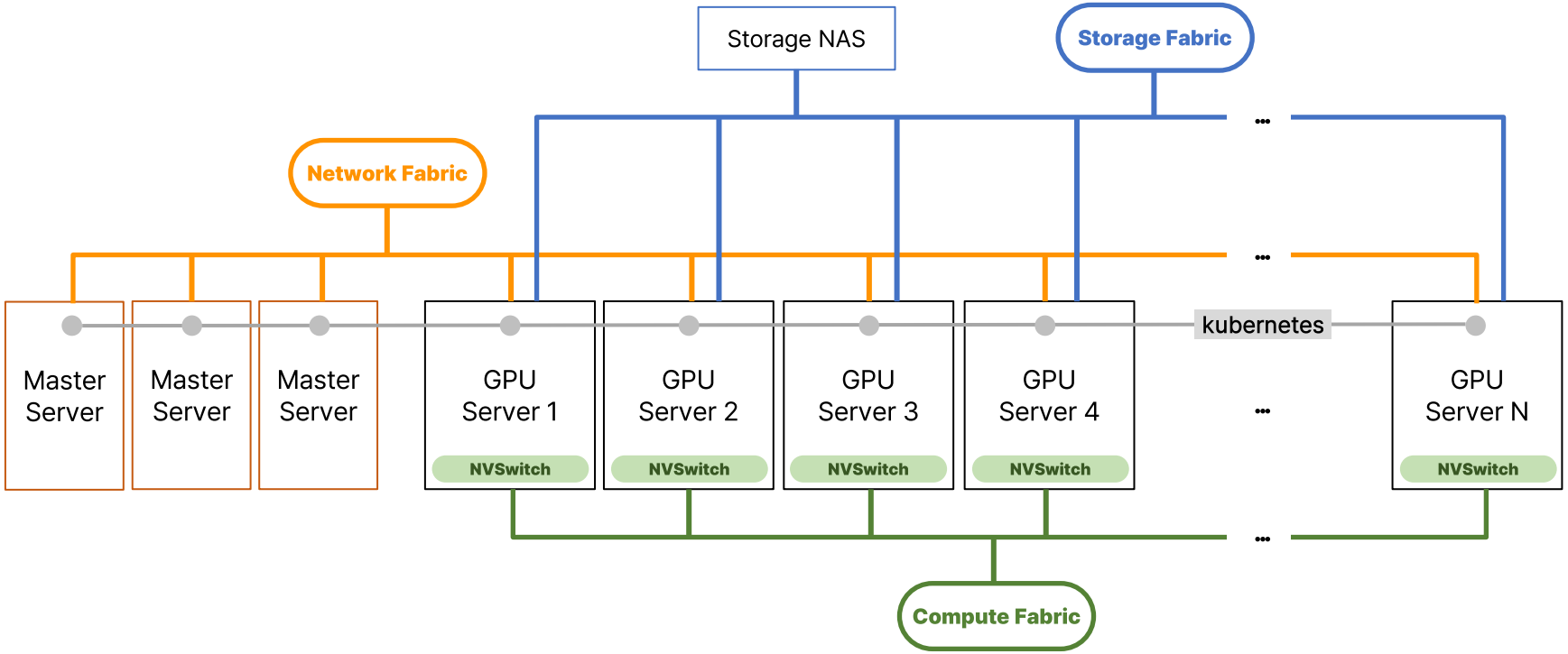
[NVIDIA GTC Keynote, Nov 2021; Log scale]





However, improving the efficiency of infrastructure resources is a challenging task, as the continuous increase in resources required for AI models makes it more complicated to build an AI infrastructure.

With more GPU machines required to cover the increased computing demand of AI models, building an AI infrastructure has become more complex and challenging.



High efficiency in infra utilization

: the key to making AI more widely available.

TEN offers efficient solutions.

TEN offers MLOps software and dedicated AI infrastructure that facilitate the development and operation of AI.

Tools

MLOps

Infrastructure

Server, GPU,
Network, Storage

Holistically served by TEN

Knowledge

Experts in the field
and data

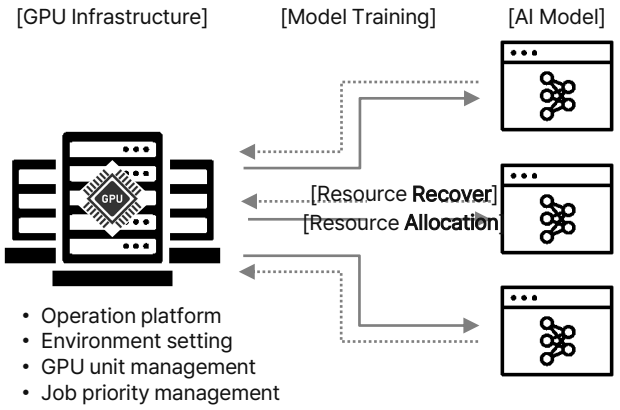
TEN provides solutions for each infrastructure management point to address issues that arise during AI model training and operation.

Problems in AI Training Process

- Operation platform optimized to the model training characteristics (resource-hungry) is needed when building on-premise infrastructure.
- As resource is allocated in server units with the conventional platform, it is not possible to know the idle and operation rates at the GPU unit level.
- Overhead is incurred in the resource allocation process to change the server settings (OS, drivers, libraries, etc.) for each developer.
- Job scheduling is required for each model training.

Solution for AI model training

Maximized GPU resource operation rate 24/7

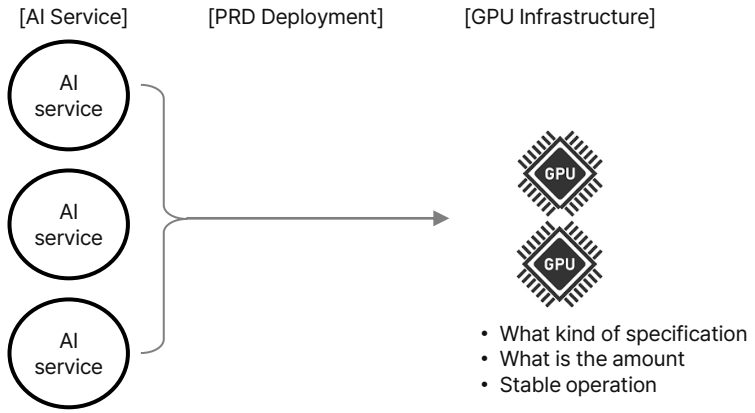


Problems in AI Operations Process

- GPU resource specification and amount required for service operation cannot be estimated for model deployment.
- Stability cannot be secured when operating multiple services due to the interference between services within the GPU resource.
- Cost goes up from the continuous increase in the amount of resources used by model.

Solution for AI model operation

Stable operation quality with minimum GPU usage



TEN offers COASTER, a container platform, and AI Pub Dev and AI Pub Ops, which are MLOps tools and also RA:X, an AI Infrastructure consulting service.

Container platform

COASTER

GUI-based MLOps service for AI development and operations

AI Pub Dev

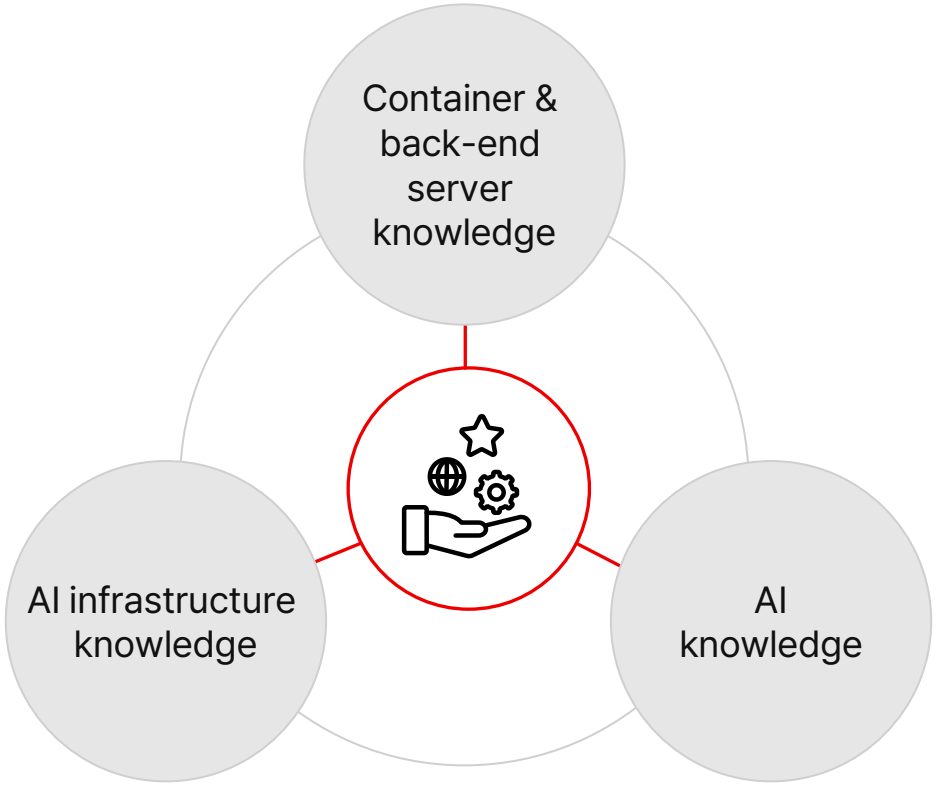
AI Pub Ops

For AI Infrastructure

RA:X

TEN's Products

AI training and operation require numerous skilled experts across various fields. TEN developed function-level products that support companies to establish their own governance structures and offers MLOps tools to operate AI services without all-rounder experts.



COASTER

for

Kubernetes-proficient developers and MLOps developers

Command Line Interface

AIPub

for

AI modelers or AI service operators who are not familiar with Kubernetes and back-end development

Graphic User Interface

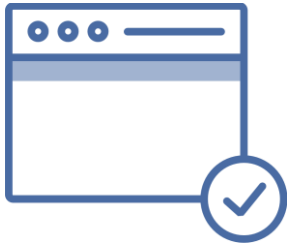
TEN's COASTER and AI Pub provide optimum solutions that encompass everything from AI service to the building and maintenance of infrastructure.

1 Day

Shortened time to reach AI distribution, which used to take 8.6 months

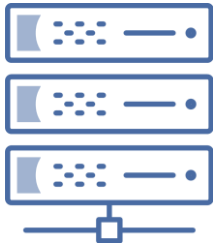
100% util & 1/10 cost

Infrastructure is used at the highest operation rate during the AI training phase, then brought to the lowest rate during the AI operation phase.



Off-the-shelf

Ready-to-use platform without any complicated development process that requires skilled experts from various fields



Tailored

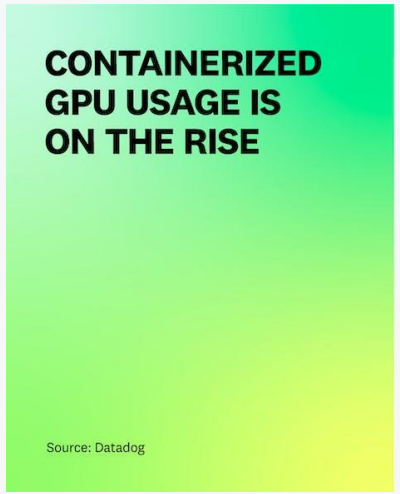
Build AI-tailored infrastructure that maximizes the performance of costly GPU resources

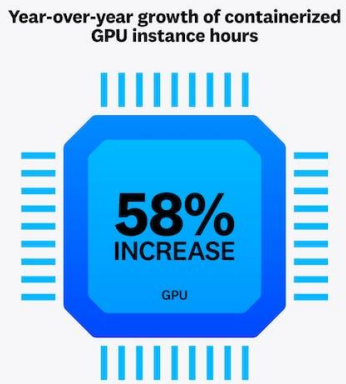
COASTER and AI Pub use Kubernetes, which is the most popular platform for container orchestration.

10 INSIGHTS ON
REAL-WORLD CONTAINER USE

FACT 2

Usage of GPU-based compute on Containerized workloads has increased





> We observed a 58 percent year-over-year increase in the compute time used by containerized GPU-based instances (compared to a 25 percent increase in non-containerized GPU-based compute time over the same time period).

1) The Cloud Native Survey, 2022 (subject: Data usage trends) https://www.datadoghq.com/container-report/

TEN listened to the needs of our clients during AI development and developed the optimum solution with AI Pub.

"We purchased servers in bulk, but we are **struggling with their operation due to the significant management overhead** required to allocate them to different teams."

"Teams purchased their own servers. Now we have **different types of GPU servers** that are **difficult to manage together** and the usage rate is very low."

"The engineering department purchased a GPU server for research work, but we do not have **an effective solution to share it between multiple labs**."

"**Multiple users are sharing the server**, but we are unsure of how to operate spaces like storage or image registry to manage the data of each user."

"We purchased A100 GPU recently, but we are finding it **difficult to configure the MIG function every time** and share it with our developers."

"We manage our development products using docker images. It is **difficult to manage the images that are used by each user** and the images that have to be shared with the team by the administrator."

TEN listened to the needs of our clients during AI operation and developed the optimum solution with AI Pub.

"We developed a speech synthesizer system, but we lack **the experience to launch the service on a GPU server**."

"We developed an AI service for identifying product defects. Is there software that can **operate and manage it automatically**?"

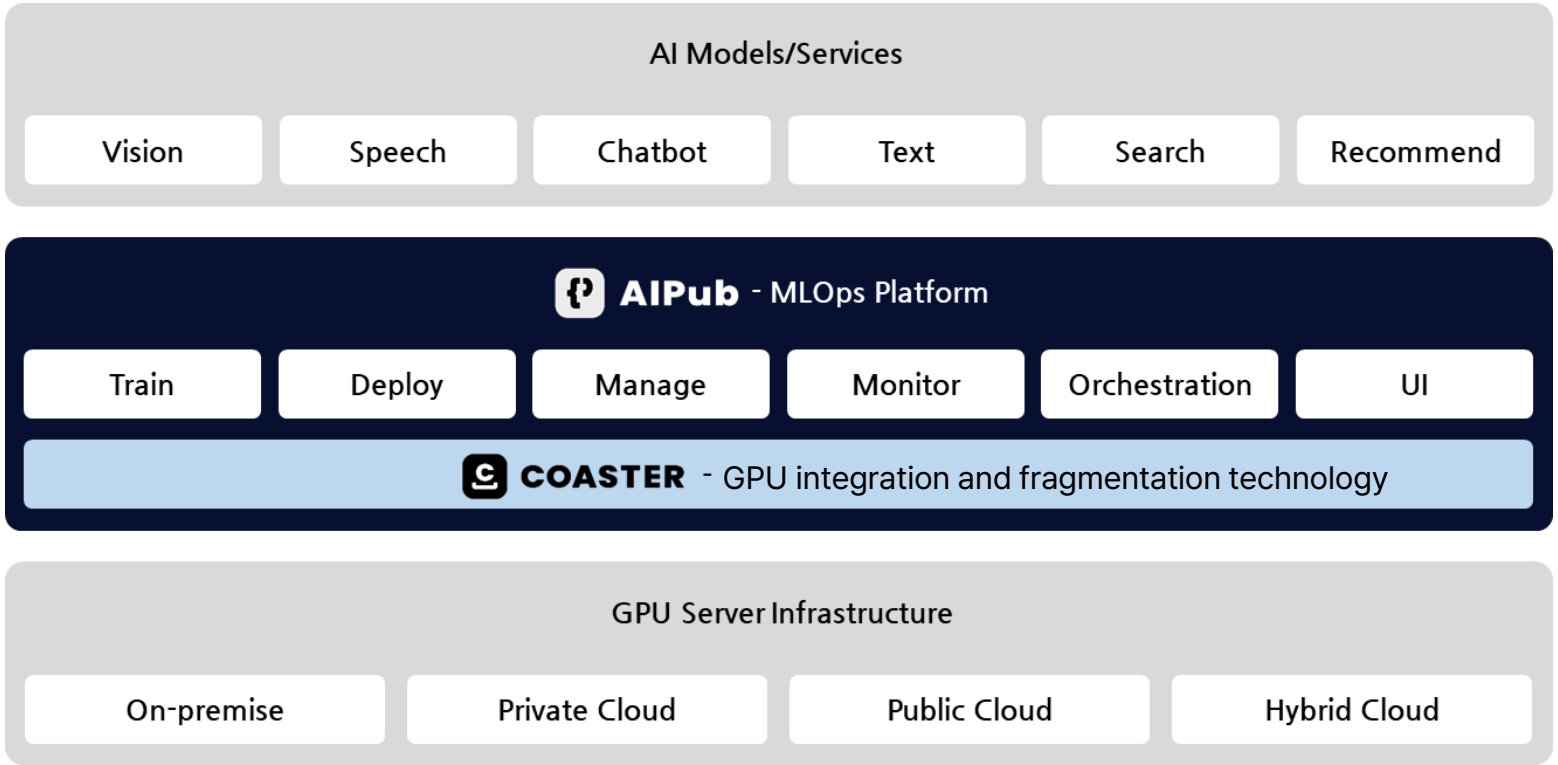
"We developed an AI service, but we are struggling with stable operation due to a **lack of operators and knowhow to operate the GPU server**."

"We do not have a solution to manage the various types of services and **different docker images for each service version**."

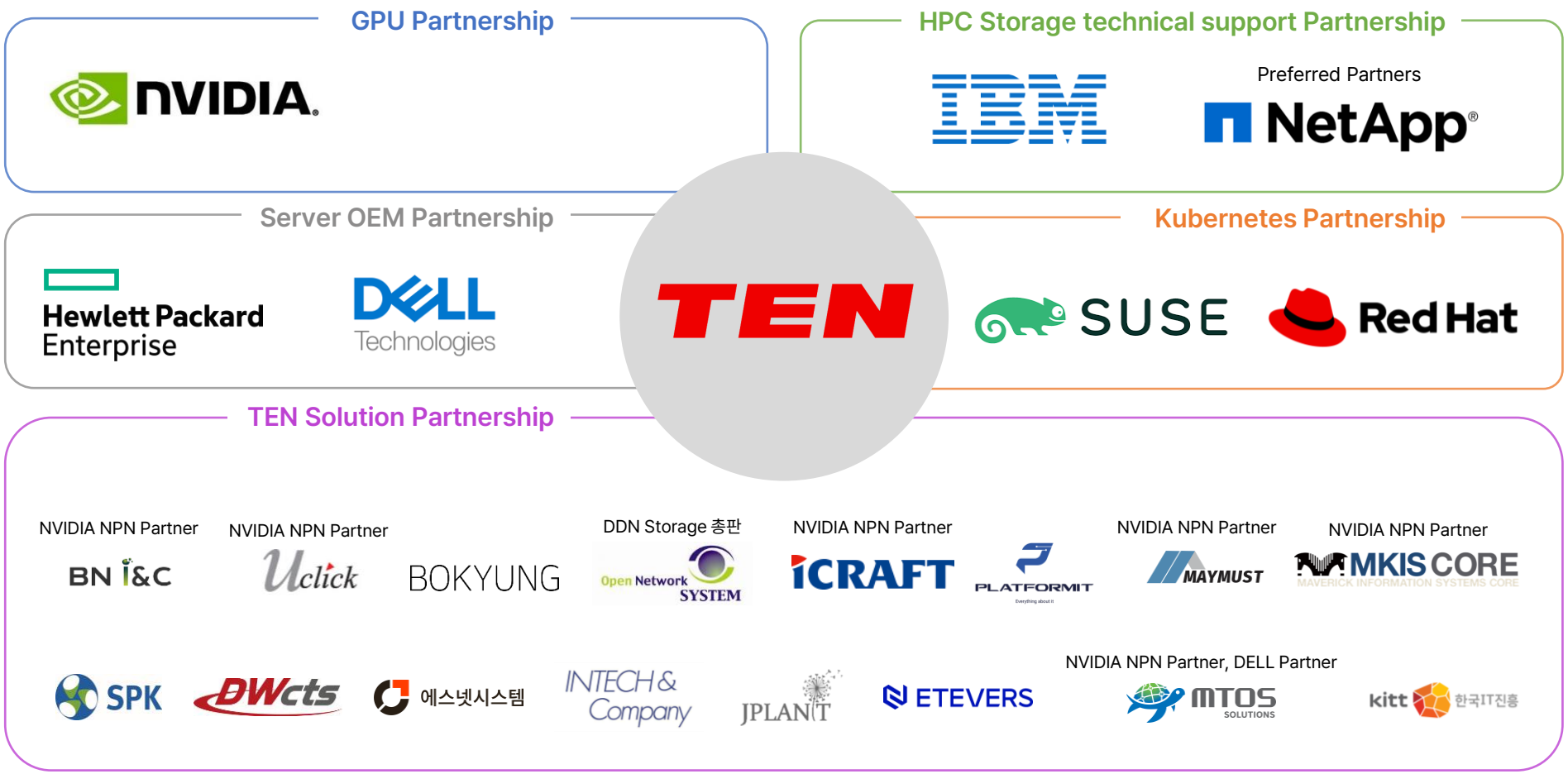
"The server resource has to be **divided and managed by each service operator or team**, but we do not know how."

"We need to **monitor the traffic of each service** and check and manage response failures. Is there an effective way to do this?"

AI Pub is a fully-managed solution that focuses on addressing issues related to the operation and infrastructure utilization throughout the AI lifecycle.
It is designed to provide service to any type of AI model in various infrastructure environments.



We offer all-in-one AI-specific services in partnership with vendors specialized in AI.



Enterprises and universities in various areas are using TEN's services.

[Client] Enterprise

국방과학연구소 AGENCY FOR DEFENSE DEVELOPMENT

세종특별자치시

LG전자

KB국민카드

한화시스템

maum.ai

AI Inventech

basgenbio

JPLANT

EJN

[Client] Academia

연세대학교 YONSEI UNIVERSITY

서강대학교 SOGANG UNIVERSITY

한양대학교 HANYANG UNIVERSITY

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Our aim was to create an AI deployment tool that can be widely used.
We envisioned a friendly and welcoming “pub-like” environment,
where individuals can come together to develop and operate AI.

AI Pub

AI Public

AI Publish

COASTER

Container platform with functions improved by extending Kubernetes

TEN

COASTER is a container platform that extends Kubernetes and enhances the operation of GPU infrastructure and user management functions.

COASTER

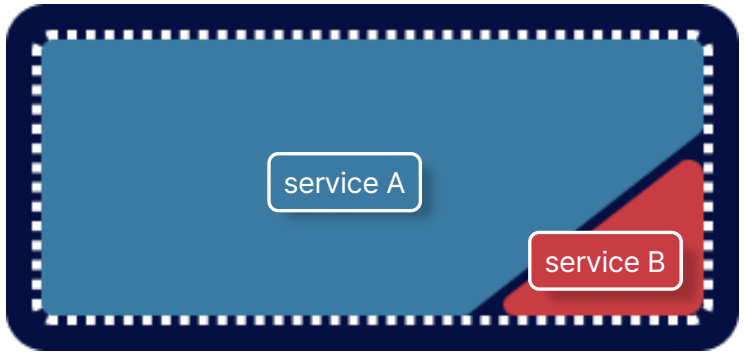
Main Services	Service Description
GPU fragmentation	Fragment the utilization and memory of one GPU unit into 100 blocks
Inquire and allocate GPU resource	Inquire the computing resource in the entire cluster using extended Kubernetes commands
User entitlement management – by individuals and groups	Assign policies to users and groups
Job scheduling and priority management	Kubernetes-based job scheduler automatically initiates jobs, while al so allowing a human operator to manually re-prioritize them through GUI.

FUNCTION of COASTER 1.

COASTER supports the usage of fragmented GPUs.

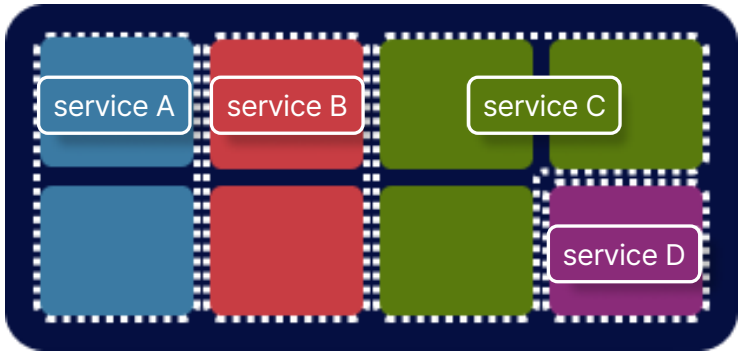
Allocating GPUs to containers in block units not only allows multiple containers to run on a single GPU
but also ensures stability by preventing the interference of resource usage between containers.

Kubernetes Native : GPU Allocation



GPUs can only be allocated to containers as a whole unit, and
multiple containers cannot run on a single GPU.

Coaster Extended : GPU Allocation



The utilization and memory of a single unit of GPU
can be fragmented into 100 blocks, each spanning 1% increments.

FUNCTION of COASTER 2.

COASTER enables the inquiring and allocation of GPU resources.

With the extended Kubernetes commands, you can inquire the computing resource of the entire cluster and allocate the required number of blocks for the appropriate type of GPU to each container. This user experience is different from the basic Kubernetes environment, where you can only inquire resource in server units and are unaware of the GPU type of each server.

Kubernetes Native : View GPU resource

```
Capacity:
  attachable-volumes-aws-ebs:  39
  cpu:                         4
  ephemeral-storage:           83873772Ki
  hugepages-1Gi:               0
  hugepages-2Mi:               0
  memory:                      16093900Ki
  nvidia.com/gpu:              1
```

Can only view the resource status of accessed node

Coaster Extended : View GPU resource

```
[root@aws-master-01 ~]# kubectl get cr -A
NAME             RESOURCE_NAME                TOTAL          FREE
block-tesla-t4   ten1010.io/block-tesla-t4    400            130
cpu              cpu                          160000m        92000m
gpu-tesla-t4     ten1010.io/gpu-tesla-t4      4              2
memory           memory                       1600017238Ki   1283834Ki
[root@aws-master-01 ~]#
```

Can view the computing resource of the entire cluster
using extended commands

Kubernetes Native : Allocate GPU resource

```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
spec:
  containers:
    - name: cuda-container
      image: nvcr.io/nvidia/cuda:9.0-devel
      resources:
        limits:
          nvidia.com/gpu: 2 # requesting 2 GPUs
```

Allocate entire GPU

Coaster Extended : Allocate GPU resource

```
apiVersion: v1
kind: Pod
metadata:
  name: block-pod
spec:
  containers:
    - name: tensorflow
      image: tensorflow/tensorflow
      resources:
        requests:
          cpu: 2000m
          memory: 4Gi
        limits:
          ten1010.io/block-tesla-t4: 70
```

Allocate divided GPU

FUNCTION of COASTER 3.

COASTER provides user entitlement management on individual and group level.

Users who share a policy can be placed in the same group, enabling collective management of their entitlement to access resources, such as namespace, shared storage, image registry and server node.

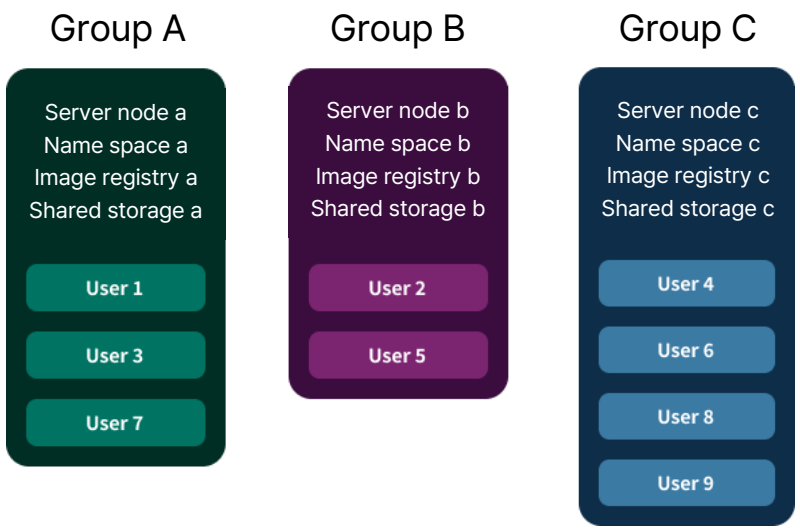
Kubernetes Native : user entitlement management

The access to resources of each user is managed separately, making the management process complicated and difficult.

Coaster Extended : user entitlement management

Users who share a policy are in the same group, and their entitlement to access resources, such as namespace, shared storage, image registry and server node, can be managed together.

Coaster's open source project : Resource-group-controller



Example for user entitlement management of COASTER

Group A :
User entitlement Only for User 1, 2, 3
Server node a / Name space a / Image registry a / Shared storage a

Group B :
User entitlement Only for User 2, 5
Server node b / Name space b / Image registry b / Shared storage b

Group C :
User entitlement Only for User 4, 6, 8, 9
Server node c / Name space c / Image registry c / Shared storage c

FUNCTION of COASTER 4.

COASTER's scheduler makes it easy to re-prioritize jobs in the queue.

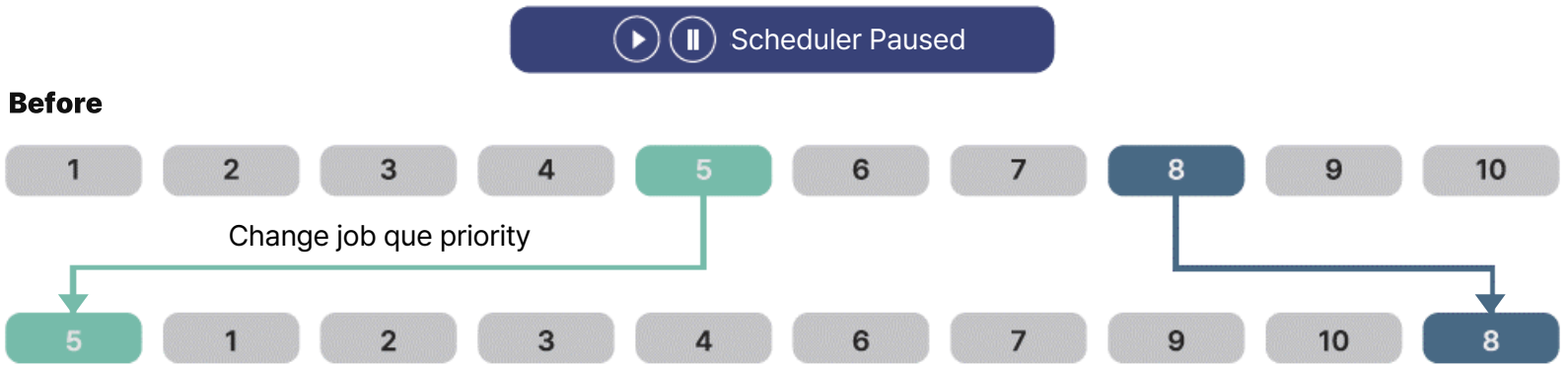
Kubernetes Native : scheduler

Basic scheduler of Kubernetes manages the queue using the FIFO (First In First Out) method.

Coaster Extended : scheduler

It is easy to change the priority of the jobs in the queue.

Example for scheduler of COASTER



After : The 5th and 8th jobs in the queue before the change have been respectively moved to the 1st (formerly 5th) and 10th (formerly 8th) positions in the updated queue.

See a demonstration of COASTER's functions in this video.

AIPub Dev

With COASTER at its core,
fully-managed service that supports AI development

With COASTER at its core, AI Pub Dev supports AI development. It provides a fully-managed service for model training, resource and workload management, etc.

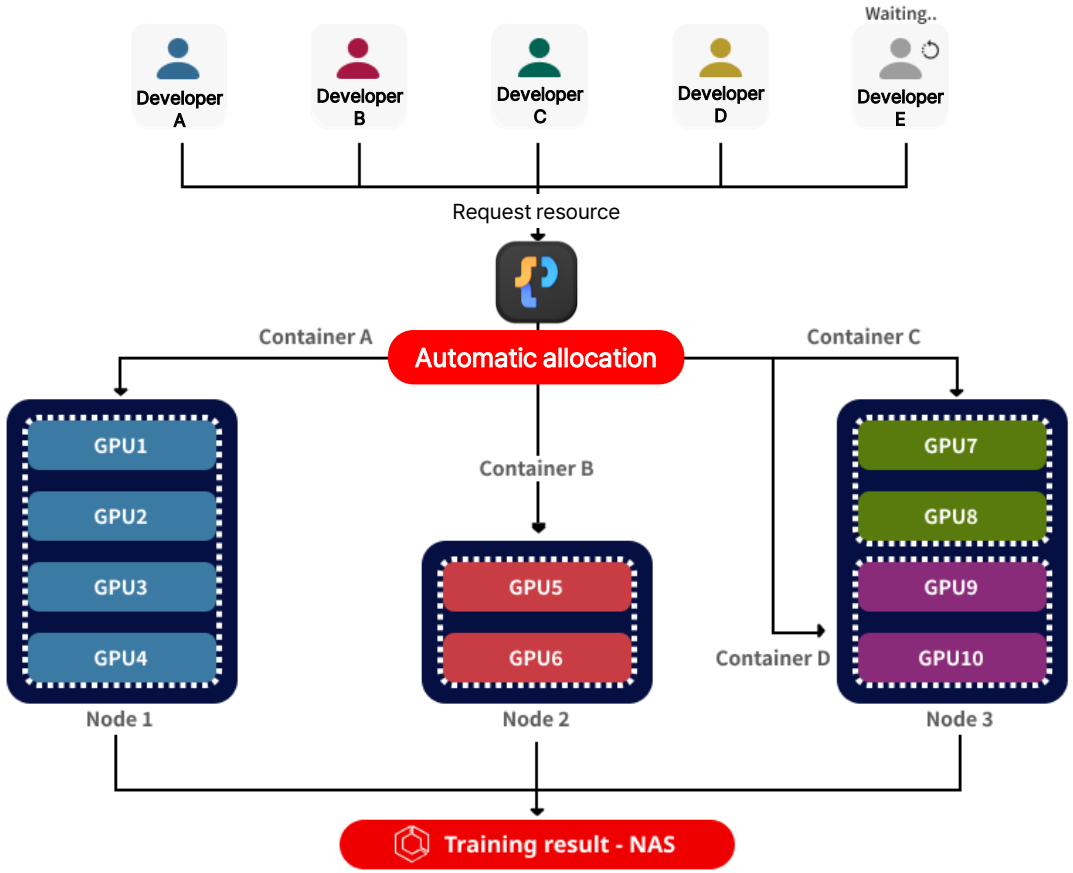
AI Pub Dev

Main services	Service description
Create workspace	Manage the user's development environment in docker image
	Create workspace based on development images
	Connect with Jupyter Notebook and TensorBoard
Model training	Automatically allocate resource required for each AI training
	Can apply for GPU resource and CPU resource
Resource management	Limit resource usage of each user account
	Withdraw idle resource
	Manage workspace of each node & Set MIG for each node
	Monitor entire infrastructure
Resource group management	(for manager) Create resource group and Set user's authorization
	Edit resource group
Workload management	Stop/resume scheduler
	Job scheduling and prioritization
Usage history management	Manage resource usage history of each user account
	Download usage history

FUNCTION of AI Pub Dev 1.

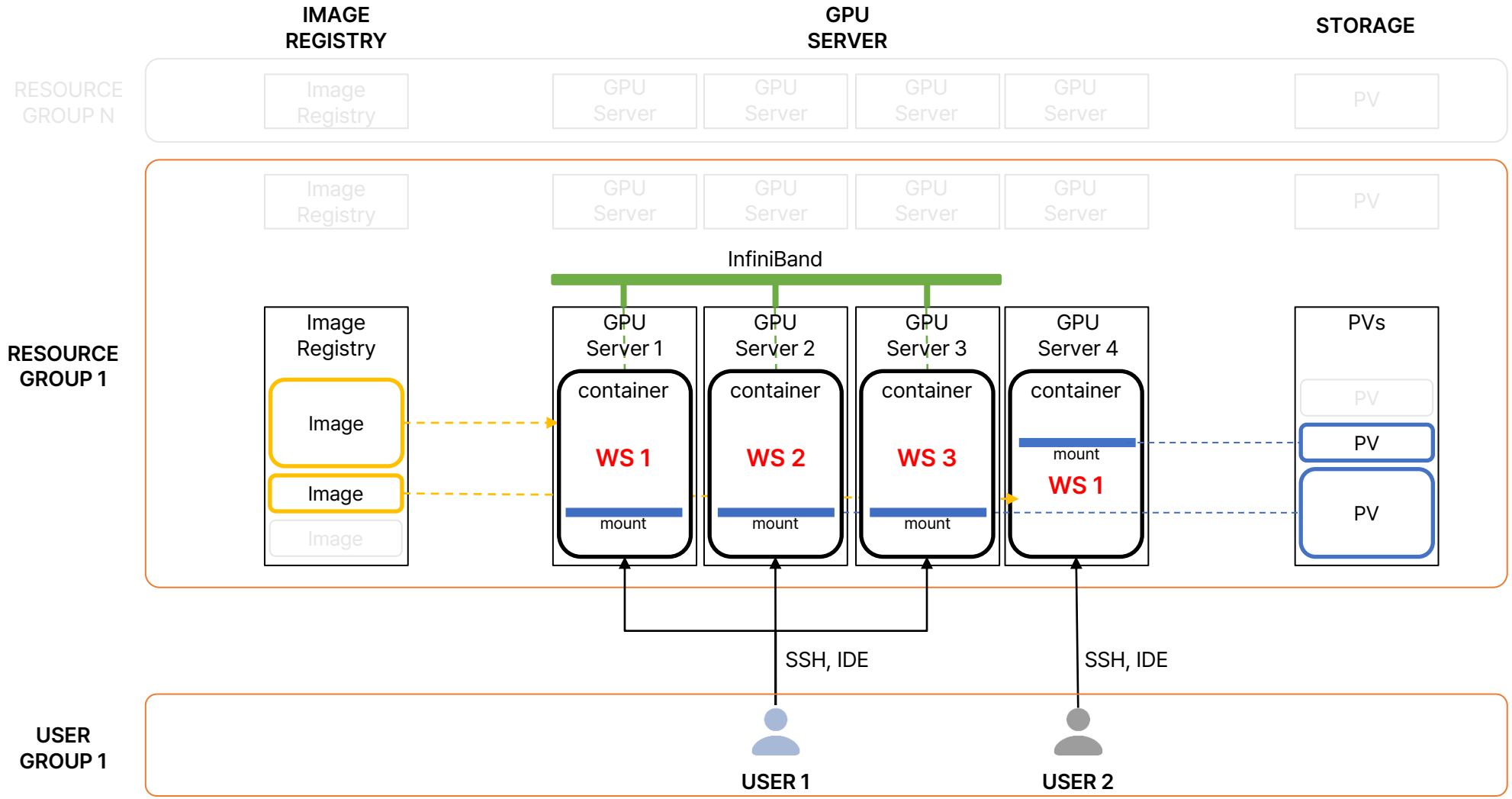
AI Pub Dev can allocate AI infrastructure to a team or an individual user.
It can also measure the amount of resource usage of each team or developer.

With AI Pub Dev, you can build an AI development infrastructure to allocate and manage resource during centralized management.
You can also measure the resource usage of each user account.



FUNCTION of AI Pub Dev 2.

AI Pub Dev allows users to create workspaces based on resources accessible within the affiliated group.



See a demonstration of AI Pub Dev's functions in this video.

AIPub Ops

With COASTER at its core,
fully-managed service that supports AI operation

AI Pub Ops offers technologies required for AI operation.
Built on COASTER with Kubernetes, it provides service mesh and application functions.

Application	Service that enables a stable operation of trained AI models
Service Mesh	Provides service discovery, load balancing, timeout and retries, and allows administrators to manage security and monitor performance of the cluster.
Kubernetes	Inquire, allocate and manage all cluster resources based on fragmented GPU units
Coaster	Allocate fragmented GPUs to containers and improve efficiency of required resource usage per service

AI Pub Ops supports AI operation with COASTER at its core. It provides a fully-managed service for AI service and resource.

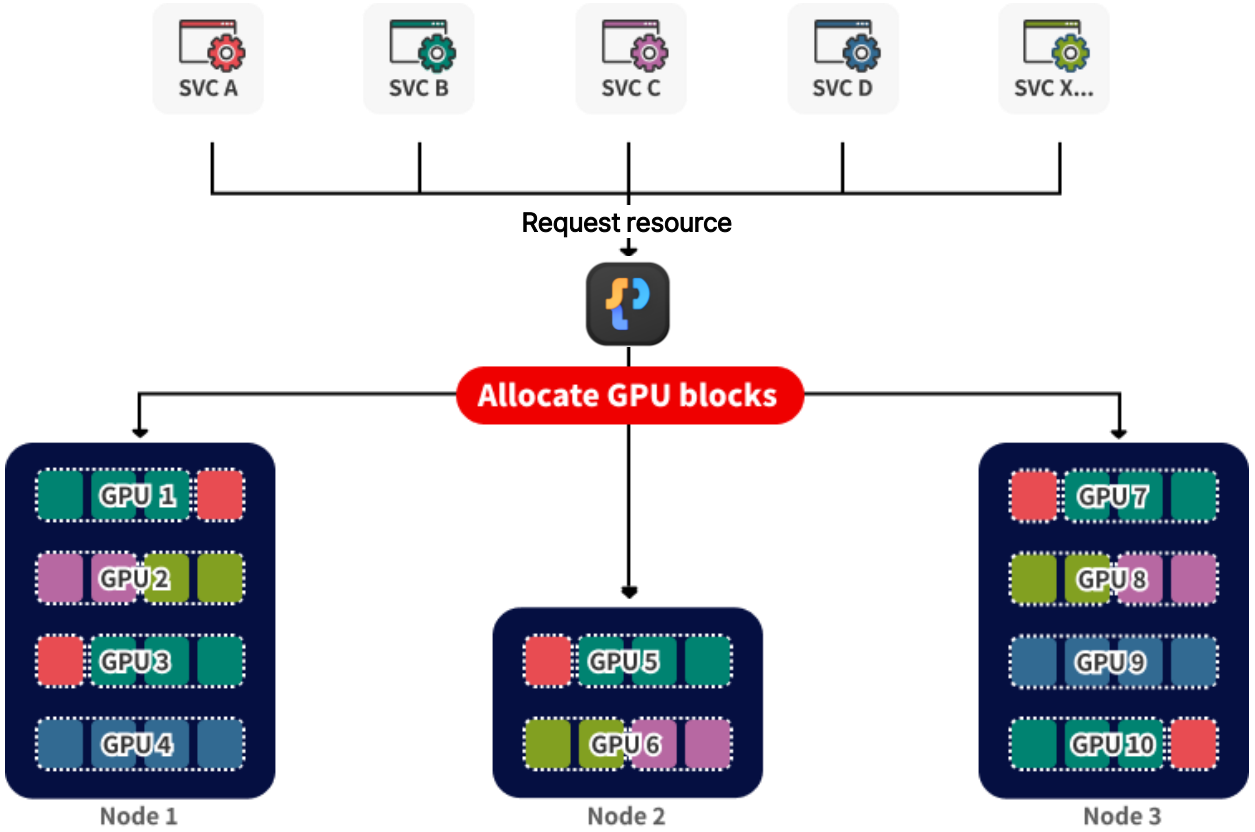
AI Pub Ops

Main services	Service description
Service creation and update	Provides UI for service creation, suspension, deletion and distribution
	Provides UI for non-stop service update
	Version management and service rollback
Service monitoring	Monitor operation status using service list and details
	Service error alert and troubleshooting by checking logs
Resource group management	Enables administrator to create resource group and set user entitlement
Resource management	Enables allocation of GPU blocks for each service
	Monitor real-time operation rate of GPU blocks and servers

FUNCTION of AI Pub Ops 1.

AI Pub Ops manages the user's services using docker images It reduces cost by fragmenting the GPU into the smallest size of 100 units.

AI Pub Ops enables non-developers to create, suspend, delete, update and rollback services.



FUNCTION of AI Pub Ops 2.

AI Pub Ops also supports AI service operation on on-premise servers and offers the same service operation and management functions that are provided in the public cloud environment.

High Availability

Eliminate single point of failure
through redundancy

Rolling Update

Update any time
without service suspension

Load Balancing

Automatic deployment by requesting
serivce to less busy servers

Scale-out

Process traffic by increasing the
number of servers automatically
according to service request

Failover Response

Detect service failover and run new
service to secure stability

Abnormal Event Alert

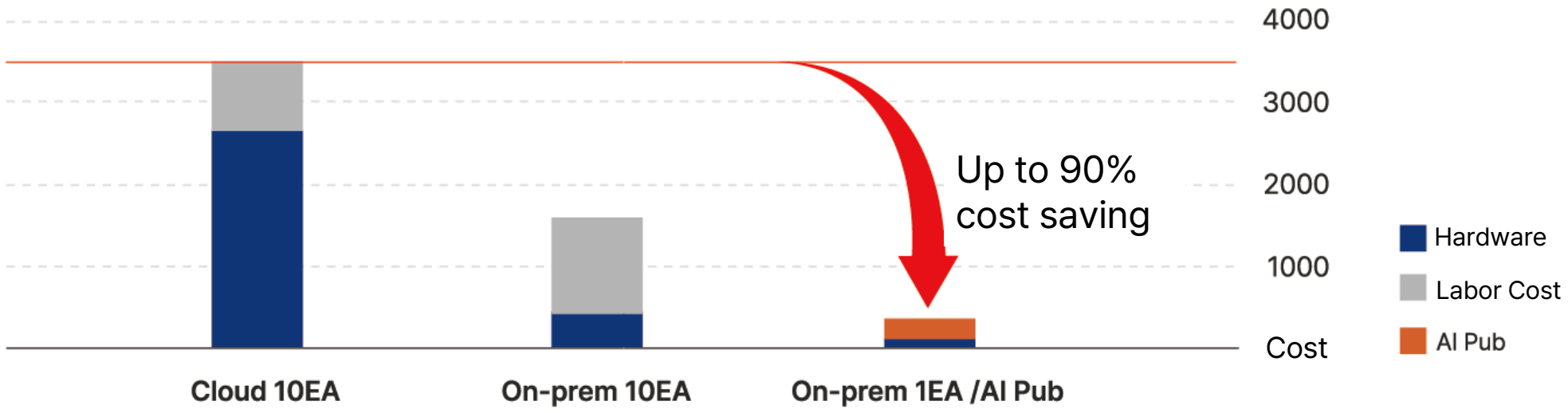
Real-time alert
through KakaoTalk or Slack
when a major operation event occurs

AI Pub Ops can reduce your AI operation cost by up to 90%.

Instead of operating your AI service on a public cloud, you can reduce cost by building a server of the same size and using AI Pub.
Your service can be operated with only 10% of the server resource by using our GPU fragmentation function, and you can also save cost by eliminating the manpower required to develop, maintain or repair functions for service operation.

[Infrastructure Cost for Five-Year Operation of 50 AI Services of Client Company]

(Actual example: 10 servers with 4 units of T4 managed by 4 mid-level developers
→ 1 server with 4 units of T4 managed by AI Pub and no managing personnel)



TEN also offers consulting services to address the difficulties in AI operation that may not be fully covered by AI Pub Ops.

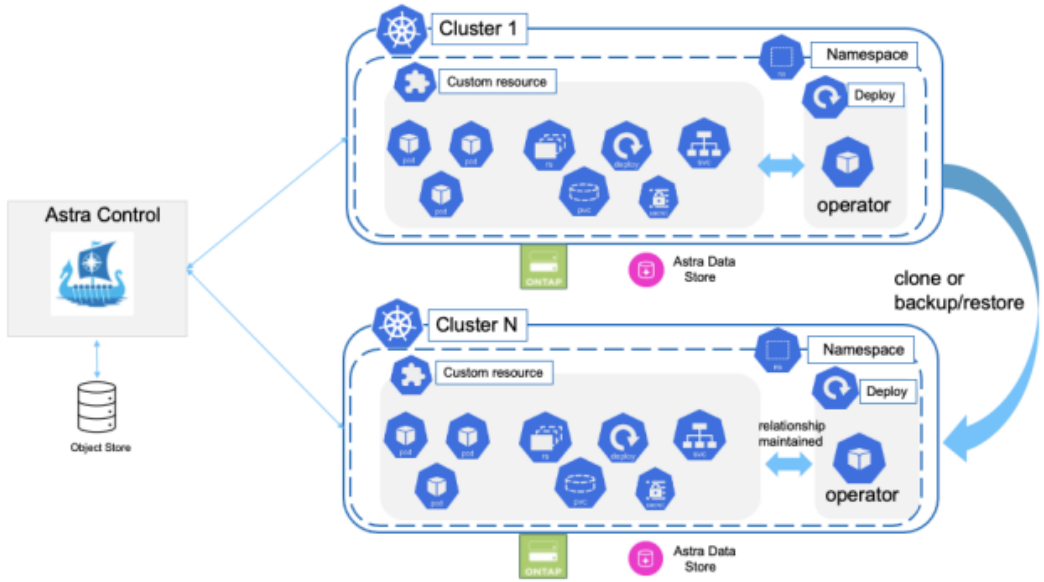
We operate your service for you using our in-depth AI operation knowhow.
With our services, you can focus solely on creating more AI values without spending extra time and money to maintain continuous service stability.

AI Pub Ops

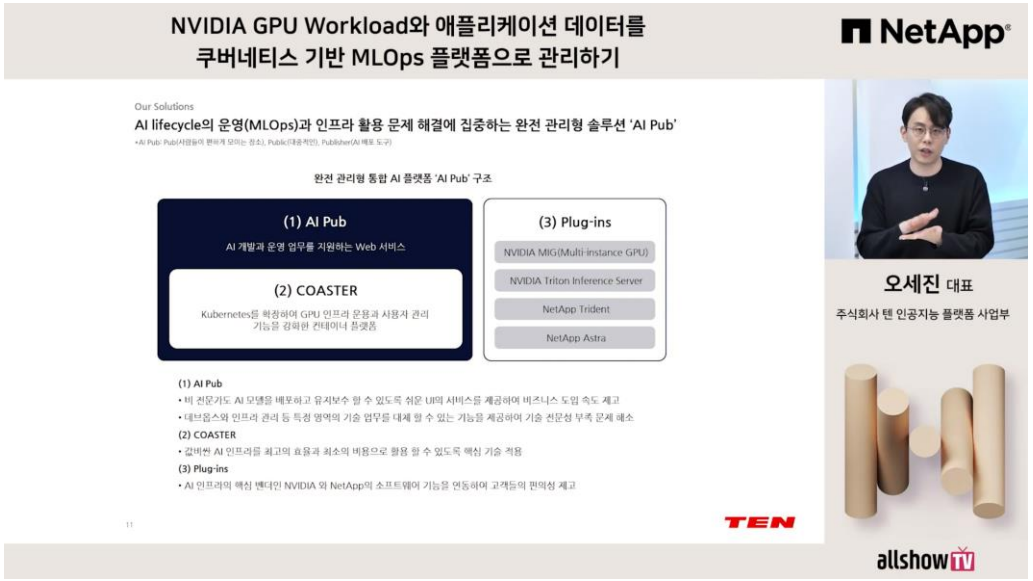
AI Operation by

TEN

AI Pub Ops integrates with NetApp Astra and supports the data management, backup, migration and rollback of applications that consist of multiple micro-services.



COASTER X NetApp



[Webinar]
'Easier management of AI services and data on MLOps platform in Kubernetes environment'

AI Pub Ops uses NeuVector's scanning function to filter docker images from various sources to ensure security.

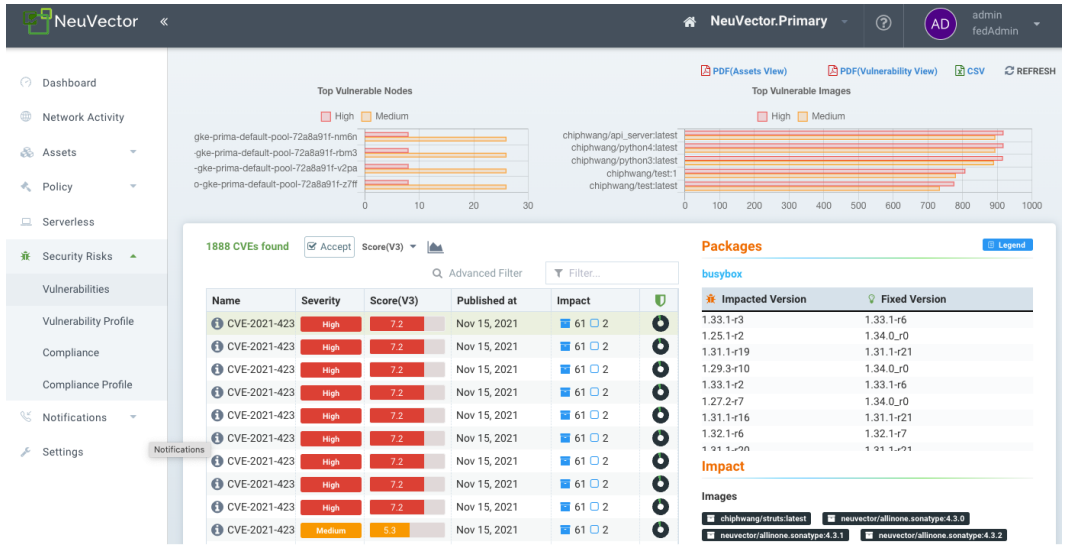
AI Pub Ops can check for vulnerabilities in container images, nodes that form the clusters, and running images or jobs in a container.

NeuVector
Full Lifecycle Container Security

AIPub

AI Pub Image Registry

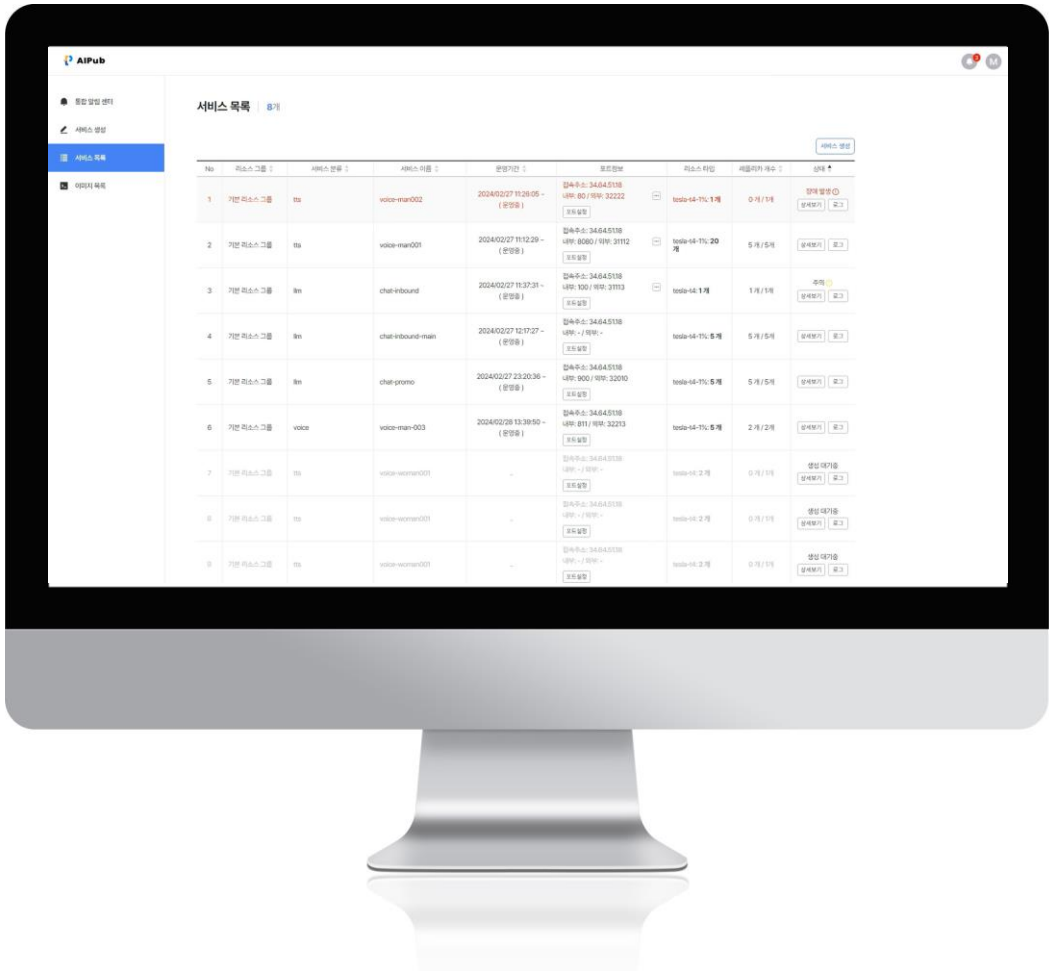
COASTER x SUSE

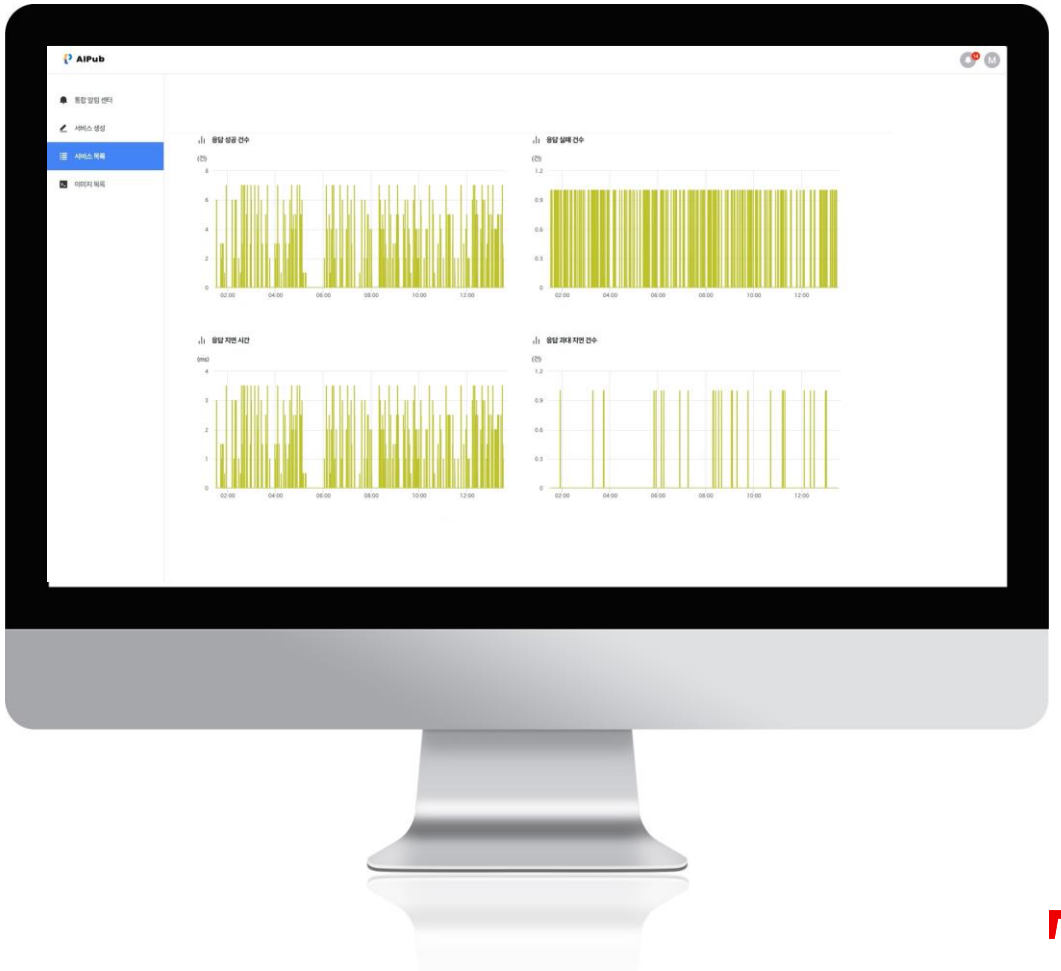


CVE Scanning	The CVE system manages unique identifiers assigned to various vulnerabilities and system defects that may expose devices, systems and programs to hacking.

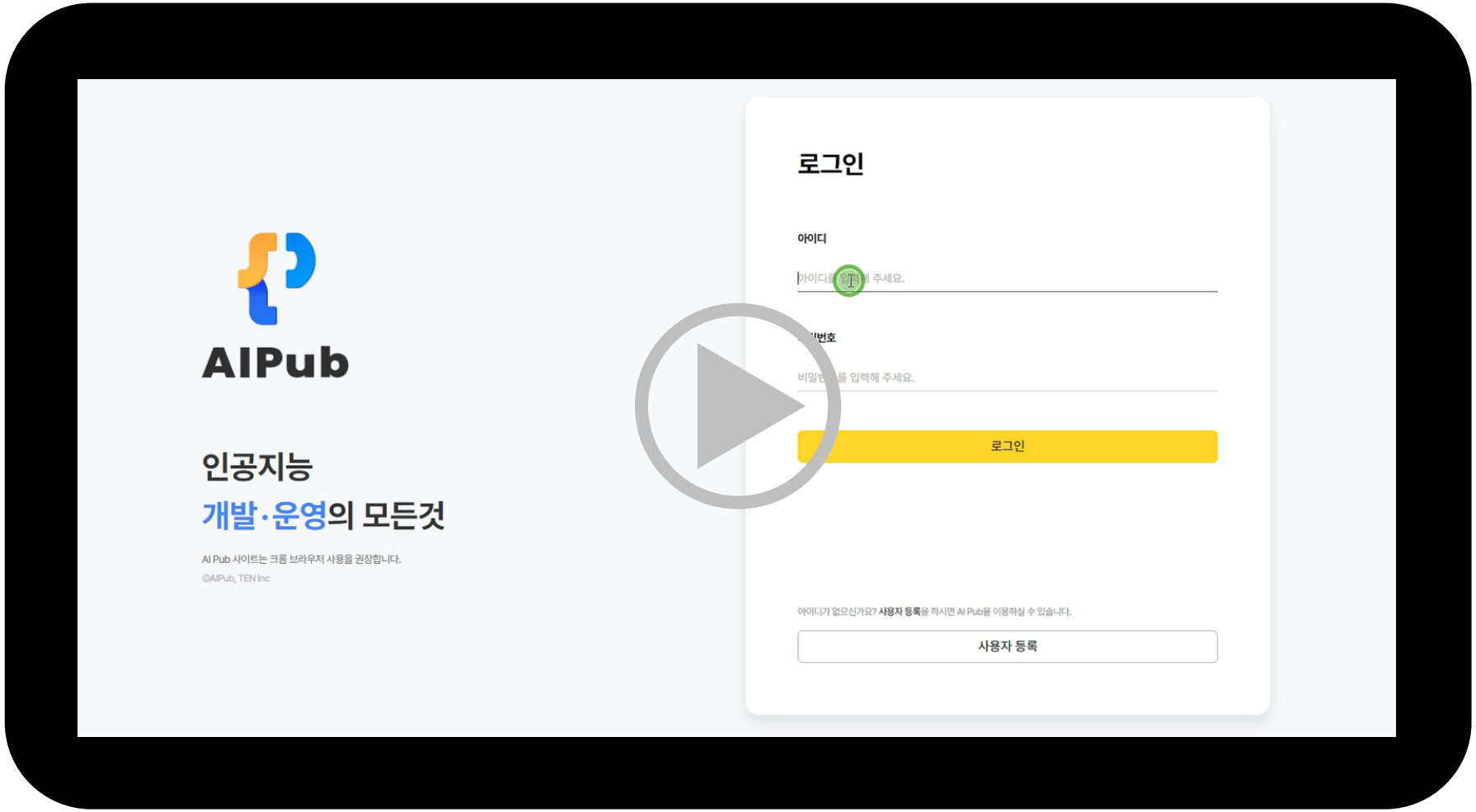
AI Pub Ops provides monitoring services for the AI resource manager and service operator.

Use AI Pub Ops to monitor the status of the system and manage risks.





See a demonstration of AI Pub Ops' functions in this video.



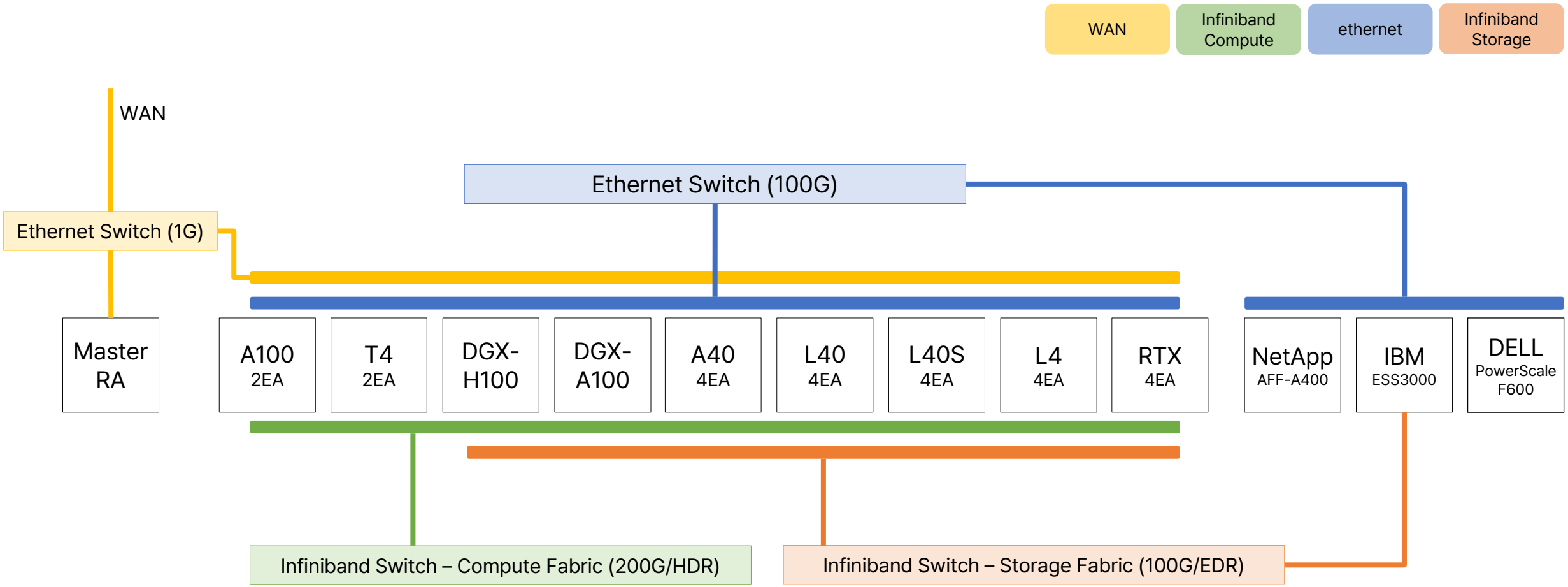
RA:X

AI Reference Architecture of TEN

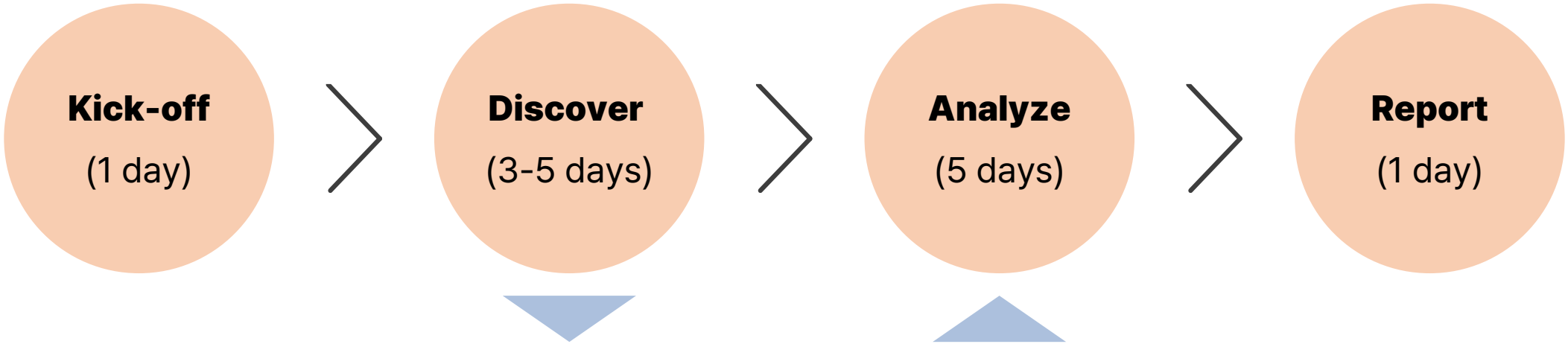
TEN's know-how and test-based consulting service
for AI infrastructure construction

RA:X does not suggest that each hardware be purchased on a budget.
RA:X makes you know the value of the entire infrastructure.

TEN introduces a Reference Architecture built and operated directly.



Infrastructure configuration consulting based on RA:X follows the process outlined below.

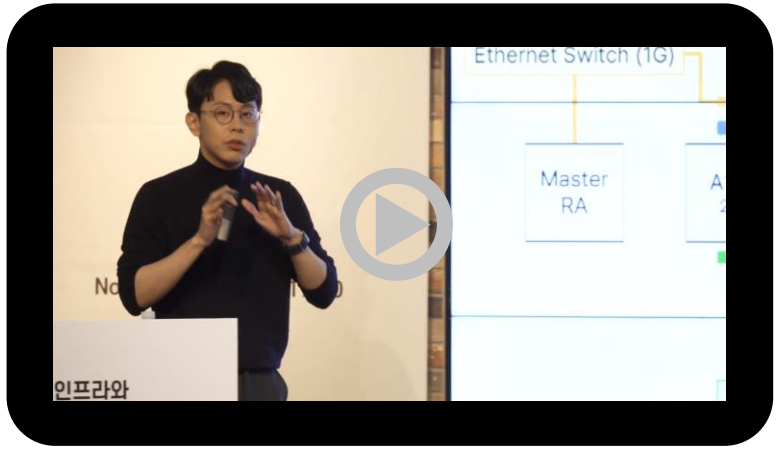


[Survey Categories]

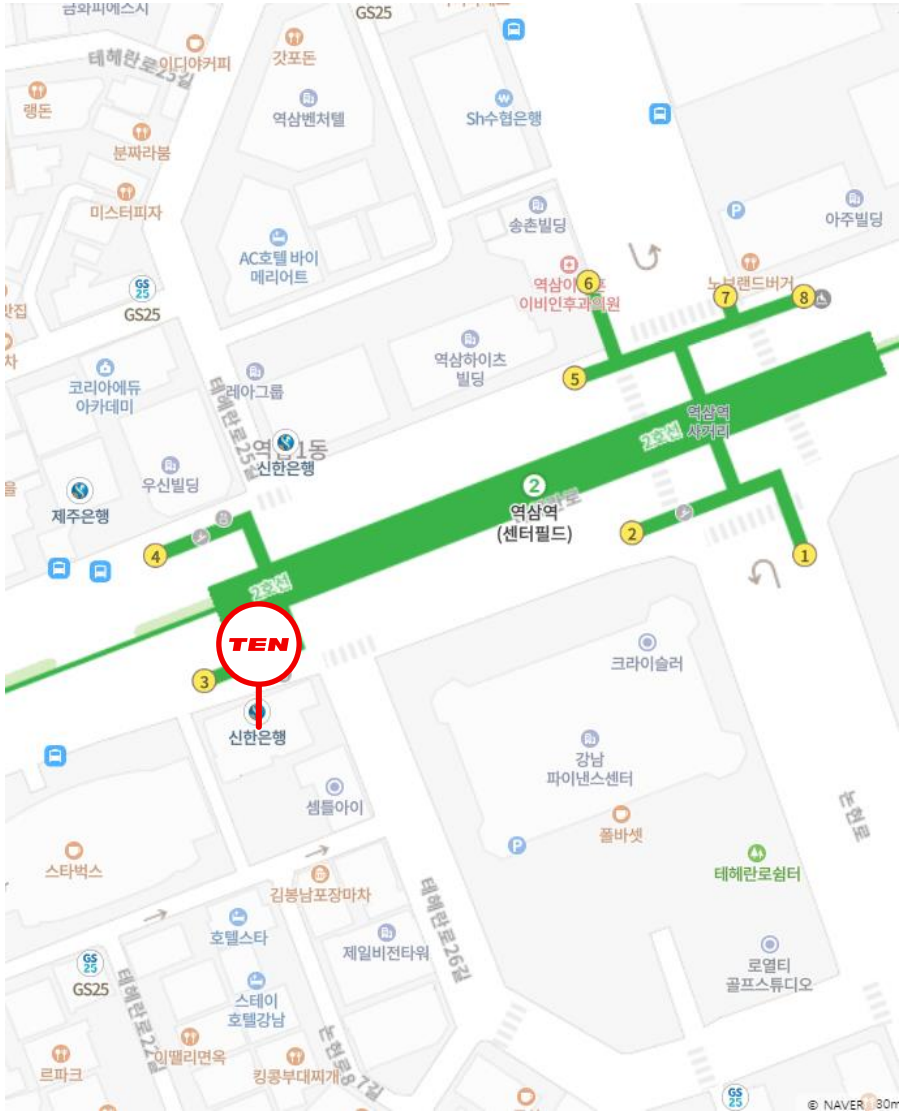
- Business Strategy
- Data Science
- Governance
- Applications etc.
- Data Engineering
- Request for Infrastructure
- Security

TEN is hosting seminars and workshops on the challenges and solutions of AI infrastructure configuration. We empathize with concerns about infrastructure and aim to share our insights in various forums to assist in finding solutions.

2023 AI Infrastructure Workshops
: New AI Infra Solution of TEN
& NVIDIA-Certified Solution
Opening Session by TEN



2023 AI Infrastructure Workshops
: New AI Infra Solution of TEN
& NVIDIA-Certified Solution
Session C, Closing Session by TEN



MORSE ME
WHEN YOU NEED
HELP WITH AI.

Are you looking for a competent partner for your AI services?

Experience a whole new level of AI development and operation by partnering with TEN Inc. We offer you highly-efficient solutions as your trustworthy partner.

A No. 1203, Hyeonik Building, Taehaeran-ro, Yeoksam-dong, Gangnam-gu, Seoul, South Korea

P +82-2-6956-1071

M helloten@ten1010.io

H https://ten1010.io/

Make AI accessible for all

THANK YOU