



Giải pháp Hạ tầng AI Doanh nghiệp

Inspur AI Server & AIStation

Tài liệu tổng thể về Inspur Group, danh mục AI Server, nền tảng AIStation và đánh giá khả năng phân phối tại thị trường Việt Nam.

Phát hành

Phiên bản 1.0 · Tháng 07/2026

Nội bộ QD.TEK – Sales, Presales & BOD

Đơn vị biên soạn

CÔNG TY CP PHÂN PHỐI CÔNG NGHỆ QUANG DŨNG

phong.dao@qdtypek.vn

Về bản trình bày này

Tài liệu tổng thể phục vụ đội ngũ QD.TEK trong hoạt động tư vấn, thiết kế giải pháp và triển khai kênh phân phối cho danh mục Inspur AI Server & nền tảng AIStation tại Việt Nam.

01 · Phiên bản v1.0 Phát hành Tháng 07/2026 Bản đầu tiên tổng hợp từ tài liệu hãng Inspur/IEIT Systems và báo cáo thị trường.	02 · Ngôn ngữ Tiếng Việt Giữ nguyên thuật ngữ EN chuyên ngành GPU, NVLink, MIG, MLOps, OAM, DDP... được giữ nguyên gốc để tra cứu.	03 · Đối tượng Sales · Presales Kiến trúc sư · BOD Dùng làm tài liệu tham chiếu nội bộ QD.TEK cho hoạt động phân phối Inspur.	04 · Phạm vi 11 phần 58 slides Bám sát cấu trúc tài liệu nguồn Inspur_AI_Server_AIStation_Solution_VN.
---	--	--	---

Lưu ý về thương hiệu

Từ năm 2023, mảng máy chủ của Inspur được tái cấu trúc và cung cấp ra thị trường quốc tế dưới thương hiệu **IEIT Systems**. Trong toàn bộ tài liệu này, "Inspur" và "IEIT Systems" đề cập đến cùng một danh mục sản phẩm máy chủ và nền tảng AI.

Mục lục – Nền tảng & Giải pháp

01	Tổng quan về Inspur Group Lịch sử, quy mô, ba công ty niêm yết, thị phần toàn cầu và mạng lưới khu vực.	Slide 06 – 11
02	Danh mục giải pháp AI & Server Kiến trúc 3 lớp: Infrastructure – Platform – Application; các dòng server chủ lực.	Slide 13 – 15
03	Inspur AI Server – Phần cứng Triết lý heterogeneous computing, chi tiết các model flagship (NF5488A5, NF5688M7...).	Slide 17 – 21
04	Nền tảng AIStation – Phần mềm Kiến trúc 4 lớp, modules, workflow, reference architecture và chỉ số hiệu năng.	Slide 23 – 28
05	Case study & khách hàng tiêu biểu SANY Heavy Industry, Autonomous Driving, Media, Alibaba, 12306 Railway.	Slide 30 – 32
06	Đánh giá toàn diện – SWOT Strengths – Weaknesses – Opportunities – Threats và ma trận tổng hợp.	Slide 34 – 38

Mục lục – Thị trường & Go-to-Market

07	Phân tích thị trường Việt Nam Quy mô data center, AI server, cấu trúc thị phần và các phân khúc khách hàng.	Slide 40 – 43
08	So sánh cạnh tranh Inspur vs Dell, HPE, Lenovo, Supermicro, Huawei – bảng key point và radar chart.	Slide 45 – 46
09	Khả năng phân phối & Go-to-Market Mô hình 3 tầng, chiến lược giá, phân khúc ưu tiên, KPI năm đầu tiên.	Slide 48 – 51
10	Kết luận & khuyến nghị 7 khuyến nghị chiến lược và roadmap triển khai 24 tháng.	Slide 53 – 55
11	Phụ lục & nguồn tham khảo Glossary, các tài liệu tham chiếu và thông tin liên hệ QD.TEK.	Slide 57 – 58

Bố cục 11 phần – được xây dựng để đội Presales có thể trích riêng từng cụm (Phần 3–4 cho kiến trúc; Phần 7–9 cho kinh doanh) mà vẫn giữ nguyên logic tổng thể.



Phần 01 · Part One

01

Tổng quan về Inspur Group

Phạm vi phần này

Từ Lãng Trào 浪潮 1960 đến số 1 AI Server toàn cầu.

Phần này tổng hợp bức tranh doanh nghiệp Inspur Group: lịch sử phát triển 65 năm, ba công ty niêm yết, quy mô R&D, vị thế thị phần theo Gartner & IDC, và mạng lưới hiện diện tại 120+ quốc gia – trong đó có Việt Nam.

- 1.1 Giới thiệu chung & tầm nhìn
- 1.2 Lịch sử phát triển (1960 – 2023)
- 1.3–1.4 Quy mô R&D & ba công ty niêm yết
- 1.6–1.8 Thị phần và vị thế toàn cầu
- 1.9–1.10 Khách hàng chiến lược & mạng lưới khu vực

Inspur – Lãng Trào 浪潮

Một trong những tập đoàn công nghệ thông tin hàng đầu Trung Quốc, chuyên cung cấp sản phẩm và dịch vụ về **cloud computing, big data, trí tuệ nhân tạo (AI), Industrial Internet, application software, new-generation communication** cùng nhiều lĩnh vực ứng dụng khác.

Hiện diện toàn cầu

120+

quốc gia và khu vực trên toàn cầu

Lâu đời nhất Trung Quốc

1960

Khởi nguồn từ nghiên cứu transistor

Tầm nhìn

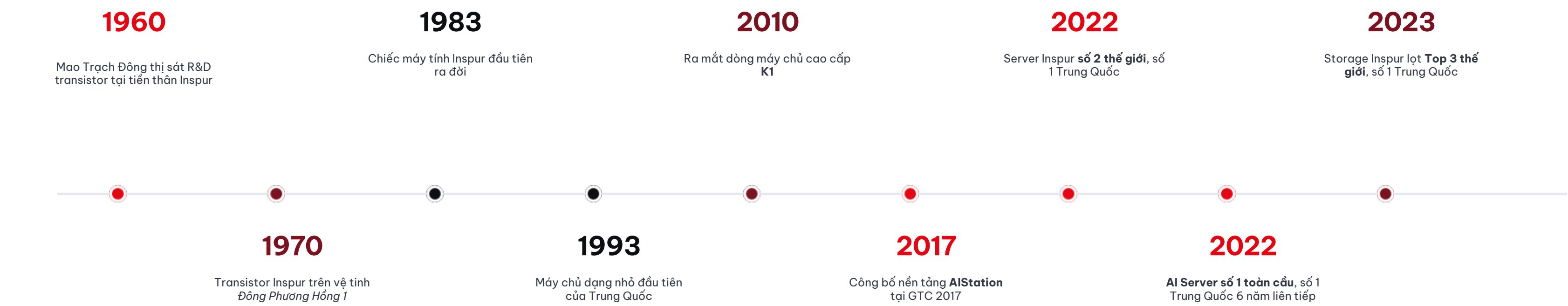
Trở thành doanh nghiệp hàng đầu thế giới trong lĩnh vực CNTT thế hệ mới – nhà cung cấp dịch vụ xuất sắc cho quá trình chuyển đổi số của nền kinh tế xã hội, và là doanh nghiệp xương sống trong xây dựng hạ tầng số mới (*new infrastructure*).

6 trụ cột kinh doanh

- 01 Computing Equipment**
Server · Storage · Network Equipment
- 02 Software**
ERP · PaaS Haiyue · Cloud OS
- 03 Cloud Computing**
Cloud Sea OS · Public & Private Cloud
- 04 New Generation Communication**
5G · Edge Computing
- 05 Big Data**
Data Platform · Data Governance
- 06 Industrial Internet**
Smart Manufacturing · IIoT

63 năm – từ transistor tới AI Server số 1

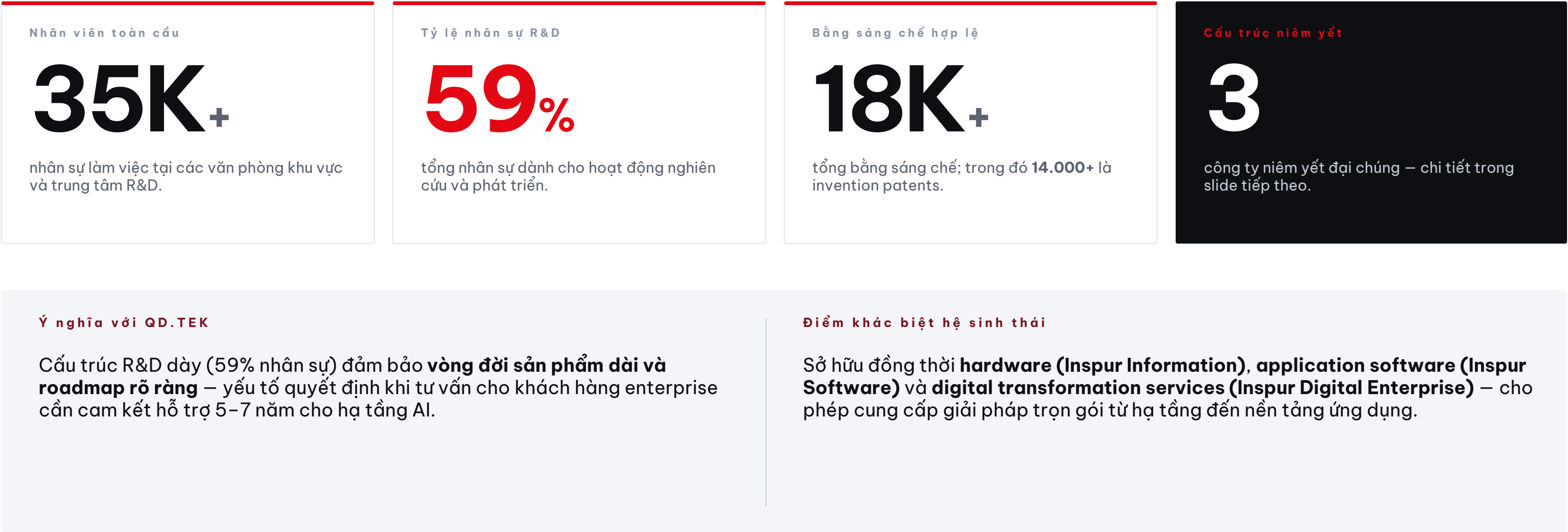
Các mốc chính trong hành trình phát triển của tập đoàn Inspur / Lãng Trào.



Chú thích	Mốc quyết định vị thế AI/Server	Mốc mở rộng danh mục sản phẩm	Mốc kỹ thuật nền tảng
-----------	---------------------------------	-------------------------------	-----------------------

Quy mô nhân lực & năng lực R&D

Bốn chỉ số phản ánh chiều sâu công nghệ và cấu trúc doanh nghiệp của Inspur Group.



Nguồn: Tài liệu chính thức Inspur Group; số liệu tổng hợp cập nhật tới 2023.

Ba công ty niêm yết chủ lực

Ba pháp nhân niêm yết công khai – phủ đủ ba lớp hardware, software và services.

Hardware Arm	Software Arm	Services Arm
Inspur Information	Inspur Software	Inspur Digital Enterprise
Mã chứng khoán 000977 · SZSE	Mã chứng khoán 600756 · SHSE	Mã chứng khoán 00596.HK · HKEX
Trọng tâm Server · Storage · AI Hardware	Trọng tâm Application Software · ERP · Cloud OS	Trọng tâm Digital Transformation Services
Pháp nhân sản xuất và phân phối toàn bộ danh mục Inspur AI Server, general-purpose server, storage & open compute; là nguồn gốc thương hiệu IEIT Systems cho thị trường quốc tế.	Cung cấp ERP cho doanh nghiệp lớn (số 1 Trung Quốc), PaaS Haiyue, Cloud Sea OS – nền tảng phần mềm hạ tầng và ứng dụng cho khối chính phủ và doanh nghiệp trung ương.	Niêm yết trên sàn Hồng Kông (HKEX) – pháp nhân cung cấp dịch vụ chuyển đổi số, tư vấn triển khai và các dịch vụ có nguồn thu ngoại tệ.

Ghi chú

Từ 2023, mảng máy chủ được tái cấu trúc thành thương hiệu **IEIT Systems** cho thị trường quốc tế nhằm thích ứng với các biện pháp kiểm soát xuất khẩu của Hoa Kỳ. Trong tài liệu này, "Inspur" và "IEIT Systems" đề cập cùng một danh mục sản phẩm.

Vị thế thị phần toàn cầu & Trung Quốc

#1

AI Server toàn cầu

Nguồn: IDC · dữ liệu 2022

#2

Server toàn cầu

9,9% thị phần · Gartner 2022

6 năm

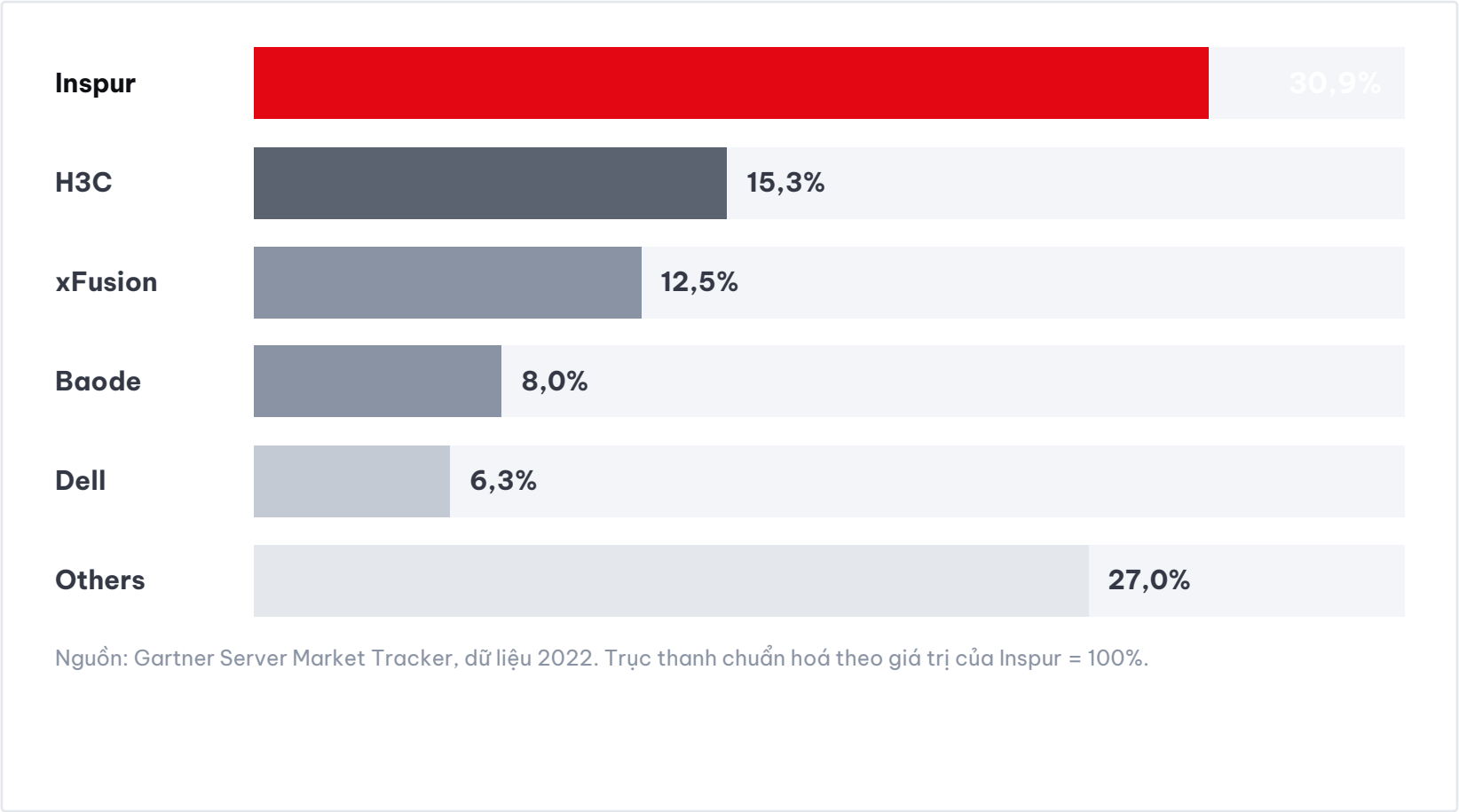
AI Server #1 Trung Quốc

liên tiếp · 46,6% share · IDC China

Bảng tổng hợp – Vị thế toàn cầu & Trung Quốc

SẢN PHẨM	XẾP HẠNG	THỊ PHẦN
Server toàn cầu (2022)	#2 thế giới	9,9%
Server tại Trung Quốc (2022)	#1 Trung Quốc	30,9%
AI Server toàn cầu	#1 thế giới	—
AI Server Trung Quốc (6 năm liên tiếp)	#1	46,6%
Storage Trung Quốc (10 quý liên tiếp)	#1	—
Storage toàn cầu (H1/2023)	Top 3	—
Cloud OS (Cloud Sea) – tỷ lệ bằng sáng chế	#1 Trung Quốc	—
Phần mềm quản trị doanh nghiệp lớn	#1 Trung Quốc	—

Thị phần Server tại Trung Quốc – 2022



Khách hàng chiến lược & mạng lưới khu vực

Khách hàng chiến lược tiêu biểu

Alibaba

10 năm liên tiếp

Đối tác hạ tầng cho sự kiện **Double Eleven (11/11)** – cung cấp >**60%** năng lực điện toán cho hạ tầng Internet Trung Quốc trong sự kiện.

12306

7 năm liên tiếp

Hệ thống đặt vé đường sắt quốc gia Trung Quốc – nâng năng lực đặt vé cao điểm từ **1,2 triệu vé/ngày** lên **20 triệu vé/ngày**; triển khai AI face recognition, giảm 3× thời gian kiểm vé thủ công.

Khách hàng khối trọng điểm

China Mobile · China Post · 75+ doanh nghiệp trung ương

Cùng các nhà cung cấp dịch vụ điện toán đám mây hyperscale toàn cầu.

Mạng lưới khu vực – trích ngang

KHU VỰC	HEAD OFFICE	L1 ENG.	L2 DED.	KHO
Singapore	Singapore	15	25	2.500 ft²
Hong Kong	Hong Kong	15	25	1.000 ft²
China Mainland	Shanghai	50 (100)	12	6.200 ft²
Indonesia	Jakarta	10 (25)	22	1.000 ft²
Thailand	Bangkok	15 (8)	15	1.800 ft²
Malaysia	Kuala Lumpur	8 (20)	24	1.600 ft²
Japan	Tokyo	10	18	1.500 ft²

Ghi chú Việt Nam · Dư địa cho QD.TEK

Việt Nam đã có hiện diện Inspur (Hà Nội & HCM), **Skillset**: Server, Storage, Network, Wireless, Security. Tuy nhiên **Cabling** và **System Integration** trực tiếp còn hạn chế so với Singapore, Malaysia, Australia.

→ Chính là **khoảng trống** mà **đối tác phân phối trong nước (QD.TEK)** có thể bổ sung: deployment, cabling, SI end-to-end.



Phần 02 · Part Two

02

Danh mục Giải pháp AI & Server

Phạm vi phần này

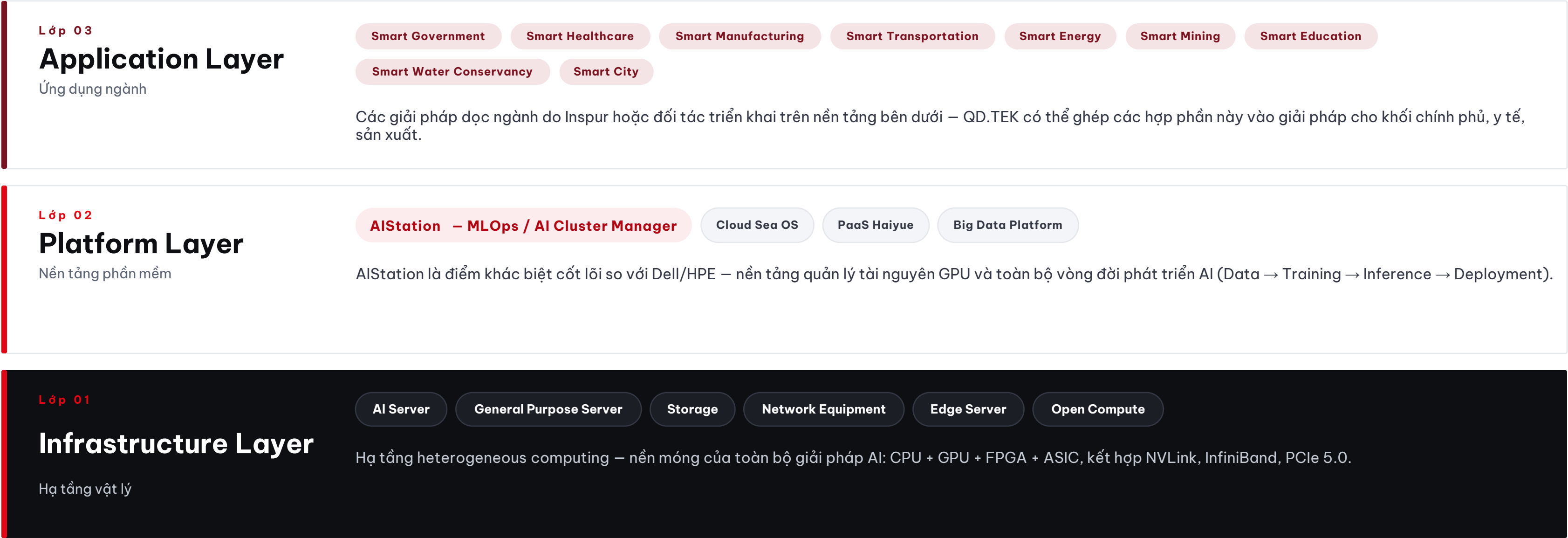
Ba lớp Infrastructure – Platform – Application.

Danh mục Inspur được tổ chức theo mô hình ba lớp – từ hạ tầng phần cứng (AI Server, Storage, Network), qua lớp nền tảng (AIStation, Cloud Sea OS), tới các giải pháp ứng dụng ngành (Smart Government, Smart Healthcare, Smart Manufacturing...).

- 2.1 Kiến trúc danh mục 3 lớp
- 2.2 Phân loại các dòng server chính
- 2.3 Định vị các dòng AI Server chủ lực

Kiến trúc danh mục ba lớp

Ba lớp xếp chồng từ hạ tầng vật lý lên tới các ứng dụng ngành – kiến trúc chuẩn để tư vấn tổng thể cho khách hàng.



Phân loại các dòng server chính

Sáu nhóm sản phẩm bao phủ đủ nhu cầu từ AI training tới edge computing.

DÒNG SẢN PHẨM	KÝ HIỆU TIÊU BIỂU	ĐỊNH VỊ	ỨNG DỤNG ĐIỂN HÌNH
AI Server (Heterogeneous) Chuyên dụng cho AI	NF5488 · NF5468 · NF5688 · NF5448	Máy chủ chuyên dụng AI training/inference với GPU & accelerator	Deep Learning · LLM training · HPC
Rack Server General purpose	NF5280 · NF5180 · NF8480	Server đa dụng cho ứng dụng doanh nghiệp	Virtualization · Database · ERP
High-end Server Mission-critical	K1 Power	Máy chủ high-end mission-critical	Ngân hàng · Bảo hiểm · Chính phủ
Edge Server Edge computing	NE5260 · NE3160	Server biên tối ưu điều kiện triển khai xa DC	5G · IoT · Smart City
Storage SAN/NAS/Distributed	AS series	Storage doanh nghiệp – Top 3 thế giới H1/2023	Big Data · Backup · Archive
Open Compute Chuẩn OCP/OAM	ORv3 · OAM	Server theo chuẩn Open Compute Project	Hyperscale Data Center

Định vị các dòng AI Server chủ lực

Danh mục AI Server bao phủ từ 2U inference edge cho tới 6U flagship LLM training.

MODEL	GPU HỖ TRỢ	CPU	FORM FACTOR	ĐỊNH VỊ
NF5488A5	8 × NVIDIA A100 (NVLink)	2 × AMD EPYC 7002/7003	4U	AI training hiệu năng cao
NF5488M6	8 × NVIDIA A800 (NVLink)	2 × Intel Xeon Ice Lake	4U	AI training – phiên bản cho khu vực có kiểm soát xuất khẩu
NF5688M6	8 × NVIDIA A100/A800 (NVSwitch)	2 × Intel Xeon Ice Lake	6U	Large-scale training
NF5688M7	8 × NVIDIA H100/H800 (NVSwitch)	2 × Intel Xeon Sapphire Rapids	6U	LLM training · Generative AI (flagship)
NF5468A5	8 × PCIe 4.0 GPU (A100/A40/T4)	2 × AMD EPYC Rome/Milan	4U	AI training/inference linh hoạt
NF5468M6/M7	8–20 × PCIe GPU	2 × Intel Xeon	4U	Cloud gaming · IVA · inference
NF5448A6	4–8 GPU PCIe	2 × AMD EPYC	4U	Light AI training · HPC (2,5 petaFLOPS/unit)
NF5280M5/M6	2–3 × T4 (inference)	Intel Xeon	2U	Inference tại edge / branch

Nguồn: Datasheet chính thức Inspur/IEIT Systems (2022–2024). Chi tiết từng model được đi sâu trong Phần 3.



Phần 03 · Part Three

03

Inspur AI Server – Phần cứng

Phạm vi phần này

Heterogeneous computing tối ưu 4 chiều.

Inspur không định vị AI Server là "máy chủ có GPU" đơn thuần, mà là heterogeneous computing platform – tối ưu đồng thời Compute (CPU+GPU+FPGA+ASIC), Interconnect (NVLink, NVSwitch, PCIe 5.0, InfiniBand), Cooling (Air + Liquid + Immersion) và Density (8 GPU trong 4U/6U).

- 3.1 Thiết lý thiết kế heterogeneous
- 3.2 Chi tiết 4 model tiêu biểu + reference cluster
- 3.3 Ecosystem GPU/Accelerator được hỗ trợ
- 3.4 Sản phẩm Open Accelerator (OCP/OAM)

Heterogeneous Computing Platform

Inspur định vị AI Server không đơn thuần là "máy chủ có GPU" mà là **heterogeneous computing platform** – được tối ưu toàn diện theo 4 chiều dưới đây.

Chiều 01

Compute

Kết hợp nhiều loại đơn vị tính toán:

CPU (x86/ARM)
+ GPU
+ FPGA / ASIC accelerator

Cho phép chọn kiến trúc tối ưu theo workload: training GPU-heavy, inference với ASIC, xử lý luồng với FPGA.

Chiều 02

Interconnect

Kết nối tốc độ cao giữa các đơn vị:

NVLink
+ NVSwitch
+ PCIe 4.0 / 5.0
+ InfiniBand HDR / NDR

Bandwidth tới **600 GB/s** giữa 8 GPU trong cùng 1 server – quyết định hiệu năng distributed training.

Chiều 03

Cooling

Ba lớp giải nhiệt sẵn có:

Air Cooling
+ Direct-to-Chip (D2C)
+ Immersion Cooling

Bắt buộc khi GPU thế hệ H100/H200 có TDP >700W – DC tại VN cần lộ trình chuyển sang liquid.

Chiều 04

Density

Mật độ tính toán trong 1 rack:

Hỗ trợ **8 GPU** trong 4U/6U
Tối ưu GPU/rack
Tối ưu W/rack

Ưu thế cạnh tranh chính khi giá thuê DC và điện năng tại VN đều cao.

Inspur NF5488A5

"Flagship AI Training Server" – 8 × NVIDIA A100 SXM4 với NVLink thế hệ 3 trong form-factor 4U.

Thông số nổi bật

Form factor	4U rack
GPU	8 × NVIDIA A100 SXM4 với NVLink thế hệ 3
CPU	2 × AMD EPYC 7002/7003 TDP 280W mỗi socket
NVLink bandwidth	600 GB/s per GPU
GPU memory	Tối đa 640 GB HBM2e với 8 × A100 80GB – aggregated
MIG partitioning	Tối đa 7 instance / GPU 8 GPU → tới 56 instance độc lập

Điểm nhấn hiệu năng

Hiệu năng gấp 56×

Khi kích hoạt MIG tối đa trên toàn bộ 8 GPU – đạt 56 GPU instance độc lập trong 1 hệ thống 4U, phục vụ multi-tenant hoặc inference lớn.

Vị trí trong danh mục · Use case

NF5488A5 là dòng flagship thế hệ Ampere – chọn khi khách hàng có bài toán training model lớn nhưng chưa nhất thiết cần Hopper.

- Deep Learning training
- Recommendation model
- Computer Vision
- HPC hybrid
- Multi-tenant AI cloud

NF5688M7 & NF5468A5

3.2.2 · LLM & Generative AI

NF5688M7

Form factor	6U
GPU	8 × NVIDIA H100/H800 full-mesh NVSwitch
CPU	2 × Intel Xeon Sapphire Rapids
Memory	DDR5 · hỗ trợ > 2 TB
Networking	Tối đa 10 × 200G HDR InfiniBand
Storage	Tối đa 16 × 2.5" NVMe SSD

Use case

LLM training, fine-tune model >70B, Generative AI stack. Là dòng cao cấp nhất Inspur cho training thế hệ Hopper.

3.2.3 · Flexible AI Platform

NF5468A5

Form factor	4U
GPU	Tối đa 8 × PCIe 4.0 GPU A100 PCIe / A40 / T4 / V100
CPU	2 × AMD EPYC Rome/Milan
Điểm khác biệt	Hỗ trợ liquid-cooled NVIDIA PCIe GPU – thế hệ đầu tiên trong ngành
Linh hoạt cấu hình	Đa dạng loại GPU trong cùng chassis
Cooling	Air / D2C Liquid tùy cấu hình

Use case

AI training vừa và nhỏ, inference density cao, phù hợp SME/mid-market và các dự án đa dạng workload.

NF5448A6 & kiến trúc cụm tham chiếu

3.2.4 · Light AI Training + HPC

NF5448A6

Hiệu năng đỉnh

2,5petaFLOPS / unit

Cao hơn 75–150× so với server 2-socket CPU truyền thống.

Form factor

4U

GPU

4–8 GPU PCIe

CPU

2 × AMD EPYC

Ứng dụng

Cơ sở nghiên cứu · HPC nhẹ · Edge training

3.2.5 · Reference Architecture – cụm AI điển hình

USE CASE	SERVER	SL	GPU TỔNG
Inference cluster Inference đa tenant	N F5280M5	200	800 × T4
Training cluster Training model doanh nghiệp	N F5468M5	40	320 × V100
LLM training Fine-tune / train LLM	N F5688M7	32+	256+ × H100

Cách đọc bảng cho QD.TEK

Ba loại cụm trên là **ba phân khúc phần cứng khác nhau** – QD.TEK có thể tham chiếu khi tư vấn scale cho khách hàng: inference 800 T4 cho công ty nội dung số; 320 V100 cho ngân hàng số; 256 H100 cho nhà mạng / DC provider triển khai LLM.

Nguồn: Reference architecture chính thức từ Inspur – số lượng ước lượng cho các dự án đã triển khai.

GPU/Accelerator ecosystem & OCP

3.3 · Accelerator được hỗ trợ trên AI Server & AIStation

NVIDIA

GPU thương mại phổ biến nhất – cả training & inference

P100

V100

V100S

A30

A40

A100

A800

H100

H800

L40

H200

Lưu ý xuất khẩu: H100/A100 chỉ cung cấp cho thị trường được phép; A800/H800 là bản dành cho thị trường có kiểm soát.

Cambricon · FPGA · ASIC

Đa dạng hoá vượt ngoài NVIDIA – giảm phụ thuộc thương mại & địa chính trị

Cambricon (China)

MLU270

MLU290

MLU370

MLU390

FPGA & ASIC

Intel FPGA

Xilinx FPGA

Custom ASIC (hyperscale)

3.4 · Open Accelerator (OCP / OAM)

Vị thế OCP

Thành viên tích cực của Open Compute Project và OAM.

Nền tảng Open Accelerator AI Server có các đặc điểm kỹ thuật vượt trội bên phải.

CPU-GPU BW

4x

so với thiết kế truyền thống

Network BW

4.0 Tbps

tối đa

System Memory

8 TB

tối đa

Local Storage

300 TB

tối đa

Nguồn: Tài liệu Open Accelerator platform của Inspur/IEIT Systems.



Phần 04 · Part Four

04

Nền tảng AIStation – Phần mềm

Phạm vi phần này

MLOps end-to-end trên hạ tầng Inspur.

AIStation là AI Development & MLOps Platform do Inspur phát triển từ 2017, quản lý toàn bộ vòng đời AI: Data → Training → Inference → Deployment. Đã tích hợp chính thức với NVIDIA AI Enterprise Software Suite (08/2022).

- 4.1 Định vị sản phẩm & tích hợp NVIDIA AI Enterprise
- 4.2 Kiến trúc tổng thể 4 lớp
- 4.3 5 modules chức năng
- 4.4 Workflow phát triển AI
- 4.5–4.7 Reference infra · Hiệu năng · Khách hàng mục tiêu

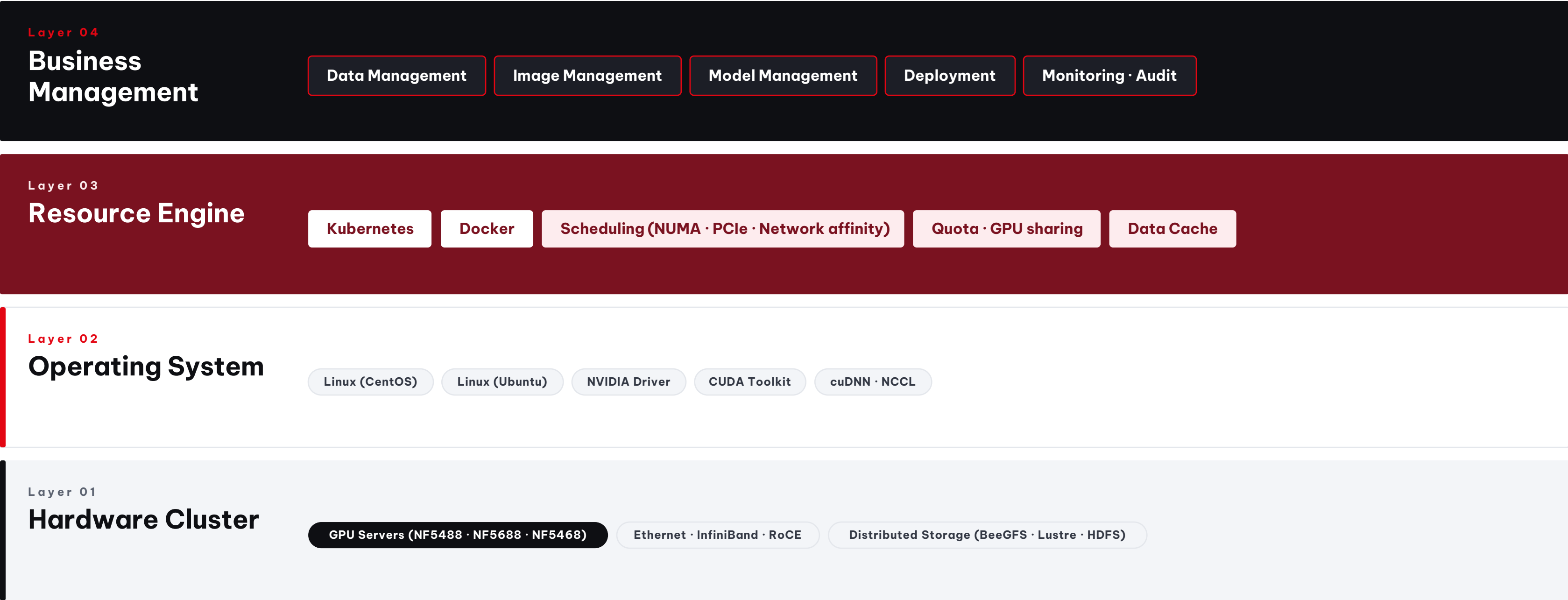
AIStation – AI Development & MLOps Platform

Nền tảng phần mềm cấp doanh nghiệp do Inspur phát triển từ năm **2017** – được công bố lần đầu tại GTC 2017 và tích hợp chính thức với NVIDIA AI Enterprise Software Suite từ 08/2022.

Giá trị 01 Quản lý tài nguyên GPU cluster Cho phép nhiều team, project, tenant cùng dùng chung cụm GPU với quota, priority, sharing – nâng hiệu suất sử dụng cụm từ ~40% lên 80%.	Giá trị 02 Toàn bộ vòng đời phát triển AI Bao phủ 4 giai đoạn: Data → Training → Inference → Deployment – cùng công cụ Jupyter, WebShell, Pipeline, Visualization tích hợp sẵn.	Giá trị 03 Rút ngắn training, tăng iteration Training time từ 2 ngày còn 4 giờ (~12×); deployment 2 ngày còn 5 phút (~576×); fault recovery tự động trong 30 giây.
Tích hợp chiến lược NVIDIA AI Enterprise Software Suite Công bố chính thức tháng 08/2022	Ý nghĩa Cung cấp bộ công cụ AI end-to-end được NVIDIA chứng nhận – không phải mô hình "third-party MLOps" như Dell (Run:ai) hay HPE (Ezmeral). Đây là điểm khác biệt cạnh tranh cốt lõi vs mọi đối thủ tại VN.	

Kiến trúc tổng thể – 4 lớp

Ngăn xếp AIStation đọc từ dưới lên: Hardware → Operating System → Resource Engine → Business Management.



Năm modules chức năng chính

1

Data Management

- Data Import từ NAS, Object Storage, HDFS
- Data Preprocessing
- Data Acceleration (caching)
- On-demand loading

2

Development Tools

- **Jupyter Notebook** tích hợp sẵn
- **WebShell** truy cập cluster từ trình duyệt
- **Pipeline** – tự động hoá workflow ML
- **Visualization** – giám sát training real-time

3

Training Framework

Hỗ trợ đầy đủ framework chủ đạo:

TensorFlow

PyTorch

Caffe

MxNet

PaddlePaddle

- Hyper-parameter tuning tự động
- Distributed Data Parallel (DDP) tự động

4

Model Serving

- TensorFlow Serving
- NVIDIA TensorRT Inference Server (**Triton**)
- PyTorch Inference Server
- Auto-scaling · Load balancing

5

Resource Scheduling

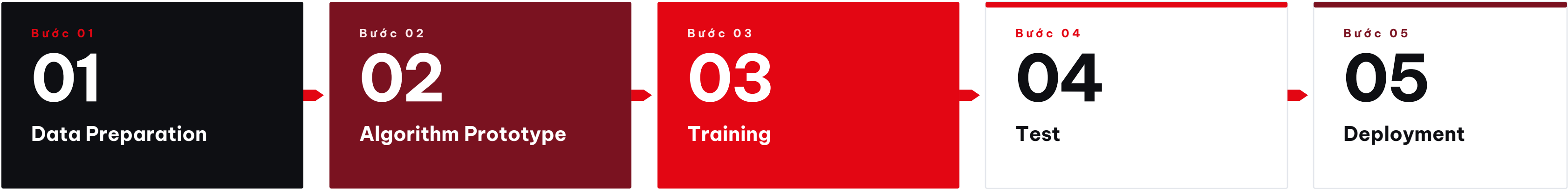
Core engine · Nơi tạo khác biệt cạnh tranh

- Scheduling theo **NUMA affinity** · **PCIe affinity** · **Network affinity**
- **GPU sharing** tới mức 1 GB memory hoặc 64 task/card

- **Quota management** theo user / team / project
- **Fault tolerance** – tự động migrate task trong **30 giây**

Workflow phát triển AI trên AIStation

Năm giai đoạn từ chuẩn bị dữ liệu tới triển khai model – mỗi giai đoạn có công cụ tương ứng của AIStation.



Công cụ AIStation tương ứng

Công cụ Data Management Preprocessing Data Cache	Công cụ Jupyter Notebook WebShell Pipeline builder	Công cụ Distributed Training DDP tự động Hyper-param tuning	Công cụ Model Registry A/B Testing Metrics Dashboard	Công cụ TensorRT (Triton) TF Serving Auto-scaling
--	--	---	--	---

Điểm bán hàng chính (Sales angle)

Với khách hàng đang tự dựng MLOps trên Kubernetes thuần: AIStation cắt **60–70% thời gian setup** và cung cấp SLA thương mại cho từng module – điều mà open-source stack không có.

Reference infra & đối tượng khách hàng

4.5 · Cấu hình hạ tầng tham chiếu – cụm AIStation tiêu biểu

THÀNH PHẦN	CẤU HÌNH
Network Data + Management	Ethernet @ 1/10 Gbps IPMI @ 1 Gbps InfiniBand @ HDR / NDR
Storage size Shared FS	4 TB – 100 TB (tùy quy mô)
Compute Node Storage Cache/Scratch	SSD 2 TB – 10 TB per node
GPU per node Cấu hình điển hình	8 × V100 / A100 / H100
Distributed FS Shared filesystem	BeeGFS · Lustre · HDFS

Gợi ý cấu hình khởi động

Cụm PoC nhỏ nhất: **4 node × 8 GPU + 20 TB storage + IB HDR** – đủ để trình diễn giá trị AIStation cho khách hàng enterprise.

4.7 · Đối tượng khách hàng mục tiêu

- 01

Finance
Ngân hàng · Bảo hiểm · Fintech – fraud detection, risk & credit scoring
- 02

Telecom
Nhà mạng – network optimization, customer intelligence
- 03

Internet
Công ty công nghệ – recommendation, ads, content moderation
- 04

Manufacturing
Sản xuất thông minh – defect detection, predictive maintenance
- 05

Medical / Healthcare
Y tế – medical imaging, drug discovery
- 06

Autonomous Driving
Xe tự lái – perception model training
- 07

Smart Transportation
Giao thông thông minh – traffic prediction, autonomous public transport

Các chỉ số hiệu năng nổi bật

Kết quả đo được trên các cụm thực tế do Inspur công bố – dùng làm tham chiếu cho tư vấn cho khách hàng.

Chỉ số	Trước	Sau khi dùng AISTATION	Cải thiện
Thời gian training	2 ngày	4 giờ	~12×
Thời gian deployment	2 ngày	5 phút	~576×
Hiệu suất sử dụng cụm	~40%	80%	2×
Multi-GPU acceleration ratio	—	> 90%	Tối ưu
Fault recovery	Thủ công	Tự động trong 30s	—

Time-to-train

12[×] giảm thời gian training

Từ 2 ngày còn 4 giờ trên cùng cấu hình cụm.

Time-to-deploy

576[×] nhanh hơn

Từ 2 ngày còn 5 phút – cho phép iterate model liên tục.

Cluster utilization

2[×] hiệu suất sử dụng GPU

40% → 80% – tối ưu ROI đầu tư phần cứng.



Phần 05 · Part Five

05

Case study & khách hàng tiêu biểu

Phạm vi phần này

Từ nhà máy 15 tỷ USD tới 20 triệu vé/ngày.

Năm case study đại diện cho 5 dạng bài toán khác nhau: sản xuất công nghiệp, xe tự lái, truyền thông, e-commerce hyperscale và hệ thống giao thông quốc gia – chứng minh giải pháp đã được validate ở nhiều quy mô.

- 5.1 SANY Heavy Industry – Sản xuất thông minh
- 5.2 Doanh nghiệp Autonomous Driving
- 5.3 Công ty Công nghệ Truyền thông
- 5.4 Alibaba – Double Eleven
- 5.5 Đường sắt Trung Quốc – 12306

SANY Heavy Industry

Ứng dụng Inspur AI Server + AIStation cho dây chuyền sản xuất máy xúc tại tập đoàn công nghiệp nặng lớn.

Bối cảnh doanh nghiệp

Doanh thu

15,7 tỷ USD

Sản lượng máy xúc

100K /năm

Nhu cầu

Tự động hoá dây chuyền sản xuất bằng AI – **computer vision detection** lỗi sản phẩm, dò khuyết tật hàn, kiểm tra ren bulon trên vỏ máy xúc.

Áp lực chính: chu kỳ phát triển model 2 tuần là quá lâu để bắt kịp thay đổi mẫu mã sản phẩm.

Giải pháp triển khai

Cụm AI Server Inspur (NF5468 series) + nền tảng AIStation với Pipeline builder, Data Cache, distributed training tự động.

Nguồn: Case study công bố chính thức của Inspur/IEIT Systems.

Kết quả đo lường

Chu kỳ phát triển model

2 tuần → 3 ngày

4,7x

API calls xử lý

25M calls/ngày

Xử lý real-time trên dây chuyền sản xuất.

Tỷ lệ tự động hoá dây chuyền

~90%

Chuyển từ kiểm tra thủ công sang AI detection.

Autonomous Driving & Media

5.2 · Case study 02

Doanh nghiệp Autonomous Driving

Vấn đề

Cần training model perception (Camera + LiDAR fusion) với dataset lớn – cụm GPU hiện tại không hiệu quả, utilization thấp, iteration chậm khiến vòng đời sản phẩm bị trễ.

Kết quả

Cluster utilization

+30%

nhờ scheduling & GPU sharing

Training time

Rút ngắn đáng kể

nhờ DDP tự động

5.3 · Case study 03

Công ty Công nghệ Truyền thông

Vấn đề

Dataset **1,5 TB** với **rất nhiều file nhỏ** – I/O trở thành bottleneck, khiến GPU chỉ đạt 30–40% utilization thực tế dù mua đầy đủ H100.

Kết quả

Giải pháp: **Data caching + distributed FS acceleration** của AIStation (BeeGFS/Lustre).

Tác động chính

Tối ưu hoá đáng kể tốc độ đọc dữ liệu, rút ngắn **epoch time** – giúp iterate model nhanh hơn nhiều lần trên cùng cụm hardware.

Alibaba & hệ thống đường sắt 12306

5.4 · E-commerce hyperscale

Alibaba

Sự kiện

Double Eleven (11 · 11)

Thời gian hỗ trợ

10

năm liên tiếp

Năng lực điện toán

>60%

hạ tăng Internet TQ trong event

Hạ tầng Inspur đóng vai trò xương sống cho sự kiện thương mại điện tử lớn nhất thế giới, với đỉnh giao dịch nhiều triệu lệnh/giây.

5.5 · National Transportation

12306

Hệ thống đặt vé đường sắt quốc gia Trung Quốc

Thời gian hỗ trợ

7

năm liên tiếp

Năng lực đặt vé cao điểm

1,2M vé/ngày → 20M vé/ngày

Ứng dụng AI

Triển khai face recognition cho toàn tuyến — tăng hiệu suất kiểm vé gấp 3x so với thủ công.

Ghi chú: Cùng với Alibaba, hai case study này là bằng chứng Inspur có kinh nghiệm hyperscale ở quy mô hàng chục triệu người dùng đồng thời.

06

Đánh giá toàn diện giải pháp – SWOT

Phạm vi phần này

Bức tranh cân bằng S · W · O · T.

Phần này trình bày cả điểm mạnh (S), điểm yếu (W), cơ hội (O) và thách thức (T) của Inspur trong bối cảnh thị trường Việt Nam – làm nền cho các quyết định về pricing, phân khúc, kênh phân phối ở Phần 9.

- 6.1 Strengths – 8 điểm mạnh cốt lõi
- 6.2 Weaknesses – 6 điểm yếu cần giảm thiểu
- 6.3 Opportunities – 6 cơ hội thị trường VN
- 6.4 Threats – 5 thách thức cạnh tranh
- 6.5 Ma trận SWOT tổng hợp

S

Điểm mạnh – 8 lợi thế cốt lõi

S1

Vị thế thị trường vững chắc

#1 AI Server toàn cầu, #2 Server toàn cầu, dẫn đầu Trung Quốc.

S2

Danh mục AI Server đầy đủ nhất

Từ 2U inference server tới 6U flagship LLM training.

S3

Ecosystem AI hoàn chỉnh

Hardware + AIStation platform + tích hợp NVIDIA AI Enterprise.

S4

Hiệu năng được chứng thực

Multi-GPU acceleration >90%, training time giảm 12×.

S5

Kinh nghiệm quy mô hyperscale

Alibaba, 12306, China Mobile – dự án hàng chục triệu users.

S6

Năng lực R&D mạnh

59% nhân sự R&D, >18.000 bằng sáng chế.

S7

Giá cả cạnh tranh

Thường thấp hơn Dell/HPE 10–20% cùng cấu hình.

S8

Tuân thủ Open Compute

Sản phẩm theo chuẩn OCP & OAM – thân thiện hyperscale DC.

W Điểm yếu – 6 rủi ro cần giảm thiểu

W1	Rào cản từ Entity List Hoa Kỳ Từ 03/2023	Inspur Group nằm trong US Entity List từ 03/2023 – ảnh hưởng cung ứng GPU cao cấp trực tiếp cho khách hàng nhạy cảm.
W2	Thương hiệu quốc tế yếu hơn Brand awareness tại VN	Nhận diện thương hiệu tại VN vẫn thấp hơn Dell/HPE/Lenovo – cần thời gian và chiến dịch xây dựng.
W3	Hệ sinh thái đối tác trong nước còn hạn chế Số SI/reseller ít	Số lượng SI, reseller có certified engineer về Inspur ít – chính là cơ hội cho QD.TEK làm hub kỹ thuật.
W4	Documentation ưu tiên tiếng Trung Tài liệu kỹ thuật	Một số tài liệu kỹ thuật, KB, portal support vẫn ưu tiên tiếng Trung – cần bản địa hoá cho thị trường VN.
W5	Đội ngũ Việt Nam còn mỏng Theo tài liệu hãng	Chỉ 2 (mở rộng 9) L1 engineer tại Hà Nội, tổng kho ~770 ft² – QD.TEK có thể bổ sung 2nd-line support.
W6	Rủi ro địa chính trị Khách hàng nhạy cảm	Ngân hàng nhà nước, quốc phòng, một số dự án nhạy cảm có thể e ngại – cần lộ trình lựa chọn phân khúc phù hợp.



Cơ hội – 6 xu thế đẩy nhu cầu tại VN

O1

Bùng nổ nhu cầu AI Server VN

Thị trường AI Data Center VN dự kiến từ **140,15 triệu USD (2025)** lên

858 triệu USD

tối năm 2035

O2

Chiến lược quốc gia về AI

Chính phủ VN đẩy mạnh chuyển đổi số, AI, chip.

Ngân sách nhà nước và ngân sách các bộ ngành cho AI/DC tăng đều 2023–2026.

O3

Data Center mở rộng

41

DC

đang hoạt động · **221 MW** tổng công suất

Viettel, VNPT, FPT, CMC đầu tư mạnh vào AI-ready DC.

O4

Lợi thế giá

SME & mid-market VN nhạy cảm về giá.

Chênh 10–20% so với Dell/HPE là ngưỡng đủ để khách hàng "chuyển cửa" nếu có phân phối, hỗ trợ tốt.

O5

Thiếu hụt GPU toàn cầu

Khách hàng cần lead time ngắn.

Inspur có năng lực cung cấp nhanh cho thị trường APAC nhờ chuỗi cung ứng gần khu vực.

O6

Tách thương hiệu IEIT Systems

Giảm áp lực địa chính trị.

Phù hợp để tiếp cận khách hàng có ràng buộc tuân thủ trong các dự án quốc tế.

T

Thách thức – 5 áp lực cạnh tranh

T1	Cạnh tranh khốc liệt từ Dell, HPE, Lenovo	Các hãng có brand mạnh, kênh phân phối hình thành lâu năm tại VN – chiếm phần lớn thị phần enterprise và government.
T2	Supermicro tăng trưởng mạnh trong AI Server	Chiếm ~9–25% thị phần AI server global 2025 – cạnh tranh trực tiếp trên SME/hyperscale cả về giá lẫn thời gian giao hàng.
T3	Chính sách xuất khẩu Hoa Kỳ	Có thể siết chặt hơn với công nghệ AI tiên tiến – tác động dây chuyền tới nguồn cung H100/H200/B200 qua kênh Inspur.
T4	NVIDIA DGX là "kim chỉ nam"	Với khách hàng AI enterprise – cạnh tranh trực tiếp về mặt "reference architecture". Viettel đã đưa NVIDIA DGX B200 vào vận hành đầu tiên tại VN.
T5	Local player mới nổi tại VN	G-Group, Saigontel đã trình làng "AI server made in Vietnam" hợp tác với đối tác Hàn Quốc – có thể là lựa chọn ưu tiên với khối chính phủ, quốc phòng.

Ma trận SWOT tóm tắt

Bảng 2×2 tổng hợp – dùng làm điểm chốt cho Phần 7 (Thị trường VN) và Phần 9 (GTM).

	Yếu tố tích cực	Yếu tố tiêu cực
Nội bộ Internal	S Strengths – Điểm mạnh S1 #1 AI Server toàn cầu S2 Danh mục đa dạng – 2U → 6U S3 AIStation – MLOps hoàn chỉnh	W Weaknesses – Điểm yếu W1 US Entity List (03/2023) W2 Đội ngũ VN mỏng W3 Brand awareness thấp tại VN
Bên ngoài External	O Opportunities – Cơ hội O1 Thị trường AI VN tăng CAGR 20%+ O2 Data center bùng nổ (41 DC · 221 MW) O3 Lợi thế giá 10–20% so với Dell/HPE	T Threats – Thách thức T1 Cạnh tranh Dell/HPE/Lenovo T2 Chính sách xuất khẩu Mỹ T3 Supermicro tăng nhanh

07

Phân tích thị trường Việt Nam

Phạm vi phần này

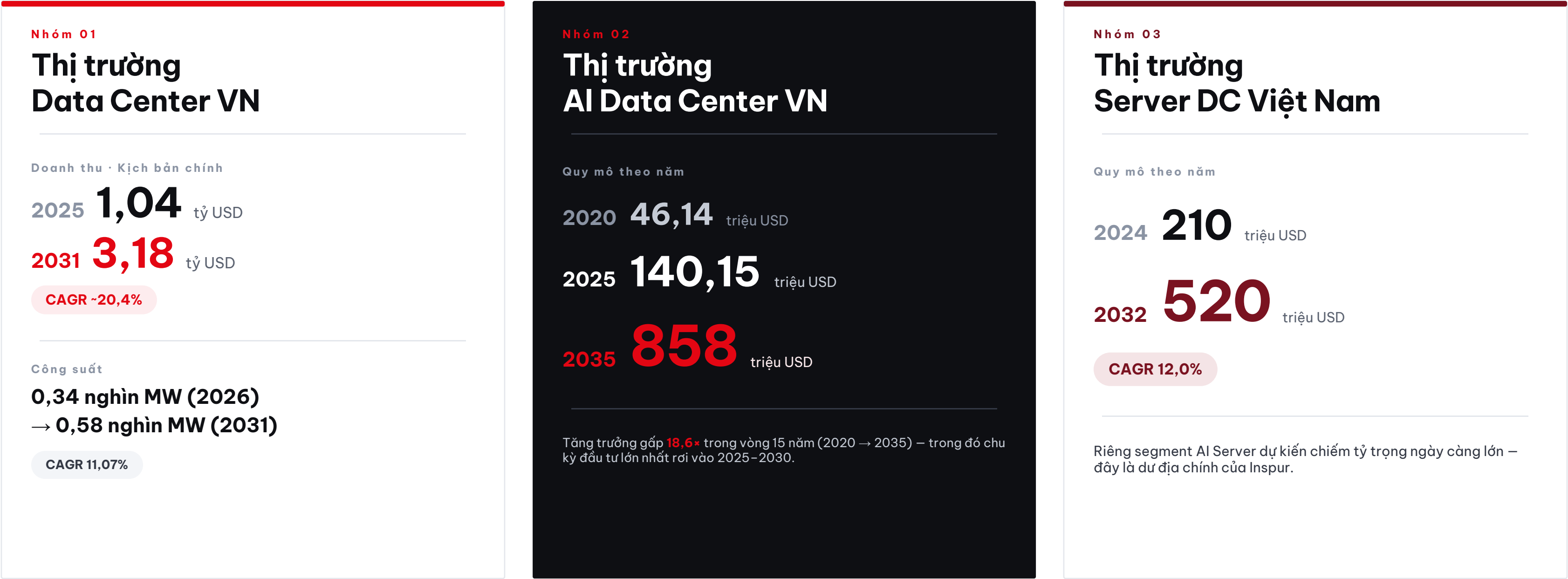
Quy mô – cấu trúc – xu hướng – phân khúc.

Phần này chuyển từ đánh giá sản phẩm sang bối cảnh thị trường – quy mô DC/AI Server tại VN, cấu trúc DC provider hiện tại, các xu hướng nổi bật (NVIDIA DGX B200 Viettel, Foxconn, made-in-Vietnam AI server) và bảng phân khúc khách hàng mục tiêu.

- 7.1 Quy mô và tăng trưởng – DC, AI DC, Server VN
- 7.2 Cấu trúc DC provider Việt Nam
- 7.3 Xu hướng nổi bật 2025–2026
- 7.4 Phân khúc khách hàng & mức ưu tiên

Quy mô và tăng trưởng thị trường Việt Nam

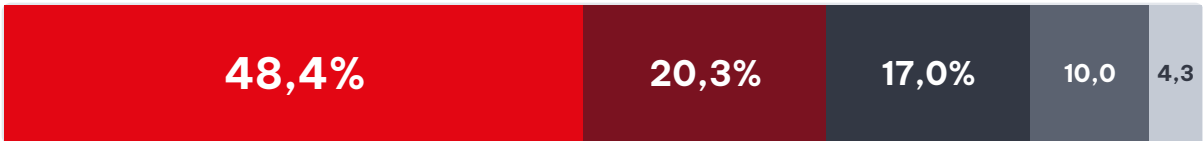
Ba nhóm số liệu: (1) Data Center tổng thể · (2) AI Data Center · (3) Server DC – tất cả theo báo cáo thị trường độc lập.



Cấu trúc Data Center Provider tại VN



Thị phần DC Provider · 2024



- Viettel IDC – 48,4%
- VNPT – 20,3%
- FPT Telecom – 17,0%
- CMC Telecom – 10,0%
- NTT DATA · ST Telemedia · Edge Centres · Hanoi Telecom · VNTT – ~4,3%

Nguồn: Báo cáo thị trường DC Việt Nam 2024.

Ý nghĩa cho chiến lược GTM của QD.TEK

Ý nghĩa 01 – Concentration

4 tên tuổi (Viettel · VNPT · FPT · CMC) chiếm ~95% thị phần – chỉ cần thắng 2-3 khách hàng lớn là đã đạt volume mục tiêu năm đầu.

Ý nghĩa 02 – Chu kỳ đầu tư

Cả 4 provider đều đang mở rộng DC AI-ready 2025-2027 – cửa sổ tiếp cận đúng thời điểm để đặt reference architecture.

Ý nghĩa 03 – Cạnh tranh

Viettel đã chọn NVIDIA DGX B200 làm reference – Inspur cần định vị ở tier tiếp theo (VNPT, FPT, CMC) với lợi thế giá + AIStation bundle.

Ý nghĩa 04 – Foreign player

NTT DATA, ST Telemedia, Edge Centres đang gia nhập – dư địa cho các dự án colocation quy mô vừa, thời gian go-live nhanh.

5 xu hướng nổi bật tại Việt Nam

01	Viettel – NVIDIA DGX B200 đầu tiên tại VN 02/2026	Đưa vào vận hành hệ thống NVIDIA DGX B200 đầu tiên tại Việt Nam – khẳng định vị thế dẫn đầu về hạ tầng AI. Đây là benchmark mà mọi provider khác sẽ so sánh.
02	Siêu máy tính AI nội địa Đang vận hành	Việt Nam sở hữu siêu máy tính AI với hiệu năng 1,5 ExaFLOPs FP8 – thuộc nhóm dẫn đầu Đông Nam Á về tính toán AI.
03	Foxconn mở rộng sản xuất máy chủ AI Chuyển dịch cung ứng	Dự kiến máy chủ AI sản xuất tại VN chiếm ~50% doanh thu server 2026 của Foxconn – thay đổi chuỗi cung ứng và giá thành khu vực.
04	AI Server "made in Vietnam" Local player mới nổi	G-Group, Saigontel giới thiệu máy chủ AI made-in-Vietnam hợp tác với đối tác Hàn Quốc (P&C) – sẽ là lựa chọn ưu tiên cho khối chính phủ / quốc phòng.
05	Chuyển đổi số ngành Digital transformation	Đẩy mạnh chuyển đổi số ngành ngân hàng, viễn thông, chính phủ – nhu cầu AI Server và MLOps platform tăng đồng loạt trên 4 khối.

Phân khúc khách hàng tại Việt Nam

Bảng xếp hạng mức độ ưu tiên (★ 1–5) – cơ sở cho chiến lược GTM ở Phần 9.

PHÂN KHÚC	ĐẶC ĐIỂM & NHU CẦU AI	MỨC ƯU TIÊN	ĐÁNH GIÁ
Nhà mạng viễn thông Viettel · VNPT · FPT · MobiFone	Nhu cầu AI cho 5G, NLP, customer intelligence – dự án quy mô lớn, cam kết đầu tư dài hạn.	★★★★★	Rất cao
Data Center Provider Viettel IDC · VNPT · CMC · FPT	Hạ tầng colocation, GPU-as-a-Service, cần lead time ngắn và giá cạnh tranh.	★★★★★	Rất cao
Ngân hàng số & Fintech TCB, MB, VPBank, Momo, ZaloPay...	AI cho credit scoring, fraud detection – nhu cầu mid-size cluster, ít nhạy cảm địa chính trị.	★★★★	Cao
Tập đoàn công nghệ VNG · Zalo · Momo...	LLM, recommendation system – sẵn sàng đầu tư GPU H100/H200.	★★★★	Cao
Doanh nghiệp sản xuất Vinfast · Samsung Vina · LG...	Computer vision defect detection, predictive maintenance – cluster inference & light training.	★★★★	Cao
Nghiên cứu & Đại học VinAI · VinBigData · Bách khoa · VNU	AI research, benchmark – dự án nhỏ nhưng có giá trị xây thương hiệu và talent pipeline.	★★★	Trung bình
Y tế Bệnh viện lớn, viện nghiên cứu	Medical imaging, drug discovery – vòng đời quyết định dài, cần chứng nhận y tế.	★★★	Trung bình
Chính phủ · Quốc phòng Bộ ngành, đô thị thông minh	Smart city, e-Government – nhạy cảm địa chính trị, ưu tiên "made in Vietnam" hoặc EU/US.	★★	Thấp

08

So sánh cạnh tranh với đối thủ

Phạm vi phần này

Inspur đối mặt với 5 đối thủ chính.

Bảng key point trên 12 tiêu chí kỹ thuật/thương mại và radar chart chấm điểm 1–5 trên 8 chiều – làm cơ sở cho positioning statement khi bán hàng.

8.1 Bảng so sánh key point tổng thể (6 hãng)

8.3 Radar chart định vị 8 chiều

Bảng so sánh key point tổng thể

Inspur vs Dell PowerEdge · HPE ProLiant · Lenovo ThinkSystem · Supermicro · Huawei FusionServer.

Tiêu chí	Inspur / IEIT	Dell PowerEdge	HPE ProLiant	Lenovo ThinkSystem	Supermicro	Huawei FusionServer
Thị phần Server toàn cầu 2022	9,9% (#2)	14,7% (#1)	9,4% (#3)	~7%	~6%	4–5%
Thị phần AI Server 2024	~12%	~20%	~15%	~11%	~9% (đang tăng)	bị hạn chế XK
Flagship AI Server	NF5688M7 (8× H100/H800)	PowerEdge XE9680	Cray XD670 · DL380a	ThinkSystem SR685a V3	SYS-821GE-TNHR	Atlas 900 A2
MLOps Platform in-house	AIStation (+ NVIDIA AI Enterprise)	Không (Run:ai, cnvrg.io)	HPE ML Dev Env · Ezmeral	Lenovo LiCO	Không có	ModelArts
Giá (cùng cấu hình)	Cạnh tranh (thấp hơn 10–20%)	Cao	Cao	Trung bình cao	Cạnh tranh	Trung bình
Kênh phân phối tại VN	Hạn chế (SLA, một số reseller)	Rất rộng (FPT, HPT, PMV, CMC...)	Rất rộng	Rộng	Trung bình	Bị hạn chế
Support / RMA tại VN	770 ft² · 2 (9) L1	Toàn quốc, warranty tại chỗ	Toàn quốc	Toàn quốc	Qua distributor	Hạn chế
Rủi ro địa chính trị (Entity List)	Có (từ 03/2023)	Không	Không	Không	Không	Có
Chuẩn OCP / Open Accelerator	Có, mạnh	Có	Có	Có	Dẫn đầu	Có
Liquid Cooling	D2C · Immersion	Có	Có (Cray)	Có (Neptune)	Rất mạnh	Có
Certified H100 / H200 / B200	Có (qua kênh được phép)	Có	Có	Có	Có	Không (H100 cấm)

Định vị 8 chiều – chấm điểm 1-5

Điểm càng cao càng tốt; 5 = tối đa. Cột "Tổng /40" là điểm tổng hợp trên 8 chiều.

CHIỀU ĐÁNH GIÁ	INSPUR	DELL	HPE	LENOVO	SUPERMICRO	HUAWEI
Hiệu năng AI Training	5	5	4	4	5	4
Danh mục sản phẩm	5	5	4	4	4	4
MLOps Platform	5	3	4	4	2	4
Giá cả (thấp = tốt)	5	3	3	4	5	4
Brand VN	2	5	5	4	3	4
Kênh phân phối VN	2	5	5	4	3	3
Support VN	2	5	5	4	3	3
Đảm bảo địa chính trị	2	5	5	4	5	2
Tổng /40	28	36	35	32	30	28

Chốt định vị

Inspur **tối đa hoá 4 chiều "sản phẩm"** (Training, danh mục, MLOps, giá) – nhưng thua ở 4 chiều "vận hành thị trường" (brand, kênh, support, địa chính trị) tại VN.

Vai trò QD.TEK

Chính là **lấp 4 chiều còn thiếu** – brand awareness, kênh phân phối, support tại chỗ và tuân thủ pháp lý – nâng "tổng thực chiến" từ 28 lên tiệm cận 36.



Phần 09 · Part Nine

09

Khả năng phân phối & Go-to-Market

Phạm vi phần này

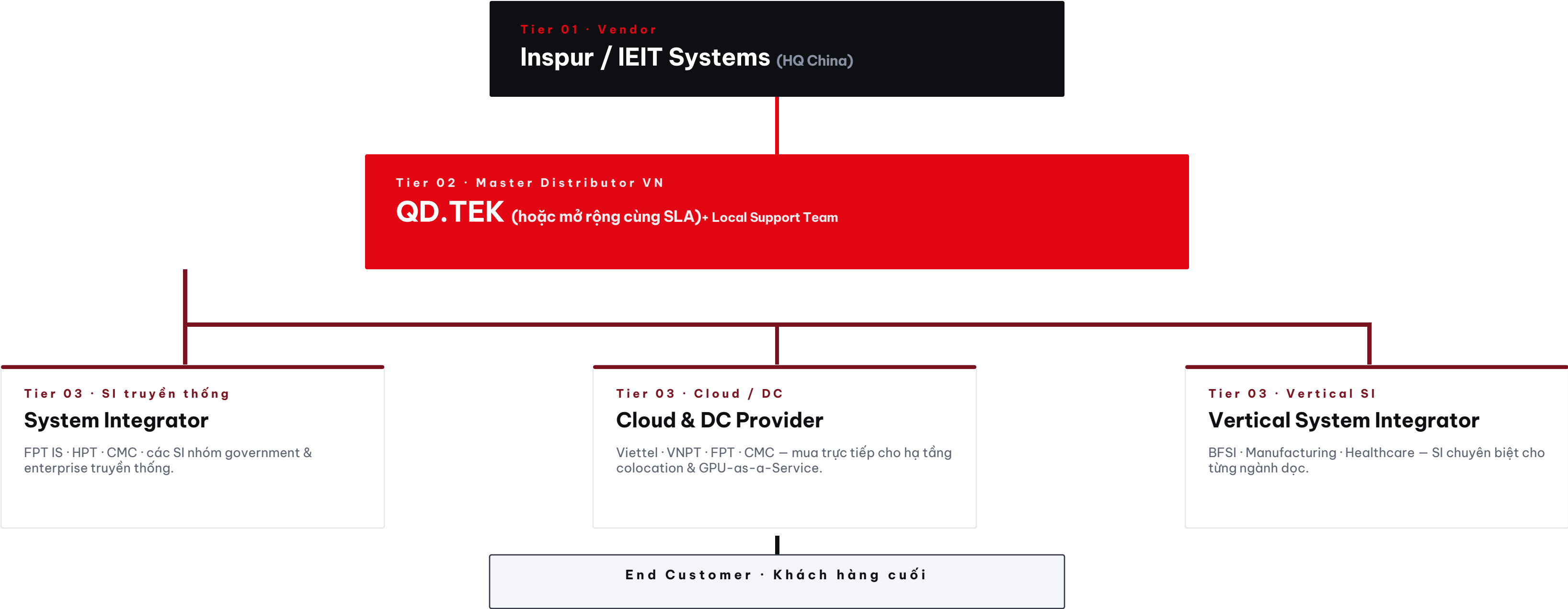
Từ hiện trạng phân phối đến KPI năm đầu tiên.

Phần này chuyển mọi phân tích ở Phần 1–8 thành hành động: mô hình phân phối 3 tầng, chiến lược giá thâm nhập, phân khúc P0/P1/P2, chiến lược marketing, và KPI cụ thể cho năm đầu tiên.

- 9.2 Mô hình phân phối 3 tầng
- 9.3 Chiến lược Go-to-Market – Pricing & Marketing
- 9.4 Phân khúc P0 · P1 · P2 (12–24 tháng đầu)
- 9.5 KPI năm đầu tiên

Mô hình phân phối 3 tầng

Cấu trúc kênh từ hãng tới khách hàng cuối, thông qua Master Distributor và 3 nhóm đối tác chuyên biệt.



Chiến lược Pricing & Marketing

9.3.2 · Chiến lược giá

Chiến thuật 01 · Penetration Pricing

Thấp hơn Dell/HPE

15–20%

Áp dụng trong 12–18 tháng đầu – chấp nhận margin thấp để chiếm reference customer P0.

Chiến thuật 02 · Volume Discount

Chiết khấu bậc thang cho dự án Data Center lớn (>50 server).

Bậc chiết khấu công bố minh bạch trong RFP để tăng khả năng thắng thầu ở Viettel/VNPT/FPT/CMC.

Chiến thuật 03 · AIStation SaaS/Subscription

Bán AIStation theo subscription cho khách hàng cần OPEX thay CAPEX.

Phù hợp fintech & SME – mở rộng phân khúc mà mô hình CAPEX truyền thống không tiếp cận được.

9.3.4 · Chiến lược Marketing & Brand

Sự kiện

Tài trợ & tham gia Vietnam AI Summit, Vietnam Data Center Summit

Tổ chức **Inspur AI Innovation Day Vietnam** hằng năm – flagship event của QD.TEK

Nội dung

Bản địa hoá toàn bộ tài liệu kỹ thuật sang tiếng Việt

Case study bản địa (khi có 1–2 khách hàng lớn tại VN)

Reference architecture bằng tiếng Việt cho từng phân khúc

PR

Truyền thông qua ICTNews, VnExpress, Nhịp Cầu Đầu Tư

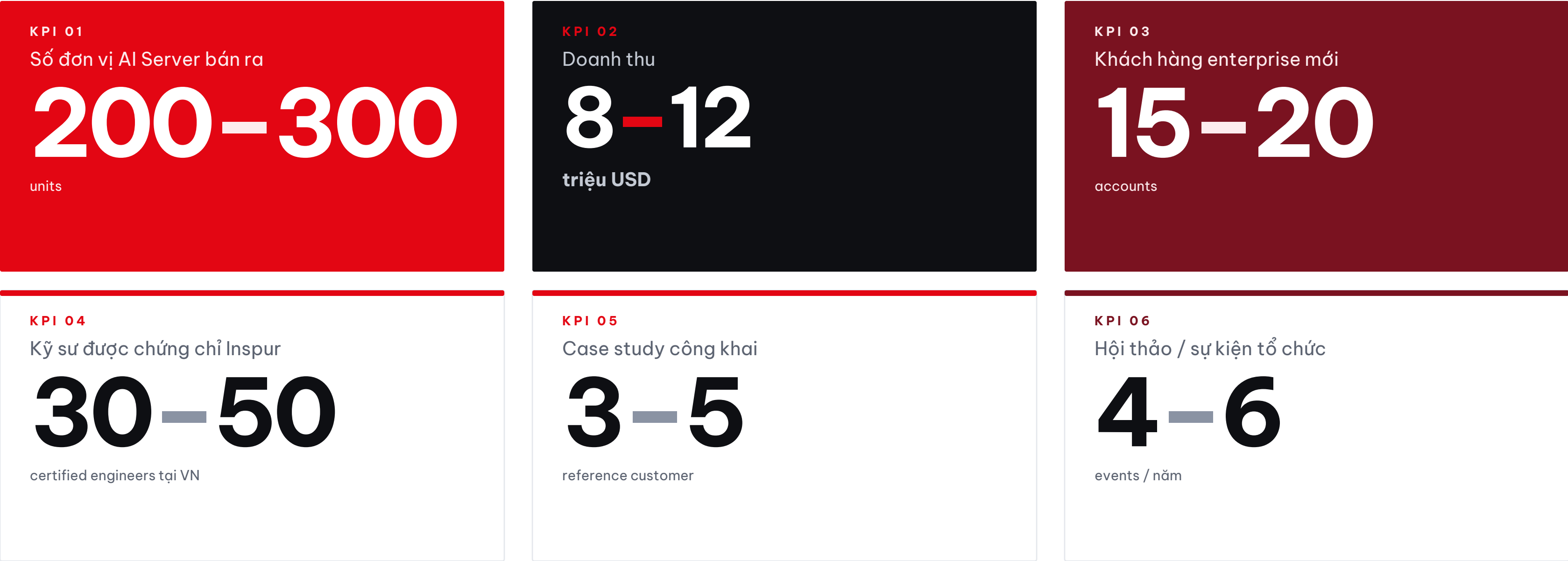
Nhấn mạnh vị thế **#1 AI Server toàn cầu**

Phân khúc P0 · P1 · P2 (12–24 tháng đầu)

Ưu tiên P0	Phân khúc Cloud & DC Provider Viettel IDC · VNPT · CMC · FPT	Lý do ưu tiên Volume lớn, chu kỳ đầu tư cao, lợi ích hiệu năng-giá thể hiện rõ ràng nhất – chỉ cần thắng 2–3 khách hàng là đủ đạt KPI năm đầu.
Ưu tiên P1	Phân khúc Fintech & Ngân hàng số TCB · MB · VPBank · Momo · ZaloPay ...	Lý do ưu tiên Nhu cầu AI cao (credit scoring, fraud detection), sẵn sàng đầu tư, ít nhạy cảm địa chính trị.
Ưu tiên P1	Phân khúc Doanh nghiệp sản xuất Vinfast · Samsung Vina · LG ...	Lý do ưu tiên Nhu cầu Computer Vision (defect detection), predictive maintenance – dễ triển khai bundle inference cluster + AIStation.
Ưu tiên P2	Phân khúc Đại học & Viện nghiên cứu VinAI · VNU · HUST ...	Lý do ưu tiên Xây thương hiệu, tạo talent pipeline; volume nhỏ nhưng giá trị dài hạn về mặt reference & nhân sự.

KPI đề xuất – năm đầu tiên (12 tháng)

Sáu chỉ số định lượng – đo hàng quý, tổng kết cuối Q4.



Nguồn: Đề xuất khung KPI của QD.TEK dựa trên phân tích thị trường tại Phần 7 và mô hình phân phối tại 9.2.



Phần 10 · Part Ten

10

Kết luận & Khuyến nghị

Phạm vi phần này

Chốt lại thông điệp và roadmap 24 tháng.

Ba slide chốt lại toàn bộ tài liệu: kết luận về vị thế giải pháp, 7 khuyến nghị chiến lược cụ thể cho QD.TEK, và roadmap triển khai 24 tháng chia thành 4 giai đoạn.

- 10.1 Kết luận tổng thể
- 10.2 7 khuyến nghị chiến lược
- 10.3 Roadmap 24 tháng – 4 phase

Kết luận – Ba thông điệp cốt lõi

01

Về giải pháp

Inspur là giải pháp **Hardware + Platform mạnh nhất** cho AI hiện nay với giá cạnh tranh.

#1 AI Server toàn cầu, danh mục 2U → 6U, AIStation MLOps hoàn chỉnh – không đối thủ nào tại VN có combo này.

02

Về thị trường

Việt Nam đang ở **chu kỳ đầu tư AI Server 2025–2030** – cửa sổ vàng để đặt reference.

Thị trường AI DC lên 858 triệu USD vào 2035 (18,6× so với 2020), 41 DC · 221 MW đang mở rộng – Inspur đủ điều kiện chiếm 5–10% AI Server VN nếu có kênh phân phối đúng.

03

Về vai trò QD.TEK

QD.TEK là mảnh ghép bù đắp **4 chiều "vận hành thị trường"** mà Inspur còn thiếu tại VN.

Brand awareness · Kênh phân phối · Support tại chỗ · Đảm bảo pháp lý – QD.TEK biến điểm 28/40 của Inspur thành 34–36/40 khi khách hàng ra quyết định.

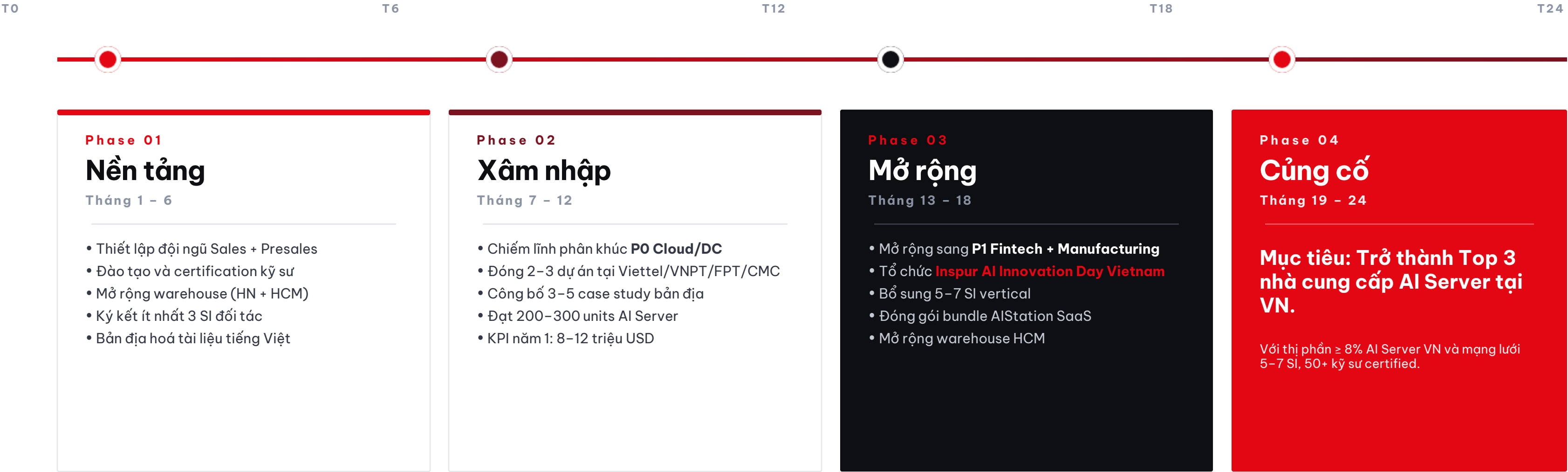
7 khuyến nghị chiến lược tổng thể

Đầu ra hành động sau toàn bộ phân tích ở Phần 1–9 – checklist thực thi cho BOD, Sales và Presales của QD.TEK.

01 Đẩy mạnh Brand Awareness Qua chuỗi sự kiện, hội thảo, PR trong 12–18 tháng đầu – mục tiêu Inspur/IEIT Systems được nhận biết ngang Dell/HPE trong khối AI/DC.	02 Bản địa hoá tài liệu Toàn bộ tài liệu kỹ thuật, dịch vụ hỗ trợ, KB, portal chuyển sang tiếng Việt – điều kiện tiên quyết để Presales địa phương hiệu quả.
03 Mở rộng kênh phân phối Không dừng ở 1 master distributor – kết nạp thêm 2–3 SI vertical chuyên biệt (BFSI, Manufacturing, Healthcare).	04 Certified Engineer Program Đầu tư đào tạo & certification – mục tiêu 30–50 kỹ sư được chứng chỉ trong năm đầu, tạo hệ sinh thái kỹ thuật viên vững chắc.
05 Bundle Hardware + AIStation Đây là điểm khác biệt cốt lõi vs Dell / HPE / Supermicro – luôn bán kèm, không tách riêng phần cứng.	06 Tập trung 4 phân khúc P0–P1 Cloud/DC Provider · Fintech · Manufacturing · Academic – không phân tán nguồn lực sang khối chính phủ trong 12 tháng đầu.
07 Minh bạch về cơ cấu IEIT Systems Chủ động truyền thông về việc mảng server đã tái cấu trúc thành thương hiệu IEIT Systems cho thị trường quốc tế từ 2023 – giảm e ngại địa chính trị và mở đường cho các khách hàng có ràng buộc tuân thủ.	

Roadmap triển khai 24 tháng

Chia làm 4 phase liên tiếp – mỗi phase có trọng tâm riêng, tránh phân tán nguồn lực.





Phần 11 · Part Eleven

11

Phụ lục & nguồn tham khảo

Phạm vi phần này

Glossary, nguồn và thông tin liên hệ.

Hai slide cuối: bảng thuật ngữ để đội Sales/Presales dễ tra cứu, danh sách nguồn tham khảo cho từng số liệu, và thông tin liên hệ đầu mối phía QD.TEK.

- 11.1 Bảng thuật ngữ (9 mục)
- 11.2 Nguồn tham khảo
- 11.3 Đầu mối liên hệ QD.TEK

Thuật ngữ & nguồn tham khảo

11.1 · Thuật ngữ (9 mục)

THUẬT NGỮ	GIẢI THÍCH
AIStation	Nền tảng MLOps / AI Development của Inspur
MIG	Multi-Instance GPU – công nghệ NVIDIA chia GPU thành nhiều instance độc lập
NVLink / NVSwitch	Interconnect tốc độ cao của NVIDIA cho multi-GPU
OAM	OCP Accelerator Module – chuẩn mở của Open Compute Project
DDP	Distributed Data Parallel – kỹ thuật training song song
HPC	High Performance Computing
BeeGFS / Lustre	Distributed File System dùng trong HPC
IEIT Systems	Thương hiệu quốc tế của mảng server Inspur từ 2023
Entity List	Danh sách kiểm soát xuất khẩu của Bộ Thương mại Hoa Kỳ (US BIS)

11.2 · Nguồn tham khảo

A · Tài liệu chính thức Inspur / IEIT Systems

- AI Server.pdf (tài liệu nội bộ)
 - Inspur Group.pdf (tài liệu nội bộ)
 - Inspur NF5488A5 Datasheet
 - Inspur NF5468A5 White Paper
 - Inspur NF5448A6 White Paper
 - AIStation Product Brief
- <https://www.inspur.com>

B · Báo cáo thị trường

- Gartner Server Market Tracker (2022)
- IDC Worldwide Quarterly Server Tracker
- Mordor Intelligence – Vietnam Data Center Market
- ResearchAndMarkets – Vietnam DC Market Investment Analysis
- DCMarketInsights – Vietnam AI DC Market Size 2025–2035
- ABI Research – AI Server Market Size & Vendor Shares

C · Chính sách

- US Department of Commerce · BIS Entity List
- Website chính thức: inspur.com.vn · sla.vn



Đầu mối liên hệ

Phong Dao

Đầu mối chính · Sales / Business Development

Email **phong.dao@qdtek.vn**

Đơn vị **CÔNG TY CỔ PHẦN PHÂN PHỐI
CÔNG NGHỆ QUANG DŨNG**

Thương hiệu **QD.TEK**

Địa chỉ trụ sở

Tầng 20, Tòa nhà Centec 72-74
Nguyễn Thị Minh Khai, Phường Xuân Hoà
TP. Hồ Chí Minh, Việt Nam

Cảm ơn quý vị · Thank you

Cùng xây dựng hạ tầng AI Việt Nam.

QD.TEK — Master Distributor tại Việt Nam cho danh mục Inspur AI Server & nền tảng AIStation.

Bước tiếp theo

- 01 Đặt lịch presales briefing cho khách hàng chiến lược (PO Cloud/DC).
- 02 Đăng ký chương trình đào tạo certified engineer Inspur / AIStation.
- 03 Yêu cầu cấu hình BOM & báo giá cho use case cụ thể.