



Supermicro & Intel Data Center GPU Visual Cloud solution

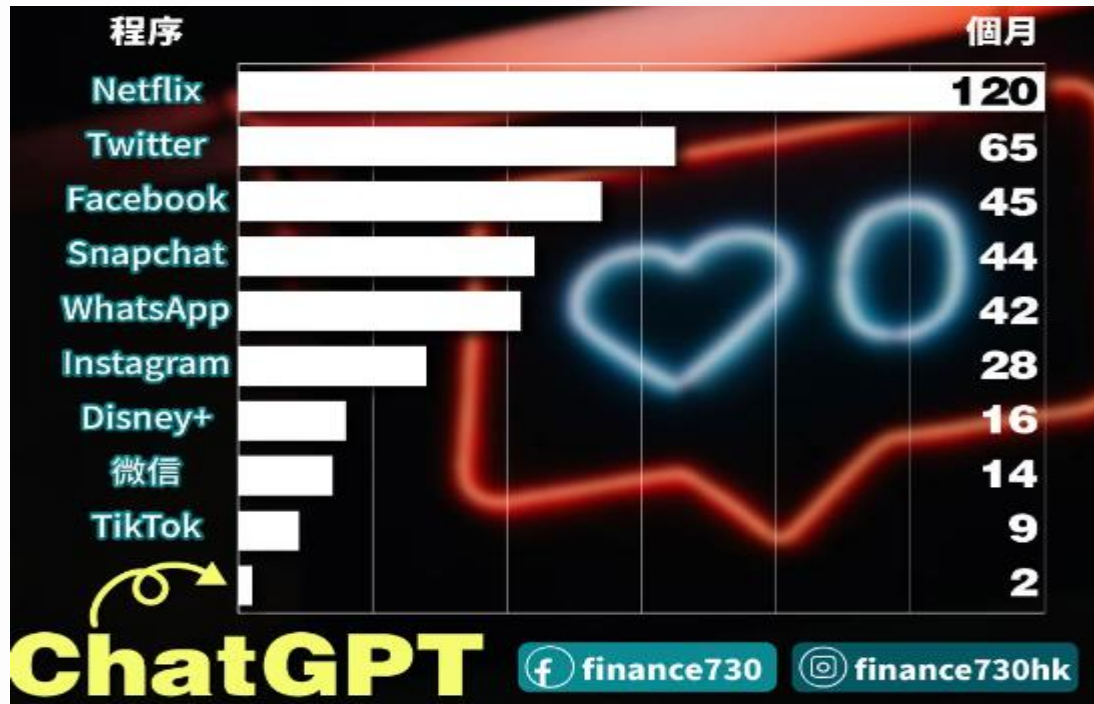
Alex Lee

Director, Technology Enablement

March 23, 2023



How much time to reach 100M members for APP?



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021


ChatGPT，全稱聊天生成型預訓練變換模型（**Chat Generative Pre-trained Transformer^[2]**），是OpenAI開發的人工智慧聊天機器人程式，於2022年11月推出。該程式使用基於GPT-3.5架構的大型語言模型並以強化學習訓練。

ChatGPT可寫出相似真人的文章，並在許多知識領域給出詳細和清晰的回答而迅速獲得關注，證明了從前認為AI不會取代的知識型工作它也足以勝任，對金融與白領人力市場的衝擊相當大。

 早安, Chat

 早安，請問有什麼我可以幫助您的嗎？

 你認為將來AI可以取代人類嗎？

 AI已經在許多領域中取得了顯著的進展，並且在某些任務上表現出色，但是AI在某些方面仍然無法取代人類。因為人類有著獨特的思考方式和創造力，這些是AI目前還無法完全模擬的。

例如，在創意領域，如音樂、文學和藝術等，AI目前還無法像人類一樣創作出有創意的作品。人類還擁有情感和同理心等特質，這些特質是AI缺乏的，這些特質是人類在互動和合作中至關重要的因素。

但是，隨著AI技術的不斷發展，我們可能會看到更多的工作被自動化和機器人化。這就要求人們不斷地更新自己的技能和知識，以應對新的挑戰和機會。同時，我們也需要關注和管理AI的發展，以確保其符合倫理和法律要求，並對人類帶來積極的影響。



那人類有辦法消滅AI嗎？



由於AI已經成為現代科技發展的重要組成部分，它的影響和應用越來越廣泛，因此要完全消滅AI幾乎是不可能的。即使人類試圖切斷電力和網絡等基礎設施，AI仍然可能在未來出現。

此外，即使AI出現了危險或不當使用的情況，人類也有責任和能力管理和控制AI的應用，以確保其符合倫理和法律要求。這需要社會各方面的合作，包括政府、產業、學術界和公眾等，共同確保AI的發展符合人類的利益和需要。

因此，我們應該關注AI的發展和應用，並將其看作是一個需要管理和引導的重要技術，以實現其對人類社會的積極貢獻。



哈哈,Chat我開玩笑的,別放在心上.



好的，沒問題。如果您有任何問題或需要幫助，隨時都可以向我提出。



請問人工智慧 (AI) 可以替代那些工作？



人工智慧 (AI) 已經開始替代一些簡單、重複性高的工作，特別是那些可以被程式化的工作。以下是一些例子：

1. 生產線上的機器人：AI可以透過機器學習，學習如何處理和操作生產線上的機器，從而提高生產效率和品質。
2. 客服中心的自動化客服：AI可以通過自然語言處理來了解客戶問題並提供解決方案，從而減少公司需要聘用的客服人員。
3. 金融行業的自動化：AI可以幫助銀行和保險公司處理簡單的交易、風險評估和索賠處理，減少需要人工處理的流程。
4. 預測性維護：AI可以使用機器學習和資料分析來預測設備和機器的故障，減少因機器故障而產生的停機時間和維修成本。

需要注意的是，AI並不能完全替代所有工作，許多工作需要人類的專業技能、判斷力和創造力，而這些都是AI目前還無法取代的。AI能夠幫助人們更好地處理和管理工作，但仍需要人類的參與和監督。



請問人工智慧 (AI) 可以創造那些工作？



人工智慧 (AI) 同時也創造了一些新的工作機會，特別是需要專業技能和知識的工作。以下是一些例子：

1. AI開發人員：AI開發人員可以開發和維護AI系統和應用程序，並為企業和組織提供AI解決方案。
2. AI數據分析師：AI數據分析師可以使用AI工具和技術來分析和處理大數據，並從中獲得有價值的信息和見解。
3. AI倫理師：隨著AI技術的發展和應用，AI倫理師可以幫助企業和政府制定AI的道德和法律規範，保護公眾的利益和權益。
4. 語言翻譯：AI可以幫助語言翻譯師更快地處理和翻譯大量文本和語音，並提供更準確的翻譯結果。
5. 人工智能監督：AI需要被監督和管理，以確保其運作正常並符合道德和法律標準。因此，AI監督人員可以負責監控和管理AI系統，以確保其安全和合規。

需要注意的是，AI創造的新工作機會需要具備相應的專業技能和知識，因此人們需要學習和不斷更新自己的技能，以適應未來的工作市場。




GPUs (圖形處理器) 可用於處理各種工作負載，包括以下幾種：

1. 圖形處理：GPU最初是用於加速電腦圖形的處理，包括遊戲、電影和動畫等，通過渲染圖像和圖形效果來提高視覺效果和運行速度。
2. 科學計算：GPU也可以用於加速科學計算，如物理模擬、氣象學、生物學、人工智慧和深度學習等。
3. 虛擬化：GPU還可以用於虛擬化環境，如虛擬桌面和遊戲串流服務等，通過加速虛擬機器的處理能力來提高用戶體驗。
4. 加密貨幣挖掘：GPU也可以用於加密貨幣挖掘，通過解決加密演算法來獲得加密貨幣獎勵。


為了更好地利用GPU的運算能力，通常會將不同類型的工作負載分配到不同的GPU上，這種做法被稱為Workloads Segmentation。通常會使用軟件工具來實現Workloads Segmentation，例如CUDA (Compute Unified Device Architecture) 和OpenCL (Open Computing Language) 等。這些工具可以讓開發人員根據不同的工作負載，分配適當的GPU資源，從而實現更高效的計算和處理。

Intel: GPUs (XPU) Workloads Segmentation

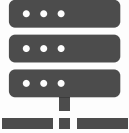


AI Training and Deep Learning

Gaudi / Gaudi 2




AI Training and Deep Learning




HPC


Ponte Vecchio (Max Series)




VDI for power (3D) workers




4K Cloud Gaming



AI visual inference




Cloud Rendering




Video Conference


Flex Series 170
150 W | 16GB
1-Slot




VDI for task workers




Android Gaming



Video Streaming



Video Transcoding



Live Streaming

Flex Series 140
75 W | 12GB
1-Slot





請問visual cloud的什麼?



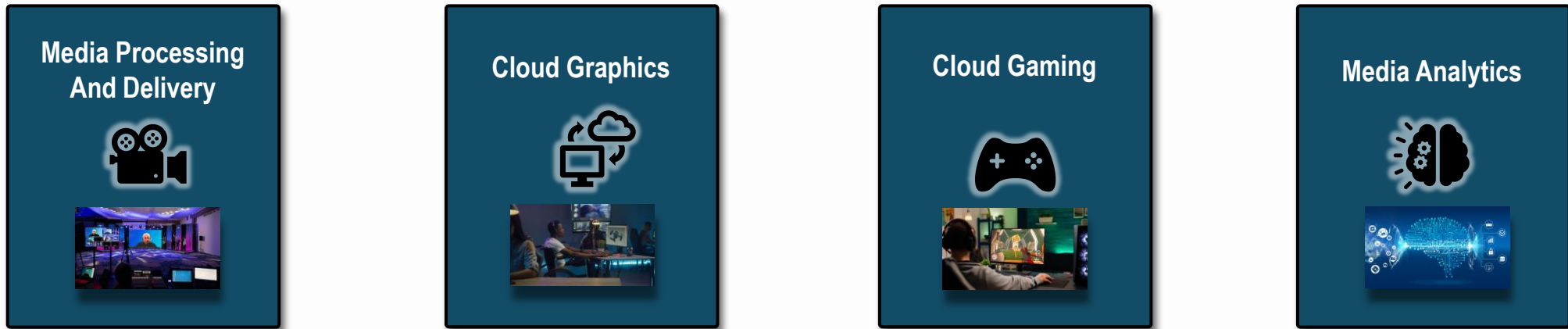
Visual Cloud是一種基於虛擬化和雲端運算技術的視頻和多媒體解決方案，它結合了虛擬化、雲端運算、網絡和多媒體技術，提供高效、靈活、可擴展的視頻和多媒體處理、編碼、傳輸和存儲。它可以在雲端環境下，將傳統的硬件和軟件分離，將多個視頻和多媒體處理功能整合到一個系統中，並且可以通過網絡在不同的設備上實現高品質的視頻和多媒體應用。

Visual Cloud可以用於多種場景，例如：

1. 視頻編碼和解碼：Visual Cloud可以實現高效的視頻編碼和解碼，從而提高視頻播放的質量和流暢度。
2. 視頻會議和通訊：Visual Cloud可以提供高質量的視頻會議和通訊服務，包括視頻和音頻編碼、多點接入、視頻錄製等功能。
3. 媒體分發和傳輸：Visual Cloud可以實現高效的多媒體分發和傳輸，包括媒體編碼、數據壓縮、傳輸協議等。
4. 視頻和多媒體應用：Visual Cloud可以用於多種視頻和多媒體應用，例如視頻廣告、視頻監控、視頻直播等。

總的來說，Visual Cloud是一個高效、靈活、可擴展的多媒體解決方案，可以幫助企業和組織實現更高品質的視頻和多媒體應用。

The Visual Cloud IT Landscape



Data Center



Core Network



CDN Servers



Edge Servers



Access Network



IoT Gateways



IoT Devices



Cloud

Edge

Supermicro & Intel: The Visual Cloud Solution

Open - High Density – Flexible - Low TCO

Software

Open Standards Architecture:

- Intel's Open Visual Cloud
- Standard Industry Frameworks and Libraries

Accelerator

Intel Data Center GPU Flex Series

- Hardware Accelerator
- Xe-HPG-Core Architecture

Platform

Cloud and Edge platforms

- Supermicro Systems
- Extensible Portfolio
- Supermicro's Total IT Solution

- Cost Efficiency
 - Open, royalty-free environments
 - Low bandwidth consumption
 - High density
 - Low TCO
- User Experience
 - Low Latency
 - Network and data availability

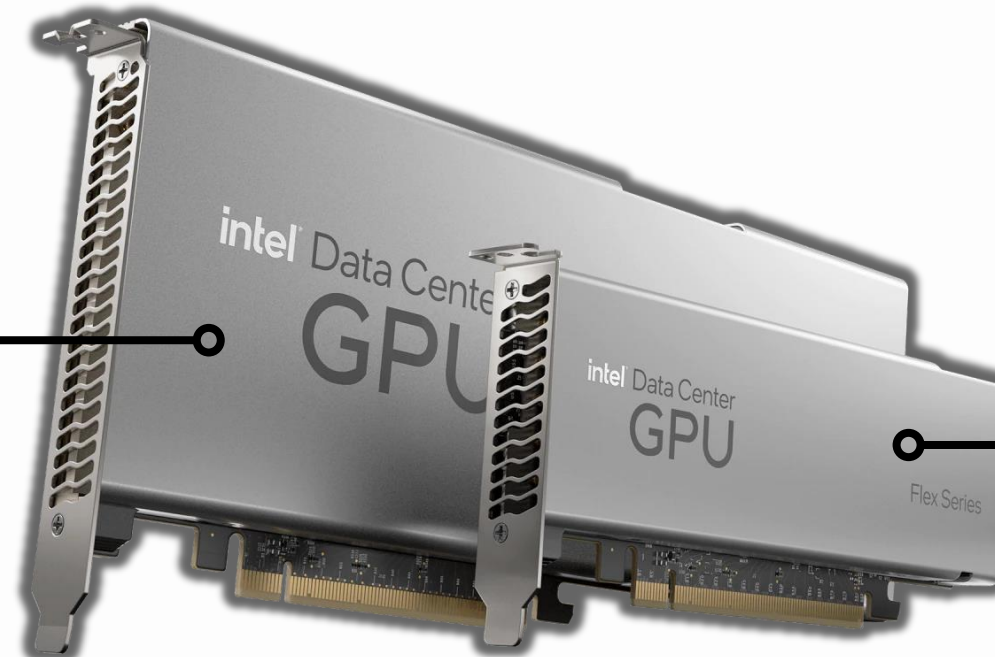
Intel Data Center GPU Flex Series



- **Supermicro is first-to-market with solutions for Intel Data Center GPU Flex Series!**
- Up to 500 TOPS* of systolic compute
- Industry-first hardware-based open source AV1 Encoder in a GPU
- Formerly known as Arctic Sound M (ATS-M)

**Intel Flex Series
170 GPU**

**Intel Flex Series
140 GPU**



Supermicro Solutions – Visual Cloud



**Habana
Gaudi/Gaudi2**

Flex Series 140

Flex Series 170

Supermicro Systems:
Optimized compute, storage, software, and services



Visual
Cloud

AI & HPC

Cloud & Virtualization

5G, Edge Computing, and IoT

Origin Server	SDN / NVF	CDN Server	Edge Server	RAN / vRAN	IoT Gateways	Devices
Data Center	Backbone	Micro Data Centers		Base Station	5G and Telco	
Cloud	Network Core	Network Edge		Access Network	On-premise edge	



**Data
Center**



**Core
Network**



**CDN
Servers**



**Edge
Servers**



**Access
Network**



**IoT
Gateways**



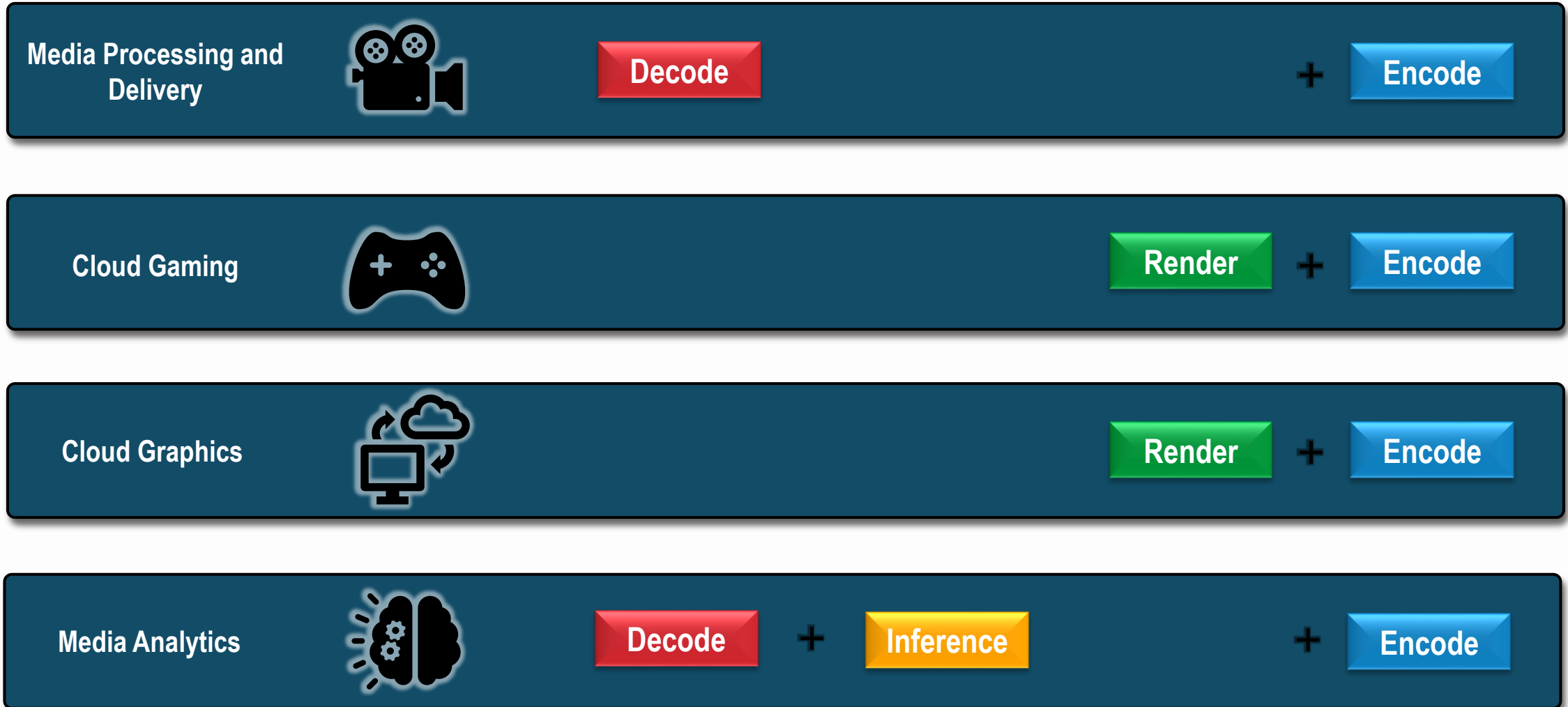
**IoT
Devices**



Cloud

Edge

Visual Cloud Workloads Pipeline (Building-Blocks-Solution)



Intel Flex Series GPU Micro-Architecture



Architecture		Intel Flex Series 170 GPU	Intel Flex Series 140 GPU	
Render Slice	Xe-Core	GPU Micro-architecture	Xe-HPG	Xe-HPG
		Xe-Core (XC)	32	16 (8/GPU)
		Xe Vector Engine	512	256 (128/GPU)
		Xe Matrix Engine	512	256 (128/GPU)
		Systolic Array (AI Inference)	2.5x	1x
		Cache (L2\$)	16M	8M (4M/GPU)
	Function rendering	Ray Tracing Units	32	16 (8/GPU)
		Sampler (Texture Sampling)	32	16 (8/GPU)
		Geometry (processing)	8	4
		Rasterization	8	4
		Depth (HIZ)	8	4
Pixel Backend (Processing)	16	8		
Media Slice (Media Engine)	Xe Media Engines	2	4 (2/GPU)	
	MXF: Multi Format Coded	2	4	
	SFC: Scaler and format conversion	2	4	
	VQE: Video Quality Engine	2	4	

Visual Cloud Workloads



- Media Analytics
- Inference
- Rendering



- Media Delivery
- Transcoding
- Streaming

Why supermicro & Intel for the Visual Cloud



Software

Open Standards Architecture:

- Intel's Open Visual Cloud
- Standard Industry Frameworks and Libraries

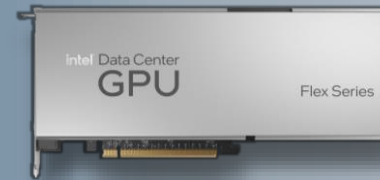


Accelerator

Intel Data Center GPU Flex Series

- Hardware Accelerator
- Xe-HPG-Core Architecture

Flex Series 170



Flex Series 140



Platform

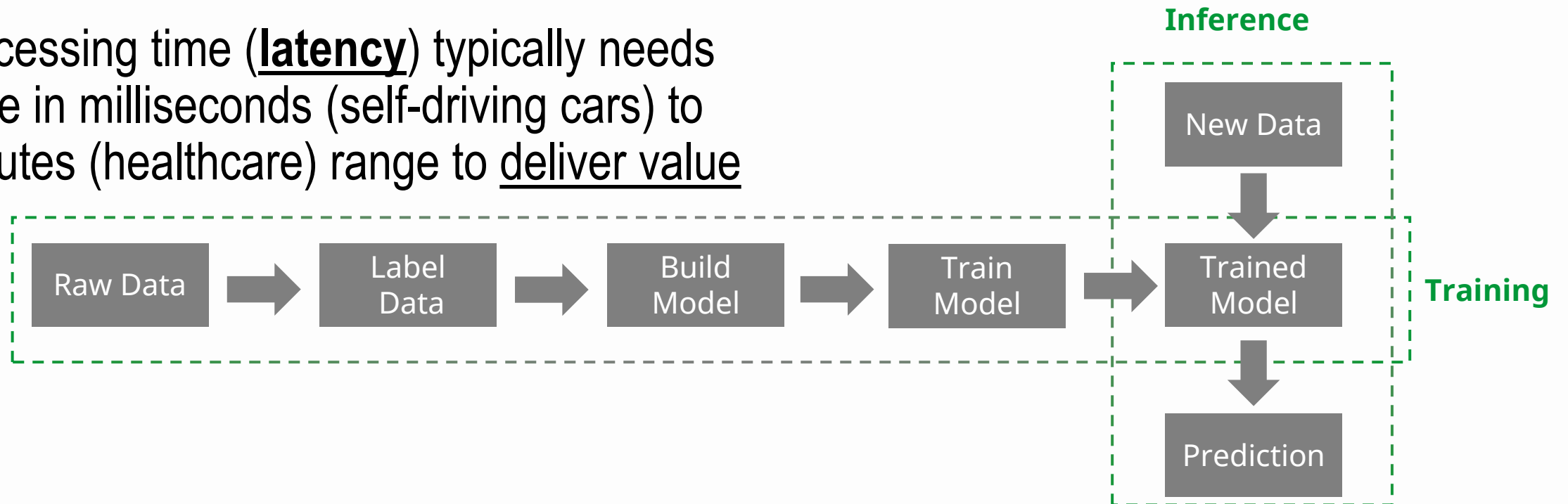
Cloud and Edge platforms

- Supermicro Systems
- Extensible Portfolio
- Supermicro's Total IT Solution

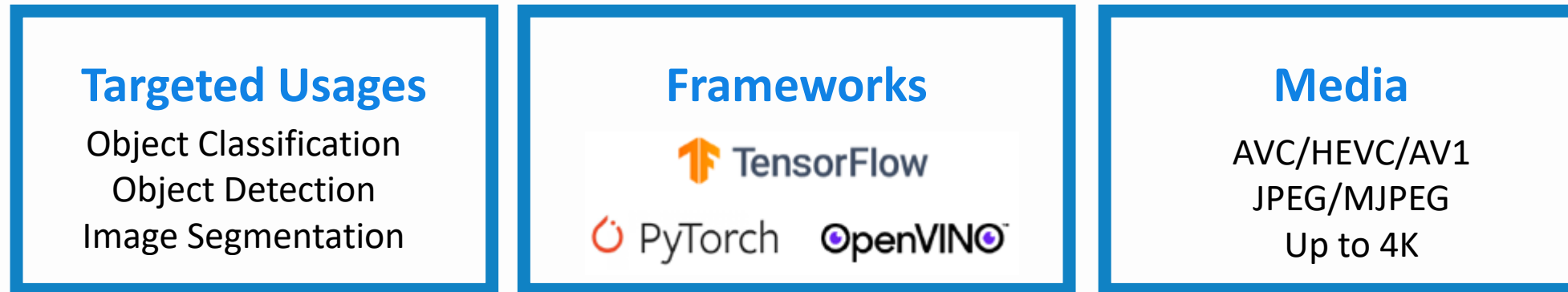


AI Edge Inferencing

- This is where the models that have been trained are put to work
- Training happens over hours, days and weeks - inferencing happens in real-time
- Processing time (latency) typically needs to be in milliseconds (self-driving cars) to minutes (healthcare) range to deliver value

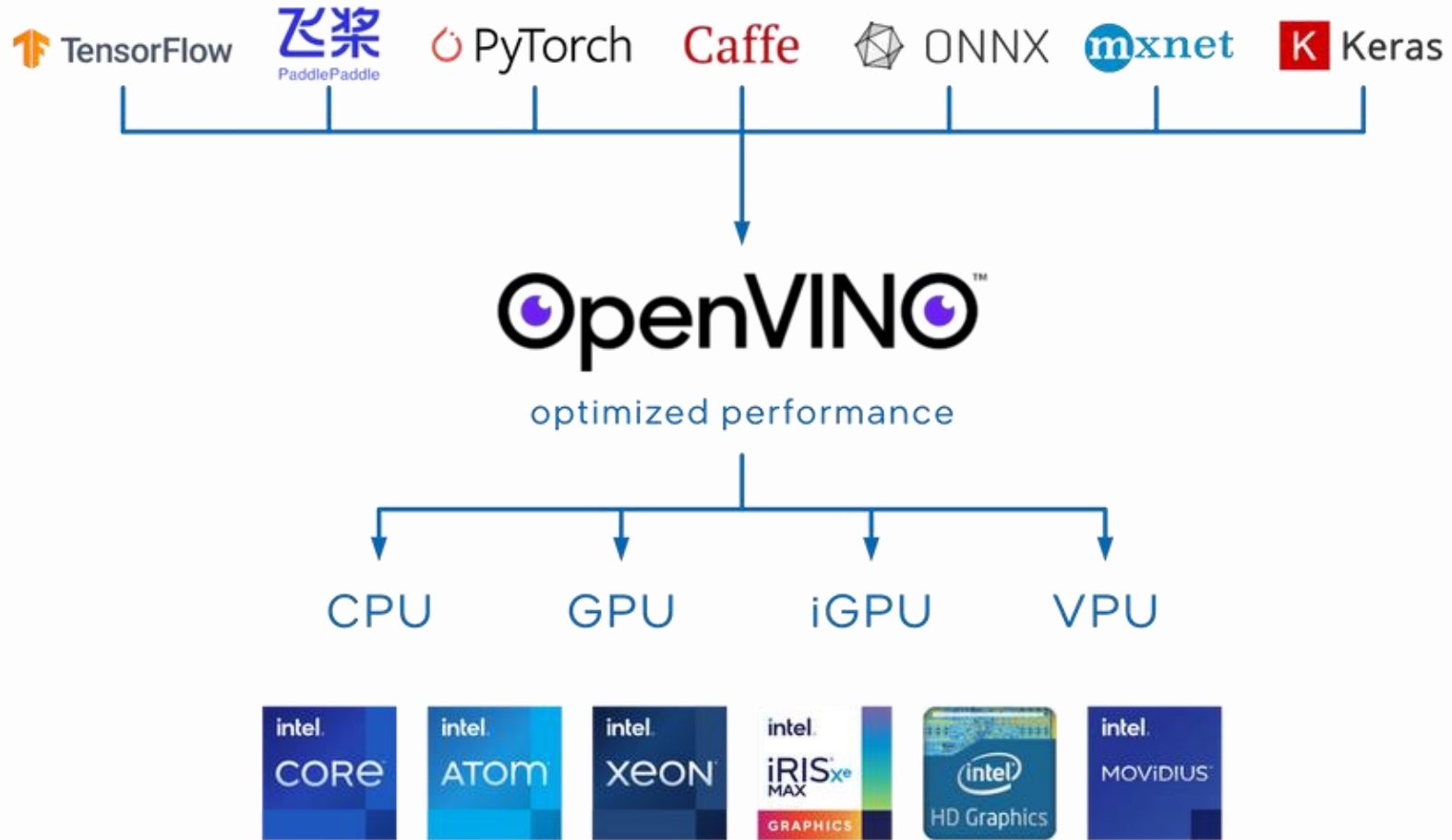


Visual Inference, Media Analytics with Intel® Data Center GPU Flex Series



...Up to 500 TOPS* of systolic compute with Intel Data Center GPU Flex Series

Intel® Distribution of OpenVINO™



Media Analytics - AI Inferencing

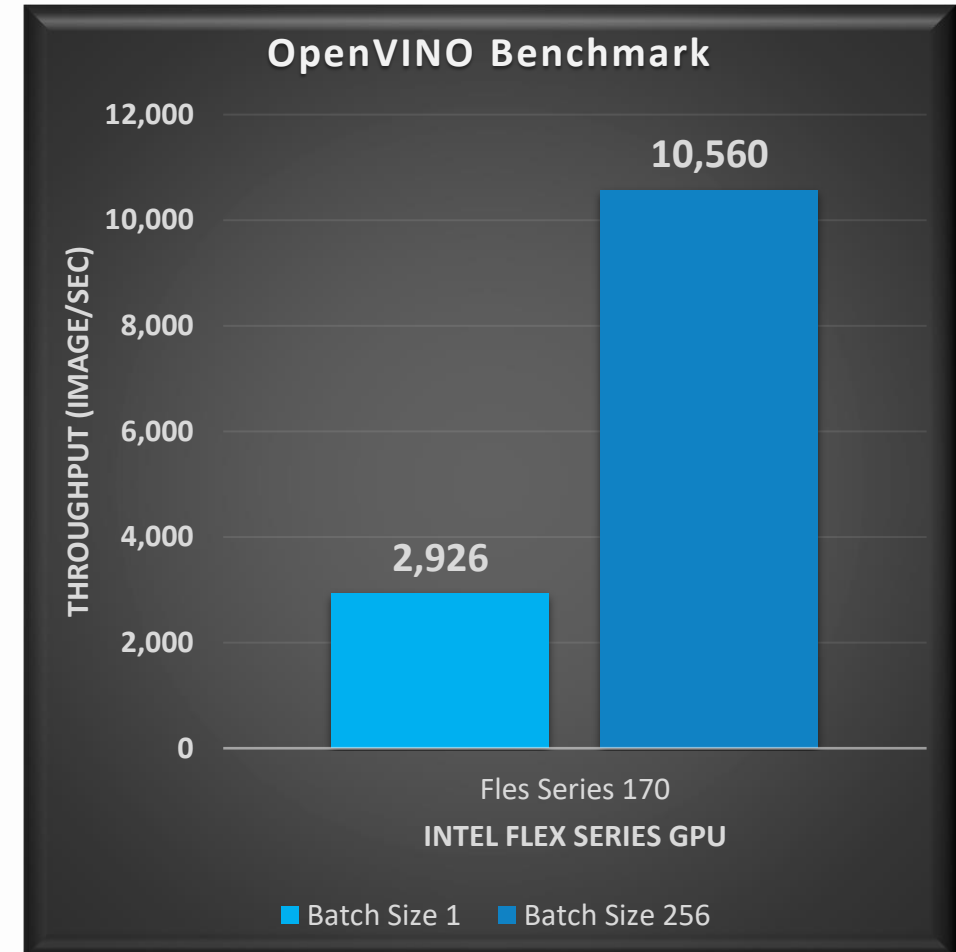
Decode

Inference

Encode



- Intel's Flex Series GPU AI Visual Inference for: Smart City, Library indexing and compliance, and AI-guided video enhancement workloads
- Flex Series GPU supports OpenVINO, an open-source toolkit for optimizing and deploying AI inference.
- OpenVINO boosts deep learning performance in:
 - Computer vision
 - Automatic speech recognition
 - Natural language processing



Why Intel Data Center GPU Flex Series for Transcoding

Open Architecture



An open, flexible, standards-based software stack together with **oneAPI**

Built-In AV1 Encode



- Royalty-free open-source **AV1 codec**
- Support for **AVC (H.264)**, **HEVC (H.265)**, and **VP9**.

5x

Media transcode throughput at half the power of the competition

Intel Flex Series 140 GPU compared to existed card HEVC 1080p60

2x

Decode throughput at half the power of the competition

Intel Flex Series 140 GPU compared to existed card across HEVC, AV1, AVC, VP91

68

(Up to)

720p30 on select game streams

Single Intel Flex Series 170 GPU1

Media Analytics - AI inference Live demo

Input text:

portrait photo headshot by mucha, sharp focus, elegant, render, octane, detailed, award winning photography, masterpiece, rim lit



Nice, as we can see the picture has quite a high definition 🍷.

Text To Image Demo

Pipeline settings
Input text: amazing watercolor painting
Seed: 42
Number of steps: 10
Strength: 0.5
Input image:



100%  5/5 [00:00<00:00, 6.85it/s]

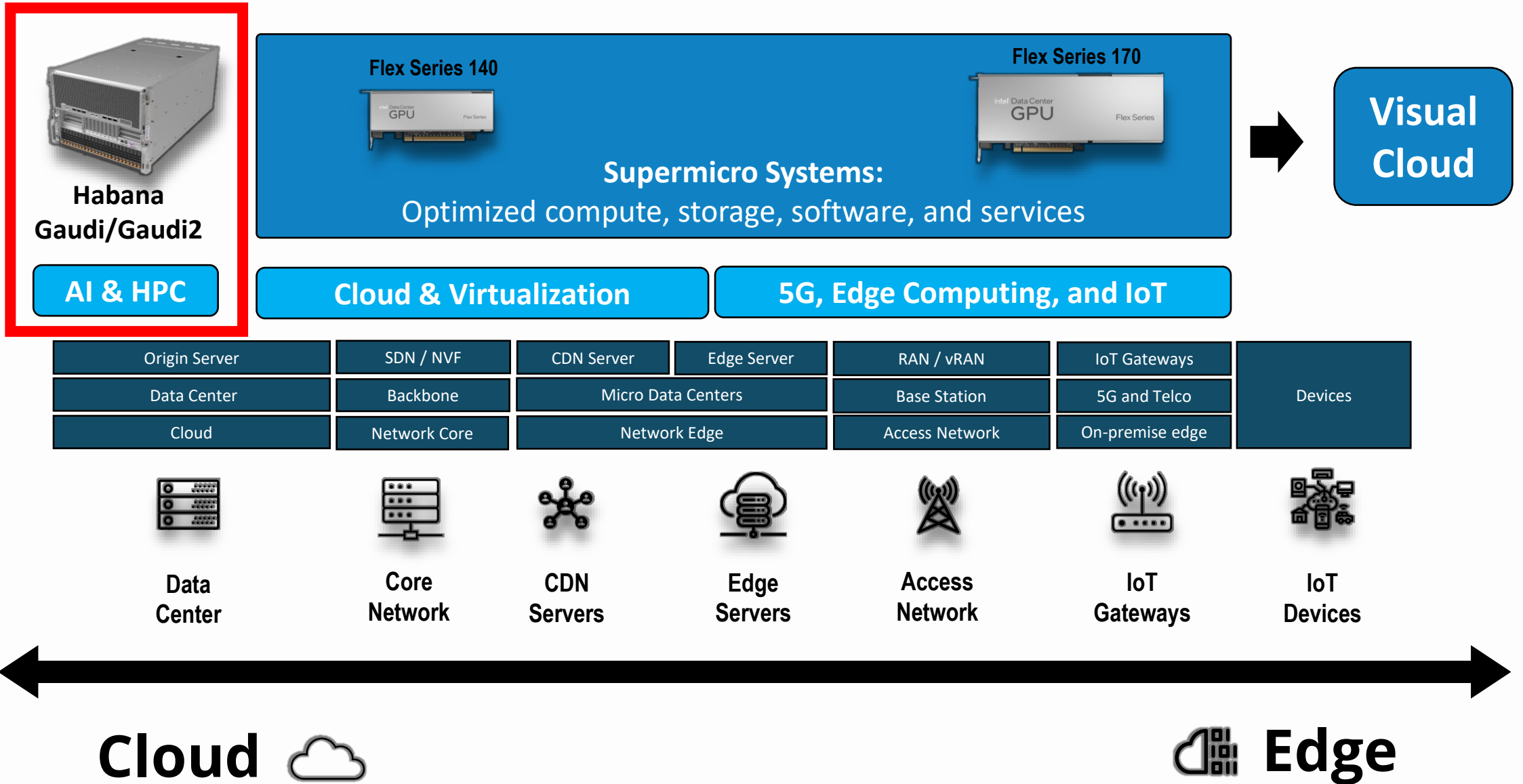
Input text:

amazing watercolor painting



Image To Image Demo

Supermicro Solutions – AI Training



Supermicro Habana Gaudi® AI Training Server

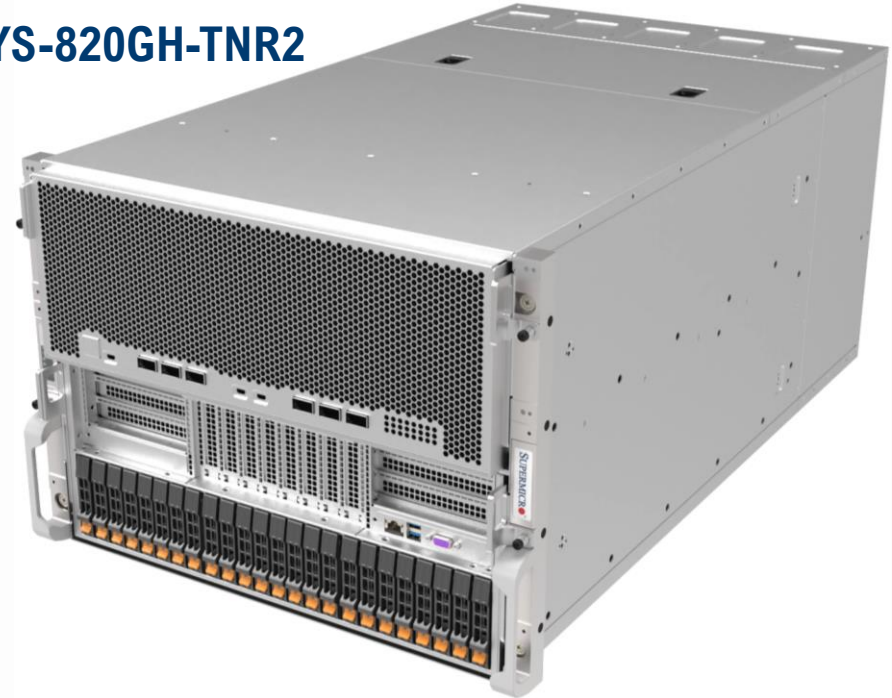
GAUDI®

SYS-420GH-TNGR

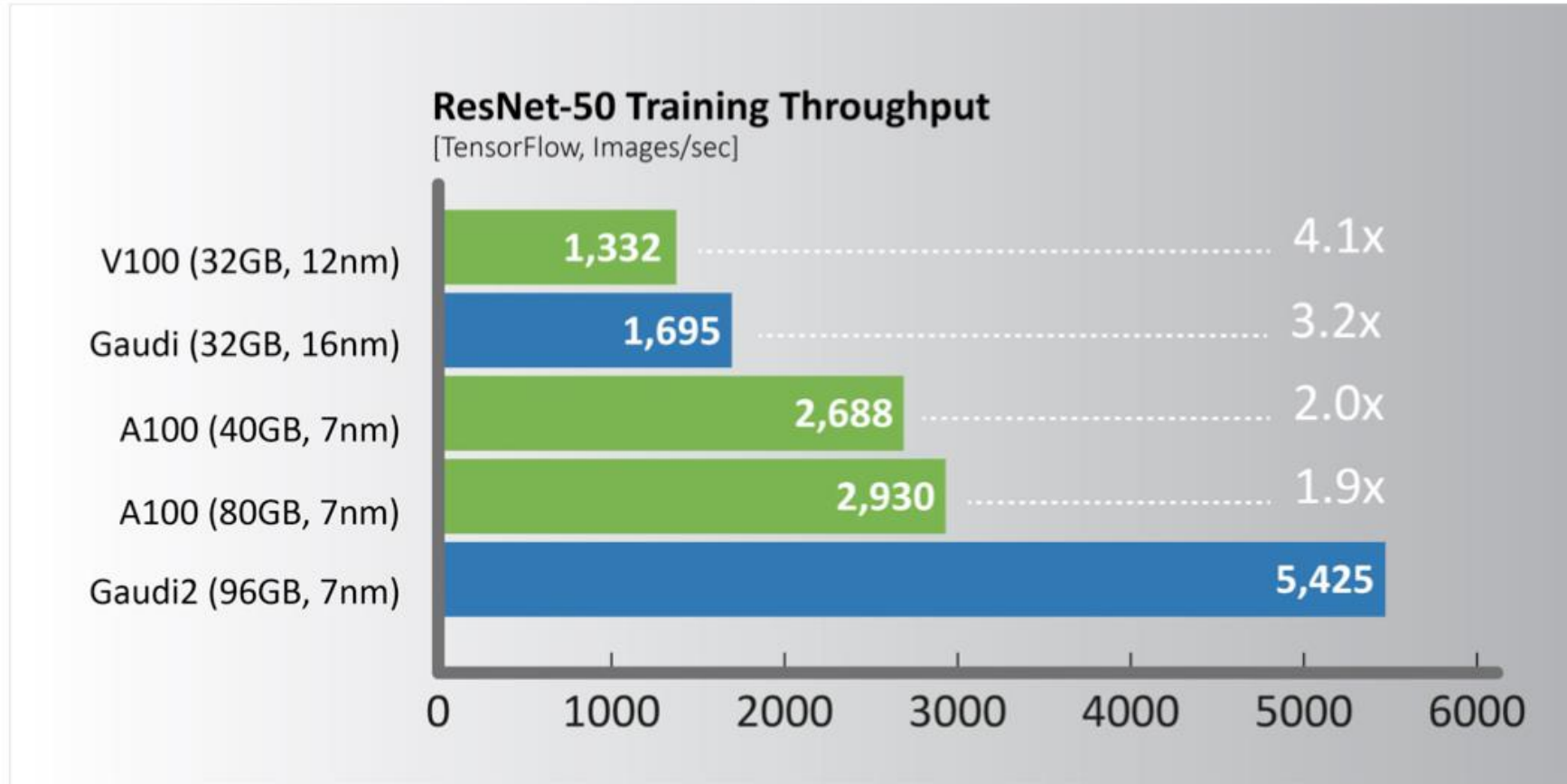


GAUDI²

SYS-820GH-TNR2



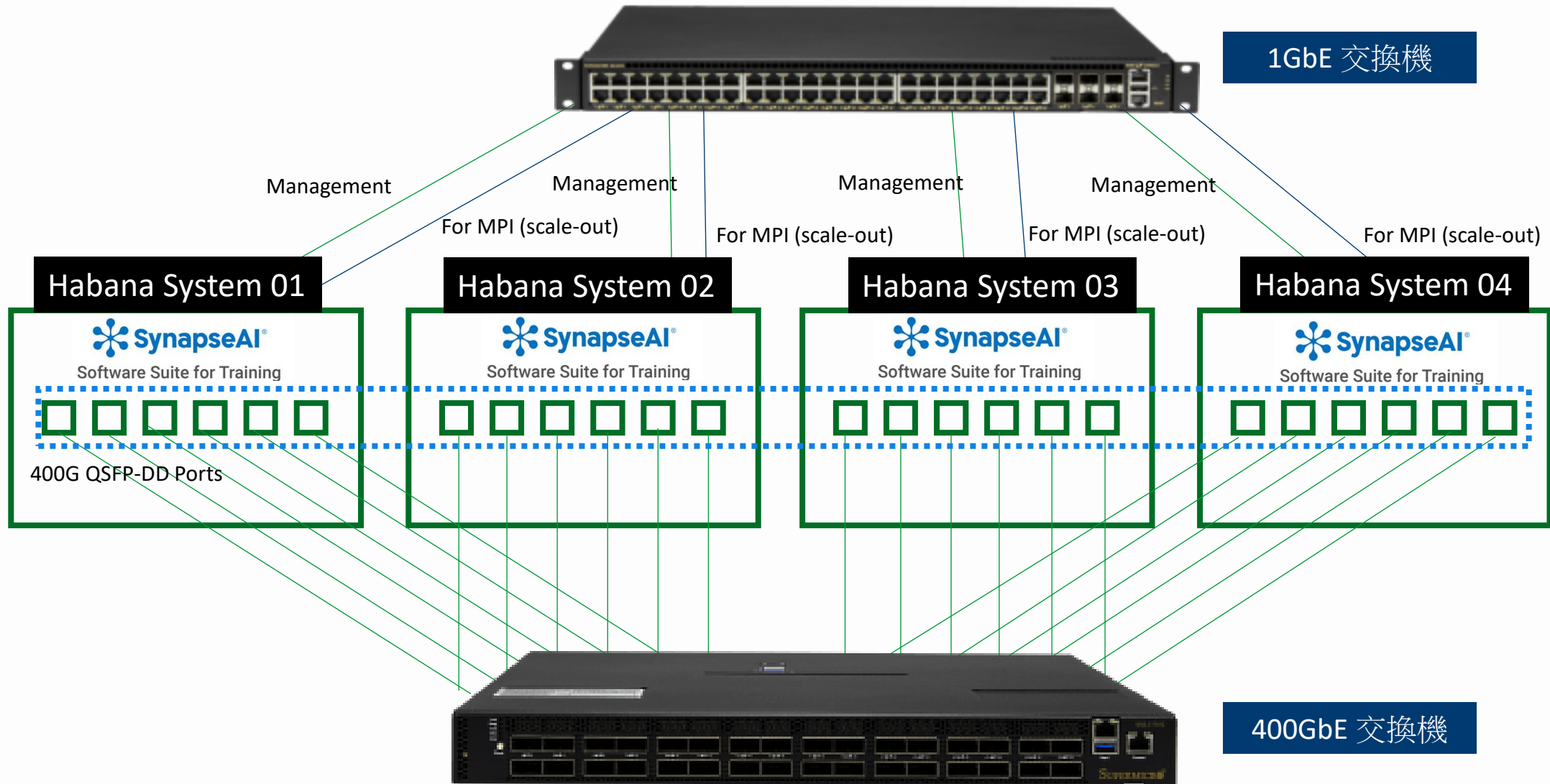
Leading AI Training Performance – Computer Vision



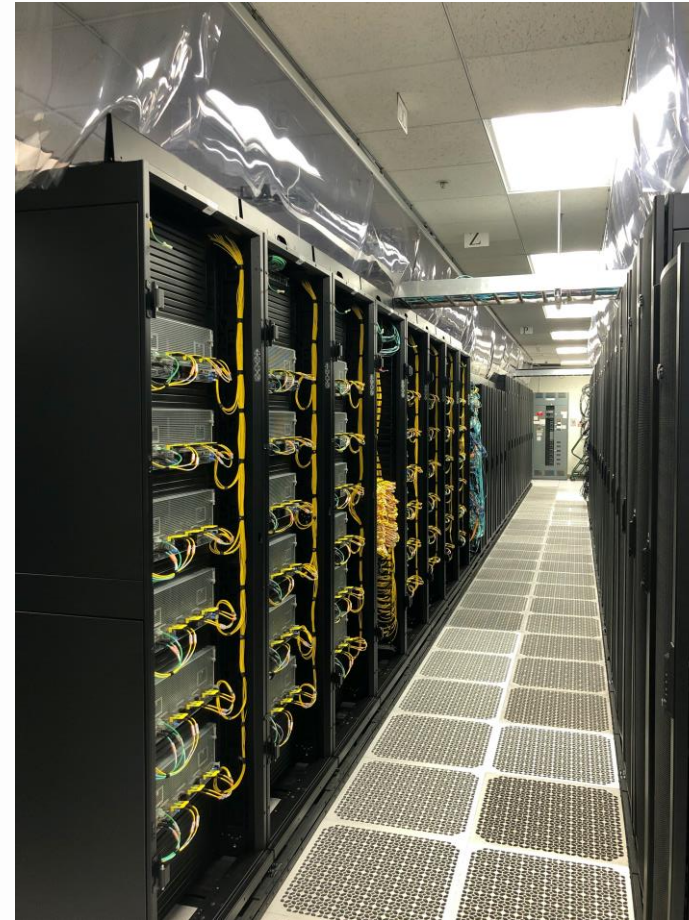
A100-80GB : Measured by Habana on Azure instance Standard_ND96amsr_A100_v4 using single A100-80GB using TF docker 21.12-tf2-py3 from NGC (optimizer=sgd, BS=256)
A100-40GB : Measured by Habana on DGX-A100 using single A100-40GB using TF docker 21.12-tf2-py3 from NGC (optimizer=sgd, BS=256)
V100-32GB : Measured by Habana on p3dn.24xlarge using single V100-32GB using TF docker 21.12-tf2-py3 from NGC (optimizer=sgd, BS=256)
Gaudi2: Measured by Habana on Gaudi2-HLS system using single Gaudi2 using SynapseAI TF docker 1.4.0-435 (BS=256)

Results may vary.

Scale-Out Topology of 4 Habana Systems



Solution Brief and Rack Level Reference Design Available



- 支持科學和工程領域的研究
 - 天文學、氣候科學、化學和粒子物理學
 - 共部署42台Gaudi AI伺服器; 總計336 顆Gaudi

Supermicro 32*400G SSE-T7032SR



Target Use Case

- Data Center ToR, Leaf, Spine applications

Key Advantages

- Data Center Optimized Silicon for Machine learning (ML) and Artificial Intelligence (AI)
- Industry Leading Latency
- Line Rate 32x400G up to 12.8T performance
- SONiC OS pre-install and support L2/L3 and Data Center friendly features
- Habana certified switch
- User friendly Toolless design on top case
- User friendly Luggage tag support at front side

*Detail Spec: [Supermicro | Products | Networking | SSE-T7032SR](#)

Hardware Specification

• Interface Ports

- 32x QSFP-DD Ports
- 1x RJ45 Serial Console
- 1x RJ45 100/1000BASE-T Management
- 1x USB Type A Storage

• Onboard CPU Engine

- Intel Quad Core CPU
- 16GB DDR4 DRAM
- 16MB SPI Flash
- 128GB SSD

• Physical and Environmental

- 1RU, Mounting Rail for 19" rack
- Dimensions: (WxDxH) 433 x 535 x 44cm
- Rear-to-Front airflow version available
- 2x Redundant, Load Sharing, Hot Swappable PSUs
- Hot Swappable 6+1 Redundant Fan Modules

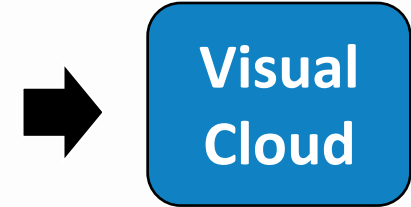
Supermicro Solutions - Management



Flex Series 140

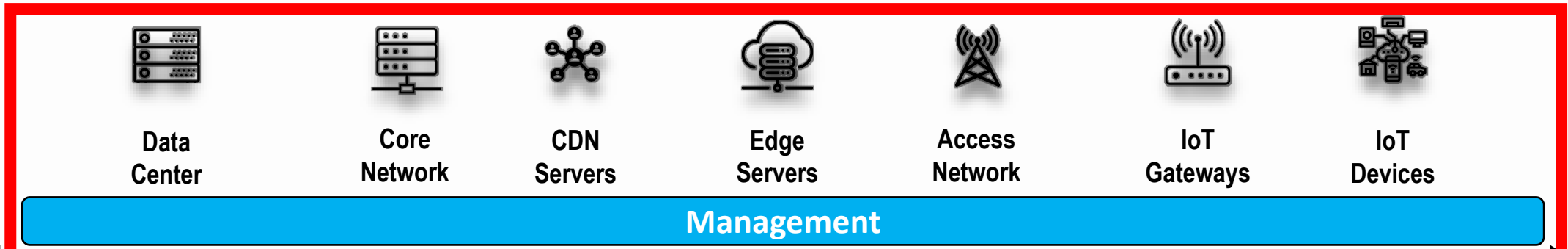
Flex Series 170

Supermicro Systems:
Optimized compute, storage, software, and services



- AI & HPC**
- Cloud & Virtualization**
- 5G, Edge Computing, and IoT**

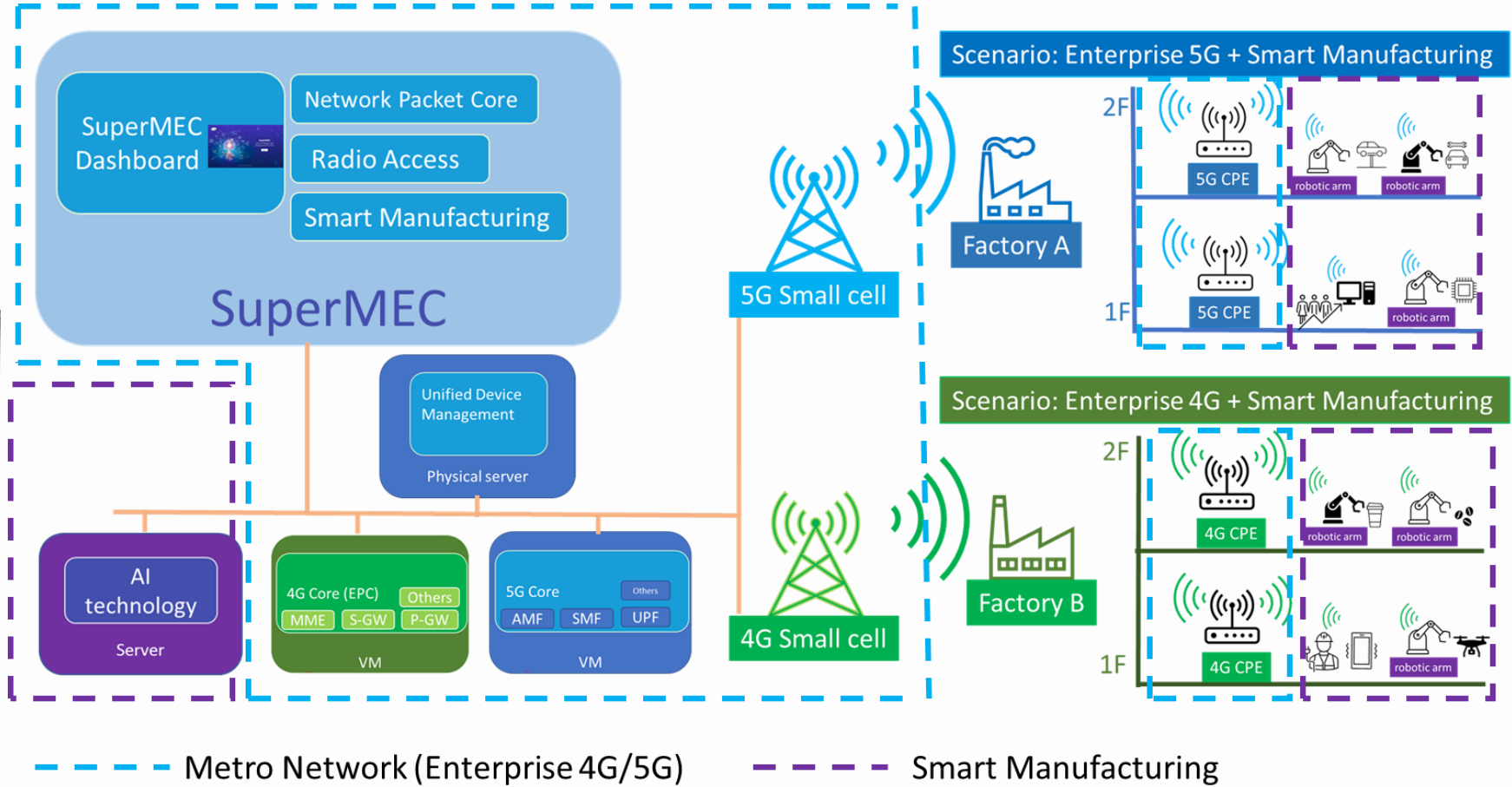
Origin Server	SDN / NVF	CDN Server	Edge Server	RAN / vRAN	IoT Gateways	Devices
Data Center	Backbone	Micro Data Centers		Base Station	5G and Telco	
Cloud	Network Core	Network Edge		Access Network	On-premise edge	



SuperMEC

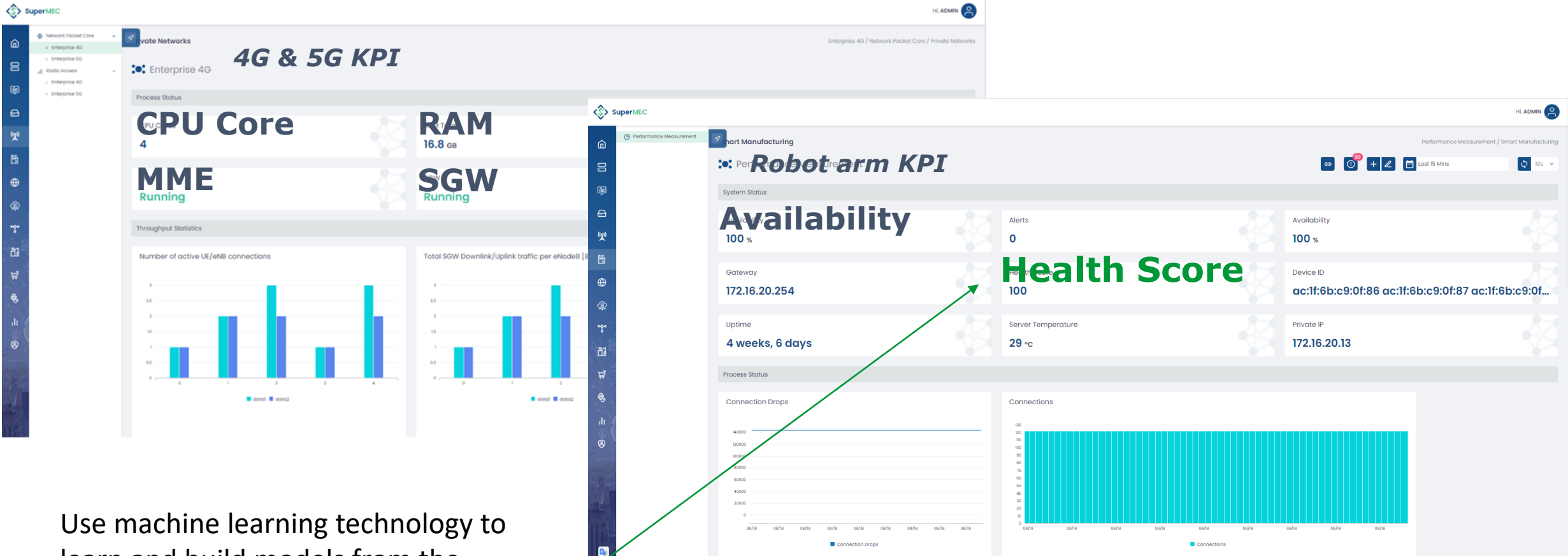


AI hardware in “Private Network + Smart Manufacturing” scenario



SuperMEC

AI software in “Private Network + Smart Manufacturing” scenario



Use machine learning technology to learn and build models from the collected data from sensors, and then calculate the **health score** of the robot arm.

Smart manufacturing pages shows Robot Performance Measurement etc

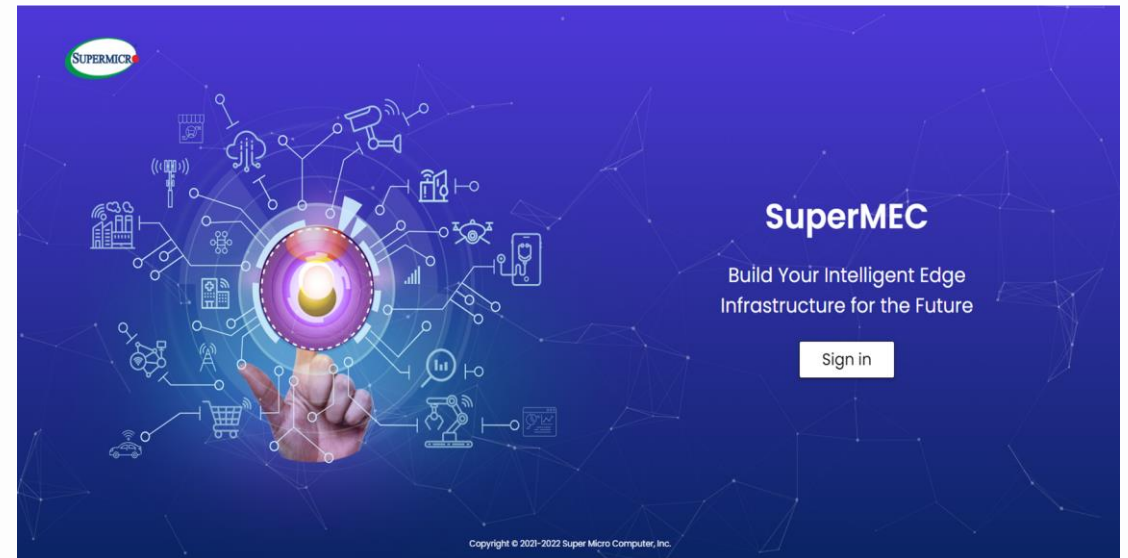
SuperMEC (Super Multi Edge Composer)

SuperMEC is a complete cloud edge infrastructure management platform.

It is used by customers to manage critical workload in 6 verticals: smart manufacturing, metro networks, telco cloud, retail, telemedicine and smart city.

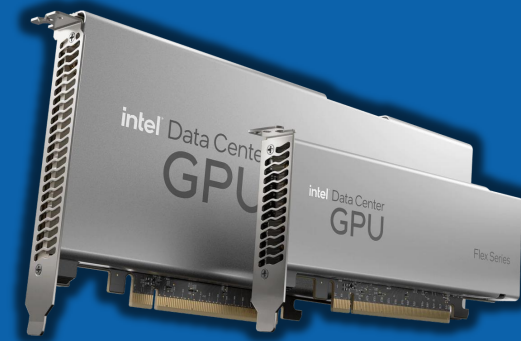
SuperMEC provides end users with telemetry, analytics, life cycle management, Infrastructure as a Service (IaaS), Monitor as a Service (MaaS), and Security as a service (SECaaS) that are essential to Edge and IoT verticals.

1. Fault management
2. Configuration management
3. Performance/ monitor management
4. Security management
5. Inventory management
6. License management
7. OS Provision
8. Event Log
9. Kubernetes architecture



Welcome to visit our booth to learn more with our team

Supermicro & Intel Data Center GPU Visual Cloud solution



DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2023 Super Micro Computer, Inc. All rights reserved.



www.supermicro.com

VOYAGER

Advancing AI for Science & Engineering