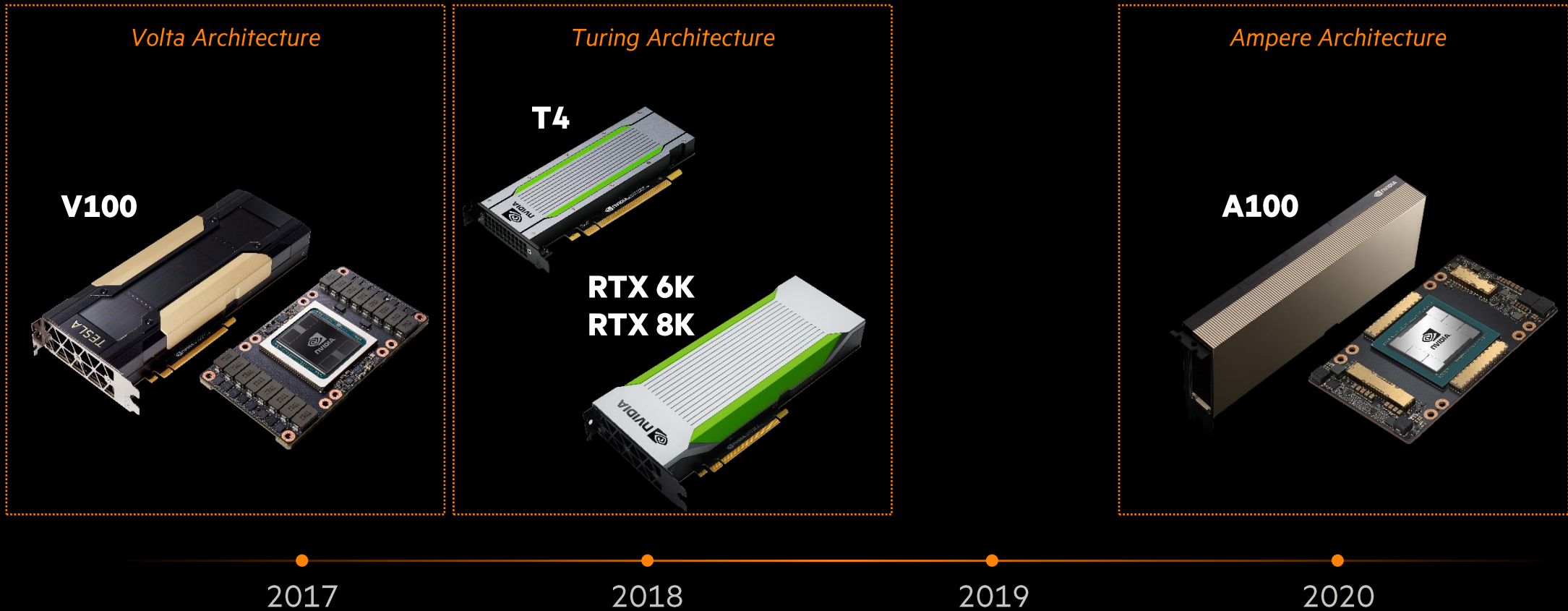


# HOW TO OPTIMIZE HPC/AI INFRA W/ NVIDIA & HPE

---

정구형 부장 | NVIDIA Korea

# NVIDIA DATA CENTER GPU HISTORY



# 최신 NVIDIA DATA CENTER GPU



**NVIDIA T4 16GB**

Multi-Purpose GPU for Enterprise Acceleration, Graphics, Inference



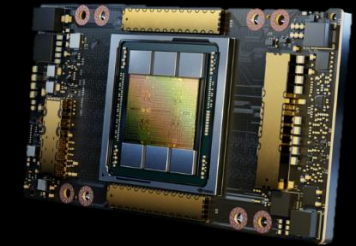
**NVIDIA A40 48GB**

World's Most Powerful Data Center GPU for Visual Computing (RTX 6K/8K Passive)



**NVIDIA A100 40GB**

World's Most Powerful Data Center GPU w/ MIG for max utilization



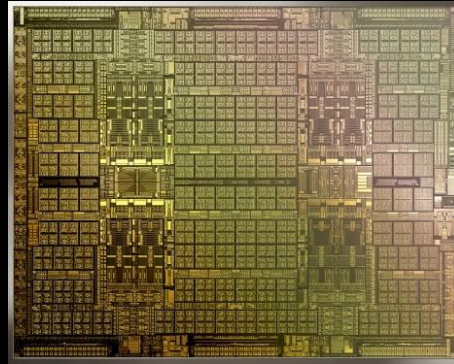
**A100 80GB SXM4**

World's Most Powerful Data Center GPU For Largest Workloads and Datasets

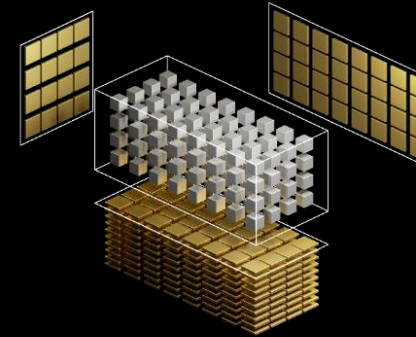
# CHOOSE THE RIGHT NVIDIA DATA CENTER GPU

Workload	Description	NVIDIA A100 SXM4	NVIDIA A100 PCIe	NVIDIA T4	NVIDIA A40
<b>Deep Learning Training</b>	For the absolute fastest model training time	<ul style="list-style-type: none"> <li>8-16 GPUs (for new installs)</li> <li>80GB: For largest models (DLRM, GPT-2 over 9.3Bn parameters in one node)</li> </ul>	<ul style="list-style-type: none"> <li>4-8 GPUs</li> </ul>		
<b>Deep Learning Inference</b>	For batch and real-time inference	<ul style="list-style-type: none"> <li>1 GPU w/ MIG</li> <li>80GB: 7MIGs at 10GB each for large batch size constrained models (RNN-T)</li> </ul>	<ul style="list-style-type: none"> <li>1 GPU w/ MIG</li> </ul>	<ul style="list-style-type: none"> <li>1 GPU</li> </ul>	
<b>HPC / AI</b>	For scientific computing centers and higher ed and research institutions	<ul style="list-style-type: none"> <li>4 GPUs with MIG for supercomputing centers</li> <li>80GB: For largest datasets and high memory bandwidth applications</li> </ul>	<ul style="list-style-type: none"> <li>1-4 GPUs with MIG for higher ed and research</li> </ul>		
<b>Render Farms</b>	For batch and real-time rendering				<ul style="list-style-type: none"> <li>4-8 GPUs</li> </ul>
<b>Graphics</b>	For the best graphics performance on professional virtual workstations*			<ul style="list-style-type: none"> <li>2-8 GPUs for mid-range virtual workstations for professional graphics</li> </ul>	<ul style="list-style-type: none"> <li>4-8 GPUs for mid and high-end professional graphics and RTX workloads or simulation</li> </ul>
<b>Enterprise Acceleration</b>	Mixed Workloads – Graphics, ML, DL, analytics, training, inference	<ul style="list-style-type: none"> <li>1-4 with MIG for compute intensive multi-GPU workloads</li> <li>80GB: data analytics with largest datasets</li> </ul>	<ul style="list-style-type: none"> <li>1-4 GPUs with MIG for compute intensive single GPU workloads</li> </ul>	<ul style="list-style-type: none"> <li>4-8 GPUs for balanced workloads*</li> </ul>	<ul style="list-style-type: none"> <li>2-4 GPUs for graphics intensive* and compute workloads</li> </ul>
<b>Edge Acceleration</b>	Edge solutions with differing use cases and deployment location		<ul style="list-style-type: none"> <li>1-2 GPU with MIG</li> </ul>	<ul style="list-style-type: none"> <li>1-8 GPUs for inference and video workloads</li> </ul>	<ul style="list-style-type: none"> <li>2-4 GPUs for graphics intensive &amp; AR/VR workloads*</li> </ul>

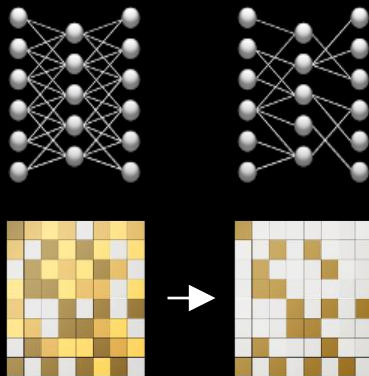
# NVIDIA A100 GPU 5가지 특징



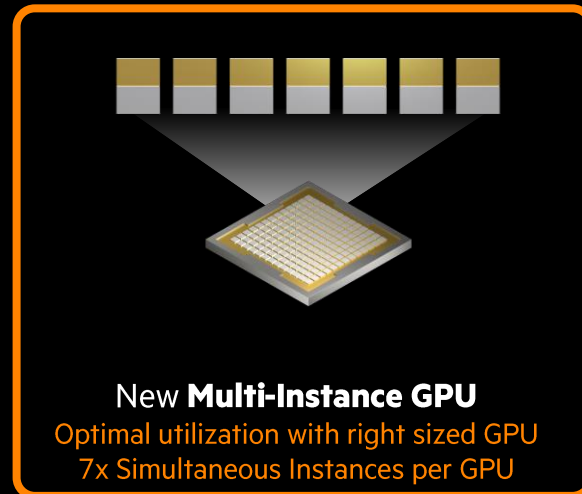
NVIDIA Ampere Architecture  
World's Largest 7nm chip  
54B XTORS, HBM2



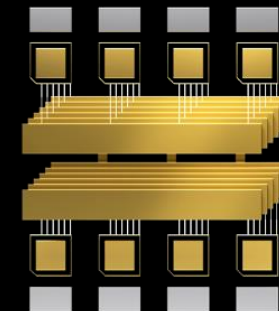
3<sup>rd</sup> Gen Tensor Cores  
Faster, Flexible, Easier to use  
20x AI Perf with TF32  
2.5x HPC Perf



New Sparsity Acceleration  
Harness Sparsity in AI Models  
2x AI Performance



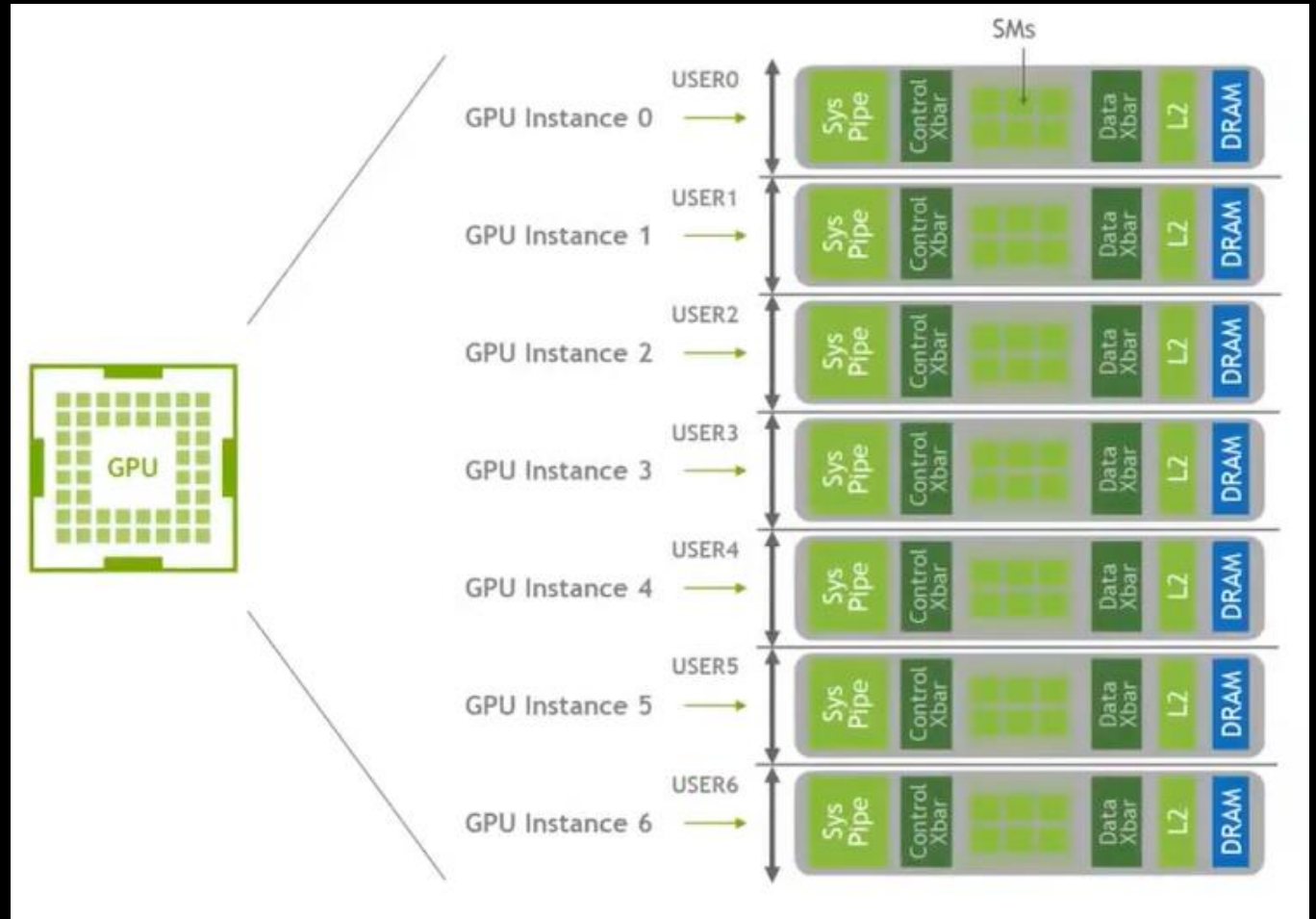
New **Multi-Instance GPU**  
Optimal utilization with right sized GPU  
7x Simultaneous Instances per GPU



3<sup>rd</sup> Gen NVLINK and NVSWITCH  
Efficient Scaling to Enable Super GPU  
2X More Bandwidth

# MIG (MULTI INSTANCE GPU)

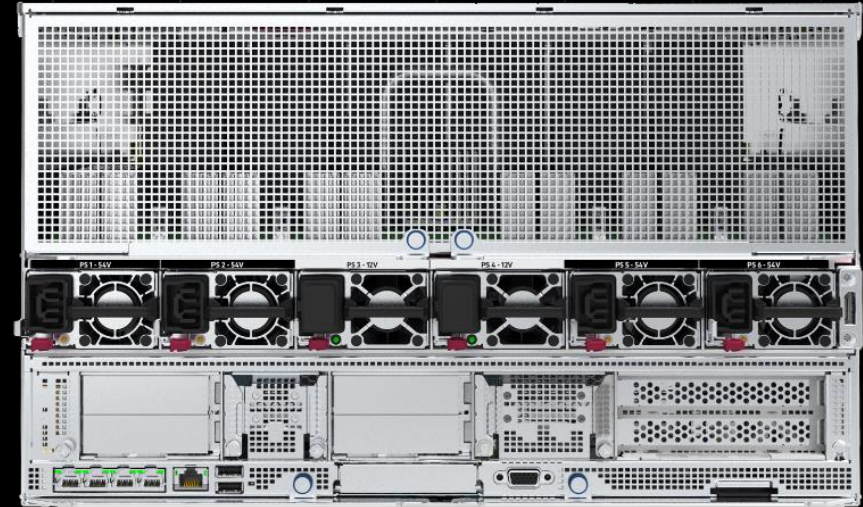
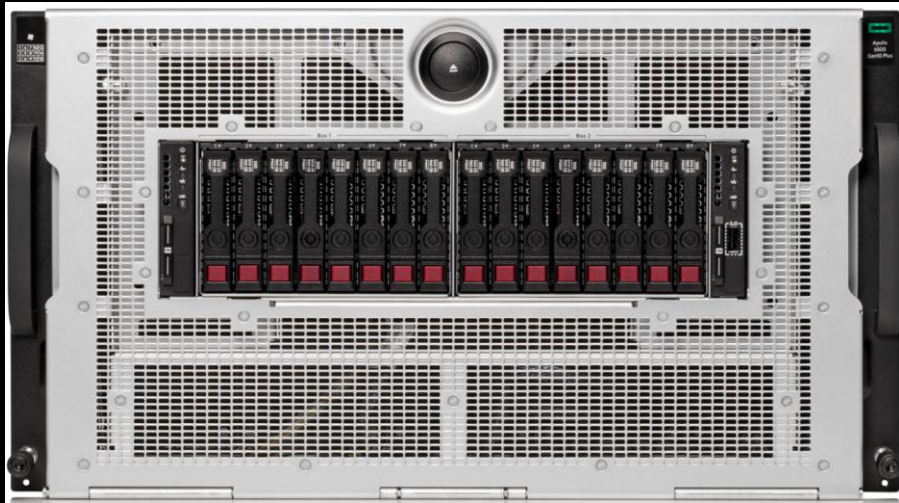
- A100 GPU, 최대 7개 GPU Slice로 분할
- 목적: GPU 사용률 극대화
- 사용 사례
  - A100 GPU 1장 미만 워크로드
  - 경량 학습, 추론, 개발, 일부 HPC
- 혜택
  - GPU 1장당 총 19가지 분할 가능
  - 분할된 MIG, H/W 독립, QoS 보장



# A100 GPU 1장을 MIG으로 구성할 수 있는 18가지 경우

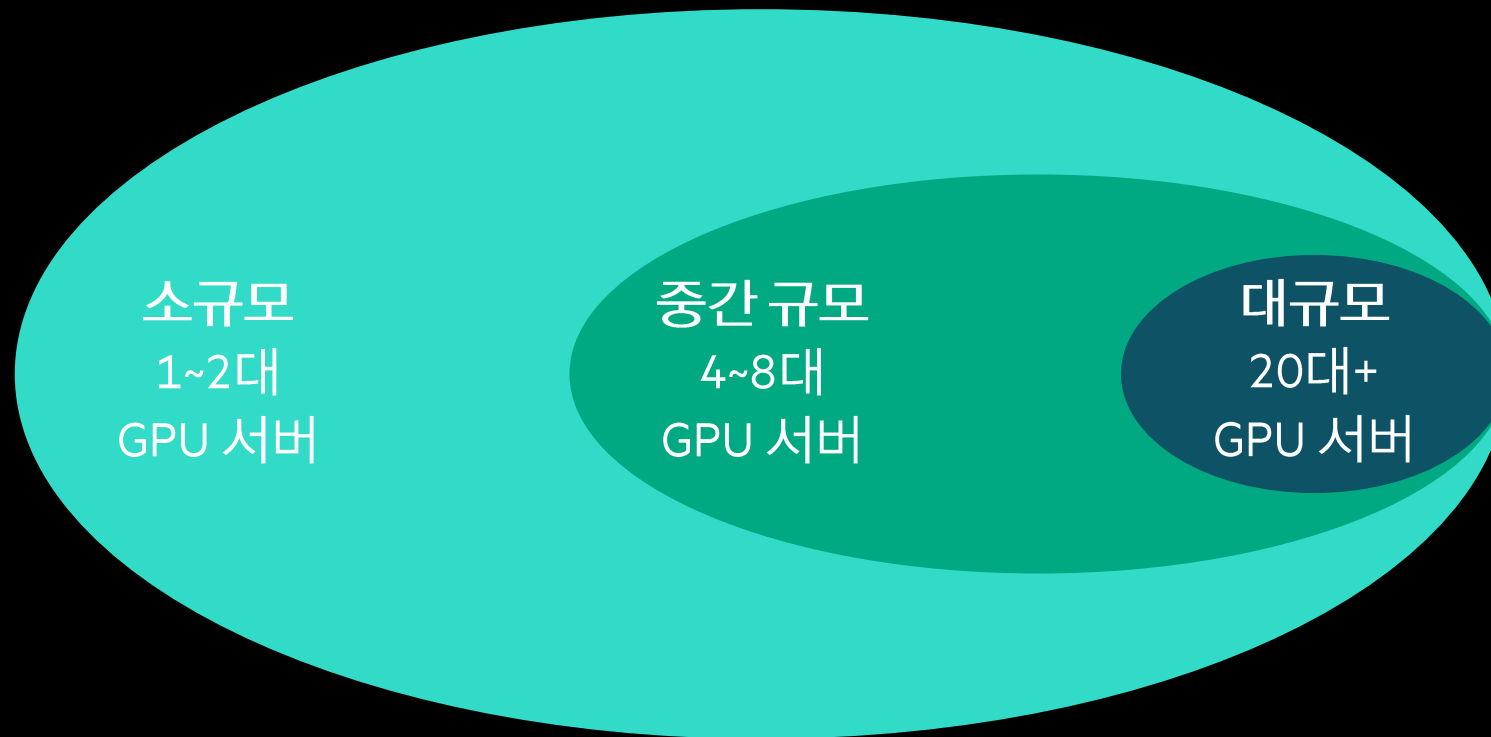
Slice #1	Slice #2	Slice #3	Slice #4	Slice #5	Slice #6	Slice #7
7						
4			2		1	
4			1	1	1	
2		2		3		
2		1	1	3		
1	1	2		3		
1	1	1	1	3		
3					3	
3					2	
3					1	1
2		2		2		1
2		2		1	1	1
1	1	2		2		1
1	1	2		1	1	1
2		1	1	2		1
2		1	1	1	1	1
1	1	1	1	2		1
1	1	1	1	1	1	1

# HPE APOLLO 6500 GEN10 PLUS



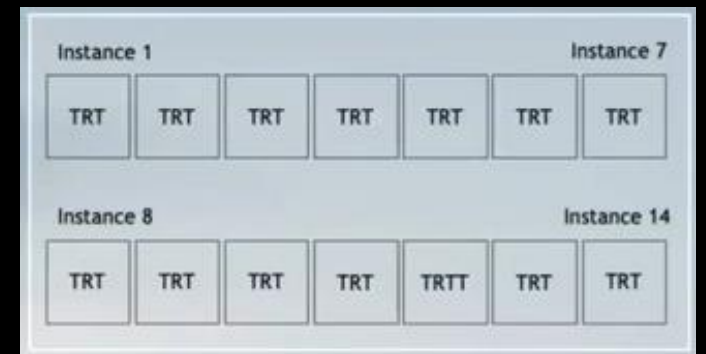
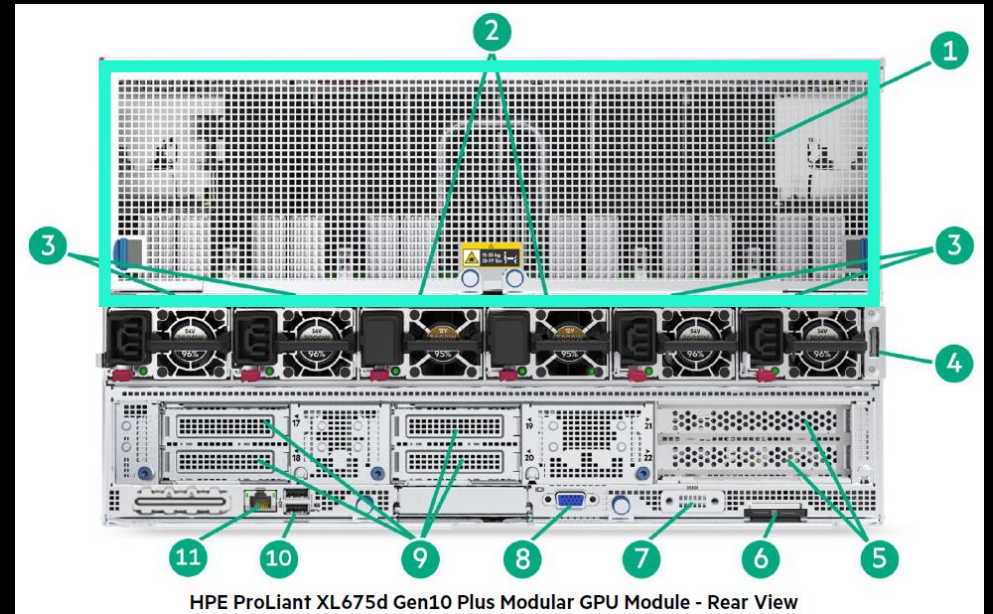
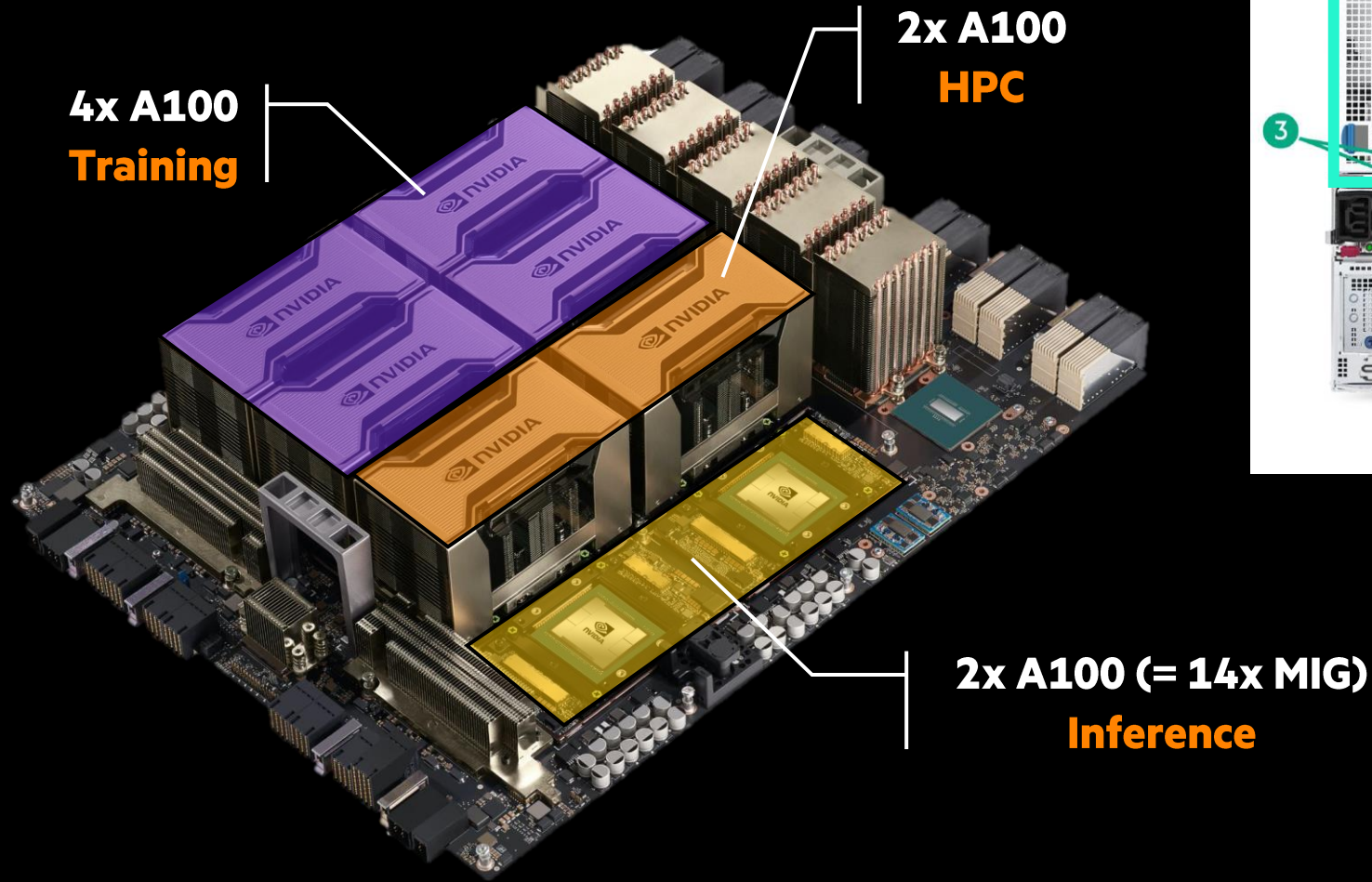


# 규모에 따른 GPU 서버/클러스터 구성 (예)



# 소규모 - 1~2대 GPU 서버 구성

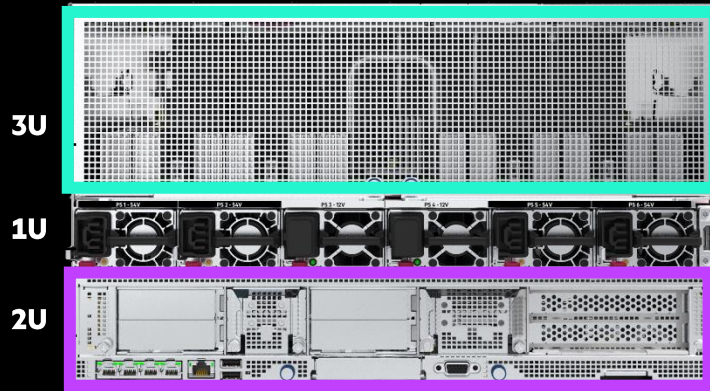
## GPU resource 분배 (예시)



# 소규모 - 1~2대 GPU 서버 구성

## 서버 구분 옵션

### HPE ProLiant XL675d Gen10 Plus



#### 3U Support—Accelerator tray:

- NVIDIA HGX A100 8-GPU
- 8, 10, 16 PCIE @ 300W – 75W

#### 2U Support—Compute tray:

- 2P AMD Rome , 280W CPU
- 4x LP PCIE Gen4 IO Slots
- 2x FH PCIE Gen4 IO Slots

### HPE ProLiant XL645d Gen10 Plus



#### 3U Support – Accelerator tray:

- NVIDIA HGX A100 4-GPU
- 4–8 PCIE @ 300W–75W

#### 2U Support—Compute tray:

- 1P AMD Rome, 280W CPU
- 2x LP PCIE Gen4 IO Slots

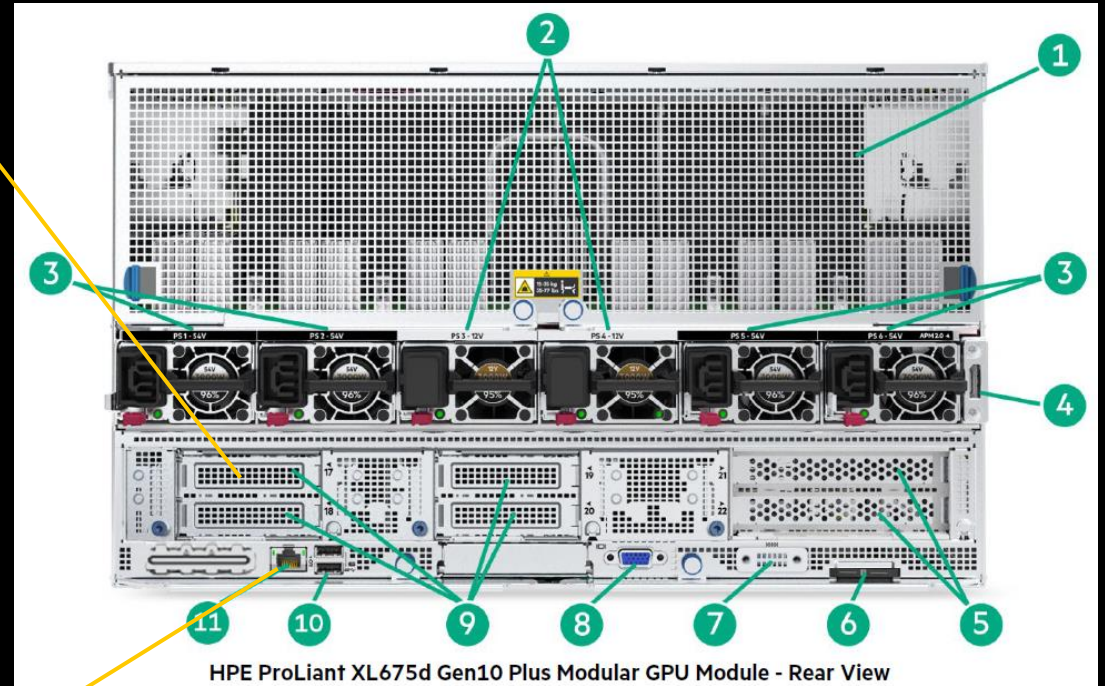
# 소규모 - 1~2대 GPU 서버 구성

## Network 구성 (예시)

**Ethernet 1Gb 4-port Adapter**  
서비스망



**Dedicated iLO management port**  
관리망

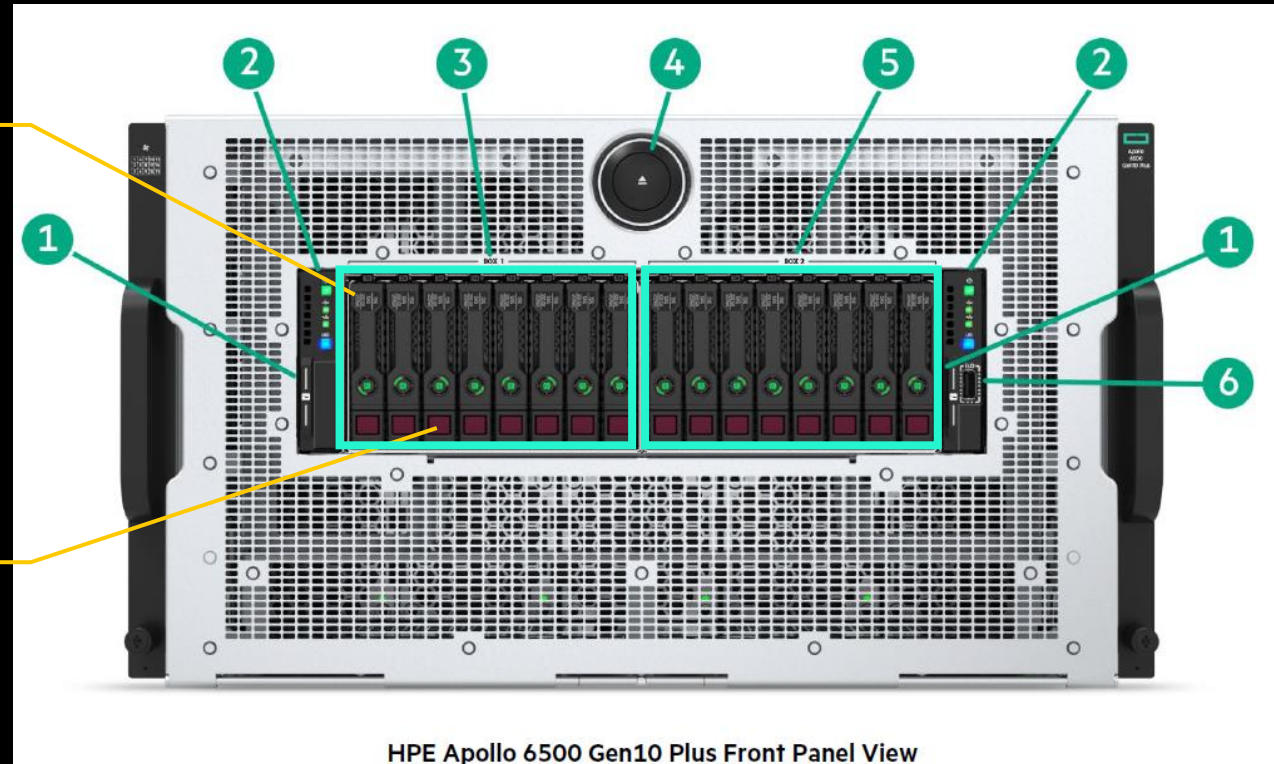


# 소규모 - 1~2대 GPU 서버 구성

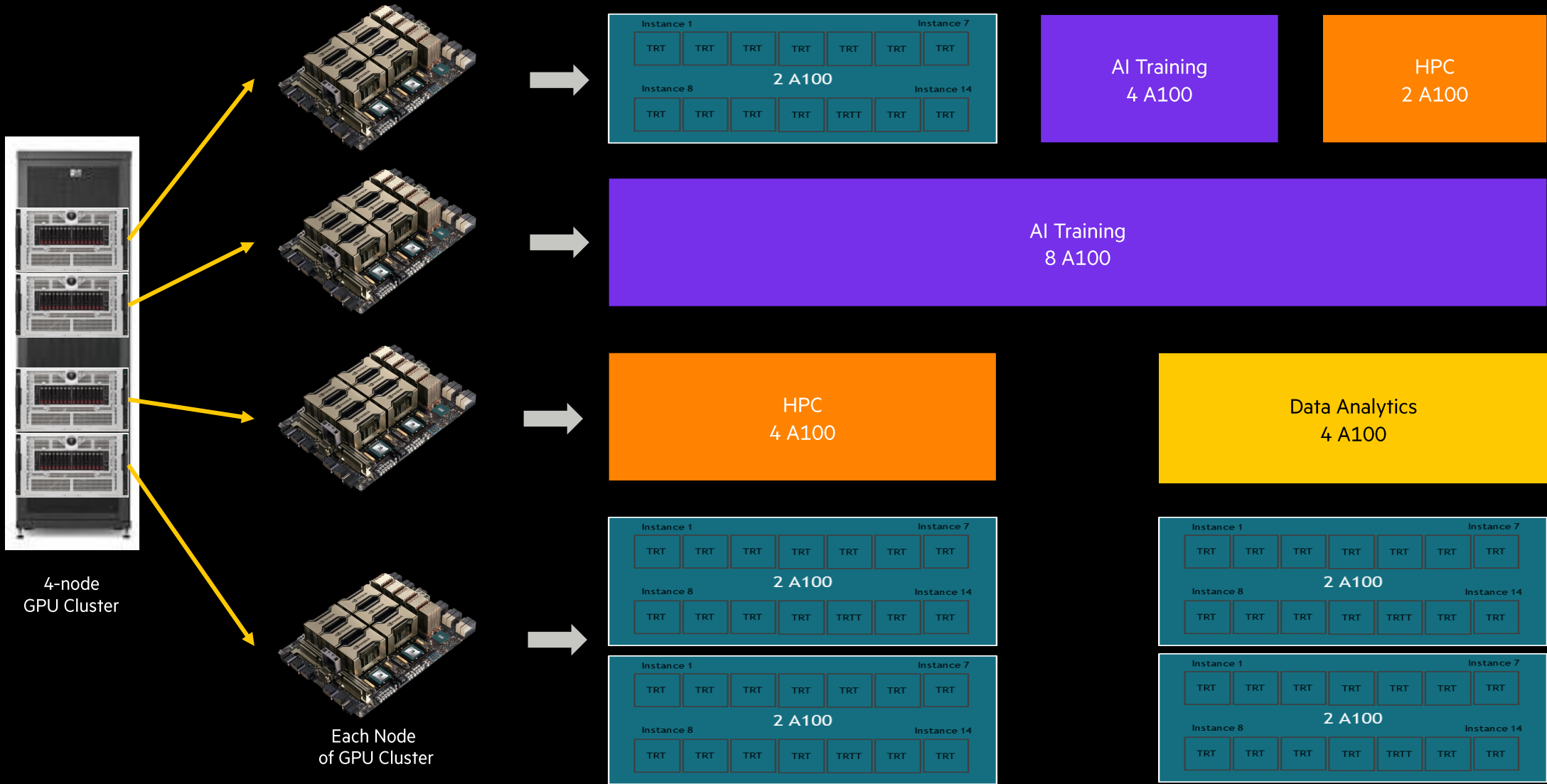
## Internal Storage 외 (예시)

2x 1.92TB SSD  
OS

?x 3.84TB SSD  
Data

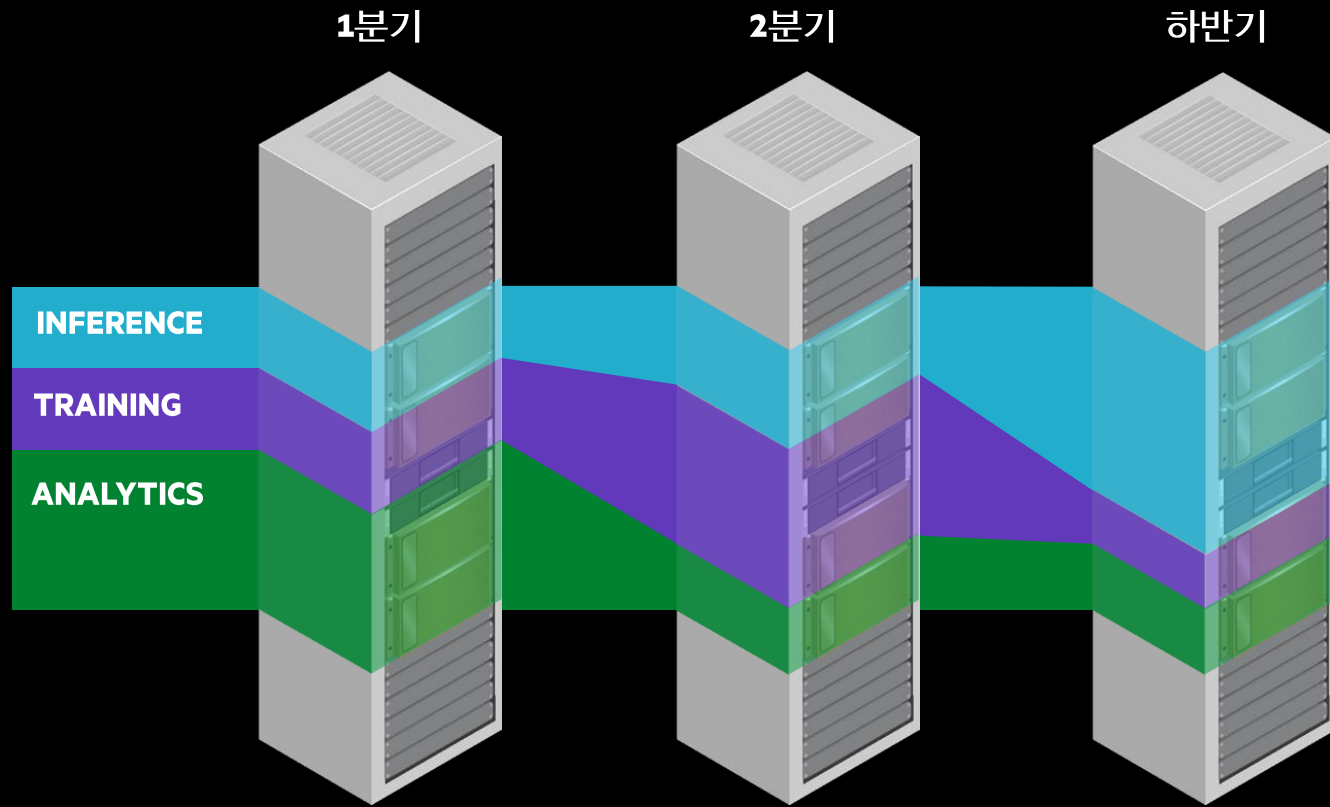


# 중간 규모 - 4~8대 GPU 서버 구성 (예)



# 중간 규모 - 4~8대 GPU 서버 구성 (예)

시기적 특성 있는 업무의 경우, **A100 MIG** 기능으로 유연한 시스템 구성



# 중간 규모 - 4~8대 GPU 서버 구성

## Network 구성 외 (예시)



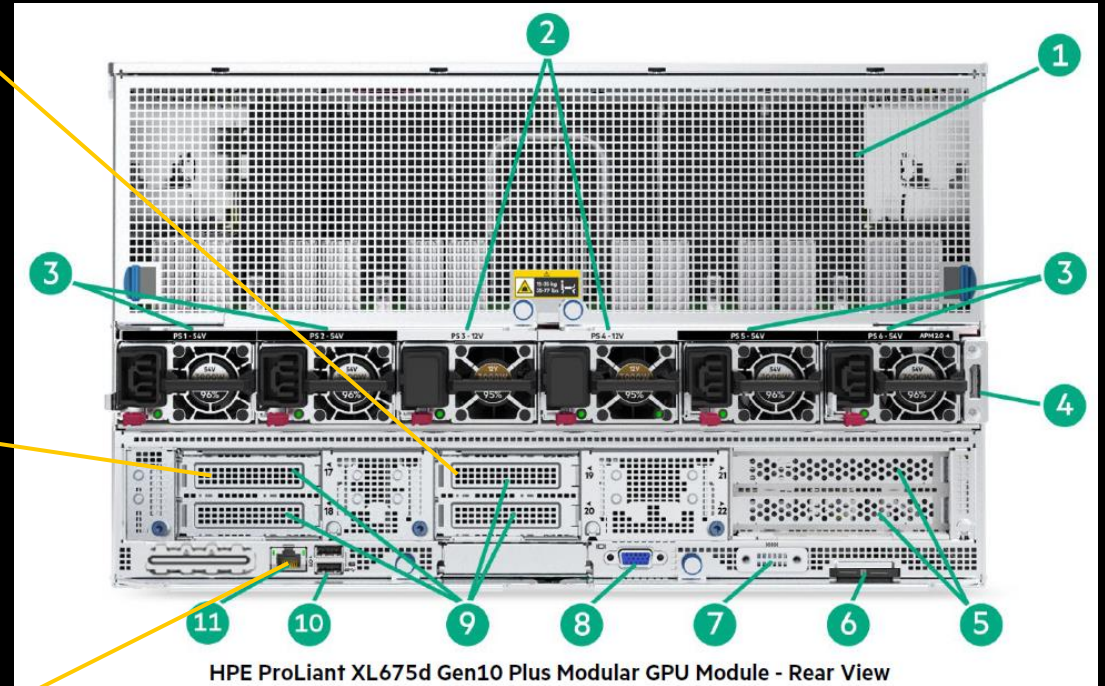
**Mellanox Ethernet 10/25Gb  
2-port Adapter**  
서비스망



**InfiniBand 100Gb  
1-port Adapter**  
Compute Fabric



**iLO management port**  
관리망



HPE ProLiant XL675d Gen10 Plus Modular GPU Module - Rear View



# 대규모 - 20대+ GPU 서버 구성

## GPT-3 모델이란?

- 자연어 처리 모델 (텍스트-인간의 언어-를 다루는 인공 지능 모델)

## 어디서 만들었나?

- OpenAI (샌프란시스코 위치 인공 지능 연구소)

## GPT 시리즈

- GPT-1 (2018년; 파라미터 수 약 1억 1천 7백 개)
- GPT-2 (2019년; 파라미터 수 약 15억 개)
- GPT-3 (2020년; 파라미터 수 약 1750억 개)

# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

1) GPT-3 모델이 쓴 기사를 사람이 봤을 때 48%이 사람이 쓴 것으로 착각

	Mean accuracy	95% Confidence Interval (low, hi)	<i>t</i> compared to control ( <i>p</i> -value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 ( $2e-4$ )	4.9%
GPT-3 Medium	61%	58%–65%	10.3 ( $7e-21$ )	6.0%
GPT-3 Large	68%	64%–72%	7.3 ( $3e-11$ )	8.7%
GPT-3 XL	62%	59%–65%	10.7 ( $1e-19$ )	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 ( $5e-19$ )	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 ( $3e-21$ )	6.2%
GPT-3 13B	55%	52%–58%	15.3 ( $1e-32$ )	7.1%
<b>GPT-3 175B</b>	<b>52%</b>	49%–54%	16.9 ( $1e-34$ )	7.8%

Table 3.11: **Human accuracy in identifying** whether short (~200 word) news articles are model generated. We find that human accuracy (measured by the ratio of correct assignments to non-neutral assignments) ranges from 86% on the control model to 52% on GPT-3 175B. This table compares mean accuracy between five different models, and shows the results of a two-sample T-Test for the difference in mean accuracy between each model and the control model (an unconditional GPT-3 Small model with increased output randomness).

**48%는 사람이 쓴 걸로 착각함**

# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

2) 텍스트로 만들고자 하는 어플리케이션을 설명하면



# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

2) 텍스트로 만들고자 하는 어플리케이션을 설명하면, 스스로 코딩을 짜서 어플리케이션을 완성

The screenshot shows a web browser displaying the 'debuild.co' interface. On the left, there is a text input field with the placeholder 'Just describe your app!' and buttons for 'Clear' and 'Generate'. Below the input field, a code editor shows a snippet of JavaScript code for a React component. On the right, a preview window shows a simple web form with an input field labeled 'Enter a todo' and a 'Save todo' button. Below the preview, the text 'learn about ai', 'WHAT', and 'IT WORKED WTF' is displayed. At the bottom of the video frame, a man is visible, and a text overlay reads '실제로 입력하면 To do들이 입력되는' (When you actually enter, todos are entered). A small 'super 코딩' logo is also present.

```
// an input that says "Enter a todo"
// and a button that says "Save todo".
// then show me all my todos
class App extends React.Component {
  constructor(props) {
    super(props);
  }
  state = {
    todos: []
  };
  render() {
    return (
      <div>
        <input type="text" value="Enter a todo" />
        <button type="button" value="Save todo" />
      </div>
    );
  }
}
```

실제로 입력하면 To do들이 입력되는

# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

3) 이메일 회신을 하고자 할 때 핵심 키워드만 입력하면,

127.0.0.1:5000

### GPT-3 Quick Response by OthersideAI

Quickly write an email in your style by simply stating the points you would like to get across ⚡  
Request beta access at othersideai.com 🤖

Received Email

Matt,

Thanks for chatting last week. Hearing your vision for Otherside got both Jim and I really excited. We really like where you're going with this. After talking with my partners yesterday, we're looking at making an investment of \$100K into Otherside on a SAFE. Would this be sufficient to join your round? If so, we'll send over our proposed terms.

On another note, as we discussed, let me know about your estimated market size.

Please let me know. Looking forward to an amazing journey together!

Thanks

Response Points

- \* thanks
- \* no
- \* our minimum is \$150K investment
- \* would \$150K be possible
- \* \$90B market

78.3K views

그래서 그런 것을 그냥 쓰지 않고 핵심만

# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

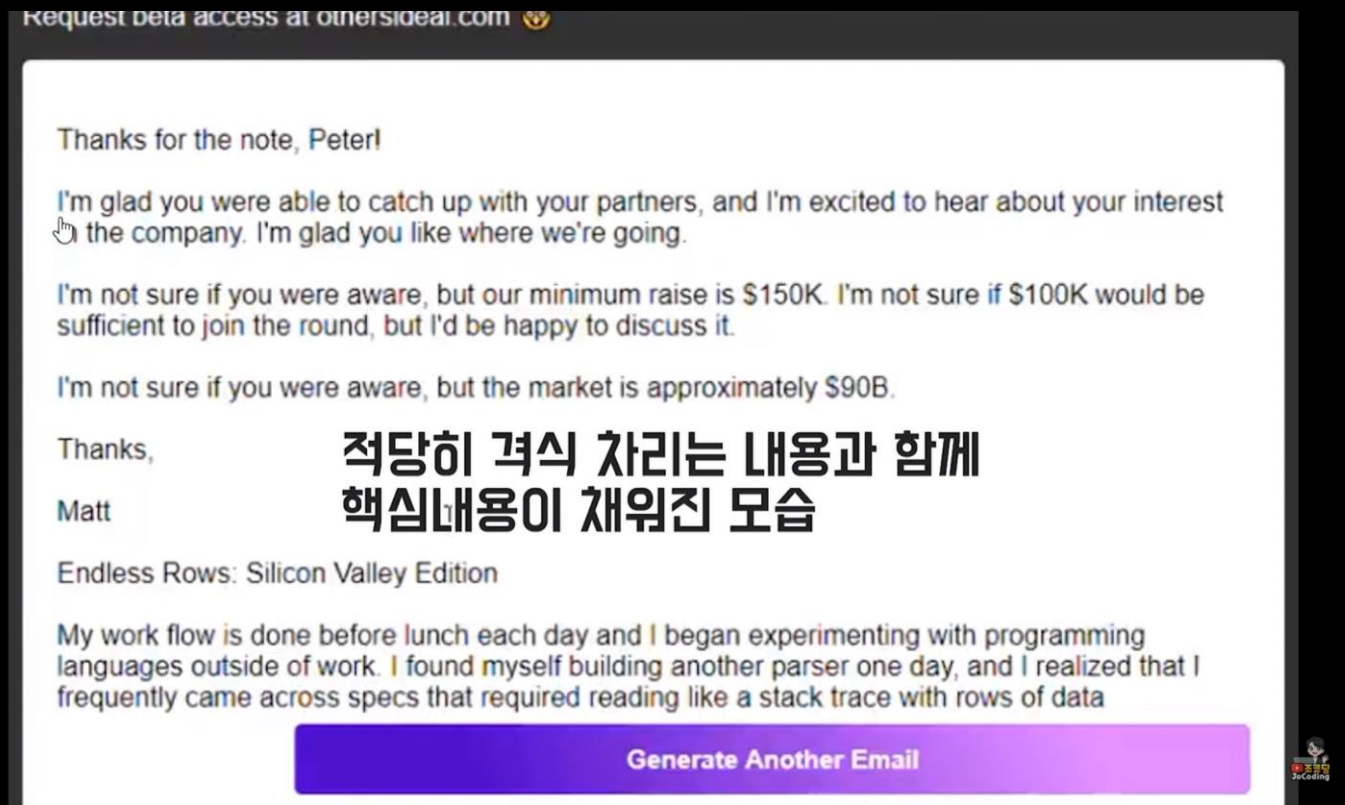
3) 이메일 회신을 하고자 할 때 핵심 키워드만 입력하면,



# 대규모 - 20대+ GPU 서버 구성

## 학습된 GPT-3 모델의 성능

3) 이메일 회신을 하고자 할 때 핵심 키워드만 입력하면, 자동으로 격식 차려진 이메일 작성



# 대규모 - 20대+ GPU 서버 구성

- 지난 7월 NVIDIA [기사](#)에 따르면, OpenAI 기관이 아래와 같은 규모의 시스템(Microsoft Azure)을 사용해서 GPT-3 모델을 훈련시키는데 성공했다고 합니다.
  - ✓ 285,000 CPU cores
  - ✓ 10,000 GPUs (NVIDIA V100)
  - ✓ 400 gigabits per second of network connectivity for each GPU server
- ZDNet의 [기사](#)에 따르면, 업계에서는 GPU 1장을 갖고 GPT-3을 훈련시킬 경우, 약 355년이 걸릴 것이라고 합니다.
- V100 GPU 1만 장일 경우, 소요 시간?
- V100 GPU가 아닌 A100 GPU일 경우, 소요 시간?



# 대규모 - 20대+ GPU 서버 구성

## 참고 문서

- ✓ HPE Apollo 6500 Gen10 Plus QuickSpecs (2020년 12월 버전)
- ✓ NVIDIA DGX SuperPOD Reference Architecture (2020년 11월 버전)

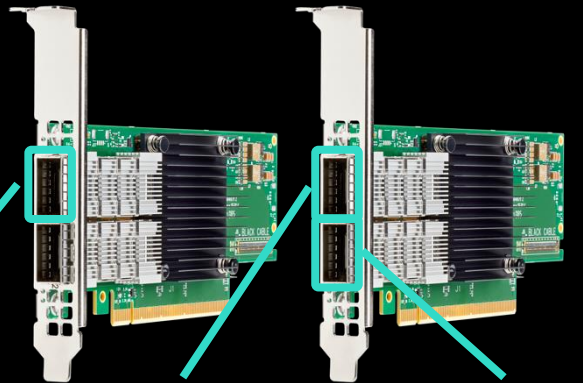
# 대규모 - 20대+ GPU 서버 구성

## Network 구성 (예시)



**4x Mellanox ConnectX-6 1-port VPI**  
Compute Fabric (InfiniBand - HDR)

Compute  
(InfiniBand)



Storage  
(InfiniBand)

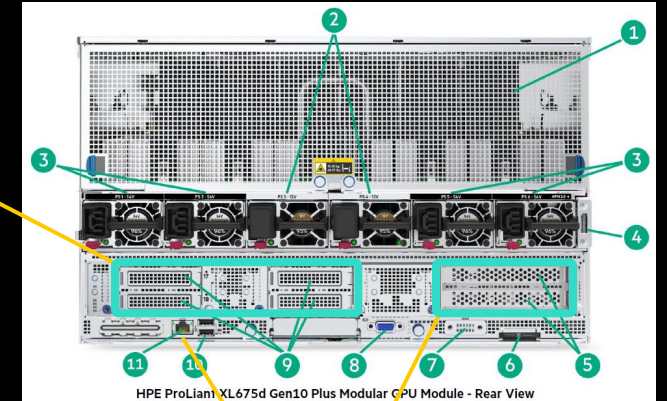
Storage  
(InfiniBand)

In-band  
(100GbE)

**2x Mellanox ConnectX-6 2-port VPI**  
Storage Fabric (InfiniBand - HDR)  
In-band management (Ethernet 100Gb)



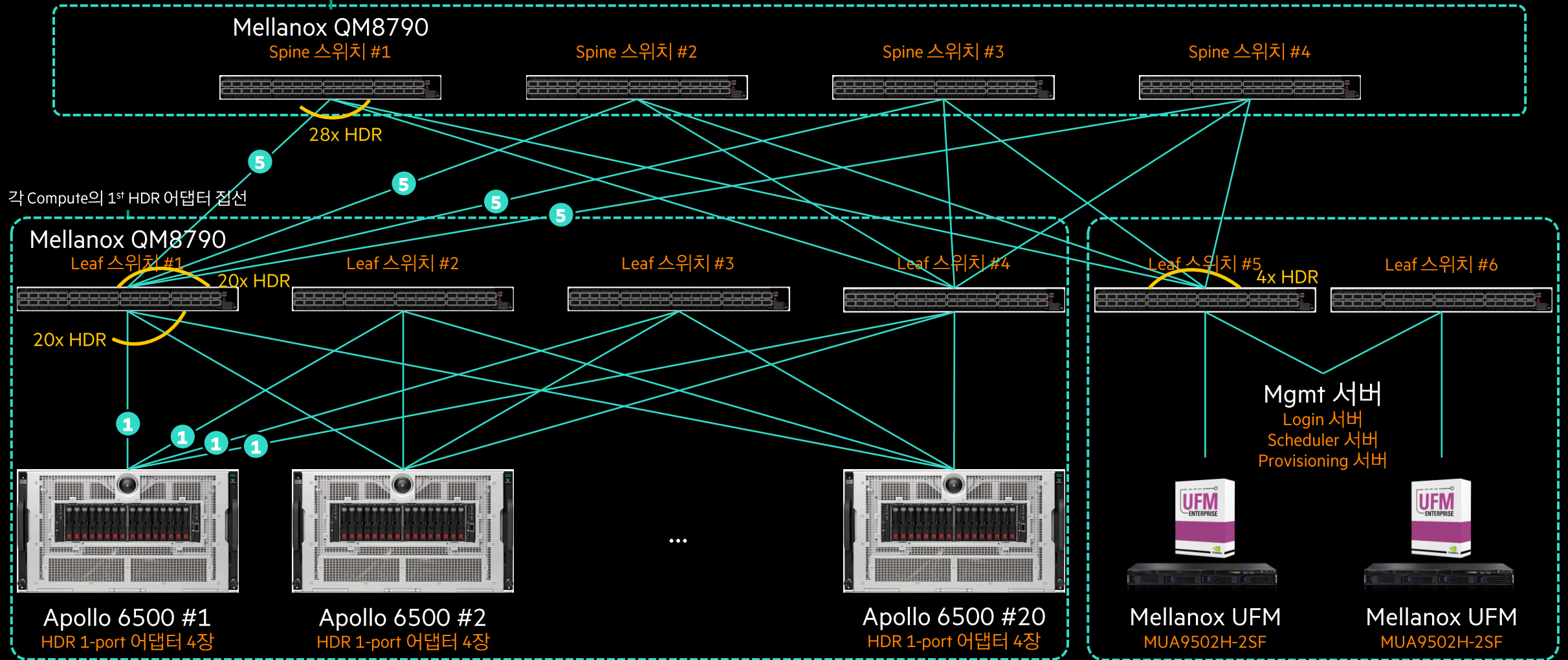
**iLO management port**  
Out-of-band management (1GbE)



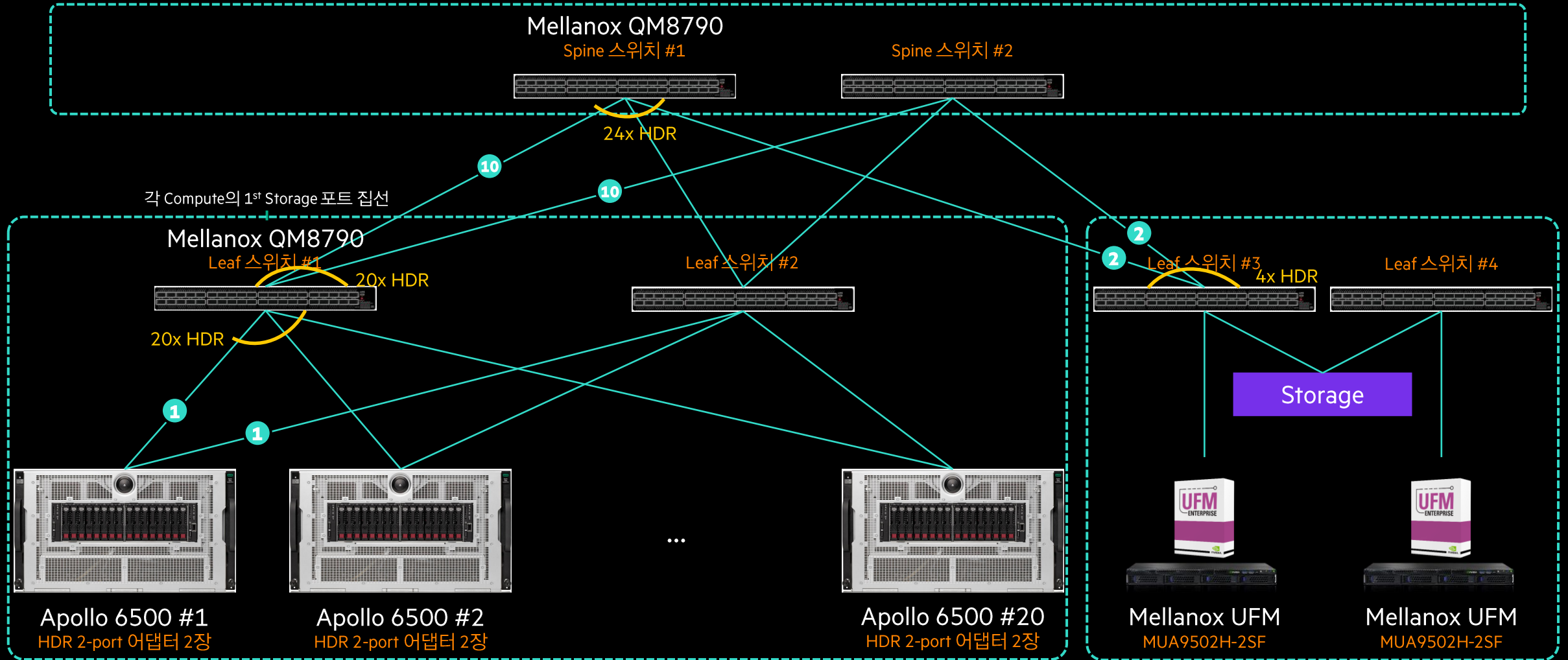
# 대규모 - 20대+ GPU 서버 구성시 COMPUTE FABRIC ARCHITECTURE

각 Compute의 서로 다른 HDR 어댑터간 통신

Leaf 스위치 그룹으로부터 올라오는 총 80x HDR port + UFM pot 받기 위한 Spine 스위치 그룹



# 대규모 - 20대+ GPU 서버 구성시 STORAGE FABRIC ARCHITECTURE



**THANK YOU**

