

목차

NVIDIA vGPU types 3

NVIDIA vGPU types

NVIDIA vGPU가 지원되는 환경에서 `mdevctl types` 명령어를 실행하면 사용 가능한 vGPU 타입들이 출력된다.

```
root@pve:~# mdevctl types
0000:01:00.0
  nvidia-156
    Available instances: 11
    Device API: vfio-pci
    Name: GRID P40-2B
    Description: num_heads=4, frl_config=45, framebuffer=2048M,
max_resolution=5120x2880, max_instance=12
  nvidia-215
    Available instances: 11
    Device API: vfio-pci
    Name: GRID P40-2B4
    Description: num_heads=4, frl_config=45, framebuffer=2048M,
max_resolution=5120x2880, max_instance=12
  nvidia-241
    Available instances: 0
    Device API: vfio-pci
    Name: GRID P40-1B4
    Description: num_heads=4, frl_config=45, framebuffer=1024M,
max_resolution=5120x2880, max_instance=24
  nvidia-46
    Available instances: 0
    Device API: vfio-pci
    Name: GRID P40-1Q
    Description: num_heads=4, frl_config=60, framebuffer=1024M,
max_resolution=5120x2880, max_instance=24
  nvidia-47
    Available instances: 11
    Device API: vfio-pci
    Name: GRID P40-2Q
    Description: num_heads=4, frl_config=60, framebuffer=2048M,
max_resolution=7680x4320, max_instance=12
  nvidia-48
    Available instances: 0
    Device API: vfio-pci
    Name: GRID P40-3Q
    Description: num_heads=4, frl_config=60, framebuffer=3072M,
max_resolution=7680x4320, max_instance=8
  ... 중략 ...
  nvidia-62
    Available instances: 0
    Device API: vfio-pci
    Name: GRID P40-1B
```

Description: num_heads=4, frl_config=45, framebuffer=1024M, max_resolution=5120x2880, max_instance=24

위에서 Name 부분에 P40-1B, P40-1Q, P40-3A 등의 형식이 나오는데 각 타입들이 의미하는 내용은 아래와 같다.

vGPU 타입	메모리	자동 할당되는 SM 수	CUDA 코어 (대략)	비고
P40-1Q	1GB	1 SM	~128개	가벼운 작업
P40-2Q	2GB	2 SM	~256개	중급 AI
P40-3Q	3GB	3 SM	~384개	-
P40-8Q	8GB	8 SM	~1,024개	고성능
P40-24Q	24GB	30 SM (전체)	~3,840개	독점 사용

- 스케줄링 정책에 따른 코어 사용 방식

정책	코어 사용 방식
Q (Fixed Share)	각 vGPU에 고정된 시간 슬라이스 → 코어는 순차적으로 사용 (성능 예측 가능)
B (Best Effort)	코어는 경쟁적 사용 → 비어 있는 SM을 먼저 점유
A (Equal Share)	모든 vGPU가 동등하게 SM 공유

따라서 엄격하게 자원을 나눠서 사용하는 경우에는 Q타입을 사용하면 되고 부하에 따라 자동으로 경쟁적으로 알아서 사용하게 하려면 B타입을 사용 하면 된다. A타입의 경우는 모든 인스턴스가 동일하게 GPU를 나눠쓰게 된다.

위 규칙에 따라 P40-2Q/B/A 형식의 인스턴스 8개를 생성했을때 타입별로 아래와 같은 효과가 난다.

상황	Q 타입 (Fixed)	B 타입 (Best Effort)	A 타입 (Equal Share)
1개 VM만 작업 중(7개 휴식)	거의 100% 성능 사용 가능	거의 100% 성능 사용 가능	무조건 1/8(12.5%)만 사용
8개 VM 모두 풀 로드	각 VM 정확히 1/8 성능 보장	먼저 시작한 VM이 더 많이 가져감 (불공평 가능)	모든 VM 정확히 1/8 성능
AI 학습 (성능 예측 중요)	◎ 최고	△ 변동 심함	△ 너무 느림
VDI/가상 데스크톱 (밀도 중요)	△ 밀도 낮음	◎ 최고 (B4까지 가능)	△ 느림
테스트/개발 환경	○ 괜찮음	○ 괜찮음	◎ 완벽 공정

From:
<https://atl.kr/dokuwiki/> - AllThatLinux!

Permanent link:
https://atl.kr/dokuwiki/doku.php/nvidia_vgpu_types

Last update: 2025/11/17 08:32

