업데이트일 : 2018-11-23

II. 컴퓨팅

2. GPU Server

목차

2.1 ucloud GPU Server 소개

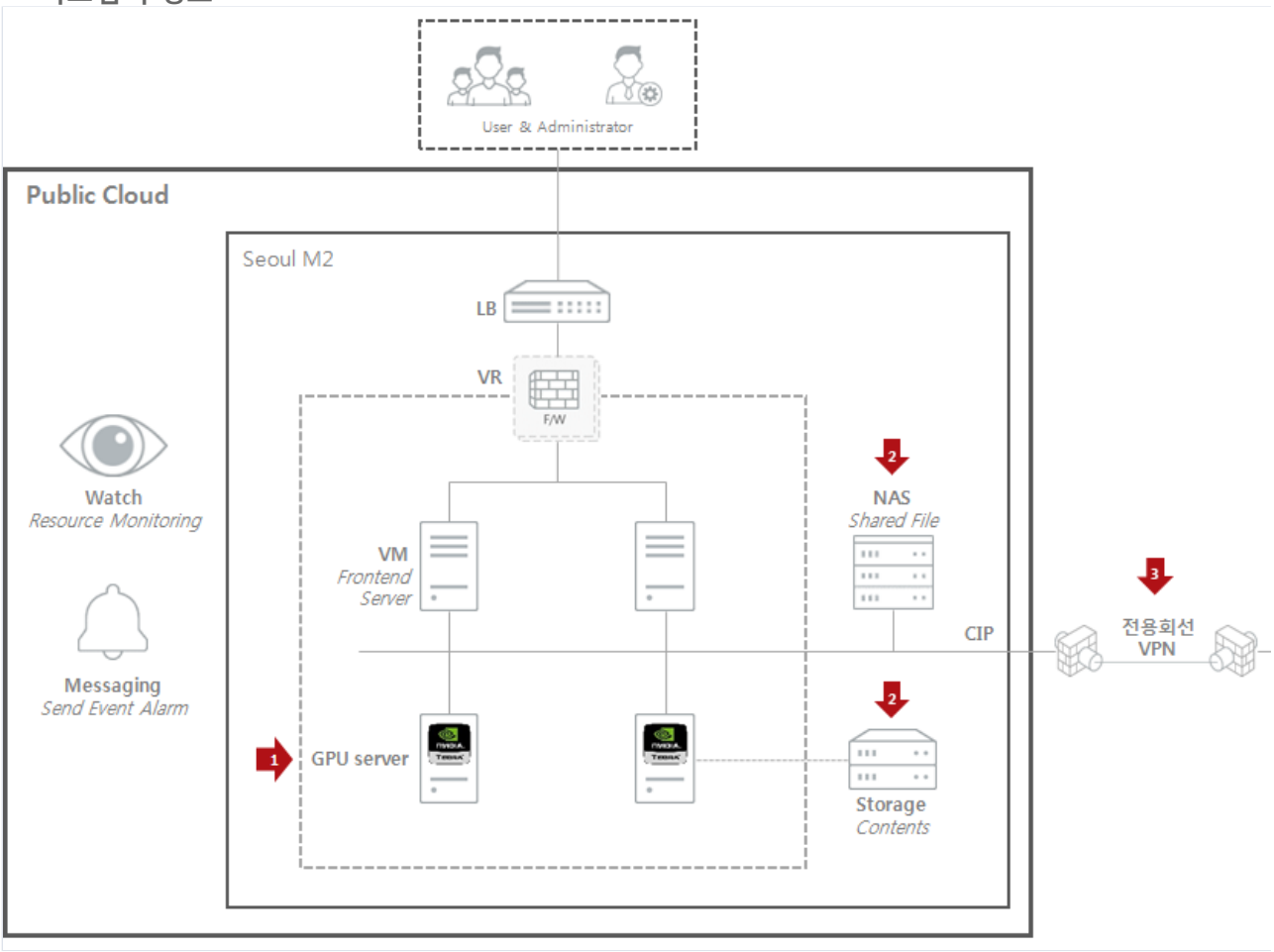
2.1.1 목적/용도

GPU server는 GPGPU(General-Purpose computing on Graphics Processing Units) 기술을 사용하여 전통적으로 CPU가 담당했던 응용 프로그램들의 계산에 이용할 수 있는 서버입니다. CPU 대비 코어의 성능은 훨씬 떨어지나 그 수가 매우 방대하여 병렬 연산에 큰 장점을 가지고 있습니다.

2.1.2 구조/원리

ucloud GPU Server는 VM 1대에 GPU 1개가 모두 할당되는 Passthrough방식을 사용하여 고성능 CUDA, AI등의 개발에 용이합니다.

▫ 시스템 구성도



1.1.3 유의사항/제약사항

▫ ucloud GPU Server Live Migration 불가

○ ucloud GPU Server는 VM 1대에 GPU 1개가 모두 할당되는 Passthrough방식을 사용하여 Live Migration이 불가하여 VM Stop 후 Migration을 해야 합니다.

▫ ucloud GPU Server 서울 M2존에서만 생성 가능

○ ucloud GPU Server는 현재 서울 M2존에서만 생성이 가능합니다.

2.2 ucloud GPU Server FAQ

2.2.1 ucloud GPU FAQ

▫ ucloud GPU Server와 일반 Server와 사용상 차이점이 있나요?

○ ucloud GPU Server도 일반 Server와 Disk 추가, Snapshot 생성 등 모든 사용법이 동일합니다. 부가서비스도 모두 동일하게 사용이 가능합니다.

▫ ucloud GPU Server를 일반 연산용이 아닌 그래픽용으로 사용할 수 있나요?

○ ucloud GPU Server는 NVIDIA Tesla K80모델을 사용합니다. 해당 GPU 모델은 그래픽용이 아닌 CUDA, AI등 연산 개발용으로 나온 GPU이므로 그래픽용 보다는 연산개발용으로의 사용을 권장합니다.

▫ ucloud GPU Server는 서울 M2존에서만 사용이 가능한가요?

○ ucloud GPU Server는 현재 서울 M2존에서만 사용이 가능합니다.

▫ ucloud GPU Server에서 GPU가 할당되었는지 어떻게 확인하나요?

○ nvidia-smi 명령어를 사용하여 확인하며, Windows의 경우는 제어판-하드웨어-장치 관리자-디스플레이 어댑터에서도 확인이 가능합니다.
ucloud GPU Server 이용방법 가이드를 참고하시면 자세하게 확인 방법이 나와있습니다.

▫ ucloud GPU Server용 이미지에 설치된 NVIDIA CUDA Driver 버전은 몇인가요?
○ NVIDIA CUDA Driver 8.0 버전을 사용합니다.

▫ucloud GPU Server가 일반 CPU Server 대비 성능이 얼마나 좋은가요?
○ucloud GPU Server는 BlackScholes 시험 시 CPU 대비 234배 우수하며 채권수익률 측정 시 2.57배 우수합니다.

▫ucloud GPU Server 소개 페이지에 GPU 제원이 CUDA Core 2* 2496, GDDR5 Memory 2 * 12GB라고 적혀 있는데 해당 내용이 GPU 1개에 대한 제원인가요?

○K80 GPU 2개에 대한 내용입니다. K80 GPU의 경우 GPU 카드 1개에 칩셋이 2개여서 GPU 2개가 사용이 가능합니다. 따라서 Passthrough 방식으로 VM에 K80 GPU 1개를 할당한다면 CUDA Core는 2496, GDDR5 Memory는 12GB입니다.이 외의 Half Precison, Single Precision 등 2*로 표시 안된 부분들은 모두 K80 GPU 1개에 대한 내용입니다.

2.2.2 ucloud GPU 용어집

▫ GPU: Graphics Processing Unit의 약자로, 컴퓨터의 영상정보를 처리하거나 화면 출력을 담당하는 그래픽카드로 CPU가 처리하기 버거워하는 3D 그래픽 작업을 처리하는 칩셋.
▫ CUDA Core: NVIDIA에서 표기하는 코어 방식으로 3D 계산과 관련된 일 처리 전문코어
▫ 3D Rendering: 3D 게임 세계의 객체들을 객체가 가지고있는 방향과 위치 정보를 이용하여 2D 화면에 출력
▫ Machine Learning: 기계학습, 인공지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
▫ GDDR5: Graphics Double Data Rate version 5의 약자로 고대역폭에 최적화된 고속 DRAM의 일종

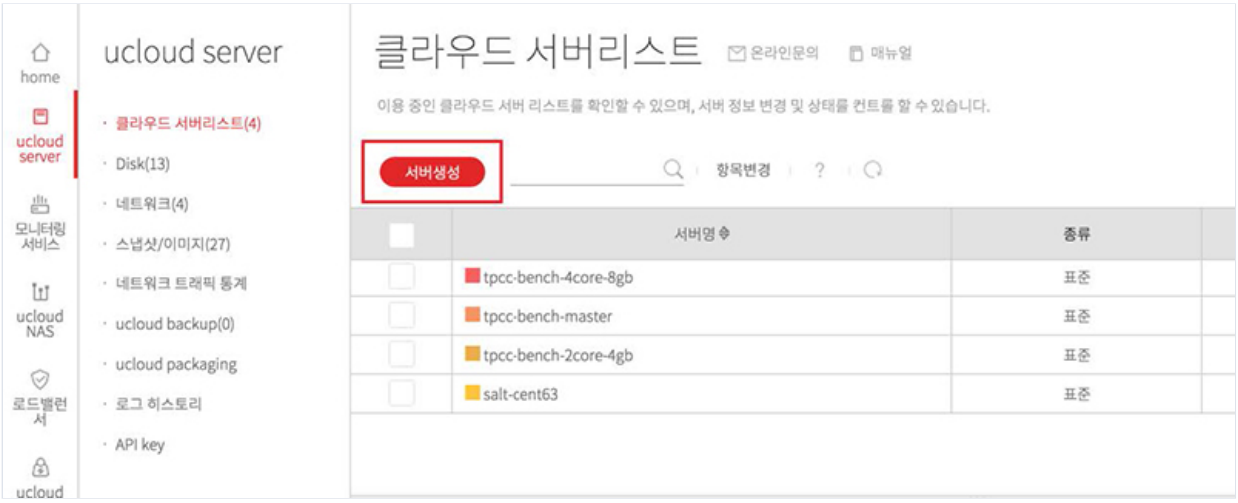
2.3 ucloud GPU Server 이용방법

2.3.1 상품 신청

ucloud GPU Server는 일반 Server와 상품 신청 방법 및 사용방법이 모두 동일합니다. 다만, GPU Server는 현재 서울 M2존에서만 생성이 가능합니다.

(1) 메인 홈페이지 우측 상단에 위치한 "클라우드 콘솔 버튼" 클릭

(2) ucloud server > 클라우드 서버리스트 > "서버 생성" 클릭



(3) (필수 입력사항 기준) 서버명 입력 및 중복확인 -> 그룹명 선택 -> 위치 선택 -> 운영 체제 선택 -> 생성 요청 서버 수 결정 -> 서버 사양 선택 -> "신청하기" 버튼 클릭 순으로 서버 신청 진행

(4) (필수사항) '서버명' 입력 및 중복 확인

(5) (필수사항) '위치'(서버가 생성될 Zone) KOR-Seoul M2존 선택

(6) (필수사항) '운영체제' 선택 (상품 종류 – 표준, High-Memory, SSD 및 이미지 선택 가능)

(7) 상품종류 선택(GPU Server 선택, KOR-Seoul M2 Zone의 표준서버는 All flash SSD 서버입니다.)

(8) 이미지선택 선택(기본이미지, 나의이미지, 공개이미지)

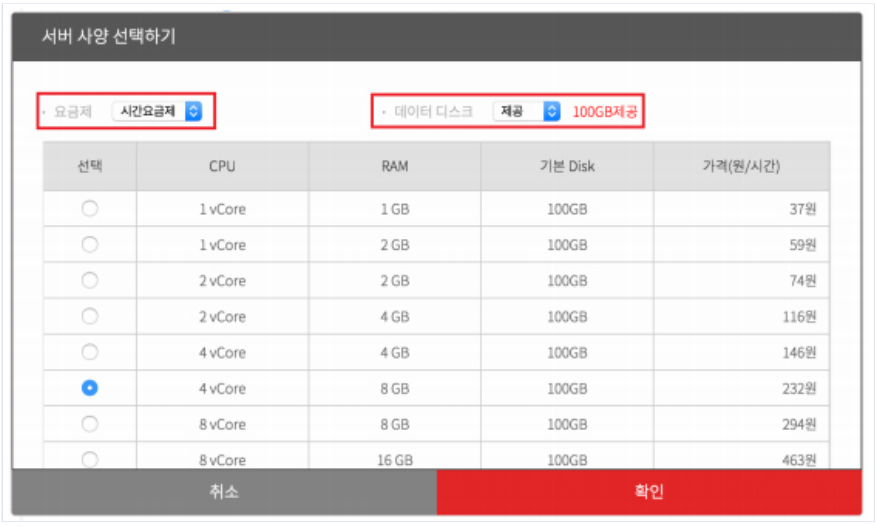
※ ucloud GPU Server는 Server 생성 즉시 추가적인 환경 구축 없이 바로 사용 가능하도록 관련 NVIDIA Driver 및 고성능 설정이 적용된 자동화 이미지를 제공합니다.
※ Centos7.2 OS 선택 시 유의사항 : Centos7.2에서 'service --status-all' 명령어 실행 시 Network 서비스가 restart 될 수 있으므로 'Systemctl' 명령어 사용을 권고 드립니다.

(9)(선택사항) 'CIP IP' 선택

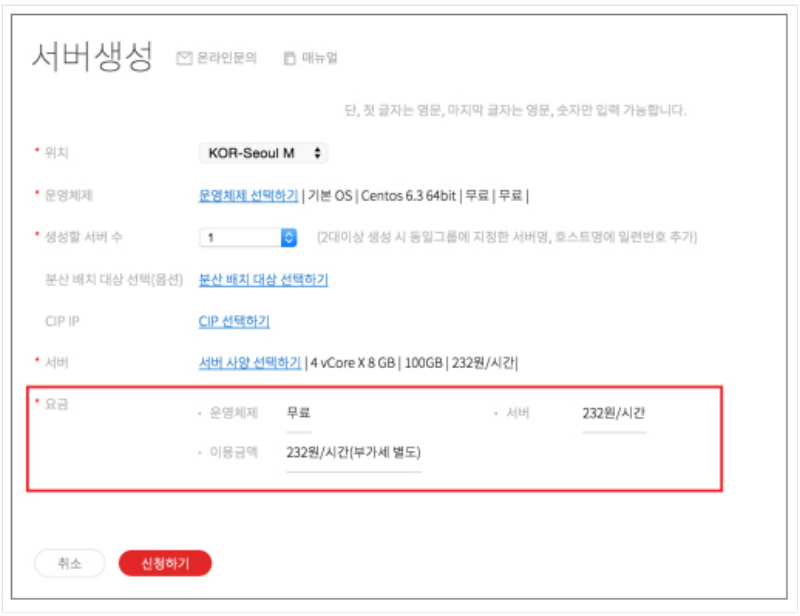
※ CIP (Cloud Internal Path)가 생성된 상태에서만 해당 기능의 사용이 가능하며 ucloud server 네트워크 탭에서 CIP 생성 및 관리 기능을 지원합니다.
※ 'IP 자동할당' 기능: CIP 네트워크 대역 내에서 IP를 자동으로 할당, 'IP 지정' 기능 : 네트워크 대역 내에서 사용자가 직접 IP 지정
※ CIP를 통해 zone간 네트워크 통신이 가능

(10)(필수사항) 서버 사양 선택 (요금제 및 데이터 디스크 제공 여부 선택 가능)

※ 데이터 디스크 제공 선택 시에는 기본적으로 100G (OS 디스크 + 데이터 디스크)가 제공되며, 미 제공을 선택 시에는 OS 디스크 (Linux 20GB, Windows 50GB)만 제공됩 니다.
※ 서버의 사양은 선택한 OS의 종류에 따라 선택이 가능한 사양만 보여지게 됩니다.

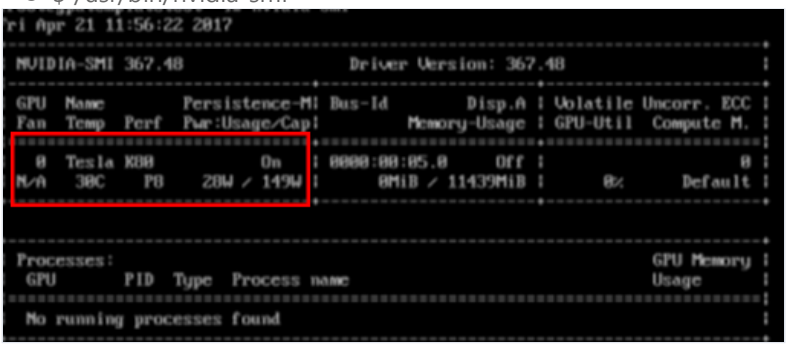


(11) 요금 정보 확인에서 자동으로 생성된 요약정보 확인 후 "신청하기" 버튼 클릭으로 서버 신청 완료



2.3.2 GPU 할당 확인

- **Linux**
  - $ /usr/bin/nvidia-smi



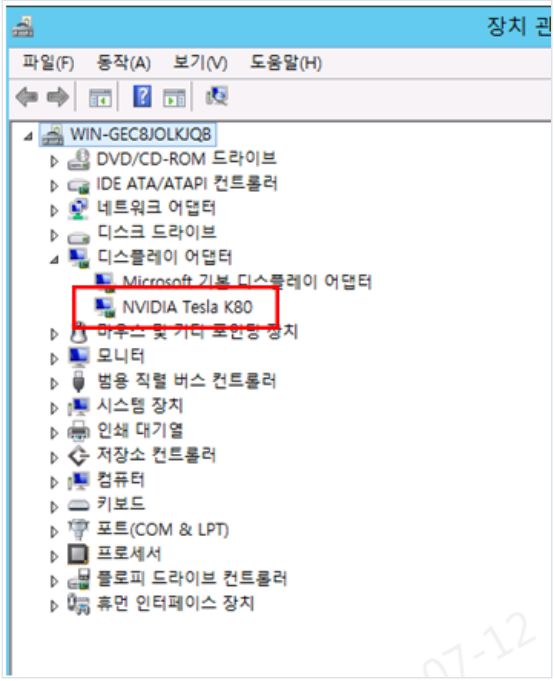
- **Windows**
  - 실행-cmd 로 cmd 창을 열고 아래 명령어 입력
  - $ cd C:₩Program Files₩NVIDIA Corporation₩NVSMI
  - $ nvidia-smi

```
C:\Users\Administrator>cd C:\Program Files\NVIDIA Corporation\NVSMI

C:\Program Files\NVIDIA Corporation\NVSMI>nvidia-smi
Fri Apr 21 14:04:05 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 369.30                 Driver Version: 369.30                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name            TCC/WDDM | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           TCC  | 0000:00:05.0     Off |                    0 |
| N/A   26C    P8    28W / 149W |      0MiB / 11423MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

ㅇ 제어판 - 하드웨어 - 장치관리자 - 디스플레이 어댑터에서 NVIDIA Tesla K80 확인



2.3.3 GPU 자동화 이미지

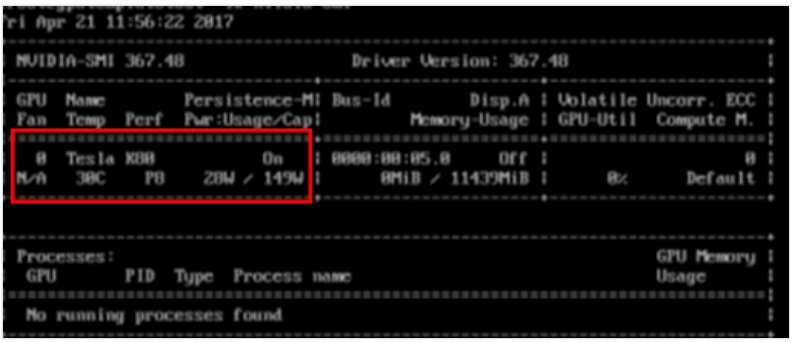
ucloud GPU Server는 Server 생성 즉시 별도의 환경 구축이 필요없이 바로 사용이 가능하도록 NVIDIA Driver 설치 및 고성능 설정이 적용된 자동화 이미지를 제공합니다.

ㅁ 자동화 이미지 적용 내용

NVIDIA CUDA Driver 8.0 이상 버전 설치 및 아래 고성능 설정
- nvidia-smi -pm 1(persistant mode on)
- nvidia-smi -auto-boost-default=0(auto boosts off)
- nvidia-smi -ac 2505, 875(최대 성능 설정)

ㅁ 자동화 설정 확인 방법

- nvidia-smi -pm1(persistent mode on)
- $ nvidia-smi 명령어 입력 후 Persistence-M이 On인 것 확인



- nvidia-smi -auto-boost-default=0(auto boosts off)
  $ nvidia-smi -q -d CLOCK 입력 후 Applications Clocks 확인
- nvidia-smi -ac 2505, 875(최대 성능 설정)
  $ nvidia-smi -q -d CLOCK 입력 후 2505, 875 확인

ㅁ 자동화 설정 해제 방법

- nvidia-smi -pm 0(persistant mode off)
- nvidia-smi -auto-boost-default=1(auto boosts on)
- nvidia-smi -ac '사용할 성능 수치 입력'(x,x 형태)