클러스터 시스템에서 GPU 사용 통계정보 획득 방안에 대한 연구

권민우*, 김성준*, 윤준원*, 홍태영*
*한국과학기술정보연구원 슈퍼컴퓨팅센터
e-mail:mwkwon81@kisti.re.kr

Study on the method of acquiring GPU usage statistics information in cluster system

Min-Woo Kwon*, Sung-Jun Kim*, JunWeon Yoon*, TaeYoung Hong*
*Dept. of Supercomputing Center, KISTI

요 약

한국과학기술정보연구원에서는 최근 빅데이터, 인공지능에 관한 연구 인프라 수요를 대응하기 위해 슈퍼컴퓨터 4호기 보조 가속기 시스템인 GPU 클러스터를 운영 중에 있다. GPU 클러스터 시스템은 사용자들 간에 효율적인 작업 배분을 위해 SLURM JOB 스케줄러를 이용하고 있다. 본 논문에서는 SLURM JOB 스케줄러를 통해 실행되는 사용자의 작업별 GPU 사용 통계 정보를 획득하는 방안에 대하여 소개한다.

1. 서론

최근 다양한 분야에서 빅데이터, 인공지능을 이용한 연구가 활발히 이루어지면서 GPU 클러스터 인프라 수요가 늘어나고 있다. 한국과학기술정보연구원에서는 이러한 수요에 대응하기 위해 슈퍼컴퓨터 4호기 보조 가속기 시스템을 운영하고 있다. 이 시스템은 32대의 서버로 구성되어 있으며 NVIDIA GPU인 TESLA V100 26대와 K40 18대가 장착된, 총 이론성능 220TFLOPS의 초고성능시스템이다[1]. 이 클러스터 시스템에는 SLURM JOB 스케줄러가 설치되어 사용자들 간에 효율적인 작업 배분이 이루어지고 있다[2]. 이러한 클러스터 시스템의 효율적인 운영을 위해 GPU 사용 통계 정보를 획득하고 관리하는 다양한 연구가 진행되고 있다[3-4]. 본 논문에서는 SLURM에서 생성되는 과금정보를 이용하여 사용자가 실행한 작업별 GPU 사용 통계정보를 획득하는 방안을 소개한다.

2. 계산노드별 GPU 사용 통계 정보 축적

NVIDIA GPU는 NVIDIA Management Library (NVML)를 이용해 모니터링이 가능하다[5].

그림 1과 같이 NVIDIA System Management Interface (nvidia-smi) 명령어는 NVML을 이용해 GPU의 모니터링 정보를 조회하고 출력한다.

nvidia-smi는 query기능을 이용하여 필요한 정보만을 출력할 수 있다. 본 논문에서는 GPU의 온도 (temperature.gpu), GPU코어 사용률 (utilization.gpu), GPU 메모리 사용률 (utilization.memory), 총 메모리 용량 (memory.total), 비어있는 메모리 용량 (memory.free), 사용 중인 메모리 용량 (memory.used)을 그림 2와 같이 획득하였다.

```
Thu Aug 23 10:54:34 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.111                 Driver Version: 384.111                  |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-PCIE...  On   | 00000000:1B:00.0 Off |                    0 |
| N/A   39C    P0    37W / 250W |  15418MiB / 16152MiB |     21%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-PCIE...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   40C    P0    36W / 250W |  15370MiB / 16152MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0     32317      C   python                                     15408MiB |
|    1     32317      C   python                                     15360MiB |
+-----------------------------------------------------------------------------+
```

(그림 1) nvidia-smi 출력 화면

```
$ nvidia-smi --query-gpu=timestamp,temperature.gpu,utilization.gpu,util
ization.memory,memory.total,memory.free,memory.used --format=csv
timestamp, temperature.gpu, utilization.gpu [%], utilization.memory [%],
 memory.total [MiB], memory.free [MiB], memory.used [MiB]
2018/08/23 11:10:37.096, 39, 20 %, 0 %, 16152 MiB, 734 MiB, 15418 MiB
2018/08/23 11:10:37.097, 40, 0 %, 0 %, 16152 MiB, 782 MiB, 15370 MiB
```

(그림 2) nvidia-smi query 출력 화면

```
20180823104001,1534988401,0.0,0.0,12205.0,12204.0,1.0
20180823104501,1534988701,0.0,0.0,12205.0,12204.0,1.0
20180823105001,1534989001,0.0,0.0,12205.0,12204.0,1.0
20180823105501,1534989301,0.0,0.0,12205.0,12204.0,1.0
20180823110001,1534989601,0.0,0.0,12205.0,12204.0,1.0
20180823110501,1534989901,0.0,0.0,12205.0,12204.0,1.0
```

(그림 3) 축적된 GPU 통계정보

위와 같은 nvidia-smi 명령어를 리눅스의 crond 데몬을 이용하여 5분마다 수행하여 계산노드 내부의 특정 위치에 저장한다. 그림 3은 저장된 GPU 사용 통계 정보를 보여준다.

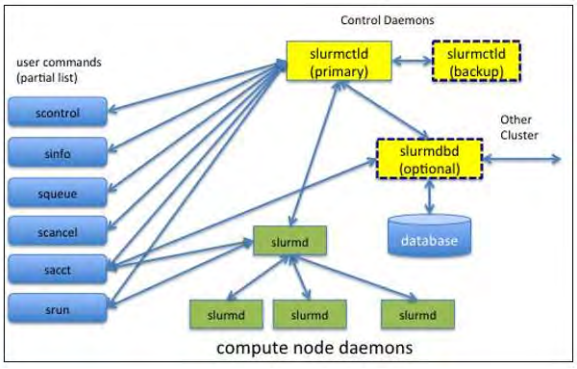
2. SLURM JOB 스케줄러의 작업 환경 정보 및 과금 정보 저장

SLURM JOB 스케줄러는 실행 중인 사용자의 작업에 대해서 scontrol 명령어를 이용해 그림 4와 같이 작업 환경 정보를 제공한다.

```
SubmitTime=2018-08-23T10:09:12 EligibleTime=2018-08-23T10:09:12
StartTime=2018-08-23T10:09:13 EndTime=2018-08-24T10:09:16 Deadline=N/A
PreemptTime=None SuspendTime=None SecsPreSuspend=0
Partition=bigmem_node AllocNode:Sid=login-tesla01:764
ReqNodeList=bigmem ExcNodeList=(null)
NodeList=bigmem
BatchHost=bigmem
NumNodes=1 NumCPUs=40 NumTasks=1 CPUs/Task=1 ReqB:S:C:T=0:0:*:*
TRES=cpu=40,node=1
Socks/Node=* NtasksPerN:B:S:C=0:0:*:* CoreSpec=*
MinCPUsNode=1 MinMemoryNode=0 MinTmpDiskNode=0
Features=(null) DelayBoot=00:00:00
Gres=(null) Reservation=(null)
OverSubscribe=NO Contiguous=0 Licenses=(null) Network=(null)
```

(그림 4) SLURM 스케줄러의 사용자 작업 환경 정보

또한 SLURM JOB 스케줄러는 SlurmDBD(Slurm Database Daemon)를 이용해 어카운팅 데이터를 아래 그림 5와 같이 데이터베이스에 저장한다[2].



(그림 5) SLURM 스케줄러 데몬 구조도

데이터베이스에서 작업별 과금 정보를 얻어오기 위해서는 SLURM JOB 스케줄러의 sacct 명령어를 사용한다. 아래 그림 6과 같이 sacct 명령어를 이용해 작업ID, 작업 시작 시간, 작업 종료 시간, 작업 상태 정보 등을 얻어올 수 있다[2].

```
$ sacct -T -X -ojobid,jobname,start,end,state
       JobID    JobName               Start                 End      State
------------ ---------- ------------------- ------------------- ----------
11282              bash 2018-08-23T00:00:00             Unknown    RUNNING
11348              bash 2018-08-23T00:00:00             Unknown  COMPLETED
11425              bash 2018-08-23T09:00:11             Unknown     FAILED
11426              bash 2018-08-23T10:37:47             Unknown    RUNNING
11434        IO_IB_BMT 2018-08-23T09:51:59             Unknown     FAILED
11435        IO_IB_BMT 2018-08-23T09:53:08             Unknown  COMPLETED
```

(그림 6) SLURM 스케줄러 sacct 출력 화면

4. Perl 스크립트를 이용한 사용자의 작업별 GPU 사용 통계 정보 획득

계산노드별로 축적된 GPU 사용 통계 정보와 SLURM JOB 스케줄러에서 제공하는 작업 환경 정보와 과금 데이터를 토대로 사용자의 작업별 GPU 사용 통계 정보를 획득할 수 있다. 본 논문에서는 Perl 스크립트를 이용하여 통계정보를 획득하였다. 작업 ID를 입력 받아서 SLURM JOB 스케줄러의 작업 환경 정보를 조회하여 작업이 동작한 계산노드 정보를 얻어온다. 또한 작업 ID를 입력 받아서 sacct 명령어를 이용해서 작업의 시작 시간, 종료 시간, 상태 정보를 얻어온다. 상태 정보가 정상적으로 종료된 작업인 경우에 작업의 시작 시간과 종료 시간을 기준으로 작업이 동작한 계산노드로부터 GPU 사용량 통계정보를 얻어온다. 모든 계산노드의 통계 정보를 합산하고 평균값을 구하여 작업의 전체 GPU 사용량 통계정보를 얻어 오게 된다. 그림 7은 15개의 작업에 대하여, GPU코어 사용률, GPU 메모리 사용률, 비어있는 메모리 용량, 총 메모리 용량, 사용 중인 메모리 용량의 평균을 구한 결과를 보여준다.

```
jobid:2203 uGPU:69.6  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2204 uGPU:66.1  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2207 uGPU:69.2  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2208 uGPU:74.4  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2209 uGPU:73.0  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2210 uGPU:76.5  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2211 uGPU:56.5  uMEM:1.3   mTOTAL:12205 mFREE:12045 mUSED:00160
jobid:2212 uGPU:86.8  uMEM:1.3   mTOTAL:12205 mFREE:12045 mUSED:00160
jobid:2213 uGPU:94.6  uMEM:1.3   mTOTAL:12205 mFREE:12045 mUSED:00160
jobid:2214 uGPU:100.0 uMEM:17.8  mTOTAL:12205 mFREE:10030 mUSED:02175
jobid:2215 uGPU:60.0  uMEM:19.2  mTOTAL:12205 mFREE:09864 mUSED:02341
jobid:2216 uGPU:70.4  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2217 uGPU:62.9  uMEM:95.6  mTOTAL:12205 mFREE:00533 mUSED:11672
jobid:2238 uGPU:0.0   uMEM:0.0   mTOTAL:12205 mFREE:12205 mUSED:00000
jobid:2255 uGPU:55.8  uMEM:19.2  mTOTAL:12205 mFREE:09864 mUSED:02341
```

(그림 7) 사용자 작업별 GPU 사용 통계 정보

참고문헌

[1] National Institute of Supercomputing and Networking, KISTI, http://www.nisn.re.kr

[2] slurm worload manager, SchedMD, https://slurm.schedmd.com/

[3] Andrew Barry “Resource utilization reporting” In Proc. Cray Users′ Group Technical Conference (CUG), 2013.

[4] Shinpei Kato, Scott Brandt, Yutaka Ishikawa, Ragunathan Rajkumar “Operating Systems Challenges for GPU Resource Management” In Proc. of the International Workshop on Operating Systems Platforms for Embedded Real-Time Applications, 23-32, 2011.

[5] NVIDIA System Management Interface, NVIDIA, https://developer.nvidia.com/nvidia-system-management-interface