클러스터 시스템의 장애 발생 계산노드 자동 복구 기능 구현

권민우*, 안도식*, 홍태영*
*한국과학기술정보연구원 슈퍼컴퓨팅인프라센터
mwkwon81@kisti.re.kr

Implementation of automatic recovery function for computiong node with failure of cluster system

Min-Woo Kwon*, Do-Sik An*, TaeYoung Hong*
*Dept. of Supercomputing Infrastructure Center, KISTI

요 약

한국과학기술정보연구원(이하 KISTI)의 국가슈퍼컴퓨팅센터에서는 슈퍼컴퓨터 5호기인 Nurion과 Neuron 시스템을 구축하여 국내 연구자들에게 서비스하고 있다. 이 중에서 Neuron 시스템은 GPU 클러스터 시스템으로 SLURM Batch Scheduler를 이용하여 공동활용서비스를 제공하고 있다. 본 논문에서는 Neuron에서 사용 중인 SLURM Batch Scheduler와 리눅스의 crontab 기능을 이용하여 소프트웨어 장애가 발생한 계산노드를 자동으로 복구시키는 기능을 구현하여 장애처리 대기시간을 단축시키는 기법에 대해서 소개한다.

1. 서론

KISTI의 국가슈퍼컴퓨팅센터에서는 슈퍼컴퓨터 5호기 메인시스템 Nurion과 보조시스템 Neuron을 운영하고 있다[1]. 이 중에서 Neuron 시스템은 GPU 기반 클러스터 시스템으로 AI 및 주요 계산과학 분야(소재, 바이오 등) 중에서 GPU 성능 가속이 뛰어난 분자동력학 및 전자구조계산 분야를 지원하고 있다. 2023년 기준 국내 사용자 495명이 AI 분야의 작업을 80%, 계산과학 분야의 작업을 20% 정도 수행하였다. 본 시스템은 국내 산·학·연·관 모든 연구자가 1년 3차례 사용자 신청(R&D 혁신지원 프로그램)을 통해 무상으로 사용가능한 시스템이다[2].

Neuron 시스템은 SLURM Batch Scheduler를 이용한 HPC 서비스와 AI 서비스 진입 장벽 해소를 위한 웹 기반 MyKSC AI 클라우드 서비스를 제공하고 있다[3]. Neuron 시스템은 CPU 기반 시스템인 Nurion과 Lnet 라우터를 이용한 인터커넥트 네트워크 및 스토리지 연동이 되어 있으며, 동일한 LDAP 서버를 사용하여 두 개의 시스템에 동시 접속이 가능하다. 이러한 환경에서 Nurion에서 수행한 HPC 관련 연구데이터를 Neuron에서 사용함으로 AI 기술을 적용한 혁신을 가속화하는 HPC+AI+Data Analytics 통합 서비스 환경을 제공하고 있다. 본 논문에서는 Neuron의 SLURM Batch Scheduler와 리눅스의 crontab 기능을 이용하여 단순 재기동을 통해 복구 가능한 소프트웨어 장애가 발생한 계산노드를 자동으로 복구시키는 기능을 구현하여 장애처리 대기시간을 단축시켜 운영 효율성을 극대화하는 기법에 대해서 소개한다.

2. 계산노드 장애 탐지 및 SLURM Partition 변경

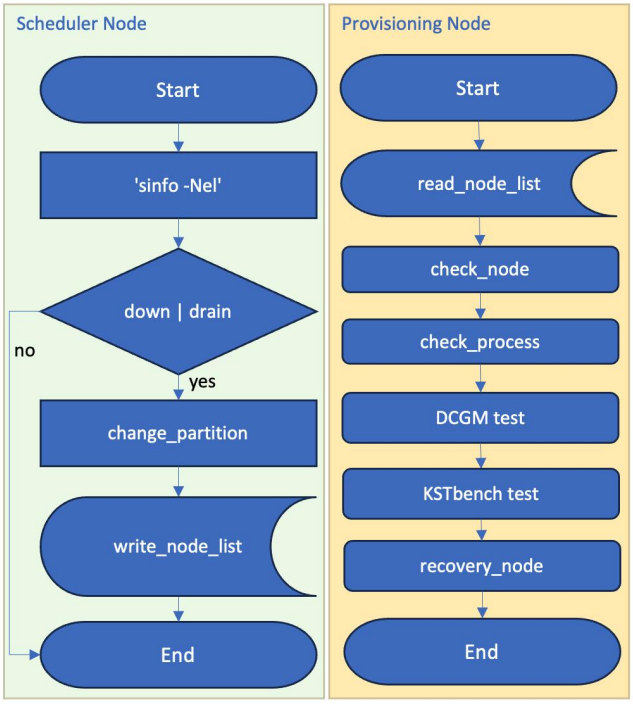
SLURM Batch Scheduler는 동일한 사양을 가진 계산노드들을 Partition이라는 명칭의 그룹으로 관리한다. Neuron 시스템에는 아래와 같은 Partition이 존재한다[4].

<표 1> Neuron SLURM Partition

Partition	설명
cas_v100nv_8	SXM V100 8GPU 장착 노드들
cas_v100nv_4	SXM V100 4GPU 장착 노드들
cas_v100_4	PCIe V100 4GPU 장착 노드들
cas_v100_2	PCIe V100 2GPU 장착 노드들
amd_a100nv_8	SXM A100 8GPU 장착 노드들
maintenance	장애처리 및 벤치마크 테스트

이 중에서 'maintenance'는 계산노드의 장애 발생 시에 임시로 이동시켜 장애 조치를 하거나 운영자가 벤치마크 테스트나 신규 시스템 SW(OS, GPU 드라이버, OFED, Lustre client 등)의 업그레이드 테스트를 수행할 목적으로 사용된다. 본 논문에서 제안하는 기법은 단순 재기동을 통해 복구가 가능한 소프트웨어 장애가 발생한 계산노드의 장애를 자동으로 탐지하고 자동으로 복구하는 기능을 이 Partition으로 이동시켜 수행한다.

그림 1은 slurmctld 데몬이 동작하고 있는 스케줄러 노드와 계산노드의 재기동을 담당하는 프로비저닝 노드에 구현되어 있는 장애탐지 및 자동복구 기능의 순서도이다. 이는 각각 리눅스 OS에서 제공하는 crontab 기능을 이용하여 주기적으로 수행되도록 구현되어 있다. 스케줄러 노드는 계산노드의 장애를 자동으로 탐지하고 탐지 시에 계산노드를 'maintenance' Partiton으로 자동으로 이동시킨다. Neuron과 같이 Batch Scheduler를 통해 운영되는 HPC 시스템은 사용자가 계산노드의 모든 자원을 극한의 상태까지 사용할 수 있는 시스템이다. 이런 HPC 시스템은 계산노드를 재기동하지 않고 오랜기간 가동 시에 노드가 hung 상태에 빠지는 단순 재기동을 통해 복구 가능한 소프트웨어 장애가 가끔 발생하며, 이는 리눅스의 ps 커맨드를 수행하여 쉽게 확인할 수 있다.



(그림 1) 장애탐지 및 자동복구 순서도

SLURM Batch Scheduler는 운영 중에 Partiton을 이동시키는 방식을 두 가지로 제공하고 있다. 첫 번째는 'slurm.conf' 파일의 내용을 수정 후에, 'scontrol reconfig' 커맨드를 수행하는 방법이 있고, 두 번째는 'scontrol update PartitionName=(이름) Nodes=(계산노드 목록)' 커맨드를 이용하는 방법이 있다. 본 논문에서는 Parititon 이동의 편의성을 위해 두 번째 방법으로 기능을 구현하였다.

'sinfo -Nel' 커맨드는 계산노드의 상태를 조회하는 커맨드이다. 계산노드의 기본적인 5가지 상태는 아래 표와 같다[5].

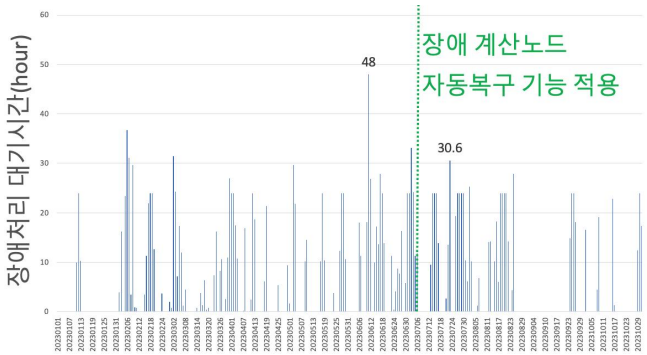
<표 2> SLURM 계산노드 상태

State	설명
idle	작업이 할당되지 않은 정상 상태
mix	작업이 자원의 일부를 점유한 상태
alloc	작업이 자원의 전체를 점유한 상태
down	스케줄러에서 빠진 상태
drain	계산노드에 장애가 발생한 상태

계산노드에 장애 발생시 'down'이나 'drain' 상태가 되는데, 이때 'maintenance'로 계산노드로 이동시키고 'node_list' 파일에 이력을 기록한다. 프로비저닝 노드는 스케줄러 노드가 기록하는 'node_list' 파일을 읽어서 장애 계산노드가 있는 경우, 계산노드의 상태와 계산노드에서 수행되는 프로세스의 상태를 체크한 후에 계산노드에 대한 recovery를 수행한다. 계산노드의 장애가 단순 재기동을 통해 복구 가능한 소프트웨어 장애인 경우, 계산노드를 리붓시켜 자동으로 복구시킨다. 계산노드 리붓이 완료되면, 서비스에 필요한 데몬을 올리고 노드의 상태를 점검하기 위한 테스트를 수행한다. Neuron의 계산노드에는 대부분 NVIDIA GPU가 탑재되어 있기에 NVIDIA 제공하는 DCGM(Data Center GPU Manager)를 이용해 GPU 카드의 하드웨어 장애 여부를 판별한다[6]. 추가적으로 국가슈퍼컴퓨팅센터에서 자체적으로 개발한 벤치마크 코드 템플릿인 KSTbench를 사용하여 종합적인 성능을 측정한다[7]. 이러한 테스트를 통해 계산노드의 상태를 점검 후에 정상상태로 판정이 되면 원래 Partition으로 이동시킨 후에 'scontrol update nodename=(계산노드) state=idle' 커맨드를 이용하여 계산노드를 'idle' 상태로 만들어 사용자의 대기 중인 작업에 할당될 수 있게 된다.

3. 장애 처리 대기시간 분석

그림 2는 장애 계산노드 자동복구 기능 적용 전후에 장애처리 대기시간 통계('23.01-10.)를 보여준다. 장애처리 대기시간이란 해당 일자에 장애가 발행한 노드들을 처리하는데 걸린 시간을 모두 합산한 시간이다. 이는 리붓을 통해 해결되는 단순 SW 장애뿐 아니라 그 밖의 SW 장애와 HW 교체가 필요한 장애처리 대기시간을 포함한다. 통계를 살펴볼 때, 자동복구 기능 적용 전보다 적용 후에 대략 5% 정도의 평균 장애처리 대기시간이 감소한 것을 볼 수 있다. Peak값의 경우는 자동복구 기능 적용전에는 50시간 가까이 나타났으나 적용후에는 30시간 이내로 40% 가까이 감소한 것을 확인할 수 있다. Neuron 시스템의 경우, 국가 주요정보 통신 시설로서, 슈퍼컴퓨터 종합상황실에서 365일 24시간 모니터링을 하고 있으며, 장애 발생 시에 운영자에게 즉각적으로 통보가 되기 때문에 복구 기능 적용 전후에 평균 장애처리 대기시간에 큰 차이를 보이지는 않는다. 그러나 이러한 시스템이 갖추어지지 않은 경우에는 운영자의 수동 작업에만 의존하기에 야간이나 주말, 공휴일에 장애처리 대기시간이 증가하고 이에 따라 시스템의 가동률이 저하될 수 있다. 그러나 본 논문에서 제안하는 자동복구 기능을 통해 이러한 문제점을 해결할 수 있을 것으로 기대한다.



(그림 1) 장애처리 대기시간 통계('23.01.-10.)

4. 결론 및 향후 연구 방향

본 논문에서는 계산자원의 가동률을 극대화하기 위하여 복구 가능한 소프트웨어 장애가 발생한 계산노드를 자동으로 복구시키는 기능을 제안하였다. 제안된 기법을 통해 장애처리 대기시간을 단축시켜 운영 효율성을 극대화할 수 있었다. GPU의 경우, 다른 부품들과 달리 상태정보를 획득하는데 어려움이 있다. 향후에는 GPU의 상태 정보를 분석하여 장애 복구 기능을 고도화시키는 연구를 수행할 예정이다.

사 사

이 논문은 2024년도 한국과학기술정보연구원의 기본사업(과제명:국가 플래그십 초고성능컴퓨터 인프라 구축 및 서비스, 과제번호:K24L2M1C1)으로 수행된 연구입니다.

참고문헌

[1] KISTI 국가슈퍼컴퓨팅센터 홈페이지, 보유자원, https://www.ksc.re.kr/byjw/sg
[2] KISTI 국가슈퍼컴퓨팅센터 홈페이지, 혁신지원, https://www.ksc.re.kr/jwsc/hsjw/hsjwan
[3] KISTI 슈퍼컴퓨터 웹 서비스 포털, MyKSC, https://my.ksc.re.kr/
[4] KISTI Neuron지침서, 스케줄러를 통한 작업실행, https://docs-ksc.gitbook.io/neuron-user-guide/undefined/running-jobs-through-scheduler-slurm
[5] SLURM Batch Scheduler Manual, sinfo, https://slurm.schedmd.com/sinfo.html
[6] NVIDIA Developer, DCGM, https://developer.nvidia.com/dcgm
[7] KSTbench Github, Benchmark Templates https://github.com/vitduck/KSTBench/blob/main/README.md