효율적인 배치 작업 정보 관리를 위한 모니터링 시스템 설계

김성준*, 이재국*,홍태영*
*한국과학기술정보연구원 슈퍼컴퓨팅인프라센터
{sjkim,jklee,tyhong}@kisti.re.kr

Design of efficiency monitoring system for managing batch job information

Sung-Jun Kim*, Jae-Kook Lee*, Tae-Young Hong*
*Dept of Supercomputing Infra, KISTI

요 약

한국과학기술정보연구원에서는 슈퍼컴퓨터 5호기 시스템 및 가속기 기반 시스템을 국내 연구자들에게 서비스를 하고 있다. 시스템 관리자들은 시스템 상태 조회 및 통계 정보 산출등의 목적으로 배치 작업 관리 솔루션에 주기적으로 다양한 정보의 요청을 수행한다. 빈번한 정보 요청은 작업관리 솔루션에 부하를 줄 수 있다. 본 논문에서는 사용자들의 배치 작업 관리를 위해 사용하는 배치 작업 관리 솔루션인 PBSPro와 SLURM을 활용한 효율적인 시스템 모니터링 기법을 설계하고자 한다.

1. 서론

한국과학기술정보연구원(이하 KISTI)에서는 슈퍼컴퓨터 5호기(이하 누리온) 시스템과 가속기 기반 시스템(이하 뉴론)을 국내 연구자들에게 서비스를 제공하고 있으며, 시스템 사양은 아래와 같다.

<표 1> 누리온(Nurion) 시스템 사양

구분		KNL	SKL
모델명		Cray CS500	Cray CS500
노드 수		8,305	132
이론 성능		25.3 PFlops	0.4 PFlops
CPU	모델명	Intel Xeon Phi 7250	Intel Xeon Gold 6148
	이론 성능	3.0464 TF	1.536 TF
	core수	68	20

<표 2> 뉴론(Neuron) 시스템 사양

모델명		Lenovo nx360		HP Apollo 6500
노드 수		21		28
메인메모리		128GB		384GB
CPU 모델		Intel Xeon IvyBridge		Intel Xeon Sky Lake, Cascade Lake
GPU	모델명	Tesla V100		Tesla V100
	HBM2	16GB	32GB	32GB
	GPU수 (노드당)	2개	2개	1x4대,2x15대, 4x4대,8x 5대

시스템 관리자들은 시스템 모니터링, 사용자 작업 현황 조회 및 기술 지원의 목적으로 배치 작업 관리 솔루션으로부터 다양한 정보를 주기적 혹은 비정기적으로 요청을 하고 있다.

또한, 사용자들이 현재 시스템의 가용 노드 수 및 큐별 대기작업 수 및 수행 작업수를 쉽게 파악하여, 빠른 작업 수행을 위해서 작업 제출을 위한 큐를 결정하는데 도움이 되는 정보를 제공한다. 이러한 정보들은 대표 홈페이지내 시스템 현황 정보 제공 및 서비스 대쉬보드에서 확인할수 있도록 하고 있다.

기본적으로 사용자 작업과 관련한 모든 정보들은 배치 작업 관리 솔루션이 제공하기 때문에 정보 요청에 대한 응답시 마스터 데몬에 대한 부하가 필연적으로 발생할 수밖에 없는 상황이다. 마스터 데몬은 사용자 작업에 대한 관리의 역할도 담당하고 있기 때문에, 부하로 인한 데몬의 성능 저하는 전체적인 서비스 성능 저화와 연계가 된다.

본 고에서는 이러한 배치 작업 관리 솔루션의 마스터 데몬의 부하 경감을 위한 효율적인 정보 관리를 위해 관리자 및 사용자를 위한 정보 요청을 체계화하고 단일화하는 모니터링 시스템을 설계함으로써,

마스터 데몬의 부하를 경감하고 보다 안정적인 서비스 환경을 제공하고자 한다.

2. 관련 연구

가.PBSPro

Altair PBSPro는 HPC 클러스터, 클라우드 및 슈퍼컴퓨터의 생산성을 개선하고 리소스 활용도와 효율성을 최적화하며 작업 부하 관리 프로세스를 간소화하도록 설계가 되었다. PBSPro는 세계 수준의 슈퍼컴퓨터 및 소규모 클러스터 소유자를 위한 신뢰할 수 있는 솔루션으로, 작업 스케줄링, 관리, 모니터링의 기능을 제공한다. [1)]

나.slurm

이전에 Simple Linux Utility for Resource Management 또는 간단히 Slurm으로 알려진 Slurm Workload Manager는 전 세계의 많은 슈퍼컴퓨터 및 컴퓨터 클러스터에서 사용되는 Linux 및 Unix 유사 커널 용 무료 오픈 소스 작업 관리 프로그램이다.[2)]

다.MongoDB

MongoDB 는 무료 오픈 소스 플랫폼 문서 지향 데이터베이스이다. NoSQL의 일종으로 확장성이 좋고 성능이 우수하며, JSON과 같은 동적 스키마형 문서들을 선호함으로써 전통적인 테이블 기반 관계형 데이터베이스들(RDMS)과는 차별화 된다.[3)]

3. 운영 현황

관리자들은 기본적으로 제공되는 명령어를 활용하기도 하지만, 원하는 형태의 정보를 표출하는 프로그램을 쉘스크립트나 파이썬 등을 활용하여 작성하여 사용한다.

가. PBSPro 모니터링 q

(1) 기본 명령어

○ qstat

qstat은 PBS 작업의 상태 정보를 조회하는 명령어로 여러 옵션을 조합하여 다양한 정보를 제공하는 명령어이다.

○ pbsnodes

pbsnodes는 계산 노드들의 유휴,장애 상태 등 노드 상태 정보를 제공하는 명령어이다.

(2) 활용 예제

다음은 주로 사용하는 기본 명령어와 함께 사용되는 옵션들의 조합과 추출 가능한 정보들에 대하여 보여준다.

○ qstat -F : 작업 상태 상세 정보 표출

- 작업아이디,소유자,큐,상태,시간(제출,시작,종료),요구자원규모,사용자원규모,사용노드정보 등

○ qstat -w -n1 : 수행중인 작업 정보

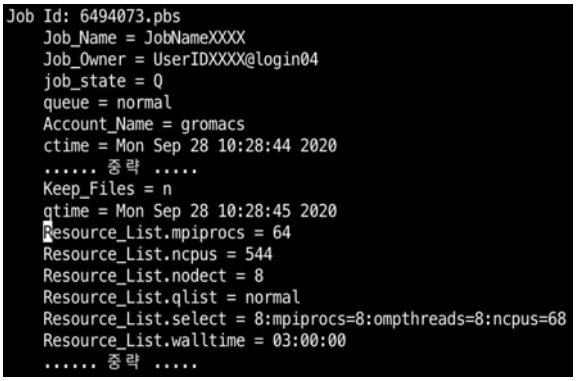
- 작업아이디,사용자,큐,사용노드정보

○ qstat -w -s1 : 대기중인 작업 정보

- 작업아이이,사용자,큐,대기 사유

○ pbsnodes -av : 노드별 상세 정보

- 노드명,수행작업아이디,관련 큐이름,상태 등



<그림 1> qstat -f 명령어 실행 예시 [PBSPro]

나. slurm 모니터링

(1) 개요

slurm은 확장성을 제공하기 위해서 개발자용 라이브러리(perl)을 제공하고 있으며, 이를 활용하여 원하는 warpper 프로그램을 개발할 수 있는 기능을 제공하고 있다.

(2) 기본 명령어

○ qstat

slurm에서 제공하는 perl 라이브러리를 활용하여 pbs의 qstat과 유사한 결과를 제공하는 명령어

○ sinfo

계산 노드의 상태(유휴,장애,할당)정보를 제공하는 명령어

(3) 활용 예제

○ qstat -f :작업 상태 상세 정보 표출

- 작업아이디,소유자,큐,상태,시간(제출,시작,종료),요구자원규모,사용자원규모,사용노드정보 등

○ sinfo -o : 노드별 상세 정보
- 노드명,수행작업아이디,관련 큐이름,상태 등

```
Job Id: 43710
  Job_Name = reverse
  Job_Owner = USERXXXX@glogin02
  job_state = R
  queue = skl_v100_2
  qtime = Mon Sep 28 13:47:56 2020
  mtime = Mon Sep 28 13:47:59 2020
  ctime = Sat Oct  3 13:48:02 2020
  Account_Name = kat_user
  exec_host = gpu43/40+gpu44/40
  Priority = 4294888305
  euser = USERXXXX(123123123)
  egroup = ACCOUNT00(1231231)
  Resource_List.walltime = 120:00:00
  Resource_List.nodect = 2
  Resource_List.ncpus = 80
```

<그림 2> qstat -f 명령어 실행 예시[slurm]

다. 배치 작업 프로그램 정보 획득 방법
- 상태 정보 백업 : cron을 활용하여 주기적인 명령어 실행을 통한 정보 수집
- 시스템 상태 조회 : 시스템 모니터링을 위해서 특정 상황 발생시 명령어를 실행하여 시스템 상황 파악에 활용

라. 모니터링시 필요 정보
○ 작업 관련
- 작업아이디, 사용자(이이디,계정), 큐, 자원정보(코어,노드,메모리),응용코드, {제출,시작,종료}시간, 사용 노드

○ 노드 관련
- 노드 상태{가용,유휴,장애}, 연관 큐정보

○ 외부 연동 정보
- 사용자별 {계약종류,소속,지역}정보

4. 시스템 설계

시스템 관리 및 시스템 상태 백업 및 모니터링등의 다양한 목적으로 배치 작업관리 프로그램으로부터 중복하여 수집되던 정보를 주기적으로 일괄 수집하여 데이터베이스에 저장함으로써, 작업관리 데몬의 부하를 경감할 수 있도록 시스템을 설계하였다.

이렇게 저장되는 정보들은 시스템 상태 정보의 아카이빙의 성격도 가지고 있으며, 향후 다양한 목적을 의해서 재활용할 수도 있을 것이다.

가. 수집부

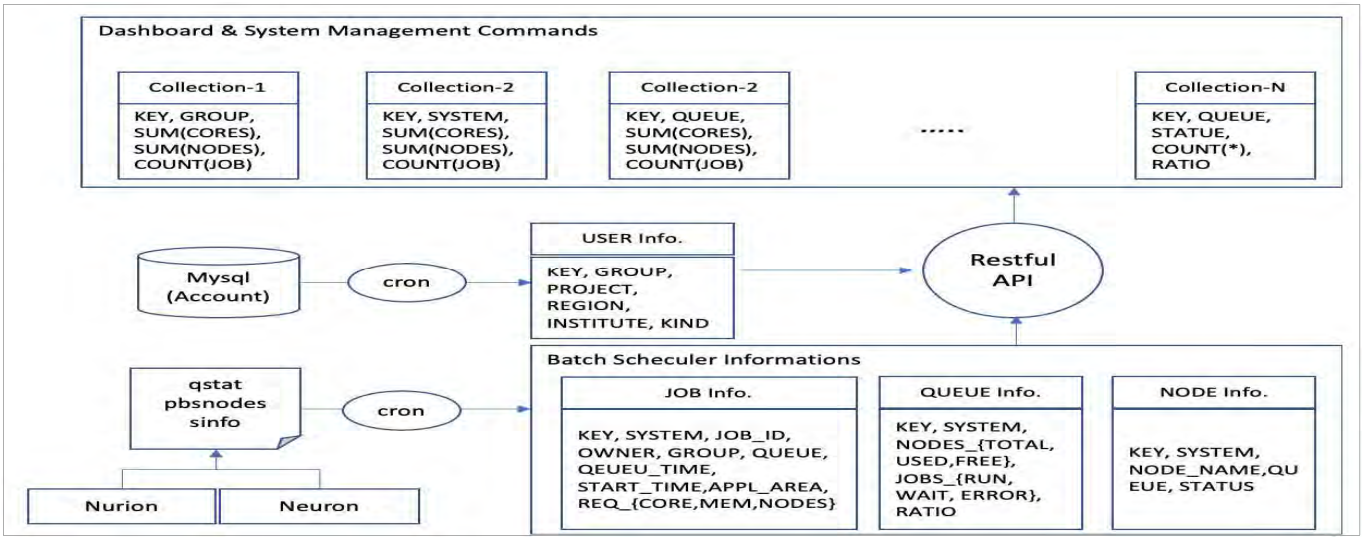
수집부는 각 시스템별로 운영되는 배치작업프로그램의 명령어를 주기적으로 수행하여, 상세 정보를 파싱하여 데이터베이스에 저장한다.

나. 저장부

데이터베이스는 다양한 정보의 저장을 위해서 비정형데이터베이스인 MongoDB를 활용하고, 기본적으로 시간의 역순으로 저장하여 관리한다.

다. 가공부

가공부는 표출하고자 하는 정보에 따라서 참조되는 컬렉션(테이블)을 조합하여 원하는 정보를 생성하며, 생성된 정보는 Restful-API의 형태로 요청/응답하는 형태로 구성한다.

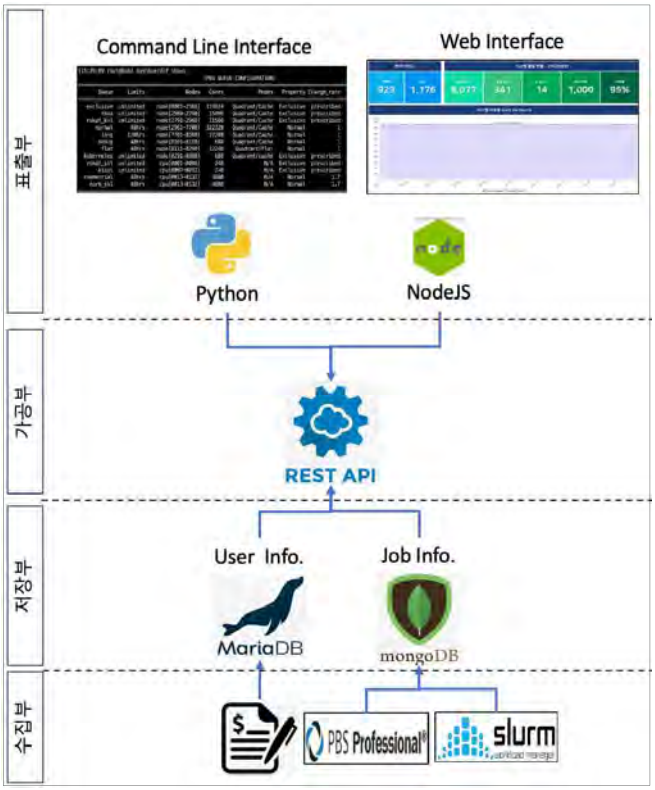


<그림 3> 데이터 흐름도

라. 표출부

표출부는 가공부와 요청/응답을 하며, 웹을 통해서 표출하는 대쉬보드나 관리자가 주로 사용하는 명령어 라인 프로그램으로 작성하여 운영한다. 표출부에서 제공할 수 있는 정보는 다음과 같다.

- o 시스템별 노드 활용 현황
  - 활용/유휴/점검노드 현황
- o 사용자별 시스템 활용 현황
  - 계약종류별/응용분야별/과제종류별
- o 시스템 사용율(Last 24Hour)
- o 큐별 사용자 작업 현황 정보 등



<그림 4> 시스템 구성도

향후 개발할 시스템은 <그림 4>와 같이 개발할 예정이 있다. 수집된 데이터는 restful-api를 이용하여 접근하여, 정보의 표출은 웹기반 혹은 명령어 기반으로 목적에 따라서 개발되어 제공될 예정이다.

5. 결론

본고에서 설계한 모니터링 시스템은 일차적으로 빈번한 정보 요청으로 인한 배치 작업 프로그램의 마스터 데몬의 부하를 경감시키는데 목적을 둔다. 하지만, 이를 통해 시스템 상태 정보 및 작업 정보들에 대한 주기적인 백업을 통해서 향후 다양한 목적으로 재활용할 수 있는 가능성을 제공할 수 있다.

참고문헌

[1] https://www.altair.co.kr/pbs-professional/
[2] https://slurm.schedmd.com/documentation.html
[3] https://commin.tistory.com/78
[4] MongoDB, http://www.mongodb.com
[5] 김성준,홍태영, “RESTful API를 이용한 슈퍼컴퓨터 서비스 데시보드 구축”, 추계학술발표대회,2019,180p-182p