국가 슈퍼컴퓨팅 공동활용 환경을 위한 통합 모니터링 환경 구축

Implementation of Resource Monitoring System for PLSI

김성준, 성진우, 장지훈, 이상동
한구과학기술정보연구원 슈퍼컴퓨팅사업팀

Sung-Jun Kim, Jin-Woo Sung, Ji-Hoon Jang, Sang-Dong Lee
KISTI Supercomputer Management Dept.

요약

국내 슈퍼컴퓨팅자원을 상호 연동하여 국가 과학기술 개발에 활용함으로써 공공 슈퍼컴퓨팅자원의 활용을 극대화 하기 위해 추진되고 있는 국가슈퍼컴퓨팅공동활용체제구축사업(PLSI)에서는 여러 기관의 자원을 국내 연구자들에게 서비스를 할 예정에 있다. PLSI사업에 참여한 기관들은 각자 환경에 적합한 시스템 모니터링 시스템을 구축하고 있으나 각 기관간에 서로 상이한 환경을 운영되고 있기 때문에 여러 기관들의 시스템 정보를 한번에 파악하기는 어려운 상태이다. 본고에서는 통합 관제센터에서 단일 GUI 환경을 통하여 PLSI 사업에 참여한 여러 기관들의 자원의 상태 및 사용자 작업 정보 등을 용이하게 파악하기 위한 통합 모니터링 시스템을 구축하였다.

Abstract

PLSI(Partership & Leadership for national-wide Supercomputing Infrastructure) Project is about providing computational resources which maintained by partner institutes to researchers. Each institutes use system monitoring tools adopted their environment different others, it is difficult to understand all system's status at a glance. In this paper, we have designed and implemented monitoring system which system administrator easily understanding all resources status like CPU Load, Job status, etc. in PLSI use web browser.

I. 서 론

국내 슈퍼컴퓨팅자원을 상호 연동하여 국가 과학기술 개발에 활용함으로써 공공 슈퍼컴퓨팅자원의 활용을 극대화 하기 위해 추진되고 있는 국가슈퍼컴퓨팅공동활용체제구축사업(PLSI)에서는 여러 기관의 자원을 국내 연구자들에게 서비스를 할 예정에 있다. 현재 부산대학교의 자원과 한국과학기술정보연구원의 자원이 연동 작업 중에 있으며 향후 연동된 자원을 공동활용하여 국내의 연구자들에게 제공할 예정에 있다.

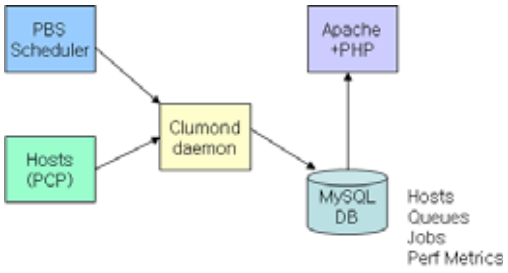
각 기관에서 운영되는 슈퍼컴퓨팅 자원들은 각 시스템 도입시에 설치한 전용 모니터링 시스템을 이용하여 로컬 사이트에서 자원에 대한 정보를 모니터링을 하고 있는 상황이다. 하지만 각 기관에서 운영하는 모니터링 프로그램이 서로 상이하여 하나의 통합 관제 센터에서 통합하여 자원의 상태를 파악하기에는 다소 어려움이 존재하고 있다. 현재 한국과학기술정보연구원의 Hamel 리눅스 클러스터는 clumon을 사용하고 있으며, 부산대학교의 Daisy 시스템은 ganglia를 이용하여 자원 사용 상황 및 시스템 상태를 모니터링하고 있다.

본 고에서는 이기종 모니터링 시스템을 통합하여 단일 GUI 환경을 통하여 여러 자원들의 자원 및 시스템 내부 상태를 모니터링 함으로써, 향후 설치되어서 운영될 PLSI 통합 관제 센터에서 PLSI 산하 기관들의 슈퍼컴퓨팅 자원들을 효율적으로 통합 관제하기 위한 통합 모니터링 시스템을 구축을 설명한다.

II. 관련 연구

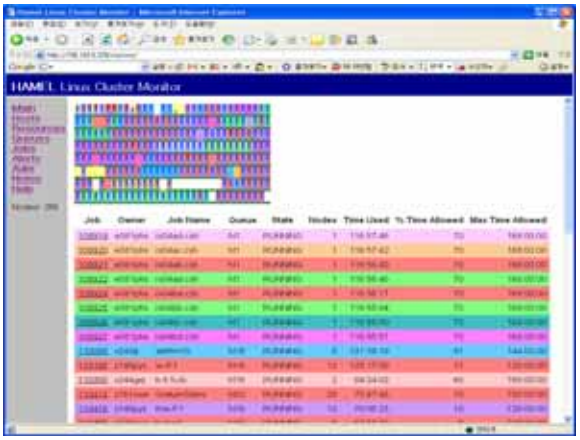
1. Clumon

Clumon은 NCSA에서 개발한 리눅스 기반 클러스터 모니터링 프로그램이다. SGI의 Performance Co-Pilot 과 PBS 스케쥴러를 기반으로 개발되었으며 MySQL+Apache +PHP를 이용하여 구현되었다. 클러스터의 호스트들의 정보와 작업 정보들을 대략적인 현재 상태를 한눈에 파악할 수 있도록 구성되었다.



▶▶ 그림 1. Clumon Architecture

Clumon 설계시 고려된 사항은 확장성, 낮은 대역폭, 구성의 자유로움, 관리의 용이성이었다. 수천 노드이상의 호스트들에서 정보를 동시에 추출하는 것은 호스트 자체나 네트워크 레이어에 많은 부하를 야기할 수 있는 문제가 내포되어 있기 때문에 확장성이 가장 큰 문제로 고려되었다. 또한 성능 데이터를 모니터링 하는 것은 정밀도, 시스템 규모, 수집 빈도와 시스템의 부하와 소모되는 네트워크 대역폭간의 반비례 관계에 있다. 모니터링 시스템의 관리자는 수집되는 데이터의 양과 수집 빈도를 운영 중인 시스템에 미치는 영향을 고려하여 조절 가능하도록 구성하였다.



▶▶ 그림 2. 하멜에서 운영 중인 Clumon 화면

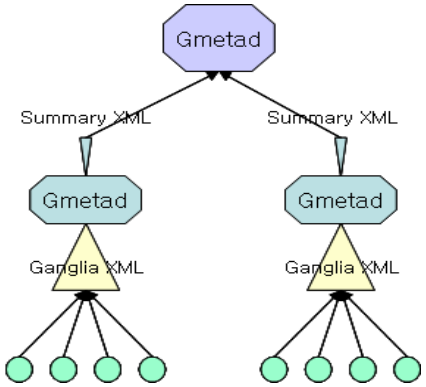
2. Ganglia

Ganglia(ganglion의 복수)는 생명체의 신경절을 의미하는 영어 단어로 클러스터의 사용 상태를 모니터링해 주는 도구이다. UC Berkerley의 Millennium Project에서 개발하였으며, 전 세계의 오픈 소스 공동체인 SourceForge.net을 통해 소스 코드가 공개되어 자유롭게 사용할 수 있는 소프트웨어이다.

Ganglia는 클러스터군과 그리드 자원들과 같은 HPC 시스템들을 위한 분산 모니터링 시스템으로 클러스터 군들의 관리에 중점을 둔 계층적 구조로 이루어져 있다.

ganglia는 외부 데이터 전송방식을 XDR을 이용하고 있으며 간소화, 이식성을 높이기 위해서 데이터의 표현을 XML형식을 취하고 있으며, 데이터의 저장과 가시화를 위해서 RRDtool을 이용하고 있다.

ganglia는 낮은 오버헤드와 높은 데이터 안정성을 제공하기 위해서 잘 설계된 데이터 구조와 알고리즘을 사용하고 있다. 이러한 이유로 많은 전 세계적으로 많은 사이트에서 클러스터를 모니터링하기 위해서 사용되고 있다.



▶▶ 그림 3. Ganglia monitoring Tree

▶▶ 그림 4. ganglia 운영 화면

III. 설계

1. 시스템 구성

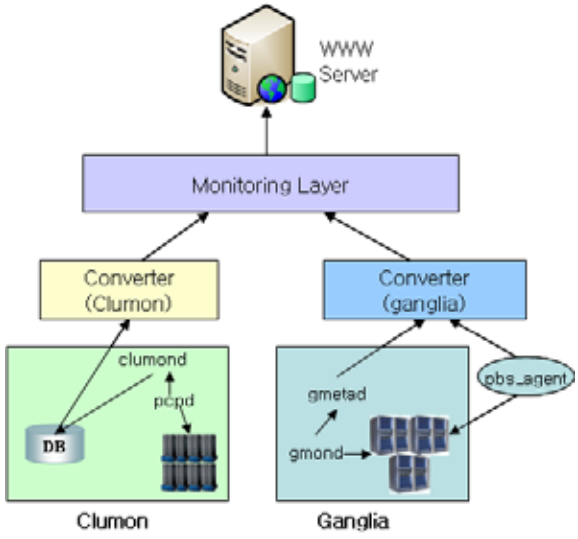
PLSI에 가입한 기관들은 이미 각자의 자원을 모니터링하기 위한 시스템 모니터링 프로그램을 사용하고 있다. 이렇기 때문에 별도의 모니터링 프로그램을 설치하는 것은 이중으로 자원 정보를 수집하면서 로컬 시스템에 부하를 줄 수도 있다. 또한, 여러 기관의 시스템 운영 정책이 서로 상이하기 때문에 별도의 모니터링 프로그램을 설치하는 것은 적합하지 않았다.

따라서, 본 고에서 구현하는 통합 모니터링 시스템은 기존에 로컬에서 운영 중인 모니터링 시스템은 수정하지 않으며 각 모니터링 프로그램 수집한 정보를 가공하여 통합관제 센터에서 각 기관의 자원 및 작업의 상태를 파악하는 것에 중점을 두었다.

아래 그림은 통합 모니터링 시스템의 구성도이다. 그림에서 보는 바와 같이 기존 clumon을 사용하는 기관의 경우 clumon이 사용하는 데이터베이스에서 통합 모니터링에 필요한 항목을 추출하여 통합 모니터링을 위한 데이터베이스로 복사해서

웹을 통해서 PSLI 통합관제 센터 쪽에서 관리 할 수 있도록 구성하였다.

ganglia의 경우 모니터링 정보를 데이터베이스를 사용하지 않으며, RRDTool을 이용하여 파일로 저장하고, 모니터링된 정보들은 gmetad 데몬으로 요청을 하면 시스템들의 상태 정보를 XML의 형태로 반환해 주는 동작한다. ganglia의 경우에는 이 gmetad로부터 시스템들의 상태 정보를 전달받아서 이를 통합모니터링에서 사용하는 데이터베이스에 시스템의 상태 정보를 추가 하였다.



▶▶ 그림 5. 시스템 구성도

작업정보의 경우 clumon은 내부 데이터베이스에 저장되어 있는 정보를 통합 모니터링의 데이터베이스로 복사해왔으며, ganglia의 경우 별도의 agent(pbs_agent)를 해당 시스템에 설치하여 시스템에서 처리 중인 사용자들의 작업 정보를 전달받아서 가공 후 통합 모니터링 데이터베이스의 작업 관련 테이블에 추가하는 형태로 구성하였다.

2. 모니터링 항목

현재 사용 중인 clumon과 ganglia의 모니터링 항목은 서로 상이하게 구성되어 있다. 각 시스템에 대한 자세한 정보는 해당 시스템의 전용 모니터링 시스템을 이용하면 가능하다. 따라서, 통합 모니터링에서 두 시스템의 모든 모니터링 항목을 모니터링 하는 대신에 전체 시스템에 대한 개략적인 상태 정보만은 모니터링 하는 형태로 구성하였다.

[표 1] 각 시스템의 모니터링 항목

ganglia	clumon
boottime	System Load
bytes_in	Free Memory
bytes_out	Total syscalls
cpu_nice	PPID
cpu_system	Defuct Process
cpu_user	Operating System
..... (중략)	Swap out
proc_run	Swap in
swap_free	Processors Speed
swap_total	..... (중략)
cpu_speed	CPU Clockspeed
disk_free	CPU Vendor
mem_free	OS Release
cpu_num	Physical Memory
load_one	Number of Processors
mem_total	FileSys Used
disk_total	FileSys Capacity

위의 표 1에서와 같이 두 모니터링 시스템은 서로 다른 형태로 모니터링 항목을 정의하고 있으며, 일부에서만 유사한 항목을 모니터링 하고 있다.

본고에서는 다음 표 2 와 같이 시스템 로드, 메모리 사용량, 파일시스템 사용량, 프로세서 정보와 같은 공통 모니터링 항목을 정의하여 각 시스템의 개략적인 상태 정보를 추출하여 통합 모니터링을 구성하도록 하였다.

[표 2] 모니터링 항목 비교

제안 시스템	clumon	ganglia
CPU Speed	CPU Clock	cpu_speed
CPU Vendor	CPU Vendor	machine_type
OS Version	OS Release	os_release
Total Memory	Total Memory	mem_total
Number of CPUs	No of CPUs	cpu_num
CPU Load	System Load	load_one
Free Memory	free memory	mem_free
Filesystem Used	filesys used	disk_free
Filesystem Capacity	filesys Capacity	disk_total

IV. 구현

1. 구현 환경

아래 표 3은 구현환경을 보여주고 있다. 데이터 추출을 위한 프로그램들은 perl을 이용하여 작성되었으며 ganglia의 작업 정보 요청을 처리하기 위한 pbs_agent는 c로 작성되어 데몬

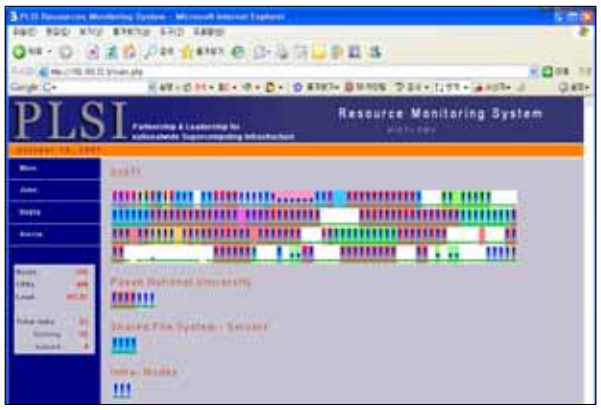
의 형태로 항상 메모리에 띄워져 사용된다.

[표 3] 시스템 구현 환경

	웹 인터페이스	Data 추출 Agent
개발 언어	php	perl / C
DataBase	MySQL	
WebServer	Apache	

ganglia의 경우 별도의 데이터베이스가 존재하지 않기 때문에 clumon의 데이터 구조를 기본으로 하여 ganglia의 정보를 통합 데이터베이스에 추가하는 형태로 시스템을 구성하였기 때문에 웹 인터페이스를 clumon의 웹 인터페이스를 기반으로 하여 수정/개발 하였다.

2. 구현 환경



▶▶ 그림 6 main 화면

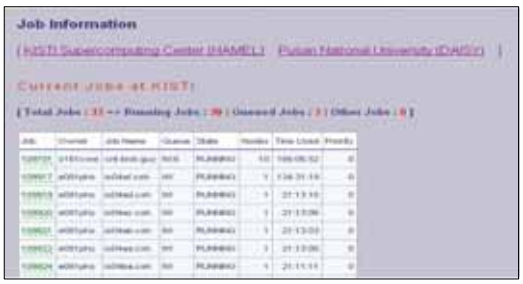
그림 6은 구현한 시스템의 메인 화면이다. 메인화면에는 현재 구성된 KISTI의 Hamel 클러스터 시스템과 부산대학교의 SGI 시스템을 포함하여 두 시스템 간의 공유 파일 시스템의 서버 장비들의 상태정보와 보조 장비인 어카운팅 서버와 통계 서버와 같은 보조 장비들의 상태 정보를 표시하고 있다.

▶▶ 그림 7. 호스트 정보 목록 페이지

그림 7은 선택한 사이트의 호스트들의 정보 목록을 보여주고 있다. 제공되는 정보는 노드명, 운영체제 종류 및 버전과 CPU 속도, 메모리 용량, CPU 개수와 1분간의 시스템 로드 정보를 보여 주고 있다.

▶▶ 그림 8. 호스트 상세 정보 페이지

그림 8을 선택한 노드의 상세 정보를 보여주고 있다. 제공되는 정보는 시스템 로드 정보와 메모리 사용량을 그래프의 형태로 보여주고 있으며, 노드의 하드웨어 정보와 시스템의 상태 정보를 보여 주고 있다.



▶▶ 그림 9. 작업 정보 페이지

그림 9는 해당 사이트에서 수행 중인 사용자들의 작업 목록을 보여주는 페이지이다. 현재 수행 중인 작업 개수와 대기 중인 작업 수 및 각 작업들이 사용하는 노드의 수량과 작업 수행 시간 등을 나열하고 있다.

▶▶ 그림 10. 작업 상세 정보 페이지

그림 10은 해당 작업의 상세 정보를 보여 주고 있다. 선택된

작업이 요구한 자원에 대한 정보 및 실제 할당된 자원의 정보들이 표시되고 있으며, 수행되는 노드들에 대한 정보가 그래프의 형태로 제공된다.

V. 결 론

본 논문은 국가슈퍼컴퓨팅공동활용체제구축사업(PLSI)에 참여한 기관들의 컴퓨팅자원의 상태를 통합관제 센터에서 효율적으로 관제하기 위한 통합 모니터링 시스템을 구현하였다. 각 기관의 모니터링 체계는 유지하면서 각 기관자원의 공통된 자원 정보만을 추출하여 일관된 정보를 제공하도록 구성하였으며 상세한 정보는 해당 기관 모니터링 시스템을 이용하도록 구현하였다.

이를 이용하여 관제센터에서는 보다 쉽게 한눈에 전체 자원에 대한 상태 정보를 파악할 수 있으며, 전체 자원에서 수행중인 사용자들의 작업 상태도 파악할 수 있다. 향후에는 이후 추가적으로 가입되는 기관들의 자원 정보 및 사용자 작업정보도 모니터링하는 기능을 추가할 예정이며 자원의 활용과 사용자 작업 패턴을 분석할 수 있는 통계 정보도 추가적으로 개발할 예정에 있다.

▌참 고 문 헌▐

[1] 서현수,“MySql 시스템 관리와 프로그래밍”,한빛미디어, 2002
[2] 박준철 역,“TCP/IP 소켓프로그래밍 Version C”,사이텍미디어, 2001
[3] Bill Evjen, Kent Sharkey, etc, “Professional XML” WROX Press, 2001
[4] Matthew L. Massie, “The ganglia distributed monitoring system:design, implementation, and experience”, PARALLEL COMPUTING, 2004
[5] http://clumon.ncsa.uiuc.edu/
[6] http://ganglia.sourceforge.net/