# HPC 클라우드 환경에서 서비스 기반의 타입에 따른 리소스 관리방안 연구

손아영,조혜영,박준영,정기문*

한국과학기술정보연구원

{ayson, chohy, jypark, kmjeong}@kisti.re.kr

# A study on Service-aware Resource Management according to resource type in HPC Cloud

A-Young Son, Hyeyoung Cho, Junyoung Park, Gi-Mun Jeong*

Korea Institute of Science and Technology Information

요 약

본 논문은 한정된 계산 자원을 여러 사용자들이 효율적으로 사용하기 위해서는 실시간으로 자원상태를 파악하여 작업 수행에 맞는 시스템을 할당해야 한다. 과학적 응용 프로그램들이 데이터 집약적으로 발전하고 있으며, 이로 인해 지연 시간이 증가하고 있다. 지연시간을 줄이고 자원관리를 위해 서비스 배치 방법이 연구되고 있으나, 대부분의 관련 연구들은 서비스에 따른 고려가 미비하다. 본 논문에서는 성능을 높이고 자원을 효율적으로 사용하기 위해 사용자 요구에 따른 본 논문에서는 클러스터 환경에서 계산 자원을 효율적으로 할당하는 역할을 수행한다. 제안하는 방법을 통해 고성능 컴퓨팅 서비스에 대한 다양한 수요에 대응하여 사용자 만족도를 향상 시킬것으로 기대된다.

## Ⅰ. 서 론

최근 계산과학 분야에서 HPC(High Performance Computing)를 필요로 하는 소프트웨어의 사용이 많아지고 있으며, 전통적인 계산과학 뿐만 아니라 최근 빅데이터나 인공지능 등 새로운 분야가 대두되고 사용자가 다변화되면서 고성능 컴퓨팅 인프라 사용 환경에 대한 다양한 요구사항이 제기되고 있다. 또한, 기관마다 유사한 시스템을 각각 구축해서 남은 자원을 활용하려고 한다면 중복 예산 낭비가 예상되며, 그에 따른 유휴 자원까지 발생하게 된다. 클라우드 환경에서 가상화 기술 등을 통해 고성능 컴퓨팅(HPC)의 사용환경을 제공함에 있어, 유휴 자원을 줄이면서 한정된 계산 자원을 사용자들이 효율적으로 사용할 수 있도록 클러스터를 구성할 필요가 있다.

특히, HPC 작업 실행에 필요한 계산 자원을 컨테이너 형태로 제공하는 쿠버네티스 클러스터를 구현 시, 자원을 고려하여 타입에 따라 실행할 클러스터를 생성 및 할당할 수 있는 기술이 요구된다.[1]

본 논문에서는 고성능 컴퓨팅을 위한 클라우드 환경에서, 실시간 자원 현황에 기반하여 리소스 타입에 따라 클러스터를 구성하고 자원을 할당하고자 한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서 관련 연구로 HPC 클라우드 동향, 활성화 방안에 대해 기술하고, 3장에서 HPC 클라우드 환경에서 타입에 따라 클러스터를 구성하여 자원 관리할수 있는 위한 구조를 제안한다. 4장 결론에서는 제안하는  향후 활용 방안과 연구 계획에 대해 제시하며 마친다.

## Ⅱ. 관련연구

HPC 아키텍처에서는 보통 수백 또는 수천의 서버가 네트워크 또는 클러스터를 형성한다. 각 서버는 노드(Node)라고 하며, 클러스터 내에서 이러한 노드들이 병렬로 작동하여 클러스터에서는 여러 소프트웨어 프로그램과 알고리즘이 동시에 실행되어 다양한 HPC 응용 프로그램을 지원한다. 계산과학 연구자의 수요에 대응할 수 있도록 HPC 기반의 클라우드 환경에서 병렬처리컴퓨팅, GPU 컴퓨팅, 스토리지 컴퓨팅, 소프트웨어 개발 등을 진행할 수 있는 서비스 기술 개발이 필요하다.

초고성능컴퓨팅 인프라 사용 환경에 대한 다양한 요구사항이 제기되고 있다. 이는 매우 데이터 집약적인 경향 있으므로 유연하고 효율적인 사용 환경에 대한 필요성이 커짐에 따라 클라우드 기술이 대안으로서 많은 관심을 받게 됐다.

기존 연구에서는 Task의 특성에 따라 자원 사용의 특성이 다르게 나타나지만,  Task의 특성이 반영되지 않아 모든 Task가 일괄 할당되고 있다.

〔표 1〕 기존 배치방법을 활용한 자원관리 방법

| | 〔1〕 | 〔2〕 | 〔3〕 |
|---|---|---|---|
| 목표 | 에너지 비용 최소화 | 리소스 효율 | 에너지 효율 |
| 제안하는 방법 | 데이터 학습 | 우선순위결정 | Fuzzy |
| 장점 | 사용자 만족 | 리소스 이용 | 다목적 고려 |
| 본 연구와 차이점 | 사용자 요구사항에 따른(서비스 타입) 리소스 관리 | | |

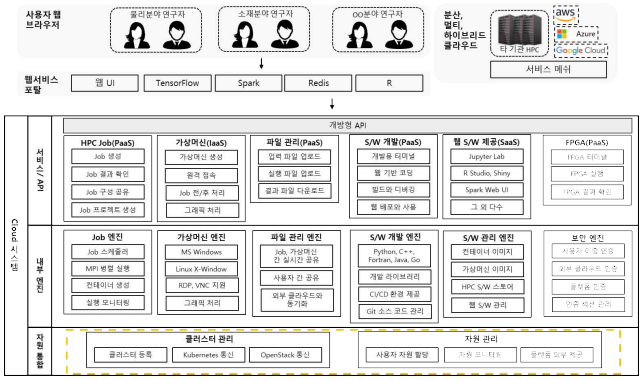
따라서 서비스를 운영할 때는 서비스 유형과 리소스 상태를 고려해야 한다. 그러나, 제안된 서비스 배치방식의 대부분은 표 1과 같이 동적 자원 상태 및 서비스 유형에 따른 고려는 미비하다.  본 논문에서는 변화하는 워크로드에 대응하여 서비스 타입을 고려하고자 한다.

## Ⅲ. 제안하는 구조

다음과 같은 기능을 포함하는 방식을 제안한다.

1.. 제안하는 구조

본 논문에서는 고성능 컴퓨팅을 위한 클라우드 환경에서, 사용자 요구사항에 적합한 자원을 제공하도록 자원관리를 위한 서비스 배치를 위한 구성도를 설계하였다.




[그림 1] 제안하는 구조

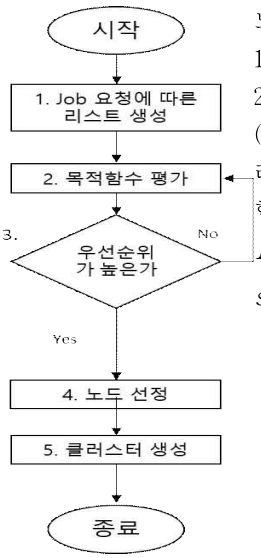
이때, 클러스터관리 , 자원관리로 구성되어 있다. 클러스터 관리를 통해 고성능 컴퓨팅을 위한 클라우드 환경에서 자원 상태를 고려하여 복수의 작업 각각을 실행할 수 있는 가상 인스턴스 클러스터를 동적으로 구성할 수 있다. 서비스 자원을 모니터링을 기반으로 job에 따라 다중 클러스터 구축 방법 및 시스템에 대해 설명 한다.

- 클러스터는 관리는 아래 그림처럼 Kubernetes와 OpenStack 플랫폼 소프트웨어를 지원하며 각 클러스터는 관리자 모드에서 1개 이상 등록 사용할 수 있다. 이후 job 서비스, 가상머신 서비스에서 클러스터 서비스를 통해서 각 API를 호출하고 응답받게 된다. 접속 인증과 통신 에러 처리 등을 담당하게 된다.

클러스터 관리부와 클라우드 자원 관리부로 구성된다. 본 시스템에서는 클러스터와 자원관리를 위해 사용자 정의 Job 어플리케이션 등록 관리, 동적 Slurm 클러스터 생성, Job실행 상태 모니터링 기능도 포함 된다.

2. 동작 과정

Job이 들어오는 데로 구성하는 것이 아닌, 자원상태에 따라 우선적으로 노드를 선정하여 구성 할 수 있다.




1. Job에 따른 노드리스트를 구성한다.
2. 서비스에 따라 목적함수를 평가한다.
(설정된 목적함수는 서비스 타입에 따라 달라질수 있으며, 해당 함수는 성능을 고려한 함수이다)

$$N_s = U + RT - W$$

$$s.t\ Max(N_s)$$

| U | 자원 사용량 |
|---|---|
| RT | 응답시간 |
| W T | 대기시간 $\min(\frac{1}{r(r-1)}),\ r = service\ rate$ |

[그림 2] 클러스터 구성방법

3. 노드의 우선순위가 높은지 확인한다

if $p(N_s) > p(N_{s+1})$ then $p = p(N_s)$

p : 우선순위 , $J_s$ : 우선순위 점수

4 높은 우선순위를 가지는 노드를 선정한다.

5. 선정된 노드에 클러스터를 생성

한정된 계산 자원을 여러 사용자들이 효율적으로 사용하기 위해서는 실시간으로 자원상태를 파악하여 작업 수행에 맞는 시스템을 할당해야 한다. 본 발명에서는 클러스터 환경에서 계산 자원을 효율적으로 할당하는 역할을 수행하는데 목적이 있다

## Ⅲ. 결론

본 논문에서는 HPC 환경에 자원관리를 위한 서비스 기반의 자원 관리 구조를 제안하였다. 사용자 요구에 맞게 최적화된 알고리즘을 서비스 형태로 제공 받음으로써 높은 접근성을 가질 것이다.

HPC 환경에서 자원 배치 방식을 활용하여 성능 최적화 하고 끊김 없는 서비스 제공을 받을 수 있도록 가상머신 할당이 필요한 시점에 수행하려는 워크로드의 특성을 파악하여 워크로드를 가장 효율적 으로 수행 시킬 수 있도록 서비스 성능을 보장하고자 하는 것이 목적이다. 모니터링을 통해 상태를 파악하고 최적의 가상머신을 선택하여 사용자의 요구에 따라 사용자의 만족도를 높이고 QoS도 보장하고자 한다. 향후 데이터 학습에 따라 다양한 분야로의 확장이 가능하며, 실제 환경에서 적용하여 활용하고자 한다.


## ACKNOWLEDGMENT

본 논문은 한국과학기술정보연구원에서의 기본사업으로 (No.K24-L2-M1-C7 ,초고성능컴퓨팅 공동 활용 기술개발) 으로 수행된 연구임. 교신저자 : 정기문*


## 참 고 문 헌


[1] Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristicsfor efficient management of data centers for cloud computing. Future generation computersystems, 28(5), 755-768.

[2]. Jennings, B., & Stadler, R. (2015). Resource management in clouds: Survey and research challenges. Journal of Network and Systems Management, 23, 567-619.

[3] Son, A. Y., & Huh, E. N. (2019). Multi-objective service placement scheme based on fuzzy-AHP system for distributed cloud computing. Applied Sciences, 9(17), 3550.

[4]초고성능컴퓨팅인프라 클라우드 서비스 구축을 위한 제언(2022), KISTI 이슈브리프 제46호