HPC 클라우드 환경에서 QoS를 위한 서비스 배치 기반의 자원 최적화 연구

손아영, 조혜영, 박준영, 정기문*
한국과학기술정보연구원
{ayson, chohy, jypark, kmjeong}@kisti.re.kr

A study on resource optimization based on Service Placement for QoS in HPC Cloud Environments

A-Young Son, Hyeyoung Cho, Junyoung Park, Gi-Mun Jeong*
Korea Institute of Science and Technology Information

요 약

최근 과학 애플리케이션은 점점 더 데이터 집약적으로 발전하고 있으며, 이로 인해 시스템 지연(latency)이 심화 되고 있다. 이러한 문제를 해결하기 위해 다양한 자원 관리 및 서비스 배치 기법이 제안 되었지만, 대부분의 기존 연구는 서로 다른 서비스 유형에 따른 효율성 요구를 충분히 반영하지 못하고 있다. 본 논문에서는 사용자 요구 기반의 다중 지표(multi-metric) 의사결정 방식을 통합한 퍼지(Fuzzy) 시스템 기반의 고성능 컴퓨팅(HPC) 클라우드 서비스 배치 기법을 제안하고자 한다. 본 기법은 실시간 자원 모니터링 및 프로파일링을 통해 시스템 자원을 동적으로 할당함으로써, 성능을 최적화하고 비용을 절감 효과가 있다. 향후 에는 지속적인 피드백 학습 메커니즘을 도입하여 배치 정확도를 향상시키고 시스템 효율을 개선하고자 한다.

Ⅰ. 서 론

최근 기후 과학, 천체물리학, 계산 생물학, 고에너지 물리학 등 다양한 과학 분야에서 활용되는 애플리케이션들은 점차 대용량 데이터를 필요로 하게 되었다. 이에 따라 컴퓨팅 자원과 저장소 간의 데이터 이동이 성능 병목의 주요 원인으로 작용하고 있다[1]. 이러한 환경 변화는 고성능 컴퓨팅(High Performance Computing, HPC) 인프라 사용에 대한 다양한 요구사항을 발생시키며, 최적화되지 못한 자원 관리는 성능 저하, 에너지 낭비, 서비스 품질 저하 등 여러 문제를 발생 시킬 수 있다.

AI 및 HPC 워크로드는 연산 유형에 따라 적합한 가속장치를 선택해야 하지만, 기존 클라우드 플랫폼에서는 이를 고려한 최적의 자원 매핑이 어렵다. 고성능 클라우드 환경에서 서비스를 제공하기 위해서는 효과적인 자원 관리가 필수적이며, 그중 서비스 배치는 자원 관리의 핵심 요소 중 하나이다.

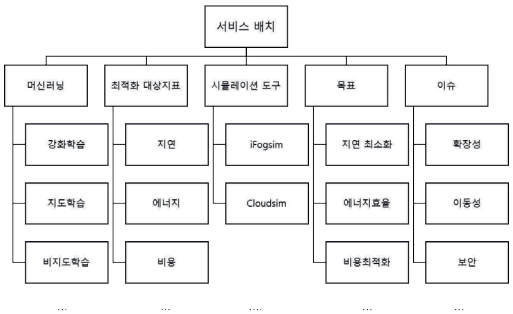
자원 할당이 필요한 시점에 서비스의 특성과 요구사항을 분석하고, 이를 바탕으로 가장 적합한 자원 위치에 서비스를 배치가 필요해 지고 있다. 최적의 자원을 선정하기 위해서는, 성능 보장, 에너지 효율성, 자원의 활용도 등 다양한 측면을 고려하는 것이 중요하다.

본 연구에서는 서비스 요청의 특성을 분석하기 위한 프로파일링 구조를 도입하고, 이를 통해 서비스 타입을 분류하여 클러스터 환경에서 자원을 효율적으로 배치하는 방안을 제안한다. 특히 분산 환경에서는 고성능 자원을 전략적으로 배치해야만 응답성 있는 서비스를 제공할 수 있으며, 사용자가 급격히 증가하거나 애플리케이션 요구가 다양해지는 경우 기존의 정적 자원 관리 방식으로는 성능 저하 문제를 해결하기 어렵다.

이를 해결하기 위해, 본 논문은 HPC 환경에 적합한 퍼지 기반의 자원 관리 및 서비스 배치 기법을 제안한다. 퍼지 논리와 입자 군집 최적화(Particle Swarm Optimization, PSO)를 기반으로 하여, 복잡하고 불확실한 자원 요구를 보다 분석하고 최적의 자원 배치를 수행하는 전략을 구성하는 것이 목표이다. 본 논문은 다음과 같이 구성되어 있다. 2장에서 관련 연구로 HPC환경에서의 서비스 배치 동향에 대해 기술하고, 3장에서 HPC 클라우드 환경에서 모니터링 기반의 프로파일러 시스템을 기반으로 자원 타입을 결정하는 방법을 제안하고, 4장 결론에서는 제안하는 향후 활용 방안과 연구 계획에 대해 제시하며 마친다.

Ⅱ. 관련연구

본 논문에서는 HPC 클라우드 환경에서는 지연 시간과 자원 할당 문제 등 여러 이슈가 존재한다. 특히 데이터 중심의 과학 애플리케이션은 자원 간의 비효율적 분배로 인해 성능 저하가 빈번하게 발생합니다.



[그림 1] 서비스 배치 기술 동향

그림1과 같이 특히 동적환경에서 강화학습 사용이 증가하고 있으며, 다중 목표 및 최적화를 위한 방향으로 연구가 진행되고 있다. 인터페이스 기술은 IoT 환경에서는 MQTT, 일반적 환경에서는 REST 가 많이 사용되고

있으며 시뮬레이션 도구는 ifogSIm 과 CloudSim 이 많이 사용되고 있다. 지연과 에너지 최적화를 목적으로 서비스 배치가 연구중에 있다. HPC 환경의 가변적인 워크로드를 고려한 MCDM 및 머신러닝 기반 최적화 기법이 제안되고 있으나, 실시간 자원 가용성과 서비스 특성을 모두 고려한 방식은 부족한 상황입니다.

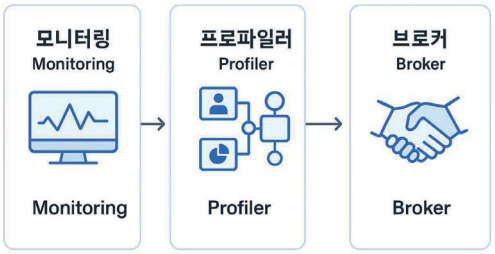
본 논문에서 사용자에 따른 서비스 배치 방법으로 사용자에게 QoS를 보장하고 자원관리를 위한 서비스 배치를 하고자 한다.

Ⅲ. 제안하는 방법

본 절에서는 자원 예측값을 기반으로 적절한 노드를 선택하는 퍼지 기반 서비스 배치 시스템을 제안한다. 주요 입력값으로 CPU 사용량, RAM 사용량 등을 활용하여, 적절한 배치 노드를 판단합하게 된다. 제안하는 구조는 모니터링, 프로파일러, 브로커로 구성되어 있다.

1.. 시스템 아키텍처 개요

제안하는 구조는 다음과 같은 모듈로 된다.



[그림 2] 제안 프로세스

1). 모니터링(Monitoring)

- 물리적 및 가상 머신에서 CPU, 메모리 등 시스템 자원 상태를 실시간으로 수집하여 전체 인프라의 사용 현황을 파악

2) 프로파일러(Profiler)

- 사용자 요청과 서비스 목적을 분류하며, 자원 요구사항을 분석

퍼지 기반 PSO(입자 군집 최적화, Particle Swarm Optimization) 모듈을 활용하여 최적의 가상 머신(VM)을 선택하고, 이에 따라 서비스 배치 위치를 결정

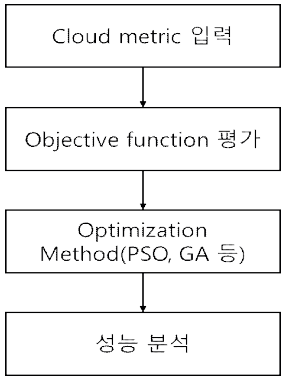
3) 브로커(Broker)

- 사용자 요구 및 서비스 유형에 따라 필요한 자원 유형을 판별하고, 적합한 자원 후보를 선정.

- 이후 우선순위 및 가용성 기준에 따라 최적의 자원을 배정.

2. . 서비스 배치 절차

배치 절차는 다음과 같은 계층적 단계를 거칩니다:



[그림 3] 서비스 재치 절차

1) 분류: 서비스 요구 조건에 부합하는 자원을 그룹화

2) 우선순위 결정: CPU, 메모리, 대역폭 등 기준으로 자원 후보 평가

3) 후보 선택: 비활성 자원이 많은 클러스터를 우선적으로 선택

4) 의사결정: 각 입력을 퍼지 시스템에 적용하여 최적 노드 도출

3.. 퍼지 의사결정 시스템

퍼지 시스템은 다음과 같은 단계로 작동한다.

1)입력 분석: 자원 사용률을 퍼지 변수로 변환

2) 퍼지 규칙 적용: 적합성 평가를 위한 퍼지 규칙 실행

3) 디퍼지화: 최종 배치 결정을 위한 명확한 수치 도출

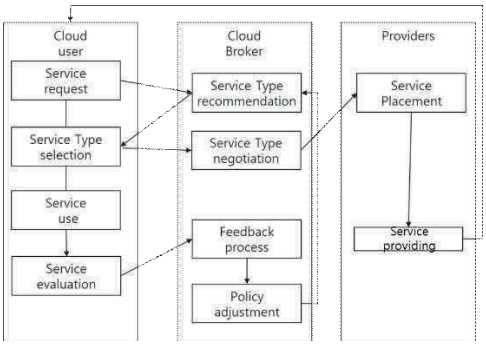
4) 예측 및 학습: 피드백 데이터를 기반으로 예측 정확도 개선

4. 피드백 및 정책 관리

1)자원 사용 중 이상 상태(Hotspot/Coldspot) 감지

2)정책 동적 조정 및 노드 재배치 수행

3) 새로운 서비스 유형에 따라 정책 및 자원 분배 갱신



[그림 4] 브로커를 통한 제공 과정

제안하는 구조를 통해 HPC 클라우드 환경에서 사용자 만족도를 높이고 자워을 관리를 통해 어플리케이션 성능 저하 문제를 해결하는데 목적이 있다.

Ⅲ. 결론

본 연구에서는 HPC 환경에서 데이터 집약형 애플리케이션의 성능 저하 문제를 해결하기 위해 퍼지 기반의 서비스 배치 시스템을 제안하엿다. 제안 기법은 실시간 자원 프로파일링과 피드백 기반 학습을 통해 서비스 연속성과 시스템 효율을 동시에 확보할 수 있다.

고성능 클라우드 환경에서는 다양한 장치에도 적용이 필요해지고 있어, 향후 에는 실 데이터셋 기반으로 성능 검증과 다양한 환경에서 확장성을 검증하고자 한다.

ACKNOWLEDGMENT

본 논문은 한국과학기술정보연구원에서의 기본사업으로 (No.K25L1M2C2,수요자 맞춤형 연구환경 제공을 위한 초고성능컴퓨팅 기술개발) 으로 수행된 연구임. 교신저자 : 정기문

참 고 문 헌

[1] K. Elissa, “Title of paper if known,” unpublisYU, Liang; JIANG, Tao; ZOU, Yulong. Fog-assisted operational cost reduction for cloud data centers. IEEE Access, 2017, 5: 13578-13586.

[2] STOJKOSKA, Biljana Risteska; TRIVODALIEV, Kire. Enabling internet of things for smart homes through fog computing. In: 2017 25th Telecommunication Forum (TELFOR). IEEE, 2017. p. 1-4.

[3]LIU, Boyun, et al. Workload forecasting based elastic resource management in edge cloud. Computers & Industrial Engineering, 2020, 139: 106136.