고성능 클라우드 컴퓨팅 환경에서 사용자 워크로드 타입에 따른 서비스 배치 방법 연구

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

A Study on Service Placement according to workload type in High Performance Cloud Computing

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

요 약

기후 과학, 천체 물리학, 생물정보학 등 주요 과학 응용 분야는 점차 데이터 집약적 특성을 보이고 있으며, 이에 따라 고성능 클라우드 컴퓨팅 환경은 자원의 유연성과 확장성을 제공하는 핵심 인프라로 주목받고 있다. 그러나 동적인 클라우드 기반 HPC 환경에서 서비스 연속성과 성능 보장은 여전히 중요한 과제로 남아 있다. 본 연구에서는 클라우드 환경의 자원 관리 방법 중 하나인 서비스 배치기법을 활용하여, 자원 할당이 필요한 시점에 워크로드의 특성을 분석하고 이를 기반으로 효율적인 서비스 배치를 수행하는 방안을 제안한다. 제안된 방법은 워크로드 특성에 적합한 자원 배치를 통해 기존 방식 대비 다양한 HPC 서비스 수요에 대응함으로써 사용자 만족도를 높일 수 있을 것으로 기대된다.

1. 서론

최근 계산 과학 분야에서는 시뮬레이션 데이터의 증가, 실험 환경의 고도화 등으로 인해, 데이터 생성 속도가 증가하고 있다[1]. 이러한 환경에서 고성능 클라우드 컴퓨팅은 계산 과학자들에게 보다 응답성이 뛰어난 서비스를 제공하기 위한 대안으로 주목받고 있다. 이를 위해서는 지속적인 자원관리가 더 중요해 지고 있다.

기존의 자원 관리 방식은 실시간 수요와 다양한 서비스 유형을 반영이 어렵다.그 중에서도 자원 이용률이 임계값을 초과하거나 이하인 경우, 서비스 품질 저하를 초래할 수 있다. 따라서 서비스 배치를 할때는 서비스 유형을 고려하는 것이 필요하다.

이에 본 논문은 퍼지 시스템(Fuzzy System)과 다중 기준 의사결정(MCDM) 기법을 기반으로 시스템 및 방법을 제안한다. 제안하는 방법은 사용 패턴을 분석하고 이를 기반으로 자원을 동적으로 할당하며, 학습을 통해 배치 프로세스의 의사결정 정확성을 향상 시키고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 기존 연구의 분석을 통해 한계점을 제시하고, 3장에서는 제안하는 서비스 배치방법과 구조에 대해 서술한다. 4장에서는 제안한 연구의 결론을 도출하고 향후 연구에 대해 제시한다.

2. 관련연구

계산과학 분야에서도 데이터 증가에 따라, 계산과학자의 요구에 맞는 서비스 제공이 필요 하다. 이때, 클라우드 자원 관리술을 활용하여 다양한 연구에서도 에너지 비용 최소화, 자원 활용 극대화, 에너지 효율 향상 등을 목표로 기법들이 제안되었다[4,7,10]. 관련 연구들은 자원관리를 서비스 배치 방법을 활용하여, 다양한 기법을 적용해 왔으며, 대표적으로 의 최적화 기법의 대표적인 GA(Genetic Algorithm) , AHP 및 TOPSIS 등 의사결정기법, 강화 학습 기반의 정책 제안 등이 활용되었다.[8,9][11]. 그러나 기존의 서비스 배치 방법을은 자원의 상태가 변화는 상황을 충분히 반영하기 어렵다는 한계적이 존재하였다. 또한 서비스 유형에 따라 요구사항을 구체적으로 고려하지 못하는 한계가 있었다.

따라서 본 논문은 이러한 한계를 극복하기 위해 의

사결정방법인 퍼지 시스템을 적용하고, 다목적 및 다중 메트릭 고려하여 사용자 요구를 만족시키는 방안을 제안하고자 한다.

3. 제안하는 구조

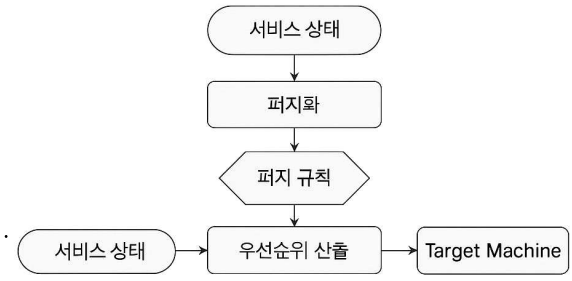
본 연구는 HPC 클라우드 환경에서 사용자 요구사항에 따라 자원을 제공할 수 있도록 서비스 배치 방법을 제공하고자 한다. 따라서, 자원상태를 고려하여 리소스 타입별로 구성하고, 이를 통해 효율적으로 자원을 할당하고자 한다.

3-1. 계층 분류

(그림1)과 같이 계층형 기반으로 기존연구[3~8]를 기반으로 워크로드 유형을 분류하고 세부 기준을 분류하였다.각 항목별로 가중치를 부여하여 브로커가 서비스 특성에 따라 자원 배치를 최적화하도록 설계하였다. 예를 들어, 데이터 집약적 워크로드의 경우 데이터 저장·이동(C1, C2) 관련 항목의 가중치를 높이고, 실시간 이벤트 탐지 워크로드에서는 지연(Locality/실시간성, C3, C6) 관련 항목의 가중치를 높여 배치 의사결정에 반영한다.

3-2. 퍼지 시스템 기반 서비스 배치

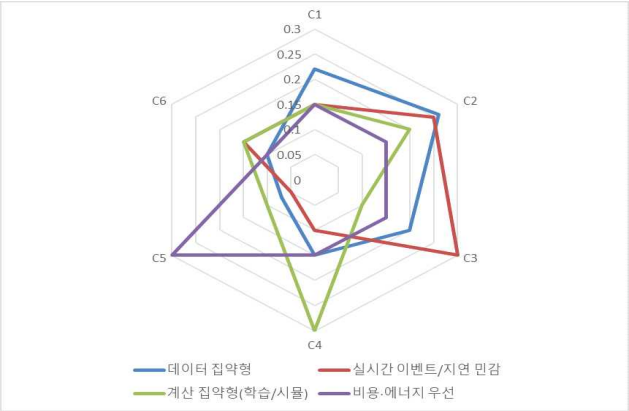
제안하는 방법은 퍼지시스템을 기반으로 (그림2)와 같이 구성된다. (그림2)는 자원상태를 예측하여 가중치를 계산하고, 타겟 머신을 선정하는 방식이다.



(그림 1) 워크로드에 따른 선정 방법

- 의사결정: 모니터링 metric을 기반으로 자원 사용량에 대해 퍼지화 하고, 사전된 정의된 규칙에 따라 우선순위를 산정 한다.
- 후보리스트 구성 : 사용자 요구사항에 따라 우선순위리스트를 구성한다. 이때 서비스 유형에 따라 구성이 달라지고 항목에 따라 선정되는 타겟 머신이 달라진다.
- 타겟 머신 선정 : 선정된 우선순위를 기반으로 최종적으러 타겟 머신을 선정하고 서비스 배치를 실시한다.
- 서비스 타입별 후보 리스트 구성: 요구 자원의 적합도가 높은 자원 목록을 서비스 유형에 따라 구성하고, 각 항목의 우선순위를 산출한다.

(그림 3), <표 1> 은 워크로드 별 가중치가 높게 나온 것을 정리한 것이고 이를 기반으로 배치를 실시하게 된다.



(그림 2) 워크로드 유형에 따른 선정 metric

<표 1> 워크로드 유형에 따른 선정 metric

워크로드 유형	metric
데이터 집약형	C1(저장·접근), C2(I/O), C3(지역성)
실시간 이벤트/지연 민감	C2(I/O), C3(지역성), C6(서비스 특성)
계산 집약형(학습/시뮬)	C4(확장성·자원)
비용·에너지 우선	C5(비용·에너지)

워크로드 유형별로 설정된 가중치를 반영함으로써, 브로커는 후보 자원 리스트 중에서 가중치가 가장 높은 자원을 선택하여 효율적인 서비스 배치를 수행할 수 있다.

4. 결론

본 논문은 고성능 클라우드 컴퓨팅 환경에서 의사결정 구조 및 방법을 제안하였다. HPC 클라우드 환경에서 자원 이용률을 높이고 서비스 안정성을 제공하고자 한다. 제안하는 방법을 통해 사용자 요구사항에 맞는 자원을 동적으로 제공할 수 있을 것으로 기대된다.

향후 연구에서는 제안된 방법의 정확도는 높이며, 실제 환경 및 대규모 환경에서 적용하여 성능을 검증하고자 한다.

ACKNOWLEDGMENT

본 논문은 한국과학기술정보연구원에서의 기본사업으로 (No.K25L1M2C2-01) 으로 수행된 연구임.
교신저자 : 정기문

참고문헌

[1] Giannakou, Anna, et al. "Understanding Data Movement Patterns in HPC: A NERSC Case Study." SC24: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2024.

[2]Alashaikh, Abdulaziz, Eisa Alanazi, and Ala Al-Fuqaha. "A survey on the use of preferences for virtual machine placement in cloud data centers." ACM Computing Surveys (CSUR) 54.5 (2021): 1-39.

[3] J. Gray, The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[4] R. Bordawekar, et al., “Data Management Challenges in Climate Science,” ACM SIGMOD Record, vol. 44, no. 3, pp. 17 – 22, 2015.

[5] L. Arm, et al., “LSST Data Management and Real-Time Event Detection,” Journal of Astronomical Telescopes, Instruments, and Systems, vol. 6, no. 1, pp. 1 – 14, 2020.

[6] E. Afgan, et al., “Galaxy Platform for Reproducible Biomedical Data Analysis,” Nucleic Acids Research, vol. 46, no. W1, pp. W537 – W544, 2018.

[7] T. Beermann, et al., “Rucio: Scientific Data Management at Exascale,” Computing and Software for Big Science, vol. 3, no. 1, 2019.

[8] G. Zhang, et al., “Scientific Computing Meets Big Data Technology,” ArXiv preprint, arXiv:1507.03325, 2015.

[9] A. K ㅣarniadakis, et al., “Physics-informed Machine Learning: Theory and Applications,” Nature Reviews Physics, vol. 3, pp. 422 – 440, 2021.

[10] K. Yelick, “Exascale Computing and Big Data,” Communications of the ACM, vol. 61, no. 6, pp. 54 – 63, 2018.

[11] M. Shankar, et al., “FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications,” ArXiv preprint, arXiv:2106.06433, 2021.