# 서버리스 환경에서의 워크로드 버스트 특성 분석 및 예측 지표 연구

추교현, 남상혁, 조민규, 안수겸, 박상오
중앙대학교

khchoo@cslab.cau.ac.kr, shnam@cslab.cau.ac.kr,
mgjo@cslab.cau.ac.kr, skahn@cslab.cau.ac.kr, sopark@cau.ac.kr

# A Study on Workload Burst Characteristics and Predictability in Serverless Environments

Kyohyun Choo, Sanghyuck Nam, Mingyu Jo, Sukyum Ahn, Sangoh Park

Chung-Ang Univ.

## 요 약

본 연구는 서버리스 컴퓨팅 환경에서 발생하는 예측하기 어려운 워크로드 급증(Burst) 현상의 특징을 정량적으로 분석하고, 이를 선제적으로 대응하기 위한 통계적 예측 지표를 제안한다. 대규모 서버리스 트레이스를 분석한 결과, 전체 워크로드 중 버스트 상태는 17.3%에 불과하여 매우 희소한 특성을 보이며, 개별 함수 수준에서도 평균 버스트 밀도가 10.2%로 낮게 형성됨을 확인하였다. 또한, 버스트 발생 전 10 분간의 피처 상관관계를 분석한 결과, 호출 횟수와 기울기 지표가 발생 직전 0.55 이상의 높은 상관관계를 보인다는 것을 입증하였다. 이러한 발견은 저비용 및 고효율의 전문가 혼합 모델(Mixture-of-Experts) 기반 선제적 오토스케일링 시스템 설계의 핵심적인 근거가 된다.

## Ⅰ. 서 론

서버리스 컴퓨팅(FaaS, Function-as-a-Service) [1]은 개발자가 인프라 관리 부담 없이 코드만 배포하면 되는 환경을 제공하며, 트래픽에 따른 유연한 확장성과 자원 효율성이라는 강력한 장점을 가진다. 그러나 이러한 장점에도 불구하고, 예측 불가능한 워크로드 폭증(Burst) 상황에서 발생하는 콜드 스타트 문제는 시스템의 신뢰성을 심각하게 위협한다.

콜드 스타트로 인한 수 초 이상의 지연 시간은 곧바로 서비스 수준 협약(SLA) 위반으로 이어지며, 기존 서버리스 시스템에서 사용되는 반응형 스케일러는 트래픽이 이미 유입된 후에 함수 인스턴스를 생성하므로 이러한 지연을 근본적으로 막을 수 없다.

따라서, 성능과 자원 효율성 사이의 최적의 트레이드오프를 달성하기 위해서는 미래 트래픽을 예측하고 준비하는 선제적 스케일링이 필수적이다. 본 연구는 실제 서버리스 환경의 데이터 분포가 지닌 극심한 불균형성에 주목하며, 모든 트래픽 상황에 무거운 단일 예측 모델을 사용하는 대신 평상시(Steady)와 버스트(Burst) 상황에 각각 특화된 전문가를 배치하는 지능형 오토스케일러의 필요성에 대해서 증명하고자 한다.

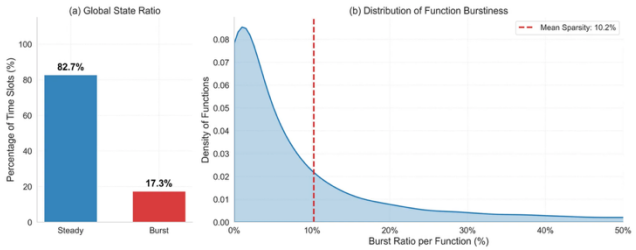


Figure 1. Serverless Workload Characteristics and Sparsity Distribution.

## Ⅱ. 워크로드 버스트 특성 분석

본 섹션에서는 대규모 서버리스 클러스터 트레이스 [2]를 분석하여 워크로드의 상태 분포와 함수별 버스트 발생 양상을 파악한다.

(a) Global State Ratio

Figure 1-(a)을 통해 전체 워크로드의 상태를 분석한 결과, 서버리스 환경의 트래픽은 정상 상태(Steady)가 82.7%를 차지한다. 반면, 급격한 트래픽이 증가하는 버스트(Burst) 상태는 17.3%에 불과하다. 이는 버스트가 매우 희소하게 발생하므로, 모든 상황에 고성능 예측 모델을 추론시키는 것은 자원 낭비가 될 수 있음을 알 수 있다.

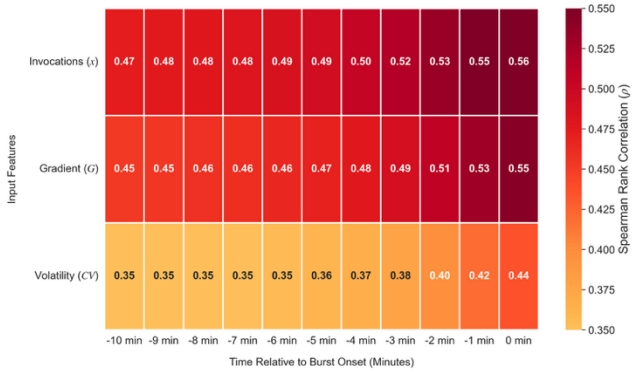


Figure 2. Serverless Workload Characteristics and Sparsity Distribution.

(b) Distribution of Function Burstiness

Figure 1-(b)은 개별 함수들이 가지는 버스트 비율의 밀도 분포를 보여준다. 워크로드 분석 결과, 전체 함수의 평균 버스트 비율은 10.2%로 나타난다.

그래프의 형태를 보면 대다수의 함수가 0~10% 사이의 낮은 버스트 비율을 가지는 Long-tail 분포를 형성하고 있다. 이는 서버리스 워크로드 예측 시 대다수의 평상시 상황은 경량화된 모델로 처리하고, 드물게 발생하는 버스트 상황에는 이에 특화된 모델을 사용하는 혼합 전문가 모델 (Mixture-of-Experts) 접근법이 논리적으로 타당함을 입증한다.

## Ⅲ. 버스트 예측을 위한 피처 상관관계 분석

버스트를 선제적으로 감지하고 함수 인스턴스를 미리 생성하기 위해서는 예측 모델 [3]이 입력 벡터로 유의미하게 활용할 수 있는 통계적 피처를 선별하는 것이 중요하다. 본 연구에서는 호출 횟수, 트래픽의 변화율, 그리고 데이터의 변동성 세 가지 피처를 선정하여 버스트 발생 시점과의 스피어먼 순위 상관계수 변화를 분석하였다.

호출 횟수와 기울기 지표는 버스트 발생 10 분 전부터 각각 0.47, 0.45 의 상관관계를 보였다. 이 수치는 버스트 발생 시점이 다가올수록 점진적으로 상승하여 발생 시점에는 각각 0.56 과 0.55 에 도달한다. 이를 통해 트래픽의 절대적인 유입량뿐만 아니라, 유입 속도의 가속도가 버스트 진입을 결정짓는 핵심 피처임을 알 수 있다.

변동성 지표는 호출 횟수나 기울기 지표에 비해 상대적으로 낮은 상관관계(0.35~0.44)를 유지하였다. 그러나 버스트 발생 2 분 전부터는 0.40 을 돌파하여 급격한 상관관계 상승을 보였다. 이는 변동성 지표가 장기적인 추세 예측보다는 버스트 직전의 급격한 불안정성을 포착하는 지표로서 가치가 있음을 의미한다.

본 분석을 통해 버스트 발생 약 10 분 전부터 유의미한 상관관계를 확인하였다. 이는 서버리스 시스템이 새로운 함수 인스턴스를 미리 생성(Pre-warming)하기에 충분한 시간을 확보할 수 있음을 입증하는 연구 결과이다.

## Ⅳ. 결론

본 연구는 서버리스 컴퓨팅 환경에서의 워크로드 특징을 정량적으로 분석하고, 이를 선제적으로 감지하기 위한 핵심 통계 지표의 유효성을 입증하였다.

전체 워크로드 중 버스트 상태는 17.3%에 불과하며, 개별 함수의 평균 버스트 밀도는 10.2%로 매우 낮음을 확인하였다. 이는 평상시 상황에는 경량 모델을, 버스트 상황에는 특화된 고성능 모델을 사용하는 혼합 전문가 모델 [4] (Mixture-of-Experts) 구조의 타당성을 뒷받침한다.

스피어먼 순위 상관계수 분석 결과, 호출 횟수와 트래픽 기울기는 버스트 발생 10 분 전부터 0.45 이상의 유의미한 상관관계를 형성하였다. 특히 발생 시점에 근접할수록 상관성이 0.55 이상으로 강화되어, 결정적인 선행 지표로 활용될 수 있음을 증명하였다.

결론적으로 본 연구에서 도출된 지표와 특성 분석 결과는 서버리스 시스템의 콜드 스타트 문제를 획기적으로 개선하고자 자원 효율성을 극대화하는 지능형 오토스케일러의 구현 가능성을 입증한다. 향후 연구에서는 실제 서버리스 시스템 환경에서 MoE 모델을 시스템에 통합하여 SLA 준수율 및 비용 절감 효과를 정량적으로 검증할 계획이다.


## ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로

한국연구재단의 지원(RS-2024-00345869)과

정부(과학기술정보통신부)의 재원으로 한국연구재단 지원(RS-2025-02217071)을 받아 수행된 연구임


## 참 고 문 헌


[1] Z. Li et al., "The Serverless Computing Survey: A Technical Primer for Design Architecture," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1- 34, Sep. 2022, doi: 10.1145/3508360.

[2] Shahrad, M., et al. "Serverless in the Wild: Characterizing and Optimizing the Serverless Workload at a Large Cloud Provider." Proceedings of the 2020 USENIX Annual Technical Conference (ATC '20), 2020.

[3] Wang et al., "Libra: Harvesting Idle Resources Safely and Timely in Serverless Clusters," in *Proceedings of the Sixteenth European Conference on Computer Systems (EuroSys '21)*, 2021, pp. 119- 134.

[4] Shazeer, N., et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." International Conference on Learning Representations (ICLR), 2017.