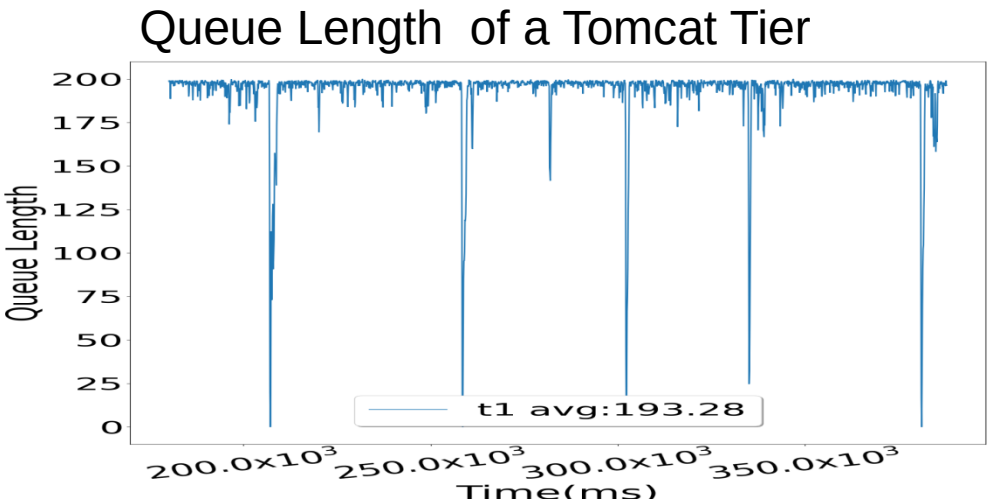
A Bi-metric Autoscaling Approach for n-Tier Web Applications on Kubernetes

Changpeng ZHU, Bo HAN, Yinliang ZHAO

Frontiers of Computer Science, DOI: 10.1007/s11704-021-0118-1

Problems & Ideas

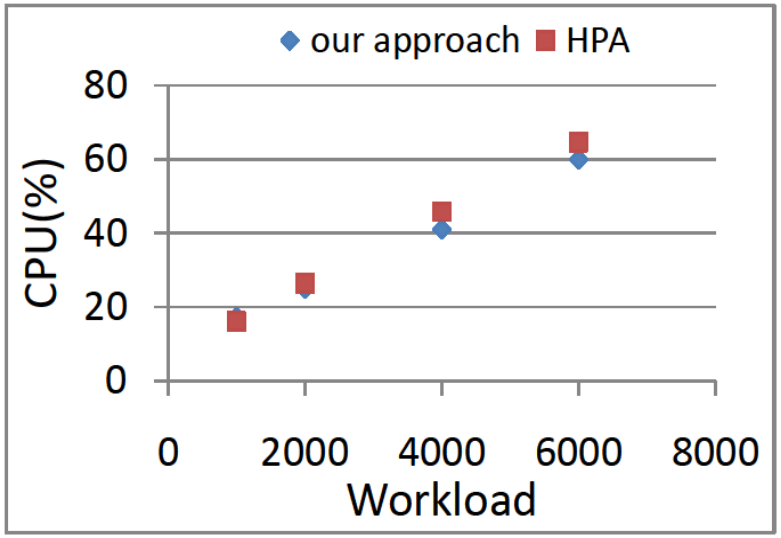
- Problems of HPA provided by Kubernetes
  - Scale more pods than expected by a single performance metric, for example, CPU utilization
  - Cost more resources
- Ideas: Scale pods by a combination of two metrics
  - A hardware performance metric, for example, CPU utilization
  - A software performance metric, for example, utilization of a thread pool in Tomcat.
    - utilization = queue length / thread pool size(a constant)
    - ELBA is used to evaluate the queue length of a tier



Main Contributions

1. Present drawbacks of HPA by an experimental study of n-tier web applications running on Kubernetes.
2. Propose a bi-metric approach to solve these drawbacks.
   - Our approach scales less pods than HPA and contributes to less resource costs.

Less CPU cost



Less Pods