Percentile-Based Approach to Forecasting Workload Growth

Alexander Gilgur, C.Stephen Gunn, Douglas Browning,
Xiaojun Di, Wei Chen, Rajesh Krishnaswamy

Google, Inc / Alphabet, Inc
alexgilgur@gmail.com

Abstract

When forecasting resource workloads (traffic, CPU load, memory usage, etc.), we often extrapolate from the upper percentiles of data distributions. This works very well when the resource is far enough from its saturation point. However, when the resource utilization gets closer to the workload---carrying capacity of the resource, upper percentiles level off (the phenomenon is colloquially known as flat--- topping or clipping), leading to under---predictions of future workload and potentially to undersized resources. This paper explains the phenomenon and proposes a new approach that can be used for making useful forecasts of workload when historical data for the forecast are collected from a resource approaching saturation.