A Study of Transient Bottlenecks in Cloud

QINGYANG WANG, COLLEGE OF COMPUTING, GEORGIA INSTITUTE OF TECHNOLOGY

An essential requirement for cloud computing environment is to achieve both good performance and high utilization. High utilization is essential for cloud vendors for high return on investment. Good performance (fast response time) is critical for business of e-commerce web applications like Amazon. Unfortunately, both practitioners and researchers have experienced large response time fluctuations of web applications deployed in cloud, especially when utilization becomes high. In this talk, I will show that transient bottlenecks are an important contributing factor to the large response time fluctuations.

Transient bottlenecks are bottlenecks with a short lifespan on the order of tens of milliseconds. Though short-lived, transient bottleneck can cause a long-tail response time distribution that spans a spectrum of 2 to 3 orders of magnitude, from tens of milliseconds to tens of seconds, due to the queuing effect propagation and amplification caused by complex inter-tier resource dependencies in the system. I will present three causes of transient bottlenecks that we found in different system layers (e.g., Java garbage collection, CPU DVFS, virtual machine consolidation). In each of these causes, I will illustrate how transient bottlenecks are created and how it can lead to requests with response time up to several seconds.

Qingyang Wang is a PhD candidate in School of Computer Science at Georgia Institute of Technology. His research interests are broadly in distributed systems and cloud computing with a current focus on performance and scalability analysis of large-scale web applications like Amazon and Google. He has been closely collaborating with the industry collaborator Fujitsu labs in Japan. His work appears in venues such as TRIOS, ICDCS, Cloud, ICAC, SRDS, and IPDPS. Qingyang received his BS degree in Computer Science from Wuhan University, China, in 2004, and MS degree in Computer Science from Chinese Academy of Sciences, in 2007.