# NSF PI Meeting: The Science of Cloud Computing

### List of Principal Investigators:

1. Ningfang Mi
   Northeastern University, PI, ningfang@ece.neu.edu

2. Waleed M Meleis
   Northeastern University, Co-PI, meleis@ece.neu.edu

### Related Research Topics:

1. Cloud Architectures and Systems

2. Cloud Self-Monitoring and Autonomic Control

### Current Research Activities:

Burstiness occurs in workloads in which bursts of requests aggregate together during short periods of time and create periods of peak system utilization. Bursty workloads have been identified in diverse domains, including financial transaction processing, web servers, and internet routing. These temporal surges in incoming requests can dramatically degrade the performance of modern distributed computing systems such as cloud computing systems. Cloud systems are now growing in popularity, and are expected to become a dominant computational platform. However previous research on resource management for clouds does not take into account the effects of burstiness on performance and system design.

In our preliminary work, we observed that burstiness does exhibit in multi-tier enterprise systems, which in turn causes a phenomenon of *bottleneck switch* between servers across time and found that system performance with a bursty workload can be up to three orders of magnitude worse than the performance with a non-bursty workload. As a result, we proposed a new capacity planning methodology to capture workload burstiness in performance models and designed a new load balancing scheme for a cluster system, which separates jobs to servers according to their sizes, but further balances bursty profiles among all servers by reducing the load of the server(s) that admit arrival streams with high burstiness. This is the first time that burstiness in the arrival process becomes a critical aspect of load balancing.

Our preliminary work further investigated the impact of burstiness on load balancer performance in cloud systems via simulations developed on the CSim library. We observed that under most load balancers (e.g., the greedy ones which always select the best computing site for job submission based on shortest queue length, least utilization, etc.), there is a significant performance degradation caused by the imbalance of load among computing sites. This is because the load balancers cannot detect system load surges on computing sites during bursty arrivals. All load balancing decisions are made based on the outdate information. As a result, bursty arrivals are highly likely to be submitted to the same site, and consequently incur significant load on that particular site, resulting the performance degradation under bursty workloads. Therefore, we argue that *such deleterious effects due to burstiness and remote submission delay must be considered in the performance evaluation and load balancer design for cloud computing*. Unfortunately, the conventional load balancers used in the present cloud platforms (e.g., Microsoft's Azure and Amazon's EC) ignore burstiness in arrivals. Motivated by this problem, we thus propose to develop new methodologies for effective resource allocation in cloud computing systems under bursty conditions.

Currently, some of our preliminary work has been accepted by IEEE International Conference on Communications (ICC2011). PIs also received the Amazon Web Service (AWS) in Education Research grant, which allows PIs to develop fundamental understanding of the sources of burstiness in clouds, to understand how complex cloud computing systems respond to burstiness, and evaluate how improved resource allocation algorithms can reduce the effect of burstiness on system performance in clouds.

## ARA-C2E: Adaptive Resource Allocation for Cloud Computing Environment under Burst Workloads

Cloud computing systems allow users to pay to temporarily lease collections of virtual machines that are then used to execute applications. These virtual machines are assigned to physical resources to satisfy resource requirements, minimize latency, and maximize throughput. Popular cloud computing frameworks include Microsoft's Azure and Amazon's Elastic Computing platforms. In cloud systems, many applications are no longer single-program-single-execution applications. These applications involve a large number of concurrent and dependent jobs, which can be executed either in parallel or sequentially. Simultaneously launching jobs from different applications during a short time period can immediately cause a significant arrival peak, which further aggravates resource competitions and load unbalancing among computing sites. Also, as the number of these applications significantly increases in recent years, the present of Internet flash-crowds and burstiness surges becomes more frequent. As a result, how to counteract burstiness and maintain high system performance and availability becomes imminently important but challenging as well in clouds. However, conventional methods only consider exponentially distributed interarrival times between jobs, which unfortunately neglect cases of bursty arrivals and cannot capture the impacts of burstiness on system performance.

Therefore, in this project, we plan to develop new methodologies for effective resource allocation under bursty conditions. We propose to design novel techniques for resource allocation in cloud systems (e.g., Amazon EC2 and Microsoft's Azure), which attempt to satisfy peak user demands, improve overall system performance and availability, and meet SLAs. In particular, we will first broadly investigate the burstiness impacts on resource in distributed or cloud computing systems. By leveraging the knowledge of burstiness, we plan to develop new predictors which can accurately forecast the changes in user demands and system loads. By using such useful predicted information, we will design new algorithms that can detect burstiness and allocate resources (i.e., computing sites) using both static and dynamic approaches. We expect that the new algorithms can balance application workloads by optimizing the assignment of applications to virtual machines, and balance workloads within applications as well by optimizing the dispatch of requests across instances. We also expect that the new methodologies can allow applications to share information about their users' demands to improve overall system performance.

The deliverables of this project include:

- Fundamental understanding of the sources of burstiness in workloads and how complex cloud computing systems respond to burstiness.
- Broad investigation of burstiness impacts on resource allocation in cloud computing systems.
- A new predictor that can accurately detect the changes in workload and provide useful information for load balancers.
- A new resource allocation methodology for cloud systems (e.g., Microsoft's Azure Cloud Platform) that attempts to reduce the effect of burstiness on system performance and optimize resource utilization..

In order to successfully accomplish this project, we have applied and received Amazon Web Service (AWS) in Education Research grant to access the cloud services for research. We expect that our new resource allocation schemes will provide insight into how the cloud computing platforms such as Amazon EC2 can be optimized for bursty workloads. By developing and deploying burstiness-aware resource allocation strategies, cloud computing frameworks will be able to make better use of its infrastructure, cloud computing applications will consume fewer resources, and users of cloud computing applications will experience better response times. Taken together, these advances will support the adoption of this new cloud computing facility. The broader impact of the proposed research will be to develop a new class of algorithms and techniques for allocating cloud computing resources. Significant effort will be made to help foster technology transfer of the results of this project.