Resource Allocation and Job Scheduling in the Cloud
Issues and Challenges

Eleni Karatza
Professor Emeritus
Department of Informatics
Aristotle University of Thessaloniki, Greece
karatza@csd.auth.gr

Abstract:
Cloud computing has been continuously growing, offering computational services to many scientists, consumers and enterprises as utilities, on a pay-per-use approach. Cloud computing is a cost-effective infrastructure for running complex and computationally intensive applications and provides great possibilities in many computation areas.

The cloud computing paradigm can offer various types of services, such as computational resources for complex applications, web services, social networking, urban mobility, health care, environmental science, etc. Furthermore, the simultaneous usage of services from different clouds can have additional benefits such as lower cost and high availability.

Cloud computing is a very important topic in academia and industry. However, while there has been substantial research already, there still remain important issues that must be addressed, such as: performance, resource allocation, efficient scheduling, energy conservation, reliability, protection of sensitive data, security and trust, cost, availability, quality, interoperability.

Effective management of cloud resources is crucial to use effectively the power of these systems and achieve high system performance. Complex multiple-task applications may have precedence constraints and specific deadlines and may impose several restrictions and QoS requirements; therefore resource allocation and scheduling is a difficult task in clouds where there are many alternative heterogeneous computers. The scheduling algorithms must seek a way to maintain a good response time to leasing cost ratio. Furthermore, energy-efficient scheduling can decrease the energy consumption in the cloud and therefore not only to reduce the cost, but also to minimize the impact of cloud computing on the environment.

Evaluating the performance of an existing cloud system is often not feasible. Simulation is a valuable alternative mean to examining cloud performance, and to also assessing the impact of workload and system changes.

In this talk we will present state-of-the-art research covering a variety of concepts on resource allocation and job scheduling in the cloud, based on existing or simulated cloud systems, that provide insight into problems solving. We will also provide future directions in the cloud computing area.