Efficient resource provisioning and Scheduling in Clouds and Grids with Uncertainty

Andrei Tchernykh

CICESE Research Center, Ensenada, Baja California, México, chernykh@cicese.mx

http://usuario.cicese.mx/~chernykh/

We discuss the role of uncertainty for different scenarios of HPC, Grid and Cloud Infrastructures. We provide some theoretical and experimental bounds and QoS, dynamic and adaptive approaches.

As system scales and energy consumption increase, such new technologies have the power to do significant damage to our ecosystems. Traditional heuristic-based approaches to resource optimization become insufficient. Efficient eco-friendly power-aware computing resources optimization should be considered both in terms of reducing the environmental impact and reducing costs.

Ecologically sustainable computing is a rapidly expanding research area spanning the fields of computer science and engineering, resource optimization as well as other disciplines. The ecofriendly management service minimizes energy consumption, and adapts to dynamic load characteristics to meet desired QoS constraints.

Management of resources and jobs in such systems usually aims to optimize few objectives of processing efficiency and energy efficiency (green computing). In view of the permanent increasing of energy used for computing (data centers, Grid, etc.) the paradigm of green computing seems very important.

Clouds differ from previous computing environments in the way that they introduce a continuous uncertainty into the computational process. The uncertainty becomes the main hassle of cloud computing bringing additional challenges to both end-users and resource providers. It requires to waive habitual computing paradigms, adapt current computing models, and design novel resource management strategies to handle uncertainty in an effective way.

We discuss several major sources of uncertainty in clouds: dynamic elasticity, dynamic performance changing, virtualization, loosely coupling application to the infrastructure, among many others. The manner in which the service provisioning can be done depends not only on the service property and needed resources, but also users that share resources at the same time, in contrast to dedicated resources governed by a queuing system.

Various solutions have already been proposed and implemented. However, there are still many open issues in this field, including the consideration of power-aware parallel job scheduling in presence of uncertainty.