SORS: Scheduling with uncertainty in Grids and Clouds

Speaker: Andrei Chernykh, CICESE, Ensenada, BC, México

Title: Scheduling with uncertainty in Grids and Clouds

Abstract: Clouds differ from previous computing environments in the way that they introduce a continuous uncertainty into the computational process. The uncertainty brings additional challenges to both end-users and resource providers. It requires waiving habitual computing paradigms, adapting current computing models to this evolution, and designing novel resource management strategies to mitigate uncertainty and handle it in an effective way.
In this talk, we address scheduling algorithms for different scenarios of HPC, Grid and Cloud

Infrastructures. We provide some theoretical and experimental bounds and QoS. Dynamic and adaptive approaches are presented.
We discuss the role of uncertainty in the resource/service provisioning, investment, operational cost, programming models, etc. that have not yet been adequately addressed in the scientific literature. We discuss several major sources of uncertainty: dynamic elasticity, dynamic performance changing, virtualization, loosely coupling application to the infrastructure, among many others. A workload in such an environment is not predictable and can be changed dramatically. It is impossible to get exact knowledge about the system. Parameters such as an effective processor speed, number of available processors, and actual bandwidth are changing over the time. Elastic escalation process has a higher repercussion on the QoS, but adds another factor of uncertainty.
Providers might not know the quantity of data and computation required by users. For example, every time when a user requires a status of his e-mail or bank account, it could generate different amount of data and take different time for delivering. A pool of virtualized, dynamically scalable computing resources, storages, software, and services add a new dimension to the problem. 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.
We also discuss a model for cloud computing applications, called CA-DAG. This communication-aware model allows making separate resource allocation decisions, assigning processors to handle computing jobs, and network resources for data transmissions. We will discuss the benefits, weaknesses, and performance characteristics of such a model and resource allocation strategies in presence of uncertainty due to dynamic behavior of the execution context, job mix workloads, or uncertainty of the workflow properties.

Short Bio: Andrei Tchernykh is a researcher in the Computer Science Department, CICESE Research Center, Ensenada, Baja California, Mexico. From 1975 to 1990 he was with the Institute of Precision Mechanics and Computer Engineering of the Russian Academy of Sciences (IPMCE, Moscow, Russia). He received the Ph.D. in Computer Science from IPMCE in 1986. In CICESE, he is a coordinator of the Parallel Computing Laboratory. He is member of the National System of Researchers of Mexico (SNI), Level II. He leads number of national and international research projects. He served as a program committee member and orginizer of professional international conferences. His main interests include scheduling, load balancing, adaptive resource allocation, scalable energy-aware algorithms, multi-objective optimization, heuristics, meta-heuristics, and incomplete information processing (http://usuario.cicese.mx/~chernykh/).

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---