Peer-to-peer Cooperative Scheduling Architecture for National Grid Infrastructure

Ludek MATYSKA, Miroslav RUDA & Simon TOTH

CESNET, CZ

For the last ten years, scheduling of computational jobs across MetaCentrum (Czech national grid) is managed by one, central PBSPro installation. This centralized approach supported the possibility to schedule jobs between different clusters (spread across whole Czech Republic), with full control of complete situation of all clusters, with shared fair-share policy for users and with better support for large jobs, running across different clusters. Development effort was concentrated on improving stability of such setup (especially in case of instability of national network connecting different clusters) and support for advanced scheduling methods and virtualization. With growing number of clusters and processor, this setup becomes problematic and may become single point of failure and scalability bottleneck.

As a result, a new scheduling architecture is proposed, which relies on higher autonomy of clusters. It is based on a peer to peer network of semi-independent schedulers for each site or even cluster. Individual schedulers maintain their assigned clusters--allowing e.g. to submit jobs locally even if the external connectivity is lost--while cooperating with its peers to support features that mimics the centralized planning. Namely, the system still supports central accounting of jobs, fair share of computational resources across all sites of the MetaCentrum and scheduling jobs across resources. As MetaCentrum provides also virtualized resources (moving towards cloud provisioning), the scheduling system is being integrated with the Magrathea system to support scheduling of virtual clusters, including the setup of their internal network.

In parallel to the change of the overall architecture, the scheduling system itself is being replaced. Instead of PBSPro, chosen originally for its declared support of large scale distributed environment, the new scheduling architecture is based on the open-source Torque system. The implementation and support for the most desired properties in PBSPro and Torque are discussed and the necessary modifications to Torque to support the MetaCentrum scheduling architecture will be presented, too.