Contribution ID: **10** Type: **Presentation**

Cloud-enabled, scalable Data Avenue service to process very large, heterogeneus data

Wednesday, 11 November 2015 17:00 (20 minutes)

Compute-intensive applications such as simulations applied in various research areas and industry require computing infrastructures enabling highly parallel, distributed processing. Grids, clusters, supercomputers and clouds are often used for this purpose. There also exist tools that allow easier design and construction of such complex applications, typically in the form of workflows, which tools can utilize various types of distributed computing infrastructures (DCIs) and provide automated scheduling, submission and monitoring of workflow tasks (jobs). Some tools support job-level granularity, that is, each job in a workflow may potentially be executed in a different computing infrastructure.
Numerous storage solutions exist, however, storage resources accessible from within a given DCI are often limited by the protocols supported by the computing elements themselves. Binding jobs to a particular storage resource makes very difficult to port the workflow to other computing resources, or exchange data between different DCIs. To alleviate this problem a data bridging solution had been proposed, called Data Avenue, through which all common storage operations (such as listing, folder creation, deletion, renaming) and data access (download/upload) can be done on a wider set of storage resources (SFTP, GridFTP, SRM, iRODS, S3, etc.) using a uniform web service (HTTP) interface. Jobs, in this way, become capable of accessing diverse storage resources regardless of the DCI where the job is currently being run, resulting in more flexible and portable workflows.
Such a mediation service however occasionally implies very high CPU and network load on the server, as data exchanged over a storage-related protocol between the Data Avenue server and the storage has to be converted to HTTP established between the Data Avenue Server and the client. On massive, concurrent use, such as running parameter sweep applications where thousands of jobs may run in parallel, a single Data Avenue server could soon become a bottleneck, and clients may experience a significant decline in transfer rate. On the other hand, such peak loads are often followed by idle periods, when Data Avenue host will be underexploited.
This presentation introduces a solution to scale Data Avenue (DA) services on-demand by multiplying the available Data Avenue servers. The solution uses cloud infrastructure (IaaS) to dynamically grow or shrink the capacity of the server depending on the current load, composed of architectural components: load balancer, cloud orchestrator, VM pool, and a common database. Load balancer is responsible for dispatching client requests to one of the servers in the VM pool, which contains virtual machines having individual Data Avenue services pre-installed. Cloud orchestrator continuously monitors the load of VMs in the pool, and based on predefined load thresholds, starts new or shuts down instances, respectively. A common database to which each DA VM connects persists data of client interactions over lifetimes of individual DA VMs. An important advantage of this solution is that clients communicate with a single Data Avenue endpoint (load balancer), whereas mechanisms behind the scenes are hidden. Details of the proposed solution and preliminary experimental results are also reported.

Primary author: KACSUK, Peter (MTA SZTAKI)

Co-authors: Dr HAJNAL, Akos (MTA SZTAKI); Dr TUSA, Francesco (University of Westminster)

Presenter: KACSUK, Peter (MTA SZTAKI)

Session Classification: Exploiting the EGI Federated clouds - Paas & SaaS workshop