Migrating Scientific Applications from Grids and Cluster Computing into the Cloud - Issues, Challenges and Best Practices

The typical large users of Grid computing today seems to be at the cross-roads. Scientists and commercial ventures (like the financial firms on Wall Street) using large scale grid systems face several dilemmas – most significantly to either use an existing scientific application package with associated libraries already successfully developed, compiled and deployed for a specific grid infrastructure configuration albeit inefficiently or rewrite from scratch the very same application for use over an existing public or private cloud offering – either mostly over an Infrastructure as a Service cloud or in the rare case, over a Platform as a Service Cloud. On the one side, is the availability of the 'legacy' grid infrastructure with perhaps outdated software lacking continuous support and improvisations while on the other hand is the allure, despite costs, of re-architecting and reimplementing the same application leveraging the abstractly defined cloud programming interfaces. In addition, and apart from the technological challenges, the primary philosophy of 'pay per use' in the world of cloud does not seem to be fully aligned with the approaches of grid users in obtaining governmental funding as well as its expenditure thereafter. In this paper, we focus on a middle path between the above two extremes - the path to migrate substantial parts of existing and specific grid based scientific applications and libraries to appropriately leverage the right cloud service abstractions. While not all such grid based applications share commonalities or the ease of migration, certain patterns of distributed computations – certain key grid patterns - emerge and can used extensively to aid the migration. We share an approach and a methodology on migrating certain classes of scientific applications stripped off the domain specificities from the world of certain classes of grid infrastructures to that of either public cloud services like the AWS from Amazon or more appropriately and easily to a privately deployed open source cloud based on Eucalyptus or perhaps using Hadoop. Core Cloud services abstractions leveraged include the compute (like AWS EC2), storage (like AWS's EBS and S3) as well as communication bandwidth to optimize on the data transfer costs. We also share the application of this approach and methodology to certain other classes of scientific applications to the open source Hadoop deployed over a cluster of networked workstations. Our approach and methodology has been formulated based on the studies of the structure and code of the classes of scientific applications and the best practices followed in using them. In this paper, we highlight the issues and challenges that throw up in these typical migration efforts and the best practices we have formulated to mitigate them. Much of the rationale for these have come from studies of certain grid applications and their executions as well as experimental migration approaches of work in progress. References 1. Using Clouds to Address Grid Limitations, Giancomo V McEvoy, Bruno Shultze, Proc ACM MGC'08, 2008 2. Transparent Grid Enablement of Weather Research and Forecasting, Masoud Sadjadi et al, Proc 15th ACM MardiGras Conference, 2008 3. An EGEE Comparative Study – Grids and Clouds – Evolution or Revolution https://edms.cern.ch/document/925013/

Primary authors : Dr. THENALAPADI, Mohan S (Infosys Technologies Ltd)