Integrating your repository with Cloud infrastructure-benefits and challenges identified in the DuraCloud pilot program Author(s):

Kimpton, Michele (mkimpton@duraspace.org)

Abstract: Cloud compute and cloud storage are now available from several commercial providers around the globe as a commodity service. Due to the extremely low cost, ease of use, and instant scalability the Academic community is taking a hard look at how to best utilize this resource as an extension of its own IT environment. Even though there has been much interest in using cloud infrastructure few organizations have seriously integrated the cloud as part of their environment due to several challenges foreseen, whether real of perceived. In a large survey undertaken by the DuraSpace organization in Winter 2010, some of the biggest challenges identified by our community were security, performance, reliability and trust. Little data has been published to either validate or discredit the key challenges noted within the academic community, although much anecdotal information exists regarding site outages, poor performance, loss of data and the like. The purpose of this presentation is to present the findings of a large scale pilot program utilizing cloud infrastructure from multiple commercial cloud providers, as a utility. The presentation will discuss the key challenges and benefits identified when using cloud storage and compute as a utility during the pilot program. The presentation will provide detailed analysis, where possible, across multiple cloud providers. The analysis presented will include, when applicable, what solutions were deployed to overcome security, reliability, performance and other identified technical issues. Fifteen institutions participated in the pilot program and utilized cloud infrastructure from up to three commercial cloud providers through a mediating platform developed by the DuraSpace organization, named DuraCloud. DuraCloud is an open technology platform the makes it easy for organizations to utilize cloud storage and compute from multiple cloud providers through one service platform. Each partner loaded data into the cloud from their repositories, ranging in size from a few hundred gigabytes to ten terabytes. Data types included images, video, audio and text documents. Over the pilot period the partners performed numerous compute operations on the data held in the cloud store. These operations were enabled through services available in the DuraCloud platform which ran in the cloud compute environment. Examples of service available were preservation support services such as bit integrity checking, file format conversion, replication to multiple cloud stores, and synchronization with local repository stores. Other services were access based services such as tagging, video streaming, image viewing and editing. Through this experience the DuraSpace team was able to analyze cost, reliability, ease of use, security issues and performance across three commercial cloud providers during the six month pilot period. Results and issues will be shared in the presentation, as well as key lessons learned integrating a local repository with cloud infrastructure.