

Session title: Tales from the Cloud: Experiences and Challenges
Presenter(s): Bridger Dyson-Smith, University of Tennessee, Knoxville
Declan Fleming, University of California, San Diego
Ryan Steans, Texas Digital Library
Andrew Woods, DuraSpace
Time of session: Sunday, November 4, 10:15 am-noon
Room: Lawrence AB

Notes on session

Tales from the Cloud: Experiences and Challenges

Ryan Steans - Texas Digital Library

It's weirdly comforting to see that the issues faced by Texas Digital Library seem familiar: too much work, too little staff; IT infrastructure woes; restrictive campus environment; etc.

Buy in cost in terms of complexity for AWS is low, i.e.- the service contracts are simple and straightforward. In spring 2011, TDL went all in with AWS. Facilitated virtualization and hence rapid deployment.

TDL has no horror story about the migration to AWS. Downtime for the switch was 10-15 minutes. Everything they do now sits on AWS; no local copies. They pay about \$5500/mo. for these services. Ironically hitting the \$5000 cap means they may have to put it out to bid.

They did it without asking permission from the university ('beg forgiveness ...'). Drove down costs: less hardware, etc. Members of the TDL have not noticed and frankly don't really care. Run the whole thing with 2 local sys admins.

TDL has not seen any performance issues (in terms of bandwidth and response/transfer times) with using AWS.

Their experience has been largely positive, but he does have concerns when it comes time to move large amounts of data.

Great line from Steans: "most university IT policies haven't caught up with the cloud." Much truth in that.

Bridger Dyson-Smith - UT Knoxville

Use Duracloud and set to expand their usage from 3TB to 15TB.

They use it as a backup solution.

They're willing to support the cost given that they don't have local staffing to support another option.

Andrew Woods - DuraSpace

Part of the implementation team for the DuraCloud system.

Has built experience working with various cloud storage providers: Amazon, SDSC, etc., including some that no longer exist.

Enumerated many of the advantages of using the cloud, one of which is faster deployment.

He has found that it's seamless to work with storage vendors to get a checksum when an item is deposited, but after time has passed it's difficult to get them to generate a checksum for ongoing bit integrity checking.

One downside is the "perceived" loss of control.

One drawback is "eventual consistency." When a file is transferred to a local rack, one can see it immediately. Using large cloud providers there is a delay as load balancers do their part to push it around.

HTTP has its limitations and transfers have to be managed. Most cloud providers allow, theoretically, to push items up to 5GB, but in reality at 1GB the transfer failure rate climbs, so objects of that size or greater must be broken up for transfer and 'reassembled' on the other end.

Something he calls "server volatility." The server that supports your app may get rebooted, hardware gets upgraded, and so forth, and that can have strong impacts on your services. Possible to deal with this, but mechanisms must be in place and be effective.

Cloud providers are beginning to offer versioning options. Mixed blessing, since multiple copies have cost implications.

DuraCloud is run as an open-source project. One can download and use it, or use it as a service.

Declan Fleming - UCSD

Issue that forced them to go to the cloud was a project to work with big data (~100TB) coming from UCSD researchers.

Transition to SDSC OpenStack was fairly easy. Has two API choices: RackSpace, RESTian. They opted for RESTian. Required no caching layer; up and downloads were fast without that.

With OpenStack, only <5GB segments work. Performance same regardless of file size; just a bitstream as he put it.

Described their container vs. subdirectory discussion. Found the sweetspot around 100 containers, which happens to jive with what AWS and others tolerate.

Showed a cost model where labour costs were included, and then local and cloud storage come out very closely in terms of overall cost.

His cost model doesn't include some local costs: cooling, power, racking, facilities mgmt, hardware purchasing decisions and actions. Add those in and cloud looks even better.