

2015-10-26 Performance - Scale Meeting

Time/Place

- Time: 1:00pm Eastern Time US (UTC-4)
- Dial-in Number: (712) 775-7035
 - Participant Code: 479307#
 - International numbers: [Conference Call Information](#)
 - Web Access: <https://www.freeconferencecallhd.com/wp-content/themes/responsive/flashphone/flash-phone.php>

Attendees

- [Andrew Woods](#)
- [Nick Ruest](#)
- [Chuck Schoppet](#) (USDA/National Agricultural Library)
- [Unknown User \(daniel-dgi\)](#) (discoverygarden)
- [Maurice York](#) (University of Michigan)
- [Robert H. McDonald](#) (Indiana University)
- [Esmé Cowles](#) ★
- [Unknown User \(bbpennel\)](#)

Agenda

1. Voice F4 performance/scale interests and/or concerns
2. Review previous [F4 performance benchmarking \(summary\)](#)
 - a. Unimplemented "Technical Working Group" [performance assessment plan](#)
3. Establish focus for next round of F4 performance benchmarking
4. Establish collaborative benchmarking plan

Note: Once we have a new set of benchmarks, we can kickstart the follow-on effort of extending F4's scale.

Next call: [2015-11-09 Performance - Scale Meeting](#)

Minutes

1. Performance interests/concerns:
 - a. Esme: want to get a good benchmark that is meaningful, can be used to test changes
 - i. Also hearing concerns about Hydra/LDP/ORE proxies being chatty and slow, so want to look at client interaction model, look at both client and server improvements
 - b. Danny: also interested in a good baseline to have solid data to work on
 - i. Also interested in large binaries, like large video files
 - ii. As a vendor, interested in handling large datasets (several terabytes+), both in terms of large numbers of objects and large total size of preservation files
 - c. Chuck: Concerned about scalability in total number of objects. Have ~6 million objects now, anticipate 10-20 million.
 - i. Transaction performance: detecting duplicates with Solr
 1. Danny: good point about the new stack: how do external services like Solr and triplestores fit into transactions, performance questions?
 - d. Maurice: Mass-migration as content is moved into Fedora, performance of ingest/migration
 - i. Also interested in repository performance during migration, and the tradeoff between that and ingest performance
 - ii. Moving towards more active resource management, want to verify resources as they move into Fedora
 - iii. Want to plan migration based on expected rate of ingest, want to know how to improve performance if rate of ingest isn't as good as we'd like
 - e. Robert: Research datasets with Amazon infrastructure, interested in scalability in particular
 - f. Nick: Interested in ingest performance esp. for migrating from Fedora 3, with fixity checking
 - i. Large files (250GB video files)
 - ii. Clustering/sharding: need to figure out what role clustering can play and what scenarios it would help improve
2. Process:
 - a. Need to revisit benchmarks that we run to make sure they address the issues raised here
 - b. Second agenda item has links to previous work
 - i. There will need to be significant changes to revamp these to use a different testing tool (previous tool is no longer supported)
 - ii. Technical Working Group performance assessment plan divided up the testing space in slight different way
 - c. Can use existing test results to identify particular performance concerns that seem more likely to be an issue versus those that look like they are performing better
 - d. It would be good to re-verify some of the prior testing in a new framework that we can support moving forward
 - e. Esme: good to review the test results to see if they address the concerns voiced here, and if so, if the results look promising or raise concerns
 - f. Danny: the TWG work would be a good starting point to determine what kind of testing we should do
 - i. Have used JMeter before – regardless of what tool is used, need to coordinate and share tools
 - g. Nick: Sharing is key, would be good to setup an environment
 - h. Esme: We could setup a Vagrant setup to make an easy-to-setup environment for consistent testing and lower barrier to performing testing

3. Actions:

- a. Esme: review concerns raised today and determine how well it addresses concerns raised today
- b. Danny: look at creating sample dataset and specifying sample data (file sizes, number of objects, etc.)
- c. Maurice: help with specifying sample data
- d. Andrew: update wiki to organize ongoing and previous work
 - i. Update community with today's discussion and next steps