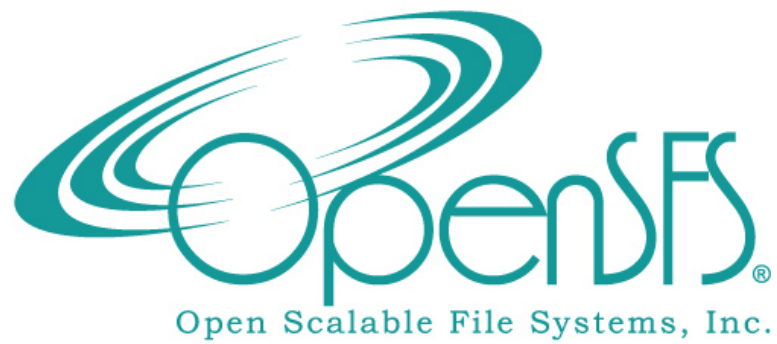




OpenSFS

Metadata Performance and Scalability

Technical Specification

Version 1.0

OpenSFS Solicitation Number CX589

April 4, 2011

Background

The work described herein is in support of continued scalability of the Lustre parallel file system. The Lustre parallel file system has been proven to scale to extremely large client counts, storage capacities, and storage bandwidth yet it faces challenges in key areas of scalability required by many HPC centers. To improve metadata scalability and performance, metadata services shall be developed which scale in performance and capacity as storage elements are added to the Lustre file system while making efficient use of the capabilities of multi-core systems.

1.0 Metadata server performance and metadata services scalability

1. Interactive workloads (ls -l, du) do not perform as well with Lustre as they do on local file systems. Furthermore, the MDS software architecture has not kept pace with the capabilities of current multi-core hardware architectures. It is a requirement that the MDS be capable of using efficiently all of the compute resources available on commodity server platforms.
2. The single metadata server is Lustre's greatest architectural liability. The file system provides horizontal scalability of the data store across multiple object storage servers, but the metadata services are still limited to a single metadata server. Lustre performance and capacity can be improved by enabling horizontal scale-out of metadata services allowing the file system namespace to be distributed. Maximal performance will result when directories themselves can be distributed across multiple MDS nodes.

The following objectives for the metadata server performance and metadata services scalability shall be met by Offeror's solution:

- Offeror shall propose a design for improved metadata server performance (1.) and metadata services scalability (2.). Offeror may alternatively elect to only address metadata server performance (1.) or metadata services scalability (2.).
- Offeror shall modify the Lustre architecture and source code to implement this design.
- All modifications shall be supported in the Offeror's product branch.

Offeror's proposal shall describe how proposed solutions will meet the requirements (and timelines) specified in Table 1 – Column Q2 2012. Proposals need not address each of the requirements set forth in Table 1 to be considered responsive but must indicate which requirements will be satisfied (or to what degree).

In addition, the offeror proposal should include ancillary features or functionality required to provide a production quality, robust solution that meets the metadata requirements. These additional utilities may not specifically result from the requirements, but are needed for a production release of Offeror's solution.

Metric	Lustre 2.0/2.1¹	Q2 2012	Q1 2014
maximum number of files in file system	4 billion	100 billion	1 trillion
maximum number of files in directory	10 million	50 million	10 billion
maximum number of subdirectories	10 million	1 million	10 million
maximum number of clients	131072	64 thousand	128 thousand
maximum number of OSS nodes	-	1 thousand	4 thousand
maximum number of OSTs	8150	2 thousand	8 thousand
maximum OST size	16 TB	32 TB	128 TB
maximum file system size	64 PB	100 PB	256 PB
maximum file size	320 TB	1 PB	-
maximum object size	2 TB	16 TB	64 TB
peak aggregate file creates/s	-	200 thousand	400 thousand
peak directory listings/s (ls -l)	-	40 thousand	100 thousand
maximum single client open files	~3 thousand ²	100 thousand	-
peak single client file creates/s	-	30 thousand	-

1.1 Deliverables

1.1.1 High-level design documents detailing Offeror's solution.

Completion: Start of Contract + 1 months

1.1.2 Detailed design documents detailing Offeror's solution.

Completion: Start of Contract + 2 months

1.1.3 Alpha delivery of Offeror's solution.

Completion: Start of Contract + 5 months

1.1.4 Demonstration of Alpha delivery. Offeror shall be provided access to a system to run the demonstration. The demonstration shall include a set of workloads provided by OpenSFS. Demonstration shall also include the validation of recovery under a variety of component and

¹ sourceSource : [Lustre 2.0 Manual, Table 5-1](#)

² Lustre does not impose a maximum for the number of open files, but the practical limit depends on the amount of RAM on the MDS. No "tables" for open files exist on the MDS, as they are only linked in a list to a given client's export. Each client process probably has a limit of several thousands of open files which depends on the ulimit.

server failure scenarios as specified by OpenSFS. Offeror's solution shall prove resilient under these failure scenarios.

Completion: Start of Contract + 6 months

1.1.5 Beta delivery of Offeror's solution.

Completion: Start of Contract + 10 months

1.1.6 Demonstration of Beta delivery. In addition to demonstration of all functional elements of the Alpha delivery, the Beta delivery shall at a minimum support all previously supported fault-tolerant mechanisms and overall stability of the Lustre file system.

Completion: Start of Contract + 11 months

1.1.7 Final delivery of Offeror's solution. The final delivery shall be a supported product by Offeror and be integrated within the canonical Lustre source.

Completion: Start of Contract + 14 months

1.1.8 Demonstration of final delivery. In addition to demonstration of all functional elements of the Alpha and Beta deliveries, the final delivery shall address any issues identified in the Alpha and Beta demonstrations.

Completion: Start of Contract + 15 months

2.0 Execution

2.1 Management and Reporting

Management oversight is important for keeping activities on track and within scope. As part of each deliverable, the Offeror shall exercise due diligence in management activities, including regular attendance and participation in management meetings as described below.

Weekly technical management: Once per week, the Offeror's technical representative(s) shall meet with OpenSFS's technical teams and Offeror's technical team to discuss progress on tasks detailed within this document. The Offeror shall provide current status on all tasks and be prepared to discuss current status in technical detail. Any resource dependencies, such as access to OpenSFS testing equipment must be identified in these meetings.

The Offeror shall submit a monthly progress report no less than two business days after the last day of each month detailing progress made on each deliverable. The Offeror shall provide documentation of completion for each deliverable prior to billing.

OpenSFS's technical representative must document completion of a deliverable in order for it to be considered complete.

The Offeror (Project Manager or Support Senior Manager) shall attend quarterly executive management meetings with OpenSFS management to discuss progress on work items as well as broader architectural issues on the Lustre technology, including technical detail and forward-looking feature development. The communication method for these meetings (e.g. face-to-face, conference call, etc.) shall be determined by mutual agreement.

2.2 Software Licensing and Source Code

All software developed and data generated within the scope of this agreement are subject to the OpenSFS contributor agreement. Offeror shall also release all developed and modified code via GPL V2 or other open source license reasonably designated by OpenSFS. Source code for all software developed under this agreement shall be hosted at a repository as specified by OpenSFS.

2.3 Lustre Software Enhancements

All changes to Lustre software shall be made available in a production release of Lustre 2.x except where otherwise noted. All changes to Lustre must pass OpenSFS's Lustre test suite (negotiated with Offeror) in order to successfully complete delivery milestones. It is desired that Offeror's changes to the Lustre file system shall be backward compatible with non-enhanced Lustre clients, servers, and routers.

2.4 Access to Hardware Resources

Access to facilities and equipment required to facilitate the work to be performed may be negotiated.