Intel® Lustre* Data on MDT/Small File I/O

Mikhail Pershin

April 8, 2014

** Some name and brands may be claimed as the property of others.*

Problem with small files

- Lustre* file system read/write performance is currently optimized for large files
- In addition to the initial file open RPC to the MDT, there are separate read/write RPCs to the OSTs to fetch the data, as well as disk IO on both MDT and OST
- This hurts small file performance significantly when there is only a single read or write RPC for the file data

*some names and brands may be claimed by others

What Data-on-MDT will give us?

The Data On MDT (DOM) project aims to improve small file performance by allowing the data for small files to be placed only on the MDT, so that these additional RPCs and I/O overhead can be eliminated, and performance correspondingly improved

What do we need to implement?

- A new layout for files with data on MDT
- Client is able to send IO requests to an MDT
- A mechanism to perform migration from MDT to OST

Project phase I

- Support basic DOM mechanism
- No auto migration
- DOM layout is set explicitly by `lfs` tool with manual migration

Project phase II

- Auto migration when file becoming too large
- Default directory or filesystem striping for new files

Project phase III

- Performance optimizations such as readahead during readdir or stat
- First write detection so files are only created on MDT when appropriate
- Reserve OST objects in advance even for small files

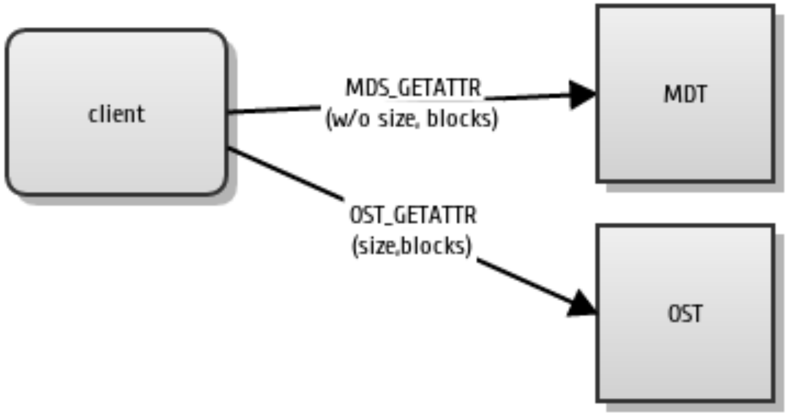
What are benefits?

There are several use cases we expect to see performance improvements:

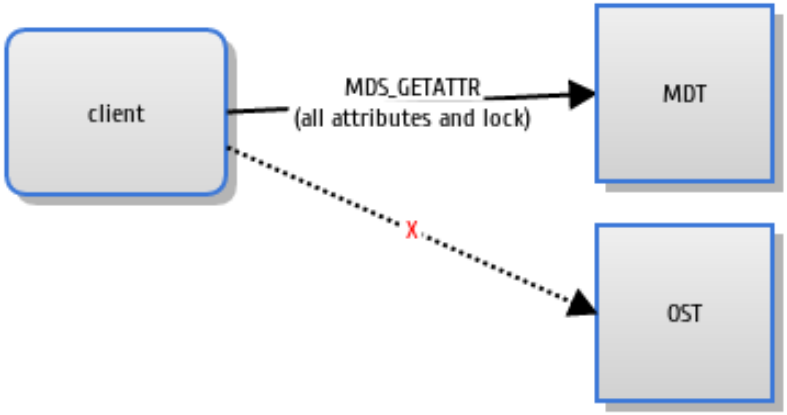
- MDS GETATTR (stat, readdir)
- MDS OPEN
- READ/WRITE
- Data read-ahead is possible during stat/readdir

STAT in Lustre* 2.5 vs Small Files

Lustre 2.5



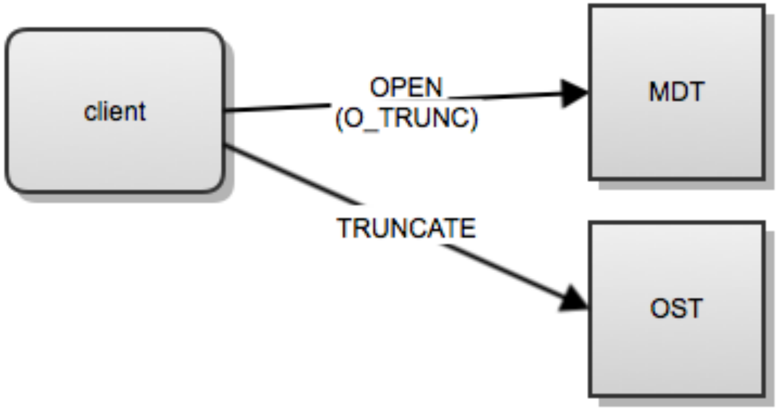
Small File IO



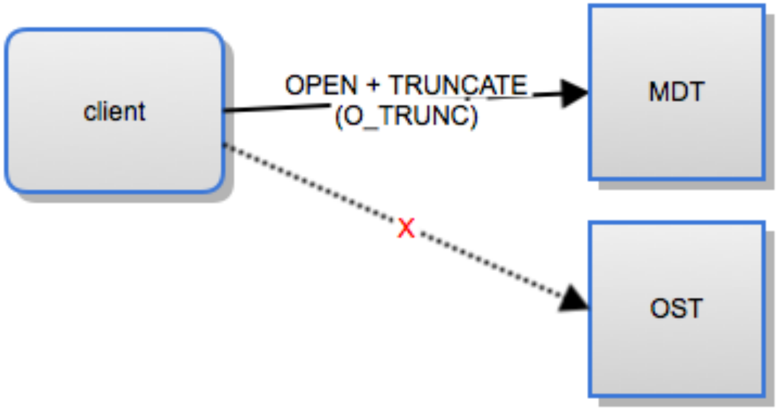
*some names and brands may be claimed by others

OPEN in Lustre* 2.5 vs Small Files

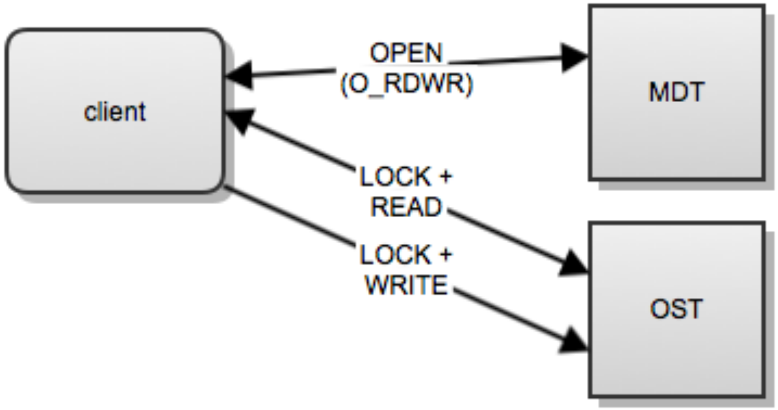
Lustre 2.5



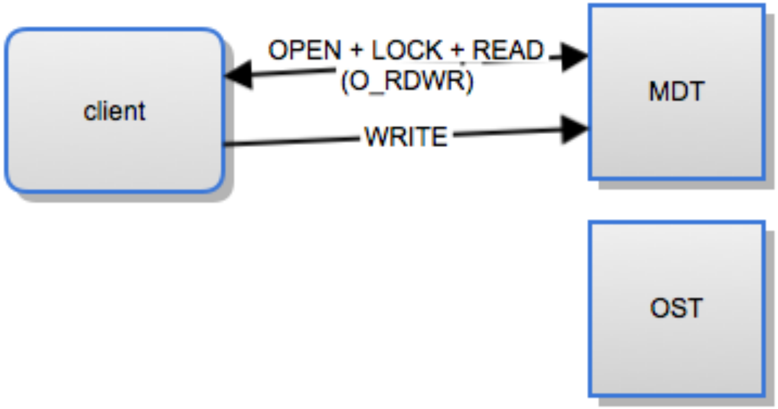
Small File IO



Lustre 2.5



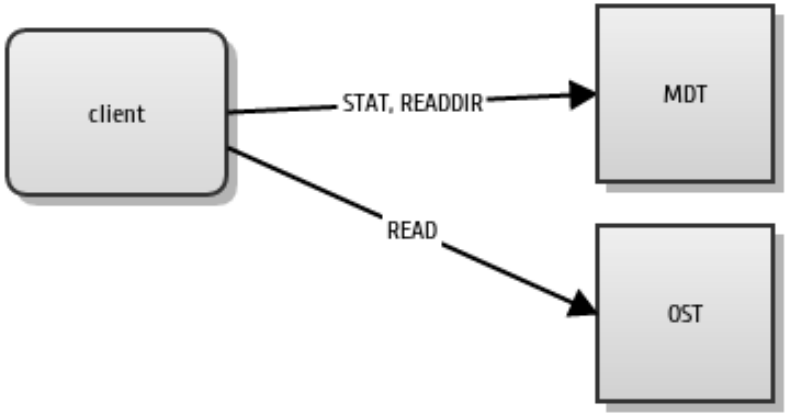
Small File IO



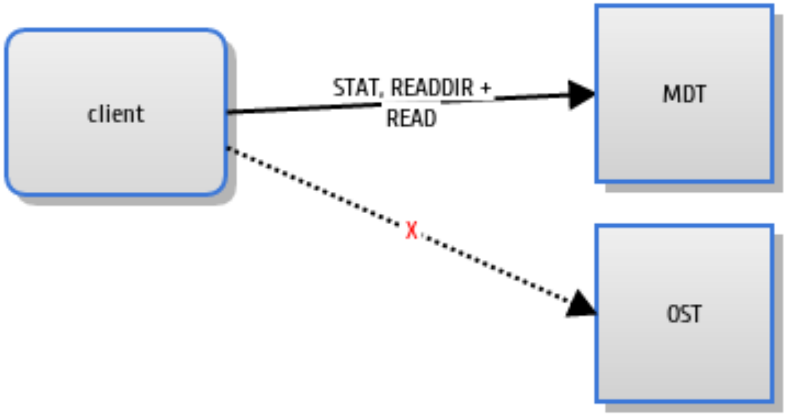
*some names and brands may be claimed by others

Small Files read-ahead vs Lustre* 2.5

Lustre 2.5



Small File IO



*some names and brands may be claimed by others

Performance benefits estimation

part of operation	server	time for operation, usec	Data-on-MDT goal
Write at the end of file			
open + lock	MDT	1296 + N	1296 + N
glimpse	OST	392 + N	0
IO lock	OST	1737 + N	0
IO	OST	1760 + N	1760 + N
		5185 + 4N	**3056 + 2N**
Read small file			
open + lock	MDT	844 + N	844 + IO read + N
glimpse	OST	333 + N	0
IO lock	OST	726 + N	0
IO read	OST	1392 + N	0
		3295 + 4N	**~ 1200 + N**
Stat of existent file with data			
getattr + lock	MDT	954 + N	954 + N
glimpse	OST	654 + N	0
		1608 + 2N	**954 + N**

Thank You