MarFS Metadata Scaling

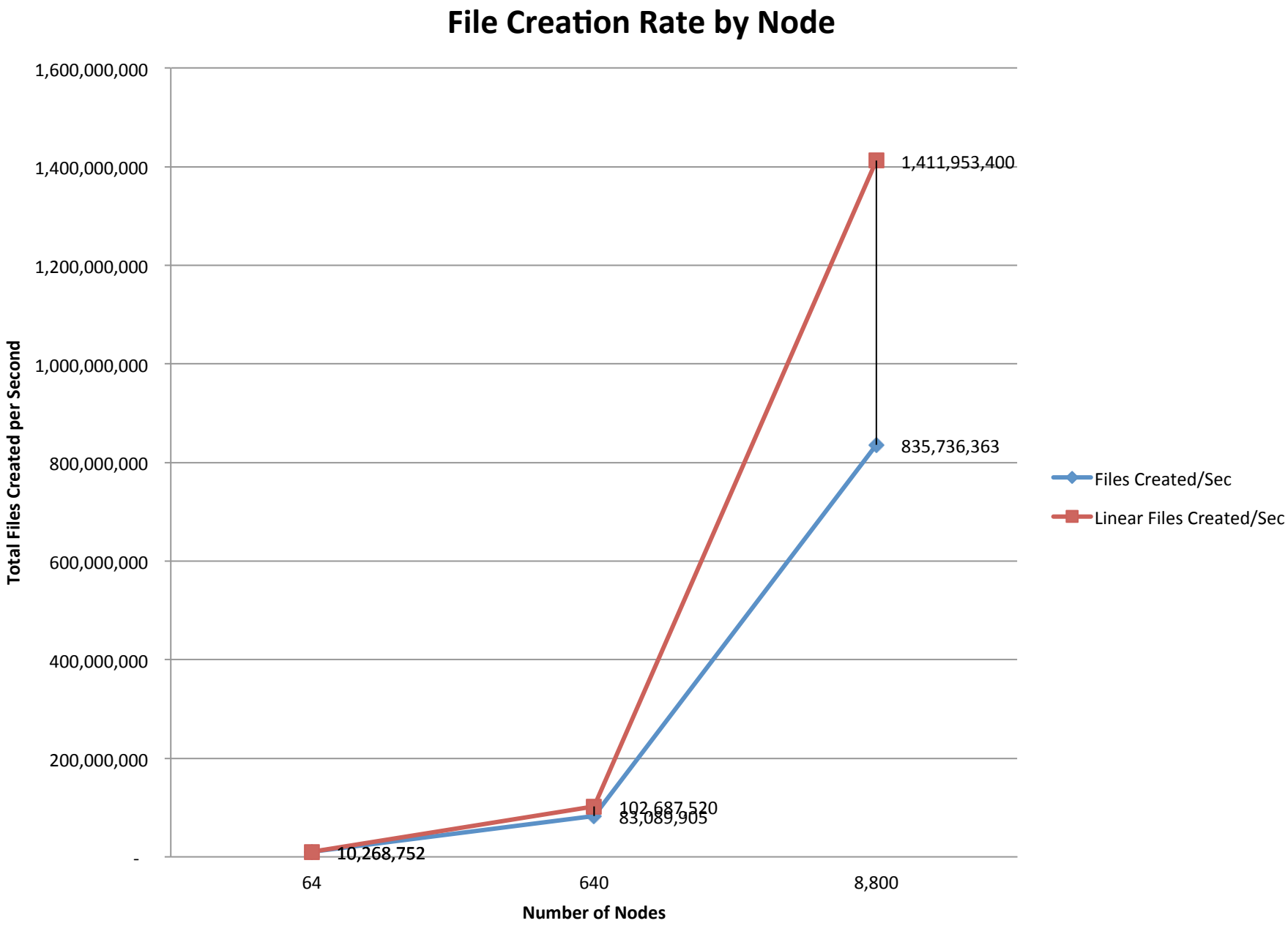
PDSW WIP Report 2016

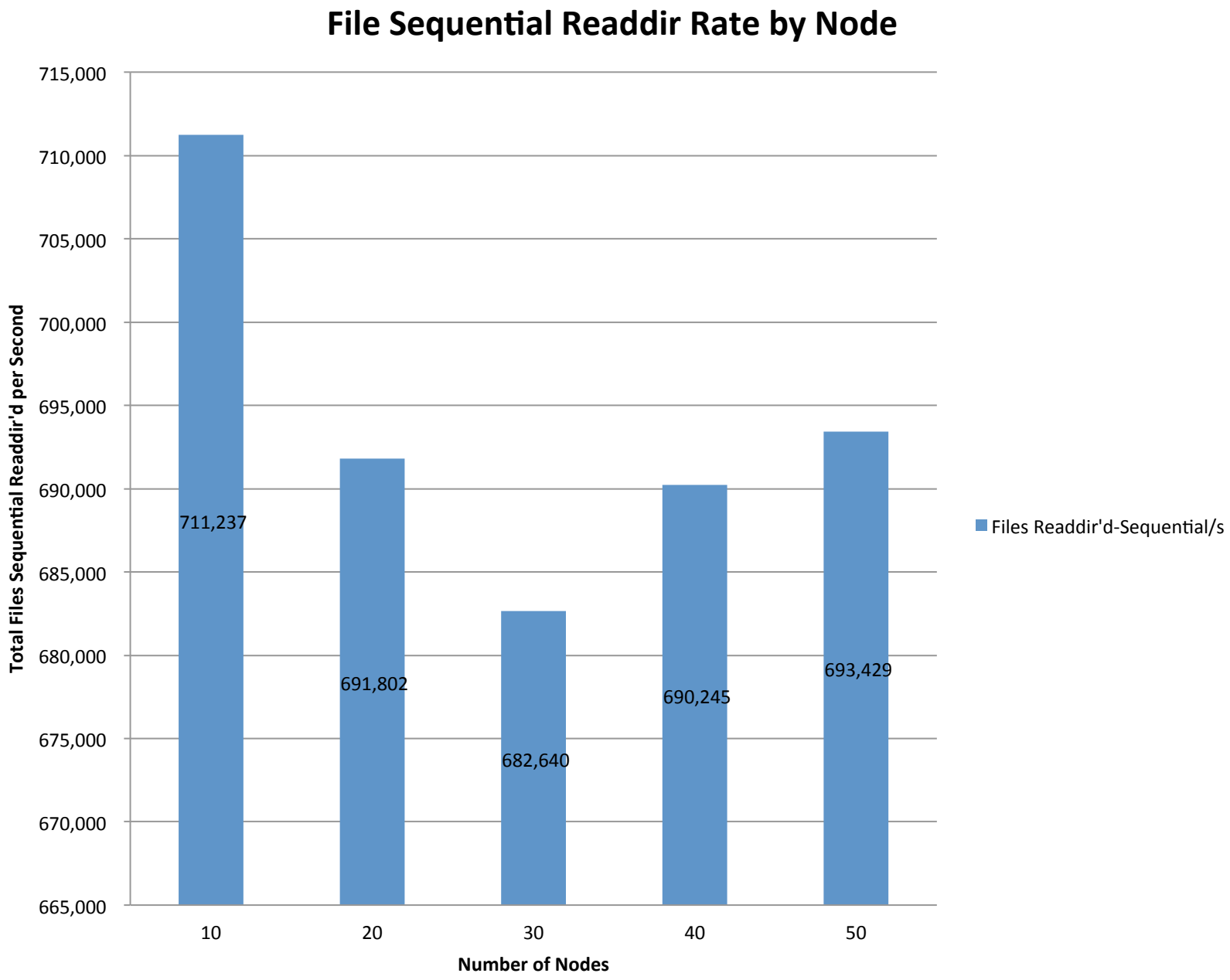
David Bonnie, Hsing-Bung Chen, Gary Grider, Jeffrey Inman, Brett Kettering, William Vining

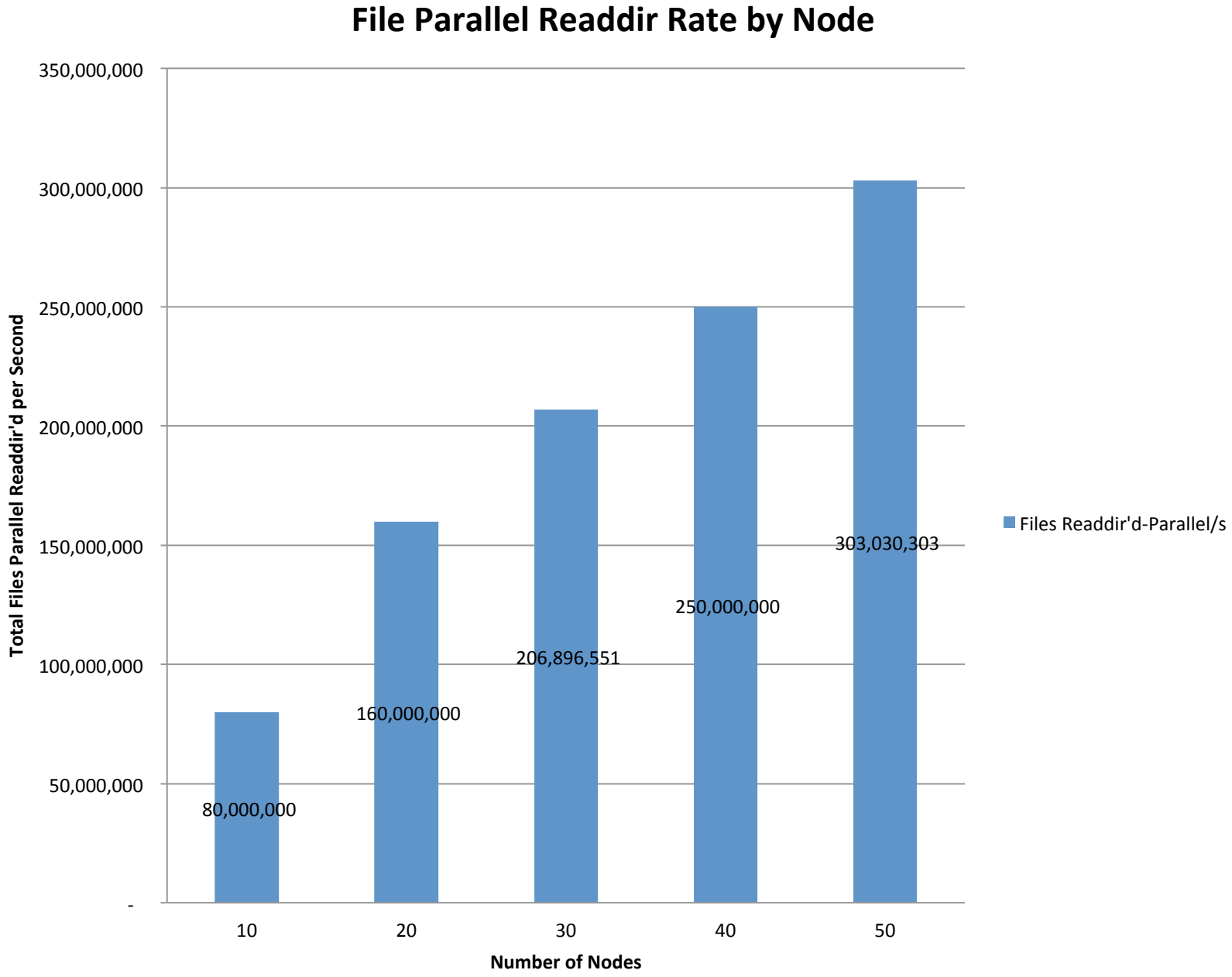
LA-UR 16-28615

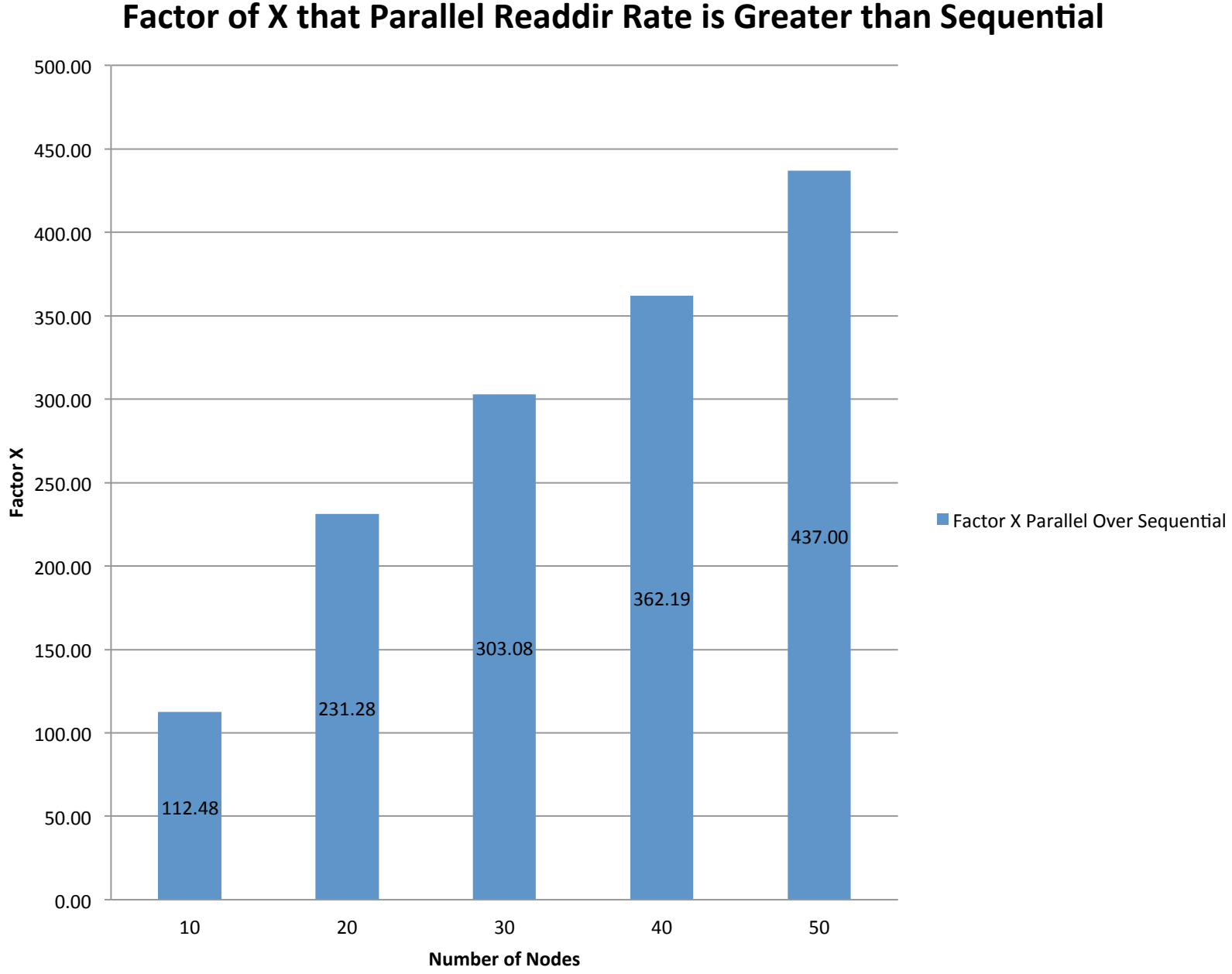
Metadata scaling components

- Deploy one drMDS per file system as rank 1 on first node
  - Make new directories & broadcast dir inode to fdMDSc's
- Deploy fsMDSc's on ¼ cores for each node in file system service
  - Handles its sharded part of distributed file metadata when broadcast commands are sent
- Deploy fsMDSp's on ¼ cores for each node in file system service
  - Handles its sharded part of distributed file metadata when command are sent to a specific fsMDSp.
- Deploy file system Clients on ½ cores for each node in file system service
  - Execute file system operations, such as create


File Creation Rate by Node
1,600,000,000
1,400,000,000
1,200,000,000
1,000,000,000
800,000,000
600,000,000
400,000,000
200,000,000
-
Total Files Created per Second
1,411,953,400
835,736,363
102,687,520
83,089,905
10,268,752
64
640
8,800
Number of Nodes
Files Created/Sec
Linear Files Created/Sec


File Sequential Readdir Rate by Node
Total Files Sequential Readdir'd per Second
715,000
710,000
705,000
700,000
695,000
690,000
685,000
680,000
675,000
670,000
665,000
711,237
691,802
682,640
690,245
693,429
10
20
30
40
50
Number of Nodes
Files Readdir'd-Sequential/s


File Parallel Readdir Rate by Node
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
-
Total Files Parallel Readdir'd per Second
80,000,000
160,000,000
206,896,551
250,000,000
303,030,303
10
20
30
40
50
Number of Nodes
Files Readdir'd-Parallel/s


Factor of X that Parallel Readdir Rate is Greater than Sequential
500.00
450.00
400.00
350.00
300.00
250.00
200.00
150.00
100.00
50.00
0.00
Factor X
112.48
231.28
303.08
362.19
437.00
10
20
30
40
50
Number of Nodes
Factor X Parallel Over Sequential

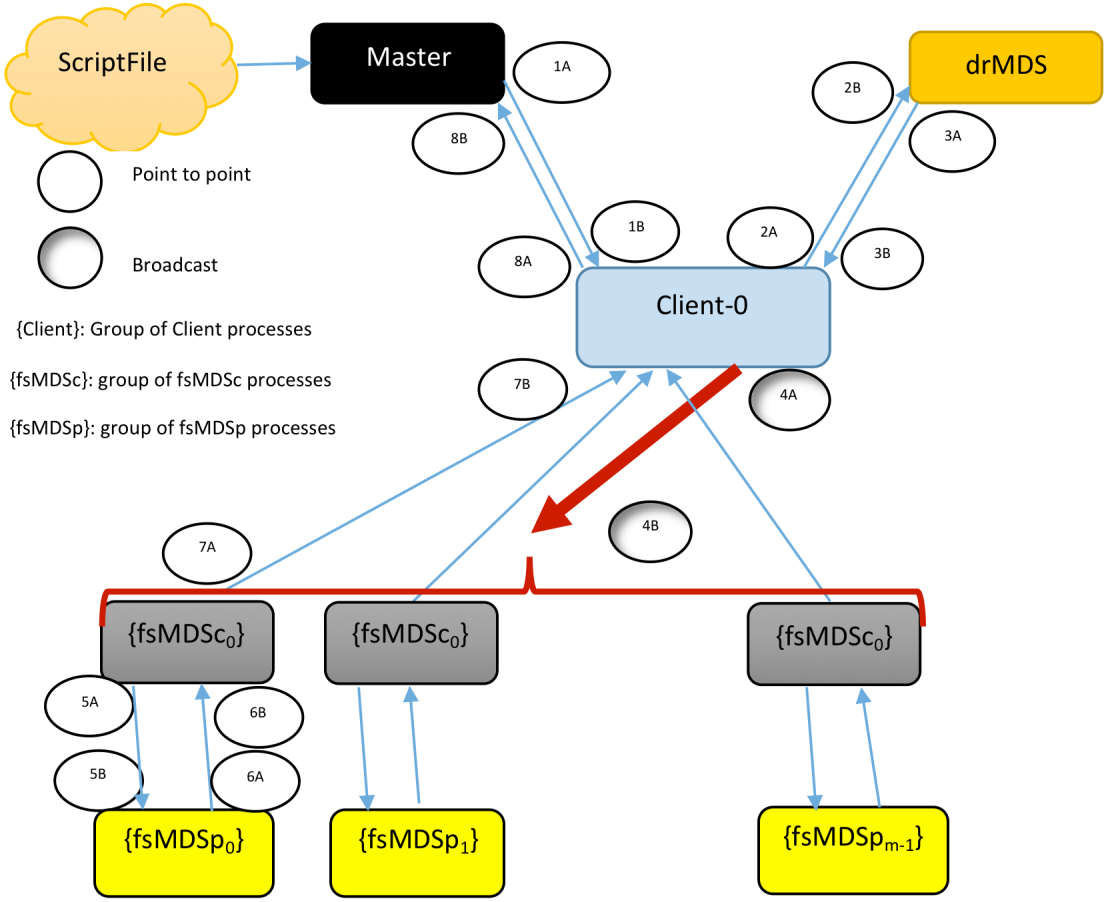
Background Information

MARFS METADATA SCALING

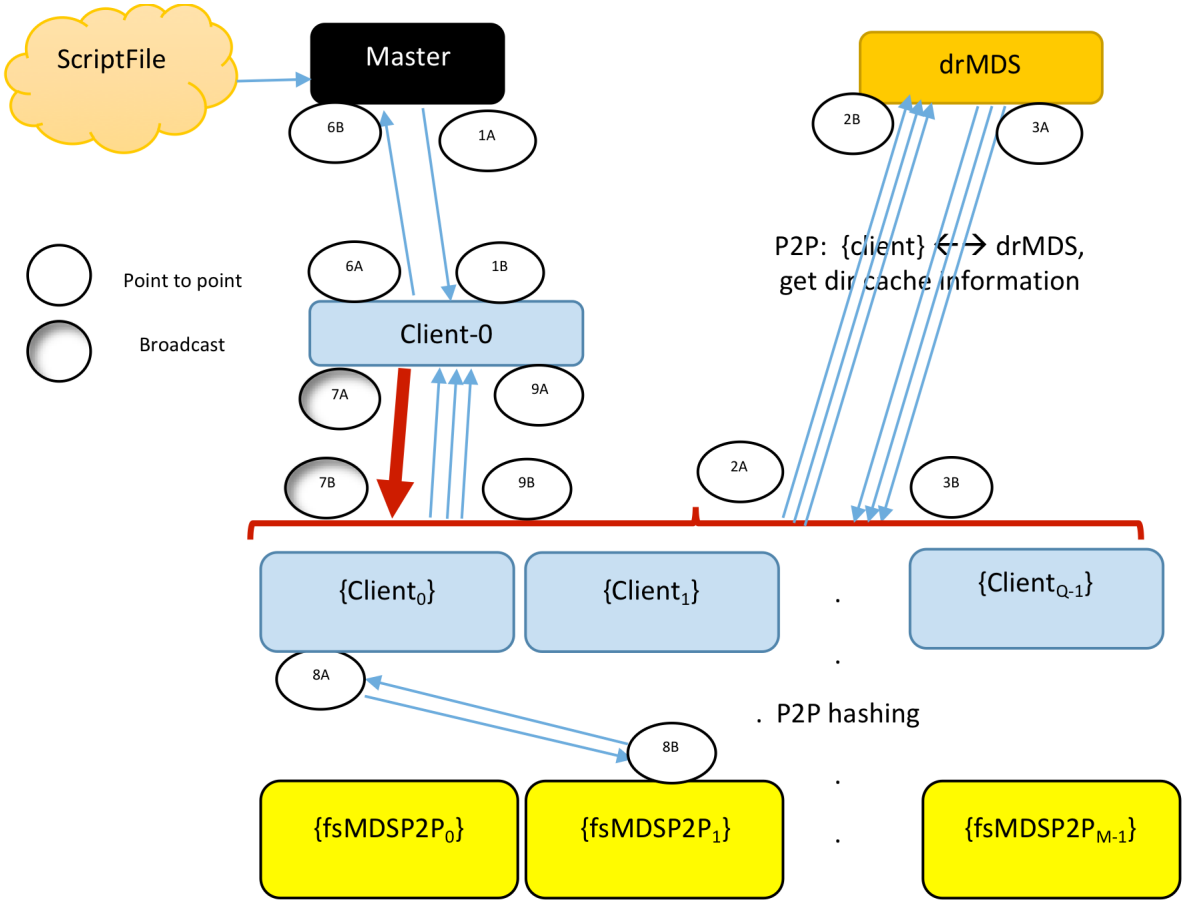
MarFS Overview

- Provides near-POSIX over cloud-style erasure and objects
  - Yields reliable storage on inexpensive disk
  - Supports legacy apps' files/folders/ownership/etc.
- Store large data sets for weeks to months on PFS, 1 TB/s
- Store data forever in archive, 10s GB/s
- Store large data sets for months to year'ish on MarFS, 100s GB/s
  - Data set O(PB), aggregate data O(EB)
- Systems growing from O(M) cores/O(PB) memory to O(B) cores/O(10s PB) memory
  - Going to O(B) files per job in one directory and O(10s T) files per file system

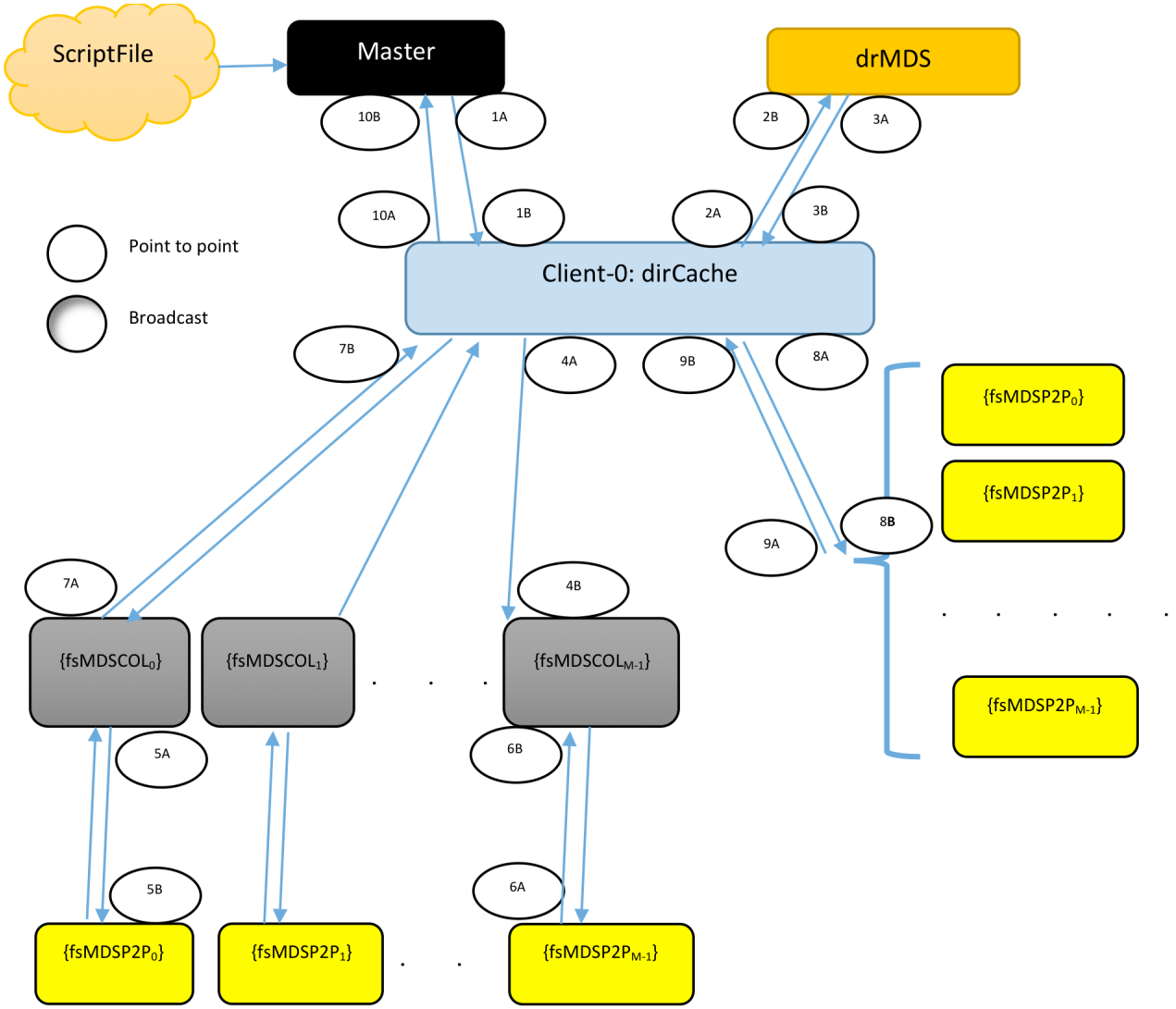
Here's a picture of creating a directory



Here's a picture of creating files



Here's a picture of sequential readdir



Here's a picture of parallel readdir