OAK RIDGE National Laboratory

Evaluation of DoM/DNE scaling performance in Lustre

James Simmons

Rick Mohr

May 2021

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Metadata EXA scaling challenges

- How many MDS servers do we need?
  - DoM increase space needs.
  - Do more MDS servers mean more performance
- Does moving to NVME devices make sense?
- What setup makes the most sense?
  - DNE Phase 1
    - Create directories on a specific remove MDT
  - DNE Phase 2
    - Enables deployment of striped directories on multiple MDS nodes
  - DNE Phase 3
    - Grow organically across MDTs

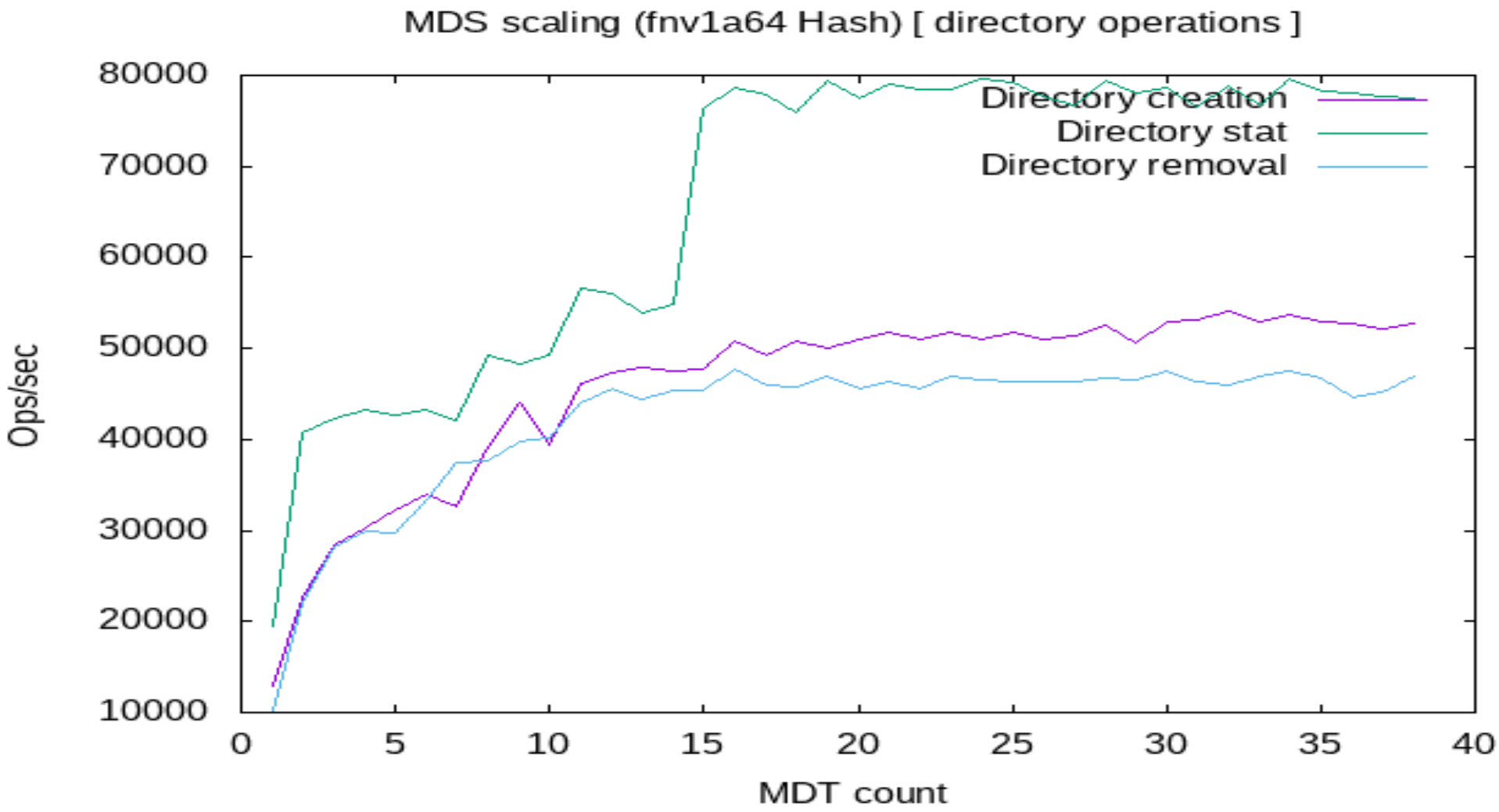
Testing Parameters

- Internal testbed cluster
  - 6 and 24 clients (Tyan B8026T70AV16E8HR / Dell PowerEdge R6415) tested
    - 7 ranks per node
  - DDN 14k mapped to 6 OSS each with 2 OSTs
  - 38 MDS servers (Dell PowerEdge R6415) with 1 NVME device 732.4G
- Directory setup (DNE2+)
  - lfs setdirstripe -c $mdt -i -1 $OUTDIR
  - lfs setdirstripe -D -c $mdt -i -1 $OUTDIR
  - lfs setstripe -c $OSTCOUNT --pool capacity $OUTDIR
  - lfs setstripe -L mdt -E 64K -E -1 $OUTDIR (DoM testing)
- Using lustre 2.14.51 from a few weeks ago. ZFS backend.

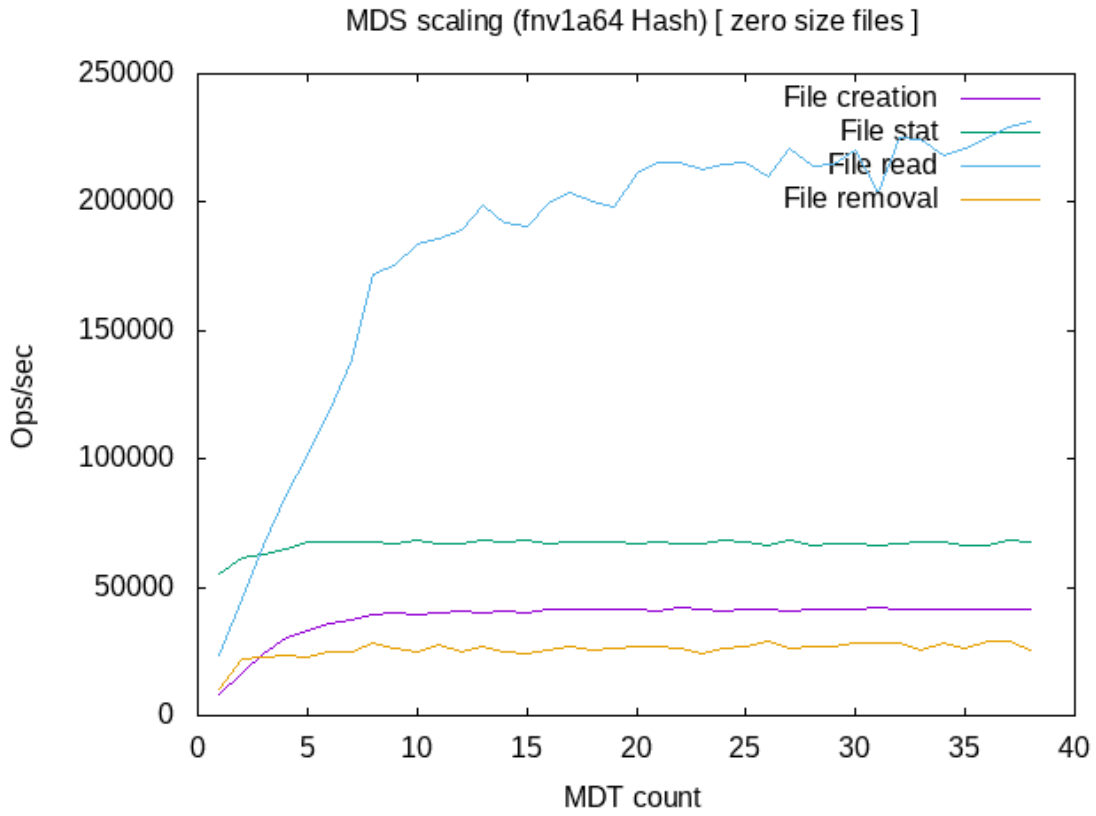
benchmarks

- mdtest
  - DNE2 shared directory testing (-g option to remove inherited striping)
    - $OUTPUT and test-dir.X-0 will be striped across several MDTs. Directories under test-dir.X-0 will be mapped randomly to one MDT of the striping set.
  - Zero size files, DoM and non-DoM testing of 32K files
  - 100K files created per node
  - mdtest -n $file_per_dir -g -i 3 -p 30 -N 1 '-w 32736 -e 32736' -d $OUTDIR
  - Ran with 7 ranks per node.

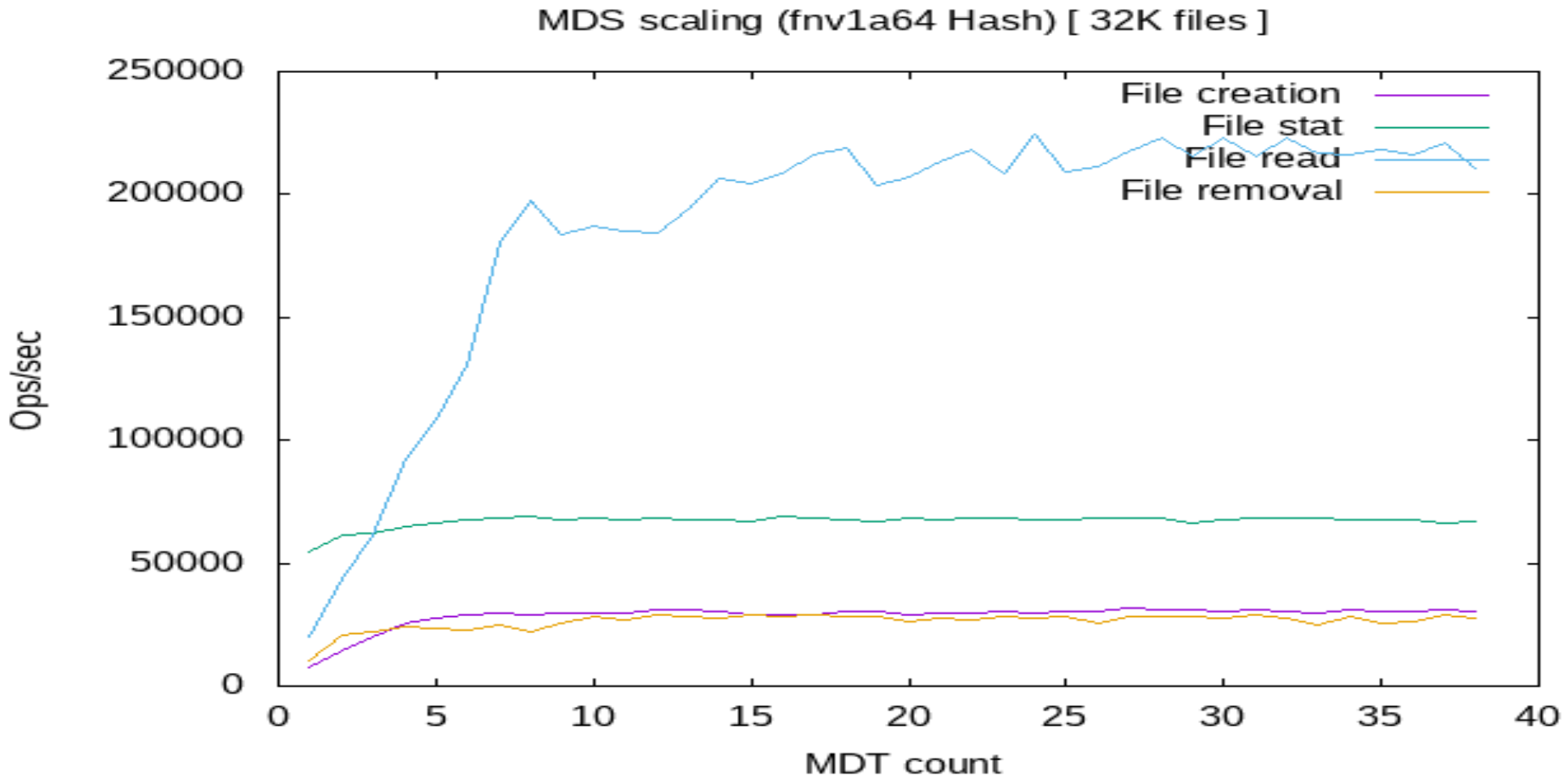
Shared directory performance



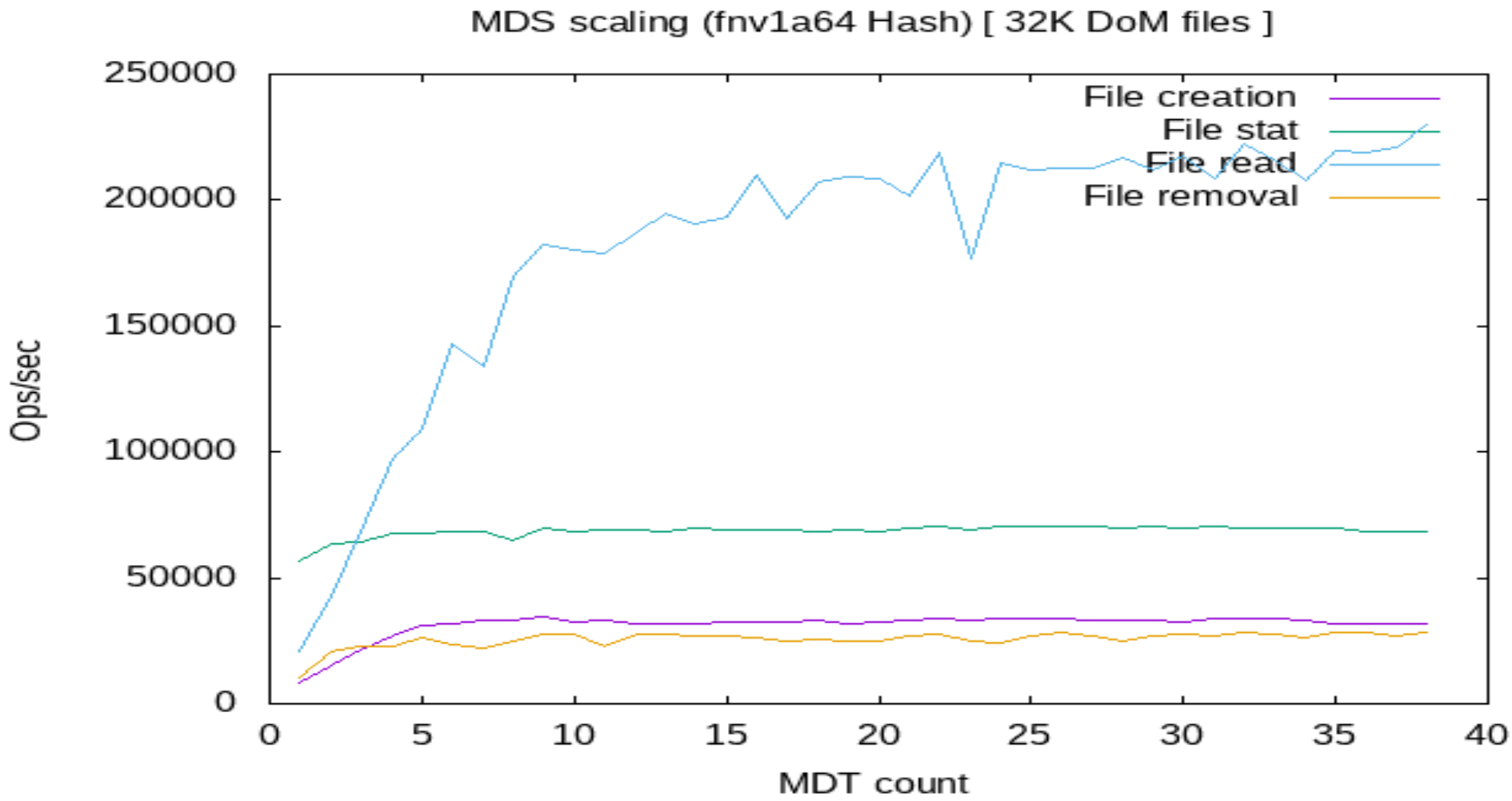
File performance for zero size files



File performance using OSTs



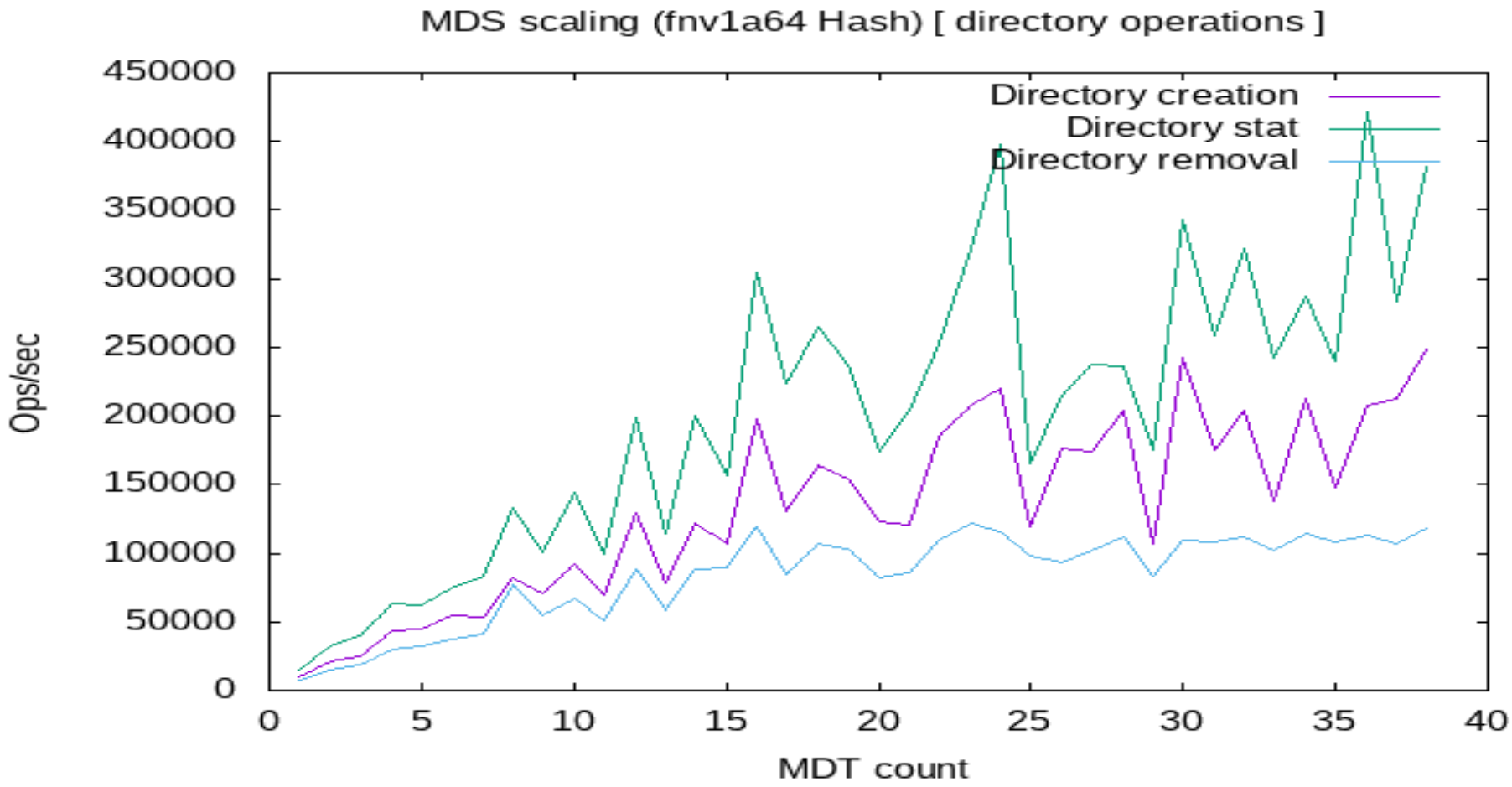
DoM file performance



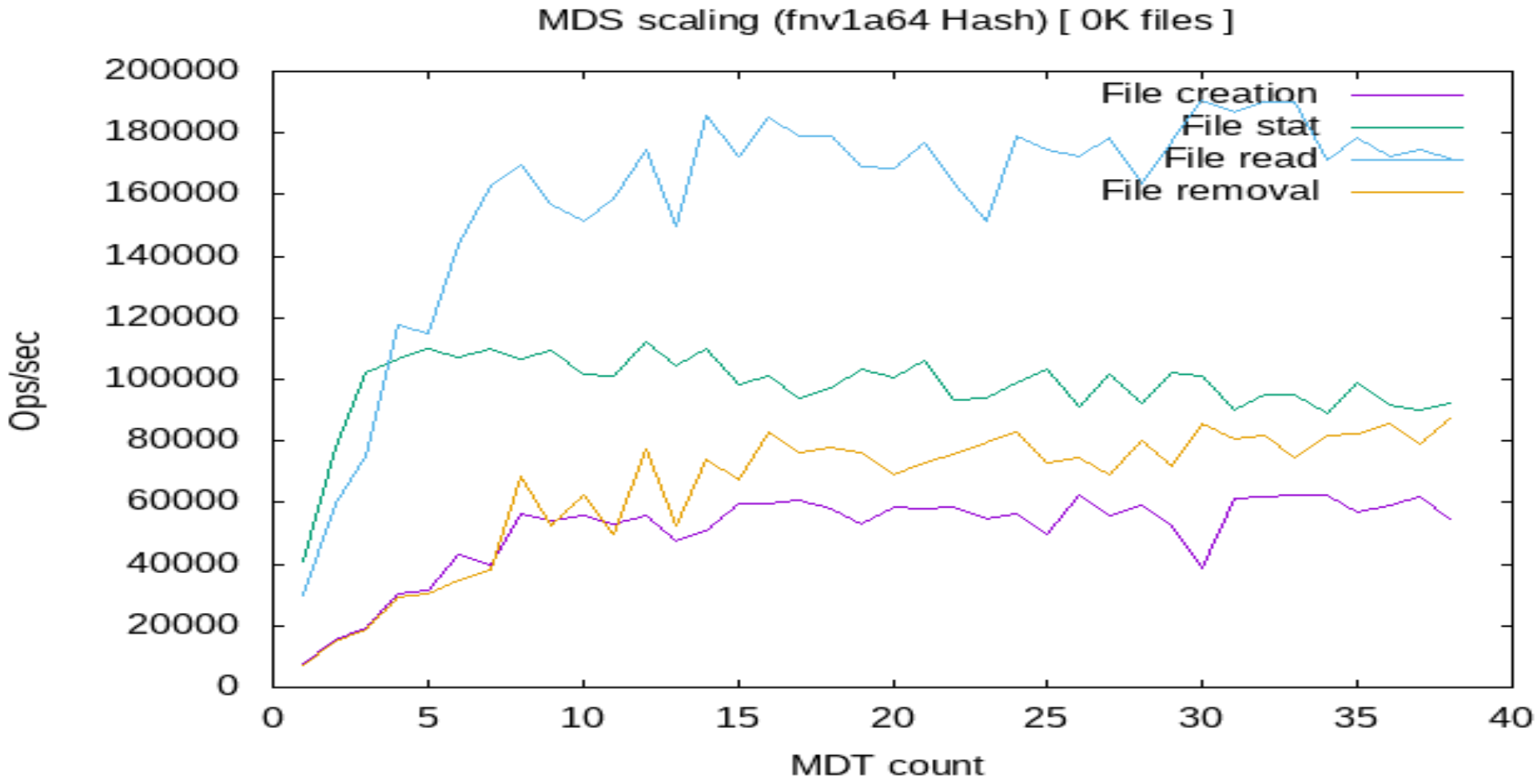
Reason for results

- lfs getdirstripe mdtest_tree.X.0 will show they are all mapped to the same MDT
  - Rank of 7 saturates data transfer to a single MDT.
  - Parallelization only from multiple client nodes.
    - Need many clients to get the most out of it.
- Next set of mdtest results are with unique directories per rank
  - Confirmed with lfs getdirstripe mdtest_tree.X.0 they are mapped to random MDT in the striping set.
  - With 7 ranks on 9 clients we out number the MDT count.

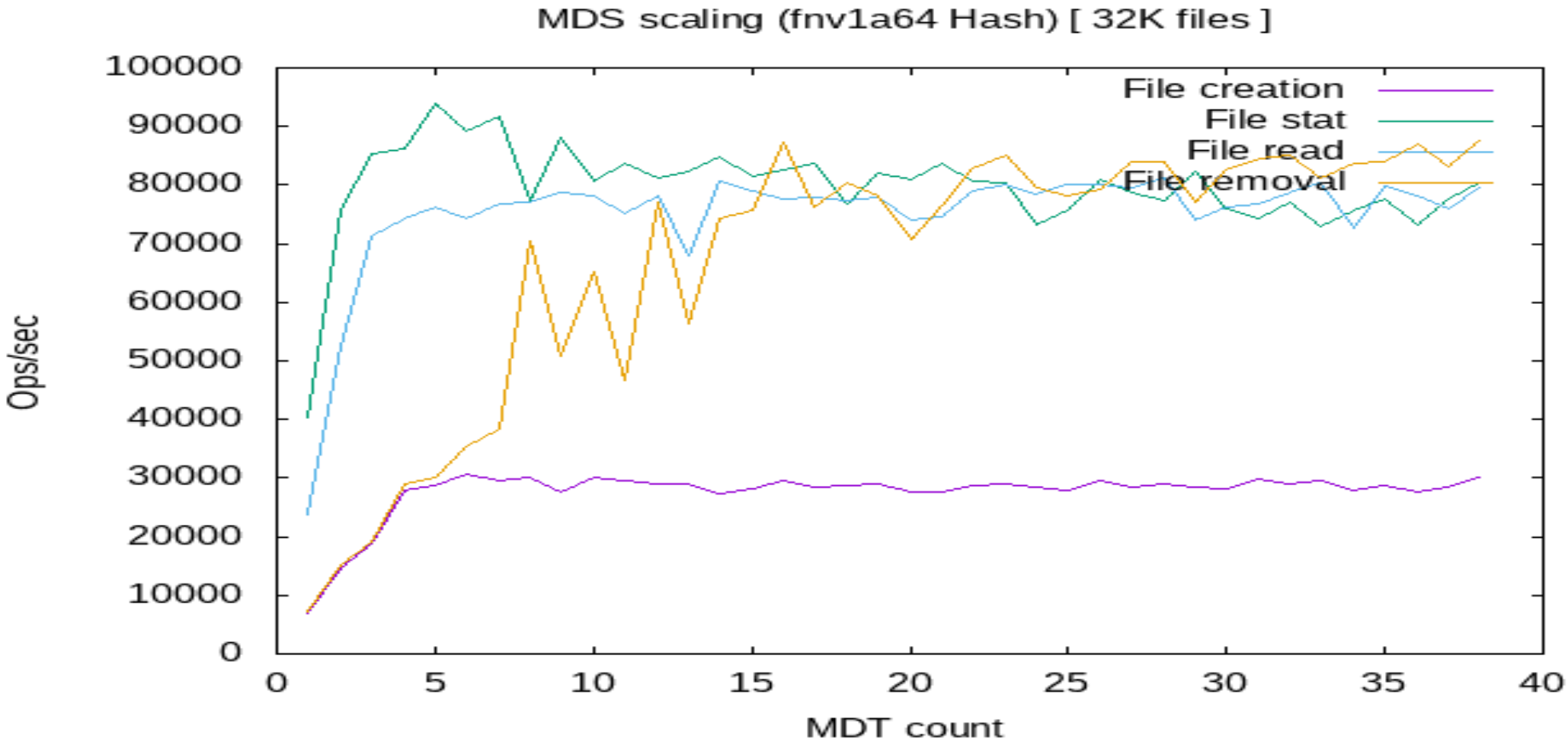
Unique directory performance



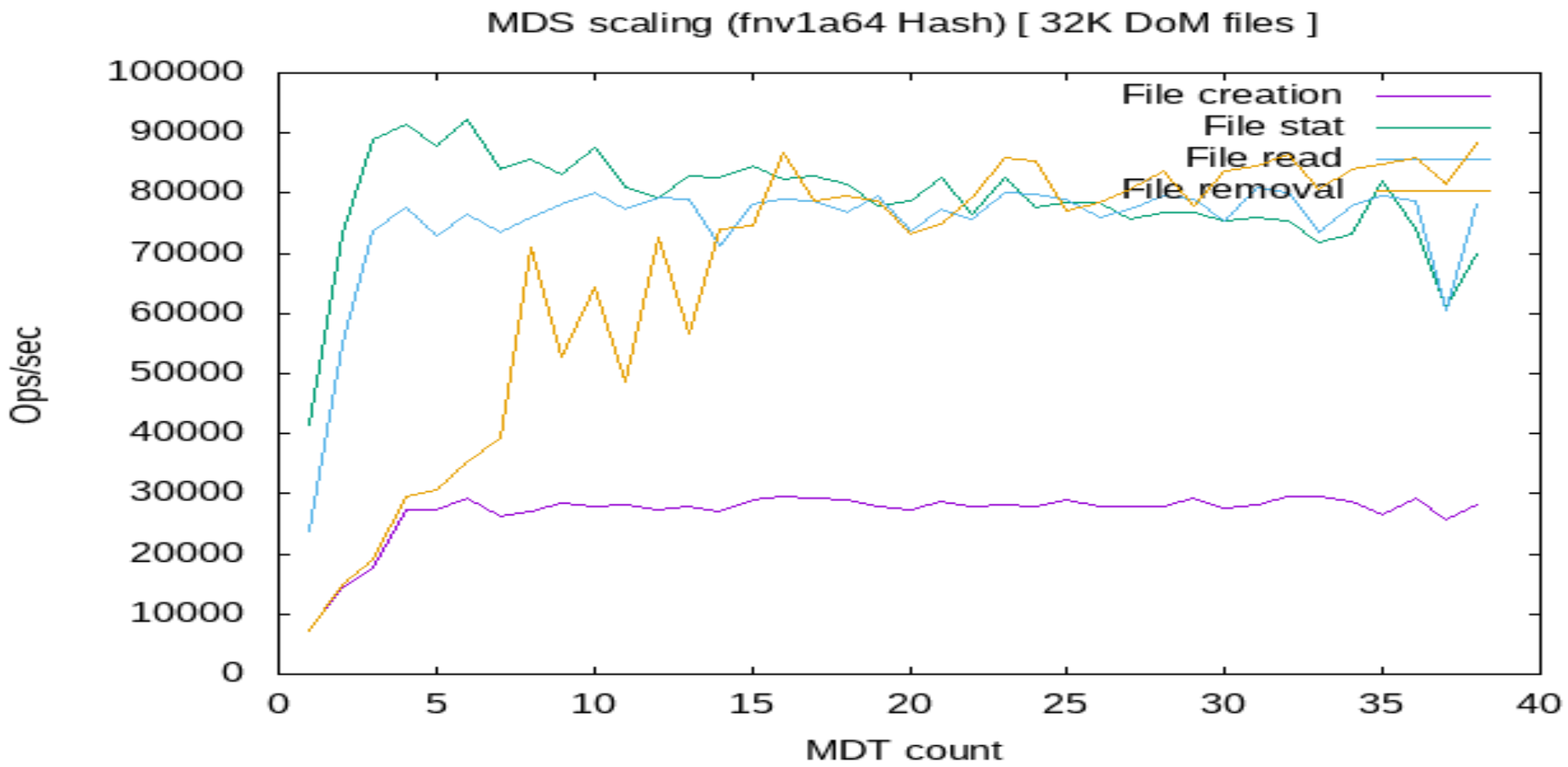
File performance for zero size files



File performance for 32k files



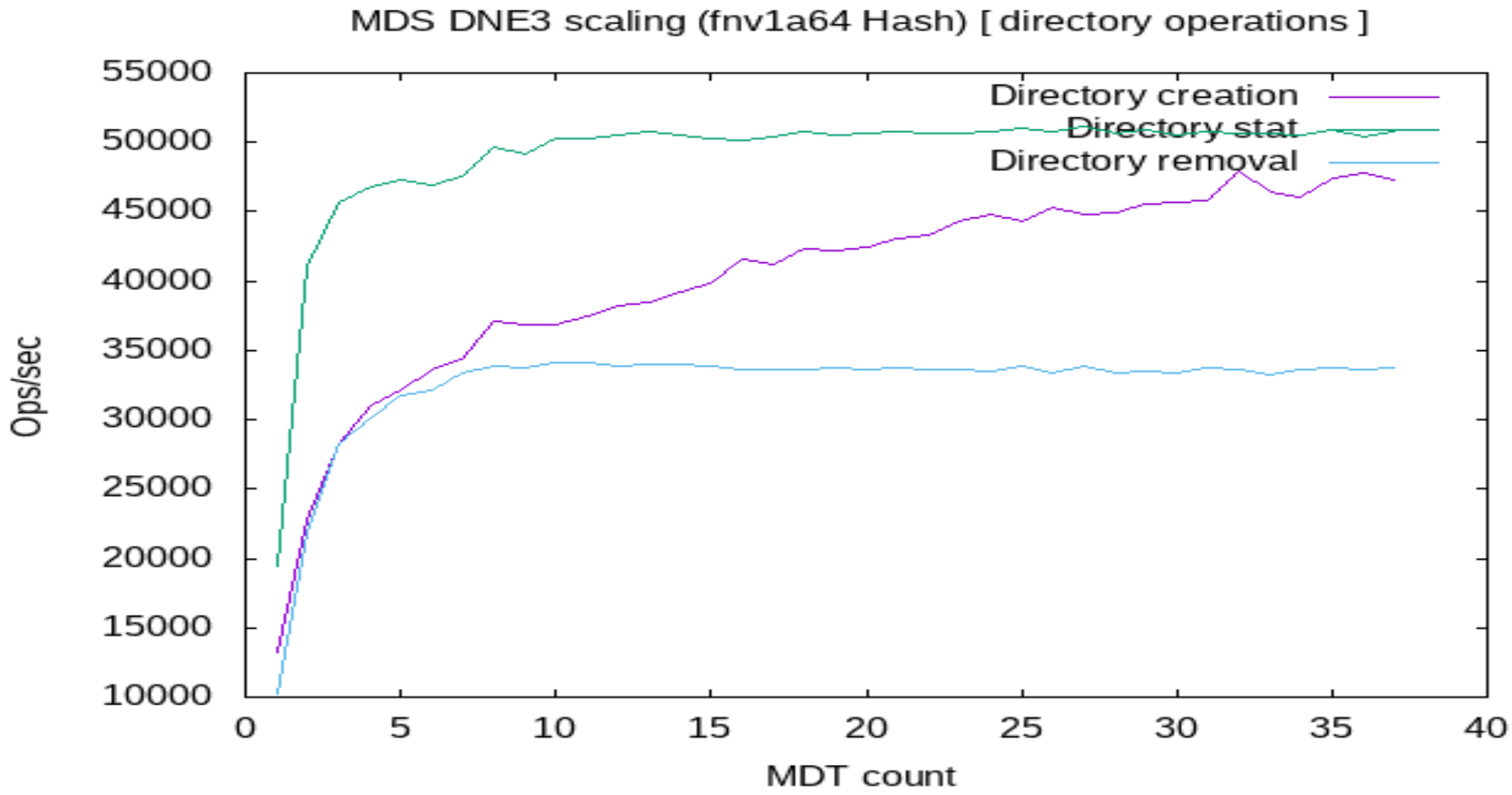
File performance for 32k DoM



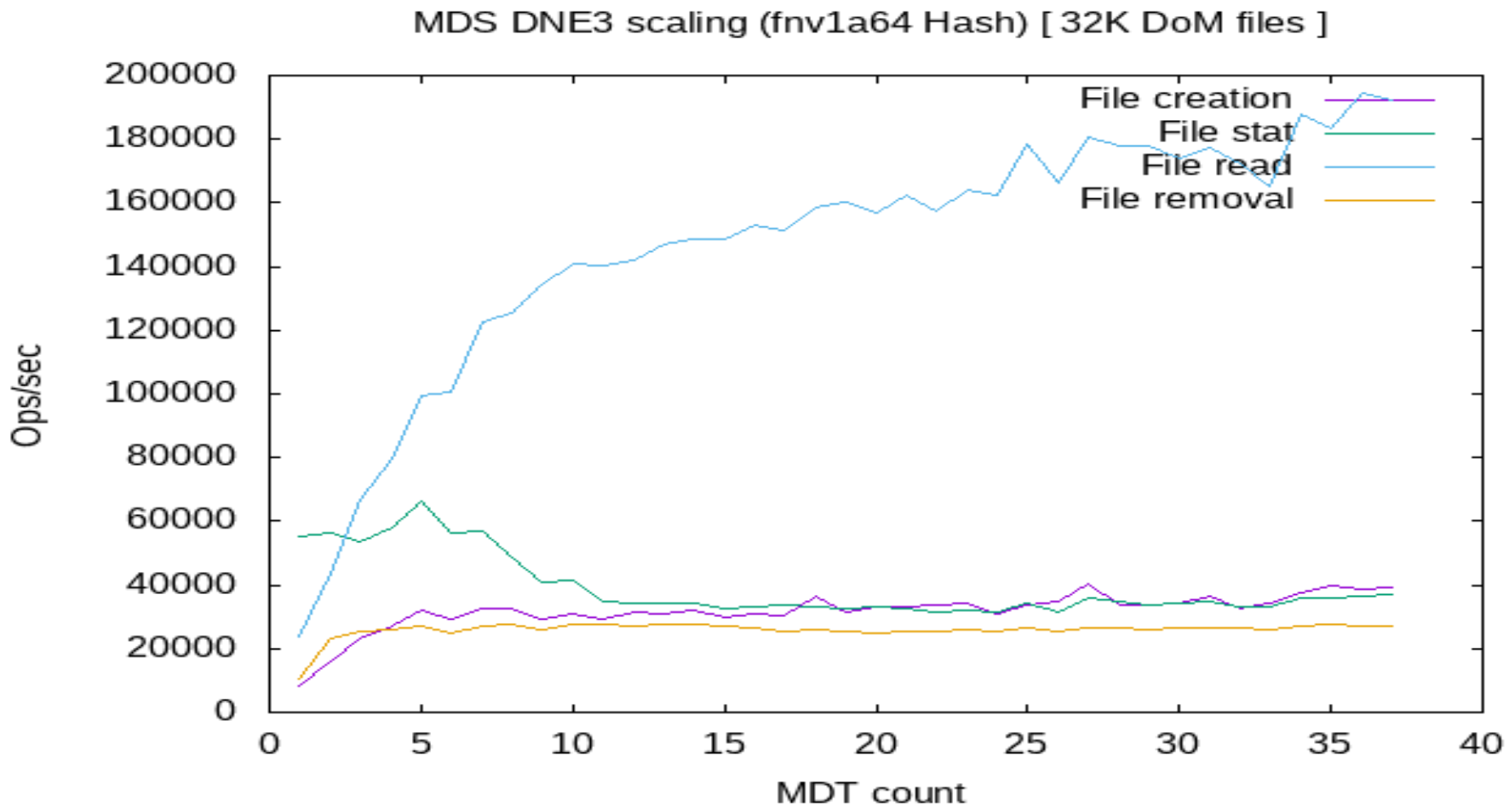
Giving DNE3 a test run

- DNE3 is new auto striping across MDTs feature
  - Files or directories get moved to other MDTs when limit is reached.
    - Ease of administration.
  - Ticket LU-11025. Landed to 2.14
- MDS setup for our testing.
  - lctl set_param mdt.*.enable_dir_auto_split=1
  - lctl set param mdt.*.dir_split_delta=1
  - lctl set param mdt.*.dir_split_count=15000
- Same mdtest script as before using shared directories.

DNE3 directory operations results



DNE3 file operations performance



Conclusions

- DNE 2 Directory operations
  - Stats scale well
  - Directory creation putters out at 16 but continues to scale
  - Directory removal flat lines after 16 MDTs
- DNE2 File operations
  - All setups (0 size, DoM, OST based) behave the same. What is best DoM size?
  - File removal flat lines at 16 MDTS
  - The rest flat lines at 4 MDTs
- DNE3 (Time limit only allowed DoM 32K file + directory testing)
  - Directory operations
    - Log raise to 4 MDT and flattened for removal and stats. Creates did have an improvement at a slower rate
    - Performance worst than DNE2 by a large margin.
  - Better file read but everything else performs worst then DNE2 setup. Stats degrade.

Acknowledgements

This work was performed under the auspices of the U.S. DOE by Oak Ridge Leadership Computing Facility at ORNL under contract DE-AC05-00OR22725.