Cluster I/O Performance On Large Data Files

UC San Diego

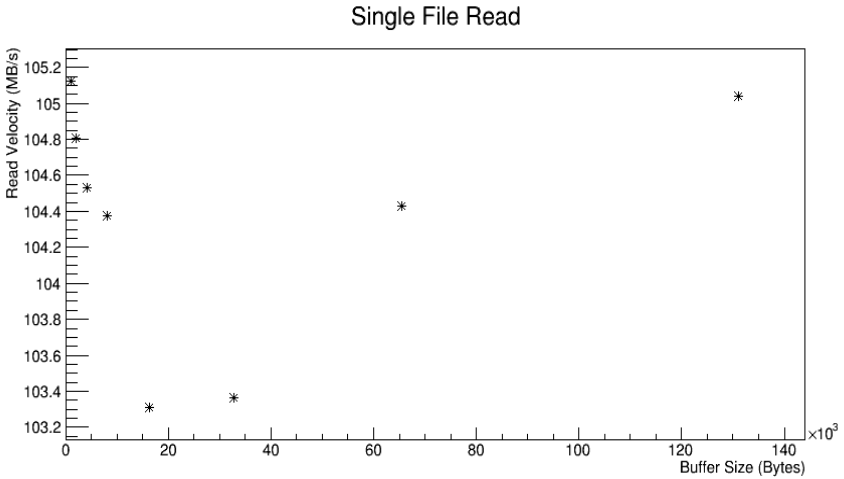
Specifications

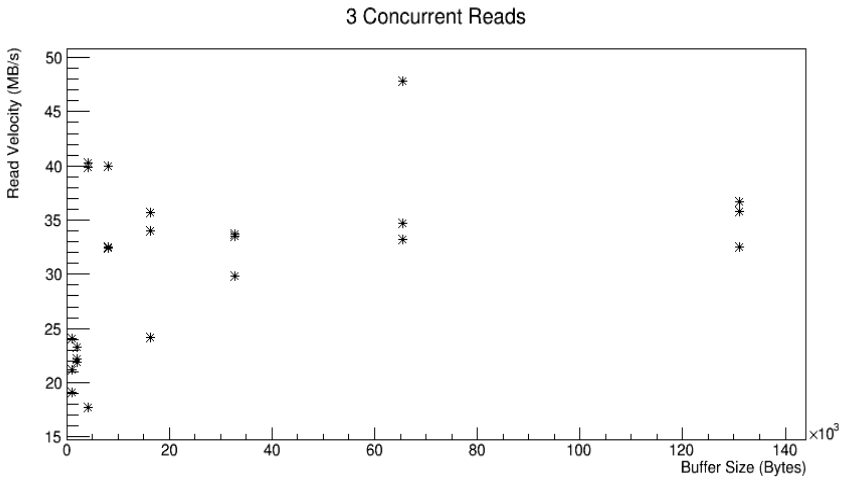
- Using CMS3 files trimmed to 2GB
- Reads compared:
  1) local disk
  2) HDFS through FUSE
  3) HDFS through C API
  4) ram cache

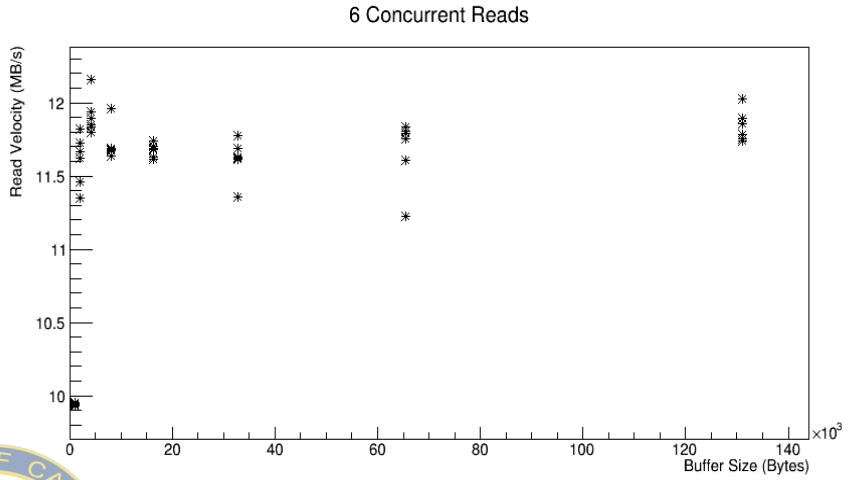
UC San Diego

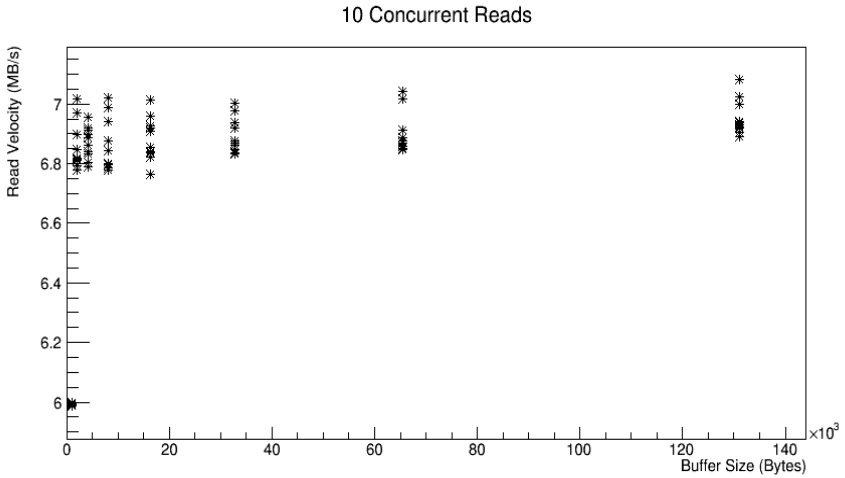
Local Disk Performance

Read From Local Disk, Cache Dropped



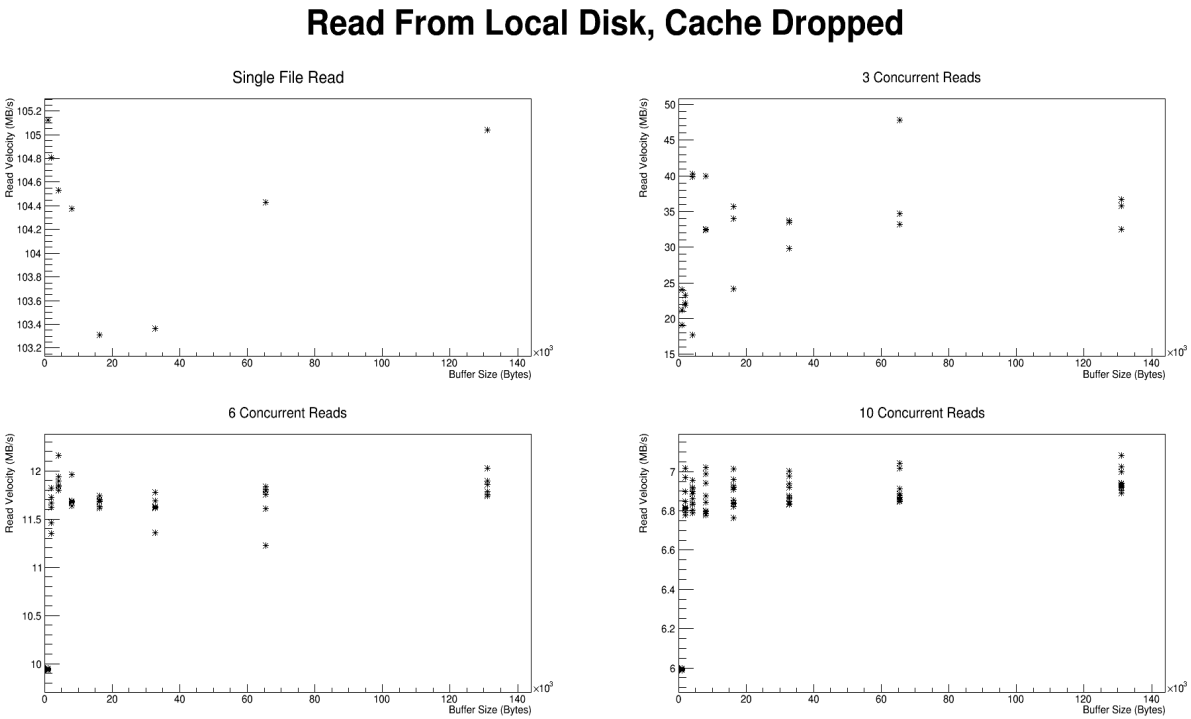






UC San Diego

Local Disk Performance



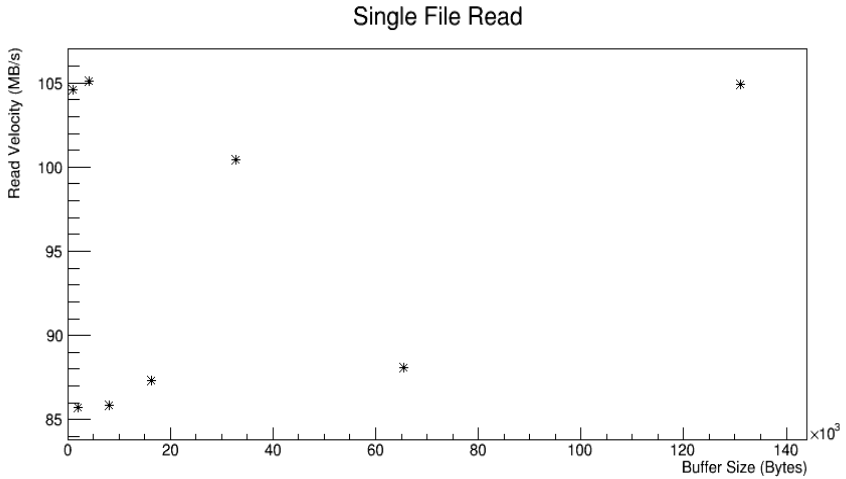
1. No files being written to disk
2. Ram cache dropped after each read
3. Tested on UAF-4

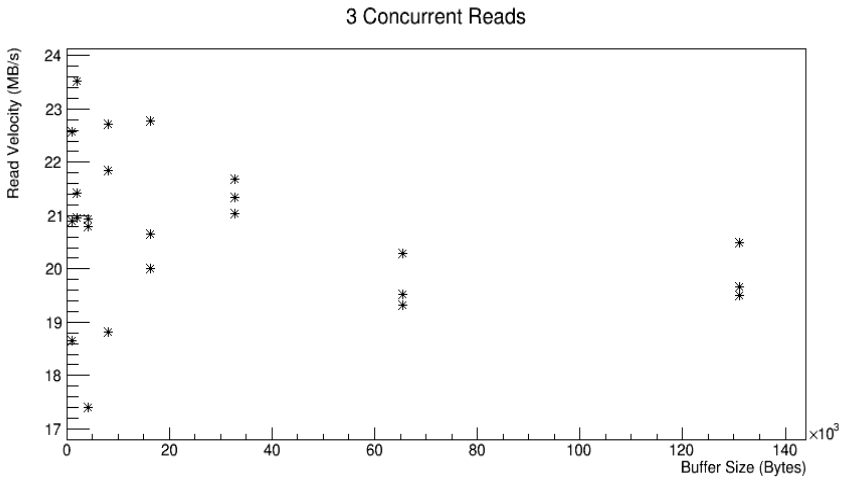
Max Read Velocity ~ 100 MB/s from 1 or 3 files with large buffer size.
30% performance hit when reading from 10 files.

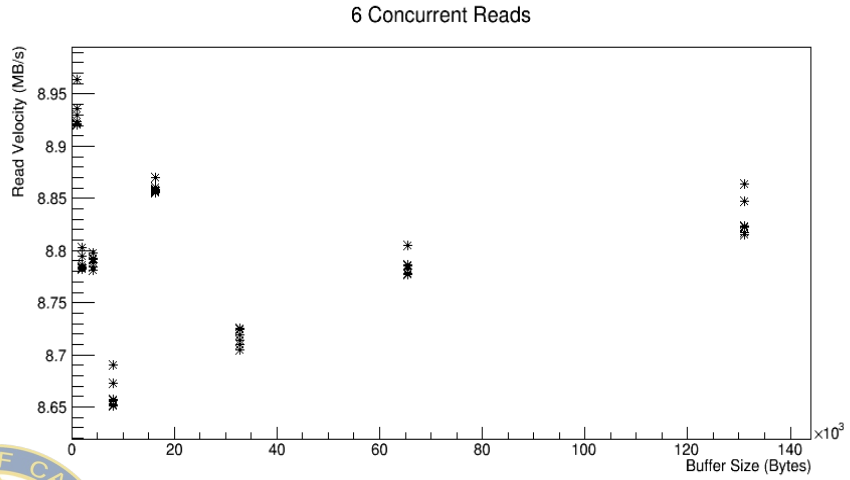
UC San Diego

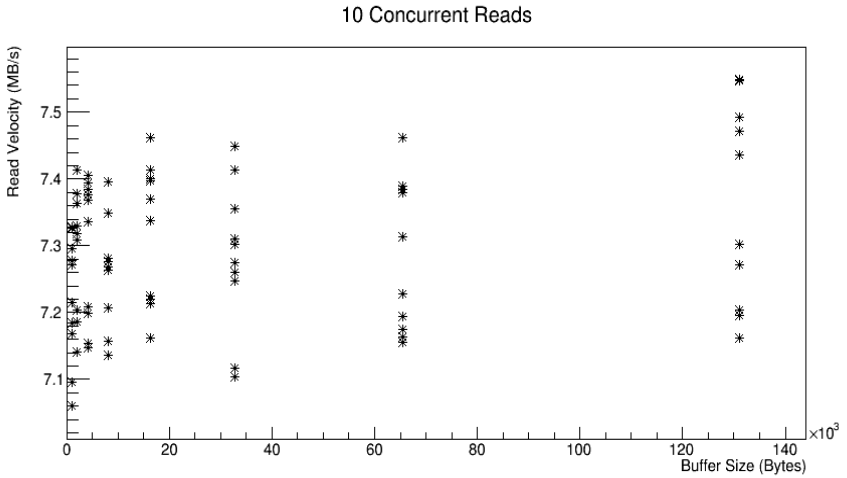
Local Disk Performance (with write)

Read Local Disk With ~10MBps Write, Cache Dropped



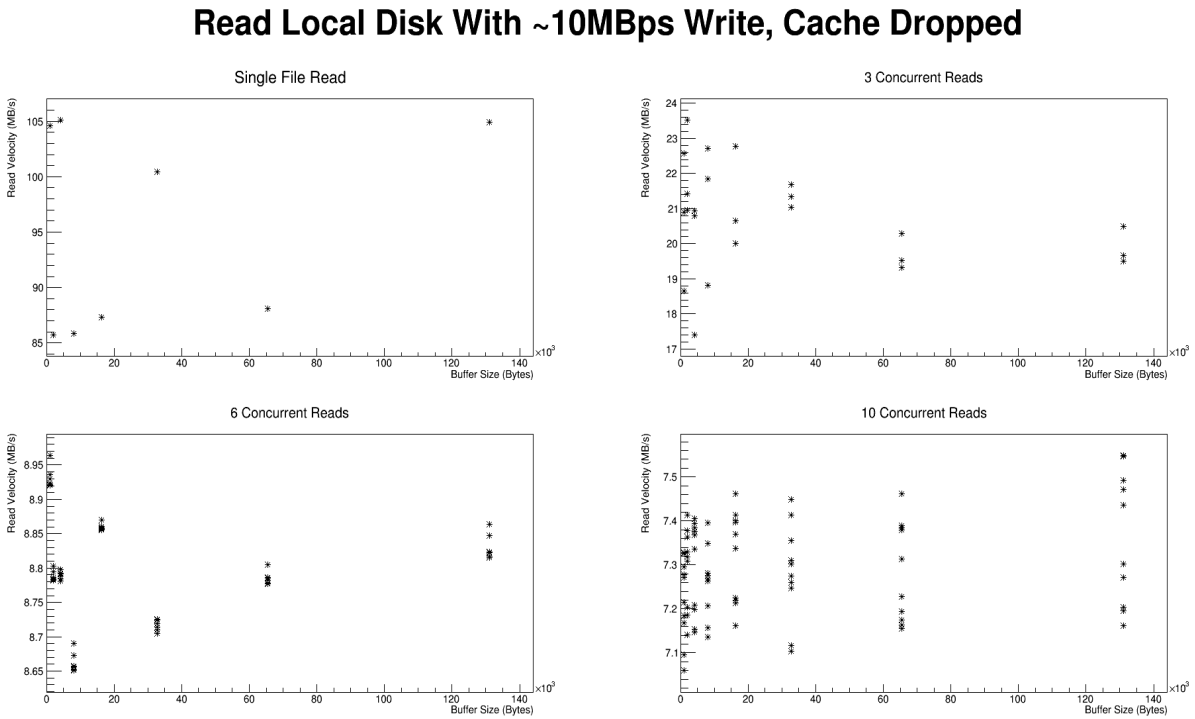






UC San Diego

Local Disk Performance (with write)



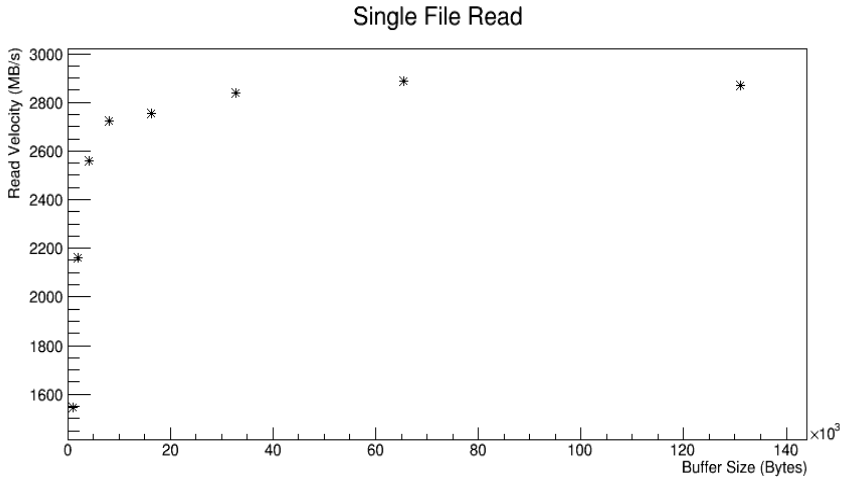
1. No files being written to disk
2. Ram cache dropped after each read
3. Continuously writing ~10MB/s to disk
4. Tested on UAF-4

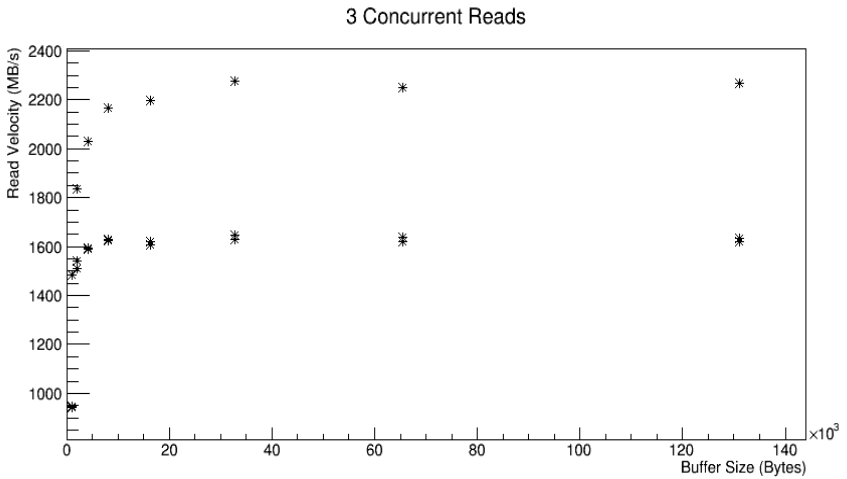
Max Read Velocity ~ 95 MB/s from 1 file, buffer size independent?
30% performance hit when reading from 10 files.

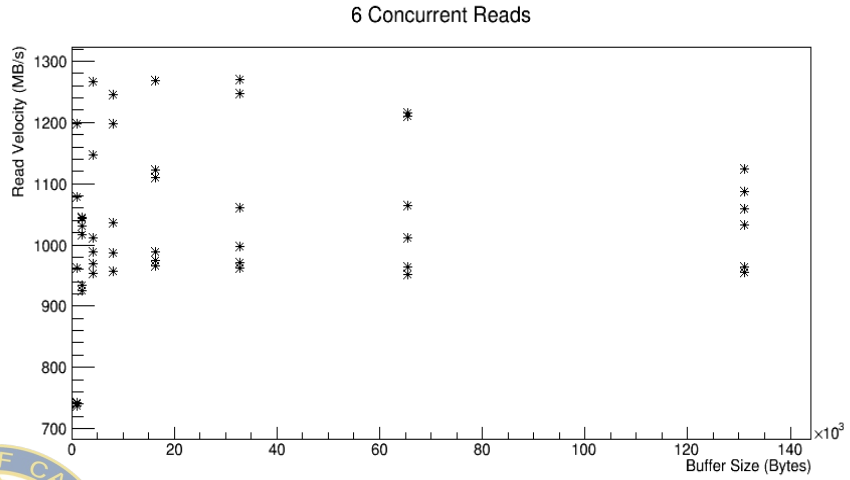
UC San Diego

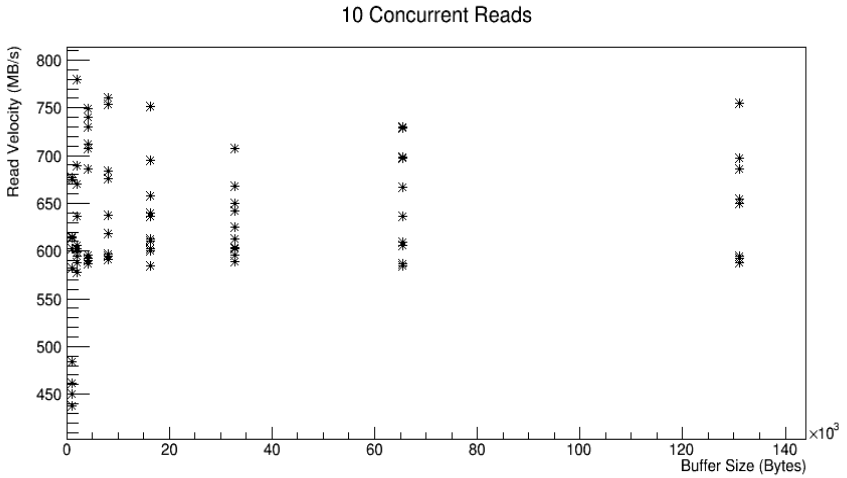
Ram Cache Performance

Read From Ram Cache



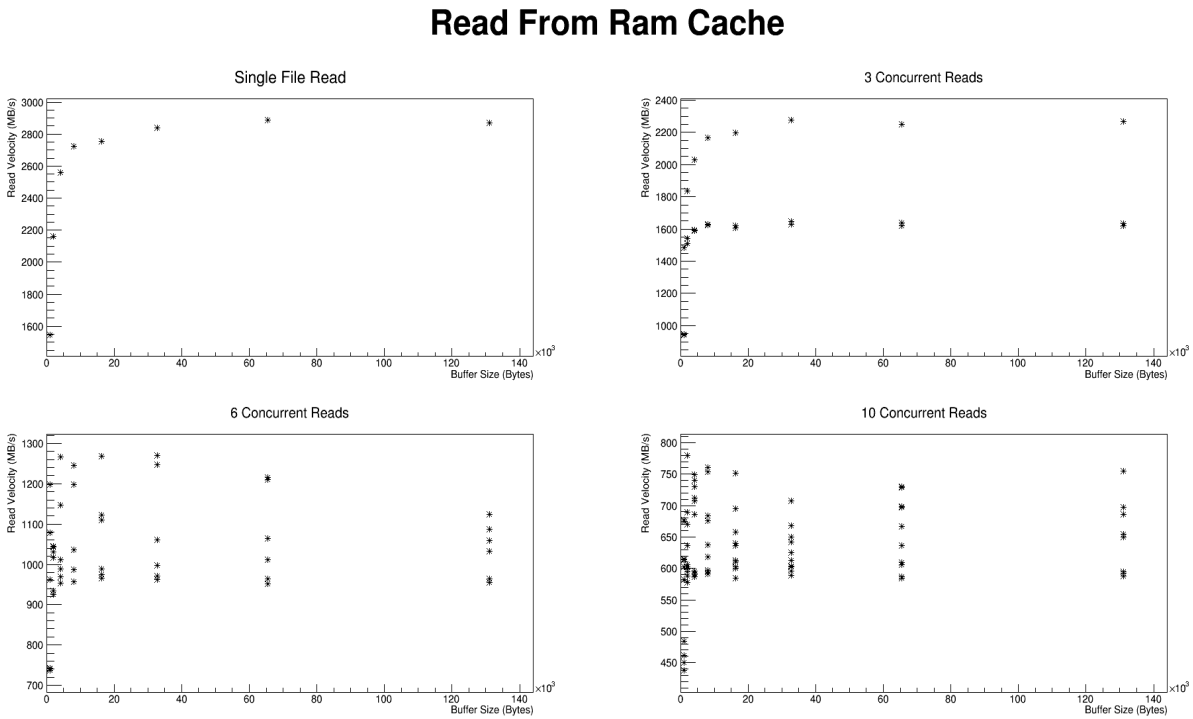






UC San Diego

Ram Cache Performance



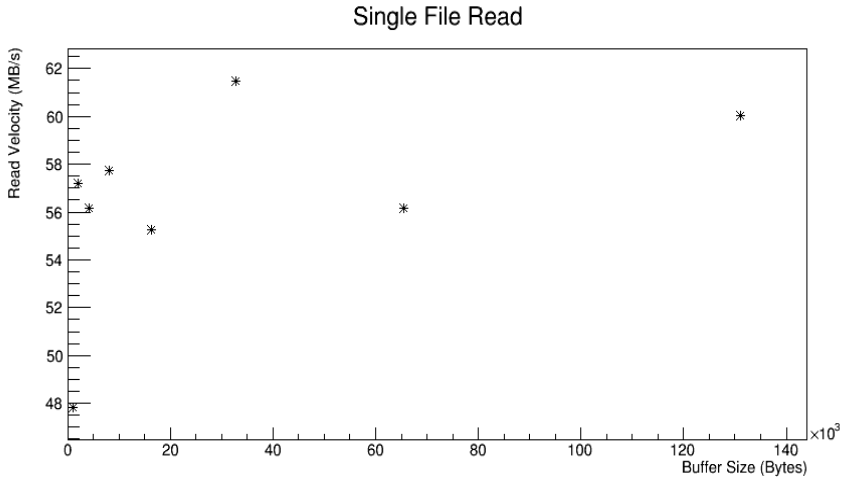
1. No files being written to disk
2. Tested on UAF-4

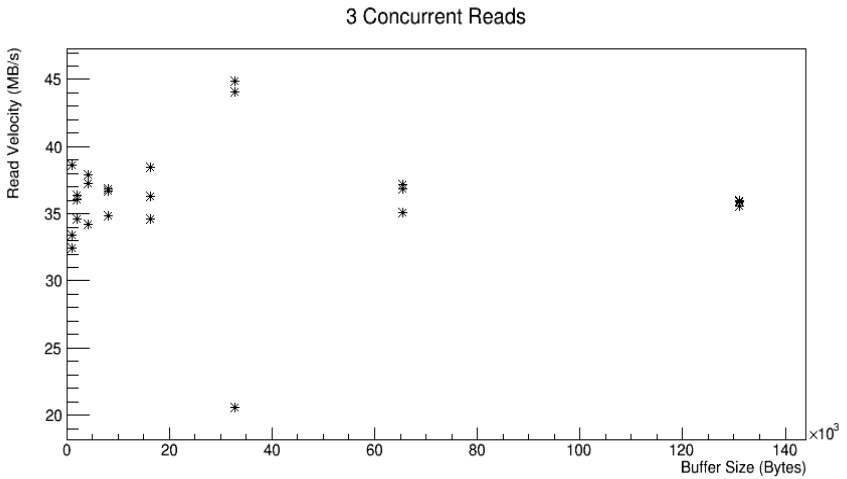
Max Read Velocity ~ 6 GB/s from 3, 6, or 10 files independent of buffer size. 50% performance hit when reading from 1 file.

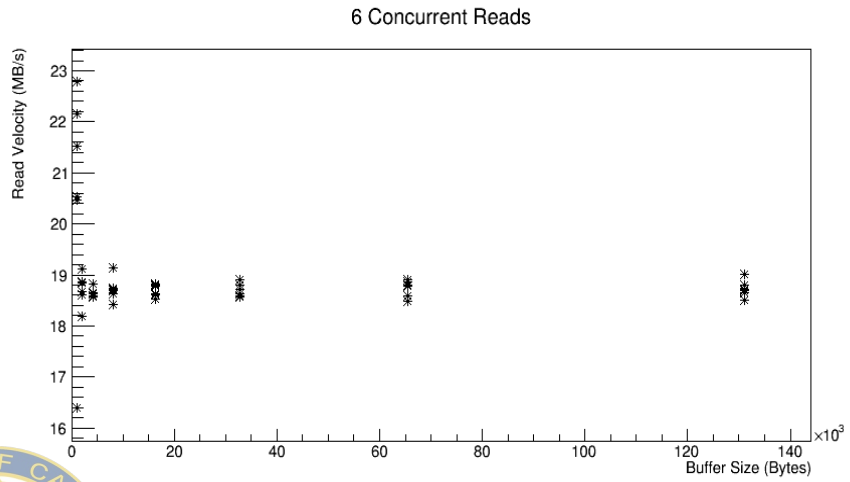
UC San Diego

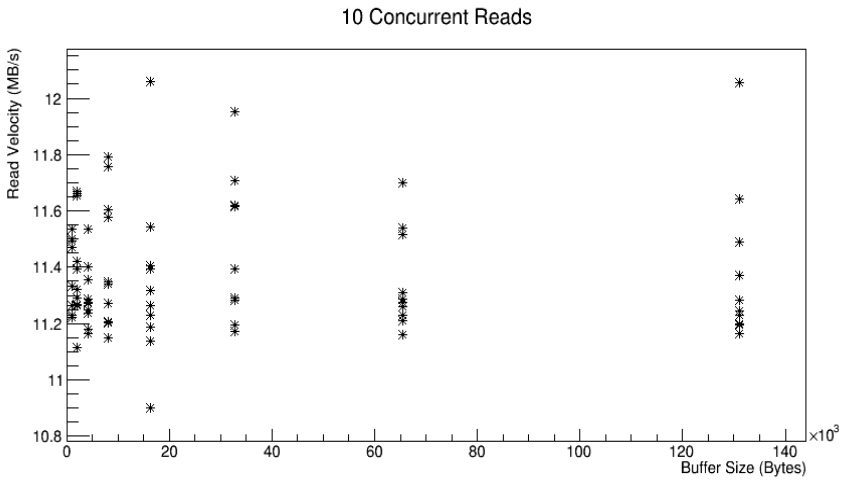
HDFS Performance (FUSE)

HDFS Through FUSE



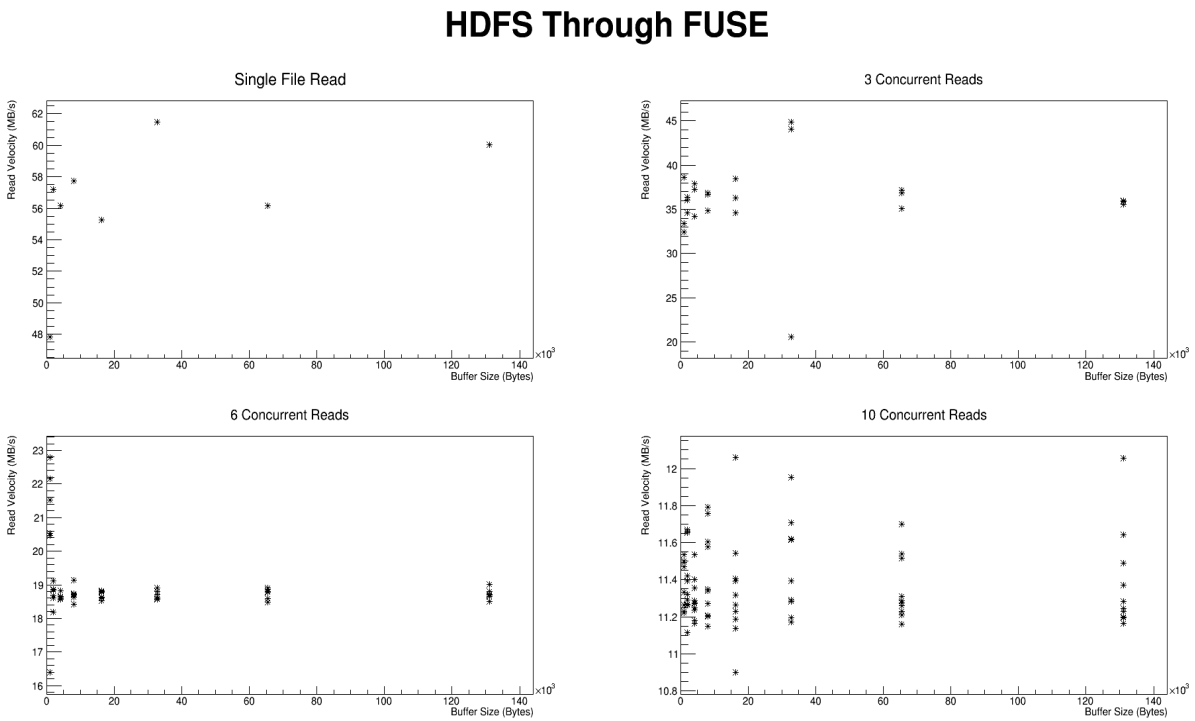






UC San Diego

HDFS Performance (FUSE)



1. No files being written to disk
2. No attempt to drop ram cache
3. HDFS mounted through FUSE
4. Tested on UAF-4

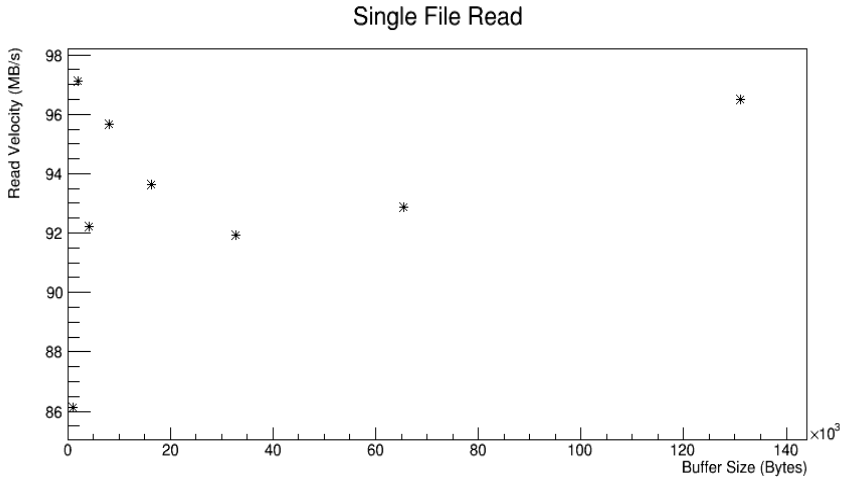
Reproduces conditions for current babymaking procedure.

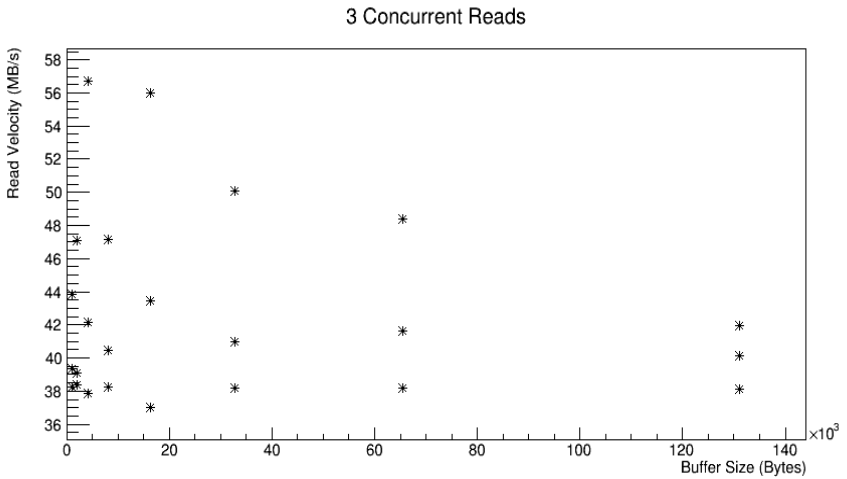
Max Read Velocity ~ 115 MB/s from 6 to 10 files with large buffer size.
50% performance hit when reading from 1 file.

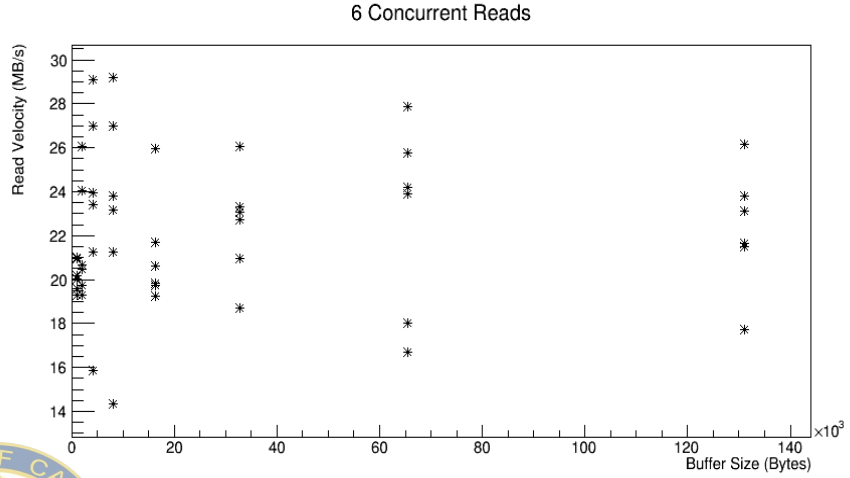
UC San Diego

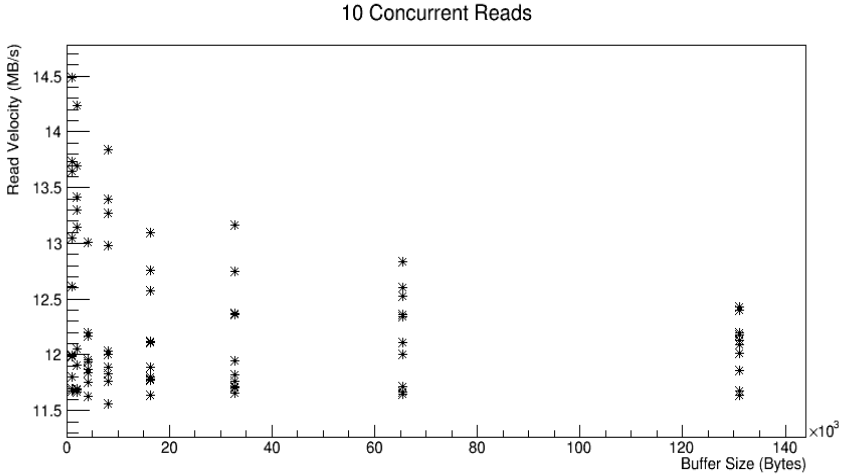
HDFS Performance (C API)

HDFS C API



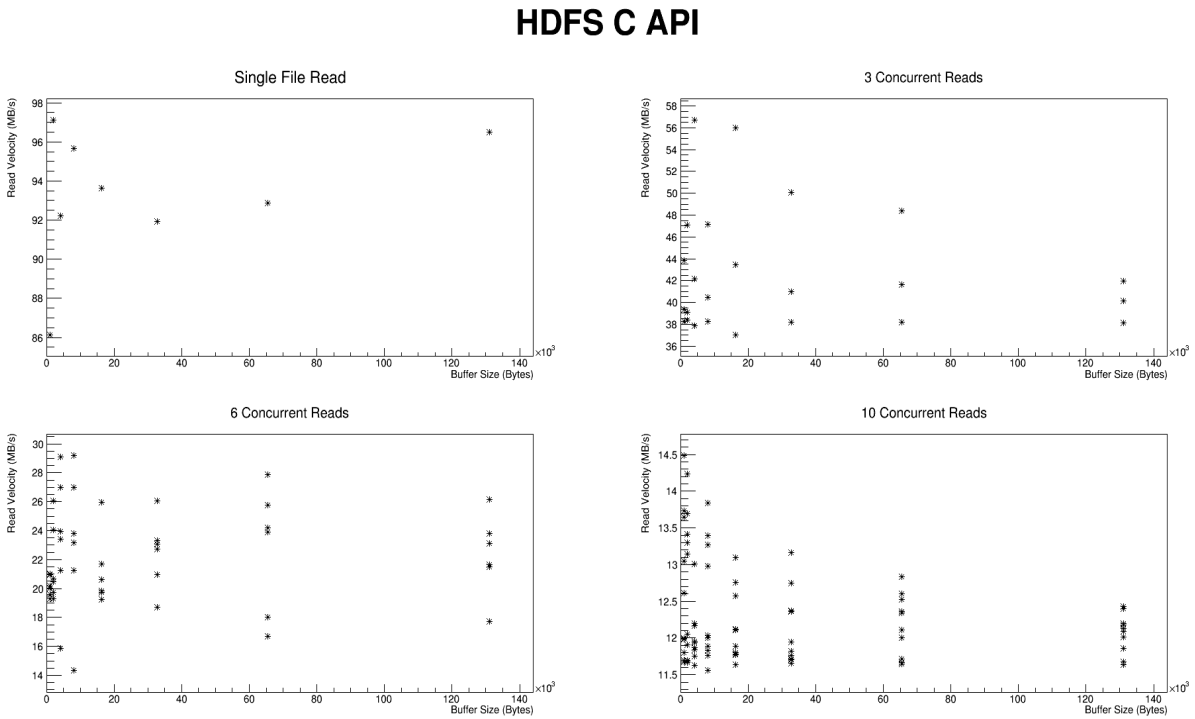






UC San Diego

HDFS Performance (C API)



1. No files being written to disk
2. No attempt to drop ram cache
3. HDFS accessed through C API
4. Tested on UAF-4

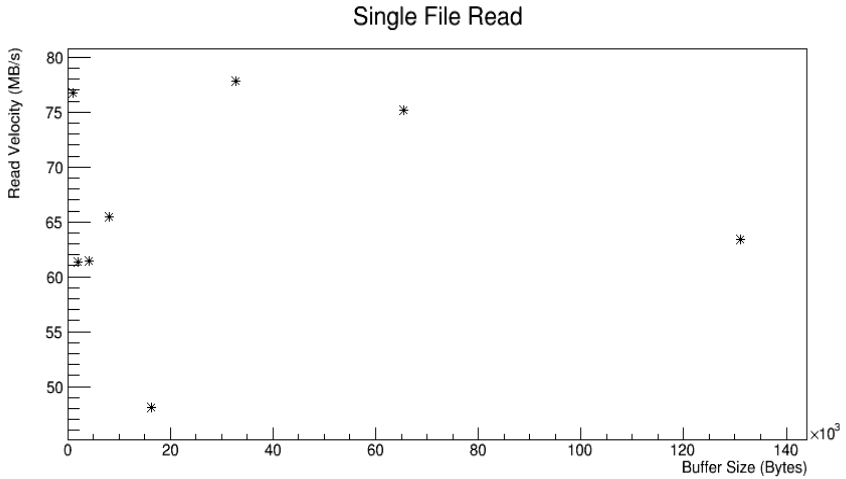
Reproduces conditions for current babymaking procedure.

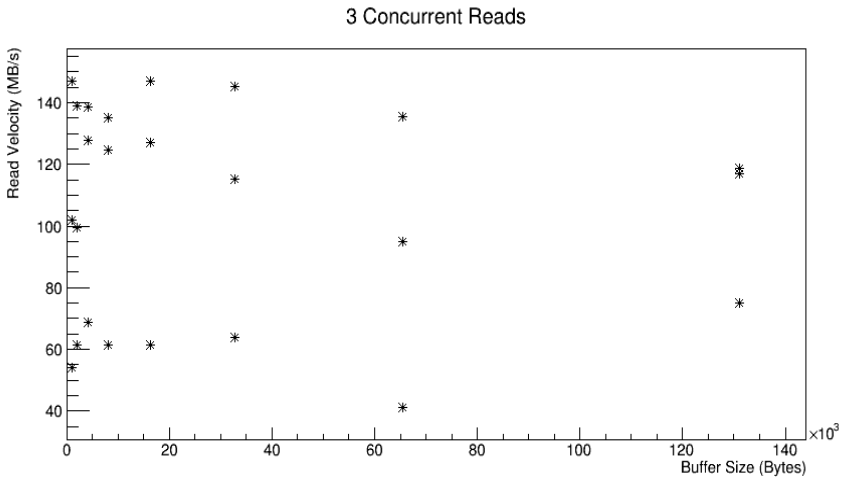
Max Read Velocity ~ 130 MB/s from 6 to 10 files with small buffer size.
25% performance hit when reading from 1 file.

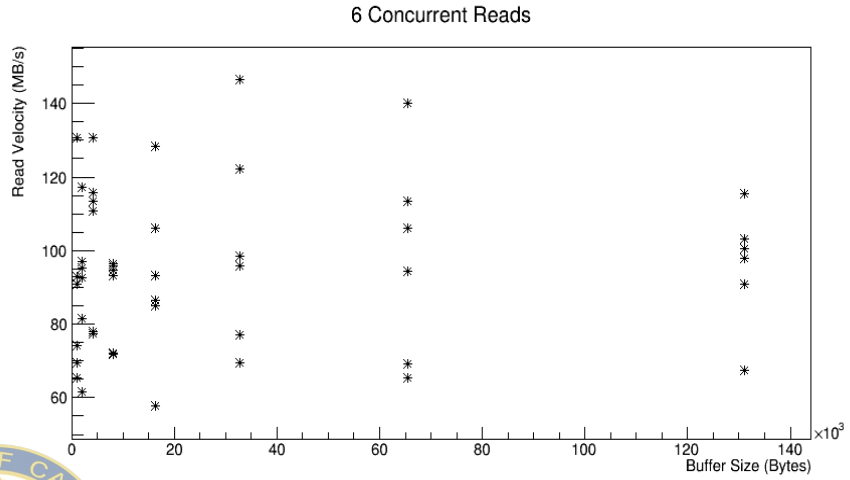
UC San Diego

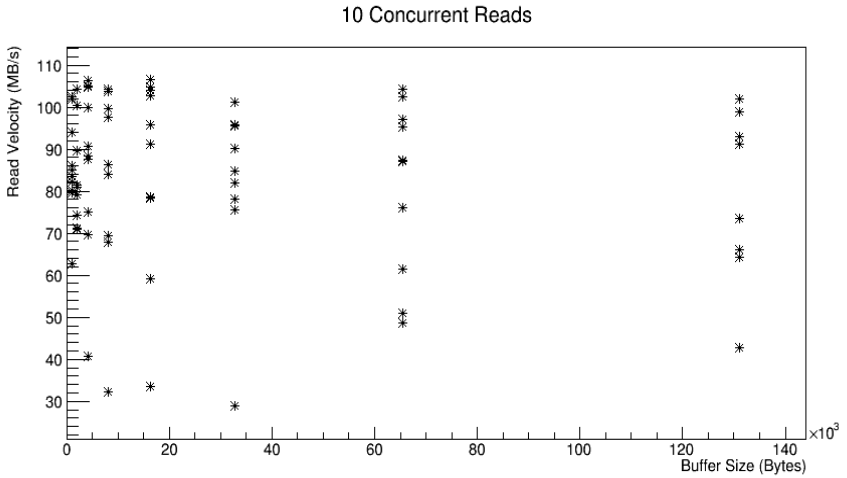
Local Disk Per File

Disk Per File, Cache Dropped, cabinet-8-8-0



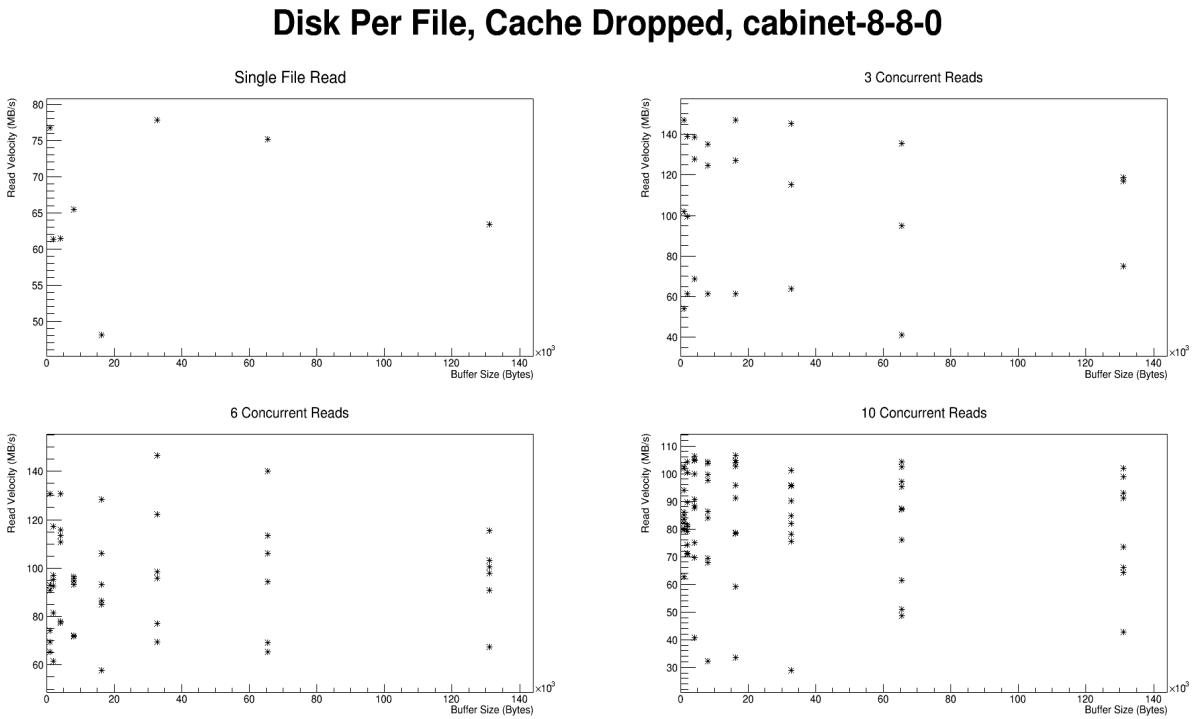






UC San Diego

Local Disk Per File



1. No files being written to disk
2. Cache cleared
3. Tested on cabinet-8-8-0 under full load (~24 jobs)

Reproduces conditions for proposed babymaking procedure.

Max Read Velocity ~ 800 MB/s from 10 files, buffer size dependence unclear.

~5x niave improvement over HDFS C API best case(no cache clearing) vs LDPF worst case on 10 file reads.

UC San Diego

Next Steps

1)Take HDFS reads when the cache has been dropped?
2)Convert babymaking code to read from local disk

UC San Diego

Conclusions

1) HDFS read is network limited near 120 MB/s which is approximately equal to the performance of a single drive.
2) By storing files in a careful ordered maner across disks, file I/O could be niavely sped up by at least half an order of magnitude.

UC San Diego