Explanations of data movement and system throughput improvement

Influence of data movement on throughput improvement

Currently we haven't considered the influence of data movement on throughput improvement because it seems that the data movement won't have a big impact on system throughput. With a reasonable internal network bandwidth, data movement introduced by our algorithm can be easily handled. I have found same assumption in similar paper [1].

I want to illustrate this problem through a simple example:

We assume SSD OSD only account for small part of the storage system. So the data movement impact will mainly focus on SSD OSDs.

We assume the following parameters:

1. SSD storage on each OSD: 500G with 300G utilized.
2. Amount of data need to be moved: 300G * 20% = 60G
3. Internal network bandwidth for each OSD: 10M/s
4. Update Period for each OSD: 2 Weeks

According to above parameters, we can see that the total available bandwidth of each OSD in one update period is

10M/s * 60s * 60m * 24h * 7w * 2w/update = 12096G/update

So the data movement only occupies 60 / 12096 = 0.00496 of total bandwidth.

Even if we move all the data on SSD, the data movement only occupies 0.0248 of total bandwidth.

Also there is no strict deadline for the data movement, so the process can move the data whenever the internal network is lightly loaded.

NOTE: Since we lacks knowledge of real HPC deployments, so red numbers are based on my previous knowledge of commercial deployments and some conjecture. But 20% data movement in 2 weeks is a pretty aggressive conjecture.

Results on data movement vs throughput improvement

I want to mention our test deployment and method to calculate system throughput first:

1. Test environment:
    a. 640 OSDs, 10% of OSDs are equipped with SSD storage
    b. 10,000 data objects, with 100,000 times access (read for read-only test, write for write-only test)
    c. We currently have not considered the constraints of network bandwidth
2. Method to calculate system throughput:

a. Read-only test: we assume the throughput when read a data object is the AVERAGE of all the replicas' throughput.

b. Write-only test: we assume the throughput when write a data object is the BEST of all the replicas' throughput, that is, if there is any replica on SSD OSD, the throughput will be the same as the throughput of SSD OSD.

Test results:

1. Read only test

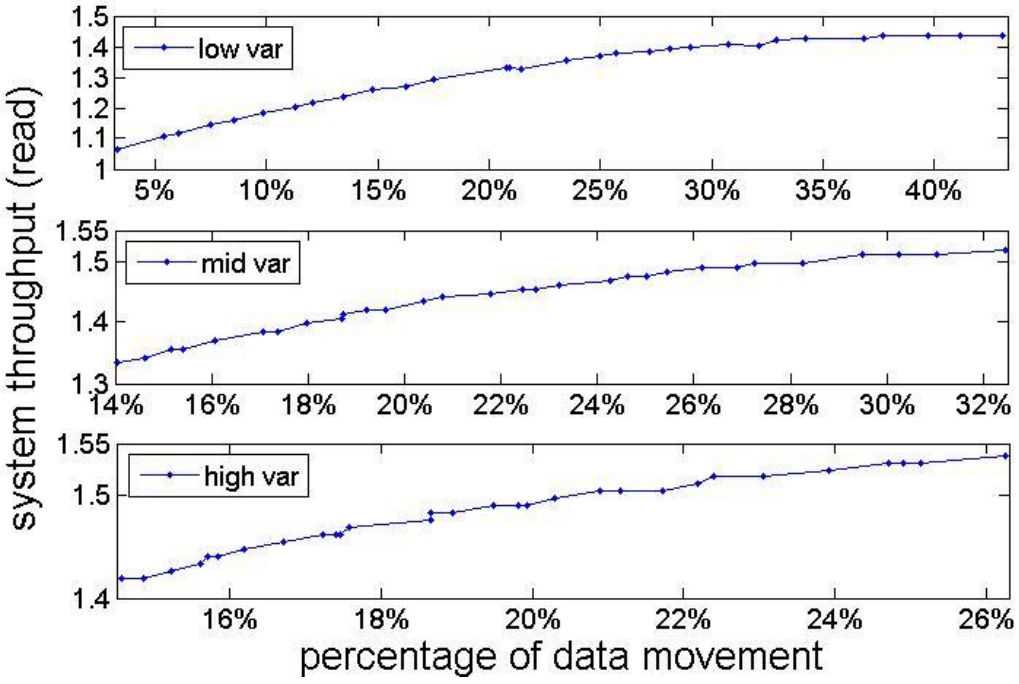

Figure 1 Results for read only test with low, mid, high variance of data objects' access (higher variance means some objects have been accessed much more times than others)

2. Write only test

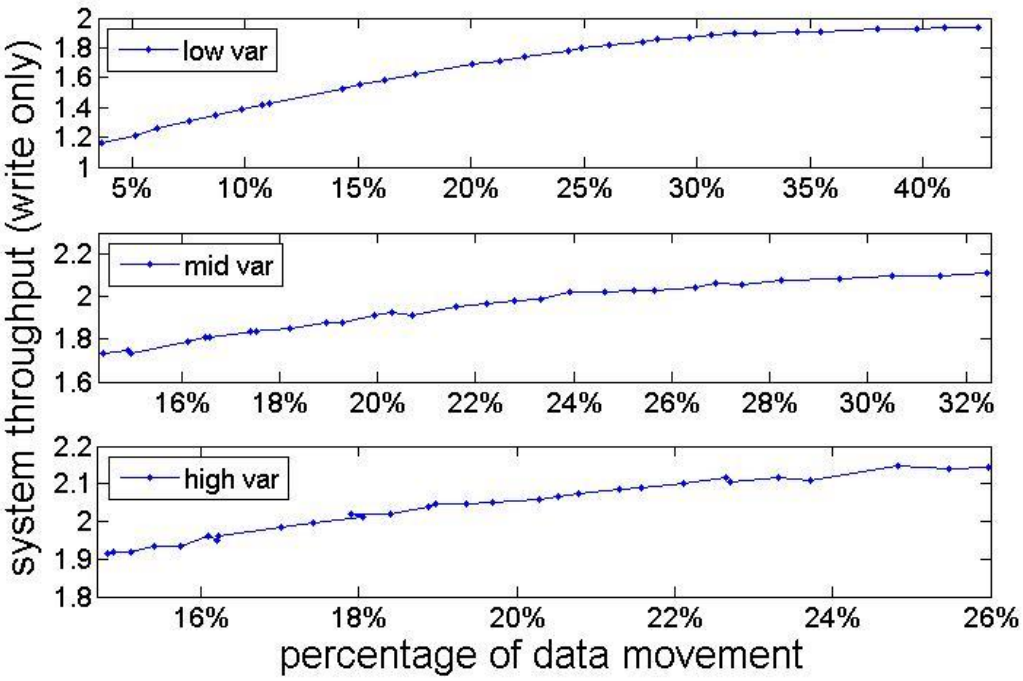

Figure 2 Results for write only test with low, mid, high variance of data objects' access (higher variance means some objects have been accessed much more times than others)

References

[1] Feng Chen, "Hystor: Making the Best Use of Solid State Drives in High Performance Storage Systems"