TR5: Scalability Study of Ceph

Xing Lin
xinglin@cs.utah.edu
University of Utah

03/15/2013

1 Introduction

This document presents the results about the performance scalability of Ceph, running with AT&T virtual machines.

2 Experiment Setup

I did my experiments with 4 virtual machines, provided by a teammate from AT&T. Three virtual machines were used as data nodes and the forth was used as the client machine for benchmarking. The configuration of each VM is presented in Table 1.

CPU	8×2.7 GHz cores
Memory	32 GB
Virtual Disk	343 GB

Table 1: Virtual machine configuration

The configuration for the Ceph cluster is presented in Table 2. The version for each program used in this measurement is presented in Table 3.

Attribute	Value
Ceph osd pool size (num of replicas)	3
Placement group number	100
Journal	10 GB ramdisk

Table 2: Ceph configurations

tool	version
OS	Ubuntu12.04
Ceph	Argoguant.0.56.3
fio	2.0.14

Table 3: Tools

3 Results

3.1 Raw Disk Throughput

I did a simply measurement with dd, to get the raw sequential read/write bandwidths for these virtual disks. The bandwidths for each virtual machine are presented in Table 4.

VM	4 MB Read	4 MB Write
vm1		
vm2		
vm3		

Table 4: Raw bandwidths of the virtual disk for each virtual machine

For scalability, I mainly looked into three aspects: IO depth, block size, and parallel reads/writes to many Rados block devices.

3.2 IO Depth

I first looked into how iodepth affects the throughput of a single sequential workload. The parameters are presented in Table 5 and the results are presented in Figure 2 and 1 .

block size	duration	directio	ioengine
4 MB/4 KB	60 s	1	libaio

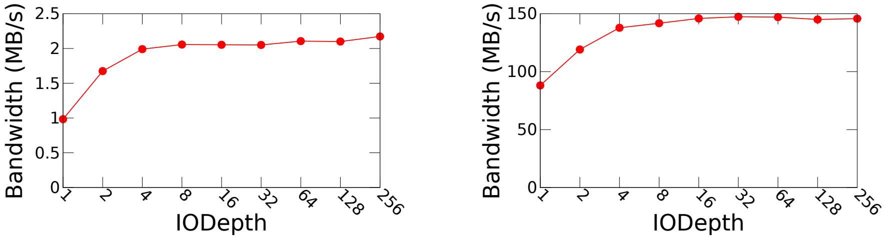
Table 5: Sequential workload parameters

3.3 Block Size

In this measurement, I fixed iodepth=1 and varied the block size. The parameters of workloads used in this set of experiments are presented in the Table 6 and results are presented in Figure 3.

duration	directio	ioengine	iodepth
60 s	1	libaio	1

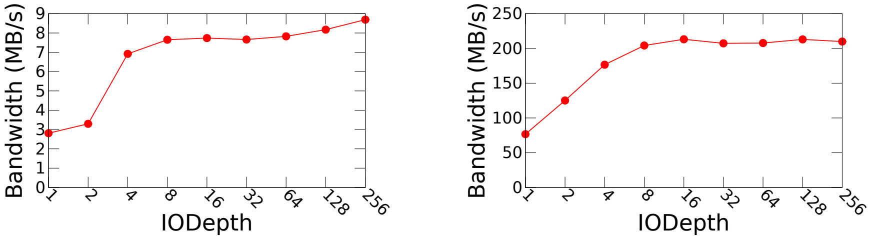
Table 6: Sequential workload parameters



(a) blocksize=4k

(b) blocksize=4m

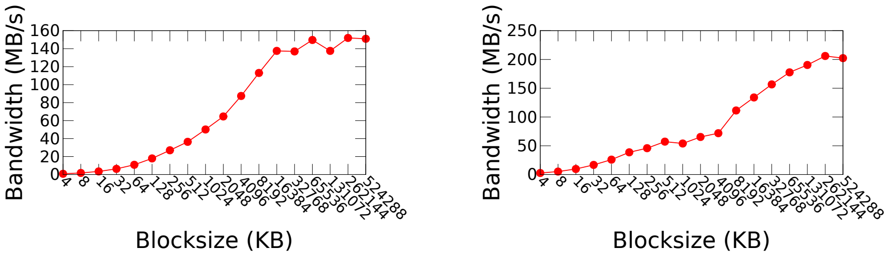
Figure 1: Average bandwiths of a SW workload, varying IO depth



(a) blocksize=4k

(b) blocksize=4m

Figure 2: Average bandwiths of a SR workload, varying IO depth



(a) Sequential Write

(b) Sequential Read

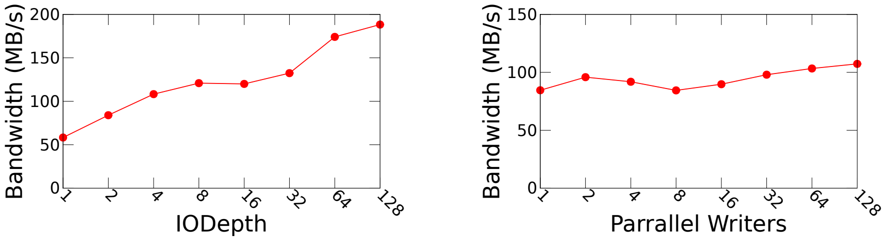
Figure 3: Average throughputs of a sequential workload, varying the block size

3.4 Parallel Writers/Readers

In this set of experiments, iodepth was fixed at 1 and the number of Rados block devices (RBD) that were concurrently read/written was varied. The parameters about the workloads are presents in Table 7 and the results are presented in Figure 4.

block size	duration	directio	ioengine	iodepth
4 MB	60 s	1	libaio	1

Table 7: Sequential workload parameters



(a) Read

(b) Write

Figure 4: Aggregated throughputs of a sequential read/write workload, when the number of parallel readers/writers is varied.

4 Discussions

What we can see from Figure 1 and 2 is that bandwidths increase as iodepth is increased.

Figure 3 shows that the block size affects the bandwidth significantly. Ceph can **not** provide a high bandwidth for small block size workloads.

Figure 4(a) shows that as we increases the number of concurrent RBDs to read from, the aggregated bandwidth increases. That illustrates that Ceph is scalable for serving parallel reads. However, for parallel writes, the aggregated bandwidth does not change that much, as shown in Figure 4(b). One aspect that likely contributes to this difference is that for each write, Ceph will replicate it to all three virtual disks we are using. For reads, Ceph only needs to read from one disk.

5 TODO

1. extended with random workloads