TR1: Stability Study of Ceph RBD

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

03/08/2013

1 Introduction

This document presents the results about the performance stability of Ceph Rados Block Device (RBD).

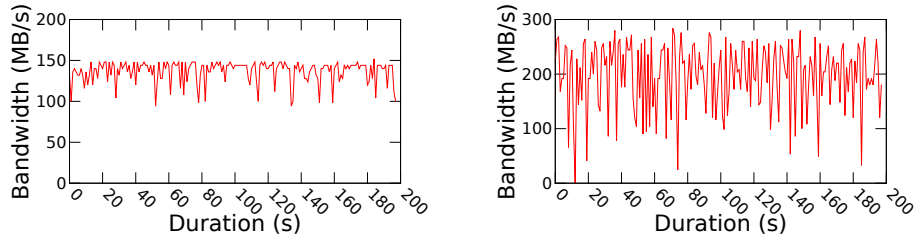
2 Experiment Setup

I did my experiments in the Emulab network testbed, hosted by the FLUX group at the University of Utah. I used 4 d820 nodes as the Ceph cluster and another d820 node as the client machine. Each d820 node has 6×600 GB SCSI disks so in total we have 4×6 (24) disks. We used xfs as the file system for each osd. The journal size is set to be 10 GB. To simulate the situation where the journal disk is hosted by a SSD, I used a tmpfs file as the journal disk (by specifying osd journal = /dev/shm/journal/$name-journal in /etc/ceph/ceph.conf). The fio tool is used to generate synthetic workloads. The configuration file for the ceph cluster, the ns file for the Emulab experiment and the job files for fio are all available in the TR1 dir in the src dir.

tool	version
ceph	argonaut.0.48.2
fio	2.0.14

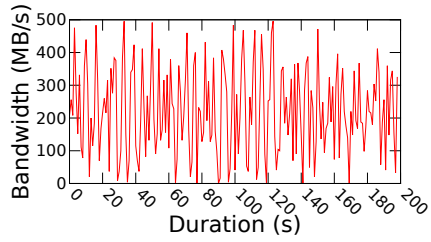
Table 1: Tools

We only consider sequential workloads. We used bandwidth as the throughput metric. For random workloads, it does not make sense to consider performance stability since IO requests themselves are random. IOPS is usually used to measure random workloads. However, for random workloads, the time to serve each request is intrinstically random: the location of each request is random.

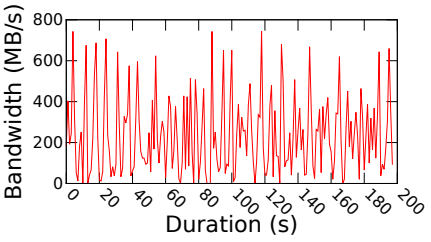


(a) IODepth=1

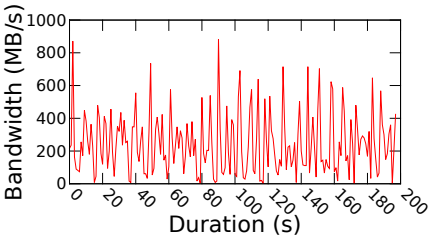
(b) IODepth=2



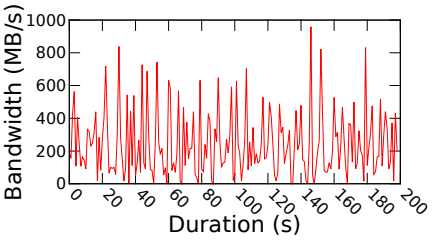
(c) IODepth=4



(d) IODepth=8



(e) IODepth=16



(f) IODepth=32

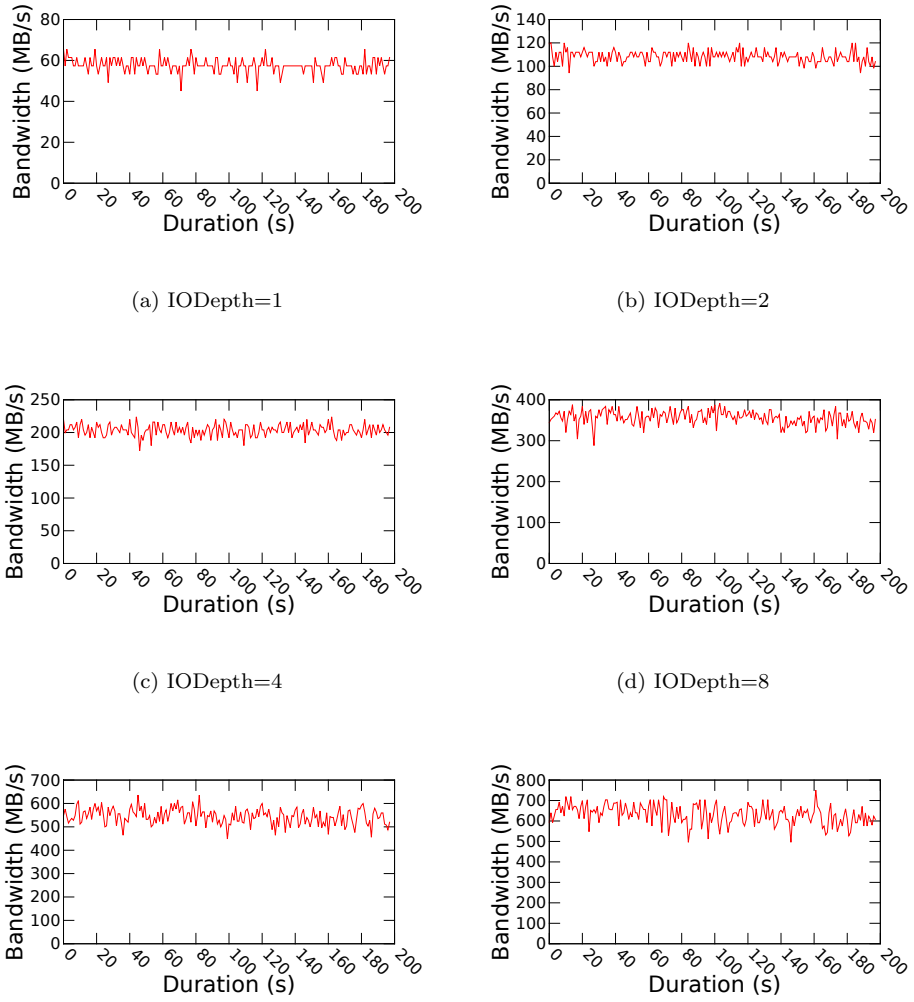
Figure 1: Instant Throughputs of a 4M Sequential Write Workload

3 Results

4 Discussions

[h] The throughput of a sequential workload changes dynamically. That seems to be fundamental. So, one question one might ask is what the criteria are to classify observed performances into stable or not?

Another observation is that sequential read workloads get more stable through-



(a) IODepth=1

(b) IODepth=2

(c) IODepth=4

(d) IODepth=8

(e) IODepth=16

(f) IODepth=32

Figure 2: Instant Throughputs of a 4M Sequential Read Workload

puts than sequential write workloads in Ceph. What are the possible causes for this?