Benchmark SeaweedFS as a GlusterFS replacement

We've recently conducted a small scale PoC for benchmarking SeaweedFS as a possible GlusterFS replacement.

The Data

Our data represents sensor sample from several sites over several time frames.

- 30 Sites
- 7000 Sensors per site
- 40 Time frames
- 1M samples per sensor per time frame, size of each sample is 1Byte

Total amount of raw data: 1M * 7000 * 40 * 30 = 8.4TB raw data

We consider each sensor data to be highly compressible for each timeframe and site.

GlusterFS Format

On the GlusterFS we store the sensor data in files for each timeframe and site, hence we have total of 1200 files (30Sites *40 Timeframes).

Each file contains 7000*1M Samples, ~7GB raw data, with additional metadata.

We compress each 1M data vector to a size of ~0.5MB resulting in files of ~3.7GB.

Total data on disk is ~4TB.

File system structure:

<gluster root>/<site>/<timeframe>/<data_file>

SeaweedFS Format

On the Seaweed we store one file per sensor per site, per timeframe. Hence the number of files is 7000 * 30 * 40 = 8,400,000 Files. The size of each file is ~0.5MB.

Total data on disk is ~4TB.

File system structure:

<seaweed root>/<site>/<timeframe>/<data_file>_<sensor_id>

The Setup

For the storage servers, we used 4 servers with the following HW spec:

- 16 CPU Cores
- 192GB RAM
- 25Gbps NIC
- 6 X 960GB SSD on RAID5

NOTE1: For the the SeaweedFS volume server we split the 6 X 960GB SSD on RAID5 drive into 3 partitions of roughly equal size, running 3 instances of SeaweedFS volume server.

NOTE2: We run all seaweed components (master, filer, volume, fuse mount) as docker containers.

NOTE3: We used leveldb as backend storage for the filer

NOTE4: We configured the weed volumes with -index=memory

The Benchmark

For the benchmark we run 300 concurrent clients. The client itself is described in the next section.

We measure the total client inner loop run time, and the end result of our benchmark is the P50 and P90 times, after the system run for a while during the steady state, where all the clients run enough cycles of the inner loop.

The Client

The client consists of outer and inner loop.

The client is the same for both file types and filesystems, and the differences are only related to the different file formats.

The Client Inner loop

The client process assigned a site and selects a subset of 500 sensors.

In a loop over the 40 timeframes, the client loads the data of the 500 sensors for the specific site and timeframe, decompress the data, and perform a dummy CPU intensive compute on it, for instance, a busy loop.

The runtime of this loop above is measured and reported.

The Client Outer loop

The outer loops simply selects a random site and runs the inner loop. This loop runs until the client stops.

The Results

Below is a short summary of the results we achieved during our benchmark,

results are seconds:

	GlusterFS	SeaweedFS
P50	1001	401
P90	1741	442
P100	2942	521

Our overall experience with SeaweedFS was good, we like it’s architecture. It performed very well in this POC benchmark, and we plan on moving to a larger scale POC.