Scaling parameters of the straw buckets in CRUSH

Kazuho Fujii

July 26, 2015

1 Background

Ceph is a remarkable distributed file system developed by Weil *et al.* It has a original algorithm named CRUSH, that decides which machine data is allocated to. Particularly, straw buckets in CRUSH is an amazing idea. Data is put into items in a bucket. Data named x is put into the item which maximizes $f_i h(x)$. $h(x)$ is the hash value of the data name; f_i is scaling parameter of the item. Scaling parameters adjust the number of data each item has to the weight of item w_i.

In the CRUSH algorithm paper, it is not specified how should we get scaling parameters. So, I calculated the optimum values of scaling parameters. I make a note about the scheme I got, because it seemed to be different from the implementation in the Ceph source code.

2 Scaling parameters

I think about a straw bucket which has N items. The weight of the item i is w_i, and items are sorted by their weights.

$$0 < w_1 < w_2 < \cdots < w_{N-1} < w_N \tag{1}$$

The scaling parameter of the item i is f_i.

$$0 < f_1 < f_2 < \cdots < f_{N-1} < f_N \tag{2}$$

f_i are normalized. The products of them are unity.

$$f_1 f_2 \cdots f_N = 1 \tag{3}$$

The probability where the item i is choices is

$$p_i = \sum_{j=1}^{i} \left(\frac{1}{N-j+1} \prod_{k=j}^{N} \frac{f_j - f_{j-1}}{f_k} \right) \tag{4}$$

; where $f_0 := 0$. It should be in proportion to the weight of the item:

$$\frac{w_i}{W} = p_i,\ W := \sum_{i=1}^{N} w_i \tag{5}$$

.

Here, I define Δw_i as

$$\Delta w_i := \frac{w_i - w_{i-1}}{W},\ w_0 := 0 \tag{6}$$

.

$$\begin{aligned} \Delta w_i &= \frac{(f_i - f_{i-1})^{N-i+1}}{(N-i+1) f_i f_{i+1} \cdots f_N} && (7) \\ &= \frac{(f_i - f_{i-1})^{N-i+1} f_1 f_2 \cdots f_{i-1}}{N-i+1} && (8) \end{aligned}$$

. The last equation is because the production of f_i is unity.

As arranging I can get the recursion about f_i:

$$f_i = f_{i-1} + \left(\frac{(N-i+1)\Delta w_i}{f_1 f_2 \cdots f_{i-1}} \right)^{\frac{1}{N-i+1}} \tag{9}$$

. With this recursion I can get all scaling parameters.

3 Conclusion

I introduced a scheme to get the optimum scaling parameters for straw buckets in CRUSH. However, the straw bucket is old now. A more sophisticated algorithm has been invented by Weil. It is called straw2. Straw2 is far better algorithm than the old straw bucket. The scheme in this note can not be used with straw2. Therefore, I think the scheme is not effective unfortunately. I expect it could be used to another field.

References

[1] S. A. Weil *et al.* Ceph: a scalable, high-performance distributed file system

[2] S. A. Weil *et al.* CRUSH: controlled, scalable, decentralized placement of replicated data