Jack Marquez

Extreme Value Theory

Extreme Value Theory (EVT) is focused on the study of the behavior of the maximum values in a set of random variables, and it is commonly used in statistics to analyze or predict when a new extreme value should happen [1].

In EVT, the values provided by a population must be independent and identically distributed (i.i.d). Also, the distribution of the extreme values tends to one of three possible unique asymptotic forms, Gumbel distribution, Fréchet distribution or Weibull distribution [2].

For the Generalized Extreme Value distribution (GEV), the probability density function (pdf) is:

$$P_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right) & \text{if } \xi \neq 0 \\ \exp\left(-e^{\left(\frac{x-\mu}{\sigma}\right)}\right) & \text{if } \xi = 0 \end{cases} \quad (1)$$

ε = shape, σ = scale, μ = location

The shape parameter in the equation 1 defines if the distribution belongs to Gumbel distribution (ε = 0), Fréchet distribution (ε > 0) or Weibull distribution (ε < 0).

EVT has two methods that allow getting the GEV distribution of the maximum values in a set of data. Those methods are *Peek Over Threshold* and *Block Maxima Method*. In the first one, it is necessary to establish a threshold value and the values that are going to be used to get the distribution are the ones who are bigger than the threshold. In the second method, the data is divided into blocks of the same size and the maximum value in each one must be determined. In this case, I used the *Block Maxima Method* to get all the maximum values in each block and then I used the *Maximum Likelihood Estimation* to find the GEV parameters that best fit for these values [3].

Modeling storage systems performance

As preliminary work, I did some benchmark tests to analyze the performance of storage technologies in a CC cluster. In this case, I developed the test using Lustre file system, which is known for being one of the most used file systems by HPC applications due to its parallel writing. A Cloud Computing cluster (CloudLab) has been configured with 11 nodes, including eight nodes as data storage nodes.

I performed a test in order to simulate a checkpointing process in which data is stored in the file system. I used three types of tests. The first is using the Linux dd command, which allows us to

create and host data of a specific size. The second is with the creation of a C script, which also creates files of a specific size and finally, a benchmark (PIOS) specially designed for Lustre is being used, which also allows the creation of these files of a specific size.

From these initial tests, I characterized the performance or time it takes for the Lustre file system to store this data and I observed that it has a behavior that can be studied through a statistical model using Extreme Value Theory (EVT).

The resulting parameters for this preliminary test were calculated using R and the values are next:

$$\varepsilon = 0.06263,\ \sigma = 0.10278,\ \mu = 3.16291$$

According to the value of the shape, which is close to cero, it is possible to say that these values have a Gumbel distribution seen in figure 1.

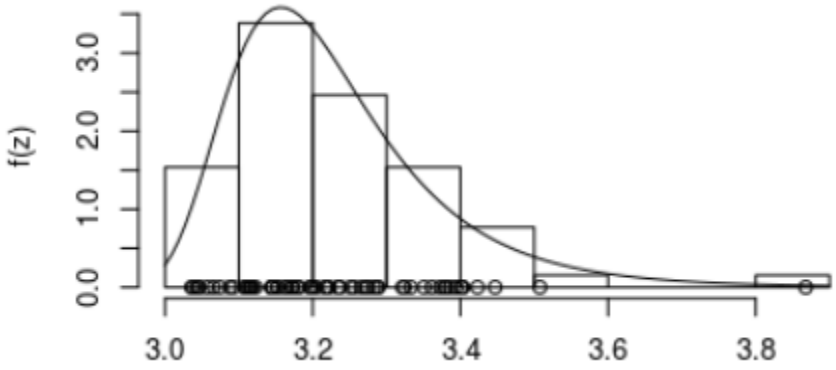

Fig 1. Density plot

Extreme value techniques include the *return level analysis*, which looks to calculate the expected value of the maximum set of variables. This return level is defined as the value that will be exceeded on average only once every N samples every *b* blocks of the distribution. This return value can be calculated as:

$$RL_i = F^{-1}(P) \qquad (2)$$

Figure 2 shows the return levels for this distribution, which is increasing once the number of samples does. The number of samples, in this case, is the number of storage nodes.

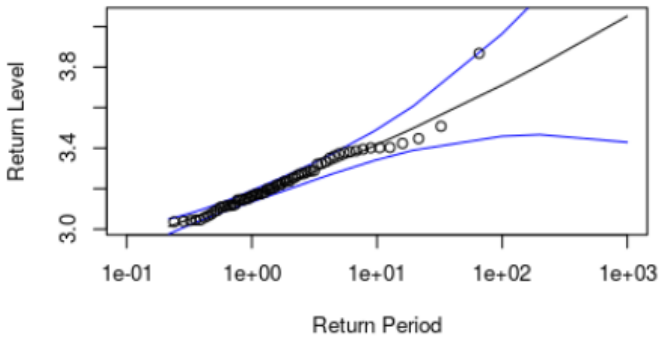

Fig 2. Return level plot

What I want to do next is to extrapolate this case to 16, 32, 64 nodes per now, and validate the behavior of the return levels. Then I will test this kind of file system with technologies that support the infrastructure of CC like Hadoop and Kubernetes.

All this test are being carried out in Chamelon cloud testbed and also in Cloudlab testbed

REFERENCES

[1] R. Gençay and F. Selçuk, “Extreme value theory and Value-at-Risk: Relative performance in emerging markets,” *Int. J. Forecast.*, vol. 20, no. 2, pp. 287–303, Apr. 2004.

[2] E. Castillo, *Extreme Value Theory in Engineering*. Elsevier, 2012.

[3] O. H. Mondragon, P. G. Bridges, S. Levy, K. B. Ferreira, and P. Widener, “Understanding Performance Interference in Next-Generation HPC Systems,” in *SC16: International Conference for High Performance Computing, Networking, Storage and Analysis*, Salt Lake City, UT, USA, 2016, pp. 384–395.