

One of CERNs detectors on the Large Hadron Collider is called ALICE, which stands for A Large Ion Collider Experiment. ALICE is a 10.000 tonne detector, which main purpose is detecting collisions, in particular lead particle collisions. Once these particles collide they become extremely hot and break-down to a matter called quarks and gluons. ALICE detects this collision and gathers the data from it.

Starting in 2018, ALICE will go offline for a two year break to upgrade it's software. A piece of this software is a load balancing algorithm. When ALICE is running, it will pump out 1.1TB/S of data. This algorithm is supposed to be able to efficiently distribute this data over a total of 1750 computers. 250 of these computers are whats called First Level Processors, and the other 1500 are called Event Processing Nodes. All of these computers are meant to process, compress and store the data so that it can be recreated later on.

This data is distributed with heartbeats. These heartbeats occur every 20ms. FLPs will gather the data between two heartbeats, and compresses them into a Sub Time Frame. These STFs are sent to the EPNs. Here they will be compressed further down into a Time Frame. When an EPN is disabled, it should not receive any data anymore, at which it should be removed from the data line. When this event happens it is referred to as a fail-over. This whole process of distributing data both efficiently and correctly is called Load Balancing.

A previous experiment done by Heiko van der Heijden showed that a Blacklist algorithm is a more efficient algorithm for this load balancing application, compared to a Re-initialization algorithm. The experiment was conducted using a cluster available on Nikhef. The amount of units used during the experiment was 2 FLPs and 12 EPNs, which is equivalent to the same ratio that is used at CERN. The availability of this cluster was uncertain, and hence a cluster was needed that was more available, and also expandable. This resulted in a cluster of Raspberry Pi's.

In order to validate this cluster compared to the one on Nikhef, a correlation needed to be found between the results from the cluster and the new one. Afterwards these experiments can be executed with an unit count of 3 FLPs and 18 EPNs, and 4 FLPs and 24 EPNs.

The software used consists of two main frameworks, FairMQ and Zookeeper. FairMQ is a data transport layer of the larger framework FairRoot. For this experiment a specially made trimmed down version was used in order to accommodate with the Raspberry Pi. Zookeeper regulates the Blacklist. Here the registration of online and offline EPNs are done so that the FLPs know what to send where. Zookeeper runs on a special computer called the Information Node.

At first this new prototype using Raspberry Pi's had to be developed and built. The complications lay in the Raspberry Pi having only one Ethernet connection natively, and the software needing changes to work with an ARM architecture. After that, three experiments were conducted that were the same as the previous experiment. The first one disables one EPN during the run, the second disables every EPN apart from one during a run, and the last one disables every EPN during a run, but also uses a random sample size. These three experiments

were then analyzed with the previous experiment to calculate a correlation between them. After that the same three experiments are conducted again with a higher amount of FLPs and EPNs. Finally two new experiments are run, where multiple EPNs will disable at once. During the first one these EPNs will all be after one another in the list of available EPNs. During the second one these EPNs will be scattered in the list of available EPNs. This to check whether the Blacklist algorithm has a different results depending on the layout of the system architecture.

In conclusion the validity of the new prototype is proven by having a Pearson correlation value of 0.95, 0.99 and 0.99. With an error rate of 0.1%, these values all strongly deny the null hypothesis of there being no correlation between the experiments. Furthermore the expansion of the experiment shows that there is no negative effect on the Information Node. TF loss goes closer to 1, which is the minimal TF loss in this current setup. Lastly the layout of the system architecture is important for individual clusters that have a fail-over, but the total TF loss when every cluster has had a fail-over is still the same.

It is recommended to further increase the amount of units to check whether or not the TF loss would eventually get to 1, if it would reach an asymptote at some point, or if it turns around at some point and becomes a sine wave possibly. It is also recommended to continue doing these experiments using Ansible induced fail-overs, as opposed to TF induced fail-overs. This to eliminate the predictability of the fail-overs, and to eliminate the minimal 1 TF loss. Lastly it is recommended to use a lowest latency approach, as opposed to a round robin approach, for the data distribution. Using this in conjunction with the Blacklist might obtain a TF loss of 0, because of the EPN that has a fail-over would not be targetted.