DFKChain Degradation Incident April 10, 2023 Post Mortem

Hello all. We had a degradation on DFKChain today and wanted to report on the cause, steps taken, and what we have learned.

Degradation:

Duration:
From: Apr 10, 2023, approximately 11:25 am EST (block 15859150)
To: Apr 10, 2023, approximately 2:52 pm EST (block 15859176)

Cause:

An unknown issue caused some of the validator nodes to run out of free memory and the operating system terminated the virtual machine process. Only a portion of the processes needed to validate the chain were terminated, which took some time to diagnose.

This issue at first only impacted 1 node, but quickly began to impact more. As the nodes were impacted, the threshold of validators dropped below 75% which is needed for consensus, and the chain stopped producing new blocks regularly.

In total, 5 nodes were impacted (2 at first, then 3 more during the efforts to restore consensus).

An investigation into the root cause is still ongoing.

Action:

The 3 nodes which crashed were immediately restarted but had to regenerate their state due to an unclean shutdown. During that time, the decision was made to upgrade our nodes to 64GB to stay well above the 80% recommended threshold, targeting a 50% buffer.

Once that decision was made, the 4th node to crash was first upgraded to a larger instance size before being restarted. Shortly afterwards, a 5th node crashed, and was upgraded to a larger instance size and restarted.

Because of the risk of our first 3 nodes crashing again after they regenerated, the decision was made to pause and upgrade the first 3 nodes that were still regenerating, so that they would be sure to be stable once that was completed. They were therefore upgraded and restarted the regeneration process.

The remaining 3 nodes were then configured to use swap memory on disk to prevent them from crashing due to the same Out of Memory issue that the other nodes experienced. This prevented them from crashing and allowed us to get the other 5 nodes back to full functionality once they regenerated fully.

Once all 8 nodes were running and the backlog of transactions were mined, the remaining 3 nodes began to be taken offline and upgraded, one by one, until all 8 nodes were running on the larger instance sizes and were stable. This is still ongoing right now.

Learned and Next Steps:

In addition to monitoring CPU and Disk Space, we are adding monitoring and alerts for high memory usage, over a threshold that we consider safe.

We are also enabling swap disk memory on all of the nodes as a fallback in case there are any future spikes.

We will continue to investigate the root cause of the memory usage spike.

We are weighing the option of creating smaller chunks for the regeneration feature to allow for faster recovery in the future in case regeneration is required.