Article

Working with customers I get this question quite often on how to size their cluster. Apart from master nodes, one question that often comes up is the size of data nodes. Hardware vendors now offer disks up to 8TB capacity that can offer customers up to 96 TB of storage per node assuming twelve spindles per node. Even if you go with 12x4TB disks, that is still a whopping 48 TB of storage per node. For most cases, I have always recommended my customers 12x2TB disks over last three years and I continue to do so, given bandwidth remains very expensive, and as we'll see below, it is a very important component when you are sizing a cluster and deciding between high and low density data nodes.

The calculations I am sharing here were done for a customer when they told me that re-replication of blocks when a node fails takes a very long time. This customer had 12x4TB disks on each node.

So rather than preferring one opinion over the other, let's do some maths and then decide what works for your use case. There is no right or wrong answer. As long as you understand what you are doing and the scenario of what's going to happen when a failure occurs is acceptable risk for your business, then choose that method. This article is to help guide you make that decision.

Let us make some assumptions about our jobs and disks.

Assume a server chassis that allows 12 spindles.

Have 2x1TB disks in RAID1 for OS.

10x2TB disks in JBOD (RAID0) for data nodes.

Assume 50 MB/s per spindle throughput.

In case of failure of one node, we can expect following traffic

10x50MB/s x 0.001 (convert MB to GB) = 0.5 GB/s x 8(convert GB to Gb) = 4.8 Gb/s

Assume 16 TB of data on the disks that needs to re replicated.

16 TB x 1000 (Convert TB to GB) = 16000 GB x 8 (convert GB to Gb) = 96000 Gb.

Time required to re-replicate lost data blocks = 96000 Gb/4.8 Gb per sec = 20000 seconds /60 = 333.33 minutes

= 5.55 hours.

Now see, what happens when you have 48 TB of storage. Assume

2x1TB disks in RAID1 for OS

10x4TB disks in JBOD (RAID0) for data nodes.

Again assume 50 MB/s per spindle throughput

In case of failure of one node, we can expect following traffic.

10x50MB/s x 0.001 (convert MB to GB) = 0.5 GB/s x 8(convert GB to Gb) = 4.8 Gb/s

Assume 36 TB of data on the disks that needs to be re-replicated. 36 TB x 1000 (Convert TB to GB) = 36000 GB x 8(Convert GB to Gb) = 288000 Gb.

Time required to re-replicate lost data blocks = 288000 Gb/4.8Gb per sec = 60000 seconds/60 = 1000 minutes/60 = 16 hours.

Now this can be improved if instead of a chassis with 12 disks, you have a server chassis that allows 24 disks. Then, instead of 10x4TB disks, you will have 22x2TB disks (given 2 disks will be used for OS). This improvement will come at the expense of higher bandwidth. Remember, there is no free lunch. Let's see what happens in this case.

2x1TB disks in RAID1 for OS

22x2TB disks in JBOD (RAID0) for data nodes.

Again assume 50MB/s spindle.

In case of failure of one node, we can expect following traffic.

22x50MB/s x 0.001(Convert MB to GB) = 1.1 GB/s x 8(convert GB to Gb) = 8.8 Gb/s

Assume 40 TB of data on the disks that needs to be re-replicated. 40 TB x 1000(Convert TB to GB) = 40000 GB x 8(Convert GB to Gb) = 320,000 Gb.

Time required to re-replicate lost data blocks = 320,000 Gb/8.8 Gb per sec = 36,363 seconds/60 = 606 minutes/60 = 10 hours.

So, the time to re-replicate lost blocks is down to 10 hours from 16 hours while you also increased the amount of data on each node by 4TB.

As you have seen that number of spindles improve performance. They also use more bandwidth. But under normal circumstances when you are not re-replicating blocks due to failure, more spindles will result in better performance.

Depending on the use case, assuming performance is desired, 12x2TB is better than 12x4TB and similarly 24 x 1TB is better than 12x2TB.

Your decision to choose number of disks should also consider other factors like MTTF of a disk which will impact the number of failures you can expect as you increase the number of disks. That discussion for some other time.