SLEC Priority Percent Calculation [WIP]

Discussion on General Ideas

The math note is an expansion on the combinatorics calculation implemented in the MLEC sim's flat DP mode. The concept is simple - when we have a disk failure, we are interested in how many data on the drive (priority percent) belongs to the disk's priority and then we can calculate the time needed for repair in relation to normal RAID. Knowing the repair time in relation to normal RAID, we can then calculate the MTTDL of network SLEC DP in relation to normal RAID and mathematically calculate the number of nines for network SLEC DP durability.

Notations

- r: number of racks
- B: number of disks per rack
- k: number of data chunks
- m: number of parity chunks
- p: the current priority of the disk
- n: stripe width $(k+m)$
- f: total number of failed disk
- f_s: number of failures in the stripe (number of disks that we need to repair)
- $f_{r_{[i]}}$: number of failures in the rack, with parameter $[i]$ indicating which rack the variable is referring to
- $\mathcal{CAP}$: capacity of the disk
- $S_{net_W} = S_{net_R} = S_{net}$: speed of network IO
- $S_{disk_W} = S_{disk_R} = S_{disk}$: speed of disk

Discussion on Speed Bottleneck

For the sake of simplicity for discussion, we currently assume that the network bandwidth $S_{net} = \infty$, so that we only consider S_{disk} during repair operations. In the future we will make more indepth analysis of the network bandwidth's impact on repair speed.

Discussion on Parallelism

Due to the constraint that network SLEC DP places on chunk placement, when there is failure, we read from one node per rack for all other racks (k in total). We use the following notation

- f: total number of failed disk

$$\text{read parallelism} = (r - f_s)B$$

When writing the reconstructed chunk to the *virtual reserved space*, we still want to utilize all of the virutal reserved spaces on all the surviving disks to maximize reconstruction. Therefore,

$$\text{write parallelism} = rB - f$$

Discussion on Priority Percent Calculation for Plain DP

We can see from the general discussion that we are interested in knowing what percentage of data on the failed disk belongs to which priority, and repair them at different speeds/priority. The following rough equation is how the calculation is carried out in plain DP. We first want to know the number of stripes that are in the degraded state (has at least one failed chunks). Therefore, we select $(n-1)$ drives from all the drives $(\text{good\_num} + \text{fail\_num} - 1)$ (minus 1 for at least one failure). The total number of stripes in the degraded state is

$$\text{number of impacted stripes} = ncr(\text{good\_num} + \text{fail\_num} - 1, n - 1)$$

We then want to calculate how many stripes are of the current priority. We first select the surviving drives in the stripe from all the surviving drives.

$$\text{all possible surviving chunks} = ncr(\text{good\_num}, (n - p))$$

Then we select all the failed chunks out of failed drives. Note that since we have already included the *at least one* failure in the total affected stripes calculation, we need to make adjustment here

$$\text{all possible failed chunks} = ncr(\text{fail\_num-1}, p - 1)$$

Combining the three equations above, we can get the priority percentage calculation for **plain DP** as follows

$$\text{prio percent} = \frac{\text{possible failed chunks} * \text{possible surviving chunks}}{\text{affected stripes}} = \frac{ncr(\text{good\_num}, (n-p)) * ncr(\text{fail\_num-1}, p-1)}{ncr(\text{good\_num} + \text{fail\_num} - 1, n-1)}$$

Discussion on Priority Percent Calculation for SLEC network DP

For priority percent calculation of SLEC network DP, we are going to follow a similar concept, but make adjustments according to the constraint of one chunk per rack placement. Due to this constraint, we can see that

- fail_num should not be the total number of failed disks, but rather the number of failed disks in the stripe
- good_num should not be the total number of surviving disks, but rather the number of surviving drives on unimpacted racks (stripe level)
- note that for the denominator, we still use $ncr(s + f - 1, n - 1)$ instead of $ncr(s + f_s - 1, n - 1)$ because we are calculating all the degraded stripes

Given these two modifications, the priority percent calculation becomes

$$\text{good\_num(SLEC)} = s = \sum_{\substack{i=0 \\ i \not\subset \text{impacted racks}}}^{r} (B - f_{r_i})$$

$$\text{prio percent} == \frac{ncr(s, (n-p)) * ncr(f_s - 1, p - 1)}{ncr(s + f - 1, n - 1)}$$

Enumerate Some Cases

I. When there is a single failure [distinct rack], this means that

- All the affected stripes must not have another drive on the same rack
- The remaining good drives to select affected stripes out of are $(r - 1)B$
- The number of failed drive is 1 and has priority of 1
- The priority percent equation then becomes

$$\frac{ncr((r-1)B, n-1) * ncr(1-1, 1-1)}{ncr((r-1)B + 1 - 1, n-1)} = \frac{ncr((r-1)B, n-1)}{ncr((r-1)B, n-1)} = 1$$

II. When there are two failures [same rack], this means that the **priority 1 stripes**

- All the affected stripes must not have another drive on the same rack
- The remaining good drives to select affected stripes out of are $s = (r - 1)B$ because damaged stripes will have other chunks sitting on other racks, and the damaged chunk be sitting on either one of the failed drive
- The the total number of failed drive is 2. However, failures with in stripe f_s is 1 for all degraded stripes due to SLEC property.
- The priority percent equation then becomes

$$\frac{ncr((r-1)B, n-1) * ncr(2-1, 1-1)}{ncr((r-1)B + 1 - 1, n-1)} = \frac{ncr((r-1)B, n-1)}{ncr((r-1)B, n-1)} = 1$$

III. When there are two failures [distinct rack], this means that **priority 1 stripe**

- The remaining good drive is still $(r - 1)B - 1$ because the priority 1 stripe will have all surviving chunks in all racks except the one that contains the failed chunk. The minus 1 is because in one of the rack containing one of the surviving chunk, there is one failed disk that happens to not impact this stripe.
- The number of failed drive is 2
- The priority percent equation then becomes

$$\frac{ncr((r-1)B - 1, n-1) * ncr(2-1, 1-1)}{ncr((r-1)B - 1 + 2 - 1, n-1)} = \frac{(r-1)B - n + 1}{(r-1)B}$$

Using the same $r = 10, B = 10, n = 10$ example, we have priority percent equals $\frac{81}{90} = 0.9$

IV. When there are two failures [distinct rack], this means that **priority 2 stripe**

- All the stripes with priority 2 have both of the chunks sitting on each of the failed drive residing in two racks. This means that the remaining good drives to select from is $(r-2)B$.
- The number of failed disk is 2
- The priority of the stripes is 2
- The priority percent equation then becomes

$$\frac{ncr((r-2)B, n-2) * ncr(2-1, 2-1)}{ncr((r-2)B+2-1, n-1)} = \frac{n-1}{(r-2)B+1}$$

Using the same $r = 10$, $B = 10$, $n = 10$ example, we have priority percent equals $\frac{9}{81}$

When there are three failures [1, 2], this means that **priority 1 stripe** containing failed drive on rack 1

-

V. When there are three failures [distinct rack], this means that **pirority 3 stripe**

- The remaining good drive is $(r-3)B$
- The number of failed disk is 3
- The priority of the stripe is 3
- The priority perent equation then becomes

$$\frac{ncr((r-3)B, n-3) * ncr(3-1, 3-1)}{ncr((r-3)B+3-1, n-1)} = \frac{(n-1)(n-2)}{[(r-3)B+1][(r-3)B+2]}$$

VI. When there are four failures [distinct rack], this means that **priority 4 stripe**

- The remaining good drive is $(r-4)B$
- The number of failed disk is 4
- The priority of the stripe is 4
- The priority percent equation then becomes

$$\frac{ncr((r-4)B, n-4) * ncr(4-1, 4-1)}{ncr((r-4)B+4-1, n-1)} = \frac{(n-1)(n-2)(n-3)}{[(r-4)B+1][(r-4)B+2][(r-4)B+3]}$$

VII. When there are n failures across n distinct racks, the stripes with priority n

The priority percent calculation should be the following. First we let the number of failures be f

$$\prod_{i=1}^{f-1} \frac{(n-i)}{[(r-f)B+i]}$$