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 (good_num + 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
Working on finding a good looking model. See enumerated cases below.

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 due to the placement constraint
- The number of failed drive is 1 and has priority of 1
- The percent of stripes of priority 1 out of all the affected stripes is

$$\text{prio percent} = \frac{ncr(r-1, n-1)B^{n-1}}{ncr(r-1, n-1)B^{n-1}} = 1$$

Explaning the equation: for the numerator, we know that there is a single failed disk. Therefore, we select the surviving chunks $(n-1)$ of stripes from the surviving **racks** (not disk because of the placement constraint) $(r-1)$. However, we also have the choice of B disks in each of the surviving racks. Therefore, we add the multiplier of B^{r-1}. For the denomenator, we have the total number of stripes that have at least one failed chunk. Therefore, the equation is identical to the numerator.

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

- We will **only** have priority 1 stripes because there cannot be any stripes that contains both of the failed disk on the same rack. This means that the percent of priority 1 stripe out of all the affected stripe should be 1.
- The percent of stripes of priority 1 out of all the affected stripes is

$$\text{prio percent} = \frac{ncr(r-1, n-1)B^{n-1}}{ncr(r-1, n-1)B^{n-1}} = 1$$

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

- For the numerator, there are two conditions.
  - stripes that have chunks sitting on both of the impacted racks (however does not contain both failed disk).
  - stripes that have the failed chunk sitting on either of the impacted chunks.
- The percent of stripes of priority 1 out of all the affected stripes is

$$\text{prio percent} = \frac{ncr(r-2, n-2)B^{n-2}(B-1) + ncr(r-2, n-1)B^{n-1}}{ncr(r-1, n-1)B^{n-1}} = \frac{B-1}{B}\frac{n-1}{r-1} + \frac{r-n}{r-1}$$

Using an example of $r = 10, B = 10, n = 9$, we can see that the priority percent would be

$$\text{prio percent} = \frac{ncr(8,7) * 10^7 * 9 + ncr(8,8) * 10^8}{ncr(9,8) * 10^8} = \frac{82}{90}$$

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

- This means that all the priority 2 stripes contains both of the failed disks
- The percent of priority 2 stripes out of all the affected stripe is

$$\text{prio percent} = \frac{ncr(r-2, n-2)B^{n-2}}{ncr(r-1, n-1)B^{n-1}} = \frac{1}{B}\frac{n-1}{r-1}$$

Using the example of $r = 10, B = 10, n = 9$, we can see that the priority percent would be

$$\text{prio percent} = \frac{ncr(8,7) * 10^7}{ncr(9,8) * 10^8} = \frac{8}{90}$$

We can see that this prio percent, combined with the prio percent of priority 1 stripes, addes up to 1. Which means that our model is correct.

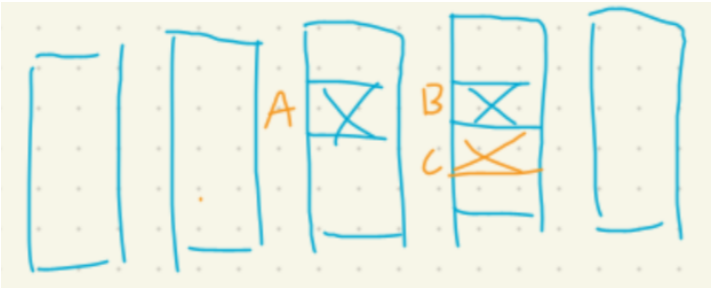
V. When there are three failures [same rack], this means that **priority 1 stripes**

- There will still be only priority 1 stripes
- This then makes the formula identical to single failure priority percent calculation

$$\text{prio percent} = \frac{ncr(r-1,n-1)B^{n-1}}{ncr(r-1,n-1)B^{n-1}} = 1$$

VI. When there are three failures [1, 2]

This graph shows an example of the failure pattern.



As we can see, due to the asymmetrical layout, we would have to discuss priority percent for each drive.

For priority 1 stripes on disk A

- It can either include a chunk on the rack with two failures, or only include chunks on unimpacted racks
- Therefore, the percentage of priority 1 stripes out of all the stripes that are impacted by disk A is

$$\text{prio percent} = \frac{ncr(r-2,n-2)B^{n-2}(B-2) + ncr(r-2,n-1)B^{n-1}}{ncr(r-1,n-1)B^{n-1}} = \frac{n-1}{r-1}\frac{B-2}{B} + \frac{r-n}{r-1}$$

For priority 2 stripes on disk A

- It can include either disk B or C as the second failed chunk, and it has to have failed chunks sitting on both of the rack.
- Therefore, the percentage of priority 2 stripes out of all the stripes that are impacted by disk A is

$$\text{prio percent} = \frac{ncr(r-2,n-2)B^{n-2} * 2}{ncr(r-1,n-1)B^{n-1}} = \frac{n-1}{r-1}\frac{2}{B}$$

For priority 1 stripes on disk B/C

- It can include a chunk on rack of disk A, or have all other chunks sitting on unimpacted racks
- Therefore, the percentage of priority 1 stripes out of all the stripes that are impacted by disk B is

$$\text{prio percent} = \frac{ncr(r-2,n-2)B^{n-2}(B-1) + ncr(r-2,n-1)B^{n-1}}{ncr(r-1,n-1)B^{n-1}} = \frac{n-1}{r-1}\frac{B-1}{B} + \frac{r-n}{r-1}$$

For priority 2 stripes on disk B/C

- It has to include disk A
- Therefore, the percentage of priority 2 stripe sout of all the stripes that are impacted by disk B is

$$\text{prio percent} = \frac{ncr(r-2,n-2)B^{n-2}}{ncr(r-1,n-1)B^{n-1}} = \frac{n-1}{r-1}\frac{1}{B}$$

VII. When there are three failures [distinct rack], this means **priority 1 stripes**

- We have three scenarios to select good ones out of
- – All the other chunks are sitting on unimpacted racks
  – One chunk is sitting on any of the other two impacted racks
  – Two chunks are sitting on the other two impacted racks
- Therefore, the percentage of priority 1 stripes out of all the stripes that are impacted by one of the failed disks is

$$\begin{aligned}\text{prio percent} &= \frac{ncr(r-3,n-1)B^{n-1} + ncr(r-3,n-2)B^{n-2}(B-1) + ncr(r-3,n-3)B^{n-3}(B-1)^2}{ncr(r-1,n-1)B^{n-1}} \\ &= \frac{(r-n)(r-n-1)}{(r-1)(r-2)} + 2\frac{(r-n)(n-1)}{(r-1)(r-2)}\frac{B(B-1)}{B^2} + \frac{(n-1)(n-2)}{(r-1)(r-2)}\frac{(B-1)^2}{B^2}\end{aligned}$$

VIII. When there are three failures [distinct] rack, this means **priority 2 stripes**

- Note that in all the previous priority percent calculation, we only select the good chunks out of the surviving disks (given the rack constraint). However, note that we derive our intuition from the priority percent calculation for flat DP, which involves selecting failed chunks out of all the failed disks as well. This was not needed in all the previous calculations because it would all result in 1, but it starts to matter now.
- There are two scenarios we need to consider
- – Stripe containing two failed chunks on two of the impacted racks, and all other chunks reside on unimpacted racks
  – Stripe containing two failed chunks on two of the impacted racks, and one of the surviving chunk reside on the other impacted rack
- For each of the above scenario, we can see that we have $ncr(2,1) = 2$ ways of choosing the two failed chunks out of the three failed disks.
- Therefore, the percentage of priority 2 stripes out of all the stripes that are impacted by one of the failed disks is

$$\begin{aligned}\text{prio percent} &= \frac{ncr(r-3,n-2)B^{n-2}ncr(2,1) + ncr(r-3,n-3)B^{n-3}(B-1)ncr(2,1)}{ncr(r-1,n-1)B^{n-1}} \\ &= \frac{(r-n)(n-1)}{(r-1)(r-2)}\frac{2B}{B^2} + \frac{(n-1)(n-2)}{(r-1)(r-2)}\frac{2(B-1)}{B^2}\end{aligned}$$

IX. When there are three failures [distinct] rack, this means **priority 3 stripes**

- Have 3 failed chunks sitting on each of the impacted racks
- Therefore, the percentage of priority 3 stripes out of all the stripes that are impacted by one of the failed disks is

$$\text{prio percent} = \frac{ncr(r-3,n-3)B^{n-3}}{ncr(r-1,n-1)B^{n-1}} = \frac{(n-1)(n-2)}{(r-1)(r-2)}\frac{1}{B^2}$$

X. When there are $\mathcal{N}$ failures, this means that priority $\mathcal{N}$ stripe

$$\text{prio percent} = \frac{ncr(r-\mathcal{N},n-\mathcal{N})B^{n-\mathcal{N}}}{ncr(r-1,n-1)B^{n-1}} = \frac{1}{B^{\mathcal{N}-1}}\frac{\prod_{i=1}^{\mathcal{N}-1}(n-i)}{\prod_{i=1}^{\mathcal{N}-1}(r-i)}$$