IBM **Spectrum Scale**

Fast File System Rebalance

Spectrum Scale User Group Meeting 2017, Denver

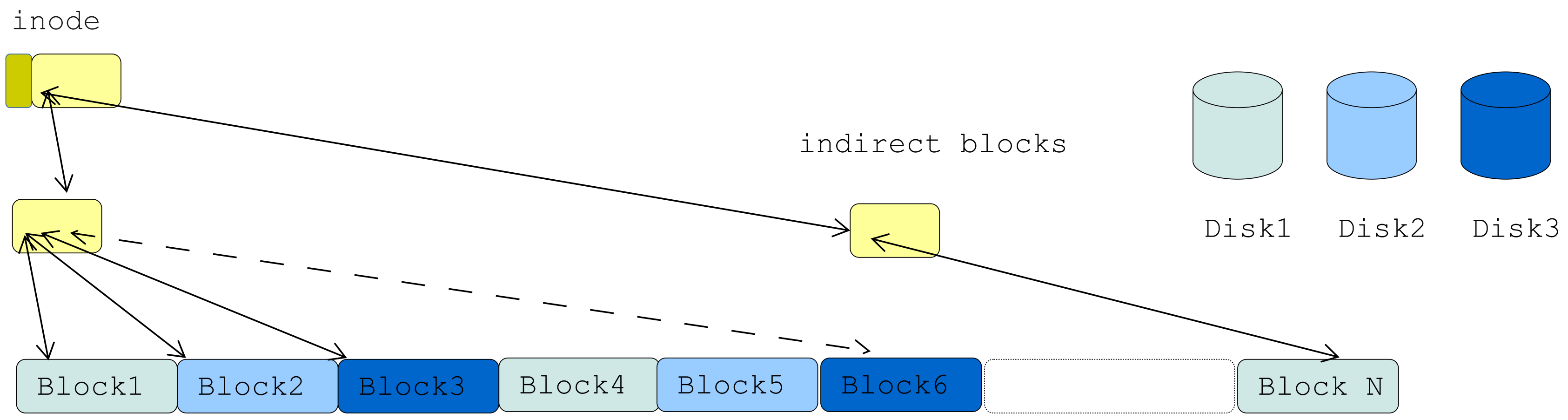
Hai Zhong Zhou

Agenda

- File System Rebalance Concept
- Current Approach & Problem
- New Approach & Benefits
- Benefits Result
- Q & A

File System Rebalance Concept

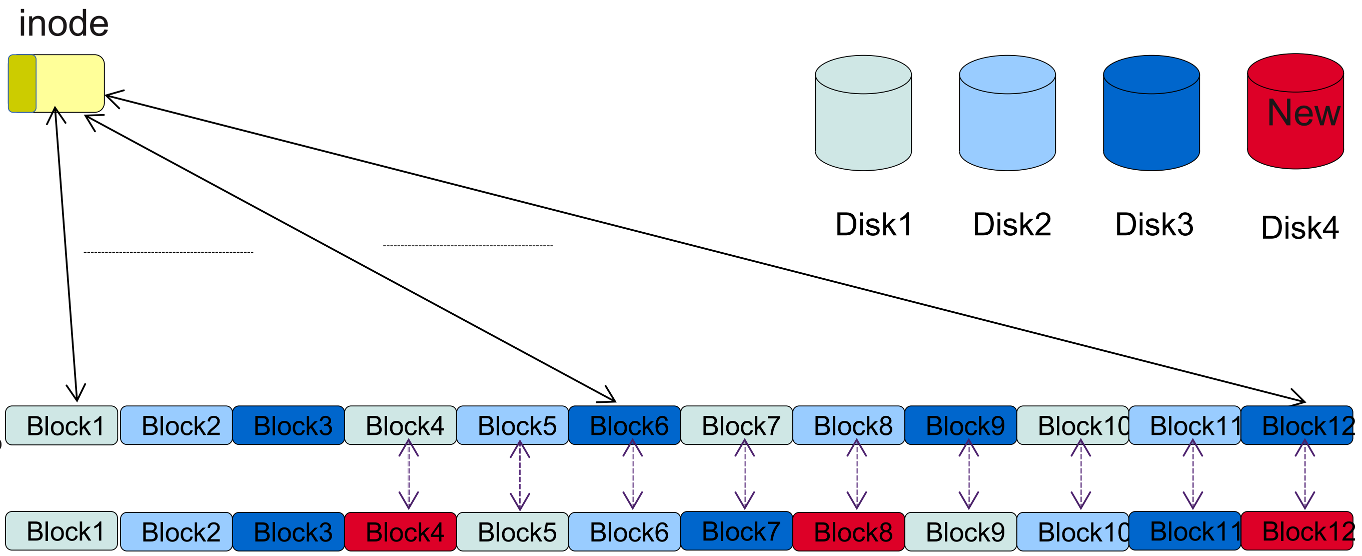
- GPFS spreads files more or less evenly across all of its disks to obtain the fullest disk bandwidth utilization and to maximize performance. However, for one reason or another, the file system could become out of balance and therefore a rebalance is needed. For example, when a new disk is added to the file system, the file system needs to be rebalanced.



Data blocks are put on available disks evenly across all available disks, with static fixed order, D1->D2->D3 in this example

Current Approach & Problem

Current Approach:
With the static fixed round robin order, the blocks are not only moved from previously existing disks to newly added disks, but also moved between previously existing disks. Basically the (N+M-1)/(N+M) percentage of total blocks would be moved (M is the number of previously existing disks, N is the number of newly added disks). Means almost 100% data in file system will be moved, when N and M are big numbers.

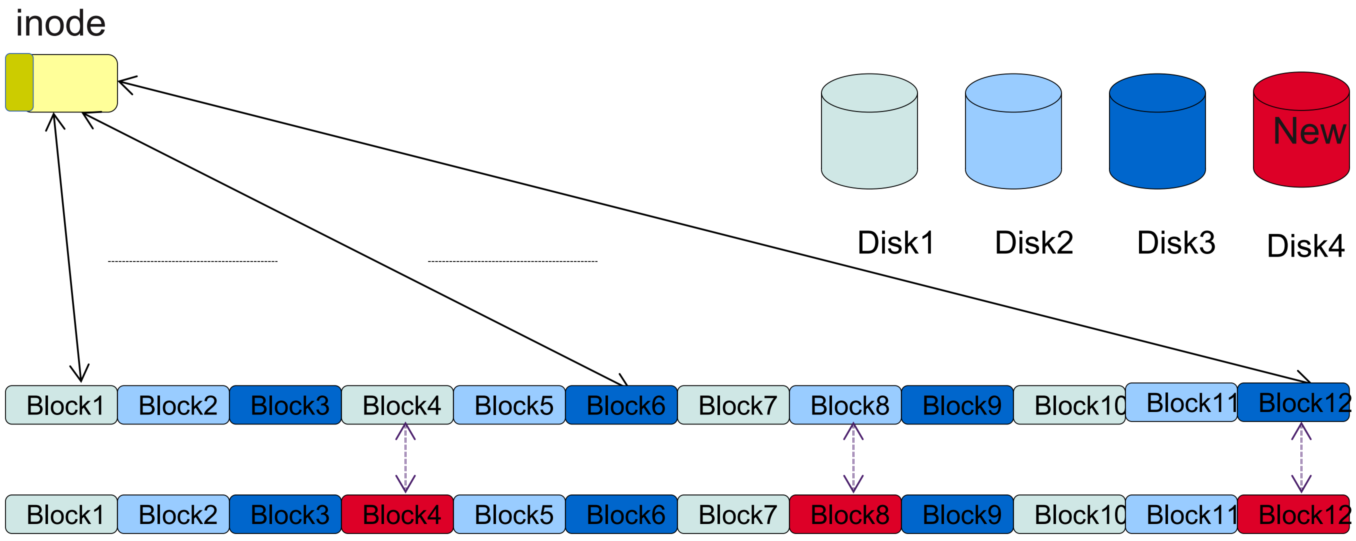


Before rebalance, data blocks are rounded robin with fixed order: D1->D2->D3
After rebalance, data blocks are rounded robin with new fixed order: D1->D2->D3->D4
9/12=75% blocks moved in this example: N=1, M=3, (3+1-1)/(3+1) =75% in theory.

Problem:
- Moving large amount of the file system data is a very I/O intensive and thus a time-consuming operation.
- Big workloads to network and disk subsystems, so big impacts to normal operations.

New Approach & Benefits

New Approach:
The new approach is only moving some blocks from the previously existing disks to the newly added disks, by introducing a dynamic round robin order for blocks placements. Basically only N/(N+M) percentage of total blocks would be moved(N is the number of newly added disks, M is the number of previously existing disks). When M is a large number, and N is a small number, then just a very small fraction of blocks would be moved.



Before rebalance, data blocks are rounded robin with fixed order: D1->D2->D3
After rebalance, data blocks are rounded robin with new dynamic order:D1->D2->D3->D4->D2->D3->D1->D4->D3->D1->D2->D4
3/12=25% blocks moved in this example: N=1, M=3, 1/(1+3)=25% in theory.

Benefits:
- Minimizing data blocks movement, then rebalance operation would be light weighted and can proceed very fast.
- Minimizing data blocks movement, then minimizing the workloads from rebalancing to network and disk subsystems, so minimizing the impacts to normal operations.

Benefits Results

- Performance improvements from fast rebalancing would be variant, could be 2 times faster or dozens of times faster, depending on the numbers of NSD and replicas configuration for the file system. See data movements percentage comparison: N/(N+M) vs (N+M-1)/(N+M), e.g. N=1, M=100, then 1/101 vs 100/101, means the data movements during fast rebalancing is only 1/100 of the movements made with old rebalancing.
- Impacts introduced by fast rebalancing is significantly reduced, happy to user's applications.

One tests example results for the old and new fast rebalance:
Configuration: 64 existed NSD disks in the file system, and 8 NSD disks were added.

Experiments	4.2.3.2 vs. 4.2.4 SpeedUp (restripefs_b_4232/ restripefs_b_424)		
	mmrestripefs -b after mmaddisk	mmrestripefs -b after mmchdisk suspend	
Replication=1 (-m 1 -r 1), 50% filled capacity, -N all	4.50 times	4.68 times	
Replication=1 (-m 1 -r 1), 90% filled capacity, -N all	5.18 times	5.73 times	
Replication=2 (-m 2 -r 2), 50% filled capacity, -N all	1.68 times	2.50 times	
Replication=2 (-m 2 -r 2), 90% filled capacity, -N all	1.74 times	2.87 times	
Replication=1 (-m 1 -r 1), 90% filled capacity, adddisk/deldisk ONE DISK	10.75 times	11.42 times	

Notes:
- The new fast rebalance will be the default option for user.
- The limitation on total number of PIT worker threads has been removed since 4.2.3 release.

Q & A

Thank You!