Answers (/community/s/group/0F90L0000001NDBSA2/) — Hadooping__ (/community/s/profile/0050L0000093B1yQAE) (Customer) asked a question.
November 11, 2012 at 6:53 PM (/community/s/question/0D50L00006BIv3WSAT/nitty-gritty-singledrive-per-volume-raid-0-vs-jbod)

Nitty Gritty: Single-Drive per Volume RAID 0 vs. JBOD

I am looking at testing MapR for my employer and we want to see if MapR can handle some of our health care data warehousing needs. I have some test servers that have drives hooked to RAID controllers. Looking at the RAID cards, they don't support jbod, so I just figured I'd present each drive as a 1 drive RAID 0 volume. Can the MapRFS, with all it's enhancements over HDFS understand the drive geometry properly when abstracted through a logical layer? Will this setup even work? If it does work, and I sacrificing some potential performance? Are there any theoretical gotcha's if I then pushed this into a production environment?

Also, please go into detail if you can. Why this will work, why performance won't be a problem, reading through MapRs documentation, I see the big advantage is direct-to-disk but can't get my head around how a "logical" disk presented would do vs a real disk (i.e. if my servers had HBAs etc).

Upvote Answer Share 8 answers 131 views

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
6 years ago

There are a few problems with RAID cards:

1. They cannot handle the I/O rates that the MapR software can throw at them. The drives themselves are perfectly capable, but the cheap I/O card usually falls over. You are better letting it by just a pass-through.

2. The bad-drive detection worsens with more intermediate hops. With direct disk access, MapR runs timers on how long i/o's are taking and will detect bad drives better. Error reporting in MapR also becomes more accurate (MapR uses hdparm to find the problems).

Another major point is the amount of gigabytes of disk that needs to be rebuilt when a drive is lost. When MapR is controlling the disks directly, MapR will only rebuild the __data__ and not the empty (unoccupied cylinders). A raid-controller will rebuild the whole drive, quite a bit of which might be useless since the disk might be only 50% full.

MapR create 3-drive RAID-0's by default. That is, if your box has 12 drives, it will create 4 raid-0 groups of 3 drives each. It is a trade-off between getting higher perf and limiting the amount of data that needs to be rebuilt when a drive fails. In your case if you have a 12-drive RAID-0, the amount of data to be rebuilt will be massive so consider using smaller RAID-0 groups, and present each one to MapR as a separate device. Then use the disksetup option to tell MapR to treat each such RAID-0 device as a separate storage pool so that MapR does not try to stripe across them.

Upvote Reply

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
6 years ago

If I understand correctly, each RAID-0 volume on the Raid card will show up as its own "drive" or "device" in Linux. As far as MapR is concerned it is being presented 10 "drives". We did test such a configuration about 2 years ago, so your Raid cards are probably newer than what we tested on. Our cards cost about $300 each at that time.

The MapR software created 4 storage pool, 2 of which were striped across 3 such "drives" each, and 2 of which used 2 "drives" each. It might have been better to create 2 storage pools of 5 "drives", but we didn't try it.

Performance was ok, not great. It was however much better than doing any sort of RAID-5 or RAID-4 at the card.

Can you clarify the comment about HBA card? Are you planning to use a SAN? We would recommend against that. It is better to use locally attached SATA or SAS drives.

Upvote Reply

Hadooping__ (/community/s/profile/0050L0000093B1yQAE) (Customer)
6 years ago

Basically we are looking at taking our current hardware config, which has a RAID Card that can not do JBOD and testing it with MapR. To get close to raw device access, we are doing 1 drive RAID0 volumes (10 of them) presented to MapR. We did get things up and running and they all show up in MapR as individual disks, but instead of the Disk models, they show up as LSI Drives (LSI is the chipset on the Raid Card I would assume) So my question comes into play here, should we spring for real HBA cards to present the drives directly to MapR if we decide to go the MapR route. I.e. will we get performance, reliability, etc improvement by going this route. And basically, I wanted to ask if MapR had done any testing on those two configurations. Single Drive Raid 0 Volumes presented to MapR vs. JBOD disks presented to Mapr. (i.e. all things being equal, was there a different between JBOD and single drive RAID0). As to the HBA, we just want to get SAS/SATA HBA and present drives directly to MapR. No SAN, just locally attached drives. The only difference is RAID Card with lots of RAID 0 volumes (one for each disk) vs HBA card with each drive presented physically to MapR.

Upvote Reply

Hadooping__ (/community/s/profile/0050L0000093B1yQAE) (Customer)
6 years ago

The second part of your answer is confusing. I am looking at only doing RAID0 in the sense that my RaiD Card doesn't support JBOD, and thus I am presenting each of the 10 drives on the card as RAID0 volume made up of one drive. (i.e. I'll have 10 disks in 10 RAID0 Volumes). While that seems like it should be ok, I was wondering if when going to production, we should switch to HBA cards that do true JBOD. I am not trying to do any RAID0 Volumes of multiple drives (like your 12 rive RAID-0 Volume

comment). So, given that, does your answer stay the same? Have you done tests of multiple drives off a RAID card with sinlge drive RAID0 Volumes vs the same hardware - RAID card + HBA card?

Upvote Reply

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
6 years ago

Based on your description, it should work fine with MapR without requiring HBA's. Our experience with LSI raid-cards is a single card can do about 800 MB/s aggregate in JBOD mode (ie, either 800 MB/s read, or 800 MB/s write, or 400 MB/s read + 400 MB/s write if the reads and writes are 50-50).

On systems with more than 12 drives, we recommended getting two of these cards per box to get better perf.

Thanks for asking us!

Upvote Reply

Hadooping__ (/community/s/profile/0050L0000093B1yQAE) (Customer)
6 years ago

Does the same hold true for a LSI based RAID Card presenting the drives as single drive RAID0 volumes? That is what's getting me. So a LSI card in JBOD mode, you have test results, but what about a card like the Dell Perc 6i where it doesn't support JBOD and the drives can only be presented as single drive RAID0 volumes. Same performance?

Upvote Reply

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
6 years ago

I would expect so, modulo the performance of the card itself. I think we might have the Dell Perc 6i cards as well. I will inquire.

Upvote Reply

John Omernik (/community/s/profile/0050L0000093BFdQAM) (Customer)
6 years ago

I'd be interested in the results of your testing. This RAID0(with a single disk per volume) vs. true JBOD has come up with our cluster as well. It's one of those things that seems like it should work, but true test results would be awesome.

Upvote Reply

Login to answer this question