Hadoop Map/Reduce / MAPREDUCE-6541

Exclude scheduled reducer memory when calculating available mapper slots from headroom to avoid deadlock

Details

Type:	Bug	Status:	RESOLVED
Priority:	Major	Resolution:	Fixed
Affects Version/s:	2.7.1	Fix Version/s:	2.8.0, 3.0.0-alpha2
Component/s:	None		
Labels:	None		
Target Version/s:	2.7.4		
Hadoop Flags:	Reviewed		

Description

We saw a MR deadlock recently:

- When NM restarted by framework without enable recovery, containers running on these nodes will be identified as "ABORTED", and MR AM will try to reschedule "ABORTED" mapper containers.
- Since such lost mappers are "ABORTED" container, MR AM gives normal mapper priority (priority=20) to such mapper requests. If there's any pending reducer (priority=10) at the same time, mapper requests need to wait for reducer requests satisfied.
- In our test, one mapper needs 700+ MB, reducer needs 1000+ MB, and RM available resource = mapper-request = (700+ MB), only one job was running in the system so scheduler cannot allocate more reducer containers AND MR-AM thinks there're enough headroom for mapper so reducer containers will not be preempted.

~~MAPREDUCE-6302~~ can solve most of the problems, but in the other hand, I think we may need to exclude scheduled reducers resource when calculating #available-mapper-slots from headroom. Which we can avoid excessive reducer preemption.

Attachments

MAPREDUCE-6541.01.patch	9 kB	12/Nov/15 14:20
MAPREDUCE-6541.02.patch	8 kB	27/Oct/16 11:00

Issue Links

relates to

MAPREDUCE-6513 MR job got hanged forever when one NM unstable for some time CLOSED

Activity

Wangda Tan added a comment - 06/Nov/15 21:47

+jlowe, kasha.

Naganarasimha G R added a comment - 06/Nov/15 22:19

Hi wangda, Case seems to be similar with ~~MAPREDUCE-6513~~ ?

Wangda Tan added a comment - 06/Nov/15 22:29

Naganarasimha, think for pointing me ~~MAPREDUCE-6513~~, I think they're similar issues, but the proposal maybe a little different:

- ~~MAPREDUCE-6513~~ is trying to make retried mappers has higher priority.
- This JIRA is trying to exclude pending reducer memory so reducer preemption will involve earlier.

I'm linking the two JIRAs.

Naganarasimha G R added a comment - 06/Nov/15 23:18

Hi wangda, Then will it be similar to ~~MAPREDUCE-6514~~ ?

Wangda Tan added a comment - 06/Nov/15 23:21

Naganarasimha, that's also different: ~~MAPREDUCE-6514~~ is to fix bug in existing reducer preemption logic (remove reducer request locally but not notify RM). This is an enhancement of how to calculate available mapper slots.

Varun Saxena added a comment - 06/Nov/15 23:53

Assigned it to myself as I am working on 2 similar JIRAs'.
Wangda kindly reassign if you will be working on it.

Haven't thought through all the cases but on the face of it, this makes sense. Because reducers will have higher priority so whether mapper has enough headroom needs to take into account that pending reducers will be assigned resources before it. Which will be done if we exclude pending reducers resources. If I am not wrong you are talking about below condition in RMContainerAllocator#preemptReducesIfNeeded

```
    // The pending mappers haven't been waiting for too long. Let us see if
    // the headroom can fit a mapper.
    Resource availableResourceForMap = getAvailableResources();
    if (ResourceCalculatorUtils.computeAvailableContainers(availableResourceForMap,
        mapResourceRequest, getSchedulerResourceTypes()) > 0) {
      // the available headroom is enough to run a mapper
      return;
    }
```

Coming to ~~MAPREDUCE-6514~~, in addition to updating ask, it will also consider whether to ramp up reduces or not if maps are hanging in scheduleReduces(). Will update JIRA description accordingly.

Wangda Tan added a comment - 07/Nov/15 00:39

Thanks for sharing your thoughts varun_saxena. And thanks for taking this, please go ahead.
Reconsidered these issues, I think 3 fixes are all required:

- ~~MAPREDUCE-6513~~: failed/killed mappers should have higher priority
- ~~MAPREDUCE-6514~~: reducer preemption should also cleanup resource requests in RM side.
- And also this one.

I think previous two are more important, this is just an optimization.

+vinodkv.

Varun Saxena added a comment - 09/Nov/15 09:04

I think it should be scheduled reducer's memory which should be excluded instead of pending.

Wangda Tan added a comment - 09/Nov/15 17:52

varun_saxena, yes you're correct, updated title/desc.

Thanks,

Hadoop QA added a comment - 12/Nov/15 14:59

☒ -1 overall

Vote	Subsystem	Runtime	Comment
0	reexec	0m 6s	docker + precommit patch detected.
+1	@author	0m 0s	The patch does not contain any @author tags.
+1	test4tests	0m 0s	The patch appears to include 1 new or modified test files.
+1	mvninstall	3m 3s	trunk passed
+1	compile	0m 17s	trunk passed with JDK v1.8.0_60
+1	compile	0m 20s	trunk passed with JDK v1.7.0_79
+1	checkstyle	0m 10s	trunk passed

+1	mvnsite	0m 24s	trunk passed
+1	mvneclipse	0m 15s	trunk passed
+1	findbugs	0m 43s	trunk passed
+1	javadoc	0m 17s	trunk passed with JDK v1.8.0_60
+1	javadoc	0m 18s	trunk passed with JDK v1.7.0_79
+1	mvninstall	0m 23s	the patch passed
+1	compile	0m 18s	the patch passed with JDK v1.8.0_60
+1	javac	0m 18s	the patch passed
+1	compile	0m 19s	the patch passed with JDK v1.7.0_79
+1	javac	0m 19s	the patch passed
+1	checkstyle	0m 10s	the patch passed
+1	mvnsite	0m 23s	the patch passed
+1	mvneclipse	0m 15s	the patch passed
+1	whitespace	0m 0s	Patch has no whitespace issues.
+1	findbugs	0m 55s	the patch passed
+1	javadoc	0m 16s	the patch passed with JDK v1.8.0_60
+1	javadoc	0m 18s	the patch passed with JDK v1.7.0_79
-1	unit	9m 21s	hadoop-mapreduce-client-app in the patch failed with JDK v1.8.0_60.
+1	unit	10m 36s	hadoop-mapreduce-client-app in the patch passed with JDK v1.7.0_79.
-1	asflicense	0m 25s	Patch generated 7 ASF License warnings.
		30m 33s	

Reason	Tests
JDK v1.8.0_60 Timed out junit tests	org.apache.hadoop.mapreduce.v2.app.TestFail

Subsystem	Report/Notes
Docker	Client=1.7.1 Server=1.7.1 Image:test-patch-base-hadoop-date2015-11-12
JIRA Patch URL	https://issues.apache.org/jira/secure/attachment/12771973/MAPREDUCE-6541.01.patch
JIRA Issue	~~MAPREDUCE-6541~~
Optional Tests	asflicense compile javac javadoc mvninstall mvnsite unit findbugs checkstyle
uname	Linux 2ff195d93dd8 3.13.0-36-lowlatency #63-Ubuntu SMP PREEMPT Wed Sep 3 21:56:12 UTC 2014 x86_64 x86_64 x86_64 GNU/Linux
Build tool	maven
Personality	/home/jenkins/jenkins-slave/workspace/PreCommit-MAPREDUCE-Build/patchprocess/apache-yetus-fa12328/precommit/personality/hadoop.sh
git revision	trunk / 9ad708a
findbugs	v3.0.0
unit	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6154/artifact/patchprocess/patch-unit-hadoop-mapreduce-project_hadoop-mapreduce-client_hadoop-mapreduce-client-app-jdk1.8.0_60.txt
unit test logs	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6154/artifact/patchprocess/patch-unit-hadoop-mapreduce-project_hadoop-mapreduce-client_hadoop-mapreduce-client-app-jdk1.8.0_60.txt
JDK v1.7.0_79 Test Results	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6154/testReport/
asflicense	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6154/artifact/patchprocess/patch-asflicense-problems.txt

modules	C: hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app U: hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app
Max memory used	228MB
Powered by	Apache Yetus http://yetus.apache.org
Console output	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6154/console

This message was automatically generated.

Vinod Kumar Vavilapalli added a comment - 18/Aug/16 01:13

2.7.3 is under release process, changing target-version to 2.7.4.

Hadoop QA added a comment - 18/Aug/16 01:20

☒ **-1 overall**

Vote	Subsystem	Runtime	Comment
0	reexec	0m 0s	Docker mode activated.
-1	patch	0m 6s	~~MAPREDUCE-6541~~ does not apply to trunk. Rebase required? Wrong Branch? See https://wiki.apache.org/hadoop/HowToContribute for help.

Subsystem	Report/Notes
JIRA Patch URL	https://issues.apache.org/jira/secure/attachment/12771973/MAPREDUCE-6541.01.patch
JIRA Issue	~~MAPREDUCE-6541~~
Console output	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6676/console
Powered by	Apache Yetus 0.3.0 http://yetus.apache.org

This message was automatically generated.

Naganarasimha G R added a comment - 27/Oct/16 10:39

Hi varun_saxena, can you please rebase the patch !

Varun Saxena added a comment - 27/Oct/16 10:57

Sure. Will do it shortly.

Varun Saxena added a comment - 27/Oct/16 11:01

Updated the patch

Hadoop QA added a comment - 27/Oct/16 11:30

☒ **-1 overall**

Vote	Subsystem	Runtime	Comment
0	reexec	0m 11s	Docker mode activated.
+1	@author	0m 0s	The patch does not contain any @author tags.
+1	test4tests	0m 0s	The patch appears to include 1 new or modified test files.
+1	mvninstall	6m 42s	trunk passed
+1	compile	0m 29s	trunk passed

+1	checkstyle	0m 18s	trunk passed
+1	mvnsite	0m 28s	trunk passed
+1	mvneclipse	0m 16s	trunk passed
+1	findbugs	0m 34s	trunk passed
+1	javadoc	0m 15s	trunk passed
+1	mvninstall	0m 22s	the patch passed
+1	compile	0m 20s	the patch passed
+1	javac	0m 20s	the patch passed
-1	checkstyle	0m 15s	hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app: The patch generated 8 new + 226 unchanged - 0 fixed = 234 total (was 226)
+1	mvnsite	0m 27s	the patch passed
+1	mvneclipse	0m 12s	the patch passed
+1	whitespace	0m 0s	The patch has no whitespace issues.
+1	findbugs	0m 42s	the patch passed
+1	javadoc	0m 13s	the patch passed
+1	unit	9m 0s	hadoop-mapreduce-client-app in the patch passed.
+1	asflicense	0m 15s	The patch does not generate ASF License warnings.
		21m 34s	

Subsystem	Report/Notes
Docker	Image:yetus/hadoop:9560f25
JIRA Patch URL	https://issues.apache.org/jira/secure/attachment/12835550/MAPREDUCE-6541.02.patch
JIRA Issue	~~MAPREDUCE-6541~~
Optional Tests	asflicense compile javac javadoc mvninstall mvnsite unit findbugs checkstyle
uname	Linux 34acfc665c56 3.13.0-93-generic #140-Ubuntu SMP Mon Jul 18 21:21:05 UTC 2016 x86_64 x86_64 x86_64 GNU/Linux
Build tool	maven
Personality	/testptch/hadoop/patchprocess/precommit/personality/provided.sh
git revision	trunk / 4e403de
Default Java	1.8.0_101
findbugs	v3.0.0
checkstyle	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6780/artifact/patchprocess/diff-checkstyle-hadoop-mapreduce-project_hadoop-mapreduce-client_hadoop-mapreduce-client-app.txt
Test Results	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6780/testReport/
modules	C: hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app U: hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app
Console output	https://builds.apache.org/job/PreCommit-MAPREDUCE-Build/6780/console
Powered by	Apache Yetus 0.3.0 http://yetus.apache.org

This message was automatically generated.

Varun Saxena added a comment - 27/Oct/16 11:37

Naganarasimha, want me to fix checkstyle ? Most of them (i.e. whitespace after {) are false negatives

Naganarasimha G R added a comment - 27/Oct/16 12:29

not required will commit the patch!

Naganarasimha G R added a comment - 27/Oct/16 12:42

Thanks for the contribution varun_saxena and review from wangda. Committed it to trunk, branch-2 & branch-2.8.

Hudson added a comment - 27/Oct/16 13:00

SUCCESS: Integrated in Jenkins build Hadoop-trunk-Commit #10703 (See https://builds.apache.org/job/Hadoop-trunk-Commit/10703/)
~~MAPREDUCE-6541~~. Exclude scheduled reducer memory when calculating (naganarasimha_gr: rev 060558c6f221ded0b014189d5b82eee4cc7b576b)

- (edit) hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app/src/test/java/org/apache/hadoop/mapreduce/v2/app/rm/TestRMContainerAllocator.java
- (edit) hadoop-mapreduce-project/hadoop-mapreduce-client/hadoop-mapreduce-client-app/src/main/java/org/apache/hadoop/mapreduce/v2/app/rm/RMContainerAllocator.java

People

Assignee:

Varun Saxena

Reporter:

Wangda Tan

Votes:

0 Vote for this issue

Watchers:

13 Start watching this issue

Dates

Created:

06/Nov/15 21:47

Updated:

06/Jan/17 08:09

Resolved:

27/Oct/16 12:43