Infrastructure Meeting 2017-05-25

Infrastructure meetings take place every other Thurs. at 9:00 Pacific on the BlueJeans infrastructure-meeting channel: https://bluejeans.com/383721668

Date

25 May 2017

Attendees

- ☐ Hsin-Fang Chiang
- ☑ Paul Domagala
- ☑ Gregory Dubois-Felsmann
- ☑ Igor Gaponenko
- ☐ Fabio Hernandez
- ☑ Brian Van Klaveren
- ☐ Simon Krughoff
- ☐ Kian-Tat Lim
- ☑ Fritz Mueller
- ☐ Donald Petravick
- ☐ John Swinbank
- ☑ Xiuqin Wu

Goals

- Ensure successful use of the current NCSA infrastructure
- Plan for near- and medium-term activities

Discussion items

Item	Who	Notes
Review of last meeting notes & action items	Unknown User (pdomagala)	• Any updates

PDAC cluster master node performance issues	Igor Gaponenko	• See attachments below
		• I've submitted a ticket on this so it can be assigned & tracked
		• two basic questions:
		  • troubleshooting: cause, mitigations & solutions
		  • mode of operation: how would we deal with such a problem in a production situation?
		
		Notes:
		
		Much of the i/o handled by root file system on the master node which is too small and too slow
		
		Prefer SSDs since performance scales linearly.
		
		Will be ramping up utilization in the June/July timeframe. More people to access Wise data.

Role of Nebula in prototyping	Simon Krughoff	
PDAC Status	Gregory Dubois-Felsmann	
Topics for next meeting		

Action items

Please enter action items in the form

Responsible Person, Due Date, Description

- [x] Fritz Mueller: send LSST Nagios info to Unknown User (pdomagala)
- [x] Unknown User (pdomagala): cross-walk Nagios instances and plan consolidation
- [x] Unknown User (pdomagala): establish monitoring working group/page: gather use cases & needs

Attachments

SLAC conversation re. lsst-dev i/o perfomance problems

Igor Gaponenko [3:29 PM]
@channel I'm not sure where should I post this complain, in this forum or in #dm-infrastructure. A problem is that the only filesystem we have in the PDAC *master* node *lsst-qserv-master01* has really horrible I/O performance. I wouldn't worry much about it unless the very same file system was not shared by the OS and *Qserv*'s MySQL/MariaDB database server. This setup bites us in two ways. Firstly we're using this file system (via the database server) to store intermediate results reported by *worker* nodes before doing the result set aggregation. In some cases the result sets could be rather large (a few *GB* per query). Secondly, the database service provides a number of key catalogs, some of which could be rather large (like the so called *secondary index*). The current disk subsystem of the node is just no match to those tasks. For example, when I'm scanning one of the *secondary index* (just to count the number of entries) then I'm seeing:
```iostat -m 1

avg-cpu: %user %nice %system %iowait %steal %idle
0.60 0.00 0.12 0.07 0.00 99.20

Device: tps MB_read/s MB_wrtn/s MB_read MB_wrtn
sda 437.00 14.94 0.01 14 0
dm-0 437.00 14.70 0.01 14 0

avg-cpu: %user %nice %system %iowait %steal %idle
0.40 0.00 0.05 0.27 0.00 99.28

Device: tps MB_read/s MB_wrtn/s MB_read MB_wrtn
sda 363.00 13.31 0.05 13 0
dm-0 364.00 13.31 0.05 13 0
```
NOTE how low is *BOTH* the CPU utilization and the disk I/O (for both IOPS and MB/s) . This looks just horrible. s there any chance we could add thw second file system based on 4 SSD in the RAID10 (0+1) configuration? That should't be super expensive. Four 0.5 TB disks would cost a couple of thousand. And this must be be the software-based RAID to allow the TRIM-ing (if that's still a problem for the newest SSD disks). If we could put the NVMe disk then it would be even better.