SIR on data loss in ASGC in Oct. 2016

Description

On October 18th, 2016, one of the storage arrays in ASGC started to misbehave. It suffered multiple hard drive failures at the same time. After we took some actions, most partitions in the storage recovered, but still one partition (one RAID) remained down and inaccessible. We called the engineer from storage vendor for help. The storage vendor engineer tried to force online this RAID and secure the data, but ended up in failure.
As a result, we lost all data in this partition and had to declare data lost.

Impact

There were about 135k files, or about 20TB lost in total. Most lost files belonged to ATLASSCRATCHDISK (~130k files) and ATLASDATADISK (~4k files), and the few rest belong to other spacetoken in ATLAS as well. Thus, this incident caused an impact on ATLAS production and user analysis.

Timeline of the Incident

Time (UTC)	Incident
18/10/16 20:07	One DPM disk server became unstable. ATLAS jobs started failing on site Taiwan-LCG2 and TW-FTT in ASGC.
19/10/16 03:15	Problems on the backend storage of the disk server were identified. Resetted the storage array controller and tried recovering RAIDs.
19/10/16 08:21	Most of the RAIDs were recovered and set online, but one of them remained unavailable due to multiple hard drive failures.
19/10/16 08:30	Escalated the problem to storage vendor.
24/10/16 03:27	Vendor engineer came and tried some recovery procedures.
24/10/16 07:16	After some tries and diagnoses, vendor declared that there were severe physical failures, so the RAID could not be properly set online.
24/10/16 07:24	Reported this situation to ATLAS. Started deleting metadata of lost files on our DPM. Started draining data on this storage array to other storage arrays.

03/11/16 10:01	Finished deleting metadata of lost files on our DPM.
04/11/16 10:00	Started data loss declaration to ATLAS DDM.
07/11/16 11:00	Finished data loss declaration to ATLAS DDM.

Follow-up

The storage array containing the lost partition has been considered unreliable since this event. Thus, we are now draining data from this storage array to the others, and will decommission this storage array.