Understanding the Purge Threat at Peta-Scale

Alexandra Holloway <fire@soe.ucsc.edu>
Jack Baskin School of Engineering • University of California Santa Cruz

The Purge Threat: A major usability problem

When the peta-scale parallel storage system fills up, the system needs to make a decision about which files to keep and which to discard.

This decision is transferred to the users.

If users do nothing, the system implements a least-recently-accessed policy, and deletes the oldest files.

Interviews

We conducted semi-structured interviews with 17 scientists (developers, researchers, users, consultants) at two national laboratories (LANL and LLNL) in July to September 2011.

① Scientist runs a simulation or job, creating tens of thousands of files on the parallel system.

② Scientist imports select results for processing and visualization.

Parallel storage

Huge

No backup

Fills up often

Transfer

Small storage (NFS)

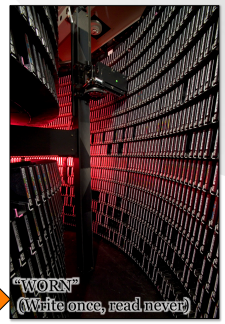
Backed up nightly

③ (Cautionary) Scientist decides which data to archive to tape. (Reactionary) Scientist does nothing until purge threat.

Protect against system crash, mediate unanticipated data loss

Cautionary archive

Archival storage (tape)



Why do we archive?

Motivation for archiving

1) Need to keep all data that resulted in a decision
2) Parallel storage system is not backed up
3) Parallel storage system runs out of space

Interfacing with the Purge Threat

Publication
On purge threat, a long list of affected files is published. People don't read this list.

Gravity
The user is not informed about how serious the purge threat really is. Important files are mixed with unimportant files in a long list of file names.

Scope
It is difficult to determine the scope of the threat. People may miss their files among other peoples' files in the published list.

Disk fills up: Purge Threat

Notify users

Guarantee no data loss as result of purge

Reactionary archive

④ System triggers a purge when almost out of space. System notifies users to read the list of files affected and decide whether to keep the files. This is the **Purge Threat:** the threat of having files deleted.

⑤ Decision-making surrounding the purge threat.

P U R G E

Decide: Do nothing

Purged data

Strategies for addressing the Purge Threat

Participants' methods

1) Analyze and move manually to tape;
2) Use a script to move files automatically to tape;
3) Refresh the timestamp on files with touch (postpone the problem and put colleagues' files at risk).

Conclusions

1) The system is difficult to use, resulting in deliberate subversion.
2) Moving files manually requires awareness of important files; fosters understanding that archival storage is expensive.
3) Keeping files local (not in archive) is preferred because archival storage is very slow.
4) It is difficult for users to determine the extent of the purge threat.

Office of Science U.S. Department of Energy

Acknowledgements: This material is based upon work supported in part by: the Department of Energy under Award Number DE-FC02-10ER26017/DE-SC0005417, the Department of Energy's Petascale Data Storage Institute (PDSI) under Award Number DE-FC02-06ER25768, and the National Science Foundation under awards CCF-0937938 and IIP-0934401 (I/UCRC Center for Research in Intelligent Storage).
Photo credit: info300.net/TapeStorageRobot.jpg