Monday, November 2. 2009

PSARC 2009/479 zpool recovery support

There was another important addition to the Opensolaris code base. PSARC 2009/479 zpool recovery support. Perhaps it isn't a feature for the front pages, because such a feature would be futile for ZFS in a world where all devices adhere to standards and definitions as they should.

The problem: Especially with cheap hardware you have a large amount of strange and interesting effects like weirdly reordered commands or commands that should only return after completion of a task, but return directly after accepting the command. Or as the PSARC states:Uncooperative or deceptive hardware, combined with power failures or sudden lack of access to devices, can result in zpools without redundancy being non-importable.The code resulting out this PSARC case implements automatic ways and mechanism to discard the last transactions to a zpool to get back to an uncorrupted, thus importable state. You were able to do similar things manually in the past, but that was a somewhat arcane science. Due to the nature of ZFS as a copy-on-write filesystem, there is a probability, that you can import the filesystem at a point in time slightly before the corrupt state. So executing this command should give you an output like this one: # zpool clear -F data
Pool data returned to its state as of Tue Sep 08 13:23:35 2009.
Discarded approximately 29 seconds of transactions.As your last backup is often much older than a few seconds, this is a nice way to handle unimportable zpools due to "uncooperative or deceptive hardware". As the complete structure of the file system can be validated by the checksums, you can even be sure that this recovered state is a consistent state.

Posted by Joerg Moellenkamp in English, Solaris at 15:32

Is there a chance to see this as putback in Solaris 10 ?
Would be a great improvement.
Anonymous on Nov 3 2009, 08:09

Agreed, this is the main feature I've been looking for before using ZFS for data I 'care' for.
Anonymous on Nov 6 2009, 16:54

Well ... when you really care for your data, you should get rid of subsubsubcomponents, that need this functionality And manually this procedure is possible since the introduction of zdb.
Anonymous on Nov 6 2009, 17:38