

Date: January 1, 2006



RE: What is Continuous Data Protection (CDP)?

Continuous Data Protection (CDP) is a backup concept that has received a lot of attention in the technical press recently. CDP has been around for a long time and largely ignored until two (2) large players (Symantec/Veritas and EMC) have stirred up the market with products touting their CDP capabilities.

But what is CDP really? It is a method or process of performing data backup and in many cases it is a feature of an existing backup product. To provide "Continuous Data Protection," a backup program must do two things:

1. Backup data as the data is written to the hard drive.
2. And, the backup program must keep revisions.

What is the goal of Continuous Data Protection?

The goal of Continuous Data Protection is to constantly backup data as the data is created or modified. So, a true CDP solution eliminates the possibility of losing data between backup sessions because in theory, no time goes by between the backup sessions. A backup is always occurring to proactively protect all new data.

What are the problems that CDP solves?

Let's look at what happens when you perform a single backup every night at midnight. Employees come in and create and modify data throughout the day. All of the work that has been performed during the day is at risk until it is backed up at midnight. What happens if you have a hard drive failure at lunch time? In this situation, all of the data created this morning is lost because you can only restore to the point of the last backup.

CDP solves this problem by backing up the data on-the-fly; as the data is created and changed. So, in this scenario, the work during the morning would be recoverable because the CDP backup features would be capturing the data throughout the morning.

What are the problems that CDP creates?

Using a CDP based solution creates several issues that can become large problems in a relatively short time frame.

Consider the following before choosing to implement a CDP solution.

Increased Load on Resources

The first issue that comes up is the increase in network and individual computer resource load. Since all data that is written to the hard drive has to be backed up to a storage device at the same time; you have effectively doubled the load on the computer running the CDP solution. In other words, every time data is written to the hard drive a backup event is occurring. An increase in the amount of data written to disk results in an equal increase in the load put on the computer because this simultaneous backup process is running.

It is important to note that the computers that have the greatest need for CDP tend to be those computers that are under the greatest load demand to begin with. The increased load caused by the CDP process may overwhelm these computer resources, resulting in a disruption in their normal operation.

Difficulty with Revision Management

Another issue that you can encounter with a CDP based solution relates to file revision management. Let's assume you have a database that is continually updated. Since this database is continually changing, the CDP system is backing up a near infinite amount of little changes, as they occur. At some point, these minute changes need to be pulled together to form a revision of the database

in order to fulfill the requirement of point-in-time restoration.

But at what point do you assemble these small changes to represent one, single revision? Let's assume your CDP solution creates restore points every 10 minutes. The question becomes, "How many restore points can be stored given our current storage capacity?" If 25 revision points may be stored, then your system will allow you to restore to any point in the last 25 revisions. These revisions are created every 10 minutes, which covers a little over 4 hours. Anything beyond the 4 hour time frame is lost. So, if your database becomes corrupted, you have a 4 hour window until every revision held by the CDP backup system is also corrupted; eliminating your ability to recover a clean copy of your database.

What is Near CDP?

You have probably heard of the term "Near CDP." Near CDP is defined as a backup solution that performs the backup process when a file closes. The key difference between CDP and Near CDP products is that True CDP backs up all of the changes to files as they occur whereas Near CDP performs one backup at the time the file is closed and editing is complete.

There are 2 key advantages that Near CDP products have over True CDP products. The first advantage is that with Near CDP products you have a significant reduction in load on the computer being backed up. The second advantage is that the problems with True CDP's revision management are cleared up because the revisions are clearly tied to file close events.

But, there are problems with Near CDP solutions. The single largest problem with Near CDP occurs when a file is continuously edited and never closes. For example, Microsoft's Exchange Server never closes the Mailbox Stores. As a result, the Near CDP backup trigger never occurs because the files are never closed and therefore they are never backed up. Another example of files that are continuously in use is seen in MS-SQL files. Again, these files are ignored by the Near CDP solution because they never close to trigger the backup process.

As a work around, the Near CDP solutions also carry a backup scheduler to backup data that is overlooked by the file closing

trigger. So effectively, the Near CDP solution operates as a traditional scheduled backup program does and you do not enjoy any of the CDP-related benefits for files that are continuously in use.

It is important to point out that Near CDP solutions can introduce a disruption in the normal operation of your production servers. Since the Near CDP backup is triggered by a file close event, if an application closes a file, and then re-opens the file, that application may be denied access to the re-opening of the file. This can happen because the first close event triggers the Near CDP solution to start the backup and then your request to re-open the file is denied because the backup program has locked the file during the backup process. For example, you may update data in an accounting program and when you post your update, the accounting program may save its files and then generate a report based on your new information. The report production may be disrupted because the Near CDP backup program has gained access to the account program data.

Compare CDP to other backup methods.

Let's compare CDP backup solutions to 2 other popular data protection methods: schedule-based backup and mirroring.

CDP and Scheduled File Backup

We've mentioned that True CDP operates on-the-fly, so therefore it is always working and does not follow a schedule to initiate the backup process. Both True

CDP and Near CDP solutions create problems of their own when it comes to backing up files that are in continuous use. Schedule based backup overcomes these issues by allowing you to determine when and how frequently you want to backup. And schedule based backup programs can more clearly correlate backup events with revision management.

CDP and Mirroring

It is important to note that while both mirroring and backup are both data protection methods, they are not one and the same. Mirroring refers to the process of making a duplicate copy of data to a second storage device. The key difference between mirroring and backup is that mirrored data is a straight copy of the original files and backup programs tend to encode the stored copy using some combination of encryption and compression which necessitates using a restore program to gain access to the backup data. Generally speaking, mirrored data is a copy of the raw data which is immediately ready for use and does not require you to use a restore program to gain access to it. Mirroring differs from CDP because mirroring typically does not contain revisioning. In a mirrored environment, when an original file changes, the mirror copy is overwritten with the new file. In a CDP environment, the new file is added to the revision history and the data may be encrypted and compressed before it is stored.

When is CDP interesting and when should it be avoided?

If you have been reading all the latest industry news about CDP, you may be attracted to these solutions because you want to minimize the amount of data that can be lost since the last backup occurred. If your business has a critical need to protect data on-the-fly, as the data changes, you may consider a CDP solution to be the answer. But, be aware that the CDP solution may also introduce new problems; such as increased load on network resources and the shortening of the time window to restore clean data in the event of corruption.

It is very common for businesses to implement multiple data protection (backup) solutions to provide a greater level of data

protection. Most companies find that one solution does not fit all of the needs of their computing environment. For example, you may have a few servers for which CDP makes sense and the rest of your servers are best protected with traditional schedule based backup. And some businesses implement more than one backup solution to protect the same computer; which provides them with several ways to restore data.