

[Home](#) > [iSGTW - 28 April 2010](#) > Feature: Back to Basics - Data Management

Back to Basics – Data management



by TRACEY WILSON

Tracey Wilson is a program manager for Avetec's HPC Research Division, DICE, the Data Intensive Computing Environment. DICE is a non-profit program that serves HPC and IT data centers in commerce, government and academia by providing independent third-party evaluations of data management practices on its geographically distributed test bed.

From smartphones to weather forecasts, data drives our world.

In the simplest terms, data is information, which can be found in many forms. Back when computer scientists used punch cards to store data, keeping them in order was a way of controlling or managing data.

We used to talk about how incredibly large a gigabyte of data was. Now terabytes and petabytes are the norm in scientific circles, and we will soon be talking about exabytes just as casually. And as the amount of data we generate grows, so does our need for data management.

Data management — the effective control of data throughout its entire lifecycle — is

getting more important every day as systems become increasingly complex, software demands increase, and the size and number of data files generated continues to grow exponentially.

File Systems in Data Management

When choosing a storage system for electronic data, you have to choose a physical medium and a logical medium. The physical medium could be a disk, a hard drive, or [tape](#). The logical medium is the file system itself, and there are many types. The most common types of files systems are:

- Local file systems used on workstations or servers — Windows NTFS or Linux GFS
- Shared file systems for large data repositories or archives — Sun SAM-QFS, Quantum StorNext, or SGI DMF
- Distributed file systems — SGI CXFS
- Parallel file systems — Lustre, IBM GPFS, or Panasas PanFS

Different mediums lend themselves to different purposes, based on their characteristics. For example, a very fast parallel file system is something you'd use for fast access by several clients at once and is typically seen on high-end or high performance computing systems.

On the other end of the spectrum, archive systems are used for longer-term storage and vary in performance. Today, they are typically divided into tiers by storage duration. For example, tiers may include:

- Tier 0 – Solid state disks. This provides the highest level of performance and would be most appropriate for data to which you need immediate and regular access.
- Tier 1 – Hard drive disks (SAS) and fiber channel disks. These have an expected lifetime of 45 days.
- Tier 2 – Higher performance serial ATA drive (SATA). This is a cheaper disk storage pool where data may stay for several months.
- Tier 3 – Lower performance SATA. Data may stay here longer than several months.
- Tier 4 – Long-term storage on tape or virtual tape library with high capacity. This is the longest-term option, where data is written once and rarely read.

Locality, Movement, Integrity, and Manipulation

There are four main areas within data management that are of major concern:

Data Locality — Locality is about the ability to archive and access data from different locations even though data may be stored at one primary location; this makes large remote scientific data sets as easy to access as if they were in a local file system. Questions that administrators have regarding locality include: Is data stored in a way that is logical and accessible? How is the data viewed from a local system?

Data Movement — This is the ability to move data efficiently and reliably to geographically dispersed systems and locations. What path or mechanism can effectively move data between locations? Can the data be sent in one large transfer stream or can it be broken up and sent in several streams in parallel? Do I need to worry about encryption and how do I verify integrity of the data once it arrives?

Data Integrity — This refers to the ability to maintain the quality and security of data during

[iSGTW 28 April 2010](#)

[Feature - Q&A: Peer-reviewed physics, at the speed of light](#)

[Back to Basics - Data management](#)

[Opinion - What would Linnaeus do?](#)

[Link of the week: Did you know?](#)

[Image of the week: Earthquake comics](#)

[Announcements](#)

[Tell us your travel tales coming back from the User Forum, win a T-shirt](#)

[e-IRG releases new Roadmap on e-Infrastructure](#)

[HD videoconference, Internet2 Meeting, DUSEL, 28 April, North Dakota, USA](#)

[Open-source release, HUBzero, Purdue University, Indiana, USA](#)

[VPH2010: Extension of Call for Abstracts to 17 May, Brussels, Belgium](#)

[Jobs in grid](#)

[Subscribe](#)

Enter your email address to subscribe to *iSGTW*.

[Unsubscribe](#)

[iSGTW Blog Watch](#)

transfer, access or storage is paramount. Once you move data, access it, or copy it, is it the same as before? Has it been corrupted or changed beyond your ability to recover?

Data Manipulation — The capability to change, search and manage data in local and distributed environments allows users to get more use out of their data. There may be better ways to search and modify data using metadata (the information that describes the data itself).

Day-to-Day Aspects of Data Management

The daily management of data may not be outwardly recognized as the key to an experiment's success, but effective control of this precious commodity requires careful administration.

Responsibilities vary greatly among data center employees. Some administer and maintain the integrity of file systems so that users will be able to access their data. Archive administrators do the same, and also have to determine how fast that data is growing and stay abreast of data locality, movement and storage trends. In addition, they must identify bottlenecks and decide how to purge data or expand infrastructure.

For example, if doing remote site transfers, the organization may be restricted to conducting backups at night. If local backups are the norm, the backup processes may have an impact on the ability of others to conduct their work by restricting access to files. Deciding how to balance these needs and concerns is part of an archive administrator's job.

Good Data Management Means Better Use of Information

Good data management helps data centers maintain better control. Users can access and move data more effectively when structure, data flow and size are correctly set up for performance. Benchmarking networks and file systems can help diagnose problems and provide keen insight into optimum performance tuning parameters.

As long as the race continues to increase computing power, data management will hold more and more of our data centers' attention as we seek to gain more control over the information we gather.

For more about data storage, see the iSGTW Nature Networks Forum [discussion about magnetic tape](#).

Tags: [Americas](#) [Cloud computing](#) [Data management](#) [Feature](#) [Grid computing](#) [Supercomputing](#)

Share this page:



[Email this page](#)



[digg.com](#)



[reddit](#)



[del.icio.us](#)



[StumbleUponMa.gnolia.com](#)



Disclaimer:

These are external Web sites and iSGTW cannot guarantee their security nor endorse their content.

[Keep up with the grid's blogosphere](#)

[Mark your calendar](#)

April 2010

19-23, [Cloud Lab 2010](#)

19-23, [IEEE IPDPS](#)

20-22, [BioIT World Conference](#)

23-24, [Health e-Child](#)

May 2010

3-4, [CCV 2010](#)

3-7, [NorduGrid](#)

10-12, [DEISA-PRACE Symposium](#)

[More calendar items...](#)