Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER
Navigating Information Technology Horizons℠
New in 2007 Report #TCG2007042 March 10, 2007

Data Copying – A Toolbox of Business Solutions

Analyst: Michael Fisch

Management Summary

A toolbox contains a variety of tools that serve different purposes. Consider a wrench, hammer, and screwdriver. They are similar in that they affix implements – bolts, nails, and screws – that hold objects together. Workers use all three of them in constructing the walls and windows of a house. At the same time, the tools are distinct in shape, size, and the respective tasks they perform. No one would use a screwdriver to drive a nail, nor would a hammer have much luck with a bolt!

The same is true among data copying solutions (think in the broadest sense of this term). **These are tools that solve business problems by copying and replicating data, but they differ significantly in *how* and *for what purpose*.** No tool does everything, though some do multiple tasks. An enterprise must first know what it wants to accomplish. These business objectives related to data copying fall into four categories:

- **Business continuity and data protection –** Keep the business running by ensuring information access.
- **Data repurposing –** Extract more utility from information assets.
- **Data migration –** Manage information dynamically over its lifecycle.
- **Data distribution and consolidation –** Move information to where it is best put to use.

With specific objectives in mind, the next step is to understand the available technologies. Data copying solutions divide into several categories:

- **Remote mirroring**
- **Point-in-time copy**
- **Continuous data protection**
- **Automated data migration**
- **Data copy and migration**
- **Backup and restore**

Finally, match the right technologies to your business objectives. You may be surprised at what these tools can accomplish – there is more to copying than first meets the eye. Read on for details.

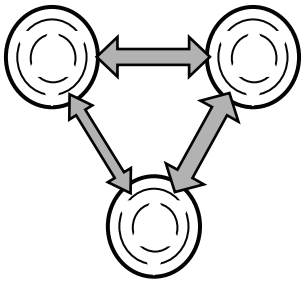
IN THIS ISSUE

The Business Objectives

Data copying is a tool for solving business problems. Consider the basic act of copying or replicating digital information. It is at the heart of a variety of IT solutions. In turn, these solutions are an integral part of business operations. Like anything else an enterprise employs – a building, patent, or advertising campaign – they are a means to accomplish certain objectives.

Therefore, in considering data copying solutions, the first question to ask is what an enterprise wants to accomplish. The answer will determine the most appropriate solution. Data copying actually supports a variety of enterprise activities:

Business Continuity and Data Protection



Business continuity and data protection go hand in hand. Consider when workers go on a strike; business activities stop. When a factory is flooded, business activities stop. And when information is unavailable or lost, business activities stop. Land, labor, and capital are no longer the exclusive economic inputs in our modern digital world. Business runs on information too.

For instance, an enterprise must be able to process a transaction to take an order. Workers must be able to send and receive e-mails to communicate with each other and with customers, suppliers, and partners. Accounts receivable must be accessible to collect bills. **The continuity of business operations depend on information access, and information access depends on effective data protection and recovery.**

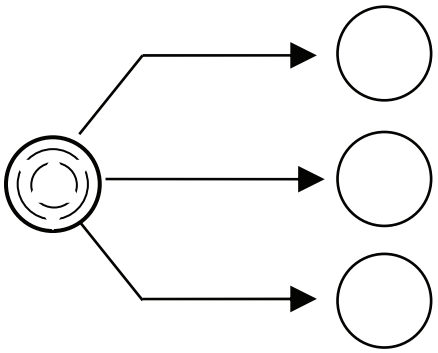
At the heart of business continuity is data copying. Extra copies of data function like the spare tire in the trunk of a car. If one fails, another is ready to take its place. In the real world of IT, systems fail, media fails, and human operators make mistakes. Data copying protects against these calamities. In the case where disaster strikes an entire location, even multiple local copies are not enough. The remedy is to store data in multiple locations, preferably a long distance away from each other. **Redundancy plus remoteness equals resilience.**[1]

When designing a system for data protection and business continuity, there are three principal factors to consider:

- **Recovery time objective (RTO)** – How long it takes to recover from a failure. This is measured in minutes, hours, or days.
- **Recovery point objective (RPO)** – How current or up-to-date the recovered system will be. Another way to look at it is how much data loss the enterprise can tolerate.
- **Resilience** – How many failures a system can sustain without suffering data loss. Multiple copies are resilient, but multiple copies in multiple locations are more so. Redundant components and systems and more-durable media are also factors.

The nature of a business will determine the appropriate level of continuity or, conversely, tolerance for downtime. For instance, a Wall Street securities broker handling hundreds of millions of dollars in transactions per hour would need the most robust solution. A bank branch office would have less-stringent requirements.

Data Repurposing



All data serves a purpose, and it can serve more than one. **The primary purpose of a**

[1] See *Business Continuity Goes Better With SANs – The 3 Rs of Resilience* in **The Clipper Group Explorer** dated January 25, 2003, at http://www.clipper.com/research/TCG2002003.pdf.

data set is to support the application that normally accesses and probably created it. For instance, databases, file servers, and messaging applications each have associated volumes that contain their data. This application-to-data relationship is normally exclusive, but what if other applications could make productive use of it? **Data repurposing is the term for making data available for alternative uses.** It allows an enterprise to extract more utility from information. In other words, data repurposing helps maximize the return on information.

An important alternative use for data is application testing and development. When enterprises introduce new or revised IT applications, they need to be sure the applications work without glitches. Otherwise, they may encounter problems and slow business operations after the rollout. The best way to ensure success is to test and develop with real enterprise data.

Another use is business intelligence and data warehousing. Better information produces better decisions and a better bottom line. Enterprise applications contain a goldmine of information about sales, customers, markets, finances, operations, etc. Business intelligence and data warehousing solutions mine it with sophisticated data collection and analysis techniques to assist in decision making.

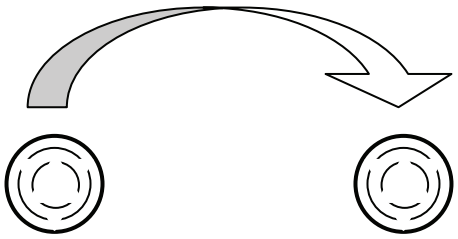
Running simultaneous processes against a data set is another possibility. If time is of the essence, an enterprise may expedite the analysis of a data set by running multiple, simultaneous processes against it. For instance, an energy company has seismic field data and wants to determine quickly if the site is worthy of further exploration.

These alternative applications normally need to run against a copy to avoid corrupting the original or lowering availability of the primary application. Hence, **copying is an indispensable part of data repurposing.**

Principal factors to consider for data repurposing are:

- **Copy speed** – The amount of time required to create a copy. This depends on the size of the data set as well as the copy technology used. It can be virtually instantaneous or take minutes or hours.
- **Primary application performance** – The effect on the performance of the primary application. This depends on the nature of the secondary application and the copy technology used.
- **Persistence** – The amount of time that the copy needs to exist. This may be temporary or indefinite.
- **Access level** – The level of data access required, in terms of read-only or read-and-write access.
- **Resilience** – The level of protection applied to the copy, such as RAID[2].

Data Migration



Think about how businesses move around goods and other assets. Retail store chains move inventory from suppliers to warehouses to shelves in individual stores. Parcel post services move packages around a country, from sender to receiver. Banks move money around the world. Likewise, **most enterprises move data during the course of operations.** They migrate data from one storage platform to another, which involves a form of copying. This may seem mundane, but if done well, it can make an enterprise more productive and efficient.

There are several reasons for data migration, starting with equipment upgrades and load balancing, both routine tasks. IT equipment typically has a lifespan of three-to-five years, after which enterprises will upgrade or replace it. Part of this process is to move data from the old platform to the new with as little disruption as possible. This is not easy, like replacing the foundation of a house without disturbing the occupants. When an enterprise has multiple shared-

[2] RAID (redundant array of independent disks) describes various techniques for writing data on disks to enhance performance and availability.

storage platforms[3], data can also become unevenly distributed, which lowers overall utilization. Some platforms are too full; some are not full enough; and only some are "just right". Load balancing is the process of intelligently migrating data between platforms to even out the distribution, increase utilization, and get more out of storage assets.

Data migration is also an integral component of information lifecycle management (ILM)[4]. ILM is a concept for dynamically managing data over its lifecycle, from creation to deletion, as its value changes over time. It takes a holistic, long-term view of information and seeks to optimize the balance between storage service levels[5] and cost. These are competing elements since, as in everything, better service costs more. Intelligent data migration among different storage tiers is the primary means to accomplish ILM. The benefits include cost containment, data retention for operational and regulatory requirements, better access performance of primary storage, and faster backup, restore, and replication.

Principal factors to consider for data migration are:

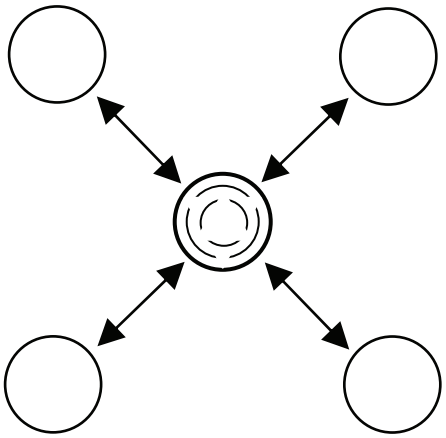
- **Transparency** – The ease of access to migrated data. If the client user or application can access migrated data without local reconfiguration, then it is transparent. If reconfiguration is required, such as updating share mappings for a file system, then it is not transparent.
- **Non-disruptiveness** – Data availability during the migration. Data is traditionally inaccessible during migration, but special tools are now available that allow continuous access.
- **Automation** – The ability to migrate data automatically based on programmed policies. This is a critical enabler for ILM and, to a lesser degree, for capacity load balancing.

[3] SAN, NAS, or DAS.

[4] See *Top 10 Things You Should Know About Information Lifecycle Management* in the issue of **Clipper Notes** dated March 10, 2007, available at http://www.clipper.com/research/TCG2007039.pdf.

[5] Describes the quality of storage applied to a data set in terms of performance, availability, and recoverability.

Data Distribution and Consolidation



Data distribution and consolidation are similar to migration, but they involve copying data to other locations without erasing the original. The above examples of the retail store chain, parcel post service, and bank also apply here. The reason for moving inventory, parcels, and money is to provide local access. For instance, a toothbrush at a warehouse is of no use to a customer who wants to purchase it. It must be on the store shelf. A letter is no use to the recipient unless it eventually arrives in his or her hand. Money is of no use to a customer unless he can withdraw it locally and spend it. In a similar way, distribution makes data more useful by providing local access.

While it is possible to access data remotely without first distributing it, the process is typically slow. The lower bandwidth of remote data links combined with transmission delays over distance (even the speed of light is a limitation) can slow real-time access to an unacceptable crawl. Again, business operations depend on timely access to information. A solution is to copy the data to local storage, like downloading a file over the Internet, where applications and users can access it quickly.

This issue is especially prevalent in enterprises with multiple sites separated by tens, hundreds, or thousands of miles. Data can reside anywhere, and users may need to access it from anywhere. Data distribution to remote sites or consolidation from them helps to manage this dilemma. Specific situations where data distribution or consolidation is useful include:

- Sending documents and content for use at remote offices
- Consolidating transactions and reporting from remote offices
- Shared directories
- Distributed development teams
- Software updates on remote servers to ensure version consistency throughout the enterprise
- Delivering digital products to customers, such as software, documents, music, and video
- Exchanging information with suppliers, partners, and customers

Principal factors to consider for data distribution and consolidation are:

- **Coherency** – The degree to which data copies are up to date and synchronized. Coherency is always an issue when dealing with copies because the original and its copies are subject to change during the course of operations. Working with stale data, such as development code, financial statements, or price lists, can obviously cause problems.
- **Distribution speed** – How quickly data is copied to remote locations. This is a function of the amount of data, distance, network bandwidth, bandwidth efficiency, full versus differential copies, and number of locations.
- **Automation** – The ability to distribute data automatically based on programmed policies. Automation helps ensure data is locally available when users need it, rather than having to request it and wait.

General Factors

In addition to the above, there are some general factors to consider for data copying solutions. The first is the inescapable factor of *cost*. Since the purpose of copying is to solve business issues, one must have a business rationale for implementing and demonstrate a plausible return on investment. All forms of copying cost money – some more than others. Certain benefits are readily quantifiable and others, like business continuity, are more subjective but equally real. The bottom line is that an enterprise should be able to answer these questions affirmatively: *Is it affordable? Is it worth it?*

Security is also a critical factor.[6] Enterprises should take active and sufficient measures to guard against theft, loss, or alteration of data. This is true for storage and data management in general, especially when data is traveling over external communication links. No enterprise wants to jeopardize the integrity of its information or allow it to fall into the wrong hands. Techniques for security include encryption, authentication, compartmentalization, and integrity checks.

Finally, there is *manageability*. This encompasses the ease of setup and administration on an ongoing basis. Enterprise-wide, centralized management with a comprehensive, intuitive user interface and minimal administrator intervention is the ideal. Something close to that is good, too.

The Technology Solutions

With an idea of the business objectives you want to accomplish, the next question is which technology solutions will support and enable them. There are many to choose from – data copying solutions are like *Baskin & Robbin's* 31 flavors. Not only are there distinct technology categories, but product features vary greatly within a category and can span multiple categories.

Remote Mirroring

Remote mirroring solutions maintain a complete physical copy of data on disk at a remote site in real time. As the source data changes, so does the target. Like the spare tire in the trunk, if the source data becomes unavailable for some reason, a current or nearly current copy is available at the remote site to resume operations. It protects from local system failures or disasters, such as fires, floods, or electricity outages.

There are many remote mirroring features.

- **Synchronous and asynchronous** – Mirroring can be synchronous or asynchronous,

[6] For more details, see *Storage Security – What Are Your Vulnerabilities?* in the issue of **Clipper Notes** dated March 10, 2007, at http://www.clipper.com/research/TCG2007040.pdf.

depending on the level of protection, performance, and distance required. Synchronous mirroring provides the highest level of protection, ensuring that no data is lost by keeping the source and target "in sync". However, it can slow application performance and is practical over distances up to about 100 km.

Asynchronous mirroring provides a good level of protection over virtually unlimited distances without affecting application performance. The tradeoff is a lag of seconds to minutes between when writes are committed to the source and target. If there is a disaster, this data lag is exposed to loss.

- **Consistency and consistency groups** – Consistency affects how readily an application can restart from a copy. Think of a data set as having "loose ends" (i.e., writes in cache or a buffer that are not yet committed to disk), and this feature ties up the loose ends in the remote mirror. It is faster to restart from a consistent copy, involves fewer steps, and removes the risk of losing any partially-committed writes or transactions. Some applications involve multiple data sets, such as distributed databases and n-tier applications. ***Consistency groups*** **ensure that all volumes related to an application are consistent at a point in time, even if they are located in different storage arrays.**

- **Multiple arrays or sites** – Enterprises may need to mirror to or from multiple systems or sites. For instance, one may want to consolidate the replication of multiple servers or arrays into a single system at a remote site. In another case, an enterprise wants superior protection and chooses to synchronously mirror to a metro-area data center and asynchronously to a remote data center.

- **Server/storage integration** – The main purpose of remote mirroring is to quickly failover in the event of a disaster. This involves the application, server, operating system, storage, and data. An integrated mirroring solution increases the speed and ease of recovery.

- **Host server, array, or network based** – A mirroring solution may run on a host server, storage array, or SAN-based platform or switch. Each involves certain tradeoffs in the scope of replication, heterogeneous support, manageability, functionality, consumption of server processing, and cost.

Point-in-time Copy

Point-in-time (PIT) copy solutions create a copy of data on disk at a specific point in time. If a remote mirror is like a window offering a real-time view, a PIT copy is like a photograph that captures a scene at a fixed point in time. These copies are useful for data repurposing, non-disruptive backups, and quick recovery to a previous point in time.

Should a data set become corrupt due to a system or operator error or virus, administrators can quickly revert to a prior "clean" copy and use transaction logs to rebuild the data back to the present. This is called a *repair*, and it is useful to recover from small and not-so-uncommon hiccups. It is faster than recovery from tape and protects against types of failure (data corruption) that mirroring cannot.

There are many PIT copy features:

- **Full and snapshot** – These are the two basic types of copies. A *full copy* is a complete duplicate. Also known as a clone, it is readable and writable and consumes the same amount of storage capacity as the original. It takes a little time to create upfront (based on the amount of data) and may slightly degrade application performance during this process. A special feature called *instant copy* makes it available almost immediately while performing the copy operation in the background.

  The other PIT type is a *snapshot*, also known as copy-on-write or differential copy. It consumes less space than the original (say, 30%) and is available immediately. The tradeoff is that it is typically read-only and, therefore, suitable for fewer activities than a full copy. It is designed to exist for a limited amount of time; it consumes capacity as the original data set changes (i.e., copy-on-write). If it affects performance, it is on an ongoing basis as the original changes and data is copied out. A recent innovation delivers a cross between full and snapshot – a snapshot copy that is both readable and writeable.

- **Consistency and consistency groups** – Consistency ensures PIT copies are readily usable and restartable by the application.
- **Application integration** – Some applications also require special integration to deliver full, consistent copies with minimal disruption. Examples include Microsoft's *Exchange* and *SQL*.
- **Resilience** – If a copy is important enough to protect, applying RAID 1 (local mirroring) or RAID 5 (parity) guards against drive failure.
- **Host server, array, or network based** – Like remote mirroring, a PIT copy solution may run on a host server, storage array, or a platform or switch in a SAN. There are pros and cons associated with each approach.

Continuous Data Protection

Continuous data protection (CDP) is a newer technology that recreates data sets to virtually any prior point in time. Though some would not categorize it among data copying solutions, it does involve copying incremental changes to primary data as they occur. If mirroring is like a window and PIT copies are like photos, then CDP is like video. It effectively snaps a series of photos for rewinding back to virtually any point. As a result, a data set can recover quickly from corruption or viruses by rebuilding from just prior to the failure. The most recent replica can be near real time, approaching asynchronous but not synchronous mirroring. Today, this technology applies to a contained set of applications or file systems.

CDP features can include:

- **Host or networked based** – CDP solutions have three fundamental solution architectures:
  - Software that runs on an appliance or intelligent switch in a SAN,
  - An agent that runs on each application server and mirrors writes to a second storage platform, and
  - An agent that runs on each application server and sends writes to a recovery server. Agents will consume some server processing resources.

  Each architecture has respective advantages.
- **File, block, or message replication** – These solutions replicate data at the file, block/volume, or message level. Block level has the broadest applicability, and file and message level can have greater specificity for replication and restore.
- **Granularity of recovery** – The point-in-time granularity for a restore can be by the individual byte, block, I/O, or time intervals like seconds, minutes, or hours.
- **Local or remote** – All solutions replicate locally, and some can send data to a remote site for extra protection.

Automated Data Migration

Automated data migration (ADM) solutions move data based on policy between storage tiers to strike a balance between information access and cost. The goal is to place data in the right tier at the right time, as its value changes, to meet business requirements. The tiers are storage platforms or partitions with different price/performance characteristics and possibly media types (i.e., disk, optical, tape). ADM effectively implements ILM at the application or file system level by using metadata (information about data) to classify data and make decisions about when and where to move it. Furthermore, it maintains a link or association between migrated data and the application, so users can still access it. ADM is useful for lowering costs, meeting regulatory and business requirements for data retention and archiving, and improving the performance of a file system or application.

ADM features can include:

- **File, record, message, or block migration** – ADM solutions act on particular types of data, as no universal solutions yet exist. For instance, one may work with databases, another file systems, and yet another messaging applications. Some even act on raw volumes by looking at usage patterns on a block-by-block basis. Look for a solution that deals with the data types and applications that you want to address.

Business Objectives and Technology Solutions

	Business Continuity & Data Protection	Data Repurposing	Data Migration	Data Distribution & Consolidation
Remote Mirroring	X		X	
Point-in-time Copy	X	X		
Continuous data Protection	X			
Automated Data Migration			X	
Data Copy and Migration	X	X	X	X
Backup and Restore	X			

- **Richness of metadata** – Metadata is the raw fuel for policy engines. Richer and more descriptive metadata enables more sophisticated and effective migration policies. ADM solutions vary in the metadata they can leverage.
- **Non-disruptive migration** – Some solutions allow clients to access data during migration; others make data unavailable during the process.
- **Transparency** – This describes how easily and quickly users can access migrated data. It may be completely transparent and automatic for users; it may require users to take a special action; or it may require administrator intervention.
- **Index and search** – Tools for search and retrieval can greatly simplify the access and use of migrated data. This is useful for activities like legal discovery and audits, for instance.

Data Copy and Migration

Data copy and migration solutions move data between systems. This category is admittedly a catchall, but it has distinguishing characteristics. These solutions copy data at a point in time (leaving the original in place) or migrate it (removing the original after copying it). Multiple systems are involved; these are not copies within a single system, like PIT copy. The systems involved are either servers or storage platforms. Unlike ADM, data copy and migration solutions are not "application aware". They do not maintain an ongoing link or association between the copy and the application or file system that created it. Their purpose is to move data within an enterprise or between an enterprise and external entities, like suppliers and customers.

Data copy and migration features can include:

- **File or block copy** – Copies are at either the file or block/volume level.
- **Heterogeneous support** – Solutions may be limited to moving data between servers of the same operating system or storage platforms from the same vendor, or they may support heterogeneous operating systems or storage platforms.
- **Non-disruptive migration** – If clients and applications can continue to access data during a copy or migration, then it is non-disruptive.
- **Full and incremental copy –** All solutions make full copies, but some can also track changes and send incremental updates for subsequent copies.

Backup and Restore

Backup and restore systems are the tried-and-true backbone of data protection. Enterprises typically backup their data on a daily basis to tape, disk, optical media, or some combination thereof. It is available to restore data in case of loss or corruption.

Some also would not categorize backup and restore systems among data copying solutions, but they actually do copy and move data. As a matter of fact, backup systems protect the widest range of applications, operating systems, and data types and write to the widest range of media (disk, tape, optical), both local and remote. None of the other technologies listed here altogether replace it. However, **backup systems do not cover all enterprise data protection needs, and the other technologies are useful and even necessary for augmenting it.** For instance, combining tape backup with point-in-time (PIT) copies can dramatically improve the resiliency of an application.

A major trend in the backup arena is the increasing use of low-cost disk as a backup target. It speeds backups and restores, helping enterprises cope with shrinking backup windows and increasingly stringent RTOs. Reliability is also better. However, tape is still a solid media with the lowest cost and the advantage of portability, and it is not going away. In many cases, enterprises backup most frequently to disk and then later migrate data to tape for long-term storage and offsite archiving.

The Match

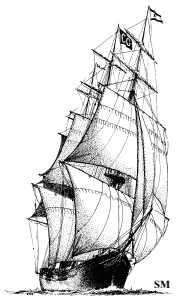
The final step is to match technology solutions with your business objectives. This is an iterative process. As you understand in more detail the business requirements, technology capabilities, and costs, there will be practical tradeoffs and modifications. The table above maps technology solutions with the business objectives they can support.

You may find it possible to do more than initially anticipated since data copying solutions can serve multiple purposes. *This solution is great for distributing files to remote offices, and it can send files back for centralized backup!* You may find that the perfect solution is beyond your means, but a good-enough alternative will suffice. *An ideal three-site mirroring solution may be too costly, but two sites plus tape vaulting is doable.* You may even find that one solution can meet several business objectives because it incorporates multiple technologies – hitting two birds with one stone, so to speak. *This solution does backup and restore as well as automated data migration.*

Conclusion

Data copying is essential. Like a plumber without a wrench in his toolbox, an enterprise could not get along without it. It supports a surprising number of enterprise activities – from enabling continuity to lowering costs to enhancing business productivity.

Data copying solutions are many and varied, in terms of both their capabilities and the problems they solve. So, take a close look at your business requirements and the available technologies. You may be surprised at what the latest tools can accomplish.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Michael Fisch **is Director of Storage and Networking for The Clipper Group.** He brings over ten years of experience in the computer industry working in sales, market analysis and positioning, and engineering. Mr. Fisch worked at EMC Corporation as a marketing program manager focused on service providers and as a competitive market analyst. Before that, he worked in international channel development, manufacturing, and technical support at Extended Systems, Inc. Mr. Fisch earned an MBA from Babson College and a Bachelor's degree in electrical engineering from the University of Idaho.

- ***Reach Michael Fisch via e-mail at mike.fisch@clipper.com or at 781-235-0085 Ext. 211. (Please dial "211" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **The Clipper Group Voyager,** **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.