



# *Storage Challenge*

By **PETER A. BUXBAUM**  
MGT CORRESPONDENT

# CAPACITY AND PERFORMANCE STORAGE GOALS CAN BE AT ODDS IN PRESERVING MASSIVE GEOSPATIAL DATA SETS.

If you digitize the information in every book in the Library of Congress, you have about 10 terabytes of information. (One terabyte equals 1,000 gigabytes.) Put every movie ever made on DVDs, and you come up with four petabytes of data. (One petabyte equals 1,000 terabytes.)

Four petabytes is also the amount of data the National Geospatial-Intelligence Agency (NGA) expects to be collecting annually in coming years.

“There are numerous sources that NGA gets data from,” said Sonny Finnegan, one of the co-leads for NGA’s Storage Services Integrated Product Team. “We have a big job ahead of us.”

The challenges NGA faces are two, according to Michael Ehman, chief executive officer of Cutting Edge Networked Storage, a provider of storage solutions. “One is the sheer volume of data being handled and where to put it,” he said. “The second is how to move the data around while getting performance out of your system.

“As satellites continue to provide better resolution, agencies like NGA don’t just want a picture of the earth, they want one every 15 minutes to see how things have changed,” Ehman added. “That, coupled with government security requirements, can accelerate these data sets into the hundreds of terabytes pretty quickly.”

Beyond that, intelligence analysts often want to scrutinize data as it is coming in. “In the geospatial data environment, there are very large files and a very rapid stream of arriving information,” said Robby Robbins, government/intelligence segment manager at SGI, a provider of high performance computing and storage solutions.

“Incoming geospatial data often must be tagged and made available for search and analysis in real time or near-real time,” Robbins explained. “A weak link in this process could cause a bottleneck of information to develop. Some software applications that are deployed by users of geospatial data require a high level of system performance. The applications must be able to quickly extract relevant data from storage, and either ship it back to storage when it is done or forward it to an alert system.”

Achieving the dual goals of providing adequate capacity for storage together with high system performance leaves agencies like NGA with a number of choices. First and foremost is the question of cost. An organization with unlimited funds might opt for the fastest, but least efficient and most expensive of storage media. That, of course, is an unrealistic option. Instead, speed must be balanced with efficiency—including the level of data density that can be handled by a particular medium—as well as other factors, such as the relative importance of different data sets to organizational operations.

## ENTERPRISE STRATEGY

NGA has not yet made all the decisions involved in considering these issues, but the agency does have a high-level vision of where it is headed. What NGA has decided to do is to take advantage of the Base Realignment and Closing (BRAC) process, through which it will be consolidating four data service centers into one, to move its data management to an enterprise solution.

“Instead of each entity having its own storage, BRAC is allowing us to start over,” said Finnegan. “It is providing us with a great opportunity to move to an enterprise storage solution.”

The ultimate system NGA is contemplating will be characterized by a tiered storage scheme. “What that means is that we have classified different types of data depending on the data users’ requirements for access to that data,” explained Finnegan.

An analysis of the hierarchy of NGA data resulted in four tiers of data, each of which will use different storage media. In descending order of priority, they are: hard drives; online storage; near-line storage such as disks; and off-line tape storage for archiving, backups and disaster recovery.

“Standardized solutions will be cheaper to run and to maintain,” said David Hopkins, who also serves as co-lead of the Storage Services Integrated Product Team. “It will be easier to scale systems as our storage requirements grow. We will also get a performance bonus on top of that. We will be able to be more responsive to our customers’ needs.”

The performance enhancement Hopkins referred to will be coming largely through the elimination of duplicate data and through the virtualization of data sets. In this context, virtualization means that the system will link users to a single source of stored data instead of having the data duplicated and stored on different systems around the network, thus freeing up storage capacity and enhancing the network’s performance.

## INTENSIVE IMAGERY

Capacity and performance are indeed the two watchwords when it comes to the storage of geospatial data, for Woody Hutsell, executive vice president of Texas Memory Systems, which makes solid-state disks and digital signal processing systems.

“We have observed two primary concerns,” said Hutsell. “One is that imagery data is very capacity intensive. The second is related to performance. Once acquired, agencies like NGA often have a simultaneous requirement to analyze the data. Viewing, searching and analyzing data contribute to

the requirements for the performance of the system and for storage.”

Texas Memory recently introduced a storage solution that attempts to balance speed, efficiency and cost-effectiveness by utilizing two kinds of solid state disks: RAM-based solid state media and flash storage media. “RAM-based solid state is the fastest storage medium, but it provides only low data density and results in a high cost per unit of storage,” Hutsell explained. “One chassis can accommodate 128 gigabytes. That would be almost nothing to someone working with geospatial data.”

Flash storage, although slower than RAM-based drives, offers a very high level of data storage density, and, at one-fifth the price per unit of storage, a cost-effective alternative to RAM-based storage media. That is why Hutsell says that a combination of flash and RAM-based storage could fit the bill for many geospatial applications.

“We use the RAM-based storage for caching,” he explained, “and flash as the other storage medium. Flash is inherently not volatile. You don’t need backup for the battery or the disk. It provides higher density and capacity at lower cost, but is not as fast as RAM. Certain geospatial data sets that are not frequently accessed could reside in traditional storage media. But if you have a hot spot of data that needs to be tracked, it could be loaded onto faster media.”

On the other hand, some applications will require a RAM-based solution, along with the higher price tag. “For some write-intensive applications, and those with a lot of online transactions or the need for many concurrent users, RAM will be the better fit,” said Hutsell.

As suggested by the NGA plan and Hutsell’s comments, storage considerations are intimately related to information management. Principles of information life cycle management (ILM), should inform storage investment decisions, said Randy Chalfant, chief technology officer at Sun Microsystems. ILM refers to a set of strategies for administering storage systems.

“The ideal situation would be to acquire a cheap, manageable, highly scalable, and an almost infinitely available capability,” Chalfant said. “It should be boundless in terms of scalability, manageable by as few people as possible, and secure.



Geospatial data, such as this overhead image generated by a combination of satellite and aerial views, requires massive storage capacity. [Image courtesy of SGI]

“Applying ILM principles and practices to storage management can lead to the discovery, recovery and reuse of a great deal of capacity,” he added. “Implementing systems that don’t scale well can lead to 70 percent of wasted capacity. The ability to reclaim capacity and put it back into the free pool for storage can be brought about through proper management practices. If you are collecting 15 terabytes of geospatial every day, the last thing you want is to field something that is going bump up against barriers.”

Chalfant suggested that an array of Sun Microsystems products, such as its high-capacity and high-performance servers, could be brought to bear to create just such a solution. Sun Microsystems technologies are all based on open standards, which boast greater reliability and lower costs, he contended.

## SPINNING DISKS

Software can be leveraged to maximize the capacity and efficiency of data stored on hard drives, noted Mark Weber, president of the U.S. public sector division at NetApp, a storage solutions company. NetApp is in the spinning disk business, Weber explained, and its systems can provide storage capacity exceeding one petabyte. But the software that the company provides is what maximizes the efficiency of its disk storage space.

“We play in the battle management area, and we have customers in the military and civilian intelligence agencies, including those dealing with geospatial data,” Weber said.

“Why put geospatial data on disks? Because tape is not fast enough. When you have a huge archive of imagery data, you want to make sure that it is reliably accessible.”

For Weber, the key consideration is deploying the right software to maximize the efficiency of storage media while assuring reliable access to data. “Without the software, the typical disk utilization is only 25 percent,” he said. “With NetApp software, we can get utilization up to 70 percent to 80 percent on the disk.”

Those facts have several implications. For one thing, it means that an organization can invest in a lower level of storage capacity. “It may seem counterintuitive since we sell storage capacity, but that is what we tell our customers,” said Weber. “With our technology they are able to drive efficiencies in their storage systems and to lower their total cost of ownership.”

The increased utilization of disks and lower investment necessary in storage capacity also yields savings in space, power and cooling requirements for storage apparatus, while providing environmental benefits at the same time, Weber noted.

NetApp also provides virtualization software, which limits the duplication of data by pointing users to an original stored copy of a data block. “This also maximizes the utilization and efficiency of disk space,” said Weber, “and also makes disk storage more cost effective.”

Cutting Edge Networked Storage has focused since the early 1990s in developing customized hardware storage solutions based on the open standards Linux operating system. “We use open standards hardware as well,” said Ehman. “We spend most of our energies on customized engagements for customers.”

Cutting Edge’s customers include international users of geospatial data. The company has also performed specialized projects for the U.S. Navy and Marine Corps.

Cutting Edge’s storage product line is an integrated solution that can accommodate as many as 48 drives supported by a single mother board. “We are working to get to 72 drives,” said Ehman. “This allows get a large amount of bulk storage on a network.”

The company also offers an enterprise class solution that links a network of servers for enhanced storage capabilities and can be optimized for throughput or capacity on a variety of different network pipes.

“Our products are designed to accommodate very large volume sizes,” said Ehman. “We’ve tested our volumes to several terabytes. That has been a good sweet spot for us in the market.”

SGI produces hardware and software that are purpose built to handle streams of large block data without dropping a single frame, said Kurt Kuchein, SGI storage manager. “The drives we are currently shipping can accommodate one petabyte. We will soon be launching a product that can accommodate one and a half petabytes in two racks. We have achieved greater density with these drives and also have achieved greater capacity by networking the drives. These systems are built for a real-time environment to get data in and out at a guaranteed speed and latency without data degradation.”

NGA officials, meanwhile, remain open-minded toward using commercial or open technologies. “Both proprietary and open technologies will be considered,” said Hopkins. “For NGA the storage solutions must be able to scale into the petabyte range.

“We don’t want to reinvent the wheel,” he added, “but NGA’s responsibility to maintain data for national archive purposes and DoD’s data management and security requirements do add a couple of layers of additional complexity to this project.”

Ultimately, the storage dilemma, from the military’s perspective, is not only about data, but about warfighting, experts say.

“When you are able to move data around faster, you are able to get information to analysts faster and they are able to complete their analysis faster,” said Robbins. “When the information in question is the location of a bad guy, this can mean the difference in the success or failure of a mission. When you are better able to make faster decisions, you are also able to save the lives of our own soldiers. This is not a trivial thing.” ★



Contact Editor Harrison Donnelly at [harrisond@kmimediagroup.com](mailto:harrisond@kmimediagroup.com). For more information related to this subject, search our archives at [www.MGT-kmi.com](http://www.MGT-kmi.com).