lexana\net

December 2007

Terabyte SneakerNet: The Carry-on Data Center

By Drew Hamre

I recently helped relocate a data center for a US government agency, moving system images from Virginia to a consolidated facility in Minnesota. During ‘cutover’ weekend, production systems in Virginia were dropped and remained unavailable until data could be refreshed at the Minnesota data center and the new systems could be brought online. To minimize downtime, the time allotted for refreshing data across locations was extremely brief.

email

This data transfer was daunting for two reasons: the amount of data being moved was large (as much as two-terabytes would be moved during cutover weekend), and the network between the two sites was slow. Portable media were the only viable alternative, but these devices would need to meet extremely difficult requirements. They would need to be fast (due to the brief transfer window), inexpensive (no hardware funds were budgeted) and portable (devices needed to be hand-carried by agents to meet security mandates).

Ideally, the team needed to find fast, cheap transfer devices that could be stowed in an airliner’s overhead bin. This paper reviews alternatives for the transfer, focusing on the latest generation of commodity Network Attached Storage (NAS) devices that allowed agents to transfer as much as 12-terabytes in a single carry-on duffle bag.

Background: Transitioning Data Centers

At the time of the transition, both data centers were fully functional: Virginia supported production, while Minnesota (a parallel environment running pre-staged system images on new hardware) supported acceptance testing. Minnesota’s parallel environment included a full snapshot of Virginia’s file system and databases (roughly 10-terabytes in total). However, this snapshot had grown progressively out of sync with production, due to the snapshot’s increasing age and the insertion of test data.

During cutover weekend, the volatile subset of data in Minnesota would need to be over-written by fresh snapshots from the production system. It was calculated that as much as 2-terabytes would need to be moved at this time.

The goal was to complete this transition over a single weekend. Virginia systems would be dropped (permanently) on Friday evening, and the synchronizing data would be transferred to Minnesota and restored. After testing, the Minnesota systems would be released to end-users via DNS changes on Monday morning.

Thus, the time to transfer 2-terabytes was ***approximately 40-hours*** (from Friday at 8PM to Sunday at noon):

Time	Planned Activity
Friday 8PM	Virginia systems down. *Transfer of volatile data begins.*
Saturday	Information is flown by courier to Minnesota on scheduled airlines
Sunday Noon	All data available on SAN in Minnesota. *End of transfer window.*
Sunday 5PM	Files reattached/restored; Minnesota systems available for testing
Monday 8AM	DNS changed. Minnesota systems released for general use.

Table 1: Planned Schedule for Cutover Weekend

The Economics of Bandwidth

In the client's original plans, it was expected that files would be transferred between data centers across their wide area network. However, the client was unaware that the link connecting their two locations was only a DS3 (T3).

While DS3s are theoretically capable of 45Mbits/sec, this particular circuit was shared among several agencies. Our client's allocation varied, but the maximum observed throughput never exceeded 17-Mbits/sec – less than half the full DS3 bandwidth and far too slow to move two terabytes of data in the 40-hours allotted.

Link	Maximum Throughput	Time to transfer 2-terabytes
Full DS1 (T1)	1.544 Mbits/sec	2878.5 hours
Client's fractional DS3 circuit	*17 Mbits/sec*	*526.1 hours*
Full DS3 (T3)	44.736 Mbits/sec	99.3 hours
Full OC3 (STS3)	155.520 Mbits/sec	28.6 hours

Table 2: Estimated Time to Transfer 2-TeraBytes of Data via Network

In our particular case, if we had dropped the Virginia systems and tried to synchronize data across the client's fractional DS3, the systems would have remained down for three weeks (to preserve data state) before the information was fully transferred to the new systems in Minnesota.

The SneakerNet Alternative

Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway. – Generally attributed to computer scientist Andrew Tanenbaum

Portable media (the 'SneakerNet') quickly became the option of choice once it was demonstrated network transfers weren't feasible. With SneakerNets, people move information from one computer to another by physically carrying removable media.

Time to transfer 2-TB using portable media

SneakerNets are characterized by high latency (the delay between the sender's transmission and the initial receipt of the message) but also by extremely high throughput (transfer bit-rate). The table below shows the total time to transfer 2 terabytes of data between Virginia/Minnesota using portable devices. In this example, multiple NAS devices are used – each capable of 35-Gbytes/hour.

Task	Using 3 NAS Devices	Using 6 NAS Devices
Copy from source to transfer media (35 Gbytes/hr/NAS)	19/Hrs	9.5/Hrs
Airline transit between Virginia/Minnesota data centers	6/Hrs	6/Hrs
Copy from transfer media to target (35 Gbytes/hr/NAS)	19/Hrs	9.5/Hrs
Total elapsed time for transfer	44 Hours	25 Hours

Table 3: Estimated Time to Transfer 2-TB via Portable Media. (In this example, the portable media are multiple NAS devices, each with a peak transfer rate of 35-GB/hr.)

As the example shows, even allowing for the high latency of portable media, the overall end-to-end throughput approximates that of a dedicated (and extremely expensive) OC3 link.

Media Options for TeraScale SneakerNets

Our relocation project used NAS devices for most data transfers (details below). However, other portable media formats were considered including a) USB hard drives, b) "shippable" disk arrays, c) and tape.

USB hard drives

Hugely popular in both home and lab, USB hard drives were our initial choice as transfer device. However, these devices suffered from data center usability problems (data centers may deactivate USB ports due to security concerns, and VMware/USB interactions can be complex). There are also issues using USB drives from the c-class blade servers deployed in our project.

Modern blade servers include an internal USB port. However, this port is intended only for small devices that can be sealed in the server chassis, such as a USB security key. For *external* USB devices, the only available connection to such blades is through the multi-function port via a 'Local I/O Cable':

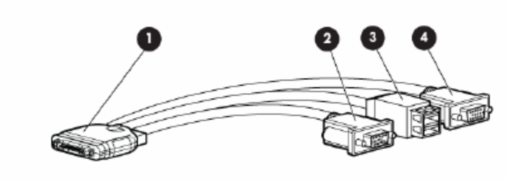


Figure 1: Blade USB Interface via 'Local I/O Cable'

To use the Local I/O Cable, its connector (1) is attached to a custom multifunction port on the blade's front panel. The cable then provides access to video (2), USB (3), and serial (4) interfaces. The USB ports on this interface don't directly support USB 2.0 hard drives.

Shippable disk array

The project team considered using a small disk array, such as a SCSI drive enclosure. Such devices are connected to a front-end server (e.g., re-purposed DL380). Files are transferred from the disk array to the front end server at full SCSI speeds; the front-end server can then share these files across the LAN. These disk arrays are shippable but are too large for transport by courier (and thus not directly usable in our project).

Figure 2: MSA30 Drive Array

An alternative was considered wherein a pair of disk array chasses would be installed, one at each location. During cutover weekend, disks would be removed from the source chassis, transported by courier (disks only), and then installed in the target chassis. However, the cost (two chasses) and risk (drive placement needed to be identical in both locations) was judged unacceptable.

Tape drives

Both Virginia and Minnesota had identical small tape robots that supported 26 SDLT-160/320 tapes (4.16 TB native capacity). The tape robots were connected via SCSI to front-end servers (similar to the disk array configuration, above). Since both locations had identical tape robots, only the SDLT tapes themselves needed to be transferred (easily managed by courier). In addition, transfer speeds were extremely fast. Using NetBackup to read and write tapes, the maximum observed throughput on our LAN neared 100-GBytes/hour.

Figure 3: Tape Unit

Because of these advantages (portability, speed) we did use tapes for some data transfers during the project. However, the team had access to only one person at each location who understood NetBackup/tape operations, and our tape usage was always contingent on their availability. Another disadvantage was scalability: with only a single tape robot at each location, we couldn't increase aggregate throughput by running multiple devices simultaneously.

Finally, we experienced reliability problems on our aged tape units. Work streams were commonly disrupted during tape transitions (though happily, these errors were re-startable). More serious errors occurred due to SCSI cabling unreliability.

NAS Device == Computer PLUS Disk Array

So lately I'm sending complete computers. We're now into the 2-terabyte realm, so we can't actually send a single disk; we need to send a bunch of disks. It's *convenient to send them packaged inside a metal box that just happens to have a processor in it.* – Jim Gray, Turing Award winner and SneakerNet advocate

Conceptually, a NAS device is akin to the MSA-30 disk array, except that NAS devices combine a computer and disk array in a single enclosure. Whereas a disk array requires a separate front-end server to be usable, a NAS device includes a server and disk array in a single box. NAS devices are almost as simple to use as a USB drive – just plug the NAS device into an Ethernet switch, rather than into a USB port. The particular NAS device used in our project was manufactured by LaCie.

The LaCie Ethernet Disk

LaCie's NAS device is small desktop unit (roughly 6 x 9 x 9 inches) that weighs 12-pounds. Each device includes four removable, hot-swappable 7200-RPM disk drives, a hardware RAID controller (RAID 0, 1, 5, 5+ spare and 10), and two Gigabit Ethernet ports. The embedded system runs Linux in 256 MB of memory on Intel's 80219 XScale processor.

The device currently comes in four capacities, with the largest providing four terabytes of storage in a single unit (RAID-0 with four 1TB drives). The chassis is sturdy, and couriers were able to hand-carry up to three devices – each safely bubble-wrapped – in a single carry-on duffle bag.

Figure 4: LaCie Ethernet Disk

Transfer speeds varied by configuration and task (for example, RAID-0 is faster than RAID-5; reading data is faster than writing). However, the maximum observed throughput on our network was 35-GBytes/hour for a single device.

Because more than one device could run in parallel on our LAN without degrading performance, multiple LaCie devices were used to improve aggregate throughput (e.g., 3-devices could provide over 100-GBytes/hour peak transfer speeds). The low cost (2TB ~ $1,100) made multiple device purchases affordable even for our budget.

Connection topologies

We used two different strategies for connecting NAS devices in the data center: NAS to switch, and NAS direct-to-server.

a) NAS to switch

Connecting the NAS device to an Ethernet switch provides the greatest flexibility. The NAS device simply appears as another system on the network, and multiple servers can connect to the NAS (and its files) as necessary. This approach requires that the NAS device be assigned an IP address on the target subnet.

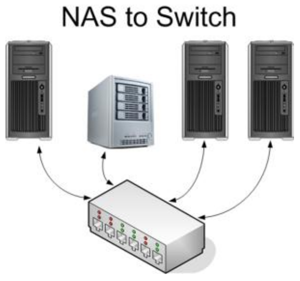


Figure 5: NAS Connected to Switch

b) NAS to Server

Connecting the NAS device directly to a server typically provides the greatest throughput. Assuming the server has an unused NIC, this secondary NIC can be set to a non-routable IP address (e.g., 192,168.1.x) and the NAS device can be set to a compatible address (e.g., 192.168.1.y).

Once the NAS and server are connected, they can exchange files at full GigE speeds with no network contention.

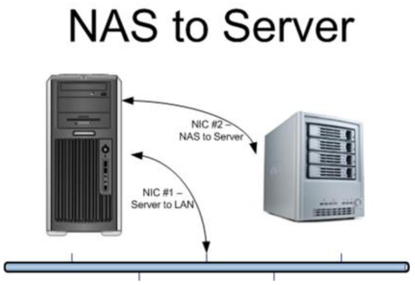


Figure 6: NAS Connected to Server

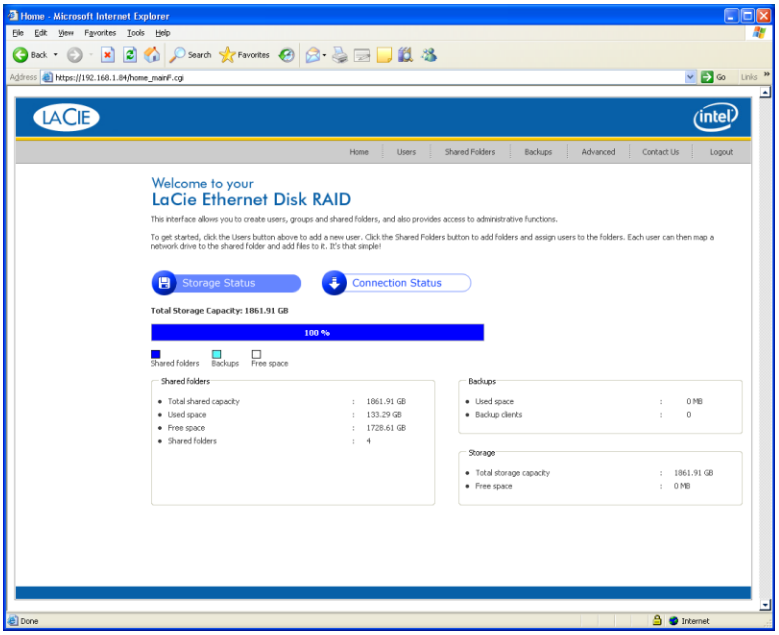
A final advantage for fully portable media

One final advantage of fully portable, non-rack mounted devices (e.g., NAS devices and USB drives) was critical for our project. Because these devices can easily be moved from server to server, it's possible to transfer data while avoiding a network hop. By contrast, disk arrays and tape drives typically incur a network hop when transferring data between the front-end server and the target system.

In addition, the non-rack mounted devices make it possible to easily service multiple subnets (a requirement of our project, which encompassed both internal networks and an enclave/DMZ). By contrast, rack mounted hardware typically can't service multiple subnets without cumbersome re-addressing and/or re-cabling schemes.

Administering the LaCie device

The LaCie device is administered via web pages that are served up by its embedded Linux OS. The following screenshots were captured after browsing to the LaCie device's IP address and logging on as administrator. Once logged on, the following summary is presented:



https://192.168.1.84/home_mainF.cgi

Welcome to your
LaCie Ethernet Disk RAID

This interface allows you to create users, groups and shared folders, and also provides access to administrative functions.

To get started, click the Users button above to add a new user. Click the Shared Folders button to add folders and assign users to the folders. Each user can then map a network drive to the shared folder and add files to it. It's that simple!

Total Storage Capacity: 1861.91 GB

100 %

■ Shared folders ■ Backups □ Free space

Shared folders

- Total shared capacity : 1861.91 GB
- Used space : 133.29 GB
- Free space : 1728.61 GB
- Shared folders : 4

Backups

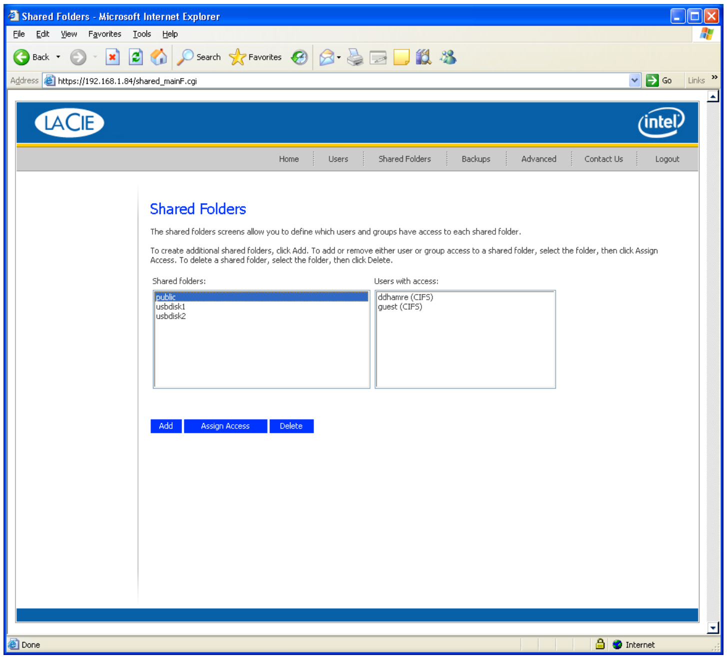
- Used space : 0 MB
- Backup clients : 0

Storage

- Total storage capacity : 1861.91 GB
- Free space : 0 MB

Figure 7: Administrative Summary for the LaCie Ethernet Disk

Administrators can define users and file shares, and link the two appropriately:


Shared Folders - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address https://192.168.1.84/shared_mainF.cgi
Go Links
LACIE
intel
Home Users Shared Folders Backups Advanced Contact Us Logout
Add Assign Access Delete
Done
Internet

Shared Folders

The shared folders screens allow you to define which users and groups have access to each shared folder.

To create additional shared folders, click Add. To add or remove either user or group access to a shared folder, select the folder, then click Assign Access. To delete a shared folder, select the folder, then click Delete.

Shared folders:

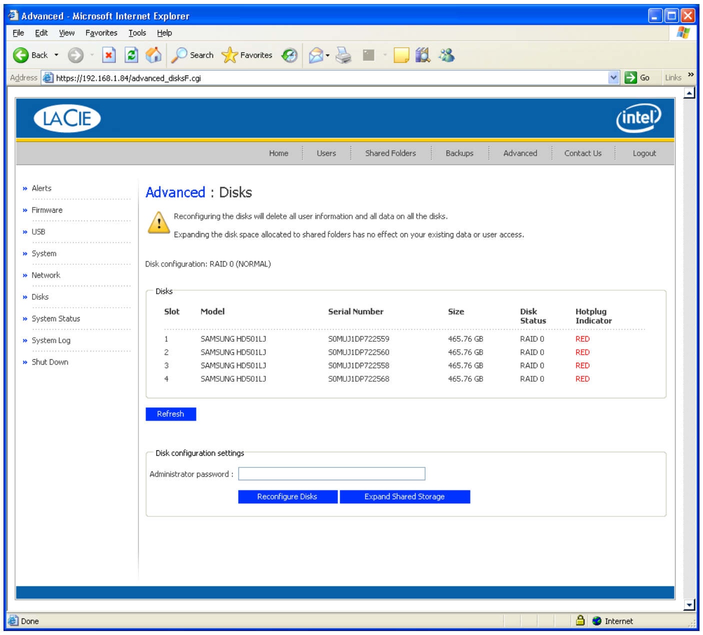
public
usbdisk1
usbdisk2

Users with access:

ddhamre (CIFS)
guest (CIFS)

Figure 8: Managing Shared Folders and User Accesss

The hardware management interfaces (below) are equally clean and straightforward. In this example, the device was configured for RAID 0 (not RAID 5), so the interface correctly shows the drives can't be hot-swapped.


Advanced - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address https://192.168.1.84/advanced_disksF.cgi
Go Links
LACIE
intel
Home Users Shared Folders Backups Advanced Contact Us Logout
» Alerts
» Firmware
» USB
» System
» Network
» Disks
» System Status
» System Log
» Shut Down
Refresh
Reconfigure Disks
Expand Shared Storage
Done
Internet

Advanced : Disks

Reconfiguring the disks will delete all user information and all data on all the disks.

Expanding the disk space allocated to shared folders has no effect on your existing data or user access.

Disk configuration: RAID 0 (NORMAL)

Disks

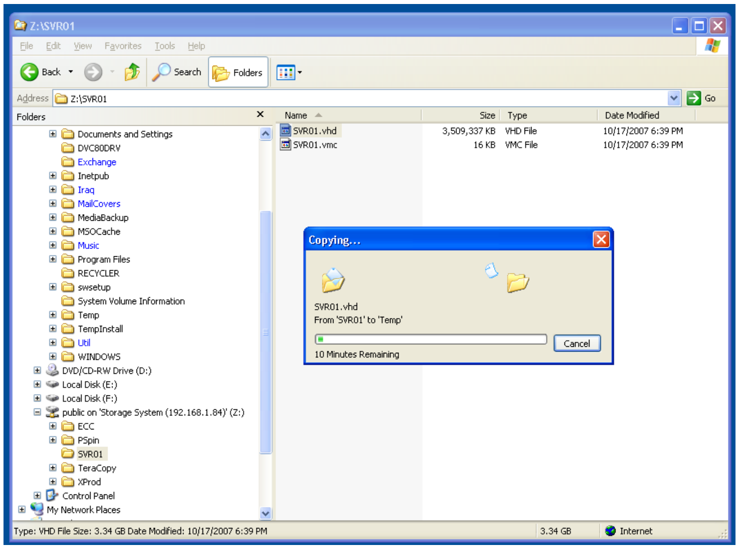
Slot	Model	Serial Number	Size	Disk Status	Hotplug Indicator
1	SAMSUNG HD501LJ	S0MUJ1DP722559	465.76 GB	RAID 0	RED
2	SAMSUNG HD501LJ	S0MUJ1DP722560	465.76 GB	RAID 0	RED
3	SAMSUNG HD501LJ	S0MUJ1DP722558	465.76 GB	RAID 0	RED
4	SAMSUNG HD501LJ	S0MUJ1DP722568	465.76 GB	RAID 0	RED

Disk configuration settings

Administrator password :

Figure 9: Administering Device Configuration

NAS devices are extremely easy to use during file transfers, functioning as mapped network drives. In the example below, the LaCie is mapped as Z: and a 3-GB file is being copied to C:\Temp. Note again that for LaCie devices configured with RAID-0, the maximum observed transfer speed was 35-GBytes per hour.



Name	Size	Type	Date Modified
SVR01.vhd	3,509,337 KB	VHD File	10/17/2007 6:39 PM
SVR01.vmc	16 KB	VMC File	10/17/2007 6:39 PM

SVR01.vhd
From 'SVR01' to 'Temp'

10 Minutes Remaining

Type: VHD File Size: 3.34 GB Date Modified: 10/17/2007 6:39 PM 3.34 GB Internet

Figure 10: Using the NAS Device as a Mapped Network Drive.

Windows ‘Copy’ and ‘XCopy’ and the Problem of Large Files

One of the primary advantages of NAS drives is ease of use. As with USB hard drives, standard Windows utilities (Explorer, copy, etc) are used to manage files (unlike tapes, where specialized software like NetBackup must be used).

Unfortunately, this ease-of-use may be treacherous because familiar commands – including Windows’ *copy* and *xcopy* - may not be suitable for data-center-scale operations. In particular, it’s not uncommon to encounter performance and reliability problems when using *copy* or *xcopy* to migrate large files.

For this reason, alternative utilities should be considered; popular choices include *RoboCopy* (formerly available in Windows’ resource kits and now bundled with Vista), Microsoft’s *ESEUTIL* (an Exchange utility), or shareware such as *TeraCopy*.

In testing with a LaCie device, TeraCopy was roughly 5% faster than simple Windows copy (an advantage that should improve with larger files). TeraCopy also includes an extremely fast CRC check to verify copy operations.

Closing Notes: Tread Softly into the Data Center

This project exploited the cost/size advantages of mass-market storage devices to help migrate a multi-terabyte data center. These devices’ low cost, speed, and reliability were instrumental to a successful relocation. However, these devices are engineered chiefly for home use, and there are restrictions that must be considered with any recommendation to use similar devices in a raised-floor environment.

Security is an obvious concern: any device capable of hauling terabytes of data into a building could leave with a comparable amount. Security teams should be involved early in the project, and will need to clear any use of large-capacity portable storage.

The NAS and USB devices discussed above are not rack-mountable and require 110V power. For these reasons, the devices are typically exposed in walkways during use, with power cords snaking across the raised floors. If you use similar devices, expect your intrusion in the data center to be tolerated only for short periods of time, and expect to know the hiding place of every extension cord in the building by the time you’re done.

Acknowledgement

Thanks to reviewers for their time and suggestions; any inaccuracies remain mine.

Related Knowledge Briefs or References

The TeraScale SneakerNet by Jim Gray
http://research.microsoft.com/~gray/papers/TeraScaleSneakerNet.doc

A Conversation with Jim Gray
http://www.acmqueue.org/modules.php?name=Content&pa=showpage&pid=43

The Economics of Bandwidth by Jeff Atwood
http://www.codinghorror.com/blog/archives/000783.html

LaCie Ethernet Disk RAID Review: No-frills small-biz RAID by Craig Ellison
http://www.smallnetbuilder.com/index.php?option=com_content&task=view&id=30009&page=5&Itemid=75

Review: LaCie Ethernet Disk RAID 2TB entry-level NAS box By Rob Kerr
http://www.reghardware.co.uk/2007/01/17/review_lacie_ethernet_disk_raid/

LaCie Documentation (2TB NAS)
Datasheet: http://www.lacie.com/download/datasheet/ethernetdiskraid_en.pdf
Manual: http://www.lacie.com/download/manual/ethernetdiskraid_en.pdf
Quick Install: http://www.lacie.com/download/qig/ethernetdiskraid.pdf

Summary

Capacious, fast, and inexpensive, the latest generation of commodity Network Attached Storage (NAS) devices is sufficiently mature to play a role in data center relocation. We recently moved multiple terabytes of data for a relocating government agency over a single weekend. Using commodity NAS devices, security agents comfortably transported up to 12-terabytes of data in their carry-on luggage. Running multiple devices in parallel, data was transferred from portable NAS devices to permanent storage at an aggregate rate that exceeded 100-gigabytes/hour.

Drew Hamre is a principal of lexana\net and lives in Golden Valley, Minnesota.