

Gnutella Peer-to-Peer File Sharing: What It Means For Data Storage

If you have never heard of Gnutella, then it is a sign that you need to talk with your kids more. As every teenager knows, Gnutella is a peer-to-peer file sharing system used mostly to distribute music files in MP3 format over the Internet. Unlike the now infamous Napster, Gnutella does not rely on a central server or service. It is a distributed system. In fact, the only thing that is close to centralized on the Gnutella network is the master directories of participating Gnutella servers, of which there are several. Even shutting down all the master directories won't shut down the systems since another one would simply pop-up in hours and large static directory lists are freely circulated around the Internet.

What Gnutella ultimately does is allow the user to search thousands of directories on thousands of disks at once. When the client software - which is available from several different vendors for free - connects to the Gnutella network, it searches the main directories for the IP addresses of the currently connected systems that are available to share files. Having access to portions of thousands of disks, the searchable space can easily reach into the terabytes. Usually, there is between three and six terabytes of searchable space at any one time available. That's **six terabytes!** All from individual disk space that is, on average, less than five gigabytes a piece.

Gnutella is a large, highly distributed, virtualized storage system. Individual disks are available or not available but some storage is always available. More storage is added and some removed moment to moment. The system cannot be destroyed unless the entire Internet and all the individual computers that participate in it are simultaneously turned off or disconnected. This makes it highly resilient. It is impossible to bring down the network via a lawsuit the way Napster was or through all even the most extreme Internet-based attack.

For most people who use Gnutella, it's a convenient way to get music without paying for it. This should put the music industry into an apoplectic state. The software industry should also be worried because quite a lot of illegally copied software is also available although not nearly to the extent of music. The implications for the IT community are different and generally positive.

It's a Good File Storage System

It is important to put the Gnutella system in perspective. Whatever it's nefarious uses it is a good storage system. Since it is highly distributed, it is incredibly invulnerable to loss of availability. By using small amounts of disk space on thousands of computers, it makes efficient use of existing resources. It provides a front end to the ever changing storage network that makes the network transparent to the end-user. In other words, it accomplishes much of what the storage industry has been attempting to do for years. It creates a

distributed, networked, virtualized, storage subsystem running over IP networks. Some of the features that are core to the Gnutella network are amazing considering that the software is free and the majority of users have been college students. For example, Quality of Service statistics are collected and reported for each node as well as overall network statistics such as connection speed. This allows users or other systems to decide to where to get a file based on search criteria *and* the best network connection.

As would be expected there are some downsides to the system. For one, it's completely file oriented. Perhaps it could be adapted to block I/O but that would be a remarkable reworking of the system.

The client software that is currently available doesn't neatly integrate with operating systems. I haven't found a Gnutella client yet that integrates into the Windows Explorer interface, a key need in the corporate environment.

There is also no way to know if the contents of a file are in fact what the file name says it is. Searches are not conducted on a more descriptive database though there is no reason why this could not be added. So, while quality of service in terms of the network can be reported, quality of content cannot. The Gnutella community is looking at adding this feature. The addition of quality of content for files in the Gnutella file space would be a significant improvement over most current files systems today. Go into any corporate LAN and look for a file whose name you do not know explicitly and see just how difficult this can be.

Uses in IT

The Gnutella system would seem at first to be the natural enemy of most IT departments. The way it is currently used, to get music files from the Internet, is a major problem for most IT managers. Besides using enormous amounts of bandwidth for non-company purposes, there are major security issues with file sharing systems in public networks like Gnutella.

The advantageous however are compelling. Originally, file servers existed because disk space was expensive and there was no other way to share files within a network. With the advent of cheap hard drives for desktop computers, the cost argument no longer holds true. Large amounts of disk space are attached to servers whose sole purpose is to allow people to share files. Files are circulated by e-mail, creating a dozens of duplicates in the process, or placed in shared folders on file servers. The inherent problem with this approach is added costs of storage equipment and on-going maintenance of the file servers. Repurposing the Gnutella technology would bring benefits to IT managers trying to manage this type of storage in the network.

In effect, costs savings are realized because the Gnutella system makes efficient use of *existing* resources. If most IT managers did a survey of all the existing disk

space in their networks, not just what is attached to servers, they would find a lot of unused space. At the same time, they often find that they don't have enough shared space available for the needs of their end-users. Part of the reason for this is that much of the existing disk space is locked up in desktop computers or underutilized server attached storage. SANs and NAS devices help to alleviate some of this inefficiency by making more of the server attached storage accessible to more servers but it does little to leverage the unused space on desktop and departmental storage. In addition, it actually encourages the duplication of files for distribution. Gnutella, by making it easy to find and share files where they already exist, makes better use of underutilized storage resources.

File servers and NAS devices create availability issues since loss of a single file server can cause loss of all access to shared files. It is rare that departmental file servers have the same levels of redundancy and fault tolerance as central IT servers. Gnutella-type networks should be attractive to IT managers because they are naturally distributed. This makes them less prone to catastrophic failure. When file space that is spread throughout a large physical space it is difficult to bring down the entire storage network. Access to a particular file is not guaranteed, but it would be rare for the entire file space to be held offline unless there is a major network outage.

In its current form usefulness is limited but with some small changes it could provide a way of sharing files in a corporate network that would make finding these files easier. In addition, it would allow a more efficient use of existing disk space spread throughout the corporation and reduce the need to deploy file servers simply to allow for shared directories. A different type of client needs to be developed. It should have a look and feel that integrates with the file management tools that are commonly found on corporate desktops. A good example to follow would be the Yahoo! Drive client. This software makes the file storage system provided by the Yahoo website (Yahoo Briefcase) appear as a drive on a desktop computer running Windows.

Another important feature that would be necessary for this technology to be deployable in a professional IT organization is some way of performing backups. In this case the backup software would need some method of backing up all the shared files in the network even though they are spread out over thousands of disks on individual desktops. To be truly effective, this capability would need to become part of the network itself.

Security would also have to be addressed. File sharing the peer-to-peer networks is remarkable egalitarian. Anything put out for sharing can be retrieved by anyone within the network. This open sharing scheme has implications in the corporate environment where files are rarely designed to be shared across the entire company. There is some rudimentary zoning in Gnutella called horizons

but a solid policy-based security scheme would be necessary to quell internal security issues.

Now, if you're a software vendor...

If your company derives significant revenue from software sales, systems like Gnutella should cause some concern. It is very easy to distribute pirated software on the public Gnutella network. Since it is nearly impossible to shut down every site, legal remedies are likely to be ineffective. Simply finding the sites - which can easily shut down and start up at a whim - is a daunting task. Copy protection schemes are stop gap measures since these could be defeated and the hacks distributed on the same networks.

On the other hand, peer-to-peer file sharing in the more controlled corporate IT environment represents a reasonable way of distributing software updates throughout a network. Once quality-of-content features are in place, it would be possible for even a fairly large IT department to find and distribute patches and updates to desktop computers in an automated fashion without having to already know where to find them in the network. Thus, as a corporate network changes, the patch directories can change but the software update system would still find them without changes to the desktop or server client software.

Alas, Gnutella will not win in the corporate IT world.

While this all sounds positive, it will probably not be Gnutella or one of its scions that implement this type of distributed file system in IT departments. Because of its roots in sharing pirated music, which are typically very large files, most IT departments are hostile to it. That hostility is also felt by the software industry which has issues with the easy pirating of their intellectual property.

The core features of the Gnutella network, peer-to-peer file sharing with QoS, will find its way into the IT department as either an add-on system overlaying existing files systems or as a core feature of file systems. The various flavors of Windows have had peer-to-peer sharing capabilities since Windows 95 but these are primitive, manual, and static. The application of Gnutella techniques are likely to be co-opted by system software companies. As Microsoft has proven time and time again, it is better to embrace new technology rather than let it challenge your core products.

Peer-to-peer file systems are still in their infancy and not corporate quality. So were web servers eight years ago and MP3 file three years ago. These were all embraced by corporate IT departments and systems manufacturers because they represent significant value to end-users and IT departments alike. Gnutella-like peer-to-peer file systems present a similar value and hence will succeed over time.