

Recent Interview with Phil Schwan, CEO of Cluster File Systems, Inc., Developer of the Lustre Open Source Cluster File System

Phil Schwan, CEO of Cluster File Systems (CFS), spoke this month with Computer Technology Review about the state of his company and the adoption and positioning of the Lustre File System. Lustre is a high-performance cluster file system developed for managing very large volumes, with very high bandwidths. Many of the world's largest supercomputers have been using Lustre in production for more than a year.

Q: What is the DNA behind Lustre?

Phil Schwan: The Lustre project was started in 1999 by CFS's founder, Dr. Peter Braam, when we were together at Carnegie-Mellon University. Peter began working on an answer to the Department of Energy's Scalable/Global/Secure file system challenge. The goal was to solve the next ten years of cluster file system problems, for systems of tens of thousands of nodes and many hundreds of petabytes.

Lustre's architecture was developed to address these massive storage needs with the expectation that if we could solve the hardest problems of the biggest clusters, then scaling down to the more typical systems would be possible. Working with the early-adopter customers and partners, we've had systems in production for about a year and a half, handling HPC workloads for government and private industries.

Q: Would you please give us an overview of the Lustre architecture?

Schwan: There are three types of nodes in a Lustre installation. Metadata Servers manage the namespace. The names of the files are stored here, but the actual file data is split up and distributed over one or more Object Storage Servers. Today's largest Lustre installations can have several hundred of these commodity servers. You can scale the bandwidth and capacity of your clusters very linearly by adding more, because they operate completely in parallel. Many shared file systems have a single pool of disk shared directly between the entire cluster. That's an approach that really doesn't scale, so we give each server its own pool of disk to manage on its own.

The third type of node is the client node. These speak a Lustre object protocol to the metadata and object servers, and provide a standard file system interface to applications. Today those native clients run only on Linux, but you could use Samba or NFS to re-export, so that you have access to your Lustre file system from whatever other OS you may happen to have.

Q: How is Lustre sold and distributed today and what is the licensing model?

Schwan: We saw that there was no other really good open source cluster file system for the Linux community, so Cluster File Systems was founded to create, distribute, and support Lustre as an open source product. Today, CFS distributes Lustre in two ways. For the first year of any new release of Lustre, like our recent Lustre 1.4 release, the software is available to our paying customers and our commercial partners under the GNU General Public License (GPL). After a year, often sooner, we make that new version available on the Web to anyone, under the GPL.

We stay in business by licensing and supporting the Lustre software directly and through partners such as HP, Cray, Sun, Linux Networkx, Verari and similar companies. Directly, we've fostered some exceptional relationships around very high-end deployments for which there are very few choices other than Lustre. For example, the largest supercomputer in the world, Lawrence Livermore's IBM-based BlueGene/L system, will use Lustre. We're also working very closely with Cray on the new Red Storm supercomputer, which will be installed in 2005 at Sandia National Laboratories.

Q: So, HPC was clearly the birthplace of Lustre, but how far off are you from being adopted in commercial enterprise environments?

Schwan: We already have some strong acceptance within the industries that resemble our engineering and scientific customers. These are industries such as oil and gas, pharmaceuticals, digital media, manufacturing and automotive, and I expect this part of our business to continue growing rapidly. Transaction processing environments, such as those found on Wall Street or in the insurance industry, are still a few years off, as they tend to be very conservative and require additional features. While these are on our roadmap, they're not finished yet.

Q: Your partners like HP, Sun, and others sell in both the HPC and commercial markets. Can we assume that while they are pursuing the engineering and scientific markets today, they will move more into the enterprise with Lustre at some point downstream?

Schwan: Yes, that's true of many of our partners. It doesn't do them any good to push Lustre where it's not yet ready. We are, however, having many discussions with partners and enterprise customers about where they want Lustre to be in the future.

Q: How far down the development path have you been able to go? How far along is Lustre today?

Schwan: Lustre is ready for prime time. It is extremely usable in production for a large number of applications on both large and small clusters. Organizations in many industries are using it very successfully with the help of CFS and our partners. We've had the fortune of being able to develop a very healthy, respectable customer base without the need of a great marketing effort. While there is still a lot more to accomplish, and our roadmap is long, Lustre is a reality today – and our customers are proving that.

Q: What is the largest Lustre system today, and what are the upper bounds of system size and file system size?

Schwan: The largest Lustre system in production today is a 4,000-CPU cluster at Lawrence Livermore National Laboratory. With Lustre 1.4, there should be little problem installing 3,000 or 4,000 nodes right out of the box. The Red Storm computer will be substantially larger and it will be online in 2005, so a 10,000 node installation is not far out.

With respect to file system size, Lustre was architected to handle 64-bit file systems with thousands of petabytes, but we didn't really expect people to do that for some time. We are finding that people are deploying petabytes now, and we are lifting all the limits.

Q: Is there life for Lustre beyond Linux? Other operating systems?

Schwan: Absolutely. In 2005, we will release a Lustre client for Apple's Mac OS X. As for other systems like AIX, Solaris, and Windows, we are considering all of them, but need to see what customers want. It seems quite likely that we will do a Windows port at some point in the near future.

Q: How did you decide to offer Lustre as open source and what is your company's profit model?

Schwan: We started the Lustre project and we founded Cluster File Systems with the goal of providing a good open source cluster file system for Linux. With that in mind, we began by doing contract development of Lustre, developing prototypes or stabilizing the features that customers wanted most. Sometimes these were for clusters that were immediately available, like at Lawrence Livermore and Pacific Northwest and NCSA, and sometimes these were very forward-looking prototypes of things they wanted in years to come.

CFS is in the middle of shifting quite substantially to a model based on services and licensing. Well more than half of our revenue now comes from services such as enterprise support, training and licensing, both through partners and directly to customers.

Q: What is your corporate growth strategy and how fast do you want to grow?

Schwan: If we had been asked that question when we started, we would have said we would become a ten-person niche supercomputing company. Today we have about 45 engineers developing and supporting Lustre. I don't expect the same kind of explosion of growth that we've had in our first three years, but I do expect that we'll continue to grow slowly. We would prefer to work through experienced partners and fulfill a role of architecture development, productization and high-level support, rather than interfacing with each individual user directly. We don't expect to be a 500-person company.

Q: What do you foresee for the company over the next five years?

Schwan: We might be very happy to stay small, but our roadmap is as good or better than that of any company in the storage industry. We will be setting the standard in cost per terabyte, reliability, performance, features, and scalability. We will probably end up competing in most major areas of storage deployment.

In five years, the system will be quite mature, but people will still need to know that if something happens they can pick up the phone and call us. They will want speak directly to the experts who designed and implemented the Lustre file system, the people who know it best.

Q: So the buck stops with you?

Schwan: Ultimately, that's true. Whether you buy Lustre support from Cluster File Systems or from one of our industry partners, the buck does stop here.

###