

Super Saver

The Race Is On To Build A Fast, Global File System For Supercomputers

Save often, especially when you run a supercomputer.

For Gary Grider, group leader of Los Alamos National Laboratory's High Performance Computing Systems Integration Group, saving data is of the highest importance. He is part of a team that is developing what may well be the world's fastest supercomputer, a petascale machine called Roadrunner with more than 32,000 processors. IBM Corp. is leading the effort.

Jobs simulating nuclear-weapon degradation could take months to run. If a single processor failed — a statistical probability given the sheer number of CPUs used — the work would be corrupted. So, naturally, the lab wants to save often, just as you might do with your PC. But in this case, the procedure involves frequently saving terabytes of data as quickly as possible — no small feat.

That's why Los Alamos specified that data must be able to flow back from the processors to the storage arrays at an unprecedented 50 Gbps, far beyond the capability of any single storage cluster. Running multiple storage arrays in parallel would do the trick, but that approach requires advanced techniques for coordinating the storage and management of data.

Roadrunner isn't alone in facing this challenge. "You can easily put a lot of CPU power in the room, but to do useful work, you also need very good I/O," said Mike Gigante, SGI's engineering director of file-serving technologies. "Unfortunately, many people don't think about the I/O until the CPU is set up, and they realize that the overall utilization efficiency of their computer is very low."

File here

Managing a computer's data is the job of the file system, and agencies, volun-

teer bodies and industry are working on a new generation of file systems, often called global or parallel file systems, that can support machines such as Roadrunner. The challenge is picking the right one for the job.

In many ways, Energy Department laboratories have been a driving force behind the development of global file systems. In 1994, Energy labs banded together to develop Lustre, a file system designed specifically for the upcoming supercomputer deployments. "We didn't see anyone out there who had what we wanted," Grider said.

In short, a file system is a data structure for storing files on disk, according to Andrew Tanenbaum and Albert Woodhull's *Operating Systems: Design and Implementation*. Generally speaking, a file is a collection of data, or a binary object that can be loaded into memory and run as a program. The file system defines the file-naming conventions as well as what operations can be done to a file (open, read, write, close, delete, append, get attributes, set attributes and rename).

Network file systems specify what operations can be done to a file over a network.

A global file system usually has extra, added duties. It must provide a means to keep track of data spanned across multiple storage arrays. For large implementations, simply buying a number of independent arrays only leads to confusion, since users must manually keep track of which array holds their information, and swapping information between arrays is a hassle.

"You definitely want all [data] shared across all the nodes in a cluster, so all the nodes see the same data, can read the data and write the data to common places," Grider said. At the same time, speed of access should not be hampered by having the pointers to the data aggregated into one large pool — a difficult problem to tackle.

Lustre attacks the problem, and it is still used in many supercomputer systems. Cluster File Systems Inc. of Boulder, Colo., now manages the Lustre code base. Lustre can be used in scenarios where hundreds of megabytes per second need to be moved.

Supercomputer manufacturer Cray Inc. of Seattle plans to use Lustre for the internal file system of its next-generation X1 supercomputer, nicknamed Black Widow, said Peter Rigsbee, product marketing manager for the company. Lustre also worked well in Cray's Red Storm XT3/4 series of supercomputers, whose nodes do not have full operating systems. "Since Lustre is an open-source product, it is far more malleable," Grider said.

Lustre might not be suitable for all implementations, though. "Lustre has some wonderful attributes, but it is pretty complicated—it is difficult to set up and difficult to maintain," one industry veteran has said. Few Lustre experts are out there, and tools are fairly rudimentary.

Mix and match

Overall, Energy labs use a mixture of the major global file systems, Grider said. Lawrence Livermore and Sandia national laboratories both use Lustre. Livermore also uses a global file system originally developed by IBM for its own systems, called the General Parallel File System (GPFS). Argonne National Laboratory and the Ohio Supercomputer Center are developing the Parallel Virtual File System (pVFS), an open-source parallel file system currently best suited for small clusters.

For the Roadrunner system, Los Alamos chose the Panasas ActiveScale File System, which will run on the ActiveScale Storage Cluster from Panasas Inc. of Fremont, Calif.

"We do things in parallel," said Larry Jones, vice president of marketing for

continued on back...

Panasas. "A typical network appliance handles things serially. Basically, you send in a request, it works really hard to process that request and send [the data] back as fast as it can." Panasas' approach to speeding the exchange is to break the job into multiple portions that can be simultaneously executed.

In the Panasas system, each pair of disk drives is housed in a blade chassis and gets a dedicated network connection. Each chassis, which can have up to 10TB of storage, has four dedicated Gigabit Ethernet connections, a number of interconnects usually reserved for arrays with 10 times that capacity.

Tackling NFS

The Network File System, originally developed by Sun Microsystems Inc., has long been the norm in Unix-based network environments. It is easy to set up and very reliable. An NFS storage server, on average, can offer 500-Mbps transfer rates. NFS serves as the basis for the Network Attached Storage systems offered by Network Appliance Inc. of Sunnyvale, Calif.

While good for normal network storage implementations, NFS has a reputation for not scaling very well to supercomputing environments. An environment that needs throughputs faster than 500 Mbps must set up multiple arrays and spread the data out among them. Software has been developed to aggregate multiple NFS servers into a common pool, though this approach tends to create a bottleneck, as all requests must go through a single server.

NFS also does not address the problem of how multiple users can access the same file. "You have problems arise when different clients talk through different nodes, but

are talking about the same file," said Peter Honeyman during a presentation at the SC06 supercomputer conference last November in Tampa, Fla. Honeyman is head of the Center for Information Technology Integration at the University of Michigan, as well as a contributor to Version 4 of NFS.

Since NFS is well-known in the network administrator community, various initiatives are under way to boost its possible output. The University of Michigan, with some Energy Department funding, is developing an extension to NFS called NFS RDMA (Remote Direct Memory Access). NFS RDMA promises to break the bottleneck by eliminating some of the work a server does when moving files, said Gigante.

When sending or writing a file, NFS typically uses a lot of its host server's processor power. The more data being sent, the more the CPU is being used. As a result, a single CPU can, at most, pump out about 1 Gbps for reading data and about half that to write material to the storage disk.

RDMA offloads most of that work to the chip on the network card itself, which means each server can deliver a lot more data. By considerably scaling back CPU overhead, an NFS RDMA storage server could host as many as 16 Infiniband network adapter cards. With each card pumping out 800 Mbps, a storage array could offer a throughput of 10 Gbps or more. This speed approaches the throughput of Lustre-based systems.

Another NFS enhancement being developed, called Parallel NFS, promises even greater throughput. Like NFS RDMA, pNFS is a planned extension to Version 4 of NFS. It would solve the bottleneck

problem by parallelizing the file services. In essence, data can be spread out across multiple servers, according to an Internet draft on pNFS authored by Panasas chief technology officer Garth Gibson and Peter Corbett of Network Appliance (see GCN.com, Quickfind 732).

Such parallelization can boost throughput way beyond what even NFS RDMA can offer, Gigante said. One advantage pNFS offers is that the individual client can pull data from a large number of servers, allowing access to far more data than any one NFS server could offer. The most a client could pull from any one NFS server may be anywhere from 50 Mbps to 90 Mbps. Yet under pNFS, a client could establish five connections to five different servers, aggregating 450 Mbps.

Most vendors see pNFS as the way forward, though researchers note that it is less mature than NFS RDMA and still two years or more from commercial deployment.

Today's supercomputer designers have available a variety of solutions and must weigh the merits and drawbacks of each. Paul Buerger, who heads up systems and operations for the Ohio Supercomputer Center, noted that the center has been experimenting with distributed, parallel file systems including GPFS, Lustre and pVFS. OSC provides supercomputing power to universities and businesses in the state and the surrounding region.

"Each of these file systems has its advantages and disadvantages," he said of today's crop of parallel file systems. "None of them has yet distinguished itself as the answer to all I/O issues in supercomputing."