HPC WIRE ADVERTORIAL

Scaling to New Heights in Linux Capability, Linux Performance

By Steve Neuner and Jason Pettit

It takes a village to raise an operating system. Or so it would seem to anyone familiar with the care and feeding of Linux®. The ever-growing community of avid Linux developers is big on collaboration. This all-for-one approach has enabled the open-source environment to evolve and improve in multiple directions faster than would otherwise be possible if all development were contained under a single roof.

There are huge advantages to this. Economically, the potential price/performance from a solution incorporating standard Linux and Intel® products has been enormously compelling. Functionally, contributions to Linux often come from far afield, embracing developments of smart individuals and innovative companies alike. And unlike with other operating systems, various computer manufacturers are welcome to have a hand in the direction Linux is taking. The devoted Linux community has, almost by sheer force of will, made Linux the emerging darling of enterprise and productivity applications.

But is Linux a viable OS for real HPC applications, real HPC users, and real HPC workloads? Is Linux finally ready for 64-bit applications that demand big scaling and big memory in physical and life sciences, manufacturing, energy, and government? In other words, is the Linux community succeeding in its efforts to scale this open-source OS beyond its widely perceived limit of eight processors, so it ultimately can drive the robust productivity tools that until now have been available only on proprietary platforms?

The answers are in, and the news is good.

Just as general commercial and enterprise applications have benefited from the unique model that drives Linux development, so will applications written for very large systems, clusters, and superclusters. In fact, important work has been under way for some time, and record-setting HPC Linux systems have recently come to market. SGI, for instance, determined early on that Linux, with a standard kernel and enhanced software and utilities, had the stuff to drive massively scalable systems to solve HPC-class problems. So the company leveraged its experience with NUMA and HPC from its systems based on the IRIX® OS and MIPS® processors and concentrated on the Linux improvements most crucial to HPC environments. SGI also is providing the robust HPC system-, resource-, and data-management tools—via IRIX libraries, tools, and

software packages ported to Linux—to provide a powerful, standards-based high-performance system using Linux and Intel® Itanium® 2 processors.

The results of that work may surprise those who still view Linux as a low-cost alternative to the Windows® operating system. On January 7, SGI announced a family of Linux servers and superclusters capable of scaling to 64 Intel Itanium 2 processors in a single OS image—and up to thousands more across nodes via global shared memory. The new SGI® Altix™ 3000 systems set performance records across the board—in processor speed and scalability, memory bandwidth, I/O throughput, and real-world application performance—proving that Linux indeed has the chops to deliver for scientists, engineers, and other users of advanced technical computing systems.

These new systems combine SGI's supercomputing and HPC expertise with an implementation of Linux that is tuned for maximum performance and scalability, while maintaining binary compatibility with the standard Linux distribution and industry-standard Linux applications. Just as crucial to the final product, however, are key OS enhancements, a high-performance I/O subsystem, and optimized software tools.

OS Enhancements

Much of SGI's work on the OS focused on memory placement, process pinning, and CPU set support. When either solving a complex problem that needs large numbers of CPUs and a lot of memory or when running multiple, unrelated problems simultaneously within the same large system, using the system's resources efficiently enables the application or batch job to complete within a predictable and consistent time period. On Linux, SGI provides the commands 'cpuset' and 'dplace' to help ensure a particular workload or batch job can more efficiently use the available CPU and memory resources. These commands also help ensure that multiple jobs can carve out and use the resources they each need without getting in each other's way and can help prevent a smaller job from inadvertently thrashing across a larger pool of resources than it can effectively use. The underlying support for dplace, cpuset, and runon is implemented using SGI open-sourced CpuMemSets support, which provides the kernel support and infrastructure for implementing these commands and functionality.

Providing extensive system accounting capabilities is often important for very large systems, especially when the system will be shared or made available for other organizations to use. The Comprehensive System Accounting (CSA) software package was an open-source collaboration between SGI and Los Alamos National Laboratory. CSA is focused on providing Linux jobs-based accounting of per-task resource and disk usage for specific login accounts. CSA uses job containers, which on Linux provide the notion of a "job." A job is an inescapable container and a

collection of processes that enable CSA to track resources for any point of entry to a machine (e.g., interactive login, cron job, remote login, or batched workload).

As the demand for larger and larger systems continues to increase for solving complex HPC problems, eventually the OS's ability to manage a large pool of CPUs, memory, and I/O resources can limit performance and become a bottleneck for some workloads. SGI provides the ability to divide a larger system into smaller system "partitions," where each partition runs its own copy of the OS kernel. This is an effective way to overcome bottlenecks, especially for parallel workloads like MPI jobs. Since each of these partitions is still connected using the high-performance SGI® NUMAlink™ interconnect technology, the low-latency and high-bandwidth benefits for communication between processes are not sacrificed when partitioning is used to scale beyond a single system image.

Partitioning not only improves performance but also provides higher availability. By reconfiguring a larger system into smaller partitions and using the partitions together as part of a shared-memory "supercluster," a hardware or an OS failure within one of the partitions, or "cluster nodes," can be contained to prevent bringing down the rest of the system and software running on the other partitions or cluster nodes.

SGI also configured and used the Linux Device File System (Devfs) on its systems to handle large numbers of disks and I/O buses. Devfs ensures that device path names remain persistent across reboots after other disks or controllers are added or removed. This keeps system administrators from having to worry about renaming or renumbering 50 or more disks if a controller fails. In tests, SGI has found Devfs to be reliable and stable in high-stress system environments with configurations of up to 64 processors and with dozens of Fibre Channel loops and hundreds of disks attached.

I/O Subsystem and Data Management Tools

Real HPC workloads require the hardware and software ability to efficiently drive and manage big data. However, such I/O throughput and data management software has been a traditional weak point for early Linux clusters in HPC. Over the past two years, SGI has aggressively tackled these bottlenecks through the systematic tuning and porting of key IRIX applications.

The SGI Altix 3000 family has demonstrated sustained I/O throughput of more than 2GB per second running a single copy of Linux, an industry-leading result for Linux systems. With data sets growing ever larger, the ability to move information from disk to memory is an increasingly important component of overall system performance. This remarkable achievement allows Linux

applications to overcome the increasing challenge of handling big data in high-performance computing by increasing throughput to levels beyond those attained by most UNIX® OS-based systems.

To optimize I/O throughput for its implementation of Linux, SGI ported its IRIX SCSI midlayer, XFS™ filesystem, XVM volume manager, and data migration facilities to provide a robust, high-performance, and stable storage I/O subsystem on Linux. The SGI XSCSI subsystem on Linux leverages IRIX production quality and commercially proven code to provide more robust error handling, failover, and storage area network infrastructure support, as well as years of large-system performance tuning.

The XFS filesystem is a fast recovery filesystem that provides direct I/O support, space preallocation, access control lists, quotas, and other commercial filesystem features. Open sourced in May 2001, XFS was the first journaling filesystem available for Linux. While other filesystems on Linux are available, the years of performance tuning and improvements leveraged from IRIX make XFS particularly well suited for large data and I/O workloads commonly found in HPC environments.

The SGI XVM and CXFS™ clustered filesystem for Linux provide software volume manager features such as disk striping and mirroring as well as a clustered filesystem enabling high-performance sharing of a filesystem among multiple systems clustered together. Since CXFS enables multiple systems to access the same filesystem data simultaneously, CXFS improves data availability in high-availability cluster configurations. Also, the CXFS filesystem's heterogeneous support among Linux, IRIX, Solaris™, and Windows NT® (with AIX® and HP-UX® on the way) enables high-speed data sharing and interoperability between mixed OS and HW environments. Also, using the SGI cluster high-availability server environment along with CXFS on a partitioned system enables filesystems to be efficiently shared between partitions and provides improved availability for other services like NFS and database servers.

The SGI® Data Migration Facility (DMF) for Linux provides the ability to work with very large data sets that can exceed a system's online disk capacity. DMF transparently migrates XFS file data between disks and offline media. DMF enables high-performance I/O subsystems to work with large data sets that would otherwise overrun a system's online disk resources. For offline storage support, DMF works with any media transport and robotic automounter supported by SGI® OpenVault™ or Tape Migration Facility for Linux.

HPC Application Tools and Support

The record-setting scalability of the Linux environment from SGI also is aided by an exclusive suite of SGI tools and features designed to maximize performance on the most demanding technical and scientific applications. The middleware layer of the SGI ProPack™ 2.1 for Linux® open-source feature includes several tools and libraries to help improve performance on large NUMA systems for solving a complex problem with an application that needs large numbers of CPUs and a lot of memory, or when multiple, unrelated applications are running simultaneously within the same large system.

A key part of SGI ProPack™, the SGI Message Passing Toolkit (MPT), provides industry-standard message-passing libraries optimized for SGI computers. On Linux, MPT contains MPI and SHMEM APIs. These transparently utilize and exploit the low-level capabilities within SGI hardware, such as its block transfer engine (BTE) for fast memory-to-memory transfers and the hardware memory controller's fetch operation (fetchop) support. Fetchop enables direct communication and synchronization among multiple MPI processes while eliminating the overhead associated with system calls to the operating system. Also, an explicit call within the MPT library, XPMEM, is part of the breakthrough ability of these new Linux clusters to share memory across nodes interconnected by NUMAlink.

SGI has also ported its famous SCSL math libraries from IRIX to Linux, enabling a broad range of optimizations on the most demanding scientific algorithms.

Parallel workloads, such as MPI jobs, can be launched, monitored, and controlled across a cluster or partitioned system using SGI Array Services software. Array Services provides the notion of an array session, which is a set of processes that can be running on different cluster nodes or system partitions. Both the job container and Array Services are implemented using process aggregates (PAGGs), which are kernel modules that provide process containers. CSA, the job container, and PAGGs have all been open sourced by SGI for Linux.

The SGI ProPack NUMA tools, HPC libraries, and additional software support layered on top of a standard Linux distribution provide a powerful HPC software environment for big compute and data-intensive workloads. Much like a custom ASIC on hardware providing the “glue logic” to leverage and use commodity processors, memory, and IO parts, SGI ProPack software provides the glue logic to leverage the Linux operating system as a commodity building block for large HPC environments.

The Proof: Real-World Performance

None of these advances would matter if they didn't enable Linux systems to achieve the scalability required by users in the physical and life sciences, manufacturing, oil and gas, and government and defense markets. Fortunately, that's precisely what SGI's implementation of Linux is designed for—and delivers. At its debut, the SGI Altix 3000 family systematically set world-record performance and price/performance marks across all four traditional dimensions of HPC: compute, memory, I/O, and scalability.

First, let's consider compute performance. In recent tests, the SPECfp_Rate_base2000 benchmark on a 64-way SGI® Altix™ 3700 supercluster with 1.0 GHz Itanium 2 processors turned in a screaming score of 862. The closest 64-way single system image competitor was the HP Superdome™, running its top-of-the-line PA-8700+ processor, with a score of 267. In the 32-way arena, SGI Altix 3000 again dominated with a score of 443.IBM's top-of-the-line pSeries™ 690 currently has a score of 251, and HP Superdome has a score of 128. For those with budgets to consider, the results reveal a 3x price/performance advantage over the IBM® Power4™ systems.

Now onto memory: The STREAM Triad Benchmark, which measures the system's memory bandwidth performance, also reveals impressive performance by the SGI system. The 64-way SGI Altix 3700 reported a bandwidth of 125GB per second on a single system image, while the 64-way HP Superdome reported bandwidth of 27GB per second. In the 32-way arena, IBM pSeries 690 reported a bandwidth of 32GB per second, while SGI doubled that number with a 32-way SGI Altix system. This represents a 3.2x price/performance advantage.

I/O results also lead the industry. The SGI Altix 3000 family is the only Linux solution to push I/O beyond 2GB per second, easily breaking the typical Linux threshold of 300–500MB per second and soundly beating many proprietary OS platforms.

And finally, there's scalability. SGI has shown yet again that its NUMAflex™ architecture and OS engineers can scale like no one else. In this case, that means taking a host of industry applications from manufacturing, sciences, and energy all the way up to 64 processors on a standard Linux distribution.

Application performance results are equally impressive. In recent tests using CD-adapco Group's STAR-CD™ test suite, the SGI Altix 3000 family demonstrated unparalleled application performance and scalability. Performing an exhaustive array of computational fluid dynamics tasks, SGI systems easily outperformed competing IBM servers by a factor of 2x in

price/performance, underscoring the SGI Altix 3000 family's competitive advantage over systems with less balanced architectures.

In other shared-memory applications that demand leading CPU compute power, memory bandwidth, and I/O throughput, SGI Altix 3000 delivered overall outstanding performance. The SGI system outperformed the IBM eServer p690 system in measuring real-world application performance of Gaussian®, an advanced technical application that allows scientists to predict energies, molecular structures, and vibrational frequencies. Comparing 32-processor systems, the SGI Altix 3000 running Gaussian v98 delivered a 50% performance increase over that of IBM eServer p690 (again, for the budget-conscious, this represents a 2x price/performance advantage).

Linux has obliterated its eight-processor limit. Thanks to continuing efforts of the Linux community, it is maturing rapidly, opening new opportunities to users and developers who want to stretch its capabilities beyond commercial and enterprise applications. SGI's work is indicative of the flexibility of this elegant operating system—and the enormous opportunities in global shared memory and breakthrough price/performance that await those who hope to deploy open-source solutions for true HPC applications.

Steve Neuner is the Linux engineering director at SGI, and is focused on deriving optimum performance from systems based on Linux and Itanium. Jason Pettit is product manager for Linux solutions at SGI.