Linux and Compaq's GS320 Server

GREASED LIGHTING

PETER RIVAL

Venturing beyond servers, computing clusters and embedded systems, Linux now reaches new domains – super computing! After several announcements from Hewlett Packard (SuperDome), SGI (Origin 3000-IA64) and IBM (Numa-Q), Compaq/DEC is one of the first manufacturers to demonstrate Linux running on a super computer platform.

The behemoth stands almost 2 meters tall and 3.5 meters wide and weighs in at over 2000 pounds. Its green panel reads "Cpu-32 Mem-256 Pci-64". Then you realize that it's running Linux. As disk lights blink and the two large drum fans hum, the reality of Linux in the enterprise hits home. The system is certainly a far cry from the 80386-based PC that Linus Torvalds first used to write Linux. But then, Linux today is a far cry from the first version posted all those years ago.

The Hardware

Compaq first introduced the AlphaServer GS series earlier this year as the new high-end of their AlphaServer line. Architecturally, the system is a unique mix of supercomputer and general business computer. It is the first ccNUMA (cache-coherent Non-Uniform Memory Access) system in the AlphaServer line, eschewing the traditional SMP model for the scalability advantages of a NUMA architecture. The higher CPU count, memory capacity and I/O bandwidth also represent quantum leaps forward in the AlphaServer line. Perhaps the most unique aspect of the new GS series is the completely modular "building block" design, enabling greater scalability and making future upgrades easy.

The cornerstone of this modularity is the Quad Building Block, or QBB. Each QBB contains up to four Alpha EV67 CPUs, four memory modules containing up to 8 GB each, an I/O port, a Global Port for attaching to other QBBs and a high-speed switch to connect all of these modules together (see Figure 1). The EV67 CPUs currently clock at 731 MHz, with upgrades in the near future to over 1GHz. Each QBB with attached PCI I/O drawer, is fully capable of running a separate instance of Tru64, OpenVMS or Linux. One of the most critical design decisions in the QBB is that there is no trade-off between CPU slots and memory slots as is common in many high-end systems - a fully

configured QBB can have the maximum of both CPUs and memory. Each QBB provides 6.4 GB/sec of memory bandwidth, 1.6 GB/sec of I/O bandwidth and 17.6 GB/sec of raw interconnect bandwidth.

Two QBBs communicate directly through their Global Ports. When connecting more than two QBBs a high-speed, low-latency Hierarchical Switch is used. This switch provides many critical functions, including maintaining a cache coherency directory and enforcing hardware partitions. The bandwidth between two QBBs is 1.6 GB/sec in each direction. The bandwidth increases by a factor of four with each QBB added; a fully configured GS320 provides an aggregate bandwidth of 51.2 GB/sec.

The two-level switch hierarchy is responsible for the linear scalability of per-CPU bandwidth, as compared to standard bus-based systems where the more CPUs that are added, the less memory bandwidth each has available to it. This switch hierarchy is then matched with a highly scalable memory interleave strategy, aggressive memory resource scheduling and aggressive data link bandwidth management to provide a huge memory system bandwidth with a capacity to handle hundreds of outstanding references. Because of this design, the impact of remote memory accesses to be lower than many other ccNUMA products - just under a factor of 3:1 over local memory access (330 ns local access latency vs. 960 ns remote latency).

These components are impressive separately, but when integrated into a fully configured GS320, they bring their separate technologies are consolidated. Combined with the QBB and Hierarchical Switch in a complete system are a fully redundant and hot-swappable power subsystem, a collection of master and slave PCI drawers including software power control and a host of environmental controls to help ensure peak availability. A comparison of various GS series configurations can be found in Table 1.

The Software

"I remember being intimidated by a machine that takes 15 minutes to power-up, and has a console that manages consoles". Perhaps this sentiment, by Jay Estabrook of Compaq, sums up the state of Linux on massively scalable systems at this point - particularly in light of the fact that the porting work has been completed. The system is currently being used for performance and NUMA optimization work, the impact of which should soon be seen in the Linux kernel itself.

Initially, the porting work was spurred on by an encounter at CeBIT in Germany between Andrea Arcangeli, Anas Nashif and Stefan Fent of SuSE and David Mitchel of Compaq. It just so happened that a prototype GS system normally used for training was available and all of the players were able to join together. Joining Andrea, Anas and Stefan were Jay Estabrook and Larry Sendlosky of Compaq. The whole process of reaching a complete boot on this system took only a single week of full time work. As Andrea mentions, "I finally reached the alpha userspace for the first time using an ATAPI bootable CDROM in the last hour of the last day we had available".

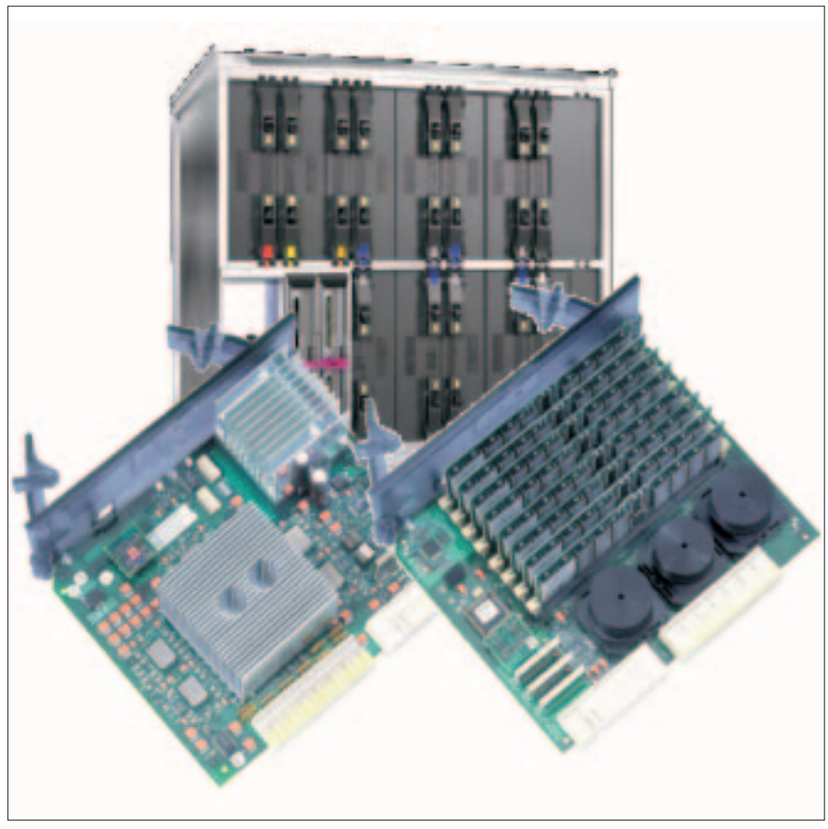

Figure 1. Quad Building Block Components (CPU-Board and Memory-Unit)

From this initial success, more work was still to be done to get the kernel to boot correctly on a revenue-class system configured with 16 CPUs and 16 GB of memory. At this point, this was the largest single Alpha system Linux had ever attempted to boot on. Finally, after a few weeks of work and some hands-on help from the GS series architects, a GS160 was fully useable under Linux. It was believed by most that knew of the achievement at the time that this was about as high as Linux would scale without heavily involved re-architecting of core components of the kernel. At this point, sights were set on discontiguous memory and proper NUMA support.

Just as coincidentally as the original port had started, the author was asked if he would like to try

Comparison of System Features

Feature	GS80	GS160	GS320
QBBs	2	4	8
CPUs	8	16	32
Partitions Supported	2	4	8
Memory (GB)	64	128	256
Bandwidth (GB/sec)	12,8	25,6	51,2
PCI boxes/slots (64 bit)	4 / 56	8 / 112	16 / 224
Bandwidth (GB/sec)	3,2	6,4	12,8

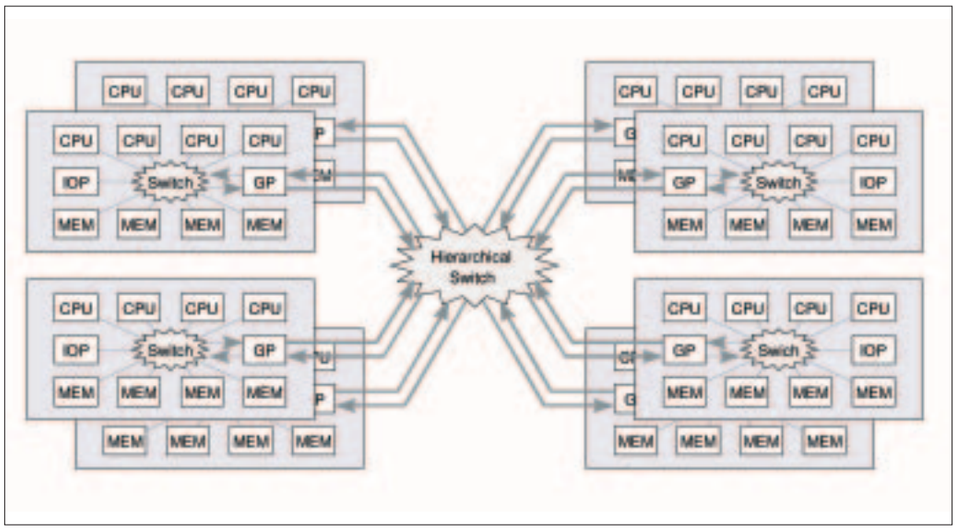

Figure 2. Eight QBBs connected via hirachical switches. With this configuration, Linux is able to manage 32 CPUs and 256 Gigabyte of main memory

to see if Linux would boot on a prototype GS320. Never one to pass up an opportunity such as this, a bootable disk was quickly created and the first attempt was quickly made. After several painful minutes as the system built up various memory management maps and other tables for a system of this size, a login prompt was on the screen. Despite even the most hopeful assertions that some changes would be necessary, the same kernel that had booted on a GS160 worked cleanly on a fully configured GS320.

Future Directions

Even as the computer industry continues to grow and shift at a blindingly rapid rate, Linux seems to grow and shift at an even faster rate. While simultaneously making strong inroads into the handheld and embedded markets, Linux is quickly evolving into a strong contender in the enterprise market. While still lacking some of the enterprise reliability and management components of other commercial UNIX offerings, the availability of Linux on such high-end systems as the GS320 and those from IBM, HP, SGI and Sun should help to rapidly increase the development of such tools.

Still to come for the GS series is support for discontiguous memory and NUMA optimizing various critical parts of the Linux kernel. Despite the fact that Linux runs on a GS320, and in fact is quite useable, the system will not be ideally used without supporting these critical technologies. While Compaq has currently not committed product support on the GS Series, Linux running on such high-end systems provides a base for further enterprise capability development and hopefully help Linux on all systems large and small.

In particular, Compaq will make use of Linux on the GS320 to analyze aspects of Linux such as lock scaling and algorithm assessment. These investigations should lead to advances in both performance and stability even in smaller systems. For example, locking designs that do not scale at the high end also usually are not scaled well on lower-end systems, but while the impact is less noticeable it is usually still present. Also, with all of the RAS features inherent in the GS series, this presents a prime opportunity to extend the support for advanced technologies such as hot-swap and hot-add components as well as system and application partitioning.

The Compaq AlphaServer GS320 has been designed with scalability and reliability as core components. In introducing the GS series of Alpha servers, Compaq has extended the high end of their server line to a new level. Linux support of this architecture will provide a critical opportunity for enhancing enterprise-level RAS features in Linux as a whole, as well as helping to bring Linux to new levels of scalability. Even as fast as the computer industry is moving, Linux is more than keeping pace. ■

```
[root@sundown2 /root]# cat /proc/cpuinfo
cpu : Alpha
cpu model : EV67
cpu variation : 7
cpu revision : 0
cpu serial number :
system type : Wildfire
system variation : 0
system revision : 0
system serial number :
cycle frequency [Hz] : 730794500
timer frequency [Hz] : 1000.00
page size [bytes] : 8192
phys. address bits : 44
max. addr. space # : 255
BogoMIPS : 1488.97
kernel unaligned acc : 0 (pc=0,va=0)
user unaligned acc : 0 (pc=0,va=0)
platform string : Compaq AlphaServer GS320
6/731
cpus detected : 31
cpus active : 31
cpu active mask : 00000000ffffff7f
```

Impressive: cpuinfo of a GS 320

Read Further

The initial boot log of Linux on a GS320 can be found here: http://www.alphanews.net/?op=displaystory&sid=2000/8/28/141244/404

More information on the AlphaServer GS series can be found here: http://www.compaq.com/alphaserver/gs_series.html

■

The Author

Peter Rival is a Software Engineer at Compaq Computer Corporation. He spends most of his time trying to wring cycles from the Tru64 and Linux kernels, as well as being a Linux advocate both within and without Compaq. He can be reached at frival@zk3.dec.com.