RETROSPECTIVE: The SGI Origin: A ccNUMA Highly Scalable Server

James Laudon and Daniel Lenoski
Google, Mountain View, CA

In the early 1990s most computers were single-processor systems. Multiprocessor systems were the province of high-end computing: supercomputers and enterprise servers. In this environment, SGI (Silicon Graphics) offered its multiprocessor supercomputers. These systems were originally designed to support the large compute and IO demands of their most advanced graphics hardware but over time became a large market of their own. Before the Origin 2000, SGI's supercomputers were bus-based systems employing snoopy cache coherence. Origin's predecessor, the SGI POWER Challenge [5] pushed the bus-based approach to its limits. With its single-core CPU chips, POWER Challenge systems were physically large. Their large size along with the limited bisection bandwidth of the shared bus made it fairly clear that significant scaling beyond POWER Challenge's maximum of 36 processors would not be possible. In 1993, a small team, including the authors, was tasked with incorporating recent directory-based scalable shared memory multiprocessor research into a commercial system that would scale 1-2 orders of magnitude beyond POWER Challenge.

As members of the Stanford DASH project [O7][1], the authors were part of the explosion of research in this area in the 1980s and early 1990s and have fond memories of their friendly rivalry with the Alewife [O1] team and late-night ISCA discussions on COMA [4] vs ccNUMA with their colleagues from Sweden. Much of the architecture of DASH was carried forward into the Origin 2000, augmented with new features including those required by a true commercial offering. Some of the more notable features include:

- The SPIDER network router ASIC [O4] with 6 high-speed bi-directional links supporting adaptive wormhole routing and four virtual channels per physical channel
- A Hierarchical Fat Bristled Hypercube network topology
- A focus on minimizing both the local memory latency and the ratio between remote and local memory latency
- Equal bandwidth to remote and local memory
- A directory format capable of maintaining cache coherence across 1024 processors (512 nodes)
- Support for efficient page migration and TLB shootdowns
- A rich set of synchronization primitives including both fetch-and-op and load-linked/store-conditional
- A network-ordering-insensitive coherence protocol with full MESI support (DASH was MSI); and including an efficient S→E transition (effectively a dynamic O state)
- The first coherence protocol to be fully formally verified for three nodes [O3]
- The nearly 1 MILLION gate Hub ASIC (raises Dr. Evil pinkie[2])
- The Xbow IO crossbar ASIC with 8 high-speed bi-drectional links, wormhole routing, two virtual channels per physical channel, full coherence with system memory, bandwidth allocation, and block transfer DMA engines

The Origin was a commercial success, selling thousands of systems to a wide range of customers, including research and financial institutions and high-performance computing centers. Notable Origin deployments include a pair of 512-processor systems at NASA Ames, and the ASCI Blue Mountain system [1], built from 48 HIPPI-connected 128-processor Origin systems and installed at Los Alamos National Laboratory to build a simulator to replace live nuclear weapons testing. The Blue Mountain system, shown in Figure 1, highlights the large multiprocessor physical size challenges of the time, where 6144 processors (which would fit in a couple of datacenter racks today) required a full machine room floor.

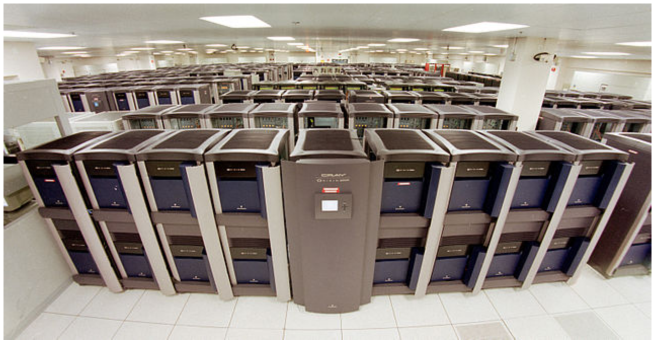

Figure 1 ASIC Blue Mountain Supercomputer[3]

Despite this commercial success, directory-based shared-memory systems were at their peak in popularity around the time of the Origin 2000. There were both technical and business reasons behind the shift away from directory-based shared-memory supercomputers.

One of the foremost goals of cache coherent shared-memory systems like Origin was to make parallel programming easier by freeing the programmer from maintaining coherence in software and from worrying about data placement. While distributed shared memory systems succeeded in eliminating the need for software coherence, they fell short on the promise of removing data placement concerns. Even with Origin's low 2:1 to 3:1 (depending on total processor count) remote to local memory latency and hardware support for efficient page migration, writing code oblivious to data placement left too

[1] [O..] will refer to the bibliography in the original paper

[2] Shamelessly borrowed from our colleague Cliff Young

[3] Los Alamos National Laboratory (https://commons.wikimedia.org/wiki/File:Blue Mountain Supercomputer.jpg), „Blue Mountain Supercomputer“

much performance on the table. Origin systems were expensive and typically purchased to tackle high-value scientific and business problems that required large amounts of compute, memory, and/or IO bandwidth. Hero programmers who were able to squeeze the last bit of performance out of the hardware, compilers, runtime systems, and OSes often programmed them. These programmers spent considerable time thinking about data placement, and for them, supercomputers built with a non-coherent global address space provided an adequate mix of performance and ease of programming. In addition, while early message-passing frameworks could be challenging to use, the ease of use of MPI [3] and other message-passing programming infrastructure improved over time.

Origin was also not immune from the attack of the killer micros [6]. At Origin's launch, small-scale two- and four-processor workstations could handle much of the lower-end work, and SGI sold separate computer lines at this scale concurrently with Origin. One- and two-socket x86-based systems were starting their long ascent into domination of the server space and with their steady growth in core count over time, two-socket Intel/AMD servers today are approaching the core count of larger Origin systems.

While directory-based shared memory systems may have been on the decline, they certainly were not out for the count. The Origin 2000 was succeeded by the MIPS-based Origin 3000 and the Itanium-based Altix 3000, and the blade-based 4000, with this final offering discontinued in 2006. In our paper, published in 1997, we compared the Origin with three of its contemporary ccNUMA systems: the SCI ring-based Sequent NUMAQ [O8] and Data General NUMALiiNE, and the Convex Exemplar X [O2]. Sequent and Data General continued to sell ccNUMA systems until they were both acquired in 1999 (Sequent by IBM, Data General by EMC) and their computer lines discontinued a few years later.

The Convex Exemplar X proved to be the hardiest of the breed. HP had acquired Convex in 1995 and incorporated its directory-based coherence into its high-end Superdome [2] servers. Superdome servers still sell today and support large database applications like SAP Hana with up to 32 sockets and up to 48 TB of main memory, maintaining coherence through a directory cache. Convex is fairly unknown today despite having a significant influence on the computer industry. Convex helped start the minisuper computer boom of the 1980s with the release of their Cray–like C1 mini-supercomputer in 1985, and then after several follow-ons which increased the maximum core count from 1 to 8 processors pivoted to a completely different directory-based ccNUMA design with the Exemplar line. As an aside James worked on systems competing with Convex twice in his career: first as an undergraduate intern working on the Astronautics ZS-1 [7] which competed with the original Convex C1 and later on the Origin which competed with the Exemplar X.

While distributed shared memory systems never took over the world like we envisioned in those late night ISCA discussions, some of Origin's technology still exists today. Directory-based coherence is employed in specific situations in today's multicore and multi-socket systems. For example, directories are often used to maintain coherence between multiple levels of caches where the higher level caches are shared by multiple caches lower in the memory hierarchy. The ccNUMA protocols supported in today's 2+ socket servers (e.g. Intel's Ultra Path Interconnect aka UPI) are another example of directory-based coherence being used at a small scale.

Today's CXL IO subsystems support a CXL.cache mode where IO devices fully participate in the memory coherence protocol as did Origin's Xbow-based IO subsystem. The latest CXL 3.0 enhanced coherence protocol has a directory based coherence protocol supporting "Back Invalidates" that distributes the coherence work across the CXL devices.

Working on Origin 2000 was fun and rewarding. Many on the team felt we were a key driver of a major shift in the computing industry. There were plenty of long nights and the occasional disappointment of having to drop a particular feature to meet chip budgets or schedule constraints, but the team made steady progress, inventing new tools and techniques along the way. After Origin 2000, while several team members embarked on the design of the 3000, many of the team members moved on to projects in other areas at SGI or other companies, including the large number of networking and graphics chips startups that were the next big thing at the time such as Juniper Networks and Nvidia. While this natural dispersion of the team continued over the years, for many years after the project finished there was an annual Origin reunion lunch held at one of the team's favorite Chinese restaurants on Castro Street in Mountain View.

We're very grateful that SGI took a large, calculated risk on an emerging technology when launching the Origin project and to have worked with such a fantastic team early in our careers. It's also gratifying to see that risk-taking spirit alive today with the large number of innovative projects and startups pursuing their particular vision for machine learning acceleration. We look forward to seeing what the next round of innovators come up with to tackle whichever major shift the computing industry holds next.

REFERENCES

[1] https://en.wikipedia.org/wiki/ASCI_Blue_Mountain

[2] Gostin, G., Collard, J.-F., Collins, K., 2005. The Architecture of the HP Superdome Shared-Memory Multiprocessor. In ICS '05: Proceedings of the 19th Annual International Conference on Supercomputing, pp. 239–245.

[3] Gropp, W., Lusk, E., Doss, N., and Skjellum, A. 1996. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, vol. 22, no. 6, pp. 789-828.

[4] Hagersten, E., Landin A., and Haridi, S., 1992 DDM-a cache-only memory architecture. *Computer*, vol. 25, no. 9, pp. 44-54.

[5] Loos T. and Bramley R., 1996. MPI Performance on the SGI Power Challenge. In MPIDC '96: Proceedings of the Second MPI Developers Conference, pp. 203-206.

[6] Markoff, J., 1991. The Attack of the 'Killer Micros'. In New York Times May 6, 1991 Section D, Page 1

[7] Smith, J.E., Dermer, G.E., Vanderwarn, B.D., Klinger, S.D., Rozewski, C.M., Fowler, D.L., Scidmore, K.R., and Laudon, J.P., 1987. The ZS-1 Central Processor. In Proceedings of the Second International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS).