

High Performance Computing Meets Databases

Gordon Haff

Copyright © 2006 Illuminata, Inc. Licensed to Silicon Graphics, Inc.
e-mail: ghaff@illuminata.com

(Rec. 17 February 2005)

Key words: hpc, databases, data analysis, data mining, optimization, top500, clusters, smp, commercial computing, enterprise datacenter

1. INTRODUCTION

Many of the earliest computers were equally at home crunching ballistics tables or census data, solving technical problems or handling business tasks. But, over time, high performance computing (HPC) came to be increasingly dominated by technical computing specialties that bore little resemblance to the business workloads and systems of the day. Once again, however, many system architectures designed for HPC workloads are converging with those used for commercial computing.

In part, this is because many of the tasks themselves have converged. Data analysis and mining, credit risk analysis, and optimizing product delivery logistics are just a few “business intelligence” jobs whose characteristics resemble many of the large simulation models run in research labs. The economics of semiconductor design and fabrication are another major factor. HPC-specific technologies such as “vector” processors have lost the war against more general-purpose CPUs. Mainstream processors can amortize development and production costs across much larger product volumes. Those superior economics, in turn, make more R&D funds available, speeding the advance of general purpose parts. As a result, specialist engines face increasingly stiff competition, especially when it comes to price/performance.

One can point to many examples of the renewed marriage of HPC and commercial IT. Many large supercomputing clusters, for example, look remarkably similar to Web farms. Both are frequently constructed from dual-processor x86 servers connected by Gigabit Ethernet¹. The trend extends up-market to SMP servers running DBMS workloads. The IBM p5 575, for example, though designed

for HPC cluster requirements, is still based on general purpose POWER5 processors and other standardized components. Moreover, it is increasingly used for commercial workloads, especially business intelligence tasks running DB2². Even SGI, a company that has for several years very deliberately focused on its technical computing sweet spot, is re-entering commercial computing specifically targeting large databases and associated applications – using its Altix SMP line³.

2. THE PENDULUM SWINGS BACK

All but gone from the TOP500 list of the world's largest supercomputers are the once-dominant vector machines. They've been replaced by systems running speedy general-purpose microprocessors, or their derivatives⁴. Most of today's systems use *scalar* processors, so called because they historically handled one item of data at a time. In contrast, *vector* processors could apply operations to a large quantity of data in a single operations⁵ – a very efficient way to handle structured number crunching.

By the early 1990s, vector supercomputers could still churn out more gigaflops than anything else, but RISC microprocessors began to approach vector machines' efficiency – especially their ability to complete an operation on every machine clock cycle. This is especially true of “superscalar” processors such as IBM POWER and Sun UltraSPARC that can finish multiple operations on every clock. Coupled with increasingly smart compilers, clever

¹To be sure, the software that schedules and integrates the jobs differs considerably even if the underlying hardware and operating systems are often the same.

² See our “POWERing the Performance Factory” and “IBM's DB2: Database for Enterprise Masses”.

³ See our “SGI Brings Big Iron to Linux” and “Altix Goes Modular”.

⁴ IBM's BlueGene, for example, uses a custom processing module built around a PowerPC core. See our “Blue Gene's Teraflop Attack”.

⁵ In truth, even vector processors don't handle every data element in parallel, but a single operation can queue up long strings of highly-efficient computation.

optimizations of caches and memory subsystems, and tightly-coded routines for key mathematical algorithms, individual off-the-shelf RISC servers performed like mini-supercomputers. Such optimizations also laid the groundwork for Intel's Itanium processor, which further shifts the optimization burden from chip hardware to compiler software. Moreover, many scalar CPUs have been augmented with SIMD instructions, which are essentially auxiliary vector units⁶.

In one sense, the latest TOP500 statistics⁷ reflect the death of the supercomputer – but only if one narrowly defines “supercomputer” in terms of past dedicated scientific computing machines from the likes of Convex, Cray and MasPar. In another broader and more meaningful sense, the supercomputer has never been healthier. Leveraging components used in at least moderate volumes elsewhere in the industry lets designers build high performance systems that are more generalized – and far cheaper – than ever before.

To be sure, TOP500 is a rarified list. Its systems scale to heights rarely if ever, seen in commercial computing. But towering scale points notwithstanding, much of what is happening at HPC's frontiers also reflects the realities of today's enterprise datacenters. The TOP500 includes plenty of plain vanilla clusters that string together off the-shelf dual-processor servers with off-the-shelf Gigabit Ethernet. Some improve the links with optimized interconnects such as InfiniBand, but the basic off-the-shelfness remains.

But the largest TOP500 entrants eschew “volume” servers entirely instead leveraging more capable SMP nodes – servers like IBM's p5 575 and SGI's Altix 300/350 that are, even individually, enormously powerful systems. Such “fat node” clusters handle some of the toughest computing problems on the planet.⁸

3. CLUSTERS ARRIVE

The yin and yang of scale-out and scale-up are hardly unique to HPC; the same alternatives play out in commercial datacenters⁹.

Oracle and IBM have been the most active developers of clustered database technology. Indeed, in 2003, Oracle made a strategic push to move away from its Big Iron roots towards a more distributed approach built around “com-

modity” servers.¹⁰ Oracle had offered a parallel version of its database – historically named Oracle Parallel Server (OPS) – since the mid-1980s. However, the product's hugely complex installation and configuration requirements, not to mention its limited scalability kept it from being more than a tangential part of Oracle's line.

With version 9i, Oracle started afresh. First it rebranded OPS as Real Application Clusters (RAC), in an effort to put all the historical baggage associated with OPS behind it. Moreover, RAC represented a considerable overhaul, including an infusion of technology and expertise from then Compaq's Tru64 cluster team. Each Oracle RAC instance runs on its own node (that is, server), which maintains a local cache of recently accessed data. The trick to making a distributed architecture work is efficiently maintaining a single, correct logical view of data across the distributed caches. It's not easy, and it gets much harder as the number of nodes grows. RAC's mechanism, Cache Fusion, replaces and dramatically improves upon the Distributed Lock Manager (DLM) of the earlier OPS architecture¹¹.

As a result of these technical advances, distributed database scalability has improved considerably. Standard interconnect technologies such as InfiniBand have also made low-latency communications – a *sine qua non* of cluster scalability when any serious amount of coordination is required – much more widely and economically available. The result is that clustered databases are much more achievable.

4. BIG IRON AND DATABASES: STILL MARRIED

However, clusters have never displaced scale-up SMP. Many datacenters are growing rapidly to handle the relentless surge in transaction rates and analysis demands created by the confluence of trends from the Web to mobility to real-time decision making to the digitization of just about everything. But they're hardly spending like drunken sailors, unlike the Internet boom spending spree. Operational efficiency and business value are watchwords. Businesses will spend a bit more initially to structurally cut costs down the road, but the value must be clear.

Unlike HPC, where clustering is often about hitting an incredibly high scale point at pretty much whatever cost, database clusters are primarily about scaling efficiently – in other words, saving money. The savings may be about limiting up-front capital expense by allowing modular expansion as data volumes and requests grow, or it may be

⁶ Intel's SSE2 and PowerPC's AltiVec are microprocessor-based examples targeted at graphics and multimedia functions.

⁷ TOP500 is an oft-quoted list that attempts to rank the world's largest supercomputers. See our “Reflections on a List”.

⁸ Even larger systems can be aggregated into clusters – the Columbia Supercomputer NASA's Ames Research Center clusters 30 512-processor Altix 3000 servers – but such installations push even the boundaries of TOP500-class scale; it's far more common for such servers to fly solo even for considerably compute-intensive HPC tasks.

⁹ See our “Scale Up vs. Scale Out: The Fallacious Dichotomy”.

¹⁰ Although this discussion centers on Oracle, much of the same storyline applies to IBM's DB2-even if, characteristically Oracle has made far more grand, public proclamations on the subject.

¹¹ It works very differently. For instance, it lets nodes semi-autonomously transfer data among themselves, not just share a centralized view of outstanding locks.

about handling particularly large database problems with attractive price-performance. After all, published OLTP benchmarks show that high-end SMP servers can exceed three million transactions per minute. Yet there are precious few (if any) organizations on the planet that need this number of commercial transactions from any single application or database. Extreme business intelligence and data warehousing workloads – think the logistics operations for a very large retailer, for example – are more likely to strain the boundaries of workaday SMP scale. And that is where DBMS clusters shine the brightest.

Clusters have always carried a complexity “tax”. However attractive the list price, the special configurations, training, and procedures represent a significant hurdle to acceptance. Over time, this tax has decreased, but it's never gone away. If nothing else, there are just more system images to manage. Clusters also still require attention to how nodes are configured, connected, and tuned. And while we're a few years beyond having to manually partition data to evenly distribute database workload, selecting the right indexing and query strategies typically requires a fair bit of analysis and ongoing adjustment¹².

Even beyond matters of money and performance, many IT shops just prefer scale-up SMP as a matter of cultural preference, technological familiarity, and simplicity. Or they may run a database such as Sybase, which is still popular in the financial sector, but has never seriously traveled down the clustering path.

5. HEFTY NODES FOR HEAVY LIFTING

The result is that by far the greatest swath of today's transactional databases run on but a single server, which is generally simpler to manage and operate¹³. For workaday applications, even basic two- or four-processor x86 servers may deliver plenty of throughput. But as workloads grow, scaling isn't as simple as adding an additional server. The server itself must grow. This scaled up commercial “Big Iron” often looks little different from the “fat nodes” of HPC.

A multitude of different system elements contributes to performance on a given database. There's the processor itself, of course, not just its raw processing speed but also other features such as caches – which help keep more data closer to the CPU and therefore more quickly accessed than main memory. Large SMP nodes tend to utilize *very* large (multi-megabyte) caches¹⁴.

Latency, the speed of accessing data, drives *many* aspects of what runs a database quickly. A package sent through the mail might contain more bits than a typical voice call, but for rapid coordination, the telephone is far more effective than the post box. SMP backplanes have access latencies measured in hundreds of nanoseconds – some 10× or more better than what one finds in the best InfiniBand connected clusters, and 100× or more better than what Gigabit Ethernet networks deliver¹⁵.

In addition to the speed of the internal links and the time that it takes to retrieve data from memory, there's the amount of physical memory configured. Touching memory is always *much* faster than going to disk – which will take milliseconds at best, more than a thousand times the pokiest memory access. Large memory configurations that keep more data in memory can therefore enhance performance considerably. How much depends a fair bit on the access patterns; in-memory buffers are most effective when the same locations are accessed multiple times. Fat SMP nodes permit very large memory configurations – tens or hundreds of gigabytes of RAM, accessed over very high-bandwidth pipes.

In Oracle's case, the System Global Area (SGA) is the primary in-memory database structure that helps reduce disk accesses. Its regions include the default buffer cache that stores data blocks when they are read from the database, the keep buffer cache that DBAs use primarily to hold frequently referenced lookup tables that should always be kept in memory for quick access, and the shared pool that holds object structures and code definitions, as well as other metadata. An even more direct way to speed database accesses using memory is to run the entire database in memory. Today, this remains a specialized niche although one that's relevant for certain “real-time” environments.

Although high latencies are often a bigger issue than insufficient bandwidth, the size, speed, and number of a system's various connecting pipes still remain important figures of merit. Transfers of large blocks of sequential data are a canonical example of a bandwidth-hungry workload – one that's been commonly more associated with business intelligence than transaction processing. However, the boundaries between traditional OLTP and real-time data analysis are blurring. Transactional databases are getting richer and bigger. Myriad endpoints like RFID tags are increasing transaction rates in some environments exponentially. All these drive commercial bandwidth needs as

¹² That is, analysis and effort beyond the already large and complex project of establishing enterprise-scale data architectures on which many data warehouses and analytic apps depend.

¹³ Many DBMS servers are, in fact, run as part of a two-node HA clusters. However, this configuration is usually primarily for high availability failover rather than performance.

¹⁴ There are many competing theories about where to put, and how to organize, cache. With Itanium, for example, Intel has put very large

caches directly on the CPU die, where they can be most rapidly accessed. IBM's POWER5 designs take a different approach; they deploy much larger caches, but then must locate them on a separate die, which entails longer access times. In both cases, however, the caches are much larger – not just percentages, but integer factors of what one finds in the x86 world.

¹⁵ See our “Latency Matters!”

well – just as huge petroleum reservoir datasets or ultra-granular weather models do in technical computing. Fat SMP nodes can move many gigabytes (or tens of gigabytes) of data to and from processors every second, and also from memory to and from I/O devices. While such rates can be seen in aggregate in large “thin node” clusters, “fat nodes” are optimized for heavy-duty data moving and I/O throughout.

Finally if more speculatively, other innovations developed for HPC are starting to garner interest in the commercial space. FPGAs (Field Programmable Gate Arrays) are used to accelerate the performance of performance – critical algorithms such as Fast Fourier Transforms (FFTs) in technical workloads¹⁶. DBMSs likewise have a variety of critical algorithms – such as those that handle indexing, text search, and pattern recognition routines – that could potentially be sped up considerably by this approach, although doing so would require ISV support. An SGI customer is experimenting with RASC¹⁷ for database acceleration, and the company plans an Altix Information Management “appliance” that will use RASC to accelerate database queries and data ingestion. Although acceleration appliances have never enjoyed the widespread adoption their makers hoped for during the Internet boom, examples such XML processing engines from DataPower – a company acquired by IBM in 2005 – nonetheless indicate continuing interest in, and development around, the approach.

6. CONCLUSION

Many commercial tasks now closely resemble their once exotic technical relations. The hardware that runs them is likewise less divergent than in times past. We’ve previously discussed, for example, IBM’s grooming of p5 575 and Cluster 1350 – both inaugurated as HPC products – for business intelligence and DBMS roles.

¹⁶ Although FPGAs are accessed like software through function calls, they’re actually hardware components and therefore much faster than software libraries. Originally implemented for HPC as PCI-attached boards, they are now starting to be directly attached to the system interconnects, such as those in the Cray XD1 and the latest members of SGI’s Altix line.

¹⁷ Reconfigurable Application Specific Computing – SGI’s name for its FPGA technology.

Linux running atop Itanium is another, less vendor-specific example; it is seen most notably in servers from HP and SGI but also Fujitsu, NEC, and Unisys. While not volume platforms on the scale of x86, they nonetheless represent in aggregate an attractive target for ISVs. Nearly two thousand applications run on Linux and Itanium. Certified database, ERP, and analysis applications include: IBM DB2 UDB, Oracle 9i and IOg (with and without RAC), IBM Informix Dynamic Server, IBM DB2 Intelligent Miner, and the full suite of SAP products. Certification for the full SAS suite, including the ETL Server (which extracts, transforms, and loads data) and the BI suite of business intelligence tools is underway. Other popular infrastructure applications such as EMC NetWorker, SteelEye LifeKeeper, and Symantec’s VERITAS NetBackup are also supported.

With applications now available (and Linux increasingly scalable), systems that were once HPC specialists are starting to find their way into the running of commercial workloads. SGI, for example, points to gaming companies using its gear to run the real-time OLTP applications that manage slot machine, manufacturers running SAP or Oracle PLM for manufacturing operations, and universities, such as Masaryk University in the Czech Republic, running Oracle 10g RAC for administrative systems.

Whether companies specialize in HPC or business computing is increasingly far more a question of deliberate market focus, expertise, and the specific applications to which they’ve devoted the energy to get certified than it is about the attributes of the hardware, operating system, or other such architectural components. That’s not to minimize the reasons that companies specialize. It’s the rare vendor that can competently hit to all fields; at the very least it takes considerable scale and resources. However, even HPC specialists have a growing opportunity to leverage products, skills, and installed base into commercial sales. After all, the needs of the research lab and the enterprise datacenter are more closely matched today than at any point in decades.

This documentation, in electronic format, comprises software developed at private expense; if acquired under an agreement with the USA government or any contractor thereto, it is acquired as “commercial computer software” subject to the provisions of its applicable license terms and conditions, as specified in (a) 48 CFR 12.212 of the FAR; or, if acquired for Department of Defense units, (b) 48 CFR 227-7202 of the DoD FAR Supplement; or sections succeeding thereto. Contractor/manufacturer is SILICON GRAPHICS, INC., 1200Amphitheatre Parkway, Mountain View, CA 94043-1351.

Silicon Graphics, SGI, IRIX, and the SGI logo are registered trademarks of Silicon Graphics, Inc., in the United States and/or other countries worldwide, used with the permission of Silicon Graphics. All other trademarks and copyrights are owned by their respective owners.