When Multicore Isn’t Enough: Trends and the Future for Multi-Multicore Systems

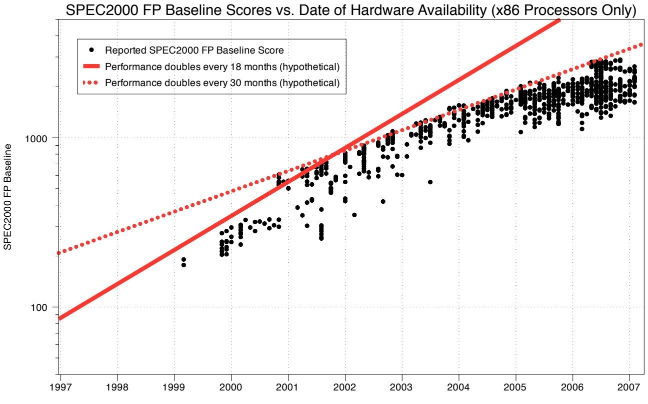
Matthew Reilly
SiCortex, Inc.

***Abstract*—Commodity multicore architectures take advantage of “Moore’s Law” by providing *more* processors in a given area, as opposed to *faster* processors.**

Systems built for high-processor-count applications must balance both per-die memory bandwidth and per-die communications bandwidth against core count and compute power per core. SiCortex builds multi-multicore systems to support high-processor count computing applications. The SiCortex SC5832 contains 972 multicore nodes of six MIPS processors each, providing 5,832 processors in less than 150 cubic feet. The system dissipates less than 18KW. Smaller configurations that fit in a single standard 19” rack are also available.

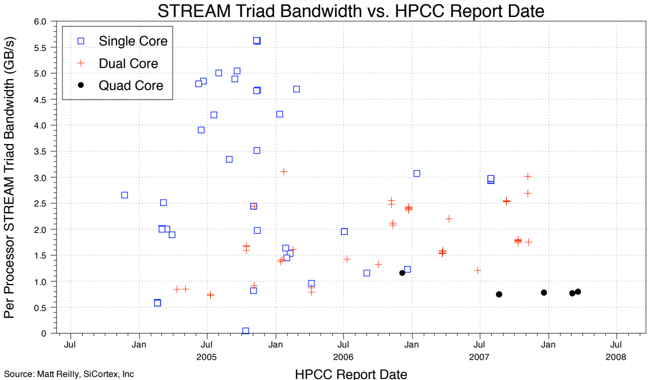
SiCortex cluster systems are built upon a multicore system-on-a-chip that comprises six cache coherent low power processors, two DDR2 DRAM ports, and a high performance message passing fabric interface.

This presentation will review the industry trends that created the need for balanced multi-multicore systems, the SiCortex architecture, and a few examples of its use and performance.



Source: Matt Reilly, SiCortex, Inc
http://www.bigncomputing.org

Fig. 1. SPEC2000 FP Score vs. Availabilty Date



Source: Matt Reilly, SiCortex, Inc
http://www.bigncomputing.org

HPCC Report Date
(Excludes Cray X1, NEC SX Series, and IBM Power4/5)

Fig. 2. Stream TRIAD Bandwidth vs. Report Date for 1, 2, and 4 Core Components

I. Trends in MultiCore Components

Multicore designs were a rational response to increasing schedule risk and decreasing architectural performance gains in microprocessor designs. By 2001 microprocessor architecture had almost fully exploited the out-of-order and speculative hardware mechanisms that had been first pioneered in the 1960’s. Each new generation of processor designs was finding less and less to be gained from more and more elaborate micro-architectural features: if a 2048 entry branch predictor gets 99% of all predictions correct, doubling its size is likely to have little impact on overall performance. Further, technology was conspiring to make the gains of 1990’s GHz wars a fond memory. With each new process improvement, transistors were getting smaller and faster, but interconnect speeds were not keeping pace. It is hard to develop a business case for pushing commodity processors to 5GHz and above given the risk to the production schedule.

Figure 1 shows the trend of SPEC2000 floating point performance vs. hardware availability date for all x86 processors in the SPEC2000 results database. The solid line shows the trend line for a 2X performance gain every 18 months. The dotted line corresponds to a doubling interval of 30 months. It is clear that the industry turned a corner around 2001.

Moore’s Law never promised faster processors: it promised more devices with each process generation. Multicore designs take advantage of Moore’s Law by replication. But this replication has its cost. Commodity multicore processors often abandon a careful balance between floating point capacity and memory bandwidth per processor. Figure 2 shows the trend in McCalpin Stream TRIAD memory bandwidth for a collection of single core and multicore processors. Note that each increase in the number of cores per unit came at the cost of memory bandwidth provided to each core.

While integrating more cores onto a die has provided a short term answer to many computing problems, its utility has limits. Multicore architectures are a stopgap. To be useful, each core must be fed by the external memory system. As the number of cores per die grows, the pin-bandwidth per die must also increase. Unfortunately, pin bandwidth is governed by factors

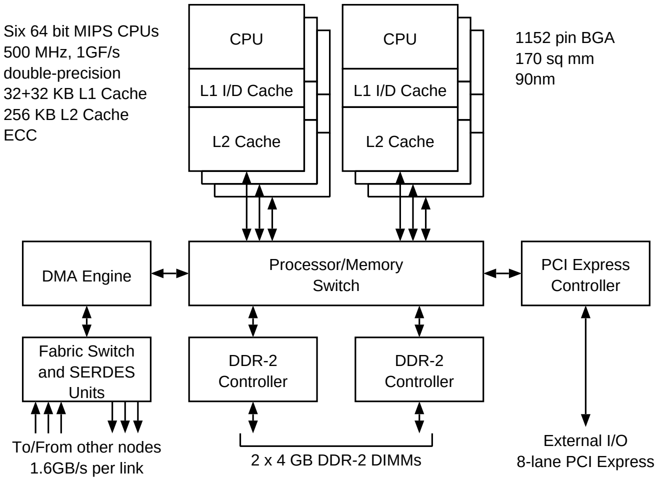


Fig. 3. The SiCortex Six Processor Cluster Node

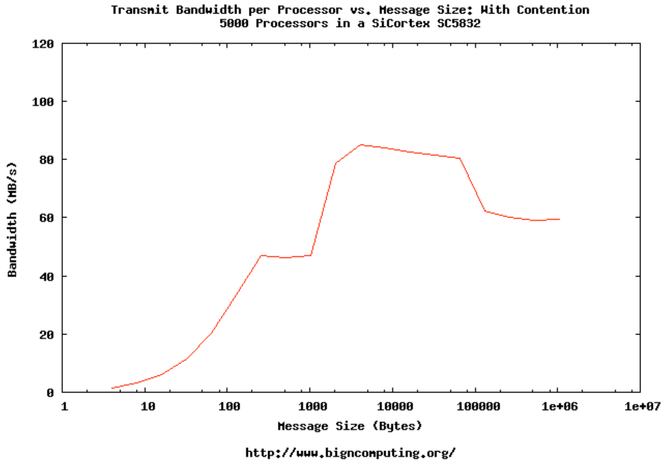


Fig. 4. MPI_Send/MPI_Recv Delivered Bandwidth: high fabric contention

that are independent of Moore's law. Eventually, the number of cores per die must be reconciled with the available off-die bandwidth.

The industry will, of course, address some of these shortcomings, but it is clear that, for a large number of reasons, the rapid increase in single processor performance is likely a thing of the past. Future advancements in computing capability must come from harnessing more cores to a single solution.

II. Enabling High Processor Counts: The SiCortex SC5832 Massive Multi-Multicore System

SiCortex builds high processor count computer systems for technical applications. The SiCortex SC5832 system comprises a cluster of 972 six-processor computing nodes connected via a high speed parallel point-to-point network. The network (fabric) topology is a *Kautz graph* which guarantees that the longest path through the network is logarithmic in the number of network nodes. In the case of the SC5832, the network diameter is 6 hops, with each node connecting three input ports and three output ports into the fabric. [1], [2], [3]

Figure 3 shows the single-chip, six-processor cluster node. Twenty-seven cluster nodes are included in each of the thirty-six modules in the SC5832, for a total of 972 nodes or 5,832 processors in the system.

Low overhead communications operations are implemented in the DMA engine and its associated Fabric Switch, both on the same die as the node's six processors. Messages are originated (*e.g.* via MPI_SEND calls) by an application running on one of the node's six processors, packetized by the DMA engine, tagged with a network routing string, and inserted into the fabric via the DMA port on the fabric switch. The packets may pass through a number of other nodes (up to six in the SC5832 fabric) before arriving at the destination where the packet is passed through the destination node's fabric switch to its DMA engine. The DMA engine then assembles the packets and delivers the message to a receiving application that has invoked the MPI_RECEIVE function.

Figure 4 shows measured message bandwidth between each of 2500 pairs of processors in an SC5832 performing pairwise message exchanges for messages of varying size. Of key interest is the performance of the fabric for small message sizes. For exchanges of 1024 bytes or fewer, the cost per exchange is fixed at 5 microseconds per 128 byte chunk. (Actual applications can improve on this, the benchmark uses synchronous message operations.) For messages longer than 1K bytes, the DMA engine switches to a "rendezvous" protocol that improves communication efficiency for large messages.

MPI PingPong Latency	1.4 μS
MPI PingPong Bandwidth	1.5 GB/s
DGEMM (Matrix Multiply)	720 MFLOPS per processor
Linpack (HPL)	3.6 TF
PTRANS (Transpose) Bandwidth	210 GB/s
STREAM TRIAD	345 MB/s per processor 1.9TB/s per system
Random Ring Bandwidth	50 MB/s
Random Access (GUPS)	0.74 GUPS
Random Access (GUPS) Optimized	5.5 GUPS

TABLE I
Microbenchmark and Kernel Speeds for the SC5832

Table I lists current results for several of the HPC Challenge benchmarks. The presentation includes other measurements of performance on other compute kernels and larger applications.

References

[1] M. Reilly, L. C. Stewart, J. Leonard, and D. Gingold. (2006, Dec.) Sicortex technical summary. [Online]. Available: http://www.sicortex.com/press/sicortex-tech_summary.pdf

[2] L. C. Stewart and D. Gingold. (2006, Dec.) A new generation of cluster interconnect. [Online]. Available: http://www.sicortex.com/press/sicortex-cluster_interconnect.pdf

[3] B. Elspas, W. H. Kautz, and J. Turner, "Theory of cellular logic networks and machines," Stanford Research Institute, Tech. Rep. AFCRL-68-0668, 1968.