Petascale Computing in a Cubic Meter by 2015

James C. Anderson, JCA@LL.MIT.EDU
MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02420

Overview

An ongoing study is being conducted to determine the feasibility of a petascale mobile signal processor by the year 2015. The main goal of this study is to develop system-level roadmap information that can be used to identify and guide development of key technologies. The signal processor is required to have a compact ($1m^3$) form factor suitable for platforms such as ships and submarines, and feature modular construction to ease on-site assembly and maintenance.

The 2015 performance goals include a throughput of 1 PFLOPS (peta, or 10^{15}, 32-bit floating-point operations per second) continuously sustained for computing 1K (1024-point) complex FFTs (fast Fourier transforms). This goal requires a simultaneous I&O (input and output) data rate of 1.28 Pbits/sec. The system must also provide 0.1 Pbyte high-speed memory (i.e., 10 FLOPS/byte), and be rapidly reconfigurable to support other general-purpose signal processing applications.

Compute nodes

In 3/05 (March 2005), a 6U form factor (16.0 x 23.4 x 2.03 cm^3 = 0.76 liter volume) COTS (commercial off-the-shelf) card was available that consumed 55 watts (using an on-board, 91% efficient DC-to-DC converter) to provide 6.68 GFLOPS (giga, or 10^9) on each of four CNs (compute nodes). Each 12.5W CN included a 1 GHz PowerPC MPC7447A general-purpose RISC (reduced instruction set computer) with on-chip AltiVec vector processor capable of computing a 1K complex FFT (the equivalent of 51,200 real operations) in 7.66 µsec (i.e., 83.6% of its theoretical 8 GFLOPS peak throughput). Unfortunately, the COTS card had only 38% of the memory size and 24% of the I/O bandwidth needed to support the target applications.

Within a year, improved memory and I/O devices were available for an 11.7W CN with the following characteristics (noting that all RF, or radio frequency, communication links are "wired," not "wireless"): 6.68 GFLOPS processor (8.0W), 8.55 Gbits/sec coaxial cable input link (1.2W, including 10dB gain RF input amplifier and demultiplexer), 8.55 Gbits/sec coax output link (0.7W, including 10dB gain RF output amp and multiplexer), "glue logic" (0.8W), 640 Mbytes DDR-SDRAM (double data rate synchronous dynamic random access memory, 1.0W) and 8 Mbytes nonvolatile flash memory (with negligible power consumption, active only at power-up, capable of holding a million lines of C code using lossless compression). All devices had an area of 15 mm x 15 mm or less, which would allow them to be integrated into a "small footprint" 3D (three dimensional) vertical stack. Using technology available as of 3/06, each CN stack would require 15 layers of the type shown in Fig. 1, and stacking is only feasible when devices with the highest power consumption are placed nearest the bottom layer, just above the 3D package's integrated metal heatsink (i.e., the processor must be on the bottom, input link on the next layer, etc.).

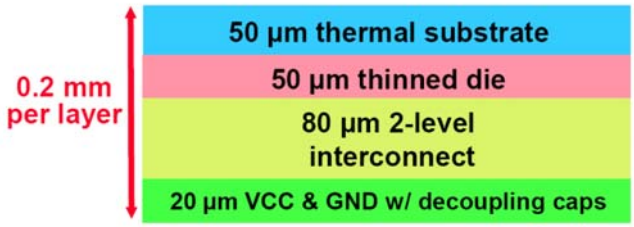

Figure 1: 3D stack layer structure (15 layers/CN)

Advanced packaging

It is presently possible to use HOPG (highly oriented pyrolytic graphite) as the central "core" heatsink of 2-sided "sandwich" construction SEM-E (standard electronic module format E, MIL-STD-1389 and IEEE-Std-1101.4-1993, 14.9 x 16.3 x 1.52 cm^3 = 0.37L). HOPG has a thermal conductivity approximately 7.8X that of aluminum [1], allowing a SEM to dissipate 375W. The combination of 3D technology on SEMs, along with future availability of low-voltage DC-to-DC converters having high current, efficiency and density, could provide nearly a 15X size reduction vs. COTS by 2015, as shown in Fig. 2.

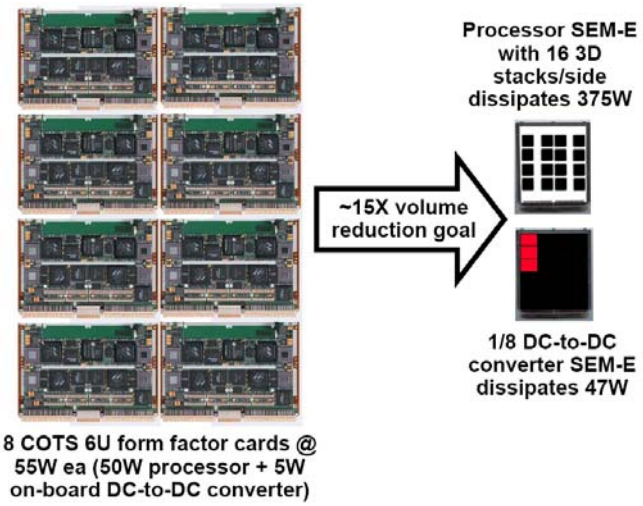

Figure 2: 15X size reduction vs. COTS by 2015

This work was sponsored by the Department of the Air Force under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author, and are not necessarily endorsed by the United States Government. The author wishes to acknowledge the contributions of Larry Retherford (mechanical/thermal) and Albert Horst (power).

Projection methodology

Using the present improvement rate of 2X every 3 years (which does not exceed the improvement rate projected by the International Technology Roadmap for Semiconductors, 2005 Edition, http://public.itrs.net), by 3/15 each 11.7W CN will provide 53.4 GFLOPS throughput, 5.12 Gbytes RAM, 64 Mbytes flash and an I&O rate of 68.4 Gbits/sec. Assuming all processor SEMs are connected to a central switch via coaxial cable (1.8 mm outside diameter, 65 GHz bandwidth and 20dB loss per 2.7 meters overcome by the RF amps), the I&O data modulation format for 3/15 must have high spectral efficiency (e.g., 8-phase shift keying or Gaussian minimum shift keying @ 2 bits/sec per Hz).

As of 3/06, 80% efficient COTS DC-to-DC converters could be used to create a one-sided converter SEM to power a pair of adjacent processor SEMs, allowing 15 converter and 30 processor SEMs per 45-slot cage (960 CNs/cage). Converter density is expected to double by 3/09, and efficiency is expected to rise to 89% by 3/12. By 3/15 the density is expected to double again, allowing 5 converter and 40 processor SEMs/cage (1280 CNs/cage).

System design

As shown in Fig. 3, the $1m^3$ system enclosure consists of 16 processor SEM cages topped by a central switch chassis and 4 FO/RF (fiber optic to and from RF) converter SEM cages. Estimated system weight is 1800 kg, and a 2 x 2 m^2 weight spreader must be used if the floor's load limit is 500 kg/m^2 (100 lbs/ft^2). The total input power budget is 340kW (20 cages x 45 SEMs/cage x 375W/SEM, although the FO/RF converter SEMs consume less).

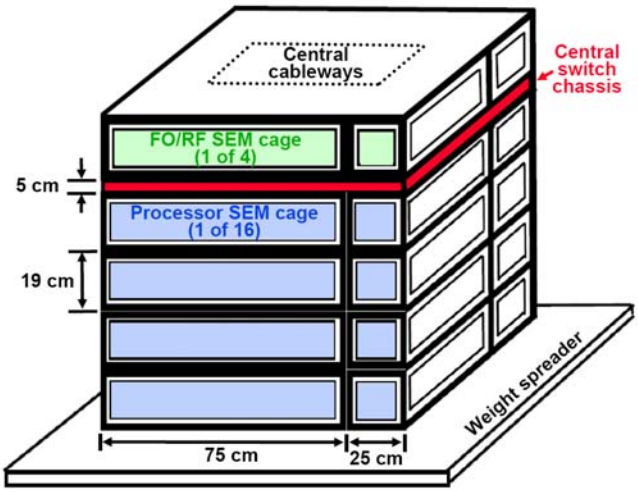

Figure 3: $1m^3$ system enclosure

The FO/RF cages are separated by a 50 x 50 cm^2 central cableway, as shown in Fig. 3, through which coax cables from the FO/RF converter SEMs travel downward to the central switch. The processor cages are similarly separated by another central cableway through which coax cables from the processor SEMs travel upward to the central switch. The central switch chassis contains a large, horizontal circuit card with a central MEMS (microelectromechanical structure) RF "butterfly" crosspoint switch array capable of connecting any input to any output. The MEMs switches only require power when changing state, and the 1.3 million packaged switches would occupy a total area of 0.36 m^2 using today's 40 GHz bandwidth devices [2]. Note that the switches are bi-directional, allowing the use of transceivers in place of input and output amps if desired.

At each corner of the central cableways is an 8 x 8 cm^2 conduit that runs from the top of the system enclosure down through the floor and carries 40 AWG #2 wires (6.5 mm diameter each) for 48VDC power, 10 pipes for liquid coolant (12.7 mm or 1/2" each) and 2560 FO cables (1 mm diameter each, with 4X the coax cable data rate over a distance of many meters) for external I/O.

System performance

Fig. 4 illustrates key figures of merit for comparison purposes. Operating points for COTS 6U processor racks (56 x 64 x 185 cm^3 containing 6 card cages with 21 slots/cage) are shown, although COTS cards do not provide the memory size and I/O bandwidth needed to support the target applications. The 3/15 goal of 1000 GFLOPS/L is shown for a comparable SEM processor rack (i.e., the processor SEM cages with 20,480 CNs and associated central cableway). Although many challenges remain, it appears that the foregoing petascale system is feasible by 2015, and a reduced-capability prototype could be built today.

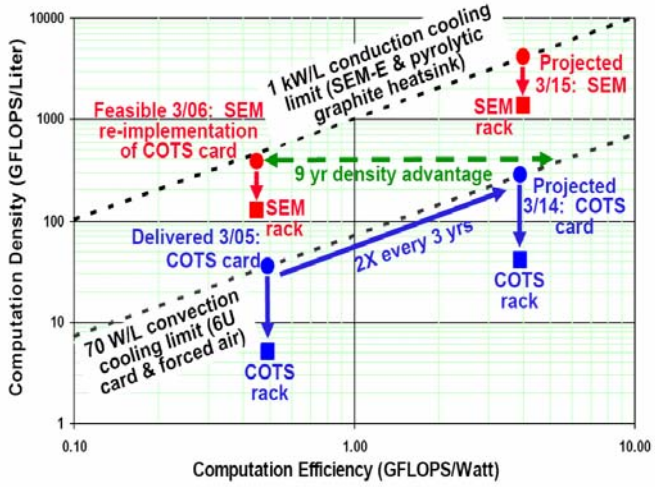

Figure 4: Figures of merit

References

[1] Carl Zweben, "Thermal Materials Solve Power Electronics Challenges," *Power Electronics Technology Magazine*, Feb. 2006, http://powerelectronics.com/mag/602PET24.pdf.

[2] S. Duffy, C. Bozler, S. Rabe, J. Knecht, L. Travis, P. Wyatt, C. Keast and M. Gouker, "MEMS Microswitches for Reconfigurable Microwave Circuitry," *IEEE Microwave and Wireless Components Letters*, Vol. 11, No. 3, March 2001, http://ieeexplore.ieee.org.