SoC for HPC: An Agile Approach to Building HPC Systems from Commodity Components

John Shalf, LBNL and Jim Ang, SNL

July 15, 2016

The current mainstream HPC ecosystem relies upon Commercial off-the-Shelf (COTS) *commodity* building blocks to enable cost-effective design by sharing Non-Recurring Engineering (NRE) costs across a larger ecosystem. Past and current HPC nodes use commodity chipsets and processor chips integrated together on custom motherboards. An emerging segment of the industry, with support from government agencies, are embarking upon a new era for commodity HPC where the chip acts as the "silicon motherboard" that interconnects building blocks of commodity IP to create a complete integrated System-on-Chip (SoC). This approach is still very much COTS, but the commodities are licensable IP for pre-verified circuit designs (the Lego-blocks for SoC designs) rather than the chips. It achieves cost-competitiveness because, as shown in Figure 1, the dominant cost of designing a chip is the cost of verifying the circuit building blocks. The cost/benefits derive from the ability to leverage a commodity ecosystem of these pre-fabricated and licensable circuit blocks where the NRE cost of designing and

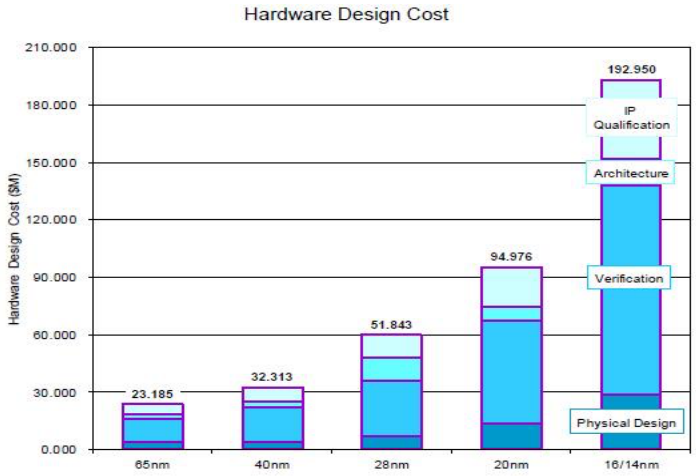

Figure 1: Verification and design costs dominate the cost of producing a new chip. The embedded System on Chip (SoC) market offers commodity Intellectual Property (IP) circuit designs that are pre-designed and pre-verified, where the verification costs are shared across a much larger market. Design time and verification costs are minimized by using these existing circuit designs (IP Blocks) and arranging them into HPC-targeted chip designs. Development time can be reduced for 4-6 years down to 18months by utilizing design-libraries of pre-existing IP, which is a vibrant commodity market.

verifying a new processor or memory controller design (an IP building block) can be amortized by licensing the technology to myriad embedded applications. The market for licensed circuit IP in the embedded space is much larger (both in volume and total revenue) than for server chips and the market segment for commodity IP building blocks for SoCs is growing at a far faster pace than the current server chip market. Figure 2 shows that the market for licensed circuit IP in the embedded space is much a larger

marketplace than for server chips (both in volume and total revenue) and the market segment for building blocks is growing at a faster pace than the current server chip market.

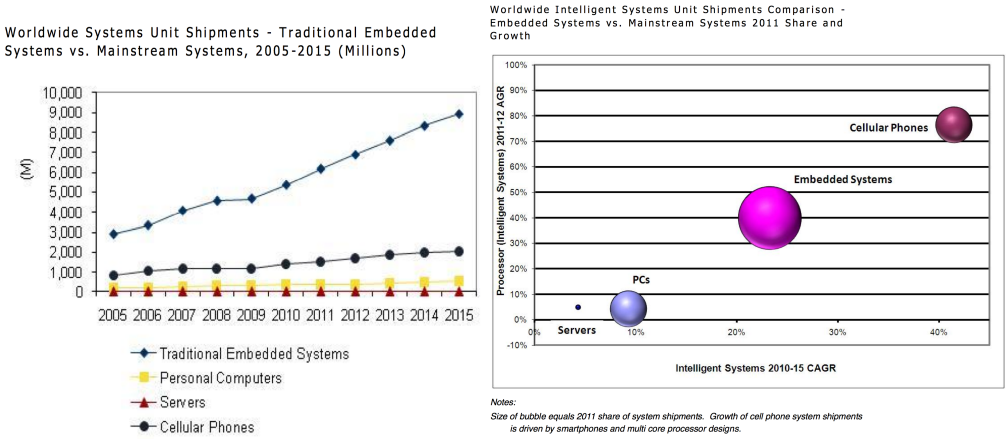

Figure 2: The market size and revenues in the high-performance commodity IP space far outpace the volumes and investments in the server chip technology that currently underpins HPC. (data from 2013 IDC report)

The SoC approach to integration involves combining these IP blocks together onto a single silicon chip. Traditionally SoC design methods have focused on low-power consumer electronics or high performance embedded applications. But now SoC design methods are moving into high-end computing due to the emergence of embedded IP offering capable double-precision floating point, 64-bit address capability, and options for high performance I/O and memory interfaces. This enables HPC chip designers to include features they need, and exclude features that are not required in a manner that is not feasible with today's commodity board-level computing system design. SoC integration is able to further reduce power, increase integration density, and improve reliability. It also can enable designers to minimize off-chip I/O by integrating peripheral functions, such as network interfaces and memory controllers. Furthermore, the embedded market has developed extraordinarily capable tools for rapidly prototyping, simulating, and synthesizing full SoC design*s,* with a much faster turn-around than we have come accustomed to for commodity server chip designs (many designs targeted at an 18 month design cycle for the hyper-competitive consumer market). By leveraging the enormous commodity IP market for design tools, processor cores, memory controllers, and I/O circuit designs, chip designers can focus their effort and NRE costs on engineering a handful of essential features that are not covered by the commodity ecosystem.

Why now?

A critical change in the cadence of computing improvement has occurred. This change both requires a response to cope with the significant impact it will have and represents an opportunity to better serve communities of interest to the government. In the past, because of the development time for custom devices and the rapid cycle of commodity processor performance advances, by the time a custom implementation was complete it was typically overcome by the advancing performance capabilities of commercial

general-purpose products. In addition, the infrastructure for custom device designs resulted in unacceptably high design costs. As a result, special purpose devices were both costly and of limited impact. Today the rate of performance advances of commodity processor roadmaps has slowed, reducing the speed of the computing improvement cadence. The effective end of Dennard scaling has dictated this cadence. Also, the opportunity to increase the development cycle speed for SoC/custom devices now seems possible, increasing the SoC/custom device cadence. These two changes could enable SoC/custom device development to step forward to have a significant impact on future architectures and for government processing needs.

Players in this emerging direction include ARM, RISC-V, Broadcom, Qualcomm, and Cavium. These are all non-traditional players in HPC, and yet the opportunity for disruption is very high (pursued vigorously by the European Union, China, and Japanese exascale programs), so it is crucial for DOE to track developments in this area.

Overview of Sunway TaihuLight Supercomputer

The specific design of the TaihuLight is not revolutionary, and not substantially better in terms of per-node peak performance compared to US designs. The TaihuLight node is more energy efficient because of the ability to eliminate unnecessary elements from the design, which is a benefit of the flexibility offered by the SoC design methodology, severe limits on the total memory capacity, and the freedom from constraints of current processor roadmaps. The real benefit of TaihuLight is to demonstrate the advantages of a more agile approach to development using commodity IP integrated into SoCs. Their design is not as fast or as programmable as US designs, but the development cycle was extremely short 18-24 months, at a fraction of the total development cost. Dongarra's report estimated the costs of Sunway TaihuLight to be $270M USD including all hardware capital costs, R&D, software, and the cost of constructing the building to house it! Even if the current instantiation of the Sunway machine design is less sophisticated than a US design, China will be able to create 3-4 generations of refinements in the time it takes to create one generation using the traditional 4-6 year development cycle.

The Chinese TaihuLight supercomputing system is a concrete example of the commodity IP/SoC agile design approach – see our discussion below. The Computing Processing Elements (CPEs) are derived from an existing embedded Digital Signal Processor core design that is typically used in high performance embedded signal processing and avionics applications. Using the SoC design methodology, in a short timeframe (estimated at 18-24 months), Sunway integrated 64 CPEs onto a chip using a simple mesh Network-on-Chip with a Management Processing Element (MPE) to achieve an aggregate performance that is competitive with US HPC chip designs that require a substantially longer development cycle. The NoC topology and the array of simple cores looks very much like a clone of US manycore designs. The CPE cores are far simpler however, which enables them to deliver performance at a lower power budget than is achievable using the more complex core designs used in typical manycore server designs despite the use of older generation 28nm lithography technology.

One purported weakness of the TaihuLight system is its poor per-node memory bandwidth and capacity due to use of older DDR3 memory technology. We project that

the next generation of Sunway will likely integrate much faster more contemporary memory technologies such as HBM (high performance stacked memory) technology that will overcome this interim limitation. HBM stacked memory is a broadly available standards-based commodity and primarily sourced from Asia, so there is no technical reason why they would not improve future memory subsystems. It is also likely (but no direct evidence yet) that future generations of Sunway will benefit from an integrated interconnection network interface, and there is significant headroom for improvement in performance and energy efficiency with the improved Fab process technology. These purported deficiencies of the Sunway TaihuLight are ultimately ephemeral.

The (now) number 2 system in China is the Tianhe-2a system at NUDT. Since the U.S. Commerce Department embargoed the sale of Intel KNL processors to China to upgrade the accelerators in Tianhe-2a system, China has been pursuing the development of an ARMv8 based accelerator to upgrade Tianhe that looks for all practical purposes like an ARM-based clone of the originally planned Intel KNL co-processors. This will constitute a completely different design, but using the same IP/SoC design methodology. In another announcement at ISC16, Fujitsu has announced their intent to license ARMv8 for their new vector processor design for the Post-K supercomputer at RIKEN. There are indications that the Japanese Exascale effort is also pursuing IP/SoC designs.

China Domestic Chip Fabrication Capability

Sunway TaihuLight processors are fabricated in China using 28nm process technology. For 2016, this is state of the art for the semiconductor foundries in China. The largest foundry company in China is Semiconductor Manufacturing International Corporation (SMIC). SMIC has 28nm/40nm capability foundries in Beijing, Shanghai, and Shenzhen capable of fabricating 200mm and/or 300mm scale wafers. From public news sources, Taiwan Semiconductor Microelectronics Corporation (TSMC) has broken ground on the construction of a 300mm Foundry targeting 16nm chips in Nanjing, China. This Fab is scheduled to be in production in late 2018. This is the process technology that NVIDIA will be using in the 2018 timeframe for Sierra and Summit (systems acquired for the CORAL collaboration of ORNL, Argonne, and LLNL). Note: NVIDIA produces its GPUs at TSMC foundries in Taiwan.