The TSUBAME Cluster Experience a Year Later, and onto Petascale TSUBAME 2.0

Satoshi Matsuoka

Global Scientific Information and Computing Center
Tokyo Institute of Technology
2-12-1 Oo-Okayama, Meguro-ku, Tokyo 152-8550, Japan
matsu@is.titech.ac.jp
http://www.gsic.titech.ac.jp

TSUBAME (Tokyo-tech Supercomputer and Ubiquitously Accessible Mass-storage Environment) is a new supercomputer installed at Tokyo Institute of Technology in Tokyo, Japan, on April 1st, 2006, and as of 2007 facilitating over 85 Teraflops of peak compute power with acceleration, 22 Terabytes of memory, and 1.6 Petabytes of online disk storage, "Fat Node" as well as fast parallel interconnect—architectural principles based on traditional supercomputers. TSUBAME became the fastest and largest supercomputer in Asia in terms of performance, memory and storage capacity etc., starting from the 38.18 Teraflops performance (7th overall) for the June 2006 Top500 announcement, and taking the fastest in Asia crown for 3 consecutive Top500s in a row (currently at 48.88 Teraflops). At the same time, being PC architecture-based, TSUBAME, being a large collection of PC servers, allows for offering much broader services than traditional supercomputers resulting in a much wider user base, including incubation of novice students. We term such architectural and operational property of TSUBAME as "Everybody's Supercomputer", as opposed to traditional supercomputers with very limited number of users, thus making their financial justifications increasingly difficult.

Tsubame is a result of collaborative effort between the univerity academia at the Tokyo Institute of Technology, and multiple industrial partners worldwide. The contract was awarded to NEC, who jointly with Sun Microsystems built and installed the entire machine, and also collaboratively provide on-site engineering to operate the machine. Other commercial partners, such as AMD (Opteron CPUs), Voltaire (Infiniband), ClearSpeed (Accelerator), CFS (LUSTRE parallel filesystem), Novell (SUSE Linux), provided their own products and expertise as building blocks. It was installed in just three weeks, and when its operation started on April 3rd, 2006

Overall, TSUBAME's installation space is approximately $350m^2$ including the service area. There are approximately 80 compute/storage/network racks, as well as 32 CRC units for cooling, that are laid out in a customized fashion to maximize cooling efficiency, instead of the machine itself merely being placed as an afterthought. This allows for considerable density and much better cooling efficiency compared to other machines of similar performance. TSUBAME occupies three rooms, where room-to-room Infiniband connections are achieved

F. Cappello et al. (Eds.): EuroPVM/MPI 2007, LNCS 4757, pp. 8–9, 2007.

via optical fiber connection, whereas uses the CX4 copper cable within a room. The total power consumption of TSUBAME is less than a Megawatt even at peak load, making it one of the most power- and space- efficient genera-purpose cluster supercomputer in the 100Teraflops performance scale.

TSUBAME's lifetime was initially designed to be 4 years, until the spring of 2010, with possible short-term extensions realized by incremental upgrades to maintain the competitiveness of the machine. However, eventually the lifetime will expire, and we are already beginning the plans for designing the next generaion "TSUBAME 2.0". Here, simply waiting for processor improvements relying on CPU vendors would not be sufficient to meet the growing computational demands, as a result of success of "Everybody's Supercomputer", in growth of the supercomputing community itself, not just the individual needs. Another requirement is not to increase the power or the footprint requirement of the current machine, resulting in a considerable challenge in supercomputer design we are researching at the current moment.

One research investment we are conducting in this regard is in the area of acceleration technologies, which will provide vastly improved Megaflops/Watt ratio. In fact, even currently, two-fifth of TSUBAME's peak computing power is provided by the ClearSpeed Advanced Accelerator PCI-X board. However, acceleration technology is still narrowly scoped in terms of its applicability and user base; as such, we must generalize the use of acceleration via advances in algorithm and software technologies, as well as design a machine with right mix of various heterogeneous resources, including general-purpose processors, and various types of accelerators. Another factor is storage, where multi-Petabyte storage with high bandwidth must be accommodated. Challenges are in devising more efficient cooling, better power control, etc. There are various challenges abound, and it will require advances in multi-disciplinary fashion to meet this challenge. This is not a mere pursuit of FLOPS, but rather, "pursuit of FLOPS usable by everyone"—a challenge worthwhile taking for those of us who are computer scientists. And the challenge will continue beyond TSUBAME 2.0 for many years to come.