Tokyo Institute of Technology Submission to the First Annual CUDA Achievement Award

Principal Investigator: Satoshi Matsuoka (GSIC Center, Tokyo Tech.)

1.1 Executive Summary

Tokyo Institute of Technology (Tokyo Tech), one of the world's premiere universities in science and engineering, was awarded Japan's first NVIDIA CUDA Center of Excellence (CCOE) status in 2009, for its innovative research and education in GPU Computing at the forefront of HPC. Since then, we have designed and constructed the Japan's first petascale supercomputer, TSUBAME2.0, and a series of various advanced software and applications research based on it. Such activities have been rewarded with numerous results presented at top academic venues as well as numerous global accolades and press reports. Here we highlight the 3 core achievements of TSUBAME / CUDA CCOE as our submission to the CUDA Achievement Award, but the results are not just limited to them.

The list of accolades for the past 18 months is as follows, achieved through the innovative R&D of TSUBAME2.0 centered around GPUs:

- The Top 500: World Rank #4, 1.192 Petaflops (Nov. 2010), #5 (June 2011, Nov. 2011)
- The Green 500: World Rank #2 (Nov. 2010), "The Greenest Production Supercomputer in the World" (Nov. 2010, June 2011)
- The Graph 500: World Rank #3 (Nov. 2011)
- 2011 ACM Gordon Bell Awards (Nov. 2011)
    - Special Achievement in Scalability and Time-to-Solution (Shimokawabe et. al.), Honorable Mention (Bernaschi et. al.)
- ACM SC11 Special Recognition Award for Perfect Score in Technical Paper (Bautista et. al.) (Nov. 2011) on GPU-accelerated, highly reliable checkpointing on Tsubame2.0
- ACM George Michael Memorial HPC Ph.D. Fellowships Honorable Mention (Bautista et. al.) (Nov. 2011)
- HPCWire Reader's Choice Awards (Nov. 2011)
    - Best HPC collaboration between government and industry: TSUBAME 2.0 project in collaboration with the Tokyo Institute of Technology, DataDirect Networks, Mellanox/Voltaire, NEC, NVIDIA, Intel, Microsoft, and Hewlett-Packard
    - Best application of "green computing" in HPC : Tokyo Institute of Technology TSUBAME 2.0.
- HPCWire Editor's Choice Award (Nov. 2011)
    - Best application of "green computing" in HPC : Tokyo Institute of Technology for TSUBAME 2.0.
- The Promotion Foundation Award for Electrical Science and Engineering for "Research and Development of the Greenest Production Supercomputer in the World". (Jan. 2012)
- Recognition as "No.1 Supercomputer in the World as Clear Overall Winner in Number Crunching, Data Crunching and Energy Efficiency: the HPC Hat Trick", HPC Wire Article (Feb. 2012).
- The Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Development Category, for "Developent of Tsubame2.0, the Greenest Production Supercomputer in the World" (April 2012)
- In addition, the PI Satoshi Matsuoka has been selected to become the ACM Fellow (Nov. 2011), only the 7th ACM fellow from Japan, and the co-PI Takayuki Aoki CUDA CCOE Fellow (April, 2012)

1.2 Design and Development of TSUBAME2.0, the first Multi-Petascale, GPU-Centric, Green, Production Supercomputer in Japan

Based on basic research on accelerators and many-core processors that started back in 2004, and actual collaboration with NVIDIA in TSUBAME1.2 which became the first large-scale supercomputer in the world to host hundreds of GPUs and be ranked on the Top500, TSUBAME2.0 was designed and deployed in Nov. 2010, sporting over 2.4 Petaflops of peak performance, of which over 90% is provided by the 4264 Fermi GPUs. It became the 4th fastest supercomputer in the world on the Top500 in Nov. 2011, recording 1.192 Petaflops, and still is 5th fastest today nearly 18 months later.

So much can be said about its design that was the world's first to be crafted as an entirely GPU centric, massively scalable supercomputer from ground-up rather being a tack on afterthoughts like other GPU supercomputers of the same era, but here is a short summary. GPUs were made to be central to computing, not only for its massive flops but also tremendous memory bandwidth, as the ratio of GPU chips to CPUs were maximized (3 GPUs to 2 CPUs) per node. The nodes were made to be "fat", i.e., accommodate as much GPU/CPU/memory/network/local SSD storage as possible for utmost flexibility, efficiency, and ease-of-programming for wide-ranging applications, with 1.7 Teraflops and 400GB/s of compute power and memory bandwidth respectively. For such a configuration, a node with massive I/O would be required to match the bandwidth of all of its components---in fact the I/O chips effectively become the central hub structure for routing the data between the numerous fat-node components above, instead of the CPU-centric design principles of tack-on machines that distances the GPUs from other intra- and inter-node resources. Moreover, the nodes needed to be designed to be as compact as possible, with effectively less than 1U space, despite the "fatness", as the design called for packing 1400 nodes into under 160m2 of space, or only about 40+ racks. This is to say a single rack is more than 50 Teraflops, essentially an "Earth Simulator in a Box". Such a design was also advantageous in implementing a fully-optical, full-bisection, multi-rail, fat-tree network, supported by the latest silicon photonics and large IB switching technologies with 3500 cables extending to mere 100kms in total; this is in stark contrast to the Earth Simulator or even the K computer which required thousands of kilometers of cabling. The aggregate bisectional bandwidth of TSUBAME2.0 being over 200Terabits/s, can transport the entire daily data flowing through the global Internet in less than half a day. To realize I/O for massive data throughput of GPUs and the network without having to implement a storage system larger than the machine itself, nearly 3000 SSDs were distributed among all the nodes, resulting in 2/3 Terabyte/s I/O bandwidth with essentially negligible power or space increase. Finally, hybrid water-air cooling with environmentally isolated racks with minimized chiller-to-cabinet distance allows for cooling the densely configured machine, with each rack consuming up to 35KWs, for typical PUEs of below 1.2, and the highest power efficiency for a petascale supercomputer in the world, as will be described later. Sensors and proactive power cap control as well as customized chassis configurations to allow for extensive and efficient thermal regulation of the GPUs, owing to the high reliability of the machine in production under any load.

Altogether, TSUBAME2.0 design has become the template of GPU-centric supercomputer design, incorporating the latest and the most advanced components and their technologies that had not been used or only used sparingly in supercomputers --- Many-core, high-performance & high-bandwidth GPUs / I/O centric high-bandwidth and ultra-compact fat node / Multi-rail full bisection fat-tree network with silicon photonics / Silicon I/O / High-efficiency environmentally contained cooling and proactive power control --- and architecturally combined them into the most cost/power/space efficient multi-petaflops supercomputer in the world. Again, this is quite contrasting to the K computer which employs components and architectural designs that are much more traditional and conservative, resulting in massive size and cost.

We have worked extensively to allow TSUBAME2.0 to be user-friendly despite its advanced features by developing a GPU-centric supercomputer software stack, some of the components taking the best of the breed while in other cases developing our own. The former are various programming tools and libraries for GPUs. The latter are more experimental programming tools and libraries, as well as covering other aspects such as resource management extensions for GPUs, resiliency, and green computing. Such sets of work not only had resulted in a number of well-known, global, and extremely honorable academic and industrial technology accolades as indicated above, but also allowed our users to quickly utilize TSUBAME2.0, especially its GPUs.

Such combined efforts in advanced hardware and software R&D for GPUs and other revolutionary aspects have demonstrated world top-class performance in benchmarks, but also in real applications (Figure 1) as we will be described later.

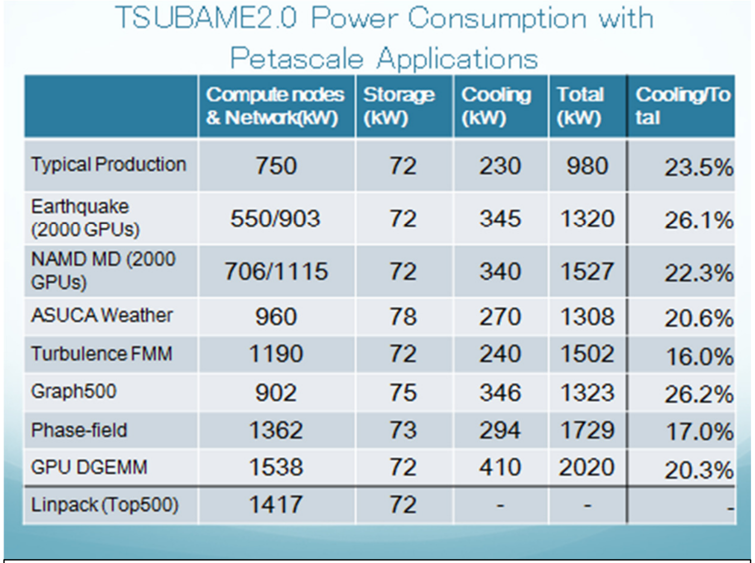
TSUBAME2.0 has been operated 24/7 in production for 16 months and now has over 5000 users, with 2000 active supercomputing users from within Tokyo Tech. as well as nationwide and internationally as one of the premiere national academic computing resources. At any time nearly 100 users are logged on to the system, and in March we observed over 90% node utilization, sometime going nearly 100%. The overall GPU utilization continued to exceed over 2000 in number, which means that effective GPU utilization in jobs are approximately 60%. We believe the latter is attributable to combined efforts of continued enhancement in the GPU-enabled software stack, resource allocation and the usage model, improvement in the reliability of GPU resource allocation and fault detection/isolation, as well as enrichment of the educational program including multitudes of classes and training courses on GPU usage as well as support.

In addition, TSUBAME2.0 has sported over 50 industry groups over the last 18 months as users in our various industry usage programs. The motivation behind the usage of what is fundamentally an academic machine is that, there is no other well-managed ultra-large-scale supercomputer built with scalable GPU infrastructure in Japan, and the industry is keen on investigating and evaluating its capability for its future business competitiveness.

1.3 "TSUBAME2.0 Greenest Production Supercomputer in the World" and the Tohoku-Kanto Earthquake

Continued push for more performance has resulted in tremendous concern for energy efficiency in supercomputing. TSUBAME2.0 realized this requirement to the ultimate, with aggressive design centered on many-core GPUs minimizing CPUs, ability to monitor and limit their power consumption, efficient cooling, as well as applications to exploit such power, has resulted in "Greenest Production Supercomputer in the World" award from the Green 500. The achieved 958GFlops/W was second in the world, and only matched or exceeded by essentially experimental machines which optimized for power and not performance or production, something that TSUBAME2.0 could not afford to do. Still, compared to laptops PCs, TSUBAME2.0 is not only 70,000 times faster, but also almost 3 times more power efficient.

Power efficiency was not limited to benchmarks alone. Figure 1 shows the 5-10 times power advantage that of conventional CPU-based machine under most practical operational settings running multitudes of real applications, say compared to ORNL Jaguar.



TSUBAME2.0 Power Consumption with Petascale Applications

	Compute nodes & Network(kW)	Storage (kW)	Cooling (kW)	Total (kW)	Cooling/Total
Typical Production	750	72	230	980	23.5%
Earthquake (2000 GPUs)	550/903	72	345	1320	26.1%
NAMD MD (2000 GPUs)	706/1115	72	340	1527	22.3%
ASUCA Weather	960	78	270	1308	20.6%
Turbulence FMM	1190	72	240	1502	16.0%
Graph500	902	75	346	1323	26.2%
Phase-field	1362	73	294	1729	17.0%
GPU DGEMM	1538	72	410	2020	20.3%
Linpack (Top500)	1417	72	-	-	-

Figure 1: Large Scale Apps and Power for TSUBAME2.0

Such power efficiency was the baseline for allowing continuous operations during the power crisis that hit Japan after the Tohoku-Kanto Earthquake that compromised Japan's power grid to the extent that, combined with almost record-breaking summer heat, forced the government to impose mandatory 15% peak power reduction for all major electricity users, including Tokyo Tech., during daytime when air conditioning and other production facilities would be at their maximum, to avoid rolling blackouts. In fact right after the disaster, 95% of Japan's supercomputing power had been compromised. To cope with the situation, we developed detailed monitoring as well as automated peak shifting resource scheduler, to allow arbitrary nodes to go into low power state or shut down and rebooted without much affecting the user experience. Moreover, extensive power measurements for our top-tier applications proved that TSUBAME2.0 would consume less power under controlled operations compared to its predecessor, TSUBAME1.2. We were thus able to run the machine almost at full capacity, demonstrating the power efficiency of GPUs, affecting many Japanese centers to also adopt GPUs, such as Univ. Tsukuba, Kyushu University, National Genomics Institute, etc.

1.4 Petascaling Applications in TSUBAME2.0

There are now over 10 real applications that scale to the entire machine, including the nanoscale dendrite formation simulation as well as the cardiovascular blood-flow simulation, the former winning the main prize and the latter honorable mention in the 2011 ACM Gordon Bell Prize. We encourage such scaling challenge for many-core applications with our biannual "Grand-Challenge" program, and many have demonstrated petascale performances, and/or superior performance and scalability compared to conventional CPU-based machines with similar petascale performance numbers such as the ORNL Jaguar Cray XT5 and Julich Jugene IBM BG/P, when running the same application. In the interest of space, we present here the dendrite application result.

The mechanical properties of metal materials largely depend on their intrinsic internal microstructures. To develop engineering materials with the expected properties, predicting patterns in solidified metals would be indispensable. The phase-field simulation is the most powerful method known to simulate the micro-scale dendritic growth during solidification in a binary alloy. To evaluate the realistic description of solidification, however, phase-field simulation requires computing a large number of complex nonlinear terms over a fine-grained grid. Due to such heavy computational demand, previous work on simulating three-dimensional solidification with phase-field methods was successful only in describing simple shapes. The new simulation techniques achieved scales unprecedentedly large, sufficient for handling complex dendritic structures required in material science. The phase-field equation is discretized by the second-order finite difference scheme for space with the first-order forward Euler time integration on a three-dimensional regular computational grid. The computational stencil accesses 19 points around the computational point .The simulations on TSUBAME 2.0 have demonstrated good strong scaling as well as weak scaling and we achieved 2.0PFlops in single precision for our largest configuration, using 4000 GPUs along with 16000 CPU cores, achieving 44.5 % efficiency. Also, the power efficiency during the run was 1468MFlops/W, being better than our Green500 result of 958MFlops/W. This and other results demonstrate that GPUs and a supercomputer properly designed to utilize its power, are truly scalable and efficient across many real-life applications, leading the way towards exascale and beyond.

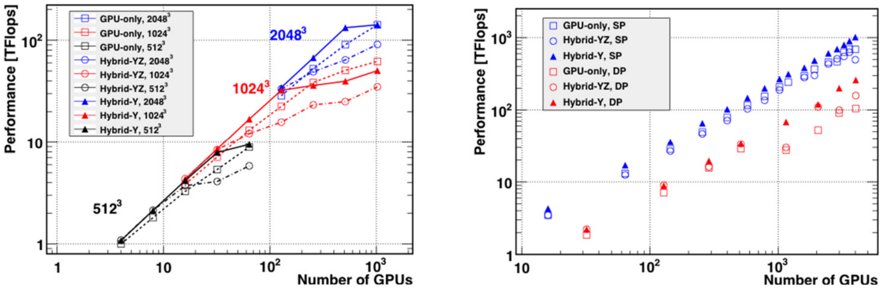


Fig.3 Performances of the phase-field simulation. Strong (left figure) and weak (right figure) scaling.

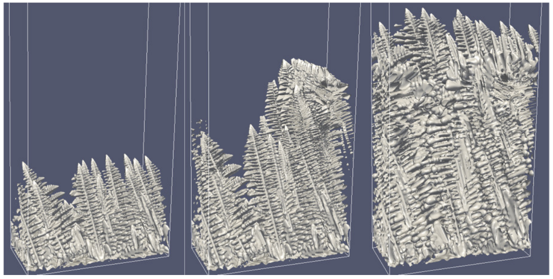


Fig.2 Growth of the metal dendrite solidification solving the phase-field model. 768x1632x 3264 using 1156 GPUs on TSUBAME 2.0.

The paper of this work is accepted as a technical paper in SC'11 conference and also won the **ACM 2011 Gordon Bell Award.**