ANALYST REPORT

GREEN REVOLUTION COOLING

The Immersion Cooling Authority

Tokyo Institute of Technology

The Immersion Cooled TSUBAME-KFC: From Exascale Prototype to The Greenest Supercomputer in The World

In Collaboration with GRC — Republished 2018

Authors:
Toshio Endo, Akira Nukada,
Satoshi Matsuoka

Conference Paper:
The 20th IEEE International Conference on Parallel and Distribution

Discussions, Stats, and Author Profiles:
at: http://www.researchgate.net/publication/275646769

10 YEARS

11525 Stonehollow Drive, Suite A-150, Austin, TX 78758
512.692.8003 • info@grcooling.com • grcooling.com

Toshio Endo Akira Nukada Satoshi Matsuoka
Global Science Information and Computing Center,
Tokyo Institute of Technology, Japan
Email: {endo, matsu}@is.titech.ac.jp, nukada@matsulab.is.titech.ac.jp

Abstract—Modern supercomputer performance is principally limited by power. TSUBAME-KFC is a state-of-the-art prototype for our next-generation TSUBAME3.0 supercomputer and towards future exascale. In collaboration with Green Revolution Cooling (GRC) and others, TSUBAME-KFC submerges compute nodes configured with extremely high processor/component density, into non-toxic, low viscosity coolant with high 260 Celsius flash point, and cooled using ambient / evaporative cooling tower. This minimizes cooling power while all semiconductor components kept at low temperature to lower leakage current. Numerous off-line in addition to on-line power and temperature sensors are facilitated throughout and constantly monitored to immediately observe the effect of voltage/frequency control. As a result,TSUBAME-KFC achieved world No. 1 on the Green500 in Nov. 2013 and Jun. 2014, by over 20% c.f. the nearest competitors.

I. INTRODUCTION

The most predominant issue towards future exascale supercomputers is power consumption, and the importance of being "green" has been, and still is a focus of many research and systems in supercomputing for the past 10 years at the least. Early systems employed low power embedded or ultra portable notebook processors such as Green Destiny [1], Blue-Gene/L[2], and MegaProto[3]. The DARPA exascale report established the goal of 20 megawatts for future exascale machines, analyzed the extreme challenges involved, and concluded that comprehensive low-power design would be required to even come close to the goal by 2018-20 time frame. This was followed on by the IESP (International Exascale Software Project) group, whose white paper was published to supplement the software requirements for achieving the 50 gigaflops/W goal[4], [5]. Nonetheless, the goal is still considered very difficult to reach without comprehensive research and engineering, with power reduction being the ultimate goal over all other objectives.

We have been extremely cognizant of the low power requirements supercomputers; there have been a number of research projects we have conducted, including ULP(Ultra Low-Power) HPC project conducted during 2007-2012, with the goal of achieving the 1000-fold improvement of power-performance efficiency(Figure 1). Many novel research achievements that harness substantial opportunity for HPC power saving. They include the use of new devices, new architectures, DVFS techniques, as well as extremely aggressive cooling strategies. As an example, the use of extreme many core processors, in the form of GPUs and Xeon Phi, have seen initial adoption as general purpose processors in the HPC arena, allowing up to several factors power-performance improvement compared to conventional CPUs [6], [7], [8]. There are now numerous supercomputers that employ many-core processors as their principle compute components, including the top two on the June 2014 edition of the Top500[10].

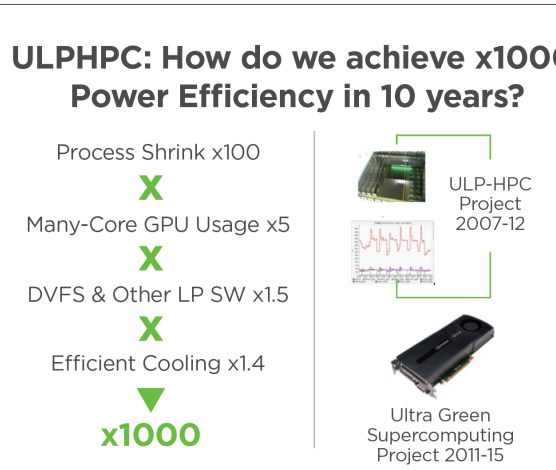

Fig. 1. The breakdown of 1000 times improvement in power efficiency in the ULPHPC project

Tokyo Tech.'s TSUBAME2.0 supercomputer was designed ground up to accommodate GPUs as the principle compute engine, commissioned in Nov. 1st, 2010 to become #4 in the world on the Top 500. Approximately 90% of the FLOPs and 80% of the memory bandwidth are provided by over 4000 NVIDIA Fermi M2050 GPUs in the system. TSUBAME2.0 became #3 in the world in the Green500[11] which ranks the Top500 supercomputers in terms of their power efficiency, with 958 megaflops/W. Also it was recognized as the "greenest production supercomputer in the world" for Nov. 2010 and June 2011 respectively, as other machines on the list were small and non-production in nature. TSUBAME2.0 has recently been upgraded to TSUBAME2.5 by replacing the GPUs to the latest Kepler K20x, improving the Green500 power efficiency to 3069 megaflops/W, more than tripling the power efficiency and being ranked 6th in the world.

Nonetheless we are still more than a factor of 15 away from the 50 gigaflops/W goal for the 20MW exaflop machine. Moreover, we have expended the onetime architectural "leap", in that we have maximized the use of the architectural transition to extreme many-cores. As such, we are required to exploit other power conservation methods, such as DVFS and novel cooling methods, as well as further enhancing the many core usage, towards exascale. Thus, the follow-on project to ULP-HPC, the Ultra Green Supercomputing was proposed and funded directly by MEXT (the Japanese Ministry of Education, Sports, Culture, Science and Technology), to integrate the results and experience obtained from both the series of basic research and production TSUBAME supercomputers, and further deploy innovative power saving elements.

As a main deliverable and research platform of this project, TSUBAME-KFC was designed, constructed, and deployed in the fall of 2013. TSUBAME-KFC serves the following purposes.

First, TSUBAME-KFC serves as a prototype for TSUBAME3.0 to be commissioned in 2016 as a successor to TSUBAME2.5, and features extensive exploitation of many core architecture in extremely dense packaging, numerous sensors and control features of power and thermals. It relies even more on many-core GPU for providing performance both in compute and memory. As such TSUBAME-KFC features over 600 Teraflops of single precision performance in a single rack, and could easily scale to a petaflop for TSUBAME3.0 while maintaining the per-lack power consumption to the level of TSUBAME2.0, or approximately 35 KWatts per rack.

Secondly, TSUBAME-KFC facilitates extremely efficient cooling technology via warm liquid immersion, developed by GRC, in an unmanned modular environment as described in Section II. This method largely reduces power consumption for cooling since power hungry chillers are removed; as demonstrated in Section IV, the power consumption for cooling is less than 10% of IT power (power usage effectiveness is less than 1.1).

The overall scheme for achieving the 1000 fold power efficiency during the period of 2006-2017 is shown in Figure 1. Process scaling (the Moore's law) allows us to achieve nearly x100 increase in power performance the use of extreme many cores such as GPUs, plus proper software adaptation such as reduced precision, will give us x5; extensive active power and thermal control of the machine could net us as much as 50%; finally, efficient cooling, with possible provisions for energy recovery, could provide 40%; these factors are independent and thus could be multiplicative, netting x1000 improvement overall.

Although the experiments with TSUBAME-KFC will continue until the spring of 2016, it has already demonstrated world's top power efficiency. On the Green500 and the Green Graph 500 lists announced in November, 2013, TSUBAMEKFC became number one in the world in both, surpassing the second ranked machines on both lists by 24% respectively. This indicates that TSUBAME-KFC represents the best of the state-of-the-art in power efficiency in supercomping, and the results could even be carried over to IDCs for their improvements.

A. Discussion on Cooling

Ahead of overview of TSUBAME-KFC's cooling technology with warm liquid immersion, we discuss the cooling methodologies; While immersion cooling has been deployed in the past in machines such as the Cray-2, the Florinate coolant utilized was extremely expensive, and moreover evaporated at low temperature of 56 degrees Celsius, and in fact the vapor was collected to be re-condensed, requiring airtight packaging. In fact all the follow-on supercomputers except for ETA10 resorted to either low temperature air or water (nonsubmersive) cooling.

II. OVERVIEW OF TSUBAME-KFC — THE STATE-OF-THE-ART POWER EFFICIENT SUPERCOMPUTER DESIGN

For example, TSUBAME2 utilizes low temperature water to cool a refrigerator-like semi-sealed rack by HP called the MCS rack, and inside the rack there is a forced circulation of cooled air [9], and the server inside is air-cooled. For TSUBAME2.0, the inlet water temperature is approximately 7–8 degrees Celsius typical, while the outlet is 15–18 degrees. The inlet air to the server matches that, while the server outlet temperature is approximately +10ΔT. Although the cooling solution is far more efficient than conventional air cooling in SC and IDC centers, due to the chilled water and rack/node fan requirements, with the observed PUE (power usage effectiveness) of 1.29 on the year average basis, in that we are losing more than 20% of energy towards cooling.

TABLE I.

AVERAGE AMBIENT TEMPERATURE IN TOKYO

Month	Average Temperature (°C)	Average Wet-Bulb Temperature (°C)
Jan	6.1	2.1
Feb	6.1	2.1
Mar	9.4	5.0
Apr	14.6	10.3
May	18.9	14.9
Jun	22.1	18.5
Jul	25.8	22.4
Aug	27.4	23.0
Sep	23.8	21.1
Oct	18.5	14.9
Nov	13.3	9.2
Dec	8.7	5.2

One of the largest sources of power consumption of TSUBAME2 was identified to be cooling and semiconductor leakage power according to earlier prototype experiments and measurements on the production machine. In particular, average power of TSUBAME2 is slightly under a megawatt, but requires 8-9 degrees chilled water to be supplied by the chillers. Except for winter, ambient temperature of Tokyo is well above such (Table I), and thus involves operating power-hungry compressors. Also, when GPUs are in full operation,their temperature rises to nearly 80-90 degrees even with our chilled water enclosed air cooling. Through preliminary studies comparing air cooling to immersion cooling, we observed about 15% increase in power.

B. TSUBAME-KFC

In order to reduce overall total power usage, removing power-hungry compressors while keeping temperature of processors lower, we have decided to build a liquid immersion cooled, highly dense cluster with extensive power and thermal monitoring, with the requirements as depicted already in Section 1, called TSUBAME-KFC (Kepler Fluid Cooling). TSUBAME-KFC was designed and built in collaboration with NEC, NVIDIA, and most importantly, GRC that provided the liquid immersion cooling technology. Figure 2 is the overview of KFC, and Figure 3 the external view of the module and the cooling tower. TSUBAME-KFC submerges the servers in a warm liquid; although previous systems such as SuperMUC[14] have employed warm liquid cooling, since the latter is water-cooled, it required custom design of liquid pipes

TSUBAME-KFC: Ultra-Green Supercomputer Testbed

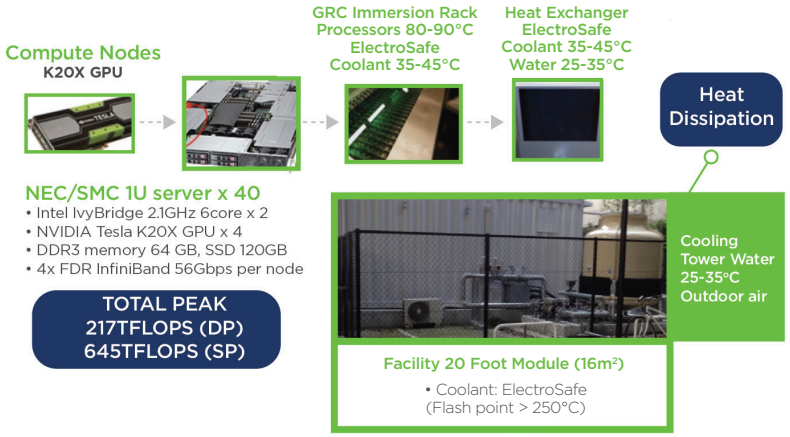

Fig. 2. TSUBAME-KFC cooling overview: the heat emitted from the server is transferred to GRC's ElectroSafe coolant, which then transfers the heat to a warm water loop, which in turn is cooled by an evaporative cooling tower.

with sufficient thermal capacity to cool the server. GRC's ICEraQ liquid immersion cooling (Figure 4) allows us to use standard servers with smaller degree of customization as is described later. Although the amount of coolant required is substantial, over 1000 liters compared to a few liters for standard pipe-based water cooling, it has the advantage of significant thermal capacity for easier control and resilience to power fluctuations. Immersion also has the advantage of effectively cooling all the components in the system, not just CPUs/GPUs or power supply, adding to increased cooling efficiency as well as long-term component stability.[1] Figure 5 shows how all the nodes are completely submerged in the ElectroSafe coolant.

In order to cool the coolant itself, there is a heat exchanger that transfers heat to the secondary water loop right next to the rack. The water in turn is cooled by ambient air / evaporative cooling tower right outside the TSUBAME-KFC 20-feet module. The cooling tower is a standard drip cooler where the water is slowly flowed to the bottom, cooling the water with ambient air through radiation and evaporation in the process. Although the cooling temperature cannot go below the dew point, because the maximum outlet target temperature of the TSUBAME-KFC rack is 35 degrees Celsius, as well as substantial capacity of both coolant and water in the cooling loop, allowing for averaging of the thermals throughout the machine without any excessive hotspot, preliminary analysis indicates that we do not require any chillers, even in the hottest summer Tokyo weather exceeding 35 degrees Celsius with high humidity. We believe the problem in the worst case scenario can be largely overcome with appropriate load control of the machine to not to reach maximum TDP. As such, the required cooling power consists of two pumps to circulate the coolant and water loops respectively, as well as a large fan internal to the cooling tower, plus replenishing of evaporated water.

C. Compute Nodes and Their Customization

TSUBAME-KFC employs 40 nodes of a customized version of standard highly-dense compute server NEC/

Fig. 3. Exterior view of TSUBAME-KFC: evaporative cooling tower right next to the 20-foot module for complete lights-out operation and low waterpump power

SMC 104Re-1G (Supermicro OEM) that embodies 2 CPUs and 4 GPUs in a dense 1U form factor:

- Intel Xeon E5-2620 v2 (IvyBridge) 6 Cores2.1GHz ×2
- DDR3 Memory 64GB
- NVIDIA Tesla K20X GPU ×4
- SATA3 SSD 120GB (Expanded with 2 × 500GB SSD March 2014).
- 4x FDR InfiniBand HCA ×1
- CentOS Linux (x86 64) 6.4
- GCC 4.4.7, Intel Compiler 2013.1.039
- CUDA 5.5
- OpenMPI 1.7.2

The theoretical peak performance of each node is 15.8 Teraflops in single precision and 5.3 Teraflops in double precision floating point respectively. With 40 nodes comprising a single rack, the combined performance is approximately 212 Teraflops in double precision and reaches over 632 Teraflops in single precision, approaching

[1]In fact, another claimed advantage of immersion is the elimination of all moving parts and sealing of the electronic contacts from the air, preventing them from long-term corrosive effects of sockets and connectors. One purpose of TSUBAME-KFC to conduct long term evaluation of this conjecture, but currently this is undergoing and will be a subject of future publications.

Fig. 4. The GRC ICEraQ system installed inside the module

Fig. 5. The compute nodes submerged in the GRC ICEraQ system

the petaflop/rack metric for exascale. Although standard servers were used as a baseline, the following customization were performed on the nodes jointly with GRC and NEC:

- Removal of moving components — In order to submerse to high viscosity liquid, all moving parts such as server fans (12 units) were removed, as well as employing SSDs for storage. This has the additional benefit of lowering the node power requirements.
- Thermal paste replacement — Since silicone grease between the processor and the passive cooler will dissolve in the liquid coolant, it was replaced with thin metallic sheets.

D. Coolant

The 40 servers are submersed in a 42U horizontal rack that is placed sideways instead of vertically. The theoretical peak power consumption of the nodes are approximately 40 kilowatts, far exceeding the standard 5~15 kilowatts per rack in IDCs, and slightly greater than TSUBAME2.0's 35 kilowatts, although the ICEraQ system has the capability to handle up to 100 kilowatts per rack.[2] In order to conform to regulations in Japan, the coolant had to be optimized carefully jointly with GRC and NEC. Regular mineral oil based coolants are non-toxic with low viscosity and fairly resistant to long-term oxidization, but nonetheless have the flash point of 177 degrees Celsius, and as a result considered as a flammable material under the Japanese fire laws subject to strict fire regulations equivalent to gasoline stands, with fairly stringent measures and licenses required for both the installation (such as comprehensive fire extinguishers and periodic inspections) as well as the operators (licenses approved through national examinations), infeasible for large-scale supercomputer center operations. After extensive research, GRC's ElectroSafe+ coolant was proposed, with flashpoint of 260 degrees, well above the threshold of 250 degrees avoiding such complications.[3] The selected coolant is also reasonable in the aspect of price.

E. Power and Thermal Measurements

TSUBAME-KFC embodies numerous power and thermal sensors, both on-line integral to the servers and other IT equipment, as well as independent off-line sensors. All the sensor streams are aggregated and can be observed in realtime as well as archived collectively. For example, the node power sensors are off-line, with individual Panasonic KW2G sensors and AKL1000 data logger, allowing server and switch measurement to be done in real time without any performance intrusion every second (Figure 6). Also infrastructural powers such as coolant and water pumps, as well as cooling tower are measured. The list of sensors and their measurement intervals are described in Table II.

Although the sensors might seem too extensive for a production supercomputer, we believe most of the sensors will be incorporated into production TSUBAME3.0, to realize finegrain control of the machine for power minimization.

The entire TSUBAME-KFC system was completed and began operation in October 2013, and will continue with various experimentations leading up to TSUBAME3.0, until Spring of 2016.

III. POWER EFFICIENCY METRICS

For completeness we describe the power efficiency being used by the Green500 list, as well as the PUE (Power Usage Effectiveness). Admittedly there have been numerous discussions regarding the metrics in the recent years, in that they only capture certain facets of power efficiency of a given machine, and should not be taken as ultimate decisive values.

Nonetheless for brevity we will not be controversial in this paper, but rather accept these as established metrics with well reported results that allows us to compare the power

[2]Some "racks" claim to have greater thermal density than TSUBAME-KFC; for example, BlueGene/Q is known to embody 60 kilowatts per rack. However, volume-wise they have a substantially bigger rack, at 2.08m×1.22m×1.22m it is 2.3 times larger than TSUBAME-KFC and 1.3 times area wise.
[3]Nonetheless, the facility was subject to inspection by the fire marshal for approval during planning and after completion.

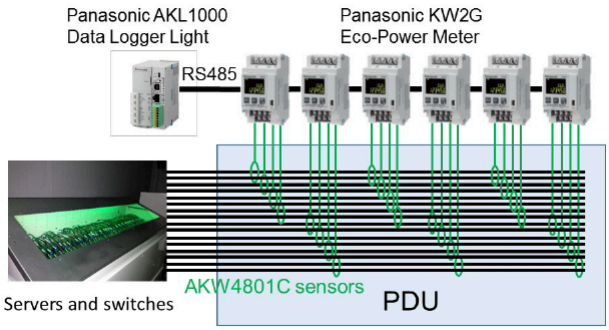

Fig. 6. Power measurement system of TSUBAME-KFC nodes

TABLE II.

Measured Component	Type	Provided By	Interval	Resolution
Computer node	Power	Panasonic data logger	1 sec.	0.1W
Network	Power	Panasonic data logger	1 sec.	0.1W
Cooling tower	Power	Panasonic data logger	1 sec.	0.1W
Pump (coolant)	Power	Panasonic data logger	1 sec.	0.1W
Pump (water)	Power	Panasonic data logger	1 sec.	0.1W
Outdoor air	Temperature	Panasonic data logger	10 sec.	0.1°C
Outdoor air	Humidity	Panasonic data logger	10 sec.	1%
Indoor air	Temperature	Panasonic data logger	10 sec.	0.1°C
Indoor air	Humidity	Panasonic data logger	10 sec.	1%
CPU, GPU	Temperature	IPMI on nodes (BMC)	<1 sec.	1°C
Pump (coolant)	Speed	GRC controller box	20 sec.	2%
Pump (coolant)	Power	GRC controller box	20 sec.	0.1W
Coolant	Temperature	GRC controller box	20 sec.	0.1°C
Water	Temperature	GRC controller box	20 sec.	0.1°C

efficiency with other systems at reasonable degrees of accuracy.

A. PUE

PUE (power usage effectiveness) is widely used in data centers as a metric to measure the power efficiency of cooling, and is given by:

$$\text{PUE} = \frac{\text{(IT Equipment Power + Infrastructure Power)}}{\text{IT Equipment Power}}$$

In this article, we equate cooling power to infrastructure power. PUE=1 indicates the ideal case of no power required for cooling, while PUE=2, which is a fairly common value in classical data centers, indicate that cooling is spending as much power as the IT equipment. As indicated,TSUBAME2.5 PUE is measured to be 1.29 on they early average, which is far superior to such values, but still is spending (1 – 1/1.29) = 23% power to cooling, and more if we account for the chassis fans. One caveat with PUE is that, low PUE may not necessarily indicate overall power efficiency. This has been a problem pointed out recently with the so called ambient air cooling, in which natural outside air is used instead of cooled air by CRAC units. Due to the higher ambient temperature, the operational temperature of machines become considerably higher than traditional cooling methods, increasing the overall IT power with increased fan speeds and higher leakage current, both of which are accounted for as IT power, thus enlarging the denominator — the end result being lower PUE but higher overall datacenter power.

Thus it is important to measure PUE in terms of baseline power consumption under traditional cooling methods, not what is measured with new cooling solutions.

B. The Power Efficiency Metric of the Green500

Green500 basically measures the power efficiency of a machine during the Linpack run as dictated by the run rules of the Top500 Linpack benchmark[12], and the measured Linpack Flops/Watt is used as a metric to rank the machine, up to the 500th on the Top 500 list. A machine has to be listed on the Top 500 list to qualify to be ranked on the Green500, but a more power efficient run could be conducted and reported; otherwise the submitted Top 500 power result will be used.

The power measurement in the denominator of the Top500 is measured under the following conditions:

1) Overall power of the compute node and the network are measured and included.
2) Storage power are not included, unless it is integral to the node such as local disks.
3) Power for cooling are not included, unless it is integral to the node such as chassis fans.
4) For the Level 1 measurement, minimum 20% or 1 minute timespan of the duration of the entire Linpack run, whichever is longer, needs to be measured and averaged. The first 10% of the beginning and the last 10% of the Linpack run cannot be included in the measurement.

As specified in condition 3 above, one of the criticisms of the Green500 is the failure to include the overall cooling power requirements, i.e., advances in efficient cooling technologies do not have direct relevance to the Green500 measurement and ranking. Nonetheless, efficient cooling could have indirect effect such as elimination of chassis-internal cooling components, as well as lower thermal operations for components leading to lower power.

Also, condition 4 indicates that, for Level 1 (and Level 0) measurement only a subset duration needs to be measured. Due to the nature of Linpack, whose power consumption gradually drops in modern architectures as the unsolved matrix becomes smaller and thus more communication intensive, the reported number is typically smaller than the overall average (Level 2 measurement) or during peak. For instance, our November 2013 submission of the TSUBAME2.5 supercomputer recorded 2.831 Petaflops, and the average power consumption for the entire run was 1125 kilowatts. On the other hand, the Level 1 measurement of the 70–90% duration of the run was 922.5 kilowatts, resulting in 3.069 GFlops/Watt submission measurement ranked 6th in the world.

Although there are some controversy as to whether the run rule constitutes a valid power measurement for Linpack, nonetheless as a ranking metric could be considered "fair" if they are measured the same way across all the machines. However, due to the one minute rule above, small machines whose power degradation timespan is much shorter than one minute is disadvantaged in principle.

TABLE III.

COMPARISON OF TSUBAME-KFC SUBMERSED NODE AND AIR-COOLED NODE

Cooling	Air-Cooled (26°C)	Immersion Coolant (29°C)	Immersion Coolant (19°C)
Temp (°C)			
CPU1	46	42	33
CPU2	50	40	31
GPU1	52	47	42
GPU2	59	46	43
GPU3	57	40	33
GPU4	48	49	42
Node Power (W)	**749**	**693**	**691**

IV. TSUBAME-KFC EVALUATION

A. Effect of Liquid Immersion Cooling on the Servers

We first measured how liquid immersion cooling affects the power and thermals of individual, densely configured TSUBAME-KFC node. We configured an air-cooled node with exactly the same CPU/GPU/memory hardware for comparative purpose. As mentioned earlier, the air-cooled node has the original 12 fans, while the immersed node have the fans eliminated.

Table III shows the comparison between air-cooled with inlet 26 degrees Celsius, versus 29 and 19 degrees inlet for the ElectroSafe coolant. The servers are continuously running double precision matrix multiply using CUBLAS to incur the highest power and thermal load.

Comparing the air-cooled versus immersion, although the former has lower temperature input, the latter exhibits substantially lower temperature, especially GPU2 and GPU3 where difference is more than 10 degrees, or ΔT of 33 degrees for air compared to 20 degrees for ElectroSafe. This is due to much higher thermal capacity of ElectroSafe, especially since these GPUs are inline to the airflow path of GPU1 and GPU4, being affected by the already warmed air. The result for 19 degree coolant inlet is even more significant, ΔT being fairly consistent 24 degrees.

Comparing the server power consumption, liquid immersion is approximately 7.8% lower than air, while inlet temperature difference has very little effect on the overall power. As mentioned earlier, this is largely the combined effect of fan removal and lower semiconductor temperature suppressing leakage current. That there is small difference between the two temperature points of liquid immersion could indicate that the former is more dominant, but the prior experiments have indicated that the latter effect is also significant. This could be possibly explained by the exponential effect of leakage current versus temperature being exponential in nature, and thus that even at 29 degrees the component temperature was too low to exhibit the difference not hitting the rising "knee" of the curve, while at higher temperature it would quickly rise. We plan to conduct more thorough experiments during summer months to investigate the temperature point at which the component temperature will start to exhibit noticeable increase in server power consumption, a valuable data for power control in that we would want to control the coolant temperature just under this point to minimize cooling power.

B. Power and PUE Measurement

In order to measure TSUBAME-KFC power consumption and PUE, we stressed the server with the highest load of CUBLAS matrix multiply as in the previous subsection for all the nodes. Figure 7 shows the results. The PUE number for TSUBAME is derived from the actual measurements from the real-time power sensors, while the air cooling was extrapolated from real power measurement of the server, with the assumption that the state-of-the-art air cooling would be as efficient as TSUBAME2.0's PUE of 1.29.

According to the measurements, power consumption of the TSUBAME-KFC IT equipment was 28.9 kilowatts while the pump and cooling tower power combined were 2.60 kilowatts for apparent PUE of 1.09. However, as mentioned earlier, the PUE comparison here is misleading, as the node server power consumption itself has gone down significantly. The total power usage of 31.5 kilowatts is 22% smaller than 40.7 kilowatts in the air-cooled case. In fact, the total power usage of 31.5 kilowatts is essentially equivalent to the IT-only power usage of air-cooled machine, being 31.2 kilowatts. As such TSUBAME-KFC efficiency cannot merely be judged by comparing PUE values alone.

C. Green500 Measurement

In order to measure the efficiency of TSUBAME-KFC under more realistic setting, we challenge the Green 500 for utmost efficiency. Although both benchmarks do not directly measure contributions from cooling, nonetheless we expect higher efficiency through improved node efficiency as described above, as well as improved tuning for power efficiency rather than absolute performance. As an initial comparative measure, we targeted the previous #1 system for the June 2013 edition of the Green 500, namely the 3.209 Gigaflops/Watt record of the CINECA/Eurotech Aurora machine.

In order to achieve the maximum power efficiency, we employed the following strategies at the software and hardware control levels:

- At the architecture level we increased the ratio of GPU to CPU ratio from 1:1 to 2:1, thus decreasing the overhead of CPU and other peripheral power consumption(The GPU was also slightly slower, being NVIDIA K20 instead of K20X for KFC, but the overall effect on the performance is believed to be low for theGreen500.)
- We employed a new, more efficient in-core Linpack kernel provided by NVIDIA for both TSUBAME-KFC and TSUBAME2.5 measurements. This version only computes Linpack using the memory space of GPUs, not of the whole node, for efficiency and better power-performance. However, this also results in a much shorter runtime, and for a small machine such as TSUBAME-KFC, this hinders Level-1 measurement due to the one minute rule as described earlier.
- We tuned the HPL parameters to the maximum extent by exhaustive search of the parameter space. This involved not only the standard tuning of HPL parameters such as the block size (NB), and process grid (P&Q), but also

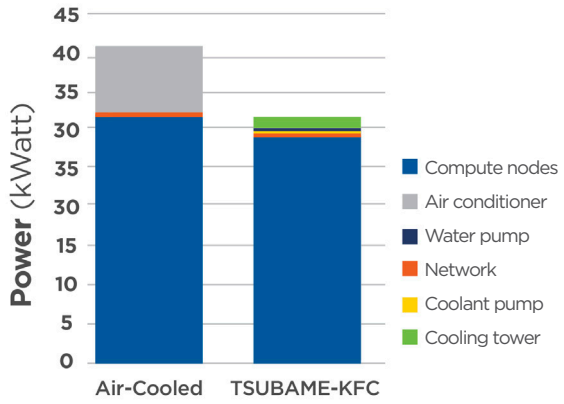

Fig. 7. PUE EvaluationofTSUBAME-KFCversus Air-CooledMachine

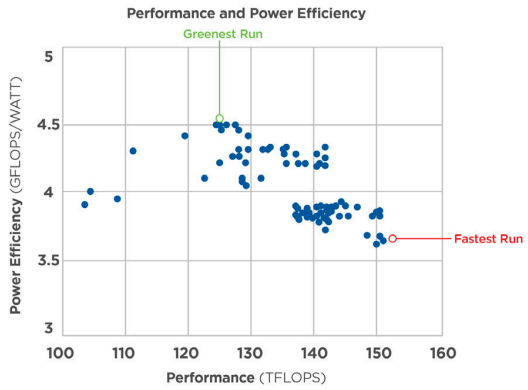

Fig. 8. Results of Linpack benchmark runs with various configurations. Each dot corresponds to a single Linpack run.

adjustment of the GPU clock and voltage, where the slowest clock of 614 MHz proved to be the most power efficient, compared to the default of 732 MHz (Available GPU clocks rates were 614(best), 640, 666, 705, 732 (default), 758, 784 (MHz)).

- We conducted measurements at the most thermally low and stable night hours.

During the tuning phase, we noticed that the best performance does not equal best power efficiency. Figure 8 shows power efficiency of Linpack runs with various configurations; we observed the best power efficiency is 24% better than the case with the best speed performance.

As a result, Flops/Watt of reached 4.503 GFlops/Watt, improving the CINECA record by over 40% (Table IV). On November 18th, 2013, the Green500 list was announced, in which TSUBAME-KFC was ranked #1 in the world, with 24% lead over Wilkes, the second ranked machine. In fact, the top three machines were similarly configured with 2:1 GPU ratio of NVIDIA K20X GPUs versus Intel Ivy Bridge Xeon CPUs and Infiniband interconnect. We attribute the difference to better cooling as well as more extensive tuning of the parameters. In the latest list announced in June 2014, TSUBAME-KFC was ranked #1 again; the efficiency number is slightly changed to 4.390GFlops/Watt, since we had to choose another Linpack run for submission in order to keep the system in the latest Top500 list. Also TSUBAME-KFC obtained #1 in another ranking, namely Green Graph500 list [13] in November 2013, designed for competition for power efficiency in big-data analysis area, although we omit details for want of space.

D. Evaluation with the Phase-Field Simulation

We also evaluated power efficiency of a real application, namely the stencil based "phase-field" simulation, which was awarded the 2011 Gordon Bell prize[15] by achieving 2 Petaflops on the TSUBAME2.0. This application simulates the micro-scale dendritic growth of metal materials during solidification phase. We have used it for evaluation of power efficiency for multiple years in our projects and have comprehensive power performance records executing on TSUBAME1.0 machine commissioned in 2006.

When we run the application on a single TSUBAME-KFC node with four GPUs, we observed 3.62 TFlops (single precision)with the power consumption of 652 Watt; as such the power efficiency is 5.55 GFlops/Watt. Table V compares the result with that on an air-cooled machine with the same configuration, demonstrating that the immersion cooled machines provides 8.5% higher efficiency, consistent with our Green500 measurement.

TABLE IV.

POWER EFFICIENCY IN THE GREEN500 METRICS

System	Time	Speed (TFLOPS)	Power (KW)	Power Efficiency (GFlops/W)
TSUBAME2.0	NOV. 2010	1192	1244	0.958
CINECA	Jun. 2013	98.51	30.70	3.209
TSUBAME2.5	Nov. 2013	2831	922.5	3.069
Wilkes	Nov. 2013	191.1	52.62	3.632
TSUBAME-KFC	Nov. 2013	125.1	27.78	**4.503**

Figure 9 also shows the development of power efficiency by comparing several machines in different generations since 2006. Here we observe a gap between the 2006 (CPU only) and 2008 (with GPUs) numbers, which is the one-time performance leap with the many-cores transition. By extrapolating the results of multiple generations of GPU machines, we estimate that the expected performance circa 2016 being as 15.5 GFlops/Watt, which is 1,200 times more efficient than the 2006 number. This result supports our initial assessment of our ability to achieve the target improvement depicted in Figure 1, i.e., x1,000 in 10 years.

V. CONCLUSION

We demonstrated a prototype supercomputer that combines most of the known the state-of-the-art architecture in both hardware and software, materialized as TSUBAME-KFC, achieving the world's top power efficiency in Green500 rankings circa November 2013 and June 2014. TSUBAME-KFC improved the previous Green500 #1 record by over 40%, and was 24% better than similarly configured machine ranked #2. The most notable technology deployed by TSUBAME-KFC is the liquid immersion cooling; the total

TABLE IV.

POWER EFFICIENCY OF THE PHASE-FIELD SIMULATION ON TSUBAME-KFC

System	Speed (SP TFLOPS)	Power (KW)	Power Efficiency (GFlops/W)
A KFC node	3.62	0.652	5.55
An air-cooled node	3.58	0.701	5.11
40 KFC nodes	126	25.5	4.93

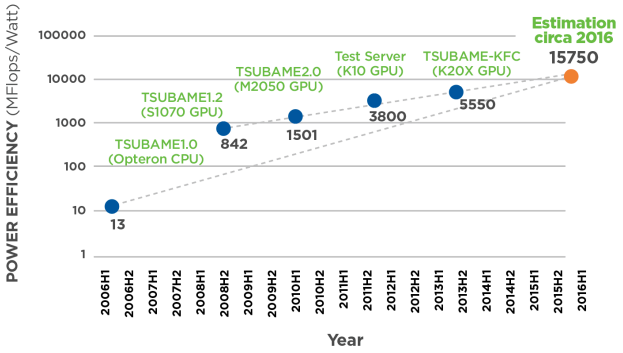

Fig. 9. Power efficiency of the phase-field simulation on machines in different generations

power consumption is reduced by 29% over compute nodes with the same configurations. These power saving features of TSUBAMEKFC will be incorporated into TSUBAME3.0, as is or with some pragmatic adaptations.

Not only as a prototype of TSUBAME3.0, TSUBAME-KFC also intends to be a platform for reproducible experiments regarding power saving. It is well known that the rise in semiconductor temperature results in substantial increase in power requirements, due to the increase in leakage current. As such, it is very difficult to conduct a reproducible experiment, maintaining constant thermal conditions. By immersion cooling with liquids of massive thermal capacity, we can control the thermals more easily; as is described later, TSUBAME-KFC is immersed in 1200 liters of liquid, allowing high thermal stability thus reproducibility all year round.

We will continue our experimentations of TSUBAME-KFC, some of the most up-to-date-result only obtainable during summer in the camera-ready version of the paper, as well as longer term data such as long-term component faults. As an experimental platform, we will conduct further customization updates either in software or hardware if affordable, to affect the design of TSUBAME3.0 as the bleeding-edge power efficient and big data supercomputer of the era. We also hope to open up TSUBAME-KFC for uses by our collaborators so as to obtain reproducible results in power efficient computing on the state-of-the-art architecture.

ACKNOWLEDGMENTS

This research is funded by Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Ultra Green Supercomputing and Cloud Infrastructure Technology Advancement. We also deeply thank the extensive collaborations with NEC, NVIDIA, GRC, Supermicro, Mellanox, as well as GSIC/Tokyo Institute of Technology.

REFERENCES

[1] W. Feng: Making a Case for Efficient Supercomputing. ACM Queue, 1(7):54-64, October 2003.
[2] IBM Journal of Research and Development, special double issue on Blue Gene, Vol.49, No.2/3, March/May, 2005.
[3] Hiroshi Nakashima, Hiroshi Nakamura, Mitsuhisa Sato, Taisuke Boku, Satoshi Matsuoka, et. al. (2 more authors), "MegaProto: 1 TFlops/10 kW Rack Is Feasible Even with Only Commodity Technology", Proc. IEEE/ACM Supercomputing 2005, the IEEE Computer Society Press, Nov. 2005.
[4] International Exascale Software Project, http://www.exascale.org/iesp
[5] Japan Science and Technology Agency, CREST, Development of System Software Technologies for Post-Peta Scale High Performance Computing. http://www.postpeta.jst.go.jp/en
[6] K. Datta, M. Murphy, V. Volkov, S. Williams, J. Carter, L. Oliker, D. Patterson, J. Shalf, and K. Yelick, "Stencil Computation Optimization and Auto-tuning on State-of-the-Art Multicore Architectures", IEEE Supercomputing (SC08), Article No. 4, pp. 1-12, 2008.
[7] Song Huang, Shucai Xiao, and Wu-chun Feng, "On the Energy Efficiency of Graphics Processing Units for Scientific Computing", IEEE International Symposium on Parallel & Distributed Processing (IPDPS 2009), pp. 1-8, 2009.
[8] Shiqiao Du, Takuro Udagawa, Toshio Endo and Masakazu Sekijima, "Molecular Dynamics Simulation of a Biomolecule with High Speed, Low Power and Accuracy Using GPU-Accelerated TSUBAME2.0 Supercomputer", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011), pp. 1-5, 2011.
[9] Satoshi Matsuoka. "TSUBAME 2.0 Begins – The long road from TSUBAME1.0 to 2.0 (Part One) –", The TSUBAME E-Science Journal, Tokyo Tech. GSIC, Vol. 2, pp. 2-10, Nov. 2010
[10] Top500 Supercomputer Sites, http://www.top500.org
[11] The Green500 List, http://www.green500.org
[12] A. Petitet, R. C. Whaley, J. Dongarra, A. Cleary, "HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers", http://www.netlib.org/benchmark/hpl/
[13] The Green Graph500 List, http://green.graph500.org
[14] Leibniz Supercomputing Centre, "SuperMUC Petascale System". http://www.lrz.de/services/compute/supermuc
[15] T. Shimokawabe, T. Aoki, T. Takaki, A. Yamanaka, A. Nukada, T. Endo, N. Maruyama, S. Matsuoka. "Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer", IEEE/ACM Supercomputing (SC11), pp. 1–11, Seattle, November 2011.

Note: The K20X GPUs were changed to K80s in 2015. The name of the system is now TSUBAME-KFC/DL.