The Total Picture of TSUBAME 2.0

GPU Computing for Dendritic Solidification based on Phase-Field Model
in terms of the number of processors, CPUs have been much more dominant. In TSUBAME 2.0, each node is equipped with three GPU accelerators, which enable a great leap not only in performance but also energy efficiency.

TSUBAME 2.0’s main compute node, manufactured by Hewlett-Packard, has two Intel Westmere EP 2.93GHz processors, three NVIDIA Fermi M2050 GPUs and about 50GB memory. Each node provides computation performance of 1.7 teraflops (TFLOPS), which is about 100 times larger than that of typical laptop PCs. The main part of the system consists of 1,408 computing nodes, and the total peak performance will reach 2.4 PFLOPS, which will surpass the aggregated performance of all the major supercomputers in Japan.

Each CPU in TSUBAME 2.0 has six physical cores and supports up to 12 hardware threads with the hyper-threading technology, achieving up to 76 gigaflops (GFLOPS). The GPU sports NVIDIA’s new processor architecture called Fermi and contains 448 small cores with 3GB of GDDR5 memory with 515 GFLOPS performance at maximum.

In general, efficient use of GPUs requires different programming methodologies than CPUs. For this purpose, CUDA and OpenCL are supported so that users can execute programs designed for TSUBAME 1.2 on the new system. Also Tesla M2050 GPUs have advantage both in performance and programmability; the adoption of the true hardware cache will make performance tuning of programs much easier.

Key Technologies toward Petaflops

Basically TSUBAME 1.0/1.2 and 2.0 are supercomputing clusters, which consist of a large number of usual processors such as Intel compatible CPUs. Additionally, they are equipped with processors designed for specific purposes, called accelerators, in order to significantly improve performance of scientific applications based on vector computing. We have experiences in operating ClearSpeed accelerators on TSUBAME 1.0, and 680 NVIDIA Tesla GPUs in TSUBAME 1.2, and learned various technologies for energy-efficient high-performance computing with accelerators. However, in terms of the number of processors, CPUs have been much more dominant. In TSUBAME 2.0, each node is equipped with three GPU accelerators, which enable a great leap not only in performance but also energy efficiency.

TSUBAME 2.0’s main compute node, manufactured by Hewlett-Packard, has two Intel Westmere EP 2.93GHz processors, three NVIDIA Fermi M2050 GPUs and about 50GB memory. Each node provides computation performance of 1.7 teraflops (TFLOPS), which is about 100 times larger than that of typical laptop PCs. The main part of the system consists of 1,408 computing nodes, and the total peak performance will reach 2.4 PFLOPS, which will surpass the aggregated performance of all the major supercomputers in Japan.

Each CPU in TSUBAME 2.0 has six physical cores and supports up to 12 hardware threads with the hyper-threading technology, achieving up to 76 gigaflops (GFLOPS). The GPU sports NVIDIA’s new processor architecture called Fermi and contains 448 small cores with 3GB of GDDR5 memory with 515 GFLOPS performance at maximum.

In general, efficient use of GPUs requires different programming methodologies than CPUs. For this purpose, CUDA and OpenCL are supported so that users can execute programs designed for TSUBAME 1.2 on the new system. Also Tesla M2050 GPUs have advantage both in performance and programmability; the adoption of the true hardware cache will make performance tuning of programs much easier.

One of key features of TSUBAME 2.0 architecture is adopting hardware with highly improved bandwidth in order to enable data communication efficiently. To keep performance of intra-node communication high, the memory bandwidth reaches up to 32 GB/s on CPU and 150 GB/s on GPU. Communication between CPUs and GPUs is supported by the latest PCI-Express 2.0 x16 technology with bandwidth of 8GB/s.

As the interconnect that combines more than 1,400 computing nodes and storage described later, TSUBAME 2.0 uses the latest QDR InfiniBand (IB) network, which features 40Gbps bandwid
As a supercomputing system that supports e-Science, large scale storage system that supports fast data access is necessary. TSUBAME 2.0 includes a storage system with capacity of 7.1 petabytes (PB), which is six times larger than that of TSUBAME 1.0. As typical usage, users can store large scale data used by their tasks; additionally, the storage system is used to provide Web-based storage service, which users in Tokyo-Tech can use easily.

The storage system mainly consists of two parts: 1.2 PB home storage volume and 5.9 PB parallel file system volume. The home storage volume is designed so that it provides high reliability, availability and performance based on redundant structure. Especially, it provides up to 1100 MB/s of accelerated NFS performance, via QDR IBs and 10Gbps networks. The volume also supports other protocols, such as CIFS, iSCSI, in order to support transparent data access from all computing nodes running both Linux and Windows. It is also used for various storage services for educational and clerical purpose in Tokyo Tech.

Despite the significant improvements in performance and capacity, the power consumption of TSUBAME 2.0 will be as comparable as the current one, thanks to a lot of research and engineering advances in power efficiency. Cooling is also greatly improved by exploiting much more efficient water-cooling systems. We expect the power usage effectiveness (PUE, an index to evaluate energy efficiency of computing facility) of TSUBAME 2.0 will be as low as 1.2 on the average.

Large Scale Storage for e-Science

As a supercomputing system that supports e-Science, large scale storage system that supports fast data access is necessary. TSUBAME 2.0 includes a storage system with capacity of 7.1 petabytes (PB), which is six times larger than that of TSUBAME 1.0. As typical usage, users can store large scale data used by their tasks; additionally, the storage system is used to provide Web-based storage service, which users in Tokyo-Tech can use easily.

The storage system mainly consists of two parts: 1.2 PB home storage volume and 5.9 PB parallel file system volume. The home storage volume is designed so that it provides high reliability, availability and performance based on redundant structure. Especially, it provides up to 1100 MB/s of accelerated NFS performance, via QDR IBs and 10Gbps networks. The volume also supports other protocols, such as CIFS, iSCSI, in order to support transparent data access from all computing nodes running both Linux and Windows. It is also used for various storage services for educational and clerical purpose in Tokyo-
We have provided the large computing service (HPC queue), for so-called capability jobs where a user group can use up to 1000 CPU cores and 120 GPUs exclusively in TSUBAME 1.2. For TSUBAME 2.0, we will further enhance this feature, providing up to 10,000 CPU cores and thousands of GPUs to selected user groups under regulated reservation and/or peer review process.

We will federate a web portal, called TSUBAME portal with Tokyo Tech portal to enhance usability such as paperless account application. For Tokyo Tech users, TSUBAME accounts are unified with accounts of Tokyo Tech portal. Additionally we will provide account services across Japan as one of the nine leading National University supercomputing centers and its national alliance.

A network storage service will be provided to Tokyo Tech users by using the large scale storage of TSUBAME 2.0. Users can easily utilize the storage from their PCs without recognizing existence of TSUBAME.

We promote data-oriented e-Science using TSUBAME 2.0’s high-end storage resources, which has been difficult for traditional Japanese supercomputing centers. For this purpose, TSUBAME 2.0 and national supercomputing centers are tightly couple with 10Gbps class networks, called SINET. This will enable data sharing and transportation service via GFarm file system and GridFTP using our newly developed RENKEI-PoP technology which are now being deployed at various national supercomputing centers.

As of writing this article, system construction and detailed investigation of new operational method are ongoing. When the operation starts in November, please utilize TSUBAME 2.0, one of the largest supercomputer in the world, for leading innovation in science and technology.

TSUBAME 2.0 News Web:
http://www.gsic.titech.ac.jp/tsubame2

Official TSUBAME 2.0 Twitter account:
Tsubame20

Cloud Operation of TSUBAME 2.0

As “supercomputer for everyone”, TSUBAME 1.0/1.2 has provided availability so that users, who have been used PCs or small clusters, can sport supercomputing power. In TSUBAME 2.0, while inheriting the availability, we promote the following services in order to expand the range of variable users.

● While the typical usage of computing power of TSUBAME 2.0 remains to be throwing jobs via a batch queue system, it supports not only Linux OS (SUSE Linux Enterprise 11), but also Windows HPC Server 2008. This new operation is supported by virtual machine (VM) technology.

● We continue the hosting service using VM technology in Tokyo Tech campus. Additionally, with VM technology, we plan to improve efficiency of computing resources by suspending some low-priority jobs and dividing a single node into several virtual nodes.
In this article, we study the growth of the dendritic solidification of a pure metal in supercooling state by solving the equations derived from the phase-field model. Finite difference discretization is employed and the GPU code developed in CUDA is executed on the GPU of TSUBAME 1.2. The remarkably high performance is shown in comparison with the conventional CPU computing. Although most GPGPU applications run on single GPU, we exploit a multiple GPU code and show the strong scalability of large-scale problems.

The phase-field model is a phenomenological simulation method to describe the microstructure evolution in sub-micron scale based on the diffuse-interface concept. In this article, to describe the interface between solid phase and liquid phase, we define a non-conserved order parameter (phase field) $\phi$ taking a value of 0 in the liquid phase and 1 in the solid phase. In the interface region, $\phi$ gradually changes from 0 to 1. The position at $\phi = 0.5$ can be defined as the solid/liquid interface. Using this diffuse-interface approach, the phase-field method does not require explicit tracking of the moving interface.

In this study, a time-dependent Ginzburg-Landau equation for the phase field $\phi$ and a heat conduction equation are solved. The governing equations for the phase field $\phi$ and the temperature $T$ are given by the following equations:

\begin{equation}
\begin{align*}
\frac{\partial \phi}{\partial t} & = \lambda f \\
& \quad + \frac{\partial}{\partial x} \left( \epsilon \frac{\partial \phi}{\partial x} + \frac{\partial}{\partial x} \left( e \phi - \frac{\partial}{\partial x} \sqrt{2\phi^2} \right) \right) \\
& \quad + \frac{\partial}{\partial y} \left( \epsilon \frac{\partial \phi}{\partial y} + \frac{\partial}{\partial y} \left( e \phi - \frac{\partial}{\partial y} \sqrt{2\phi^2} \right) \right) \\
& \quad + \frac{\partial}{\partial z} \left( \epsilon \frac{\partial \phi}{\partial z} + \frac{\partial}{\partial z} \left( e \phi - \frac{\partial}{\partial z} \sqrt{2\phi^2} \right) \right) \\
& \quad + 4W\phi \left( 1 - \phi \right) \left( 1 + 6\phi - \frac{3}{2} \right)
\end{align*}
\end{equation}
\[ \frac{\partial T}{\partial t} = \kappa \nabla^2 T + 30 \psi^2 (1 - \psi) \frac{T}{C} \frac{\partial \psi}{\partial t} \]  

Here, \( M \) is the mobility of the phase field \( \psi \), \( \varepsilon \) is the gradient coefficient, \( W \) is the potential height and \( \beta \) is the driving force term. These parameters are related to the material parameters as follows.

\[ M = \frac{b^2 \varepsilon}{3N} \]  

\[ W = \frac{6\varepsilon b}{\delta} \]  

\[ \varepsilon = \sqrt{\frac{5b\varepsilon b}{b} \left( 1 - 3b + 4b \frac{\psi^*}{\psi^*} + \psi^* \right) \psi^*} \]  

\[ \beta = -\frac{15L}{2W} \frac{T - T_m}{T_m} \psi (1 - \psi) \]  

In this article, the material parameters for pure nickel are used. The melting temperature \( T_m = 1728 \) K, the kinetic coefficient \( \mu = 2.0 \) m/Ks, the interface thickness \( \delta = 0.08 \) \mu m, the latent heat \( L = 2.35 \times 10^9 \) J/m\(^3\), the thermal conductivity \( \kappa = 1.55 \times 10^{-5} \) m\(^2\)/s and the specific heat \( C = 5.42 \times 10^3 \) J/kg\(^\circ\)C. Furthermore, the strength of interface anisotropy \( \gamma = 0.04 \) in the interface region \( \lambda < \varphi < 1 - \lambda \), we obtained \( b = \tanh^{-1}(1 - 2\lambda) \). \( \lambda \) is set to be 0.1, so that \( b \) reduces to 2.20. \( \chi \) is a random number distributed uniformly in the interval [-1, 1]. \( \alpha \) is the amplitude of the fluctuation and set to be 0.4.

### 3-1 Tuning Techniques

Equations (1) and (2) are discretized by the second-order Finite Difference Method and time-integrated with the first-order accuracy (Euler scheme). The arrays for the dependent variables \( \varphi \) at the \( n \) and \( n+1 \) time steps are allocated on the VRAM, called the global memory in CUDA. We minimize the data transfer between the host (CPU) memory and the device memory (global memory) through the narrow PCI-Express bus, which becomes a large overhead of the GPU computing.

In CUDA programming, the computational domain of a \( nx \times ny \times nz \) mesh is divided into \( L \times M \times M \) smaller domains of a \( MX \times MY \times MZ \) mesh, where \( MX = nx/L, MY = ny/M, MZ = nz/N \). We assign the CUDA threads \( (MX, MY, 1) \) to each small domain in the \( x \)- and \( y \)-directions. Each thread computes \( MZ \) grid points in the \( z \)-direction using a loop. The GPU performance strongly depends on the block size and we optimize it to be \( MX = 64 \) and \( MY = 4 \).

The discretized equation for the phase field variable \( \varphi \) is reduced to the stencil calculation referring to 18 neighbor mesh points. In order to suppress the global memory access, we use the shared memory as

![Fig.1 Snap shots of the dendritic solidification growth](image-url)
a software managed cache. Recycling three arrays with a size of \((MX + 2) \times (MY + 2)\) saves the use of the shared memory. In computing the temperature, the shared memory is used similarly, however in this case the time derivative term \(\partial \varphi / \partial t\bigg|_{n+1}^{n} \) appears in the right-hand side of Eq. (2). We fuse the computational kernel function \( \varphi_{n+1}^{i,j,k} \rightarrow \varphi_{n}^{i,j,k} \) with the kernel function of \( T_{n+1}^{i,j,k} \rightarrow T_{n}^{i,j,k} \), so that it becomes unnecessary to access the global memory by keeping the value \( \partial \varphi / \partial t\bigg|_{n+1}^{n} \) on a temporal variable in the kernel function.

3-2 Performance of Single GPU

In order to evaluate the performance of the GPU computing and check the numerical results in comparison with CPU, we also built the CPU code simultaneously. Since integer calculations are also done by checking the numerical results in comparison with CPU, we also built the hardware counter of the PAPI (Performance API) operation of the calculation for the dendrite solidification by using the CPU code.

The maximum mesh size of the run is \(640 \times 640 \times 640\) on single GPU, because one GPU board of Tesla S1070 has 4 GByte VRAM (GDDR3). By changing the mesh size, we measured the performance of the GPU computing and we had 116.8 GFlops for \(64 \times 64 \times 64\) mesh, 161.6 GFlops for \(128 \times 128 \times 128\) mesh, 169.8 GFlops for \(256 \times 256 \times 256\) mesh, 168.5 GFlops for \(512 \times 512 \times 512\) mesh and 171.4 GFlops \(640 \times 640 \times 640\) mesh. The performance of the single CPU core (Opteron 2.4 GHz) is 898 MFlops, and it was found to be a 190x-speedup on TSUBAME1.2.

The phase-field calculation consists of 373 floating point operations and 28 times global memory access (26 reads and 2 writes) per one mesh point. The same calculation is carried out on every mesh point in single precision. The arithmetic intensity is estimated to be 3.33 Flop/Byte. In the case when using the shared memory, the number of memory read reduces to 2 and the arithmetic intensity increases up to 23.31 Flop/Byte. It is understood that the calculation is compute-intensive by Computational Fluid Dynamics standards. Therefore, such high performances as 171.4 GFlops can be achieved in the GPU computing.

**Multiple-GPU Computing**

4-1 GPU Computing on Multi-node

Multiple GPU computing is carried out for the following two purposes: (1) enabling large-scale computing beyond the memory limitation on a single GPU card and (2) speedup the fixed problem pursuing strong scalability. Multiple GPU computing requires GPU-level parallelization constructing a hierarchical parallel computing, since the blocks and the threads in CUDA have already been parallelized inside the GPU. The computational domain is decomposed and a sub-domain is assigned to each GPU.

Using the MPI library for the communication between GPU nodes, we run the same process number as the GPU number. Direct data transfer of GPU-to-GPU is not available and a three-hop communication is required: global memory to host memory, MPI communication, host memory to the global memory. This communication overhead of the multi-GPU computing is relatively much larger than that of CPU. In this article a 1-dimensional domain decomposition is examined for simplicity.

4-2 Overlap between Communication and Computation

In order to improve the performance of the multiple-GPU computing, an overlapping technique between communication and computation is introduced. Since Eqs. (1) and (2) are explicitly time-integrated, only the data of one mesh layer at the sub-domain boundary is transferred. The GPU kernel is divided into two and the first kernel computes the boundary mesh points. At this moment, the data is ready to be transferred and the CUDA memory copy API from the global memory to host memory starts asynchronously as the stream 0. Simultaneously the second kernel that computes the inner mesh points starts as the stream 1. The overlapping of stream 0 with the stream 1 can hide the communication time.

4-3 Performance of Multiple GPU Computing

In the four cases: \(512 \times 512 \times 512\) mesh, \(960 \times 960 \times 960\) mesh, \(1920 \times 1920 \times 1920\) mesh and \(2400 \times 2400 \times 2400\) mesh, their performances are examined with changing the number of GPUs for both the overlapping and the non-overlapping cases. The strong scalabilities are shown in Fig.2.

In every case, the performance of the overlapping computation is greatly improved compared with that of the non-overlapping. The ideal strong scalabilities are achieved up to 8 GPUs for \(512 \times 512 \times 512\) mesh, from 4 to 24 GPUs for \(960 \times 960 \times 960\) mesh, from 30 to 48 GPUs for \(1920 \times 1920 \times 1920\) mesh. We have a perfect weak scalability in the extent of the GPU number used in our runs.

In the overlapping cases, the ideal strong scalabilities are suddenly saturated by increasing the GPU number. For a shorter computational than communication time is not possible to hide the communication time any more.

It should be highlighted that the performance of a 10 TFlops is achieved with 60 GPUs, which is comparable performance of the application running on world top-class supercomputers. We directly compare the GPU performance with the CPU on TSUBAME 1.2 for the same test case of \(960 \times 960 \times 960\) mesh. In the overlapping case, 24 GPUs show the performance of 3.7 TFlops in Fig.3, and it is noticed that 24 GPUs are comparable with 4000 Opteron (2.4 GHz) CPU cores, even if we assume the perfect strong scalability.
The GPU computing for the dendritic solidification process of a pure metal was carried out on the NVIDIA Tesla S1070 GPUs of TSUBAME 1.2 by solving a time-dependent Ginzburg-Landau equation coupling with the thermal conduction equation based on the phase-field model. The GPU code was developed in CUDA and a performance of 171 GFlops was achieved on a single GPU. It is found that the multiple-GPU computing with domain decomposition has a large communication overhead. Both the strong and the weak scalabilities were shown. A performance of 10 TFlops was achieved with 60 GPUs, when the overlapping technique was introduced. The GPU computing greatly contributes to low electric-power consumption and is a promising candidate for the next-generation supercomputing.

Acknowledgements

This research was supported in part by KAKENHI, Grant-in-Aid for Scientific Research (B) 19360043 from The Ministry of Education, Culture, Sports, Science and Technology (MEXT), JST-CREST project “ULP-HPC: Ultra Low-Power, High Performance Computing via Modeling and Optimization of Next Generation HPC Technologies” from Japan Science and Technology Agency (JST) and JSPS Global COE program “Computationism as a Foundation for the Sciences” from Japan Society for the Promotion of Science (JSPS).

References

International Research Collaboration

The high performance of supercomputer TSUBAME has been extended to the international arena. We promote international research collaborations using TSUBAME between researchers of Tokyo Institute of Technology and overseas research institutions as well as research groups worldwide.

Recent research collaborations using TSUBAME

1. Simulation of Tsunamis Generated by Earthquakes using Parallel Computing Technique
3. GPU computing for Computational Fluid Dynamics

Application Guidance

Candidates to initiate research collaborations are expected to conclude MOU (Memorandum of Understanding) with the partner organizations/departments. Committee reviews the “Agreement for Collaboration” for joint research to ensure that the proposed research meet academic qualifications and contributions to international society. Overseas users must observe rules and regulations on using TSUBAME. User fees are paid by Tokyo Tech’s researcher as part of research collaboration. The results of joint research are expected to be released for academic publication.

Inquiry

Please see the following website for more details.