SCIENCE CHINA
Information Sciences

July 2020, Vol. 63 171001:1–171001:14
https://doi.org/10.1007/s11432-020-2861-0

• REVIEW •
From CAS & CAE Members

Research trend of large-scale supercomputers and applications from the TOP500 and Gordon Bell Prize

Weimin ZHENG

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, *China*

Received 5 March 2020/Accepted 24 March 2020/Published online 8 June 2020

Abstract China is playing an increasingly important role in international supercomputing. In high-performance computing domain, there are two famous awards: The TOP500 list for the fastest 500 supercomputers in the world and the Gordon Bell Prize for the best HPC (high-performance computing) applications. China has been awarded in both TOP500 list and Gordon Bell Prize. In this paper, we review the supercomputers in the latest TOP500 list and seven Gordon Bell Prize applications to show the research trend of the large-scale supercomputers and applications. The first trend we observe is that heterogeneous architectures are widely used in the construction of supercomputing systems. The second trend is that artificial intelligence applications are expected to become one of the main stream applications of supercomputing. The third trend is that applying heterogeneous systems to complex scientific simulation applications will be more difficult.

Keywords TOP500, Gordon Bell Prize, large-scale, supercomputer

Citation Zheng W M. Research trend of large-scale supercomputers and applications from the TOP500 and Gordon Bell Prize. Sci China Inf Sci, 2020, 63(7): 171001, https://doi.org/10.1007/s11432-020-2861-0

1 Introduction

In high-performance computing (HPC) domain, two rankings are very important. One is the TOP500 list for supercomputers around the world, and the other one is the Gordon Bell Prize for recognizing outstanding HPC applications. In this review, we are trying to give an overview of the development trend of the high-performance computer industry from the TOP500 and Gordon Bell Prize.

The TOP500 is a well-known ranking list of supercomputers all over the world. Since 1993, the rank list is updated and released twice a year. It creates a solid statistical foundation of research on supercomputing power on earth. In 2017, Sunway TaihuLight, located in Wuxi, China, won the top supercomputer on the TOP500 list. TH-2A, another Chinese supercomputer, is on rank 4.

The Gordon Bell Prize was established in 1987, which is an award for recognizing the outstanding HPC applications. The Gordon Bell Prize is presented at the SC conference (SuperComputing Conference) each year by ACM, and it is usually awarded to the applications of the supercomputers on the top ranks of TOP500. Since 2011, the award winner will also get a cash award of $10000, funded by Gordon Bell, a pioneer in high-performance computing.

In 2016, Chinese researchers, Chao YANG et al. won the Gordon Bell Prize with their study "10M-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics", which is an application running on Sunway TaihuLight. This is the first time that China got this prize since this prize was established

Email: zwm-dcs@tsinghua.edu.cn

30 years ago. One year later, Haohuan FU et al. won the Gordon Bell Prize with their study again with a nonlinear earthquake simulation application running on Sunway TaihuLight.

China is playing an increasingly important role in international supercomputing. We analyze the supercomputers and seven Gordon Bell Prize applications to show the development trend of large-scale supercomputers and applications.

We organize the paper as follows. Section 2 reviews the TOP500 list and provides our insights. Section 3 describes the Gordon Bell Prize and analyzes Gordon Bell Prize applications. Section 4 discusses the trends of supercomputers and applications. Section 5 gives the conclusion.

2 The TOP500 insights

2.1 Introduction of the TOP500

The TOP500 is a well-known rank list of supercomputers all over the world. The rank list is based on the LINPACK benchmark, a widely-used benchmark that measures a system's ability to solve dense linear systems. The list candidates are free to port the code to fit into their own platform and gain the best performance. They can also choose the most suitable problem size that is suitable for the system [1]. The benchmark can effectively measure the float computation power of any system from embedded chips to the world's largest supercomputers.

In the newest TOP500 list revealed in December, 2019, two supercomputers from the United States occupy the top two positions of the list. Sunway TaihuLight, located in Wuxi, China, which was the top supercomputer on the 2017 list, now becomes rank 3. TH-2A, another Chinese supercomputer is on rank 4.

2.2 Plan of CORAL

CORAL is the collaboration between the DOE Office of Science Leadership Computing Facility centers, Oak Ridge Leadership Computing Facility and Argonne Leadership Computing Facility, and the National Nuclear Security Administration Laboratory, Lawrence Livermore National Laboratory. The plan aims at increasing the computing power of the top supercomputing systems by 4 to 6 times against the previous generation supercomputers including the well-known TITAN and Sequoia.

In 2014, as a part of the plan, two systems were built by Oak Ridge and Lawrence both choosing the latest IBM systems with IBM Power9 CPUs, NVIDIA's newest V100 GPUs and Infiniband interconnection provided by Mellanox. The systems are released on June, 2018, with Summit ranking the first and Sierra ranking the second. After the effort for half of a year, they ranked top 2.

2.2.1 *IBM Summit supercomputer*

IBM Power 9 CPU consists of 22 cores with 4 threads per core. It also supports NVLink, the fastest inter-GPU connection. On each dual-socket node of Summit, there are 6 NVIDIA V100 SMX GPUs whose peak float point performance reaches 7 TFLOPs. In total, 4608 compute nodes interconnected by Mellonx Infiniband make the whole system with additional side hardware and software support. The Internet topology is a fat tree, which is classical but of a reasonable budget. It is also more stable in performance during runtime, as MPI computation becomes irrelevant to its physical location.

2.2.2 *IBM Sierra supercomputer*

Similar to Summit, 4320 nodes with 4 NVIDIA V100 GPUs make the main body of the IBM Sierra supercomputer. The theoretical peak performance is 125 PFLOPs, while the actual performance measured by the LINPACK benchmark is 96 PFLOPs, 75.3% efficientive. As a part of the CORAL plan, this supercomputer is in charge of simulating nuclear explosions.

2.3 HPCG benchmark

As LINPACK is decades old, new benchmark closer to real applications has been long demanded. HPCG, a newer benchmark that not only measures float point number computation but also memory access are introduced to the TOP500 list, currently as a non-ranking optional column on the list. It consists of multiple tasks including SpMV (sparse matrix-vector multiplication), vector updates, sparse triangular solving, etc.

Although on the LINPACK benchmark, many systems on the TOP500 list can reach an efficiency of over 70%, surprisingly, they only achieve around 1% efficiency on the HPCG benchmark. The HPCG benchmark introduces more challenging problems faced by supercomputer designers, and is directing the future of supercomputers.

2.4 Summary

In 2019, the world's top supercomputers are fast developing. Based on the recently blooming heterogeneous accelerators, especially GPUs, the new generation of supercomputers are more powerful in the LINPACK benchmark, as well as perform better on the HPCG benchmark. Half of the top 10 supercomputers are based on heterogeneous accelerator architecture.

Generally, the top 10 supercomputers in this version of the TOP500 list are similar to the last TOP500 list. However, from the perspective of the total number of supercomputers on the list, Chinese list entries increase from 219 to 227, enlarging the gap above the United States (118). From the view of total computation power, the United States makes up 37.8%, while the Chinese increases from 29.9% to 31.8%.

Excitingly, the sum of the computation power of the top 100 systems on the list, for the first time, breaks EB level, reaching the point of 1004 PFLOPs. This shows that the global supercomputer power is steadily increasing. Especially, the number of large-scale systems is raising.

3 The Gordon Bell Prize

3.1 Introduction of the Gordon Bell Prize

The Gordon Bell Prize was established in 1987, which is an award for recognizing the outstanding HPC applications. It is usually awarded to the applications of the supercomputers among the top ranks of TOP500. People can submit a technique paper during the SC conference submitting process to apply the Gordon Bell Prize. If their study gets into the finalist, they can present their study in the SC conference and their paper will appear in the conference proceedings.

3.2 Gordon Bell Prize in China

In recent decades, China has achieved great achievements in supercomputing. Tianhe-2 and Sunway TaihuLight achieve the first place of TOP500 supercomputers successively, however, China never got the Gordon Bell Prize before 2016. An important activity of CNCC (CCF China National Computer Congress) in 2014 is to discuss "How far we are from the Gordon Bell Prize in China".

In 2016, at the SC Conference, Chinese researchers won the Gordon Bell Prize with the name "10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics". This is the first time that China got this prize since this prize was established. After one year, another Chinese team won the Gordon Bell Prize about earthquake simulationon Sunway TaihuLight.

3.3 2018 Gordon Bell Prize finalist: ShenTu-processing multi-trillion edge graphs on millions of cores in seconds

3.3.1 *Application introduction*

Graphs are an important abstraction used in many scientific fields. With the magnitude of graph-structured data constantly increasing, effective data analytics requires efficient and scalable graph pro-

cessing systems. Although HPC systems have long been used for scientific computing, people have only recently started to assess their potential for graph processing, a workload with inherent load imbalance, lack of locality, and access irregularity.

The authors proposed ShenTu as the first general-purpose graph processing framework that can efficiently utilize an entire Petascale system to process multi-trillion edge graphs in seconds. It can traverse a record-size 70-trillion-edge graph in seconds. Furthermore, ShenTu enables the processing of a spam detection problem on a 12-trillion edge Internet graph, making it possible to identify trustworthy and spam web pages directly at the fine-grained page level.

ShenTu was built and tested on Sunway TaihuLight, the homegrown supercomputer in China. A grand new application type, graph computing, is introduced to supercomputers, breaking the wall between traditional supercomputers and big data.

3.3.2 *Innovation points*

Graph data is extremely irregular, with the highly-skewed power-law distribution and highly random edge destination. In the contrast, supercomputers are mostly highly optimized for regular tasks like dense algebra or stencil operators. The random accesses graph computing requires are not as efficient as those architecturally optimized computations. Also, massive parallelism brings challenges to graph processing, both at chip scale and at machine scale.

Sunway TaihuLight is equipped with 40960 SW26010 chips, which have 4 MPEs (management processing engines) and 256 CPEs (computation processing engines) each. A 4-times over-subscribed fat-tree network is involved in the interconnect, which places even more challenges.

To address these challenges, ShenTu proposed multiple novel approaches to efficient processing of irregular data on massively parallel architecture.

(1) Hardware specialization. ShenTu specializes each hardware components to the different capabilities of the heterogeneous processors, implementing an innovative spatial pipeline strategy that utilizes heterogeneous many-core chips for streaming graph processing.

(2) Supernode routing. To scale the heterogeneous spatial pipeline to full system, ShenTu proposed a topology-aware message relaying method that seamlessly couples with the hardware-specialized pipeline, excavating the potential of intra-supernode network affinity.

(3) On-chip sorting. Most computation steps in the heterogeneous pipeline involve sorting of messages, which is usually considered not as efficient on massive-core chips. Based on the explicit on-chip network feature of CPE clusters on SW26010, ShenTu designed a kernel template that sort messages on chip without writing intermediate messages into main memory, delivering maximum flexibility and performance for fine-grained message forwarding and processing.

(4) Degree aware messaging. To address the skewness of vertices degree in real-world graphs, ShenTu proposed to directly produce messages for edges corresponded to low-degree vertices and setup mirrors globally for high-degree vertices. This simple method both improves load balance and reduces communication during graph processing.

3.3.3 *Performance*

ShenTu advances the problem scale of graph processing by one order of magnitude and performance by 2–3 orders of magnitude. It enables in-memory processing of synthetic graphs with up to 70 trillion edges, demonstrating performance up to 1984.8 GPEPS on 38656 compute nodes, utilizing 40% of the bisection bandwidth of Sunway TaihuLight interconnect with highly irregular tasks. In addition, ShenTu shows the first large-scale spam-score study with 271.9 billion web pages and 12.3 trillion links between them.

3.4 2018 Gordon Bell Prize finalist: 167-PFLOPs deep learning for electron microscopy: from learning physics to atomic manipulation

3.4.1 *Application introduction*

Scanning transmission electron microscopy (STEM) is a very important equipment for controlling matter. It can not only visualize material structure at an atomic scale, but also be used for atomic-level modification of matter. However, there are tens of thousands of STEM platforms all over the world, generating huge volumes of images. How to deal with these images becomes a main problem. Deep learning is one of the most successful methods for computer vision and image analysis, but defining network topology and hyperparameter become a new problem.

3.4.2 *Innovation points*

The project MENNDL [2] focuses on large scale deep learning training. To solve the mentioned problems, MENNDL uses a genetic algorithm to find an optimal network topology with corresponding hyperparameters which can reduce the wasting computation time on poorer performing networks. To fit the IBM Summit, MENNDL utilizes an asynchronous genetic algorithm to ensure the maximum utilization of computational resources. Once establishing a network topology and an initial set of hyperparameters, MENNDL uses a support vector machine to fine tune hyperparameters. This project is expected to promote the realization of fully automatic training, ultimately accelerate the training phase and improve the training effect.

3.4.3 *Performance*

With 4200 nodes on IBM Summit, MENNDL can reach 152.5 PFLOPs performance. The authors predict that when using 4600 nodes, MENNDL can achieve approximately 167 PFLOPs. Compared with human experts, MENNDL achieved a validation accuracy of 99.51%.

3.5 2018 Gordon Bell Prize finalist: a fast scalable implicit solver for nonlinear time-evolution earthquake city problem on low-ordered unstructured finite elements with artificial intelligence and transprecision computing

3.5.1 *Application introduction*

As more people move to cities, vulnerabilities of cities in terms of natural disasters such as earthquakes have been a hot research topic for a long time. Earthquake city simulation is an effective way to improve the design of buildings and estimate seismic damages. These simulations require the support of powerful supercomputers and on which more than four Gordon Bell Finalist papers focus. In order to get more precise results, this paper utilizes computer-aided engineering CAE and introduces a new simulation condition that couples ground and urban structure.

The major computation-consuming procedure in seismic simulation is solving a large-scale nonlinear equation. MOTHRA [3], an urban seismic problem solver, is proposed in this paper and designed for a situation where both the ground and urban building are targeted. This complex model introduces worse characteristics for computation such as poor convergence. And this hinders current methods from effectively computing the simulation results.

Before MOTHRA is proposed in this paper, the state-of-the-art solver is GAMERA, which enters the SC14 Gordon Bell Finalist, and the standard solver is PCGE. However, neither GAMERA nor PCGE can fulfill computation resources in current supercomputers when facing such a complex problem.

3.5.2 *Innovation points*

MOTHRA is a nonlinear dynamic low-order finite-element solver. It provides two methods, artificial intelligence and transprecision computing, to optimize preconditioners in this problem.

Artificial intelligence optimizing preconditioner. In this problem, the irregularity of graph connectivity results in poor convergence. We can use several local operations to change the worse characteristics in the equation and improve convergence. However, detecting these characteristics is difficult for traditional methods. MOTHRA adopts AI, specifically artificial neural network, to identify these characteristics. The training dataset contains small graphs with similar worse properties as large ones. Thus, the model can be applied to large equations. With the help of AI, the authors can effectively find these regions and operate them to eliminate iterations in the following computation. Meanwhile, the quality of solutions still remains the same.

With the emergence of FP16 and FP8 numbers, mixed-precision computing has become more prevalent. For problems that do not require high accuracy, a short float point can achieve better performance at the cost of tolerable result errors. MOTHRA uses FP32 for preconditioning because it has a relatively low precision requirement. As for the element-by-element (EBE) method, FP16 is used for matrix-vector multiplications.

Low-precision floating numbers accelerate not only computation but also communication. Today, accelerators such as GPU play a more important role. But their bandwidth is relatively slow compared with their powerful computation ability. So, MOTHRA stores FP64 in FP21 or FP16 to reduce data size. In this way, it reduces the data size to about 1/3 to 1/4 compared with FP64.

3.5.3 *Performance*

In the experiments, the AI optimized preconditioning can effectively reduce the number of iterations for convergence. Without the acceleration of AI, 132664 iterations of conjugate gradient (CG) are needed for 25 time steps simulation. However, as for the same problem, it only requires 88 iterations for MOTHRA with AI. And the total number of floating-point operations decreases from 184.7 PFLOP to 33.2 PFLOP, which is 18.0% of the former one.

MOTHRA has a significant performance improvement on the K supercomputer, Piz Daint and Summit. It outperforms PCGE by 18.6 times, 24.9 times and 25.3 times on these three supercomputers respectively. And MOTHRA also has a 3.99× speedup compared with GAMERA.

The transprecision communication also shows benefits. The evaluation shows that transprecision communication attains 1.05× overall speedup on Piz Diant and 1.10× on Summit. And these results prove that this optimization is more effective on a low computation-bandwidth ratio supercomputer, because Piz Diant has a larger computation-bandwidth ratio and less speedup than Summit. This method makes MOTHRA have a better scalability in a large-scale situation.

In summary, MOTHRA gets 19.8% and 14.7% of FP64 peak performance on 4608 nodes on Piz Diant and 4096 nodes on Summit respectively. It has significant improvement compared with the standard solver PCGE and the state-of-the-art solver GAMERA.

3.6 2018 Gordon Bell Prize finalist: simulating the weak death of the neutron in a femtoscale universe with near-exascale computing

3.6.1 *Application introduction*

Quantum chromodynamics (QCD) simulations are traditional HPC tasks. This project [4] focuses on computing the lifetime of a neutron. The current precision of computing this lifetime is low, which is not even enough to discriminate between two different experimental results. However, a better precision requires much heavier computations.

This project ports the problem to Summit and Sierra, and it focuses on optimizing QCD simulations for high-density nodes in these supercomputers. In a so-called high-density node, there are proportionally more GPUs per CPU, and per NIC. For example, there are 6 GPUs in a node of Summit, and 4 GPUs in a node of Sierra.

3.6.2 *Innovation points*

There are three major concerns in this project.

(1) As there are proportionally fewer CPUs to manage GPUs, the CPU overhead of intra-node and inter-node communications should be as little as possible. Direct communications among GPUs make a significant contribution, which maximizes the GPU communication performance, and reduces the contention of CPUs.

(2) The configuration space of communication policies and parameters is large, especially in high-density nodes. Possible configuration options include whether to use GPU DMA engines, whether to use zero-copy reads or writes, whether to use GPU Direct RDMA, and whether to pack communications together, and the optimal decision depends on the complex hardware and software environment. This project adopts an auto-tuning mechanism to maximize the communication performance.

(3) Taking the advantage of the dynamic process management of MPI 3.1, this project adopts a novel task scheduling tool, which makes it possible to run CPU-bound tasks and GPU-bound tasks in a single node, concurrently, which makes it possible to utilize the originally idle CPU time. This tool also makes sure the nodes belonging to the same tasks are close to each other, so as to maximize the communication performance.

3.6.3 *Performance*

As a performance result, this project achieves a 20-PFLOPs peak sustained performance, which is 15% of the peak performance. At low node count, the peak sustained performance is as high as 20%.

Besides detailed techniques, one should notice that these optimizations are all based on existing software frameworks, which results from the long-term efforts on optimizing GPU accelerated architectures. This project also shows the necessity to optimize an application workflow in a whole, instead of optimizing its isolated parts.

3.7 2018 Gordon Bell Prize: exascale deep learning for climate analytics

3.7.1 *Application introduction*

Exascale deep learning training [5] by Lawrence Berkeley National Laboratory and NVIDIA is the first study successfully scaling distributed deep learning training to 27360 GPUs with exascale peak performance. This study aims to help solve one important problem in climate analytics, i.e., recognizing extreme weather patterns. Except for that, the study also pushes the frontier of the solution to be high-quality and pixel-level.

In this study, the team chooses two convolutional neural networks of very different architectures, Tiramisu and DeepLabv3+, and leverages Horovod, a high performance distributed deep learning training framework on top of Tensorflow to perform large-scale data-parallel distributed training.

To achieve as high performance and efficiency as possible, the team carries out a lot of innovative optimizations on both system and algorithm level.

The system-level optimizations focus on IO and communication.

(1) IO. Since training on the whole system will bring huge IO pressure, e.g., training Tiramisu on the full Summit system requires about 5.23 TB/s, which is more than twice the performance of the global file system of Summit, reading from the global file system is unrealistic. The team instead partitions the whole dataset and stages them into local SSD storage separately. Moreover, the team uses an optimized data ingestion pipeline, which overlaps the data loading with GPU computation and utilizes multiple processes to read simultaneously.

(2) Communication. Allreduce in each iteration becomes a bottleneck in large scale training. Considering ring-based allreduce performs better with NVLink while tree-based allreduce has better scalability, the team proposes a hierarchical allreduce scheme, which uses NCCL for ring-based allreduce within one node and MPI for tree-based allreduce across different nodes. Besides, the team also solves a bottleneck

caused by Rank 0, the centralized scheduler in Horovod, by substituting a hierarchical aggregation and broadcast of the control messages for the flatten pattern.

3.7.2 *Innovation points*

The algorithm level innovations mainly include weighted loss, layer-wise adaptive rate control and gradient lag.

(1) Weighted loss. Since extreme weather pattern rarely happens, the labels are highly biased in the dataset, which will lead to terrible accuracy if training directly. For the sake of high accuracy, the team assigns a weight to each class of label in loss function to ensure their contribution is roughly equal.

(2) Layer-wise adaptive rate control (LARC). LARC [6] is an effective approach to solve the divergence problem when training with a large batch size. The team also adopts this approach in this study.

(3) Gradient lag. Allreduce operations are on the critical path of synchronous distributed training, which significantly limits the performance. To fulfill the potential, the teams allow lagging of gradient, i.e., the finish of the allreduce operations in the current iteration can be delayed until the update of the next iteration begins. With gradient lag, the communication is fully overlapped with computation, resulting in doubled performance measured by FLOP/s.

3.7.3 *Performance*

With the innovative optimizations on both system and algorithm level, Tiramisu scales to 5300 P100 GPUs with a sustained performance of 21.0 PFLOPs/s and 79.0% parallel efficiency, and DeepLabv3+ scales up to 27360 V100 GPU with a peak performance of 1.13 EFLOPS/s using FP16 precision. Besides, the optimizations in this study not only address the common problems in large-scale training, but also compose a full set of solutions for training deep neural networks on large-scale heterogeneous supercomputers.

To summarize, this study is a distinguished milestone of applying HPC resource and techniques for AI development. It also shows the potential of HPC in AI areas at a large scale for the first time. The study reveals that scientific big data analysis, especially those for high-resolution and massive scientific simulation data, is an opportunity of the combination of scientific computing and AI, and will also be one of the core applications of supercomputing.

3.8 2018 Gordon Bell Prize: attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction

3.8.1 *Application introduction*

Opioid misuse and addiction [7] are having huge impacts on public health and social and economic welfare. It is important to understand the genetic structure of how individuals develop chronic pain and respond to opioids. However, chronic parin and opioid addiction are highly complex disorders, they are likely manifested due to the combined actions of multiple interacting genetic factors.

Studies have made it clear that risks for many complex human diseases derive from non-additive interactions between multiple genes. It is hard for people to be aware of human disease without non-additive effects if high order, because existing methods are incapable of finding out such interactions from available samples.

The study builds up an application called CoMet, which is used to find out the higher-order combinations of single nucleotide polymorphisms (SNPs). CoMet brought up a network-based approach to build sets of SNPs representing higher-order combinations. In this CoMet, two methods are mainly used to get results: custom correlation coefficient (CCC) and proportional similarity (PS).

Custom correlation coefficient. It is common in science domains to compute for the mathematical relationships between pairs of vectors. In this field, the CCC was brought up to calculate the correlation between mutations across a population of individuals. It can be used to find out groups of SNPs which tend to co-occur in a population, and finally can be used to find combinations of SNPs which associate

with certain phenotypes. Furthermore, in order to capture more complex relationships, the study had previously introduced a new ternary network definition, namely 3-way networks based on the concept of hypergraphs.

Pleiotropy is the phenomenon in which a gene is involved in multiple phenotypes. The discovery of pleiotropy can be reduced to a vector comparison problem. CoMet constructs a matrix with SNPs as row and phenotypes as columns. The number in the matrix represents the rate of association between the SNP and a particular phenotype.

3.8.2 *Innovation points*

The main innovation points are listed below.

(1) 2-way methods recast as modified GEMM operations. To achieve high performance on supercomputer equipped with GPUs, CoMet had cast the CCC and PS operations to modified GEMM. They replace the `c+=min(a,b)` GEMM scalar to `c+=min(a,b)` in PS. For CCC, a more complex modification is needed involving bit-level operations.

(2) Making use of hardware features. For PS methods, the CUDA `fminf` and `fmin` functions quickly take the minimum of two values. CoMet dynamically chooses from the intrinsic according to the problem precision, gaining the best performance for each condition while avoiding critical loss of significance effects for long vectors.

(3) 3-way methods via multiple modified GEMMs. To apply GEMM on 3-way methods, CoMet transfers the problem to a sequence of modified GEMM operations each for a plane of the cube, which explores the high performance of modified GEMM used in the 2-way GEMM case. Moreover, there is no off-node communication when computing the block of results, which further improves the performance.

(4) Removing redundant computations, achieving load balance. The project transfers the problem to distributed modified GEMM. Efficient distributed dense linear algebra often requires multidimensional parallelism. CoMet implemented three parallelism axes: partitioning the set of input vectors, decomposing the vectors along the length and replicating the vectors to distribute computation of result blocks. CoMet is performing an all-to-all vector comparison, so communication plays an important part. CoMet implements asynchronous pipelining to hide the communication under computation. CPU/GPU transferring is well scheduled during GPU computation.

3.8.3 *Performance*

CoMet formulates vector similarity methods as generalized distributed dense linear algebra operations. They can reach up to 98% weak scaling efficiency at full system relative to single GPU kernel performance. They got 189 PFLOPS single precision for PS method and CC method at 2.36 ExaOps, with over 10^{18} mixed-precision floating-point operations per second. The performance improvement is four to five orders of magnitude beyond state of the art.

3.9 2019 Gordon Bell Prize finalist: 46 PFLOPS simulation of a metallic dislocation system

3.9.1 *Application introduction*

First principles calculations based on quantum mechanics [8] have achieved impressive results in predicting the properties of various materials. However, it requires a huge amount of computation due to the high computational complexity. Kohn-Sham density functional theory calculations (DFT) have provided many key insights for materials behavior (mechanical, chemical, electronic and optical properties). The exponential computational complexity (in number of electrons) of solving the many-electron Schrödinger equation (SE) is reduced to cubic computational complexity by his approach to DFT [9].

In order to compute meaningful material properties, DFT has stringent accuracy requirements, which is about $O(10^{-5})$. Owing to the cubic-scaling computational complexity, materials systems are usually limited to a maximum of several thousands of electrons. For this reason, researchers have made many attempts over the past two decades. However, the approaches they proposed to DFT are not very suitable

due to robustness, accuracy and some other reasons. Therefore, the traditional method of solving the Kohn-Sham equation with a plane-wave discretization [10–12] is still the preferred method for metallic systems despite its limitations.

3.9.2 *Innovation points*

DFT-FE employs the finite-element (FE) discretization of the Kohn-Sham equations. There are four major innovations in this study.

(1) A significant degrees of freedom (DoFs) will be introduced by simply implementing FE discretization. To overcome this challenge, an error-analysis informed adaptive higher-order FE discretization has been employed to reduce the DoFs needed to achieve chemical accuracy. DFT-FE also employs spectral finite-elements with nodal points coincident with Gauss-Lobatto-Legendre (GLL) points, which in conjunction with GLL quadrature for numerical integration renders M diagonal and the generalized Hermitian eigenvalue problem is transformed into a relatively simple problem.

(2) All computationally intensive steps in the ChFSI procedure are ported to GPUs and data transfers between CPU and GPU are minimized. For some steps that are not suitable for execution on GPU, like CholGS-CI and RR-D, they are performed on CPUs in parallel using the ELPA library [13].

(3) Instead of global sparse matrix approaches, DFT-FE uses dense matrix operations with a blocked approach to reduce the memory access costs in Chebyshev polynomial filtering procedure. An interesting observation is that the communication pattern of all wavefunction vectors across the FE domain decomposition partition boundaries is identical to MPI point-to-point communication pattern. Based on this observation, DFT-FE performs the MPI communication for all wavefunction vectors simultaneously to minimal network latency.

(4) DFT-FE develops and validates mixed-precision strategies for some steps. After using mixed precision strategies in the ChFSI procedure, about 85% of the total FLOPs count is performed in FP32.

3.9.3 *Performance*

This paper has demonstrated the parallel scaling performance, time-to-solution, and sustained performance of DFT-FE on large-scale metallic systems. Simulations are executed on 3800 GPU nodes of the Summit supercomputer. On a dislocation system in Magnesium containing 105080 electrons, DFT-FE has achieved an unprecedented sustained performance of 46 PFLOPS (27.8% peak FP64 performance) which is 14.9× greater than that of any previous study.

3.10 2019 Gordon Bell Prize: a data-centric approach to extreme-scale Ab initio dissipative quantum transport simulations

DaCe [14], a data-centric approach, is presented to scale ab initio quantum transport simulations to extreme-scale, reaching an extraordinary performance of 85.45 PFLOPs/s (42.55% of the peak) in double precision on 4560 nodes of Summit. This study optimizes Ab initio quantum transport solver by analyzing data dependence.

3.10.1 *Application introduction*

Quantum transport simulation is a classic and non-trivial domain scientific application. Many studies have optimized it and get good results. Most of the existing quantum transmission simulators use simple models that require a small amount of calculation. The accuracy is not high enough in some complex simulations. A state of the art solver, OMEN [15], is picked out twice as a Gordon Bell Prize Finalist. However, the implementation of OMEN is very complicated. For large-scale heterogeneous computing platforms, further and deeper adaptation optimization is required.

3.10.2 *Innovation points*

This study presents DaCe, a data-centric programming framework, which allows domain scientists to program with high-level languages, i.e., Python. The stateful dataflow multigraph (SDFG) is extracted by analyzing the data dependence in the DaCe framework. Then performance engineers transform the SDFG to develop the performance. Finally, the transformed SDFG is translated into codes and binaries.

Ab initio quantum transport simulations are programmed in DaCe framework by domain scientist, and DaCe scales it to extreme-scale by analyzing and transforms the SDFG. There are two major transformations on SDFG: communication avoidance and dataflow optimizations.

(1) Communication avoidance reduces communication volume by transferring broadcast and P2P to all-to-all communications.

(2) Inefficient vector operations are aggregated into dense matrix operations by transforming the dataflow of SSE.

3.10.3 *Performance*

In the experiments, the performance of 85.45 PFLOPs (42.55% of the peak) is achieved on 4560 nodes of Summit in double precision, and 90.89 PFLOPs in mixed precision, which is over 100× faster than OMEN [15] for each atom.

4 Research trend of supercomputers and applications

In this section, we show the important development trend of supercomputing and applications.

Trend 1. Heterogeneous architectures are widely accepted in the construction of supercomputing systems. For the top ten systems on the latest TOP500 list, seven systems use heterogeneous supercomputing architectures, which implies that heterogeneous architecture has become the general trend of building top-level supercomputing systems. However, the argument of whether to use heterogeneous accelerators or heterogeneous many cores will still continue. Among the seven heterogeneous supercomputing systems in the TOP10, five systems are built with NVIDIA GPUs. In contrast to GPU accelerators, Tianhe 2A, a supercomputing heterogeneous system, chooses Matrix-2000 as acceleration computing devices, and NEC's SX-AURORA TSUBASA utilizes vector processing units as accelerators. The future trend of related technologies deserves attention. Heterogeneous many-core architectures that integrate different computing cores in the same chip are also worth noting. Such architecture has been realized in China's Sunway TaihuLight System and has been proven to be effective. Additionally, Aurora, the first Exa-scale computing system supported by the US CORAL program (the Collaboration of Oak Ridge, Argonne and Livermore) will also adopt a similar architecture design, and it is expected to be completed in around 2021.

At present, heterogeneous architectures have become the general trend of building top-level supercomputing systems, and the heterogeneity debate of whether using heterogeneous accelerators or heterogeneous many cores will continue.

Trend 2. Artificial intelligence applications are expected to become one of the mainstream applications of supercomputing. Computing power has been considered as one of the most important foundations for artificial intelligence's popularity. With the expansion of the scale of deep neural networks, the training of the latest network often requires thousands of GPU hours (such as BERT, NASNet, etc.), or even more. Supercomputing systems with top computing capabilities will be used to support large-scale artificial intelligence applications, and expand the technological boundaries. For example, as more and more artificial intelligence technologies are widely used in scientific and engineering computing, GPU accelerators that supporting high-performance tensor computing may also be increasingly preferred by decision makers in supercomputing centers. The Gordon Bell Prize in 2018 is awarded to a large-scale deep learning application, and artificial intelligence-related applications have unprecedentedly occupied half of the nominations, all of which indicate that the combination of artificial intelligence and supercomputing will be combined with increasingly tight.

At present, there are not many artificial intelligence algorithms and applications that can enable high scalability. Taking the most widely used deep learning applications as an example, increasing the batch size to improve data parallelism may lead to convergence problems, thereby limiting the total available amount of parallel resources, and the model parallelism has a communication bottleneck.

Trend 3. Applying heterogeneous systems to complex scientific simulation applications will be more difficult. Only two finalists of the Gordon Bell Prize in 2019 are the in-depth adaptation of complex scientific computing applications on large-scale heterogeneous high-performance computers (named Summit). Researchers use various techniques to improve the scalability of the applications (hybrid accuracy, data tracking, etc.). Moreover, these top machines in the United States are relatively easy to be used to develop machine learning applications, and several nominated applications in 2018 have proven such insight. Compared to machine learning applications, applying the heterogeneous supercomputers to complex scientific simulation applications is more difficult, though the two nominated applications in 2019 are a good progress. We can expect that how to deeply integrate complex scientific simulation applications with machine learning in the future is a worthy research direction.

5 Conclusion

In this review, we discuss the usage of supercomputers, key techniques such as large memory and computing capabilities, and the related investment and supercomputing's development. The Gordon Bell Prize provides a chance for us to think about these topics. It can be seen that in addition to traditional numerical simulation-based scientific computing applications, these listed applications provide new directions in the aspects of artificial intelligence and big-data processing. How to better integrate artificial intelligence technologies with existing scientific applications, with innovative scientific discovery methods, and with scientific computing models, shall create new opportunities for building future supercomputing applications and developing highly scalable parallel systems.

In the past several Gordon Bell Prize, the applications on the Sunway TaihuLight superomputer in China have won a total of six nominations and two awards, indicating that China's high-performance computing researchers have the ability to carry out the most cutting-edge research work in the world. The award itself is not the purpose. Our goal is to promote the widespread usage of high-performance computing applications and systems with technology progress and people's livelihood services. For example, we need to consider how to convert the awarded high-precision weather forecast algorithm into a practical numerical weather forecast application to make real contributions to disaster reduction and prevention. The science and technology departments also should provide a long-term and stable support to related research, promote the interactive leadership of application research and system development, form a positive feedback mechanism, and promote the healthy and sustainable development of China's supercomputing field.

References

1 Zhai J, Chen W, Zheng W. PHANTOM: predicting performance of parallel applications on large-scale parallel machines using a single node. In: Proceedings of the 15th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2010
2 Patton R M, Johnston J T, Young S R, et al. 167-PFLOPs deep learning for electron microscopy: from learning physics to atomic manipulation. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, 2018. 50
3 Ichimura T, Fujita K, Yamaguchi T, et al. A fast scalable implicit solver for nonlinear time-evolution earthquake city problem on low-ordered unstructured finite elements with artificial intelligence and transprecision computing. In: Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, 2018. 627–637
4 Berkowitz E, Clark M A, Gambhir A, et al. Simulating the weak death of the neutron in a femtoscale universe with near-exascale computing. In: Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, 2018. 697–705
5 Kurth T, Treichler S, Romero J, et al. Exascale deep learning for climate analytics. In: Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, 2018. 649–660

6 Ginsburg B, Gitman I, You Y. Large batch training of convolutional networks with layer-wise adaptive rate scaling. In: Proceedings of International Conference on Learning Representations, 2018

7 Joubert W, Weighill D, Kainer D, et al. Attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, 2018. 57

8 Das S, Motamarri P, Gavini V, et al. Fast, scalable and accurate finite-element based ab initio calculations using mixed precision computing: 46 PFLOPs simulation of a metallic dislocation system. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2019. 1–11

9 Kohn W, Sham L J. Self-consistent equations including exchange and correlation effects. Phys Rev, 1965, 140: A1133–A1138

10 Giannozzi P, Andreussi O, Brumme T, et al. Advanced capabilities for materials modelling with quantum ESPRESSO. J Phys-Condens Matter, 2017, 29: 465901

11 Gygi F. Architecture of Qbox: a scalable first-principles molecular dynamics code. IBM J Res Dev, 2008, 52: 137–144

12 Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B, 1996, 54: 11169–11186

13 Marek A, Blum V, Johanni R, et al. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. J Phys-Condens Matter, 2014, 26: 213201

14 Ziogas A N, Ben-Nun T, Fernández G I, et al. A data-centric approach to extreme-scale ab initio dissipative quantum transport simulations. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2019

15 Luisier M, Schenk A, Fichtner W, et al. Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: from boundary conditions to strain calculations. Phys Rev B, 2006, 74: 205323

Profile of Weimin ZHENG

Prof. Zheng obtained his B.S. degree from the Department of Automatic Control, Tsinghua University in 1970, and M.S. degree from the Department of Computer Science and Technology, Tsinghua University in 1982. He was a visiting scholar at the State University of New York in Stony Brook from 1985 to 1986, and at the University of Southampton in UK from 1989 to 1991.

Prof. Zheng has long been engaged in the research of high-performance computer architecture as well as parallel algorithms and systems. He led the establishment and application of the cluster architecture of high-performance computers in China, and participated in the development of the extremely large-scale weather forecast application based on the domestic Sunway TaihuLight, which won the ACM Gordon Bell Prize in 2016. He has served as the director of the 863 High-performance Computer Evaluation Center. His contributions to some scientific problems and engineering techniques such as the scalability, reliability and cost-efficiency of storage systems are highly praised by both domestic and international peers, and the network storage system, disaster-tolerant system, and self-maintenance system developed by his research team are playing important roles in multiple grand projects.

Prof. Zheng was elected a member of the Chinese Academy of Sciences in 2019. He is currently a professor and a doctoral supervisor at the Department of Computer Science and Technology, Tsinghua University. Among his many awards and honors, he was the president of China Computer Federation, and he received the Beijing Excellent Teacher Award, and the title of Beijing Famous Teacher, Special Allowance of the State Council, the State Science and Technology Progress Award (one 1st and two 2nd prizes), the State Technological Invention Award (2nd prize), He Liang He Li Science and Technology Progress Award, and the first China Storage Lifetime Achievement Award. Prof. Zheng and his collaborators published more than 500 papers and more than 10 books. The course of Computer Architecture given by Prof. Zheng was selected as a quality course in Tsinghua University, and was selected as a national quality course in 2008. He is now the editor in chief of the Journal *Big Data*.