Xiaobin HE, Xin CHEN, Heng GUO, Xin LIU, Dexun CHEN, Yuling YANG, Jie GAO, Yunlong FENG, Longde CHEN, Xiaona DIAO, Zuoning CHEN, 2023. Scalability and efficiency challenges for the exascale supercomputing system: practice of a parallel supporting environment on the Sunway exascale prototype system. *Frontiers of Information Technology & Electronic Engineering*, 24(1):41-58.
https://doi.org/10.1631/FITEE.2200412

Scalability and efficiency challenges for the exascale supercomputing system: practice of a parallel supporting environment on the Sunway exascale prototype system

Key words: Parallel computing; Sunway; Ultra-large-scale; Supercomputer

Xiaobin HE; Xin CHEN
E-mail: hexiaobin_1984@163.com; ischen.xin@foxmail.com
ORCID: https://orcid.org/0000-0001-6785-1561
https://orcid.org/0000-0002-0562-0319

Motivation

- The growth in the scale of the exascale system will release huge computing power. Taking into account the scalability and efficiency of parallel applications under such a huge parallel scale raises great challenges to the design of the exascale system.

- In addition, with the increasing demand for computing power in artificial intelligence (AI) applications, the fusion of AI and high-performance computing (HPC) applications has become an important breakthrough direction for HPC applications. It is also a challenge to support the high scalability and efficiency of AI applications in the exascale era of HPC.

1) System and framework

- The Sunway exascale prototype system (SEPS) is a small-scale verification system designed according to the exascale application requirements, and can be scaled up to the exascale level.
- The parallel supporting environment is composed of the parallel operating system, distributed storage system, debugging and tuning system, scientific computing parallel framework, and AI ecosystem.

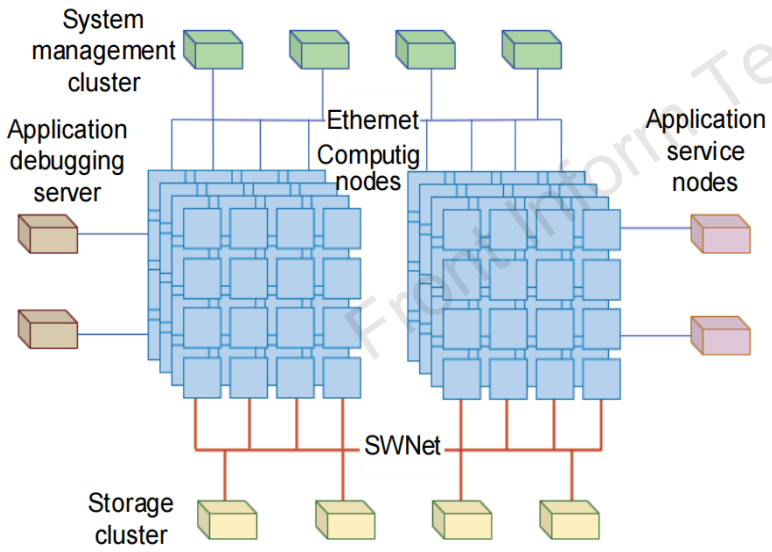


Fig. 1 Architecture of the Sunway exascale prototype system

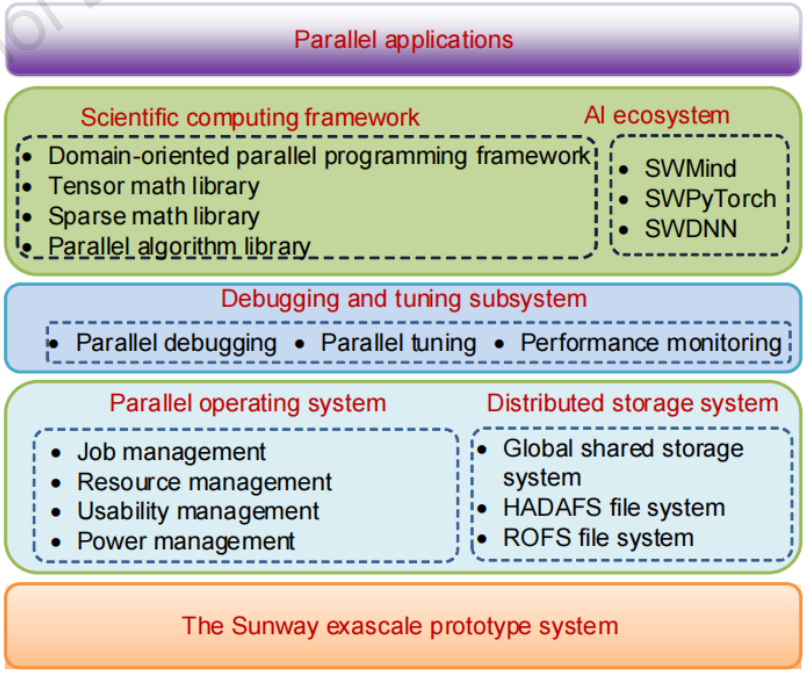


Fig. 2 Components of the parallel supporting environment

2) Design of the parallel supporting environment

- The resource management for the whole system on SPES enables users to execute their jobs concurrently and achieve the unified management, monitoring, and on-demand allocation for many-core computing nodes.
- SEPS carries a hybrid storage system, consisting of a global file system (GFS), a burst buffer file system (HADAFS), and a read-only file system (ROFS).

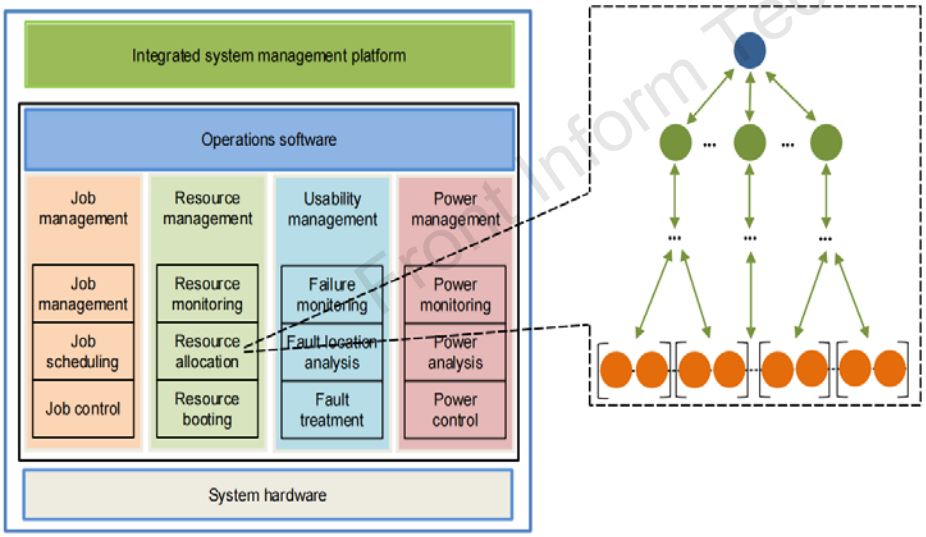


Fig. 3 Parallel operating system structure

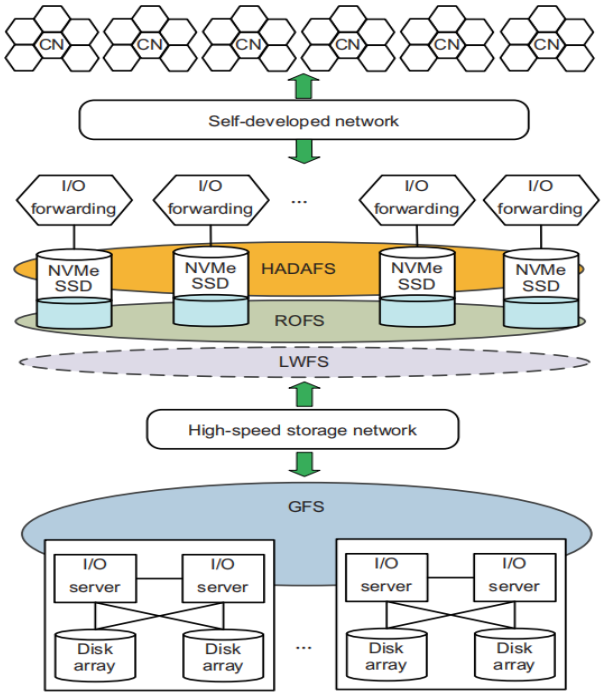


Fig. 4 Sunway storage architecture

2) Design of the parallel supporting environment

- The debugging and tuning subsystem provides developers with tools to develop a large-scale application in less time. Debugger, parallel debugging tool, and large-scale lightweight debugging tool make up the debugging part. They support single-node applications, medium-scale applications, and large-scale applications, respectively.
- The framework is composed mainly of two parts. One part is a domain-oriented parallel programming framework. The other part is the kernel-level parallel programming framework, including the tensor math library, sparse math library, and parallel algorithm library.

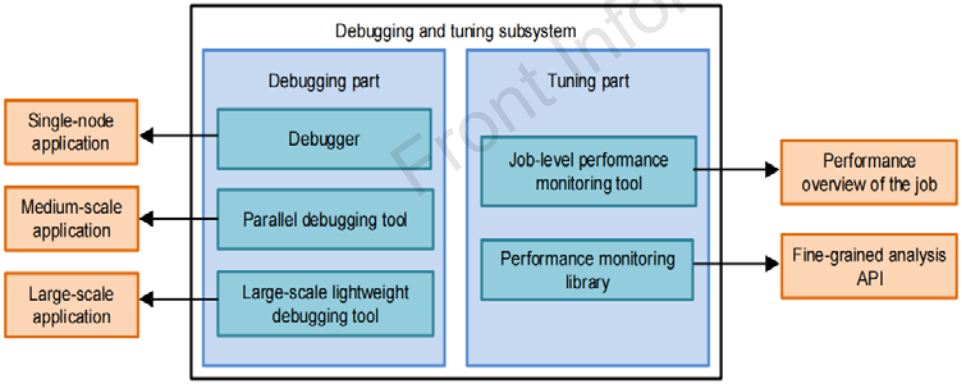


Fig. 5 Debugging and tuning subsystem

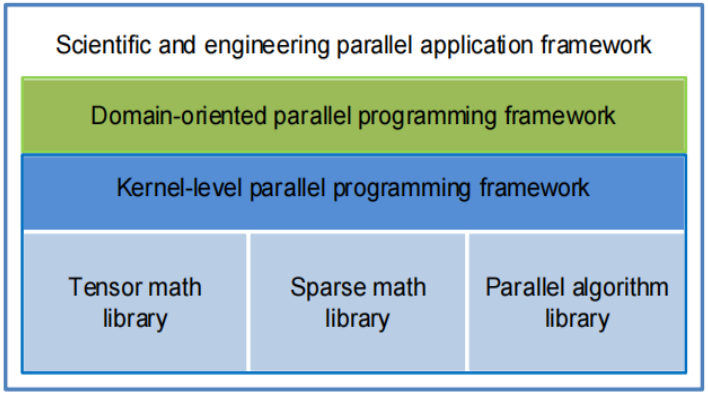


Fig. 6 Scientific computing parallel framework

3) Optimization technologies (I/O optimization)

- LIBIO calls the functions of the LWFS client port in a library manner, including an application request intercepting component and a standard LWFS client port component.
- When LIBIO is used, the user does not need to modify the code; the user needs to just link the LIBIO library in the compilation stage.
- The single-process bandwidth of a computing node is increased by more than twice, and the aggregate bandwidth of a single computing node is increased by more than five times.

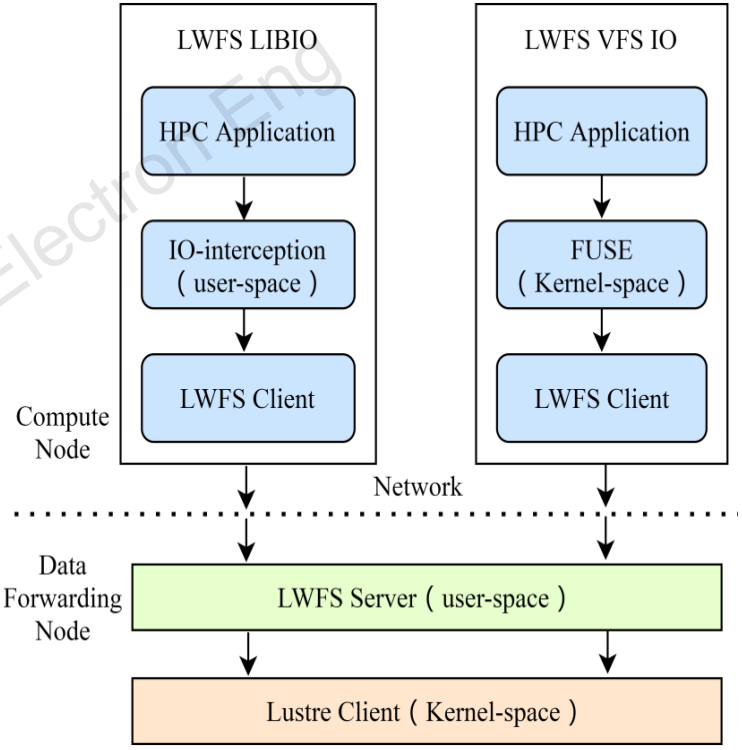


Fig. 7 Software architecture of LWFS LIBIO

3) Optimization technologies (multi-level parallelization mode)

- At the first level, the computing tasks in the specific application are divided into multiple independent tasks according to physical or geometry characteristics.
- At the second level, i.e., processlevel (MPI) parallelism, the independent task within the subgroup is mapped to each MPI process.
- The third level is thread-level (CPE) parallelism.

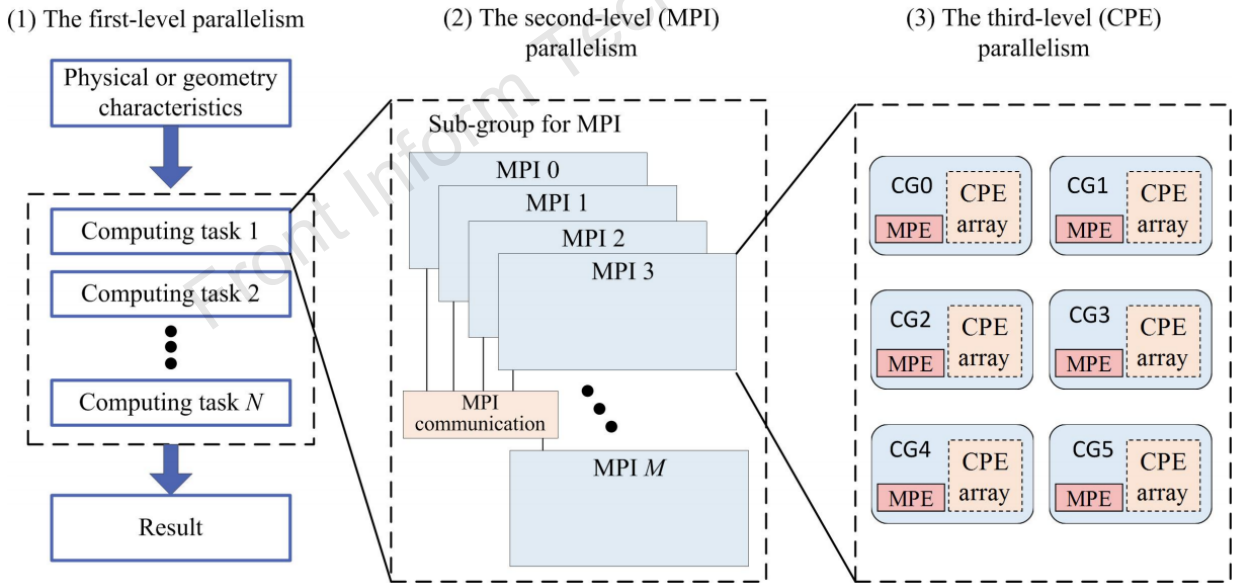


Fig. 8 Multi-level parallelization mode

4) Major results

- These applications show great scalability potential in large-scale systems.
- Among them, the random quantum circuit (RQC) simulation, Raman spectra simulation, and tokamak plasmas simulation have all been shortlisted for the 2021 Gordon Bell Prize due to significant progress in these fields.

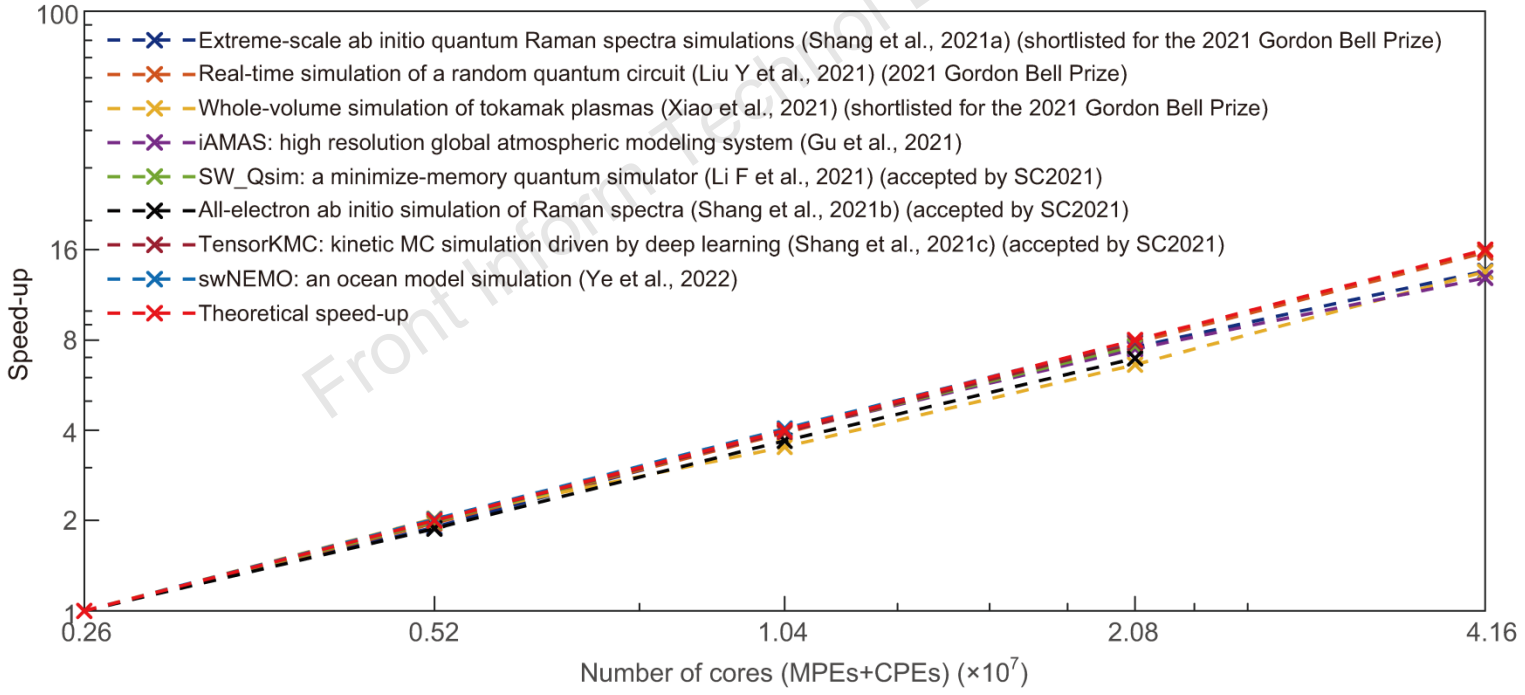
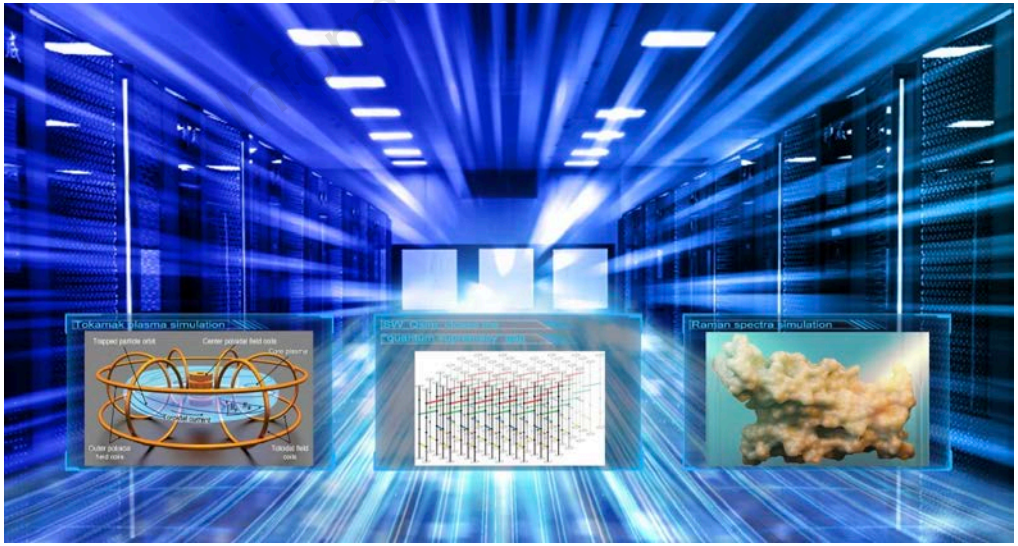
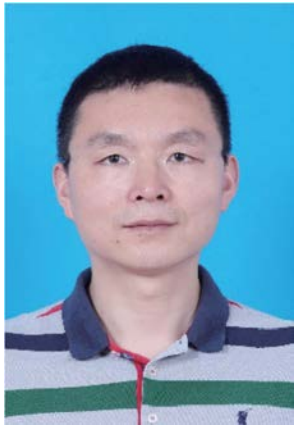


Fig. 9 Expected scalability and efficiency of multiple outstanding applications on SEPS

Conclusions and future outlook

- This paper introduces the design of a parallel application support environment of SEPS and proposes the key technologies of software in guaranteeing the scalability and efficiency of exascale applications.
- Looking to the future, new complex applications reveal novel characteristics, such as computation and data movement varying with time and discreteness and sparseness caused by data non-locality. Given these new features, we will carry out research on novel parallel algorithms and parallel supporting software design and optimization methods.





Xiaobin HE received his BE degree from the Harbin Institute of Technology, Harbin, China, in 2006, and his MS degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently an associate researcher at the National Research Center of Parallel Computer Engineering and Technology, Beijing, China. His main research interests include high-performance computing and distributed storage systems.

Xin CHEN received his BE degree from the National Digital Switching System Engineering & Technological Research Center (NDSC), Zhengzhou, China, in 2016, and his MS degree from NDSC in 2018. He is a research assistant at the National Research Center of Parallel Computer Engineering and Technology, Beijing, China. His research activities focus on high-performance parallel computation and applications.

Xin LIU received her PhD degree from PLA Information Engineering University, Zhengzhou, China, in 2006. She is currently a research fellow at the National Research Center of Parallel Computer Engineering and Technology, Beijing, China. She is a designer of the scientific and engineering application platform of the Sunway TaihuLight System, responsible for the large-scale parallel algorithm research and application software development. Her research interests include parallel algorithms and parallel application software.

Dexun CHEN received his PhD degree from Tsinghua University, Beijing, China, in 2021. He is currently a research fellow at the National Research Center of Parallel Computer Engineering and Technology, Beijing, China. His research interests include high-performance computing and parallel application software.