Particuology 7 (2009) 332–335

Contents lists available at ScienceDirect

Particuology

journal homepage: www.elsevier.com/locate/partic

Brief communication

Multi-scale HPC system for multi-scale discrete simulation—Development and application of a supercomputer with 1 Petaflops peak performance in single precision

Feiguo Chen, Wei Ge*, Li Guo, Xianfeng He, Bo Li, Jinghai Li*, Xipeng Li, Xiaowei Wang, Xiaolong Yuan

State Key Laboratory of Multi-Phase Complex System, Institute of Process Engineering, Chinese Academy of Sciences, P.O. Box 353, Beijing 100190, China

ARTICLE INFO

Article history:
Received 3 June 2009
Accepted 5 June 2009

Keywords:
High-performance computing (HPC)
Graphic processing unit (GPU)
Lattice Boltzmann method
Molecular dynamics
Multi-phase flow
Multi-scale method
Particle method

ABSTRACT

A supercomputer with 1.0 Petaflops peak performance in single precision, designed and established by Institute of Process Engineering, Chinese Academy of Sciences, is introduced in this brief communication. A designing philosophy utilizing the similarity between hardware, software and the problems to be solved is embodied, based on the multi-scale method and discrete simulation approaches developed at Institute of Process Engineering (IPE) and implemented in a graphic processing unit (GPU)-based hybrid computing mode. The preliminary applications of this machine in areas of multi-phase flow, molecular dynamics and so on are reported, demonstrating the supercomputer as a paradigm of green computation in new architecture.

A supercomputer with 1.0 Petaflops peak performance in single precision was formally put into use on April 17, 2009. It was designed and established by the Institute of Process Engineering (IPE), Chinese Academy of Sciences, following the strategy of structural similarity between hardware, software and problems to be solved. Two major HPC (high-performance computing) system producers in China, Lenovo and Dawning, also participated in the establishment. The supercomputer embodies an endeavor to merge a variety of computational problems and algorithms under the general software and hardware framework of multi-scale simulation, so as to improve the efficiency and reduce the cost of HPC systems without sacrificing their applicability in most strategic areas.

The study of multi-scale models at IPE can be dated back to the 1980s (Li, 1987; Li, Tung, & Kwauk, 1988) in establishing the so-called energy-minimization multi-scale (EMMS) model. For verifying the model, fully discrete particle simulation of gas–solid flow has been conducted since the 1990s (Ge & Li, 1996, 2001, 2003). In the process, the common nature of different particle methods on different scales has been observed (Ge & Li, 2000), that is, strong local interactions among the constituent elements with weak long-range correlations, and the pair-wise additive nature of most interactions. This nature was reflected in the framework of the EMMS model, that is, the long-range correlations were simplified to a variational criterion. In extending the EMMS model to a general approach to explain and quantify the interplay between elements on different scales (Li & Kwauk, 1994, 2003) by analyzing the so-called "compromise" among the dominant mechanisms in the systems, the multi-scale strategy in supercomputing was gradually established, featuring the structural similarities between the hardware, the software and the problem to be solved, thus leading to the birth of an algorithm framework applicable to most physical systems (Ge & Li, 2002; Tang, 2005; Wang, 2004), and the computer architectures capable of taking full advantage of the perfect parallelism and scalability of this framework have been patented.

However, implementation of this strategy was deferred by the lack of suitable hardware components. The release of CUDA 1.0 from Nvidia in June 2007 (NVIDIA, 2007) gave IPE an opportunity to practice this strategy with commercial hardware components. The first system was established in February 2008 using 126 HP xw8600 workstations with 200 Nvidia Tesla C870 GPGPU (General Purpose Graphic Processing Unit) cards (20 GTX9800GX2 cards were added later). The workstations were connected by Gigabit Ethernet, through 2D torus meshes of 12 × 10 and 2 × 3 for neighborhood communication on two levels, and a switch for nonlocal communications and management (Ge, Li, & Guo, 2008). This net topology was realized by an elaborate configuration of IPs and partition of the subnets (Guo, He, & Yan, 2008) and Cent OS 5.1 of the Linux family was chosen as the operating system. The system was installed in a limited space of about 50 m^2 and cooled by home-use

* Corresponding authors. Fax: +86 10 62558065.
E-mail addresses: wge@home.ipe.ac.cn (W. Ge), jhli@home.ipe.ac.cn (J. Li).

1674-2001/$ – see front matter © 2009 Chinese Society of Particuology and Institute of Process Engineering, Chinese Academy of Sciences. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.partic.2009.06.002

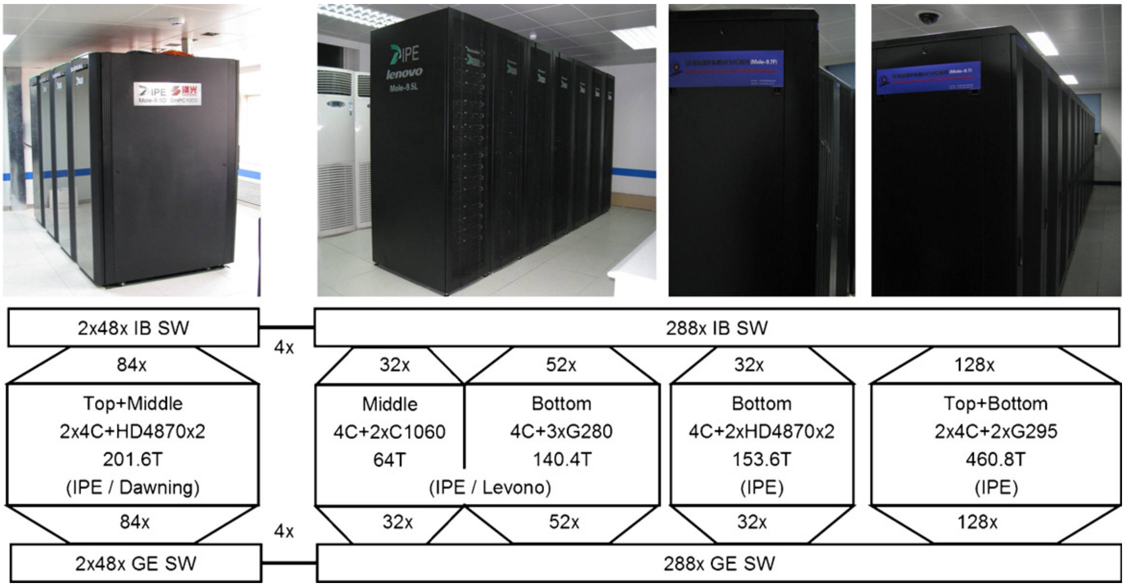

Fig. 1. General architecture of the *Mole-8.7* system.

air-conditioners. The electrical consumption of the system under typical load (including cooling) was about 70 kW. The system may serve as a traditional PC cluster with a peak performance over 10 Teraflops. However, the capability of this system was mainly from its GPGPUs, which delivered over 100 Teraflops peak performance in single precision. For this reason, it was named *Mole-9.7*, which means "$10^{-9.7}$ mol flops". The real performance of this system in typical discrete particle simulations was over 20 Teraflops (Chen, Ge, & Li, 2008), and it has been employed to carry out simulations on different scales, from molecules to industrial reactors (Chen, Ge, & Li, 2009; Chen, Ge, Li, Wang, et al., 2009).

The *Mole-9.7* system was recently upgraded to over 450 Teraflops peak performance using new GPUs of Nvidia GT200 family, and a 150 Teraflops system was newly established by IPE using AMD RV770 GPUs. IPE also developed two 200 Teraflops systems jointly with Lenovo and Dawning, respectively, using both Nvidia and AMD GPUs. As shown in Fig. 1 (Ge, Wang, He, & Li, 2009; Wang, He, et al., 2009), the four systems have now been furnished with 20 Gbps InfiniBand networking and interconnected to construct a multi-scale architecture: The first level employs high-performance CPUs for computation, the second level employs both high-performance CPUs and GPUs, while the third level employs GPUs only, with CPUs just acting as controllers. Three of the four systems consist two levels in themselves, and different combinations of these systems may give optimized configuration for specific computational problems. The whole system is named *Mole-8.7* for its 1.0 Petaflops peak performance. It is installed in rooms of about 120 m^2 and consumes about 300 kW electricity under typical operation, including cooling. A centralized monitoring system based on Ganglia is now in operation, and a more comprehensive management system is under development (Yuan et al., 2009).

The *Mole-8.7* system has run a single program on over 95% of its GPUs, simulating flow in grooved channels using lattice Boltzmann method, which was compiled under different programming environments of Nvidia/CUDA1.1 and AMD/Brook+ (SDK1.4beta) (Li et al., submitted for publication) and then coupled through MPI in runtime. Over 10% peak performance is reached in this run and triple rate can be expected if proper load balancing is introduced, which outperforms the efficiency of corresponding serial program on mainstream CPUs (Li et al., submitted for publication). Over 20% peak performance on 100–400 GPUs was reached in typical discrete particle simulations such as Lenard–Jones molecular dynamics (Xu, 2009) and explicit diffusion-equation solver (Chen et al., 2008).

The *Mole-8.7* system has already been carrying out simulations of high scientific significance and industrial interest. Fig. 2 shows

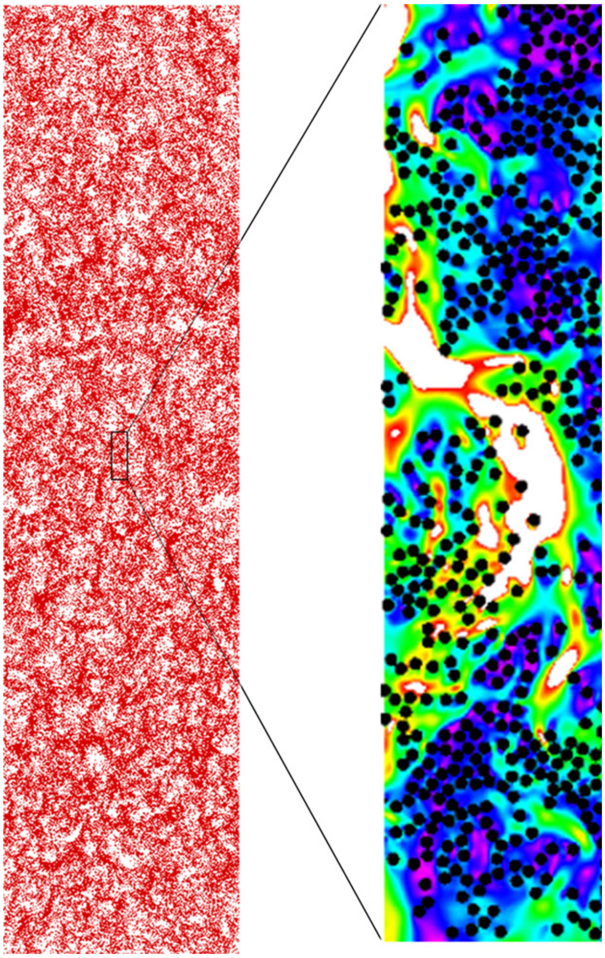

Fig. 2. Instant solid particle distribution and local flow structure from the DNS of gas–solid flow with 0.1 million solid particles.

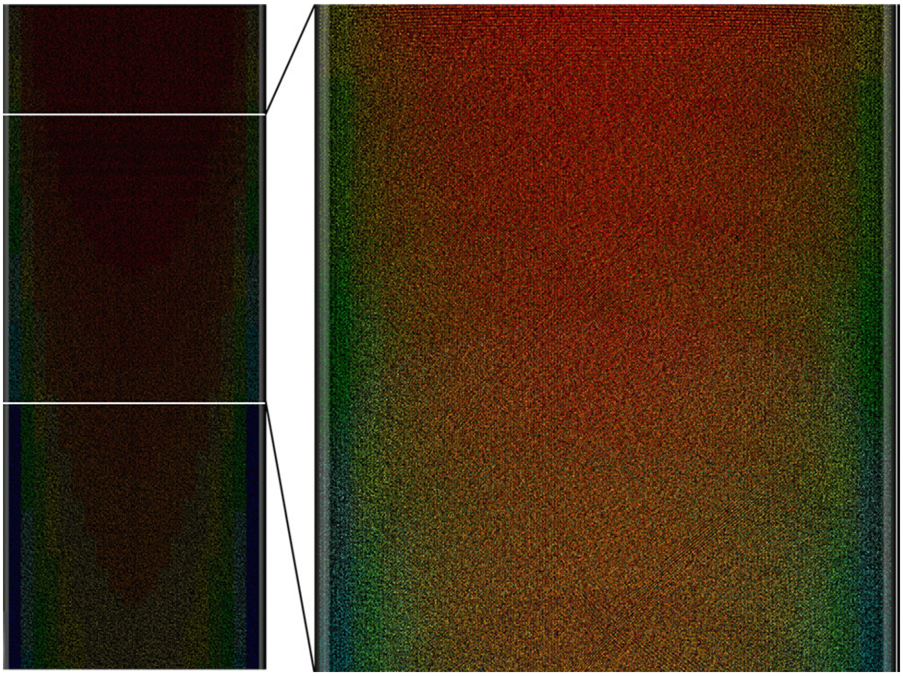

Fig. 3. Snapshot from a multi-scale simulation for a reactor with 60 million particles (the colors present solids concentration, red—low, blue—high). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

snapshots from a direct numerical simulation of 0.1 million solid particles in gas flow, using a macro-scale particle method (Ma, Ge, Wang, Wang, & Li, 2006) and 256 GPUs. The simulation, involving 0.37 billion fluid particles, evolves at a speed of 5 steps/s, nearly 2600 times faster than we have ever had on a single CPU core (Xiong, Li, Chen, Ge, & Li, submitted for publication), reveals the fundamental dynamic behavior in gas–solid systems to a scale and resolution unprecedented in both experimental measurements and traditional simulations, and is expected to provide unique information on the heterogeneity, stability and constitutive correlations of such complex flow.

For industrial applications, *Mole-8.7* is now capable of simulating, in 2D, meter-scale fluidized beds on sub-millimeter particle scales in nearly real-time, that is, on the order of one minute computing for 1 s physical operation. This is achieved not only through progress in hardware capability, but more importantly, through optimization of simulation method (Li & Ge, 2009). The Energy Minimization Multi-Scale (EMMS) model (Li et al., 1988; Li & Kwauk, 1994) is employed on the first and second levels to give the coarse-grid flow structure within a few seconds, thanks to the acceleration of GPUs (Chen, Yang, Ge, & Li, submitted for publication). Discrete particle simulation is then performed on the second and third levels to visualize the sub-grid dynamic flow structure, as shown in the snapshot of Fig. 3 (Xu et al., submitted for publication) from a simulation using 480 GPUs and 60 million solid particles. Though it is a demonstration only and the particle distribution is not realistic yet, it sheds light on the prospect of virtual process engineering, which calls for the designers to specify the configuration and operation conditions of equipments and to examine their real performance promptly.

Mole-8.7 is also powerful in simulating oil/water flow in porous rocks, which is important for oil exploitation and recovery. As shown in Fig. 4 (Wang, Zhang, & Ge, 2009), an area of 0.16 m^2 with fractures below 0.1 mm is simulated directly using lattice Boltzmann method on only 96 GPUs of the system. *Mole-8.7*, therefore, provides a highly competitive alternative to physical experiments which are both expensive and difficult. Another example is from the metallurgical industry, as shown in Fig. 5 (Qi, Chen, Ge, Xu, & Li, submitted for publication): the bedding of feed-stocks in a blast furnace can be simulated on particle level in 3D within a computing time of 5 h using 120 GPUs of the system.

The system is still being expanded, to reach 2 Petaflops peak and 1 Petaflops attainable performance by the end of 2009, for which IPE has developed new building blocks with up to 96 GPU cards on 16 nodes with QDR InfiniBand networking. Meanwhile, applications

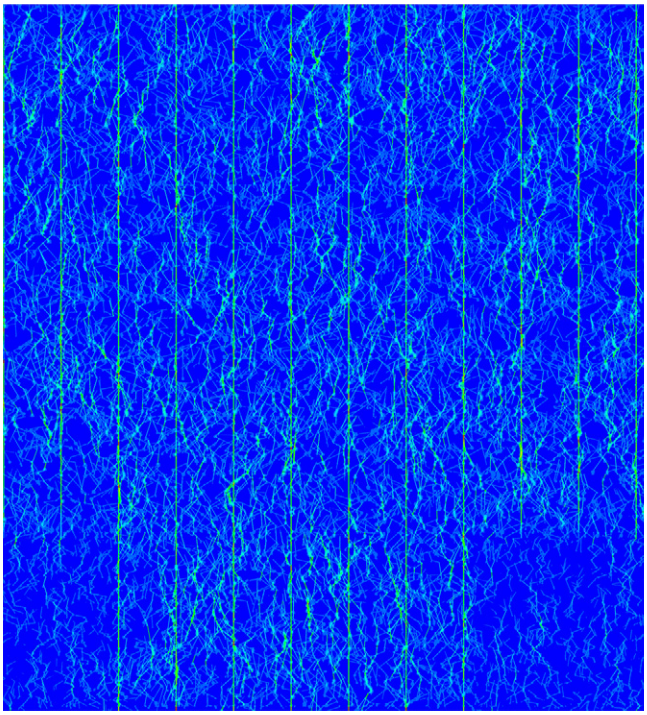

Fig. 4. Direct simulation of flows in porous media using lattice Boltzmann method.

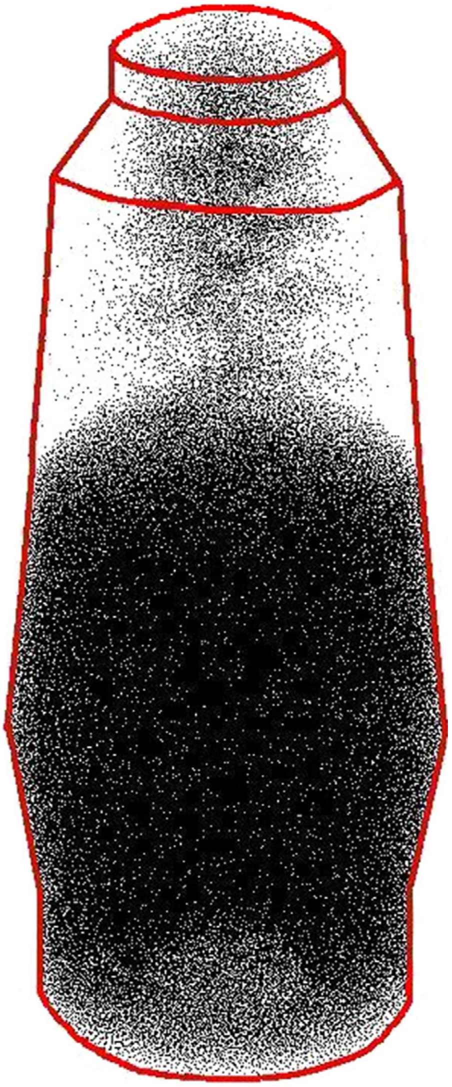

Fig. 5. DEM simulation on the bedding of feed-stocks in a blast furnace.

are being extended to material science (with CPU computing for crystal interface and GPu computing for the bulk of crystal grain), biochemistry, data and image processing (Meng, Chen, Wang, & Li, 2008), and even to social problems, making use of its multi-scale discrete structure in both software and hardware. The application-oriented development mode and the extremely low cost-effect ratio demonstrated by the *Mole-8.7* system, as compared to the current mainstream HPC systems, may set a new paradigm for green computation. Optimization of the computing systems on more fundamental levels to reflect the multi-scale structure and discrete nature of the physical world is also under way.

References

Chen, F., Ge, W., & Li, J. (2008, April). *GPU-accelerated simulation on global virus spreading*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Chen, F., Ge, W., & Li, J. (2009). Molecular dynamics simulation of complex multiphase flow on a computer cluster with GPUs. *Science in China Series B: Chemistry*, *52*(3), 372–380.

Chen, F., Ge, W., Li, X., Wang, X., Xu, J., Hou, C., et al. (2009). *GPU-based parallel computation for multi-scale discrete simulation*. Beijing: Science Press. [in Chinese].

Chen, J., Yang, N., Ge, W., & Li, J. (submitted for publication). *Numerical investigation on the EMMS model implemented on GPUs*.

Ge, W., & Li, J. (1996). Pseudo-particle approach to hydrodynamics of gas/solid two-phase flow. In J. Li, & M. Kwauk (Eds.), *Proceedings of the 5th international conference on circulating fluidized bed* (pp. 260–265). Beijing: Science Press.

Ge, W., & Li, J. (2000). Simulation of discrete systems with local interactions: A conceptual model for massive parallel processing. *Computers and Applied Chemistry*, *17*(5), 385–388 [in Chinese].

Ge, W., & Li, J. (2001). Macro-scale pseudo-particle modeling for particle-fluid systems. *Chinese Science Bulletin*, *46*(18), 1503–1507.

Ge, W., & Li, J. (2002). General approach for discrete simulation of complex systems. *Chinese Science Bulletin*, *47*(14), 1172–1175.

Ge, W., & Li, J. (2003). Macro-scale phenomena reproduced in microscopic systems—pseudo-particle modeling of fluidization. *Chemical Engineering Science*, *58*(8), 1565–1585.

Ge, W., Li, J., & Guo, L. (2008, February). *Logical designing and physical layout of the Mole-9.7 system*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Ge, W., Wang, X., He, X., & Li, J. (2009, March). *Logical designing and physcial layout of the Mode-8.7 system*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Guo, L., He, M., & Yan, L. (2008, February). *Implementation of the Mole-9.7 system software and networking*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Li, J. (1987). *Multi-scale modeling and method of energy-minimization in two-phase flow*. Unpublished doctoral dissertation. Beijing: Institute of Chemical Metallurgy, Chinese Academy of Sciences.

Li, J., & Ge, W. (2009). *Scheme for multi-scale simulation on computers with multi-scale architecture—speed-up of computation with stability conditions*. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences, personal communication.

Li, J., & Kwauk, M. (1994). *Particle-fluid two-phase flow: The energy-minimization multi-scale method*. Beijing: Metallurgical Industry Press.

Li, J., & Kwauk, M. (2003). Exploring complex systems in chemical engineering: The multi-scale methodology. *Chemical Engineering Science*, *58*, 521–535.

Li, B., Li, X., Zhang, Y., Chen, F., Ge, W., & Wang, X., et al. (submitted for publication). *Lattice Boltzmann simulation of grooved channel flows on the Mole-8.7 system with hybrid GPUs*.

Li, J., Tung, Y., & Kwauk, M. (1988). Multi-scale modeling and method of energy minimization in particle-fluid two-phase flow. In P. Basu, & J. F. Large (Eds.), *Circulating Fluidized Bed Technology II* (pp. 89–103). New York: Pergamon Press.

Ma, J., Ge, W., Wang, X., Wang, J., & Li, J. (2006). High-resolution simulation of gas–solid suspension using macro-scale particle methods. *Chemical Engineering Science*, *61*, 7096–7106.

Meng, F., Chen, F., Wang, W., & Li, J. (2008). A CUDA-based parallel image reconstruction method for CTs. *China Patent No. 200810114478.1*.

NVIDIA. (2007). *CUDA 1.0 release*. Retrieved June 2007, from http://developer.nvidia.com/object/cuda_1_0.html.

Qi, H., Chen, F., Ge, W., Xu, M., & Li, J. (submitted for publication). Simulation of granular material using the Discrete element method—a GPU-based implementation.

Tang, D. (2005). *A general method of parallel computation of particle method and its preliminary applications*. Unpublished doctoral dissertation. Beijing: Institute of Process Engineering, Chinese Academy of Sciences.

Wang, X. (2004). *A framework for parallel simulation of particle systems with pair-additive local interactions—toward a general approach*. Unpublished doctoral dissertation. Beijing: Institute of Process Engineering, Chinese Academy of Sciences.

Wang, X., He, X., Li, X., Li, B., Chen, F., & Yuan, X., et al. (2009, March). *Implementation of the Mole-8.7 system software and networking*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Wang, X., Zhang, Y., & Ge, W. (2009, April). *Lattice Boltzmann simulation of the flow in porous media using GPUs*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Xiong, Q., Li, B., Chen, F., Ge, W., & Li, J. (submitted for publication). *Direct numerical simulation of gas–solid flow using macro-scale particle methods on GPU-based HPC system*.

Xu, J. (2009, April). *Test of molecular dynamics simulation on the Mole-8.7 system using Lenard–Jones potential*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.

Xu, M., Liu, Y., Chen, J., Shen, F., Zhang, Y., & Meng, F., et al. (submitted for publication). *Application of multi-scale method in virtual process engineering*.

Yuan, X., Guo, L., He, X., & Wang, X. (2009, March). *System monitoring of large-scale HPC clusters with hybrid GPUs*. IPE internal report. Beijing, China: Institute of Process Engineering, Chinese Academy of Sciences.