Teracluster LSSC-II——Its designing principles and applications in large scale numerical simulations

ZHANG Linbo[1], CHEN Shiyi[2], LI Xinliang[3], CAO Jianwen[4], ZHANG Wensheng[5] & GONG Xingao[6]

1. State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China;
2. Department of Mechanical Engineering, The Johns Hopkins University and Center for Computational Science and Engineering, Peking University, Beijing 100875, China;
3. Institute of Mechanics, Chinese Academy of Sciences, Beijing 100080, China;
4. Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100080, China;
5. Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China;
6. Surface Physics Laboratory (National Key Laboratory), Fudan University, Shanghai 200433; Institute of Solid State Physics, Chinese Academy of Sciences, Hefei 230031, China

Correspondence should be addressed to Zhang Linbo (email: zlb@lsec.cc.ac.cn)
Received October 15, 2003; revised December 23, 2003

Abstract The teracluster LSSC-II installed at the State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences is one of the most powerful PC clusters in China. It has a peek performance of 2Tflops. With a Linpack performance of 1.04Tflops, it is ranked at the 43rd place in the 20th TOP500 List (November 2002), 51st place in the 21st TOP500 List (June 2003), and the 82nd place in the 22nd TOP500 List (November 2003) with a new Linpack performance of 1.3Tflops. In this paper, we present some design principles of this cluster, as well as its applications in some large-scale numerical simulations.

Keywords: high performance computing, PC cluster, large scale numerical simulations.

DOI: 10.1360/04za0005

1 Introduction

The teracluster LSSC–II was the result of a joint effort by the Lenovo Group Ltd, the major state basic research project "Large Scale Scientific Computation" (LSSC), and the State Key Laboratory of Scientific and Engineering Computing (LSEC). The primary goal in building the cluster is to provide research groups of LSSC and LSEC as well as other leading scientific computing research teams in the country a good platform for doing researches on large scale numerical simulations. The main research areas include weather and climate modelling, direct simulation of turbulent flows, oil reservoir simulation, seismic structure imaging and parameter inversion, computational problems in material science, and high performance computing algorithms and software engineering.

With the second phase of the LSSC project being funded in September 2001, the plan to build this cluster was initiated and it was soon approved in October 2001 by the Ministry of Science and Technology of China. In April 2002, after a bidding process, the Lenovo Group Ltd was selected as the hardware supplier and integrator of the cluster. The on-site installation and integration were finished by the end of August 2002 and after a month of configuring, testing and tuning, the cluster was put into service since October 2002.

The rest of the paper is organized as follows: in section 2, an overview of the cluster is given, including its hardware components and system tools, in section 3, some basic results on the performance of the cluster are discussed, then in sections 4—7, results of some numerical simulations using the cluster are presented, and finally in section 8 are some concluding remarks.

2 System overview

2.1 Hardware

The hardware part of LSSC-II consists of 1 master node (the *console node*), 1 monitoring node, 4 login nodes, 2 I/O nodes, and 256 compute nodes. All nodes are connected with either Gigabit Ethernet or fast Ethernet, and the compute nodes are further connected to a Myrinet-2000 switches system. The detailed hardware data of the system are as follows: Console node: Dual 700 MHz Xeon CPU, 1 MB L2 cache, 4 GB RAM, dual Gigabit Ethernet adapters; I/O nodes: Dual 700 MHz Xeon CPU, 1 MB L2 cache, 4 GB RAM, Gigabit Ethernet adapter, Lenovo SureFiber200R disk array; Compute nodes: Dual 2.0 GHz Xeon CPU, 512 KB L2 cache, 1 GB DDR RAM, 18 GB SCSI disk, fast Ethernet adapter, Myrinet M3F-PCI64B host interface; Login nodes: Same as the compute nodes, except that they are not equipped with a Myrinet host interface; Myrinet switches: 6 M3-E128 switch enclosures which form a full-bisection bandwidth 256-port Clos network. A diagram of the system is shown in fig. 1.

2.2 System layout and tools

The operating system on LSSC-II is based on RedHat Linux 7.2 and the Linux kernel that is currently used is the version 2.4.21.

All system-wide services, including NIS, NTP, LVS, OpenPBS, etc., are running on the master node which also serves as the gateway to the Internet. External users connect to the master node using ssh. They are automatically redirected to one of the login nodes via LVS, the Linux Virtual Server software. Outgoing connections also go through the master node and are masqueraded using NAT, the Network Address Translation. Ordinary users are only allowed to access the login nodes, the compute nodes are allocated through OpenPBS, and a user can only access a compute node when he has a job running on that node, it permits complete control of the computing resources in the cluster which is important for a multi-user, multi-job execution environment.

The system tools running on LSSC-II are in fact those that have been developed in

LSEC since 1998 and are still running in other older machines, including LSSC-I[1]. They were successfully ported to and adapted for LSSC-II in a short time, providing an identical computing environment for our users.

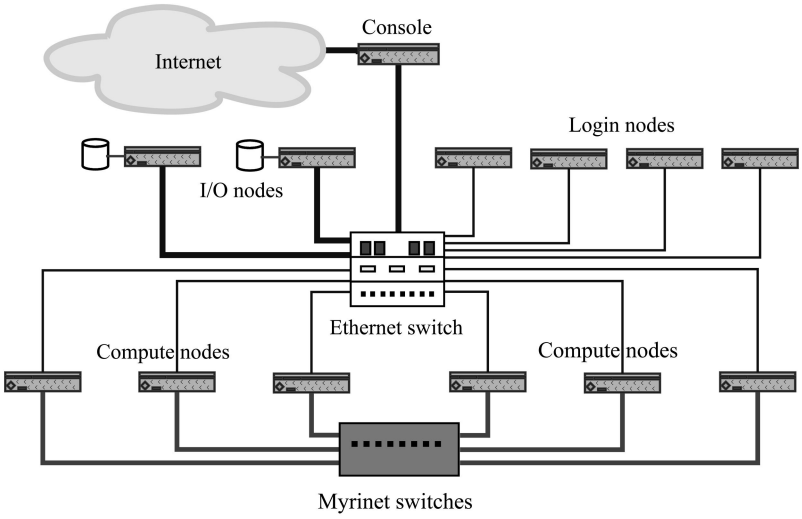

Fig. 1. Diagram of LSSC-II.

To effectively manage such a large cluster and run parallel jobs, some new tools have also been designed. For example, the bash script `rspawn`[2] allows to quickly and concurrently start remote processes on a large number of nodes, and it is extensively used in many system management tools as well as in the startup of parallel programs.

2.3 Compilers and parallel programming environments

The compilers on LSSC-II are GNU C/C++/Fortran, the Portland Group Inc.'s C/C++/Fortran compilers (version 3.1), and the Intel Fortran compiler. The base communication library is GM-1.6.4 and the main parallel programming environment is MPICH-GM 1.2.5..10, both supplied by Myricom. The MPICH-GM on LSSC-II has a good communication performance: the latency for short messages is about 10 μs and the bandwidth for large messages is about 240 MB/s, it allows parallel programs with intensive communications to achieve good parallel efficiency. Other message passing libraries, such as PVM over GM, MPICH over TCP/IP, etc., are also available on LSSC-II.

2.4 Job control

OpenPBS 2.3.16 is used on LSSC-II for job queuing and scheduling. A set of scripts are provided for job submission and execution of parallel programs within a job. Restrictions on access to different resources for different users are achieved through a set of properly designed queues.

Computational jobs are submitted using an LSF[3]-like command `bsub`, which selects appropriate resources requirements and creates an OpenPBS job script according to a set of

command-line options similar to the original LSF command. Another LSF-like command, `bjobs`, is used to display status of jobs. The reason for providing LSF-like commands is that we have been running LSF on the very first PC cluster in the lab, the users do not have to learn new commands in order to use LSSC-II ——we will continue this policy to provide an identical or at least similar job submission environment to our users when installing new computers.

Interactive jobs are also submitted and executed through OpenPBS. Various techniques are used to give users more interactive working environment in the process of running a program, and to solve problems in input/output redirections.

2.5 Testing and tuning

After hardware installation and integration by Lenovo Group, we have run a series of tests and benchmarks to ensure correct functioning and good performance of all system components. This step turns out to be important because through the process we have found some hardware defects, improper BIOS settings, and wiring problems, and has improved the system's overall performance and stability. For example, with exactly the same code, data, and parameters, the execution time of the oil reservoir simulation for the Daqing oil field (BLK model, 1.16 M grid points, 291 wells, 31.5 years, simulation performed on 32 compute nodes, 64 processors) has been reduced from 4.13 to 1.91 h after system tuning (the simulation time given by fig. 4 in section 5 is different because it uses a further improved nonlinear solver).

3 Performance

The best Linpack performance measured when the cluster was built was 1.046 Tflops, using GM-1.5.1, MPICH-GM 1.2.1..7b, and Intel Math Kernel Library 5.2. This performance has been used for the 20th and 21st TOP500 lists. In September 2003, a new Linpack performance of 1.297 Tflops was obtained by using GM-1.6.4, MPICH-GM 1.2.5..10, and Kazushige Goto's BLAS[4]. The improvement in the Linpack performance mainly comes from the new BLAS library and it allows LSSC-II to occupy the 82nd place in the 22nd TOP500 list[5].

The second benchmark result that we present here is obtained with the mixing layer turbulence simulation code (the MX code, see section 4.2). Table 1 and fig. 2 show the timing results on LSSC-II of the MX code on a $110 \times 220 \times 110$ grid. In fig. 2, the "+" line corresponds to the case in which only one CPU on each node is used, the "$\times$" line corresponds to the case in which both CPUs of each node are used, and the "*" line gives the results obtained on an SGI Origin 3800 (600 MHz MIPS R14000 CPU) for comparison purpose. For this relatively small grid size, we get excellent parallel speedup for up to 64 nodes, thanks to the low communication latency of Myrinet on LSSC-II. With the same number of CPUs, the performance of LSSC-II is slightly higher (resp. lower) than the SGI Origin 3800 when using one CPU per node (resp. two CPUs per node), and their

parallel efficiencies are roughly the same. In table 1 and fig. 2 we can observe superlinear speedup on LSSC-II if the number of nodes is smaller than 16 when running one process per node, it is due to better cache hit rates because the memory size of the program on each node becomes smaller with increasing number of nodes. In table 1 we also see another interesting phenomenon: running two processes per node (i.e., using both processors of each node) only reduces the execution time by about 35% with respect to running one process per node if the number of nodes is smaller than 8, and only by 21% if the number of nodes is greater than or equal to 8. We think that the reason for this loss in efficiency is that when using both processors of a node, they compete for both memory access (memory contention) and for the communication bandwidth. The memory contention effect can be more clearly seen in table 2 in which the performance of a Fortran Linpack code is shown. The second row in table 2 is the performance obtained by running one instance of the code on one processor while keeping another processor idle and the third row is the average performance of running two instances of the code at the same time, one on each processor.

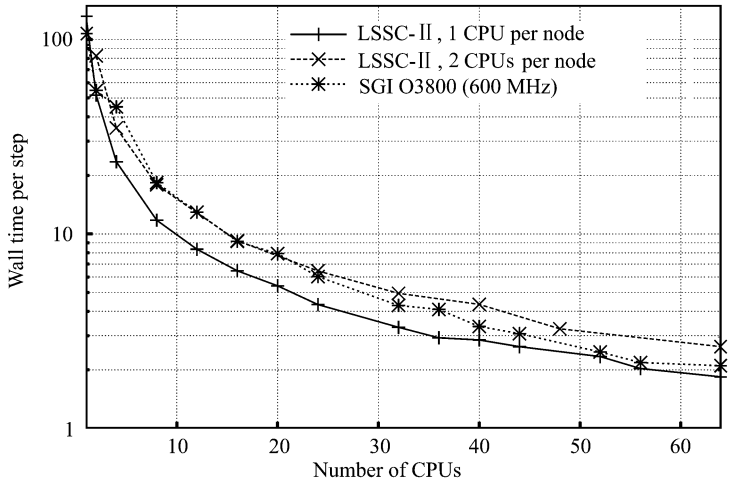

Fig. 2. Timing results of the MX code.

Table 1 Wall time (seconds) per time step of the MX code

No. of nodes	1	2	4	8	16	32	64
1 process/node	132	51.8	23.5	11.8	6.46	3.31	1.84
2 processes/node	82.3	35.1	17.9	9.23	4.95	2.63	1.58

Table 2 Performance in Mflops of the Fortran Linpack code

Order of the matrix	100	400	1000	2000	3000	4000	5000	6000
Running one instance	553	210	191	194	198	188	198	200
Running two instances	553	116	93	93	97	91	97	98

4 Direct numerical simulation of turbulence

Direct numerical simulation (DNS) becomes an important tool in recent research of

turbulence. In this section, we present some interesting DNS results for both incompressible and compressible turbulence obtained on LSSC-II.

4.1 Direct simulation of incompressible turbulence

Several direct numerical simulations (DNS) of incompressible turbulence by solving the two- or three-dimensional incompressible Navier-Stokes equations have been performed on LSSC-II, including isotropic turbulence in a box under periodic boundary conditions, three-dimensional homogeneous turbulence under rapid rigid rotation, and two-dimensional turbulence (see refs. [6—10] for details of these simulations).

In these simulations, Navier-Stokes equations are solved using a spectral scheme. In order to perform fast Fourier transform efficiently, data are repeatedly tranposed on the processors grid, leading to intensive communications.

Some timing results of the isotropic turbulence simulations are given in fig. 3(a) and (b). The "new scheme" in fig. 3(b) refers to a new algorithm with reduced memory usage (it only requires 6 arrays instead of 12.5 for three-D spectral codes). For the case of 256^3 grid, we get almost linear speedup when the number of nodes is increased from 4 to 32 using the old scheme, showing good communication performance of LSSC-II (fig. 3(a)). While for larger grids (512^3—2048^3), with both the old and the new schemes, we observe loss of parallel efficiency, due to a really large amount of data in the communications.

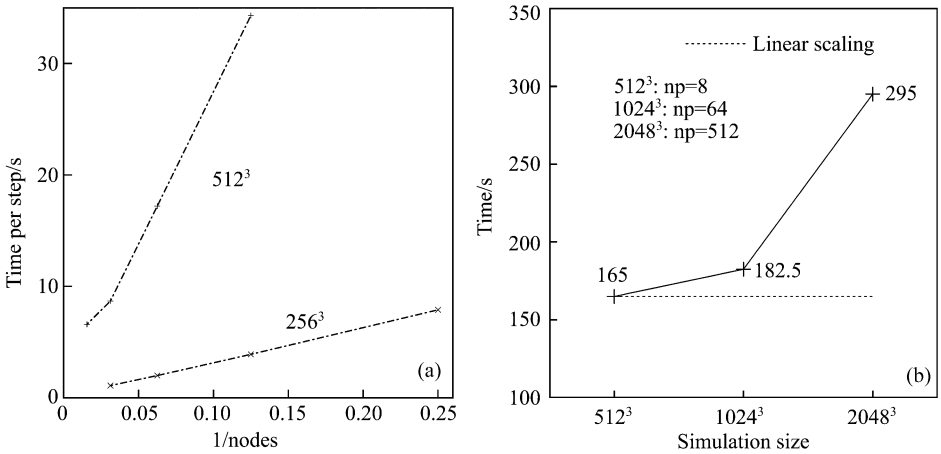

Fig. 3. Isotropic turbulence simulation. (a) Old scheme; (b) new scheme.

4.2 Direct simulation of compressible turbulence

Compact and super-compact finite difference schemes are widely used in the DNS of compressible turbulence because of their good properties in resolving high wave numbers. The compact difference schemes require solution of tri- or penta-diagonal linear systems of equations which can be efficiently implemented on distributed memory parallel computers using the block-pipelined method[11].

In 1998, using a 6th-order centered and a 5th-order upwind compact difference schemes, we performed the direct numerical simulation of transition and turbulence in compressible mixing layer, on a 32-node PC cluster in LSEC[12]. It was the first result of DNS of compressible turbulence in China and the first success story of PC clusters in DNS of turbulence (while at that time DNS of compressible turbulence was considered by some people impossible in China due to lack of supercomputers).

The computing power of LSSC-II allows larger simulations with finer resolution to be carried out. Here we present some recent simulations and their performance on LSSC-II.

4.2.1 DNS of decaying compressible turbulence of relative high turbulent mach number. DNS of compressible turbulence is more difficult than that of the incompressible turbulence. When the turbulent Mach number is greater than 0.3 the shocklets may appear in the compressible turbulent flow fields. The underlying physics and mechanisms of shocklet formation are still unknown[13]. The turbulent Mach number in DNS cannot be very high with the present existing numerical methods and computer resource. To overcome the difficulty in the DNS of compressible turbulence at high turbulent Mach number, a new difference scheme called GVC8 was developed. We have succeeded in direct numerical simulation of decaying compressible turbulence up to turbulent Mach number 0.95. The statistical quantities thus obtained at lower turbulent Mach number agree well with those from previous authors start with the same initial conditions, but they are limited to simulate at lower turbulent Mach numbers due to so-called start-up problem. Energy spectrum and coherence structure of compressible turbulent flow are analyzed. The scaling law of compressible turbulence is studied. The computed results indicate that the extended self-similarity holds in decaying compressible turbulence despite of occurrence of shocklets, and compressibility has little effects on relative scaling exponents when turbulent Mach number is not very high. Details of this work can be found in ref. [14], and the computing performance of the DNS cases are discussed.

Table 3 shows the DNS cases we studied. The program was coded using MPI and Fortran77. Performance of each case is shown in table 4 in which SMC-cluster is a 16-node PC cluster of the SMC Center in Tsinghua University, the CPU of SMC-cluster is PIII 733 MHz. The case F2 in table 4 has also been performed on 64 CPUs, 32-nodes of LSSC-II, the performance is about 5.7 s/step.

4.2.2 DNS of passive scalar in decaying compressible turbulence. There are many problems related with the scalar flux in turbulence, such as the pollutant density in the air, chemical or biological species concentration and salinity in the ocean, etc. Since the 1990s, much work has been done in studying the passive scalars of turbulent flow. In recent years some results of DNS for passive scalars with relative high Schmidt number ($\mathrm{Sc} > 1$) are reported. All those DNS results of passive scalars are for the incompressible

turbulent flows, but the scalars in compressible turbulent flows is more interesting in aeronautics and astronautics, for example, the mixing of fuel and air in supersonic ramjet is a typical problem of scalars in compressible turbulent flow.

Table 3 The computational conditions of DNS of decaying compressible turbulence

CASE	Re_λ	Mt	Scheme	Mesh size
D1	72	0.5	GVC8	128^3
D2	72	0.8	GVC8	128^3
D3	72	0.9	GVC8	128^3
D4	72	0.95	GVC8	128^3
D5	72	0.5	WENO5	128^3
D6	72	0.9	GVC8	256^3
E1	153	0.6	GVC8	256^3
E2	153	0.8	GVC8	256^3
E3	153	0.6	WENO5	256^3
F1	153	0.3	UD7	256^3
F2	153	0.6	UD7	256^3

Table 4 The computing performance of DNS of decaying compressible turbulence

CASE	CPUs/nodes	Time per step	Computer
D1—D4	8/8	34	SMC-cluster
D5	8/8	86.7	SMC-cluster
D6	32/32	135	LSSC-I
E1—E2	64/32	6.8	LSSC-II
E3	64/32	37	LSSC-II
F1—F2	32/32	105	LSSC-I

In our work, the passive scalars in decaying compressible turbulence are solved using DNS of the 7th-order accuracy upwind difference scheme and 8th-order accuracy group velocity control (GVC8) scheme. The start Reynolds number Re=72, the turbulent Mach numbers Ma=0.2—0.9, and the Schmidt numbers of passive scalars Sc=2—10 are used in the computations. In order to validate the computed results, the numerical experiments are made with different simulation methods but with the same fluid parameters. In order to check if the small structures we are interested in are captured, the same problems are solved with mesh doubling. Numerical experiments show that the results given in this work are reliable. The Batchelor κ^{-1} range in scalar spectrum is found in our simulations. The effect of compressibility on the flow structures is discussed. The results show that the high wave number spectrum decays faster with increasing turbulent Mach number. Details of this work can be seen in ref. [15], and the performance of the DNS cases are discussed in this paper.

The performance of each DNS case is shown in tables 3 and 4. The difference approximation for the passive scalar is the same as that for the Navier-Stokes equations. To save computing resource the Reynolds number in computation is not very high, and the

grid number for the passive scalar is doubled in each direction comparing with the grid number for other flow parameters. 8th-order Langrangian interpolation is used to transport the value of coarse mesh to the fine mesh.

The parameters used in the DNS cases are shown in table 5.

Table 5 The parameters used in the DNS of passive scalar in decaying compressible turbulence[a)]

Case	Re_λ	Mt	Sc	Numerical method	Grid size for fluid	Grid size for passive scalar
D1	72	0.2	5	UD7	$256 \times 256 \times 256$	$512 \times 512 \times 512$
D2	72	0.5	5	UD7	$256 \times 256 \times 256$	$512 \times 512 \times 512$
D3	72	0.7	5	UD7	$256 \times 256 \times 256$	$512 \times 512 \times 512$
D4	72	0.9	5	GVC8	$256 \times 256 \times 256$	$512 \times 512 \times 512$
E1	72	0.5	2	UD7	$256 \times 256 \times 256$	$512 \times 512 \times 512$
E2	72	0.5	10	UD7	$256 \times 256 \times 256$	$512 \times 512 \times 512$
E1T	72	0.5	2	UD7	$128 \times 128 \times 128$	$256 \times 512 \times 512$
D2Ta	72	0.5	5	UD7	$128 \times 128 \times 128$	$256 \times 256 \times 256$
D2Tb	72	0.5	5	GVC8	$256 \times 256 \times 256$	$512 \times 512 \times 512$

a) Re_λ is the initial Reynolds number, Mt is the initial turbulent Mach number, and Sc is the Schmidt number. UD7 denotes 7th-order upwind difference method and GVC8 denotes 8th-order group velocity control scheme.

The program is coded using MPI Fortran 77, and the DNS results are computed on LSSC-II. The averaged performance is shown in table 6.

Table 6 Performance of DNS of passive scalar in decaying compressible turbulence

Case	CPUs/nodes	Seconds/time per step
E1T, D2Ta	8/8	11.0
D1—D3, E1—E2	64/32	18.6
D4, D2Tb	64/32	20.5

5 Oil reservoir simulation

In this section, we present the performance of two large scale industrial test cases of oil reservoir simulation on LSSC-II. The first case is a three-dimensional and three-phase (oil-water-gas) black oil model problem from the Daqing Exploration Development Research Institute in China with a $199 \times 87 \times 67$ grid system (1.16 M grid cells) and 291 wells, it was simulated for 31.5 years. The second case is another black oil model from the seventh oil region of west Gudong in the Chinese Shengli Oilfield, the grid dimensions are $160 \times 320 \times 27$, or 1382400 grid cells and 4147200 unknowns, there are 326 wells in the simulation region, it was simulated for 14 years. The parallel oil reservoir simulator we developed[16] is designed to run on distributed-memory machines using MPI for message-passing. Refer to refs. [17,18] for details of parallel solver based on preconditioned Krylov subspace iterations.

Simulating the first test case takes a total of 126 time steps. Each time step consists of 3.58 inexacted Newton steps in average for solving the nonlinear system of equations resulting from an implicit time discretization scheme. Each Newton step consists of 5.92 inner-outer pattern FGMRES(12) iterations in average. The preconditioning process in each FGMRES iteration needs 7.77 ILU-GMRES(12) iterations in average to solve a small system. And each ILU-GMRES iteration needs 11.44 iterations in average to be used as components of the whole preconditioner.

Simulation of the first test case is done using a much bigger time step (91.25 d per time step in average) than the time step used for the second test case (30.78 d per time step in average). The choice of smaller time steps for the second test case is due to more frequent injection/production operations to the water/oil wells, and larger time step length leads to divergence of Newton-like methods.

The second test case takes a total of 166 time steps for 14 years simulation, each time step needs 4.33 Newton steps in average, each Newton step needs 6.99 FGMRES iterations in average, each preconditioning process needs 4.94 ILU–GMRES iterations in average, and each ILU–GMRES iteration needs 5.18 iterations in average.

Table 7 gives some statistics of the two industrial cases simulated. In fact, the simulation time of the two cases is roughly the same.

Comparing the two sets of iteration numbers, we can see that larger time steps produce more workloads in the solution of linear systems and in the preconditioning. From this we state that the nonlinear equations for the second test case are easier to be solved than for the first test case.

Table 7 Statistics of the two industrial test cases

	The lst test case	The 2nd test case
Number of nonlinear systems	451	718
Number of FGMRES iterations	20744	24831
Number of ILU-GMRES iterations	237383	128590
Elapsed hours on 8 nodes/16 CPUs	5.59	5.34
Elapsed hours on 16 nodes/32 CPUs	2.99	2.91

The first test case, i.e. the Daqing black oil model has been simulated using up to 128 processors on LSSC-II with our parallel simulator. Fig. 4 gives elapsed wall-clock times and relative speedups of this data model. The elapsed time is given in hours, and the relative speedup is computed with respect to the case of 8 processors. Due to limitation by the physical memory of the compute nodes, a minimum of 4 nodes are required to perform this simulation. The relative parallel efficiencies on 16, 32, 64 and 128 processors with respect to 8 processors are 78%, 73%, 75%, and 63%, respectively. The communication costs are 0.24 (8 CPU), 0.71 (16 CPU), 0.45 (32 CPU), 0.41 (64 CPU), and 0.45 (128 CPU) hour, respectively. The communication time is relatively small compared with the

computation time. The parallel efficiencies are quite satisfactory, considering the communication complexity of the parallel nonlinear solver. The communication:computation ratio is almost 1:1 in the case of 128 processors, indicating that 8 to 128 processors are suitable for one million-grid cell problems of black oil model on this kind of machines, which allow to accomplish the simulation within a working day.

Fig. 5 gives timing results of the first test case running one process per node (One NodeOneProc) and two processes per node (OneNodeTwoProc) respectively. In fig. 5 y-axis is elapsed wall-clock time (hours) of various computing stages. We see similar effects of memory contention as in tables 1 and 2, i.e. loss in parallel efficiency when using both processors of the compute nodes.

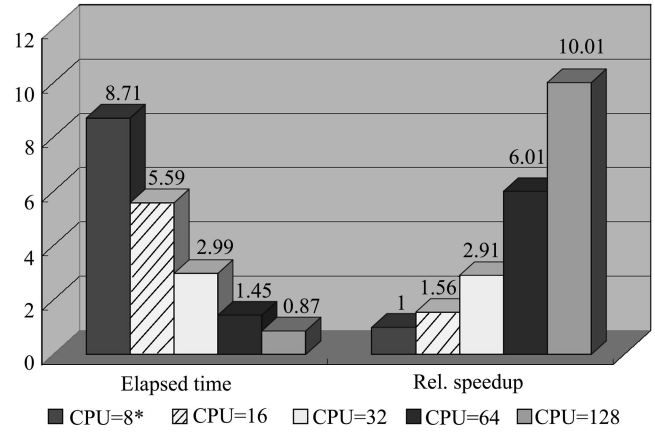

Fig. 4. Computation time and speedup.

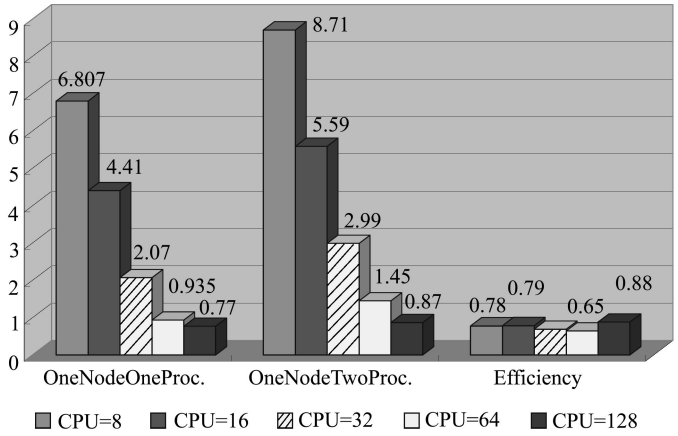

Fig. 5. One CPU/node v.s. two CPUs/node.

6 Parallel performance analysis for 2D and 3D prestack depth migration

Large scale computation in 3D prestack migration has important significance in the

oil industry. In this section, the prestack depth migrations for 2D Marmousi model and 3D SEG/EAEG model are implemented and tested on the PC clusters LSSC-I and LSSC-II[19,20]. The parallel performance of these simulations, especially parallel speedup and parallel efficiency, is presented. Both models are international standard models, and the numerical computations have important practical values.

The Marmousi model has maximum physical dimension of $x = 9200m$ and $z = 3000m$. For numerical experiments considered here, the scope of x ranges from $3000m$ to $9000m$. The model has 369 and 751 grid points in the lateral direction (x) and vertical direction (z) respectively. The data consist of 240 common-shot records with 96 receiver traces each and 725 samples/trace. Therefore, the total data amount is about 64 MB. The number of grid cells in migration reaches the order of one million. The migration method is shot-profile migration by Fourier finite-difference method, and the frequency number in computation is 512. Table 8 is the speedup and parallel efficiency on LSSC-I using different number of nodes. Tables 9 and 10 are the results obtained on LSSC-II. From these tables we see that the shot-profile migration has near linear speedup. However, the parallel efficiency drops a little when using both processors of the computational nodes. For example, with 24 nodes, the parallel efficiency is 97.50% when using one processor on each node, whereas it is 94.42% when using both processors on each node.

Table 8 Speedup and parallel efficiency on LSSC-I

Nodes	2	4	8	16	24	48
Time/min	406.86	203.72	101.95	51.06	34.05	17.15
Speedup	2.00	3.99	7.97	15.92	23.87	47.39
Efficiency(%)	99.89	99.74	99.66	99.49	99.46	98.74

Table 9 Speedup and parallel efficiency on LSSC-II using one processor of each node

Nodes	2	4	8	16	24
Time/min	143.68	72.14	36.31	17.87	12.29
Speedup	2.00	3.98	7.92	16.09	23.40
Efficiency(%)	100.08	99.66	99.00	100.58	97.50

Table 10 Speedup and parallel efficiency on LSSC-II using two processors of each node

Nodes	2	4	8	16	24
Time/min	114.03	72.09	36.05	18.75	12.69
Speedup	2.52	3.99	7.98	15.34	22.66
Efficiency(%)	126.10	99.73	99.72	95.86	94.42

The 3D SEG/EAEG model is a complex benchmark model for testing the ability of 3D migration or inversion. The data here have 50 shot lines with $160m$ line space, and each line has 96 shots with $80m$ shot space. Each shot has 68×6 receivers with $4992ms$ recording length and $8ms$ time sampling. The grid points of velocity model is 10^7. The data amount is 6.23 GB. And the grid number in migration computation reaches the order

of 10 to 100 millions. The migration has been performed on LSSC-I using 50 processors and on LSSC-II using 50 processors on 25 nodes. For this computation, the performance per processor on LSSC-II is about 1.42 times faster than that on LSSC-I. In the computation, the shot-profile Fourier finite-difference method is used.

Both 2D and 3D prestack migrations have perfect parallel performance. The parallel efficiency is almost linear because there are only very few communications in the algorithms.

7 Study on the structural and physical properties of nano-materials

Research of materials on the nano-scale has been an interesting subject for the physicists, chemists and materials scientists. Characterization and manipulation of materials at nano-scale is usually difficult, while the computational analysis has offered a unique way to study the physical and chemical properties of nano-materials with a fast development of computer technique, since the nano-materials can be modelled and simulated by using the modern computational method. However, it needs supercomputers, the conventional desktop computers is too slow.

Large scale PC clusters, such as LSSC-I and LSSC-II, have significantly narrowed the gap of the computer power between China and the developed countries, and also have made it possible to perform world-class research on the nano-system in China. By using such supercomputers, we have set up a few projects to study the structural and physical properties of nano-materials, some of the results have been published in more than ten research papers[21].

The VASP code (Vienna *Ab-initio* Simulation Package) has been implemented on LSSC-II and we have found that its scalability is overall better than its implementation on other commercial supercomputers such as a number of IBM, HP and SGI systems for up to 32 CPUs (see fig. 6).

We now describe a couple of computations performed on LSSC-II. The first example is a new constant pressure MD method for the nano system[22], where the interaction potentials are calculated based on the density functional potentials (VASP). This method is first applied to the single carbon nanotube under external pressure, which is inaccessible without supercomputer. All the calculations were done on LSSC-II. We chose a very large supercell ($17\text{Å} \times 17\text{Å} \times 17\text{Å}$, with 48 atoms). The long molecular dynamics simulation was performed at 300 K, we found that the carbon nano-tube shows hard-soft transition under the external pressure, with a shape changing from circular-like to elliptic-like. This finding will be helpful to further understand the physical properties of carbon tube and also be important to the application of the device.

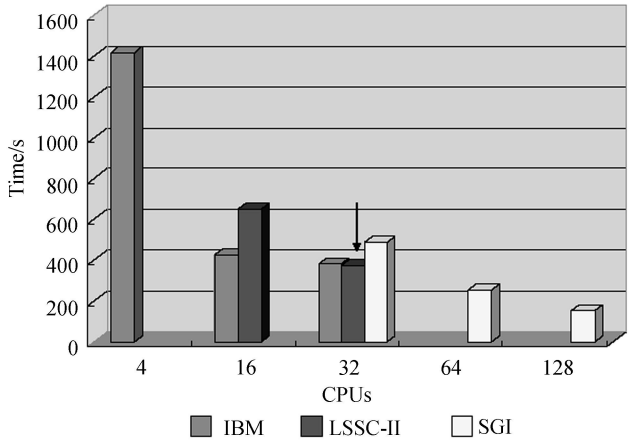

Fig. 6. Comparison of CPU time (second).

The second example is the problem of surface diffusion which is one of the most important kinetic processes controlling surface growth and thin film morphology. It has been a subject of extensive experimental and theoretical studies, both for its technological, as well as its basic scientific interest. To quantitatively study the physical relationship between surface diffusion barriers and external strain, thanks to LSSC-II, we have performed a comprehensive first-principles investigations of the strain-dependent adatom diffusion for Si on Si(001), Ge on Si(001), Si on Ge(001), and Ge on Ge(001) surfaces. The essentially linear dependence of the diffusion barriers on the external strain is recovered in all of these diffusion processes within small strain regime, giving strong evidence that a *priori* quantitative prediction of the effect of external strain on surface diffusion can be achieved by the linear approximation. A comparison between the surface diffusion processes on Si(001) and Ge(001) shows that the diffusion barriers on Ge(001) are lower, and the anisotropy of the diffusion is smaller, in good agreement with experiments.

8 Conclusion

PC clusters are becoming one of the most popular high performance computing platforms for scientific and engineering computing. The current high performance interconnecting network devices available, such as Myrinet, QsNet, etc., typically have a few microseconds in communication latency and 200—300 MB/s in communication bandwidth, and the floating point performance of today's microprocessors is about 20—60 Gflops. Our experiences with LSSC-II show that PC clusters with this kind of high performance interconnection are suitable for a large class of application problems, mainly those leading to numerical solution of PDEs. For the problems requiring more intensive communications, such as fast fourier transform, faster communication devices are required. While

rapid increase in communication bandwidth is expected, for example, the bandwidth of infiniBand can reach 700—800 MB/s and it begins to appear in the market, further reducing the communication latency seems to be much harder with present interconnecting hardware.

Another bottleneck on the performance of PC clusters is the slow memory access speed, the data in table 2 being a typical example, which limits the real efficiency of typical application programs to be between 3%—5% of the peak performance. The situation is becoming more severe with the increasing power of micro processors (Moore's law).

At the time of the writing of this paper, the cluster has been serving the LSSC project and LSEC for more than one year, running 24 h a day and 7 d a week. Apart from a few hardware failures in various hardware components, the system is stable and suits well to the need of the project. We are very satisfied with the design and performance of the system, and the only weak point that we have found in the system architecture is the relatively low I/O bandwitdh and small capacity (only 1 TB) of the RAID arrays——this is not a design problem, but the result of the limited budget for building the cluster.

Building a large cluster with hundreds of nodes or more is totally different from building smaller clusters. The system has to be carefully designed to yield good balance in different parts (communication, computation, filesystems, etc.). Many issues, including effectively deploying and managing large amount of nodes, limitations on the number of nodes/CPUs in some system software packages, stability of the hardware components, should be carefully resolved. Environmental issues such as wiring, cooling, power supply, etc., are also to be taken into account. Building and running LSSC-II have provided us with many valuable knowledge and experiences which will help in building and running larger clusters in the future.

Finally, we think that the importance of LSSC-II is not only in its successful construction and application, but also in its significance of the first joint work between a leading information technology company and advanced research institutions in the country, facilitated by the crucial support, initiative and input from the LSSC project, in building high performance computers. We believe that this is a big step forward and a milestone in China's high performance computing technology development.

Acknowledgements This work was supported by the Special Funds for the Major State Basic Research Projects (Grants No. G19990328), and partly supported by the National Natural Science Foundation of China (Grant No. 40004003).

References

1. Zhang Linbo, Design, configuration and some performance results of the PC cluster of the LSSC project, J. on Numer. Meth. and Comput. Applicat. (in Chinese), 2003, 24(1): 68—75.
2. Zhang Linbo, Simple and Effective Management of Large Clusters——a Tutorial, International Conference on Parallel Algorithms and Computing Environments, the Chinese University of Hong Kong, October 2003;

http://www.
sc.ac.cn/ICPACE/lecture/Linbo Zhang.tgz.
3. Platform Inc., The Load Sharing Facilities, http://www.platform.com
4. High-Performance BLAS by Kazushige Goto, http://www.cs.utexas.edu/users/flame/goto
5. TOP500 Supercomputing Sites, http://www.top500.org
6. Chen Shiyi, Ecke, R. E., Eyink, G. L. et al., Physical mechanism of the two-dimensional entropy cascade, Physical Review Letters, 2003, in press.
7. Chen Qiaoning, Chen Shiyi, Eyink, G. L. et al., Kolmogorov's third hypothesis and sign statistics, Physical Review Letters, 2003, 90: 214501.
8. Chen Qiaoning, Chen Shiyi, Eyink, G. L. et al., Intermittency in the joint cascade of energy and helicity, Physical Review Letters, 2003, 90: 214503.
9. Chen Qiaoning, Chen Shiyi, Eyink, G. L., The joint cascade of energy and helicity in three-dimensional turbulence, Physics of Fluids, 2003, 15: 361.
10. Chen Qiaoning, Chen Shiyi, Eyink, G. L. et al., Resonant interactions in rotating homogeneous three-dimensional turbulence, submitted to J. Fluid Mechanics, 2003.
11. Zhang Linbo, On pipelined computation of a set of recurrences on distributed memory systems, Chinese Journal of Numer. Math. and Appl., 2000, 22(1): 22—30.
12. Fu Dexun, Ma Yanwen, Zhang Linbo, Direct numerical simulation of transition and turbulence in compressible mixing layer, Science in China, Ser. A, 2000, 43(4): 421—429.
13. Moin, P., Mahesh, K., Direct numerical simulation: A tool in turbulence research, Annu. Rev. Fluid Mech., 1998, 30: 535—578.
14. Li Xinliang, Fu Dexun, Ma Yanwen, Direct numerical simulation of compressible isotropic turbulence, Science in China, Ser. A, 2002, 45(11): 1452—1460.
15. Li Xinliang, Fu Dexun, Ma Yanwen, Direct numerical simulation of passive scalar in decaying compressible turbulence, Science in China, Ser. G, 2004, to appear.
16. Liu Wei, Cao Jianwen. Mezzatesta A. et al., Parallel reservoir simulation on shared and distributed memory system, SPE, 2000, 64797.
17. Cao Jianwen, Pan Feng, Yao Jifeng et al., The implementation of a parallel software of petroleum reservoir simulation and its application on home-made high performance computers, Journal of Computer Research and Development, 2002, 39(8): 973—980.
18. Cao Jianwen, Lai Choihong, Numerical experiments of some Krylov subspace methods for black oil model, International Journal of Computers and Mathematics with Applications, 2002, 44(1/2): 125—141.
19. Zhang Guanquan, Zhang Wensheng, Parallel implementation of 2-D prestack depth migration, The Fourth International Conference/Exhibition on High Performance Computing in Asia-Pacific Region, Beijing, China, May 14—17, 2000, Expanded abstracts, 970—975.
20. Zhang Wensheng, Zhang Guanquan, 3-D prestack depth migration for SEG/EAEG subsalt with the SSF method, SEG International Exposition and Seventy-First Annual Meeting, San Antonio, TX, USA, September 9—14, 2001, Expanded abstracts, 1061—1064.
21. Chan, S. P., Chen, G., Gong, X. G. et al., Oxidation of carbon nanotubes by singlet O-2, Physical Review Letters, 2003, 90: 086403.
22. Sun, D. Y., Gong, X. G., A new constant-pressure molecular dynamics method for finite systems, Journal of Physics-Condensed Matter, 2002, 14: L487—L493.