MA6252 Homework1

Liu Zhaoqiang A0109983

1. Computer Performance and its evaluation

According to Wikipedia, computer performance is characterized by the amount of useful work accomplished by a computer system or computer network compared to the time and resources used. Computer performance metrics (things to measure) include availability, response time, channel capacity, latency, completion time, service time, bandwidth, throughput, relative efficiency, scalability, performance per watt, compression ratio, instruction path length and speed up. CPU benchmarks are available.

Depending on the context, good computer performance may involve one or more of the following:

Short response time for a given piece of work
High throughput (rate of processing work)
Low utilization of computing resource(s)
High availability of the computing system or application
Fast (or highly compact) data compression and decompression
High bandwidth / short data transmission time

2. Benchmarking computer performance for scientific computing

Because there are so many programs to test a CPU on all aspects of performance, benchmarks were developed. The most famous benchmarks are the SPECint and SPECfp benchmarks developed by Standard Performance Evaluation Corporation and the Consumer Mark benchmark developed by the Embedded Microprocessor Benchmark Consortium.

Fathoming the chief performance characteristics of a processor or system is one of the purposes of low-level benchmarking. A low-level benchmark is a program that tries to test some specific feature of the architecture like, e.g., peak performance or memory bandwidth. The purpose of this benchmark is to measure the performance of data transfers between memory and arithmetic units of a processor.

Low-level benchmarks are powerful tools to get information about the basic capabilities of a processor. However, they often cannot accurately predict the behavior of “real” application code. In order to decide whether some CPU or architecture is well-suited for some application (e.g., in the run-up to a procurement or before writing a proposal for a computer time grant), the only safe way is to prepare application benchmarks. This means that an application code is used with input parameters that reflect as closely as possible the real requirements of production runs. The decision for or against a certain architecture should always be heavily based on application benchmarking. Standard benchmark collections like the SPEC suite can only be rough guidelines.

What is the LINPACK performance of your own PC/Laptop and predict how long will it therefore take to solve a dense system on your PC with N=1000 unknowns:

Intel(R) LINPACK data

Current date/time: Sun Jan 25 12:15:05 2015

CPU frequency: 3.287 GHz
Number of CPUs: 1
Number of cores: 4
Number of threads: 8

Parameters are set to:

Number of tests	: 5				
Number of equations to solve (problem size)	: 1000	2000	5000	10000	15000
Leading dimension of array	: 1000	2000	5008	10000	15000
Number of trials to run	: 4	2	2	2	2
Data alignment value (in Kbytes)	: 4	4	4	4	4

Maximum memory requested that can be used = 1800304096, at the size = 15000

============= Timing linear equation system solver =================

Size	LDA	Align.	Time(s)	GFlops	Residual	Residual(norm)
1000	1000	4	0.021	32.2365	1.029343e-12	3.510325e-02
1000	1000	4	0.020	34.1878	1.029343e-12	3.510325e-02
1000	1000	4	0.020	32.8533	1.029343e-12	3.510325e-02
1000	1000	4	0.020	33.8446	1.029343e-12	3.510325e-02
2000	2000	4	0.143	37.2965	4.298950e-12	3.739560e-02
2000	2000	4	0.143	37.3387	4.298950e-12	3.739560e-02
5000	5008	4	2.267	36.7775	2.581643e-11	3.599893e-02
5000	5008	4	2.244	37.1656	2.581643e-11	3.599893e-02
10000	10000	4	17.913	37.2274	9.603002e-11	3.386116e-02
10000	10000	4	19.199	34.7342	9.603002e-11	3.386116e-02

3. Critique of LINPACK and TOP500

Since 1993, the fastest supercomputers have been ranked on the TOP500 list according to their LINPACK benchmark results. The list does not claim to be unbiased or definitive, but it is a widely cited current definition of the "fastest" supercomputer available at any given time. The

TOP 10 Sites for November 2014 can be found in http://www.top500.org/lists/2014/11/ .

In general, the speed of supercomputers is measured and benchmarked in "FLOPS" (FLoating point Operations Per Second), and not in terms of "MIPS" (Million Instructions Per Second), as is the case with general-purpose computers.

The LINPACK Benchmarks are a measure of a system's floating point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering.

No single number can reflect the overall performance of a computer system, yet the goal of the LINPACK benchmark is to approximate how fast the computer solves numerical problems and it is widely used in the industry. The FLOPS measurement is either quoted based on the theoretical floating point performance of a processor (derived from manufacturer's processor specifications and shown as "Rpeak" in the TOP500 lists) which is generally unachievable when running real workloads, or the achievable throughput, derived from the LINPACK benchmarks and shown as "Rmax" in the TOP500 list. The LINPACK benchmark typically performs LU decomposition of a large matrix. The LINPACK performance gives some indication of performance for some real-world problems, but does not necessarily match the processing requirements of many other supercomputer workloads, which for example may require more memory bandwidth, or may require better integer computing performance, or may need a high performance I/O system to achieve high levels of performance.

4. Technological trends

As described in a recent international assessment of simulation-based engineering and science by the World Technology Evaluation Center (WTEC, 2009), the world of computing is flattening, and any country with interest and minimal resources can have access to HPC. Several countries have the technology, resources, and desire to be first to achieve exaflop (10^{18} FLOPS) computing. Japan, Germany, France, and China are all committed to HPC, with ready access to world-class resources, faculty, and students. Germany is investing nearly U.S. $1 billion toward next-generation HPC hardware in partnership with the European Union (WTEC, 2009). Japan has always been a leader in HPC, and as the first to achieve 40 teraflops in the past decade with the Earth Simulator built by Japanese computer company NEC, Japan has an industry-university-government roadmap to reach exascale computing by 2025.

The pace of the next major leap in computing power, an exaflop computer, will be significantly challenging to accomplish in a cost-effective and scalable way. The top obstacles are power consumption, resilience, the memory wall, and scalability. Kogge et al. (2008) discuss these areas in great detail. Using computer vendor roadmaps suggests that an exaflop computer would require finding 10^3 to 10^4 times more parallelism in existing applications, hundreds of megawatts of power, and a mean time to failure of only a few minutes. Power requirements and resiliency obstacles are a significant challenge and will need to be resolved by the HPC industry with government-supported research and development.

Finding the increased levels of parallelism for strong scaling will require new approaches to the underlying algorithms. Strong scaling holds the problem size fixed and improves computational speed by increasing the number of processors used to solve the problem, while weak scaling allows the problem size to increase by increasing the number of processors. In the era of teraflop to petaflop computing, weak scaling has dominated because of the requirement of realistic, numerically converged, three-dimensional simulations. In addition, for much of this era the processors were single-core chips and increased processing speeds provided much of the increased HPC capability. In exaflop computing, strong scaling, which improves computational speed, will be an important consideration for many time-dependent applications, as opposed to weak scaling that improves the ability to process larger problems.