High Performance Computers

- From the beginning of the digital age, supercomputers have been time machines that let researchers peer into the future, both intellectually and temporally.
  - Intellectually bring to life models of complex phenomena when economics and other constraints preclude experimentation.
  - Temporally, they reduce the time to solution by enabling us to evaluate larger and more complex models than would be possible on more conventional systems.

High Performance Computers

- ~ 25 years ago
  - 1x10^8 Floating Point Ops/sec (Mflop/s)
    - Scalar based
- ~ 10 years ago
  - 1x10^9 Floating Point Ops/sec (Gflop/s)
    - Vector & Shared memory computing, bandwidth aware
    - Block partitioned, latency tolerant
- Today
  - 1x10^12 Floating Point Ops/sec (Tflop/s)
    - Highly parallel, distributed processing, message passing, network based
    - data decomposition, communication/computation
- ~ 5 years away
  - 1x10^13 Floating Point Ops/sec (Pflop/s)
    - Many more levels MH, combination/grids&HPC
    - More adaptive, LT and bandwidth aware, fault tolerant, extended precision, attention to SMP nodes

What is a Supercomputer?

- A supercomputer is a hardware and software system that provides close to the maximum performance that can currently be achieved.
- Over the last 10 years the range for the Top500 has increased greater than Moore’s Law
  - 1993: ≈ 59.7 Gflop/s
  - 500 = 422 MFlop/s
- 2005:
  - ≈ 1x10^13 Floating Point Ops/sec (Pflop/s)
  - Many more levels MH, combination/grids&HPC
  - More adaptive, LT and bandwidth aware, fault tolerant, extended precision, attention to SMP nodes

Top 500 Computers

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP
  \[ \text{Ax} = b, \text{dense problem} \]
- Updated twice a year
  - SC'xy in the States in November
  - Meeting in Germany in June

24th List: The TOP10

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Computer</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>BlueGene, 512 MHz</td>
<td>1</td>
</tr>
<tr>
<td>Intel</td>
<td>Itanium2, 3.2 GHz</td>
<td>2</td>
</tr>
<tr>
<td>Sun</td>
<td>UltraSPARC, 1.6 GHz</td>
<td>3</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Earth Simulator, 4.3 TFlops</td>
<td>4</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>Supercomputer, 5.0 TFlops</td>
<td>5</td>
</tr>
<tr>
<td>SGI</td>
<td>Altix, 1.6 GHz</td>
<td>6</td>
</tr>
<tr>
<td>NEC</td>
<td>Earth Simulator, 5.0 TFlops</td>
<td>7</td>
</tr>
<tr>
<td>IBM</td>
<td>Power, 1.6 GHz</td>
<td>8</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Earth Simulator, 3.5 TFlops</td>
<td>9</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>Supercomputer, 4.3 TFlops</td>
<td>10</td>
</tr>
</tbody>
</table>
Top500 Conclusions

- Microprocessor based supercomputers have brought a major change in accessibility and affordability.
- MPPs continue to account of more than half of all installed high-performance computers worldwide.

High-Performance Computing Directions: Beowulf-class PC Clusters

**Definition:**

- COTS PC Nodes
  - Pentium, Alpha, PowerPC, SMP
- COTS LAN/SAN Interconnect
  - Ethernet, Myrinet, Gigabit, ATM
- Open Source Unix
  - Linux, BSD
- Message Passing Computing
  - MPI, PVM
  - HPF

**Advantages:**

- Best price-performance
- Low entry-level cost
- Just-in-place configuration
- Vendor invulnerable
- Scalable
- Rapid technology tracking

Enabled by PC hardware, networks and operating system achieving capabilities of scientific workstations at a fraction of the cost and availability of industry standard message passing libraries. However, much more of a contact sport.
Virtual Environments

Do they make any sense?

Performance Improvements for Scientific Computing Problems

Different Architectures

Different Architectures

- Parallel computing: single systems with many processors working on same problem
- Distributed computing: many systems loosely coupled by a scheduler to work on related problems
- Grid Computing: many systems tightly coupled by software, perhaps geographically distributed, to work together on single problems or on related problems

Distributed and Parallel Systems

Types of Parallel Computers

- The simplest and most useful way to classify modern parallel computers is by their memory model:
  - shared memory
  - distributed memory
Shared vs. Distributed Memory

- **Shared memory** - single address space. All processors have access to a pool of shared memory. (Ex: SGI Origin, Sun E10000)
- **Distributed memory** - each processor has its own local memory. Must do message passing to exchange data between processors. (Ex: CRAY T3E, IBM SP, clusters)

Distributed Memory: MPPs vs. Clusters

- Processors-memory nodes are connected by some type of interconnect network
  - Massively Parallel Processor (MPP): tightly integrated, single system image.
  - Cluster: individual computers connected by switch

Processors, Memory, & Networks

- Both shared and distributed memory systems have:
  1. **processors**: now generally commodity RISC processors
  2. **memory**: now generally commodity DRAM
  3. **network/interconnect**: between the processors and memory (bus, crossbar, fat tree, torus, hypercube, etc.)

Standard Uniprocessor Memory Hierarchy

- **Intel Pentium 5** 3.2 - 3.46 GHz processor
  - 130 nm technology
  - 12 KBytes of 4 way assoc. L1 instruction cache with 32 byte lines
  - 8 KBytes of 4 way assoc. L1 data cache with 32 byte lines
  - 512 KBytes of 8 way assoc. L2 cache, 32 byte lines
  - 2 MB L3 cache
  - 400 MB/s bus speed
  - 3.2 GBs
  - SSE2 provide peak of 6.4-6.92 Gflop/s
Processor-Related Terms

Clock period (cp): the minimum time interval between successive actions in the processor. Fixed, depends on design of processor. Measured in nanoseconds (~1-5 for fastest processors). Inverse of frequency (MHz)

Instruction: an action executed by a processor, such as a mathematical operation or a memory operation.

Register: a small, extremely fast location for storing data or instructions in the processor.

Functional Unit: a hardware element that performs an operation on an operand or pair of operations. Common FUs are ADD, MULT, INV, SQRT, etc.

Pipeline: technique enabling multiple instructions to be overlapped in execution.

Superscalar: multiple instructions are possible per clock period.

Flops: floating point operations per second.

Cache: fast memory (SRAM) near the processor. Helps keep instructions and data close to functional units so processor can execute more instructions more rapidly.

TLB: Translation-Lookaside Buffer keeps addresses of pages (block of memory) in main memory that have recently been accessed (a cache for memory addresses).

SRAM: Static Random Access Memory (RAM). Very fast (~10 nanoseconds), made using the same kind of circuitry as the processors, so speed is comparable.

DRAM: Dynamic RAM. Longer access times (~100 nanoseconds), but hold more bits and are much less expensive (10x cheaper).

Memory hierarchy: the hierarchy of memory in a parallel system, from registers to cache to local memory to remote memory. More later.

Latency: How long does it take to start sending a "message"? Measured in microseconds.

(Also in processors: How long does it take to output results of some operations, such as floating point add, divide etc., which are pipelined?)

Bandwidth: What data rate can be sustained once the message is started? Measured in Mbytes/sec.

Topology: the manner in which the nodes are connected.

Best choice would be a fully connected network (every processor to every other). Unfeasible for cost and scaling reasons.

Instead, processors are arranged in some variation of a grid, torus, or hypercube.
Highly Parallel Supercomputing: Where Are We?

- **Performance:**
  - Sustained performance has dramatically increased during the last year.
  - On most applications, sustained performance per dollar now exceeds that of conventional supercomputers. But...
  - Conventional systems are still faster on some applications.

- **Languages and compilers:**
  - Standardized, portable, high-level languages such as HPF, PVM, and MPI are available. But...
  - Initial HPF releases are not very efficient.
  - Message passing programming is tedious and hard to debug.
  - Programming difficulty remains a major obstacle to usage by mainstream scientists.

- **Operating systems:**
  - Robustness and reliability are improving.
  - New system management tools improve system utilization. But...
  - Reliability still not as good as conventional systems.

- **I/O subsystems:**
  - New RAID disks, HiPPI interfaces, etc. provide substantially improved I/O performance. But...
  - I/O remains a bottleneck on some systems.

The Importance of Standards - Software

- Writing programs for MPP is hard...
- But... one-off efforts if written in a standard language
- Past lack of parallel programming standards...
  - Has restricted uptake of technology (to "enthusiasts")
  - Reduced portability (over a range of current architectures and between future generations)
- Now standards exist: (PVM, MPI & HPF), which...
  - Allows users & manufacturers to protect software investment
  - Encourage growth of a "third party" parallel software industry & parallel versions of widely used codes

The Future of HPC

- The expense of being different is being replaced by the economics of being the same
- HPC needs to lose its "special purpose" tag
- Still has to bring about the promise of scalable general purpose computing...
- But it is dangerous to ignore this technology
- Final success when MPP technology is embedded in desktop computing
- Yesterday’s HPC is today’s mainframe is tomorrow’s workstation

Achieving TeraFlops

- In 1991, 1 Gflop/s
- 1000 fold increase
  - Architecture
    - Exploiting parallelism
  - Processor, communication, memory
    - Moore’s Law
  - Algorithm improvements
    - Block-partitioned algorithms
Future: Petaflops ($10^{15}$ flop pt ops/s)

Today $\approx \sqrt{10^{15}}$ flops for our laptops

- A Pflop for 1 second $\approx$ a typical workstation computing for 1 year.
- From an algorithmic standpoint:
  - concurrency
  - data locality
  - latency & sync
  - floating point accuracy

Petaflop ($10^{15}$ flop/s) Computers
Within the 5 Years

- Five basis design points:
  - Conventional technologies
    - 4.8 GHz processor, 8000 nodes, each w/16 processors
  - Processing-in-memory (PIM) designs
    - Reduce memory access bottleneck
  - Superconducting processor technologies
    - Digital superconductor technology, Rapid Single-Flux-Quantum (RSFQ) logic & hybrid technology multi-threaded (HTMT)
  - Special-purpose hardware designs
    - Specific applications e.g. GRAPE Project in Japan for gravitational force computations
  - Schemes utilizing the aggregate computing power of processors distributed on the web
    - SETI@home ~26 Tflops

Petaflops ($10^{15}$ flop/s) Computer Today?

- 2 GHz processor ($O(10^9)$ ops/s)
  - 1/2 Million PCs
  - $18$ ($2K$ each)
  - 100 Mwatts
  - 2.5 acres
  - 1/2 Million Windows licenses!!
  - PC failure every second

---

5. A Pflop for 1 second $\approx$ a typical workstation computing for 1 year.
6. From an algorithmic standpoint:
   - concurrency
   - data locality
   - latency & sync
   - floating point accuracy

---

9. Today $\approx \sqrt{10^{15}}$ flops for our laptops
10. Five basis design points:
    - Conventional technologies
        - 4.8 GHz processor, 8000 nodes, each w/16 processors
    - Processing-in-memory (PIM) designs
        - Reduce memory access bottleneck
    - Superconducting processor technologies
        - Digital superconductor technology, Rapid Single-Flux-Quantum (RSFQ) logic & hybrid technology multi-threaded (HTMT)
    - Special-purpose hardware designs
        - Specific applications e.g. GRAPE Project in Japan for gravitational force computations
    - Schemes utilizing the aggregate computing power of processors distributed on the web
        - SETI@home ~26 Tflops

---

13. Petaflops ($10^{15}$ flop/s) Computer Today?
14. 2 GHz processor ($O(10^9)$ ops/s)
   - 1/2 Million PCs
   - $18$ ($2K$ each)
   - 100 Mwatts
   - 2.5 acres
   - 1/2 Million Windows licenses!!
   - PC failure every second