Supercomputer Architecture

B V V Sri Raj Dutt

Third Year Undergraduate
Computer Science and Engineering
Indian Institute of Technology, Kanpur

Tutors : Prof. Wellein, Prof. Stümer

Indo-German Winter Academy 2009, IIT Roorkee
Outline

1. Introduction
2. History
3. Present day supercomputers
4. Parallelism
5. Memory Hierarchy
6. Top supercomputers
Definition - Supercomputer

Computer that is at the frontier of the current processing capacity (speed of calculation)

Used for solving highly calculation intensive tasks (complex problems) like in

- Simulation and modelling of physical phenomena
  - Climate change
  - Explosions - detonation of nuclear weapons
  - Airplanes in wind tunnels
  - Behaviour of molecules
- Analysis of data
  - National Security Intelligence
  - Genome sequencing
  - Astronomical observations
- Intricate design of engineering products
- Cryptographic Analysis
Introduction

Characteristics

Comparison with normal PC

- Supercomputer - executing few programs as fast as possible
- Normal PC - execute many programs concurrently

Characteristics

- High costs
- Specialized
- Carefully designed Memory Hierarchy
- Controlled environment
- Operating system - variants of Linux
- Programming Language - Fortran, C

Speed of the supercomputer is measured in peak FLOPS.
A supercomputer is composed of processors, memory, I/O system, and an interconnect.
A supercomputer is composed of processors, memory, I/O system, and an interconnect.
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

- The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.
- The memory bandwidth however has increased only at rate of 38% per year overall.
- From 1995 this rate has decreased to 23% per year.
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall

From 1995 this rate has decreased to 23% per year
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall.

From 1995 this rate has decreased to 23% per year.
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall.

From 1995 this rate has decreased to 23% per year.
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall

From 1995 this rate has decreased to 23% per year
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall.

From 1995 this rate has decreased to 23% per year.

Some Concepts

- Instruction Level Pipelining
- Cache
As semiconductor and packaging technology improves, different aspects of a computer improve at different rates.

The floating-point performance of microprocessors improved by 59% per year during the period of past 20 years.

The memory bandwidth however has increased only at rate of 38% per year overall.

From 1995 this rate has decreased to 23% per year.

**Some Concepts**

- Instruction Level Pipelining
- Cache
Introduced mainly in 1960s - primarily by Seymour Cray at CDC

Example: CDC6600

- 1-10 Mflops
- Single complex CPU
- Differences compared to then computers
  - handle only logic and data - so memory latency cut down
  - 60-bit word length
  - RISC instead of CISC
- Central Processor
  - 10 parallel functional units - superscalar
  - no pipelining
- Implemented virtual peripheral processors
Example: CDC7600

- (1969-1975) 10-100 Mflops
- Instruction Pipelining - 3 times boost in performance
- Instruction -
  - fetch
  - decode
  - execute
- Pipelined execution on functional units in parallel

Cray moved on to form his own company Cray research and started working on Cray-1
Cray series

- 1975-1985
  - Worked on new design - vector processors
    - load large data at once
    - perform operations simultaneously

- Mid 1980s - modest number of vector processors in parallel (4-16)
- Late 1980s and 1990s - Massively parallel processing systems with thousands of ordinary CPUs - either off-the-shelf or custom designs
Parallel designs based on *off-the-shelf* RISC microprocessors are mainly used.

Computer clusters of MIMD processors each processor of which is SIMD.

Each processor does a portion of the work and uses custom interconnections to interact with its peers.

Processors - vector, scalar or superscalar.

Different memory hierarchies - shared or distributed.

**Classification of supercomputers**

- Custom-based
- Commodity-based
- Hybrid
Parallel computing

- Parallel computing
  - divide large problem into smaller ones
  - solve each of them concurrently on different processors

- Parallelization strategies
  - Data parallelism
  - Functional parallelism
Parallelism

Scalability

- Parallelization is the process of formulating a problem in a way that lends itself to concurrent execution by several execution units of some kind.
- Using $N$ processors, we expect the execution time to come down $N$ times ideally.
- This is termed as speedup of $N$.
- Reality is not so perfect.

Limitations

- Shared resources.
- Dependencies between different processors.
- Pipeline startups.
- Communication.
- Load imbalance.

The serial part limits speedup.
- 1 processor - \( T(1) = s + p = 1 \) i.e., \( p = 1 - s \)
- \( n \) processors - \( T(n) = s + \left( \frac{p}{n} \right) \)
- Scalability (Speed Up) = \( \frac{T(1)}{T(n)} = \frac{1}{s + \frac{1-s}{n}} \)

\[\text{Figure: [7] Amdahl’s Law}\]
Different architectures are at programmer’s disposal to implement parallel algorithm -

1. Shared-memory systems
   1. Uniform Memory Access (UMA)
   2. cache coherent Non-Uniform Memory Access (ccNUMA)

2. Distributed-memory systems
   1. Massively Parallel Processing
   2. Cluster computing

Each has a standardized programming model
Shared Memory

- Number of systems work on shared physical address space
- Interaction by modifying data in the shared space
- Relatively easy to program as all processors have same view of data
- OpenMP Programming
- Problems -
  - Memory-CPU bus bandwidth is the bottleneck
  - cache Coherence
Uniform Memory Access

- also called Symmetric Multiprocessing (SMP)
- Flat-memory model: Bandwidth and latency same for all processors for all memory locations - easy to use
- Disadvantage: Limited scalability since a single resources (CHIPSET) has to manage all data traffic and cache coherence
- Simplest example - dual core processor
Memory Hierarchy

Shared Memory Computing

(cc)NUMA

- Memory is physically distributed but a shared address space exists
- Physical Layout is similar to distributed systems
- Thus memory access performance varies on which processor accesses which part of memory
- Two locality domains linked through a high speed connection called Hyper Transport (AMD), QPI (Intel Core i7 series)
- Easier to scale to large configuration
- Disadvantage: data locality issues
Distributed Memory

- Each processor is connected to exclusive local memory
- Each node has at least one Network Interface to mediate communication with communication network
- Different processors communicate with each other through the network
- Communication can be blocking or non-blocking
- Programming - Message Passing - MPI
Massively Parallel Processing (MPP)

- Independent microprocessors that run in parallel
- Each processor has its own memory
- Individual operating systems
- All processing elements are connected to form one large computer
- A number of processors (can be all!) work on the same task
- Messaging Interface is used for communication
- Examples: Blue Gene
Distributed Computing Architectures: Clusters

- Completely independent computers combined to a unified system through software and networking
- Components connected through fast LANs - relies heavily on network speed
- Easily upgraded with addition of new systems
Hybrid Architecture

- The largest and fastest computers of the present day employ hybrid of shared and distributed memory architectures.
- At low level compute nodes comprise multiple processors sharing same address space (shared memory).
- At higher level Distributed Memory.
**Top 5 Supercomputers - November 2009**

- www.top500.org - List released twice per year based on Linpack benchmark

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Peak Performance (petaFlops)</th>
<th>Some characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jaguar</td>
<td>2.3</td>
<td>Cray XT5 - HE Opteron Six Core 2.6 GHz</td>
</tr>
<tr>
<td>2</td>
<td>Roadrunner</td>
<td>1.37</td>
<td>BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband</td>
</tr>
<tr>
<td>3</td>
<td>Kraken XT5</td>
<td>1.028</td>
<td>Cray XT5-HE Opteron Six Core 2.6 GHz</td>
</tr>
<tr>
<td>4</td>
<td>JUGENE</td>
<td>1.002</td>
<td>Blue Gene/P Solution</td>
</tr>
<tr>
<td>5</td>
<td>Tianhe-1</td>
<td>1.002</td>
<td>NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870, Infiniband</td>
</tr>
</tbody>
</table>
Jaguar supercomputer

- 84 cabinet quad-core Cray XT4 system and 200 upgraded Cray XT5 cabinets, using six-core processors
- Total - 362 terabytes of memory
- The two systems are connected to the Scalable I/O Network (SION), which links them together and to the Spider file system

**Jaguar Specifications**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>XT5</th>
<th>XT4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Teraflops</td>
<td>2,332</td>
<td>263</td>
</tr>
<tr>
<td>Six-Core AMD Opterons</td>
<td>37,376</td>
<td></td>
</tr>
<tr>
<td>Quad-Core AMD Opterons</td>
<td>7,832</td>
<td></td>
</tr>
<tr>
<td>AMD Opteron Cores</td>
<td>224,256</td>
<td>31,328</td>
</tr>
<tr>
<td>Compute Nodes</td>
<td>18,688</td>
<td>7,832</td>
</tr>
<tr>
<td>Memory (TB)</td>
<td>299</td>
<td>62</td>
</tr>
<tr>
<td>Memory Bandwidth (GB/s)</td>
<td>478</td>
<td>100</td>
</tr>
<tr>
<td>Disk Space (TB)</td>
<td>10,000</td>
<td>750</td>
</tr>
<tr>
<td>Interconnect Bandwidth</td>
<td>374</td>
<td>157</td>
</tr>
<tr>
<td>Floor Space (ft²)</td>
<td>4,352</td>
<td>1,344</td>
</tr>
<tr>
<td>Cooling Technology</td>
<td>Liquid</td>
<td>Air</td>
</tr>
</tbody>
</table>
Four nodes on both the XT4 and XT5 boards (blades).

The XT4 nodes have a single AMD quad-core Opteron 1354 Budapest processor coupled with 8 gigabytes of DDR2-800 memory.

The XT5 is a double-density version of the XT4. It has 3.7 times the processing power and twice the memory and memory bandwidth on each node.

The XT5 node has two Opteron 2435 Istanbul processors linked with dual HyperTransport connections. Each Opteron has directly attached 8 gigabytes of DDR2-800 memory.

The result is a dual-socket, twelve-core node with 16 gigabytes of shared memory and a peak processing performance of 125 gigaflops.
Jaguar contd...

- The Cray XT5 and XT4 blades incorporates a high-bandwidth, low-latency interconnect composed of Cray SeaStar2+ chips and high-speed links based on HyperTransport and proprietary protocols.
- The interconnect directly connects all compute nodes in a 3D torus topology,
- Use Cray’s version of Suse Linux as operating system

**Figure:** HyperConnect to SeaStar2+ in Cray XT4
Simulations

- Provide First Simulation of Abrupt Climate Change
- Help Studies of Supernovas, Space
- Simulations that will help reveal the detailed workings of cellulose, a potential biofuel material

and many more...
Roadrunner

- Second Fastest supercomputer
- Hybrid of 3 different architectures
- Consists of dual-core Opteron server processors (based on AMD-64 architecture) and attached to each is a Cell processor (IBM - Power Architecture Technology)
- Overall specs -
  - 6480 Opteron processors
  - 12960 IBM Cell processors
  - Infiniband Network
  - Linux - Red Hat Enterprise Linux and Fedora
Roadrunner contd...

- Triblade - 2 dual-core Opterons with 16GB RAM, four PowerXCell 8i CPUs with 16 GB Cell RAM

Triblade Configuration of Roadrunner - 400 GigaFlops
Opetron Processor

- First 64-bit instruction set x86 architecture
  - Introduced in 2003 by AMD
  - CISC instruction set on a superscalar RISC micro architecture
  - Complex instructions are internally translated to a RISC-like micro-ops RISC instruction set
  - Higher performance than earlier CISC architectures
- Each Opetron chip contains two separate processor cores
- Double computing-power at each motherboard processor socket
- Cost effective
- ccNUMA memory architecture
- Available memory bandwidth can be scaled with special care in programming and usage
Memory Hierarchy

CB2 Cell blades

- 256 KB of “working” memory (per SPE)
- 25.6 GB/s off-SPE BW
- ~200 GB/s per Cell on EIB bus
- 4 GB of shared memory (per Cell)
- 21.3 GB/s/chip
- 8 GB of NUMA shared memory (per node)
- 16 GB of distributed memory (per node)
- 2 GB of memory (per core)
- 4 GB of shared memory (per socket)
- 5.4 GB/s/core
- 8 GB of NUMA shared memory (per node)

One Cell chip per Opteron core

LS21 Opteron blade

- PCIe x8 (2 per blade) (2 GB/s, 2 us)
Roadrunner contd...

TriBlade:
- 4x PowerXCell 8i
- 2x Opteron 2210
- 32 GB RAM
- ~400 GFlops

Infiniband 4x

BladeCenter H:
- 3x TriBlade
- 12x PowerXCell 8i
- 6x Opteron 2210
- 96 GB RAM
- ~1.2 Tflops

Connected Unit:
- 15x Racks
- 60x BladeCenter H
- 180x TriBlades
- 720x PowerXCell 8i
- 360x Opteron 2210
- 5.76 TB RAM
- ~72 Tflops

Com & Service node:
- 12x I/O nodes, IBM x3655
- ~250 Gflops
- Service node, IBM x3655
- Infiniband switch, Voltaire ISR2012

Connected Unit switch
Conclusions

- Supercomputers of today are the general purpose computers of tomorrow.
- Commodity clusters satisfy the needs of many supercomputer users.
- Some important applications need the better main memory bandwidth and latency hiding that are available only in custom supercomputers.
- Many need the better global bandwidth and latency interconnects that are available only in custom or hybrid supercomputers and most would benefit from the simpler programming model that can be supported well on custom systems.
- The increasing gap between processor speed and communication latencies is likely to increase the fraction of supercomputing applications that achieve acceptable performance only on custom and hybrid supercomputers.
- The local and global memory bandwidth bottleneck is expected to become a more serious problem in the future due to the nonuniform scaling of technology.
References


*TOP 500*, http://www.top500.org


Cray Inc. www.cray.com

Acknowledgement

I would really like to thank my tutors and all the organizers and members of Indo-German Winter Academy for giving me this wonderful opportunity.

Thank You