Maximising Performance

How can performance be maximised?

- Reduce the time per instruction (cycle time) [1]: clock rate.
- Increase the number of instructions executed per-cycle [2]: pipelining.
- Allow multiple processors to work on different parts of the same program at the same time [3]: parallel computing.

Limiting factors for approach [1] and [2]

- There is a limit to how quick CPU can tick.
- Heat dissipation.
- A instruction processing procedure cannot be divided into infinite stages
- Power consumption

Limiting factors for approach [3]

- Overhead of communications
Metrics

- **FLOPS (FLoating point Operations Per Sec)** - a measure of the numerical processing of a computer which can be an indicator of its scientific computing (real numbers) capability.

- The floating-point format is a variation of scientific notation - the real number is represented using a mantissa, base, and exponent.

**Storing real number in computers:**

- use the fixed length of word as the storage space for a real number (e.g. 64bits)
- The number is converted to base-2
- Mantissa is normalised (1.61 is normalised, 16.1 is not)
- Some parts of the word are used to store the mantissa, 1bit to store sign, and the rest to store the exponent
Example of Floating Point Numbers

172.625  base 10

10101100.101 X 2^0  base 2

1.0101100101 X 2^7  base 2 normalised

Using 32 bits (4 bytes) to store the number in computers, in which the first bit for sign, the next 8 bits for exponent, and the rest for Mantissa

0 00000111 0000000000000010101100101

S Exp  Mantissa
Floating point format

- **Advantages**
  - Using a fixed-length space to store a wide overall range of values
    - If 64 bits are used to store the real numbers, in which 11 bits are used to store exponent and 52 bits to mantissa (the remaining 1 bit used to store sign). We can derive the range of numbers this storage layout can represent
    - More bits are used to store mantissa, higher precision, but smaller range
    - More bits are used to store exponent, wider range, but lower precision

- **Disadvantages**
  - The difference between two successive numbers is not uniform
  - When the numbers cannot be perfected converted to base-2 numbers, they must be rounded to be stored in the format, leading to some problems where algebraic rules do not appear to apply

- The LINPACK benchmark produces a FLOPS results. It solves a dense system of linear equations.
Metrics

- MIPS (Millions of Instructions Per Second) - a measure of the speed of a processor.
  - Peak MIPS rates (usually vendor supplied) can be misrepresentative
  - Meaningless Information on Performance for Salespeople
  - People seldom refer to it
Metrics

- **SPECint** - measures a processor’s integer processing capabilities.
  - Latest version SPECint2006
  - Can test cpu, memory, compiler, but cannot test networking, I/O
  - Consists of a series of benchmarks (12, including compression, compilation)

- **Generating SPECint benchmark results**
  - each benchmark has a reference time
  - The reference time is divided by the measured runtime of the benchmark
    
    For example, if we run the benchmark 401.bzip2 to test the system, whose reference time is 1400. The actual runtime of the benchmark is 140 sec. then the benchmarking result for this benchmark is calculated as 1400/140=10
  - These are averaged to produce a final integer performance of the system.
## SPECint2006 benchmark suite

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Language</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>400.perlbench</td>
<td>C</td>
<td>Programming Language</td>
</tr>
<tr>
<td>401.bzip2</td>
<td>C</td>
<td>Compression</td>
</tr>
<tr>
<td>403.gcc</td>
<td>C</td>
<td>C Compiler</td>
</tr>
<tr>
<td>429.mcf</td>
<td>C</td>
<td>Combinatorial Optimization</td>
</tr>
<tr>
<td>445.gobmk</td>
<td>C</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>456.hmmer</td>
<td>C</td>
<td>Search Gene Sequence</td>
</tr>
<tr>
<td>458.sjeng</td>
<td>C</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>462.libquantum</td>
<td>C</td>
<td>Physics / Quantum Computing</td>
</tr>
<tr>
<td>464.h264ref</td>
<td>C</td>
<td>Video Compression</td>
</tr>
<tr>
<td>471.omnetpp</td>
<td>C++</td>
<td>Discrete Event Simulation</td>
</tr>
<tr>
<td>473.astar</td>
<td>C++</td>
<td>Path-finding Algorithms</td>
</tr>
<tr>
<td>483.xalancbmk</td>
<td>C++</td>
<td>XML Processing</td>
</tr>
</tbody>
</table>
Metrics - Communications

- **Bandwidth (bytes/sec)**
  - How much data can be sent per second over the network

- **Latency (seconds)**
  - The time between the source processor sending a message and the destination processor receiving the message

- **Influential factors**
  - **Interconnection type:**
    - On-board interconnection or over-networks.
  - **Topologies:**
    - bus (advantages: low cost, easy to maintain consistency; disadvantages: contention)
    - Crossbar (advantages: less contention; disadvantages: high cost,)
    - hub, switch
  - **Protocols:** stacks
  - unicast, multicast, broadcast
Metrics

Storage capabilities:

- **Bandwidth and Latency.**
  - Bandwidth: how much data can be accessed per second in a certain storage facility
  - Latency: the time between sending a data accessing request and receiving the requested data

- **Storage facilities:** register, cache, memory, hard disk

- **Memory hierarchies:** cpu register -> cache -> main memory -> remote memory

- **Hard disk hierarchies:** Local and remote file systems