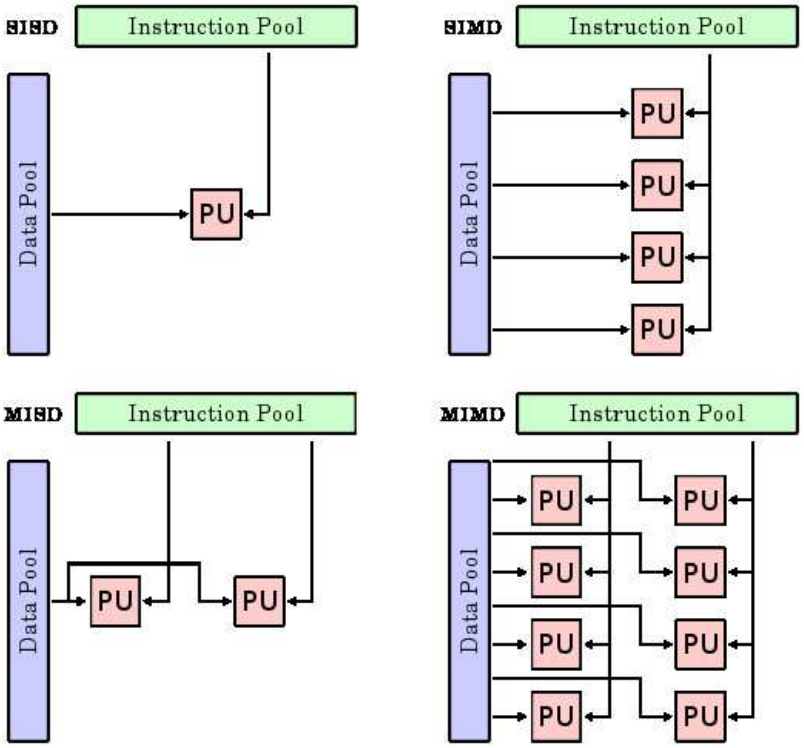
Miscellaneous High-Performance Features

Over the past several decades, numerous ideas have been used to increase computational speed. Here is a brief summary.

Variations in Computer Architecture (Flynn's Taxonomy)



SIMD "SIMD" is an acronym for "Single Instruction Multiple Data". The idea here is that a single instruction is applied to multiple data items simultaneously. Historically, "vector supercomputers", (e.g., Cray Y-MP, 1988) used SIMD.. The IBM PowerPC series chips (used in Macintosh computers until 2005) included both traditional (scalar) instructions, and vector instructions. Apple's Altivec library gave impressive performance for linear algebra calculations. The disadvantage of this approach is that some algorithms can not be easily written in terms of vector operations.

MIMD "MIMD" is an acronym for "Multiple Instruction Multiple Data". (Example Intel Xeon Phi). The OpenMP shared memory programming model is based on multiple instruction streams operating independently on multiple (private) data. One of the key issues in MIMD is the memory system, and here there are several variations:

Bus-based In the simplest form, all processors are attached to a bus which connects them to memory.

Hierarchical A hierarchy of buses gives processors access to each other's memory. Processors on different boards may communicate through inter-nodal buses. Example: Sun V490.

Distributed Shared Memory Each processor has no direct connection to other processor's memory. For data to be shared, it must be passed from one processor to another as a message. However, the message passing is transparent to the user process space. I.e., a programmer writes a shared memory program (e.g., OpenMP), and the OS/hardware does the message passing "behind the scenes". The SGI Origin series of computers were based on distributed shared memory and a proprietary network interconnect (Cray Link).

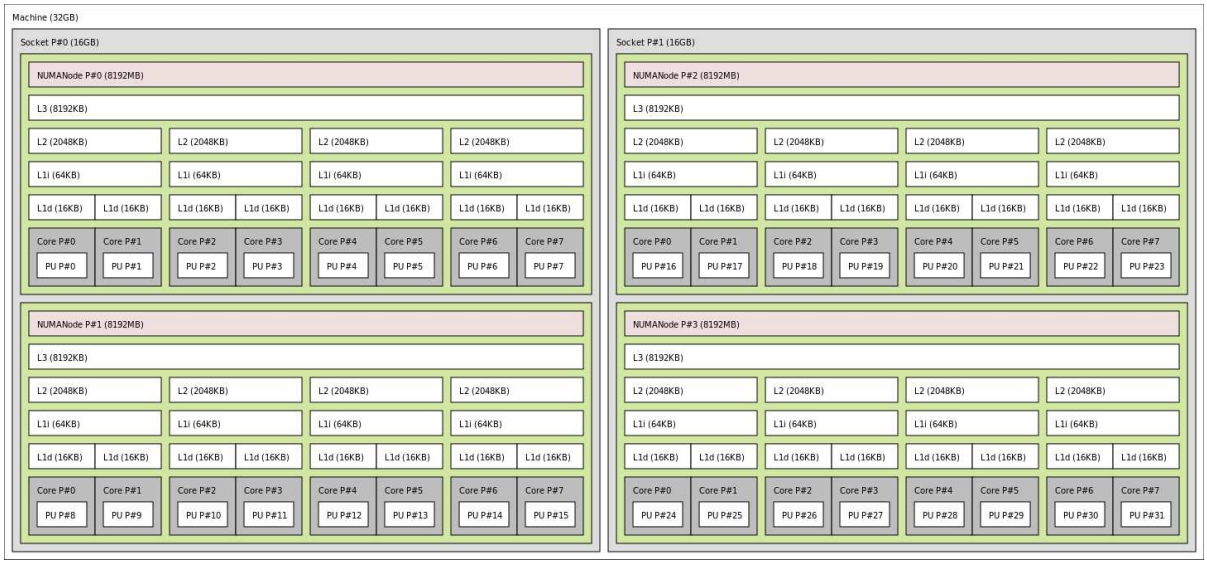
NUMA "NUMA" is an acronym for "Non-Uniform Memory Access". The idea is that the time needed to fetch a piece of data in shared memory depends on where it is located. Distributed shared memory systems are usually also NUMA. Most Linux distributions support NUMA.

SMT SMT is an acronym for Simultaneous Multi-Threading. The idea here is to allow more than one thread to run on the same processor core. Additional functional units, and/or additional processor registers are needed to support multiple threads. Example: SPARC T series processors.

Pipelining Pipelining can be thought of as an "assembly line" style of work. In its most common form, the fetch, decode and execute phases of instruction processing can be overlapped. The SPARC architecture made extensive use of pipelining. A simplified (RISC) instruction set design allowed an instruction to be retired at every clock cycle. This design created the infamous "delay slot" for branching instructions: the instruction following a branch instruction will be executed whether the branch is taken or not, because the instruction is already "in-flight" on the bus at the time the branch instruction is changing the program counter (PC register).

Super-scalar Many modern processors have multiple functional units per core. Some floating point units can be designed to operate in multiple modes. For example the floating point unit could perform one 64-bit floating point multiply, or it could perform two simultaneous 32-bit floating point multiplies. Example: AMD bulldozer module.

AMD Bulldozer Memory Organization



Example Architecture AMD Bulldozer