P6 Technical Glossary

Architectural State. The value of registers, flags and memory as viewed by the programmer.

Associative Memory. A table that is accessed not via an explicit index, but by the data it contains. If no entries of the associative memory match the input data, a "miss" signal is asserted. If any entries of the associative memory match the input data, the associative memory indicates the match, and produces any related data that was stored with that entry. This is often termed Content Addressable Memory.

Branch Prediction. Pipelined machines must fetch the next instruction before they have completely executed the previous instruction. If the previous instruction was a branch, then the next-instruction fetch could have been from the wrong place. Branch prediction is a technique that attempts to infer the proper next instruction address, knowing only the current one, typically using an associative memory called a BTB.

Branch Recovery. When a branch is mispredicted, the speculative state of the machine must be flushed and fetching restarted from the correct target address. We call this activity branch recovery.

BTB. Branch Target Buffer. A small (typically 128-512 entry) associative memory that watches the ICache index and tries to predict which ICache index should be accessed next, based on branch history. Optimizing the actual algorithm used in retaining the history of each entry is an area of ongoing research. P6 uses a variant of Yeh's algorithm (IEEE Micro-24 conference proceedings, 1991.)

CAM. Content Addressable Memory. Synonym for Associative Memory. A table that is accessed not via an explicit index, but by the data it contains. If no entries of the associative memory match the input data, a "miss" signal is asserted. If any entries of the associative memory match the input data, the associative memory indicates the match, and produces any related data that was stored with that entry.

Dynamic Execution. Dynamic Execution involves optimally adjusting instruction execution by prediciting program flow, analysing the program's dataflow graph to choose the best order to execute the instructions, then speculatively executing instructions in the preferred order.

ICache. Instruction Cache. A fast local memory that holds the instructions to be executed. When a program tries to access an instruction that is not yet (or no longer) in the cache, the CPU must wait until hardware fetches the desired instructions from another cache (or memory itself) downstream. These stalls in the fetch/decode unit of the P6 are typically overlapped by the other units that are processing independently.

Instruction Pool. A metaphor used to describe the mechanism used by three independent P6 units to communicate. The pool is implemented as a CAM called the ReOrder Buffer (ROB).

L2 Cache. Caches exist in a "memory hierarchy." There is a small but very fast L1 cache; if that misses, then the access is passed on to the bigger but slower L2 cache, and if that misses, the access goes to main memory (or L3 cache if the system has one).

Pipelining. A microarchitecture design technique that divides the execution of an instruction into sequential steps, using different microarchitectural resources at each step. Pipelined machines have multiple IA instructions executing at the same time, but at different stages in the machine.

RAT. Register Alias Table. Renames programmer visible register references to internal physical registers. This mapping is done at run time.

Reservation Station. A generalized mechanism where uops wait for their dependent component parts. Once a uop has all of its operands, the uop is dispatched to a resource unit for execution.

Resource Unit. The p6 ahs multiple resources that are scheduled by the Reservation Station: they include 2 integer units, a full floating point arithmetic unit, a floating point multiplier, divider, and shifter, and two address generation units.

Retirement. A generalized mechanism that removes a completed uop from the ROB and commits its state to whatever permanent architectural state was designated by the Intel architecture instruction.

ROB. ReOrder Buffer. The P6 functional unit where initial uops wait, speculative results are collected, and then are retired.

Speculative Execution. A generalized mechanism that permits instructions to be started "early," i.e., ahead of their normal execution sequence. Results of this speculation are stored temporarily (in the ROB) since they may be discarded due to a change in program flow.

Store Buffer. A queue that receives write requests from the CPU and sends them to the memory subsystem. This store buffer is snooped by pending loads.

Superscalar. The ability to process more than one instruction per clock. The Pentium processor has two execution pipes (U and V) so it is superscalar level 2. The P6 can dispatch and retire 3 instructions per clock so it is superscalar lever 3.

UOP. A micro operation. The three decoders translate Intel architecture instructions into fixed length uops that are easier to schedule by the dispatch/execute unit. Most IA instructions translate to single uops, some need up to four, and the complex instructions (e.g., Enter, Leave) need microcode support.