Pentium III Instruction Stream

Introduction

Pentium III uses several key features to exploit ILP

This part of our presentation will cover the methods that the third generation P6/IA32 architecture uses and their advantages/disadvantages.

Features

- Completely speculative execution
- superscalar issue
- Speculative register renaming
- Deeply pipelined execution
- Large branch prediction unit

Pentium III Execution

- Deeply Pipelined
  - Over 30 stages for many ops (without miss penalties)
  - Several tradeoffs for deeply pipelined models
    - Stall penalties
    - Clock rate

Pentium III Execution Model

- Consists of
  - In-order front end/issue
  - Out of order execution core
  - In order retirement unit (non-speculative)

Front End Execution

- ICache access
- Branch prediction
- Decode
- Issue

ICache

- Icache is
  - 16KB , 4 way set associative, 32 byte cache lines
- L2 (unified)

Branch Prediction

- BTB (branch target buffer) decides address of next executed instruction
- Speculative state advantages
  - Less complicated recovery
  - Less Mispredict costs
- BTB runs off of prefetch

Branch Prediction (Cont.)

- Dynamic predictor
  - Yeh's algorithm
  - last 4 directions available per branch address
  - One cycle disadvantage on taken branches
  - RSB

Branch Prediction (Cont.)

- Static predictor
  - 6 cycle penalty
  - Forward branches(not taken)
  - Backward branches(taken)

Decode

- Three decode units
  - Two simple, one complex
- Micro ops
  - RISC type operations
    - Can be 1-4 per CISC operation

Decode (Cont.)

- Issue problems arise
  - Program instruction ordering very important
- Tradeoff
  - Issue of 4-wide instructions improves compiler performance by allowing more optimization

Decode (Cont.)

- Williamette (last IA32 architecture) has
  - Execution trace cache
    - Immediately accessible (no cache hit delay)
    - Exploits temporal locality

Execution

- Micro-ops follow distinct trails
  - RAT (register alias table)
  - ROB (re-order buffer)
  - Reservation station
  - Execution units

RAT

- Register Mappings (source, destination)
  - Eliminates false dependencies
    - In-Order Retirement
  - Allows out of order execution from ROB
- Issues up to 3 micro-ops to ROB per cycle
  - See any throughput problems?

RAT (cont.)

- Can access either ROB or RRF
  - Solves true dependencies
  - State bits required
- Branch Mispredicts?
  - Flush all state(mappings) older than branch
  - No new mappings until all current instructions retired

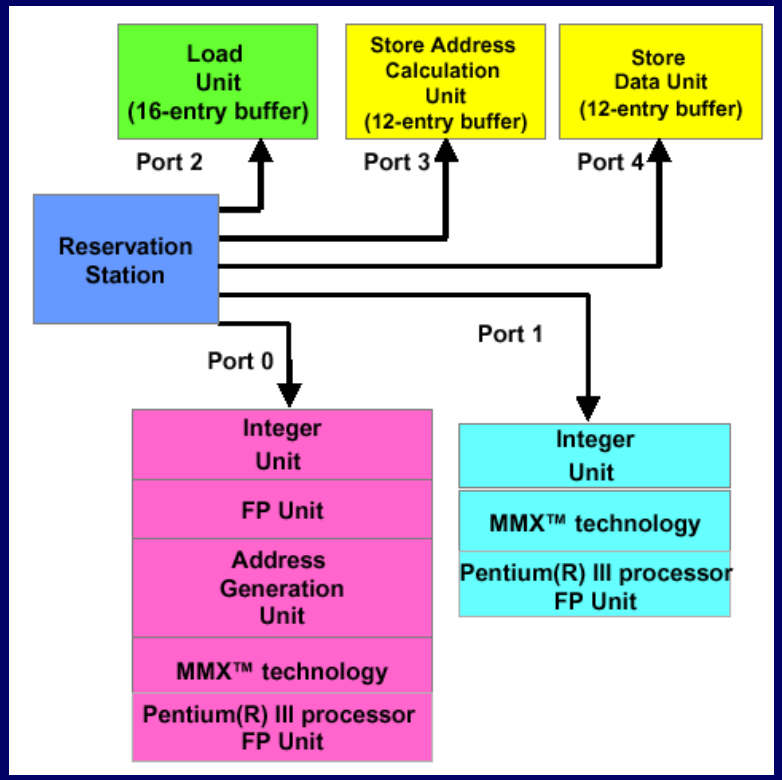
ROB

- ROB is temporary location of queued micro-ops
- 40 entries
  - Contain micro-ops, state, and results

ROB states

- SD
  - Scheduled for execution
- DP
  - Micro-op is at head of dispatch queue
- EX
  - Currently being executed
- WB
  - Completed execution; waiting for results
- RR, RT
  - Ready for retirement, being retired

Reservation Station



Reservation Station (Cont.)

- 5 ports for different ops
  - FP, Int, MMX, SSE, LSQ ops
  - More throughput problems?
- 20 entry queue
  - Organization not specified

Execution

- Scheduling
  - One scheduler for each port
  - 20 entry queue optimized by priority algorithm
- Dispatch
  - All 5 ports can be dispatched every clock cycle

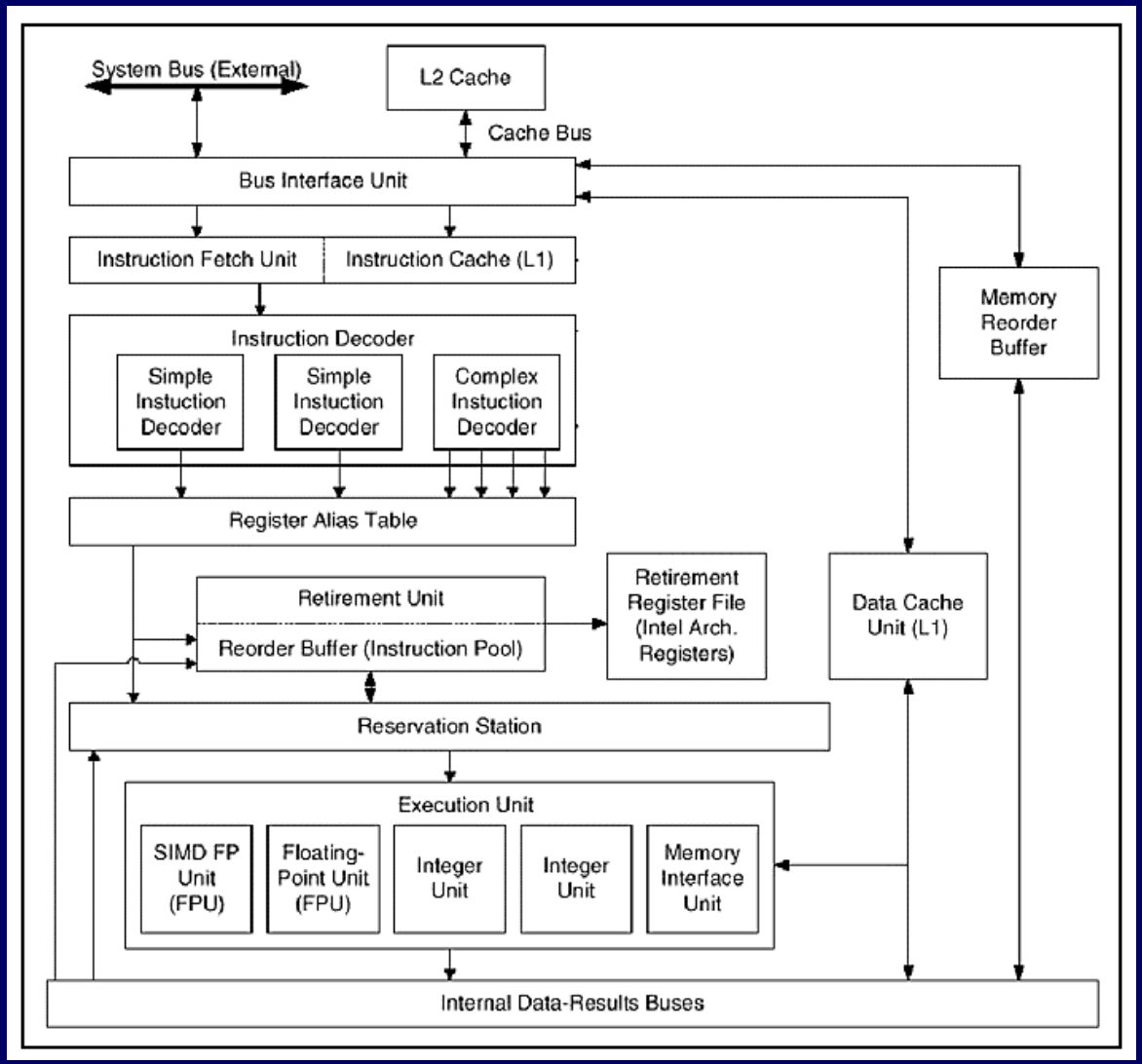
Execution (Cont.)

- Dispatch
  - Dcache misses, hazards resolved
  - Results written back to ROB
    - Resolves dependency chain

Retirement

- Results written to RRF
  - Non-speculative state
  - Register maps deleted, if possible

Throughput


System Bus (External)
L2 Cache
Cache Bus
Bus Interface Unit
Instruction Fetch Unit
Instruction Cache (L1)
Instruction Decoder
Simple Instuction Decoder
Simple Instuction Decoder
Complex Instuction Decoder
Register Alias Table
Retirement Unit
Reorder Buffer (Instruction Pool)
Retirement Register File (Intel Arch. Registers)
Memory Reorder Buffer
Data Cache Unit (L1)
Reservation Station
Execution Unit
SIMD FP Unit (FPU)
Floating-Point Unit (FPU)
Integer Unit
Integer Unit
Memory Interface Unit
Internal Data-Results Buses

Area Considerations

- As it turns out
  - IA32 architecture doesn't scale entirely well
    - Die area a large problem
    - Bus / logical complexity grows in non linear fashion

Finally

- It seems that
  - IA32 is at an end
  - VLIW is next