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 units
- Reservation station
- ROB (re-order buffer)
- RAT (register alias table)

Micro-ops follow distinct trails
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

- Load Unit (16-entry buffer)
- Store Address Calculation Unit (12-entry buffer)
- Store Data Unit (12-entry buffer)

Reserve Station

- Port 0
- Port 1
- Port 2
- Port 3
- Port 4

- Integer Unit
- FP Unit
- Address Generation Unit
- MMX™ technology
- Pentium(R) III processor
  - MMX™ technology
  - Pentium(R) III processor
  - FP Unit
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
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

- ILIW is next
- IAZ2 is at an end