The Future of AMD Processors: Zen

Thomas Guerin & Jason Skrien
Agenda

- AMD’s Market Position

- Bulldozer
  - Architecture
  - Benefits
  - Flaws

- Zen
  - Architecture
  - Enhancement Goals
  - New Memory Hierarchy
  - Scheduling
  - Improved Branch Prediction
  - Micro-op Cache
  - Simultaneous Multi-threading (SMT)
  - Power Savings
  - New Instructions
  - Release Timeline

- Conclusions & The Future of AMD
AMD’s Market Position

- AMD’s market share has declined from nearly 50% in Q1 2006 to a mere 20% in Q1 2016
- In May 2015, Kerrisdale Capital Investment claimed that AMD would be bankrupt by the year 2020
AMD Bulldozer: The Good

- 40-entry scheduler gives good instruction level parallelism
- Inexpensive relative to current Intel products
AMD Bulldozer: The Bad

- Designed to increase frequencies while maintaining IPC of previous architecture
  - Long pipeline resulted in high latency

- Branch Prediction
  - 5% miss rate (13% on 7zip Benchmark)
  - 20 cycle penalty

- 2-way cache associativity is low

- Power inefficient
Bulldozer Architecture

- AMD implemented a 2 core per module scheme
  - Two integer schedulers
  - One floating point scheduler
  - L1 cache shared across “cores”
  - L2 and L3 cache shared across modules
Bulldozer Flaws

- Software scheduling disaster
  - Windows sees the load of eight modules rather than the four cores
  - Tasks may be scheduled to run on free cores of busy modules - not ideal
  - CPU clocked higher in 4M/4C often faster than normal clock in 8C/4M

The consumers look unhappy
Add more cores
Introducing: Zen

- Extracting instruction-level parallelism
- Promises of 40% IPC improvement (28.6% reduction in CPI)
- Improved Instruction Level Parallelism
- Lower power
- New ISA Extension
- Scales from low-power notebooks to servers
- PCI Express 3.0: Rumored 36 lanes
  - Up to 24 Lanes on Intel Kaby Lake
  - Up to 38 lanes of 2.0 on Bulldozer
General Architecture

- Cores grouped per CPU Complex (CCX)
  - 4 cores per CCX
  - Private L1/L2 caches
  - Shared L3 cache
New Memory Hierarchy

- **L1:** (private)
  - 64k 4-way I-Cache
  - 32k 8-way D-Cache
  - Switched to write-back
- **L2:** 512k 8-way (private)
- **L3:** (shared)
  - 8MB per CCX
  - 16-way
  - Holds blocks evicted from L1 & L2
  - No redundant data from L2
  - High associativity: reduce conflicts
  - Claims of 5x L3 bandwidth
Scheduling

- Dual schedulers - one for int, one for floating point
- Integer Rename Space 168 registers
- 6x14 scheduling queues
- Increased size of scheduler register files
  - 160 floating point entries
  - 192 integer entries
Improved Branch Prediction

- Two branches per Branch Target Buffer
- Neural network machine learning methodology
- Strided Sampling Hashed Perceptron Predictor
  - In existing microprocessors
    - Oracle SPARC T4
    - AMD Bobcat APU
    - Samsung Exynos 8890 (Galaxy S7)
  - Keeps a history of instructions
  - Samples bits within that history
  - Training finds correlations between history sample and outcome
Hashed Perceptron Branch Predictor

1. Hash segments of branch history into multiple tables
2. Apply a threshold to the sum of the weights selected by the hash functions. A threshold is applied to predict the branch.
3. Update weights of neural network using “perceptron learning”

\[ y = w_0 + \sum_{i=1}^{n} x_i w_i \]
Hashed Perceptron Methods

Naïve Sampling:

Geometric Sampling:

Strided Sampling:
Micro-op Cache

- x86 is CISC not RISC
- Complex variable-length instructions
- Instructions decoded into micro-ops
- Cache offloads the complex decode hardware
- Decreases power consumption
Simultaneous Multi-threading (SMT)

- Intel® Hyper-Threading™ is an example
- Two threads per physical core
- Keeps the FUs busy, increases IPC
- If one thread is blocked, the other can continue to use the core
Power Savings

- “Aggressive” clock gating
  - Disables unused portions of the circuitry
  - Stop components from switching states

- Changing from 28 nm planar to 14 nm FinFET
  - Improved transconductance
  - Reduces power consumption
  - Reduces required die size
## New Instructions

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADX</td>
<td>(Add extended) Add two unsigned multiprecision integers plus carry</td>
</tr>
<tr>
<td>RDSEED</td>
<td>Complements RDRAND instruction, generates seed for PRNG</td>
</tr>
<tr>
<td>SMAP</td>
<td>Supervisor Mode Access Protection</td>
</tr>
<tr>
<td>SHA1</td>
<td>SHA1 encoding</td>
</tr>
<tr>
<td>SHA256</td>
<td>SHA256 encoding</td>
</tr>
<tr>
<td>CLFLUSHOPT</td>
<td>Cache Line Flush</td>
</tr>
<tr>
<td>XSAVEC</td>
<td>Save Compact</td>
</tr>
<tr>
<td>XSAVES</td>
<td>Save Supervisor</td>
</tr>
<tr>
<td>XRSTORS</td>
<td>Save Restore</td>
</tr>
<tr>
<td><strong>CLZERO</strong></td>
<td>Clear line of cache</td>
</tr>
<tr>
<td><strong>PTE Coalescing</strong></td>
<td>Merge 4K page tables into 32K page tables</td>
</tr>
</tbody>
</table>
Zen Release Timeline

- “Summit Ridge” expected Late 2016/Early 2017
- Low-end 8 cores/16 threads $200 - 300 MSRP
- High-end 8 cores/16 threads ~$500
Conclusions & The Future of AMD

- Hopefully Zen is the release AMD needs to pull it back from the edge of bankruptcy
- #MakeAMDGREAT AGAIN
- AMD processors in the latest generation of game console promise to provide revenue in the meantime
References

- http://www.theregister.co.uk/2016/08/22/samsung_m1_corerelease/