<table>
<thead>
<tr>
<th>Feature</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThunderX2®</td>
<td>32 core die 128 threads Arm®v8.1</td>
</tr>
<tr>
<td></td>
<td>First Arm-based Top 500 System</td>
</tr>
<tr>
<td></td>
<td>First non-x86 CPU in Microsoft Azure</td>
</tr>
<tr>
<td></td>
<td>Most widely deployed Arm-based server processor</td>
</tr>
<tr>
<td></td>
<td>Industry-leading performance at time of release</td>
</tr>
<tr>
<td></td>
<td>Proven production quality solution at scale</td>
</tr>
</tbody>
</table>
ThunderX3™ overview

- Single die: Up to 60 cores
- Dual die: Up to 96 cores
- Arm v8.3 with select v8.4/v8.5 features
- 30% single thread gain at equal frequency over ThunderX2
- Up to four threads per core
- High bandwidth switched ring interconnect
- Up to 8 DDR4-3200 channels
- Single die: 2X-3X perf over ThunderX2 at equal power
  - Further gains from dual die
- Up to 64 PCIe Gen4, 16 PCIe controllers
- Fine grain power monitoring/management
- TSMC 7nm
ThunderX3 core block diagram

Conditional Branch predictors
Indirect predictor
DBTB
RAS
Next PC
L2DBTB
Loop Buffer + Predictor

Indirect predictor
DBTB
8 way L1I
Loop Buffer + Predictor

Fetch

4 micro-ops

8 x Decode

8 micro-ops

Rename and Dispatch

2048 entry L2 TLB

64 KB 8 way L1I

Next Way Predictor

220 Entry ROB

70 Entry Unified Scheduler

512 KB 8 way L2

256 Entry GPR

ALU

ALU

ALU/branch

ALU/branch

LSA

LSA

ST Data

32 KB 8 way L1D

80 Entry LRQ

52 Entry SRQ

256 Entry FPR

NEON

NEON

NEON

NEON

HW pre-fetcher
Core microarchitecture – Fetch

- 64KB Icache, 8 way set associative, 64B line size, next line pre-fetch, way prediction
- Decoupled fetch for large instruction footprint codes
- 8-wide instruction fetch
- Fetch breaks on 64B line boundary, or on a taken branch
- Large condition branch predictors, indirect and return address predictors
- Fetched bundle is decoded 8 instructions at a time
- Decode breaks a few instruction types into multiple micro-ops
Core microarchitecture – Decode/dispatch

- Decoded micro-ops enter skid buffer –
  - Up to 8 micro-ops per cycle
- Each thread has a 32 micro-ops skid buffer – 8 four micro-ops bundles
- 4 micro-ops dispatched per cycle to scheduler
- NOP not dispatched to scheduler – Go to ROB and retire
- Some merging between bundles in skid buffer
Core microarchitecture – Scheduler

- Out of order issue from unified issue queue
  - 70 entries
- Seven issue ports:
  - Port 0: ALU, FP/SIMD
  - Port 1: ALU, FP/SIMD, integer mul/div
  - Port 2: ALU, Branch, FP/SIMD
  - Port 3: ALU, Branch, FP/SIMD
  - Port 4: Ld/St
  - Port 5: Ld/St
  - Port 6: Store data
Core microarchitecture – D-cache / DTLB / L2-cache

- 32KB D-cache, 8-way associative, 64B line size, write back
- Small L1TLBs for zero impact translation in the common case
- 2K entry L2 TLB, 8-way associative
- 512KB L2-cache, 8-way associative – private to core
  - Larger L2-cache increases area and latency with minor incremental performance benefit
- Hardware prefetcher into L2-cache
  - Next line
  - Strides
  - Region
ThunderX3 core performance enhancements over ThunderX2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Approx. pct gain over ThunderX2 (SPECInt)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td></td>
</tr>
<tr>
<td>Icache Size</td>
<td>0.5%</td>
</tr>
<tr>
<td>512KB L2-cache</td>
<td>2.5%</td>
</tr>
<tr>
<td>Larger out-of-order structures</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Width</strong></td>
<td></td>
</tr>
<tr>
<td>Wider decode</td>
<td>2%</td>
</tr>
<tr>
<td>Additional ALU port</td>
<td>1.5%</td>
</tr>
<tr>
<td>Two branches per cycle</td>
<td>1.5%</td>
</tr>
<tr>
<td><strong>Algorithm</strong></td>
<td></td>
</tr>
<tr>
<td>Branch prediction enhancements</td>
<td>3%</td>
</tr>
<tr>
<td>Front end resteer enhancements</td>
<td>1%</td>
</tr>
<tr>
<td>Reduce micro-op expansion</td>
<td>6%</td>
</tr>
<tr>
<td>D-cache bank conflict reduction</td>
<td>0.5%</td>
</tr>
<tr>
<td>Reduce FP structural hazards</td>
<td>1%</td>
</tr>
<tr>
<td>Prefetch enhancements</td>
<td>1.5%</td>
</tr>
<tr>
<td><strong>Latency</strong></td>
<td></td>
</tr>
<tr>
<td>FP latency reduction</td>
<td>0.5%</td>
</tr>
</tbody>
</table>
ThunderX3 performance – Single die
Substantial performance gains
Multithread execution

- Four hardware threads per core
- Each thread includes full copy of Arm architecture state
- Threads share core pipeline resources
- To OS each thread appears as a regular Arm CPU – So four CPUs per core
- Area impact of 4-way SMT relative to no SMT: ~5%
- ThunderX3 has 60 cores / 240 threads per die
Thread arbitration

Goals
- Fair sharing of pipeline resources among threads
- Maximize pipeline utilization

Four points of arbitration:
- **Fetch**: Prioritize threads with fewer instructions in pipeline over threads with more instructions
- **Dispatch**: Similar to Fetch but just considering stages after Dispatch
- **Scheduler (issue)**: Age based priority
- **Retire**: Favor threads with more instructions to retire

Dynamic sharing of caches, branch predictor structures

Thr0 PC  →  ARB  →  FETCH  →  Thr0 ISB  →  ARB  →  DISPATCH  →  SCHEDULER  →  ARB  →  EXECUTE  →  Thr0 CQ  →  ARB  →  RETIRE
Thr1 PC  →  ARB  →  Thr1 ISB  →  ARB  →  DISPATCH  →  SCHEDULER  →  ARB  →  EXECUTE  →  Thr1 CQ  →  ARB  →  RETIRE
Thr2 PC  →  ARB  →  Thr2 ISB  →  ARB  →  DISPATCH  →  SCHEDULER  →  ARB  →  EXECUTE  →  Thr2 CQ  →  ARB  →  RETIRE
Thr3 PC  →  ARB  →  Thr3 ISB  →  ARB  →  DISPATCH  →  SCHEDULER  →  ARB  →  EXECUTE  →  Thr3 CQ  →  ARB  →  RETIRE
Multithread scaling performance – Single core

<table>
<thead>
<tr>
<th>Low IPC (~0.5)</th>
<th>MySQL</th>
<th>1 thread</th>
<th>2 threads</th>
<th>4 threads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relative performance</td>
<td>1.00</td>
<td>1.79</td>
<td>2.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium IPC (~1.25)</th>
<th>Leela</th>
<th>1 thread</th>
<th>2 threads</th>
<th>4 threads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relative performance</td>
<td>1.00</td>
<td>1.38</td>
<td>1.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High IPC (&gt;2)</th>
<th>X264</th>
<th>1 thread</th>
<th>2 threads</th>
<th>4 threads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relative performance</td>
<td>1.00</td>
<td>1.18</td>
<td>1.28</td>
</tr>
</tbody>
</table>
Socket level performance – MySQL

- Roughly linear scaling in Core region
- Scaling flattens out in threaded region, but still good gains
- Net 89x over single thread
L3-Cache and interconnect

- Cores / L3-caches organized as switched rings
- DDR channels, I/O tap into rings
- L3-cache organized as tiles that are cache line striped
  - 1 ½ MB per core
  - No notion of L3 cache affinity to cores
  - Good for shared text and shared data
- Exclusive L3-cache – filled on evict from L2-cache
- Snoop based coherence with snoop filters
  - Single socket and two socket
### ThunderX3 Arm-based server processor summary

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>Up to 3x performance over industry-leading ThunderX2 within same power envelope</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Evolutionary design approach – leverage ThunderX2 platform and proven production quality solution at scale</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Four-way threading provides ~50% performance advantage on data center codes over competitor systems</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>Marvell offers a competitive solution with technology roadmap built on legacy of processor expertise</td>
</tr>
</tbody>
</table>
Thank You