Tech Insights

MICROPROCESSOR report

Insightful Analysis of Processor Technology

CORTEX-X3 POWERS UP

Arm Adds Cortex-A715 and Refreshes Cortex-A510 to Complete Triumvirate

By Linley Gwennap (July 1, 2022)

This week, Arm introduced its Makalu generation of licensable CPU cores, featuring Cortex-X3 as the high-performance model and Cortex-A715 as the midrange one. Taking aim at Apple's M2, the X3 offers double-digit performance gains over its predecessor, even before accounting for any benefit from transistor shrinks. For this upgrade cycle, the A715 focuses on power efficiency, delivering 5% greater throughput while using the same power as the previous generation. Makalu is the first generation that's fully optimized for 64-bit code, dropping compatibility with legacy 32-bit applications.

The Makalu CPUs typically pair with Cortex-A510 in heterogeneous Arm v9 configurations. Although the A510 shipped last year, Arm refreshed it with a few minor changes, including new security features. The updated version can operate 5% faster at the same power or use 5% less power at the same speed. It's configurable with 32-bit compatibility for low-end smartphones and other devices that run old software.

As usual, the new Cortex CPUs target client devices such as smartphones, tablets, and Chromebooks as well as high-performance embedded and IoT processors. They can operate in clusters of up to 12 in any combination—50% larger than before—using an updated version of the DSU-110 cluster controller. Production RTL is already available. We expect the first phones using these designs to ship early next year.

The new CPU cores follow last year's introduction of the Matterhorn generation, featuring Cortex-X2 and Cortex-A710 (see *MPR May 2021*, "Arm v9 Yields Three New CPUs"). As Figure 1 shows, the Matterhorn designs, along with the A510, were the first Arm v9 cores; of the three, only the A710 initially offered 32-bit compatibility. Whereas Arm updates its high-performance and midrange cores every year, it usually skips a few years between low-end cores. Thus, a new low-end CPU wasn't expected this year, so even a few new features are a bonus.

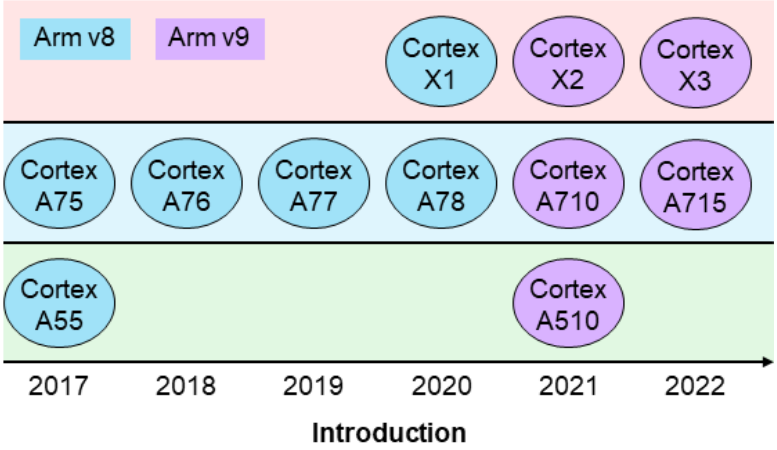

Figure 1. Cortex-CPU timeline. The two newest cores constitute the second Arm v9 generation. They can cluster with Cortex-A510, which is software compatible. These cores typically ship in products the year after their introduction.

Cortex-X3 Bulks Up Branch Predictors

To deliver an 11% gain in per-clock performance (IPC), Arm had to enhance Cortex-X3's front end, execution engine, and memory subsystem to maintain balanced performance. Particularly on large benchmarks, the front end was most often the bottleneck, so the designers made the most improvements in that area.

Like its predecessors, the X3 implements a decoupled front end that can run ahead of the decoders, fetching instructions before they're needed. Branch-prediction accuracy is critical to prefetching useful instructions. The X3 enlarges the branch target buffer (BTB) by 50% to increase the hit rate. The bigger BTB, however, requires an extra cycle of latency, so the new design splits it into a large slower level-two (L2) BTB and a smaller L1 BTB with the same latency as in the X2. An even smaller L0 BTB provides zero-delay predictions. The total BTB size is massive; even L0 has 1,024 entries.

Although the BTB is the primary branch predictor, the X3 enhances other predictors as well. It doubles the size of the return address stack to 32 entries and adds a predictor dedicated to indirect branches, which take the target address from a register and thus are difficult to predict. Arm also improved the branch history table (BHT), which predicts conditional branches, although (like most CPU designers) it withheld prediction-algorithm details.

The result of these changes is 12% less taken-branch latency relative to the previous generation, as Figure 2 shows. This improvement reduces total front-end stalls by 3%. The number of mispredicted branches drops 6% and is now about 3 per 1,000 instructions.

In Cortex-A77, Arm added an L0 cache for decoded instructions and later enlarged it to 3,072 entries. In an unusual move, the X3 cuts the L0 cache size to 1,536 entries. The company found that although each hit in the L0 saved time and power, the hit rate was below 25% on many workloads, particularly large programs. Instead of storing every decoded instruction into the L0, the new design stores only instructions it predicts will be needed again. This approach reduces the power necessary to update the L0 cache while achieving a similar hit rate even with half the entries.

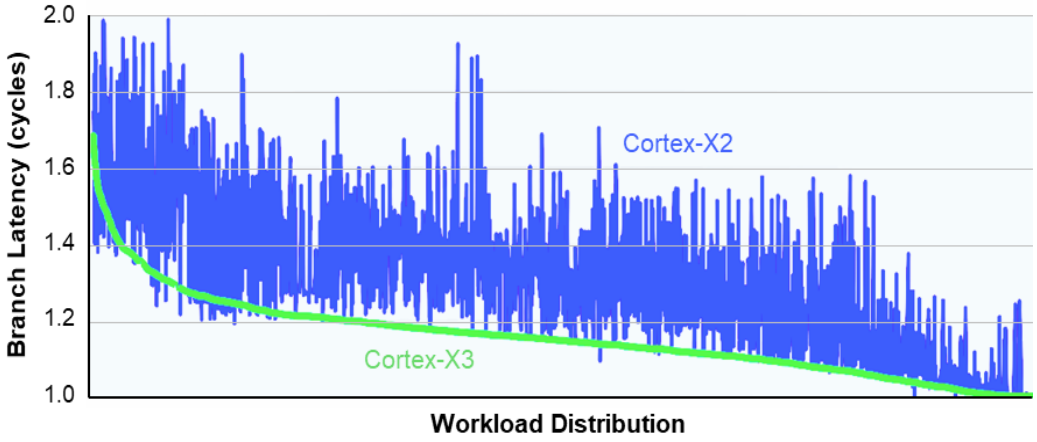

Figure 2. Average latency of predicted-taken branches. Across hundreds of test kernels (horizontal axis), Cortex-X3 reduces latency by an average of 12%. The X2's latency appears to vary more, but only because the results are sorted by the X3's latency. (Data source: Arm)

Fetch and Decode Are Now Six Wide

Instructions from the current program address (PC) are taken from the L0 cache if possible. On an L0 miss, instructions instead flow from the prefetched instruction buffer to the decoders. The X3 adds a sixth instruction decoder to increase throughput; this change benefits large workloads that don't fit into the L0 cache. In either case, the decoded instructions (macro-ops, or Mops) have their registers mapped to the physical register file. The X3 can dispatch up to eight Mops per cycle to the two instruction queues.

The queues issue instructions to the execution units when their operands are available, regardless of program order. To increase throughput, the X3 adds two integer units (ALUs). The remaining execution units are the same as in the previous generation. For example, the CPU includes four 128-bit-wide floating-point units: two that perform any operation and two that perform only multiply and addition. These units can process both Neon and SVE2 SIMD operations. As in the X2, two address-generation units (AGUs) process load and store instructions while a third handles only loads, as Figure 3 shows.

After instructions execute, the reorder buffer (ROB) holds their results until all previous instructions (in program order) are complete. The ROB has 320 entries—11% more than in the X2—and each can hold one complex instruction or two simple instructions. On average, the ROB can hold about 450 in-flight instructions, enabling greater parallelism and allowing additional cache misses without stalling the CPU.

Although the cache sizes and structures remain the same, the X3 boosts hit rates through replacement-policy changes. For example, Arm increased the accuracy of the algorithm that predicts which cache lines to evict from the L2 cache. The design also improves data prefetching, which raises hit rates by loading data before it's needed. The X3 features a dozen prefetch engines, two more than the X2. Each looks for a different access pattern and begins prefetching as soon as it recognizes that pattern. One new engine looks for sequences of indirect loads while the other seeks three-dimensional (spatial) patterns.

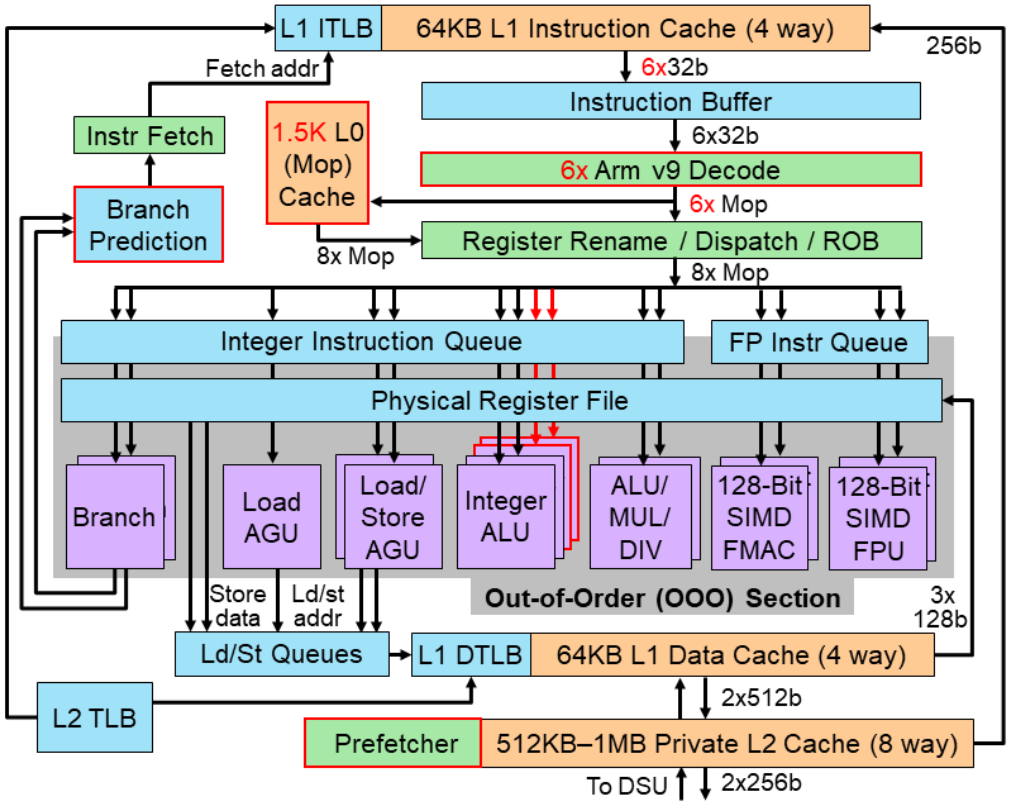

Figure 3. Cortex-X3 microarchitecture. Changes from Cortex-X2 (shown in red) include a wider front end, a smaller L0 cache, and two more integer units.

The X3 supports a private L2 cache of either 512KB or 1MB. This cache consumes a lot of die area, so most customers choose 512KB to reduce cost. Few workloads see much of a performance gain with the larger L2, but a small number gain 10–20% or more. Thus, the 1MB L2 makes sense for devices that frequently run applications with large data sets. But even for smaller applications, the 1MB cache reduces power by limiting transfers from the L3 cache. Premium smartphones may therefore implement the larger cache to extend battery life.

Cortex-A715 Wipes Out Mop Cache

Cortex-X1 derived from Cortex-A77, and the high-end and midrange cores continue to closely parallel each other. Although the X-team is in Texas and the Cortex-A715 designers are from France, the design teams share branch-prediction and prefetching advances. But the A715 must fit in a smaller area and power budget while still delivering strong performance.

The most significant changes relative to Cortex-A710 affect instruction decoding. In previous designs, the decoders can “fuse” certain two-instruction sequences into a single macro-op to save ROB entries and issue slots. The A715 adds a pre-decoder that performs fusion before storing instructions in the cache, simplifying the decode phase.

The A710 was Arm’s final CPU with mandatory 32-bit (Aarch32) compliance, so the A715’s decoders perform only 64-bit (Aarch64) instructions. This change reduces the decoder size by 75%, according to the company. The A715 reuses some of this area by enabling all decoders to handle complex instructions, rather than just two in the A710; the result is greater perfor-

mance on code with Neon and SVE instructions. Taking advantage of their svelte dimensions, the new design adds a fifth decoder as well.

Whereas the X3 halves its L0 cache (also called the Mop cache), the A715 eliminates it completely. The wider front end lets the CPU maintain the same issue rate (five instructions per cycle) as the L0 cache, and eliminating the cache saves considerable area, although power rises because extra decoding is necessary.

Better Branch Prediction

The A715 shares many of the X3's branch-prediction capabilities while implementing them in smaller, faster structures. It also moves to a three-level BTB while adopting better prediction algorithms. Like the X3, the midrange design can now predict two branches per cycle. It doubles the BHT size relative to the A710.

The A715 maintains the same set of execution units as its predecessor, as Figure 4 shows. Relative to the X3, it has two fewer integer ALUs and half as many SIMD units, which are bulky. As in the X3, each 128-bit SIMD unit can execute Neon or SVE2 instructions. The reorder buffer (ROB) grows 11% to 192 entries, expanding the out-of-order window.

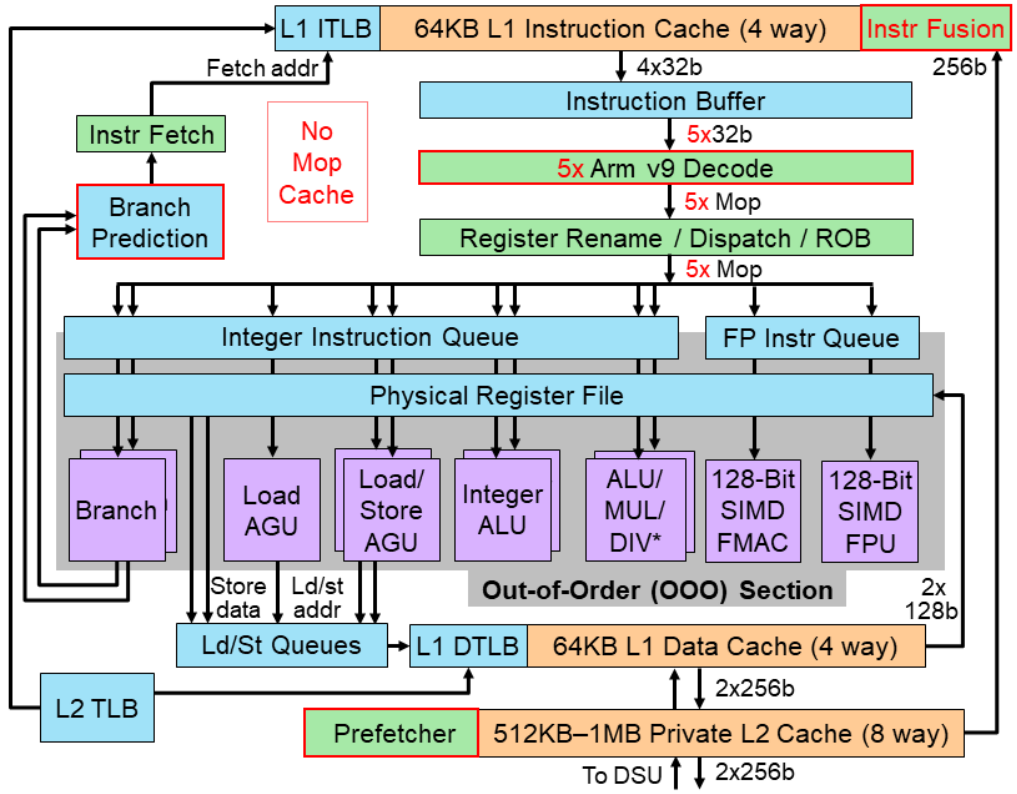

Figure 4. Cortex-A715 microarchitecture. Relative to Cortex-A710, the new design adds instruction fusion before the instruction cache, widens the decoder to five instructions, and eliminates the L0 (Mop) cache. Changes appear in red. *Only one of the two units handles division.

The new design doubles the number of data-cache banks, allowing it to handle more read and write transactions at the same time. This change reduces access conflicts, boosting performance and saving power. The A715 also

enlarges the load queue, increasing parallelism. It raises the number of L2 TLB entries by 50% to 1,536; in addition, each entry can now store two pages if they're contiguous, further expanding the translation capacity.

Cortex-X3 Raises IPC and Clock Speed

Maintaining the same clock frequency and configuration (as Arm says, "iso-everything"), Cortex-X3 is 11% faster than Cortex-X2 on the SPECint_2006 benchmark. This rating is based on gate-level emulation using a large FPGA system. The IPC gain is slightly lower on other common benchmarks: 10% on Geekbench and 8% on the larger SPECint_2017. Across a broad suite of test kernels, the average gain is about 9%.

The CPU pipeline remains the same as in the X2, and Arm expects the new design to achieve the same clock speeds: up to 3.3GHz in smartphones and up to 3.6GHz in laptops owing to their greater thermal capacity. Speeds could be slightly faster in next-generation 3nm technology. Arm withheld the new design's relative power, but a conceptual power/performance chart showed that the X3 requires more power to attain its greater IPC. We expect little power-efficiency improvement at peak performance.

For Cortex-A715, the picture is much simpler. The new core runs at the same clock speed and power as the previous generation, since the added features offset the power savings from eliminating Aarch32 compatibility. But these features boost IPC by 5%, yielding slightly greater performance at every point on the power curve.

Arm withheld the area of the new cores. We expect the A715 to be about the same size as the A710, with the big savings in decoder size and the Mop-cache removal offsetting the added features. The X3 will continue the growth trend in Arm's high-end cores; we expect about a 10% expansion, slightly less than the IPC boost. Cortex-X2 delivers less performance per unit area than its predecessor, so we expect little improvement in this metric.

Apple Hasn't Fallen

Cortex-X3 offers sizable improvements over the previous generation in an attempt to catch up to the world's fastest shipping Arm-compatible CPU: the Avalanche core in Apple's A15 and M1 processors (see *MPR Oct 2021*, "Apple A15 Extends Battery Life"). Although that company doesn't license its CPU cores, it competes against Arm's customers' customers.

The shipping Cortex-X2 delivers 22% less performance than Avalanche on the Geekbench 5 single-core test, as Figure 5 shows. Applying Arm's projection of 10% greater IPC on that benchmark, Cortex-X3 will reduce the gap. But Apple has already announced an upgraded Arm CPU in its soon-to-ship M2 processor that's 12% faster than Avalanche, thus maintaining the same advantage over the X3. Although Geekbench isn't the best benchmark for CPUs, it's representative of many smartphone apps and is widely quoted.

Cortex-X CPUs also appear in PCs, where they compete against both Apple's M-series and various x86 cores. Intel's Golden Cove CPU, which serves in the Core i9-12900KS, holds the Geekbench 5 performance record (see *MPR Sep 2021*, "Golden Cove Adds Matrix Units"). At its peak speed of 5.5GHz, however, that processor carries a 150W TDP rating, far more than any Arm processor. In the 28W Core i5-1250P, a typical laptop processor, the x86 CPU is limited to 4.4GHz; at this speed, it trails Avalanche but still leads the projected Cortex-X3. Although the X3 undoubtedly requires much less

area and power than Golden Cove, it must match that CPU's performance if it's to gain PC market share.

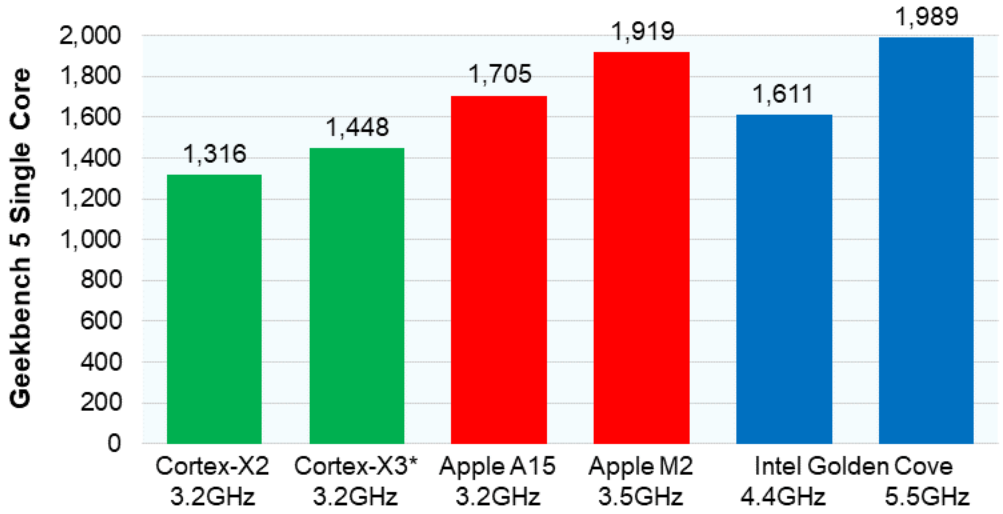

Figure 5. CPU-performance comparison. Cortex-X3 will reduce the gap relative to custom Apple and Intel CPUs but won't catch up. (Source: third-party testing, except *TechInsights estimate based on Arm projection)

Arm intends to continue boosting performance in future CPUs. After Makalu, the next generation is code-named Hunter, targeting processors for early-2024 shipment. At the same time, the company plans to replace Cortex-A510 with a new model code-named Hayes, representing an unusually short lifetime for one of its low-end cores. Continuing the mountain theme, the following year should see the Chaberton generation, which updates the high-performance and midrange lines while continuing to pair with Hayes at the low end. The company withheld performance targets and other information about these upcoming CPUs.

Almost Everyone Approves

Cortex-X3 offers a substantial performance upgrade over Cortex-X2, delivering about 10% better IPC across leading benchmarks and faster clock speeds to boot. To achieve these gains, the new CPU implements a wider front end, additional execution units, and a larger reorder window. Because a single misprediction can jettison hundreds of partially completed instructions, Arm continues to invest considerable design effort and die area into improving branch prediction, which is particularly important for leading-edge smartphone apps as well as PC and server software. These performance gains come at the cost of power and area increases, although customers can dial down the clock speed to save power.

Cortex-A715 stays within its predecessor's power budget while offering a 5% performance gain. This efficiency gain is aided by discarding compatibility with old 32-bit code. Arm has been warning developers for years of this impending change, so we see little downside. Customers that still want 32-bit compatibility can implement the feature in Cortex-A510, which provides plenty of performance for ancient apps.

The new CPU designs should satisfy most Arm customers, but Qualcomm appears unhappy with Cortex-X performance, as it plans to move from Cortex-

X3 to an in-house CPU from its Nuvia team. Even though Arm is boosting the performance of its high-end CPUs, they continue to lag Apple's, creating a problem for Qualcomm customers whose products must compete against iPhones. If this move succeeds, the two remaining Cortex-X licensees, MediaTek and Samsung, will have difficulty competing against not only Apple but also Qualcomm in flagship phones and other premium devices.

Arm can afford to lose a bit of Qualcomm's business (the latter company will continue to use Cortex-A CPUs), but the design loss is a red flag. Makalu provides a strong upgrade for mainstream Cortex licensees, but competing against Apple remains a challenge. ♦

Price and Availability

Arm has released production RTL for Cortex-X3 and Cortex-A715, along with the revised Cortex-A510 and DSU-110, to lead customers. The first chips using these cores are expected to ship in early 2023. All are available for general licensing except Cortex-X3, whose availability is restricted. Arm withheld license fees. For more information, access the company's *blog post* or its *A715 web page*.