For COMP60001/COMP70086 Advanced Computer Architecture (Imperial College London 2024)
This article is provided for the final examination, with the question paper.

ANANDTECH

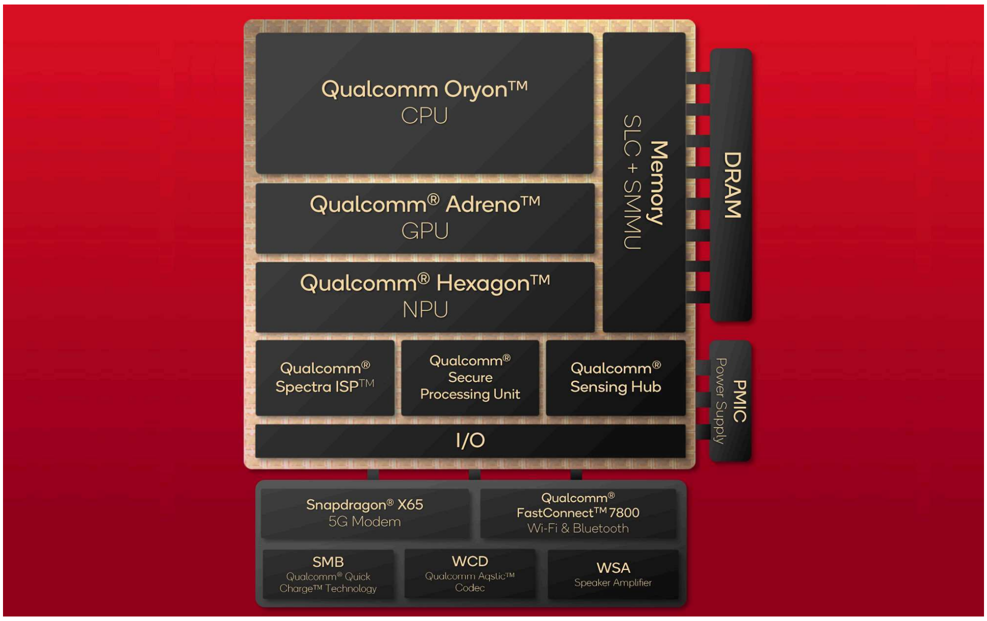
Original Link: https://www.anandtech.com/show/21445/qualcomm-snapdragon-x-architecture-deep-dive

The Qualcomm Snapdragon X Architecture Deep Dive: Getting To Know Oryon and Adreno X1

52
Comments

by Ryan Smith on June 13, 2024 9:00 AM EST

Posted in CPUs Qualcomm Adreno GPUs NUVIA Snapdragon X Elite Oryon Snapdragon X Snapdragon X Plus Adreno X1



The curtains are drawn and it's almost showtime for Qualcomm and its Snapdragon X SoC team. After first detailing the SoC nearly 8 months ago at the company's most recent Snapdragon Summit, and making numerous performance disclosures in the intervening months, the Snapdragon X Elite and Snapdragon X Plus launch is nearly upon us. The chips have already shipped to Qualcomm's laptop partners, and the first laptops are set to ship next week.

In the last 8 months Qualcomm has made a lot of interesting claims for their high-performance Windows-on-Arm SoC – many of which will be put to the test in the coming weeks. But beyond all the performance claims and bluster amidst what is shaping up to be a highly competitive environment for PC CPUs, there's an even more fundamental question about the Snapdragon X that we've been dying to get to: *how does it work?*

Ahead of next week's launch, then, we're finally getting the answer to that, as today Qualcomm is releasing their long-awaited architectural disclosure on the Snapdragon X SoC. This includes not only their new, custom Arm v8 "Oryon" CPU core, but also technical disclosures on their Adreno GPU, and the Hexagon NPU that backs their heavily-promoted AI capabilities. The company has made it clear in the past that the Snapdragon X is a serious, top-priority effort for the company – that they're not just slapping together a Windows SoC from their existing IP blocks and calling it a day – so there's a great deal of novel technology within the SoC.

And while we're excited to look at it all, we'll also be the first to admit that we're the most excited to finally get to take a deep dive on Oryon, Qualcomm's custom-built Arm CPU cores. The first new high-performance CPU design created from scratch in the last several years, the significance of Oryon cannot be overstated. Besides providing the basis of a new generation of Windows-on-Arm SoCs that Qualcomm hopes will vault them into contention in the Windows PC marketplace, Oryon will also be the basis of Qualcomm's traditional Snapdragon mobile handset and tablet SoCs going forward.

So a great deal of the company's hardware over the next few years is riding on this CPU architecture – and if all goes according to plan, there will be many more generations of Oryon to follow. One way or another, it's going to set Qualcomm apart from its competitors in both the PC and mobile spaces, as it means Qualcomm is moving on from Arm's reference designs, which by their very nature are accessible Qualcomm's competition as well.

So without further ado, let's dive in to Qualcomm's Snapdragon X SoC architecture.

Setting The Stage: Elite, Plus, & Currently Announced SKUs

As a quick refresher, Qualcomm has announced 4 Snapdragon X SKUs thus far, all of which have been made available to device manufacturers for next week's launch.

Qualcomm Snapdragon X (Gen 1) Processors

AnandTech	CPU Cores	All Core Max Turbo	Two Core Max Turbo	GPU TFLOPS	NPU TOPS	Total Cache (MB)	Memory
			Snapdragon X Elite				
X1E-84-100	12	3.8 GHz	4.2 GHz	4.6	45	42	LPDDR5X-8448
X1E-80-100	12	3.4 GHz	4.0 GHz	3.8	45	42	LPDDR5X-8448
X1E-78-100	12	3.4 GHz	3.4 GHz	3.8	45	42	LPDDR5X-8448
			Snapdragon X Plus				
X1P-64-100	10	3.4 GHz	3.4 GHz	3.8	45	42	LPDDR5X-8448

Three of these are "Elite" SKUs, which are defined by their inclusion of 12 CPU cores. Meanwhile Qualcomm is offering a single "Plus" SKU (thus far), which cuts that down to 10 CPU cores.

Officially, Qualcomm isn't assigning any TDP ratings to these chip SKUs, as, in principle, any given SKU can be used across the entire spectrum of power levels. Need to fit in a top-tier chip in a fanless laptop? Just turn down the TDP to match your power/cooling capabilities. That said, to hit the highest clockspeed and performance targets of Qualcomm's chips, a good bit of cooling and power delivery are required. And to that end we aren't likely to see X1E-84-100 show up in fanless devices, for example, as its higher clockspeeds would largely be wasted by a lack of thermal headroom. This won't stop lower-performance chips from being used in bigger devices as budget options, but the SKU table can also be considered as being roughly sorted by TDP.

And while not the subject of today's disclosure, don't be surprised to see further Snapdragon X chip SKUs further down the line. It's become a poorly kept secret that Qualcomm has at least one further Snapdragon X die in development – a smaller die with presumably fewer CPU and GPU cores – which we expect would make up a more budget-focused set of SKUs farther down the line. But for now, Qualcomm is starting with their big silicon, and consequently their highest-performing options.

Even though the first Snapdragon X devices won't reach consumers until next week, it's already clear that, judging by OEM adoption, this is going to be Qualcomm's most successful Windows-on-Arm SoC to date. The difference in adoption compared to the Snapdragon 8cx Gen 3 is practically night and day; Qualcomm's PC partners have already developed over a dozen laptop models using the new chips, whereas the last 8cx could be found in all of two designs. So with Microsoft, Dell, HP, Lenovo, and others all producing Snapdragon X laptops, the Snapdragon X ecosystem is starting off much stronger than any Windows-on-Arm offering before it.

Snapdragon Compute (Windows-on-Arm) Silicon

AnandTech	Snapdragon X Elite	Snapdragon 8cx Gen 3	Snapdragon 8cx Gen 2	Snapdragon 8cx Gen 1
Prime Cores	12x Oryon 3.80 GHz 2C Turbo: 4.2GHz	4x C-X1 3.00 GHz	4 x C-A76 3.15 GHz	4 x C-A76 2.84 GHz
Efficiency Cores	*N/A*	4x C-A78 2.40 GHz	4 x C-A55 1.80 GHz	4 x C-A55 1.80 GHz
GPU	Adreno X1	Adreno 8cx Gen 3	Adreno 690	Adreno 680
NPU	Hexagon 45 TOPS (INT8)	Hexagon 8cx Gen 3 15 TOPS	Hexagon 690 9 TOPS	Hexagon 690 9 TOPS
Memory	8 x 16-bit LPDDR5x-8448 135GB/sec	8 x 16-bit LPDDR4x-4266 68.3 GB/sec	8 x 16-bit LPDDR4x-4266 68.3 GB/sec	8 x 16-bit LPDDR4x-4266 68.3 GB/sec
Wi-Fi	Wi-FI 7 + BE 5.4 (Discrete)	Wi-Fi 6E + BT 5.1	Wi-Fi 6 + BT 5.1	Wi-Fi 5 + BT 5.0
Modem	Snapdragon X65 (Discrete)	Snapdragon X55/X62/X65 (Discrete)	Snapdragon X55/X24 (Discrete)	Snapdragon X24 (Discrete)
Process	TSMC N4	Samsung 5LPE	TSMC N7	TSMC N7

A big part of that, no doubt, comes down to the strength of Qualcomm's architecture. The Snapdragon X packs what Qualcomm promotes as a vastly more powerful CPU than the Cortex-X1 core found on the most recent (circa 2022) 8cx chip, and it's being built on a highly competitive process with TSMC's N4 node. So if all of the stars are properly aligned, the Snapdragon X chips should be a massive step up for Qualcomm.

Meanwhile, there are two other pillars that are helping to hold up this launch. The first, of course, is AI, with the Snapdragon X being the first Copilot+ capable SoC for use with Windows. Requiring a 40+ TOPS NPU, the 45 TOPS Hexagon NPU in the Snapdragon X makes the SoC the first such chip to offer this much performance for neural network and other model inference. The second pillar, in turn, is power. Qualcomm is promising nothing short of amazing battery runtimes with their SoC, leveraging their years of experience producing mobile SoCs. And if they can deliver on it while also hitting their performance goals – allowing users to have their cake and eat it, too – then it will setup the Snapdragon X chips and the resulting laptops nicely.

Ultimately, Qualcomm is looking for their Apple Silicon moment – a repeat of the performance and battery life gains that Apple reaped when switching from Intel's x86 chips to their own custom Arm-based Apple Silicon. And partner Microsoft, for their part, really, **really** wants a MacBook Air competitor in the PC ecosystem. It's a tall order, not the least of which is because neither Intel or AMD have been sitting still over the past few years, but it's not out of reach.

With that said, Qualcomm and the Windows-on-Arm ecosystem do face some obstacles that means Snapdragon X's launch trajectory can never quite match Apple's. Besides the obvious lack of a single unified party developing the hardware and software ecosystem (and all but shoving developers forward to produce software for it), Windows comes with the expectations of backwards compatibility and the legacy baggage that entails. Microsoft, for its part, has continued to work on their x86/x64 emulation layer, which now goes by the name Prism, and the Snapdragon X launch will be the first time it really gets put to the test. But even with years of Arm support within Windows, the software ecosystem is still slowly taking shape, so Snapdragon X will be more reliant on x86 emulation than Apple ever was. Windows and macOS are very distinct operating systems, both in terms of their histories and their owners' development philosophies, and this is going to be especially apparent in the first years of the Snapdragon X's lifetime.

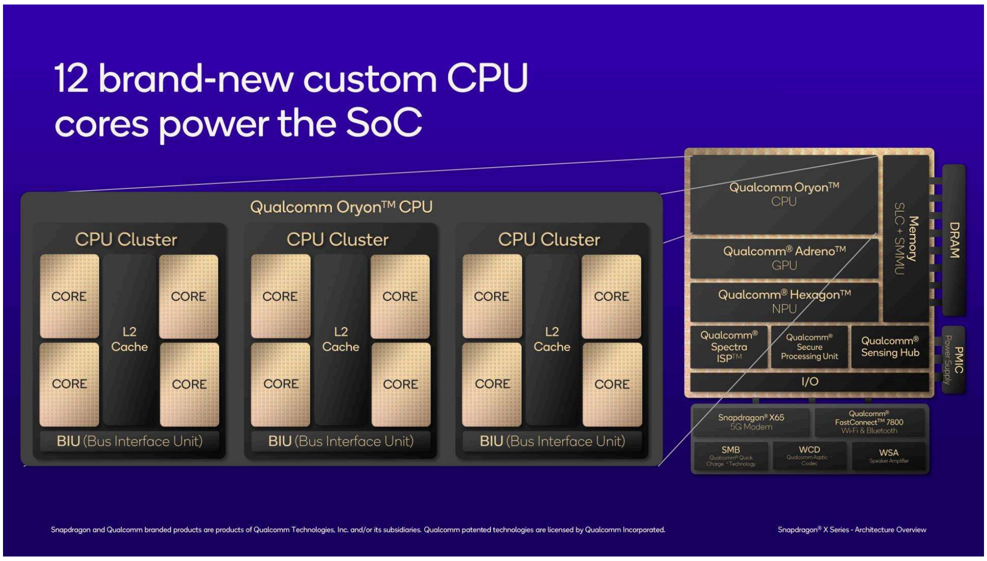
Oryon CPU Architecture: One Well-Engineered Core For All

For our architectural deep dive, we'll start with the star of the show: the Oryon CPU core.

As a quick refresher, Oryon is essentially a third-party acquisition by Qualcomm. The CPU core began life as "Phoenix", and was being developed by the chip startup NUVIA. Comprised of numerous ex-Apple staffers and other industry veterans, NUVIA's initial plan was to develop a new server CPU core, the likes of which would compete with the cores in modern Xeon, EPYC, and Arm Neoverse V CPUs.

However, seizing the opportunity to acquire a talented CPU development team, Qualcomm purchased NUVIA in 2021. And Phoenix was re-tasked for use in consumer hardware, reborn as the Oryon CPU core.

And while Qualcomm isn't focusing too much on Oryon's roots, it's clear that the first-generation architecture – employing Arm's v8.7-A ISA – is still deeply rooted in those initial Phoenix designs. Phoenix itself was already intended to be scalable and power efficient, so this is not by any means a bad thing for Qualcomm. But it does mean that there are a number of client-focused core design changes which didn't make it into the initial Oryon design, and that we should expect to see in future generations of the CPU architecture.



12 brand-new custom CPU cores power the SoC

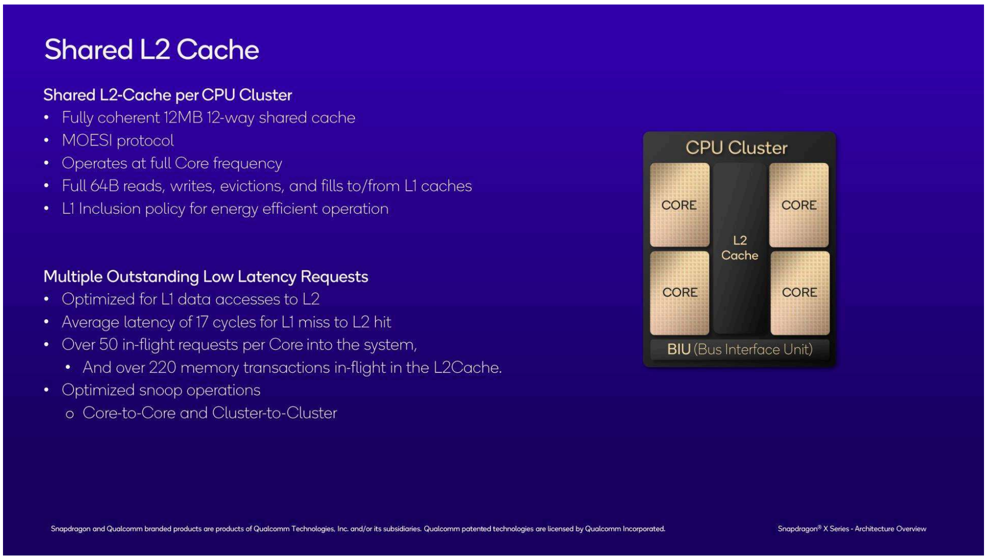
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

Diving in, as previously disclosed by Qualcomm, the Snapdragon X uses three clusters of Oryon CPU cores. At a high level, Oryon is designed to be a full-scale CPU core, capable of delivering both energy efficiency and performance. And to that end, it's the only CPU core that Qualcomm needs; there aren't separate performance-optimized and efficiency-optimized cores like there are on Qualcomm's previous Snapdragon 8cx chips, or Intel/AMD's most recent mobile chips, for that matter.

As far as Qualcomm is disclosing, all of the clusters are equal as well. So there isn't an "efficiency" cluster that's tuned for power efficiency over clockspeeds, for example. Still, only 2 CPU cores (in different clusters) can hit any given SKU's top turbo boost speeds; the rest of the cores top out at the chip's all-core turbo.

Each cluster, in turn, has its own PLL, so each cluster can be individually clocked and powered on. In practice this means that two of the clusters can be put to sleep during light workloads, and then roused from their sleep when more performance is needed.



Shared L2 Cache

Shared L2-Cache per CPU Cluster

- Fully coherent 12MB 12-way shared cache
- MOESI protocol
- Operates at full Core frequency
- Full 64B reads, writes, evictions, and fills to/from L1 caches
- L1 Inclusion policy for energy efficient operation

Multiple Outstanding Low Latency Requests

- Optimized for L1 data accesses to L2
- Average latency of 17 cycles for L1 miss to L2 hit
- Over 50 in-flight requests per Core into the system,
  - And over 220 memory transactions in-flight in the L2Cache.
- Optimized snoop operations
  - o Core-to-Core and Cluster-to-Cluster

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

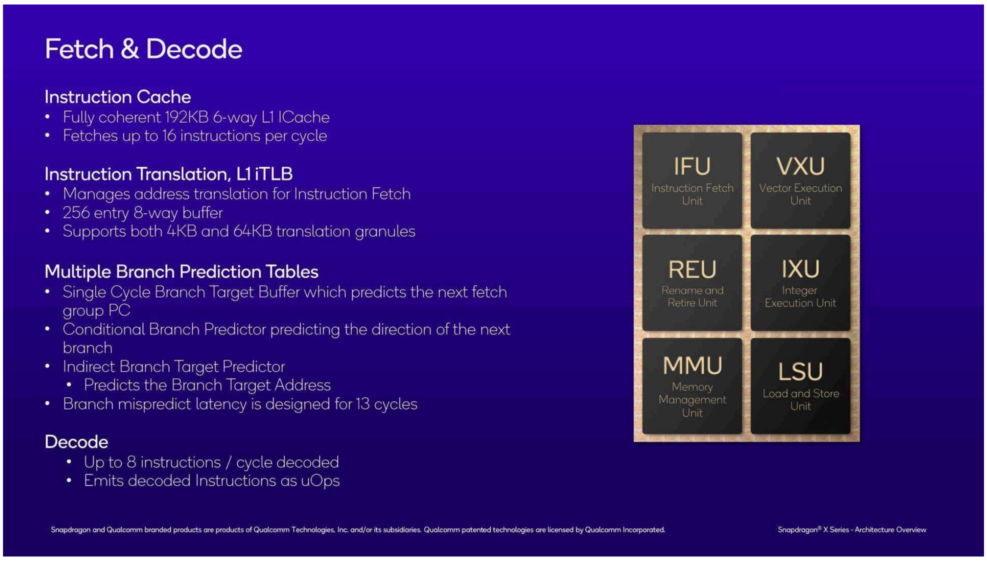
Snapdragon® X Series - Architecture Overview

Unlike most CPU designs, Qualcomm is going with a slightly flatter cache hierarchy for Snapdragon X and the Oryon CPU core clusters. Rather than having a per-core L2 cache, the L2 cache is shared per 4 cores (this being very similar to how Intel shares the L2 cache on its E-core clusters). And this is a rather huge L2 cache, as well, at 12MB in size. The L2 cache is 12-way associative, and even with its large size, there's only a 17 cycle latency to access the L2 cache after an L1 miss.

This is an inclusive cache design, so it contains a mirror of what's in the L1 cache as well. According to Qualcomm they're using an inclusive cache for energy efficiency reasons; an inclusive cache means that eviction is much simpler, as L1 data doesn't need to be moved to L2 to be evicted (or removed from L2 when being promoted to L1). Cache coherency, in turn, is maintained using the MOESI protocol.

The L2 cache itself runs at the full core frequency. L1/L2 cache operations, in turn, are full 64 byte operations, which amounts to hundreds of gigabytes per second of bandwidth between the cache and CPU cores. And while the L2 cache is mostly in place to service its own, directly-attached CPU cores, Qualcomm has implemented optimized cluster-to-cluster snooping operations as well, for when one cluster needs to read out of another.

Interestingly, the Snapdragon X's 4 core cluster configuration is not even as big as an Oryon CPU cluster can go. According to Qualcomm's engineers, the cluster design actually has all the accommodations and bandwidth to handle an 8 core configuration, no doubt harking back to its roots as a server processor. In the case of a consumer processor, multiple smaller clusters offers more granularity for power management and as a better fundamental building block for making lower-end chips (e.g. Snapdragon mobile SoCs). But it will come with some trade-offs, with slower core-to-core communication when those cores are in separate clusters (and thus having to go over the bus interface unit to reach another core). It's a small but notable distinction, since both Intel and AMD's current designs place 6 to 8 CPU performance cores inside the same cluster/CCX/ring.



Fetch & Decode

Instruction Cache
- Fully coherent 192KB 6-way L1 ICache
- Fetches up to 16 instructions per cycle

Instruction Translation, L1 iTLB
- Manages address translation for Instruction Fetch
- 256 entry 8-way buffer
- Supports both 4KB and 64KB translation granules

Multiple Branch Prediction Tables
- Single Cycle Branch Target Buffer which predicts the next fetch group PC
- Conditional Branch Predictor predicting the direction of the next branch
- Indirect Branch Target Predictor
  - Predicts the Branch Target Address
- Branch mispredict latency is designed for 13 cycles

Decode
- Up to 8 instructions / cycle decoded
- Emits decoded Instructions as uOps

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

Diving into an individual Oryon CPU core, we quickly see why Qualcomm has gone with a shared L2 cache: the L1 instruction cache in a single core is already massive. Oryon ships with a 192KB L1 I-Cache, three-times the size of the Redwood Cove (Meteor Lake) L1 I-Cache, and even larger still than Zen 4's. Overall, the 6-way associative cache allows Oryon to keep a lot of instructions very local to the CPU's execution units. Though unfortunately, we don't have the L1I latency on-hand to see how it compares to other chips.

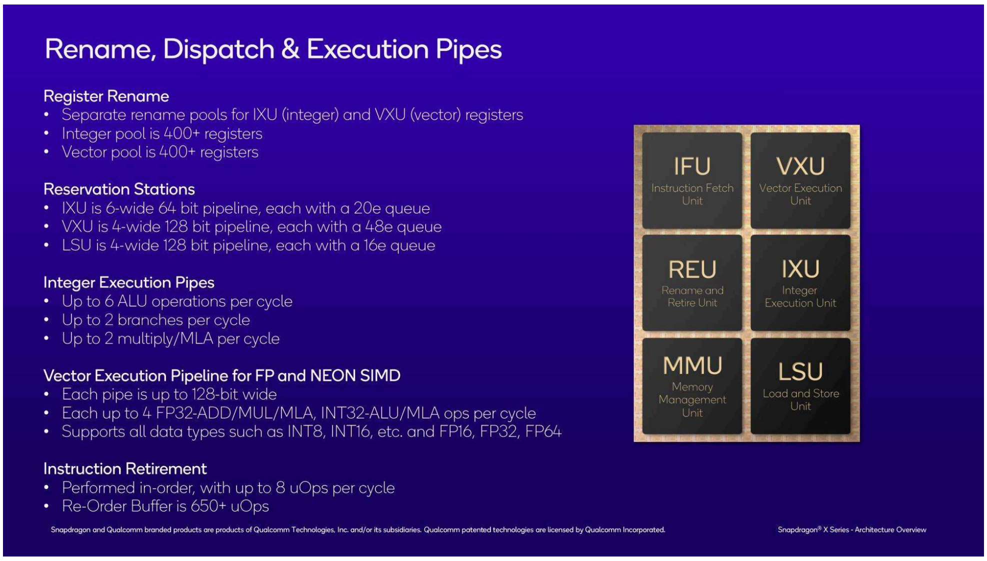
Altogether, the fetch/L1 unit of Oryon can retrieve up to 16 instructions per cycle.

That, in turn, feeds a very wide decode front-end. Oryon can decode up to 8 instructions in a single clock cycle, an even wider decode front-end than Redwood Cove (6) and Zen 4 (4). And all of the decoders are identical (symmetrical), so there are no special cases/scenarios required to achieve full throughput.

As with other contemporary processors, these decoded instructions are emitted as micro-ops (uOps), for further processing by the CPU core. Each Arm instruction can technically decode for up to 7 uOps, but according to Qualcomm, Arm v8 in general tends to be much closer to a 1-to-1 ratio of instructions-to-decoded micro-ops.

Branch prediction is another major driver of CPU core performance, and this is another area where Oryon doesn't skimp. Oryon features all the usual predictors: direct, conditional, and indirect The direct predictor is single-cycle; meanwhile, a branch mispredict carries a 13 cycle latency penalty. Unfortunately, Qualcomm is not disclosing the size of the branch target buffers themselves, so we don't have a good idea of just how big those are.

We do, however, have the size of the L1 translation lookaside buffer (TLB), which is used for virtual-to-physical memory address mapping. That buffer holds 256 entries, supporting both 4K and 64KB pages.



Rename, Dispatch & Execution Pipes

Register Rename
- Separate rename pools for IXU (integer) and VXU (vector) registers
- Integer pool is 400+ registers
- Vector pool is 400+ registers

Reservation Stations
- IXU is 6-wide 64 bit pipeline, each with a 20e queue
- VXU is 4-wide 128 bit pipeline, each with a 48e queue
- LSU is 4-wide 128 bit pipeline, each with a 16e queue

Integer Execution Pipes
- Up to 6 ALU operations per cycle
- Up to 2 branches per cycle
- Up to 2 multiply/MLA per cycle

Vector Execution Pipeline for FP and NEON SIMD
- Each pipe is up to 128-bit wide
- Each up to 4 FP32-ADD/MUL/MLA, INT32-ALU/MLA ops per cycle
- Supports all data types such as INT8, INT16, etc. and FP16, FP32, FP64

Instruction Retirement
- Performed in-order, with up to 8 uOps per cycle
- Re-Order Buffer is 650+ uOps

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

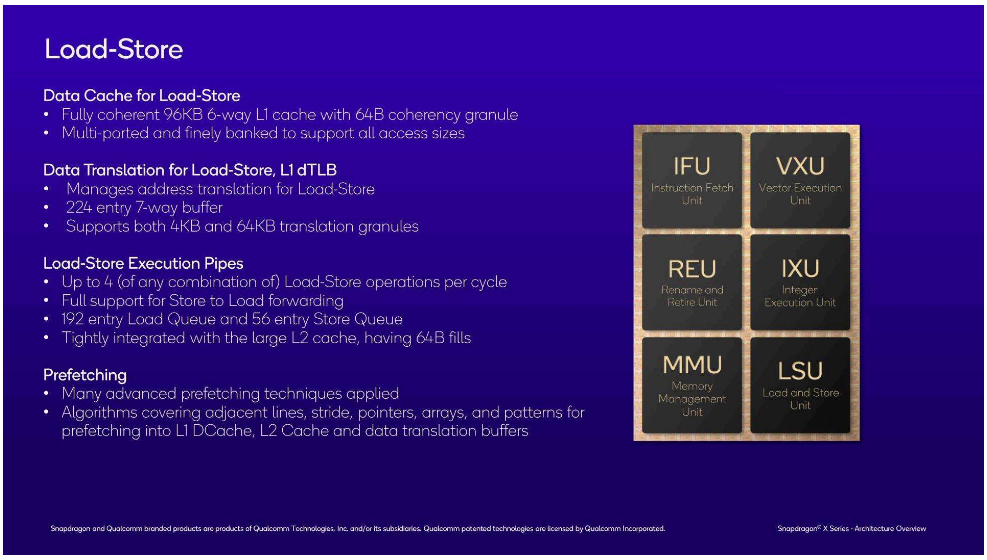
Flipping over to the execution backend of Oryon, there's a lot to talk about. In part because there's a lot of hardware and a lot of buffers here. Oryon features a sizeable 650+ re-order buffer (ROB) for extracting instruction parallelism and overall performance through out-of-order execution. This makes Qualcomm the latest CPU designer to throw traditional wisdom out the window and ship a massive ROB, eschewing claims that larger ROBs deliver diminishing returns.

Instruction retirement, in turn, matches the maximum capability of the decoder block: 8 instructions in, 8 uOps out. As noted before, the decoders can technically emit multiple uOps for a single instruction, but most often it's going to be perfectly aligned with the instruction retirement rate.

The register rename pools on Oryon are also quite massive (are you sensing a common theme here?). Altogether there's over 400 registers available for integers, and another 400 registers for feeding the vector units.

As for the actual execution pipes themselves, Oryon offers 6 integer pipes, 4 FP/vector pipes, and another 4 load/store pipelines. Qualcomm hasn't provided a full mapping of each pipeline here, so we can't run through all the possibilities and special cases. But at a high level, all of the integer pipelines can do basic ALU operations, while 2 can handle branches, and 2 can do complex multiply-accumulate (MLA) instructions. Meanwhile, we're told that the vast majority of integer operations have a single cycle latency – that is, they execute in a single cycle.

On the floating point/vector side of things, each of the vector pipelines has its own NEON unit. As a reminder, this is an Arm v8.7 architecture, so there aren't any vector SVE or Matrix SME pipelines here; the CPU core's only SIMD capabilities are with classic 128-bit NEON instructions. This does limit the CPU to narrower vectors than contemporary PC CPUs (AVX2 is 256-bits wide), but it does make up for the matter somewhat with NEON units on all four FP pipes. And, since we're now in the era of AI, the FP/vector units support all the common datatypes, right on down to INT8. The only notable omission here is BF16, a common data type for AI workloads; but for serious AI workloads, this is what the NPU is for.



Load-Store

Data Cache for Load-Store
- Fully coherent 96KB 6-way L1 cache with 64B coherency granule
- Multi-ported and finely banked to support all access sizes

Data Translation for Load-Store, L1 dTLB
- Manages address translation for Load-Store
- 224 entry 7-way buffer
- Supports both 4KB and 64KB translation granules

Load-Store Execution Pipes
- Up to 4 (of any combination of) Load-Store operations per cycle
- Full support for Store to Load forwarding
- 192 entry Load Queue and 56 entry Store Queue
- Tightly integrated with the large L2 cache, having 64B fills

Prefetching
- Many advanced prefetching techniques applied
- Algorithms covering adjacent lines, stride, pointers, arrays, and patterns for prefetching into L1 DCache, L2 Cache and data translation buffers

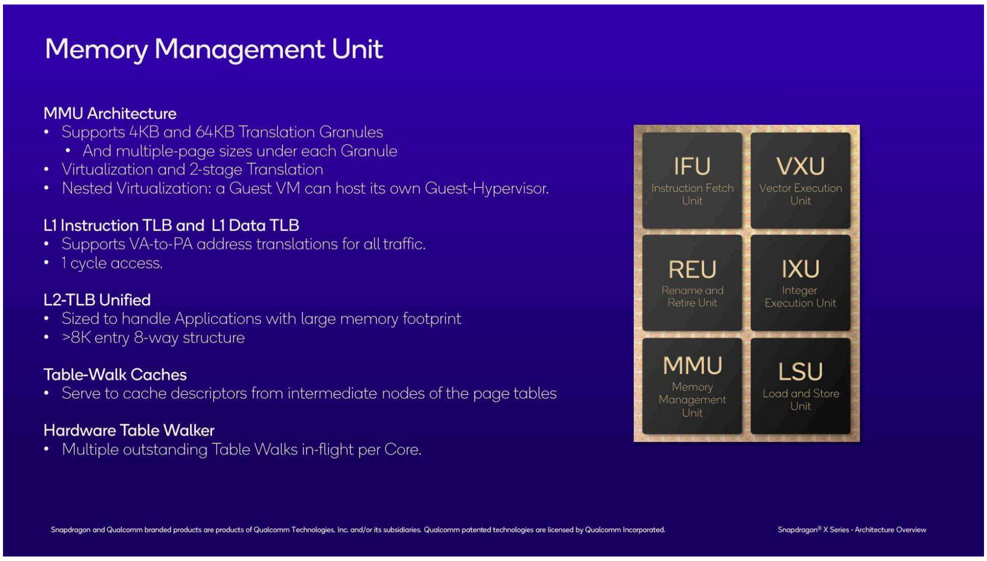
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

Branching off to its own slide, we have the data load/store units on Oryon. The core's load/store units are fully flexible, meaning that the 4 execution pipes can do any combination of loads and stores per cycle as needed. The load queues themselves can go up to 192 entries deep, while the store queues can go up to 26 entries. And all fills are the full size of a cache line: 64 bytes.

The L1 data cache supporting the load/store units is also quite sizable in its own right. The fully coherent 6-way associative cache is 96KB in size, twice the size of what you'll find on Intel's Redwood Cove (though the upcoming Lion Cove will significantly change this). And it's finely banked, in order to efficiently support a wide variety of different access sizes.

Otherwise, Qualcomm's memory prefetcher wanders a bit into "secret sauce" territory, as the company says the relatively complex unit contributes a great deal to performance. Consequently, Qualcomm isn't saying too much about how their prefetcher works, but it goes without saying that its ability to accurately predict and prefetch data can have a huge impact on the CPU core's overall performance, especially with how long a trip is to DRAM at modern processor clockspeeds. Overall, Qualcomm's prefetch algorithms seek to cover multiple cases, ranging from simple adjacencies and strides up to more complex patterns, using past access history to predict future data needs.



Memory Management Unit

MMU Architecture
- Supports 4KB and 64KB Translation Granules
  - And multiple-page sizes under each Granule
- Virtualization and 2-stage Translation
- Nested Virtualization: a Guest VM can host its own Guest-Hypervisor.

L1 Instruction TLB and L1 Data TLB
- Supports VA-to-PA address translations for all traffic.
- 1 cycle access.

L2-TLB Unified
- Sized to handle Applications with large memory footprint
- >8K entry 8-way structure

Table-Walk Caches
- Serve to cache descriptors from intermediate nodes of the page tables

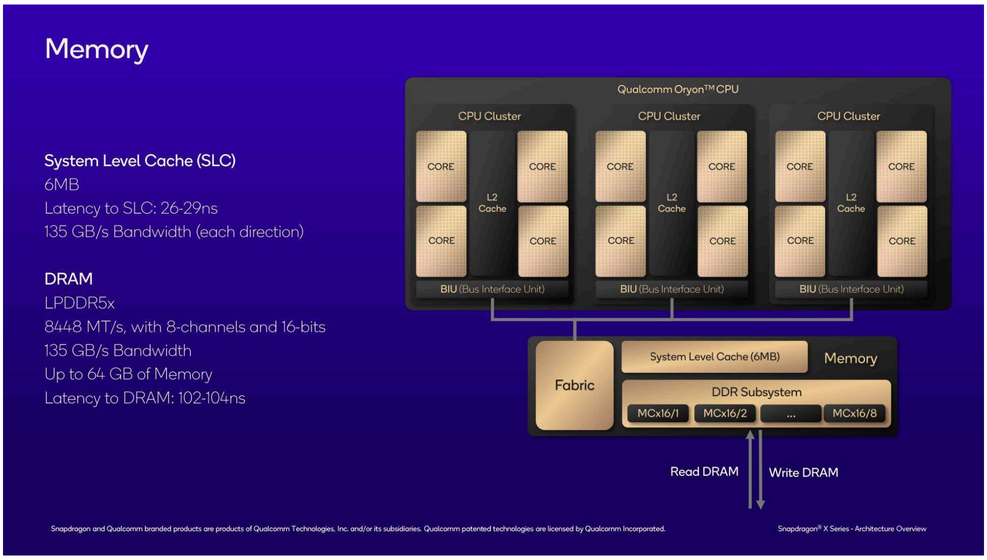
Hardware Table Walker
- Multiple outstanding Table Walks in-flight per Core.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

Conversely, Oryon's memory management unit is relatively straightforward. This is a fully-featured, modern MMU, and it supports even more esoteric features such as nested virtualization – which allows a guest virtual machine to host its own guest hypervisor for even more virtual machines farther down.

Of other notable capabilities here, the hardware table walker is another special mention. The unit, responsible for going out to DRAM if a cache line isn't in either the L1 or L2 caches, supports up to 16 concurrent table walks. And keep in mind this is per core, so a complete Snapdragon X chip can be doing upwards of 192 table walks at a time.



Memory

System Level Cache (SLC)
6MB
Latency to SLC: 26-29ns
135 GB/s Bandwidth (each direction)

DRAM
LPDDR5x
8448 MT/s, with 8-channels and 16-bits
135 GB/s Bandwidth
Up to 64 GB of Memory
Latency to DRAM: 102-104ns

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

Finally, going beyond the CPU cores and the CPU clusters, we have the highest level of the SoC: the shared memory subsystem.

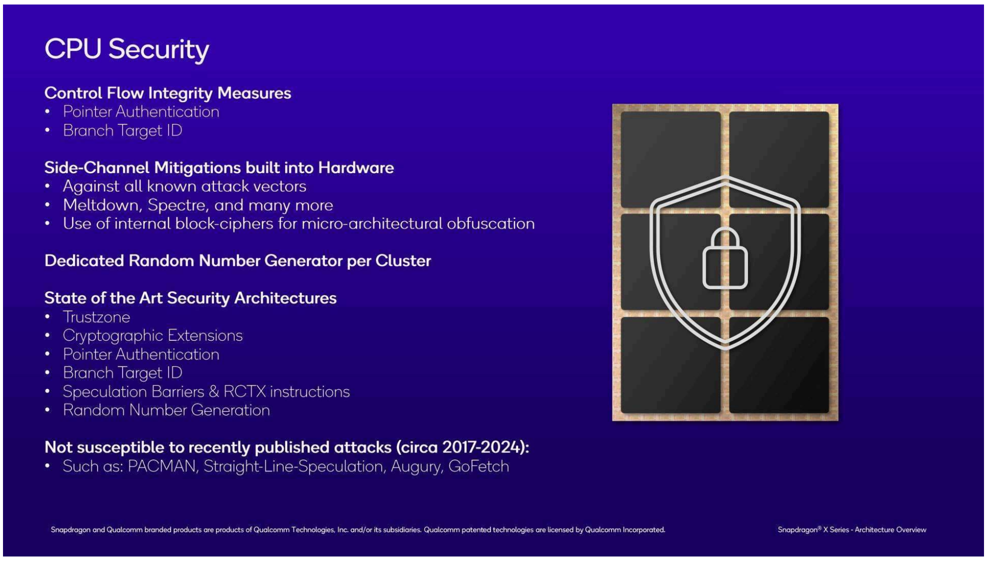
It's here where the final level of cache resides, with the chip's shared L3 cache. Given how big the L1 and L2 caches are for the chip, you might think that the L3 cache would also be quite sizeable. And you'd be wrong. Instead, Qualcomm has outfit the chip with just 6MB of L3 cache, a fraction of the size of the 36MB of L2 cache that it's backstopping.

With the chip already being cache-heavy at the L1/L2 level, and with the tight integration between those caches, Qualcomm has gone with a relatively small victim cache here to serve as the last stop before going out to system memory. Coming from traditional x86 CPUs, it's quite a significant change, though it's very on-brand for Qualcomm, whose Arm mobile SoCs also normally feature relatively small L3 caches. The upside, at least, is that the L3 cache is quite quick to access, at only 26-29 nanoseconds of latency. And it has the same amount of bandwidth as the DRAM (135GB/sec) to pass data between the L2 cache below it and the DRAM above it.

As for memory support, as noted in previous disclosures, Snapdragon X features a 128-bit memory bus with LPDDR5X-8448 support, giving it a maximum memory bandwidth of 135GB/second. At current LPDDR5X capacities, this allows Snapdragon X to address up to 64GB of RAM, though I wouldn't be too surprised down the line if Qualcomm validates it for 128GB once higher density LPDDR5X chips start shipping.

Notably, unlike some other mobile-focused chips, Snapdragon X does not use on-package memory of any kind. So LPDDR5X chips will go on the device motherboard itself, and it's up to device vendors to choose their own memory configurations.

With LPDDR5X-8448 memory, Qualcomm tells us that DRAM latency should be just over 100ns, at 102-104ns.



CPU Security

Control Flow Integrity Measures
- Pointer Authentication
- Branch Target ID

Side-Channel Mitigations built into Hardware
- Against all known attack vectors
- Meltdown, Spectre, and many more
- Use of internal block-ciphers for micro-architectural obfuscation

Dedicated Random Number Generator per Cluster

State of the Art Security Architectures
- Trustzone
- Cryptographic Extensions
- Pointer Authentication
- Branch Target ID
- Speculation Barriers & RCTX instructions
- Random Number Generation

Not susceptible to recently published attacks (circa 2017-2024):
- Such as: PACMAN, Straight-Line-Speculation, Augury, GoFetch

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

And because this is the last CPU architecture slide, we may as well throw in a quick mention of CPU security. Qualcomm supports all the security features you'd come to expect from a modern chip, including Arm TrustZone, per-cluster random number generators, and security-hardening features such as pointer authentication.

Notably, Qualcomm is claiming that Oryon has mitigations for all known side-channel attacks, including Spectre, an attack that has earned a reputation as "the gift that keeps on giving." This is an interesting claim as Spectre isn't really a hardware vulnerability itself, but rather is an inherent consequence of speculative execution. Which in turn is why it's so difficult to fully defend against (and the best defense is having sensitive operations fence themselves off). None the less, Qualcomm believes that by implementing various obfuscation tools within the hardware, they can protect against these kinds of side-channel attacks. So it will be interesting to see how this plays out.

A Note on x86 Emulation

And finally, I'd like to take a moment to make a quick note on what we've been told about x86 emulation on Oryon.

The x86 emulation scenario for Qualcomm is quite a bit more complex than what we've become accustomed to on Apple devices, as no single vendor controls both the hardware and the software stacks in the Windows world. So for as much as Qualcomm can talk about their hardware, for example, they have no control over the software side of the equation – and they aren't about to risk putting their collective foot in their mouth by speaking in Microsoft's place. Consequently, x86 emulation on Snapdragon X devices is essentially a joint project between the two companies, with Qualcomm providing the hardware, and Microsoft providing the Prism translation layer.

But while x86 emulation is largely a software task – it's Prism that's doing a lot of the heavy lifting – there are still certain hardware accommodations that Arm CPU vendors can make to improve x86 performance. And Qualcomm, for its part, has made these. The Oryon CPU cores have hardware assists in place to improve x86 floating point performance. And to address what's arguably the elephant in the room, Oryon also has hardware accommodations for x86's unique memory store architecture – something that's widely considered to be one of Apple's key advancements in achieving high x86 emulation performance on their own silicon.

Still, no one should be under the impression that Qualcomm's chips will be able to run x86 code as quickly as native chips. There's still going to be some translation overhead (just how much depends on the workload), and performance-critical applications will still benefit from being natively compiled to AArch64. But Qualcomm is not fully at the mercy of Microsoft here, and they have made hardware accommodations to improve their x86 emulation performance.

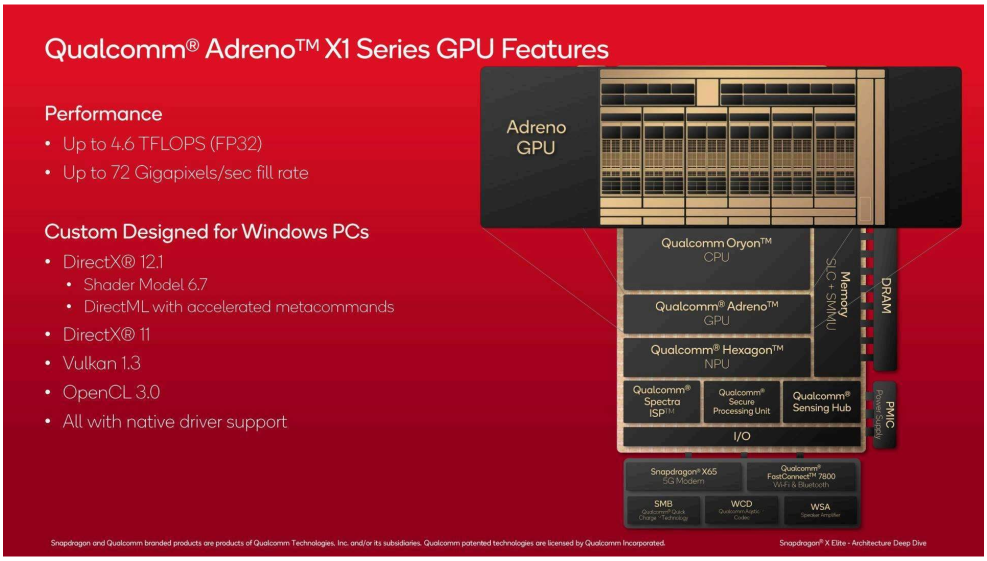
In terms of compatibility, the biggest roadblock here is expected to be AVX2 support. Compared to the NEON units on Oryon, the x86 vector instruction set is both wider (256b versus 128b) and the instructions themselves don't perfectly overlap. As Qualcomm puts it, AVX to NEON translation is a difficult task. Still, we know it can be done – Apple quietly added AVX2 support to their Game Porting Toolkit 2 this week – so it will be interesting to see what happens here in future generations of Oryon CPU cores. Unlike Apple's ecosystem, x86 isn't going away in the Windows ecosystem, so the need to translate AVX2 (and eventually AVX-512 and AVX10!) will never go away either.

Adreno X1 GPU Architecture: A More Familiar Face

Shifting gears, let's talk about the Snapdragon X SoC's GPU architecture: Adreno X.

Unlike the Oryon CPU cores, Adreno X1 is not a wholly new hardware architecture. In fact, with 3 generations of 8cx SoCs before it, it's not even new to Windows. Still, Qualcomm has been notoriously tight-lipped about their GPU architectures over the years, so the GPU architecture may as well be new to AnandTech readers. Suffice it to say, I've been trying to get a detailed disclosure from Qualcomm for over a decade at this point, and with Snapdragon X, they're finally delivering.

At a high level, the Adreno X1 GPU architecture is the latest revision of Qualcomm's ongoing series of Adreno architectures, with the X1 representing the 7th generation. Adreno itself is based on an acquisition from ATI over 15 years ago (Adreno is an anagram of Radeon), and over the years Qualcomm's Adreno architecture has more often than not been the GPU to beat in the Android space.



Qualcomm® Adreno™ X1 Series GPU Features

Performance

- Up to 4.6 TFLOPS (FP32)
- Up to 72 Gigapixels/sec fill rate

Custom Designed for Windows PCs

- DirectX® 12.1
  - Shader Model 6.7
  - DirectML with accelerated metacommands
- DirectX® 11
- Vulkan 1.3
- OpenCL 3.0
- All with native driver support

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

Things are a bit different in the Windows space, of course, as discrete GPUs push integrated GPUs well off to the side for workloads that absolutely need high GPU performance. And because game development has never become fully divorced from GPU architectures/drivers, Qualcomm's miniscule presence in the Windows market over the years has led to them going often overlooked by game developers. Still, Qualcomm isn't new to the Windows game, which gives them a leg-up as they try to move into taking a larger share of the Windows market.

From a feature standpoint, the Adreno X1 GPU architecture is unfortunately a bit dated compared to contemporary x86 SoCs. While the architecture does support ray tracing, the chip isn't able to support the full DirectX 12 Ultimate (feature level 12_2) feature set. And that means it must report itself to DirectX applications as a feature level 12_1 GPU, which means most games will restrict themselves to those features.

That said, Adreno X1 does support some advanced features, which are already being actively used on Android where DirectX's feature levels do not exist. As previously noted, there is ray tracing support, and this is exposed on Windows applications via the Vulkan API and its ray query calls. Given how limited Vulkan use is on Windows, Qualcomm understandably doesn't go into the subject in too much depth; but it sounds like Qualcomm's implementation is a level 2 design with hardware ray testing but no hardware BVH processing, which would make it similar in scope to AMD's RDNA2 architecture.

DirectX 12 Feature Levels

	12_2 (DX12 Ult.)	12_1	12_0
Ray Tracing (DXR 1.1)	**Yes**	No	No
Variable Rate Shading (Tier 2)	**Yes**	No	No

Mesh Shaders	**Yes**	No	No
Sampler Feedback	**Yes**	No	No
Conservative Rasterization	Yes	Yes	No
Raster Order Views	Yes	Yes	No
Tiled Resources (Tier 2)	Yes	Yes	Yes
Bindless Resources (Tier 2)	Yes	Yes	Yes
Typed UAV Load	Yes	Yes	Yes

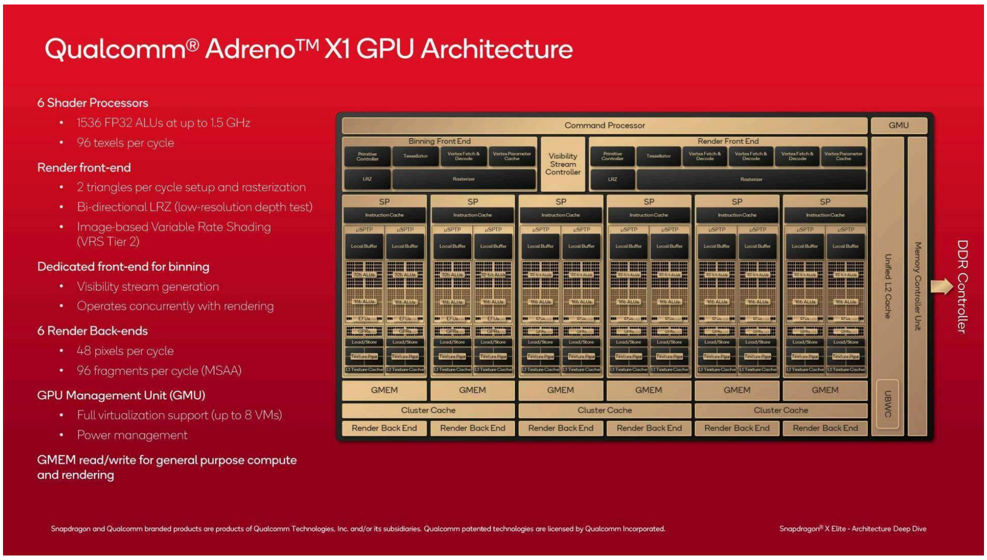
Otherwise, variable rate shading (VRS) tier 2 is supported as well, which is critical for optimizing shader workloads on mobile GPUs. So it appears the missing features holding the X1 back from DirectX 12 Ultimate support are mesh shaders and sampler feedback, which are admittedly some pretty big hardware changes.

In terms of API support, as previously noted, the Adreno X1 GPU supports DirectX and Vulkan. Qualcomm offers native drivers/paths for DirectX 12 and DirectX 11, Vulkan 1.3, and OpenCL 3.0. The only notable exception here is DirectX 9 support, which like fellow SoC vendor Intel, is implemented using D3D9on12, Microsoft's mapping layer for translating DX9 commands to DX12. DX9 games are far and few these days (the API was supplanted by DX10/11 over 15 years ago), but since this is Windows, backwards compatibility is an ongoing expectation.

Conversely, on the compute front Microsoft's new DirectML API for low-level GPU access for machine learning is supported. Qualcomm even has optimized metacommands written for the GPU, so that software tapping into DirectML can run more efficiently without knowing anything else about the architecture.

Adreno X1 GPU Architecture In Depth

High-level functionality aside, let's take a look at the low-level architecture.



Qualcomm® Adreno™ X1 GPU Architecture

6 Shader Processors

- 1536 FP32 ALUs at up to 1.5 GHz
- 96 texels per cycle

Render front-end

- 2 triangles per cycle setup and rasterization
- Bi-directional LRZ (low-resolution depth test)
- Image-based Variable Rate Shading (VRS Tier 2)

Dedicated front-end for binning

- Visibility stream generation
- Operates concurrently with rendering

6 Render Back-ends

- 48 pixels per cycle
- 96 fragments per cycle (MSAA)

GPU Management Unit (GMU)

- Full virtualization support (up to 8 VMs)
- Power management

GMEM read/write for general purpose compute and rendering

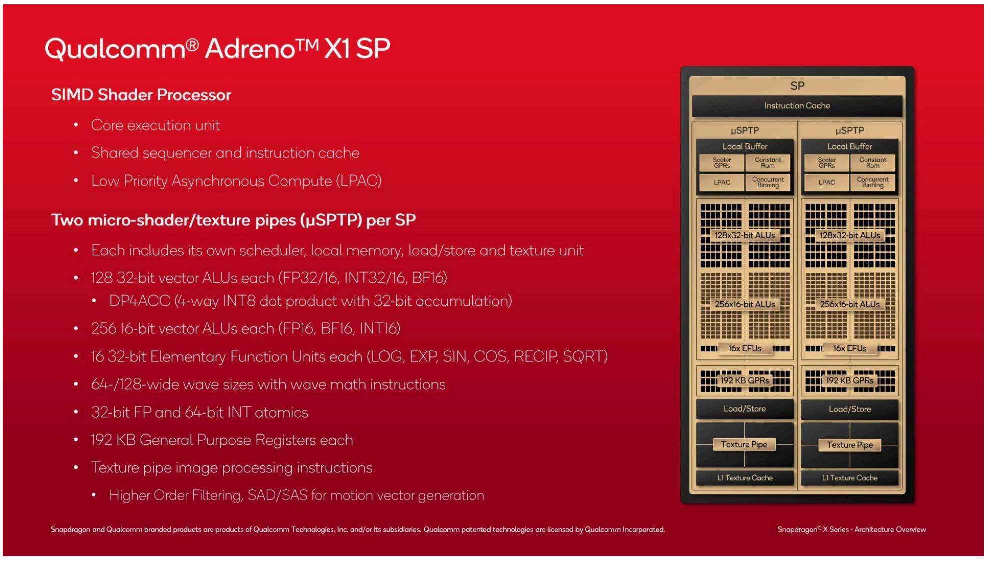
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

The Adreno X1 GPU is split up into 6 shader processor (SP) blocks, each offering 256 FP32 ALUs for a total of 1536 ALUs. With a peak clockspeed of 1.5GHz, this gives the integrated GPU on Snapdragon X a maximum throughput of 4.6 TFLOPS (with lesser amounts for lower-end SKUs).

Split up into the traditional front-end/SP/back-end setup we see with other GPUs, the front-end of the GPU handles triangle setup and rasterization, as well as binning for the GPU's tile-based rendering mode. Of note, the GPU front-end can setup and raster 2 tringles per clock, which is not going to turn any heads in the PC space in 2024, but is respectable for an integrated GPU. Boosting its performance, the front-end can also do early depth testing to reject polygons that will never be visible before they are even rasterized.

Meanwhile the back-end is made from 6 render output units (ROPs), which can process 8 pixels per cycle each, for a total of 48 pixels/clock rendered. The render back-ends are plugged in to a local cache, as well as an important scratchpad memory that Qualcomm calls GMEM (more on this in a bit).



Qualcomm® Adreno™ X1 SP

SIMD Shader Processor

- Core execution unit
- Shared sequencer and instruction cache
- Low Priority Asynchronous Compute (LPAC)

Two micro-shader/texture pipes (µSPTP) per SP

- Each includes its own scheduler, local memory, load/store and texture unit
- 128 32-bit vector ALUs each (FP32/16, INT32/16, BF16)
  - DP4ACC (4-way INT8 dot product with 32-bit accumulation)
- 256 16-bit vector ALUs each (FP16, BF16, INT16)
- 16 32-bit Elementary Function Units each (LOG, EXP, SIN, COS, RECIP, SQRT)
- 64-/128-wide wave sizes with wave math instructions
- 32-bit FP and 64-bit INT atomics
- 192 KB General Purpose Registers each
- Texture pipe image processing instructions
  - Higher Order Filtering, SAD/SAS for motion vector generation

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

The individual shader processor blocks themselves are relatively customary, especially if you've seen an NVIDIA GPU architecture diagram. Each SP is further subdivided into two micro-pipes (micro shader pipe texture pipe, or uSPTP), which is helmed by its own dedicated scheduler and other resources such as local memory, load/store units, and texture units.

Each uSPTP offers 128 FP32 ALUs. And, in a bit of a surprise, there is also a separate set of 256 FP16 ALUs, meaning that Adreno X1 doesn't have to share resources when processing FP16 and FP32 data, unlike architectures which execute FP16 operations on FP32 ALUs. Though for good measure, the FP32 units *can* be used for FP16 operations as well, if the GPU scheduler determines it's needed.

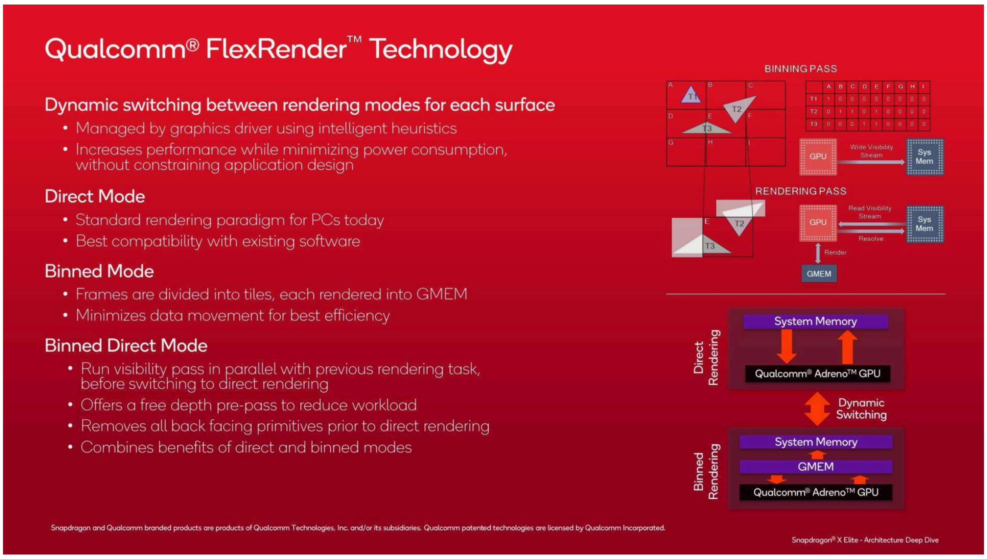
Finally, there are 16 elementary functional units (EFUs), which handle transcendental functions such as LOG, SQRT, and other rare (but important) mathematical functions.

Surprisingly here, the Adreno X1 uses a rather large wavefront size. Depending on the mode, Qualcomm uses either 64 or 128 lane wide waves, with Qualcomm telling us that they typically use 128-wide wavefronts for 16bit operations such as fragment shaders, while 64-wide wavefronts are used for 32bit operations (e.g. pixel shaders).

Comparatively, AMD's RDNA architectures use 32/64 wide wavefronts, and NVIDIA's wavefronts/warps are always 32 wide. Wide designs have fallen out of favor in the PC space due to the difficulty in keeping them fed (too much divergence), so this is interesting to see. Ultimately, despite the usual wavefront size concerns, it seems to be working well for Qualcomm given the high GPU performance of their smartphone SoCs – no small feat given the high resolution of phone screens.

ALUs aside, each uSPTP includes their own texture units, which are capable of spitting out 8 texels per clock per uSPTP. There's also limited image processing functionality here, including texture filtering, and even SAD/SAS instructions for generating motion vectors.

Finally, there's quite a bit of register space within each uSPTP. Along with the L1 texture cache, there's a total of 192KB of general purpose registers to keep the various blocks fed and to try to hide latency bubbles in the wavefronts.



Qualcomm® FlexRender™ Technology

Dynamic switching between rendering modes for each surface

- Managed by graphics driver using intelligent heuristics
- Increases performance while minimizing power consumption, without constraining application design

Direct Mode

- Standard rendering paradigm for PCs today
- Best compatibility with existing software

Binned Mode

- Frames are divided into tiles, each rendered into GMEM
- Minimizes data movement for best efficiency

Binned Direct Mode

- Run visibility pass in parallel with previous rendering task, before switching to direct rendering
- Offers a free depth pre-pass to reduce workload
- Removes all back facing primitives prior to direct rendering
- Combines benefits of direct and binned modes

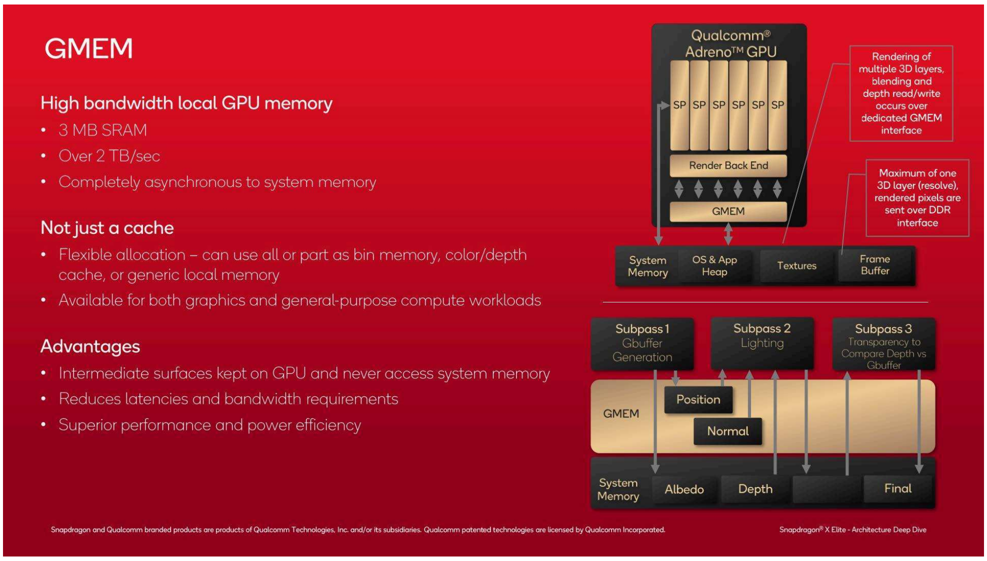
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

As noted earlier, the Adreno X1 supports multiple rendering modes in order to get the best performance possible, which the company calls their FlexRender technology. This is a subject that doesn't come up too often with PC GPU designs, but is of greater importance in the mobile space for historical and efficiency reasons.

Besides the traditional direct/immediate mode rendering method (the typical mode for most PC GPUs), Qualcomm also supports tile-based rendering, which they call binned mode. As with other tile-based renderers, binned mode splits a screen up into multiple tiles, and then renders each one separately. This allows the GPU to only work on a subset of data at once, keeping most of that data in its local caches and minimizing the amount of traffic that goes to DRAM, which is both power-expensive and performance-constricting.

And finally, Adreno X1 has a third mode that combines the best of binned and direct rendering, which they call binned direct mode. This mode runs a binned visibility pass before switching to direct rendering, as a means to further cull back-facing (non-visible) triangles so that they don't get rastered. Only after that data is culled does the GPU then switch over to direct rendering mode, now with a reduced workload.



GMEM

High bandwidth local GPU memory

- 3 MB SRAM
- Over 2 TB/sec
- Completely asynchronous to system memory

Not just a cache

- Flexible allocation – can use all or part as bin memory, color/depth cache, or generic local memory
- Available for both graphics and general-purpose compute workloads

Advantages

- Intermediate surfaces kept on GPU and never access system memory
- Reduces latencies and bandwidth requirements
- Superior performance and power efficiency

Qualcomm® Adreno™ GPU
SP
SP
SP
SP
SP
SP
Render Back End
GMEM
Rendering of multiple 3D layers, blending and depth read/write occurs over dedicated GMEM interface
Maximum of one 3D layer (resolve), rendered pixels are sent over DDR interface
System Memory
OS & App Heap
Textures
Frame Buffer
Subpass 1 Gbuffer Generation
Subpass 2 Lighting
Subpass 3 Transparency to Compare Depth vs Gbuffer
GMEM
Position
Normal
System Memory
Albedo
Depth
Final

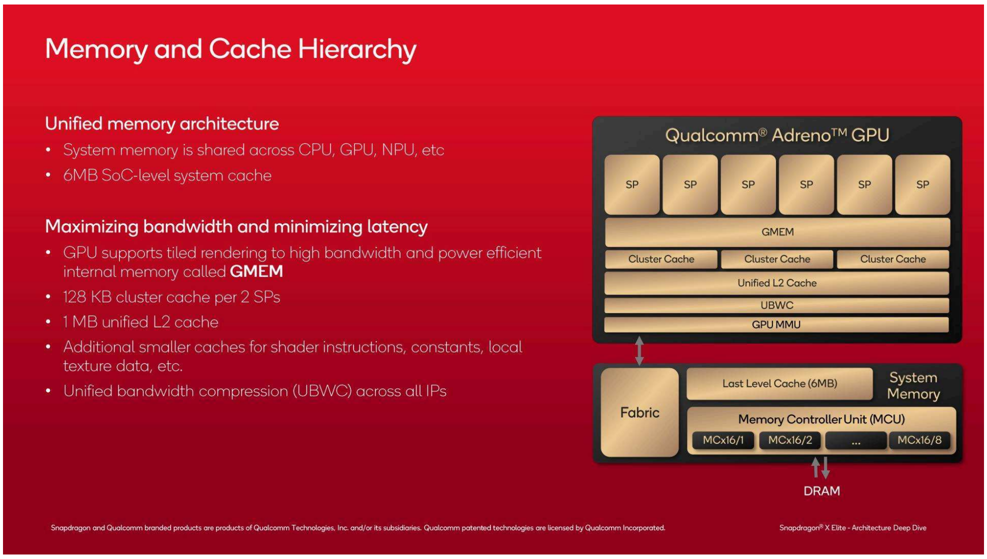
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

Key to making the binned rendering modes work is the GPU's GMEM, a 3MB SRAM block that serves as a very high bandwidth scratch pad for the GPU. Architecturally, GMEM is more than a cache, as it's decoupled from the system memory hierarchy, and the GPU can do virtually anything it wants with the memory (including using it as a cache, if need be).

At 3MB in size, the GMEM block is not very large overall. But that's big enough to store a tile – and thus prevent a whole lot of traffic from hitting the system memory. And it's fast, too, with 2.3TB/second of bandwidth, which is enough bandwidth to allow the ROPs to run at full-tilt without being constrained by memory bandwidth.

With the GMEM block in place, in an ideal scenario, the GPU only needs to write out to the DRAM once per title, when it finishes rendering said tile. In practice, of course, there ends up being more DRAM traffic than that, but this is one of Qualcomm's key features for avoiding chewing up memory bandwidth and power with GPU writes to DRAM.



Memory and Cache Hierarchy

Unified memory architecture

- System memory is shared across CPU, GPU, NPU, etc
- 6MB SoC-level system cache

Maximizing bandwidth and minimizing latency

- GPU supports tiled rendering to high bandwidth and power efficient internal memory called **GMEM**
- 128 KB cluster cache per 2 SPs
- 1 MB unified L2 cache
- Additional smaller caches for shader instructions, constants, local texture data, etc.
- Unified bandwidth compression (UBWC) across all IPs

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

And when the Adreno X1 does need to go to system memory, it will go through its own remaining caches, before finally reaching the Snapdragon X's shared memory controller.

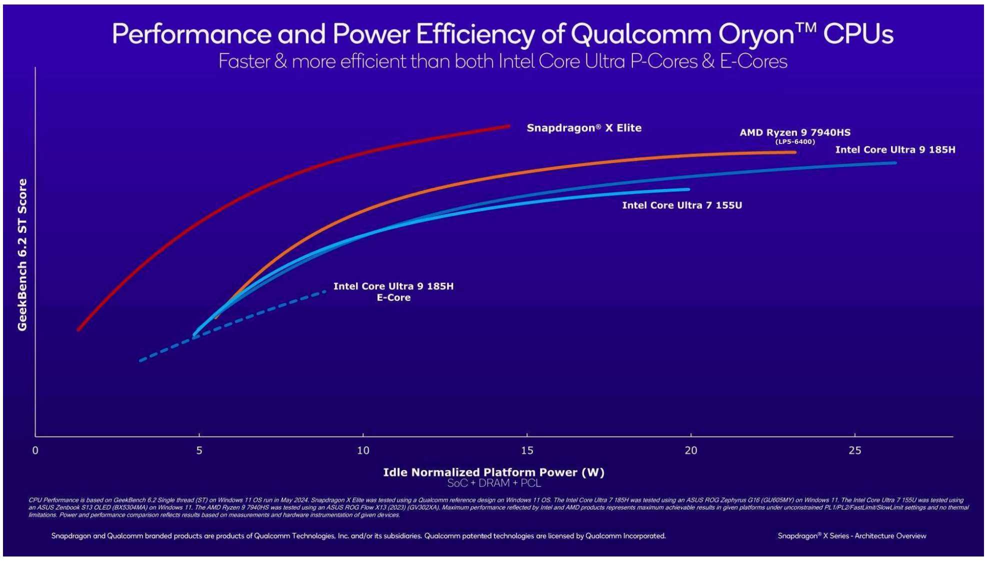
Above the GMEM, there is a 128KB cluster cache for each pair of SPs (for 384KB in total for a full Snapdragon X). And above that still is a 1MB unified L2 cache for the GPU.

Finally, this leaves the system level cache (L3/SLC), which serves all of the processing blocks on the GPU. And when all else fails, there is the DRAM.

On a concluding note, while Qualcomm doesn't have a dedicated slide for this, it's interesting to note that the Adreno X1 GPU also includes a dedicated RISC controller within the GPU which serves as a GPU Management Unit (GMU). The GMU provides several features, the most important of which is power management within the GPU. The GMU works in concert with power management requests elsewhere in the SoC, allowing the chip to reallocate power between the different blocks depending on what the SoC decides is the most performant allocation method.

Performance Promises and First Thoughts

Wrapping things up, let's touch upon a couple of Qualcomm's performance slides before closing out this architectural deep dive. While the whole world will get to see what the Snapdragon X can do first-hand next week when retail devices launch, until then it gives us a bit more insight into what to expect. Just be sure to take it with the requisite grain of salt.



Performance and Power Efficiency of Qualcomm Oryon™ CPUs

Faster & more efficient than both Intel Core Ultra P-Cores & E-Cores

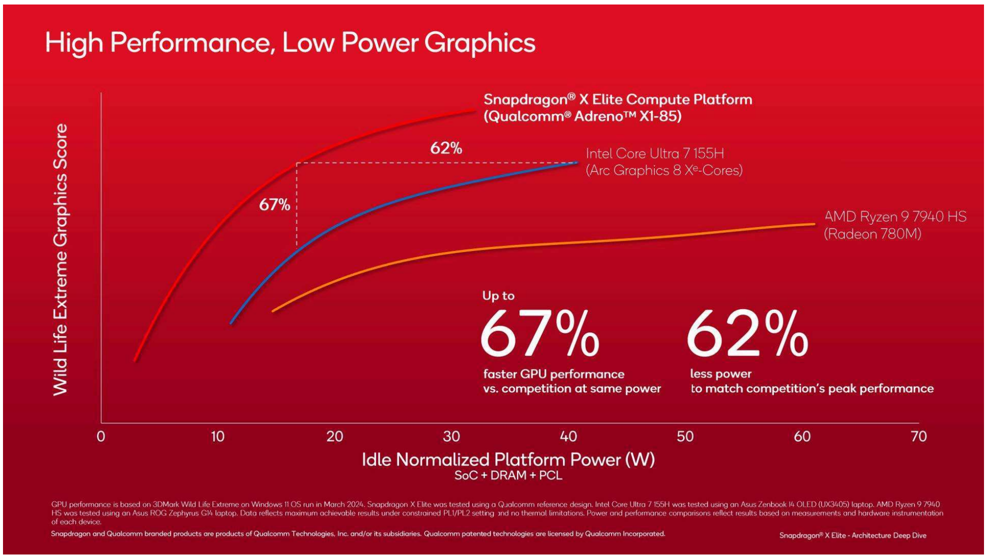
CPU Performance is based on GeekBench 6.2 Single thread (ST) on Windows 11 OS run in May 2024. Snapdragon X Elite was tested using a Qualcomm reference design on Windows 11 OS. The Intel Core Ultra 7 185H was tested using an ASUS ROG Zephyrus G16 (GU605MY) on Windows 11. The Intel Core Ultra 7 155U was tested using an ASUS Zenbook S13 OLED (BX5304MA) on Windows 11. The AMD Ryzen 9 7940HS was tested using an ASUS ROG Flow X13 (2023) (GV302XA). Maximum performance reflected by Intel and AMD products represents maximum achievable results in given platforms under unconstrained PL1/PL2/FastLimit/SlowLimit settings and no thermal limitations. Power and performance comparison reflects results based on measurements and hardware instrumentation of given devices.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

On the CPU side of matters, Qualcomm is claiming that the Snapdragon X Elite can beat the entire field of contemporary PC competitors in GeekBench 6.2 single-threading. And by a significant degree, too, when power efficiency is taken into account.

In short, Qualcomm claims that the Oryon CPU core in the Snapdragon X Elite can beat both Redwood Cove (Meteor Lake) and Zen 4 (Phoenix) in absolute performance, even if the x86 cores are allowed unrestricted TDPs. With mobile x86 chips turboing as high as 5GHz it's a bold claim, but not out of the realm of possibility.



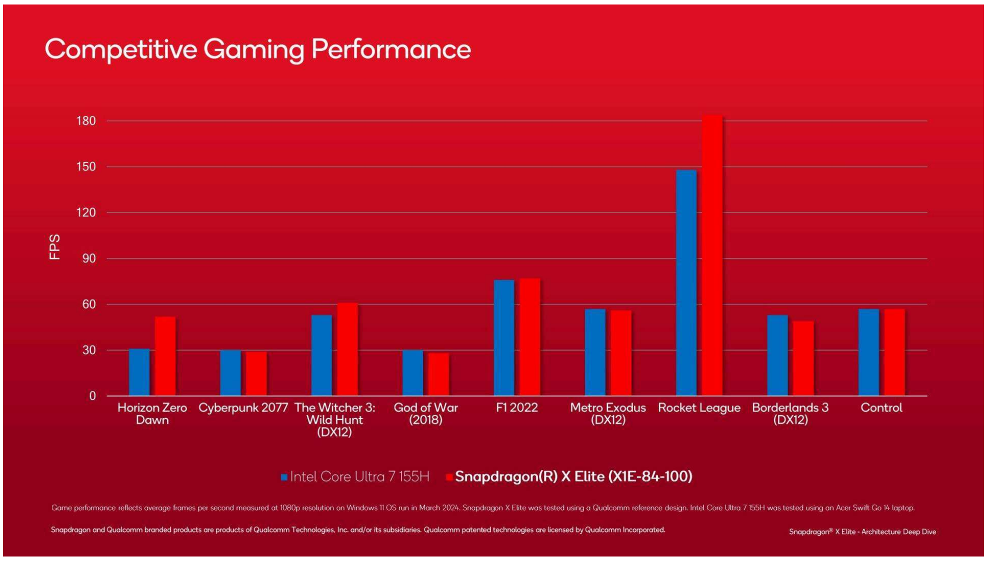
High Performance, Low Power Graphics

GPU performance is based on 3DMark Wild Life Extreme on Windows 11 OS run in March 2024. Snapdragon X Elite was tested using a Qualcomm reference design. Intel Core Ultra 7 155H was tested using an Asus Zenbook 14 OLED (UX3405) laptop. AMD Ryzen 9 7940 HS was tested using an Asus ROG Zephyrus G14 laptop. Data reflects maximum achievable results under constrained PL1/PL2 setting and no thermal limitations. Power and performance comparisons reflect results based on measurements and hardware instrumentation of each device.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Elite - Architecture Deep Dive

Meanwhile on the GPU front, Qualcomm is making similar energy efficiency gains. Though the workload in question – 3DMark WildLife Extreme – is not likely to translate into most games, as this is a mobile-focused benchmark that has long been optimized to heck and back within every mobile SoC vendor's drivers.



Competitive Gaming Performance

Game performance reflects average frames per second measured at 1080p resolution on Windows 11 OS run in March 2024. Snapdragon X Elite was tested using a Qualcomm reference design. Intel Core Ultra 7 155H was tested using an Acer Swift Go 14 laptop.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

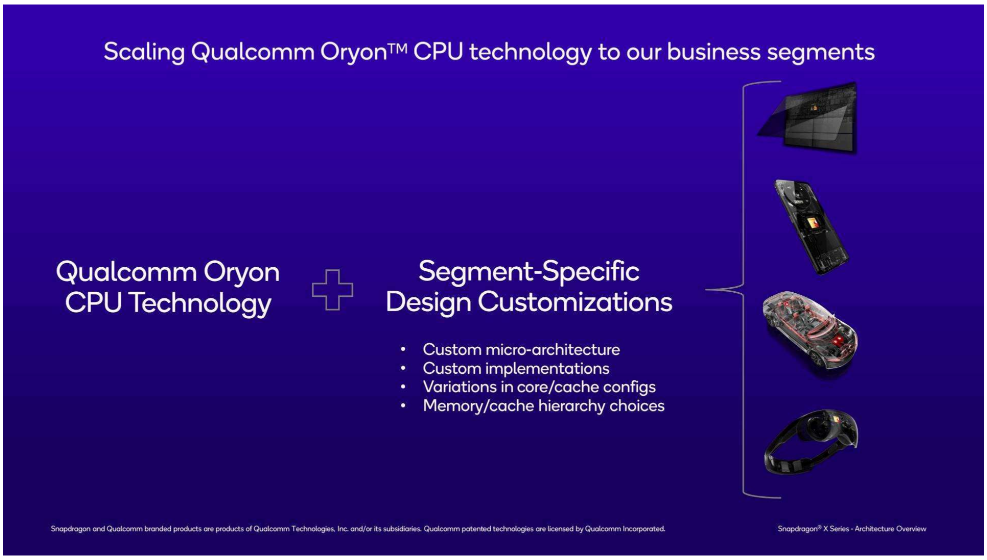
Snapdragon® X Elite - Architecture Deep Dive

Performance benchmarks using actual games are arguably more useful here. And even though Qualcomm is probably doing some cherry-picking, the top Snapdragon X SKU is often trading blows with Intel's Core Ultra 7 155H. It admittedly makes for a less impressive showing overall, but it's good to see where Qualcomm is currently landing on real games. And in this case, even just a mix of ties/beats of one of Intel's better mobile chips is not a bad showing.

First Thoughts

And there you have it, our first deep dive into a Qualcomm Snapdragon X SoC architecture. With Qualcomm investing into the Windows-on-Arm ecosystem for the long haul, this will hopefully be the first of many, as the company seeks to become the third major Windows CPU/SoC vendor.

But the ultimate significance of the Snapdragon X SoC and its Oryon CPU cores goes beyond just mere SoCs for PC laptops. Even if Qualcomm is wildly successful here, the number of PC chips they'll ship will be a drop in the bucket compared to their true power base: the Android SoC space. And this is where Oryon is going to be lighting the way to some significant changes for Qualcomm's mobile SoCs.



Scaling Qualcomm Oryon™ CPU technology to our business segments

Qualcomm Oryon CPU Technology

+

Segment-Specific Design Customizations

- Custom micro-architecture
- Custom implementations
- Variations in core/cache configs
- Memory/cache hierarchy choices

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Snapdragon® X Series - Architecture Overview

As noted by Qualcomm since the start of their Oryon journey, this is ultimately the CPU core that will be at the heart of all of Qualcomm's products. What starts this month with PC SoCs will eventually grow to include mobile SoCs like the Snapdragon 8 series, and farther along still will be Qualcomm's automotive products, and high-end offshoots like their XR headset SoCs. And while I doubt we'll really see Oryon and its successors in Qualcomm's product in a true top-to-bottom fashion (the company needs small and cheap CPU cores for their budget lines like Snapdragon 6 and Snapdragon 4), there is no doubt that it's going to become a cornerstone of most of their products over the long run. That's the value of differentiation of making your own CPU core – and getting the most value out of that CPU core by using it in as many places as possible.

Ultimately, Qualcomm has spent the last 8 months hyping up their next-generation PC SoC and its bespoke CPU core, and now it's time for all of the pieces to fall into place. The prospect of having a third competitor in the PC CPU space – and an Arm-baesd one at that – is exciting, but slideware and advertising aren't hardware and benchmarks. So we're eagerly awaiting what next week will bring, and seeing if Qualcomm's engineering prowess can live up to the company's grand ambitions.

Gallery: Qualcomm Snapdragon X CPU & GPU Architecture Slide Deck