SOLVING THE NEED FOR HIGH-PERFORMANCE, LOW-POWER COMPUTING

HOW INTEL'S TREMONT LOW-POWER X86 ARCHITECTURE HELPS SOLVE MULTI-CORE COMPUTING PROBLEMS

EXECUTIVE SUMMARY

In the world of computing, low-power performance is increasingly important. There are a multitude of trends driving this shift, which likely began in mobile but now affects all types of computing across the ecosystem.

Market Trends

PCs and tablets are thinner and lighter than ever before, and users demand longer battery life and performance. As PCs continue to emulate the personal nature of the mobile experience of smartphones, responsiveness and endurance become increasingly important while high performance remains a key value proposition. Tablets face the same predicament as they become thinner and essentially transform into PCs.

Cloud infrastructure and data center operators, especially hyperscale vendors, are looking for more power-efficient, denser servers that can scale quickly with dynamic user demands. As technologies like AI mature, the need for low-power computing will only increase. This new standard will be especially necessary as the topology of cloud infrastructure moves closer to the edge; it will need to operate in lower power envelopes without sacrificing performance.

Another trend driving this shift is the imminent growth of 5G. The increased data density of 5G will require the topology of both the network infrastructure and the cloud infrastructure to change drastically. These changes will drive the cloud closer to the network edge to reduce latency while also increasing the amount of processing in the RAN. The power envelopes that the edge cloud and RAN operate in are still very low, but network infrastructure will require higher performance levels than exist today to handle standalone 5G network features (like network slicing for the RAN and low-latency cloud applications for the edge cloud).

Current Challenges

Multi-core computing is evolving to match the needs of the market, but designs with multiple types of cores have varying processing and energy profiles. The workloads that are pushing the latest trends in computing towards low power are also compute-intensive, and computing performance cannot be allowed to suffer because of lower power envelopes.

Heterogeneous computing, in which different types of processing cores co-exist on the same chip, is increasingly common across the industry. Many of these chips combine different types of CPU cores in an attempt to strike a balance between the highest-possible performance and low power consumption. This trend of heterogeneous multi-core computing started by blending large, highly performant cores with very-low-power cores. However, as the market matured, the capabilities of low-power cores increased steadily.

The balance between different cores is necessary because there was historically a considerable performance delta between high performance and low-power cores in chip designs. Many designs came from the ARM world, which was traditionally limited to phones and tablets (although the architecture is diversifying into servers and infrastructure). Historically, these ARM cores have also been lower performing than many x86 cores, albeit at a lower power level. That said, the x86 ecosystem is a key factor driving many of the new edge-usage models; many of the frameworks and operating systems already work well on x86 at higher power envelopes, and the ecosystem is familiar with the development environment. Current x86 core designs aim for higher performance workloads; however, the net market requirements of multi-core heterogeneous CPU designs are driving the market towards new architectures.

Potential Solutions

Naturally, many would first look to redesign the larger cores to be more power efficient. However, the cost of the compromises needed would defeat the purpose of having larger, higher performance cores. Qualcomm attempted this strategy with its custom ARM cores, putting the Krait 450 architecture inside of the Snapdragon 805. However, this design strategy had a limited scale. Eventually, the company adopted a heterogeneous, multi-core design that paired larger high-performance cores with smaller low-power cores.

A new optimized CPU core design is likely to be a key component of a dynamic, multi-core heterogeneous solution. This new CPU core design would need to lower the x86

power envelope while increasing the performance compared to previous x86 low-power designs. There are very few companies with x86 licenses that could execute such a design. Intel's approach is to blend its high-performance (Core) and low-power x86 architecture cores to create a highly functional and efficient chip for many different kinds of workloads (like mobile devices, cloud infrastructure and carrier networks). Intel's recently introduced Lakefield processor – a complete multi-core heterogeneous chip design – combines the company's 'Sunny Cove' Core architecture with 'Tremont' low-power architecture to maximize performance and power across varied workloads.

Processing That Matches Workloads

Due to evolving user needs and workload changes, devices and infrastructure are getting increasingly smaller with lower power envelopes.

Mobile Device Growth

A growing demand for thinner, larger, bezel-less notebook PCs and tablets that behave more like notebooks is driving the need for an upgraded processor. The processor must deliver improved battery life and performance from generation to generation while also supporting new display resolutions and color depths like HDR. The processor and the platforms it enables also need to meet consumers' expectations that their devices will be as responsive as their most personal computing devices: their smartphones. Cutting-edge devices like Microsoft's upcoming Surface Neo with Intel Lakefield push the envelope of what's possible, with its foldable dual-display design. Devices like this need high performance at low power.

Cloud Infrastructure Growth

Major changes are underway in the realm of cloud infrastructure. A growing densification of nodes and clusters is increasing the total compute density available to applications on demand. The need for high-performance, low-power computing only increases with the cloud's constantly changing workloads and the movement towards on-demand scalable computing. In addition to the needs of the centrally located cloud datacenter, cloud infrastructure must move closer to the edge due to demands for lower latencies for future workloads driven by 5G. As the cloud evolves, the constraints for cooling and power will continue to pressure computing towards smaller, denser power envelopes.

Networking & 5G

The growing global deployments of 5G will require increased densification of cellular networks, especially as frequencies like mmWave continue to gain momentum. Early 5G networks rely heavily on 4G infrastructure for quick deployment and signal availability. However, the rollout of standalone (SA) 5G networks will demand a complete rebuilding of network infrastructure to 5G new radio (NR) standalone requirements. The RAN will require higher performance to handle the new 5G radios and support key 5G operator use cases like network slicing. The increased bandwidth fed into each new 5G node will also put added performance and power requirements on existing network infrastructure. Increased throughputs and higher bandwidth will require operators and their suppliers to consider higher performance infrastructure that can meet the expectations and specifications of 5G NR.

Lower Power Without Compromise

Smaller power envelopes from all of these markets push chip makers towards lower power cores and more efficient computing. Increased efficiency means more than just completing the task on the lowest power core; tasks must be completed competently based on the best core for that task. Mobile device constraints push the chip industry towards these new chip architectures. On the other side of the coin, hyperscale's highest long-term cost is electricity. When choosing a processor for cloud nodes and clusters, hyperscalers must seek to minimize the power consumed per bit processed. The networks that sit between mobile devices and the cloud infrastructure will also have to move towards a more efficient architecture to address increased density and higher network throughputs.

Many processor designs in the industry, especially in mobile, now employ multi-core heterogeneous structures that combine high-performance cores with low-power efficiency cores to maximize performance per watt dependent on the workload. Large, high-performance cores alone are not enough to address low-power needs. Conversely, low-power-efficiency cores are not enough to execute all workloads quickly. The heterogeneous strategy produces very efficient, highly optimized cores that help to execute tasks quickly at low power. New designs for power-efficient cores will address the need for low power and high performance with single-thread and multi-threaded workloads.

INTEL TREMONT

To address the current and future needs of the market, Intel created a new efficiency-focused, low-power x86 core codenamed Tremont. Tremont complements Sunny Cove and other types of high-performance cores to balance power and execution for different workloads. This means pairing larger cores, with their different capabilities and designs, with newer, smaller cores to right-size the chip for the application. Tremont has quite a few design features that set it apart from previous generations of Intel's low-power x86 architectures. These include new power management advances, a new clustered front end, new back-end features, added security, and cryptography improvements. Additionally, Tremont boasts 30% better IPC over Goldmont+ (the previous generation) and 70% better IPC over the generation before that.

Improved Power Management

In order to improve Tremont's compute efficiency, Intel made advances to the low-power x86 core's power management features. Intel enhanced its SpeedStep technology, which allows the clock speed of the processor to be dynamically changed based on factors like power, thermals and workload. This is done by adjusting the processor voltage and core frequency, which allows for decreased average power consumption and heat generation. The new core also supports deep power-gated states, meaning each individual core (or entire modules) can be power gated. This allows the processor to isolate power consumption to each individual core or module, giving each module independent frequency control.

Additionally, Tremont supports Intel's Speed Shift technology, initially introduced in Intel's Skylake desktop and server processors. Speed Shift allows the operating system to hand over complete or partial control of the processor's performance back to the CPU. According to Intel, this allows the processor to change frequency more quickly, with less latency, and at shorter intervals (changing frequency more often, dependent on the workload's needs). This is especially useful for smaller workloads that only require a quick burst of performance. By shortening the time needed to ramp up and change frequency, workloads are finished faster and with less power. Additionally, it allows the processor to return to a low-power, idle state more quickly.

Front-End Improvements

Intel also beefed up the front-end microarchitecture on Tremont. Intel says it improved Tremont's branch prediction to the point where it can be considered 'Core-class'—in other words, on par with Intel's larger and more high-performance Core architecture

processing cores. Intel accomplished this by increasing the size of the prediction arrays with multiple prediction tables, giving it better prediction accuracy. Prediction accuracy is important because it reduces misses, wasted CPU cycles, power consumption, and latency.

In addition to the front end's improved branch prediction, Intel added a clustered front end to the Tremont architecture. Tremont features two parallel decode pipelines so that every time the processor predicts a branch, it can send that address off to either cluster and have the cluster start decoding the instructions out of order. Each cluster has three parallel decoders, so the peak throughput for decoding instructions is six instructions per cycle. For comparison, the previous generation's peak throughput was three instructions per cycle.

More Balanced Back End

The Tremont architecture allows the front end to read instructions straight out of the cache and send them to the back end. This new design avoids the penalties of using a monolithic design to support wide instruction decode, without requiring a second cache. The new front end is paired up with a beefed-up back end and helps balance the capabilities of the processing pipeline—more on this later. Intel also implemented multiple load pipelines in the Tremont architecture to improve parallelism and utilize more of the processor at one time.

Intel also increased the execution ports in Tremont to 10 (up from eight in Goldmont+). This increases the total potential compute ability of the Tremont architecture, which is key to improving its efficiency. The two extra execution ports add an integer and vector port. These consume more work than in previous low-power x86 architectures, thanks to the enhanced front-end design. The dual load/store pipelines can do two loads, two stores or a store and a load, depending on the bandwidth that the workload needs, and also has a second AGU (address generation unit). This can further reduce the number of CPU cycles required to execute instructions by more quickly calculating the memory addressed, potentially increasing performance and saving power. Much like Intel's previous x86 low-power architectures, Tremont also features three integer ALUs (arithmetic logic unit).

Security & Cryptography

In addition to these improved capabilities, Intel also made improvements to Tremont's cryptography capabilities. Intel made considerable improvements to AES (advanced encryption standard) encryption and decryption in the previous generation. In Tremont,

Intel continued to improve performance for the AES-NI (New Instructions) instruction set and SHA (secure hash algorithm) 256.

Increased Cache

Tremont also features an improved cache subsystem configurable from 1.5 MB to 4.5 MB of L2 cache, depending on the product implementation. For comparison, Goldmont+ featured 2 MB of L2 cache in virtually all implementations of the chip (except for one server where L2 cache went up to 16 MB but scaled up to 16 cores and a much higher TDP). Having more cache allows for a higher hit rate, reduces the amount of data that needs to access memory, and adds latency to the system.

Improved Total IPC

With all of these improvements, Intel claims that Tremont's new architecture delivers at least a 30% uplift on IPC compared to the Goldmont+ architecture. Furthermore, Intel claims Tremont's enhancements result in a total IPC improvement of 70% over the previous two generations of the company's low-power x86 architectures.

Conclusion

As the IT industry moves towards lower power, higher efficiency devices, chip suppliers must rethink how they deliver the right balance between performance and energy use. High-efficiency cores like those based on Intel's new Tremont architecture can be paired with high-performance cores like Intel's Sunny Cove to deliver balanced processing and power capabilities. This is ideal for applications and workloads such as consumer client computing, infrastructure (both data center and cloud) and networking.

Call to Action

Moor Insights & Strategy believes that Intel's Tremont architecture is a balanced design that will allow the company's low-power cores to have the best possible IPC while maintaining power efficiency. Pairing these cores with Intel's 'Core' architecture should result in an entirely new generation of high-performance, low-power computing. With Intel's current rate of growth in IPC, generation over generation, expect improvements in subsequent designs down the road.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR(S)

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy
Anshel Sag, Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by Intel. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.