
ALIF
SEMICONDUCTOR
Introducing the Balletto ™ family of MCUs: The industry's first BLE MCU with integrated DSP function and NPU

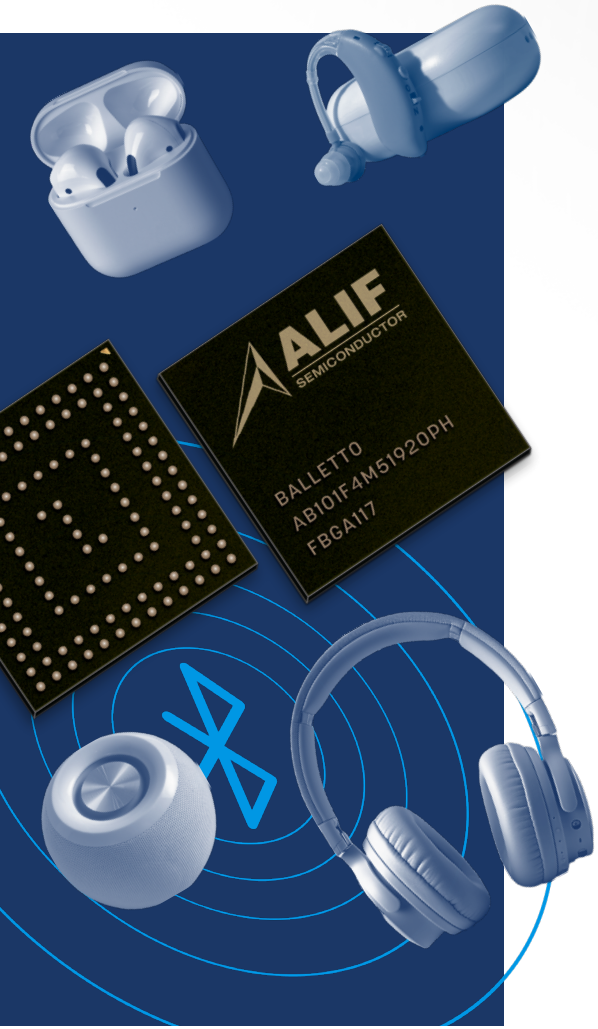

Image credit: Adobe Stocks

Introductions

From true wireless stereo (TWS) earbuds to wireless Bluetooth speakers, the Bluetooth protocol has dominated the consumer market for wirelessly streaming audio to headphones, hearables, or speakers. As such, Bluetooth audio technology is continually enhanced to support these use cases with techniques such as Advanced Audio Distribution Profile (A2DP) and more recently, the Bluetooth Low Energy (BLE) audio feature, Auracast. Seamless audio experiences have relied upon the microcontroller (MCU), audio codec IC, digital signal processing (DSP), audio digital-to-analog converters (DACs), amplifiers, etc. — which are all building blocks within these devices.

An excellent wireless audio user experience would require the integration of these components with Bluetooth capabilities while keeping in mind the severe power, size, and cost constraints. These considerations are further challenged with the introduction of compute-intensive algorithms to improve audio quality such as acoustic echo cancellation, keyword spotting, adaptive noise suppression, beamforming etc. Alif Semiconductor has entered the wireless connectivity industry as a serious contender in delivering superior wireless MCUs that actively optimize all aspects of its core functions including processing, hardware acceleration of AI/ML, power management, security, memory , and software reuse. Alif Semiconductor's Balletto family, which they announced at Embedded World 2024, is the industry's first multiprotocol (BLE 5.3 + 802.15.4) wireless MCU with integrated DSP capabilities and a neural co-processor, which eliminates the need for a separate audio codec and DSP chip. As this whitepaper will discuss, these features greatly reduce the design challenges involved in building audio applications from the bottom up, simplifying the design process while optimizing for power, space, and cost of the end-product.

The importance of Bluetooth for audio

The proliferation of Bluetooth in consumer devices

Annual Bluetooth device shipments exceeded 7 billion in 2023, with the number expected to rise to over 10 billion units by 2028. Central or platform products (e.g., smartphones, tablets, PCs, and TVs) account for a third (2.79 billion units) of the volume but are forecasted to decrease to less than a quarter of the volume in 2028 with the growth in the number and types of peripheral devices (e.g., earbuds, microphones, speakers, cameras, etc.), which will expand from two-thirds of the market in 2023 to over three-quarters (7. 32 billion units) in 2028. More than 95 percent of all Bluetooth devices will include BLE in 2028. BLE streaming audio (LE Audio) enabled by the Low Complexity Communication Codec (LC3) and its broadcast audio feature, Auracast are key enablers of this growth in BLE devices. LE Audio is expected to grow from less than 2% of the total volume of Bluetooth Audio devices in 2024 to almost 20% of the volume in 2028, registering a 10x growth rate. Although OEMs will toe the line so as not to limit their addressable market to those with LE Audio source devices only, by supporting dual-mode (BT Classic + LE Audio) until most source devices (central or platform products) have transitioned to support LE Audio, LE only single-mode audio devices will still grow to account for a significant volume of 500 million units. However, the limitations of Bluetooth Classic in audio quality and higher power consumption all but guarantee a future of Bluetooth Audio that is based on the LE Audio standard.

The advantages of the LC3 audio codec and Auracast

LE audio — with Auracast broadcast capabilities, isochronous channel support, and low complexity communication codec (LC3) codec support — offers performance leaps when compared to Bluetooth Classic. The purpose of the audio codec is to compress an audio stream from the source to ease either the transmission or storage of the data for playback (i.e., transmit audio data wirelessly over a Bluetooth link). A higher compression ratio, for instance, may ease transmission but will take more processing power and inevitably increase latency. In the case of a person listening to the audio stream of a speaker, the delay could introduce lip sync error. The increase in computational power will also drain the battery on the wearable device rapidly. The LC3 codec offers a good bit rate, high-quality compression, low computational complexity, low latency, and improved interoperability.

Image credit: Adobe Stocks

Compared to SBC, the standard codec used with Bluetooth Classic, LC3 offers more compression without degrading audio quality [2]. The combination of these factors results in a more optimal codec for consumer earbuds that are increasingly adopting more complex features such as true wireless stereo (TWS) — an application that would require as much power optimization as possible.

Additional features such as Auracast will also encourage the adoption of LE audio. This newly introduced function enables audio sources such as smartphones, laptops, TVs, PA systems, etc., to broadcast audio to an unlimited number of receivers such as earbuds, headphones, and speakers. The use cases for this particular feature are endless: from consumers simultaneously listening to the audio stream from a movie on their respective BLE-enabled audio devices, to PA systems in retail centers streaming critical announcements in public spaces for safety purposes.

Balletto: Features and capabilities

A quick overview

Alif Semiconductor's Balletto wireless MCU has BLE and 802.15.4 (e.g., Thread, Matter, Zigbee) capabilities, meaning it has full support for the audio quality improvements and power optimization found in LE audio. As shown in **Figure 1**, some key features are as follows:

- BLE 5.3 + 802.15.4 radio subsystem
- 160 MHz Arm® Cortex®-M55 CPU with Helium™ M-profile vector extension DSP
- Dedicated Arm Ethos™-U55 Neural Processing Unit with 128 MACs per cycle, up to 46 GOPs performance
- Large on-chip memory, 2 MB MRAM and 2 MB SRAM, plus an Octal SPI interface for external memory expansion
- Digital interfaces: I3C, I2C, SPI, USB HS, 2x CAN-FD, MIPI-DSI, SDIO, SD/MMC, and more
- Analog Interfaces: 24b Sigma-Delta ADC, 12b SAR ADCs, 12b DAC, and Comparators
- Package sizes sub-16 sq.mm available.

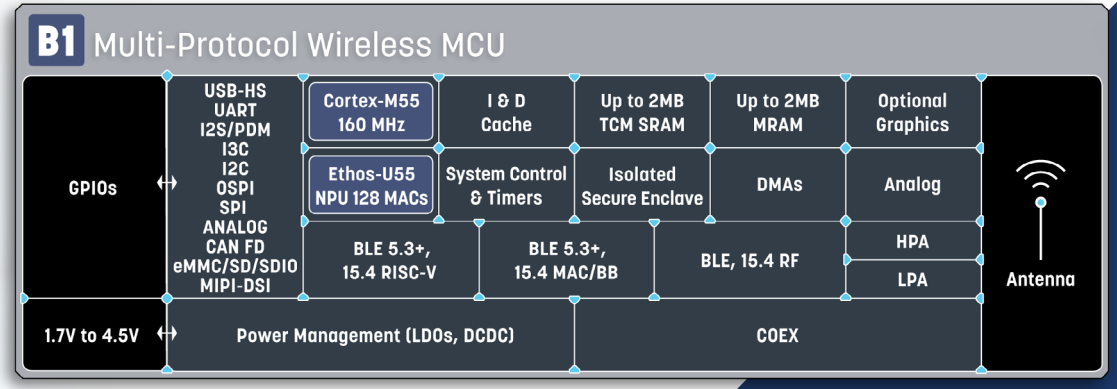

Figure 1: Block diagram of Balletto wireless MCU.

Large on-die memory and GPIO

A large on-die memory with up to 2MB of high-speed MRAM and 2MB of zero-wait state SRAM has enough memory to store and process pre-trained AL/ML models. With support for up to 77 GPIOs that can be multiplexed with many different interfaces to host a wide array of sensor functions such as digital sensing (gyroscope, accelerometer, IR, audio, LEDs, etc.) and analog sensing (light, blood pressure, EKG, skin temperature, etc.), all manner of sensor fusion is possible.

The power of integrating DSP

Instead of incorporating a separate audio codec chip to encode and decode compressed audio data, the integrated Cortex-M55 core with its integrated Arm Helium™ M-profile vector extension (MVE) provides instructions that are capable of digital signal processing such as Fast Fourier Transforms (FFT), which form the basis of audio functions resulting in a 500% improvement in DSP algorithm performance compared to a standard Cortex-M class core without Helium support. This eliminates cost and additional board complexity, lowers power consumption, and helps extend the battery life of Bluetooth LE earbuds, headphones, and other LE audio devices.

The benefits of a built-in NPU

The MCU's Ethos-U55 NPU can perform up to 46 giga-operations per second (GOPS) to extend its capabilities, with the acceleration of AL/ML models to support functions such as echo cancellation, noise suppression, keyword spotting, beam forming, and voice assistance — functions that would typically require a separate neural net processor for hardware acceleration. This particular core (Ethos-U55) excels at handling these AI/ML models with major performance improvements over other traditional MCUs (See Section 4).

Cost, space, and power-optimized

Finally, Balletto comes in an ultra-small sub-16 sq.mm package that ensures the device can more easily be integrated into consumer-wearable applications that have some of the most space- and power-constrained requirements in the industry. The Bluetooth radio receive-current of 1.5 mA and transmit-current of 2.0 mA in combination with the low system power consumption driven by Alif Semiconductor's Autonomous Intelligent Power Management (aiPM™) technology further bolsters Balletto's power optimization and help extend battery life between charges by dynamically powering only the logic and associated memory that are in use at any given time. This technology has several power modes including run mode (22 µA/MHz), ready mode, idle mode, standby mode, and stop mode (700nA). With on-board DSP, NPU, and memory blocks, the high integration of this device lowers BOM count, saving both space and cost.

Figure 2: A small package ensures Balletto can fit into some of the most space-constrained audio applications

A glance at the Balletto family's audio performance metrics

AudioMark benchmark

Built around the Arm® Cortex®-M55 CPU with Helium M-profile vector extensions (MVE) and the Ethos-U55 NPU, the Balletto wireless MCU outperforms many standard MCUs in audio performance metrics. **Figure 3** shows the AudioMark relative cycle per the specific

KWS neural net function normalized to the Cortex-M4 showing massive acceleration on the M55 core.

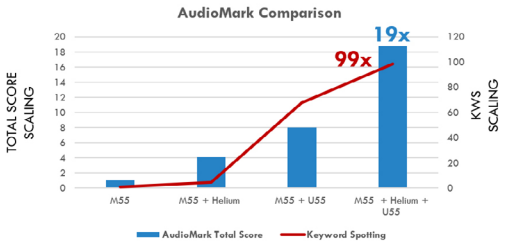

Figure 3: AudioMark total score benchmark shows a stunning 19x improvement.

The AudioMark benchmark provides a sophisticated porting layer that enables it to utilize available compute hardware in a way that reflects real-world product architecture. As shown in **Figure 4**, the benchmark takes into consideration the range of audio technology today from beamforming and direction of arrival, as well as modern filters such as KWS, acoustic echo cancellation and noise suppression. As an added benefit, AudioMark allows for the integration of acceleration such as DSPs or other dedicated audio hardware without allowing any singular technology to dominate. For example, the main function can run on a host MCU while offloading the signal-processing functions to a DSP co-processor, and then send the results to a neural net processor for inference. It is important to note that **Figure 3** shows only the improvements with the KWS function.

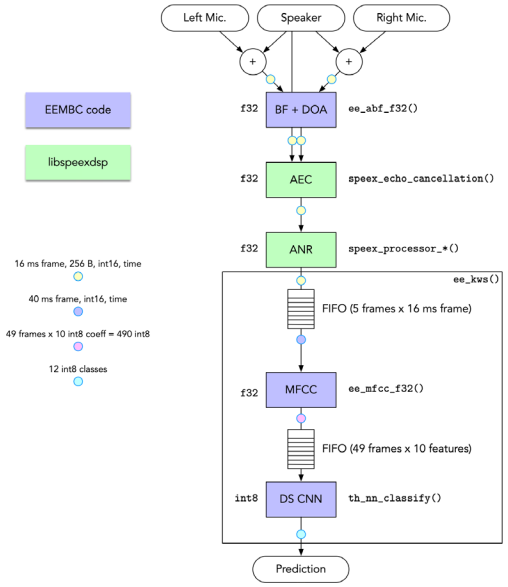

Figure 4: A rough outline of the AudioMark benchmark's pipeline (excluding physical transducers) where key components consist of: spectrum decomposition analysis, direction of arrival (DOA), beamforming (BF), acoustic echo cancellation (AEC), single-channel noise suppression, feature extraction, and neural net classification. Source: EEMBC

A basic rule of thumb is that Helium can accelerate most DSP operations by 2.5x to 3.5x. This information aligns with the 4.1x gain in speed seen in **Figure 3** when the M55 core is used with the Helium MVEs. In other words, the benchmark is accurately portraying system improvements.

The addition of the Ethos™-U55 NPU then takes the system far beyond the 4x improvement, as portions of the audio pipeline can be executed as a neural network rather than as an algorithm.

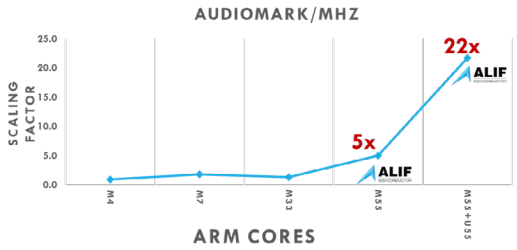

Figure 5: AudioMark/MHz performance metrics of M55 core versus other Arm cores on the market.

The AudioMark per MHz score of various Arm cores including the M4, M7, M33, M55, and M55+U55 is a good way to benchmark the performance of each core on the same workload as shown in **Figure 5**. The AudioMark/MHz score is the normalized AudioMark score obtained by dividing it by the fastest execution clock used in the system, which characterizes the overall compute efficiency of the platform. The results show that the M55 core yields a 5.8x improvement over the baseline M4 CPU; the true power of the NPU comes into the picture when combined with the M55 core, delivering a truly astounding 22x improvement in the AudioMark/MHz score.

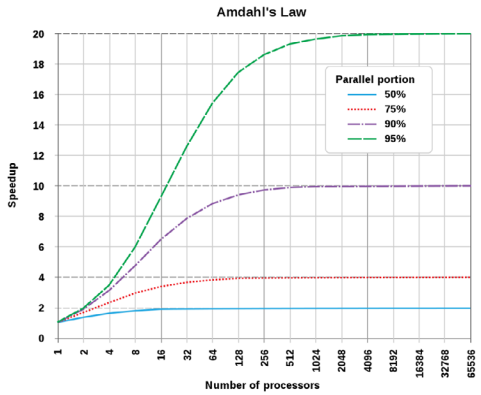

Figure 6: Theoretical speedup improvements to the execution of a task with an increasing number of processors, according to Amdahl's law.

The theoretical latency improvements to the execution of a task as a function of the number of processors executing it is governed by Amdahl's law (a formula). As shown in the graph in **Figure 6,** with the Ethos-U55, the system can take about 90% of the AudioMark work and speed it up by ~100x, achieving an 8x overall speed (see purple line). However, combining this improvement with the 3x speedup on the rest brings the system to a 18.8x speedup (~19x improvement seen with M55 + Helium + U55).

ML inference benchmarks

Increasingly, clear voice call quality and exceptional sound quality is driven by eliminating background noise as well as enhancing input audio streams by using advanced AI/ML algorithms. These machine learning algorithms are able to provide enhanced capabilities and optimally run on-device in the powerful neural processor and deliver orders of magnitude speedup and therefore lower overall system power consumption.

Table 1 shows interference benchmarks using open source models including keyword spotting (KWS) tasks with the MicroNet Medium model supplied by Arm, object detection with YOLO-Fastest, and auto speech recognition with Tiny ASR. As shown in the far right side of the table, the combination of the Arm® Cortex®-M55 CPU with Helium M-profile vector extensions (MVE) and Ethos™-U55 NPU offers up to a 109-fold improvement in inference time over Cortex-M55 alone, while the combination of the three (Ethos + M55 + Helium) offers up to a 104-fold reduction in energy efficiency over using the Cortex-M55. The combination of these three functional blocks is essential in minimizing both latency and power consumption of key ML algorithms.

High Efficiency (HE) System: Cortex-M55 and Ethos-U55 128MAC at 160MHz										
Model	Accelerated inferencing				CPU-bound inferencing (on Cortex-M55)				Improvement w. Acceleration	
	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time	Energy efficiency
KWS[1]: MicroNet Medium (ARM)	15.9	8.8	2.6	0.14	326	3.4	1.0	1.27	21x	19x
Object Detection[2]: YOLO-Fastest (face trained)	18.6	14.2	4.2	0.27	1373	5.4	1.6	8.3	74x	67x
Auto Speech Recognition[4]: Tiny ASR (Wav2letter)	78.6	10.0	3.0	0.69	8562	7.4	2.2	62.5	109x	104x

Table 1: Time, power, current, and energy used with a combination of the Cortex-M55 core, Ethos-U55 NPU, and Helium performing inferences for different ML models.

Why use the Balletto in audio wearables and speakers?

Bluetooth audio has permeated most consumer wearables and speakers, granting wireless access to music, voice recordings, video streams, and more. In the quest to remove wires and encourage a more ergonomic, hands-free design, TWS earbuds have come to the forefront of many wearable designs. TWS earbud solutions typically carry distinct functional blocks including the Li-ion battery, charger/PMIC, audio codec IC/DSP, MCU, wireless chipset (e.g., Bluetooth radio), external storage, various status LEDs, as well as sensors/transducers such as microphones, speakers, and accelerometers. Many of these blocks can be seen in Bluetooth speakers as well, with the addition of amplifiers before transmitting the audio through the speakers.

The Balletto Family manages to incorporate the MCU with DSP capabilities, an NPU, and a PMIC all in a singular chip, as well as large amounts of memory to run complex modern audio algorithms — successfully combining the bulk of the TWS earbud BOM cost and real-estate into a smaller than 16 sq.mm footprint. Moreover, the utilization of the Cortex-M55 core with Helium and the Ethos-U55 NPU massively accelerates the audio codec processing as well as key algorithms that allow for a high bitrate, high quality, and a more seamless sound experience (e.g., AI noise cancellation, background noise suppression, keyword spotting, echo cancellation, and beamforming). The underpinning aiPM™ technology in tandem with the already low-power BLE protocol encourages smart power utilization, ensuring that the battery life between charges is maximized. All of these factors combined ensure the Balletto family of BLE-enabled microcontrollers facilitates next generation Bluetooth-enabled audio devices.

Image credit: Adobe Stocks

References

1. LE Audio: The Future of Bluetooth® Audio, Market Research Note

2. Bluetooth® Technology Website. (2020, November 2). A technical overview of LC3. Retrieved from https://www.Bluetooth.com/blog/a-technical-overview-of-lc3/

https://alifsemi.com/

ALIF SEMICONDUCTOR