Sensor Fusion
Intelligence at the micro-edge
Intelligence at the micro-edge

The Internet of Things is an evolutionary transformation that couples the “things” all around us to cloud-based applications and services. Today, most connected things are, in effect, peripherals of cloud computing. Consequently, most IoT solutions require continuous connectivity and move lots of data to the cloud. The need for cloud connectivity restricts IoT use-cases to those that are networked, tolerant of occasional service interruptions, and OK with sending data to the cloud.

“Intelligence at the micro-edge” removes these restrictions – it’s the next step beyond IoT. The “micro-edge” is all the small, low-power devices like sensors and actuators at the edge of the network. “Intelligence,” in this context, means micro-edge devices that are capable of running autonomously with minimal cloud service dependencies. Cloud-based IoT services such as data aggregation, analysis, and correlation are already beginning to migrate from large server clusters to small micro-edge devices. These services increasingly use advanced AI techniques to correlate multiple sensors, learn normal behavior, and detect anomalies.

New, power-efficient chips capable of running computationally intensive AI-based applications, including machine learning and training, are powering this transformation. On-device processing eliminates the need to move sensor data to the cloud, thereby improving data privacy, response time, and reliability while reducing network bandwidth and cost.
Introducing the Ambient Scientific GPX-10

The GPX-10 is the first low-power chip explicitly designed for AI-enabled applications at the micro-edge. The chip brings on-device AI inference and training to a broad range of devices, even battery-powered ones, with no need for continuous cloud connectivity. It’s a game-changer because product developers can now integrate high levels of edge intelligence into a broad range of real-world products at low cost.
Sensor Fusion

Sensor Fusion combines inputs from multiple sensors into a unified view of a physical environment. Making sense of multiple inputs requires correlating and modeling combinations of sensor data streams, learning normal behavior, and looking for patterns and anomalies. AI techniques are the best way to do this, but today it’s usually done in the cloud. Moving Sensor Fusion to small devices at the micro-edge requires new silicon capable of performing AI inference and learning. Let’s take a closer look at two types of Sensor Fusion applications – consumer wearables and industrial monitoring.
Wearables measure many biometric and environmental parameters. Transforming these measurements into useful information requires analyzing combinations of data from multiple sensors – AI-enabled Sensor Fusion. For instance, changes in pulse and blood pressure might or might not be alarming depending on the user’s activity level. Adding machine learning to Sensor Fusion personalizes the analysis for each user to more accurately detect unusual patterns. Here are some specific product categories that are already adding advanced, AI-enabled Sensor Fusion capabilities:

**Smart watch** – Today’s smart watches are bristling with sensors for motion and biometrics. Although these watches have large processors that are capable of Sensor Fusion, product developers want to reduce power consumption during routine monitoring – and that requires new, AI-enabled, low-power silicon.

**Smart shoes** – By fusing data from multiple sensors, smart shoes can continuously measure stride length, gait, gait symmetry, foot angle, and ground reaction forces. In-shoe electronics must be extremely small, rugged, and power-efficient. This requires a single chip, AI-enabled solution that runs for weeks on a tiny battery.

**Activity monitor** – Armbands and rings with biometric and motion sensors detect and measure walking, running, and sleeping at price points well below smartwatches. AI powered Sensor Fusion is powering the next generation of applications such as learning and tracking any exercise activity, detecting falls and accidents, recognizing gestures, and analyzing temperature trends to recognize illnesses such as COVID-19. Battery life is critical because these devices are always on, and recharging isn’t convenient.

**Pet monitor (tag)** – Monitoring changes in dog and cat activity over time helps detect early-onset health and nutrition problems. Data collection, fusion, and reduction must be done on the tag at very low power with minimum smart phone interaction.

**Biometric earbuds** – Earbuds can collect surprisingly large amounts of biometric data including heart rate, blood pressure, blood oxygen, temperature, respiration, and footsteps. Earbuds typically depend on smartphones for data collection and analysis, but in-ear Sensor Fusion allows untethered operation while dramatically increasing battery life for both the earbud and the phone.
Equipment and building monitoring applications use Sensor Fusion to combine individual measurements into comprehensive models that accurately reflect real-world situations. Similar to wearables, these applications use AI techniques to infer actionable events. They also learn normal behavior to better detect anomalies with fewer false alarms. Here are just a few of the many product categories that are using these techniques.

- **Predictive maintenance** – machinery, manufacturing, appliances – It’s much cheaper to fix equipment at the first hint of a problem rather than wait for complete failure. Unusual combinations of noise, vibration, temperature, and power consumption provide early warnings for maintenance personnel. AI-based training continuously models normal operational characteristics to recognize anomalies more accurately.
• **Fire alarms** in commercial buildings also benefit from Sensor Fusion. Combining sensor data such as smoke, heat, and CO improves first responders' situational awareness while reducing costly false alarms. Adaptive training allows sensors to learn what's normal for a specific area, maintaining sensitive detection thresholds while lowering the risk of false alarms.

• **Automotive** – Modern cars and trucks have hundreds of sensors. Sensor Fusion aggregates related sets of measurements into a comprehensive model with a single connection to the car network. Fusion simplifies wiring can remain active at all times, drawing very little power.

These micro-edge applications all require a silicon platform capable of bringing low-power, high performance, autonomous AI inference, and learning to small devices at low cost. No such chip has been available – until now.
The Ambient Scientific GPX-10 is the first chip that addresses the full set of requirements for “intelligence at the micro-edge.”

- High AI performance – Cubic Computing for on-chip inference and training
- Complete solution – Single-chip with all subsystems needed for sensor fusion
- Always on – Low power inference (suitable for battery-powered devices)
- Autonomous – On-chip AI for inference, training, and re-training
- Programmability and flexibility – Adapt to a wide range of AI algorithms
- Low platform cost – High-volume CMOS silicon

Let’s take a closer look at how the Ambient Scientific GPX-10 uniquely addresses each of these requirements.

**High AI performance – Cubic Computing**

The heart of the GPX-10 is the DigAn Cubic Computing Engine, a custom digital-analog AI processor designed to accelerate neural network algorithms (NNs) at disruptively low power levels (Figure 2). Cubic computing accelerates two characteristics of NN workloads – repetitive mathematical computations and matrix data structures.

![Figure 2: DigAn Cubic Compute Engine](image-url)
**Cubic Matrix Processor** – The fundamental NN mathematical computation is matrix multiplication, which performs multiply-and-accumulate (MAC) operations on arrays of operands. NN acceleration requires doing many MAC operations simultaneously. Traditional digital multipliers can do the job, but these complicated circuits consume significant power and chip area, both undesirable for embedded applications. The GPX-10 minimizes power and area by using analog techniques for MAC computations. The fundamental science of analog multiplication is solid, dating back to Kirchhoff’s current law in the 19th century. Although other AI accelerator chip manufacturers also use analog MAC techniques, Ambient Scientific’s implementation is uniquely suited for NN applications, delivering deterministic results at variable resolutions up to 32-bit with superior performance, power efficiency, and flexibility.

**Cubic Memory** – The operands and results of neural network MAC computations are matrices. SIMD engines (i.e., GPUs) accelerate matrix math by processing rows and columns of operands in parallel. The GPX-10’s DigAn Cubic Computing Engine takes this idea to the next level by simultaneously processing arrays of operands. This requires a unique memory architecture capable of dispersing operands to the entire compute matrix in a single cycle. Here’s a simple way to imagine the concept – SIMD is “planar” computing on rows and columns while DigAn is “cubic” computing on entire arrays.

The DigAn Cubic Computing Engine combines low power analog matrix processing with high-performance Cubic Computing to run sensor fusion applications at exceptionally low power levels. Each DigAn core can perform 256 MACs per cycle. The GPX-10 has 10 cores, so it can do 2,560 MACs per cycle, yielding an impressive 512 peak GOPS and, more importantly, 4.3 TOPS per watt.
Complete solution

A single GPX-10 has all the processing functions required to implement complete Sensor Fusion applications. As shown in Figure 3, there are two subsystems on the chip. The “Always ON” subsystem has a multi-channel input block suitable for multiple analog and digital sensors. Digitized inputs flow into a 5-core AI engine that runs neural network applications at extremely low power. The “Processor Subsystem” uses an Arm M4F to run non-AI application code, while a second 5-core AI engine can enhance inference performance. The Processor Subsystem is powered down during routine sensor monitoring when only the Always ON block is needed.

Figure 3: GPX-10 SoC architecture
**Always-on**

Sensor devices must be on all the time. For battery-powered products, this requires power consumption measured in microwatts, not milliwatts. Low power is also desirable in mains-powered devices because manufacturers want to minimize power consumption in all products.

The GPX-10’s extremely low-power inference engine is ideal for sensor fusion applications. The Always ON block runs continuously, even on battery power, collecting multiple sensor data streams and analyzing them for unexpected or out-of-range behavior. When the Always ON block recognizes an anomaly, it wakes up the M4-based Processor Subsystem for in-depth analysis and event triggering. With the M4F asleep, power consumption during anomaly detection is an industry-leading ~80uW, low enough for small, battery-powered devices.

**Autonomous inference and training**

Today, most IoT sensors at the micro-edge depend on cloud-based services. Autonomous operation is highly desirable for four key reasons.

1. Monitoring must work all the time, even if network connectivity is not available.
2. Processing sensor data on-device eliminates the privacy problem of sending data to the cloud.
3. Autonomous devices can respond immediately to anomalies without a round-trip to the cloud.
4. For low-power devices, radio subsystems consume a surprisingly large amount of power. Autonomous devices can leave the radio off during routine operation.

For these reasons, AI-enabled micro-edge processors must be powerful enough to run stand-alone at the micro-edge, even for training tasks such as learning local conditions and adapting to them.

The GPX-10 is powerful enough to run complete AI inference applications on-chip. The chip also has unique features that enable ML model training and re-training, not just inference. For training, the ability to handle high-resolution input data (32-bit) is critical. The AI engine and its Cubic Memory have a unique design capable of scaling up to 32 bits with back-propagation for adaptive training, making the complete machine learning solutions autonomous.
The same features that enable training are also useful for DSP computations. The matrix math is similar, 32-bit resolution is also essential, and the Cubic Computing cores can accelerate AI and signal processing in the same application without halting the engine. For example, the Always ON block can perform input data conditioning and AI-based anomaly detection continuously without waking the Arm processor.

**Development flexibility**

Software development is always a large part of the total solution cost, often a much larger cost for AI-based applications. As shown in Figure 4, developers can use standard, off-the-shelf Arm compilers and toolchains for the M4F-based Processor Subsystem. The Ambient Cubic Compute engine has its own compiler, driven by a “Segregator” front-end that simplifies AI core management and integrates the two environments.
Using multiple models requires considerable architectural flexibility. The DigAn cores are software-programmed and capable of accelerating a wide variety of AI algorithms. Programmers can dynamically configure AI features at run-time, including operand resolution and power level, without stopping the machine. Multiple applications (i.e., inference and re-training) can run simultaneously on different cores. The GPX-10’s flexibility reduces development costs by empowering AI programmers to use familiar languages, tools, and techniques without worrying too much about the AI engine’s implementation details.

There are countless examples of AI-enabled Sensor Fusion from every vertical industry – automotive, smart city, healthcare, energy, manufacturing, transportation, and agriculture are just a few. Most of these sensor applications share similar requirements for high AI performance, low power operation, autonomy, and low cost. A platform designed around the GPX-10 can cover many applications across these diverse verticals because it can adapt to many AI algorithms without hardware-specific tuning.
Low platform cost

The GPX-10 is a complete single-chip sensor fusion solution with an application processor (Arm M4F), an always-on AI engine (DigAn), and multi-channel analog sensor inputs. Very few external components are required. The silicon process node is TSMC 40nm with no special transistors, capacitors, or inductors. It's just plain CMOS, designed for low-cost, high-volume production.

Summary

The GPX-10 is ideal for moving Sensor Fusion intelligence from the cloud to the micro-edge. It’s the first embedded chip that delivers the computational power required to run advanced AI models on-chip, without cloud connectivity, and at power levels low enough for battery-powered devices. Solution cost is minimal because it's a flexible, easy to program, single-chip AI processor built on a standard 40nm semiconductor process node.

Ambient Scientific enables the next step beyond IoT – intelligence at the micro-edge.