cadence®

Keeping Always-On Systems On for Low-Energy Internet-of-Things Applications

By Gerard Andrews and Larry Przywara, Cadence Design Systems

"Always on" processing is no longer a buzz phrase—it's a reality and a necessity in an array of applications, from mobile to health and fitness to smart homes. One of the main challenges in developing always-on systems is ensuring continuous function while minimizing power consumption. This paper examines how a novel processor offloading technique, along with configurable digital signal processing (DSP) technologies, can enable always-on computing within today's practical constraints on power consumption, enriching the user experience in the process.

Contents

Introduction

You greet your personal digital assistant by its given name; hearing your voice, it wakes up, says hello, and reminds you of your day's scheduled meetings and appointments.

You look at your smart watch, which is in a power-saving sleep mode, and it turns on when the watch detects your face.

At the end of each month, your phone, having monitored stats such as your activity level, sleep patterns, and heart rate, provides you with a report outlining your overall well-being.

All of these systems are driven by sensor data that is enabled by always-on processor technology—where some compute resources in a system are "always on" to process audio, visual, or other sensor data while the more powerful compute resources in the system are turned off. This is a common requirement in mobile systems, wearables, and for Internet of Things (IoT) devices. A new generation of products is redefining the man-machine interface and creating compelling new user experiences.

From a design standpoint, developing always-on systems carries the challenge of delivering the right level of processing while minimizing power consumption. Always-on functionality is simply not possible if these applications must run on powerful general-purpose processors that are often running a high-level OS (HLOS). The power consumption is prohibitive. Running these always-on functions on low-power DSPs provides the appropriate performance and power consumption to enable designers to incorporate this capability into their designs. In this paper, we'll discuss a unique processing paradigm—called cognitive layering—that addresses the key challenges of implementing always-on technology and helps designers create the systems that are at the heart of a rich, engaging user experience.

Always-On Architecture

A typical system on chip (SoC) for an IoT application consists of rich analog and optimized digital components. On the analog side, think radio, media access layer (MAC), and baseband; low noise amplifier (RX); power management unit; and integrated power amplifier. On the digital side, there's the low-power processor; on-chip memory; digital baseband hardware blocks; and rich sensor I/Os.

Sensors are at the heart of the connected devices that are driving the IoT. Microphones, cameras, accelerometers, gyroscopes, temperature, and pressure are just a handful of the many sensor types that are becoming increasingly pervasive in the things that we use in the home, at work, and at play. These sensors are collecting digital data that is sent to the SoC for analysis and interpretation. As a result of this activity, we can use a voice trigger to get driving directions from our phones; our home heating systems turn on when the equipment senses we are home and the room temperature has dropped below a set threshold; and we can unlock our office doors by peering into a security screen.

To reduce energy usage in IoT applications, designers can develop their system architectures with techniques including:

- Cognitive layering
- Power, clock, and data gating
- The use of sensor fusion algorithms to determine device context and intelligently manage the system power based on this context
- Optimization for certain "hot" functions, for example:
  - Reducing cycles and lowering MHz at the instruction set level
  - Bypassing the general on-chip bus at the interconnect level
  - Localizing traffic in memory partitioning
  - Accelerating communications standards as well as performance for voice algorithms, encryption, and the like

Typical IoT Components

1. Wireless transmitter
2. Off-chip memory (DDR, Flash)
3. Sensor data filtering, fusion, and reduction
4. Data compression, encoding, encryption
5. User interface/display
6. Wireless receiver
7. Communications protocol processing
8. Higher level control processing

Let's take a closer look at cognitive layering, the approach that addresses the functionality vs. power consumption challenge.

What is Cognitive Layering?

Cognitive layering is the partitioning of a processing task into layers or states that can be addressed by an appropriate processing engine. A result of this approach is the offloading of a host processor with lower power always-on processors. As shown in Figure 1, each layer has just enough processing to support the level of alertness required by the system at that point. This type of parallel processing architecture improves latency, energy, and throughput.

In these systems, optimized processing engines are often used in lieu of general-purpose processors, especially in the lower layers. The optimized processors and their software are adapted to the specific computing and interface requirements of those lower layers, offering higher performance, shorter response latency, lower energy, and lower cost, compared to using general-purpose processors. The architecture is designed so that as much of the silicon is kept "dark", or off, as possible. This approach moves computation closer to the data and minimizes energy by keeping the system at the lowest activity level required. In fact, this approach also is aligned with the new cloud-influenced data imperative to "compute locally and share globally." According to this imperative, latency-driven computing remains on the device (local), while throughput computing along with shared data-aggregation often migrate into the cloud (global).

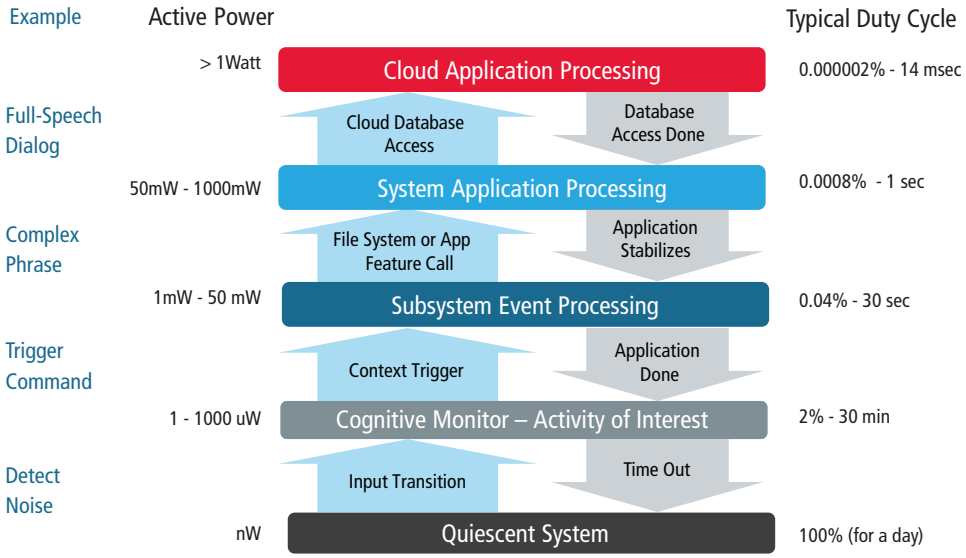

Figure 1: Cognitive layering minimizes power consumption

Considering the flow diagram in Figure 1, let's look at how this layered structure could apply to a voice-triggering application on, say, a smartphone. At the bottom step in this flow is low-energy noise detection, consuming just nanoWatts of power. Once noise is detected, this action triggers a series of system actions up the processing chain – detection of a trigger phase, recognition of a command, interpretation of the commands in the context of the current application – occasionally culminating in the higher energy act of full speech dialogue via a cloud service that calls for multiple watts of power to access and interact with remote servers and databases. This essentially illustrates how low energy drives "layered cognition" – the next level of power consumption is triggered only when it is needed. Going back to the new data imperative, the noise detection and command triggering essentially involves computing locally, while the full speech dialogue level involves sharing globally.

The same layering opportunity appears in many other applications domains – in inertial navigation, in computer vision, in interface to real-time sensors and actuators, in interactive graphics, and in local wireless communications. The same opportunities for dramatic performance and energy improvement recur again and again.

What to Look for in an Always-On Processor for IoT

A typical always-on application is wake-up processing, implemented for a function such as voice trigger. An always-on voice trigger will generally have two modes: sound detection and keyword detection. The voice trigger functionality can be paired with command and control voice recognition that's processed locally. What type of DSPs are ideal for these applications?

Clearly, very efficient, small, low-power processors are ideal to support a cognitive layering approach. "One size does not fit all"—multiple specialty DSPs would generally be needed to offload the host processor in the design. For example, a DSP that's suited for low-power voice triggering probably won't support more intensive audio processing like, say, decoding surround sound effects. Ideally, there should be one dedicated processor to provide the always-on capability in a device and this should be versatile enough to support a range of always-on functions. So, in addition to supporting voice trigger, the processor should also be efficient at functions such as sensor processing and low-resolution image processing.

When seeking processors to fill these roles, important criteria to consider include:

- Low power consumption
- Configurability (only include what's needed)
- Timing closure, typically no more than a couple hundred MHz
- Gate count vs. speed over the intended operating range
- Availability of high-precision MACs for DSP operations

Single, Scalable DSP for Low-Energy, Always-On IoT Functions

Cadence's new Tensilica® Fusion DSP uniquely meets the criteria for the always-on functionality required by IoT applications. Derived from the highly successful Cadence® Tensilica HiFi 3 DSP for audio/voice/speech, the configurable Tensilica Fusion DSP delivers ultra-low energy implementations with minimum clock rate requirements and reduced code size.

Designed to be flexible and scalable, the Tensilica Fusion DSP is also suited for wearable and wireless connectivity applications, providing an ideal platform for innovation. Voice triggering, also commonly called keyword spotting, provides an example of the energy efficiency of the Tensilica Fusion DSP. In the case of Sensory's always-on voice trigger, the Tensilica Fusion DSP uses less than 80% of the energy running the solution, compared to the current industry-leading Tensilica HiFi Mini DSP.

Configurable elements in the Tensilica Fusion DSP include:

- Single-precision floating-point unit (FPU), where floating-point instructions are issued concurrently with 64-bit load/store, speeding software development of algorithms created in MATLAB or in standard C code
- Audio/voice/speech (AVS), which has software compatibility with the Tensilica HiFi DSP and is backed by access to more than 140 HiFi audio/voice software packages
- 16-bit Quad MAC (adds a four 16x16 operation), which further accelerates communications standards like Bluetooth Low Energy and Wi-Fi, along with voice codec/recognition algorithms
- Encryption acceleration for Bluetooth Low Energy/Wi-Fi AES-128 wireless operations
- Advanced bit manipulation, which accelerates implementations of baseband MAC and PHY, including LFSR (linear feedback shift register), CRC (cyclical redundancy checking), convolutional encoding, and bit select operation for interleaving/de-interleaving
- Flexible memory architecture that works with caches and/or local memories of various sizes, depending on application

Summary

As evidenced by the popularity of voice-controlled smartphones, consumers are embracing always-on functionality. Devices that tap into this functionality—such as wearables and mobile and Internet of Things products—are proliferating. However, the traditional approach of using a host processor on a chip for a wide array of functions is not realistic for this new generation of applications.

Design engineers can deliver always-on functionality with minimum power consumption by offloading this capability to low-energy processors, such as Cadence's Tensilica Fusion DSP. The main processor in their systems can then handle the bigger processing tasks. A cognitive layering approach supports this architecture by providing a means to use only the energy necessary for a given task; the next level of power consumption is triggered only when needed by the given task at that layer. While engineers reap performance/power benefits from this approach, consumers come away with compact, energy-efficient products that deliver increasingly rich, engaging user experiences.

Built Upon the Xtensa Innovation Platform

All Tensilica DSPs are built upon the Xtensa® configurable processor, which gives you the opportunity to further customize the processor to execute your own algorithms more efficiently. Highly optimized professional development tools and hardware implementation deliverables are created specifically for your processor using the same technology that has created more than 1,000 different designs in production, with more than 2B cores shipped per year.

For Further Information

Learn more about the Cadence® Tensilica Fusion DSP: http://ip.cadence.com/fusion/

cadence®

Cadence Design Systems enables global electronic design innovation and plays an essential role in the creation of today's electronics. Customers use Cadence software, hardware, IP, and expertise to design and verify today's mobile, cloud and connectivity applications. **www.cadence.com**