Voice Assistant Devices

By

Ming Lei

Foundations of Computer Architecture

Table of Contents

List of Figures

List of Tables

Introduction

Communication is an integral part of everyday life. Even though some people only concern with people to people communication, machine to machine communication and people to machine communication are becoming increasing important. Communication is the process of sharing of information, knowledge and experiences. Many famous IT companies, such as Apple, Microsoft, Google and Samsung are spending millions of dollars on studying how to help consumers communicate computer or machine effectively. In the recent years, voice assistant or voice control devices become very popular. Voice assistant devices allow consumers to use their voice to command the device to do many things, such as control light and answer basic questions like date and weather.

Due to a wide range of voice assistant devices in the market, it is difficult to cover all the brands of voice assistant devices. This paper will focus on Amazon Echo and discuss the following parts: (1) overview of Amazon Echo; (2) processor used in these devices; (3) architecture and design; and (4) security concerns in Amazon Echo.

Amazon Echo

Amazon Echo, a smart speaker designed by Amazon Lab 126, initially released on November 2014. It looks like a cylindrical Bluetooth speaker at first glance. Actually, Amazon Echo can do more than the speaker. By using Echo, consumers may perform many different useful tasks (Haack, Severance, Wallace &Wohlwend, 2017). The following list are some well-known performance:

- Inquiring news, date, weather and other information
- Placing order from online store
- Controlling other smart devices (lights, door, window blind and thermostats)
- Interacting with third party application in your other devices

Amazon does not release the sale data of Echo, but Consumer Intelligence Research Partners estimates 20 million Echo devices are being used in the market. Therefore, Echo is considered as one of the most successful products in Amazon. The Amazon Echo is controlled by Texas Instruments DM3725 ARM Cortex-A8 Core Digital Media Processor (Johnson, 2016).

ARM Cortex-A8

The ARM Cortex-A8 is 32-bit processor core designed by ARM Holding. Due to its high performance and low power usage, ARM Cortex-A8 is widely used in mobile and consumer embedded applications, including but not limited to gaming consoles, set-top boxes, GPS systems, and mobile phones. The Cortex-A8 processor can use less than three hundred mW only for mobile devices and deliver more than two thousand Dhrystone MIPS of performance for consumer applications (Williamson). It indicates the Cortex-A8 has the increased processing capability, but keeps the power consumption of previous generations of mobile devices. Besides Amazon Echo, Cortex-A8 is used in many other popular mobile devices, such as iPhone 4 and

Samsung Galaxy S. The Figure 1 below is a block diagram of the Cortex-A8 chip. By reading this diagram, it is not difficult to understand how each component works together within the microprocessor. The ARM Cortex-A8 was considered as the first ARM Cortex design to be adopted on a large scale in consumer devices. It incorporates some new technologies in the ARMv7 architecture as shown in the following list:

- Thumb®-2 technology
- NEON™ Data Engine
- VFPv3 floating point

The Cortex-A8 also incorporates other technologies, such as TrustZone for data privacy (Aneesh). However, this paper will focus on the above three and discuss the details in next sections.

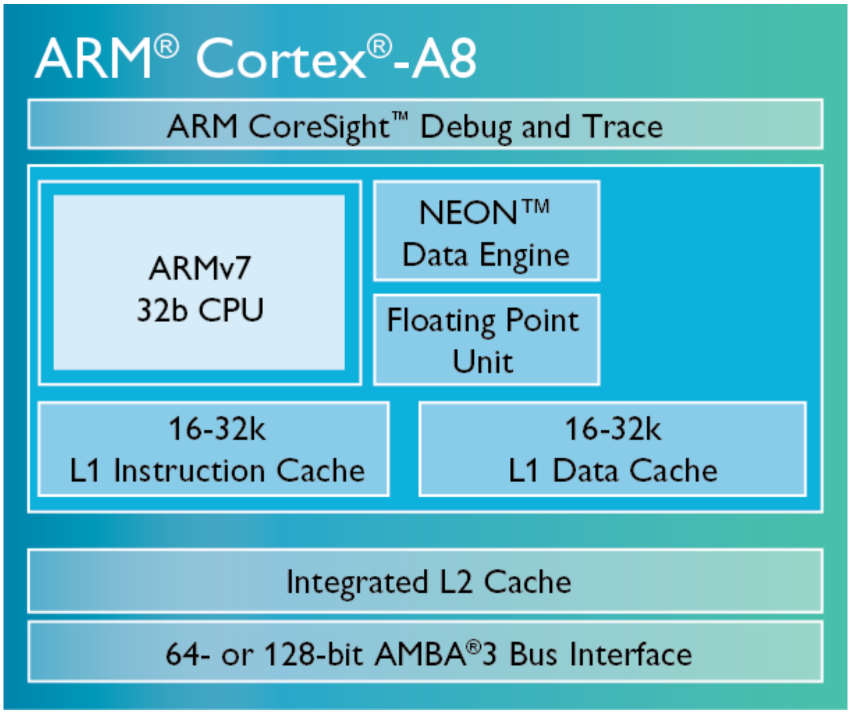


Figure 1: Cortex-A8 Chip Diagram

Architecture

The ARM Cortex-A8 is a family member of reduced instruction set computing architectures (RISC) for processors. ARM stands for Advanced RISC Machine. In general, RISC architecture needs fewer transistors compared with the complex instruction set computing architecture (CISC). Cortex-A8 is a 32-bit RISC architecture with 16 registers and a Harvard memory architecture (Hoffman & Hedge, 2009). The Harvard architecture was developed at Harvard University by Howard Aiken and others. The unique feature of Harvard architecture is the presence of separate instruction and data memories. This allows one instruction to be fetched while another stores or reads an operand. The low power consumption and good heat dissipation are the reasons why ARM is widely used in terms of quantity produced.

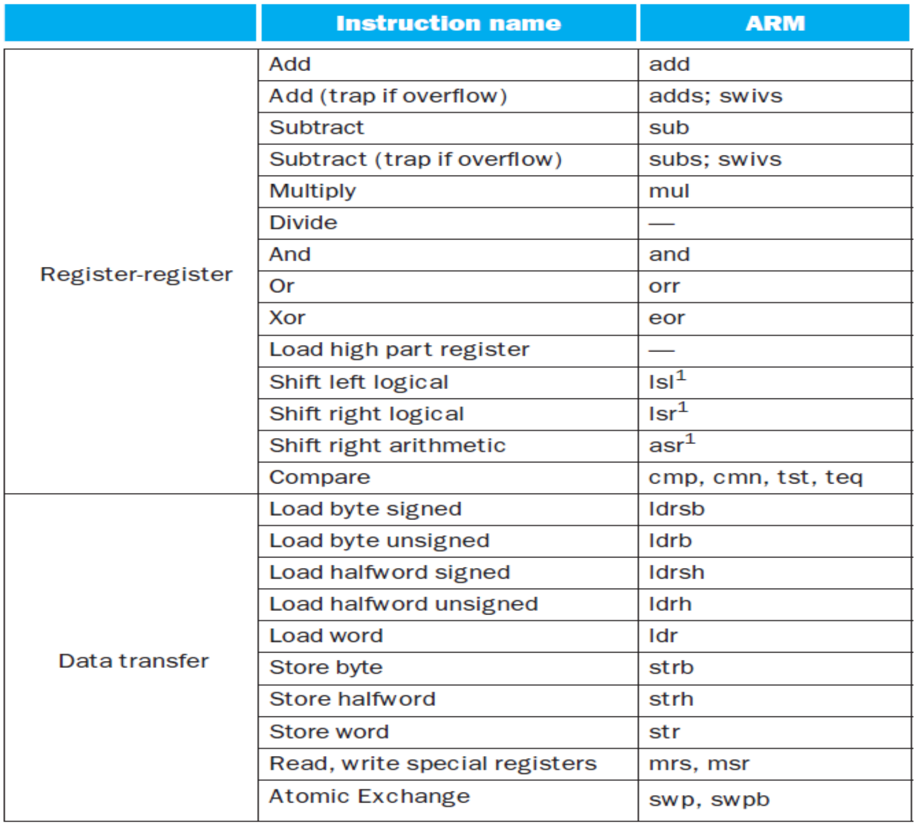
According to an article “Architecture and Implementation of the ARM Cortex-A8 Microprocessor”, the differences between Cortex-A8 and other ARM processors are the update and implementation of new technologies. First of all, Thumb-2 technology is the unifying technology of Cortex processors. To improve code density and performance, thumb-2 instruction set combines 16 bit and 32 bit instruction. The initial ARM instruction set contains 32 bit instructions only and the processors had to switch between ARM and Thumb modes during writing. By adding approximate 130 additional instructions to Thumb, Thumb-2 technology removes the mode switching between thumb and ARM to service interrupts.

Second, as you can see from Figure 1 Cortex-A8 Chip Diagram, NEON™ Data Engine is part of Cortex A8 chip. It is used for media and signal processing. In practice, NEON has good processing performance at 3D graphics, video and audio. According to ARM Limited, NEON is a mixed 64/128-bit SIMD (Single Instruction Multiple Data) architecture. SIMD indicates the process for operating on multiple data items using the same instruction. The register file and execution pipeline of NEON are discrete from the main ARM integer pipeline. Integer and single precision floating point values can be both handled by NEON technology.

Finally, Cortex A8 upgrades Floating Point Unit from VFPv2 to VFPv3. Figure 1 also shows Floating Point Unit is another important part of Cortex A8 chip. Floating Point architecture provides hardware support for floating point operations. VFPv3 is compatible with VFPv2, but it includes some enhancements, such as doubling the number of double-precision registers to 32 and adding instructions to convert between scalar, float and double (Aneesh).

Instruction Set

According to the book *Computer Organization and Design*, professor Patterson and Hennessy emphasize that it is very important to speak computer’s language properly to command the computer’s hardware effectively. The words of the computer’s language are defined as instructions. The vocabulary of commands is called an instruction set. Cortex-A8 implements 32-bit architecture known as ARMv7. ARMv7 is considered as one of the most popular instruction set architectures today. In 2011, more than 9 billion devices are using ARM. At the very beginning, ARM was abbreviation form of Acorn RISC Machine, but later changed to Advanced RISC Machine. The Table 1 on next page summarized by professor Patterson and Hennessy shows the core instruction sets for arithmetic-logical and data transfer instructions for ARM.



	Instruction name	ARM
Register-register	Add	add
	Add (trap if overflow)	adds; swivs
	Subtract	sub
	Subtract (trap if overflow)	subs; swivs
	Multiply	mul
	Divide	—
	And	and
	Or	orr
	Xor	eor
	Load high part register	—
	Shift left logical	lsl[1]
	Shift right logical	lsr[1]
	Shift right arithmetic	asr[1]
	Compare	cmp, cmn, tst, teq
Data transfer	Load byte signed	ldrsb
	Load byte unsigned	ldrb
	Load halfword signed	ldrsh
	Load halfword unsigned	ldrh
	Load word	ldr
	Store byte	strb
	Store halfword	strh
	Store word	str
	Read, write special registers	mrs, msr
	Atomic Exchange	swp, swpb

Table 1: ARM Core Instruction Set

As discussed earlier in this paper, because Cortex-A8 implements new technology Thumb-2, Cortex-A8 can carry out both 16 and 32-bit data types. The common syntax for ARM and Thumb instructions is Unified Assembly Language. According to the book *Professional Embedded ARM Development*, James Langbridge, an embedded systems consultant in computer architecture industry for more than 20 years, points out Unified Assembly Language supports generation of either Thumb or ARM instructions from the same source code. As an extension to Thumb, Thumb-2 not only add both 32 bit and 16 bit instructions but also add instruction for both DSP and floating-point calculations. These benefits allow developers to write programs for both Thumb and ARM modes. However, as you can see from Table 1 ARM Core Instruction set, ARM does not support all arithmetic operations. For example, there is no divide instruction in ARM. The data operation instructions contain shifts. The superscript 1, such as asr[1], indicates it is a variation of a move instruction (Patterson & Hennessy, 2014).

Data Path

The Cortex-A8 has been added new microarchitecture features to achieve the high levels of performances. Some of the features are not seen in other ARM architecture, such as dual-issue superscalar design. In the book *Computer Organization and Design*, professor Patterson and Hennessy points out the ARM Corxtex-A8 runs at 1GHz with a 14-stage pipeline. The significant feature of Corxtex-A8 is the dynamic multiple issue with 2 instructions per clock cycle. As a static in-order pipeline, instructions issue, execute, and commit in order. Previous

ARM processors only have a single integer execution pipeline. Issuing two data processing instructions obviously increases the number of instructions executed per cycle. As shown in Figure 2, Cortxtex-A8 pipeline contains three sections: instruction fetch, instruction decode and instruction execute.

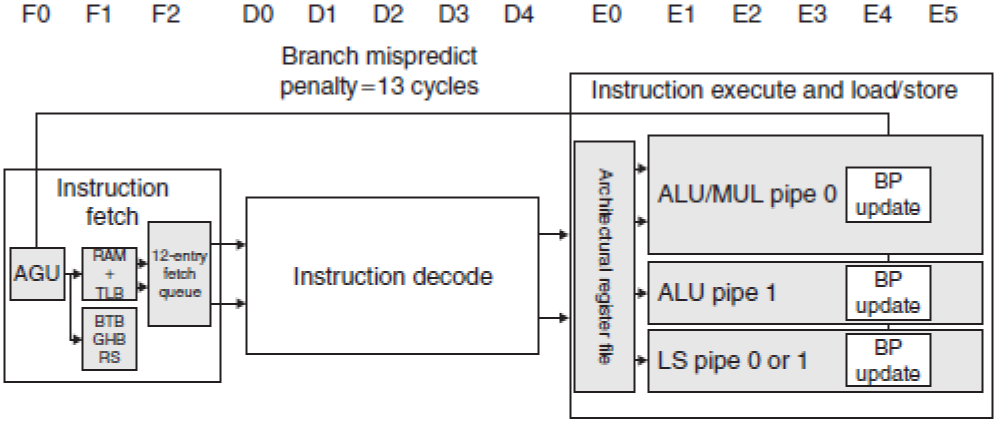


Figure 2: Cortex-A8 Pipeline

At the instruction fetch section, two instructions are fetched at a time. Fetched instructions in the first three stages are sent to a 12-entry instruction fetch buffer. Branch Target Buffer, Global History Buffer and a Return Stack are used to predict branches to keep the fetch queue full. At the instruction decode section, the five stages of the decode pipeline determine whether there are dependences among instructions. It affects the sequential execution. At the instruction execution section, there are one pipeline for load and store instructions and two pipelines for arithmetic operations. The arithmetic logic units (ALU 0 and ALU 1) are symmetric. The load-store pipeline can be coupled with instructions in either ALU0 or ALU 1, but the multiplier pipeline can only be coupled with instructions in ALU 0 (Williamson).

There are situations that prevent starting the next instruction in the next cycle. These situations are called hazards. Professor Patterson and Hennessy define three types of hazards in another book *Computer Architecture: A Quantitative Approach*:

- **Functional hazards**: functional hazards occur if two instructions selected for issue concurrently use the same functional pipeline.
- **Data hazards**: data hazards indicate an instruction depends on completion of data access by a previous instruction.
- **Control hazards**: control hazards indicate deciding on control action depends on previous instruction. It arises if branches are mispredicted.

The Cortex-A8 has an ideal CPI of 0.5 due to its dual-issue structure.

According to a published paper by consulting engineering David Williamson at ARM, besides the 14 stage integer pipeline disscuessed earlier, Cortex-A8 also uses a 10 stage NEON pipeline for accelerating multimedia and signal processing appliations. These 10 stage pipeline start at the end of ARM integer pipeline. All insturctions in the NEON media engine must be completed and it cannot throw exceptions. This is because the mispredicts and exceptions have been resolved in the ARM iteger unit already. As a result, the complexity of the NEON unit is reduced and it allows for a zero cycle load-use penalty in most cases. Figure 3 below includes Cortex-A8 all stages in the full pipeline by adding NEON's own 10 stage pipeline. Figure 3 shows there are 4 decode stages, M0-M3, and 6 execute stages, N1-N6, in NEON.

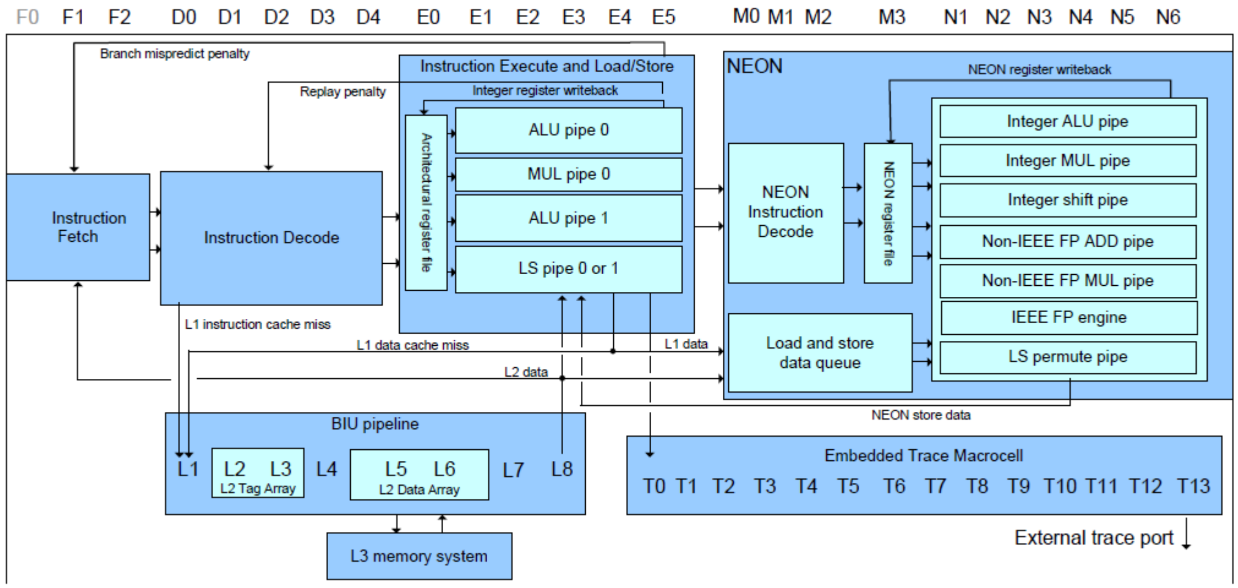
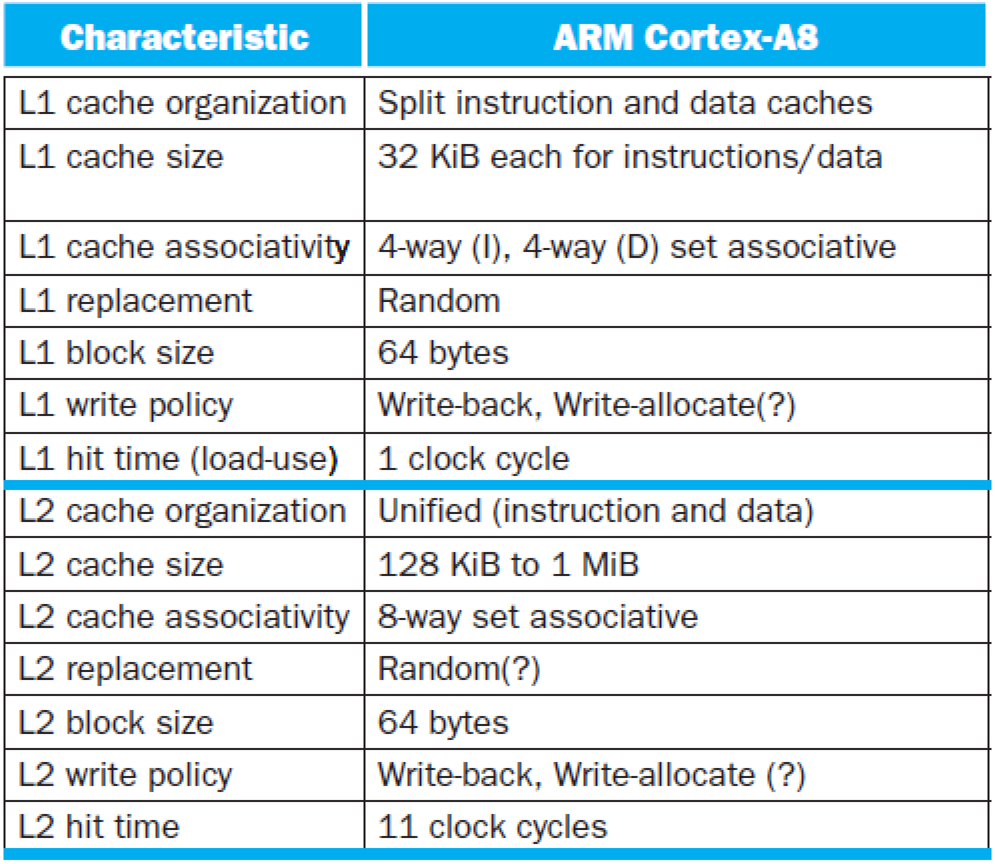


Figure 3: Cortex A8 Full Pipeline

Memory

In the published paper "ARM Cortex A8: A High Performance Processor for Low Power Applications," consulting engineering David Williamson at ARM points out Cortex A8 contains the integer load/store pipeline, the L1 cache, the L2 cache and the bus interface unit. Memory system handles all data memory transfers, such as NEON and floating-point load/store operations. The Table 2 below shows the ARM Cortex-A8 memory hierarchies summarized by professor Patterson and Hennessy in the book *Computer Organization and Design*. The Cortex-A8 only has one single core. L1 cache has 32 KiB cache size, 64-byte block size, and 4-way set associative with random replacement. The replacement for the data cache is write-back, but there is no write allocates. For fast access, A8 processor has a single-cycle load-use penalty to the L1 cache. The Cortex-A8 is famous for low latency due to the integrated L2 cache. L2 cache provides high bandwidth interface to L1 cache. The high bandwidth improves the latency of L1 cache line fills and traffic on the main system bus



Characteristic	ARM Cortex-A8
L1 cache organization	Split instruction and data caches
L1 cache size	32 KiB each for instructions/data
L1 cache associativity	4-way (I), 4-way (D) set associative
L1 replacement	Random
L1 block size	64 bytes
L1 write policy	Write-back, Write-allocate(?)
L1 hit time (load-use)	1 clock cycle
L2 cache organization	Unified (instruction and data)
L2 cache size	128 KiB to 1 MiB
L2 cache associativity	8-way set associative
L2 replacement	Random(?)
L2 block size	64 bytes
L2 write policy	Write-back, Write-allocate (?)
L2 hit time	11 clock cycles

.

Table 2: Cortex-A8 Cache

The ability to process two instructions at the same time significantly increases a processor's performance. How to execute more than one memory instruction per clock cycle to support processors like A8 is an important question for designer to think about. Professor Patterson and Hennessy indicate one technique is to partition the cache into multiple banks to enable parallel operations. This technique is similar to interleaved dynamic RAM banks.

Why It Fits

Amazon Echo is intended to use consumers' voice to instantly connect to Alex to control smart home devices, play music and get information, such as news and weather. At the first sight, Amazon Echo is just a speaker. Actually, it is an intelligent computer which collects consumers' audio data, transform audio to text and forwards to Amazon's Web Service(AWS). Once AWS has figured out answers, the result will be sent back to the device as a message. This process requires Amazon Echo provides high performance. In addition, Amazon wants to keep Echo at low-cost and low power consumption such that more consumers are able to afford them. Everything about the ARM Cortex-A8 is designed to fit these considerations.

First, as a family member of RISC, Cortex-A8 implements the processor design principle of simplified instructions. The simpler nature of the RISC instructions means that the control unit that decodes them is less complex and consumes less space on the CPU chip. The extra available space can be used to implement additional registers. Second, Cortex-A8 help reduce production costs and increase the profitability of Amazon. Assuming Amazon Echo is selling at $1,000 per unit, there may not be so many Echo in the market because consumers do not want to pay expensive price. According to Mouser Electronics, an online distributor of electronic components in Texas, a Cortex-A8 processor is selling at $5.9 per unit if the order requests approximate 1000 units. The price may be different based on various specifications and order quantity. Amazon most likely get lower price as Amazon orders more. The low cost of Cortex-A8 allow Amazon to invest peripherals to enhance the value of Echo and sell it at a profit. Finally, Cortex-A8 provides high performance at low power consumption by using the super-scalar architecture with in order instruction issue and support for forwarding path. As discussed earlier in this paper, the in-order instruction is less complex than out of order instruction. As a result, Cortex-A8 has relative low power consumption demand.

Security Concerns

Even though Amazon Echo is popular, consumers are concerned with security while enjoying the convenience and efficiency. According to a study conducted by National Chiao Tung University in Taiwan, some security vulnerabilities in Amazon Echo can threaten consumers' life. First, Amazon Echo uses weak single-factor authentication known as "wake word" for consumers to access the voice service. By default, it is "Alexa." "Wake word" works as a password and allows Amazon Echo to record what users say and send it the Amazon cloud for processing. Even though "wake word" can be configured, the options are very limited. Users may only choose one of three options "Computer", "Echo", and "Amazon." Second, Amazon Echo does not support voice authentication. In other words, Amazon Echo is unable to recognize owners' voice. Amazon Echo accepts anyone's voice commands no matter who you are. Finally, Amazon Echo does not require physical presence when using it. The study by National Chiao Tung University indicates Amazon Echo work well even if the users are not nearby as long as the sound pressure level is greater than 60 dB. This means users may control Amazon Echo if users can send voice to a speaker and the speaker is near to Amazon Echo.

The above three vulnerabilities can lead to home burglary and fake order. Nowadays many windows and doors are controlled by smart devices. Since Amazon Echo can control smart home devices, burglar may speak loudly and simply ask Echo to disarm the system and open the door. Burglar may also place order on Amazon.com because Echo is linked to Amazon account. Users can suffer financial loss from both cases.

Conclusion

As one of the fastest and power-efficient microprocessors developed by ARM, Cortex-A8 is ideal for Amazon Echo. To achieve best performance. Cortex-A8 incorporates many significant new features. These new features involve the NEON single instruction multiple data (SIMD) unit for media processing, Thumb-2 instruction set for reduced code size, superscalar pipeline for full dual-issue capability and enhanced Floating-Point Unit for floating point conversion. Technology bring human convenience, but the convenience may come with security threats. A responsible company should listen market's feedback and meet consumers' needs. Amazon may need to improve Echo to enhance its security performance and provide proper instructions to protect users' rights.

Reference

Aneesh. R. Architecture and Implementation of the ARM Cortex-A8 Microprocessor. Design & Resue. Retrieved April 26, 2018.

Johnson, B. (2016, November 16). *How Amazon Echo Works.* Retrieved April 22, 2018 from Howstuffworks Web site: https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/amazon-echo.htm

Hegde, P. and Hoffman, K.R.(2009). *ARM Cortex-A8 vs. Intel Atom: Architectural and Benchmark Comparisons.* Retrieved April 26, 2018 from Semantic Scholar Web site: https://pdfs.semanticscholar.org/0303/99fd8a3f623ddd107f5e85500cc56f777576.pdf

Haack, W., Severance, M., Wallace, M. and Wohlwend, J. (2017). *Security Analysis of the Amazon Echo.* Retrieved April 23, 2018, from Massachusetts Institute of Technology Web site: https://courses.csail.mit.edu/6.857/2017/project/8.pdf

Langbridge, J.A. (2014). *Professional Embedded ARM Development*. Indianapolis, Indiana: John Wiley&Sons,Inc.

Lei, X., Tu, G., Liu, A.X., Li,C.Y., and Xie, T. (2017, December 9). *The Insecurity of Home Digital Voice Assistants-Amazon Alexa as a Case Study*. Retrieved April 16, 2018, from arXiv Web site: https://arxiv.org/pdf/1712.03327.pdf

Patterson, D. A. and Hennessy, J. L.(2014). *Computer Organization and Design: The Hardware/software Interface*. 5th Edition. Waltham, MA: Morgan Kaufmann.

Patterson, D. A. and Hennessy, J. L.(2011). *Computer Architecture: A Quantitative Approach*. 5th Edition. Waltham, MA: Morgan Kaufmann.

Williamson,D. *ARM Cortex A8: A High Performance Processor for Low Power Applications*. Retrieved April 26,2018, from ARM Web site: http://www.arm.com/files/pdf/A8_Paper.pdf